Multimedia Tools and Applications (2022) 81:32265–32286
https://doi.org/10.1007/s11042-022-13091-9

A multimodal emotion recognition model integrating speech, video and MoCAP

Ning Jia[1] · Chunjun Zheng[1,2] · Wei Sun[1]

Received: 13 April 2021 / Revised: 3 November 2021 / Accepted: 3 April 2022 /
Published online: 13 April 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

As one of the core technologies in the field of human-computer interaction, emotion recognition focuses on the simulation of human emotion perception and understanding process. Emotion recognition is widely used in medical, education, life, transportation and other fields. At present, the emotion recognition is still a challenging topic. The accuracy of emotion recognition in multimodal is discussed, different emotion features are extracted from speech, video and motion capture (MoCAP) by using deep learning methods, and a matching emotion recognition model called facial motion speech emotion recognition (FM-SER) model is designed. Local and global information of speech, dual spectrograms are designed in audio mode to choose the time-domain and frequency-domain information, and convolutional neural networks (CNN), gated recurrent unit (GRU) and attention models are used to realize speech emotion recognition. A 3D CNN model based on attention mechanism is used in the video mode to capture the potential emotional expression. The sequential features of hand and head movements are extracted from MoCAP, and import into a bidirectional three-layer long short-term memory (LSTM) model with the attention mechanism. Based on the complementary relationship between multimodal, the decision level integrating scheme is designed with higher-precision, stronger generalization ability of emotion recognition. Through a lot of experiments, we compared the results of several popular emotion recognition models on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus. The results showed that the proposed method had higher recognition accuracies in single modality and multimodal, and the average accuracies of one modality and multimodal were improved by 16.3% and 9%. The effectiveness of FM-SER model in emotion recognition was proved.

Keywords Facial motion speech emotion recognition · Dual spectrograms · 3D convolutional neural networks · The attention mechanism · Long short-term memory

✉ Ning Jia
jianing@neusoft.edu.cn

[1] School of Software, Dalian Neusoft University of Information, Dalian, China

[2] Information Science and Technology College, Dalian Maritime University, Dalian, China

1 Introduction

In recent years, the researches of emotion recognition based on biological signals (facial image, speech signal, EEG signal, etc.) has been widely concerned by researchers. It has become a research hotspot in the fields of affective computing, pattern recognition and computer vision [9]. Facial, speech and other biological signals are important media to express emotions. Researchers designed some features and algorithms to analyze these biological signals, which help the computer recognize human emotional state [23]. The application fields of emotion recognition are very broad, including medical treatment, education, life and transportation. The source of emotion data involves all aspects of life.

The concept of emotion recognition comes from emotion computing, which was formally proposed in 1995. Emotional computing is required to give computers the ability to observe, understand and generate emotional features like people, and finally make computers interact naturally and vividly like people. Affective computing has gradually evolved into a key technology of advanced human-computer interaction. As a sub field of emotion computing, emotion recognition has attracted more and more attention in the field of artificial intelligence.

There are many ways to express emotion, and the carrier of emotion is also richer. The process of emotional expression is different from other psychological processes. In the process of activity, there are some specific behaviors, and the generation of emotion is accompanied by some physical changes and special vocal behaviors, including facial expression, muscle tension, hand movements, head rotation, pronunciation rules, etc.

At present, the research of single modality emotion recognition has made great progress [16]. However, the recognition accuracy is still less than that of more modalities. This is because two or more modalities have more emotional information, which can help researchers to deeply mine and fuse multi-modal biological signals. These messages are effective ways to enrich the researches of emotion recognition.

In order to reduce the lack of emotional expression in a single modality, evaluate emotional expression accurately and objectively, and realize effective human-computer interaction, many multimodal data sets have been developed for the research of multimodal emotion recognition. For example, the data set we used: interactive emotional dynamic motion capture (IEMOCAP) [4], which is the most popular open-source multimodal dataset. IEMOCAP contains ten people's video, audio, motion data of head and hands, including angry, happy, excite, sad, frustration, fear, surprise, other and neutral emotions. Deep learning method can learn effective nonlinear representation of speech signal from different modalities (speech, video and motion). At present, deep learning has been widely used in the emotional model of this data set. Among these modalities, acoustic features [11, 20], facial expressions [12, 18], and body movements [1, 5] have achieved good results in emotion recognition.

2 Related work

For speech emotion recognition, researchers have made many valuable achievements. Considering that feature construction and feature selection have a great impact on the performance of emotion recognition, scholars found that the correlation between speech signals in frequency domain and time domain plays an important role in speech emotion recognition [15]. SATT et al. [24] used convolutional neural network (CNN) and recurrent neural network (RNN) to extract features from logarithm spectrogram of speech signal. However, the research on the correlation between speech signals only focuses on the frequency domain or time domain, and the research on the both of them is less. Spectrogram [30, 31] is a visual expression of the time-frequency energy distribution, which can use image features to explore the relationship between adjacent frequency points. Spectrogram also provides a new idea for studying the correlation between time-frequency domain and frequency domain.

Aiming at the problem of expression recognition in video mode, facial expression is the most effective information, and the extraction and classification of expression feature is the core task. Many researchers use deep learning model to construct classifier to achieve end-to-end facial expression feature extraction and recognition from picture [8]. Gupta et al. [6] used RGB and depth image, combined with the teacher student method to migrate RGB image network to process depth image, in order to achieve high-precision emotion recognition. Xu et al. [28] established a double active layer based CNN, which realized high-precision facial expression recognition by learning robust and distinctive features from the data, and enhanced the robustness of the network. Video mode is mainly based on streaming images. How to efficiently transfer face recognition methods from images to video mode is a problem to be solved.

Aiming at the problem of emotion recognition of human motion and posture, the traditional method is based on the theoretical system of human motion and the model of human skeleton motion structure to analyze the relationship between body motion and emotion expression. For example, Luo [14] extracted the emotional features of posture, constructed the body language dataset (BoLD), and tested it on a variety of learning models. Ajili [2] proposed a human action description language to quantify different types of human emotions. This method has strong universality, but it cannot meet the needs of personalized emotional expression.

In order to achieve the effective emotion extraction of motion information, we can use the popular device -motion capture (MoCAP) [26] to assist in this function. MoCAP is a kind of technology which can record the change rule of the wearer's movement in real time in the form of data and distinguish the wearer's movement track through the wearing of sensor equipment. The problem of this emotion recognition scheme is that its accuracy is limited by the secondary development of motion model, which is also a problem to be solved at present.

One modality used in the above scheme has limited recognition ability. Therefore, many researchers used multimodal emotion recognition models to integrate features from various behaviors [10, 19, 22]. Samarth et al. [25] used data from voice, text, facial expression, rotation and hand motion to perform multimodal emotion recognition on IEMOCAP dataset. Ren et al. [21] proposed a new multimodal network for emotion recognition, which combined information from audio and video channels to achieve more robust and accurate detection results. Pan et al. [17] proposed a new multimodal attention mechanism, called cLSTM-MMA, which made the attention of the three modalities (speech, video and text) focus on the important information and integrate the important information selectively.

Bertero et al. [3] extended the long short-term memory network (LSTM) for emotion recognition of multimodal content. Zadeh et al. [29] proposed the graph memory fusion network, which stores the internal information of modes and the interactive information between modes through the gated memory unit, and adds the dynamic fusion graph to reflect the effective emotion. Hazarika et al. [7] used multi-layer gated recurrent unit (GRU) to store the current multimodal content, including speaker information and context information, and took the output of GRU as global information to analyze the emotional information contained in the multimodal content.

