IET Intelligent Transport Systems

Research Article

# Fatigue driving recognition network: fatigue driving recognition via convolutional neural network and long short-term memory units

ISSN 1751-956X
Received on 26th August 2018
Revised 30th April 2019
Accepted on 13th May 2019
E-First on 6th June 2019
doi: 10.1049/iet-its.2018.5392
www.ietdl.org

*Zhitao Xiao*[1,2], *Zhiqiang Hu*[1,2], *Lei Geng*[1,2] ✉, *Fang Zhang*[1,2], *Jun Wu*[1,2], *Yuelong Li*[2,3]

[1] *Tianjin Polytechnic University, School of Electronics and Information Engineering, No. 399 Binshui West Street, Xiqing District, Tianjin 300387, People's Republic of China*

[2] *Tianjin Key Laboratory of Optoelectronic Detection Technology and Systems, No. 399 Binshui West Street, Xiqing District, Tianjin 300387, People's Republic of China*

[3] *Tianjin Polytechnic University, School of Computer Science and Software Engineering, No. 399 Binshui West Street, Xiqing District, Tianjin 300387, People's Republic of China*

✉ *E-mail: genglei@tjpu.edu.cn*

**Abstract:** Fatigue driving has become one of the major causes of traffic accidents. The authors propose an effective method capable of detecting fatigue state via the spatial–temporal feature of driver's eyes. In this work, the authors consider fatigue detection as image-based sequence recognition and an end-to-end trainable convolutional neural network with long short-term memory (LSTM) units is designed. First, the authors apply a deep cascaded multi-task framework to extract eye region from infrared videos. Then the spatial features are learned by deep convolutional layers and the relationships between adjacent frames are analysed via LSTM units. Finally, through authors' model, a sequence-level prediction for driving state is produced. The proposed method achieves superior accuracy over the state-of-the-art techniques on authors' own dataset. Experimental results demonstrate the feasibility of authors' method.

## 1 Introduction

With the increase of vehicle holdings and the acceleration of living rhythm, road traffic accidents have become one of the social problems that seriously threaten the safety of human life and property. According to the statistics of National Bureau of Statistics of China, traffic accidents caused by fatigue driving account for >20% of the total number of traffic accidents and account for >40% of serious traffic accidents, which shows that fatigue driving is a common cause of traffic accidents. It is particularly critical to automatically warn drivers when they are drowsy. Accordingly, research on driver fatigue detection approaches and development of automatic detection system have extremely significant social value. At present, the mechanisms in fatigue driving detection have been categorised into three broad methods [1, 2], including physiological-based, vehicle-based, and vision-based approaches.

In earlier fatigue detection studies, researchers mainly used biological knowledge to detect fatigue. A series of physiological signals such as electroophthalmogram [3], electrocardiogram (ECG) [4], electroencephalogram, electromyogram [5], electrooculogram [5, 6], and respiratory rate are acquired from various sensors attached to the driver's body during driving. Various physiological parameters differences between fatigue state and normal state are analysed in order to judge whether the driver is tired or not. For example, a real-time driver's health condition monitoring system with drowsiness alertness was proposed by Jung *et al.* [4]. They collected ECG with embedded sensor on the steering wheel. The driver's health condition such as the normal, fatigued and drowsy states was finally analysed by evaluating the heart rate variability in the time and frequency domains; Correa *et al.* [7] got 83.6% accuracy using time, spectral and wavelet analysis to process EGG signals, which finally were fed into a neural network classifier for drowsiness detection. Usually, the detection methods based on driver's physiological characteristics have high detection accuracy and can objectively and accurately reflect the actual driving state of the driver, but the support of professional and expensive signal acquisition devices are needed. This type of method even requires contact with the driver's body, which may bring discomfort to the driver in the actual driving process and affect the driver's normal control of the vehicle.

The fatigue detection method relying on vehicle operating status information is used to measure the behaviours of vehicles, such as speed, driving trace, lane departure etc. Among them, steering wheel movement and lateral distance are two important characteristics. Ma *et al.* [8] used discrete wavelet analysis and neural network to process the lateral distance of the vehicle aimed to see a pattern for detecting drowsiness. Lateral distance is a feature that acquired from fusing lane position, lane curvature, and lane curvature derivative. Three of those are raw features obtained from video camera installed on car's hood or front bumper. The biggest challenge in this method is to distinguish between lane change, merge, exit etc. with lane changing caused by fatigue; Li *et al.* [9] yielded an average fatigue detection accuracy of 78.01% using the steering wheel angles and a well-designed binary decision classifier. This method is affected by vehicle type, drivers' individual differences, especially road conditions.

A third way to identify the drowsy driver is to apply cameras and machine vision algorithms for capturing and analysing face visual information while driving, such as blink frequency, yawning [10], head movement [11], and gaze direction [12], which can provide observable cues in changing facial features [13, 14]. A significant evaluation index called 'PERCLOS [15]', referring to the percentage of the eye closing time over a specific time period, is widely used to identify fatigue. Mandal *et al.* [16] presented a vision-based fatigue detection system for bus driver monitoring, which is easy and flexible for deployment in buses and large vehicles. The system consists of modules of head-shoulder detection, face detection, eye detection, eye openness estimation, fusion, PERCLOS estimation, and fatigue level classification; In addition, yawning detection is also one of the main means to judge fatigue driving. With accuracy up to 98%, a yawning detection algorithm proposed by Alioua *et al.* [10] utilised SVM detector for face extraction and extracted mouth regions via circular Hough transform. This type method has been well received, due to its non-intrusive, low-cost and friendly peculiarity in monitoring the driving state. The visual behaviour of drowsy drivers is significantly different compared to sober-minded drivers.

