IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED COMMUNICATIONS AND NETWORKING TECHNIQUES FOR WIRELESS CONNECTED INTELLIGENT ROBOT SWARMS

Received December 18, 2019, accepted December 30, 2019, date of publication January 8, 2020, date of current version January 21, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2964794*

# Non-Contact Emotion Recognition Combining Heart Rate and Facial Expression for Interactive Gaming Environments

**GUANGLONG DU, SHUAIYING LONG, AND HUA YUAN**
School of Computer Science and Engineering, South China University of Technology, Guangzhou 510000, China

Corresponding author: Hua Yuan (hy_scut@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61973126, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2017A030306015, in part by the Pearl River S&T Nova Program of Guangzhou under Grant 201710010059, in part by the Guangdong Special Projects under Grant 2016TQ03X824, in part by The Fundamental Research Funds for the Central Universities under Grant 2019ZD27, in part by The Science and Technology Planning Project of Guangdong Province under Grant 2017B090914002, and in part by the Innovation Team of the Modern Agriculture Industry Technology System in Guangdong Province under Grant 2019KJ139.

**ABSTRACT** A key to optimize a user's entertainment or learning experience when playing interactive games is to understand his emotional responses. Current methods mostly exploit intrusive physiological signals to detect a player's emotions. In this study, we proposed a method to detect a player's emotions based on heart beat (HR) signals and facial expressions (FE). In this work, a continuous recognition of HR and FE through videos captured by Kinect2.0 is conducted considering the continuous perception of the human emotion. Bidirectional long and short term memory (Bi-LSTM) network is used to learn the HR features, and convolutional neural network (CNN) is trained to learn the FE features. To further meet the demands for real-time, the SOM-BP network is employed to fuse the HR and FE features, which can perfectly recognize the player's emotion. Experimental results demonstrate our model has high accuracy and low computation time for four emotions of ''excitement'', ''anger'', ''sadness'' and ''calmness'' in different games. Moreover, the emotion's intensity can be estimated by the HR value.

**INDEX TERMS** Contactless emotion recognition, facial expression, heart rate, game evaluation.

## I. INTRODUCTION

Nowadays more and more users are attracted by computer games owing to their ability to present information interactively and playfully. The game was originally designed to increase the user's entertainment experience. The game is becoming more abundant as time goes by, which is gradually used to help users solve practical problems such as work, education and life. These can be attributed to the game can provide users with an emotional experience such as fun and excitement to reach a ''teach on the happy'' effect. These emotions can be detected and used to provide real-time adjustment to either the game difficulty or the gameplay. Therefore, the research of emotional recognition during games can maintain user's involvement and enhance their gaming experience. For this purpose, automatic emotion recognition for game users is mandatory to maintain his/her involvement without interrupting his/her gaming process [1]. Emotion recognition is mainly achieved in two ways, one is by obtaining the emotional behavior of the player, such as facial expression, facial micro-action, speech, body movements, etc. The other is to detect the physiological signals of the player, such as heart beat (HR), HR variability, electrocardiogram (ECG), and electroencephalogram (EEG). Among various emotion signals, the speech signals are the most easily available signals for emotion recognition. Shen *et al.* [2] used Support Vector Machine (SVM) as a classifier to classify happy, sad, neutral, fear and other states in the research of automatic speech emotion recognition. Yang and Lugger [3] proposed a speech emotion recognition method based on the psychoacoustic harmony perception known in music theory, and the performance was reliable. Ramakrishnan and Emary [4] employed acoustic features to recognize emotion and introduced 10 interesting applications of speech emotion

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang.

recognition. But these methods usually do not work due to the background music of games interferes with the player's voice.

Affective body movements provide important visual cues to distinguish emotion [5]–[9]. Yang and Narayanan [10] considered a statistical framework for the kinetic modeling of body movements in binary interactions. The framework identified participants' emotional states from body language. A system for recognizing emotions through full-body movements was proposed by Camurri [11], which can be used to recognize and express emotion for children with autism. Reference [12] proposed a method for automatic emotion detection based on a player's body movements in the sports game. This method, however, can't generalize to other scenarios as most games currently do not require a full-body movement from the player.

Facial expressions are the most widely employed modality for emotion recognition. A method called Facial Dynamics Map was able to aware of people's emotion correctly through a video sequence of microexpression in [13]. A probabilistic method based on 2D geometrical features for pose-invariant facial emotion recognition was proposed in [14]. Shojaeilangari *et al.* [15] employed a unified probabilistic framework based on the dynamic Bayesian network to simultaneously and coherently represent the facial evolvement in different levels to recognize emotion. In method [16], a classifier with deep convolutional network features could track the player's facial expressions in real time with an optimal recognition rate of 94.4%. These methods based on facial expressions show great performance. However, individuals have more control and can manipulate their facial expressions which can make the truly felt affective state difficult to measure from their facial expressions. Recently, the focus has shifted to using physiological signals, which can provide continuous measurements and are out of an individual's control [17], [18]. Therefore combining facial expressions and physiological signals is the best solution in an interactive gaming environment. Moreover, multimodal approaches have also been proved to improve the accuracy of affect detection [19], [20].

Previous researches showed that HR was a good indicator for discriminating between different affective states [21]. Valenza *et al.* [22] pointed out that different emotional states can trigger different HR frequencies. Moreover, we use a video-based method for measuring HR without interrupting the player's gaming process. Therefore HR is selected as the physiological signals in this study.

Although these methods have achieved notable performance, they are still necessary to be improved.

1) Continuous emotion recognition can achieve higher accuracy due to the importance of contextual information in sequence data [23], [24]. However, most of the existing methods focused on dealing with discrete signals [25].

2) Measuring heart rate is highly intrusive, which interferes with the game process of the users. Therefore, their real gaming feelings cannot be obtained.

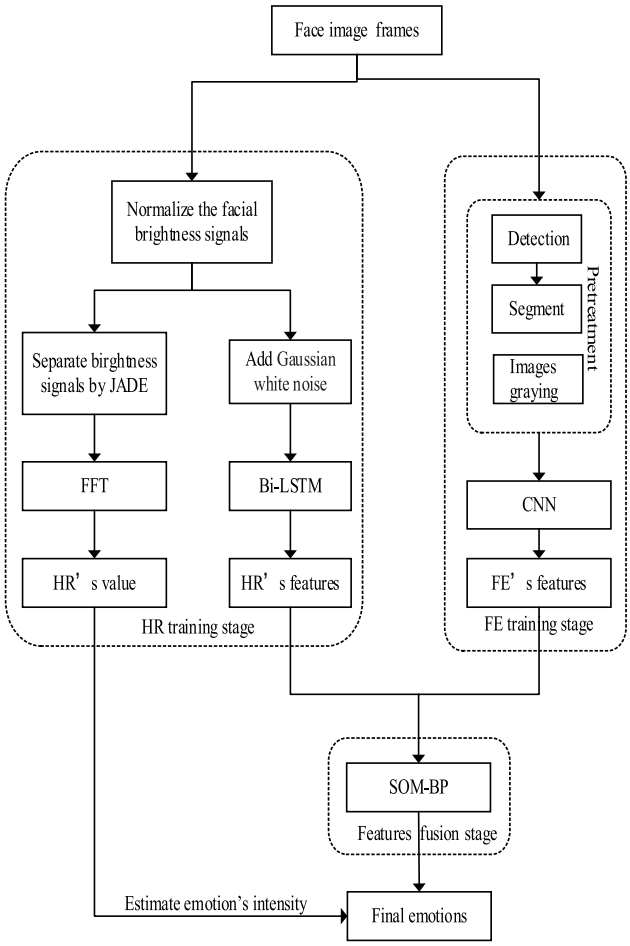



**FIGURE 1.** Process chart of the proposed method.

3) The emotion's intensity is quite useful and can be used to adjust the gameplay or the game difficulty in real time. Most current methods only recognize emotional categories.

