Driver Yawning Detection Based on Subtle Facial Action Recognition

Hao Yang, Li Liu, Weidong Min, *Member, IEEE*, *Xiaosong Yang*, *Xin Xiong*

***Abstract*—Various investigations show that driver fatigue is one of the main causes of traffic accidents. Yawning is a typical sign of fatigue. Due to complicated facial actions and expressions of drivers in real driving environment, it is difficult to detect yawning accurately and robustly, especially when there are some facial actions and expressions with the same deformation of mouth as yawning. To alleviate these problems, a novel approach to detect yawning based on subtle facial action recognition is proposed in this paper. A 3D deep learning network with Low Time Sampling rate (3D-LTS) is proposed for subtle facial action recognition, which uses 3D convolution network as base network for spatial and temporal feature extraction and adopts softmax for classification. A keyframe selection algorithm is designed to make the above method more effective. It rapidly eliminates redundant frames using simple calculation of image histogram and meanwhile effectively reject outliers using Median Absolute Deviation. A series of experiments have been conducted on both our self-collected dataset and YawDDR dataset built based on the standard YawDD dataset. Compared with several state-of-the-art methods, the experiment results showed that the proposed method performs better than the existing methods, not only effectively distinguishing yawning from similar facial actions, but also robustly detecting yawning in various external environments.**

***Index Terms*—driver fatigue; yawning detection; keyframe selection; subtle facial action; 3D deep learning network.**

I. Introduction

Intelligent driving is a hot topic in recent years. Providing early-warning signals, monitoring and assisting vehicle control are main research topics in intelligent driving [1]. Intelligent driving is important to improve road safety. Road safety is a major concern all over the world. Thousands of people are killed, or seriously injured due to drivers falling asleep at the wheels each year. Road safety is seriously threatened by driver fatigue. An investigation was conducted by the National Highway Traffic Safety Administration in the USA, and the results showed that more than 1-in-3 respondents confessed to having experienced fatigue when driving. Among those who had drowsy driving incidents, 10% confessed that they had such incidents during the past month and the past year. A study of driving in a natural state found that driver fatigue led to 22% of traffic accidents. Without any warning, driver fatigue leads six times more likely to collide or approach a collision than normal driving. Therefore, research on methods to recognize driver fatigue is important to improve road safety. Over the past decades, many driver fatigue detection methods [2]-[5] are proposed to assist drivers to drive safely and improve traffic safety. The behavioral characteristics of drivers in fatigue driving include blinking, nodding, eyes closing and yawning. In these behaviors, yawning is one of the main forms of fatigue manifestations [6]. Therefore, due to the important role of yawning in fatigue driving detection, researchers have done a lot of research on yawning detection. Compared with traditional action recognition, the facial actions are subtle and can be regarded as subtle actions.

Although many researchers have proposed different methods [6]-[9] to detect yawning, the existing yawning detection methods still face great challenges. Due to complicated facial actions and expressions of drivers in real driving environment, it is difficult to detect yawning accurately and robustly, especially when there are some facial actions and expressions with the same deformation of mouth as yawning. To alleviate these problems, a novel approach to detect yawning based on subtle facial action recognition is proposed in this paper.

The main contributions of the paper are as follows:

1) This paper proposes a new keyframe selection algorithm based on histogram similarity and outlier detection, which can effectively eliminate redundant frames and lead to a huge improvement in subtle facial action recognition.

2) A 3D deep learning network with Low Time Sampling rate (3D-LTS) is proposed to effectively classify facial subtle actions, such as yawning, singing and talking.

The rest of the paper is organized as follows. Section II gives a review of the related works. Section III is an overview of our proposed method in the whole detection framework. Section IV describes the face detection method and discusses our proposed keyframe selection algorithm. Section V discusses yawning detection in various situations using 3D-LTS network we

This work was supported by the National Natural Science Foundation of China under Grant 61762061 and Grant 61603256and in part by the Natural Science Foundation of Jiangxi Province, China under Grant 20161ACB20004. *(Corresponding author: Weidong Min.)*

Author Contributions: Hao Yang and Li Liu contributed equally.

H. Yang is with School of Information Engineering, Nanchang University, 999 Xuefu Avenue, Honggutan New District, Nanchang 330031 China. (e-mail: 15079072351@163.com).

L. Liu is with School of Information Engineering, Nanchang University, 999 Xuefu Avenue, Honggutan New District, Nanchang 330031 China. (e-mail: liuli_033@163.com).

W. Min is with School of Software, Nanchang University, 235 Nanjingdong Road, Qingshanhu District, Nanchang 330047 China. (e-mail: minweidong@ncu.edu.cn).

X. Yang is with the National Centre for Computer Animation, Bournemouth University, UK. (e-mail: xyang@bournemouth.ac.uk).

X. Xiong is with School of Information Engineering, Nanchang University, 999 Xuefu Avenue, Honggutan New District, Nanchang 330031 China. (e-mail:15070017693@163.com).

proposed in this paper, followed by the experiment results in Section VI. This paper is concluded in Section VII.

II. Related Work

The existing yawning detection methods can be roughly clas sified into the following three types, the Appearance-based Me thods, the Handcraft Feature-based Methods and the Deep Lear ning-based Methods.

A. The Appearance-based Methods

Yawning detection methods based on appearance mainly rely on the extraction of geometric and color features. Some researchers used the histogram features of the mouth area to detect yawning, but this method had obvious shortcomings such as low accuracy and poor robustness. Petropoulakis L *et al*. [7] proposed a yawning detection method that detected yawning by the variation of aspect ratio of mouth profile. They carried out series of experiments to find the critical value of aspect ratio. When the aspect ratio of mouth is smaller than this value, the state of mouth can be classified as yawning. Nasrollahi K *et al*. [8] used facial feature points to detect yawning. Vertical distance between the midpoints of the upper and lower lips is used as a yawning indicator. The threshold distance is determined by series of experiments on different testers' images. This method relied on the detection of facial feature points. When the tester's head rotated more than a certain angle, the facial feature points will become difficult to detect. Appearance-based methods have the advantage of easy calculation, but the calculation of threshold is very sensitive to external factors and has poor anti-noise capability. Another limitation is that these methods are confined to where the cameras are installed during data collection. Besides, the low resolution and jitter of the camera may result in missed or false detection.

