frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 24 August 2022
DOI 10.3389/fnbot.2022.998568

Global and local feature fusion *via* long and short-term memory mechanism for dance emotion recognition in robot

Yin Lyu[1] and Yang Sun[2]*

[1]College of Music, Huaiyin Normal University, Huai'an, China, [2]College of Software, Shenyang Normal University, Shenyang, China

Check for updates

OPEN ACCESS

EDITED BY
Peng Li,
Dalian University of Technology, China

REVIEWED BY
Asif Ali Laghari,
Sindh Madressatul Islam
University, Pakistan
Shahid Karim,
University of Sialkot, Pakistan

*CORRESPONDENCE
Yang Sun
352720214@qq.com

RECEIVED 20 July 2022
ACCEPTED 08 August 2022
PUBLISHED 24 August 2022

CITATION
Lyu Y and Sun Y (2022) Global and local feature fusion *via* long and short-term memory mechanism for dance emotion recognition in robot. *Front. Neurorobot.* 16:998568.
doi: 10.3389/fnbot.2022.998568

In recent years, there are more and more intelligent machines in people's life, such as intelligent wristbands, sweeping robots, intelligent learning machines and so on, which can simply complete a single execution task. We want robots to be as emotional as humans. In this way, human-computer interaction can be more natural, smooth and intelligent. Therefore, emotion research has become a hot topic that researchers pay close attention to. In this paper, we propose a new dance emotion recognition based on global and local feature fusion method. If the single feature of audio is extracted, the global information of dance cannot be reflected. And the dimension of data features is very high. In this paper, an improved long and short-term memory (LSTM) method is used to extract global dance information. Linear prediction coefficient is used to extract local information. Considering the complementarity of different features, a global and local feature fusion method based on discriminant multi-canonical correlation analysis is proposed in this paper. Experimental results on public data sets show that the proposed method can effectively identify dance emotion compared with other state-of-the-art emotion recognition methods.

KEYWORDS

dance emotion recognition, robot, LSTM, feature fusion, linear prediction coefficient

Introduction

Today is an era of artificial intelligence technology explosion, the demand for human-computer interaction (HCI) technology (Yu et al., 2020; Liu et al., 2022) is also increasing. Among them, emotion recognitionis an indispensable part of this technology. Facial expression is an important signal of a person's emotional state. Together with speech, hand and body posture, it forms the basic communication system of human beings in social environments. Whether we can provide perfect service for human beings according to human emotions, the key problem is to accurately identify human emotions, so as to meet human needs more intelligently (Chowdary et al., 2021; Kashef et al., 2021). Therefore, the direction of emotion recognition attracts many scholars to conduct research.

Of course, human emotions not only contain facial expressions. In real life, people can express their emotions in various forms, such as voice information, music information, physiological signals and text information, etc., which are more conducive to emotion recognition to some extent (Abbaschian et al., 2021). For example, when people is excited, people speak faster and may be accompanied by dancing gestures; When they are sad, people will droop their face and eyes, speak slowly, and may support their face with their hands. At this point, when the emotion occurs, it will also cause a certain degree of physiological changes. In addition, the corresponding emotions can also be identified through text messages. It is not enough to identify emotions only by one feature. Human emotions are inherently diversified, and features extracted by multiple modes of multiple features are more comprehensive.

Emotions can be recognized from the so-called body language, face-play, and speech. Most of their characteristics are changing with age, education, experience, etc. Moreover, there is variability among speakers, their body language, and facial expressions (Kacur et al., 2021). Emotion recognition can greatly promote the integration and development of many different disciplines, such as graphics and image processing, artificial intelligence, human-computer interaction and psychology (Jiang and Yin, 2021; Shen et al., 2021). In human-computer interaction scenes with many different modes, the combination of emotion, posture, sound and other modes can make human-computer interaction experience more real. In addition, the study of dance emotion has great application value in many fields. For example:

(1) Game development. Game developers can identify players' facial expressions and determine the preferences of the majority of players, so that they can change the design scenario, difficulty or scheme of a game to provide a better experience for players.

(2) Online teaching. Through the terminal operating system real-time acquisition of the students in the class facial expressions, timely detection of students interested in the teacher teach content, state of the students in class lectures are in good condition (Yu, 2021), whether the student to the teacher speak content understanding and grasp, and can identify to feedback the result to the teaching system, so convenient teacher in time according to the results of the identification of teaching activities and scheme adjustment to develop more effective learning strategies for students.

(3) Safe driving. Sensors installed in the car can monitor the owner's facial expressions in real time, detect the current driving state of the driver, if the driver is detected in the state of fatigue driving, will timely alarm sound, remind the driver to stay awake, to avoid the occurrence of tragedy.

(4) Medical system (Chen et al., 2021a). Design a medical machine that can recognize facial expressions, and timely tracking and detecting the patients' facial expressions. When a patient's facial expression is recognized as pain, the machine system can sound an alarm to call the medical staff, so that the patient care is more efficient, more intelligent and humane.

The emotional features contained in voice signals in audio can be expressed from the speaker's pitch, accent weight and speed, etc. Audio features reflecting certain emotions can be roughly divided into three categories: spectral features, prosodic features and tone quality features (Wang and Wang, 2021). Most of the methods to identify emotions through speech signals adopt the prosodic features of sound, among which the fundamental frequency and amplitude of sound are the most effective for emotion recognition (Murugappan and Mutawa, 2021). However, in the actual research process, it is not accurate to make judgment only by using a certain feature. The characteristics of speech emotion are not only prosodic, but also tone quality and spectrum. Asghar et al. (2022) proposed that using amplitude and frequency spectral features (MSFs) and mel-frequency cepstral coefficients (MFCCs), perceptual weighted linear predictive (PLP) and perceptual features had achieved good speech emotion recognition effects. Chouhan et al. (2021) used CNN and SVM to classify and recognize speech emotion and achieved good results. Kaur and Kumar (2021) adopted CNN for speech emotion recognition in the data set, which greatly improved the ability of speech emotion recognition. At present, the popularity of CNN model is also applied in the field of speech emotion recognition, including short and long short-term memory network (LSTM), recurrent neural network (RNN) (Yadav et al., 2021), etc,. Mohanty and Palo (2020) proposed to extract prosodic and spectral parameters of audio, and then used probabilistic neural network (PNN) and hidden markov model (HMM) to extract prosodic and spectral parameters of audio. HMM processed these two kinds of parameter features (Dai et al., 2021).