In our proposed method (as shown in FIGURE 1), the video sequences containing the player's face are collected by Kinect2.0 to make contactless emotion recognition to maintain players' involvement. Firstly, the joint approximation diagonalization of eigen-matrices (JADE) algorithm is used to perform independent component analysis (ICA) on four channels signals of red-green-blue (RGB) and infrared (IR). And Fast Fourier Transform (FFT) is performed on the obtained independent components to match the heart rate range to obtain the player's HR value. The HR value can well reflect the player's emotional intensity. Then Bi-LSTM is employed to extract the HR features as it can take context information into account and is ideal for modeling time series data. Secondly, considering that CNN network is widely used for image processing, the facial region of interest (ROI) is input to our CNN model for extracting facial expressions (FE) features after detecting, segmenting, graying the face image frame and subtracting its mean value. Finally, the Self-Organizing Map (SOM) network don not need pre-specified the category of the input data and can make the input data for cluster analysis, realizing the preliminary classification of

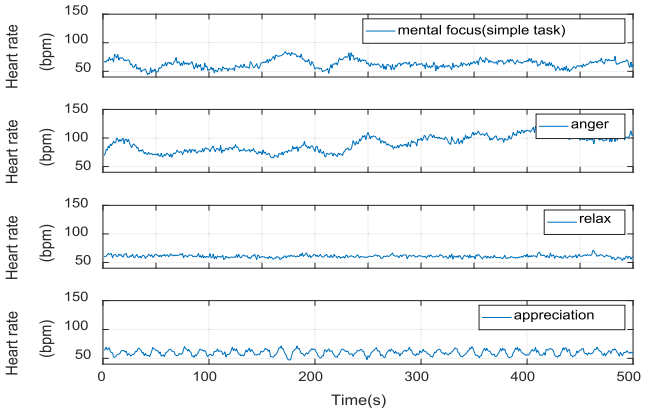



**FIGURE 2.** An example of the relationship between HR and emotion states is remade according to [21].

the data. And back propagation (BP) neural network has the ability of nonlinear mapping to complete the final classification. Our SOM-BP network is employed to fuse HR features and FE features. The fused features can well recognize the player's emotion. Our model provides a non-contact way to make use of HR and FE for emotion recognition.

The main contributions of this paper are as follows:

1) The emotion recognition for 30 consecutive seconds in our study is consistent with the continuous perception of the human emotion, which reduces the camouflage. Moreover, our method can detect the emotional intensity.

2) We use video-based detection of HR as a channel for emotion recognition, which realizes a contactless measurement for HR.

3) Our model can meet the demands for real-time due to the use of SOM-BP network.

The rest of this paper is organized as follows: Section II introduces the extraction of HR features and FE features and how to fuse them based on SOM-BP network. In Section III, we present experimental results to evaluate the proposed method. Section IV makes a conclusion.

## II. METHODS OF RECOGNITION EMOTIONS

### A. ACQUISITION OF HEART RATE VALUE AND FEATURES

People's emotions have a strong correlation with their HR. From a medical point of view, when people's emotions change, their HR will also change accordingly. McCraty [21] observed that certain emotional states are always related to different psychological and behavioral factors, and correspond to specific HR patterns. FIGURE 2 shows HR changes under a certain emotion. The HR value also is affected by emotions. Human's resting heart rate is the number of heart beats per minute (bpm) when he/she is at rest. For most of us, between 60 and 100 beats per minute is normal [26].

As the heart beats, blood is pumped into the facial muscles, causing subtle changes in facial brightness values. These subtle changes can be analyzed to obtain the HR value. Kinect2.0 can detect periodic changes in facial brightness through built-in cameras. We use Kinect2.0 to collect facial brightness signals every 30 seconds.

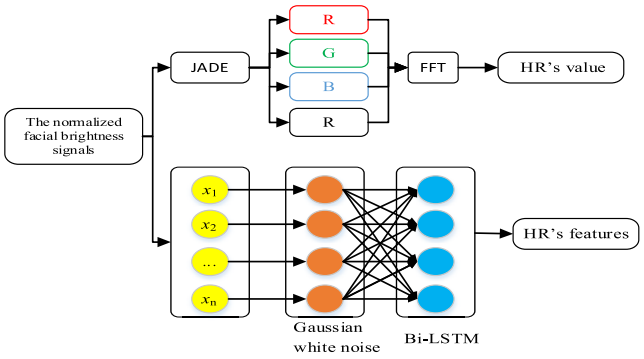



**FIGURE 3.** The flow chart of HR processing.

As shown in FIGURE 3, we process the captured facial brightness in two different ways. According to the above way (FIGURE 3), after making Independent Component Analysis (ICA) and Fast Fourier Transform (FFT) on the signals, we can calculate the heart rate value. While according to the following way (see FIGURE 3), we employ Bi-LSTM on the signals after noise reduction by Gaussian white noise. Then the emotional features represented by the HR signals can be obtained.

Here we introduce the way of calculating HR value. We apply ICA on the normalized signals. ICA is a signal analysis method based on high-order statistical features of signals. The observed random signals follow (1).

$$s = wx \tag{1}$$

where $x$ is the observation signal matrix and there is a statistical correlation between each observation signal. After the transformation of the separation matrix $w$, the correlation between the individual signal components of the signal matrices decreases. The JADE algorithm [27] belongs to the batch algorithm in the ICA algorithm, which can calculate $w$. The calculation steps are as follows.

Step 1. Calculate the covariance $\hat{R}_x$ of the signals from four channels and compute a whitening matrix $\hat{W}$.

Step 2. Calculate the fourth-order cumulants $\hat{Q}_z = \sum_{k,l=1}^{n} Cum(Z_i, Z_j, Z_k, Z_l) m_{lk}$ of the signals whitening process $\hat{z}(t) = \hat{W}x(t)$; compute the $n$ most significant eigenpair $\hat{\mathcal{N}}^e = \left\{\hat{\lambda}_r, \hat{M}_r | 1 \le r \le n\right\}$.

Step 3. Jointly diagonalize the set $\hat{\mathcal{N}}^e = \left\{\hat{\lambda}_r \hat{M}_r | 1 \le r \le n\right\}$ by a unitary matrix $\hat{U}$.

Step 4. An estimate of A is $\hat{A} = \hat{W} * \hat{U}$.

After separation, the signals from four channels (RGB and IR) are shown in FIGURE 4. Then the signals are extracted using FFT to find the matching heart rate range [28].