We find that there is no unified conclusion for multimodal fusion and selection scheme, and researchers only select valuable information from existing modality. Most of them consider the emotional information in multimodal content, but ignore the relationship between emotions in multimodal content. In fact, the key to improve the accuracy and generalization ability of emotion recognition is to extract features, design single- modality and modalities fusion strategy. Therefore, a multimodal emotion recognition model FM-SER based on deep learning technology is proposed, which combines speech, video and MoCAP. There are three sub models: speech emotion recognition (SER), facial emotion recognition (FER) and motion emotion recognition(MER). For each model, feature extraction scheme and matching model structure are designed respectively. All of the models carry attention mechanism to increase effective information, and realize emotion recognition with high precision and strong generalization ability through user-defined decision fusion scheme. The main structure of this paper is as follows. Section 2 mainly introduces the overall scheme of multimodal emotion recognition model and the design idea of each modal. Section 3 lists the results of training and testing. Section 4 is the summary of this work and the prospect of future work.

3 Proposed method

3.1 System overview

Human expression of emotion is a process in which various modalities are complemented with each other. It is a key problem to extract the emotional information of various modes and realize effective integrating. Therefore, an automatic emotion recognition method for speech, video and MoCAP was proposed. The classifiers for all the modalities were designed, then the classification of a sample is predicted by all the classifiers. Finally, the multimodal level integrating scheme is carried out by using the custom criteria. The overall structure is shown in Fig. 1.

The design ideas of all modalities are as follows. For the speech modality, the dual spectrogram of single speech is estimated, and the speech emotion recognition is realized by combining CNN, gated recurrent unit (GRU) and attention mechanism. This scheme combines the time and frequency domain information, local and global information of speech, which can effectively improve the recognition accuracy.

For the video modality, 3D CNN model based on attention mechanism is constructed to capture potential emotion expression information and complementary information from input video, which can improve the generalization ability of emotion recognition.

For the MoCAP modality, the sequential features of hand and head action are extracted from IEMOCAP dataset, and a bidirectional three-layer long short term memory (LSTM) model with attention mechanism is designed. The model can realize the two-way mapping between the action information and the emotional expression, and realize the powerful supplement for the emotional expression.

3.2 Speech emotion recognition model based on dual spectrogram

The speech emotion recognition(SER) model consists of speech feature extraction and emotion recognition. In the process of feature extraction, a dual spectrogram is designed to extract speech emotion features. The traditional recognition scheme is to transform speech into spectrogram, which can express the time-frequency distribution of speech energy in the form of visualization. When existing noise in the speech, its spectrogram is always disordered. In

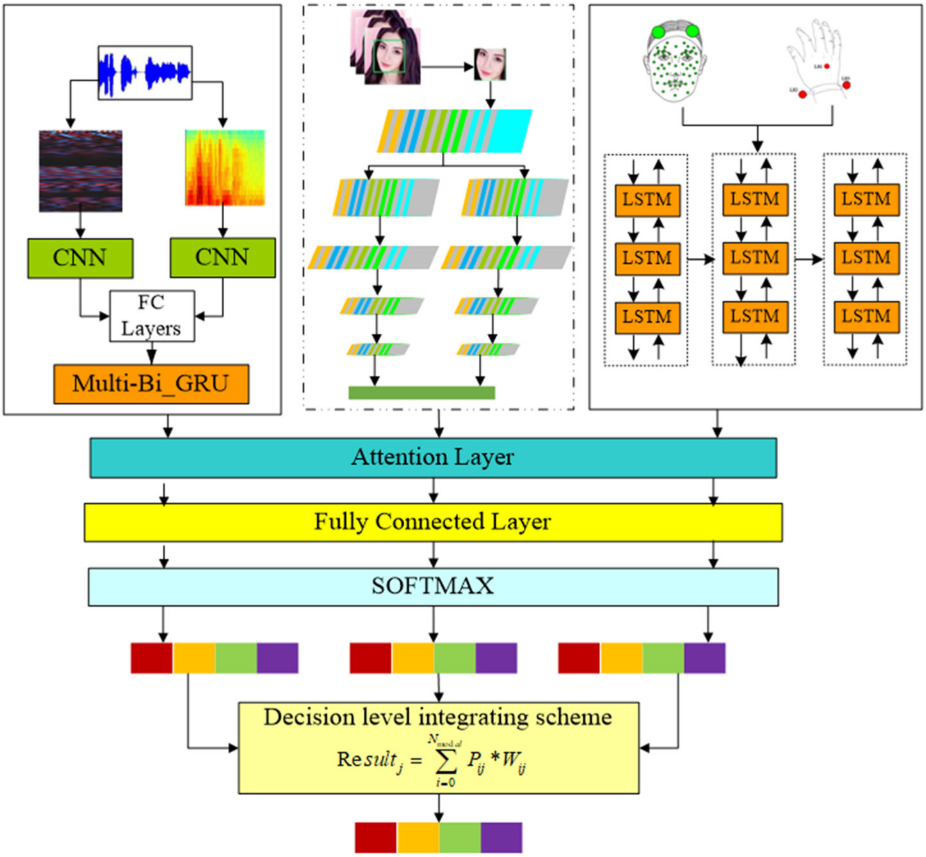
CNN
CNN
FC Layers
Multi-Bi_GRU
LSTM
Attention Layer
Fully Connected Layer
SOFTMAX
Decision level integrating scheme
$Result_j = \sum_{i=0}^{N_{modal}} P_{ij} * W_{ij}$

Fig. 1 Overall design scheme of the model

order to avoid the problem above, a dual spectrogram is designed for SER. The first layer uses the traditional spectrogram, which can reflect the correlation between time domain and frequency domain. In the second layer, the spectrogram is used to extract the salience region in the form of digital signal with the form of image.

As the most important deep learning model, GRU is widely used in natural language processing and time series related tasks. It is usually used to solve the gradient disappearance and gradient explosion problems in RNN. There is a typical temporal relationship in speech, and there is an image form in the dual spectrogram. It can be considered that the features above are existing in the dual spectrogram at the same time. In the emotion recognition stage, CNN and GRU models based on attention mechanism are designed to realize high accuracy SER.

3.2.1 Extraction of double spectrogram

The extraction of dual spectrogram is used to extract effective features in emotion. The emotional expression of speech may last for the whole audio, or may be concentrated in a certain period of time. Because of this uncertainty, we use spectrogram to extract the global expression of the whole speech, and use custom spectrogram to highlight the expression of emotion burst region, which is essentially to reflect the local effectiveness of the speech.

The extraction process of custom spectrogram is divided into two stages: speech image reconstruction and salience region extraction. The extraction process of custom spectrogram is described in detail in the following.

a. Speech image reconstruction

Speech image reconstruction is used to convert speech into standardized image, as shown in Eq. (1)

$$t_{i,j} = p_{i*\sqrt{N}+j} \quad i,j < \sqrt{N} \tag{1}$$

where N in Eq. (1) is the original numbers of the current speech, *(i, j)* is the coordinates of each pixel, $t_{i,j}$ is the matrix expression of the original signal. After normalizing $t_{i,j}$, a pixel matrix $t^{'}_{i,j}$ can be generated, which is the image expression of speech.

b. Salience Region Extraction

The image expression of speech can intuitively show whether there is signal mutation in the current region or keep smooth. When the local mutation of the image occurs, it is often the beginning or ending stage of the emotional fluctuation of the speech. Based on this, the time of mutation can be used to judge the change interval of speech energy. It can be considered that the first mutation is the beginning of the fluctuation, and the second mutation is the end of the fluctuation, and so on. The following methods are designed to find the salience region in the image.

