IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 22, 2020, accepted July 2, 2020, date of publication July 8, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007956

Dance Emotion Recognition Based on Laban Motion Analysis Using Convolutional Neural Network and Long Short-Term Memory

SIMIN WANG[1], JUNHUAI LI[1,2], (Member, IEEE), TING CAO[1,2], HUAIJUN WANG[1,2], PENGJIA TU[1], AND YUE LI[1]

[1]School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
[2]Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an 710048, China

Corresponding author: Junhuai Li (lijunhuai@xaut.edu.cn)

This work was supported in part by the National Key Research and Development Plan under Grant 2017YFB1402103, in part by the National Natural Science Foundation of China under Grant 61971347, and in part by the project of the Xi'an Science and Technology Planning Foundation under Grant 201805037YD15CG21(4).

ABSTRACT Dance emotion recognition technology is of great significance for the digitalization, virtual performance, inheritance and protection of folk dance. Based on the mechanism that emotion expression in dance performance can be fully expressed through the strength and rhythm of dance movements, a novel dance emotion expression method is proposed to train hybrid deep learning neural network, to effectively identify the seven basic dance emotions of fear, anger, boredom, excitement, joy, relaxation and sadness. First, in order to fully express the emotions contained in the dance movements, this paper defines a dance emotion expression method through Laban Movement Analysis (LMA) method, which includes the characteristic parameters of the three aspects of body structure, spatial orientation and force effect, and converts the original dance movement data into three characteristic expression parameters to obtain dance emotion data. Then, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) hybrid neural network models are used to test and train dance emotion data. Finally, in order to verify the applicability of the CNN-LSTM model, decision tree, random forest, CNN and LSTM are established and compared for accuracy. The results show that it is feasible to identify dance emotion from the perspective of dance movement, and the CNN-LSTM model is of high accuracy.

INDEX TERMS Dance emotion recognition, Laban motion analysis, CNN, LSTM.

I. INTRODUCTION

In recent years, emotion recognition has gradually become an important research direction in the field of humancomputer interaction. It is widely used for video, audio, etc. [1]. Emotion recognition is the process of analyzing various emotion information with a computer, extracting feature values describing emotions, establishing a mapping relationship between feature values and emotions, and then classifying emotions to infer the emotional state. At the same time, because emotional computing focuses on technologies and theories that promote understanding of human emotions, it has also attracted extensive attention and research [2].

At present, emotion recognition has been widely studied in the face [3], speech [4], and physiological signals [5], however, it is challenging to recognize facial and speech emotion information [3], [4]. With the extensive research of physiological signals in emotion recognition [5]–[7], it makes up for the shortcomings of emotion recognition in face and speech, but emotion recognition through brain wave signals has limitations and cannot be conveniently applied to human in everyday life.

Humans express emotions and communicate mainly through physical movements in daily life, the use of this kind of physical movement is very obvious in the emotional expression of dance, and emotion recognition is less applied in dance emotion recognition. Because dance moves are the external expression of dance emotions [8]. The emotion of dance movements is reflected in the body language and

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

movement form of the dancer [9]. In [10], the relationship between gestures and emotions in drama was analyzed, and the correlation between actions and emotions was verified. Therefore, this paper studies dance emotion from the perspective of dance movements.

For dance problems that are difficult to describe and analyze, laban action analysis method provides a good solution. Laban [11] proposed a method to describe dance movements, namely Laban Movement Analysis (LMA), in order to analyze dance movements scientifically. At present, the LMA method is widely used in emotion recognition as a description method of limb movements, for example, Ajili I [12] proposed a new human action description vector based on Laban action analysis method to recognize human expressions and actions on video images. Reference [13] based on LMA, the relationship between human movement and emotion was studied, and the results showed that there was a good correlation between LMA characteristics and emotion. In [14], Aristidou and Chrysanthou used various LMA features to classify dance performances with different emotions, and analyzed the changes of these characteristics in sports with different emotions, and found that there are different similarities in sports between different emotional states. Reference [15] with laban force characteristics as the input of neural network and establish famous russell circumplex model, realize the continuous human emotion recognition. Therefore, based on the applicability of Laban motion analysis method in body movements and emotional expression, this paper uses LMA method to analyze dance movements, extract LMA characteristic values from the movement data and obtain dance emotion data.

In emotion recognition, the common traditional algorithms are mainly regression analysis, support vector machine, k-means, apriori algorithm, etc. These algorithms can realize data processing and output of results faster, but cannot play a good role in identification when processing massive data. With the application and development of artificial intelligence in different fields, deep neural network is introduced into the field of emotion recognition, which provides a new solution for processing a large amount of emotion data and improving the accuracy of emotion recognition. Zhang et al. [4] applied CNN to image recognition and speech emotion recognition. The experimental results show that the accuracy of CNN in image recognition is 95.5%. In speech emotion recognition, the accuracy of CNN is 97.6%, both of which are higher than the traditional svm method. Reference [16] Using CNN to perform sentiment recognition on static images. Finally, the classifier showed 68.32% accuracy on the image set. C. Cheng proposed an emotion recognition algorithm based on convolutional neural network for emotional recognition of brainwave signals [5]. The experimental results show that the accuracy of the 2-category emotion recognition algorithm reaches 83.45% (the highest accuracy is 98.8%) and the highest accuracy of the 3-category recognition algorithm is 68.8%.

The main contributions of this paper are as follows:

1) A dance movement expression method based on LMA is proposed. This method is used to extract features from dance movement data to express dance emotions, and describes characteristic parameters from three aspects: limb structure, spatial orientation and force effect.
2) Due to the wide application and high efficiency of deep neural networks in emotion recognition, this paper combines CNN network and LSTM network to form a CNN-LSTM hybrid deep learning model for training dance emotion data, extracting dance emotion features and achieving dance emotion recognition.
3) The method proposed in the paper has achieved higher recognition efficiency by identifying the seven basic dance emotions of fear, anger, boredom, excitement, happiness, ease, and sadness, which is superior to traditional machine learning methods.