B. The Handcraft Feature-based Methods

To alleviate the inherent problems of Appearance-based methods, more and more researchers have begun to use Handcraft Feature-based methods [9]-[11]. Compared with Appearance-based methods, the handcraft multi-level features can make detection algorithm robust to various environments and the accuracy has been greatly improved. The algorithm proposed in reference [12] evaluated facial representation based on statistical local features. AdaBoost algorithm was adopted to learn the most discriminating fatigue facial LBP feature from a large LBP features pool, and SVM was used as the classifier. Their experiments obtained excellent accuracy in most situations, but the algorithm they proposed was not robust to various light environments. Anitha C *et al*. [13] presented a method that used the histogram values taken from the vertical projection of the lower part of the face. The presence of yawning would be indicated by a black blob in the mouth region of the binary image. This method did not use any classifier that reduced the computational time and complexity, but there may be multiple blobs present in the image which may be due to the presence of non-skin like regions around the driver's face. This method placed high demands on accurately positioning the mouth area. Ding W. Y. *et al*. [14] developed a yawning detection framework based on mouth inner contour corner detection and curve fitting of those corner points. They established a two-stage mathematical discrimination model of the mouth inner contour to detect yawning. Akrout B *et al*. [15] proposed a method that rested on the study of the spatiotemporal descriptors of a non-stationary and non-linear signal. They used the strength of the signal to determine if testers were in yawning state. Du Y *et al*. [16] first introduced the image sequence into yawning detection. They extracted the features in a sequence of images and 7 features including the height of the lip gap, the width of a mouth, grey level co-occurrence matrix, *etc*. were extracted from each image. They evaluated features and selected the useful subset of candidate features based on kernelized fuzzy rough set technique. This method got the best performance in traditional methods. But for humans, the mouth deformation of many facial actions is similar with yawning such as shouting and singing, and these actions may be incorrectly classified as yawning. Simply using multi-level features cannot effectively solve this problem.

C. The Deep Learning-based Methods

In recent years, with the development of deep learning methods, more and more deep learning frameworks are designed for image classification such as GoogleNet [17], ResNet [18] and object detection such as Faster-RCNN [19], SSD [20]. Some researchers attempt to use Deep Learning-based methods to solve the problem of yawning detection [21]-[25]. A progressive yawning detection framework [21] has been proposed, which combined the network features with traditional features to improve the detection accuracy and reduce the time cost of extraction of handcraft features. Ma Sugang *et al*. [25] proposed the use of 2D convolution network to extract facial expression features for classification. They input one face image directly into the network. This method can improve the classification accuracy obviously, but it has also ignored the temporal information and made false detection when classifying some expressions which are similar with yawning. Unlike simple static image classification, yawning is a continuous action. Simply using one frame image for classification lose the important temporal information. These existing Deep Learning-based methods performed well in classification but to a certain extent, their robustness and reliability were poor compared with traditional methods.

D. Keyframe selection in video processing

In the action recognition field, keyframes are frames that can effectively represent the action characteristics in a video segment. The selection of keyframes can greatly improve the speed and accuracy of action recognition. The algorithm proposed in reference [26] extracted pyramid motion feature (PMF) of each frame, and then the keyframes are selected by

the AdaBoost algorithm according to the extracted pyramid feature. N. Azouji *et al* [27] used benchmarking tool proposed by Lux *et al*. [28] to select keyframes. In this tool, keyframes are selected by a clustering algorithm. In the existing keyframe selection algorithms, low-level features such as color histograms and texture features are used to select keyframes. These algorithms extract the low-level features of each frame and use the classifier to select keyframes. For these algorithms, only using single low-level feature may cause poor selection, while using classifier increases the computational complexity. The benchmarking tool proposed by Lux *et al* [28] considered different combinations of low-level features. However, the use of clustering for keyframe selection still has the disadvantage of complicated calculation.

From the previous analysis, we know that Appearance-based methods have the disadvantage of being interfered by the external environment, and the Handcraft Feature-based methods are easy to be influenced by the factors that interfere with the appearance of the face region, such as the brightness of light. These methods also have a high demand on the accuracy of face detection algorithm. With the application of deep learning in yawning detection, the problems in Appearance-based methods and Handcraft Feature-based methods had been greatly improved. However, the existing Deep Learning-based methods for yawning detection are still based on one frame image for feature extraction and classification. The use of one frame image leads to a problem that false detection may occur when classifying some facial actions with the same deformation of mouth as yawning. In general, the existing yawning detection methods use simple and rough features, and few of them consider using temporal features to recognize yawning. Temporal information is important to yawning detection, and the facial action features which incorporate temporal features can effectively reduce false detection of yawning in various situations. In order to extract temporal features effectively, we designed a new keyframe selection algorithm. Our algorithm uses the histogram similarity to select the candidate frames, and then rejects the outliers in the candidate frames. It has the advantages of simple calculation because no classifier is used meanwhile it can effectively select facial action keyframes.

III. Overview of Our Proposed Method For Yawning Detection

As shown in Fig. 1. Our proposed method for yawning detection consists of three parts. The first part is data preprocess module. This part can be divided into face detection and segmentation, image size normalization and remove noise. The second part is Keyframe Selection module which selects keyframes by calculating the image similarity between adjacent frames and removing the frames with outlier similarity. The third part is Facial Action Classification module. In this module, according to the keyframes selected by the second part, a 3D deep learning network with Low Time Sampling rate (3D-LTS) is built to detect various yawning behavior.