(1) Prepare a sliding window with width $\sqrt{N}$ and height $\lambda\sqrt[4]{N}$. Where, λ determines the fineness of the salience region. With the increase of λ, the segmentation becomes rougher. λ is a hyperparameter.

(2) Starting from the initial position of the image, the sliding window scans all the image regions in turn and differentiates them from the original image, as shown in the following:

$$^{\triangle}t_{i,j} = t^{'}_{i+\lambda\sqrt[4]{N},j} - t^{'}_{i,j} \quad i,j < \sqrt{N} \tag{2}$$

$$^{\triangle\triangle}t_{i,j} = {}^{\triangle}t_{i+1,j} - {}^{\triangle}t_{i,j} \tag{3}$$

$$\begin{gathered} t_{start} = i_1 \\ if\, {}^{\triangle\triangle}t_{i1,j} > \frac{1}{N}\sum_{i,j=0}^{\sqrt{N}} {}^{\triangle\triangle}t_{i,j},\ {}^{\triangle\triangle}t^{2}_{i1,j} > \frac{1}{N^2}\sum_{i,j=0}^{\sqrt{N}} {}^{\triangle\triangle}t^{2}_{i,j} \end{gathered} \tag{4}$$

$$\begin{gathered} t_{end} = i_2 \\ if\, {}^{\triangle\triangle}t_{i2,j} > \frac{1}{N}\sum_{i,j=0}^{\sqrt{N}} {}^{\triangle\triangle}t_{i,j},\ {}^{\triangle\triangle}t^{2}_{i2j} > \frac{1}{N^2}\sum_{i,j=0}^{\sqrt{N}} {}^{\triangle\triangle}t^{2}_{i,j} \end{gathered} \tag{5}$$

where, $t'_{i,j}$ is the pixel matrix obtained in Eq. (1) $\triangle t_{i,j}$ is the matrix after difference, and the significance region is the effective area in the matrix. Quadratic difference $\triangle \triangle t_{i,j}$ is used to find the start and end positions of the salience region. In Eqs. (4) and (5), constraint conditions are added to distinguish flat and fluctuating regions. At this time, t_{start} and t_{end} are the specific positions of the final significant region.

(3) The image information in the interval $[t_{start}, t_{end}]$ is the custom spectrogram.

The second layer is spectrogram, which combines the characteristics of frequency and time domain, and dynamically shows the change of speech spectrogram with time. The color represents the signal energy of the given frequency component at the corresponding time.

When the forms of speech are changed, such as tone, intonation and speed, the spectrogram changes accordingly. With the color area changes in the spectrogram, the horizontal and vertical stripes fluctuate greatly. The extraction process of dual spectrogram is shown in the Fig. 2 below.

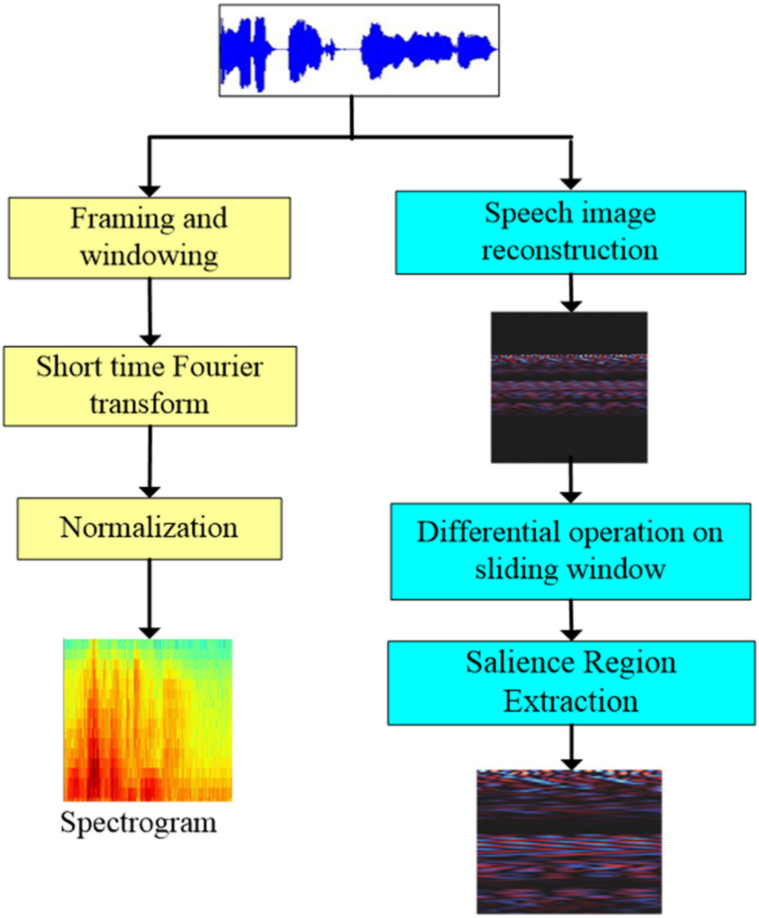
Framing and windowing
Short time Fourier transform
Normalization
Spectrogram
Speech image reconstruction
Differential operation on sliding window
Salience Region Extraction

Fig. 2 Extraction process of dual spectrogram

3.2.2 Designing of speech emotion recognition model

As the features use the dual spectrogram with image form, CNN and GRU model based on attention mechanism is selected here. CNN model is used for abstract feature extraction of dual spectrum, and GRU model is used for processing time series composed of abstract feature arrangement. The model structure is shown in Fig. 3.

As shown in Fig. 3, each layer of the dual spectrogram was put into the CNN model. Since each spectrogram has three channels, there are six channels in total.