The article is outlined as follows, section II introduces a dance emotion expression method defined based on LMA method, and processes the dance movement data set to obtain the dance emotion data set. In section III, CNN and LSTM are introduced. In section IV, the results of dance emotion recognition obtained by using the proposed method are introduced and analyzed comprehensively. Finally, recommendations for possible future studies are presented in section V.

II. EXPRESSION OF DANCE MOVEMENTS BASED ON LMA

A. DANCE MOVEMENT DATA

Part of the collected dance movements is shown in Figure 1, the saving format of dance movements is bvh motion capture file. As shown in Figure 2, it is a schematic diagram of some human bone nodes in the bvh file.

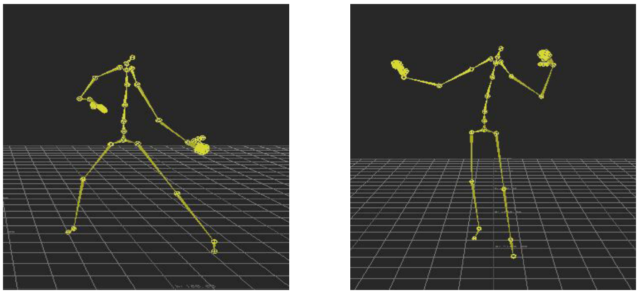

FIGURE 1. Dance moves.

This article uses the labeled dance emotion dataset of the University of Cyprus [17]. The dataset contains dance movement data of multiple different performers. Each bvh file describes 54 human bone nodes, and records the world coordinates and euler angle coordinates of each frame corresponding to 54 bone nodes in the data block part of the motion capture data area. In the experiment, this paper calculates the corresponding feature data of world coordinates and euler angles respectively, and uses the trained model to compare and verify the feature data of world coordinates, euler angles

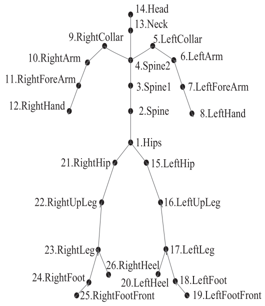

FIGURE 2. Part of human bone node.

and the combination of the two. The experimental results show that the feature data using world coordinates and euler angles simultaneously has a high accuracy.

B. EXPRESSION METHOD OF DANCE MOVEMENTS BASED ON LMA

Dance emotions are identified through dance action data. Therefore, it is necessary to analyze the dance movements. Because the traditional acceleration-based method only pays attention to the speed of limb movement changes in describing human movements, there are disadvantages of insufficient expression of motion. Therefore, this paper refers to the laban framework described by Aristidou *et al.* [18] to analyze the movements, and describes the LMA characteristic parameters from the aspects of limb structure, spatial orientation and force efficiency, and obtains the dance emotion data by calculating the corresponding LMA characteristic parameters.

1) SPATIAL ORIENTATION

The spatial orientation is shown in Figure 3, which is mainly divided into three vertical spatial orientations and nine horizontal spaces. The vertical orientation is centered on the hips node, offset up and down by a certain distance, and the human body is divided into three vertical orientations: low, medium and high. The horizontal orientation is divided into nine orientations, namely, home position, right front, right, right rear, rear, left rear, left, left front, front, each of which is divided by 22.5 degrees.

2) LIMB STRUCTURE

Limb structure includes relative distance between bone nodes, gait length, etc. Among them, this paper calculates the relative distance and gait between bone nodes, and compares the accuracy of these two eigenvalues in the following experiments. By comparison, this paper chooses to use the distance feature. The characteristic equation of the skeleton to distance is shown in Equation 1, where d is the bone pair distance,the coordinate of node i is (x_i, y_i, z_i), the coordinate of the neighbor node j of node i is (x_j, y_j, z_j).

$$d = \sqrt{(x_i - x_j)^2 - (y_i - y_j)^2 - (z_i - z_j)^2} \tag{1}$$

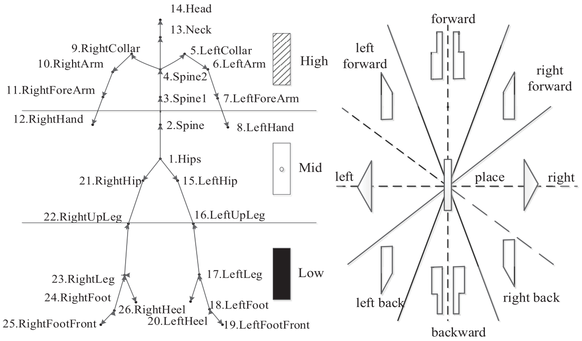

FIGURE 3. The spatial orientation.

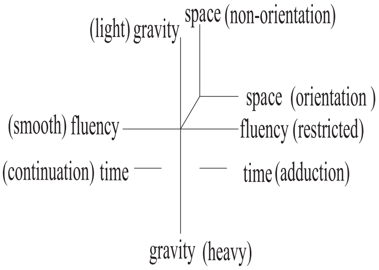

FIGURE 4. DSchematic diagram of force.

3) FORCE EFFECT

Force effect is a module related to emotional changes during the action. The force effect consists of four parts: gravity, time, fluency, and space. As shown in Figure 4, each attribute in the figure has two polarities, gravity includes light and heavy, expressing the gravity of gravity on the limb; time includes continuation and adduction, reflecting the speed of dance movements; fluency indicates the fluency and limitations of dance movements. Force effect describes gravity, time, and fluency characteristics by analyzing limb velocity and acceleration values. The space is divided into orientation and non-orientation, indicating the interaction between the limb and the external environment. Force effect describes the spatial characteristics by describing the orientation of the limb parts such as the head.