At present, there are two kinds of emotion feature extraction methods: static texture feature based and dynamic texture feature based. Emotion recognition based on static texture features is to extract the key frame of the video expression first, and replace the whole video expression recognition result with the key frame recognition result. Although it improves the speed of emotion recognition and eliminates a lot of redundant information, it lacks the time domain information. The method based on dynamic texture features contains this time domain information, and its research data is a video sequence or dynamic video. Feature extraction methods based on static texture are representative of PCA, LDA, ICA, LBP, Gist, and Gabor transform (Karim et al., 2019; Shafiq et al., 2021; Yin et al., 2021). Feature extraction methods based on dynamic texture include LBP-TOP, PHOG-TOP, LPQ-TOP, etc.

At present, although great progress has been made in the field of dance emotion recognition, there are still some problems that can not be ignored, that is, low recognition efficiency, different results with disunity of database. In order to better solve the problems faced by dance emotion recognition and further improve its practical application, this paper proposes a new dance emotion recognition based on global and local feature fusion method.

The structure of this paper is organized as follows. Section "Proposed dance emotion recognition" introduces the proposed dance emotion recognition method in detail. Then, we conduct rich experiments for the proposed method in section Experiments and analysis. There is a conclusion in section Conclusion.

Proposed dance emotion recognition

The occurrence of dance emotion is a dynamic process, which contains both time domain information and space domain information. Considering the temporal and spatial characteristics of audio features, a new feature extraction algorithm based on linear prediction Mayer frequency cepstrum coefficient (LPMFCC) is proposed. At present, the extraction of dance emotional features is mostly based on a single voice feature, which can only reflect one attribute of voice information, not the global information of expression, and the dimension of data features is very high. In this paper, we adopt LPMFCC to extract the local feature and LSTM to extract global feature. Considering the complementarity of different features, this paper proposes a dance emotion recognition system based on global and local feature fusion. The adopted feature fusion method in this paper is the latest feature fusion framework based on kernel entropy component analysis+ discriminant multiple canonical correlation analysis (KECA + DMCCA).

KECA works by projecting raw data into higher-dimensional space to Eigen decomposing the Kernel matrix (Chen et al., 2021b). The eigenvector with the maximum eigenvalue is selected to form a new data space. It is underpinned by Renyi entropy and Parzen window. KECA can resolve the problem of the linear inseparability of the other model and enhances the separability between features.

The framework diagram of proposed dance emotion recognition in this paper is shown in Figure 1. The proposed algorithm includes three main steps: preprocessing, feature extraction and classification. The feature extraction process extracts a set of global features and a set of local features respectively. After feature extraction, feature dimension is higher and invalid information is more, so the effective feature fusion framework KECA+DMCCA is adopted after feature extraction. This framework can not only fuse multiple groups of information, but also greatly reduce the feature dimension.

Dance emotion preprocessing

In order not to cause a lot of information redundancy, and not to lose the corresponding key information, but also to retain certain emotional time domain information, here, we adopt a face detection scheme based on the HSV color model (Bobbadi et al., 2022). In the HSV color model, H and S components represent color information, and V represents brightness information. The HSV color model is closely related to human's intuition on color. The RGB component of an image can be converted to HSV color space using the following formula:

$$H = \{H_1 \ if \ B \leq G; 360^\circ - H_1 \ if \ B > G\} \tag{1}$$

where

$$H_1 = \cos^{-1}\{\frac{0.5[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\} \tag{2}$$

$$S = \frac{\max(R,G,B)-\min(R,G,B)}{\max(R,G,B)} \tag{3}$$

$$V = \frac{\max(R,G,B)}{255}. \tag{4}$$

We use the plane envelope approximation (Lee and Pietruszczak, 2021) to approximate human skin color. In the planar envelope method, a pixel is considered a skin pixel if its color meets the following two conditions:

$$\begin{aligned} & S \geq Th_s;\ S \leq -H - 0.1V + 110 \\ & H \leq -0.4V + 75;\ V \geq Th_v && (5) \\ & If H \geq 0, S \leq 0.08(100-V)H + 0.5V && (6) \\ & \text{Otherwise } S \leq 0.5H + 35 && (7) \end{aligned}$$

LPMFCC for local feature extraction

Linear prediction is a common method for speech analysis. It can not only get the prediction waveform of speech signal, but also provide a very good channel model. The main idea is that there is correlation between sampling points of speech signal. The sampled values of the speech signal at a certain time can be approximated by the linear combination of the sampled values at the previous time so that the waveform of the speech signal can be estimated and predicted. In order to determine the linear prediction coefficient of speech samples, it is necessary to minimize the mean square error between the linear prediction

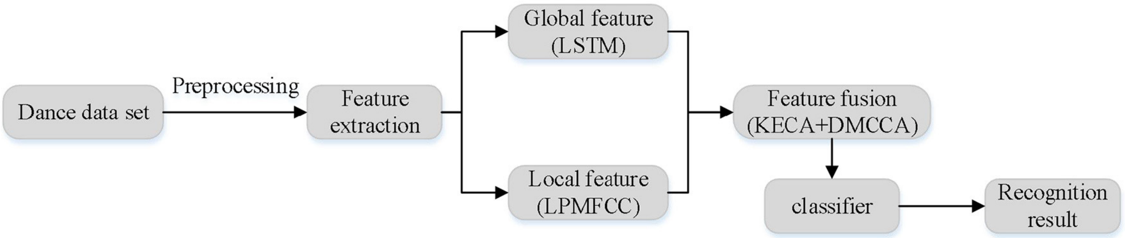


FIGURE 1
Flow chart of proposed dance emotion recognition.

sample value and the actual speech sample value. The linear prediction coefficient reflects the characteristics of speech signal.