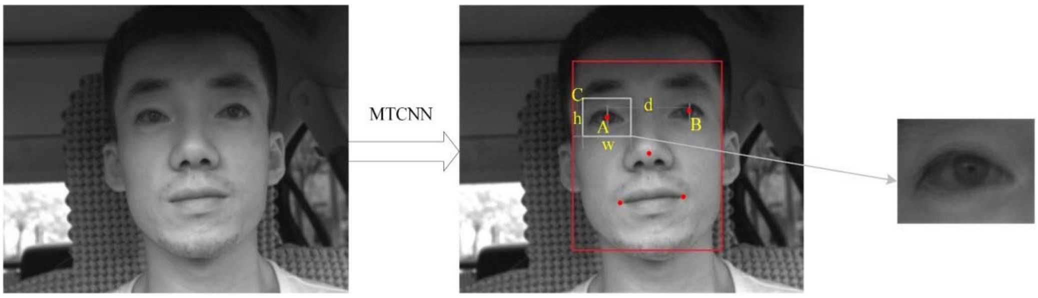

**Fig. 1** *Extracting the region of interest by MTCNN. The coordinates of the five red feature points of the face are output by MTCNN. The eye feature points are the centre of the eye region. The width of the eye region is 0.6 times of the horizontal distance between the two eye feature points and the height is 0.8 times of the width*

In recent years, deep learning [17] algorithms represented by convolutional neural networks (CNNs) have been successfully applied in substantial visual tasks: target categorisation, object segmentation, target detection etc., due to its strong feature extraction ability and robustness. Deep learning with visual features to judge fatigue driving has become a popular trend. Based on this method, researchers have achieved some good results. For instance, a yawning detection algorithm based on convolution neural network was proposed by Ma *et al.* [18] The driver's facial image is directly used as the input of the neural network to avoid complex feature extraction of facial images. Softmax classifier is used to classify the features extracted by the neural network and determine whether yawning occurs. The algorithm achieves 92.4% yawning detection accuracy on YawDD data set; Zhang *et al.* [19] developed a real-time fatigue recognition system and yielded a high accuracy of 95.81%, which identify whether the eyes are open or closed at any time via CNN and judge the driving state according to PERCLOS criterion, but the system ignores the intermediate state during eyes closure.

Currently, there are some problems with the method based on Deep learning and face visual features, which are shown in the following: (i) Lack of professional data sets for fatigue driving detection. (ii) Great interference in extracting the driver's visual features is caused by the change of light during actual driving. (iii) When the driver wears myopia glasses or sunglasses, the eye area will be occluded.

In this paper, we first attempt to consider fatigue detection as a classical problem in computer vision: image-based sequence recognition. On one hand, we produce the specialised dataset named ‘TJPU-FDD’ for fatigue detection. Capturing facial images by Infrared camera can reduce the interference of light change and obtain clear eye images through sunglasses; On the other hand, we design a novel end-to-end trainable CNN with long short-term memory (LSTM) units [20] [fatigue driving recognition network (FDRNet)] for fatigue driving judgement, through analysing spatial–temporal feature of eyes to identify sleepiness state. In FDRNet, the SE block [21] is embedded into the residual learning module [22] in order to make the network learns the residual features which are re-calibrated according to the importance of the channel. The pre-trained model is used to fine-tune the first few layers of FDRNet. The classifier for optimal feature extraction can be automatically obtained from the training data with little pre-processing. Compared to traditional methods, a higher classification accuracy is achieved.

## 2 Materials and methods

In this section, we will describe our approach to detect fatigue driving.

### 2.1 Data preparation

The success of deep learning in the field of object recognition is based on abundant experimental data distributed independently. Due to the lack of public datasets, 26 subjects, which all had a current driver's license, participated in the driving simulations experiment. Under the condition of wearing myopia glasses, wearing sunglasses, and not wearing glasses, respectively, the subjects simulate two kinds of driving states (fatigue and normal) on the passenger seat because fatigue driving behaviour is dangerous and illegal. In order to reduce the impact of illumination changes and get clear eye images when drivers wear sunglasses, we use infrared camera with filters to capture face videos at 30 fps with a resolution of 1920 × 1080 as experimental dataset named ‘TJPU-FDD’, which is divided into two classes: fatigue driving and normal driving. All samples from TJPU-FDD consist of 500 video clips, each lasting ∼6 s. In our samples labelled fatigue, the driver blinks more slowly and eyes close for a long time. In addition, the driver cannot keep eye open like normal. The phenomena such as slow eyeball movement, yawning, and eyelid closure occur. The ratio of training set, validation set and test set is four to one to one.

### 2.2 Extracting the region of interest

At present, fatigue detection relying on driver's face images has become the mainstream. The location of driver's eyes is the key in drowsiness recognition. Due to various postures, illuminations and occlusions, it is challenging to detect and align face in an unconstrained environment. A deep-cascaded multi-task framework which utilises their intrinsic connections to improve performance is proposed by Zhang *et al.* [23]. Particularly, this framework adopts a cascaded structure with a deep three-stage convolution network carefully designed to predict face and landmark location in a coarse-to-fine manner, which also yields good results when detecting side faces.