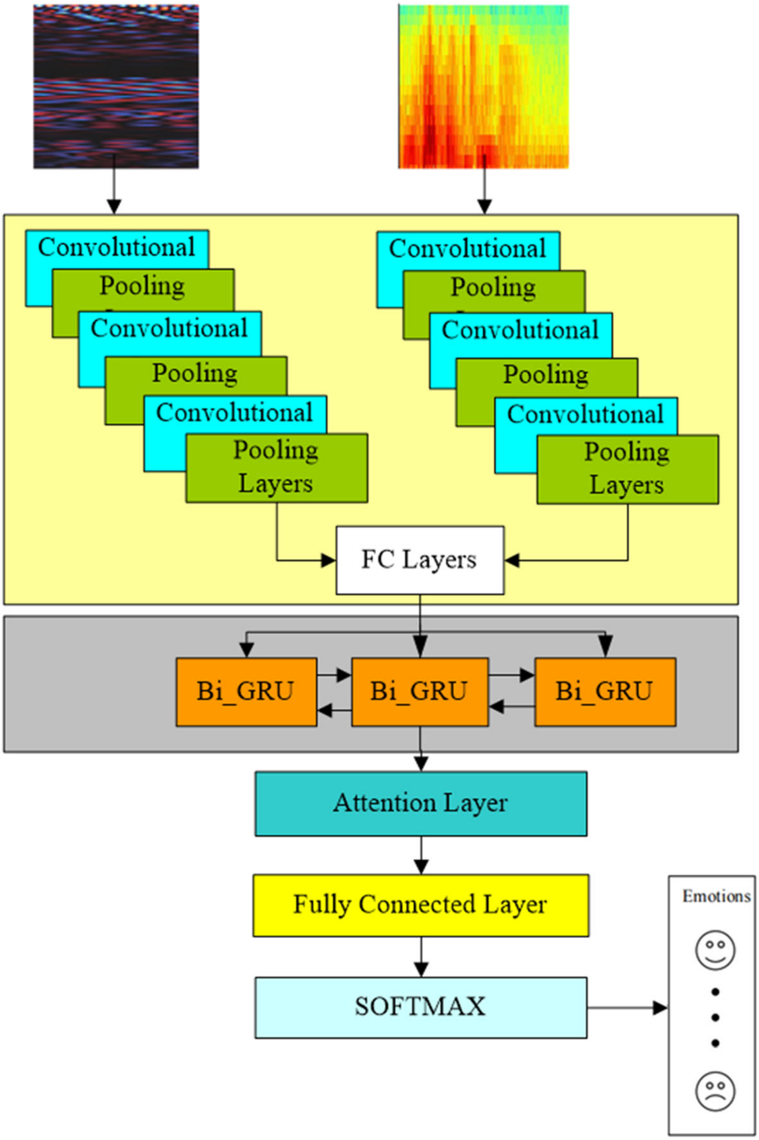

Fig. 3 Speech emotion recognition model

As a supervised model, CNN is composed of one input layer, three convolutional layers, three pooling layers and one fully connecting layer. The weights of all the feature maps are equal, which can greatly reduce the size of parameters and speed up the training. The maximum pooling method is used to make it robust to small changes of input.

The spectrogram of each layer is reshaped to a scale of 128 * 128. Therefore, the first layer is equivalent to the scale of the spectrogram. After 64 convolutional kernels with size of 7 * 7 and convolutional operation with step size of 1, the input data is conversed to 64 feature maps, and then uses the Rectified Linear Unit (ReLU) activation function after maximum pooling. The second layer of convolution layer selects all the feature maps of the upper layer as the input. The number of convolutional kernels in this layer is changed to 96, the size of convolution kernels remains unchanged, and the calculation process is similar to that in the first layer. In the third layer, the number of convolutional kernels is changed to 128, and the size of convolutional kernels is adjusted to 9*9. By alternately performing the computation on the convolution layer and pooling layer, we build a deep network model to obtain more representative features.

Next is the fully connected layer, which maps the learned emotional feature representation to the speech signal. On this basis, the dropout operation is used to prevent over fitting. Different from the traditional CNN model, the output of the fully connected layer is not the classification result, but 1024*2 dimension features, which is another expression of the effective features based on the global and local fields. Based on this, the dual spectrogram is input into the CNN model respectively, and the output features of each spectrogram in the fully connected layer are combined and jointly injected into the GRU model.

The GRU model designed here contains three bidirectional recurrent layers. The expression of emotion can be determined by the information of several frames before and after. Therefore, two opposite recurrent layers are designed for information transmission. The first layer transmits information in positive time order, and the second layer transmits information in reverse time order. Finally, through the fully connected layer and softmax layer, the predicted emotion classification result is output.

The SER model is the core of multimodal emotion recognition model. Among them, the spectrogram representing global information and the custom spectrogram representing local features are used as input, CNN and GRU models with the attention mechanism are designed. This model skillfully combines the temporal features of speech and the characteristics of spectrogram, and increases the effective information of emotion expression.

3.3 3D CNN model for facial expression recognition

Facial expression recognition(FER) model uses video sequence information to achieve effective emotion recognition. This is a classification task based on face feature deformation or face motion. Facial expression can be produced by the movement of facial muscles, and cause deformation of eyes, nose, mouth and other parts. If this change is extracted and classified, facial expression recognition can be realized. Facial expression recognition consists of three parts: face image acquisition, expression feature extraction and expression classification.

3.3.1 Face image acquisition

This process obtains the face image from the input image data. Video is an image sequence with temporal characteristics, which contains more information than static images. In order to improve the performance of face detection in video stream, the face detection algorithm of adaptive boosting (AdaBoost) [13] based on Haar feature is selected. The steps of the method are as follows:

(1) The video image is extracted in a fixed time period. The minimum unit of sample segmentation in IEMOCAP is 100 ms, so 100 ms is selected as the basic unit of sample segmentation.
(2) The coarse-grained region with head is extracted from the video image.
(3) Haar-like feature is used to represent the face, and integral graph is used to calculate the features quickly.
(4) The AdaBoost algorithm is used to select rectangular features (weak classifiers) which can best represent the face, and the weak classifiers are constructed into a strong classifier by weighted voting.
(5) The trained classifiers are connected in series to form a cascade classifier which can effectively improve the detection speed of classifier.
(6) According to the position of the face in the image, the video is re cut, and the face part of the image is retained, so that the video segment can be mapped to the speech segment. The final face image size is 70*70.

3.3.2 Expression feature extraction and classification

The facial expression feature extraction and classification model adopts 3D CNN architecture to capture facial expression discrimination features along the spatial and temporal dimensions, and realize facial expression classification. As is shown in Fig. 4, The model can generate multiple information channels from adjacent video frames, and perform convolution and down sampling in each channel. The feature map in convolutional layer is connected with several adjacent frames in the previous layer, and the size of each layer has been marked in the Fig. 4. The specific process is as follows.

The model includes one hardwired layer, three convolutional layers, two down sampling layers, one attention layer and one fully connected layer. In the first layer, a fixed hardwired core is used to process the original frame and generate five channels of information. These channels are: gray, horizontal and vertical gradient, horizontal and vertical optical flow. The first three channels are calculated in frames. The horizontal and vertical optical flow fields are calculated every one frame.

In the first layer, the 3D convolutional kernel with a size of 7*7*3 is used to convolute the five channels. In order to extract different types of features, two convolutional kernels with the same size and different kernel features are used in each position. At this time, two feature maps groups with the same size are formed in the second layer. The third layer is down sampling layer, namely Max pooling, which can down sampling the feature maps of the second layer to get the feature maps with reduced spatial resolution.

The fourth layer is convolution layer, which uses the 9*9*3 3D convolutional kernel in five channels. The fifth layer is used for down sampling, and the sixth layer is the convolutional layer, which convolutes only in the spatial dimension. The seventh layer uses the attention mechanism, which amplifies the effective signals and suppresses the invalid features through the feature maps with different weights. Finally, through the fully connected layer and Softmax layer, the predicted emotion classification results are output.

