Hindawi
Journal of Electrical and Computer Engineering
Volume 2021, Article ID 9955079, 13 pages
https://doi.org/10.1155/2021/9955079

Research Article

Research on Safe Driving Evaluation Method Based on Machine Vision and Long Short-Term Memory Network

Dongmei Shi[1] and Hongyu Tang[2]

[1]*Department of Computer Science and Technology, Suzhou College of Information Technology, Suzhou, China*
[2]*School of Electrical and Information, Zhenjiang College, Zhenjiang, China*

Correspondence should be addressed to Hongyu Tang; t_redrain@126.com

Received 19 March 2021; Revised 29 March 2021; Accepted 1 April 2021; Published 15 April 2021

Academic Editor: Yang Li

The rapid development of transportation industry has brought some potential safety hazards. Aiming at the problem of driving safety, the application of artificial intelligence technology in safe driving behavior recognition can effectively reduce the accident rate and economic losses. Based on the presence of interference signals such as spatiotemporal background mixed signals in the driving monitoring video sequence, the recognition accuracy of small targets such as human eyes is low. In this paper, an improved dual-stream convolutional network is proposed to recognize the safe driving behavior. Based on convolutional neural networks (CNNs), attention mechanism (AM) is integrated into a long short-term memory (LSTM) neural network structure, and the hybrid dual-stream AM-LSTM convolutional network channel is designed. The spatial stream channel uses the CNN method to extract the spatial characteristic value of video image and uses pyramid pooling instead of traditional pooling, normalizing the scale transformation. The time stream channel uses a single-shot multibox detector (SSD) algorithm to calculate the adjacent two frames of video sequence for the detection of small objects such as face and eyes. Then, AM-LSTM is used to fuse and classify dual-stream information. The self-built driving behavior video image set is built. ROC, accuracy rate, and loss function experiments are carried out in the FDDB database, VOT100 data set, and self-built video image set, respectively. Compared with CNN, SSD, IDT, and dual-stream recognition methods, the accuracy rate of this method can be improved by at least 1.4%, and the average absolute error in four video sequences can be improved by more than 2%. On the contrary, in the self-built image set, the recognition rate of doze reaches 68.3%, which is higher than other methods. The experimental results show that this method has good recognition accuracy and practical application value.

1. Introduction

China's manufacturing industry has entered a period of rapid growth, along with the logistics and transportation industry rising. With the improvement in people's living standards, cars have become the main means of transportation and caused increasingly busy transportation and increasing traffic accidents, which also cause people's lives, production, and property losses. According to statistics data, the annual traffic accident rate is slowly increasing, and the situation is not optimistic. Research shows that traffic accidents are mainly caused by people, cars, roads, and environmental factors, among which fatigue driving and unsafe behaviors are the main causes of traffic accidents [1], accounting for 69% of traffic accidents. Long-time fatigue driving is easy to cause traffic accidents [2]. Unsafe behaviors mainly include illegal actions, calling, smoking, inattention, eating, and fatigue driving [3, 4]. Therefore, how to use modern scientific and technological means to reduce traffic accidents and losses is worthy of study for ensuring life safety.

At present, the camera installed in front of the cab is mainly used to collect the driver's real-time image information, and the deep learning theory is used to recognize the face and behavior state information, so as to judge whether the driver has unsafe behavior [5]. In face recognition and behavior recognition, scholars have done a lot of research. Face detection methods are mainly divided into two

categories: knowledge-based methods and statistics-based methods [6]. These two methods extract the features of the face region and judge by calculating the similarity of the face features or the response value of the classifier [7]. The knowledge-based methods have the features of skin color, texture, structure, edge, and shape. The statistical methods have been deep learning, such as artificial neural network (ANN), AdaBoost method, support vector machine (SVM) [8], feature space method, long-term recurrent convolutional networks (LRCNs), and convolutional neural networks (CNNs) [9, 10]. Methods based on feature space include principal component analysis (PCA), linear discriminant analysis (LDA), and local binary pattern (LBP) [11]. Their common feature is to use the mapping image of space vector in one of the feature spaces to distinguish and judge face and nonface. Gabor and Hog methods are usually used for face recognition [12]. Combining global features with local effective features to form the final features of face recognition, fisher coding weighting criteria are used to recognize face in the video [13]. Combining Gabor features with collaborative representation, a face recognition algorithm is proposed, which improves the speed of face recognition to a certain extent [14]. The driver face detection method is an extension or extended application of the face recognition method. Driver face recognition needs to estimate the position of the face region in each frame of the input video, labeling the face region through visualization, and then to carry out machine learning. In video image behavior recognition, face feature extraction and appearance expression are the important basis of behavior recognition algorithm. The common appearance expression methods include contour template, optical flow, and feature points. Generally, color histogram, Haar feature or Haar-like feature, histogram of oriented gradient (HOG) feature operator [15], and wavelet algorithm [16] are used to extract feature from the candidate region of video image. Then, machine learning classifiers such as softmax, SVM, boosting, or random forest are used for fast classifier learning. Classifiers usually use the target information of the previous frame or several consecutive frames for online learning and obtain the target state information and constantly update it.

In general, the most direct method for driving safety behavior is to use machine vision real-time monitoring. Through the analysis of video information, the detection methods mainly include the physiological parameter detection method, vehicle trajectory behavior detection method, and driver behavior characteristics detection method [17]. Because of the need for medical detection equipment, the application value of the physiological parameter detection method is not high. The method of vehicle trajectory behavior detection has some misjudgment and hysteresis. The method of the driver behavior feature detection uses machine vision to judge safety behavior and detects fatigue state information by detecting percentage of eye closure over the pupil over time (PERCLOS) [18], blinking frequency, gaze direction, mouth features, facial expression, and other features [19]. The multiscale retinex (MSR) filtering image enhancement algorithm [20] is used to enhance the image captured in the complex environment and the driver's eye state and mouth action through visual positioning [21]. It can calculate the eye aspect ratio to describe the eye-opening degree and analyze the driver's fatigue state [22]. The PERCLOS value has the best correlation with fatigue driving, but it is mainly for the area of the rectangle where the eyes are located, but the size of the eyes is different, resulting in affected calculation accuracy. The methods of judging unsafe behaviors of drivers can be divided into two categories [23]: (1) optical flow field method. The optical flow method is used to detect the moving target and calculate the optical flow field and then extract the optical flow features, which can represent the movement from the optical flow field. (2) Using video sequence information to identify, due to the existence of a large number of mixed information under the spatiotemporal background in the video, the judgment ability of behavior expression is affected by environmental factors, and the misjudgment of driving behavior classification may occur. For ordinary image data processing, CNN is a better choice. For the data with video image sequence, RNN and LSTM are good choices, so LSTM is selected as the data classifier in this paper. Long short-term memory (LSTM) model is usually used for classification [24]. LSTM is an improvement in recurrent neural network (RNN). The bidirectional LSTM unit is used as the main framework to capture the bidirectional spatiotemporal characteristics of video sequences [25]. In order to meet the processing requirements of different length videos, a segmentation strategy is used to build a behavior recognition framework, and a LSTM network model is built based on the spatiotemporal attention of dual-stream features, which are used for human behavior recognition in videos [26]. In [27], a novel deep learning framework is proposed, which combines CNN with LSTM cell for real-time facial expression recognition (FER). In [28], a novel deep learning model called LCED which consists of one LSTM-based encoder, features image presentation, and one CNN-based decoder is proposed to weaken the accuracy differences among individuals on activity recognition. In [29], a deep learning solution called OSLCFit (organic simultaneous LSTM and CNN fit) is proposed, by using transfer learning to tune the specific task of polarity classification. The LSTM is adopted to create a model that analyzes skeleton sequence [30]. In [31], a method for process-focused assessment (PFA) is developed by learning facial expressions, using a deep neural network model. The model learns and classifies facial expressions into three categories. In [32], this study was aimed at developing a technique based on a CNN and LSTM (CNN-LSTM) model by using a deep learning approach to detect psychopaths.