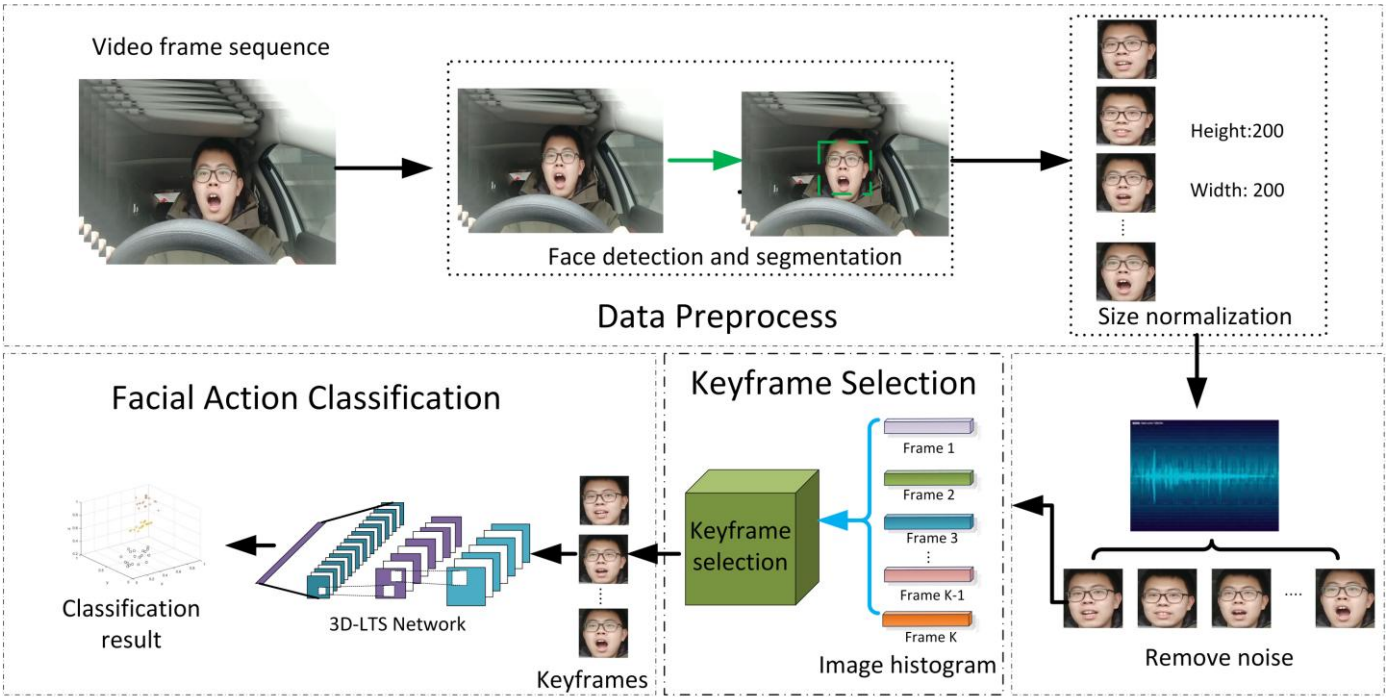

Fig. 1.The framework of our proposed yawning detection method

IV. KEYFRAME SELECTION ALGORITHM

In order to eliminate the influence of the irrelevant parts in video frames and improve the recognition rate, we need an efficient face detection algorithm. The existing face detection methods [29]-[30] an detect faces accurately and efficiently. Viola-Jones face detection algorithm [31] is a commonly used face detection algorithm, which was proposed in 2004. This face detection algorithm has the characteristics of being real-time and robust. Viola-Jones algorithm is a face detection algorithm based on Haar-like feature, Adaboost classifier and cascade classification strategy. We adopt the accelerating Viola-Jones algorithm [32] as the face detector, which is a multi-GPU implementation of the Viola-Jones algorithm. The accelerating Viola-Jones face detection algorithm is used to segment the face region from the video frame sequences. After segmentation, we uniform the size of the consecutive frames to 200×200.

In the preprocessing stage, videos are divided into frame sequences at 30 fps. To learn the spatiotemporal characteristics and effectively classify subtle facial actions from consecutive video frames, we need a keyframe selection algorithm. The algorithm must have the advantages of being real-time and high-efficiency and can select representative frames from original sequences which have little difference between two adjacent frames. In this paper, we proposed a new keyframe selection algorithm to select the keyframes that can effectively represent subtle actions. Fig.2. shows a keyframe sequence that is selected by our algorithm.

In our designed algorithm, histograms of video frames are computed to extract candidate keyframes. In order to eliminate the interference factors such as noise from the original frame sequence, we first use the Fast Median Filtering algorithm [33] to remove the noise in the original frame sequences. Fast Median Filtering is commonly used in image processing. It removes most of the noise information especially useful for speckle noise and salt-and-pepper noise. Algorithm1 provides a brief overview of Fast Median Filtering algorithm.

Algorithm 1 Fast Median Filtering

1. **Input:** Image X of size $m \times n$, kernel radius r
2. **Output:** Image Y of the same size as X

3. Initialize kernel histogram H
4. **for** i =1 to m **do**
5. **for** j =1 to n **do**
6. **for** $k = -r$ to r **do**
7. **Remove** $X_{j+k,j-r-1}$ from H
8. **Add** $X_{j+k,j+r}$ to H
9. **end for**
10. $Y_{i,j} \leftarrow median(H)$
11. **end for**
12. **end for**

Our approach is used to extract a set of keyframes $K = K_i, i = 1, \cdots M$ from a set of video frames $F = \{F_j, j = 1, \cdots N\}$, where M denotes the number of keyframes we selected from original frames and N denotes the number of original frames. We combined threshold-based histogram similarity filtering with outlier detection. Image histogram has the advantages of simple calculation. Compared to local features, global features such as image distances and histograms in classification can effectively reduce the false positives. Algorithm2 summarizes the proposed algorithm. First, we calculate the color histograms of each frame from videos to extract candidate keyframes; then these candidate keyframes are filtered base on the MAD of their RMSE and Euclidean Distance. In the first selection stage, RGB color histogram of each video frame is computed. Then, the similarity between the color histograms (γ_j and γ_{j+1}) of two consecutive frames is calculated using Euclidean Distance:

$$S_i = S(\gamma_j, \gamma_{j+1}) = \sqrt{\sum_{1}^{n} (\gamma_{j_n} - \gamma_{j+1_n})^2} \quad (1)$$

where n is the dimension of image color histogram. After calculation, we obtained a set S consisting of the similarities between F_j and F_{j+1}. We need to determine a similarity threshold T_s for the selection of keyframes, which can represent the average of the similarity between frames. We considered two threshold calculation methods, viz. the half of maximum and minimum similarity and average similarity. We used these two thresholds to select keyframes from our self-collected dataset and processed YawDD dataset. Results are shown in Table I, where s represents the similarity set. From the results we can see that using the average similarity as the metric threshold allows our yawning detection method to achieve the best overall results. The half of the maximum frame similarity and the minimum frame similarity fused two extreme similarities. This threshold cannot represent the average of similarities for these facial actions. Average similarity as threshold can select the most representative keyframes.

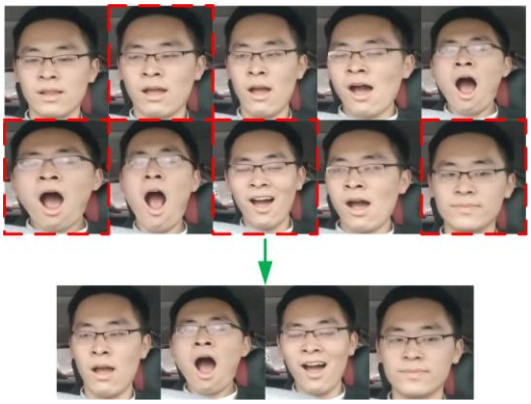

Fig. 2. Keyframes selected by our algorithm.