As shown in Fig. 1, we apply multi-task cascaded convolutional network (MTCNN) to calibrate feature points of eyes and then the corresponding image sequences of eyes are extracted from each video clip in the TJPU-FDD dataset according to the geometric relations between the feature points, which are shown as follows:

$$\begin{cases} d = x_{\rm B} - x_{\rm A} \\ w = d \times 0.6 \\ h = w \times 0.8 \\ x_{\rm C} = x_{\rm A} - w \times 0.5 \\ y_{\rm C} = y_{\rm A} - h \times 0.5 \end{cases}, \tag{1}$$

where $(x_{\rm A},\ y_{\rm A})$and $(x_{\rm B},\ y_{\rm B})$ denote the pixel coordinates of eye feature point A, B, respectively. Point C, whose pixel coordinate is$(x_{\rm C},\ y_{\rm C})$, refers to the upper-right vertex of right eye area. In addition, $d$ is the horizontal distance between point A and B. $w$ and $h$ denote the width and the height of the right eye area, respectively. See Fig. 1 for more details.

### 2.3 Fatigue driving detection

*2.3.1 Overall framework:* In the real world, a stable visual object such as fatigue state often appears in the form of a sequence, rather than in isolation. Unlike general target recognition, identifying such sequence object often requires to systematically predict a series of object labels, not a single label. Fatigue driving is a continuous and dynamic process. Generally, the blink frequency

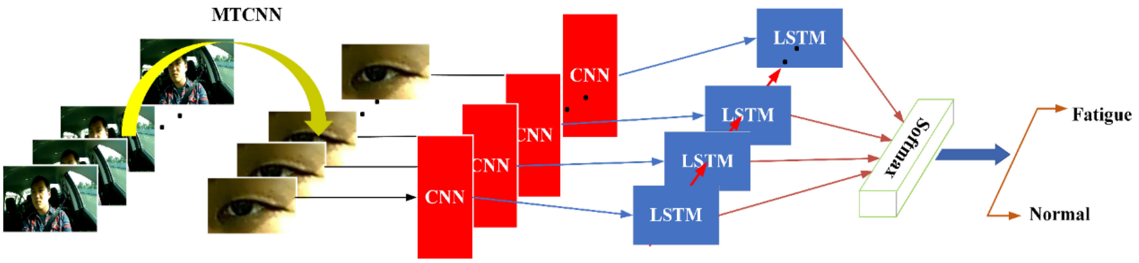

**Fig. 2** *Overall framework for fatigue detection*

decreases and the duration of closing eyes increases when the driver feels tired.

CNNs are 'deep in space' and adept at extracting images spatial information, but unable to model time-series changes. For example, we can easily judge the state of the driver's eyes at every moment only through CNN, but it is almost impossible to predict whether the driver is fatigued by the change of eye state over a period of time. In contrast, recurrent neural networks (RNNs) are 'deep in time', which means the network structure can be unfold in the time domain. The chain-like nature of RNNs reveals that they are intimately related to sequences and lists. Besides, LSTM unit, a special kind of RNNs, really enables long-range learning and prevents 'vanishing and exploding gradients' effect compared to traditional RNN architectures. LSTM seems to be more suitable for drowsiness detection task.

In the movie, the causal relationship between adjacent frames can predict the progression of the plot. Fatigue detection should be performed in time series for the same reason rather than in a single image, which could be conducive to more accurate prediction. In this paper, we consider fatigue detection as the classification of driving state according to time-series changes of eyes.

We propose a convolutional recurrent network architecture which is end-to-end trainable and suitable for sleepy detection task. This framework utilises CNN and LSTM units to extract spatial features of eyes and analysis their sequential relations, respectively. In contrast to standard CNNs, our model is deeper in that compositional representations in space and time are learned.

First, the continuous frames sequence of the eye is obtained from video clips by MTCNN. After the previous processing, we regard the continuous $T$ frames of each video as a time-step. Each input $x_t$ in a time-step is passed through a deep hierarchical CNN module aimed to learn convolutional perceptual representations and produce a fixed-length vector representation $y_t$. In order to capture temporal state dependencies, the outputs of the CNN module are then sent to a recurrent sequence learning models: LSTM units, which finally produce a multidimensional vector prediction. We can optimise model's weights in the form of end to end. Furthermore, it is critical that the weights are reused at per time-step in order to prevent the parameter size from growing in proportion to the sequence length. Two modules are connected directly and can be trained jointly to predict the driving state class at each time-step. With the softmax layer, video-level prediction is made at each time-step in the end of network. To make a single label prediction for an entire video clip, we average the label probabilities across all time-steps and choose the most probable class as the final video-level prediction for driving state. The overall framework, as applied to fatigue detection task, is depicted in Fig. 2.

The excellences of LSTM units for modelling sequential data in fatigue detection task are twofold. First, when integrated with deep convolutional layers, LSTMs are straight forward to fine-tune end-to-end. Second, LSTMs enable simple modelling for eye's image sequence of varying lengths.