Using the description based on spatiotemporal information to extract features from the driver's behavior, especially facial information, construct a feature vector, which can save the process of key frame alignment and effectively improve the image recognition rate. This kind of method improves the accuracy of face recognition but brings a certain boundary effect. When the face target is occluded or changed, the recognition speed and accuracy are affected, but at present, there are still the following deficiencies in driving behavior recognition. One is the generalization

ability of the recognition algorithm. The complexity of the physical scene reduces the definition of the video image and affects the accuracy of the driving behavior recognition algorithm. The second is the real-time problem of behavior recognition algorithm. There are two main reasons for the poor real-time performance of behavior recognition algorithm: one is that the behavior recognition algorithm cannot deal with the complex physical environment background, and the other is that the calculation of behavior feature extraction method is too large or too complex. Therefore, building a more lightweight network structure and reducing the amount of computation, the algorithm with high recognition rate has become one of the key issues of current researchers.

In this paper, this team has carried out the research on safe driving behavior recognition algorithm. The algorithm is mainly used to identify the driver's fatigue state and the main unsafe driving behavior in video image and give an alarm in time to reduce the occurrence of traffic accidents. Contributions include the following:

(1) Several deep learning algorithms and network structures are designed, and the CNN structure is improved. The convolution layer is introduced into the LSTM network structure to reduce the dimension of images. The subregion features of images are extracted by convolution layer of CNN, and then, the importance of each region is calculated. The important image feature description statement is generated, which can greatly reduce the redundant information of image sequence.

(2) In the network structure, the attention mechanism is integrated, and a dual flow channel AM-LSTM is proposed for behavior recognition and classification, which is divided into time flow and space flow. The spatial pyramid pooling (SPP) layer is adopted to scale the convolution graph, instead of traditional pooling. Different network structures are adopted, respectively, in the dual flow channel, and the time stream uses the SSD algorithm to calculate the optical flow image of two adjacent frames in the video sequence, aiming at the detection of small targets such as human eyes and mouth targets, maintaining the temporal characteristics of the behavior sequence. CNN is used to extract the spatial characteristics of RGB image at each t frame time. The recognition efficiency of the algorithm is improved.

(3) Three groups of experiments were carried out, including face recognition in the FDDB database, video face recognition in OTB100 data set, and safe driving behavior recognition in self-built driving image set.

2. Long and Short Term Memory Network

2.1. Improved LSTM Network Structure. In the face image recognition algorithm, the CNN avoids the complex preprocessing of the image and can directly process the original image, so it has been more widely used. The sparse connection and weight sharing of CNN can reduce the training parameters and the computational complexity and make the generalization ability of the model stronger, but CNN does not have the function of information memory. The current network output of a sequence of recurrent neural network (RNN) structure is not only related to the input but is formed by the interaction of the previous input and the current input. However, RNN can only remember the short distance information in the information sequence. The special structure of LSTM network [33] makes the network have the ability to memorize long-distance information. RNN neurons store effective information in uncontrollable form in each time step, while the LSTM network uses the special learning mechanism to integrate and update the information of the last time point, effectively avoiding the phenomenon of gradient explosion and gradient loss. Compared with RNN, LSTM has a state parameter to store the time-domain information. As the input is sent to the LSTM network one by one, the useful information will be filtered and stored in the state parameter. It can also be said that, in the whole training process of LSTM network, the state quantity always exists with the system and updates with time. The LSTM adds mechanisms such as memory unit c, input gate i, forgetting gate f, and output gate o. The structure diagram of LSTM at the t-th time is shown in Figure 1.

In order to better extract the key of spatial structure information, the LSTM structure is improved. The input of the current time is convoluted by a single layer and then combined with the short-term memory unit. At the t-th time, the network reads in the t-th input x_t and the state value h_{t-1} of the hidden layer at the previous time and calculates the state value h_t of the hidden layer at this time. Repeat this step until all the inputs are read. If the function represented by the RNN is denoted as f, the forgetting gate of LSTM with convolution calculation can be expressed as

$$f_t = \sigma\left(W_f * \left[h_{t-1}, x_t\right] + b_f\right). \tag{1}$$

The input gate is represented as

$$\begin{aligned} i_t &= \sigma\left(W_i * \left[h_{t-1}, x_t\right] + b_i\right), \\ \widetilde{C}_t &= \tan \mathrm{h}\left(W_c * \left[h_{t-1}, x_t\right] + b_c\right). \end{aligned} \tag{2}$$

The output gate is represented as

$$\begin{aligned} o_t &= \sigma\left(W_o * \left[h_{t-1}, x_t\right] + b_o\right), \\ C_t &= f_t \times C_{t-1} + i_t \times \widehat{C}_t, \\ h_t &= o_t \times \tanh\left(C_t\right), \end{aligned} \tag{3}$$

where * stands for convolution and W_f, W_i, and W_o stand for the gate weight matrix, respectively. The b_f, b_c, and b_o, stand for bias, respectively, and h_{t-1} stands for the state quantity at the last time. We use tanh nonlinear function because its output is in [−1, 1], which can adjust the mean value of input to 0. σ is the gate function, and the σ output is between 0 and 1, which can play the role of scaling. In order to select the key target information from the complex information, the attention mechanism [34, 35] is introduced into the LSTM network structure. Assuming that the target feature $\{x_i\}$ extracted by ResNet for each frame of video sequence is n_1,

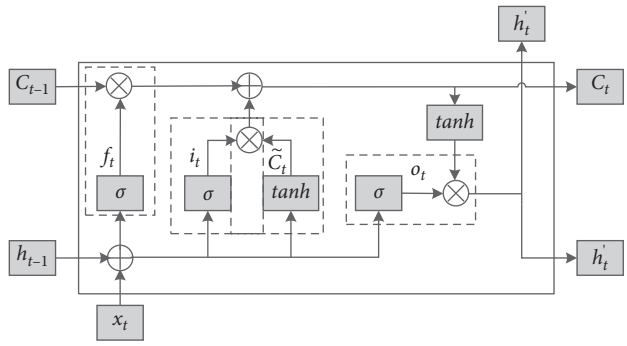

Figure 1: LSTM structure diagram at the t-th time.

the sum of all features dynamically weighted by the attention mechanism is as follows:

$$\varphi(X) = \sum_{i=1}^{n_1} \alpha_i^t x_i. \tag{4}$$

At the t-th time, the weight of the feature x_i is α_i^t, and then, the weight of the feature vector x_i is calculated as follows:

$$\begin{aligned} e_i^t &= f_{att}(x_i, h_{t-1}), \\ \alpha_i^t &= \frac{\exp\left(e_i^t\right)}{\sum_{k=1}^{n_1} \exp\left(e_k^t\right)}, \end{aligned} \tag{5}$$

where f_{att} means using the multilayer perceptron and h_{t-1} means the state of the last time. e_i^t is an intermediate variable, and k is the subscript of the eigenvector. After the weight is calculated, the attention mechanism starts to select the input sequence e_i^t and get the selected item sequence h_t'. The image description framework of attention mechanism is shown in Figure 2.