We calculate the velocity and acceleration of the joint points of the limbs. The characteristic equation of the velocity of the bone node is shown in Equation 2, where v is the speed of the bone nodethe coordinate of the current time of the i node is (x_i, y_i, z_i), the coordinate of the previous moment is $(x_{i-1}, y_{i-1}, z_{i-1})$, t is the time per frame. As shown in Equation 3, the characteristic equation of the acceleration of the bone node is shown, where a is the acceleration of the bone node, Δv is the speed increment, t is the time per frame.

$$v = \frac{\sqrt{(x_i - x_{i-1})^2 - (y_i - y_{i-1})^2 - (z_i - z_{i-1})^2}}{t} \tag{2}$$

$$a = \frac{\Delta v}{t} \tag{3}$$

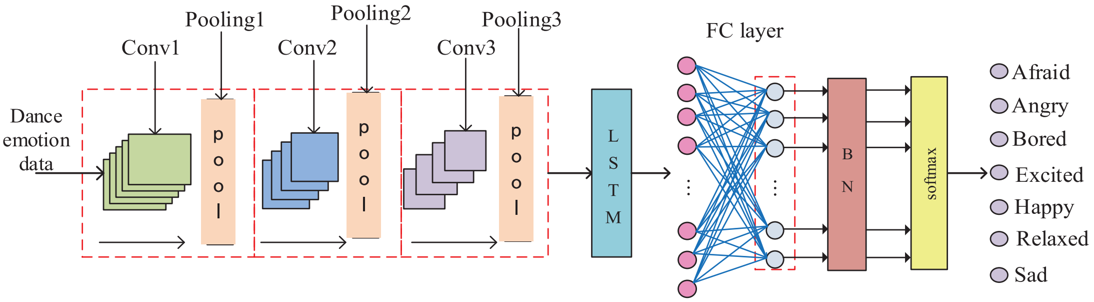

FIGURE 5. The general framework of the dance action emotion recognition method.

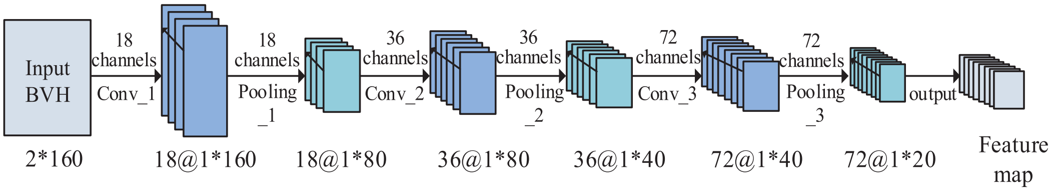

FIGURE 6. CNN model.

This article uses the above formula to calculate the relative distance between the limbs, the velocity and acceleration of the limb joint points for the world coordinates and euler angle data of each frame in the original bvh file data block, to obtain a data set expressing dance emotion. Among them, because the bvh file contains 54 human bone nodes, for each frame of data, the relative distance feature value has 52 data (there are actually 53 distance feature value data, but the head movement distance is limited in the experiment, so the relative distance of the head node is not considered), the velocity characteristic value has 54 data, and the acceleration characteristic value has 54 data. These three feature data are used to express the features of the dance movement of the current frame. Therefore, there are 160 data per frame. Since the world coordinate data and euler angle data are calculated separately, the amount of data per frame is 320.

Due to the short time of one frame, this article defines a time window of one second, that is, 30 frames, to process the distance, speed, and acceleration obtained above, that is, to calculate the average distance, average speed, and average acceleration every 30 frames. This paper uses these mean data as an emotion data set for experiments.

III. DANCE EMOTION RECOGNITION METHOD BASED ON CNN-LSTM HYBRID DEEP LEARNING MODEL

The overall framework of the dance emotion recognition method proposed in this paper is shown in Figure 5. First, the world coordinate point and euler angle data in the bvh motion capture file are calculated by LMA feature to obtain the dance emotion data set. Then, the 320 data of each frame in the dance emotional data set is dimensionally transformed into 2*160 input dimension, and input into the CNN to extract the dance emotional features. After processing by the threelayer convolution pooling layer, the eigenvalues are converted into feature vectors and input into the LSTM neural network for processing, then, feature fusion is performed through the fully connected layer, and the feature data is normalized using the bn layer. Finally, the softmax function is used to perform dance emotion classification recognition.

A. CNN-BASED EMOTION FEATURE EXTRACTION

The CNN network includes classic network models such as AlexNet [19], VGGNet [20], and GoogLeNet [21].

AlexNet Network is the first modern deep convolutional network model. Because the network size exceeds the memory limit of a single GPU at the time, the network uses two GPUs for parallel training. The VGGNet network is improved on the basis of the AlexNet network, which reduces the size of the convolution kernel and increases the number of convolution layers. GoogLeNet network is designed around network depth and convolution kernel width. Considering that the gradient disappears due to the excessive depth of the network, GoogLeNet adds two loss at different depths to ensure that the gradient back disappears.

The above classic CNN models all process color pictures with a depth of three, and the network structure is deep, which is not suitable for processing dance emotion data.

The convolutional neural network can be divided into five parts: input layer, convolution layer, pooling layer, full-connection layer and output layer [22], [32]. The structure of the convolutional neural network in this paper is shown in Figure 6. It consists of an input layer, a 3-layer convolution layer, and a 3-layer pooling layer.

The convolution layer performs convolution on the input sentiment data by using the size and number of custom convolution kernels, and performs convolutional layer processing to obtain the same feature map as the convolution kernel. Among them, after the dance emotion data is dimension converted, the input data size is 2*160. Since the general convolution layer mostly uses N*N sized convolution kernels, considering that each input data in this paper may contain dance emotional features, therefore, this paper uses the network adaptation data to change the shape of the convolution kernel to a long strip to accommodate the sample data. At the same time, in order to prevent the loss of important information caused by the incomplete extraction of emotional features of the dance, this article uses the "SAME" method to fill in the latter two layers of convolutional layers. Since convolution is a linear operation, in order to speed up feature extraction, an excitation function is added as an excitation layer after the convolutional layer of each layer, the excitation layer mainly increases the nonlinear operation. The excitation layer uses the relu activation function.

The pooling layer reduces the dimension of the feature map obtained by the convolution layer by downsampling operation to reduce the size of neurons and the number of parameters, and maintain the invariance of translation, rotation, and expansion of the network. Common pooling operations are average-pooling and max-pooling. Because the max-pooling method is able to select a feature value that is more recognizable for dance emotions, therefore, the max-pooling method is used for dimensionality reduction. Considering that the filter may not be able to process the data in a certain direction, it will lose the data containing important features, therefore, in the pooling operation, the input data is filled in the "SAME" mode, that is, the data edge is filled with 0. The convolutional layer and the pooled layer are stacked to form a deep structure, and features can be automatically extracted from the original data [23].