*2.3.2 Network architecture:* Basically, deeper networks perform better. Yet a notorious problem of vanishing gradients emerges with the network depth increasing, which shackles convergence from the beginning. This problem, nevertheless, has been largely settled by ResNet, whose core lies in the convolution layers with 'shortcut connection'. Shortcut connections skip one or more layers, transferring information from the lower layer to the top without extra parameter and computational complexity. Compared to improving the representational power of a network in the perspective of space, SE-Net shows the 'Squeeze-and-Excitation' (SE) block (Fig. 3*a*) that providing the network with a mechanism to explicitly model dynamic, non-linear dependencies between channels. Using global information in the SE block can ease the learning process, and significantly enhance the representational power of the network. Assuming that the input dimension of SE blocck is $H \times W \times C$, first, each two-dimensional feature is transformed into a real number through the global average pooling layer. Then the input feature dimension is reduced to $1/R$ times by one fully connected layer, which is called 'squeeze operation'. After activated by ReLU, the feature dimension is restored through another fully connected layer and then the sigmoid activation layer is used to normalise the weights to (0, 1), which is called 'excitation operation'. Finally, after passing through a Scale layer, the feature dimension remains unchanged but is recalibrated between channels, which is called 'scale operation'.

The CNN with LSTM units (FDRNet) proposed in this paper is a variant of ResNet-10. We combine Residual module with SE block, which we name 'Res-SE module', for better performance in fatigue driving recognition. In Fig. 3*b*, assuming that the dimension of input fed into the Res-SE module is $H_{\rm r} \times W_{\rm r} \times C_{\rm r}$, the dimension of output will be $H_{\rm r}/2 \times W_{\rm r}/2 \times 2C_{\rm r}$. One branch of Res-SE module where Conv_3 is located is called 'shortcut connection'. Another branch where Conv_1 and Conv_2 are located is called 'residual connection'. In addition, [24] shows that placing BN [25] layer and ReLU layer in the pre-activation area can improve the regularisation of the model.

Through Res-SE module, the learning task of the network is simplified. Moreover, the residual feature after re-calibration is learned, which enhances the network performance. The whole process of processing features by Res-SE module is shown in Formula (2)

$$\begin{cases} X' = F_{\rm conv\_1}(X) \\ \bar{x} = F_{\rm scale}(F_{\rm excitation}(F_{\rm squeeze}(X'))) \\ X'' = F_{\rm conv\_3}(X) \\ \bar{X} = X'' + \bar{x} \end{cases}, \qquad (2)$$

where $X$ denotes the input of the Res-SE module and $X'$ represents the output of 'conv_1'. $F_{\rm squeeze}(\cdot)$, $F_{\rm excitation}(\cdot)$ and $F_{\rm scale}(\cdot)$ refer to a series of operations described above in the SE module. $\bar{x}$ denotes the residual feature after re-calibrated. $X''$ represents the output of 'shortcut connection' and $\bar{X}$ represents the output of the Res-SE module. See Fig. 3*b* for more details.

Fig. 4 depicts the network architecture, where the spatial feature extractor is constructed by a convolution layer, a pooling layer with a residual mapping, three Res-SE modules and ends with a global pooling operation. A reshaping layer named 'Reshape_CNN' in Fig. 4 reshapes the output of CNN part to sequential representations, which is one input of LSTM units. Another input reshaped by the reshaping layer called 'Resahpe_markers' is the continuous markers. The final fully connected layer and softmax activation produce a distribution over the two output classes (fatigue driving and normal driving) for each time-step. See more details and other variants in Table 1.

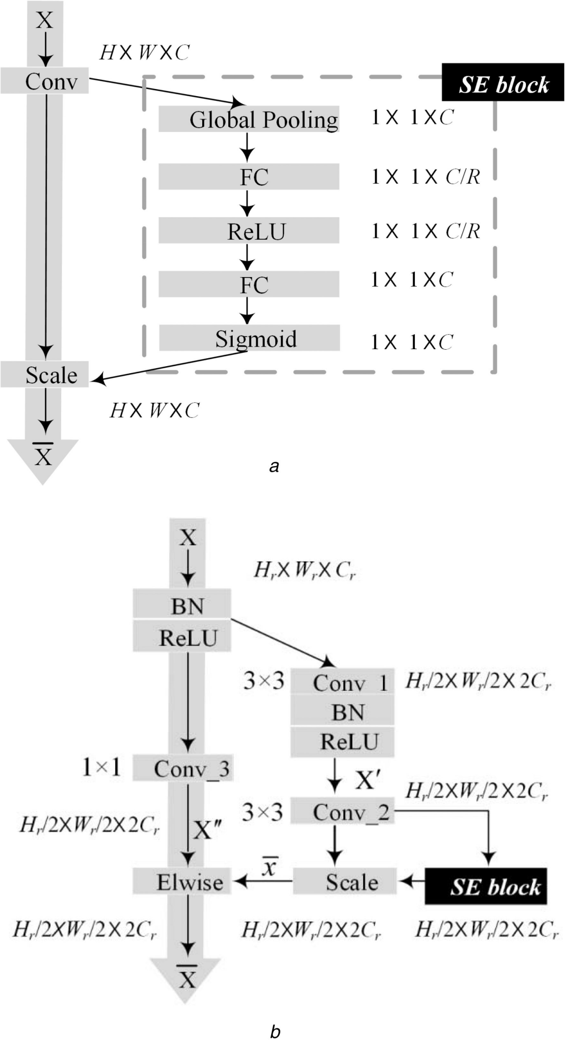

**Fig. 3** *Schema of the SE block*
***(a)*** And the Res-SE module, ***(b)*** Here Conv, Conv_1, Conv_2, and Conv_3 all represent the convolution layer, where Conv_1 has the 3 × 3 filter with a stride of 2, Conv_2 has 3 × 3 filters with a stride of 1, and Conv_3 has 1 × 1 filter with a stride of 2. FC refers to the fully connected layer and Global Pooling denotes the global pooling layer

In our network structure, LSTMs play a vital role in connecting spatial information in series, allowing our models to learn long-term dependencies. The core of LSTM is the three non-linear gates carefully designed, respectively, called input gate, output gate and forget gate, which are a way to optionally let information through and control cell states to be forgotten, updated, or retained during propagation.

