Multimedia Tools and Applications (2021) 80:8213–8240
https://doi.org/10.1007/s11042-020-10030-4

Video multimodal emotion recognition based on Bi-GRU and attention fusion

Ruo-Hong Huan[1] · Jia Shu[1] · Sheng-Lin Bao[1] · Rong-Hua Liang[1] · Peng Chen[1] · Kai-Kai Chi[1]

Received: 22 March 2020 / Revised: 25 August 2020 / Accepted: 6 October 2020 /
Published online: 31 October 2020

Abstract
A video multimodal emotion recognition method based on Bi-GRU and attention fusion is proposed in this paper. Bidirectional gated recurrent unit (Bi-GRU) is applied to improve the accuracy of emotion recognition in time contexts. A new network initialization method is proposed and applied to the network model, which can further improve the video emotion recognition accuracy of the time-contextual learning. To overcome the weight consistency of each modality in multimodal fusion, a video multimodal emotion recognition method based on attention fusion network is proposed. The attention fusion network can calculate the attention distribution of each modality at each moment in real-time so that the network model can learn multimodal contextual information in real-time. The experimental results show that the proposed method can improve the accuracy of emotion recognition in three single modalities of textual, visual, and audio, meanwhile improve the accuracy of video multimodal emotion recognition. The proposed method outperforms the existing state-of-the-art methods for multimodal emotion recognition in sentiment classification and sentiment regression.

Keywords Video emotion recognition · Multimodal · Bi-GRU · Attention mechanism · Fusion

1 Introduction

Usually, the ways humans naturally communicating and expressing emotions are multimodal [23]. That means we can express emotions either verbally or visually. When more emotions are expressed with tones, the audio data may contain major cues for emotion recognition; and when more facial expressions are used to express emotions, it can be considered that most of the clues needed for mining emotions exist in facial expressions. Identifying human emotions

✉ Ruo-Hong Huan
huanrh@zjut.edu.cn

Extended author information available on the last page of the article

using multimodal information such as human facial expressions, phonetic intonation, and linguistic content is an interesting and challenging issue.

Videos provide multimodal data in both acoustic and visual modalities. Facial expressions, vocal tones and text data in the video data can provide important information to recognize the true emotion state of a person better. Therefore, analyzing videos can create better models for emotion recognition and sentiment analysis. Textual, visual, and audio are often regarded as the main multimodal information in the research of multimodal emotion recognition about videos. The three modalities of textual, visual, and audio are simultaneously recognized and utilized, which can effectively extract the semantic and emotional information conveyed during the communication process.

It is necessary to simultaneously establish the emotion recognition models for the textual, visual and audio three modalities to utilize the three-modality data simultaneously. In the single-modality emotion recognition of textual [12, 13, 32–34], visual [1, 2, 6, 16, 42] and audio [17, 19, 25, 38, 40], some researches have achieved good recognition performance using deep learning. The recognition and utilization of the textual, visual and audio three-modality information requires the seamless integration of the three-modality information. The purpose of multimodal fusion is to combine information of multiple modalities, utilize the complementarity of heterogeneous data, provide more robust predictions, and improve the accuracy and reliability of recognition. Multimodal fusion is usually performed at the feature layer. Multiple high-dimensional features are computed into a fused feature, which is then input into a model for training. Morency et al. [23] first proposed a joint model of three modalities of textual, visual and audio for multimodal sentiment analysis, and conducted verification experiments. Poria S et al. [30], and Zhao J et al. [46] implemented fusion by concatenating the feature vectors of all three modalities to form a single long feature vector. The shortcoming of the above methods for extracting fused feature vectors is the weight consistency of each modality in the multimodal fusion. That is to say, the fact that the importance of each modality is not equal has not been taken into account. To overcome the weight consistency, Poria et al. [29] used a convolutional neural network (CNN) for multimodal sentiment analysis and proposed the convolutional MKL-based (C-MKL) model. Wang et al. [39] proposed the selective-additive learning-CNN (SAL-CNN) for multimodal sentiment analysis. Zadeh et al. [42] proposed a new model of tensor fusion network (TFN). Because of the introduction of the tensor representation, the costs of calculation and memory increase exponentially, which severely limits the application of the model, especially when there are more than three modalities in the dataset.

To make better use of textual, visual and audio three-modality data for video emotion recognition, a video multimodal emotion recognition method based on bidirectional gated recurrent unit (Bi-GRU) and attention fusion is proposed in this paper. The contributions of our work include: (1) A time-contextual learning method based on Bi-GRU is proposed. The Bi-GRU can improve the accuracy of video emotion recognition in the time-contextual learning. (2) A new network initialization method is proposed and applied to the network model. This initialization method can optimize the initialization parameters of the network model, improve the robustness of the Bi-GRU in the training and then improve the accuracy of emotion recognition. (3) A video multimodal emotion recognition method based on the attention fusion network is proposed. The attention mechanism is used to deal with the variation of the contextual state at each moment of multiple modalities. The distribution of attention at each moment of multiple modalities is calculated in real-time, so that the network model can learn the multimodal contextual information in real-time, thereby improving the accuracy of video emotion recognition under the multimodal fusion.

This paper is organized as follows: Related work is presented in Section 2. The video multimodal emotion recognition method based on Bi-GRU and attention fusion is described in Section 3. Section 4 presents experimental results and analysis, and Section 5 presents conclusions and discusses future research directions.

2 Related work

2.1 Single-modality emotion recognition

2.1.1 Textual modality

Research on textual emotion recognition has always been an active and extremely successful field. Notable works include the automatic recognition of opinionated words and their emotion polarity [11, 36], methods using n-grams and more complex language models [37, 41], and methods using polarity transfer rules or detailed feature engineering to solve the problem of emotion composition [22, 28]. Li et al. [18] proposed a hybrid approach to recognize word emotion in the dimension of eight emotion categories with corresponding intensities based on the Chinese emotion corpus. They explored approaches to identify word emotion from the aspect of general emotion attribute for a word. Experimental results showed that the integration of morpheme characteristics and semantic relations can improve the classification accuracy efficiently. These methods have been applied in many different areas, including mining opinions in Twitter and other online forums, analyzing political debates, answering questions, summarizing dialogues, and detecting citation emotion.

Research on textual emotion recognition based on deep learning has also been successful. Socher et al. [34] introduced recursive neural tensor networks and the Stanford sentiment treebank. The combination of a new model and data results for single sentence sentiment detection pushed positive/negative sentence classification and fine-grained sentiment prediction. Their research showed that the sentiment analysis for texts is far from solved. Iyyer et al. [12] introduced a deep averaging network (DAN) for textual emotion recognition. This was a simple and effective sentiment analysis model that used only the distribution of words to represent information, rather than the combined information of sentences, thus reducing the computational complexity. The model performed better than syntactic models on datasets with high syntactic variance. Kalchbrenner et al. [13] described a convolutional architecture called the dynamic convolutional neural network (DCNN) that was adopted for the semantic modelling of sentences. The network handled input sentences of varying length and induced a feature graph over the sentence that was capable of explicitly capturing short and long-range relations. The network did not rely on a parse tree and was easily applicable to any language. Seyeditabari et al. [32] formulated emotion recognition in text as a binary classification problem and presented a new network based on a Bi-GRU model to capture more meaningful information from text. They reported the results for two word embedding models which had the best performance. Shrivastava et al. [33] proposed a sequence-based CNN with word embedding to detect the emotions. An attention mechanism was applied in the proposed model which allowed CNN to focus on the words that had more effect on the classification or the part of the features that should be attended more.

2.1.2 Visual modality

Emotion recognition based on visual information is a research focus in the field of emotion computing and computer vision. Human facial expression is one of the most powerful means for humans to exchange emotions and intentions. Face analysis and video analysis methods based on deep learning have recently shown good performance on various key tasks such as face recognition, emotion recognition and activity recognition. In the previous work, the CNN mainly relies on time averaging and pooling to handle time-series sequences in video emotion recognition. The recurrent neural network (RNN) shows more advanced performance in time-series sequence analysis tasks, which has attracted great interest in recent years.

Byeon et al. [2] used 3D convolutional neural networks (3D-CNN) to extract facial features from speakers, and reduced dimensionality of the extracted features to simultaneously recognize continuous frames of facial expression images obtained by camera. It used local receptive fields and spatial down-sampling to achieve a certain degree of displacement and deformation invariance. Ebrahimi Kahou et al. [6] used both CNN and long short-term memory (LSTM) to propose a CNN-LSTM recurrent model. The face area of a speaker was convoluted into the LSTM at each timestamp. The face expression processing of speakers was similar to 3D-CNN. This architecture was superior to the CNN method that used time-average aggregation. Zadeh et al. [42] extracted facial expression features of speakers through FACET facial expression analysis framework, and proposed a RNN model based on FACET. It used FACET features every 6 frames as input information to the RNN with a memory dimension of 100 neurons, which was used as a baseline model for their follow-up experiments. Kumawat S et al. [16] proposed a novel 3D convolutional layer that called local binary volume (LBV) layer. LBV layer reduced the number of trainable parameters by a significant amount when compared to a conventional 3D convolutional layer. The LBVCNN network achieved comparable results compared to the state-of-the-art (SOTA) landmark-based or without landmark-based models on image sequences from CK+, Oulu-CASIA, and UNBC McMaster shoulder pain datasets. Bairaju et al. [1] utilized combination of CNNs and auto encoders to extract features for facial emotion detection and got considerable classification accuracy.

2.1.3 Audio modality

Automatically identifying spontaneous emotions from speech is a challenging task. On the one hand, acoustic features need to be powerful enough to capture the emotional content of various speaker styles. On the other hand, machine learning algorithms need to be insensitive to outliers while be able to model contexts. In recent years, research on audio emotion recognition based on deep learning has made great progress.

Lee et al. [17] presented a speech emotion recognition system using an RNN model trained by an efficient learning algorithm. The proposed system took into account the long-range contextual effect and the uncertainty of emotion label expressions. To extract high-level representation of emotion states with regard to its temporal dynamics, a powerful learning method with a bidirectional long short-term memory (BLSTM) architecture was adopted. Trigeorgis et al. [38] proposed a solution to the problem of 'context-aware' emotional relevant feature extraction, by combining CNNs with LSTM networks to automatically learn the best representation of the speech signal directly from the raw time representation. Lim et al. [19] proposed the speech emotion recognition (SER) method based on CNNs and RNNs. By applying the proposed methods to an emotional speech database, classification result was verified to have better accuracy than that achieved using conventional classification methods. Orjesek et al. [25] stacked convolution layer with Bi-GRU and had shown exceptional performance using only raw audio signals without any need for pre-processing. Wu et al. [40] presented a novel architecture based on the capsule networks (CapsNets) for SER. The proposed system took into account the spatial relationship of speech features in spectrograms, and provided an effective pooling method for obtaining utterance global features. The paper demonstrated the effectiveness of the CapsNets for SER.