The image description of attention mechanism adopts the framework of the soft attention model. Firstly, $16 * 16$ image subregion features are extracted by convolution layer of CNN networks, and then, the importance of each region is calculated. Finally, the important image features are generated into image description statements. The feature $\{x_i\}$ attention mechanism of the video target at the t-th moment is expressed as follows:

$$\varphi(X) = \text{Attention}\left(\{x_t\}, h_{t-1}\right). \tag{6}$$

The structure of LSTM is further improved. The convolution operation of forgetting gate and output gate of LSTM is changed to full connection operation. Because the input x_t and the last time state short-term memory h_{t-1} are two-dimensional vector feature maps, they are transformed into one-dimensional vector by global maximum pooling, and the attention mechanism is added in the input gate i_t and output gate:

$$\begin{cases} f_t = \sigma\left(W_f * \left[\overline{H}_{t-1}, \overline{X}_t\right] + b_f\right), \\ i_t = \sigma\left(W_i * \left[\overline{H}_{t-1}, \overline{X}\right]_t + b_i\right), \\ o_t = \sigma\left(W_o * \left[\overline{H}_{t-1}, \overline{X}_t\right] + b_o\right), \end{cases} \tag{7}$$

where $\overline{H}_{t-1}, \overline{X}_t$ are the maximum pooling values of h_{t-1} and x_t. The structure of AM-LSTM with attention mechanism is shown in Figure 3.

2.2. Dual-Channel AM-LSTM Structure Algorithm. LSTM network has achieved excellent performance in most video timing recognition. In order to further improve the recognition accuracy, a hybrid dual-stream channel AM-LSTM structure algorithm in combination with attention mechanism is designed for safe driving behavior recognition. In order to distinguish the size targets in different image regions, the dual channel is divided into two branches: time stream and spatial stream convolution neural network. The detection algorithms of the two branches are slightly different. Based on the inception and VGG-16 network structure, the SSD algorithm is used to calculate the optical flow image of two adjacent frames in the video sequence, and then, the timing information is extracted from the stacked optical flow image of multiple frames, mainly for the detection of small targets such as human eyes. CNN is used to extract the spatial features of RGB image at each t frame time. Finally, the semantic information obtained from the two networks is fused, and then, AM-LSTM is used for recognition and classification [36]. In the recognition process of driving behavior video sequence, the dual-stream network has better recognition performance than the single stream network because the dual-stream network uses the image and optical flow field to represent temporal and spatial information, respectively. Although it increases a certain amount of calculation, it can reduce the noise of complex environment in the video sequence.

In recognition of driver's face and action, it is necessary to classify all kinds of actions. The actions of face, eyes, and mouth belong to small target detection, and the target area is small, but it is very important. Other actions such as using mobile phone and eating belong to a large target area. CNN algorithm uses different scale feature maps of different convolution layers for target detection, and SSD algorithm is an end-to-end image target detection method; the network of time flow channel SSD adopts the VGG-16 structure with the addition of inception network. From the beginning of image data input, this algorithm will directly obtain image features and classification information through network training and only need one stage to complete the target detection. The front feature map is used to detect small targets, and the back feature map is used to detect large targets; it can improve the recognition accuracy. Driver's face, especially eyes and mouth action recognition, is very important information, which belongs to the detection range of small targets. Due to light, occlusion, and other environmental factors, there is little feature information, resulting in misjudgment. In order to further improve the accuracy of recognition, this paper uses SPP (spatial pyramid pooling) instead of traditional pooling and unifies the scale transformation of convolution image so as to change the network output to a fixed scale. The dimension of the transformed image features is reduced, and then, the AM-LSTM network is used for spatiotemporal information

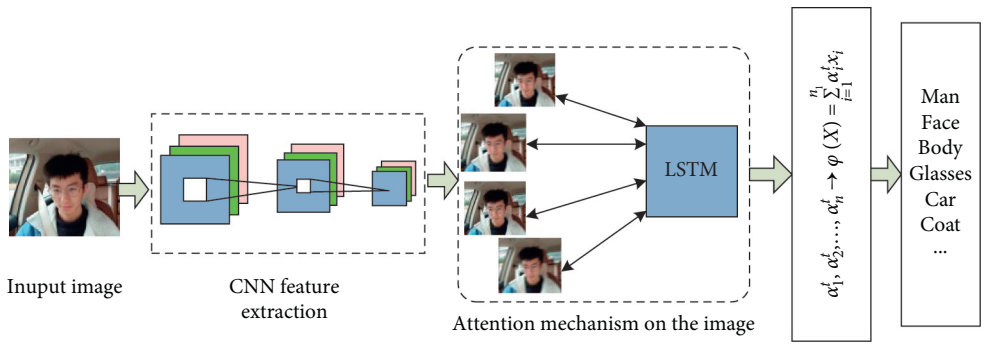

Figure 2: Image description framework of attention mechanism.

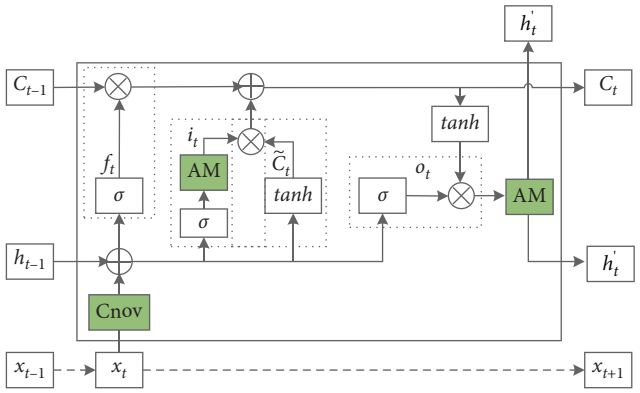

Figure 3: Network structure of improved AM-LSTM.

fusion, extraction, and recognition. Finally, softmax is used to classify various actions to realize multiscale information extraction of driving behavior features. The hybrid dual-stream AM-LSTM channel is shown in Figure 4.

The time step of AM-LSTM model is set to n, which is equal to the number of pixels in the local space rectangular window. The number of final output nodes and internal mapping nodes are set to l, which is the length of local space sequence feature matrix and the length of fusion low-level feature vector. By maximizing the output probability, the optimal output at the current time is expressed as

$$\begin{aligned}&\max \log\left(p\left(y_t \mid x_1, x_2, \ldots, x_t\right)\right)\\&\quad=\sum_{t=1}^{n_1} \max \log\left(p\left(y_t \mid y_{t-1}, \ldots, y_1, c\right)\right),\end{aligned} \tag{8}$$

where $p(y_t \mid y_{t-1}, \ldots, y_1, c)$ is the probability value of the current decoding output judged as semantic elements and the hidden layer state output of the last time step is regarded as the most representative and discriminative high-level semantic feature. Then, a softmax classification layer is connected after the AM-LSTM network for classification to get the classification label and complete the final classification task. Therefore, for the final time step, the formula can be modified as follows:

$$y^n = \text{soft max}\left(W_o h^n + b_o\right), \tag{9}$$

where w_o is the weight matrix from the hidden layer to the output layer, b_o is the deviation, and y^n is the final classification result. The loss function of this model adopts the cross-entropy function, and then, the model parameters are trained by backpropagation through time (BPTT). The cross entropy function can be expressed as

$$\begin{aligned}J(\theta) &= -\frac{1}{n}\sum_{k=1}^{n_1}\left[y_k \log\ p_\theta^k + (1-y_k)\log\left(1-p_\theta^k\right)\right]\\&\quad+\lambda\frac{\exp\left(W_{h,i}h_{t-1}+b_i\right)}{\sum_{i=1}^{n_1}\exp\left(W_{h,i}h_{t-1}+b_i\right)},\\p_\theta^k(x) &= \frac{1}{1+\exp\left(-\theta^T x\right)},\end{aligned} \tag{10}$$

where λ is the regularization parameter, θ is the training model parameter, p_θ^k indicates the probability that the eigenvalue belongs to class k, and y_k is a one hot code vector.

