



# Driver drowsiness recognition via transferred deep 3D convolutional network and state probability vector

Lei Zhao<sup>1</sup>  · Zengcai Wang<sup>2</sup> · Guoxin Zhang<sup>3</sup> · Huanbing Gao<sup>4</sup>

Received: 3 August 2019 / Revised: 2 May 2020 / Accepted: 24 June 2020 /  
Published online: 17 July 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

Driver drowsiness is a major cause of road accidents. In this study, a novel approach that detects human drowsiness is proposed and investigated. First, driver face and facial landmarks are detected to extract facial region from each frame in a video. Then, a residual-based deep 3D convolution neural network (CNN) that learned from an irrelevant dataset is constructed to classify driver facial image sequences with a certain number of frames for obtaining its drowsiness output probability value. After that, a certain number of output probability values is concatenated to obtain the state probability vector of a video. Finally, a recurrent neural network is adopted to classify constructed probability vector and obtain the recognition result of driver drowsiness. The proposed method is tested and investigated using a public drowsy driver dataset. Experimental results demonstrate that similar to 2D CNN, 3D CNN can learn spatiotemporal features from irrelevant dataset to improve its performance obviously in driver drowsiness classification. Furthermore, the proposed method performs stably and robustly, and it can achieve an average accuracy of 88.6%.

**Keywords** Driver drowsiness detection · 3D convolution neural network · State probability vector · Transfer learning · Residual learning

## 1 Introduction

Driver fatigue is a major cause of accidents. Compared with drunk driving, speeding, and other risky driving behaviors, driver drowsiness is difficult to detect and prevent, thereby

---

✉ Lei Zhao  
leizhao1219@sdjzu.edu.cn

Zengcai Wang  
wangzc@sdu.edu.cn

Guoxin Zhang  
zhanggx@polyu.edu.hk

Huanbing Gao  
gaohuanbing2004@sdjzu.edu.cn

Extended author information available on the last page of the article

compromising the security and safety of drivers, passengers, and pedestrians. In recent years, with the development of artificial intelligence technology, autonomous driving technology has progressed rapidly. However, road vehicles still need drivers. Furthermore, with conditional autonomous driving systems, driver should still be ready to take over if these systems do not operate correctly [1]. Therefore, research on effective methods for detecting driver drowsiness is important to improve transport safety. Detection methods can be divided into three categories according to basis: driver physiological parameters, operational behavior, and physical conditions [2].

Physiological signals change evidently when drivers feel drowsy. In recent years, many physiological signal detection methods for recognizing driver drowsiness have been studied [3]. Physiological signal tests include electroencephalogram, electrocardiogram, electrooculogram, and electromyogram [4]. Physiology-based methods have higher recognition accuracy compared with other methods [3]. However, most physiological acquisition sensors need to be attached to the driver's head or skin [3].

When drivers become drowsy, his or her operational behaviors change [5]. Therefore, the operational behaviors of drivers have been employed recently to detect driver drowsiness [3]. Various driver operational behaviors, such as steering wheel movement, speed, acceleration, and braking, are used to recognize driver drowsiness. However, different driving experiences, vehicle types and road conditions often affect driver operational behavior, thereby affecting the performance of these methods.

Facial behavior can reflect the changes in a driver's mental state, such as alert and drowsy. Variations in facial information, such as eyelid movement, eye openness, and facial expression, can reflect driver drowsiness [3]. Physical condition-based recognition methods adopted driver facial images to detect human facial fatigue [6]. Compared with the first two types of methods, this method is nonintrusive and not influenced by external interference. In the present study, we adopt physical condition-based methods to detect driver drowsiness.

Deep learning methods have greatly promoted the development of computer vision [7, 8]. Meanwhile, many deep learning models are used to solve the problems of driver drowsiness recognition [2, 9]. Many deep learning-based systems use convolutional neural networks (CNNs) to detect driver drowsiness because of the advantages of image recognition [7]. Some methods have used the CNN model to classify single driver facial image, whereas some methods have used the recognition outputs of static image to detect driver drowsiness in videos [9–12]. However, a static image only represents the instantaneous movement of the face because human face behavior is dynamic. Thus, the driver's real mental state is difficult to express.

In this study, we propose a novel method based on transferred deep 3D CNN and state probability vector. First, facial landmarks are detected to extract the facial region of drivers from each frame in the video. Then, a residual-based deep 3D CNN model is used to classify driver facial image sequences for obtaining its drowsiness output probability in a relatively short period of time. The difficulty in training deep 3D CNN is addressed by pretraining the constructed 3D CNN model in a human action dataset and finetuning it in a driver drowsiness dataset for recognition performance improvement. The facial movements of a person under drowsiness are slower, and they change less frequently compared with the facial movements of an alert person. Thus, a certain number of output probability values is concatenated to obtain the state probability vector, and a recurrent neural network model is adopted to classify the constructed vector and obtain the recognition result of driver drowsiness. The probability

vector not only discriminates driver drowsiness and alert states but also filters out some recognition errors produced by the 3D CNN.

The contributions of this study come in four aspects: a) a driver drowsiness recognition system with transferred deep 3D CNN and state probability vector; b) an analysis of the effect of different 3D CNN models for driver facial video classification; c) a transfer learning strategy based on uncorrelated data sets to improve the performance of the 3D CNN for driver drowsiness recognition; and d) a method of cascading the codified probability values of facial dynamic drowsiness status in a video to construct a state probability vector for representing driver facial drowsiness behavior.

The remainder of this paper is structured as follows: In Section 2 we introduce several representative related studies. In Section 3, we describe the overall architecture of the proposed method. In Section 4, we present the experiments conducted to evaluate the performance of the proposed method. In Section 5, we provide the conclusions and future research directions.

## 2 Related works

In recent years, image-based technologies have considerably developed because of the advances in related research subareas, such as face detection, tracking, and facial expression recognition, as well as in the area of machine learning, such as supervised learning, feature extraction, and deep learning. These systems can be divided into three major categories: eye-, mouth-, and fusion-based methods.