The use of the LSTM networks solves the problem of speech context modelling, but how to capture the emotional features of the speech still needs to be actively studied, although more than a decade of research provides a large number of acoustic feature descriptions.

2.2 Video multimodal emotion recognition

Multimodal research has shown great progress in a variety of tasks as an emerging research field of artificial intelligence. It is an interesting and challenging problem to identify human emotions using human facial expressions, phonetic intonation and body gestures. Many people only studied emotional content in the language, or just used images to identify human facial expressions. Therefore, there were relatively few studies that combined multiple modalities to recognize human emotions. Textual, visual, and audio are often regarded as the main multimodal information in the research of multimodal emotion recognition about videos. The purpose of fusion is to improve the accuracy and reliability of recognition. The main advantage of analyzing emotions by analyzing videos rather than just texts is the abundance of behavioral cues. Text analysis requires the use of words, phrases, and dependencies between them, but it is known that only that information is not sufficient to extract relevant emotional content. Videos provide multimodal data in both acoustic and visual modalities. Facial expressions, vocal tones and text data in video data can provide important information to recognize the true emotion state of a person better. Therefore, analyzing videos can create better models for emotion recognition and sentiment analysis.

An important challenge for multimodal fusion is how to extend the fusion to multiple modalities while maintaining reasonable model complexity. Morency et al. [23] demonstrated a joint model that integrated visual, audio, and textual features can be effectively used to identify sentiment in Web videos. They used the joint model for sentiment analysis of product and movie reviews. They also identified a subset of audio-visual features relevant to sentiment analysis and presented guidelines on how to integrate these features. But their method was to directly connect the modal information in the early fusion representation and did not study the relationship between different modalities. Their experiments were conducted in a speaker-dependent manner, without analyzing the intensity of emotions. Park et al. [26] studied the persuasiveness of communication in social activities. They demonstrated that computational descriptors derived from verbal and nonverbal behavior can be predictive of persuasiveness. At the same time, they further proved that combining descriptors from multiple communication modalities (audio, text and visual) improved the prediction performance compared to using those from single modality alone. The C-MKL model proposed by Poria et al. [29] was a multimodal emotion classification model. They used the combined feature vectors of textual, visual, and audio modalities to train a classifier based on multiple kernel learning. However, their experiments focused on discourse rather than commentary, and their research methods depended on the emotion polarity rather than emotion intensity. Nojavanasghari et al. [24] also studied persuasion. They used a deep multimodal fusion architecture which was able to leverage complementary information from individual modalities for predicting persuasiveness. They trained single neural networks for each view's input and combined the views with a joint neural network. This baseline is the SOTA in the POM dataset. Wang et al. [39] used a select-additive learning (SAL) procedure that improved the generalizability of trained neural networks for multimodal sentiment analysis. In their experiments, they showed that their SAL approach improved the prediction accuracy significantly in all three modalities, as well as in their fusion. Zadeh et al. [44] presented a novel neural architecture for understanding human communication called the multi-attention recurrent network (MARN) for sentiment analysis. The main strength of this model came from discovering interactions between modalities through time using a neural component called the multi-attention block (MAB) and storing them in the hybrid memory of a recurrent component called the long-short term hybrid memory (LSTHM). Zadeh et al. [43] introduced a novel approach for multi-view sequential learning called memory fusion network (MFN) for multi-view sequential learning, which accounted for both view-specific and cross-view interactions. It continuously modeled them through time with a special attention mechanism and summarized through time with a multi-view gated memory. Liu Z et al. [20] proposed the low-rank multimodal fusion method, which performed multimodal fusion using low-rank tensors to improve efficiency. They performed experiments on other methods [24, 26, 31] under the CMU-MOSI dataset and POM dataset to prove the effectiveness of the proposed method. Ma L et al. [21] proposed an emotion computing algorithm based on cross-modal fusion and edge network data incentive. Deep cross-modal fusion can capture the semantic deviation between multiple modalities and design fusion methods through non-linear cross-layer mapping. The results of simulation experiments and theoretical analysis showed that the proposed algorithm was superior to the edge network data incentive algorithm and the cross-modal data fusion algorithm in recognition accuracy, complex emotion recognition efficiency, computation efficiency and delay.

The summary of related work in emotion recognition using deep learning is shown in Table 1.

3 The proposed method

The video multimodal emotion recognition method based on the Bi-GRU and attention fusion is shown in Fig. 1. The main steps of the method include: extract the high-dimensional features of the three modalities of textual, visual and audio from the inputting videos and then align and normalize the feature vectors according to the word level. Input them into the Bi-GRU network for training, using a new network initialization method to initialize the weights of the Bi-GRU network and the fully connected network in the initial training of each single-modality subnetwork. The state information output of the Bi-GRU network is processed by the maximum pooling layer and the average pooling layer. Splice two pooled feature vectors as input features. The input features in the three single-modality subnetworks are then used to calculate the correlation between the multimodal state information. Then the attention distribution of each modality at each moment is calculated, that is, the weight of the state information at each moment is calculated. The input features of the three single-modality subnetworks are weighted averaged with the corresponding weights to obtain the fused feature vector as the input of the fully connected network. Input the video to be recognized into the network after training to obtain the final emotion intensity output of the video.

Table 1 Summary of related work in emotion recognition using deep learning

Authors	Year	Modality	Approaches
Socher et al. [34]	2013	textual	recursive neural tensor networks
Iyyer et al. [12]	2014	textual	deep averaging network
Kalchbrenner et al. [13]	2015	textual	dynamic convolutional neural network (DCNN)
Seyeditabari et al. [32]	2019	textual	Bi-GRU
Shrivastava et al. [33]	2019	textual	sequence-based CNN with attention mechanism
Byeon et al. [2]	2014	visual	3D convolutional neural networks (3D-CNN)
Ebrahimi et al. [6]	2015	visual	CNN-LSTM recurrent model
Zadeh et al. [42]	2017	visual	RNN model with FACET
Kumawat S et al. [16]	2019	visual	LBVCNN network
Bairaju et al. [1]	2019	visual	CNNs and auto encoders
Lee et al. [17]	2015	audio	BLSTM
Trigeorgis et al. [38]	2016	audio	CNNs and LSTM
Lim et al. [19]	2016	audio	CNNs and RNNs
Orjesek et al. [25]	2019	audio	stacked convolution layer with Bi-GRU
Wu et al. [40]	2019	audio	capsule networks (CapsNets)
Poria et al. [29]	2015	multimodal	convolutional multiple kernel learning-based (C-MKL)
Nojavanasghari et al. [24]	2016	multimodal	deep multimodal fusion architecture
Wang et al. [39]	2017	multimodal	select-additive learning (SAL)
Zadeh et al. [44]	2018	multimodal	multi-attention recurrent network (MARN)
Zadeh et al. [43]	2018	multimodal	memory fusion network (MFN)
Liu Z et al. [20]	2018	multimodal	low-rank multimodal fusion
Ma L et al. [21]	2019	multimodal	cross-modal fusion with edge network data incentive

3.1 Video feature extraction for three modalities

3.1.1 Textual features

Global Vectors (GloVe), a word representation tool based on global word frequency statistics, can express a word as a vector of real numbers that captures some semantic features between words, such as similarity and analogy. The textual features of the video are defined as $l = \{l_1, l_2, l_3, \ldots, l_{T_l}; l_t \in \mathbb{R}^{300}\}$, where T_l is the number of words in the video, l_t represents a sequence of 300-dimensional GloVe word-vector feature [27].

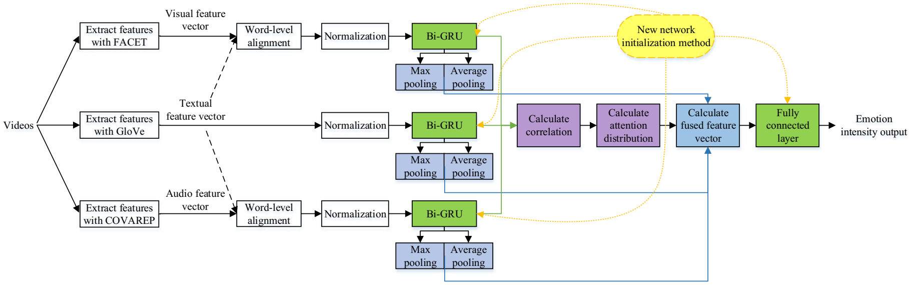

Fig. 1 Video multimodal emotion recognition based on Bi-GRU and attention fusion

3.1.2 Visual features

We use the FACET facial expression analysis framework to detect the face of the speaker in each frame, and extract seven basic emotions (anger, contempt, disgust, fear, joy, sadness, and surprise) and two advanced emotions (frustration and confusion) [7] from the speaker. Using the FACET can also extract a set of 20 facial action units [8] to indicate detailed muscle movements on the face.

We define the visual features as $v = \{v_1\ , v_2\ , v_3, \ldots, v_{T_v}\}$. The visual feature of the jth frame is $v_j = \left[v_j^1, v_j^2, v_j^3, \ldots, v_j^p\right]$, which contains a set of p visual features, where T_v is the total number of frames in the video. We use v as the input of the visual subnetwork. Since the information extracted by the FACET from videos is very rich, inputting them into the Bi-GRU can produce meaningful time-contextual high-dimensional features in the visual modality.

3.1.3 Audio features

For the audio portion of each video, using the COVAREP acoustic analysis framework to extract a set of acoustic features, including 12 mel-frequency cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmentation features, glottal source parameters, peak slope parameters [3], maxima dispersion quotients (MDQ) [14], and Liljencrants-Fant (LF) estimation of the parameters of the glottal model [9]. The voiced/unvoiced segmentation feature is a summation of residual harmonics (SRH) with robust additive noise [4], and the glottal source parameter is estimated by glottal back-filtering based on GCI synchronous IAIF [5]. These extracted features capture different features of the human voice and have been proven to be related to emotions [10].

Each segment is sampled at 100 Hz with T_a audio frames. We extract the set of q acoustic features $a_j = \left[a_j^1, a_j^2, a_j^3, \ldots, a_j^q\right]$ from the jth frame. The audio features of each segment are $a = \{a_1\ , a_2\ , a_3, \ldots, a_{T_a}\}$. Here we take a as the input of the audio subnetwork. Since the COVAREP extracts rich features from audio, using the Bi-GRU can extract continuous time-contextual high-dimensional features better in the audio modality.