TABLE I
COMPARISON OF DIFFERENT KEYFRAME SELECTION THRESHOLD (%)

Threshold	Ave	Y	YT	T
(min(s)+max(s))/2	79.5	88.2	72.9	77.2
mean(s)	**81.7**	**90.8**	**76.4**	**80.4**

Algorithm 2 Proposed Keyframe Selection Algorithm

1: **Input:** RGBframes $F=\{F_j, j=1,\cdots N\}$

2: **Output:** Set of keyframes $K=\{K_i, i=1,\cdots M\}$

Stage One:

3: **for** $j=1$ to $N-1$ **do**

4: Read F_j, F_{j+1}

5: $\gamma_j = CalculateHistogram(F_j)$

6: $\gamma_{j+1} = ClaculateHistogram(F_{j+1})$

7: $S_j = S(\gamma_j, \gamma_{j+1}) = \sqrt{\sum_1^n (\gamma_{j_n} - \gamma_{j+1_n})^2}$

8: **end for**

9: $\mu_s = Mean(S)$

10: $T_s = \mu_s$

11: **for** $j=1$ to $N-1$ **do**

12: **if** $(S_j > T_s)$ **then**

13: $K \leftarrow K \cap \{F_j\}$

14: **end if**

15: **end for**

Stage Two:

16: **for** $i=1$ to $M-1$ **do**

17: $RMSE(i) = CalcuteRMSE(K_i, K_{i+1})$

18: $ED(i) = CalcuteED(K_i, K_{i+1})$

19: **end for**

20: $\alpha = MadRMSE(RMSE)$

21: $\beta = MadED(ED)$

22: **for** $i=1$ to $M-1$ **do**

23: **if** $(RMSE(K_{i,i+1}) < \alpha) \| (ED(K_{i,i+1}) < \beta)$ **then**

24: $K \leftarrow K - \{K_i\}$

25: **end if**

26: **end for**

We compute the similarity threshold T_s in (2) by calculating mean μ_s of S.

$$T_s = \mu_s \tag{2}$$

Then F_j is added to the set K as follows,

$$S_j > T_s \therefore K \leftarrow K \cap \{F_j\} \tag{3}$$

In the second selection stage, we filter these candidate frames by removing those similar with their immediate frames. To this end, we use two image distance: Euclidean Distance (ED) and Root Mean Square Error (RMSE) as the image distance. Outlier detection has been adopted to filter frames with outlier distance metrics. In our algorithm, we use Median Absolute Deviation (MAD) to detect outliers. The MAD of unary sequence X is calculated according to formula (4).

$$MAD = median(|X_i - median(X)|) \tag{4}$$

We denote two consecutive keyframes as $K_{i,i+1}$. Obtaining two similarity vectors RMSE and ED for all $K_{i,i+1}$, we compute the MAD for each vector denoted $\alpha = MAD(RMSE)$ and $\beta = MAD(ED)$ .The $RMSE(K_{i,i+1})$ is shown in formula (5). Pairs with $RMSE$ less than β or ED less than α , are considered as similar and the frame K_i is removed from candidate keyframes according to formula (6). α and β are used to reject outliers and MAD is a robust estimator of outliers.

$$RMSE(K_{i,i+1}) = \sqrt{\frac{1}{n}\sum (K_i - K_{i+1})^2} \tag{5}$$

Here, n represents the array size of K_i.

$$\begin{gathered}(RMSE(K_{i,i+1}) < \alpha) \vee (ED(K_{i,i+1}) < \beta) \\ \therefore \mathrm{K} \to \mathrm{K} - \{K_i\}\end{gathered} \tag{6}$$

V. The Proposed Subtle Action Recognition Network

A. 3D Convolutional Network

In our method, another important contribution is to introduce action recognition mechanism into yawning detection. In recent years, action recognition has got great progress both in accuracy and speed. Researchers have proposed various networks to recognize actions. The widely used action recognition frameworks are two-stream fusion networks [34]-[36] and 3D convolutional networks [37], [38].

3D convolutional network has attracted a lot of attention on action recognition [39], scene and target recognition [40] and action similarity analysis [41]. Compared with other spatiotemporal feature extraction methods based on two-stream networks, 3D convolutional networks possess the advantages of fast calculation and high accuracy. Some researchers attempted to superimpose the 2D convolved consecutive feature maps to classify video actions, but the temporal information has also been lost during 2D convolution. By contrast, 3D convolutional network uses multiple consecutive video frames as input, as shown in Fig.3. 3D convolutional network enables better modeling of temporal information through 3D convolution and 3D pooling operations. Experiments in paper [38] found that 3×3×3 size of 3D convolution kernel can extract the most representative spatiotemporal features.

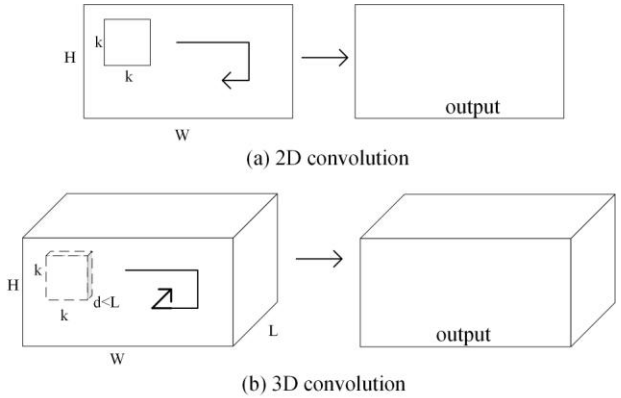

Fig. 3. Comparison of 2D convolution and 3D convolution

B. *Architecture of Our Proposed 3D-LTS Network*

A 3D network with Low Time Sampling rate (3D-LTS) is proposed for spatiotemporal feature extraction and recognition of subtle actions. Our 3D-LTS network uses 3D convolution for extraction of spatiotemporal features and softmax layer is adopted for classification. After data preprocess and keyframe selections, it is very important to decide how many consecutive frames to use as input to our 3D-LTS network in order to get the best recognition performance. We compared the results of 3D-LTS network with different input frame numbers. Our network is trained on self-collected dataset and tested on the processed YawDD dataset. The experiment results are presented in Table II and Fig. 4. From the results of the overall recognition, the results indicate that our 3D-LTS network is not highly sensitive to the number of input frames. Our network, which uses 8 non-overleap consecutive frames as input, exhibits better performance. 3D-LTS uses four3D convolution layers to extract spatiotemporal features from consecutive frames. The structure of 3D-LTS is shown in Fig. 5. From the structure diagram, we can see that all the convolution filters are $3\times3\times3$ with stride $1\times1\times1$. All pooling layers are max pooling. The deep convolution layers can extract more representative temporal features from shallow convolution layers if we slow down the pooling rate of the shallow pooling layer in time dimension. It is very important to the recognition of subtle actions. Based on this theoretical analysis, the first and second pooling layers in our 3D-LTShavekernel size of $1\times2\times2$. The number of filters for the first four convolution layers are 32,64, 128 and 256.The size of the third pooling kernel layers is $2\times4\times4$. The convolution layers are followed by a fully connected layer which is used to map features. We used one fully connected layer with 1024 outputs. The fully connected layer is used to integrate the distribution of features. We found our 3D-LTS got the best recognition performance when it is followed by one fully connected layer.