This method of calculating the HR value was later compared with the value provided by a smart wristband with strong reliabilities in measuring HR. In the experiment, we use a Kinect2.0 and a smart wristband to measure a player's HR in his random emotional states at the same time. Twenty comparison measurements are taken and results are recorded from each set, which may last 30 seconds. As shown

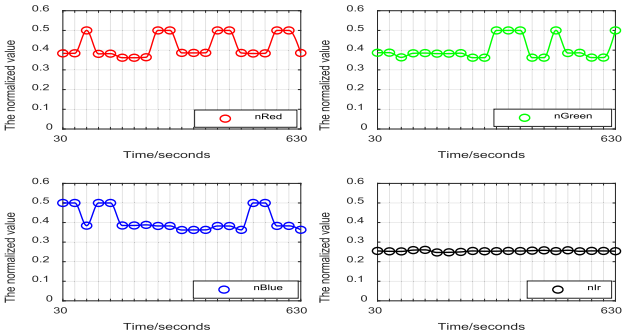



**FIGURE 4.** The normalized value of four independent components.

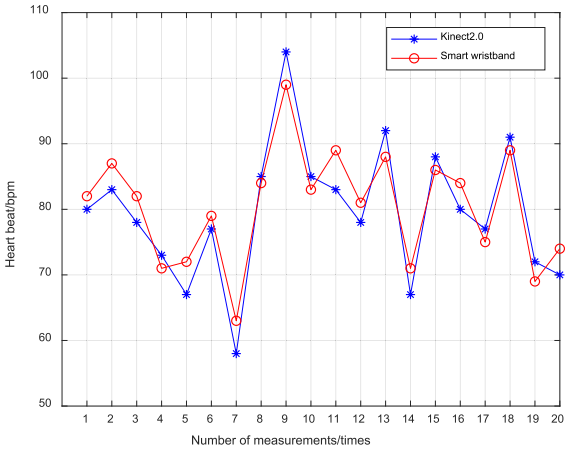



**FIGURE 5.** The comparison of heart rate measurement by Kinect2.0 and a smart wristband.

in FIGURE 5, it can be seen that the measurement error is within 6 bpm. After gaining the HR values, we can make a preliminary judgment on the player's emotional state according to our emotional judgment rules.

Next, we extract the features represented by the HR signals. Due to the influence of angle or ambient lights, there is noise in the captured signals. The captured signals cannot be directly used for HR feature extraction. We add Gaussian white noise $f_N(0, 1)$ to the signals [29]. The formula for noise can be written as

$$\xi_i = x_i / 10^{\frac{s}{10}} * f_N(0, 1) \tag{2}$$

where $\xi_i$ represents the $x_i (i = 1, 2, \ldots, n)$ signal containing Gaussian white noise, and $s$ represents the degree of noise, which is a constant. $f_N(0, 1)$ refers to a number randomly extracted from the standard positive distribution.

In order to calculate the membership of every $\xi_i$ signal, we make a full connection to this layer of signal with Gaussian white noise added according to the formula below. The formula can be written as

$$\boldsymbol{\varphi} = \hat{F} * \boldsymbol{\xi}(\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_n), \boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n))$$

$$\hat{F} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1m} & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2m} & f_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ f_{j1} & f_{j2} & \cdots & f_{jm} & f_{jn} \\ f_{n1} & f_{n2} & \cdots & f_{nm} & f_{nn} \end{bmatrix}$$

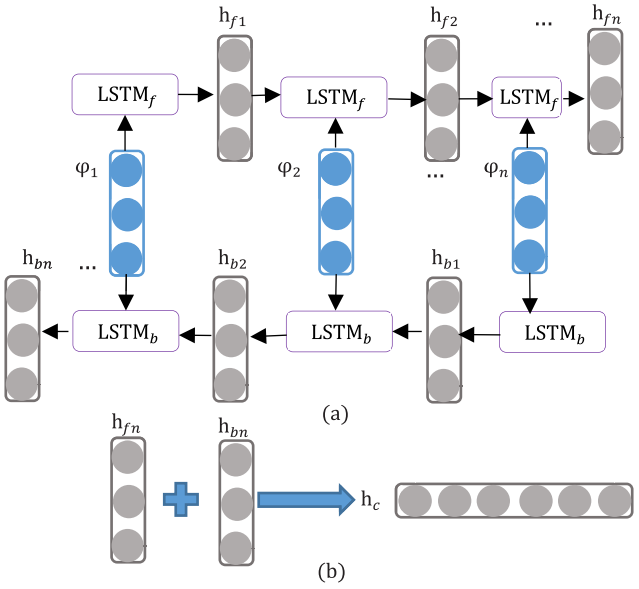



**FIGURE 6.** Bi-LSTM for extracting HR features. (a) HR signal coding. (b) Splicing of encoded HR vectors.

$$(j = 1, 2, \ldots, n; \quad m = 1, 2, \ldots, n) \tag{3}$$

The HR signals are time-series signals and are related in time. Bi-LSTM [30]–[33] can take context information into account and is ideal for modeling time series data. As shown in FIGURE 6, Bi-LSTM is used to process the HR signals. The forward LSTM ( LSTM$_f$) inputs $\varphi_1, \varphi_2, \ldots, \varphi_n$ in sequence, the encoded vectors are $h_{f1}, h_{f2}, \ldots, h_{fn}$. And the backward LSTM ( LSTM$_b$) inputs $\varphi_n, \ldots, \varphi_2, \varphi_1$ in sequence, the encoded vectors are $h_{b1}, h_{b2}, \ldots, h_{bn}$. Considering that $h_{fn}$ and $h_{bn}$ contain all the information for forward and backward, we splice them together for emotion classification (As shown in FIGURE 6(b)). The model is trained by minimizing loss function with back propagation and stochastic gradient descent method. As a result, we can label the HR signals with emotional features.

Each LSTM unit will selectively forget the information in the cell state and remember new information. This allows useful information to be passed and useless information to be discarded. The LSTM unit outputs the hidden layer status $h_t$(t= $1, 2, \ldots, n$) at each time step. How to forget, remember and output are controlled by the forgetting gate, the input gate and the output gate calculated by the hidden layer state at the last moment $h_{t-1}$ and the current input $\varphi_t$. The forgetting gate choose the information to forget according to the formula below, which determines how much of the cell state $C_t$ at the last moment is retained to the current moment. The formula can be written as

$$f_t = \sigma(W_f * [h_{t-1}, \varphi_t] + b_f) \tag{4}$$

where $f_t$ denotes the output of the forgotten gate, $W_f$ is the weight matrix of the forgetting gate, $[h_{t-1}, \varphi_t]$ is the concatenation of the two vectors, $b_f$ denotes the bias of the forgotten gate, $\sigma$ is Sigmoid function.