### 2.1 Eye-based methods

Eye-based driver drowsiness detection methods have been widely used. Mandal [13] proposed an eye-based drowsiness detection system for bus driver monitoring. This system detects eyes using two different methods and estimates the continuous level of the two eyes' openness by using spectral regression. Then, the level of openness of two the eyes are fused using adaptive integration to extract the percentage of eyelid closure for estimating the drowsiness of a bus driver. Cyganek [14] proposed a method for detecting the states of driver drowsiness by using two cameras. In the system, cascading two recognition models are used to locate the eye regions; one model is used to detect the eye location, whereas the other model is used to verify the eye coordinate. You [15] used a near-infrared camera to capture a driver image at night and then a spline function to fit the eyelid curve for driver drowsiness evaluation. Ibrahim [16] used Haar feature classifiers to detect the driver's face and adopted a correlation matching algorithm to locate and track the driver's eyes according to the shape, intensity, and size of the pupils. Song [17] proposed multiscale histograms of principal-oriented gradients to extract the texture information of eye regions for recognizing whether the driver eyes are closed. Gou [18] proposed a cascade regression framework to detect driver eye and estimated eye state. At each iteration of cascaded regression, the image pixel value of the eye center, as well as the image contextual features from the eye corners and eyelids are cascaded to locate eye regions and openness probability. Zhao [19] fused deep neural network and deep CNN to construct a deep integrated neural network (DINN) for recognizing driver eye state. The performance of the DINN is improved by using a transfer learning strategy based on an irrelevant dataset to pretrain the integrated model. The recognition rate of the DINN can achieve 97% in the driving environment experiment.

## 2.2 Mouth-based methods

Yawning often occurs when drivers feel drowsy. In recent years, many researchers have studied driver fatigue detection by extracting and analyzing driver mouth state. Omidyeganeh [20] used a modified implementation of the Viola-Jones detector for face and mouth detections and then adopted a back projection theory for extracting the changes in the driver mouth state to recognize yawning. Zhang [21] used deep learning and Kalman filter-based algorithms to detect and track driver face and nose and then adopted a neural network to detect yawning by using nose tracking confidence value, face motion features, and gradient features around the corners of the mouth. Zhang [22] used CNNs to extract spatial image mouth states and long short-term memory (LSTM) network to classify driver yawning behavior in a video. Akrouf [23] used active contours algorithm to fit driver lip and then calculated the lip area to represent the state of the mouth to judge whether the driver yawned or not.

## 2.3 Fusion-based method

The two detection methods above are widely used to detect human drowsiness because eye or mouth behavior changes considerably when drivers are fatigued. However, some crucial cues of the other organs are often overlooked by relying only on eyes or mouth, thereby affecting the recognition accuracy [24]. The study found that detecting drowsiness by fusing the behaviors of the eyes, mouth, and other regions of the face can considerably improve recognition accuracy [2]. Weng [9] proposed a hierarchical temporal deep belief network (DBN) to extract high-level features and output motion probabilities of driver face in each frame and then adopted two continuous hidden Markov models to classify driver drowsiness state. The experiment on a self-built dataset showed that the recognition accuracy of the proposed model could achieve 84.82%. Guo [10] proposed a hybrid VGG-based CNN model to extract driver eye and mouth state from each video frame and then used a time skip combination LSTM model on top of the CNNs to recognize driver states in the video. Zhao [2] fused the landmarks and image paths of driver eyes and mouth and placed them into a DBN model to classify three driver drowsiness expressions (alert, moderate drowsiness, and severe drowsiness). The experimental results in an actual driving environment showed that the recognition rate of the proposed model achieved 96.7%. Park [11] input driver face image and its optical-flow image into three CNN model (AlexNet, VGG-FaceNet, and FlowNet) and then concentrated and fed the outputs of the three models into SoftMax classifier for drowsiness classification. Shih [12] proposed a spatial-temporal network to detect driver drowsiness. First, a spatial CNN model was used to extract drowsiness-related features from each facial image in a video. Then, an LSTM model was used to construct the temporal variation of drowsiness status. Finally, a temporal smoothing method was adopted to smooth the predicted driver state scores. The proposed model achieved 82.61% average accuracy on the NTHU dataset [9]. Yu [25] used 3D CNN to learn scene understanding knowledge and extract spatiotemporal representations (eye, mouth, and head pose) from a video; then, the author fused and placed the two features into deep neural network to recognize driver drowsiness state. Although our method also used 3D CNN to detect driver drowsiness, our method is different from the method proposed by Yu [25]. We used 3D CNN to detect driver drowsiness probability in a short time video. Then, we integrated the probability value to model the state probability vector of long time videos. Finally, we fed the state probability vector into a recurrent neural network to classify driver drowsiness. In

addition, deep 3D CNN pretrained in other irrelevant dataset was used to improve the performance of the model.

### 3 Method

The proposed system consists of driver face detection and facial landmark detection, 3D CNN development, transfer learning, and state probability vector construction. The exhaustive illustration of the proposed method is presented in Fig. 1.

#### 3.1 Face detection and facial landmark detection

When a video is inputted into a driver drowsiness recognition system, the facial regions should be detected initially as a preprocessing step. In this study, the S<sup>3</sup>FD face detector is adopted to detect the facial region, and a ResNet-based facial alignment model is applied to locate and track driver facial landmarks. Driver facial regions are extracted as follows: a) extracting driver facial regions by using S<sup>3</sup>FD face detector roughly; b) detecting and tracking facial landmarks with a facial alignment model; and c) obtaining driver facial regions accurately according to the coordinates of facial landmarks.

#### 3.2 Spatiotemporal drowsiness recognition based on 3D CNN

3D CNN [26] is an extension of the CNN [27] model. The two models are composed of (3D or 2D) convolutional layers, (3D or 2D) pooling layers, and fully connected layers. The mathematical representation of 2D convolution operation is shown as follows:

$$v_{ij}^{xy} = f \left( b_{ij} + \sum_{k=0}^{M-1} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijk}^{pq} v_{(i-1)k}^{(x+p)(y+q)} \right) \tag{1}$$

where  $v_{ij}^{xy}$  is the value of the unit at location  $(x, y)$  in the  $j$ th feature map in the  $i$ th layer;  $f$  is the nonlinear function, such as tanh, sigmoid, or rectified linear unit function;  $b_{ij}$  is the bias for the feature map;  $M$  is the number of feature maps in the previous layer connected to the current feature map;  $w_{ijk}^{pq}$  is the value at  $(p, q)$  of the convolutional kernel connected to the previous