In this paper, the CNN consists of four layers: one-dimensional convolution layer and pooling layer. The size of convolution layer is 3, and the pooling layer adopts maximum pooling, with a pooling size of 2 and step size of 2. Starting from the first layer, the number of convolution cores in each layer is 32, 64, 128, and 256. The dropout layer is added at the output end of the fourth one-dimensional convolution layer, and the probability of the dropout layer is set to 0.5; that is, the input value is set to 0 with a probability of 0.5. The main network parameters include the following: the network momentum parameter can accelerate convergence; momentum is set to 0.85; weight attenuation coefficient is set to 0.001; the number of iterations is 500; the initial learning rate is 0.001; every 50 generations, the decrease is 0.0001; and the batch size is set to 50. The convolution process of CNN and SSD channels is shown in Table 1.

After CNN layer and SSD processing, a single frame image is fused, and then, the AM-LSTM layer is connected. The AM-LSTM network includes the forward layer and backward layer, and the number of hidden nodes is set to 256. The AM-LSTM network layer is followed by the full connection layer, and the category is output through softmax classification. The process is as follows.

First, for real-time video image acquisition, filtering, and noise reduction, the CNN network is used to extract the features of video image and constitute the feature set of video

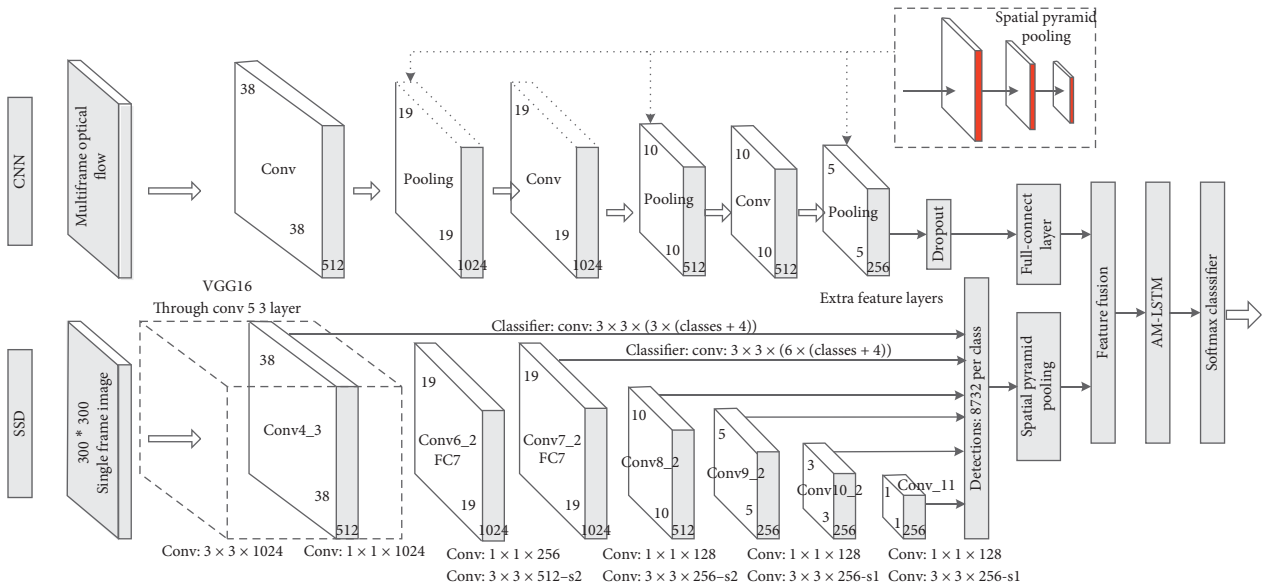

Figure 4: Hybrid dual-channel CNN + AM-LSTM network structure.

Table 1: Parameters of dual-stream convolution network.

SSD channel			CNN channel		
Convolution layer	Convolution kernel size	Size	Convolution layer	Convolution kernel size	Size
Conv4_3	3 × 3 × 1024	1	Conv	3 × 3 × 32	1
Conv6_2	1 × 1 × 1024	1	Pooling	2 × 2	2
Conv7_2	3 × 3 × 512	1	Conv	3 × 3 × 64	1
Conv8_2	1 × 1 × 512	1	Pooling	2 × 2	2
Conv9_2	3 × 3 × 256	1	Conv	3 × 3 × 128	1
Conv10_2	1 × 1 × 256	1	Pooling	2 × 2	2
Conv11_2	3 × 3 × 128	1	Conv	3 × 3 × 256	1
Dropout	—	—		√	1
SSP	2 × 2				
AM-LSTM	√	1		√	1

sequence; second, according to the window size, the AM-LSTM network with input and output attention mechanisms is constructed, and ReLU is selected as the activation function; third, the spatial-temporal flow characteristic sequences of training samples and corresponding standard class labels are input into the AM-LSTM network, and BPTT is used to train network parameters to obtain the final AM-LSTM network structure; last, the trained AM-LSTM network model is used to extract the local spatial-temporal series features of the test samples, and the classification results are obtained by softmax to obtain the judgment results of driving behavior in the video.

3. Identification of Unsafe Driving Behaviors

3.1. Global Feature Recognition. The difficulty in driving video image recognition lies in the uncertainty and diversity of action duration, as well as the complexity of background, angle, and environmental differences. Global feature recognition of driving video image mainly includes nonstandard body movements and face recognition, which are relatively large and easy to recognize. In this paper, based on the hybrid dual-stream channel detection technology, the dual-stream CNN + AM-LSTM algorithm is used to identify the video image. Generally, the video contains color RGB image information and motion optical flow information. One channel is to input RGB video frames as the carrier of spatial information into CNN and then to extract shape and other feature information. Another channel is to use optical flow information as the carrier of temporal information, which is called the time information network. SSD algorithm is used to extract features and sampling and tracking in continuous video frames, and the position of tracking point in the next frame is determined by optical flow to extract action information. Then, the data information is fused, and driving behaviors are identified by the AM-LSTM network.

3.2. Recognition of Local Features. The recognition of local features mainly refers to the actions of eyes and mouth. When people are in a state of fatigue, the eyes and mouth can be most reflected, such as yawning and squinting. Compared with the global features, this part only occupies a small part of the video image and combined with the differences of

hair, posture, artificial occlusion, and wearing glasses, so it is easy to cause misjudgment and missed judgment. In the whole traffic accident, fatigue driving accounted for 69%. In order to effectively recognize the driver's eyes and mouth action, the first step is to detect the face in the video image and narrow the detection range. We can regard the eyes and mouth action as small target detection, so this paper uses the SSD algorithm because the SSD algorithm only needs one stage to realize the detection, and its rapidity is better than the R-CNN algorithm. The common eye state recognition methods include Hough transform, human eye template, and statistical learning. In this paper, a simple method is designed to judge the degree of fatigue by detecting the opening and action frequency of human eyes and mouth. The detection points of human eyes are shown in Figure 5. For a kind of eye, there are eight detection points, and each point corresponds to a coordinate, which is expressed by (x_{Ei}, y_{Ei}). First, the distance between E_1 and E_5, E_2 and E_4, and E_6 and E_8 is calculated when opening eyes, and the average value is adopted as the normal driving standard value. Because each driver's state is different and the same vehicle may be driven by multiple drivers, only the basic model is provided. The parameters of the specific model need to be learned adaptively, according to the eye size of different people, environmental changes, and whether to wear glasses and other factors. E_{open} is used to express the eye opening.