3.1.4 Alignment and normalization

The dimension of the GloVe features extracted by the textual modality subnetwork of each segment is (T_l, 300), the dimension of the FACET features extracted by the visual modality subnetwork is (T_v, p), and the dimension of the COVAREP features extracted by the audio modality subnetwork is (T_a, q). The alignment of multimodal high-dimensional features is required [42], which is usually done in the word level. In this paper, the high-dimensional features of the visual and audio modalities are respectively aligned with the GloVe features of the textual modality according to T_l words in each segment. Specifically, record the start time and the end time of the ith word of the speech, and take the high-dimensional features of all frames in this period from the visual and audio modalities respectively. It is necessary to obtain the average features of each modality as the high-dimensional features of the corresponding modality according to the total sample number of each modality in this period. At this time, the high-dimensional features of the three modalities of textual, visual and audio are aligned in each segment. Define the number of high-dimensional features for three modalities to the number of high-dimensional features of the pre-aligned textual modality subnetwork, which is T_l.

Since the high-dimensional features extracted differ in the amplitudes, normalization is required. The normalization is to find the maximum values of the high-dimensional features of the three modalities, and the values of all the high-dimensional features are respectively divided by the maximum values in the corresponding modality. Normalization can map data to numbers in the range from 0 to 1. In the training process of neural networks, normalization can speed up network training and improve the convergence speed of the network.

3.2 Bi-GRU with new network initialization

3.2.1 Bi-GRU

The Bi-GRU network combines the model architectures of both GRU and BRNN networks. Replacing the network nodes in the RNN with the network nodes in the GRU makes it easier for the network to learn the time-contextual information. It can overcome the problem that the RNN cannot handle the long-term dependency well and causes gradient vanishing or gradient exploding in the back propagation. The BRNN network architecture can simultaneously access the information of the past time and the future time. Replacing the GRU network nodes with the network nodes in the BRNN, the new network architecture can fully learn and utilize the contextual information of the past and future moments.

The high-dimensional features of the three modalities after the word-level alignment and normalization are respectively used as the inputs of the Bi-GRU network. Take textual modality subnetwork as an example. Textual features $l = \{l_1\ ,l_2\ ,l_3,\ldots,l_{T_l}; l_t \in \mathbb{R}^{300}\}$ are input into the Bi-GRU network, where l_t represents a 300-dimensional GloVe word-vector feature. We define $\overrightarrow{G}(\cdot)$ as the forward calculation formula of the Bi-GRU network and $\overleftarrow{G}(\cdot)$ as the backward calculation formula, which are as follows:

$$\begin{aligned} \overrightarrow{h}_t &= \overrightarrow{G}\left(l_t, \overrightarrow{h}_{(t-1)}\right) \\ \overleftarrow{h}_t &= \overleftarrow{G}\left(l_t, \overleftarrow{h}_{(t+1)}\right) \end{aligned} \tag{1}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the forward state output and the backward state output respectively at the moment t of the Bi-GRU network, $\overrightarrow{h}_{(t\text{-}1)}$ is the forward state output of the moment t - 1, and $\overleftarrow{h}_{(t+1)}$ is the backward state output of the moment $t + 1$. The Bi-GRU network model architecture is shown in Fig. 2.

After the contextual information of the high-dimensional features is fully learned by the Bi-GRU network, the state information output of the network $H = [[\overleftarrow{h}_1\ , \overrightarrow{h}_1]\ [\overleftarrow{h}_2\ , \overrightarrow{h}_2], \ldots[\overleftarrow{h}_{T_l}\ , \overrightarrow{h}_{T_l}]]$ is obtained. The maximum pooling layer and the average pooling layer are used to extract features from the state information output of the Bi-GRU network. The pooling layers use overlapping aggregation technology. Pooling can reduce the feature vector dimension of the Bi-GRU network output. We extract high-dimensional representation vectors *max*(H) and *avg*(H) respectively, as follows:

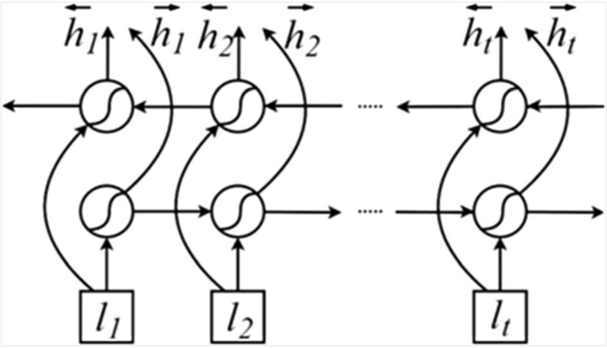

Fig. 2 Architecture of the Bi-GRU network model inputting with textual high-dimensional features

$$
\begin{aligned}
max(H) &= \left[\max_{1\le i\le T_l}\left(\overrightarrow{h}_i\right), \max_{1\le i\le T_l}\left(h_i\right)\right] \\
avg(H) &= \left[\underset{1\le i\le T_l}{avg}\left(\overrightarrow{h}_i\right), \underset{1\le i\le T_l}{avg}\left(h_i\right)\right]
\end{aligned}
\tag{2}
$$

The feature vector h^+ can be obtained by splicing the two pooled feature vectors, which is shown in the following formula:

$$h^+ = [max(H), avg(H)] \tag{3}$$

h^+ is considered as an input feature of the fully connected layer in the single-modality subnetwork. The fully connected layer maps the learned high-dimensional features to the sample label space as follows:

$$y = W_y h^+ + b_y \tag{4}$$

where W_y is the weight associated with h^+, b_y is the bias associated with h^+, and y is the emotion intensity output of a single-modality subnetwork.

The loss function of the training network is L1 Loss. L1 Loss can be used to create a standard that measures mean absolute error between each element in the input X and the target Y. The formula for calculating L1 Loss is as follows:

$$L(X, Y) = \{l_1, \ldots, l_N\}, l_N = |x_n – y_n| \tag{5}$$

where N is the number of elements in the input X, x_n is the nth element of the input X, y_n is the nth element of the target Y, and $| a - b |$ refers to the absolute value of the difference between a and b.

3.2.2 Network initialization

The Bi-GRU network is used in our core network layers and the ReLU activation function is used in the fully connected layer of our network model. Orthogonal initialization is more suitable for the Bi-GRU network, while Kaiming parameter initialization is more suitable for initializing the neuron parameters of the network for the ReLU activation function. We adjust the parameter initialization methods for the Bi-GRU network model and the fully connected layer simultaneously. Kaiming parameter initialization is used with the weights conforming to the normal distribution, and orthogonal initialization is also applied for a part of weights of the Bi-GRU to keep the eigenvalue of the orthogonal matrix to an absolute value of 1.

In our network model, the neuron parameters of the fully connected layer include weight W and bias b. Their default initialization methods are the same, which are shown as follows:

$$\begin{aligned} &W\sim U\left(-\sqrt{k},\sqrt{k}\right)\\ &b\sim U\left(-\sqrt{k},\sqrt{k}\right) \end{aligned} \tag{6}$$

where $U(-a, a)$ is the uniform distribution in the interval over $(-a, a)$, $k=\frac{1}{n_{in}}$, and n_{in} is the number of the input neurons.

We initialize the weight W according to the Kaiming initialization method and make it conform to the normal distribution, and set the bias b to a constant 0, which are shown as follows:

$$\begin{aligned} &W\sim N\left(0,\sqrt{\frac{2}{n_{in}}}\right)\\ &b=0 \end{aligned} \tag{7}$$

where $N(\mu, \sigma^2)$ means the standard normal distribution with expectation μ and standard deviation σ.

There are four kinds of neuron parameters in the Bi-GRU network, which are the weight of the input layer to the hidden layer W_{ih}, the bias of the input layer to the hidden layer b_{ih}, the weight of the hidden layer to the hidden layer W_{hh}, the bias of the hidden layer to the hidden layer b_{hh}. By default, the initialization methods for the four different neuron parameters are the same, which are shown as follows:

$$\begin{aligned} &W_{ih}\sim U\left(-\sqrt{k},\sqrt{k}\right)\\ &W_{hh}\sim U\left(-\sqrt{k},\sqrt{k}\right)\\ &b_{ih}\sim U\left(-\sqrt{k},\sqrt{k}\right)\\ &b_{hh}\sim U\left(-\sqrt{k},\sqrt{k}\right) \end{aligned} \tag{8}$$

where $k=\frac{1}{hiddensize}$, *hiddensize* is the number of features of the hidden state of the Bi-GRU network.

We initialize the weight W_{ih} in the Bi-GRU network according to the Kaiming initialization method and make it conform to the normal distribution. We use orthogonal initialization to initialize the weight W_{hh}, and set the bias b_{ih} and b_{hh} to a constant 0. The initialization methods are as follows:

$$\begin{aligned} &W_{ih}\sim N\left(0,\sqrt{\frac{2}{hiddensize}}\right)\\ &W_{hh}\sim Q\\ &b_{ih}=0\\ &b_{hh}=0 \end{aligned} \tag{9}$$

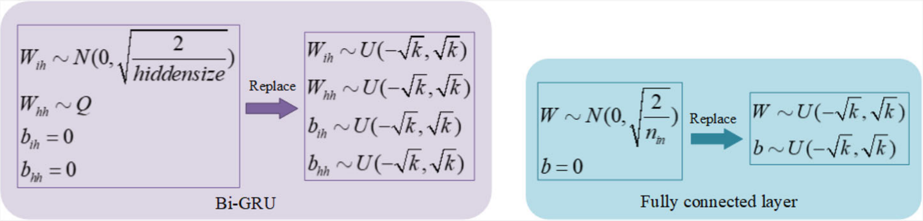
$W_{ih} \sim N(0, \sqrt{\frac{2}{hiddensize}})$
$W_{hh} \sim Q$
$b_{ih} = 0$
$b_{hh} = 0$
Replace
$W_{ih} \sim U(-\sqrt{k}, \sqrt{k})$
$W_{hh} \sim U(-\sqrt{k}, \sqrt{k})$
$b_{ih} \sim U(-\sqrt{k}, \sqrt{k})$
$b_{hh} \sim U(-\sqrt{k}, \sqrt{k})$
Bi-GRU
$W \sim N(0, \sqrt{\frac{2}{n_{in}}})$
$b = 0$
Replace
$W \sim U(-\sqrt{k}, \sqrt{k})$
$b \sim U(-\sqrt{k}, \sqrt{k})$
Fully connected layer

Fig. 3 The new network initialization method compared with the default

where Q is an orthogonal matrix unit, whose absolute value of the eigenvalue is 1. The new network initialization method is compared with the default as shown in Fig. 3.