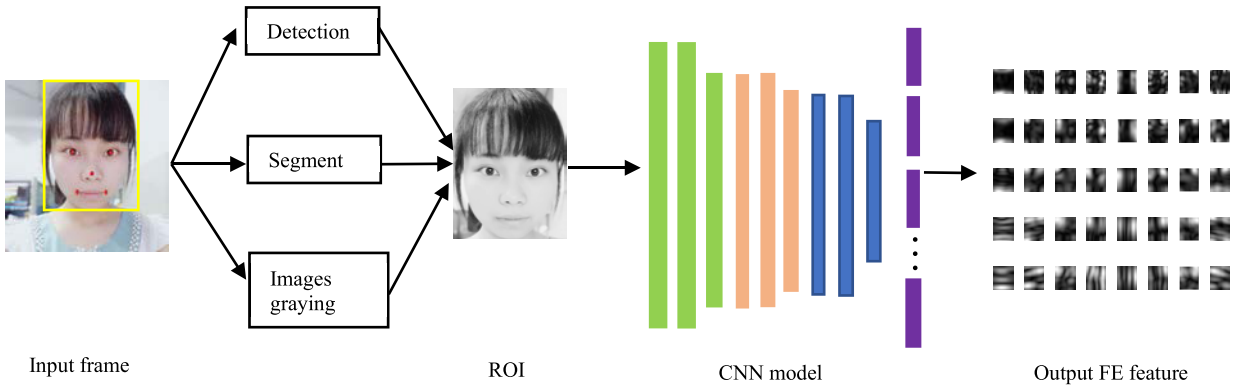



**FIGURE 7.** CNN for extracting FE features.

The input gate selects the current input $\varphi_t$ to remember according to the formula below, which determines how much of the current input $\varphi_t$ is saved to the cell state $C_t$.

$$i_t = \sigma(W_i * [h_{t-1}, \varphi_t] + b_i) \tag{5}$$

where $i_t$ denotes the output of the input gate, $W_i$ is the weight matrix of the input gate, $[h_{t-1}, \varphi_t]$ is the concatenation of the two vectors, $b_i$ represents the bias of the forgotten gate, $\sigma$ is Sigmoid function.

The temporary cell state $\tilde{C}_t$ is shown in (6), which denotes the current memory.

$$\tilde{C}_t = tanh(W_c * [h_{t-1}, \varphi_t] + b_c) \tag{6}$$

where $W_c$ is the weight matrix of the *tanh* gate, $[h_{t-1}, \varphi_t]$ is the concatenation of the two vectors, $b_c$ denotes the bias of the *tanh* gate, *tanh* represents limiting the cell state to a value between -1 and 1.

The current cell state $C_t$ is as shown in (7), which can combine current memory with previous memory to form a new cell state.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

where $f_t$ denotes the output of the forgotten gate, $C_{t-1}$ denotes the cell state of the last moment, $i_t$ represents the output of the input gate, $\tilde{C}_t$ is the temporary cell state.

The output gate controls how many the cell states are available as the current output of the LSTM according to the formula below. The formula can be written as

$$o_t = \sigma(W_o * [h_{t-1}, \varphi_t] + b_o) \tag{8}$$

where $o_t$ denotes the output of the output gate, $W_o$ is the weight matrix of the forgetting gate, $[h_{t-1}, \varphi_t]$ is the concatenation of the two vectors, $b_o$ denotes the bias of the forgotten gate, $\sigma$ is Sigmoid function.

Then we process the cell state through *tanh* to get a value between -1 and 1. Multiply the obtained value by the output of the output gate to get a new hidden layer state $h_t$.

$$\mathrm{h}_t = o_t * tanh\,(C_t) \tag{9}$$

**TABLE 1.** Our CNN model architecture.

| Type | kernel | stride | pad | output | dropout |
|---|---|---|---|---|---|
| input | | | | 48*48*1 | |
| convolution1 | 5*5 | 1 | 2 | 48*48*32 | |
| convolution2 | 4*4 | 1 | 1 | 47*47*32 | |
| pooling1 | 3*3 | 2 | | 23*23*32 | |
| convolution3 | 3*3 | 1 | 1 | 23*23*64 | |
| convolution4 | 4*4 | 1 | 1 | 22*22*64 | |
| pooling2 | 3*3 | 2 | | 10*10*64 | |
| convolution5 | 5*5 | 1 | 2 | 10*10*64 | |
| convolution6 | 5*5 | 1 | 2 | 10*10*64 | |
| pooling3 | 3*3 | 2 | | 4*4*64 | |
| FC | | | | 1*1*1024 | 0.7 |
| output | | | | 1*1*4 | |

### B. ACQUISITION OF FACIAL EXPRESSION FEATURES

HOG feature descriptor with a linear classifier is employed to complete the face detection. Then the Kinect Active Appearance Model (AAM) algorithm is used to segment our region of interest (ROI) in real time. As shown in FIGURE 7, the ROI includes five feature points (the left and the right eye, the nose, and the left and right mouth corners). Due to the influence of angle or background, there is noise in the ROI and cannot be directly used for FE feature extraction. These ROIs must be preprocessed. After filtering, denoising and gray-scale equalization, the original obtained images become grayscale images.

After detecting, segmenting, graying the face image frame and subtracting its mean value, its dimension is reduced to 48px×48px. The 48px×48px face image becomes the input of our CNN model.

Considering that CNN has perfect performances in image feature extraction, the FE feature extraction task is implemented by our CNN model. As shown in Table 1, the entire model is consisted of six convolutional layers, three pooling layers and finally one fully connected layer.

The first layer of our model is the convolution layer, which is the feature extraction layer. We perform the convolution operation on the convolution kernel and the upper layer with

all the feature maps. The output of the convolution operation is activated by the activation function, thus forming a feature map of the current convolution layer. The operation is as follows

$$\mathrm{net}_j^l = \sum_{i \epsilon M_j} a_i^{l-1} \otimes w_{i,j}^l + w_b \tag{10}$$

$$a_{i,j}^l = \mathrm{ReLU}\left(\mathrm{net}_{i,j}^l\right) \tag{11}$$

where $\mathrm{net}_j^l$ denotes the weighted input of layer 1. $a_i^{l-1}$ represents the feature map of the output of the l–1 layer. $w_{i,j}^l$ is a convolution kernel matrix, it includes the connection weights between the l–1 layer of neurons and the l layer of neurons. $w_b$ represents the offset term of the j-th feature map. $a_{i,j}^l$ denotes the j feature map of the convolution l layer. ReLU()(Rectified Linear Units) is the activation function. ReLUs are tended to be several times faster than their equivalents in training. The main advantage of using ReLUs is that it can alleviate the vanishing gradient problem which is very common in using other two activation functions (Sigmoid, Tanh). ReLU() is defined by

$$\mathrm{ReLU}(x) = \max(0, x) \tag{12}$$

where x is the input to the neuron.

The pooling layer of the CNN model can avoid a dimension of disaster brought by the increasing number of convolution layers. In our CNN model, the down-sampling is performed by max-pooling. After down-sampling, the number of feature maps is the same as before, but the number of parameters reduces as it removes unnecessary information from each feature map. The operation is as follows

$$\mathrm{net}_j^l = \mathrm{down}(a_j^{l-1}) \tag{13}$$

where $a_j^{l-1}$ is the j feature map of the pool l-1 layer. down() denotes the down-sampling function.

The fully connected (FC) layer acts as a classifier by learning all the weights to integrate the "good" features and reduce other features. After the FC layer, the output becomes a one-dimensional array. The calculation is as follows

$$\mathrm{net}_j^l = \sum_i w_{i,j}^l a_j^{l-1} + w_b^l \tag{14}$$

$$y = \mathrm{ReLU}(\mathrm{net}_j^l) \tag{15}$$

where $\mathrm{net}^l$ is the output of the FC 1 layer. $w^l$ represents the weight matrix between neurons. $a^{l-1}$ denotes the input feature vector of the upper layer. $w_b^l$ is the offset term of the fully connected 1 layer.

The detail specification of parameters is listed in Table 1.