Therefore, for a specific person, the i-th eye opening can be expressed as

$$\text{Eopen}(i) = \frac{\left\|E_1^i - E_5^i\right\| + \left\|E_2^i - E_4^i\right\| + \left\|E_6^i - E_8^i\right\|}{\left\|E_1 - E_5\right\| + \left\|E_2 - E_4\right\| + \left\|E_6 - E_8\right\|}. \tag{11}$$

In the formula, i is the i-th detection, which is used as the index of eye opening because the eye opening is bound to decrease when the driver is fatigue. The eyelid is easy to close when the driver is fatigue, and the cumulative closing time of the eyes increases. There will be a mechanism here. It can be judged by calculating the cumulative time of eye opening reduction. Considering the environmental factors such as light, the test result is that the opening is between 0.9 and 0.7, which indicates mild fatigue, between 0.7 and 0.5 is fatigue, and below 0.5 is severe fatigue. When people are fatigue, the blink frequency will become faster. As a parameter to judge driver fatigue, for the convenience of calculation, the time is converted into the number of video frames in the detection unit:

$$\text{FE}_{\text{open}}(i) = \frac{\text{frames of eye closure in 1 unit}}{\text{total frames of video sequence in 1 unit}} \times 100\%, \tag{12}$$

where FE_{open} as a threshold is used to judge the fatigue state. Generally, when a person is in a normal state, the FE_{open} value is less than 0.5. If the FE_{open} value is between 0.5 and 0.6, a person is in a mild fatigue state. If the FE_{open} value is between 0.6 and 0.8, a person is in a fatigue state. If the FE_{open} value is above 0.8, a person is in a severe fatigue state. The threshold is dynamic because each person's eye size is different and the absolute value of the opening is not the same. After the driver gets on the car, the threshold is obtained through AM-LSTM algorithm adaptive learning. In the monitoring process, the point information is constantly detected and analyzed. Then, the fatigue state is judged by it. The comparison of eye opening judgment is shown in Figure 6.

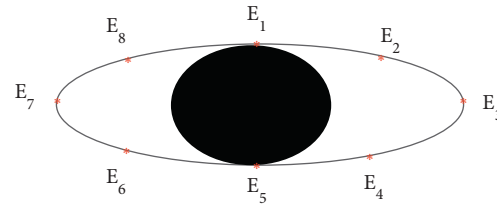

Figure 5: Schematic diagram of human eye detection.

Similarly, the fatigue state analysis of mouth movement is designed. The mouth is open, and there are eight detection points, as shown in Figure 7. Each point corresponds to a coordinate, which is expressed by (x_{Mi}, y_{Mi}). The principle is the same as the calculation of eye opening. H is the height of mouth opening, and L is the width of mouth M_{open} and is used to describe the degree of mouth opening. The schematic diagram of mouth is shown in Figure 7.

$$Mopen(i) = \frac{H}{L} = \frac{\left\|M_1 - M_3\right\|}{\left\|M_2 - M_4\right\|}. \tag{13}$$

Yawning is an indicator of fatigue, and the duration time of yawning is generally 4s-5s. Therefore, when it is detected that the driver's mouth is opened to the maximum and held at the maximum for a long time, it can be judged that the driver is yawning at this time. Generally speaking, the mouth opening threshold M_{open} is as follows: when the mouth is completely closed, the M_{open} value is 0; when the mouth is open, the M_{open} value is 0.7–1.0; when yawning, H can reach 40–60 mm, and the M_{open} value is 1.3–2.5. The larger the M_{open}, the greater the mouth opened. When the M_{open} exceeds the threshold and lasts for a certain time, it is judged as yawning. The duration time of mouth opening was $M_?$. According to $M_?$, yawning was divided into normal yawning and deep yawning. Two regular yawns or one deep yawn within one minute were mild fatigue, and three regular yawns or more than two deep yawns were severe fatigue. When $M_?$ is greater than 2 seconds and less than 4 seconds, it is normal yawn. When $M_?$ is greater than 4 seconds, it is deep yawn. Comparison of mouth movement is shown in Figure 8.

As shown in Figure 8, M_{open} of Figure 8(a) reached 1.0 and 0.1, respectively. M_{open} of Figure 8(b) reached 0.9 and 0, respectively. M_{open} of Figure 8(c) reached 0.2 and 0.5, respectively. From this, we can judge people's behavior state, and open mouth means speaking, eating, or yawning.

The VGG-16 network model is used to extract and classify them, and then, the video coding vector can be constructed according to the order of each frame. After obtaining the fatigue features of eyes and mouth based on feature point analysis, the SSD algorithm is used to divide

(a) (b) (c)

Figure 6: Comparison of eye judgment.

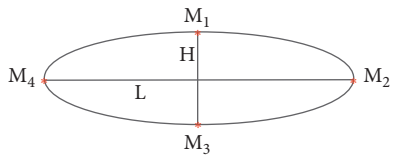

Figure 7: Schematic diagram of mouth.

(a) (b) (c)

Figure 8: Comparison of mouth movement.

the fatigue state into nonfatigue, light fatigue, and heavy fatigue according to these features. When there is a severe fatigue in unit time, it is considered as fatigue. When there is no severe fatigue, the extracted features are fused to judge fatigue. When there is fatigue in several consecutive units time, a warning message is produced.

4. Analysis of Experimental Results

This experiment is implemented on 4.0GHZ Intel CPU and GeForce GTX 1080 graphics card. The software adopts Python 3.0, Matlab2016a and Tensor flow1.3 framework. In the process of driving safety behavior recognition, the system uses near infrared camera with 850 nm wavelength to collect video sequence, the sampling rate is 30 fps, and the video size is 320×240. Three experiments are designed in this paper, which are face recognition in the FDDB database, video face recognition in the OTB100 data set, and safe driving behavior recognition in the self-built driving image set.

4.1. FDDB Database Experiment. In order to verify the effectiveness of this method, this method is compared with several other common face recognition methods, such as CNN, LRCN, and SSD. FDDB database is an unconstrained natural scene face detection data set, which contains 5171 faces in 2845 images taken from different natural scenes and faces. In the experiment, 2000 images were selected, of which 1500 were used for training and 500 for testing. In order to analyze the test data, the receiver operating characteristic (ROC) curve is drawn according to the test data. The ROC analysis is a binary classification model. According to the classification results and the area under curve (AUC), the larger the AUC area is, the better the method is. The experimental results are shown in Figure 9.

It can be seen from Figure 9 that the AM-LSTM structure integrated with attention mechanism has good performance, and the ROC curve in the FDDB database is significantly higher than that of other methods. The AUC area of the algorithm in this paper reaches 0.8658, which is 8.86% and 6.64% higher than CNN and SSD methods, respectively, and 1.41% higher than the dual-stream method of 0.8517. These all indicate that this method has strong competitiveness in face recognition.

4.2. OTB100 Data Set Experiment. In order to prove the effectiveness of attention mechanism and LSTM feature fusion, comparative experiments are conducted on different methods in OTB100 data set, using CNN, DDS, LRCN, and other methods, respectively. The results are shown in Figure 10.

It can be seen from Figure 10 that, in the video sequence, the accuracy and success rate of face target recognition of this method have reached 80.5% and 79.2% separately, which are 3.6% and 2.9% higher than that of CNN and SSD. The outcome proves the superiority of this method.

4.3. Experiment of Self-Built Driving Video Image Set. In order to verify the accuracy and practicability of the method proposed in the text, a self-built video image set was added in the experiment, as shown in Figure 11. A total of 10 situations were collected, including calling, eating, both hands leaving the steering wheel, talking, turning, squinting, yawning, normal, and other illegal actions. Three videos were collected in each situation, a total of 30 videos, numbered S1–S30. In the training, the initial learning rate is set to 0.001 and the ReLU is used as the activation function, and the stochastic gradient descent (SGD) method is used to optimize. The number of AM-LSTM nodes is 64. The training video data without secondary sparse optical flow extraction are input into the network with 2 consecutive frames, 4 consecutive frames, and 10 consecutive frames. Comparison and analysis with other methods on the self-built data set are as follows.