*2.3.3 Implementation:* During training, we follow standard practice and perform data augmentation with random-size cropping to 227 × 227 pixels. Additionally, input images are normalised through mean channel subtraction. Optimisation is performed using SGD with a mini-batch size of 256, a weight decay of 0.0005 and a momentum of 0.9. The initial learning rate is set to 0.01 and decreased by a factor of 10 every 40 epochs. We set the length of the time-step to 16 for LSTM units. The entire network is trained end-to-end for 200 epochs on the TJPU-FDD training set and training is performed on a server with a NVIDIA GTX 1080Ti GPU. We fine-tune the first two layers weights of FDRNet via the ResNet-10 model pretrained on the ImageNet dataset aimed to speed up model convergence and prevent overfitting on our small-scale dataset. The rest layers use the weight initialisation strategy described in [27]. A gradient norm clipping scheme [28] is adopted to deal with gradients exploding. We save the best model as evaluated on the validation set during the optimisation process. At test time, the predicted results are averaged across all time-steps as the final output for per clip of the test set.

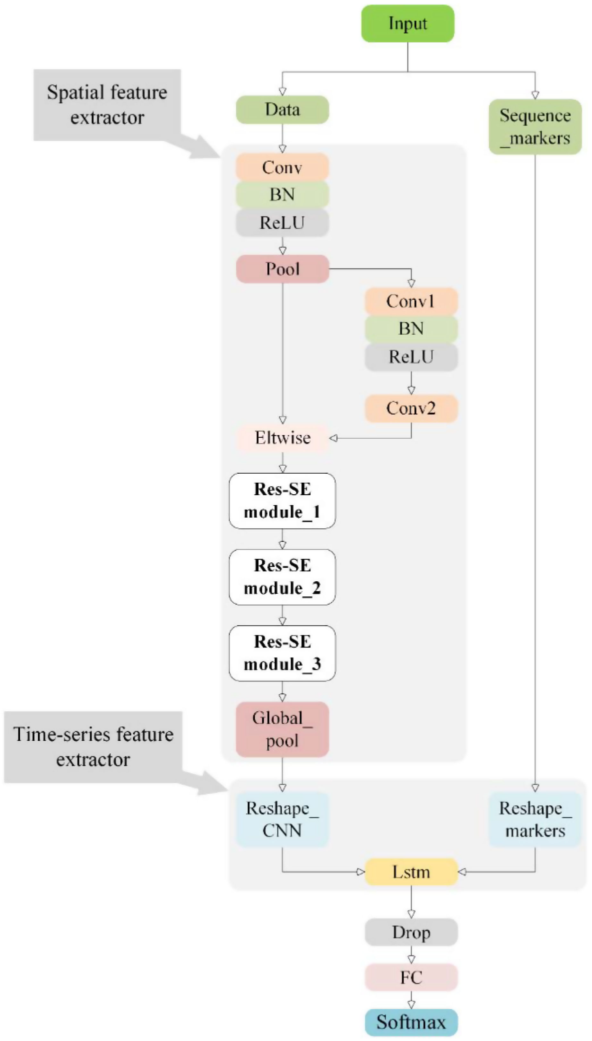

**Fig. 4** *Architecture of FDRNet*

**Table 1** Related parameters of different layers

| Layer name | Layer type | Related parameters |
|---|---|---|
| Conv | convolution | 7 × 7, 64, stride 2 |
| Pool | pooling | 3 × 3, max pool, stride 2 |
| Conv1 | convolution | 3 × 3, 64, stride 1 |
| Conv2 | convolution | 1 × 1, 64, stride 1 |
| Res-SE module | \ | $C_r$ 64\128\256, $R$ 16 |
| Global_pool | pooling | global average pool |
| Lstm | LSTM | $L$ 256 |
| Drop | dropout [26] | dropout-ratio 0.25 |
| FC | fully connected | 2-d |

Here $C_r$ refers to the number of input channels for the Res-SE module, and $L$ denotes the number of hidden units for the LSTM units

## 3 Experiments

In this section, we will compare FDRNet with other networks on TJPU-FDD validation set and test the performance of each model on our test set. Besides, the comparison between our approach and other previous methods will be done.

### *3.1 Comparing the performance of different networks on the validation set*

Convolutional neural networks have proven to be impactful models for tackling a variety of visual tasks. However, the proposed