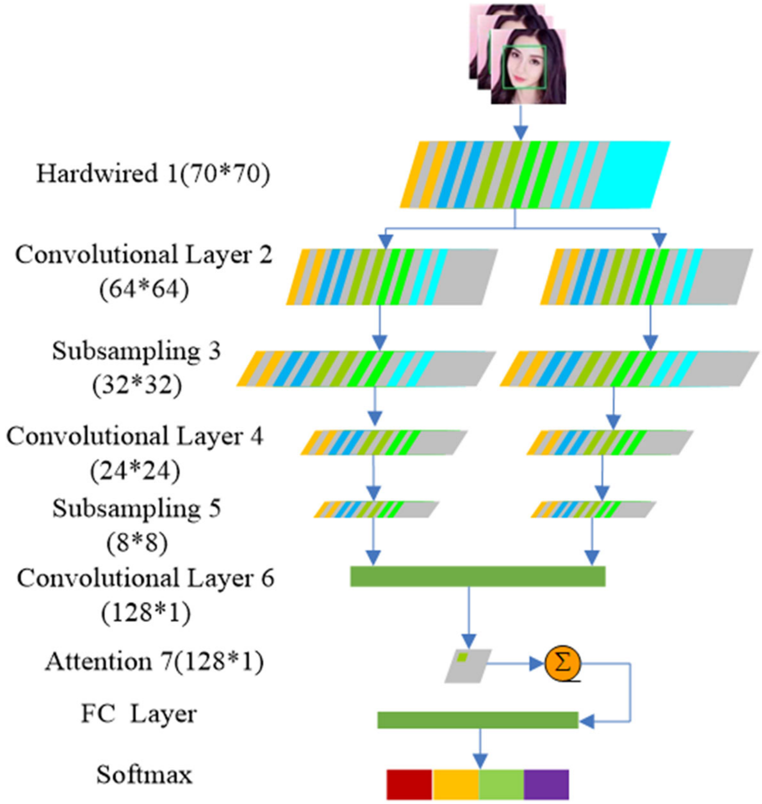

Fig. 4 The Model of Expression Feature Extraction and Classification

The FER model is the core of the multimodal in emotion recognition task. It uses the 3D CNN model to highlight the image sequence features in the video, and combines the attention mechanism to capture the high-level facial motion information in the video, so as to achieve FER.

3.4 Motion emotion recognition based on MoCAP

In IEMOCAP corpus, each actor wears a motion capture camera to record the facial expression, head and hand motion information from multiple dimensions. Motion capture data contains multiple tuples. For example, facial expression data contains 165 dimensions, hand position information contains 18 dimensions, and head rotation data contains 6 dimensions. Head and hand data is used to form the basis of motion emotion recognition(MER) model based on the existing data collection in the database. Using the above data, we construct a MER model after processing multi-dimensional hand and head motion data. The specific process is as follows:

3.4.1 Preprocessing and feature extraction of hand data

The hand data contains a total of 20 dimensional data of six cameras. In addition to the Number of frame and time starting point, the remaining 18 dimensional data are three items of each camera, namely the rotation data of X, Y and Z axes. In hand data, we need to purify the motion sequence of a single sample, which mainly involves the following operations:

a. There are a lot of null values in the data of left (LH1) and right (RH1) cameras. It is necessary to clear the six dimensional spatial information including LH1 and RH1. Only the remaining four camera data are retained.
b. There are still a few null values in the remaining four camera data, and only the row information with null values in cameras of LH2, LH3, RH2, RH3 is cleared.
c. Taking the action data of single speech as the unit, the subspace information with the largest absolute difference of hand data are extracted.

The specific process is as follows. First, a sliding window is set, the window size is D, and the window is moved from left to right. According to Eq. (6), the maximum value of the absolute difference of the data in the sliding window and its starting point are calculated.

$$\text{argmax}\left(\sum_{i=start}^{start+d}(M_{i+1}-M_i)\right), start\in[0, len(M)-d] \tag{6}$$

where, M_i is the *ith* data of MoCAP, *start* is the starting position of the sliding window, and *d* is the super parameter, which can be adjusted by developers according to the actual situation. In the current sliding window area, the sum of the absolute differences and the mean value of the absolute variance in the sliding window are calculated, the 12 dimensional data is used to replace the hand data of the whole sample.

3.4.2 Preprocessing and feature extraction of head data

The head data contains 8-dimensional data of two cameras. In addition to the number of frame and time starting point, the remaining 6-dimensional data are the information of *pitch, roll, yaw, tra_x, tra_y and tra_z*. In the head data, we need to reduce the trajectory of single sample, which mainly involves the following two types of operations:

a. Clear the two dimensional information including the columns of frame and time starting point.
b. The subspace information with the largest absolute difference of current sample is extracted.

Similarly, by setting the sliding window and using Eq. (6), the maximum value of the absolute difference of the data and its starting point position inside the sliding window are obtained. In the current sliding window area, the sum of the absolute differences and the mean value of the absolute variance in the sliding window are calculated, the 6 dimensional data is used to replace the hand data of the whole sample.

Due to the limited dimensions of hand and head data, we choose to connect the data of hand and head to form a new feature combination. However, it is found that part of the head data has a negative impact on emotion recognition. Therefore, based on all the hand features, some effective head data are fused as the final set of motion features.

3.4.3 Motion emotion recognition model

On the basis of hand and head feature extraction, a motion emotion recognition model (MER) based on MoCAP is designed. Because the action sequence is rich in the spatiotemporal relationship of emotional expression, the three-layer bidirectional LSTM model is adopted, and the combined features are used as the input. Each layer of LSTM has 256 units. Considering that emotional expression is a process of concentrated outburst in a short time, feature selection and training can be carried out by focusing on the key areas of outburst to increase the contribution of such areas. Therefore, after the last layer of LSTM model, the attention layer is added, which helps to improve the emotional attention ability and computing ability of the model. Next, the output of the attention layer is injected into the fully connected layer and Softmax to output the predicted emotion classification results. The MER model is shown in Fig. 5.

MER model is a powerful supplement to FM-SER model. The model only considers the temporal characteristics of action information, adopts three-layer bidirectional LSTM model to highlight the effectiveness of long-term information, selectively saves long short-term memory, and then combines with attention mechanism to enhance the effective information.

Based on the realization of SER, FER, MER model, the final emotion recognition result can be obtained by using integrating scheme. The overall model structure can refer to Fig. 1.

3.5 Design of Decision Integrating Scheme

The prediction results of the above three modalities are integrating together. The decision level method is adopted here. The specific integration ideas are as follows.

Suppose N_{class} is the number of emotion class and $N_{\mathrm{moda}l}$ is the number of modalities. P_{ij} is the prediction probability of the j-th emotion of the i-th modality, and its value range is [0,1],

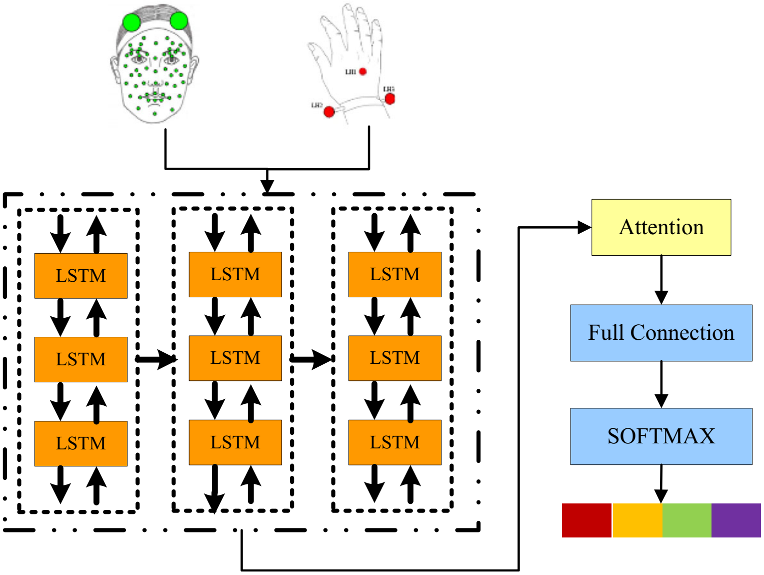

Fig. 5 Motion Emotion Recognition Model

and $\sum_{j=0}^{N_{class}} P_{ij} = 1$. W_{ij} is the confidence of the i-th modality and the j-th emotion, and its value range is [0,1]. The process is as follows.