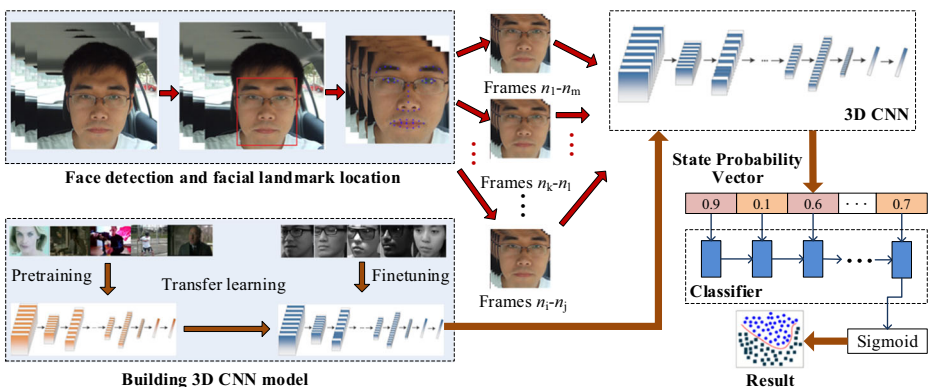


Fig. 1 Overall structure of the proposed method

layer’s  $k$ th feature map;  $P_i$  and  $Q_i$  are the height and width of the convolutional kernel, respectively. The mathematical representation of the 3D convolution operation is shown as follows:

$$v_{ij}^{xyz} = f \left( b_{ij} + \sum_{k=0}^{M-1} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijk}^{pqr} v_{(i-1)k}^{(x+p)(y+q)(z+r)} \right) \tag{2}$$

where  $R_i$  is the temporal dimension of the convolutional kernel, and  $w_{ijk}^{pqr}$  is the value at  $(p, q, z)$  of the kernel connected to the previous layer’s  $k$ th feature map. Pooling layers are used to reduce the size of the feature map. The extension strategy from 2D pooling to 3D is the same as that of convolutional layers. The full connection layers of 2D and 3D CNN are the same.

In this study, residual-based architectures [28] are adopted to construct 3D CNN. As shown in Fig. 2, residual learning block provides shortcut connections that allow a feature signal to bypass one weight layer and move to the next weight layer in the network. The mathematical representation of residual learning is shown as follows:

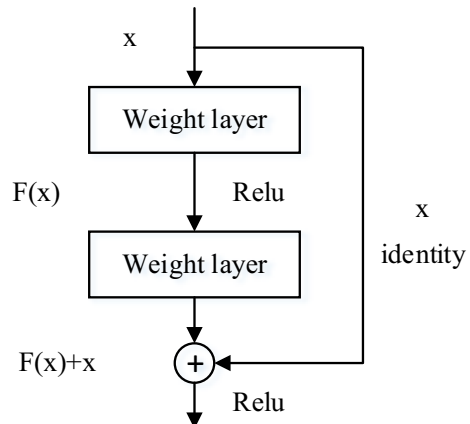
$$F(x) = H(x) + x \tag{3}$$

where,  $H(x)$  is the desired output mapping,  $x$  is the input signal of a residual unit, and  $F(x)$  is the stacked nonlinear layers. The original mapping is recast into  $F(x) + x$ . ResNet [28], which consists of residual learning units, is a successful architecture in image and human action recognition because of its outstanding performance in large-scale image and video processing. Furthermore, some derivative models based on residual-based architectures, such as DenseNet [29] and ResNeXt [30], have also been proposed successively.

Three kinds of extended 3D CNN models based on ResNet, DenseNet, and ResNeXt architectures are used to detect driver drowsiness from face image sequences. The residual learning block of these models are shown in Fig. 3 [31]. In Fig. 3, conv,  $a^3$ , and  $N$  are the parameters of convolutional kernel;  $a \times a \times a$  represents the kernel size of the convolutional filter;  $N$  and *group* are the number of groups of the convolutions’ group, which divides the feature maps into groups; BN refers to batch normalization; ReLU is the activation function.

To conduct experiments, we use seven 3D CNN models, namely named ResNet3D-18, ResNet3D-34, ResNet3D-50, ResNet3D-101, ResNet3D-152, DenseNet3D-121, and ResNeXt3D-101, to evaluate the proposed system [31]. The number of the models’ name

Fig. 2 Residual learning block architecture



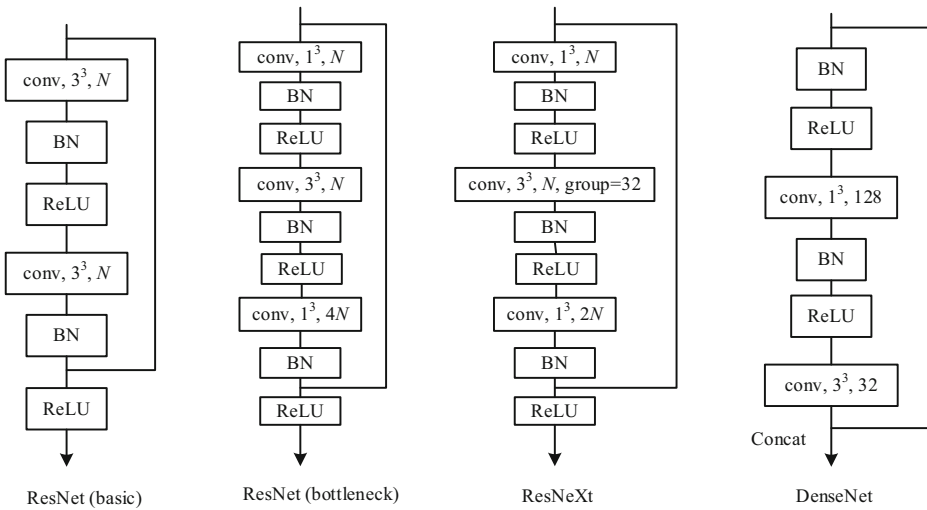


Fig. 3 Block of each architecture [31]

denote the layer number of networks. The architectures of these networks are shown in Table 1.

In the table, block name denotes the block architectures presented in Fig. 3. For each block,  $N$  is the number of the convolutional feature map, and  $M$  is the number of the blocks. In output layers, spatiotemporal down-sampling is performed by average pool, except for DenseNet3D-121. FC-TRANS represents that the convolution layer is transformed into a full connection layer. As shown in the table, ResNet3D-18 and 34 consist of the basic block of ResNet; meanwhile, ResNet3D-50, 101, and 152 consist of the bottleneck block of ResNet.

The output layers of these models are shown as follows:

$$p(A) = \sigma \left( \sum_{k=0}^K A_k W_{k,i} \right) \tag{4}$$

where  $A$  is the top hidden layer of 3DCNN;  $p(A)$  is the probability for driver drowsiness given  $A$ ;  $\sigma$  is the output classifier;  $W = [W_{1i}, W_{2i}, \dots, W_{Ki}]$  is the parameter set of the output layers.