3.3 Video multimodal emotion recognition based on attention fusion

In the video multimodal emotion recognition method based on attention fusion, the attention distribution of three modalities needs to be calculated at each moment, and the attention distribution is used as the weight of the state information output of the

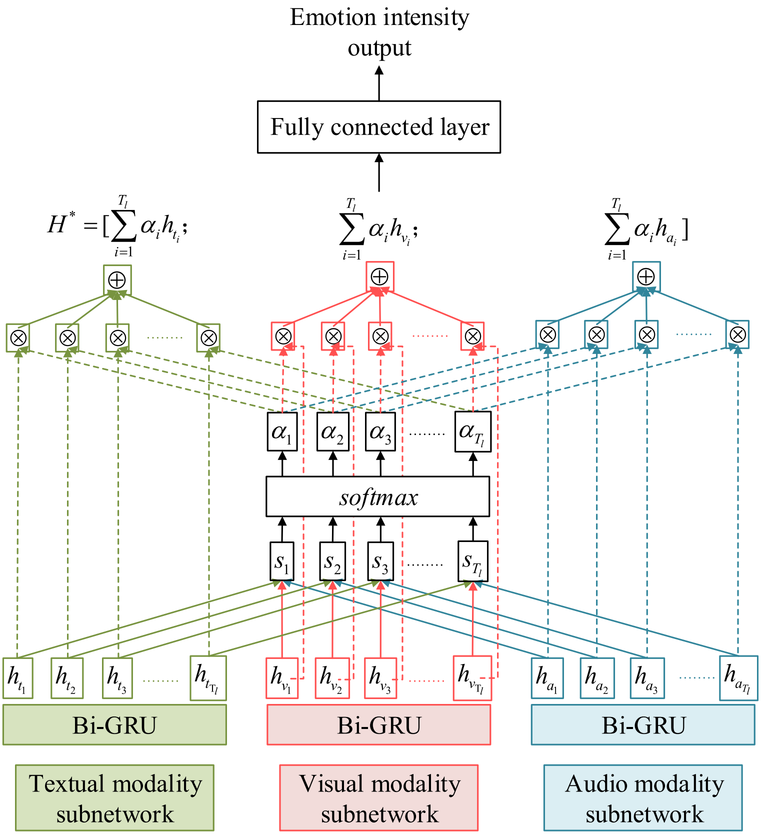
Emotion intensity output
Fully connected layer
$H^* = [\sum_{i=1}^{T_l} \alpha_i h_{t_i};$
$\sum_{i=1}^{T_l} \alpha_i h_{v_i};$
$\sum_{i=1}^{T_l} \alpha_i h_{a_i}]$
α_1 α_2 α_3 α_{T_l}
softmax
s_1 s_2 s_3 s_{T_l}
h_{t_1} h_{t_2} h_{t_3} $h_{t_{T_l}}$
h_{v_1} h_{v_2} h_{v_3} $h_{v_{T_l}}$
h_{a_1} h_{a_2} h_{a_3} $h_{a_{T_l}}$
Bi-GRU
Bi-GRU
Bi-GRU
Textual modality subnetwork
Visual modality subnetwork
Audio modality subnetwork

Fig. 4 The attention fusion network

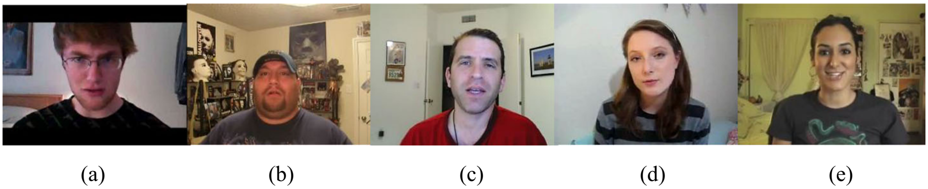

Fig. 5 Example snapshots of videos from CMU-MOSI dataset. (**a**) Highly negative, (**b**) Negative, (**c**) Neutral, (**d**) Positive, (**e**) Highly positive

Bi-GRU network in the corresponding modality subnetwork. The state information output of the Bi-GRU network is weighted averaged with the corresponding weight to obtain the fused feature vector. The fused feature vector is used as the input feature for the next fully connected layer, and the multimodal emotion intensity is finally obtained.

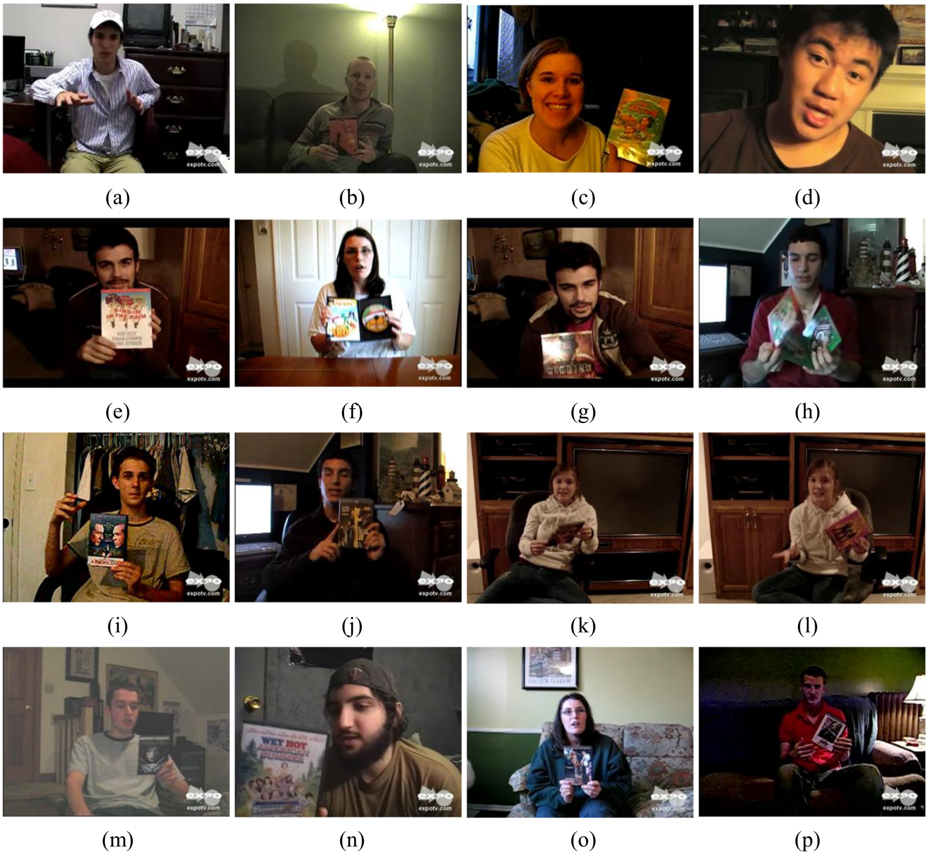

Fig. 6 Example snapshots of videos from POM dataset. (**a**) Confidence, (**b**) Passionate, (**c**) Voice pleasant, (**d**) Dominant, (**e**) Credible, (**f**) Vivid, (**g**) Expertise, (**h**) Entertaining, (**i**) Reserved, (**j**) Trusting, (**k**) Relaxed, (**l**) Outgoing, (**m**) Thorough, (**n**) Nervous, (**o**) Persuasive, (**p**) Humorous

3.3.1 Correlation calculation of the state information between multiple modalities

We define that the state information output by the Bi-GRU network of the high-dimensional features as $H = [[\overleftarrow{h}_1, \overrightarrow{h}_1]\ [\overleftarrow{h}_2, \overrightarrow{h}_2], \dots [\overleftarrow{h}_{T_l}, \overrightarrow{h}_{T_l}]]$ after learning the contextual information sufficiently, where $\overrightarrow{h}_t$ and h_t are the forward state output and the backward state output of the Bi-GRU network at the moment t. Thus, we define the state information of the textual modality subnetwork as $H_t = [[\overleftarrow{h}_{t_1}, \overrightarrow{h}_{t_1}], [h_{t_2}, \overrightarrow{h}_{t_2}], \dots [\overleftarrow{h}_{t_{T_l}}, \overrightarrow{h}_{t_{T_l}}]]$. The state information of the visual modality subnetwork is $H_v = [[\overleftarrow{h}_{v_1}, \overrightarrow{h}_{v_1}]\ [\overleftarrow{h}_{v_2}, \overrightarrow{h}_{v_2}], \dots [\overleftarrow{h}_{v_{T_l}}, \overrightarrow{h}_{v_{T_l}}]]$. And the state information of the audio modality subnetwork is $H_a = [[\overleftarrow{h}_{a_1}, \overrightarrow{h}_{a_1}]\ [\overleftarrow{h}_{a_2}, \overrightarrow{h}_{a_2}], \dots [\overleftarrow{h}_{a_{T_l}}, \overrightarrow{h}_{a_{T_l}}]]$. $\overrightarrow{h}_{t_{T_l}}$ and $h_{t_{T_l}}$ respectively are the forward state output and the backward state output of the Bi-GRU network in the textual modality subnetwork at the moment t. $\overrightarrow{h}_{v_{T_l}}$ and $h_{v_{T_l}}$ respectively are the forward state output and the backward state output of the Bi-GRU network in the visual modality subnetwork at the moment t. $\overrightarrow{h}_{a_{T_l}}$and $h_{a_{T_l}}$ respectively are the forward state output and the backward state output of the Bi-GRU network in the audio modality subnetwork at the moment t. In the previous section, we performed the word-level alignment on the high-dimensional features of the visual and audio modalities with the high-dimensional features of the textual modality. Therefore, the time step of the state information of three single-modality subnetworks is all T_l.

The state neurons of the Bi-GRU network are formed by a part of forward calculation and a part of backward calculation at the hidden layer respectively. The current time step of the state information of each single-modality subnetwork is T_l, that is, the state neurons go through T_l time steps for forward calculation and T_l time steps for backward calculation. The essence of the attention fusion network is to extract the useful fused feature vector H^* from the state information of the Bi-GRU network output H_t, H_v and H_a in the three single-modality subnetworks.