1. Given the initial value of W_{ij}, it is the training emotion recognition accuracy of the corresponding modality.
2. Calculate the $\mathrm{Re}sult_j$, which is the prediction probability of each class of emotion, the calculation method is as follows:

$$\mathrm{Re}sult_j = \sum_{i=0}^{N_{\mathrm{mod}al}} P_{ij} \cdot W_{ij} \tag{7}$$

3. Calculate the emotion recognition result of the current speech, namely $\arg\ \max(\mathrm{Re}sult_j)$.
4. Calculate the unweighted accuracy (UA) index of one modality [27] as the evaluation index of emotion recognition accuracy. The calculation method is as follows:

$$UA_{ij} = \frac{TP_{ij}}{TP_{ij} + FP_{ij}} \tag{8}$$

where, UA_{ij} represents the recognition accuracy of the i-th modality and the j-th emotion, TP_{ij} represents the number of real cases of each type in the current modality, and FP_{ij} represents the number of false positive cases of each class in the current mode.

5. Update W_{ij}. the calculation method is as follows:

$$W_{ij}^{'} = \lambda_1 W_{ij} + \lambda_2 UA_{ij} \tag{9}$$

where, λ_1 and λ_2 as adjustable weight coefficients, are also hyperparameters, which are dynamically adjusted by researchers according to multiple groups of experiments and risk experience to obtain the best coefficient.

4 Experiments and results

4.1 Datasets

All the experiments in this paper are carried out on the 8-GPU NVIDIA 1080Ti server. We use the interactive emotional dynamic motion capture (IEMOCAP) corpus to carry out the experiments. As the most popular open source corpus, it consists of discrete tags of ten emotions (angry, happy, excitement, sad, frustration, fear, surprise, other and neural). There are 10,039 speeches in this corpus, with a total duration of nearly 12 h. Each sample also contains video, motion capture and text information. We adopt a general processing scheme for IEMOCAP corpus. Four classes of emotional samples are used here: happy (merged with the exited), sad, angry and neutral. The scales of the classes are: angry(1103), happy(1636), neutral(1708), sad(1084). The scales of angry and sad are small, while that of happy and neutral are large. The five-fold cross validation method is used in the experiment. 80% of the data is used to train the deep neural network, and the remaining data is used for verification and accuracy testing.

4.2 Network parameters and evaluation criteria

In the training process, the end-to-end method is used to optimize the parameters. Some hyperparameters in the model can refer to Table 1. For video segmentation, the duration of each video cutting is 100 ms. For the MoCAP data, the size of the window for output processing is set to 128. If MoCAP data of the sample is less than the sliding window, all the valid action data of the current sample is used.

Tensorflow framework is used to build the network model structure. In LSTM and GRU models, the Epochs is 10,000. The dropout is set to 0.5. ReLU is used as the activation function and Adam as the optimizer. In 3D CNN, the cube with each 3D convolution is seven consecutive frames. Batch size, maximum number of cycles and the learning rate are the same as LSTM model.

Weighted accuracy (WA) and unweighted accuracy (UA) are used as the evaluation indexes of recognition accuracy. WA is used to monitor the overall performance of the model. It is the quotient of the correct prediction and the total number of samples. The calculation of WA completely depends on the calculation of positive examples, and the negative effect of data skew has not been considered in WA. In order to solve the problem of unbalanced sample distribution, UA is introduced to comprehensively determine the emotion recognition accuracy of each class. UA is the mean value of all the classification accuracy. Therefore, UA is a more relevant mark for imbalanced data sets.

4.3 Experimental setup

For three different modalities, based on IEMOCAP data set, experiments are designed to realize the training and testing of single model. For the multimodal model, experiments are designed to train and test, and the results are compared with other classical multimodal experiments. Finally, for the hyperparameters in the model, experiments are designed to verify the effectiveness of the parameters.

4.3.1 The effectiveness experiment for SER model

This experiment was used to verify the effectiveness of the SER model proposed in part II. We designed the following models to compare the accuracy. In this experiment, the experimental

Table 1 Hyper-parameters of proposed system

Hyper-parameters	Range	Ours
λ1 in System overview	0~1	0.5
λ2 in System overview	0~1	0.5
λ in SER	0.25~0.75	0.5
d in MER	32~256	128
Learning rate	1e-3~1e-6	1e-5
Batch size	32~256	128
Workers number	4	4

Table 2 Test results of speech emotion recognition

Model	WA	UA
Baseline: Reference [30]	68.8%	59.4%
Model 1: CNN+GRU(Using traditional spectrogram)	62.2%	62.9%
Model 2: CNN+GRU (Using Custom spectrogram)	61.2%	62.1%
Model 3:CNN(Using dual- spectrogram)	67.5%	68.3%
Model 4: 3-layer bidirectional GRU(Using dual- spectrogram)	66.7%	66.2%
Model 5: Ours(Using dual- spectrogram))	69.2%	69.6%

results in [30] were used as the baseline, and Table 2 showed the accuracy of different SER models after experimental verification.

It could be seen from Table 2 that the proposed method was compared with the method in the references. In the reference, only spectrogram was used, while in this paper, double spectrogram was used for SER, so the recognition accuracy of ours was higher with the same model. Compared with the reference, the accuracy of UA was improved by 10.2%. In addition, we also compared the recognition accuracy of using custom spectrogram alone in model 2. Because the custom spectrogram emphasized local features, the recognition accuracy of model 2 was not high. Due to the dual properties of image and time sequence, only image features were emphasized in model 3, and only time sequence features were emphasized in model 4. The results of the two models were not the best. The validity of CNN and GRU model in dual spectrogram was directly proved by model 5.

4.3.2 The effectiveness experiment for FER model

This experiment was used to verify the effectiveness of the FER model proposed in part II. We designed the following models respectively to compare the accuracy. In this experiment, the method of FER in [18] were used as the baseline, and Table 3 showed the accuracy of different FER models after experimental verification. As only UA was provided in the reference, WA was not provided, so we used * to indicate WA.

It can be seen from Table 3 that we compared the proposed method with the method in the reference. The traditional 3D CNN was used in the reference. In this paper, the extraction of human image was combined with 3D CNN and the attention mechanism. It could be found that the recognition accuracy of ours was the highest. Compared with the reference, the accuracy was improved by 14.8%. In addition, models 1–3 were a part of model 4. By comparing the experimental results, it could be found that the recognition accuracy of models 1–3 were not the best. Only the fusion of the models 1–3 could achieve the best accuracy.

Table 3 Test results of facial emotion recognition

Model	WA	UA
Baseline: the method in [18]	*	53.2%
Model 1: CNN(not extracting face)	54.2%	55.1%
Model 2: CNN(extracting face)	56.8%	57.3%
Model 3: using 3D CNN without adding attention mechanism	65.2%	65.7%
Model 4: Ours (using 3D CNN with attention mechanism)	67.9%	68.4%

Table 4 Test results of motion emotion recognition

Model	WA	UA
Baseline: reference [25] MoCap-head Model2	40.28%	*
Model 1: reference [25] MoCap-hand Model2	36.94%	*
Model 2: reference [25] MoCap-combined Model1	51.11%	*
Model 3: Ours(only using MoCap-head)	38.4%	37.6%
Model 4: Ours(only using MoCap-hand)	60.9%	61.2%
Model 5: Ours(both MoCap head and hand)	56.8%	56.4%
Model 6:single layer LSTM(effective MoCap head and hand)	61.8%	62.1%
Model 7: 3-layer bidirectional LSTM(effective MoCap head and hand)	62.7%	63.3%
Model 8: Ours(effective MoCap head and hand)	64.2%	64.5%

4.3.3 The effectiveness experiment for MER model

This experiment was used to verify the effectiveness of the MER model proposed in part II. We designed the following models to compare the accuracy. In this experiment, the head experimental results in [25] were used as the baseline, and Table 4 showed the accuracy of different MER models after experimental verification.