According to the above ideas, the linear prediction coefficient is calculated. After preprocessing the speech signal, the p-order linear prediction is to predict the sampling value $\{s(n-1), s(n-2), \cdots, s(n-p)\}$ at this moment by using the linear combination of sampling values at the previous p times of the speech signal $s(n)$, and the obtained prediction signal $\hat{s}(n)$ is:

$$\hat{s}(n)=\sum_{k=1}^{p} a_k s(n-k) \tag{8}$$

where a_k is the linear prediction error formed by the linear prediction coefficient.

$$e(n)=s(n)-\hat{s}(n)=s(n)-\sum_{k=1}^{p} a_k s(n-k). \tag{9}$$

In order to optimize the prediction effect, it is necessary to minimize the mean square value of the prediction error. The formula for the mean square value of the prediction error is:

$$\varepsilon=E[e^2(n)] \tag{10}$$

In order to minimize the mean square value of the prediction error, it is necessary to take the partial derivative of the mean square value of the prediction error formula and make it zero, as shown in Formula (11).

$$\frac{\partial[e^2(n)]}{\partial a_k}=0, k=1,2,\cdots,p \tag{11}$$

And we can get:

$$s(n-i)(n)=\sum_{k=1}^{p} a_k s(n-k)s(n-i), i=1,2,\cdots,p \tag{12}$$

If we define:

$$\varphi(i,k)=s(n-i)s \tag{13}$$

Then equation (12) can be changed as the formula (14).

$$\varphi(i,0)=\sum_{k=1}^{p} a_k \varphi(i,k), i=1,2\cdots p \tag{14}$$

Obviously, the linear prediction coefficient a_k can be obtained by solving the equation obtained by Formula (14). In this paper, the auto-correlation method and Levinson-Durbin recursion method are used to solve the equations. The prediction coefficients obtained by the above algorithms represent the feature vectors of speech frames, namely LPC feature parameters, and its extraction process is shown in Figure 2.

Linear prediction Mayer frequency cepstrum coefficient is a new characteristic parameter combining LPC and MFCC characteristic parameters. LPC parameters reflect the linear characteristics of speech, but have the disadvantage of being greatly disturbed by environmental noise. The MFCC parameters reflect the nonlinear characteristics of speech, and transform the actual frequency of speech to the Merle frequency that conforms to the auditory characteristics of human ear (Sirimontree et al., 2021). When the actual frequency is <1 kHz, the relationship between Mayer frequency and actual frequency is approximately linear. When the actual frequency is >1 kHz, the relationship between the Meir frequency and the actual frequency can be approximated as a pairwise number. The general expression of the relationship between Mayer frequency and actual frequency is:

$$f_{mel}=2958\cdot\log_{10}(1+f/700) \tag{15}$$

Where f_{mel} represents the Mayer frequency and f represents the actual frequency. Figure 3 shows that MFCC parameters are relatively sensitive to the low-frequency part of speech. However, ambient noise is in the high frequency part of speech.

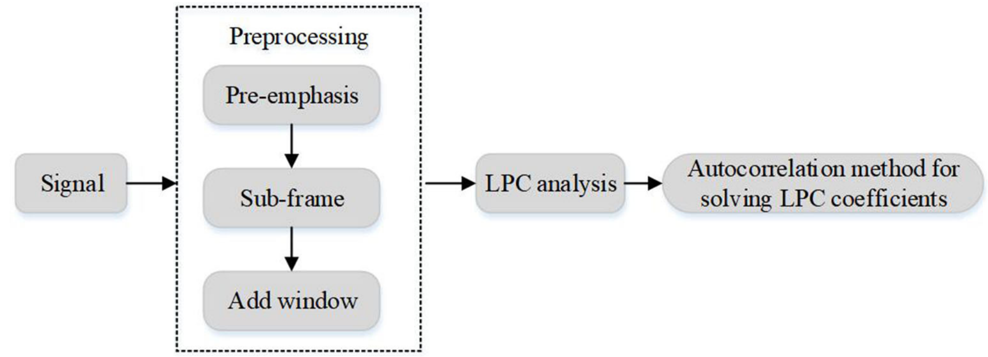


FIGURE 2
LPC coefficient extraction process.

Therefore, MFCC parameters have strong anti-interference ability and good robustness to environmental noise. The LPMFCC parameter is actually the LPC cepstrum parameter that converts the LPC parameter into Meyer frequency.

The LPMFCC feature extraction of speech first needs to extract the LPC coefficient of speech. After the preprocessing of speech signal $x(n)$, such as pre-emphasis, subframe and adding window, the LPC coefficient $x_a(n)$ of each speech frame is calculated. The order of the LPC coefficient should be set equal to the number of voice samples in a frame. Secondly, the cepstrum of LPC coefficient is calculated on Meyer frequency. First, Fourier transform is made for LPC coefficient, then LPC coefficient is executed by DFT to obtain the corresponding discrete spectrum $X_a(k)$, namely:

$$X_a(k) = \sum_{n=0}^{N-1} x_a(n)e^{-(j2pnk/N)}, 0 \leq k \leq N-1. \tag{16}$$

We will take square amplitude spectrum calculation for $X_a(k)$, and obtain the discrete energy spectrum $|X_a(k)|^2$. Where N is the point number of the Fourier transform. Then a set of meyerscale triangular filters are used to filter the discrete energy spectrum. The logarithmic operation is performed on the output result to obtain the logarithmic energy $Z_a(m)$, and the formula is as follows.

$$Z_a(m) = In(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)), 0 \leq m \leq M \tag{17}$$

The $H_m(k)(0 \leq m \leq M)$ is a number of band pass filter. M is the number of filters. Finally, a new characteristic parameter LPMFCC is obtained by calculating the logarithmic energy by discrete cosine transform.

$$C_a(n) = \sum_{m=0}^{M-1} Z_a(m) \cos\left[\frac{pn(m+0.5)}{M}\right]. \tag{18}$$

To sum up, it can be seen that the calculation method of LPMFCC characteristic parameters refers to the calculation method of MFCC coefficient and carries out cepstrum calculation of LPC coefficient under Mayer frequency. The specific extraction process is shown in Figure 3. In addition, the LPMFCC feature Y_i extracted from voice signal S_i is denoted as $Y = \{Y_1, \cdots, Y_T\}$. The average eigenvector $\hat{Y}$ is used to represent the features of speech signal S, where $\hat{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$. T represents the frame number of speech signal S.