We use the attention mechanism to consider the importance of each state information and calculate the attention distribution of each state information as the weight α_i of the corresponding state information. Since the state information between multiple modalities is taken into consideration, the weight α_i will simultaneously pay attention to the state information of the three modalities, that is, the correlation of the state information between multiple modalities s_i is related to the state information of each moment of the three single-modality subnetworks. The correlation s_i is calculated as follows:

$$s_i = V\tanh(W_t \cdot h_{t_i} + W_v \cdot h_{v_i} + W_a \cdot h_{a_i} + b_1) + b_2 \tag{10}$$

Table 2 The sample numbers of the datasets

Dataset	CMU-MOSI	POM
Training set	1284	600
Validation set	229	100
Testing set	686	203

Table 3 Comparison of emotion recognition results of various network models in the textual modality under the CMU-MOSI dataset

Method	Binary (%)				5-class (%)				Regression	
	PR	FPR	RE/Acc	F1	PR	FPR	RE/Acc	F1	MAE	Corr
LSTM	71.8	28.6	71.9	71.8	44.5	17.1	35.9	34.8	1.03	0.60
GRU	72.8	28.4	72.7	72.3	45.5	17.2	35.7	34.2	1.03	0.61
BLSTM	74.1	27.2	73.9	73.5	48.3	16.6	37.9	36.8	0.98	0.64
Bi-GRU	74.9	25.4	74.9	74.9	48.6	15.8	40.2	39.7	0.95	0.66
BLSTM$_{init}$	74.4	26.9	74.2	73.8	49.3	16.4	38.8	37.9	0.96	0.65
Bi-GRU$_{init}$	**75.7**	**24.6**	**75.7**	**75.7**	**49.4**	**15.6**	**41.1**	**40.7**	**0.93**	**0.66**

where $h_{t_i} = [\overrightarrow{h}_{t_i}, \overleftarrow{h}_{t_i}]$ is the state information output by the Bi-GRU network in the textual modality subnetwork at the moment i, including the forward state output $\overrightarrow{h}_{t_i}$ and the backward state output h_{t_i}. W_t is the weight associated with h_{t_i}. $h_{v_i} = [\overrightarrow{h}_{v_i}, \overleftarrow{h}_{v_i}]$ is the state information output by the Bi-GRU network in the visual modality subnetwork at the moment i, including the forward state output $\overrightarrow{h}_{v_i}$ and the backward state output h_{v_i}. W_v is the weight associated with $\overrightarrow{h}_{v_i}$. $h_{a_i} = [\overrightarrow{h}_{a_i}, \overleftarrow{h}_{a_i}]$ is the state information output by the Bi-GRU network in the audio modality subnetwork at the moment i, including the forward state output $\overrightarrow{h}_{a_i}$ and the backward state output h_{a_i}. W_a is the weight associated with h_{a_i}. b_1 is the bias associated with h_{t_i}, h_{v_i} and h_{a_i}. tanh is the activation function. V is the weight of multimodal fusion. b_2 is the bias of multimodal fusion.

3.3.2 Generation of the fused feature vectors

According to the current correlation of multimodal state information s_i, we can calculate the attention distribution at each moment in multiple modalities, that is, the weight α_i corresponding to the state information. The calculation of weight α_i is as follows:

$$\alpha_i = \text{softmax}(s_i) = \frac{exp(s_i)}{\sum_{j=1}^{T_l} exp(s_j)} \tag{11}$$

where softmax is a normalized exponential function.

Table 4 Comparison of emotion recognition results of various network models in the visual modality under the CMU-MOSI dataset

Method	Binary (%)				5-class (%)				Regression	
	PR	FPR	RE/Acc	F1	PR	FPR	RE/Acc	F1	MAE	Corr
LSTM	67.7	35.7	66.6	64.8	63.4	15.4	33.7	31.7	1.12	0.47
GRU	69.5	34.2	68.1	66.3	62.9	15.3	34.5	33.2	1.09	0.50
BLSTM	69.4	33.9	68.2	66.6	65.4	15.0	35.1	33.5	1.08	0.51
Bi-GRU	70.2	33.6	68.7	66.9	63.9	15.1	35.4	34.3	1.07	0.52
BLSTM$_{init}$	70.0	33.3	68.8	67.3	65.2	14.8	36.3	35.0	1.05	0.54
Bi-GRU$_{init}$	**71.4**	**32.4**	**69.8**	**68.3**	**65.8**	**14.6**	**37.6**	**37.2**	**1.03**	**0.55**

Table 5 Comparison of emotion recognition results of various network models in the audio modality under the CMU-MOSI dataset

Method	Binary (%)				5-class (%)				Regression	
	PR	FPR	RE/Acc	F1	PR	FPR	RE/Acc	F1	MAE	Corr
LSTM	67.7	33.7	65.0	64.7	52.9	16.5	32.9	31.3	1.20	0.40
GRU	68.0	33.1	65.9	65.8	54.2	16.3	34.5	33.6	1.18	0.40
BLSTM	67.7	33.7	65.0	64.7	54.7	16.3	33.8	32.4	1.19	0.40
Bi-GRU	**69.1**	**32.1**	**66.8**	**66.6**	**56.0**	16.0	**35.4**	**34.6**	**1.15**	**0.44**
BLSTM$_{init}$	69.1	32.3	66.5	66.3	55.9	**15.9**	35.4	34.6	1.15	0.44
Bi-GRU$_{init}$	68.1	32.9	66.2	66.2	53.2	16.2	35.0	34.3	1.17	0.40

The state information output by the Bi-GRU network are weighted averaged with the corresponding weight α_i to obtain the fused feature vector H^* as the input feature of the fully connected layer. The calculation of the fused feature vectors H^* is as follows:

$$H^* = \left[\sum_{i=1}^{T_l} \alpha_i h_{t_i}; \sum_{i=1}^{T_l} \alpha_i h_{v_i}; \sum_{i=1}^{T_l} \alpha_i h_{a_i} \right] \tag{12}$$

The architecture of the attention fusion network is shown in Fig. 4.

4 Experimental results and analysis

4.1 Datasets

To verify the validity of the proposed method, we use the Carnegie Mellon University multimodal opinion sentiment intensity (CMU-MOSI) dataset [45] and persuasion opinion multimodal (POM) dataset [26] for video multimodal emotion recognition experiments.

CMU-MOSI is an annotated dataset of video comments from YouTube providing three modality data of textual, visual and audio. The annotation of sentiment of CMU-MOSI closely follows the annotation scheme of the Stanford sentiment treebank [35], where sentiment is annotated on a seven-step Likert scale from highly negative to highly positive. The emotion intensity annotation is done by online staff on Amazon Mechanical Turk website. Emotion intensity ranges from −3 to +3. There are 93 different

Table 6 Comparison of emotion recognition results of various network models in the textual modality under the POM dataset

Method	Multi-label classification (%)				Regression	
	PR	FPR	RE/Acc	F1	MAE	Corr
LSTM	14.8	33.2	32.9	19.0	0.87	0.18
GRU	28.7	30.5	34.2	28.3	0.89	0.12
BLSTM	24.3	31.2	33.4	25.7	0.85	0.26
Bi-GRU	28.6	29.9	33.4	27.2	0.87	0.18
BLSTM$_{init}$	27.9	30.4	34.4	26.7	**0.84**	0.28
Bi-GRU$_{init}$	**41.8**	**26.8**	**42.1**	**38.2**	0.85	**0.28**

Table 7 Comparison of emotion recognition results of various network models in the visual modality under the POM dataset

Method	Multi-label classification (%)				Regression	
	PR	FPR	RE/Acc	F1	MAE	Corr
LSTM	23.8	32.0	34.6	24.7	0.88	0.09
GRU	32.1	30.9	34.2	27.3	0.86	0.21
BLSTM	32.2	31.8	35.0	27.8	0.86	0.17
Bi-GRU	32.2	30.4	35.5	28.6	0.87	0.19
BLSTM$_{init}$	27.7	31.7	35.4	26.6	0.86	**0.23**
Bi-GRU$_{init}$	**39.6**	**28.3**	**41.0**	**34.0**	**0.86**	0.22

speakers in the CMU-MOSI dataset, and 2199 opinion speech videos. There are 26,295 words in the commentary video. There is an average of 23.2 opinion segments per video, and the average length of each video is 4.2 s. The example snapshots of videos from CMU-MOSI dataset are shown in Fig. 5.

POM is a dataset for analysis of persuasion on online social media. It has annotations for personality and sentiment as well, which makes it very compelling for large numbers of tasks. Each video is annotated on a seven-step Likert scale with 1 being the least descriptive of the trait and 7 being the most descriptive. The speaker traits are listed as follows: confident (con), passionate (pas), voice pleasant (voi), dominant (dom), credible (cre), vivid (viv), expertise (exp), entertaining (ent), reserved (res), trusting (tru), relaxed (rel), outgoing (out), thorough (tho), nervous (ner), persuasive (per) and humorous (hum). The short forms of these speaker traits are indicated inside the parentheses and used for the rest of this paper. The example snapshots of videos from POM dataset are shown in Fig. 6.

In the experiments, we implemented binary sentiment classification, 5-class sentiment classification and sentiment regression on the CMU-MOSI dataset. The regression range is [−3, 3]. We also implemented multi-label classification of different speaker traits and speaker traits regression on the POM dataset. For classification, we use precision (PR), false positive rate (FPR), recall (RE), accuracy (Acc) and F1 score for evaluation; and for regression, mean absolute error (MAE) and Pearson product-moment correlation coefficients (Corr) between model predictions and real values are used for evaluation. Higher values denote better performance for all metrics except for FPR and MAE.

Table 8 Comparison of emotion recognition results of various network models in the audio modality under the POM dataset

Method	Multi-label classification (%)				Regression	
	PR	FPR	RE/Acc	F1	MAE	Corr
LSTM	21.1	33.3	34.1	22.1	0.88	0.16
GRU	25.1	32.2	34.1	24.8	0.87	0.19
BLSTM	24.5	32.7	33.8	23.4	0.87	0.15
Bi-GRU	28.1	32.1	34.3	23.3	0.87	0.19
BLSTM$_{init}$	24.0	33.2	33.7	22.6	0.87	0.17
Bi-GRU$_{init}$	**31.7**	**30.8**	**38.9**	**29.4**	**0.86**	**0.22**

Table 9 Comparison with the SOTA network models in the textual modality under the CMU-MOSI dataset

Method	Binary (%)		5-class (%)	Regression	
	Acc	F1	Acc	MAE	Corr
RNTN [34, 42]	73.7	73.4	35.2	0.99	0.59
DAN [12, 42]	73.4	73.8	39.2	–	–
D-CNN [13, 42]	65.5	66.9	32.0	–	–
C-MKL-T [29, 42]	71.2	72.4	34.5	–	–
SAL-CNN-T [39, 42]	73.5	–	–	–	–
SVM-MD-T [42]	70.6	71.2	33.1	1.18	0.46
TFN-T [42]	74.8	75.6	38.5	0.98	0.62
Bi-GRU$_{init}$ textual	**75.7**	**75.7**	**41.4**	**0.93**	**0.66**

4.2 Experimental setup

In single-modality experiments, different network models use the same hyper-parameter settings for the convenience of comparison. The models are trained using the Adam optimizer [15] with the epoch size 50. The early stopping training method is used to monitor the loss value of the validation set. That is, when the loss value of the validation set is not reduced for 10 consecutive times, the training process will stop. While compared with the SOTA network models, the best hyper-parameters are chosen using grid search based on model performance on a validation set. The training, testing and validation folds are exactly the same for all network models. The sample numbers of the datasets are shown in Table 2.