In this experiment, comparing the effect of the reference and current model, we could see that baseline and model 1 were the head and hand recognition results in the reference respectively, which were consistent with the data sources of model 3 and model 4 proposed by us. It could be seen that the accuracy of model 3 was not high, and that of model 4 was higher. When model 5 was used to integrate all features for recognition, the accuracy was decreased, which proved that some features in the head had errors. Therefore, through many experiments, we selected some effective hand and head data, that is, all the hand data and the third, fourth and fifth dimensional head data. We input them as features into model 6–8. The results showed that when the input features were the same and effective, the effects of single-layer LSTM and three-layer LSTM were not as good as the current bidirectional three-layer LSTM model. Model 2 was the recognition result of hand, head and rotated data in the reference. Through comparison, it could be found that the current model had the best recognition accuracy. Compared with the reference, our model had 24% accuracy improvement.

4.3.4 The effectiveness experiment for FM-SER model

This experiment was used to verify the effectiveness of the FM-SER model proposed in part II. We designed the following models to compare the accuracy. In this experiment, the experimental results of SER in Experiment 1 were used as the baseline, and Table 5 showed the accuracy of different emotion recognition models after experimental verification.

Table 5 Test results of FM-SER emotion recognition

Model	WA	UA
Baseline: SER	69.2%	69.6%
Model 1: SER+MER	73.4%	74.6%
Model 2: MER+FER	72.8%	73.4%
Model 3: SER+FER	75.4%	76.4%
Model 4: Ours(SER+FER+MER)	79.2%	80.1%

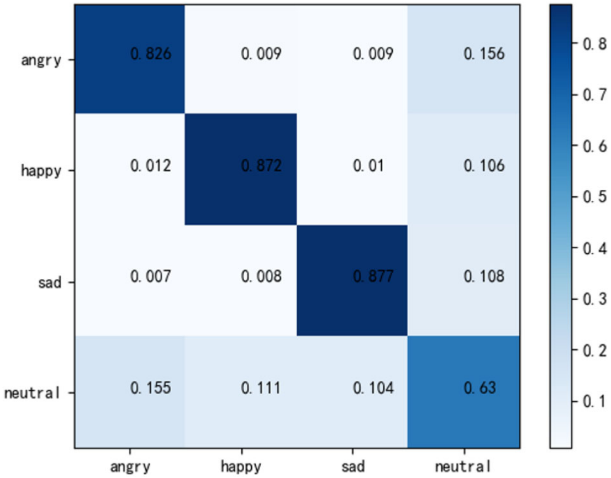

Fig. 6 Confusion Matrix of Emotional classes

Comparing the integrate effect of different modalities in this experiment, we can see that models 1–3 were the integrates of two modalities respectively, and their recognition accuracies were significantly improved than that of one modality. Model 4 had the best recognition accuracy by integrating the data of three modalities. It could be seen that these modalities complement each other and their coupling was not high.

Figure 6 was the confusion matrix of emotion recognition using the current model for multimodal emotion recognition. It could be seen that for angry, happy and sad classes, the recognition accuracies were higher, while the recognition accuracy of neutral category was lower. We found that the accuracies of SER and FER models in the neutral class were not high, which directly lead to the poor results of multi modalities. The main reason was that the features of the dual spectrograms and facial expressions of neutral were not obvious enough. In speech and video, we classified the video and speech without significant features into neutral.

In addition, in IEMOCAP, the amounts of samples in angry and sad classes are less. By observing the above experimental results, the accuracies of these two classes were not affected, which verified the effectiveness of this model for emotion recognition in the corpus which had skew data (Table 6).

Based on the IEMOCAP corpus, we compared the accuracy of emotion recognition models in several multimodal domains related papers, as shown in Table 7. We found that the multimodal recognition scheme proposed in this paper had the best recognition accuracy, which was higher than the model accuracy in the following papers, and the average accuracy was improved to 9%. This proved the effectiveness of the proposed one modality and multimodal integrating scheme.

Table 6 Confusion Matrix of Emotional classes

Accuracy	Angry	Happy	Sad	Neutral
Angry	82.6%	1.2%	0.7%	15.5%
Happy	0.9%	87.2%	0.8%	11.1%
Sad	0.9%	1.0%	87.7%	10.4%
Neutral	15.6%	10.6%	10.8%	63%

Table 7 Accuracy comparison with other popular multimodal emotion recognition models

Model	Reference	Accuracy
Text+speech+MoCap	Samarth [25]	71.04%
3D-CNN + text-CNN + openSMILE	Soujanya [19]	71.59%
RF + XGB + MLP + MNB + LR	Gaurav [22]	70.1%
2D-CNN + 3D-CNN + LSTM	Ren [21]	60.59%
FM-SER	Ours	80.1%

4.3.5 Experiment of parameter sensitivity

In order to test the validity of the hyperparameters in the model, the effects before and after adjusting the hyperparameters were compared: λ_1 and λ_2 in the system overview, λ in the custom spectrogram, and d in the sliding window of the MoCAP model. The specific design scheme was shown in Table 8. It could be seen that the adjustment of λ_1 and λ_2 had little effect on the performance of FM-SER, because the variables they decorated were the accuracies of training set and test set respectively, which could prove that the distributions of the two sets were almost the same. The parameters of λ and d in the MoCAP model need to be adjusted to a suitable value repeatedly, otherwise the recognition accuracy of the model would be affected.

4.4 Result analysis

Through the above experiments, we found that the FM-SER model we proposed had the best recognition accuracy, which was mainly reflected in the following aspects:

1. The recognition accuracy of single modality was higher.

The dual spectrogram generated from speech extracted global and local features at the same time, and the recognition effect was the best after integrating the two groups of features. Based on the attention mechanism in the FER model, the deformation of facial action was aggravated, and the recognition accuracy was improved. The head data in the MoCAP was less effective, so the integrating of all hand data and some head data could be used as a favorable model supplement. The experimental results showed that the accuracies of SER, FER, MER model were improved by 10.2%, 14.8% and 24%, and the average accuracy was improved by 16.3%. It showed that the emotion recognition accuracy of a single modality has been improved.

Table 8 Test results of Parameter Sensitive Analysis

Parameter	Setting	FM-SER UA
λ_1, λ_2	Ours, $\lambda_1=\lambda_2=0.5$	80.1%
	$\lambda_1=0.2$, $\lambda_2=0.8$	79.6%
	$\lambda_1=0.7$, $\lambda_2=0.3$	79.4%
λ	Ours, $\lambda=0.5$	80.1%
	$\lambda=0.3$	79.4%
	$\lambda=0.8$	74.5%
d	Ours, $d=128$	80.1%
	$d=32$	75.2%
	$d=256$	76.3%

2. Multimodal decision level integrating was more effective. The main modalities were speech and video, supplemented by MoCAP modality. The confidences of each modality were adjusted dynamically during fusion.