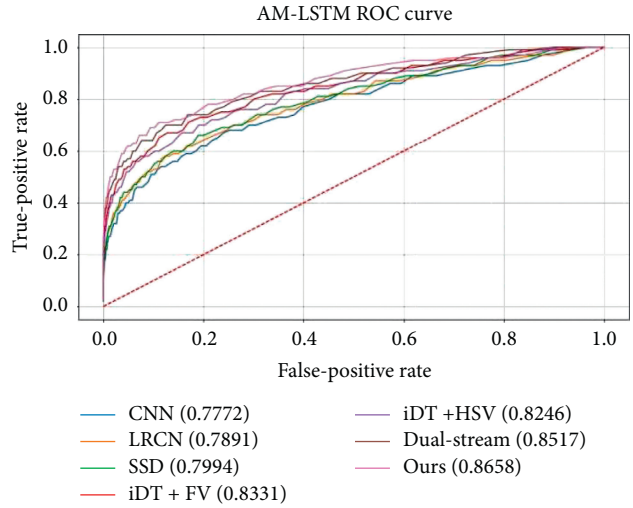

Figure 9: ROC curve of different methods.

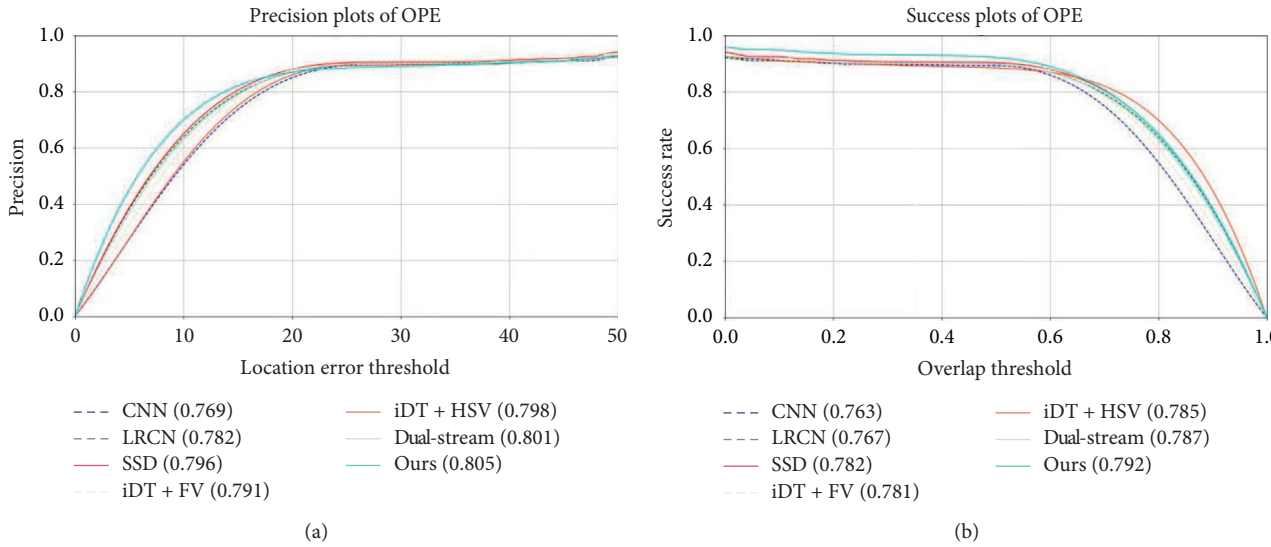

Figure 10: Comparison results of OTB100 with different methods: (a) accuracy rate; (b) success rate.

Figure 11: Self-built video image set.

Among them, the test design for eyes and mouth is to select 8 video sequences in 4 states for test, the number of samples is S16–S27, and the number of frames is 800. The results are shown in Table 2.

The AM-LSTM method is compared with the improved dense trajectories (IDTs), dual-stream video sequence recognition algorithm, SSD, CNN, and AM-LSTM algorithms. Five behaviors are tested in the self-built database, and the results are shown in Table 3.

It can be seen from Table 3 that, in order to improve the performance of video image recognition, this method also integrates the IDT trajectory features. Fisher vector (FV)

Table 2: Test results of eye and mouth movements.

Video number	FE_{open}	M_{open}	Recognition state	Normal state	Test sample number	Accuracy (%)
S16	0.35	—	Mild fatigue	Mild fatigue	763	78.38
S17	0.75	—	Fatigue	Fatigue	755	77.38
S18	0.91	—	Severe fatigue	Severe fatigue	762	78.25
S19	0.42	—	Mild fatigue	Mild fatigue	720	74.00
S20	0.55	—	Fatigue	Mild fatigue	726	74.75
S21	0.87	—	Severe fatigue	Severe fatigue	730	75.25
S22	—	0.52	Normal	Mild fatigue	765	79.63
S23	—	1.25	Fatigue	Fatigue	769	79.13
S24	—	2.12	Severe fatigue	Severe fatigue	763	80.33
S25	0.19	0.15	Normal	Normal	767	78.88
S26	0.26	0.23	Normal	Normal	766	78.75
S27	0.35	0.29	Mild fatigue	Normal	769	79.13

Table 3: Test results of different methods in self-built database.

Method	Network structure	Accuracy (%)					Parameter (MB)	Calculations (flops)
		Fatigue (%)	Calling (%)	Eating (%)	Turn (%)	Talk		
CNN	VGG-16	65.56	95.22	94.65	98.88	93.55%	1.185	1.963
LRCN	VGG-16	70.78	94.53	96.55	99.12	94.17%	1.177	2.101
SSD	Google	73.66	89.82	96.35	98.25	94.28%	1.165	1.995
iDT + FV	ResNet-50	76.56	94.76	95.82	98.81	93.55%	1.135	1.926
iDT + HSV	ResNet-50	77.53	88.71	95.52	98.18	93.32%	1.123	1.928
Dual-stream	VGG-16	78.22	93.65	94.8	98.22	92.52%	1.193	1.875
Ours	VGG-16+ Inception	80.33	97.35	96.22	98.30	95.52	1.056	1.736

method is used to encode, train, and classify the IDT features. Meanwhile, the network structure of VGG-16 and inception are adopted, which improves the network learning ability, so the recognition rate of this method is higher than that of other methods. The most difficult is to identify the fatigue state. Dual-stream method is only 78.22%, and iDT + HSV is 77.53%, while this method reaches 80.33%, higher than 2.11% and 2.8%, respectively.

In order to further verify the correctness of this model, based on the recognition results of AM-LSTM structure model, the ROC curves of eight driving behavior results are drawn, as shown in Figure 12.

It can be seen from the overall trend of the ROC curve of the eight driving behaviors in Figure 12. The LSTM structure of the hybrid dual-channel flow with pyramid pooling and attention mechanism has good classification performance; especially, the AUC value of calling, talking, eating, and so on has reached more than 81.3%. The recognition of dozing behavior in fatigue driving has also reached 68.2%.

In order to further verify the accuracy of this method, the recognition error of several methods on four different driving action sequences is analyzed, and the results are shown in Table 4.

It can be seen from Table 4 that the average absolute error of AM-LSTM method on four driving video sequences is between 9.213 and 12.313, which achieves the optimal result and has certain advantages over other methods.