4.3 Results and analysis

4.3.1 Experimental results for single modality

The emotion recognition results of various network models in the textual, visual, and audio modalities under the CMU-MOSI dataset are compared in Tables 3, 4 and 5 respectively. It can be seen from Tables 3 and 4 that in the binary sentiment classification, 5-class sentiment classification and sentiment regression, the emotion recognition results based on the Bi-GRU with new network initialization in the textual modality and visual modality are superior to the results based on the methods of LSTM, GRU, BLSTM, Bi-GRU, and BLSTM with new network initialization. However, it can be seen from Table 5 that in the binary sentiment

Table 10 Comparison with the SOTA network models in the visual modality under the CMU-MOSI dataset

Method	Binary (%)		5-class (%)	Regression	
	Acc	F1	Acc	MAE	Corr
3D-CNN [2, 42]	56.1	58.4	24.9	1.31	0.26
CNN-LSTM [6, 42]	60.7	61.2	25.1	1.27	0.30
LSTM-FA [42]	62.1	63.7	26.2	1.23	0.33
C-MKL-V [29, 42]	52.6	58.5	29.3	–	–
SAL-CNN-V [39, 42]	63.8	–	–	–	–
SVM-MD-V [42]	59.2	60.1	25.6	1.24	0.36
TFN-V [42]	69.4	**71.4**	31.0	1.12	0.50
Bi-GRU$_{init}$ visual	**69.8**	68.3	**37.6**	**1.03**	**0.55**

Table 11 Comparison with the SOTA network models in the audio modality under the CMU-MOSI dataset

Method	Binary (%)		5-class (%)	Regression	
	Acc	F1	Acc	MAE	Corr
HL-RNN [17, 42]	63.4	64.2	25.9	1.21	0.34
Adieu-Net [38, 42]	59.2	60.6	25.1	1.29	0.31
SER-LSTM [19, 42]	55.4	56.1	24.2	1.36	0.23
C-MKL-A [29, 42]	52.6	58.5	29.1	–	–
SAL-CNN-A [39, 42]	62.1	–	–	–	–
SVM-MD-A [42]	56.3	58.0	24.6	1.29	0.28
TFN-A [42]	65.1	**67.3**	27.5	1.23	0.36
Bi-GRU$_{init}$ audio	**66.2**	66.2	**35.0**	**1.17**	**0.40**

classification, 5-class sentiment classification and sentiment regression, the emotion recognition results based on the Bi-GRU in the audio modality (except for FPR in 5-class classification) are superior to the results based on the methods of LSTM, GRU, BLSTM, BLSTM with new network initialization, and Bi-GRU with new network initialization.

The emotion recognition results of various network models in the textual, visual, and audio modalities under the POM dataset are compared in Tables 6, 7 and 8 respectively. For classification results, we use the average of the results of all labels. It can be seen in multi-label classification of speaker traits, the results based on the Bi-GRU with new network initialization in three single modalities are superior to the results based on the methods of LSTM, GRU, BLSTM, Bi-GRU and BLSTM with new network initialization. In regression of speaker traits, the results based on the Bi-GRU with new network initialization in the audio modality are superior to the results based on the methods of LSTM, GRU, BLSTM, Bi-GRU and BLSTM with new network initialization. But in the textual and visual modalities, the results based on the Bi-GRU with new network initialization are slightly worse than the results based on the BLSTM with new network initialization in a certain metric.

The experimental results obtained under the CMU-MOSI dataset are compared with the results of the SOTA network models in the textual modality, visual modality and audio modality, as shown in Tables 9, 10 and 11. It can be known from Table 9 that in the textual modality, the proposed method (Bi-GRU$_{init}$ textual) is superior to the SOTA network models in the binary sentiment classification, 5-class sentiment classification and sentiment regression. It can be seen from Tables 10 and 11 that in the visual and audio modalities, the proposed methods (Bi-GRU$_{init}$ visual and Bi-GRU$_{init}$ audio) are superior to the SOTA network models in the 5-class sentiment classification and sentiment regression. The F1 scores of TFN [42] method in the binary sentiment classification are better than the proposed methods, while the proposed methods are the best in the accuracy of the binary sentiment classification among all the SOTA network models.

The experimental results obtained under the POM dataset are compared with the results of MFN [43] method in the textual modality, visual modality and audio modality, as shown in Tables 12, 13 and 14. It can be known that in the three single modalities, for most different speaker traits, the proposed methods (Bi-GRU$_{init}$ textual, Bi-GRU$_{init}$ visual and Bi-GRU$_{init}$ audio) are superior to the MFN method in multi-label classification and regression, except for Thorough (Tho) regression (MAE) in the textual modality, Humorous (Hum) classification (Acc) in the visual modality, and Trusting (Tru) regression (Corr) in the visual modality.

Table 12 Comparison with the SOTA network models in the textual modality under the POM dataset

Task k-class	Con 7	Pas 7	Voi 7	Dom 7	Cre 7	Viv 7	Exp 7	Ent 7	Res 5	Tru 5	Rel 5	Out 5	Tho 5	Ner 5	Per 7	Hum 5	Avg
Metric(%)									**Acc**								
MFN-T[43]	26.6	31.5	21.7	34.0	25.6	28.6	26.6	30.5	29.1	34.5	39.9	31.5	30.5	34.0	24.1	42.4	30.7
Bi-GRU$_{init}$ textual	**31.0**	**36.0**	**37.9**	**42.9**	**39.9**	**41.4**	**36.9**	**37.4**	**39.9**	**58.1**	**53.7**	**43.3**	**47.3**	**45.8**	**35.0**	**46.3**	**42.1**
Metric									**MAE**								
MFN-T[43]	1.07	1.15	1.03	0.88	1.07	1.11	1.14	0.99	0.92	0.59	0.61	0.79	**0.75**	0.72	1.13	0.84	0.92
Bi-GRU$_{init}$ textual	**1.00**	**0.96**	**0.94**	**0.86**	**0.92**	**0.95**	**0.97**	**0.97**	**0.87**	**0.52**	**0.60**	**0.70**	0.79	**0.69**	**1.04**	**0.81**	**0.85**
Metric									**Corr**								
MFN-T[43]	0.22	0.28	−0.01	0.12	0.14	0.19	0.19	0.23	−0.17	−0.06	0.13	0.10	0.17	0.02	0.18	0.22	0.12
Bi-GRU$_{init}$ textual	**0.42**	**0.45**	**0.12**	**0.23**	**0.40**	**0.37**	**0.33**	**0.31**	**0.28**	**0.21**	**0.17**	**0.14**	**0.29**	**0.19**	**0.31**	**0.32**	**0.28**

Table 13 Comparison with the SOTA network models in the visual modality under the POM dataset

Task k-class	Con 7	Pas 7	Voi 7	Dom 7	Cre 7	Viv 7	Exp 7	Ent 7	Res 5	Tru 5	Rel 5	Out 5	Tho 5	Ner 5	Per 7	Hum 5	Avg
Metric(%)									**Acc**								
MFN-V[43]	25.6	23.6	26.6	31.5	25.1	28.6	25.6	26.6	32.5	48.3	43.3	36.9	42.4	33.5	24.1	**37.4**	34.0
Bi-GRU$_{init}$ visual	**38.9**	**33.0**	**38.9**	**40.4**	**32.0**	**36.0**	**35.0**	**37.9**	**37.4**	**58.1**	**51.7**	**45.3**	**46.8**	**48.3**	**33.0**	33.5	**40.4**
Metric									**MAE**								
MFN-V[43]	1.08	1.15	1.01	0.93	1.09	1.07	1.14	1.03	0.93	0.66	0.68	0.77	0.77	0.77	1.14	0.79	0.94
Bi-GRU$_{init}$ visual	**0.95**	**1.06**	**0.92**	**0.86**	**1.04**	**0.98**	**0.97**	**0.95**	**0.87**	**0.54**	**0.59**	**0.69**	**0.76**	**0.69**	**1.11**	**0.79**	**0.86**
Metric									**Corr**								
MFN-V[43]	0.15	0.09	−0.08	−0.01	0.02	−0.04	0.01	0.04	0.00	**−0.17**	0.03	−0.03	0.05	0.06	0.08	0.16	0.02
Bi-GRU$_{init}$ visual	**0.35**	**0.31**	**0.18**	**0.20**	**0.18**	**0.21**	**0.27**	**0.31**	**0.23**	0.09	**0.10**	**0.26**	**0.15**	**0.22**	**0.16**	**0.23**	**0.22**

Table 14 Comparison with the SOTA network models in the audio modality under the POM dataset

Task k-class	Con 7	Pas 7	Voi 7	Dom 7	Cre 7	Viv 7	Exp 7	Ent 7	Res 5	Tru 5	Rel 5	Out 5	Tho 5	Ner 5	Per 7	Hum 5	Avg
Metric(%)									**Acc**								
MFN-A[43]	27.1	26.1	29.6	34.5	24.6	29.6	26.6	31.0	32.5	35.0	45.8	37.4	35.0	40.4	28.1	36.5	32.5
Bi-GRU$_{init}$ audio	**30.5**	**27.6**	**35.0**	**36.5**	**31.0**	**36.9**	**30.0**	**32.5**	**36.9**	**56.7**	**53.7**	**46.8**	**42.9**	**47.8**	**33.0**	**44.3**	**38.9**
Metric									**MAE**								
MFN-A[43]	1.09	1.15	0.94	0.89	1.10	1.03	1.08	1.01	0.88	0.59	0.61	0.72	0.76	0.70	1.16	0.78	0.91
Bi-GRU$_{init}$ audio	**1.01**	**1.10**	**0.91**	**0.86**	**1.03**	**1.00**	**1.04**	**0.95**	**0.84**	**0.53**	**0.57**	**0.66**	**0.74**	**0.68**	**1.08**	**0.78**	**0.86**
Metric									**Corr**								
MFN-A[43]	0.09	0.13	−0.02	0.05	0.02	−0.01	0.04	0.13	0.15	−0.07	0.02	0.00	−0.02	0.11	0.02	0.06	0.04
Bi-GRU$_{init}$ audio	**0.31**	**0.22**	**0.15**	**0.14**	**0.15**	**0.11**	**0.17**	**0.27**	**0.35**	**0.22**	**0.26**	**0.29**	**0.15**	**0.28**	**0.18**	**0.23**	**0.22**

Table 15 Comparison of the emotion recognition results for three single modalities and multimodal fusion under the CMU-MOSI dataset

Method	Binary(%)				5-class(%)				Regression	
	PR	FPR	RE/Acc	F1	PR	FPR	RE/Acc	F1	MAE	Corr
Bi-GRU$_{init}$ textual	75.7	24.6	75.7	75.7	49.4	15.6	41.1	40.7	0.93	0.66
Bi-GRU$_{init}$ visual	71.4	32.4	69.8	68.3	**65.8**	14.6	37.6	37.2	1.03	0.55
Bi-GRU$_{init}$ audio	68.1	32.9	66.2	66.2	53.2	16.2	35.0	34.3	1.17	0.40
Bi-GRU$_{init}$ multimodal	**78.8**	**22.0**	**78.7**	**78.6**	53.7	**14.5**	**45.8**	**46.0**	**0.84**	**0.71**

4.3.2 Experimental results for multimodal emotion recognition

The comparison of the emotion recognition results based on the methods for three single modalities (Bi-GRU$_{init}$ textual, Bi-GRU$_{init}$ visual and Bi-GRU$_{init}$ audio) and for multimodal fusion (Bi-GRU$_{init}$ multimodal) under the CMU-MOSI dataset and the POM dataset are shown in Tables 15 and 16 respectively. It can be seen from Table 15 that the sentiment classification and sentiment regression results of the multimodal fusion are significantly better than those of three single modalities under the CMU-MOSI dataset. It can be seen from Table 16 that in multi-label classification and regression, for most speaker traits, the results of the multimodal fusion (including PR, FPR, RE, Acc, F1, MAE, and Corr) are better than those of three single modalities under the POM dataset. Thus, the average (Avg) of all metrics of the multimodal fusion are all better than those of the three single modalities.