TABLE II
COMPARISON OF DIFFERENT INPUT FRAME NUMBER (%)

Frame Number	Ave	Y	YT	T
8	**81.7**	**90.8**	**76.4**	80.4
16	79.5	88.5	74.8	78.1
32	81.2	89.2	75.4	**82.5**
64	78.7	86.5	73.2	76.5

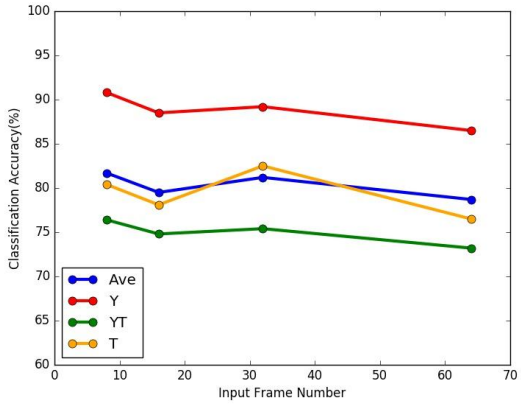

Fig .4. Accuracy of different input frame number

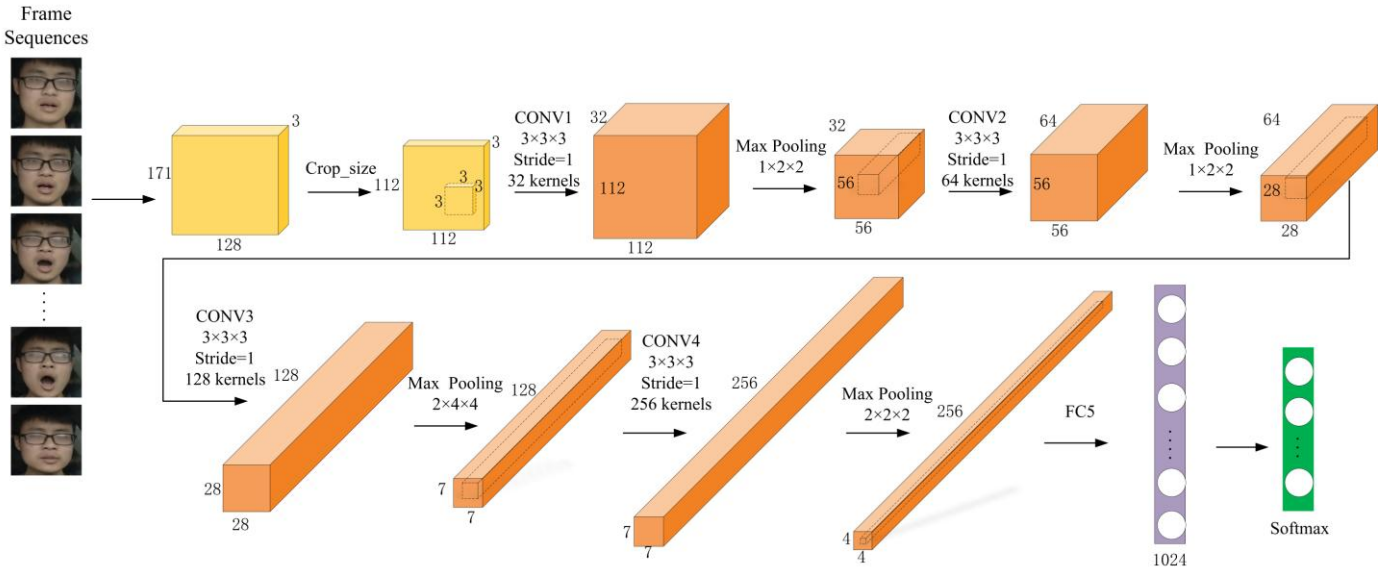

Fig. 5. The structure of our proposed 3D-LTS network

3D-LTS uses the softmax function, which is defined as: (5).

$$\mathrm{E}(d)=\frac{e^{d_j}}{\sum_{g} e^{d_j}} \quad (5)$$

Here, g represents the numbers of class, and d is function of g .Every d corresponds to a vector.

VI. EXPERIMENTS

The proposed method is implemented using Caffe deep learning framework [42]. All the experiments are conducted on a workstation with a 3.3GHz Intel(R) Xeon(R) E-2136 CPU, 16GB RAM, a NVIDIA QUDARO P5000 (16G) GPU, and Ubuntu 16.04. The algorithm was developed in Matlab2016b and VS2017, using OpenCV 3.2.0. The network parameters for training the 3D-LTS network model are as follows. The maximum number of iterations is 15000 and the initial learning rate is 0.001. We have conducted the following three parts experiments based on the datasets described in section IV.A.

1) In section IV.B, we demonstrated the effectiveness of our keyframe selection algorithm. Our network is trained on MFAY dataset and part of the YawDDR dataset, tested on YawDDR dataset.
2) We used our keyframe algorithm to select the video keyframes from the MFAY dataset and YawDDR dataset. Our network and other state-of-the-art methods were evaluated on these selected keyframes. This experiment is elaborated in section IV.C.
3) We have conducted some comparative experiments in section IV.D to compare image-based and video-based methods for yawning detection.

A. Datasets

1) YawDDR Dataset

YawDDR dataset is built based on standard YawDD dataset. YawDD is a public yawning detection dataset [43]. It can be used to verify face detection, face feature extraction, yawning detection and other algorithms. The dataset collected a series of action videos from volunteers of different genders, ages, countries and ethnicities. The dataset contains 351 videos. It recorded three or four videos for each driver, including different mouth conditions such as talking, yawning and yawning while talking.

Fig. 6.Some frame samples from YawDDR dataset

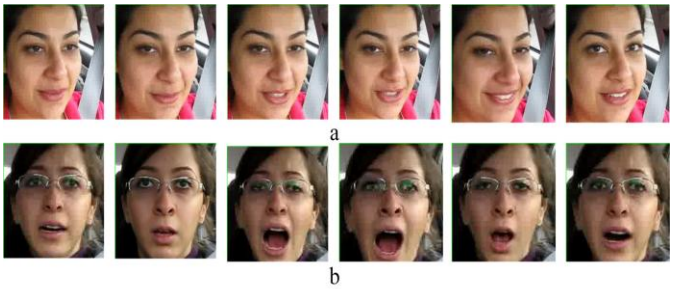

Fig.7. Some image sequences of two types of action in YawDDR dataset (a) Talking (b) Yawning

Since most of the video segments in the YawDD dataset last more than one minute, and a video segment contains more than one facial action, we need to divide the video segments in the YawDD dataset into video segments that contain only a single action. In this way we built our YawDDR dataset based on YawDD dataset. The length of videos in YawDDR dataset is about 8 seconds. There are three kinds of actions in this dataset: Talking (T), Yawning(Y) and Yawning while Talking (YT). As shown in Table III, 486 image sequences are collected in YawDDR. Some examples (before face segmentation and after segmentation) from the dataset are shown in Fig. 6 and Fig. 7. We used this dataset to verify the effectiveness of our method.