LSTM for global feature extraction

In this section, by constructing the basic model of distorted FRI signals, the characteristic sequence of distorted signals is determined to be the weighted sum of multiple copies of different delay in original signals. Therefore, LSTM network is considered to be used to construct an auto-encoder to obtain the feature sequence estimation of distorted FRI. We design a novel LSTM to extract the global features.

Distorted FRI signal model

The FRI distortion signal is the weighted sum of several known pulses in different delay copies. Multipath effect is caused by echo in real scene. Therefore, distorted FRI signal $x'(t)$ can be expressed as:

$$x'(t) = x(t) + \sum_{i=0}^{l-1} a_i x(t - t_i) \tag{19}$$

$$= \sum_{p\in Z}\left(\sum_{k=0}^{K-1} c_k\phi\left(t - t_k - pT_\tau\right) + \sum_{i=0}^{l-1} a_i\left(\sum_{k=0}^{K-1} c_k\phi\left(t - t_k - t_i - pT_\tau\right)\right) + \varepsilon(t)\right) \tag{20}$$

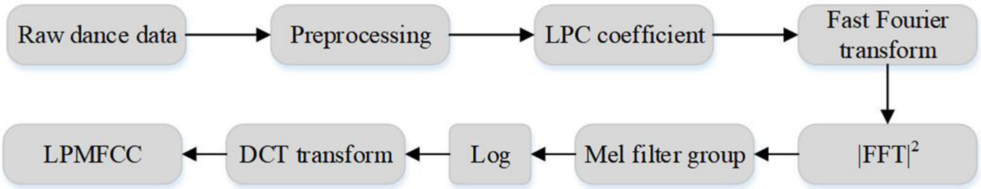


FIGURE 3
LPMFCC parameter feature extraction process.

where l indicates that there are a total of l paths to reflect the original FRI signal. a_i represents the reflection coefficient of path i. t_i represents the delay of path i. $\varepsilon(t)$ is additive White Gaussian noise.

The distorted FRI signal shown in Equation (19) is sent to the FRI sampling system, and $n_u = T_\tau / T_s$ is defined. Sub-nyquist sampling samples obtained by the sampling system can be expressed as:

$$y_n = \sum_{i=0}^{l-1} a_i \left(\sum_{k=0}^{K-1} c_k \delta(t - t_k - t_i - pT_\tau) \right). \tag{21}$$

LSTM network for FRI reconfiguration

In this paper, LSTM network is considered to be used to encode FRI distorted signals, extract signal feature sequences, and train network parameters by minimizing the cost function shown in Equation (19), as shown in Figure 4. LSTM network model consists of input layer, LSTM layer, full connection layer and output layer. The input of the network is the sample y_n obtained from FRI sampling, and the length of the sample is $N = 4K + 1$. The LSTM layer is composed of several LSTM units, which mainly learn the hidden features contained in sample y_n. The full-connection layer maps and reduces the dimension of the waveform features learned by LSTM layer, and the output layer outputs the features estimated by LSTM network. The LSTM model $f(W, b, y_n)$ is jointly determined by the three basic gate units in the LSTM structure and the cell state and output at the last moment. Under the supervision of the expected feature sequence y_m^{Sinc}, according to the cost function shown in Equation (19), parameters W and b in the model can be updated by the stochastic gradient descent algorithm to obtain the mapping weighting coefficient $\beta_{m,n}^{f} = \{\beta_{m,n}^{0}, \beta_{m,n}^{l\neq 0}\}$ between the input sample y_n and the feature sequence y_m^{Sinc}.

The forgetting gate in the LSTM model determines the retention and discarding of waveform information in the cell state of the LSTM unit at the last moment (Wang et al., 2022), and reads the output h_{t-1} of the LSTM unit at the last moment and the input $y_{n,t}$ of the LSTM unit at the current moment. The information is then filtered through the activation function *sigmoid(x)*. According to equation (20), the output of the forgetting gate can be expressed as:

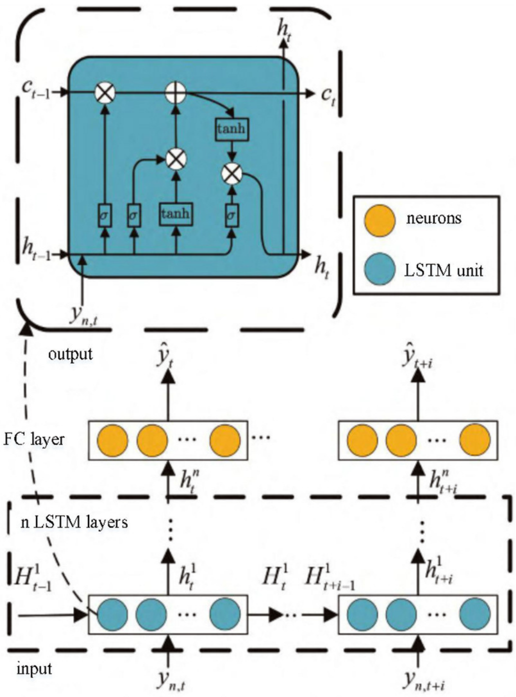


FIGURE 4
Waveform prediction model based on LSTM.

$$f_t = \sigma(W_f \cdot [h_{t-1}, y_{n,t}] + b_f) \tag{22}$$

where σ represents the sigmoid function.