The loss functions of CNN, LRCN, SSD, and AM-LSTM are analyzed in the self-built video image set. The first 75% of the data is used for training, and the second 25% is used for testing. The learning rate is set at 0.001, and the cross-

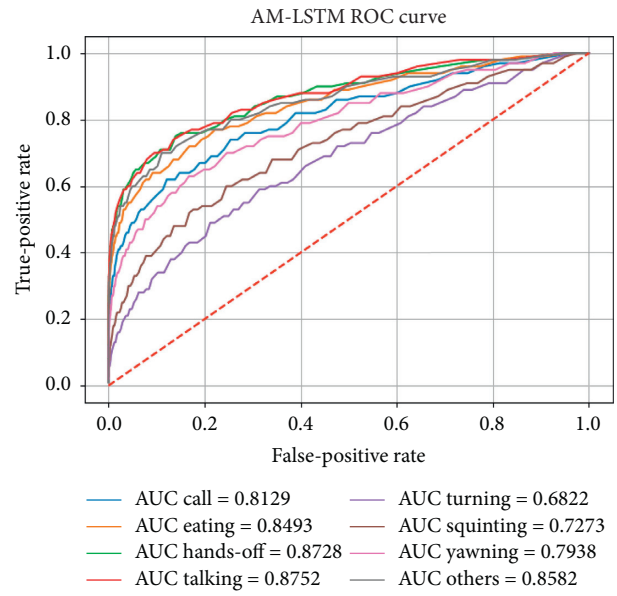

Figure 12: ROC curve of different methods.

entropy loss function is set for the loss function. The training model is trained by using the Adam linearizer, and 500 iterations are carried out. The training loss function curves of the four models are shown in Figure 13.

It can be observed from Figure 13, the loss functions of the four networks continue to decrease with the increase in the number of iterations. Among them, the CNN network loss function value decreases the slowest, and the final loss

Table 4: Average absolute error of different methods on different sequences.

Method	Sequence1	Sequence2	Sequence3	Sequence4
CNN	15.233	19.761	20.388	17.812
LRCN	13.212	12.891	16.456	17.123
SSD	11.245	11.378	14.612	13.752
iDT + FV	17.210	19.289	21.365	27.359
iDT + HSV	14.182	16.564	15.219	18.329
Dual-stream	12.576	14.253	17.498	16.367
Ours	9.213	11.012	10.255	12.213

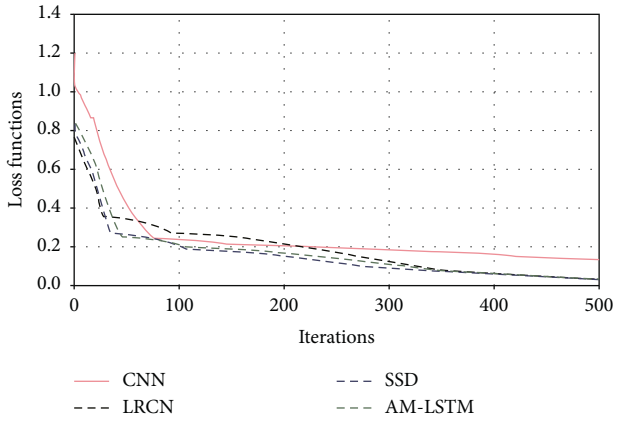

Figure 13: Loss function curves of the four models training.

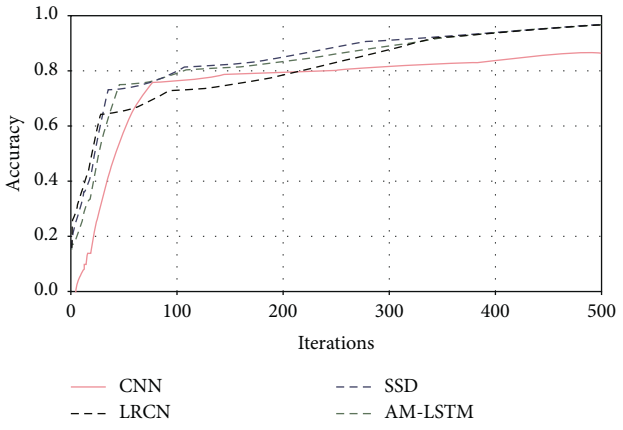

Figure 14: Accuracy curve of four models.

value is the largest. In the initial stage, SSD converges faster than the AM-LSTM network. This is because the AM-LSTM network has a forward layer and backward layer, and the initial network training is relatively slow. But after a period of iteration, the SSD network convergence speed becomes slower, while the AM-LSTM network convergence speed is still very fast, and the final convergence loss function value is smaller than that of the SSD network. The loss function of AM-LSTM model is the smallest, and the detection performance of AM-LSTM network is better than SSD and CNN networks.

The accuracy curves of the four models are shown in Figure 14.

As can be seen from Figure 14, with the increase in the number of iterations, the accuracy rates of the four networks are increasing. In the initial stage, the accuracy of LRCN network is the fastest rising than that of AM-LSTM network. After a period of iteration, the accuracy of LRCN network increases slowly, and the accuracy of AM-LSTM network is still improving. Finally, the AM-LSTM network has the highest accuracy.

5. Conclusion

Fatigue driving and driver's illegal behavior are the main sources of traffic accidents. In this paper, an improved hybrid dual-channel CNN network is proposed for driving safety behavior recognition by using artificial intelligence technology and deep learning theory. This algorithm is based on the fusion of CNN and LSTM network structures and integrating attention mechanism. Based on VGG-16 and inception network structure, pyramid pooling is used instead of traditional pooling layer in the network to unify image size, and the attention mechanism is introduced into the LSTM network to select effectively key information. In the dual-channel design, one channel uses the CNN algorithm to extract the spatial features of RGB video image, and the other channel uses the SSD algorithm to calculate the optical flow image of two adjacent frames in the video sequence, which is used to detect small targets such as eyes and mouth. After dual-channel information fusion, it is classified after AM-LSTM recognition, and the recognition results are obtained. Three types of experiments were carried out in FDDB, VOT100 database, and self-built driving video image set, and the ROC curves and curves of different methods and different driving behaviors were obtained. The results show that the method has good advantages and can effectively improve the recognition rate of fatigue driving and illegal driving behavior.

6. Future Prospects

Fatigue driving is an important cause of road traffic accidents, and unsafe driving behavior recognition is the preliminary work of driving fatigue detection based on vision, which aims at real-time detection and real-time detection and recognition of unsafe behaviors and provides data basis and warning for driving fatigue discrimination. The unsafe driving behavior recognition algorithm proposed in this paper has a certain reference value. The deep learning algorithm and network structure designed in this paper need to be further improved, such as target detection in the infrared scene, and the phenomenon of missing detection and inaccurate positioning exists. When the driver wears glasses, it produces high brightness area, which affects the accuracy of recognition. When the light is too strong or the head moves greatly, there will be misjudgment.

Although the algorithm integrates multiple parameters, increases the complexity of the model to a certain extent, and reduces the real-time performance, the recognition accuracy is improved. In practical applications, driving safety behavior recognition is essentially the real-time recognition of

various behaviors by online classifiers. The performance of classifiers directly determines the recognition performance of video targets. Although the AM-LSTM recognition and classification method designed in this paper has a large amount of calculation, it can solve this problem by using the advantages of ARM and FPGA in the machine vision development platform of NVIDIA GPU. Through this embedded platform, the safe driving behavior recognition method proposed in this paper can be applied to the actual scene.

Data Availability

At present, most of the research data come from the network public data sets, and some self-built data sets are used for experiments, which cannot be made public for the time being.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Jiangsu Natural Science Foundation of China (Project no. BK20191225) and the second batch of production-university-research cooperation bases in Suzhou Higher Vocational College in 2020 (Project no. 2020-5).

References

[1] L. X. Zhang, T. Liu, F. Q. Pan et al., "Analysis on the influence of driver factors on road traffic accident indexes," *Chinese Journal of Safety Sciences*, vol. 24, no. 5, pp. 79–84, 2014.

[2] B. Y. Zhang, *Research on Driving Safety Assessment Method based on Deep Learning*, Xi'an University of Science and Technology, Xi'an, China, 2019.