The emotion recognition results of the proposed multimodal fusion (Bi-GRU$_{init}$ multimodal) and the SOTA multimodal network models under the CMU-MOSI dataset and the POM dataset are compared in Tables 17 and 18 respectively. It can be seen from Tables 17 and 18 that the emotion recognition performance of the proposed video multimodal emotion recognition based on the Bi-GRU with new network initialization and the attention fusion network is superior to that of the listing multimodal emotion recognition methods in sentiment classification and sentiment regression.

5 Conclusions

A time-contextual learning method based on the Bi-GRU network is proposed in this paper. In the process of video emotion recognition, the output of the current moment is not only related to the previous state, but also related to the state after it. The Bi-GRU can improve the accuracy of emotion recognition in the time-contextual learning. In order to further improve the accuracy of video emotion recognition, a new network initialization method is proposed and applied to the network model. This initialization method can optimize the initialization parameters of the ReLU network model, improve the robustness in the training of the Bi-GRU network and improve the accuracy of emotion recognition. A video multimodal emotion recognition based on the attention fusion network is proposed to overcome the weight consistency of each modality in multimodal fusion. The attention mechanism is used to process the variation of multimodal context state at each moment, and the attention distribution

Table 16 Comparison of the emotion recognition results for three single modalities and multimodal fusion under the POM dataset

Task k-class	Con 7	Pas 7	Voi 7	Dom 7	Cre 7	Viv 7	Exp 7	Ent 7	Res 5	Tru 5	Rel 5	Out 5	Tho 5	Ner 5	Per 7	Hum 5	Avg
Metric(%)									**PR**								
Bi-GRU$_{\text{init}}$ textual	31.6	**41.6**	25.8	36.4	**47.3**	37.6	34.2	34.3	41.5	50.0	49.0	31.1	**55.2**	34.0	**34.4**	**47.2**	39.5
Bi-GRU$_{init}$ visual	**49.0**	37.6	35.4	36.6	34.9	28.9	**39.2**	30.4	40.0	49.4	41.0	**47.4**	51.6	44.6	29.2	37.6	39.6
Bi-GRU$_{init}$ audio	36.3	14.8	24.4	27.8	25.0	28.7	24.2	25.5	35.5	46.8	**50.4**	34.0	39.4	36.4	23.4	34.1	31.7
Bi-GRU$_{init}$ multimodal	39.4	37.4	**35.5**	**38.4**	29.6	**38.3**	37.4	**37.2**	**46.7**	**58.2**	49	44.5	52.7	**54.9**	30.4	39.9	**41.8**
Metric(%)									**FPR**								
Bi-GRU$_{\text{init}}$ textual	20.6	19.9	**27.2**	25.9	**20.3**	**19.2**	20.7	23.0	25.1	40.9	42.2	34.1	**26.4**	34.7	**20.4**	31.6	27.0
Bi-GRU$_{init}$ visual	**19.3**	21.5	28.0	28.2	24.4	24.4	22.5	24.5	26.5	37.9	46.2	**30.6**	32.1	34.6	21.3	**31.3**	28.3
Bi-GRU$_{init}$ audio	22.1	24.7	31.0	32.4	23.8	25.6	24.8	26.3	28.2	44.5	46.0	33.0	38.0	34.9	23.8	33.9	30.8
Bi-GRU$_{init}$ multimodal	19.8	**19.5**	27.4	**24.5**	22.5	22.1	**19.8**	**22.2**	**24.5**	**35.5**	**42.0**	31.1	30.8	**32.1**	21.0	33.8	**26.8**
Metric(%)									**RE/Acc**								
Bi-GRU$_{\text{init}}$ textual	31.0	36.0	37.9	42.9	**39.9**	**41.4**	36.9	37.4	39.9	58.1	53.7	43.3	47.3	45.8	**35.0**	46.3	42.1
Bi-GRU$_{init}$ visual	**38.9**	33.0	**38.9**	40.4	32.0	36.0	35.0	37.9	37.4	58.1	51.7	45.3	46.8	48.3	33.0	33.5	40.4
Bi-GRU$_{init}$ audio	30.5	27.6	35.0	36.5	31.0	36.9	30.0	32.5	36.9	56.7	53.7	46.8	42.9	47.8	33.0	44.3	38.9
Bi-GRU$_{init}$ multimodal	36.5	**36.9**	38.4	**43.3**	34.5	40.9	**39.9**	**40.4**	**41.4**	**62.6**	**54.2**	**46.8**	**47.3**	**50.7**	34.5	**47.8**	**43.5**
Metric(%)									**F1**								
Bi-GRU$_{\text{init}}$ textual	28.7	31.3	29.7	37.4	**36.3**	**38.6**	33.8	33.8	**36.4**	52.4	**46.6**	35.9	**43.9**	38.8	**31.8**	**41.3**	37.3
Bi-GRU$_{init}$ visual	**34.5**	29.8	30.1	33.4	24.4	28.0	29.8	30.1	32.1	53.0	40.2	40.5	40.2	41.3	28.0	28.4	34.0
Bi-GRU$_{init}$ audio	20.2	19.1	23.2	22.2	22.4	27.1	20.3	25.1	28.4	48.4	41.3	38.3	32.4	40.3	22.8	38.5	29.4
Bi-GRU$_{init}$ multimodal	33.7	**34.8**	**30.8**	**39.7**	28.2	35.0	**36.5**	**36.2**	35.9	**58.1**	46.0	**40.7**	41.2	**44.2**	29.8	40.6	**38.2**
Metric									**MAE**								
Bi-GRU$_{init}$ textual	1.00	0.96	0.94	0.86	**0.92**	0.95	0.97	0.97	0.87	0.52	0.60	0.70	0.79	0.69	1.04	0.81	0.85
Bi-GRU$_{init}$ visual	0.95	1.06	0.92	0.86	1.04	0.98	0.97	0.95	0.87	0.54	0.59	0.69	0.76	0.69	1.11	0.79	0.86
Bi-GRU$_{\text{init}}$ audio	1.01	1.10	0.91	0.86	1.03	1.00	1.04	0.95	0.84	0.53	0.57	0.66	0.74	0.68	1.08	0.78	0.86
Bi-GRU$_{init}$ multimodal	**0.94**	**0.94**	**0.87**	**0.82**	0.94	**0.90**	**0.95**	**0.86**	**0.77**	**0.52**	**0.57**	**0.66**	**0.66**	**0.65**	**1.01**	**0.77**	**0.80**
Metric									**Corr**								
Bi-GRU$_{\text{init}}$ textual	**0.42**	0.45	0.12	0.23	**0.40**	0.37	0.33	0.31	0.28	0.21	0.17	0.14	0.29	0.19	0.31	0.32	0.28
Bi-GRU$_{init}$ visual	0.35	0.31	0.18	0.20	0.18	0.21	0.27	0.31	0.23	0.09	0.10	0.26	0.15	0.22	0.16	0.23	0.22
Bi-GRU$_{init}$ audio	0.31	0.22	0.15	0.14	0.15	0.11	0.17	0.27	0.35	0.22	0.26	0.29	0.15	0.28	0.18	0.23	0.22
Bi-GRU$_{init}$ multimodal	0.41	**0.48**	**0.20**	**0.35**	0.31	**0.43**	**0.38**	**0.47**	**0.42**	**0.27**	**0.26**	**0.29**	**0.31**	**0.37**	**0.32**	**0.34**	**0.35**

Table 17 Comparison of multimodal emotion recognition methods under the CMU-MOSI dataset

Multimodal fusion method	Binary(%)		5-class(%)	Regression	
	Acc	F1	Acc	MAE	Corr
C-MKL [29, 42]	73.1	75.2	35.3	–	–
SAL-CNN [39, 42]	73.0	–	–	–	–
DF [24, 42]	72.3	72.1	–	1.14	0.52
BC-LSTM [42]	73.9	73.9	–	1.08	0.58
MFN [43]	77.4	77.3	–	0.97	0.63
SVM-MD [42]	71.6	72.3	32.0	1.10	0.53
RF-MD [42]	71.4	72.1	31.9	1.11	0.51
TFN [42]	77.1	77.9	42.0	0.87	0.70
Bi-GRU$_{init}$ multimodal	**78.7**	**78.6**	**45.8**	**0.84**	**0.71**

at each moment in multiple modalities is calculated in real-time. So that the network model can learn multimodal contextual information in real-time, thereby improving the accuracy of video emotion recognition under the multimodal fusion.

The main work that can be further carried out is summarized in the following four aspects:

(1) Increase the high-dimensional features in the audio modality subnetwork. The audio modality subnetwork mentioned in this paper contains COVAREP acoustic features. Some other effective acoustic features, such as the acoustic features extracted by open speech and music interpretation by large-space extraction (OpenSMILE) may be taken into consideration. OpenSMILE combines functions such as music information retrieval and voice processing to automatically analyze audio signals in real-time, and automatically extract emotional features from speech and music signals. Adding other effective acoustic features can further improve the accuracy of emotion recognition in the audio modality subnetwork and then the fused multimodal network.