2) Multi-facial Action Yawning Dataset (MFAY)

Many facial datasets are used for identity recognition, facial expression recognition and face detection. However, no one public driver yawning detection video dataset includes various facial actions. The purpose of collecting this data set is to verify the efficiency of our method for driver yawning detection in various facial actions. Therefore, a dataset is built by using an HD charge-coupled device camera in an actual driving environment. We divided the various facial actions into six classes that may occur during driving. They are Talking (T); Yawning while Talking (YT); Yawning(Y); Singing(S); Yawning while Singing (YS); Shouting (ST). Given the danger of fatigue driving, our collection site was chosen on a wide road with few pedestrians. Without affecting driving, a mini HD camera is installed in front of driver to capture their facial actions. During the experiment, the driver drives the car under different illumination and road conditions. In the copilot of the vehicle, a researcher continuously monitors the facial action changes of each subject to annotate the ground truth of each facial action. The location of monitor and driver is shown in Fig. 9.

The facial videos of 20 subjects (with ages ranging from 20 to 46) are obtained in diverse scenarios when the car is in motion. Sample images of our dataset are presented in Fig. 8. All the videos are converted into audio-video interleave format, with the video rate of 30 fps. Finally, as shown in Fig. 10,347 image sequences (53,652 images) are extracted from the obtained videos. The length of each image sequence is about 5 s (150frames).

Fig. 8. Image sequences of three facial actions in our dataset (a) Yawning (b) Singing(c) Shouting

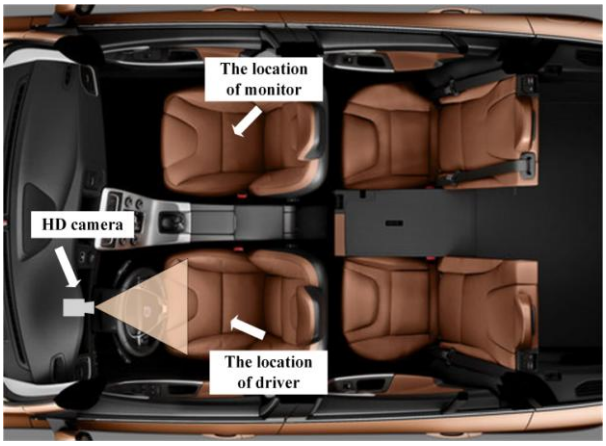

Fig. 9. HD camera installation position

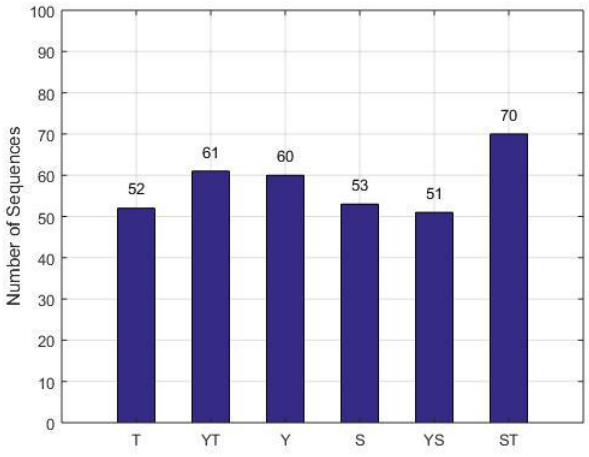

Fig.10. Number of video sequences in MFAY dataset

TABLE III
NUMBER OF SEQUENCES IN YAWDDR DATASET

Class	Y	YT	T
Male	92	67	79
Female	102	74	72

B. Results on Different Keyframe Selection Stage

To demonstrate that our keyframe selection algorithm can effectively select keyframes in the driving video frame sequence, we conducted the following experiment on YawDDR dataset and MFAY dataset. First, image histogram is used to remove the frames with little difference and select candidate keyframes, we record this operation as Stage1. In order to verify that our algorithm can effectively improve the recognition rate of various facial actions, we also provided recognition result, which did not use any keyframe selection algorithm. We called this case as Not Use. The results were shown in Table IV. After the satge1, the accuracy has been improved. On the basis of stage1, we use the MAD to reject the outliers in the candidate keyframes. After this processing, we leave the keyframes we need. From Table IV, our keyframe selection algorithm achieves the best recognition performance compared with Stage1 and Not Use. The validity of our algorithm for keyframe selection is verified.

TABLE IV
COMPARISON OF DIFFERENT KEYFRAME SELECTION STAGE (%)

Class	Ave	Y	YT	T
Not Use	72.2	80.3	64.5	69.2
Stage1	79.5	88.5	74.6	79.4
Proposed algorithm	**81.7**	**90.8**	**76.4**	**80.4**

C. Comparison with State-of-the-art Methods

In this set of experiments, we focus on comparative experiments between our proposed method and some other existing state-of-art image-based methods. We compared our method with the method based on kernelized fuzzy rough set proposed by Du Y *et al*. [16]; the algorithm based on two fold agent expert system proposed by Anitha C *et al.* [13] and two methods based on convolutional neural networks [21], [25].In order to verify the effectiveness of our method, we adopt the following model training and testing algorithm: the training set consists of the video clips which are randomly extracted from the MFAY dataset and the YawDDR dataset according to the category they belong to. The remaining video clips were used to test the model. All the video clips were processed by the keyframe selection algorithm we proposed. The selected keyframes were used to train and test the network model. As

demonstrated in Table V and Fig. 11, the recognition rate of our yawning detection method based on subtle facial actions and video keyframe has been substantially improved compared with other methods. Our method for the recognition of various facial actions outperformed the existing methods, effectively reducing the false detection. Video-based methods can effectively extract sufficient spatiotemporal action features and achieve dynamic yawning detection. This further validated the robustness of our proposed method.

D. Image-based Method Versus Video-based Method

In this section, we compared image-based and video-based methods. Our method uses consecutive frames as input, which is a video-based method. For image-based methods, the frame images in YawDDR dataset and our MAFY dataset are used for training and testing. We evenly extract some frames from two dataset and assign labels to them according to the class they belong to. The data processing step and validation algorithm for these experiments are the same. The experiment results are shown in Table VI. The results show that the video-based methods have better performance than image-based methods, because yawning is a continuous action instead of static state. Video-based methods can detect yawning in various facial situations. If only one frame is used for recognition, important temporal action information between frames will be lost. Features that indicate yawning may be confused with those that indicate the action singing or shouting. By contrast, video-based methods can provide sufficient spatiotemporal action information, which can classify an action by action frame sequences. Classifying yawning as an action rather than static state can significantly improve the false detection problem in image-based methods.