The forgetting gate outputs a number between 0 and 1, and controls the forgetting degree of the cell state C_{t-1} at the previous moment by multiplying it by the cell state C_{t-1} at the previous moment. When the forgetting gate output is equal to 1, it means that the cell state information of the last moment is completely retained. When the output is equal to 0, it means that the cell state information at the last moment is completely forgotten. If distorted waveform information exists in the sampled samples, the interaction between distortion free pulse and distortion pulse in FRI signal expression (20) can be fully utilized to eliminate the riding variable $\sum_{i=0}^{l-1} a_i(\sum_{k=0}^{K-1} c_k\delta(t - t_k - t_i - pT_\tau)) \cdot \phi'(t/T_s - n - n_l)$ in the waveform through the selection of forgetting gate. So that distorted mode waveform information does not affect the cell state at the current time.

The input gate in the model determines to add new information to the cell state of the LSTM unit at the last moment. It reads the output h_{t-1} of the LSTM unit at the last moment and the input $y_{n,t}$ of the LSTM unit at the current moment, activates it through the activation function Sigmoid and obtains the candidate vector through the activation function $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. The input layer expression can be expressed as:

$$i_t = \sigma(W_i \cdot [h_{t-1}, y_{n,t}] + b_i) \tag{23}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, y_{n,t}] + b_c) \tag{24}$$

The input gate extracts y_m^{Sinc}, the characteristic sequence of delay information and amplitude information in the sample, and records the characteristics of delay information and amplitude information to generate candidate vector $\tilde{C}_t$. Delay information and amplitude information in cell state were updated through the interaction of candidate vector and input gate output. The LSTM unit decides to add part of y_m^{Sinc} in the sample input at this time to the cell state through the combined action of candidate vector and input gate output. And update the cell state through the information of partial distorted waveform at the forgetting time of the forgetting gate, specifically expressed as:

$$C_t = f_t^* C_{t-1} + i_t * \tilde{C}_{t-1} \tag{25}$$

The output gate determines the final output of the LSTM unit at that moment. It is determined by the updated cell state, the output of LSTM unit at the previous moment and the input at the current moment, and its expression is:

$$o_t = \sigma(W_o \cdot [h_{t-1}, y_{n,t}] + b_o) \tag{26}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{27}$$

H_* in Figure 4 includes cell state C_* and output h_*. According to the model structure, the final result can be estimated as:

$$\hat{y}_m^{Sinc} = w_1 h_t + b_1 \tag{28}$$

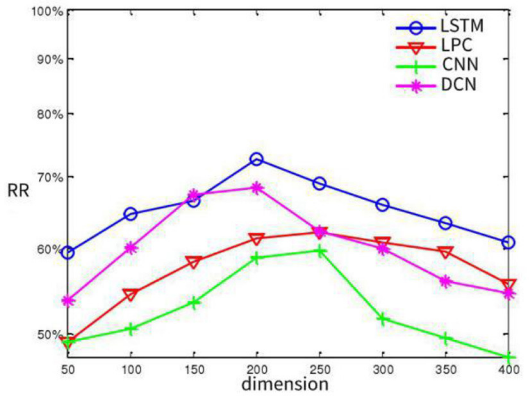
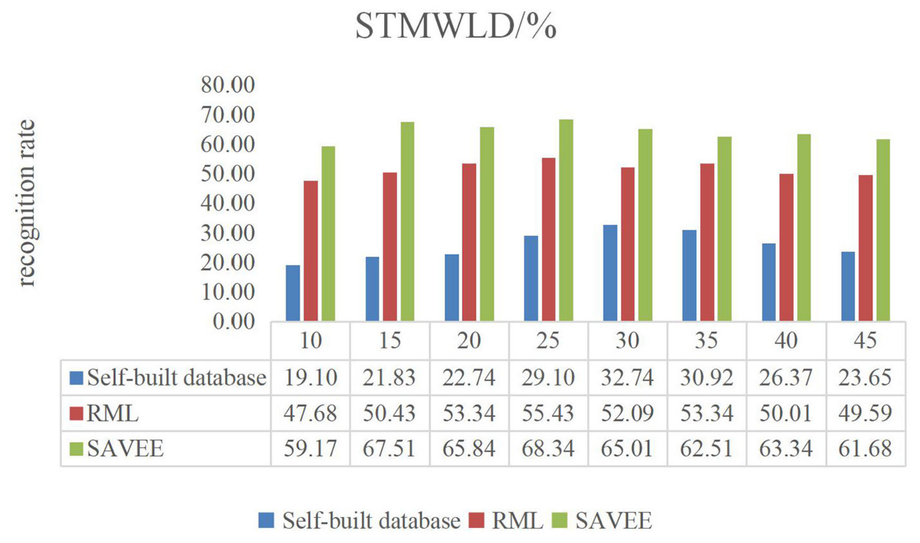


FIGURE 5
The relation between recognition rate and dimension diagram.

Experiments and analysis

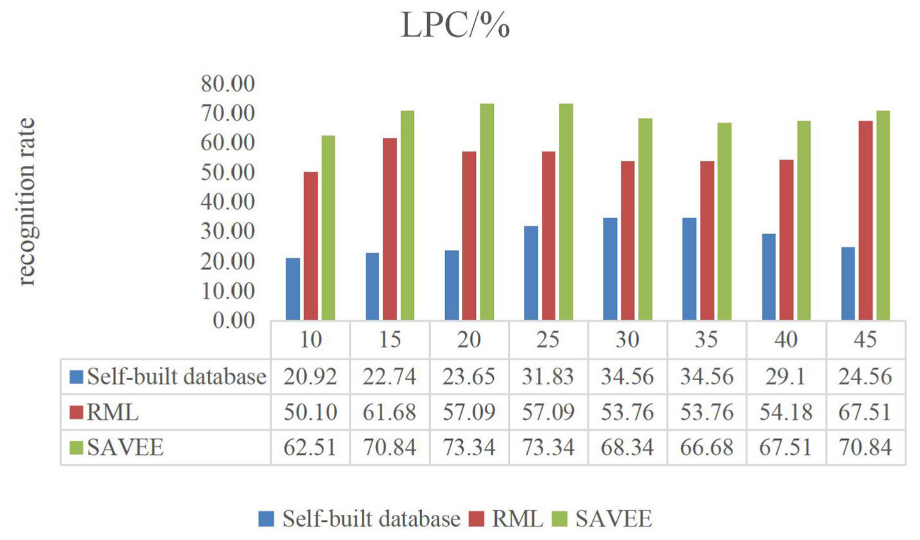
Based on the proposed mentioned algorithm in this paper, we use RML, SAVEE and self-built dance video database to make experiments. The RML database contains 720 samples, 480 short dance videos are used as training samples, and 240 short dance videos are used as test samples. SAVEE database has a total of 480 samples, among which there are 120 neutral expressions. This paper only studies the basic six types of expressions excluding neutral expressions. Then 240 short videos are used as training samples and 120 short videos are used as test samples in the experiment based on SAVEE database. In the experiment of self-built database, 240 short videos are used as training samples, among which 110 short videos are used as test samples. The proposed multi-feature extraction and fusion algorithm and support vector machine are used to achieve sentiment classification. All the experimental simulation environment in this paper is based on the experimental results of Windows 10 and MATLAB 2017a. The final experimental results and analysis are described in detail below.