(2) Research on contextual learning emotion recognition method based on the stacked Bi-GRUs. The time-contextual learning method based on the Bi-GRU used in this paper overcomes the problem that the BRNN cannot deal with long-term dependency well and causes gradient vanishing or gradient exploding in the back propagation. Next, we can consider stacking Bi-GRUs and apply them to video emotion recognition. The stacked Bi-GRUs can be defined as a model consisting of multiple Bi-GRU layers, which makes the network model deeper. Thus, we will extract features directly from the network without any manual work. It can make better use of the input data and more complex and

Table 18 Comparison of multimodal emotion recognition methods under the POM dataset

Multimodal fusion method	Multi-label classification (%)	Regression	
	Acc	MAE	Corr
SVM [20, 26]	33.9	0.89	0.10
DF [20, 24]	34.1	0.87	0.14
BC-LSTM [20]	34.8	0.84	0.28
MV-LSTM [20, 31]	34.6	0.89	0.27
TFN [42]	31.6	0.87	0.09
MFN [43]	41.7	0.81	0.35
Bi-GRU$_{init}$ multimodal	**43.5**	**0.80**	**0.35**

comprehensive features can be learned to further improve the accuracy of video emotion recognition in the single-modality subnetwork and then the fused multimodal network.

(3) Research on video multimodal emotion recognition based on hierarchical attention network. We will consider applying the idea of hierarchical attention networks to video multimodal emotion recognition. Intra-modality attention network can extract important information in the single modality. Inter-modality attention network can capture significant information globally. Thus, the accuracy of video multimodal emotion recognition can be further improved.

(4) Research on video multimodal emotion recognition based on other fusion methods. The video multimodal emotion recognition in this paper is based on the attention fusion network, which can calculate the attention distribution of three single-modality subnetworks and calculate the attention distribution of each moment in multiple modalities in real-time. Next, we will consider other fusion methods or architectures to further improve the accuracy of video multimodal emotion recognition.

Acknowledgements This work was supported by the Zhejiang Provincial Natural Science Foundation of China [grant number LY19F020032], and National Natural Science Foundation of China [grant number U1909203, 61872322].

Funding This study was funded by the Zhejiang Provincial Natural Science Foundation of China (grant number LY19F020032), and National Natural Science Foundation of China (grant number U1909203, 61872322).

Availability of data and material Data and material are fully available without restriction.

Compliance with ethical standards

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

Code availability Custom code is available without restriction.

References

1. Bairaju SPR, Ari S, Garimella RM (2019) Emotion detection using visual information with deep auto-encoders[C]//2019 IEEE 5th international conference for convergence in technology (I2CT). IEEE:1–5
2. Byeon YH, Kwak KC (2014) Facial expression recognition using 3D convolutional neural network[J]. Int J Adv Comput Sci Appl 5(12):107–112
3. Degottex G, Kane J, Drugman T et al (2014) COVAREP—A collaborative voice analysis repository for speech technologies[C]// IEEE international conference on acoustics, speech and signal processing. IEEE, Florence, pp 960–964
4. Drugman T, Alwan A (2011) Joint robust voicing detection and pitch estimation based on residual harmonics[C]// Twelfth Annual Conference of the International Speech Communication Association : 1973–1976.
5. Drugman T, Thomas M, Gudnason J, Naylor P, Dutoit T (2012) Detection of glottal closure instants from speech signals: a quantitative review[J]. IEEE Trans Audio Speech Lang Process 20(3):994–1006
6. Ebrahimi Kahou S, Michalski V, Konda K, et al (2015) Recurrent neural networks for emotion recognition in video[C]// Proceedings of the 2015 ACM on international conference on multimodal interaction, ACM, Seattle, Washington, USA, Nov 09-13: 467–474.
7. Ekman P (1992) An argument for basic emotions[J]. Cognit Emot 6(3–4):169–200
8. Ekman P, Freisen WV, Ancoli S (1980) Facial signs of emotional experience[J]. J Pers Soc Psychol 39(6): 1125–1134

9. Fujisaki H, Ljungqvist M (1986) Proposal and evaluation of models for the glottal source waveform[C]// ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, Tokyo, Japan, 11: 1605-1608.
10. Ghosh S, Laksana E, Morency L P, et al. (2016) Representation Learning for Speech Emotion Recognition[C]// Interspeech : 3603–3607.
11. Hatzivassiloglou V, McKeown K R (1997) Predicting the semantic orientation of adjectives[C]// proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics. Assoc Comput Linguist Madrid, Spain, July 07 : 174–181.
12. Iyyer M, Manjunatha V, Boyd-Graber J, et al (2015) Deep unordered composition rivals syntactic methods for text classification[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 1: 1681-1691.
13. Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188.
14. Kane J, Gobl C (2013) Wavelet maxima dispersion for breathy to tense voice discrimination[J]. IEEE Trans Audio Speech Lang Process 21(6):1170–1179
15. Kingma D P, Ba J (2014) Adam: a method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980.
16. Kumawat S, Verma M, Raman S (2019) LBVCNN: local binary volume convolutional neural network for facial expression recognition from image sequences[C]//proceedings of the IEEE conference on computer vision and pattern recognition workshops.
17. Lee J, Tashev I (2015) High-level feature representation using recurrent neural network for speech emotion recognition[C]// Sixteenth Annual Conference of the International Speech Communication Association.
18. Li J, Ren F (2013) A hybrid approach for word emotion recognition[J]. IEEJ Trans Electr Electron Eng 8(6):616–626
19. Lim W, Jang D, Lee T (2016) Speech emotion recognition using convolutional and recurrent neural networks[C]// Asia-Pacific signal and information processing association annual summit and conference (APSIPA). IEEE, Jeju, pp 1–4
20. Liu Z, Shen Y, Lakshminarasimhan V B, et al (2018) Efficient low-rank multimodal fusion with modality-specific factors[C]// proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: long papers).
21. Ma L, Ju F, Wan J, Shen X (2019) Emotional computing based on cross-modal fusion and edge network data incentive[J]. Pers Ubiquit Comput 23(3–4):363–372
22. Moilanen K, Pulman S (2007) Sentiment composition[C]// Proceedings of RANLP , 7: 378–382.
23. Morency LP, Mihalcea R, Doshi P (2011) Towards multimodal sentiment analysis: harvesting opinions from the web[C]// proceedings of the 13th international conference on multimodal interfaces, ICMI 2011, Alicante, Spain, Nov 14-18, 2011. ACM:169–176
24. Nojavanasghari B, Gopinath D, Koushik J, et al (2016) Deep multimodal fusion for persuasiveness prediction[C]// international conference on multimodal interfaces (ICMI). ACM.
25. Orjesek R, Jarina R, Chmulik M, et al (2019) DNN based music emotion recognition from raw audio signal[C]// 2019 29th international conference RADIOELEKTRONIKA (RADIOELEKTRONIKA). IEEE: 1–4.
26. Park S, Shim HS, Chatterjee M et al (2014) Computational analysis of persuasiveness in social multimedia: a novel dataset and multimodal prediction approach[C]// proceedings of the 16th international conference on multimodal interaction. ACM, Istanbul, Turkey, Nov 12-16:50–57
27. Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation[C]// Proceedings of the 2014 conference on empirical methods in natural language processing : 1532–1543.
28. Polanyi L, Zaenen A (2006) Contextual valence shifters[M]// computing attitude and affect in text: theory and applications. Springer, Dordrecht, pp 1–10
29. Poria S, Cambria E, Gelbukh A (2015) Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]// Proceedings of the 2015 conference on empirical methods in natural language processing : 2539–2544.
30. Poria S, Cambria E, Howard N, et al (2015) Fusing audio, visual and textual clues for sentiment analysis from multimodal content[J]. Neurocomputing: S0925231215011297.
31. Rajagopalan S S, Morency L P (2016) Tadas Baltrusaitis, et al. Extending Long Short-Term Memory for Multi-View Structured Learning[M]// Computer Vision – ECCV 2016. Springer International Publishing.
32. Seyeditabari A, Tabari N, Gholizadeh S, et al (2019) Emotion Detection in Text: Focusing on Latent Representation[J]. arXiv preprint arXiv:1907.09369.
33. Shrivastava K, Kumar S, Jain DK et al (2019) An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network[J]. Multimed Tools Appl 78(20):29607–29639
34. Socher R, Perelygin A, Wu J, et al (2013) Recursive deep models for semantic compositionality over a sentiment treebank[C]// proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, WA, USA, Oct 18-21: 1631-1642.

35. Socher, R, et al (2013) Recursive deep models for semantic compositionality over a sentiment treebank[C]// Proceedings of the 2013 conference on empirical methods in natural language processing.
36. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis[J]. Comput Linguist 37(2):267–307
37. Takamura H, Inui T, Okumura M (2006) Latent variable models for semantic orientations of phrases[C]// 11th conference of the European chapter of the association for. Comput Linguist:201–208
38. Trigeorgis G, Ringeval F, Brueckner R et al (2016) Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network[C]// IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Shanghai, pp 5200–5204
39. Wang H, Meghawat A, Morency L P, et al (2017) Select-additive learning: improving generalization in multimodal sentiment analysis[C]//2017 IEEE international conference on multimedia and expo (ICME). IEEE : 949–954.
40. Wu X, et al (2019) Speech Emotion Recognition Using Capsule Networks[C]// ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.
41. Yang B, Cardie C (2012) Extracting opinion expressions with semi-Markov conditional random fields[C]// proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Assoc Comput Linguist, Jeju Island, Korea, July 12-14 : 1335–1345.
42. Zadeh A, Chen M, Poria S, et al (2017) Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv:1707.07250.
43. Zadeh A, Liang P P, Mazumder N, et al (2018) Memory fusion network for multi-view sequential learning[C]//thirty-second AAAI conference on artificial intelligence.
44. Zadeh A, Liang P, Poria S, et al (2018) Multi-attention recurrent network for human communication comprehension[C]//thirty-second AAAI conference on artificial intelligence.
45. Zadeh A, Zellers R, Pincus E, et al (2016) MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[J]. arXiv preprint arXiv:1606.06259.
46. Zhao J, Chen S, Wang S, et al (2018) Emotion recognition using multimodal features[C]//2018 first Asian conference on affective computing and intelligent interaction (ACII Asia). IEEE: 1–6.

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

Affiliations

Ruo-Hong Huan [1] • Jia Shu [1] • Sheng-Lin Bao [1] • Rong-Hua Liang [1] • Peng Chen [1] • Kai-Kai Chi [1]

Jia Shu
Shujia1218@qq.com

Sheng-Lin Bao
baoshenglin1994@gmail.com

Rong-Hua Liang
rhliang@zjut.edu.cn

Peng Chen
chenpeng@zjut.edu.cn

Kai-Kai Chi
kkchi@zjut.edu.cn

[1] College of Computer Science and Technology, Zhejiang University of Technology, Zhejiang, Hangzhou, China