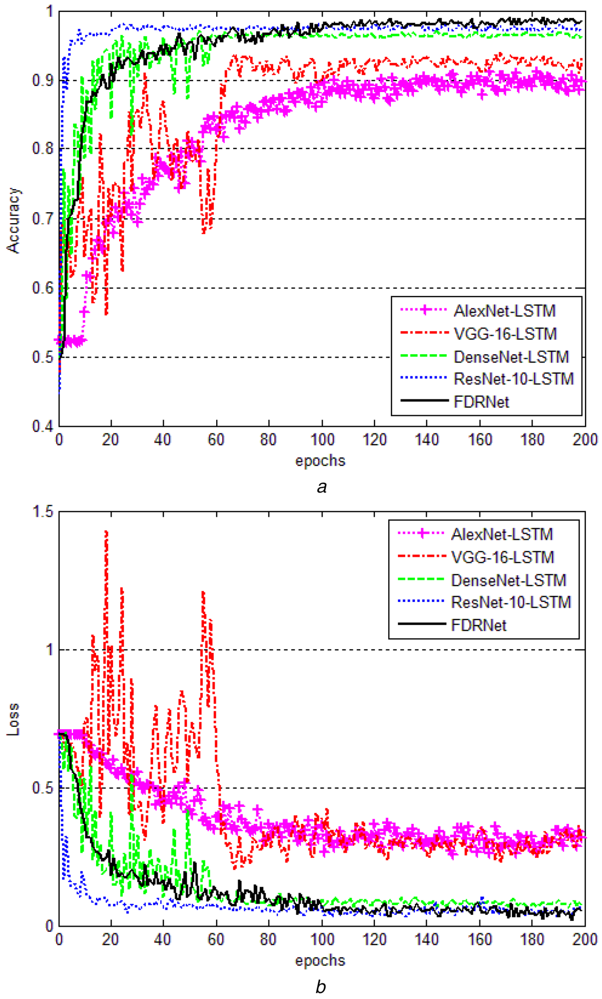

**Fig. 5** *Varying curves of*
***(a)*** Accuracy, and ***(b)*** Loss

network is, to the best of our knowledge, the first attempt to apply convolutional neural network with LSTM units for drowsiness recognition via extracting sequential-spatial features of drive's eyes. In our experiments, we try to combine some current networks with LSTM units, respectively, aimed to fully utilise the characteristics of each network and achieve the best performance in fatigue driving detection. We choose a few previous classification networks, e.g. AlexNet [29], VGG-16 [30], DenseNet [31] and ResNet-10, to replace separately the spatial feature extractor (shown in Fig. 5) with the first part of each network before the final fully connected layer. For instance, AlexNet-LSTM represents the combination of AlexNet and LSTM units.

We connect DensNet containing four Dense blocks with LSTM units in DenseNet-LSTM network where each layer has direct access to the gradients from the loss function and the original input signal via Dense block. The pre-trained model of ResNet-10 is applied to fine-tune the CNN part of ResNet-10-LSTM. Before the feature is fed into LSTM units, we reshape the output of different spatial feature extractor to the corresponding dimension. Besides, in order to achieve the best accuracy, the hyperparameter of some networks are slightly adjusted before training.

CNNs with LSTM units differs in the classification performance due to discrepant network structures and training strategy etc. Among these networks, FDRNet achieves the highest classification accuracy. Table 2 shows the comparison of the fatigue driving recognition accuracy of different hybrids on the TJPU-FDD validation set. The varying curves of accuracy and loss are depicted in Fig. 5.

**Table 2** Comparison of recognition accuracy between different hybrids on the validation set

| Hybrid | Accuracy, % |
|---|---|
| AlexNet-LSTM | 91.28 |
| VGG-16-LSTM | 93.83 |
| DenseNet-LSTM | 97.12 |
| ResNet-10-LSTM | 98.01 |
| FDRNet (Ours) | 98.96 |

**Table 3** Performance comparison of FDRNets with different T values on the validation set

| *T* | Accuracy, % |
|---|---|
| 8 | 90.51 |
| 16 | 98.96 |
| 24 | 95.33 |

**Table 4** Performance comparison of FDRNets with different R values on the validation set

| *R* | Accuracy, % |
|---|---|
| 4 | 95.52 |
| 8 | 98.15 |
| 16 | 98.96 |
| 24 | 97.47 |

### *3.2 Comparing the performance of FDRNet with different parameters*

Usually, it takes ∼0.3 s for the driver to blink. When the driver is tired, blinking becomes slower and lasts longer. In FDRNet, different lengths $T$ (8, 16, 24) of the time step are selected for LSTM units, respectively, that is, we select continuous $T$ frames as the detection window to experiment. The experimental results are shown in Table 3. It can be seen that when the time step length is 16, the network performs best on the validation set and the highest accuracy can reach 98.96%. Besides, different values of $R$ influence FDRNet accuracy. Table 4 shows the performance comparison of FDRNets with different $R$ values on the validation set. When the value of $R$ is 16, the highest accuracy can be achieved.

During the network design, the amount of LSTM units is considered as a vital factor. Indeed, we find that with the increasing number of LSTM units, the classification accuracy of each network drops to some extent and the network even does not converge. So, the number of LSTM units is finally set to 1.

### *3.3 Comparing the performance of different models on the test set*

TJPU-FDD test set consists of 103 video clips. Table 5 compares the performance of each model on the test set in terms of recall rate, precision, F1 score, wrong recognition rate, average test time, where N represents the number of wrong recognition videos. We can see that FDRNet achieves highest recall rate, precision, F1 score, lowest wrong recognition rate, less detection time, which proves that the effectiveness and feasibility of FDRNet. Besides, the detailed recognition results of FDRNet is shown in Table 6, where the first two columns refer to data distribution of TJPU-FDD test set, and the last two columns represent the number of wrong recognition videos. By observing the wrong recognition videos, we find that there are two common factors among them which could lead to erroneous recognition. One is interference from spectacle frames and the other is reflection of eyeglasses lenses.