In order to reduce the loss of feature information, feature fusion is performed at the end of the network with a fully connected layer, and connect the fully connected layer to the softmax layer, seven different dance emotions are classified by softmax function, and probability values are output to realize dance action emotion recognition. The network parameters are set as follows:

The first layer of convolutional layer selects a one-dimensional convolution kernel with a size of 1*2, the number of convolution kernels is 18, and the convolution kernel step size is 1; the size of the one-dimensional convolution kernel of the second layer is 1*2, the number of convolution kernels is 36, and the convolution kernel step size is 1; the size of the one-dimensional convolution kernel selected by the third layer is 1*2, the number of convolution kernels is 72, and the convolution kernel step size is 1.

The first layer of the pooling layer has a window size of 1*2 and the pooling layer has a step size of 2, the window size and step size of the second and third pooled layers are the same as the first layer parameters, and each layer is filled with "SAME".

Through the CNN network, high-level dance emotional characteristics can be obtained, but because CNN is a deep feed-forward neural network, it does not have the ability to process time series data. From the time dimension, dance emotion expression is related, so the LSTM save and analyze the previous emotional information, realize the modeling of the time dependence of the data, and improve the accuracy of dance emotion recognition.

B. LSTM-BASED EMOTION FUSION AND RECOGNITION

RNN is a neural network with short-term memory function. When the input data sequence is long, there will be gradient explosion and disappearance problems. In order to solve this problem, a gating mechanism was introduced. Common gate-based cyclic neural networks for LSTM networks and gated loop unit (GRU) networks [24]. Although the GRU parameter is less than the LSTM, it is easy to converge, but in the case of a large input data set, the expression performance of the LSTM is better than that of the GRU. A variant of the three common LSTM networks is the LSTM model with no forgetting [25], the peephole connection [26], the coupled input gate and the forgetting gate. The LSTM model without forgetting gates has the problem that when the length of the input sequence is very large, the capacity of the memory unit will be saturated, thereby greatly reducing the performance of the LSTM model. In the peephole connection, the three gates depend not only on the implicit state of the input and the previous moment, but also on the memory unit at the previous moment, and the network structure is complicated. Coupling input gates and forget gates combines input gates and forget gates in the LSTM network into one gate, reducing the complexity of the network structure, but there are situations where new information may be added incompletely.

Therefore, based on the above comparison, we use a conventional LSTM model for modeling. And since the LSTM network can be viewed as a deep network in the time dimension, this paper chooses to use a layer of LSTM network.

The LSTM [27] includes input gates, forgetting gates, and output gates. Through the interaction between these gating units, the LSTM network has long-term memory function. However, there is a gradient disappearance problem in deep neural networks [28]. As the hidden layer increases, this phenomenon may even lead to a decrease in accuracy [29]. The intelligent design of the storage unit in the LSTM network can effectively solve the problem of gradient disappearance in back propagation.

The LSTM has three inputs: the input value of the network at the current time, the output value of the LSTM at the previous moment, and the unit status at the previous moment; the LSTM has two outputs: the current time LSTM output value, and the current unit state. Among them, the forgetting gate, the input gate, and the output gate together control the inflow of information of the neurons, and then obtain the predicted value of the LSTM unit through the tanh function. The working steps of LSTM are as follows:

The first step is to update the forgotten gate, the forgetting gate determines how much "memory" of the unit state at the previous moment can be retained to the current moment. As shown in Equation 4. In the following formula, σ and tanh are Activation function, w is input vector corresponding weight, h_{t-1} is the output of a neuron on a moment, x_t is input of the current moment of the neuron, b is offset.

$$f_t = \sigma(w_f \cdot [h_{t-1}, x_t + b_f]) \tag{4}$$

The second step updates the input gate two parts output, the input gate determines how much of the input at the current time is saved to the unit state. As shown in Equations 5 and 6.

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t + b_i]) \tag{5}$$

$$\tilde{c_t} = \tanh(w_c \cdot [h_{t-1}, x_t + b_c]) \tag{6}$$

The third step combines the formulas 5 and 6 with the output f of the forgetting gate to update the unit status. As shown in Equation 7.

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c_t} \tag{7}$$

The fourth step updates the output gate, and how much information is output from the output gate control unit. As shown in Equations 8.

$$o_t = \sigma(w_o \cdot [h_{t-1}, x_t + b_o]) \tag{8}$$

Finally, a plurality of vectors including the time and dance emotional data sequences outputted by the LSTM layer are input to the fully connected layer to implement feature fusion.

Since the output value of the nerve cell exceeds the appropriate range of the activation function itself before passing the activation function, the nerve cell may fail to work [30]. In order to solve this problem, this article introduces the bn [31], the batch normalization method is as follows:

```
**Algorithm 1**: BN
  **Input**: input dance emotional feature data $x_1, x_2, ..., x_m$,
        $B = \{x_1, x_2, ...x_m\}$
  **Output**: output $y$
  Calculated data mean, $\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$;
  Calculated data variance, $\sigma_B^2 \leftarrow \frac{1}{m} \sum_{i=1}^{m}(x_i - \mu_B)^2$;
  Data standardization, $\hat{x}_i \leftarrow \frac{(x_i - \mu_B)}{\sqrt(\sigma_B^2 + \varepsilon)}$;
  Training parameter $\gamma, \beta, y_i \leftarrow \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$;
  Output $y$;
  Return $\gamma, \beta$;
```

After the bn layer processes, the output data is passed to the softmax layer for classification. The softmax activation function is shown in Equation 9 and is used to calculate the probability value for the multi-class output. Softmax processed multiple values obtained by the neural network to make the probability value between [0, 1] and get the predicted tag value of the data, among which the largest tag value is the classification result.

$$softmax(y_i) = \frac{exp(y_i)}{\sum_j exp(y_j)} \tag{9}$$

C. NETWORK MODEL TRAINING

Network model training is a supervised learning process, divided into forward and backward propagation.