Table 1 Network architectures

Layers	ResNet3D-{18, 34}	ResNet3D-{50, 101, 152}	DenseNet3D-121	ResNeXt3D-101
Block name	Basic	Bottleneck	DenseNet	ResNeXt
Block_0	conv, $7 \times 7 \times 7$ , $N$ : 64, stride: $2 \times 2 \times 1$ . BN, ReLU, maxpool: $3 \times 3 \times 3$ .			
Block_1	$N$ 64	64	64	128
Block_2	$M$ {2, 3}	3	{6, 6}	3
Block_2	$N$ 128	128	128	256
Block_2	$M$ {2, 4}	{4, 4, 8}	{12, 12}	24
Block_2	$N$ 256	256	256	512
Block_3	$M$ {2, 6}	{6, 23, 36}	{24, 48}	36
Block_3	$N$ 512	512	{512, 896}	1024
Block_3	$M$ {2,3}	3	{16, 32}	3
output	average pool: $4 \times 4 \times 1$ , FC-TRANS, sigmoid function			

### 3.3 Transfer learning

Acquiring a sufficient number of datasets that meet the training requirement is difficult to achieve. Thus, training our 3D CNN model, whose parameters are randomly initialized, is unrealistic. The lack of training data can render our network prone to overfitting. In recent years, transfer learning methods have been used to solve this problem and improve the recognition capability of CNNs [32, 33]. In transfer learning, a base CNN is pretrained on a dataset, which has a sufficient number of samples by using supervised learning. Then, the parameters of the learned base CNN are repurposed or transferred to a second model to be trained on a target dataset and task. This method can improve recognition accuracy if the transferred features are meaningful, general, and suitable for the tasks. Furthermore, transfer learning among unrelated data sets is still effective. However, studies on transfer learning are focused on the 2D CNN model. Hara [31] successfully applied transfer learning to the 3D CNN model to improve the accuracy of human behavior recognition. However, the experiment is only implemented in a relevant dataset. The 3D CNN model are pretrained and finetuned in different human behavior recognition datasets. In this paper, the 3D model pretrained in human behavior recognition is adopted to be finetuned in the driver drowsiness dataset to analyze whether the performance of the model is improved. The pretraining dataset is Kinetics [34], a human behavior recognition dataset. Kinetics has 400 human behaviors, of which 400 image sequences are contained. The total number of the samples is not less than 300,000, and these samples have an adequate amount of data to pretrain the 3D CNN model.

### 3.4 State probability vector

In driver face and facial landmark detection, the failure of location often occurs. To solve these problems, we adopt the integration strategy of frame substitution [2]. As shown in Fig. 4, the frames where no face or facial landmarks is detected are discarded, and the other adjacent images are extended to the position of the discarded frames. Then, a time window is used to extract image sequences with fixed frames in a video. These image sequences are placed into the proposed 3D CNN model to obtain the state probability value. As shown in Fig. 4 and Eqs. 5–7, time windows can overlap over time. Finally, the state probability values of the video are integrated into a vector. A recurrent model is used to classify the state probability vector for obtaining the driver drowsiness state of the video. The mathematical representation of the formation of state probability vector is shown in Eq. 8.

$$v_1 = [f_1, f_2, \dots, f_{N_f}] \quad (5)$$

$$v_n = [f_{nS+1}, f_{nS+2}, \dots, f_{nS+N_f}] \quad (6)$$

$$S = \frac{N_F - N_f}{n_s - 1} \quad (7)$$

$$V = [P_1, P_2, \dots, P_i, \dots, P_{n_s}] \quad (8)$$



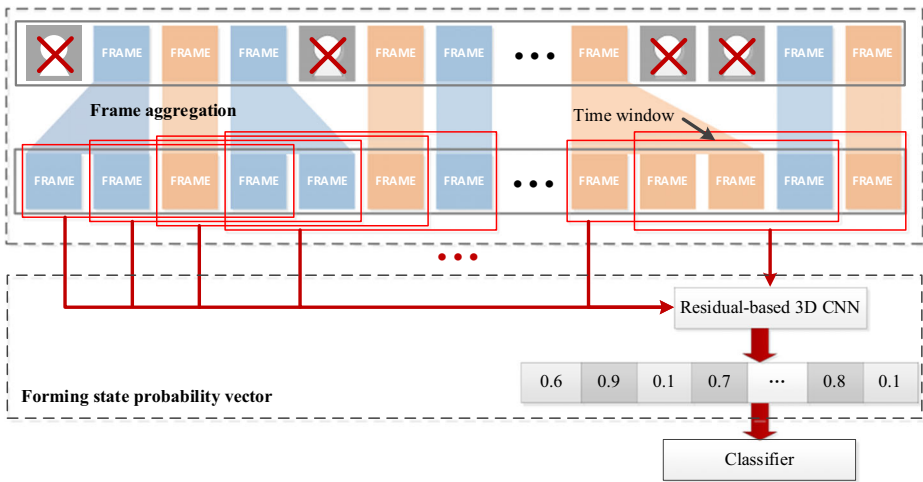


Fig. 4 State probability vector

where,  $v_n$  is the  $n$ th image sequence extracted from the driver facial video;  $f_i$  denotes the  $i$ th frame of an image sequence;  $S$  is the sliding step of time windows;  $N_F$  and  $N_f$  are the frame numbers of the video and the extracted image sequence, respectively;  $n_s$  represents the number image sequences extracted from a video;  $V$  is the state probability vector of the video;  $P_i$  is the state probability value of the  $i$ th image sequence in the video.

In a short period of time (several frames), driver drowsiness and alertness often produce similar facial behaviors, such as closing eyes and blinking, yawning, and talking. In actual driving, when the driver is alert, the speed of facial muscle movement is often faster and more frequent than when the driver is drowsy. If features are extracted in a relatively short time for drowsiness detection, then a high false detection rate will be obtained. Therefore, a long period of time is required to obtain sufficient driver facial movement information. In this study, we extract the drowsiness probability vector of drivers facial videos with 240 frames.

Finally, LSTM and gated recurrent unit (GRU) network are applied to classify the vector and analysis effectiveness of the drowsiness probability vector of drivers. LSTM and GRU are recurrent-based neural networks that have been widely used in speech recognition, natural language processing, and video recognition. Each value of the vector is placed into the corresponding input nodes of the model in chronological order. The output classifier is added to the top of the last hidden layer of the model to obtain the recognition result. The number layers of the two models are set to 2, the size of hidden layers is 10, and the output function is logistic regression.

## 4 Experimental results

In this section, we elaborate the information of the driver drowsiness dataset, platform, and data preprocessing method in Sections 4.1 and 4.2. Then, we present the experimental results to prove the effectiveness of the proposed method. We present the results of the performance evaluation of the 3D CNN in Sections 4.3–4.5. We present the results of the

performance evaluations of the proposed state probability vector and the entire system in Sections 4.6–4.10. In Section 4.11, we compared other state-of-the-art methods with the proposed system.