[3] L. Wang and R. S. Sun, "Analysis on flight fatigue risk and the systematic solution," in *Proceedings of the International Conference on Ergonomics & Health Aspects of Work with Computers*, Orlando, FL, USA, July 2011.

[4] C. Bila, F. Sivrikaya, M. A. Khan et al., "Vehicles of the future: a survey of research on safety issues," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 5, pp. 1046–1065, 2017.

[5] X. Q. Gong, Y. F. Pu, Z. Y. Yang et al., "PERCLOS algorithm for human eye opening based on finite state automata," *Computer Application Research*, vol. 31, no. 1, pp. 1046–1065, 2014.

[6] Q. Cai, Y. B. Deng, H. S. Li et al., "Review of human behavior recognition methods based on deep learning," *Computer Science*, vol. 47, no. 4, pp. 85–93, 2020.

[7] M. I. C. Murguia, R. P. Pablo, and G. R. Alonso, "A fuzzy clustering approach for face recognition based on face feature lines and eigenvectors," *Engineering Letters*, vol. 15, no. 1, pp. 35–44, 2007.

[8] Y. Gao, C. H. Zhou, and F. Z. Su, "Study on SVM calssifications with multi-features of OLI images," *Engineering of Surveying & Mapping*, vol. 47, no. 11, pp. 3084–3086, 2014.

[9] W. H. Tian, K. M. Zeng, Z. Q. Mo et al., "Driver unsafe behavior recognition based on convolutional neural network," *Journal of University of Electronic Science and Technology*, vol. 48, no. 3, pp. 381–387, 2019.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 8, pp. 84–90, 2017.

[11] A. Alahmadi, M. Hussain, H. A. Aboalsam et al., "PCAPooL: unsupervised feature learning for face recognition using PCA, LBP, and pyramid pooling," *Pattern Analysis and Applications*, vol. 23, no. 7, pp. 673–682, 2019.

[12] H. Liu, T. Xu, X. D. Wang et al., *Related HOG Features for Human Detection using Cascaded Adaboost and SVM Classifiers*, Springer, Berlin, Germany, 2013.

[13] Y. Wang, X. J. Shen, and H. P. Chen, "Multi instance learning video face recognition algorithm based on improved fisher criterion," *Acta Automatica Sinica*, vol. 44, no. 12, pp. 69–77, 2018.

[14] A. M. Hormat, K. Faez, Z. Shokoohi et al., "The new method of Extraction and Analysis of Non-linear Features for face recognition," *Journal of Electrical and Computer Engineering*, vol. 2, no. 6, pp. 766–773, 2012.

[15] H. Tan, B. Yang, and Z. Ma, "Face recognition based on the fusion of global and local HOG features if face images," *Compute Vision*, vol. 8, no. 3, pp. 224–234, 2014.

[16] W. J. Li, J. Wang, Z. H. Huang et al., "LBP-like feature based on Gabor wavelets for face recognition," *International Journal of Wavelets, Multi Resolution and Information Processing*, vol. 15, no. 5, Article ID 1750049, 2017.

[17] N. K. Kurian and D. Rishikesh, "Real time based driver's safeguard system by analyzing human physiological signals," *International Journal of Engineering Trends and Technology*, vol. 1, no. 1, pp. 41–45, 2013.

[18] B. Mandal, L. Li, G. S. Wang et al., "Towards detection of bus driver fatigue based on robust visual analysis of eye state," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 3, pp. 1–13, 2016.

[19] Z. J. Li, S. B. Li, R. J. Li et al., "Online detection of driver fatigue using steering wheel angles for real driving conditions," *Sensors*, vol. 17, no. 3, pp. 495–507, 2017.

[20] L. F. Liu, S. Q. Wu, and W. M. Xu, "Real-time fatigue driving detection based on analysis of facial landmarks," *Television Technology*, vol. 42, no. 12, pp. 27–30, 2018.

[21] J. J. Li, H. M. Yang, S. Y. Zhang et al., "Identification of driver's violation behavior based on neural network fusion," *Computer Applications and Software*, vol. 35, no. 12, pp. 222–228, 2018.

[22] M. F. Yang, K. Ren, and Z. Y. Zhao, "Studies on stress concentration and fatigue damage for ferromagnetic material based on permeability testing technology," in *Proceedings of the International Workshop on Materials Science and Engineering*, pp. 585–592, Kunming, China, October 2017.

[23] F. Xie, R. J. Wang, S. B. Shen et al., "Driving behavior recognition algorithm based on mobile inertial sensor and multi feature CNN," *Chinese Journal of inertial technology*, vol. 27, no. 3, pp. 288–294, 2019.

[24] M. Jain, AV. Subramanyam, S. Denman et al., "LSTM guided ensemble correlation filter tracking with appearance model pool," *Computer Vision and Image Understanding*, vol. 195, Article ID 102935, 2020.

[25] R. Ge, Z. H. Wang, and X. Xu, "Action recognition with hierarchical convolutional neural networks features and bi-directional long short-term memory model," *Control Theory & Applications*, vol. 34, no. 6, pp. 790–796, 2017.

[26] Z. Xie, Y. Zhou, K. W. Wu et al., "Activity recognition based on spatial-temporal attention LSTM," *Chinese Journal of Computers*, vol. 42, no. 130, pp. 1–16, 2019.

[27] S. Rajan, C. Poongodi, S. Devaraj et al., "A novel deep learning Model for facial expression recognition based on Maximum boosted CNN and LSTM," *IET Image Processing*, vol. 14, no. 7, 2020.

[28] A. Lg, Z. B. Hang, W. B. Chao et al., "Towards CSI-based diversity activity recognition via LSTM-CNN encoder-decoder neural network-ScienceDirect," *Neurocomputing*, vol. 11, 2020 in Press.

[29] R. Kiran, P. Kumar, and B. Bhasker, "OSLCFit (organic simultaneous LSTM and CNN Fit): a novel deep learning based solution for sentiment polarity classification of reviews," *Expert Systems with Applications*, vol. 157, Article ID 113488, 2020.

[30] R. Cui, A. Zhu, G. Hua, H. Yin, and H. Liu, "Multisource learning for skeleton-based action recognition using deep LSTM and CNN," *Journal of Electronic Imaging*, vol. 27, no. 4, p. 1, 2018.

[31] H.-J. Lee and D. Lee, "Study of process-focused assessment using an algorithm for facial expression recognition based on a deep neural network model," *Electronics*, vol. 10, no. 1, p. 54, 2020.

[32] F. M. Alotaibi, M. Z. Asghar, and S. Ahmad, "A hybrid CNN-LSTM Model for psychopathic class detection from tweeter users," *Cognitive Computation*, pp. 1–15, 2021.

[33] F. Xiao, X. Gong, Y. Zhang, Y. Shen, J. Li, and X. Gao, "DAA: dual LSTMs with adaptive attention for image captioning," *Neurocomputing*, vol. 364, pp. 322–329, 2019.

[34] L. Wang, J. P. Wang, P. Wang et al., "Research on twin network target tracking algorithm integrating attention mechanism," *Computer Engineering and Design*, vol. 11, 2020, https://kns.cnki.net/kcms/detail/11.2127.TP.20201030.1937.013.html.

[35] J. M. Liu, Y. Q. Su, and P. P. Wei, "Video EEG interactive collaborative emotion recognition based on long short memory and information attention," *Acta Automatica Sinica*, vol. 46, no. 10, pp. 137–147, 2020.

[36] M. Zhang, J. Li, Y. Li, and R. Xu, "Deep learning for short-term voltage stability assessment of power systems," *IEEE Access*, vol. 9, pp. 29711–29718, 2021.