In the forward propagation, the dance motion data is first preprocessed, and the dance emotion data is obtained. Then, the data is sequentially extracted through the CNN network and the LSTM network, and the obtained output feature sequence is normalized. Finally, the softmax function is used to classify and output an n-dimensional vector (n is the classification result). The category corresponding to the element with the highest probability is the classification result of the current network on the input sample.

In the case of backpropagation, the cross-entropy loss function is used to calculate the error value of the actual classification of the prediction classification and the input data as shown in Equation 10, in the formula, y_i is the actual label of the sample, $\hat{y}_i$ is the calculated sample predicted value, N is the total number of samples. At the same time, with the chain rule, the initial parameter learning rate is set, the gradient descent algorithm is used to adjust the weight and offset in the network, and the adam optimizer is used to dynamically adjust the learning rate of each parameter to optimize the network parameters to minimize the objective function.

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i log\hat{y}_i + (1 - y_i) log(1 - \hat{y}_i) \tag{10}$$

Due to the complexity of the network structure, in order to prevent over-fitting, this article adds the dropout layer behind the CNN and LSTM layers respectively. By setting the dropout-rate, it is determined whether the current neuron is discarded, thereby finding a more ''thin'' network in the original network. Dropout achieves good collaboration by forcing a neuron to work with other randomly selected neurons, attenuating and eliminating the joint adaptability between neuron nodes, and enhancing the generalization ability of the model.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL DATA

The development of dance emotion recognition research is inseparable from the support of the dance action database. Currently, the publicly available dataset of dance actions with emotion labels is collected by the University of Cyprus [17]. Therefore, this paper uses this data set for experiments. The experiment selects 7 dance data sets with different dance emotion labels for 6 people, the data set contains dances of seven emotional markers, afraid, angry, bored, excited, happy, relaxed, and sad. Since the calculated dance sentiment data contains tens of thousands of frames of data, it cannot be entered into the network model all at once. Therefore, this paper enters according to the amount of data per frame. However, since the data of each frame is a one-dimensional sequence, dimension transformation is required before inputting the network model. Considering that there are 320 data in each frame, this article performs dimensional conversion on the data of each frame according to the size of 2 * 160, and enters the neural network.

In the experiment, the dance emotional data set has a total of 1050 data. First, the data set is divided into 900 training sets and 150 test sets, and training samples and test samples are obtained respectively. Then the training samples are further divided into 70% training set and 30% verification set, that is, the original data is divided into three parts: namely: 630 training set, 270 verification set and 150 test set. The training set is used for model training, the verification set is used to adjust parameters, and the test set is used to measure the quality of the final model.

B. CNN MODEL TRAINING

The parameters adjusted in the CNN network mainly include: convolution kernel size, learning rate, and batch size.

According to the structure of Figure 6, the initial setting learning rate is 0.002, each batch sample is 100, epoch is set to 400. The recognition rate of the CNN initial model based on the above settings on the test set is 90%.

In order to set a reasonable size convolution kernel in the CNN model, experiments were performed on convolution kernels of sizes 1, 2, 3, 4, 5, 6, 7 and 8 respectively. The experimental results are shown in Figure 7. It can be obtained that when the convolution kernel size is 3, the accuracy of the model in the test set is higher.

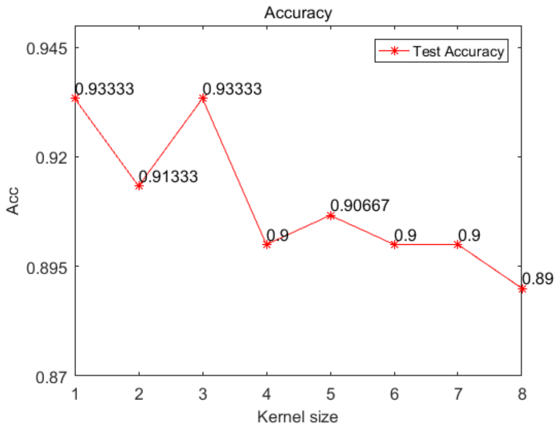

FIGURE 7. Results of different convolution kernels.

The learning rate determines the speed at which the network reaches its optimal state. In order to obtain the optimal learning rate, experiments were carried out at different learning rates of 0.0001, 0.0002, 0.0004, 0.0006, 0.0008, 0.001, 0.002, 0.005 and 0.01 respectively. As shown in Figure 8, it can be concluded from the figure that the highest accuracy rate is when the learning rate is 0.0001.

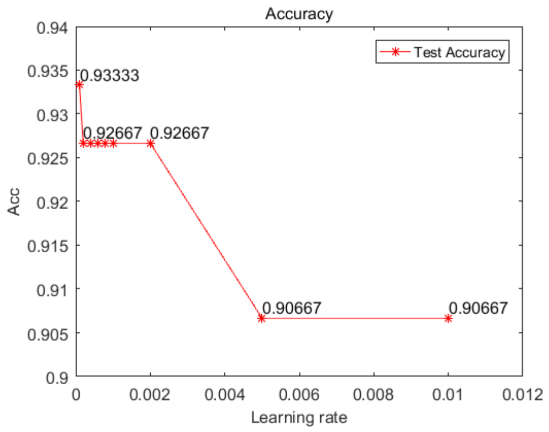

FIGURE 8. Results of different learning rates.

Batch size is the batch sample size, and its maximum value is the total number of samples in the training set. When the amount of data is small, the batch data is a full data set, which can more accurately approach the extreme value direction. However, in practical applications, due to the large amount of data used for deep learning, the principle of small batch processing is generally adopted. Less memory is required when using small batch processing, and training networks are faster. In the appropriate range, increasing the batch size value can more accurately determine the direction of the gradient descent, resulting in less training oscillations. However, when the batch size increases to a certain value, the determined downward direction does not change substantially, and the correction of the parameters is also significantly slowed down.