#### 4.1 Dataset

To evaluate the effectiveness of our method and compare with other state-of-the-art methods, we use the NTHU driver drowsiness dataset to conduct experiments [9]. The dataset contains 22 subjects and 380 videos. In the dataset, all subjects with various facial ornaments (eyeglasses and sunglasses) are captured in various environments (day and night). The subjects sit in front of simulated driving system and perform driving operations. In the meantime, each participant acts several actions that usually happen during actual driving, such as talking, blinking, yawning, and drowsiness. The video is captured by D-Link DCS-932 L with infrared LED at 30 frames per second. The size of each image is 640×480. The dataset is divided into training and validation sets. In this experiment, the validation set is adopted to test our model. Sample images of the dataset are presented in Fig. 5. The top row of Fig. 5 shows the sample frames of the alert state of the subjects, whereas the bottom row of Fig. 5 shows the sample frames of the drowsiness state of the participant. For the training and evaluation of the 3D CNN model, certain image sequences with 16 frame numbers are randomly extracted from the video-based dataset to construct the training and testing datasets. The frame rate of these image sequence datasets is 15 frames per second. The labels of the dataset are obtained in accordance with the original labels of the images in the NTHU dataset.

#### 4.2 Platform and data preprocessing

The proposed method is trained and tested using Pytorch 0.4 and OpenCV 2.4.8 on two Intel Xeon E5–2630 v4 CPU at 2.50 GHz with 256 GB RAM and 2 NVIDIA Tesla K80 graphic cards with 24 GB RAM running on Windows 10.

Normalization process is essential for training and testing. Initially, the mean values of the image sequences and vectors are obtained. Then, the difference between the mean and real values is calculated to remove the individual difference for the input data. Finally, the standard deviation is divided to render the data normally distributed. The values of all data are normalized as follows:



Fig. 5 Sample images of the NTHU dataset

$$\mu = \frac{\sum_{j=1}^J p_j}{J} \tag{9}$$

$$\sigma = \sqrt{\sum_{j=1}^J (p_j - \mu)^2} \tag{10}$$

$$P_j = \frac{p_j - \mu}{\sigma + C} \tag{11}$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation, respectively;  $p_j$  is the original input vector;  $C$  is a constant avoiding the numerator to be divided by zero;  $P_j$  is the normalized input data;  $J$  is the number of input vectors.

### 4.3 Evaluation of the different number layers of the ResNet3D model

In this experiment, the performance of the different ResNet3D number layers are compared to analyze the effect of layer’s number on recognition accuracy. The number layers of ResNet3D used in this experiment are 18, 34, 50, 101, and 152. The curves of recognition accuracies and the area under the receiver operating characteristic curve (AUC) are shown in Figs. 6 (a) and (b), respectively. On the basis of the experimental results, the recognition accuracy and AUC of the model are improved with the increase in the model layer’s number. In Fig.6 (a), the growth of AUC slows down with the increase in layers. Compared with AUC, the improvement of recognition accuracy is limited. In accordance with recognition results and model complexity, 3D CNN models with approximately 100 layers are adopted to detect driver drowsiness.

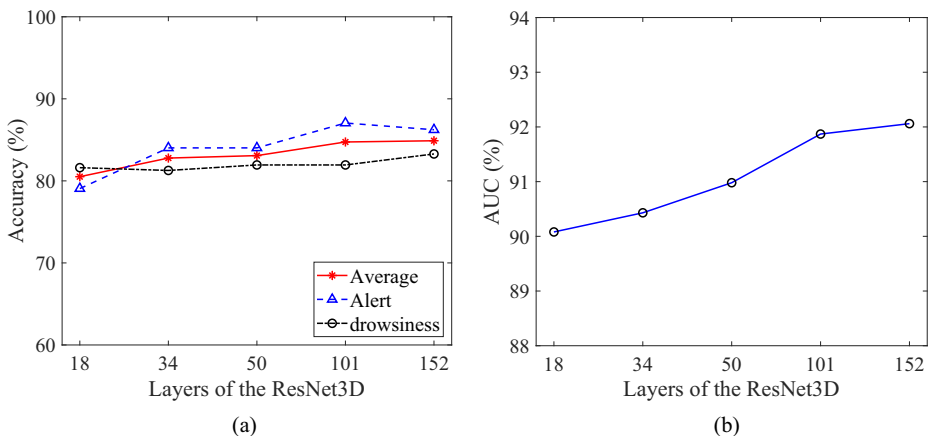


Fig. 6 Experimental results of ResNet3Ds with different numbers of layers

#### 4.4 Evaluation of different 3D CNN models

The performance of different 3D CNNs with about 100 layers (ResNet3D-101, DenseNet3D-121, and ResNeXt3D-101) are compared in this section. The recognition accuracies and AUC are shown in Table 2. The curves of the receiver operating characteristic (ROC) are shown in Fig. 7. As shown in Table 2, ResNext3D-101 achieves the best drowsiness recognition rate. In Table 2 and Fig. 7, the DenseNet3D-121 outperforms the others in terms of AUC. However, compared with the recognition accuracy, the advantages of AUC of the ResNext3D-101 and DenseNet3D-121 are not obvious. Furthermore, ResNeXt-based architecture has fewer parameters compared with DenseNet, thereby reducing the operation time of systems. Therefore, ResNeXt3D-101 is used to obtain the driver drowsiness probability value.

#### 4.5 Evaluation of transfer learning strategies for the 3D CNN models

In this section, 3D CNN models with and without a transfer learning strategy are compared. The comparison results of the two strategies are shown in Fig. 8. The experimental results indicate that when the 3D CNN models pretrained in the Kinetics human behavior dataset are used to perform classification, the accuracy of drowsiness recognition improves evidently. Same as the traditional CNN model, backpropagation algorithm is adopted to minimize the prediction error with gradient descent. If the weights and bias of the 3D CNN initialize randomly, then they may cause local optimization for the excessive number of parameters of deep 3D CNN model. Hara [31] successfully used transfer learning to improve the performance of 3D CNN in human behavior recognition detection. However, the pretraining and finetuning processes were all conducted in human behavior datasets. While these experimental results in Fig. 8 demonstrate that 3D CNN can learn spatiotemporal features from irrelevant datasets similar to 2D CNN in drowsiness detection.