TABLE V
COMPARISON OF RESULTS BETWEEN THE PROPOSED METHOD AND FOUR STATES-OF-THE-ART METHODS ON YAWDDR DATASET (%)

Methods	Ave	Y	YT	T
KFRS [16]	-	87.4	-	**82.2**
TFES [13]	-	83.3	-	78.2
2DCNN+RT [21]	-	86.8	-	77.3
2DCNN [25]	-	84.4	-	75.4
Proposed	81.7	**90.8**	**76.4**	80.4

TABLE VI
COMPARISON OF THE ALGORITHMS USING IMAGE-BASED AND VIDEO-BASED METHODS(%)

Methods	Ave	Y	YT	T
Image-based	-	87.4	-	**82.2**
Proposed	81.7	**90.8**	**76.4**	80.4

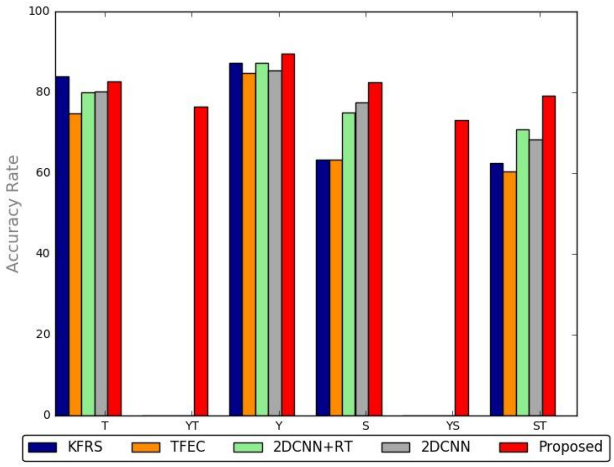

Fig.11. Recognition results of the proposed method and four states-of-the-art methods on MFAY dataset (%)

VII. CONCLUSIONS

The existing yawning detection methods only use spatial features based on static images. Most of them lack of temporal features and are not robust. They do not perform well in some situations such as some facial actions and expressions with the same deformation of mouth as yawning. To alleviate these problems, a new approach to yawning detection based on subtle action recognition is proposed in this paper. The main contributions are two folds. Firstly, a new keyframe selection algorithm is proposed, which has the advantage of simple calculation and can effectively extract the subtle action's keyframes from original frame sequences. Secondly, a subtle action recognition network based on 3D convolution network is proposed to extract spatiotemporal features and detect yawning. We conducted extensive experiments using self-collected dataset and YawDDR dataset, which we constructed based on the standard YawDD dataset. The proposed method outperformed the existing methods in recognition rate and overall performance. Effectively recognize various facial actions and reduce the false detection rate of yawning.

However, low image resolution and large camera vibration reduce the effectiveness of the proposed method. We will address this limitation in future study by using better image preprocessing methods. We will also continue to improve our keyframe selection algorithm to get more representative frames.

REFERENCES

[1] C. Marina Martinez, M. Heucke, F. Wang, B. Gao and D. Cao, "Driving style recognition for intelligent vehicle control and advanced driver Assistance: A Survey," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 666-676, March 2018.

[2] A. D. McDonald, J. D. Lee, C. Schwarz, and T. L. Brown, "Steering in a random forest: ensemble learning for detecting drowsiness-related lane Departures," *Human Factors*, vol. 56, no. 5, pp. 986-998, Aug 2014.

[3] H. A. Kholerdi, N. TaheriNejad, R. Ghaderi, and Y. Baleghi, "Driver's drowsiness detection using an enhanced image processing technique inspired by the human visual system," *Connection Science*, vol. 28, no. 1, pp. 27-46, Jan 2 2016.

[4] W. Sun, X. Zhang, S. Peeta, X. He and Y. Li, "A real-time fatigue driving recognition method incorporating contextual features and two fusion levels," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3408-3420, Dec. 2017.

[5] B. Mandal, L. Li, G. S. Wang and J. Lin, "Towards detection of bus driver fatigue based on robust visual analysis of eye state," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 545-557, March 2017.

[6] L. Li, Y. Chen and Z. Li, "Yawning detection for monitoring driver fatigue based on two cameras," *2009 12th International IEEE Conference on Intelligent Transportation Systems*, St. Louis, MO, 2009, pp. 1-6.

[7] M. M. Ibrahim, J. J. Soraghan, L. Petropoulakis, and G. Di Caterina, "Yawn analysis with mouth occlusion detection," *Biomedical Signal Processing and Control*, vol. 18, pp. 360-369, Apr 2015.

[8] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Facial video-based detection of physical fatigue for maximal muscle activity," *Iet Computer Vision*, vol. 10, no. 4, pp. 323-330, Jun 2016.

[9] S. Abtahi, B. Hariri and S. Shirmohammadi, "Driver drowsiness monitoring based on yawning detection," *2011 IEEE International Instrumentation and Measurement Technology Conference*, Binjiang, 2011, pp. 1-4.

[10] M. M. Ibrahim, J. S. Soroghan and L. Petropoulakis, "Mouth covered detection for yawn," *2013 IEEE International Conference on Signal and Image Processing Applications*, Melaka, 2013, pp. 89-94.

[11] M. Omidyeganeh, A. Javadtalab and S. Shirmohammadi, "Intelligent driver drowsiness detection through fusion of yawning and eye closure," *2011 IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems Proceedings*, Ottawa, ON, 2011, pp. 1-6.

[12] Y. Zhang and C. J. Hua, "Driver fatigue recognition based on facial expression analysis using local binary patterns," *Optik*, vol. 126, no. 23, pp. 4501-4505, 2015.

[13] C. Anitha, M. K. Venkatesha, and B. S. Adiga, "A two fold expert system for yawning detection," *2nd International Conference on Intelligent Computing, Communication & Convergence, Iccc 2016*, vol. 92, pp. 63-71, 2016.

[14] W. Y. Ding, L. Zhang, and Y. H. Chen, "Yawning detection based on mouth feature points curve fitting," *Advanced Designs and Researches for Manufacturing, Pts 1-3*, vol. 605-607, pp. 2227-2231, 2013.

[15] B. Akrout and W. Mahdi, "Yawning detection by the analysis of variational descriptor for monitoring driver drowsiness," *2016 International Image Processing, Applications and Systems (IPAS)*, Hammamet, 2016, pp. 1-5.

[16] Y. Du, D. G. Chen, Q. H. Hu, and P. J. Ma, "Kernelized fuzzy rough sets based yawn detection for driver fatigue monitoring," *Fundamenta Informaticae*, vol. 111, no. 1, pp. 65-79, 2011.

[17] C. Szegedy *et al.*, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1-9.