### *3.4 Comparing with other approaches*

Many researchers have proposed different approaches for fatigue detection and achieved good results. Compared to those approaches, the scheme based on deep learning and visual features

**Table 5** Comparison of performance between different models on TJPU-FDD test set

| Models | Data distribution of test set | Recognition results |  | *N* | Precision, % | Recall rate, % | F1 score, % | Average test time, ms |
|---|---|---|---|---|---|---|---|---|
|  |  | Fatigue | Normal |  |  |  |  |  |
| AlexNet-LSTM | fatigue (48) | 40 | 8 | 14 | 86.96 | 83.33 | 85.11 | 245.87 |
|  | normal (55) | 6 | 49 |  |  |  |  |  |
| VGG-16-LSTM | fatigue (48) | 41 | 7 | 11 | 91.11 | 85.42 | 88.17 | 497.54 |
|  | normal (55) | 4 | 51 |  |  |  |  |  |
| DenseNet-LSTM | fatigue (48) | 44 | 4 | 7 | 93.62 | 91.67 | 92.63 | 91.73 |
|  | normal (55) | 3 | 52 |  |  |  |  |  |
| ResNet-10-LSTM | fatigue (48) | 44 | 4 | 6 | 95.65 | 91.67 | 93.62 | 98.66 |
|  | normal (55) | 2 | 53 |  |  |  |  |  |
| FDRNet (Ours) | fatigue (48) | 46 | 2 | 4 | 95.83 | 95.83 | 95.83 | 132.34 |
|  | normal (55) | 2 | 53 |  |  |  |  |  |

**Table 6** Detailed recognition results of FDRNet on TJPU-FDD test set

| Real data distribution of test set |  | Number of wrong recognition videos |  |
|---|---|---|---|
|  |  | Fatigue | Normal |
| no glasses | Fatigue (20) | — | 0 |
|  | Normal (22) | 0 | — |
| myopia glasses | Fatigue (19) | — | 2 |
|  | Normal (23) | 1 | — |
| sunglasses | Fatigue (9) | — | 0 |
|  | Normal (10) | 1 | — |

**Table 7** Comparing the performance with the work in [19]

| Method | Classifier | *N* | Precision, % | Recall rate, % | F1 score, % | Average test time, ms |
|---|---|---|---|---|---|---|
| Zhang *et al.* | CNN and PERCLOS | 7 | 95.56 | 89.58 | 92.47 | 41.58 |
| FDRNet (Ours) | CNN with LSTM units | 4 | 95.83 | 95.83 | 95.83 | 132.34 |

is non-intrusive, low-cost, independent of individual factors and increasingly popular. Due to the limitations of hardware and software and the lack of professional data sets, the fatigue detection algorithms based on driver's physiological characteristics and vehicle behaviour characteristics are difficult to reproduce. In addition, there are not many existing studies based on deep learning. To further prove the superiority of our method, we only reproduce the fatigue detection method in [19] based on our data set and the same experimental environment as ours. As the experimental results shown in Table 7, our method achieves lower wrong recognition rate, higher presicion, recall rate and F1 score than [19], although fatigue driving detection using our method takes longer, which validates the idea that incorporating information across video sequences will enable better prediction.

## 4 Conclusion

In this paper, a novel method for fatigue driving recognition based on driver's eyes has been presented. We design an end-to-end trainable network combining convolutional layers with LSTM units, which are capable of learning spatial representations and modelling temporal dynamics. Aimed to reduce the influence of illumination, we use an infrared camera with filters to capture infrared videos towards driver's face, and MTCNN is used to obtain the eye area as experimental data. Residual learning module with SE block and transfer learning [32] are introduced to accelerate convergence and improve classification accuracy in this network. Via integrating the feature in time and space domain, our model can make a video-level prediction for driving state. Extensive experiments demonstrate the effectiveness of our architecture which achieves state-of-the-art performance in sleepiness detection with little input pre-processing and no hand-designed features. However, the reflective spot affects the eyes clarity when the drive wears glasses. We will next focus on perfecting image acquisition system for avoiding the above scenario. In addition, the accuracy of eye positioning directly affects the performance of fatigue detection model, which is a significant aspect that should be improved. Fusion with yawn detection is also a good choice to further boost up the performance in the future work.

## 5 Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC), under grant no. 61601325; Tianjin Science and Technology Major Projects and Engineering, under grant nos. 17ZXHLSY0040, 17ZXSCSY0060 and 17ZXSCSY0090; Tianjin Natural Science Foundation, under grant no. 17JCQNJC01400; and The Program for Innovative Research Team in University of Tianjin, under grant no. TD13-5034.

## 6 References

[1] Pratama, B.G., Ardiyanto, I., Adji, T.B.: ‘A review on driver drowsiness based on image, bio-signal, and driver behaviour’. IEEE Int. Conf. on Science and Technology-Computer, Yogyakarta, Indonesia, 2017, pp. 70–75

[2] Triyanti, V., Iridiastadi, H.: ‘Challenges in detecting drowsiness based on driver's behaviour’, *Mater. Sci. Eng. Conf. Series*, 2017, **277**, (1), p. 012042

[3] Mardi, Z., Ashtiani, S.N.M., Mikaili, M.: ‘EEG-based drowsiness detection for safe driving using chaotic features and statistical tests’, *J. Med. Signals Sens.*, 2011, **1**, (2), pp. 130–137

[4] Jung, S.J., Shin, H.S., Chung, W.Y.: ‘Driver fatigue and drowsiness monitoring system with embedded electrocardiogram sensor on steering wheel’, *IET Intell. Transp. Syst.*, 2014, **8**, (1), pp. 43–50