#### 4.6 Evaluation of the proposed method with different numbers of frames of input video

The proposed method with different numbers of image frames of input video ( $F$ ) is compared. The experimental results are shown in Table 3. The 3D CNN model used in this experiment is ResNeXt3D-101, and the classifier of the probability vector is LSTM neural network. When image sequences are extracted, the number of overlapped frames is 6. The recognition rate and AUC of the system show a marked overall change with the increase in number of frames. The recognition rate and AUC of the model are highest when  $F$  is 240. We can infer that the proposed model will perform better if  $F$  continues to rise. However, the complexity of the model increases with  $F$ . Thus, the proposed model with 240 frames is adopted in subsequent experiments.

**Table 2** Experimental results of different 3D CNN models

Model	Alertness (%)	Drowsiness (%)	Total (%)	AUC (%)
ResNet3D-101	87.05	81.94	84.74%	91.87
DenseNet3D-121	82.09	88.63	85.04	92.44
ResNext3D-101	88.71	82.94	86.10	92.35

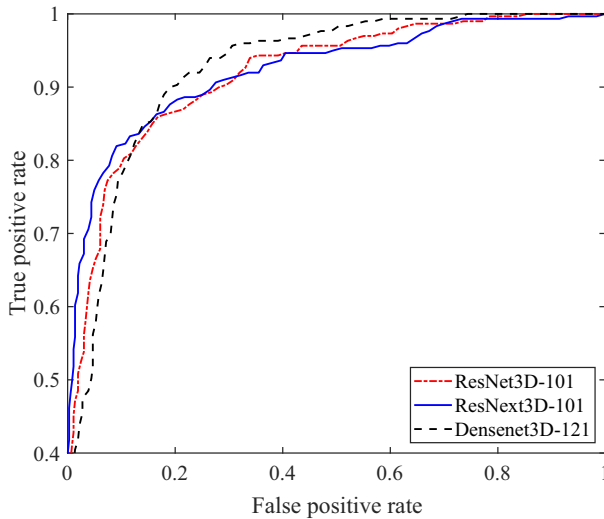


Fig. 7 ROC curves of different 3D CNN models

### 4.7 Evaluation of the proposed method with different numbers of probability vector

The proposed method with different numbers of probability vector ( $N_F$ ) is compared. The experimental results are shown in Table 4. The 3D CNN model and classifier of the probability vector are the same as Section 4.6. The recognition rate and AUC of the system show a limited overall change with the increase in  $N_F$ . When  $N_F$  is 40, the accuracy of driver drowsiness recognition is highest. When  $N_F$  is 30, the AUC of the proposed model is better than that of others. The improvement of recognition accuracy is more evident than that of AUC. Thus, the drowsiness probability vector with 40 probability values is adopted in subsequent experiments.

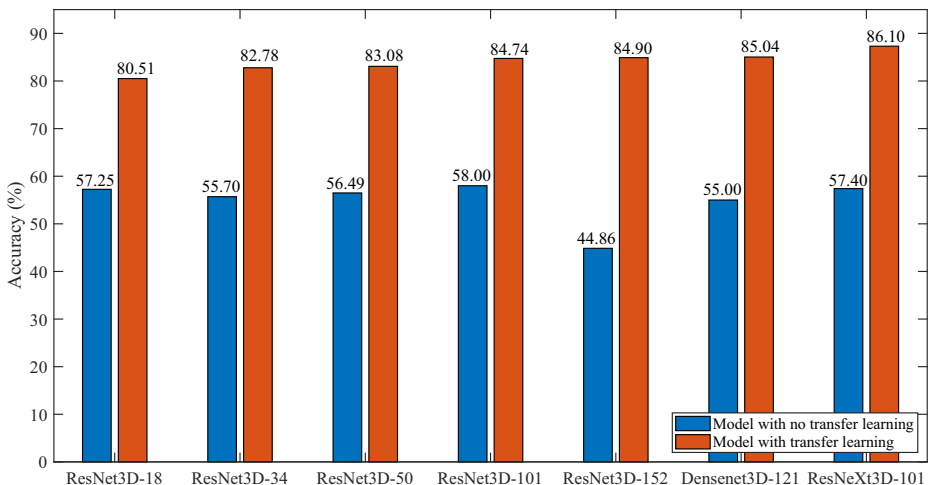


Fig. 8 Experimental results of 3D CNN with different training strategies

**Table 3** Experimental results of models with different number of frames

$F$	Alertness (%)	Drowsiness (%)	Total (%)	AUC (%)
64	79.40	83.90	81.80	89.55
96	84.50	86.52	85.80	91.79
128	83.70	87.27	85.60	92.63
160	82.83	88.01	85.60	93.36
192	87.56	86.52	87.00	94.22
240	87.12	89.89	88.60	95.25

#### 4.8 Evaluation of the proposed method with different classifiers

This set of experiments compares the performance of the proposed system with different classifiers, namely support vector machine (SVM), random forest (RF), decision tree (ID3),  $k$ -nearest neighbor ( $k$ -NN), LSTM, and GRU, which are widely applied in many machine learning fields. The 3D CNN model used in this experiment is ResNeXt3D-101, and the  $N_F$  of the state probability vector is 20. The experimental results are presented in Table 5 and Fig. 9. The classification results shown in Table 5 illustrate that the LSTM neural network outperforms other classifiers, whereas the recognition rate of SVM and RF are the worst. However, as shown in Fig. 9 and Table 5, the AUC and ROC of the model using the GRU neural network is better than those of others. The experimental results indicate that the performance of the recurrent-based neural networks (LSTM and GRU) are similar, and these networks outperform the other classifiers (SVM, RF, and  $k$ -NN).

#### 4.9 Evaluation of the proposed method with different accuracies of 3D CNN

The different accuracies of the 3D CNN in the proposed system are compared in this experiment. The comparison results are presented in Table 6. As shown in the table, model-based accuracy and AUC denote the recognition rate and AUC value of probability vector extraction models, respectively. Meanwhile, video-based accuracy and AUC denote the recognition rate and AUC value of LSTM neural networks with video probability vector as input respectively. The experimental results show that the video-based accuracy and AUC are improved with the increase in model-based accuracy and AUC. Furthermore, compared with model-based recognition results, video-based accuracy and AUC increased by 1.85% and 2.46%, respectively. The video-based

**Table 4** Experimental results of models with different numbers of probability vector

$N_F$	Alertness (%)	Drowsiness (%)	Total (%)	AUC (%)
16	82.40	90.63	86.80	95.61
20	84.12	89.89	87.20	95.39
24	84.12	90.26	87.40	95.48
30	82.40	91.01	87.00	95.72
40	87.12	89.89	88.60	95.25
60	83.26	91.39	87.60	95.56
120	83.26	91.01	87.40	95.30