In order to obtain a suitable batch size value, this article experiments and compares the recognition accuracy of different batch sizes under the test set. As shown in Table 1, when the Batch size is 150, the accuracy is the highest on the test set. Therefore, the batch size is 150.

TABLE 1. Test set accuracy under different batch sizes.

Batch size	Accuracy	Batch size	Accuracy
25	0.9266667	225	0.93333334
50	0.9266667	250	0.93333334
75	0.92	275	0.94
100	0.93333334	300	0.9266667
125	0.93333334	325	0.92
150	0.94	350	0.9266667
175	0.92	375	0.92
200	0.9266667	400	0.9266667

Based on the above experimental analysis and results, the experimental parameters of the CNN model are determined as shown in Table 2. After the test, the accuracy of the characteristic values of gait length, speed, and acceleration on the test set reached 75.33%, and the accuracy of the characteristic values of distance, speed, and acceleration on the test set reached 94%.

In this paper, experiments are carried out to extract corresponding eigenvalues from world coordinates and euler angles. In the above CNN model, the accuracy of world coordinates is 89.14%, and the accuracy of euler angles is 91.97%. Through comparison, it is verified that using both coordinate data at the same time can obtain higher accuracy.

TABLE 2. CNN model parameters.

Parameters	Values
Input vector size	160
Number of input channels	2
Convolution kernel size	3
Pool size	2
Activation function	Relu function
Learning rate	0.0001
Batch sample size	150
Epoch	400

C. LSTM MODEL TRAINING

The parameters adjusted on the LSTM mainly include: bn layer, number of neurons, learning rate.

LSTM takes each window data as the input of the network, that is, the input of LSTM is a sequence of 2 * 160. The LSTM layer initially has a dropout of 0.5, setting the learning rate to 0.0015, the batch size to 50, and the epoch to 400. Based on the above parameters, the accuracy of the model in the test set is 92%.

In order to verify the influence of the number of neurons in the LSTM layer on the recognition results, this article verifies the accuracy of the number of different neurons in 6, 12, 18, 24, 30, 36, 42 on the test set. As shown in Figure 9, the accuracy is highest when the number of neurons is 36.

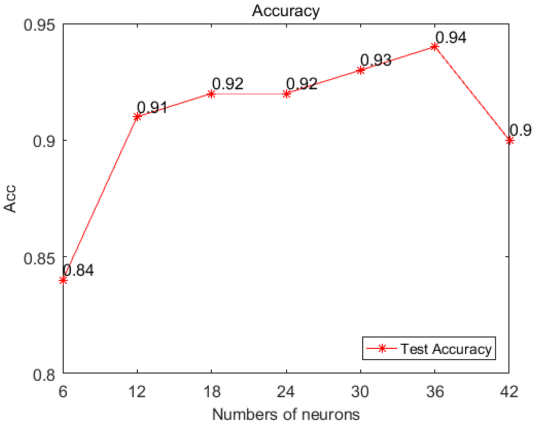

FIGURE 9. Accuracy of different neuron numbers.

Accuracy experiments were performed on the LSTM model at different learning rates of 0.0015, 0.0025, 0.0035, 0.0045, 0.0055, 0.0065. As shown in Figure 10, when the learning rate is 0.0015, the model has the highest accuracy on the test set, so the learning rate is set to 0.0015.

According to the analysis of the above experimental results, the parameters of the LSTM model are determined as shown in Table 3. After testing, the accuracy of the characteristic values of gait length, speed, and acceleration on the test set reached 84.67%, and the accuracy of the characteristic values of distance, speed, and acceleration on the test set reached 94%.

In this paper, experiments were carried out to extract the corresponding eigenvalues of world coordinates and euler angles. In the above LSTM model, the accuracy of world coordinates is 88%, and the accuracy of euler angles is 91.71%. Through comparison, it is verified that using both coordinate data at the same time can obtain higher accuracy.

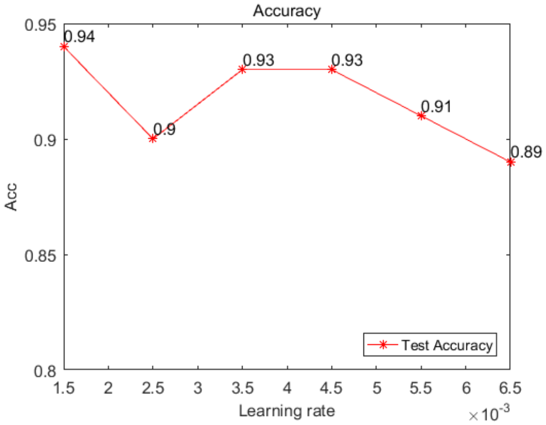

FIGURE 10. Accuracy of different learning rates.

TABLE 3. LSTM model parameters.

Parameters	Values
Input vector size	160
Number of input channels	2
LSTM layer number	2
Number of neurons per layer	36
Dropout	0.5
Learning rate	0.0015
Batch sample size	50
Epoch	400

D. CNN-LSTM MODEL TRAINING

By training the CNN and LSTM models separately, they have better recognition accuracy in the test set, but the recognition performance needs to be improved. Because the CNN-LSTM model combines the advantages of CNN and LSTM, it not only extracts deep features, but also preserves the temporal relationship of data, which has great quality for feature extraction of dance emotions. Therefore, this article combines the CNN model with the LSTM model to establish the CNN-LSTM model for dance emotion recognition.

After the same adjustment method as above, the accuracy of using the characteristic values of gait length, speed and acceleration on the test set reaches 75%, and the accuracy of using the characteristic values of distance, speed and acceleration on the test set reaches 97%.

The parameter values corresponding to the CNN-LSTM model are shown in table 4. Figure 11 shows the confusion matrix of 7 different dance emotion recognitions. The loss curve and acc curve of the CNN-LSTM model are shown Figure 12 and 13.

In this paper, experiments are carried out to extract the corresponding eigenvalues of world coordinates and euler angles. In the above CNN-LSTM model, the accuracy of world coordinates is 89%, and the accuracy of euler angles

TABLE 4. CNN-LSTM model parameters.