Back propagation [34] and stochastic gradient descent method [35] are applied to train our CNN model by minimizing loss function. Dropout is performed on the FC layers to prevent overfitting [36].

### C. FEATURES FUSION BASED ON SOM-BP

The Self-Organizing Map (SOM) proposed by Dutch in 1981 [37] forms a one or two-dimensional presentation from multi-dimensional dimensional data. The presentation keeps the topology of the data. In this way, the data vectors which closely resemble one another can be located next to each other on the map. The SOM network is a competitive learning network, which is consisted of an input layer and a competition layer.

BP neural network is a multilayer feed-forward neural network for training network according to error back propagation algorithm. As one of the most widely used neural networks, its structure includes an input layer, hidden layers, and an output layer.

The emotion recognition in interactive gaming environments is expected to quickly adjust the difficulty of the game, so real-time is very important. SOM networks can meet the demand of real-time as it does not need large amounts of data for training. However, some neurons whose initial weights are too far away from the input vector will never win in the competition and become dead neurons. To overcome this drawback, combining the BP neural network that works well in fault diagnose with the SOM network is a perfect choice. SOM network has the ability of self-learning, which can make cluster analysis on unclassified samples and implement their preliminary classification. Then the position of the winning neuron in the SOM network is input into the BP network to avoid dead neurons. So we combine the character of the SOM and BP network to fuse HR features and FE features.

As shown in FIGURE 8, our SOM-BP model includes an input layer, and a competitive layer, and a hidden layer, and finally an output layer. That is, a SOM competitive layer is added to a traditional three-layer BP network. First, the SOM network implements the preliminary recognition of the features automatically by mapping the linearly inseparable features of the high-dimensional space to the linearly separable

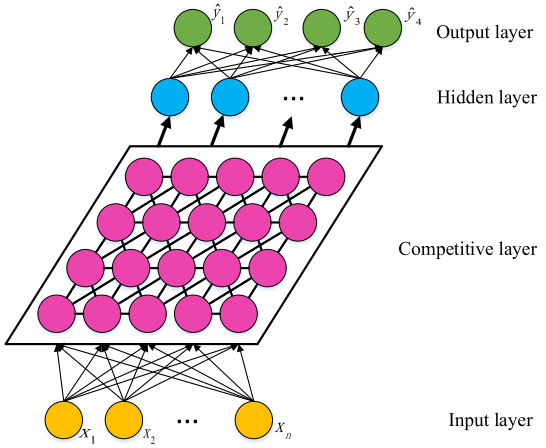



**FIGURE 8.** Our SOM-BP model. ($x_i$ ($i = 1, 2, \ldots, n$) denotes $\left(f_{HR,C} f_{FE}\right)$, which is a point in a two-dimensional plane determined by HR features and FE features. $\hat{y}_i$ ($i = 1, 2, \ldots, 4$) denotes the score for four emotion categories.).

**TABLE 2.** SOM-BP algorithm.

| **SOM-BP algorithm** |
|---|
| **Input:** feature set X = $\{x_1, x_2, x_3, \ldots, x_n\}$ |
| **Process:** |
| 1: Perform initialization on the SOM neural network; |
| 2: Calculate the Euclidean Distance $d_j$ from the input layer neurons to the competition layer neurons according to (16); |
| 3: Define the competition layer neuron with the smallest distance as the winning neuron; |
| 4: Update the weights $W_{ij}(t+1)$ between all input layer neurons to the competitive neurons according to (17); |
| 5: Update learning rate $\eta(t)$ and domain function $h_{c,j}$ (t) according to (18); |
| 6: Calculate the output $Y_k$ according to (21); |
| 7. Normalize the results of the previous step 6; |
| 8: Perform initialization on the BP neural network; |
| 9: The result of processing in step 8 is input into the BP neural network to calculate the output; |
| **repeat:** |
| 10: Calculate the error according to (22), then update the weights and thresholds of the BP neural network. |
| **until** the error reaches the predetermined range. |
| **Output:** the player's emotion category |

features of low-dimensional space. The operation makes the BP network less stress and less difficult to recognize the features.

Then the clustering features are transferred to the hidden layer from the competition layer. Finally, the BP network completes the nonlinear mapping from input to output with supervise learning mode and classifies the player's emotions.

SOM-BP network has an excellent performance in feature fusion. The HR features and the FE features are input to the neurons of the input layer. The adjacent features in the input space will be mapped to adjacent neurons in the competition layer, which is a two-dimensional plane capable of maintaining the topology of the input space. In this way, the two-dimensional features are classified, thus completing the preliminary classification of the input features. If the emotion category $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4\}$ of the output layer does not match the expected emotion category $Y = \{y_1, y_2, y_3, y_4\}$, it enters the phase of back propagation of the error, thus finishing the nonlinear mapping from the input features to the player's emotion. The fusion process can be written as the steps shown in Table 2.

The equations involved in Table 2 are described in detail here.

Equation (16) in Step 2:

$$d_j = \|X - w_j\| = \sqrt{\sum_{i=1}^{m} (X_i(t) - w_{ij}(t))^2} \tag{16}$$

where $w_{ij}$ denotes the weight between the input neuron i and the mapping neuron j.

Equation (17) in Step 4:

$$W_{ij}(t+1) = W_{ij}(t) + \eta_t h_{c,j}(t)\left(x_t - W_{ij}(t)\right) \tag{17}$$

where $\eta(t)$ denotes the learning rate, $0 < \eta(t) < 1$, $h_{c,j}$ (t) represents the domain function.

Equation (18) in Step 5:

$$h_{c,j}(t) = \exp\left(-\frac{d_{cj}^2}{2r^2(t)}\right) \tag{18}$$

where $d_{c,j}$ represents the distance from the winning neuron c to any activated neuron j in the neighborhood, r is the radius of neighborhood, whose updated rules are as follows

$$r(t+1) = \mathrm{INT}((r(t) - 1) * (1 - \frac{t}{T})) + 1 \tag{19}$$

$$\eta(t+1) = \eta(t) - \frac{\eta(0)}{T} \tag{20}$$

where INT () rounds a number to the nearest integer, T denotes the total number of iterations.

Equation (21) in Step 6:

$$Y_k = f(\min \|X - w_j\|) \tag{21}$$

where f() means a nonlinear function, $0 < f(\cdot) < 1$.

Equation (22) in Step 10:

$$\text{error} = {}^1\!/_2 \sum_{i=1}^{i=k} (\hat{y}_i - y_i)^2 \tag{22}$$

where error is the difference between the output of the SOM-BP network $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4\}$ and the expected output $Y = \{y_1, y_2, y_3, y_4\}$.

## III. EXPERIMENT

In this section, we first introduce the procedure of collecting datasets to test our model, then analyze the results of our method on the datasets and compare it with the method in [2], [11], [13], [19]. To compare which modality is more fit in an interactive gaming environment, methods in [2], [11], [13], [19] are selected. Methods in [2], [11], [13] employ speech, body movements, microexpressions respectively. Method [19] uses electroencephalogram (EEG), pupillary response and gaze distance. Since methods in [2], [11], [13] only use a modality, we also can verify the significance of combining modalities from different sources of information.

### A. PARTICIPANTS

Twelve volunteers (seven males and one female), ranging in age from 19 to 23 years old, were recruited for the experiments. They have no cardiovascular disease and are in good health. They were all university students. All volunteers were gamers with more than two years of gaming experience.