According to the feature extraction method mentioned above, the experiment on SAVEE database is taken as an example to determine the appropriate feature dimension reduction. We extract 1002-dimensional LSTM feature, 512-dimensional LPC feature, 753-dimensional CNN feature and 786-dimensional DCN feature. In order to fuse suitable effective features and facilitate subsequent data fusion, each feature extraction algorithm is adopted separately. Observing the relationship between dimensionality reduction and recognition rate to determine the appropriate dimensionality reduction and reduce the overall system computation, we conduct experiments on two databases respectively to observe the relationship between the highest recognition rate and the dimensionality reduction of each feature, and the results are shown in Figure 5. On the



	10	15	20	25	30	35	40	45
🟦 Self-built database	19.10	21.83	22.74	29.10	32.74	30.92	26.37	23.65
🟥 RML	47.68	50.43	53.34	55.43	52.09	53.34	50.01	49.59
🟩 SAVEE	59.17	67.51	65.84	68.34	65.01	62.51	63.34	61.68

FIGURE 6
STMWLD dance video experiment result.



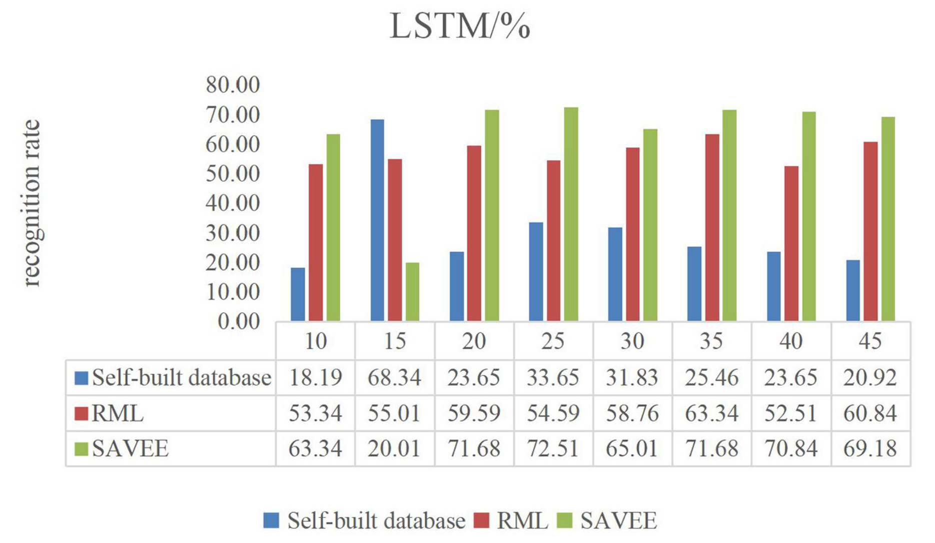
	10	15	20	25	30	35	40	45
🟦 Self-built database	20.92	22.74	23.65	31.83	34.56	34.56	29.1	24.56
🟥 RML	50.10	61.68	57.09	57.09	53.76	53.76	54.18	67.51
🟩 SAVEE	62.51	70.84	73.34	73.34	68.34	66.68	67.51	70.84

FIGURE 7
LPC dance video experiment result.

whole, it can be observed from the experimental figure that when the number of dimensionality reduction features are about 200 dimensions, the recognition rate of each feature extraction is the best.

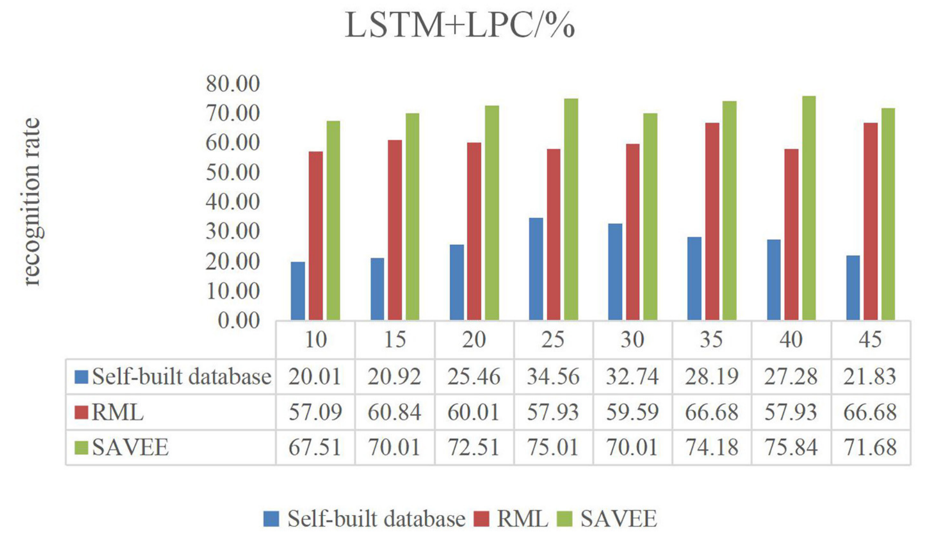
Figures 6–9 show the experimental results of STMWLD feature extraction algorithm alone, LPC-based local information, LSTM-based global information and LSTM + LPC based global information respectively.