[5] Fatourechi, M., Bashashati, A., Ward, R.K.*, et al.*: ‘EMG and EOG artifacts in brain computer interface systems: A survey’, *Clin. Neurophysiol.*, 2007, **118**, (3), pp. 480–494

[6] Gasser, T., Sroka, L., Möcks, J.: ‘The transfer of EOG activity into the EEG for eyes open and closed’, *Electroencephalogr. Clin. Neurophysiol.*, 1985, **61**, (2), pp. 181–193

[7] Correa, A.G., Orosco, L., Laciar, E.: ‘Automatic detection of drowsiness in EEG records based on multimodal analysis’, *Med. Eng. Phys.*, 2014, **36**, (2), pp. 244–249

[8] Ma, J., Murphey, Y.L., Zhao, H.: ‘Real time drowsiness detection based on lateral distance using wavelet transform and neural network’. Computational Intelligence. 2015 IEEE Symp., Cape Town, South Africa, 2015, pp. 411–418

[9] Li, Z., Li, S.E., Li, R.*, et al.*: ‘Online detection of driver fatigue using steering wheel angles for real driving conditions’, *Sensors*, 2017, **17**, (3), p. 495

[10] Alioua, N., Amine, A., Rziza, M.: ‘Driver's fatigue detection based on yawning extraction’, *Int. J. Vehicular Technol.*, 2014, **2014**, (1), pp. 47–75

[11] Mittal, A., Kumar, K., Dhamija, S., *et al.*: 'Head movement-based driver drowsiness detection: A review of state-of-art techniques'. IEEE Int. Conf. on Engineering and Technology, Coimbatore, India, 2016, pp. 903–908

[12] Choi, I.H., Kim, Y.G.: 'Head pose and gaze direction tracking for detecting a drowsy driver'. IEEE Int. Conf. on Big Data and Smart Computing, Bangkok, Thailand, 2014, vol. 9, no. 2, pp. 241–244

[13] Karchani, M., Mazloumi, A., Saraji, G.N., *et al.*: 'Presenting a model for dynamic facial expression changes in detecting drivers' drowsiness', *Electron. Physician.*, 2015, **7**, (2), pp. 1073–1077

[14] Tao, H., Zhao, Y.: 'Real-time driver fatigue detection based on face alignment'. Int. Conf. on Digital Image Processing, Hong Kong, 2017, (10420), p. 1042003

[15] Dinges, D.F., Grace, R.: 'PERCLOS: a valid psychophysiological measure of alertness as assessed by psychomotor vigilance'. US Department of Transportation, Federal Highway Administration, Publication Number FHWA-MCRT-98-006, 1998

[16] Mandal, B., Li, L., Wang, G.S., *et al.*: 'Towards detection of bus driver fatigue based on robust visual analysis of eye state', *IEEE Trans. Intell. Transp. Syst.*, 2017, **18**, (3), pp. 545–557

[17] Yann, L.C., Yoshua, B., Geoffrey, H.: 'Deep learning', *Nature*, 2015, **521**, (7553), pp. 436–444

[18] Ma, S.G., Zhao, C., Sun, H.L, *et al.*: 'Yawning detection algorithm based on convolutional neural network', *Comput. Sci.*, 2018, **45**, (6A), pp. 227–241

[19] Zhang, F., Su, J., Geng, L., *et al.*: 'Driver fatigue detection based on eye state recognition'. IEEE Int. Conf. on Machine Vision and Information Technology, Singapore, Singapore, 2017, pp. 105–110

[20] Hochreiter, S., Schmidhuber, J.: 'Long short-term memory', *Neural Comput.*, 1997, **9**, (8), pp. 1735–1780

[21] Hu, J., Shen, L., Sun, G.: 'Squeeze-and-excitation networks'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7132–7141

[22] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778

[23] Zhang, K., Zhang, Z., Li, Z., *et al.*: 'Joint face detection and alignment using multitask cascaded convolutional networks', *IEEE Signal Process. Lett.*, 2016, **23**, (10), pp. 1499–1503

[24] He, K., Zhang, X., Ren, S., *et al.*: 'Identity mappings in deep residual networks'. European Conf. on computer vision, Amsterdam, Ther Netherlands, 2016, pp. 630–645

[25] Ioffe, S., Szegedy, C.: 'Batch normalization: accelerating deep network training by reducing internal covariate shift', arXiv preprint arXiv:1502.03167, 2015

[26] Hinton, G.E., Srivastava, N., Krizhevsky, A., *et al.*: 'Improving neural networks by preventing coadaptation of feature detectors', *Comput. Sci.*, 2012, **3**, (4), pp. 212–223

[27] He, K., Zhang, X., Ren, S., *et al.*: 'Delving deep into rectifiers: surpassing human-level performance on ImageNet classification'. Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 1026–1034

[28] Pascanu, R., Mikolov, T., Bengio, Y.: 'On the difficulty of training recurrent neural networks'. Int. Conf. on Machine Learning, Atlanta, GA USA, 2013, pp. 1310–1318

[29] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'Imagenet classification with deep convolutional neural networks'. Advances in neural information processing systems, Lake Tahoe, NV, USA, 2012, pp. 1097–1105

[30] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', *Comput. Sci.*, 2014, arXiv:1409.1556

[31] Huang, G., Liu, Z., Maaten, L.V.D., *et al.*: 'Densely connected convolutional networks'. IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 2261–2269

[32] Oquab, M., Bottou, L., Laptev, I., *et al.*: 'Learning and transferring mid-level image representations using convolutional neural networks'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1717–1724