### B. APPARATUS

As shown in Fig. 9, the experimental equipment used in this study is Kinect2.0, which can record 32-bit color video frames at 1920×1080 resolution, IR camera can record 16-bit video at 521×424 resolution, whose working frequency is 50 frame per second (fps). When the test was started, the participants were asked to keep their bodies upright in front of the computer, and Kinect2.0 was placed approximately 0.6 meters in front of the volunteers. The experiment

was performed in a separate room with constant lighting and temperature.

### C. DESIGN

When 12 volunteers are playing games, we use Kinect2.0 to record their face videos, the difficulty of games. After the game is over, the volunteers' feedbacks are recorded as the ground truth. We deal with the face videos in three different ways to get the FE feature, HR feature and HR value. Firstly, to prove the effectiveness of the combination of FE and HR in an interactive gaming environment, the emotions recognized by the FE feature and HR feature respectively are compared to the emotions recognized by the fused features. Secondly, the HR value is compared to the game difficulty to verify that the emotional intensity is related to HR value. Thirdly, the degrees of excitement measured by the FE and HR were compared proving that HR is out of human control. Finally, our method is compared with the methods in [2], [11], [13], [19] and different feature fusion methods are also compared.

### D. PROCEDURE

Before the experiment, 12 volunteers were rested indoors for 5 minutes to calm their emotions. In the 5-minute resting status, volunteers were told to keep eyes closed and relax, during which HR signals were recorded as well [38]. Once their heart rate have stabilized, they were invited to seat in front of a computer and play games, meanwhile, their facial images were collected using the Kinect2.0 through RGB color and IR cameras for 30 seconds (s) at 50 fps, as shown in FIGURE 9. From start to end, every 30s, we perform a 30-second long emotion recognition on the volunteers. We saved some intermediate results in our emotional recognition program as the basis for preliminary results. The preliminary results are shown below: the HR recognition saved the HR mean within 30 seconds. The facial expression recognition obtained the following recognition results: 1)marked as *IsHappy* when it detected that the face was smiling; 2)marked as *IsNotHappy* when it detected that the face was not smiling and had a frowning action; 3)marked as *IsNeutral* when it detected that the face had no obvious smile or frowning action.

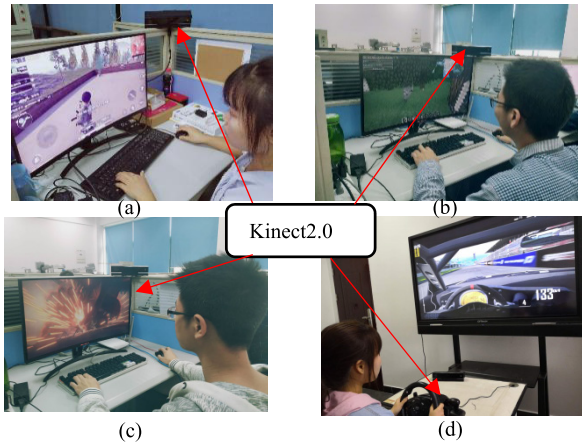



**FIGURE 9.** Emotional recognition for volunteers who are playing games. (a). the volunteer is playing PUBG. (b). the volunteer is playing Minecraft. (c). the volunteer is playing NieR: Automata. (d). the volunteer is playing Need for Speed.

We chose the most appearing expression in 30 seconds as the facial expression recognition results. At this time, we could have a preliminary judgment on volunteers' emotions according to our emotional judgment rule shown in Table 3. The preliminary results could detect whether the final emotion is misjudged in advance.

**TABLE 3.** Our emotional judgment rules.

| | | Recognized FE | | |
|---|---|---|---|---|
| | | *IsHappy* | *IsNotHappy* | *IsNeutral* |
| HR's frequency (bpm) | ≥90 | excited | angry | ~ |
| | 70~90 | calm | calm | calm |
| | ≤70 | calm | sad | sad |

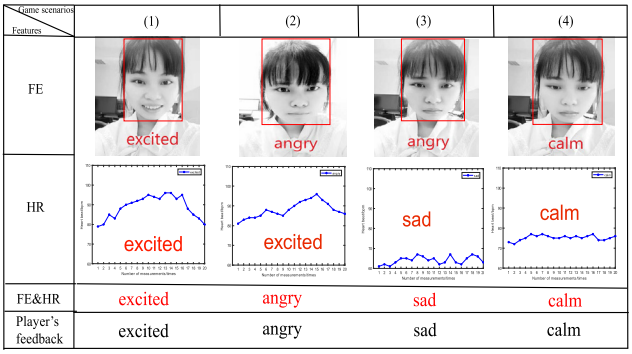
When the game was over, we immediately confirmed whether the volunteers felt the corresponding emotion during their gameplay. If the corresponding emotion was genuinely felt, it was counted once in the statistical number. If the volunteers' response was ''cannot recall'' or ''does not feel these four emotions'', this record would be regarded as an invalid measurement and discard it. In addition, according to the volunteers' feedback, we found that the intensity of emotion has a strong correlation with the frequency of HR. So our proposed method could also reflect the volunteers' emotion intensity. Then we compared whether the effective emotional recognition results were consistent with the volunteers' feedbacks. If they matched, the recognition was accurate. After each round, the subject had to take a 5 minutes rest to regain his mindset, then start a new round.

### E. RESULTS AND DISCUSSION

From the above experiences, 240 measurements were recorded, some were invalid and got discarded. We were then left with 153 measurements.

We recorded every volunteer's gameplay in four different game scenarios, respectively (1) Teammates match perfectly, (2) Fight alone and teammates don't provide support, (3) Improper operation makes the game fail, (4) Almost equal to the opponent's level. As shown in Table 4, one of the volunteers has the corresponding recognized emotions. The last row is the feedback emotion of the player in four game scenarios, which are consistent with the recognized emotions by fusing FE and HR. The first row is the emotion only recognized by FE. It can be seen that ''sad'' is misjudged as ''angry'' as the facial expressions of ''sad'' and ''angry'' are a bit difficult to distinguish. The second row is the emotion only recognized by HR. We can see that it misjudged ''angry'' as ''excited'', as the frequency of ''angry'' is closed to the one of ''excited''. Therefore, it can be concluded that the recognized results are more accurate after fusing two channel signals to reduce the ambiguity brought by FE or HR.

**TABLE 4.** The emotions from FE, HR, and FE&HR, and player's feedback emotions in four different game scenarios, respectively (1) Teammates match perfectly, (2) Fight alone and teammates don't provide support, (3) Improper operation makes the game fail, (4) Almost equal to the opponent's level.