The experimental results of the above four figures show the relationship between the recognition rate of dance expression and the number of beats when the dimensionality reduction is 200. In the video, one frame is selected every five frames. The purpose of selecting the frame number is to determine the maximum period of expression from the recognition rate on the one hand. On the other hand, the trend of the experimental results indicates that the occurrence of expression is a process from beginning to maximum and then to end to some extent. In general, although the feature algorithm STMWLD alone has the highest recognition rate of 32.74% in the self-built database, it has the highest recognition rate of 55.43 and 68.34% in the other



	10	15	20	25	30	35	40	45
🟦 Self-built database	18.19	68.34	23.65	33.65	31.83	25.46	23.65	20.92
🟥 RML	53.34	55.01	59.59	54.59	58.76	63.34	52.51	60.84
🟩 SAVEE	63.34	20.01	71.68	72.51	65.01	71.68	70.84	69.18

FIGURE 8
LSTM dance video experiment result.



	10	15	20	25	30	35	40	45
🟦 Self-built database	20.01	20.92	25.46	34.56	32.74	28.19	27.28	21.83
🟥 RML	57.09	60.84	60.01	57.93	59.59	66.68	57.93	66.68
🟩 SAVEE	67.51	70.01	72.51	75.01	70.01	74.18	75.84	71.68

FIGURE 9
LSTM + LPC dance video experiment result.

two standard databases respectively, indicating the effectiveness of our proposed algorithm. The recognition rates of LPC and LSTM fusion are 34.56, 62.51, and 73.34%, respectively. When LSTM was used to extract features, the recognition results of these three databases were 33.65, 63.34, and 72.51%, respectively. The recognition rates of LSTM and LPC fusion extraction are 34.56, 66.68, and 75.84%, respectively. This also shows that different databases have a certain influence on the experimental recognition results, and the fusion of two features is better than the recognition effect of a single feature. According to the average recognition rate obtained in Figures 8, 9, the recognition rate of facial expressions using only global features is higher than that using only local features, and about 3% higher than that using only local features.

Our method is compared with other methods, including STMWLD, LPC, LPC+STMWLD, and LSTM + LPC, as shown in Table 1. As can be seen from the data in Table 1, when a single feature is adopted, some discriminative facial expression information may be lost, resulting in low recognition rate and unsatisfactory recognition effect. Compared with other methods

TABLE 1 Expression recognition rate table of each feature extraction algorithm/%.

Feature extraction method	Self-built	RML	SAVEE
STMWLD	32.74	55.43	68.34
LPC	33.65	63.34	72.51
LPC + STMWLDCNN	34.56	62.51	73.34
LSTM + LPC	**55.46**	**76.26**	**85.84**

The bold values indicate the best values obtained by proposed method.

TABLE 2 Comparison between different methods and feature fusion methods in this paper.

Method	Recognition rate/%	Recognition time/s
HOG	35.81	4.6
CNN	63.21	2.5
Att-Net (Kwon, 2021)	75.13	2.1
CTNet (Lian et al., 2021)	76.89	1.7
CCML (Zehra et al., 2021)	73.54	1.3
Proposed	**85.84**	**0.5**

The bold values indicate the best values obtained by proposed method.

in the table, the method proposed in this paper has the best recognition rate, which is 76.26 and 85.84% for the two databases respectively. The recognition rate of natural expressions is 55.46%, which shows the effectiveness of the proposed method in real natural scenes. The biggest advantage of this method is that it combines local features with global features, and includes dynamic time domain feature information. These complementary features are more conducive to facial expression recognition in video. On the other hand, the experimental results also demonstrate the effectiveness of the proposed method.

Table 2 shows the comparison of the experimental results of feature extraction methods in this paper and those in other references. The results of experiments with different databases are also different. The table is to compare with the experimental results of feature extraction methods in other references under the condition that the selected data sets are consistent with the public data sets used in this paper as far as possible, and the relatively new references are selected for comparison with the features in this paper under the condition that the comparison standards are consistent as far as possible. It is obvious from the table that the proposed feature fusion method is superior to other feature extraction methods.

Conclusion

Dance emotion recognition based on video is a challenging and long-term problem. The emotion in video is easily disturbed by various factors. This paper proposes an effective multi-feature fusion framework to solve the problem of video expression recognition, and studies the recognition effect of video expression in natural and real scenes. The system framework of LPC algorithm and LSTM fusing complementary and multi-feature is introduced, and then these features are fusing with KECA+DMCCA framework. Finally, SVM classifier is used to realize the recognition of six basic expressions. Experiments on two public databases (RML, SAVEE) and self-built databases prove the effectiveness of the proposed feature extraction algorithm, and the experimental results also show that the recognition effect of multi-feature fusion is better than that of single feature. In the future works, we will research more advanced deep learning methods to improve the emotion recognition.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbaschian, B. J., Sierra-Sosa, D., and Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 21, 1249. doi: 10.3390/s21041249

Asghar, A., Sohaib, S., Iftikhar, S., Shafi, M., and Fatima, K. (2022). An Urdu speech corpus for emotion recognition. *PeerJ Comput. Sci.* 8, e954. doi: 10.7717/peerj-cs.954

Bobbadi, C., Nalluri, E., Chukka, J., Wajahatullah, M., and Sailaja, K. L. (2022). *HsvGvas: HSV Color Model to Recognize Greenness of Forest Land For the Estimation of Change in The Vegetation Areas. Computer Vision and Robotics.* Singapore: Springer. p. 265–280. doi: 10.1007/978-981-16-8225-4_21

Chen, X., Cao, M., Wei, H., Shang, Z., and Zhang, L. (2021a). Patient emotion recognition in human computer interaction system based on machine learning method and interactive design theory. *J. Med. Imaging Health Inf.* 11, 307–312. doi: 10.1166/jmihi.2021.3293