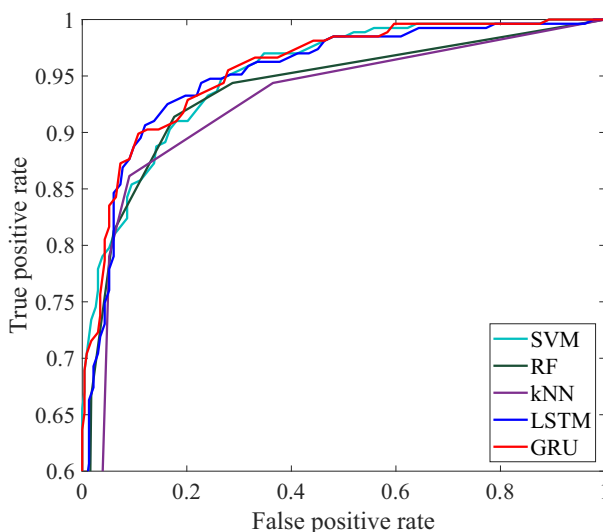
**Table 5** Experimental results of the proposed method with different classifiers

Model	Alertness (%)	Drowsiness (%)	Total (%)	AUC (%)
SVM	83.69	90.26	87.20	95.29
RF	82.40	<b>91.38</b>	87.20	93.83
<i>k</i> -NN	86.69	89.51	88.20	92.19
LSTM	<b>87.12</b>	89.89	<b>88.60</b>	95.26
GRU	86.27	90.26	88.40	<b>95.65</b>

performance of the ResNet3D-101-based system improves the most (3.46% and 3.19%). Meanwhile, the system with a few layer numbers of 3D CNN barely improves. The experimental results indicate that the recognition ability of the system will be improve remarkably when the output probability values of the 3D CNN models are concatenated into a drowsiness state vector, and the recognition accuracy and the AUC achieve 88.6% and 95.26%, respectively.

#### 4.10 Evaluation of methods with different state probability vector extraction models

CNN- and 3D CNN-based state probability vector extraction model are compared in this section. ResNeXt-101 (CNN) and ResNeXt3D-101 (3D CNN) are used to conduct the experiment. ResNeXt-101, which is pretrained in ImageNet, is finetuned in static images extracted from the NTHU dataset. LSTM is applied to classify the state probability vectors extracted using the two models. The comparison results are presented in Table 7. The experimental results show that the model-based accuracy and AUC of ResNeXt3D-101 are higher than those of ResNeXt-101. Furthermore, the accuracy and AUC of the ResNeXt-101-based method are less than 50% when static state probabilities of the driver are concatenated into the vector. We can conclude that dynamic features are more discriminative than static features for driver drowsiness classification.

**Fig. 9** ROC curves of the proposed method with different classifiers

**Table 6** Experimental results of the proposed method with different accuracies of 3D CNN

Model	Model-based accuracy (%)	Model-based AUC (%)	Video-based accuracy (%)	Video-based AUC (%)
ResNet3D-18	80.51	90.43	82.20 (1.69)	91.40 (0.97)
ResNet3D-34	82.78	90.08	83.40 (0.62)	92.65 (2.57)
ResNet3D-50	83.08	90.98	84.20 (1.12)	93.26 (2.28)
ResNet3D-101	84.74	91.87	88.20 (3.46)	95.06 (3.19)
DenseNet3D-121	85.04	92.44	86.80 (1.76)	95.26 (2.82)
ResNext3D-101	86.10	92.35	88.60 (2.50)	95.26 (2.91)

#### 4.11 Comparison of the proposed method and state-of-the-art methods

In this experiment, the performance of the proposed systems is compared with that of the state-of-the-art methods proposed by Park [11], Yu [25], Shih [12], Wen [9], and Guo [10] using the NTHU dataset. In the proposed system, ResNeXt3D-101 is used to predict the probability value of the image sequences with 16 frames for constructing the state probability vector of the video with 240 frames. Meanwhile, GRU and LSTM are adopted to classify the vector for obtaining the driver drowsiness state.  $N_F$  is 20. The other experiment results presented in this experiment are the best performance of the corresponding state-of-the-art methods. The comparison results in Table 8 indicate that the proposed methods outperform other state-of-the-art methods. The recognition accuracy of the system with ResNeXt3D-101 and LSTM achieves 88.6%.

## 5 Conclusion

In this study, a novel approach for detecting human drowsiness is proposed and investigated. First, the S<sup>3</sup>FD face detector and ResNet-based facial alignment model are adopted to detect the facial region, locate and track driver facial landmarks, respectively, and extract the driver's facial region. Then, a 3D residual-based CNN model that learned from a human action dataset is constructed to classify driver facial image sequences with a certain number of frames and obtain its drowsiness output probability value. Finally, a certain number of output probability values is concatenated to obtain the state probability vector of a video, and a recurrent neural network is adopted to classify the constructed probability vector and obtain the recognition result of driver drowsiness. The proposed method is tested and investigated on a public NTHU drowsy driver dataset. Experimental results demonstrate that 3D CNN can learn spatiotemporal features from irrelevant datasets to improve its performance obviously similar to 2D CNN in driver drowsiness classification. The recognition results also show that the AUC and recognition rate improve with the layer's number of 3D CNN; however, compared with that

**Table 7** Experimental results of different extraction models of state probability vector

Model	Model-based accuracy (%)	Model-based AUC (%)	Video-based accuracy (%)	Video-based AUC (%)
ResNeXt-101	77.07	86.38	53.80	47.56
ResNeXt3D-101	86.10	92.35	88.60	95.26



**Table 8** Comparison results of the proposed method and state-of-the-art methods

Research	Method	Accuracy(%)
Park [11]	DDD-IAA	73.06
Yu [25]	3D Future fusion	71.20
Shih [12]	MSTN	77.81
Wen [9]	Hierarchical Temporal DBN	84.82
Guo [10]	CNN + TSC-LSTM	84.85
Ours	ResNeXt3D-101 + GRU	88.40
	ResNeXt3D-101 + LSTM	<b>88.60</b>

of AUC, the improvement of recognition accuracy is limited. Therefore, that the proposed method outperforms other state-of-the-art systems and it can achieve an average accuracy of 88.6%.

However, large head rotations and occlusions can reduce the effectiveness of the proposed method. We will address these limitations in future studies by combining other useful information (e.g., gaze estimation and pupil movements). In addition, 3D CNN has more parameters than 2D CNN, thereby greatly increasing the running time of the system. Therefore, we will attempt to establish a simple, efficient, and accurate approximate model of the 3D CNN used in this study to reduce the computational complexity of our method.