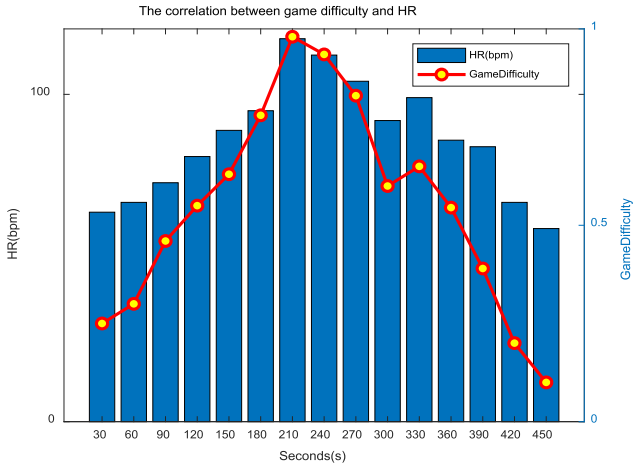




**FIGURE 10.** The correlation between game difficulty and HR.

Moreover, our method not only recognizes these different types of emotions but also measures their intensity through the player's HR value. As shown in FIGURE 10, the change in the player's HR value is positively related to the game difficulty (game difficulty is quantified with a value of 0-1.). The reason for that is, as the game level increases, the player's emotions become stronger, thus increasing his HR value.

FIGURE 11 shows the recorded gaming process of four of the players. As seen, the game difficulty was divided into six levels, with 0 being the easiest and 5 being the most difficult. We measured the degree of excitement only through HR or FE. We compared the degree of excitement measured by the two signals as the game difficulty changes. It can be seen that the degree of excitement detected by HR is more consistent with the difficulty of the game. Therefore, HR signals can reflect changes in the player's degree of excitement throughout the game more objectively.

As shown in Table 5, our method performs perfectly not only in terms of reliability but also in terms of efficiency. This can be attributed to the fact that both features are extracted from the homologous videos with appropriate algorithms and fused by the SOM-BP network. Homology videos reduce

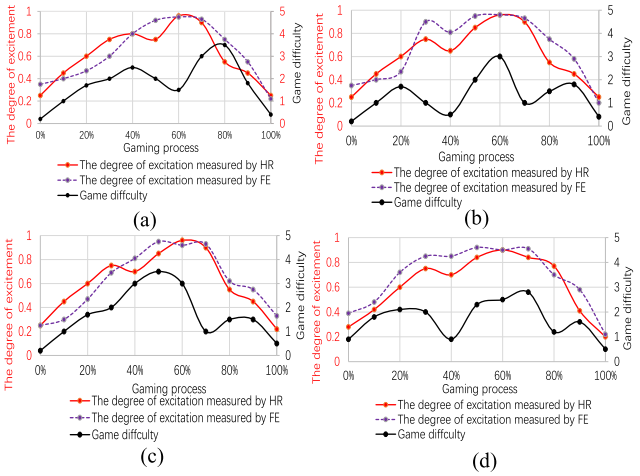



**FIGURE 11.** The Comparison of excitement degree measured by different signals with game difficulty changing. (a). The volunteer is playing PUBG. (b). The volunteer is playing Minecraft. (c). The volunteer is playing NieR: Automata. (d). The volunteer is playing Need for Speed.

**TABLE 5.** Performance comparison based on different feature sets (act denotes average computational time).

| Feature selection | | Accuracy rate (%) | ACT (ms) |
|---|---|---|---|
| Including FE | Without HR | 82.1 | 476 |
| | With HR | 87.3 | 452 |
| Excluding FE | With HR | 81.8 | 469 |

**TABLE 6.** The confusion matrix of recognition results combining FE with HR.

| | Excited | Angry | Sad | Calm |
|---|---|---|---|---|
| Excited | 90.2% | 4% | 4.3% | 1.5% |
| Angry | 6% | 83.4% | 8.9% | 1.7% |
| Sad | 7.5% | 10.2% | 80.1% | 2.2% |
| Calm | 1.4% | 1.6% | 1.5% | 95.5% |

**TABLE 7.** The confusion matrix of recognition results without HR.

| | Excited | Angry | Sad | Calm |
|---|---|---|---|---|
| Excited | 81.6% | 8.4% | 7.9% | 2.1% |
| Angry | 11.3% | 75.4% | 10.6% | 2.7% |
| Sad | 13.1% | 14.5% | 71.9% | 0.5% |
| Calm | 4.5% | 4.7% | 3.7% | 87.1% |

the amount of processed data. This demonstrates that how to extract and fuse features is necessary and effective for real-time emotion recognition.

Table 6 shows the confusion matrix of recognition results combining FE with HR. The average recognition accuracy is 87.3%. Our method has a good performance for recognizing "calm", and the accuracy of "excited" is also quite high. Since "angry" and "sad" can correspond to similar facial expressions, the probability of misjudgment is larger and their recognition accuracy is relatively low. While Table 7 shows the confusion matrix of recognition results without HR, it can be found that when combined with HR, the accuracy rate increased and more reliable judgment results are obtained.

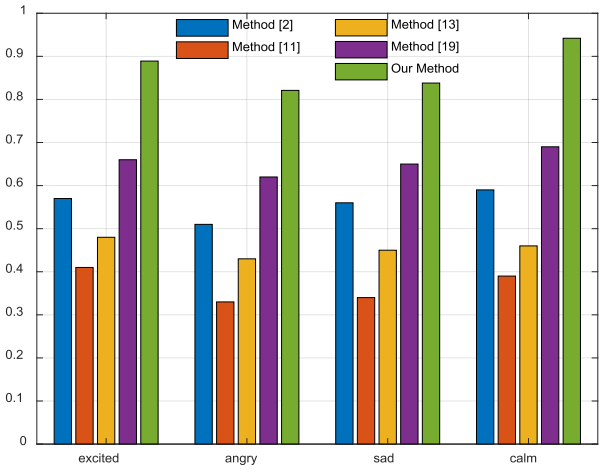



**FIGURE 12.** Accuracy comparison of five methods.

Therefore, HR makes up for the camouflage and deception brought by FE.

As shown in FIGURE 12, we compared the recognition accuracy using our method with methods in [2], [11], [13] and [19]. The results show that our method performs better in recognizing excitement, anger, sadness and calmness. Shen *et al.* [2] extracted features of speech (energy, pitch, linear prediction cepstrum coefficients (LPCC), Mel Frequency cepstrum coefficients (MFCC), Linear Prediction coefficients and Mel cepstrum coefficients (LPCMCC)) from human's utterances to automatically classify five emotional states. Method [11] proposed a computational model for the automated emotions recognition according to full-body movements. The emotion identification is completed by a method called the Facial Dynamics Map which characterizes the movements of a microexpression in different granularity in method [13]. Method [19] presents a user-independent emotion recognition method with the goal to recover affective tags for videos using electroencephalogram (EEG), pupillary response and gaze distance.

The performance improvement can be attributed to fusing two signals (FE and HR) and processing them without interfering user's gameplay. The first three methods only use one signal channel to detect emotion, the speech signal is used in [2], the full-body movement in [11] and the micro expressions in [13]. Compared to our method, the three methods mentioned above are less suitable for the game scenarios. The details can be described as follows. Players do not always make emotion-related sounds during their gameplay, and when they do, the sounds are very likely to be mixed with other sounds in the environment, making the method [2] slightly less performant. Moreover, the games played in our experiences do not require full-body movement, using method [11] will not work well in this scenario. Finally, in method [13], microexpressions are fleeting, lasting only a few frames within a video sequence. So they are difficult to perceive and interpret correctly. While HR signals and facial image frames in our method were continuously collected within 30 seconds. It can assure the higher accuracy owing to people's emotions are perceived through continuous means. Method [19] performs better than the first three methods, but is worse than our method although it uses players' EEG signals, gaze distance and pupillary response. The reason is that the collection of EEG signals makes method [19] intrusive, which disrupts the player's game process.