Chen, X., Ke, L., Du, Q., Li, J., and Ding, X. (2021b). Facial expression recognition using kernel entropy component analysis network and DAGSVM. *Complexity.* 2021, 6616158. doi: 10.1155/2021/6616158

Chouhan, K., Singh, A., Shrivastava, A., Agrawal, S., Shukla, B, D., Tomar, P. S., et al. (2021). Structural support vector machine for speech recognition classification with CNN approach. In: *2021 9th International Conference on Cyber and IT Service Management (CITSM).* Bengkulu: IEEE, 1–7. doi: 10.1109/CITSM52892.2021.9588918

Chowdary, M. K., Nguyen, T. N., and Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Comput. Appl.* 1–18. doi: 10.1007/s00521-021-06012-8

Dai, Z., Zhang, S., Wang, X., Wang, H., Zhou, H., Tian, S., et al. (2021). Sub-second transient activated patterns to sad expressions in major depressive disorders discovered *via* hidden Markov model. *J. Neurosci. Res.* 99, 3250–3260. doi: 10.1002/jnr.24942

Jiang, M., and Yin, S. (2021). Facial expression recognition based on convolutional block attention module and multi-feature fusion. *Int. J. Comput. Vis. Robot.* doi: 10.1504/IJCVR.2022.10044018

Kacur, J., Puterka, B., Pavlovicova, J., and Oravec, M. (2021). On the speech properties and feature extraction methods in speech emotion recognition. *Sensors* 21, 1888. doi: 10.3390/s21051888

Karim, S., Zhang, Y., Yin, S., Laghari, A. A., and Brohi, A. A. (2019). Impact of compressed and down-scaled training images on vehicle detection in remote sensing imagery. *Multimed. Tools Appl.* 78, 32565–32583. doi: 10.1007/s11042-019-08033-x

Kashef, M., Visvizi, A., and Troisi, O. (2021). Smart city as a smart service system: human-computer interaction and smart city surveillance systems. *Comput. Human Behav.* 124, 106923. doi: 10.1016/j.chb.2021.106923

Kaur, J., and Kumar, A. (2021). *Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest. Computer Networks and Inventive Communication Technologies.* Singapore: Springer, 499–509. doi: 10.1007/978-981-15-9647-6_39

Kwon, S. (2021). Att-Net: enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput.* 102, 107101. doi: 10.1016/j.asoc.2021.107101

Lee, Y. K., and Pietruszczak, S. (2021). Limit equilibrium analysis incorporating the generalized hoek-brown criterion. *Rock Mech. Rock Eng.* 54, 4407–4418. doi: 10.1007/s00603-021-02518-8

Lian, Z., Liu, B., and Tao, J. (2021). CTNet: conversational transformer network for emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 29, 985–1000. doi: 10.1109/TASLP.2021.3049898

Liu, Y., Sivaparthipan, C. B., and Shankar, A. (2022). Human-computer interaction based visual feedback system for augmentative and alternative communication. *Int J Speech Technol.* 25, 305–314. doi: 10.1007/s10772-021-09901-4

Mohanty, M. N., and Palo, H. K. (2020). Child emotion recognition using probabilistic neural network with effective features. *Measurement* 152, 107369. doi: 10.1016/j.measurement.2019.107369

Murugappan, M., and Mutawa, A. (2021). Facial geometric feature extraction based emotional expression classification using machine learning algorithms. *PLoS ONE* 16, e0247131. doi: 10.1371/journal.pone.0247131

Shafiq, M., Tian, Z., Bashir, A. K., Du, X., and Guizani, M. (2021). CorrAUC: a malicious Bot-IoT traffic detection method in iot network using machine-learning techniques. *IEEE Internet Things J.* 8, 3242–3254. doi: 10.1109/JIOT.2020.3002255

Shen, W., Chen, J., Quan, X., and Xie, Z. (2021). "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35 (San Diego, CA: AAAI). p. 13789–13797.

Sirimontree, S., Keawsawasvong, S., and Thongchom, C. (2021). Flexural behavior of concrete beam reinforced with GFRP bars compared to concrete beam reinforced with conventional steel reinforcements. *J. Appl. Sci. Eng.* 24, 883–890. doi: 10.6180/jase.202112_24(6).0009

Wang, J., and Wang, M. (2021). Review of the emotional feature extraction and classification using EEG signals. *Cogn. Robot.* 1, 29–40. doi: 10.1016/j.cogr.2021.04.001

Wang, L., Shoulin, Y., Alyami, H., Laghari, A. A., Rashid, M., Almotiri, J., et al. (2022). A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images. *Geosci. Data J.* doi: 10.1002/gdj3.162. [Epub ahead of print].

Yadav, O. P., Ray, S., and Yadav, Y. (2021). Enhancement and analysis of ECG signals using combined difference total variation optimization. *J. Appl. Sci. Eng.* 24, 845–852. doi: 10.6180/jase.202112_24(6).0004

Yin, S., Li, H., Laghari, A. A., Karim, S., and Jumani, A. K. A. (2021). Bagging strategy-based kernel extreme learning machine for complex network intrusion detection. *EAI Endorsed Trans. Scalable Inf. Syst.* 21, e8. doi: 10.4108/eai.6-10-2021.171247

Yu, H. (2021). Online teaching quality evaluation based on emotion recognition and improved AprioriTid algorithm. *J. Intell. Fuzzy Syst.* 40, 7037–7047. doi: 10.3233/JIFS-189534

Yu, J., Li, H., and Yin, S. (2020). Dynamic gesture recognition based on deep learning in human-to-computer interfaces. *J. Appl. Sci. Eng.* 23, 31–38. doi: 10.6180/jase.202003_23(1).0004

Zehra, W., Javed, A. R., Jalil, Z., Khan, H. U., and Gadekallu, T. R. (2021). Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex Intell. Syst.* 7, 1845–1854. doi: 10.1007/s40747-020-00250-4