The experimental results showed that the recognition accuracy of angry, happy and sad was higher than that of neutral. The average accuracy was improved to 9%. Because the original dataset had data skew in angry and sad classes, and the recognition results were not affected by the skew data, it showed that the current model could avoid the skew problem. In popular corpora, there was often uneven data distribution. The model proposed in this paper would not have a harmful impact on these corpuses.

3. In addition to the modality information used in this paper, the modalities of emotion recognition were also included text, rotated features and others, which were not used in the current model. Through the comparison in the Table 7, we found that the current model could obtain higher recognition accuracy. It was not that the more the number of modalities, the higher the recognition accuracy obtained. From another point of view, it was proved that the feature extraction in one modality was very important, and the recognition model suitable for the current feature should be provided. Therefore, the relationship between emotion and modality is far more important than which features to choose.

5 Conclusions

In this paper, a multimodal emotion recognition model FM-SER is proposed, which combines audio, video and action. It contains three modalities models, namely SER, FER and MER. For each model, feature extraction scheme and matching model structure are designed respectively. In order to pickup the time and frequency information, local and global information in the speech, dual spectrogram are designed in SER model. The task of video modality is to extract and classify facial expression features based on face image extraction. It can combine attention mechanism to capture potential emotional expression. The effective sequence features of hand and head movements are extracted from the motion modality, combined with the bidirectional three-layer LSTM model with the attention mechanism. The integrating of the three modalities is reflected in the decision-making level, and the confidences of all modalities are dynamically adjusted to achieve high-precision emotion recognition with strong generalization ability. Experimental results on the IEMOCAP emotional corpus show that the proposed method can improve the average accuracy of one modality and multimodal by 16.3% and 9%.

However, we find that the recognition effect of the current model for the neutral class is not well, which is related to the fact that the feature highlighting the neutral class is not used in the current emotion recognition model. In the future research process, we will continue improve the SER model, design the features and models of text modality, and seek a more general multimodal network structure. Researchers can select a part of the modalities according to the needs to achieve efficient speech emotion recognition and simultaneously improve the accuracy of other speech related tasks.

Acknowledgements This paper was funded by the Liaoning Province Key Laboratory for the Application Research of Big Data, the Dalian Science and Technology Star Project, grant number 2019RQ120, and the Intercollegiate cooperation projects of Liaoning Provincial Department of Education, grant number 86896244.

Declarations

Conflict of interest The authors declare that they have no competing interest.

References

1. Ahmed F, Bari ASMH, Gavrilova ML (2020) Emotion recognition from body movement[J]. IEEE Access 8:11761–11781
2. Ajili I, Mallem M, Didier JY (2019) Human motions and emotions recognition inspired by LMA qualities[J]. Vis Comput 35(10):1411–1426
3. Bertero D, Siddique FB, Wu CS et al (2016) Real-time speech emotion and sentiment recognition for interactive dialogue systems. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, pp 1042–1047
4. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: interactive emotional dyadic motion capture database[J]. Lang Resour Eval 42(4):335–359
5. Ding IJ, Hsieh MC (2020) A hand gesture action-based emotion recognition system by 3D image sensor information derived from leap motion sensors for the specific group with restlessness emotion problems[J]. Microsyst Technol 3
6. Gupta S et al (2016) Cross modal distillation for supervision transfer. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2827–2836
7. Hazarika D, Poria S, Mihalcea R et al (2018) ICON: interactive conversational memory network for muitimodal emotion detection. In: Proceedings of the 2018 Conference on empirical methods in natural language processing, Brussels, pp 2594–2604
8. Huang L, Xie F, Shen S et al (2020) Human emotion recognition based on face and facial expression detection using deep belief network under complicated backgrounds[J]. Int J Pattern Recognit Artif Intell 1
9. Jiahui PAN, Zhipeng HE, Zina LI et al (2020) A review of multimodal emotion recognition[J]. CAAI Trans Intell Syst 15(4):1–13
10. Kan W, Longlong M (2020) Research on design innovation method based on multimodal perception and recognition technology[J]. J Phys Conf Ser 1607(1):012107 (6pp)
11. Latif S, Rana R, Khalifa S (2019) Direct modelling of speech emotion from raw speech[C]. In: Interspeech 2019
12. Li J, Mi Y, Li G, Ju Z (2019) CNN-based facial expression recognition from annotated RGB-D images for human–robot interaction[J]. Int J Humanoid Robot 16(04):504–505
13. Lin M, Chen C, Lai C (2019) Object detection algorithm based AdaBoost residual correction fast R-CNN on network[C]. In: The 2019 3rd international conference
14. Luo Y, Ye J, Adams RB et al (2019) ARBEE: towards automated recognition of bodily expression of emotion in the wild[J]. Int J Comput Vis:1–25
15. Mohammed SN, Karim A (2020) Speech emotion recognition using MELBP variants of spectrogram image[J]. Int J Intell Eng Syst 13(5):257–266
16. Nie W, Yan Y, Song D et al (2020) Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition[J]. Multimed Tools Appl 4
17. Pan Z., Luo Z., Yang J, et al (2020) Multi-modal attention for speech emotion recognition. InterSpeech, 2020
18. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P (2017) Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 873–883
19. Poria S, Majumder N, Hazarika D, Cambria E, Gelbukh A, Hussain A (2018) Multimodal sentiment analysis: addressing key issues and setting up the baselines. IEEE Intell Syst 33(6):17–25
20. Ramanarayanan V, Pugh R, Qian Y, Suendermann-Oeft D Automatic turn-level language identification for code-switched Spanish-English dialog. In: Proc. of IWSDS 2018, International workshop on spoken dialog systems, Singapore, Singapore, vol 2018
21. Ren M, Nie W, Liu A et al (2019) Multi-modal correlated network for emotion recognition in speech[J]. Vis Inform 3(3)
22. Sahu G (2019) Multimodal speech emotion recognition and ambiguity resolution
23. Salama ES et al (2020) A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition[J]. Egypt Inform J

24. Satt A et al (2017) Efficient emotion recognition from speech using deep learning on spectrograms. Interspeech:1089–1093
25. Tripathi S, Tripathi S, Beigi H (2018) Multi-modal emotion recognition on IEMOCAP dataset using deep learning
26. Wang W, Enescu V, Sahli H (2015) Adaptive real-time emotion recognition from body movements[J]. ACM Trans Interact Intell Syst 5(4):1–21
27. Wu S, Li F, Zhang P (2019) Weighted feature fusion based emotional recognition for variable-length speech using DNN[C]. In: 2019 15th international wireless communications and Mobile computing conference (IWCMC)
28. Xu Y, Liu J, Zhai Y, Gan J, Zeng J, Cao H, Scotti F, Piuri V, Labati RD (2020) Weakly supervised facial expression recognition via transferred DAL-CNN and active incremental learning[J]. Soft Comput 24(8): 5971–5985
29. Zadeh A, Liang P, Mazumder N et al (2018) Memory fusion network for multi-view sequential learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence New Orleans, pp 5634–5641
30. Zhang L, Wang L, Dang J et al (2018) Convolutional neural network with spectrogram and perceptual features for speech emotion recognition[C]. In: International conference on neural information processing. Springer, Cham
31. Zhao J, Mao X, Chen L (2019) Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomed Signal Process Control 47(JAN.):312–323

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.