**TABLE 8.** Computational time comparison of three fusion method (act denotes average computational time).

| | Fuzzy Integral | MFB | SOM-BP |
|---|---|---|---|
| ACT (ms) | 758 | 613 | **452** |
| Accuracy (%) | 85.2 | 83.1 | **87.3** |
| F1(%) | 84.8 | 83.7 | **85.9** |
| RMSE (%) | 21.5 | 16.3 | **10.6** |

In addition, we compare the average computational time based on two different fusion methods in [39], [40], namely the fuzzy integral and MFB. As can be seen from Table. 8, the average computational time of our method is minimal as the SOM network does not need large amounts of data for training. The accuracy and F1 based on SOM-BP are a little higher than other methods. The RMSE based on SOM-BP is smaller than other methods.

## IV. CONCLUSION

In this paper, we propose a non-contact method for emotion recognition based on FE and HR signals. First, we used video-captured data as a means to detect emotion, making the measurement process contactless and does not interfere with the player's activity. Unlike other signals, HR can't be made up, therefore, we are sure to get genuine data. FEs are also captured by the camera. Second, our method detects the player's emotions during 30 seconds long in order to get more reliable results. Finally, the intensity of emotion can be measured using HR values, which can help game designers to design games that can maximize the users' experiences. However, our system is only trained to recognize four basic emotions and has high demands on the lighting and temperature, in future works, improvements will be made with the goal of recognizing more emotions and designing an incremental model suitable for common scenarios.

## REFERENCES

[1] B. Kumar, "Flow: The psychology of optimal experience," *Inf. Des. J.*, vol. 16, no. 3, pp. 75–77, 2008.

[2] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Proc. Int. Conf. Electron. Mech. Eng. Inf. Technol.*, Aug. 2011.

[3] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal Process.*, vol. 90, no. 5, pp. 1415–1423, May 2010.

[4] S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommun. Syst.*, vol. 52, no. 3, pp. 1467–1478, Mar. 2013.

[5] B. De Gelder, "Towards the neurobiology of emotional body language," *Nature Rev. Neurosci.*, vol. 7, no. 3, pp. 242–249, Mar. 2006.

[6] E. Crane and M. Gross, "Motion capture and emotion: Affect detection in whole body movement," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2007, pp. 95–101.

[7] D. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Understand.*, vol. 73, no. 1, pp. 82–98, Jan. 1999.
[8] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, to be published.
[9] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 15–33, Jan. 2013.
[10] Z. Yang and S. S. Narayanan, "Modeling dynamics of expressive body gestures in dyadic interactions," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 369–381, Jul. 2017.
[11] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *TiiSACM Trans. Interact. Intell. Syst.*, vol. 6, no. 1, pp. 1–31, Mar. 2016.
[12] N. Savva, A. Scarinzi, and N. Bianchi-Berthouze, "Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience," *IEEE Trans. Comput. Intell. AI in Games*, vol. 4, no. 3, pp. 199–212, Sep. 2012.
[13] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 254–267, Apr. 2017.
[14] X. Zhang, U. A. Ciftci, and L. Yin, "Mouth gesture based emotion awareness and interaction in virtual reality," in *Proc. ACM SIGGRAPH Posters (SIGGRAPH )*, 2015, p. 1.
[15] S. Shojaeilangari, W.-Y. Yau, K. Nandakumar, J. Li, and E. K. Teoh, "Robust representation and recognition of facial emotions using extreme sparse learning," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2140–2152, Jul. 2015.
[16] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features," 2014, *arXiv:1408.3750*. [Online]. Available: https://arxiv.org/abs/1408.3750
[17] H. Yoon, S.-W. Park, Y.-K. Lee, and J.-H. Jang, "Emotion recognition of serious game players using a simple brain computer interface," in *Proc. Int. Conf. ICT Converg. (ICTC)*, Oct. 2013.
[18] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 6, pp. 1052–1063, Nov. 2011.
[19] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr. 2012.
[20] Z. Zeng, M. Pantic, G. I. Roisman, and T. H. Huang, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Mar. 2009.
[21] R. McCraty, *The Coherent Heart*. Boulder Creek, CA, USA: Institute of HeartMath, 2006.
[22] G. Valenza, A. Lanata, and E. P. Scilingo, "Oscillations of heart rate and respiration synchronize during affective visual stimulation," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 4, pp. 683–690, Jul. 2012.
[23] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 827–834.
[24] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.
[25] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion Recognition: A review," in *Proc. IEEE 7th Int. Colloq. Signal Process. Appl.*, Mar. 2011.
[26] *All About Heart Rate (Pulse)*, Amer. Heart Assoc. Website. [Online]. Available: https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/all-about-heart-rate-pulse
[27] J. Hu and L. Fan, "Application of JADE to separate complex-valued sources," in *Proc. Int. Conf. Comput. Sci. Service Syst. (CSSS)*, Jun. 2011.
[28] W.-B. Kong, H.-X. Zhou, K.-L. Zheng, X. Mu, and W. Hong, "FFT-based method with near-matrix compression," *IEEE Trans. Antennas Propag.*, vol. 65, no. 11, pp. 5975–5983, Nov. 2017.
[29] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[30] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*. [Online]. Available: https://arxiv.org/abs/1508.01991
[31] E. Kiperwasser and Y. Goldberg, "Simple and accurate dependency parsing using bidirectional LSTM feature representations," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 313–327, Dec. 2016.
[32] F. Huang, X. Zhang, Z. Zhao, and Z. Li, "Bi-directional spatial-semantic attention networks for image-text matching," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2008–2020, Apr. 2019.
[33] R. Wang, X. Liang, X. Zhu, and Y. Xie, "A feasibility of respiration prediction based on deep bi-LSTM for real-time tumor tracking," *IEEE Access*, vol. 6, pp. 51262–51268, 2018.
[34] Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[35] K. Cohen, A. Nedic, and R. Srikant, "On projected stochastic gradient descent algorithm with weighted averaging for least squares regression," *IEEE Trans. Autom. Control.*, vol. 62, no. 11, pp. 5974–5981, Nov. 2017.
[36] N. Srivastava, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[37] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.
[38] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.
[39] H. Tahani and J. Keller, "Information fusion in computer vision using the fuzzy integral," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 3, pp. 733–741, May/Jun. 1990.
[40] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017.

**GUANGLONG DU** received the Ph.D. degree in computer application technology from the South China University of Technology, Guangzhou, China, in 2013. He is currently an Associate Professor with the Computer Science and Engineering School, South China University of Technology. His research interests include intelligent robotics, human–computer interaction, artificial intelligence, and machine vision.

**SHUAIYING LONG** received the B.S. degree from the School of Electronic Information Engineering, Xiangtan University. She is currently pursuing the master's degree with the School of Computer Science and Engineering, South China University of Technology. Her research interests include human–computer interaction, machine vision, and image generation.

**HUA YUAN** received the B.S. degree from Harbin Engineering University, Harbin, China, and the M.S. and Ph.D. degrees from Sichuan University, Chengdu, China. She is currently an Associate Professor with the Computer Science and Engineering School, South China University of Technology, Guangzhou, China. Her research interests include image processing, video communication, big data processing, and the next generation network architecture.