**Acknowledgments** This work was supported by the Doctoral Foundation of Shandong Jianzhu University (China, Grant no. X18039Z), the Natural Science Foundation of Shandong Province (China, Grant no. ZR2018MEE015) and the Open Foundation of State Key Laboratory of Automotive Simulation and Control (China, Grant no. 20161105).

## Compliance with ethical standards

**Conflict of interest** None.

## References

1. Deo N, Trivedi MM (2018) Looking at the driver/rider in autonomous vehicles to predict take-over readiness. arXiv preprint arXiv:181106047
2. Zhao L, Wang Z, Wang X, Liu Q (2018) Driver drowsiness detection using facial dynamic fusion information and a DBN. *IET Intell Transp Syst* 12(2):127–133
3. Sikander G, Anwar S (2018) Driver fatigue detection systems: a review. *IEEE Trans Intell Transp Syst* 20(6):2339–2352
4. Mårtensson H, Keelan O, Ahlström C (2018) Driver sleepiness classification based on physiological data and driving performance from real road driving. *IEEE Trans Intell Transp Syst* 20(2):421–430
5. McDonald AD, Lee JD, Schwarz C, Brown TL (2018) A contextual and temporal algorithm for driver drowsiness detection. *Accid Anal Prev* 113:25–37
6. Ou C, Ouali C, Bedawi SM, Karray F Driver Behavior Monitoring Using Tools of Deep Learning and Fuzzy Inferencing. In: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2018. IEEE, pp 1–7
7. Lecun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444
8. Tong M, Chen Y, Zhao M, Bu H, Xi S (2019) A deep discriminative and robust nonnegative matrix factorization network method with soft label constraint. *Neural Comput & Applic* 31(11):7447–7475
9. Weng C-H, Lai Y-H, Lai S-H Driver drowsiness detection via a hierarchical temporal deep belief network. In: *Asian Conference on Computer Vision*, 2016. Springer, pp 117–133
10. Guo J-M, Markoni H (2018) Driver drowsiness detection using hybrid convolutional neural network and long short-term memory. *Multimed Tools Appl*:1–29

11. Park S, Pan F, Kang S, Yoo CD Driver drowsiness detection system based on feature representation learning using various deep networks. In: Asian Conference on Computer Vision, 2016. Springer, pp 154–164
12. Shih T-H, Hsu C-T MSTN: Multistage spatial-temporal network for driver drowsiness detection. In: Asian Conference on Computer Vision, 2016. Springer, pp 146–153
13. Mandal B, Li L, Wang GS, Lin J (2017) Towards detection of bus driver fatigue based on robust visual analysis of eye state. *IEEE Trans Intell Transp Syst* 18(3):545–557
14. Cyganek B, Gruszczyński S (2014) Hybrid computer vision system for drivers' eye recognition and fatigue monitoring. *Neurocomputing* 126:78–94
15. You F, Y-h L, Huang L, Chen K, R-h Z, Xu J-m (2017) Monitoring drivers' sleepy status at night based on machine vision. *Multimed Tools Appl* 76(13):14869–14886
16. Ibrahim LF, Abulkhair M (2014) Using Haar classifiers to detect driver fatigue and provide alerts. *Multimed Tools Appl* 71(3):1857–1877
17. Song F, Tan X, Liu X, Chen S (2014) Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. *Pattern Recogn* 47(9):2825–2838. <https://doi.org/10.1016/j.patcog.2014.03.024>
18. Gou C, Wu Y, Wang K, Wang K, Wang F-Y, Ji Q (2017) A joint cascaded framework for simultaneous eye detection and eye state estimation. *Pattern Recogn* 67(1):23–31
19. Zhao L, Wang Z, Zhang G, Qi Y, Wang X (2018) Eye state recognition based on deep integrated neural network and transfer learning. *Multimed Tools Appl* 77(15):19415–19438
20. Omidyeganeh M, Shirmohammadi S, Abtahi S, Khurshid A, Farhan M, Scharcanski J, Hariri B, Laroche D, Martel L (2016) Yawning detection using embedded smart cameras. *IEEE Trans Instrum Meas* 65(3):570–582
21. Zhang W, Murphey YL, Wang T, Xu Q Driver yawning detection based on deep convolutional neural learning and robust nose tracking. In: 2015 International Joint Conference on Neural Networks (IJCNN), 2015. IEEE, pp 1–8
22. Zhang W, Su J Driver yawning detection based on long short term memory networks. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), 2017. IEEE, pp 1–5
23. Akrouf B, Mahdi W Yawning detection by the analysis of variational descriptor for monitoring driver drowsiness. In: 2016 International Image Processing, Applications and Systems (IPAS), 2016. IEEE, pp 1–5
24. Zhao L, Wang Z, Wang X, Qi Y, Liu Q, Zhang G (2016) Human fatigue expression recognition through image-based dynamic multi-information and bimodal deep learning. *J Electronic Imaging* 25(5):053024
25. Yu J, Park S, Lee S, Jeon M Representation learning, scene understanding, and feature fusion for drowsiness detection. In: Asian Conference on Computer Vision, 2016. Springer, pp 165–177
26. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. *IEEE Trans Pattern Analysis Machine Intell* 35(1):221–231
27. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324
28. 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, 2016. pp 770–778
29. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. pp 4700–4708
30. Xie S, Girshick R, Dollár P, Tu Z, He K aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. pp 1492–1500
31. Hara K, Kataoka H, Satoh Y Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018. pp 6546–6555
32. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T decaf: a deep convolutional activation feature for generic visual recognition. In: International conference on machine learning, 2014. pp 647–655
33. Yosinski J, Clune J, Bengio Y, Lipson H How transferable are features in deep neural networks? In: Advances in neural information processing systems, 2014. pp 3320–3328
34. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P (2017) The kinetics human action video dataset. arXiv preprint arXiv:170506950

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Lei Zhao<sup>1</sup> · Zengcai Wang<sup>2</sup> · Guoxin Zhang<sup>3</sup> · Huanbing Gao<sup>4</sup>

<sup>1</sup> School of Mechanical and Electronic Engineering, Shandong Jianzhu University, No. 1000 of Fengming Road, Jinan, China

<sup>2</sup> Vehicle Engineering Research Institute, School of Mechanical Engineering, Shandong University, Jinan, China

<sup>3</sup> Department of Mechanical Engineering, Hong Kong Polytechnic University, Hung HomKowloon, Hong Kong

<sup>4</sup> School of Information and Electrical Engineering, Shandong Jianzhu University, No. 1000 of Fengming Road, Jinan, China