Parameters	Values
Input vector size	160
Number of input channels	2
Convolution kernel size	2
Pool size	2
Activation function	Relu function
LSTM layer number	1
Number of neurons per layer	42
Dropout	0.08
Learning rate	0.001
Batch sample size	50
Epoch	400

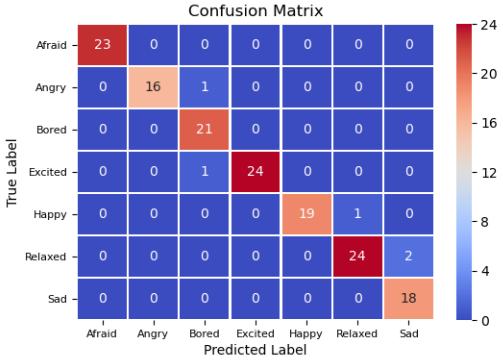

FIGURE 11. CNN-LSTM model confusion matrix.

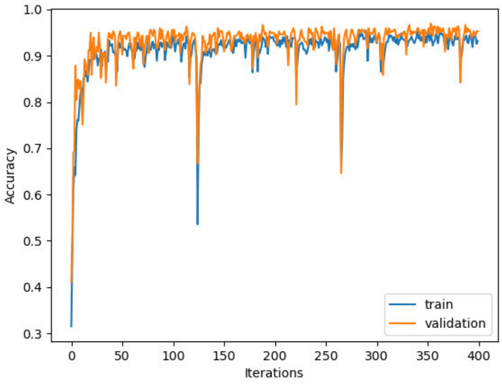

FIGURE 12. The accuracy of CNN-LSTM model.

is 90%. Through comparison, it is verified that using both coordinate data at the same time can obtain higher accuracy.

E. EXPERIMENTAL RESULT

Table 5 shows the recognition rates of decision trees, random forests, CNN, LSTM, and CNN-LSTM models on the test set. It can be seen from the table that the recognition rates of decision trees, random forests, CNN, LSTM and CNN-LSTM models are all above 90%, and the recognition accuracy of CNN-LSTM models is the highest. As shown in Table 6, there are three types The accuracy rate of the model under 7 different dance emotions, as can be seen from Table 6, the CNN-LSTM model's recognition effect on a single dance emotion is slightly higher than that of the CNN model and LSTM model.

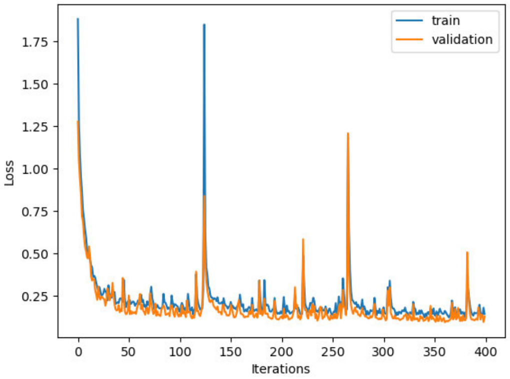

FIGURE 13. The loss rate of the CNN-LSTM model.

TABLE 5. Recognition rate of different models on the test set.

Method	Average accuracy
Decision tree	94.48%
Random forest	92.95%
CNN	94%
LSTM	94%
CNN-LSTM	97%

TABLE 6. Accuracy of different models in 7 dance emotions.

	CNN-LSTM	CNN	LSTM
Afraid	0.94	0.88	0.90
Angry	0.93	0.86	0.89
Bored	0.95	0.90	0.90
Excited	0.95	0.89	0.88
Happy	0.95	0.90	0.87
Relaxed	0.95	0.88	0.82
Sad	0.95	0.86	0.85
Average	0.95	0.88	0.87

V. CONCLUSION

Dance emotions are influenced by many factors such as actors, movements, scenes, etc. Identifying emotions from dance is a challenging question. This paper proposes a dance emotion recognition method based on CNN-LSTM hybrid deep learning model, which combines the two network structures of LSTM and CNN, and verifies the effectiveness of the method through contrast experiments. The experimental results show that the CNN-LSTM model has the highest accuracy rate of 97%, which indicates that dance emotion recognition should focus on analysis of dependencies instead of simply judging characteristic data. For the recognition of single dance emotions, the average accuracy of CNN-LSTM model is 95%. In both respects, the accuracy of the CNN-LSTM model is higher than that of the CNN model and the LSTM model. The effect of dance emotion recognition is greatly affected by the quality and quantity of training samples. Improving the robustness of dance emotion recognition and expanding the dance emotion data set need to be studied urgently.

ACKNOWLEDGMENT

The authors are grateful for the anonymous reviewers who made constructive comments.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. 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, Jan. 2009, doi: 10.1109/TPAMI.2008.52.

[2] F. Guo, F. Li, W. Lv, L. Liu, and V. G. Duffy, "Bibliometric analysis of affective computing researches during 1999 2018," *Int. J. Hum.–Comput. Interact.*, vol. 36, no. 9, pp. 801–814, 2020, doi: 10.1080/10447318.2019.1688985.

[3] T.-H.-S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, "CNN and LSTM based facial expression analysis model for a humanoid robot," *IEEE Access*, vol. 7, pp. 93998–94011, 2019, doi: 10.1109/ACCESS.2019.2928364.

[4] B. Zhang, C. Quan, and F. Ren, "Study on CNN in the recognition of emotion in audio and images," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2016, pp. 1–5, doi: 10.1109/ICIS.2016.7550778.

[5] C. Cheng, X. Wei, and Z. Jian, "Emotion recognition algorithm based on convolution neural network," in *Proc. 12th Int. Conf. Intell. Syst. Knowl. Eng. (ISKE)*, Nanjing, China, Nov. 2017, pp. 1–5, doi: 10.1109/ISKE.2017.8258786.

[6] S. Ramírez-Gallego, A. Fernández, S. García, M. Chen, and F. Herrera, "Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce," *Inf. Fusion*, vol. 42, pp. 51–61, Jul. 2018.

[7] M. Chen, Y. Qian, Y. Hao, Y. Li, and J. Song, "Data-driven computing and caching in 5G networks: Architecture and delay analysis," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 70–75, Feb. 2018, doi: 10.1109/MWC.2018.1700216.