[18] He, K.M., Zhang, X.Y., Ren, S.Q., and Sun, J.: 'Deep residual learning for image recognition', Proc Cvpr Ieee, 2016, pp. 770-778.

[19] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[20] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., and Berg, A.C.: 'SSD: Single shot multi box detector', *Lect Notes Comput Sc*, 2016, 9905, pp. 21-37.

[21] Weiwei Zhang, Y. L. Murphey, Tianyu Wang and Qijie Xu, "Driver yawning detection based on deep convolutional neural learning and robust nose tracking," *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, 2015, pp. 1-8.

[22] L. Zhao, Z. Wang, X. Wang and Q. Liu, "Driver drowsiness detection using facial dynamic fusion information and a DBN," in *IET Intelligent Transport Systems*, vol. 12, no. 2, pp. 127-133, 3 2018.

[23] In-Ho Choi, Sung Kyung Hong and Yong-Guk Kim, "Real-time categorization of driver's gaze zone using the deep learning techniques," *2016 International Conference on Big Data and Smart Computing (Big Comp)*, Hong Kong, 2016, pp. 143-148.

[24] H. Han and U. Chong, "Neural network based detection of drowsiness with eyes open using AR modelling," *Iete Technical Review*, vol. 33, no. 5, pp. 518-524, Sep-Oct 2016.

[25] Su-Gang M A , Chen Z , Han-Lin S , et al. "Yawning detection algorithm based on convolutional neural network," *Computer Science*, 2018.

[26] Liu, L., Shao, L., and Rockett, P.: 'Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition', *Pattern Recogn*, 2013, 46, (7), pp. 1810-1818.

[27] N. Azouji and Z. Azimifar, "A new approach to speed up in action recognition based on key-frame extraction," *2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP)*, Zanjan, 2013, pp. 219-222.

[28] M. Lux, K. Schoffmann, O. Marques, and L. Boszormenyi, szormenyi, and L. Boszormenyi, zormenyi, nyi, i, i, cheme for key frame extraction and *Proceedings of the 9th Workshop on Multimedia Metadata (WMM)*. CEUR Workshop Proceedings, 2009, vol. 441, pp. 19–20.

[29] Jing Li, Weidong Min, MengdanFan, Qing Han. "Real-time face recognition based on face pre-identification detection and multi-scale classification". *IET Computer Vision*, DOI: 10.1049/iet-cvi.2018.5586, 2018.

[30] Weidong Min, Jie Shi, Qing Han, and Wei Wang. "A distributed face recognition method and performance optimization". *Optics and Precision Engineering*, 25(3) : 779 - 785, 2017.

[31] P. Viola and M. J. Jones, "Robust real-time face detection, "*International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, May 2004.

[32] D. Hefenbrock, J. Oberg, N. T. N. Thanh, R. Kastner and S. B. Baden, "Accelerating Viola-Jones face detection to FPGA-Level using GPUs," *2010 18th IEEE Annual International Symposium on Field-Programmable Custom Computing Machines*, Charlotte, NC, 2010, pp. 11-18.

[33] T. Huang, G. Yang and G. Tang, "A fast two-dimensional median filtering algorithm," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 1, pp. 13-18, February 1979.

[34] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[35] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann. Hidden two-stream convolutional networks for action recognition. arXiv preprint arXiv:1704.00389, 2017.

[36] Feichtenhofer, C., Pinz, A., and Zisserman, A.: 'Convolutional two-stream network fusion for video action recognition', Proc Cvpr Ieee, 2016, pp. 1933-1941.

[37] H. Xu, A. Das and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 5794-5803.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 4489-4497.

[39] P. Wang, Y. Cao, C. Shen, L. Liu and H. T. Shen, "Temporal pyramid pooling-based convolutional neural network for action recognition," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2613-2622, Dec. 2017.

[40] J. Hou, X. Wu, Y. Sun and Y. Jia, "Content-attention representation by factorized action-scene network for action recognition," in *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1537-1547, June 2018.

[41] S. Zhou, J. Wang, R. Shi, Q. Hou, Y. Gong and N. Zheng, "Large margin learning in set-to-set similarity comparison for person re identification," in *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 593-604, March 2018.

[42] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *ACM Conference on Multimedia*, New York, 2014, pp. 675–678.

[43] S Abtahi, M Omidyeganeh, S Shirmohammadi, B Hariri, *et al*., "YawDD: A yawning detection dataset," *ACM Multimedia Systems*, Singapore, pp. 24-28, March 2014.

Hao Yang received the B.E. degree in computer science and technology from Taiyuan University of Technology, China in 2016. He is currently pursuing the M.E. degree at Nanchang University, China. His research interests include machine learning and computer vision.

Li Liu received the Ph.D. degree in pattern recognition and intelligent system from East China Normal University, Shanghai, China, in 2015. She was with the Centre for Pattern Recognition and Machine Intelligence, Concordia University, Montreal, QC, Canada, from 2013 to 2014 as a visiting doctoral student, and in2016 as a visiting scholar. She is currently a lecturer with Nanchang University. Her research interests include pattern recognition, computer vision, and document image analysis.

Weidong Min (M'12) received the B.E., M.E. and Ph.D. degrees in computer application from Tsinghua University, China in 1989, 1991 and 1995, respectively. He is currently a Professor and the Dean, School of Software, Nanchang University, China. He is an Executive Director of China Society of Image and Graphics. His current research interests include image and video processing, artificial intelligence, big data, distributed system and smart city information technology. Since 2015 he has been a Professor with Nanchang University, China. From 2011 to 2014 he cooperated with School of Computer Science & Software Engineering, Tianjin Polytechnic University, China. From 1998 to 2014 he worked as a Senior Researcher and Senior Project Manager at Core land other companies in Canada. From 1995 to 1997 he was a Post-Doctoral Researcher at the University of Alberta, Canada. From 1994 to 1995 he was an Assistant Professor at Tsinghua University, China.

Xiaosong Yang is currently an Associate Professor in the National Centre for Computer Animation, Bournemouth University, UK. He received his bachelor (1993) and master degree (1996) in computer science from Zhejiang University (P. R. China) and Ph.D. (2000) in computing mechanics from Dalian University of Technology (P. R. China). He worked as Post Doc (2000–2002) in the Department of Computer Science and Technology of Tsinghua University for two years and as Research Assistant (2001–2002) at Chinese University of Hong Kong. His research interests include deep learning, computer vision, computer animation, motion capture and synthesis, VR&AR, special effects and game development, digital health, data mining, medical visualization. He has published more than 70 papers in journals and refereed conferences.

Xin Xiong received the M.E. degree in control theory and control engineering from Nanchang Hangkong University, China in 2015. He is currently pursuing the Ph.D. degree at Nanchang University, China. His current research focuses on computer vision.