[8] N. Shikanai and K. Hachimura, "Evaluation of impressions and movements related to negative emotional expressions in dance," in *Proc. 15th Int. Conf. Control, Autom. Syst. (ICCAS)*, Busan, South Korea, Oct. 2015, pp. 657–660, doi: 10.1109/ICCAS.2015.7365000.

[9] R. Fan, S. Xu, and W. Geng, "Example-based automatic music-driven conventional dance motion synthesis," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 3, pp. 501–515, Mar. 2012, doi: 10.1109/TVCG.2011.73.

[10] M. Kipp and J.-C. Martin, "Gesture and emotion: Can basic gestural form features discriminate emotions?" in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Amsterdam, The Netherlands, Sep. 2009, pp. 1–8, doi: 10.1109/ACII.2009.5349544.

[11] L. R. Von and L. Ullmann, *The Mastery of Movement*. London, U.K.: Northcote House, 1988.

[12] I. Ajili, M. Mallem, and J.-Y. Didier, "Human motions and emotions recognition inspired by LMA qualities," *Vis. Comput.*, vol. 35, no. 10, pp. 1411–1426, Oct. 2019, doi: 10.1007/s00371-018-01619-w.

[13] J. Morita, Y. Nagai, and T. Moritsu, "Relations between body motion and emotion: Analysis based on Laban Movement Analysis," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, Berlin, Germany, 2013, pp. 1026–1031.

[14] A. Aristidou and Y. Chrysanthou, "Motion indexing of different emotional states using LMA components," *SIGGRAPH Asia Tech. Briefs*, vol. 2013, p. 21, Nov. 2013.

[15] S. Senecal, L. Cuel, A. Aristidou, and N. Magnenat-Thalmann, "Continuous body emotion recognition system during theater performances," *Comput. Animation Virtual Worlds*, vol. 27, nos. 3–4, pp. 311–320, May 2016.

[16] H. Abanoz and Z. Cataltepe, "Emotion recognition on static images using deep transfer learning and ensembling," in *Proc. 26th Signal Process. Commun. Appl. Conf. (SIU)*, Izmir, Turkey, May 2018, pp. 1–4, doi: 10.1109/SIU.2018.8404346.

[17] *Motion Captured Performances*. Accessed: Jul. 8, 2019. [Online]. Available: http://dancedb.eu/main/performances?tdsourcetag=s_pcqq_aiomsg

[18] A. Aristidou, P. Charalambous, and Y. Chrysanthou, "Emotion analysis and classification: understanding the performers' emotions using the LMA entities," *Comput. Graph. Forum*, vol. 34, no. 6, pp. 262–276, Sep. 2015, doi: 10.1111/cgf.12598.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1106–1114.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, 2014.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[22] J.-T. Huang, J. Li, and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4989–4993, doi: 10.1109/ICASSP.2015.7178920.

[23] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016, doi: 10.3390/s16010115.

[24] K. Cho, B. V. Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. Neural Comput.: New Challenges Perspect. New Millennium (IJCNN)*, Como, Italy, 2000, pp. 189–194, doi: 10.1109/IJCNN.2000.861302.

[27] A. Graves, "Generating sequences with recurrent neural networks," 20113, *arXiv:1308.0850*. [Online]. Available: https://arxiv.org/abs/1308.0850

[28] E. Kanjo, E. M. G. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Inf. Fusion*, vol. 49, pp. 46–56, Sep. 2019, doi: 10.1016/j.inffus.2018.09.001.

[29] H. Chiou-Jye and K. Ping-Huan, "A deep CNN-LSTM model for particulate matter (PM$_{2.5}$) forecasting in smart cities," *Sensors*, vol. 18, no. 7, p. 2220, 2018, doi: 10.3390/s18072220.

[30] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U..K., May 2019, pp. 5866–5870, doi: 10.1109/ICASSP.2019.8682283.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 20115, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[32] Y. He, Z. Zhang, F. R. Yu, N. Zhao, H. Yin, V. C. M. Leung, and Y. Zhang, "Deep-Reinforcement-Learning-Based optimization for cache-enabled opportunistic interference alignment wireless networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10433–10445, Nov. 2017, doi: 10.1109/TVT.2017.2751641.

SIMIN WANG is currently pursuing the master's degree in computer science and technology with the Xi'an University of Technology. Her research fields include the Internet of Things technology and emotion recognition.

JUNHUAI LI (Member, IEEE) received the B.S. degree in electrical automation from the Shaanxi Institute of Mechanical Engineering, Xi'an, China, in 1992, the M.S. degree in computer application technology from the Xi'an University of Technology of China, Xi'an, in 1999, and the Ph.D. degree in computer software and theory from the Northwest University of China, Xi'an, in 2002. He is currently a Professor with the School of Computer Science and Engineering, Xi'an University of Technology, China. His research interests include the Internet of Things technology and network computing.

TING CAO received the Ph.D. degree from Chang'an University, Xi'an, China, in 2018. From 2016 to 2017, he was a Visiting Scholar funded by the China Scholarship Council with the University of Waterloo, Canada. He is currently a Lecturer with the Xi'an University of Technology. His research interests include artificial intelligence, computer vision, and pattern recognition.

HUAIJUN WANG received the B.Sc. and M.Sc. degrees in computer science from the Xi'an University of Technology, in 2005 and 2010, respectively, and the Ph.D. degree from Northwest University, Xi'an, China, in 2014. He is currently a Lecturer with the Xi'an University of Technology. His research interests include the application and security of CPS and modeling of effectiveness evaluation of security.

PENGJIA TU received the B.S. degree in network engineering from the Shaanxi University of Technology, China, in 2016, and the M.S. degree in computer system architecture from the Xi'an University of Technology, Xi'an, China, in 2019, where she is currently pursuing the Ph.D. degree in computer science and technology.

YUE LI is currently pursuing the master's degree in software engineering with the Xi'an University of Technology. Her research interests include the Internet of Things technology and human behavior recognition.