Multi-Modal Emotion Recognition From Speech and Facial Expression Based on Deep Learning

Linqin Cai
School of Automation
Chongqing University of Posts and Telecommunications
Chongqing, China
iamlqcai@163.com

Jiangong Dong
School of Automation
Chongqing University of Posts and Telecommunications
Chongqing, China
2621532463@qq.com

Min Wei
School of Automation
Chongqing University of Posts and Telecommunications
Chongqing, China
waoumin@163.com

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327178

***Abstract*—The rapid development of emotion recognition contributes to the realization of highly harmonious human-computer interaction experience. Taking into account the complementarity of the emotional information of speech and facial expressions, and breaking through the single modal emotion recognition limitation of single emotional features, this paper proposed a method that combines speech and facial expression features. We CNN and LSTM to learn speech emotion features. Simultaneously, multiple small-scale kernel convolution block was designed to extract facial expression features. Finally, we used DNNs to fuse speech and facial expression features. The multimodal emotion recognition model was tested on the IEMOCAP dataset. Compared with the single modal of speech and facial expression, the overall recognition accuracy of our proposed model has been increased by 10.05% and 11.27%, respectively.**

Keywords—human-computer interaction, speech emotion recognition, multimodal emotion recognition

I. INTRODUCTION

Emotion plays a very critical role in everyday life. By expressing emotions, people can communicate and understand each other more easily. Emotion recognition can enable machines to understand human emotions and has extremely important application prospect. For example, in human-computer interaction (HCI), emotion enables the robot to make corresponding feedback according to the user's emotional state so as to improve the quality of human-computer interaction. In the medical industry, through the emotional recognition of the daily life of patients with mental illness, doctors can diagnose and treat the illness more effectively. With the fast development of communication technology, the rapid popularization of smartphones and social media, Internet users upload a large amount of video data to express their views. And different views can lead to unequal emotional expression. Through the emotional recognition of users' multimedia videos, we can understand netizens' attitude towards network events and products more accurately and comprehensively.

In recent years, emotion recognition has made a great process. But most scholars are devoted to unimodal emotion recognition. In particular, speech and facial expressions in most cases share the same thematic and temporal characteristics and often occur in human emotional interactions and are increasingly discussed within the emotion computing research community. However, external factors such as occlusion and lighting may affect the accuracy of facial emotion recognition. The differences between the voices of different subjects and environmental noise will also affect the final result of speech emotion recognition. By introducing more emotional features and making full use of complementary emotional information, the recognition result is more accurate, and the recognition performance has obvious advantages over single modal [3][5].

In addition, one of the most important concepts in the convolutional neural network (CNN) is receptive field. Each convolution kernel corresponds to a receptive field. The larger the convolution kernel is, the larger the value of the neuron's receptive field will be, which means more globe and semantic level features can be extracted. However, this will also lead to increased network parameters and computational complexity. Therefore, we came up with the method using multiple small convolution kernels instead of large convolution kernels to reduce the number of parameters while ensuring the same receptive field size in the paper.

For speech emotion feature extraction, common methods can only model the limited context information, but cannot make full use of the features that human emotion usually changes slowly and depends on the semantic contexts. In contrast, deep learning algorithm can calculate the hidden internal relationship between data and extract intrinsic features from large-scale training data, thus making the result of emotion classification more accurate. LSTM and CNN are very efficient network models in deep learning, which have been used in unimodal emotion recognition to learn advanced features from original data, and their performance is obviously better than other common feature extraction methods.

A multi-modal emotion recognition method based on IEMOCAP [1] dataset for speech and facial expression features was proposed in this paper. We combine CNN and LSTM to extract both global and contextual temporal acoustic features. Besides, we design a multiple small-scale kernel convolution block to extract facial expression features. Compared with unimodal emotion recognition, this model can obtain more accurate recognition results.

II. PROPOSED APPROACH

This paper proposed the multimodal emotion recognition system whose fusion benefits from the complementary information of audiovisual features. The summary of our system is shown in Fig.1.

Firstly, we preprocess the speech and facial expression data. Secondly, we use deep neural networks (DNN) to extract high-level features. 1D convolutional neural network (1D CNN) and bi-directional long short-term memory (Bi-LSTM) are used to extract global and contextual temporal acoustic features. And multiple small-scale kernel convolution blocks are used to extract facial expression high-level features. Finally, we use DNNs to fuse acoustic emotional features and facial expression features. Joint representation features are classified by the SoftMax function.

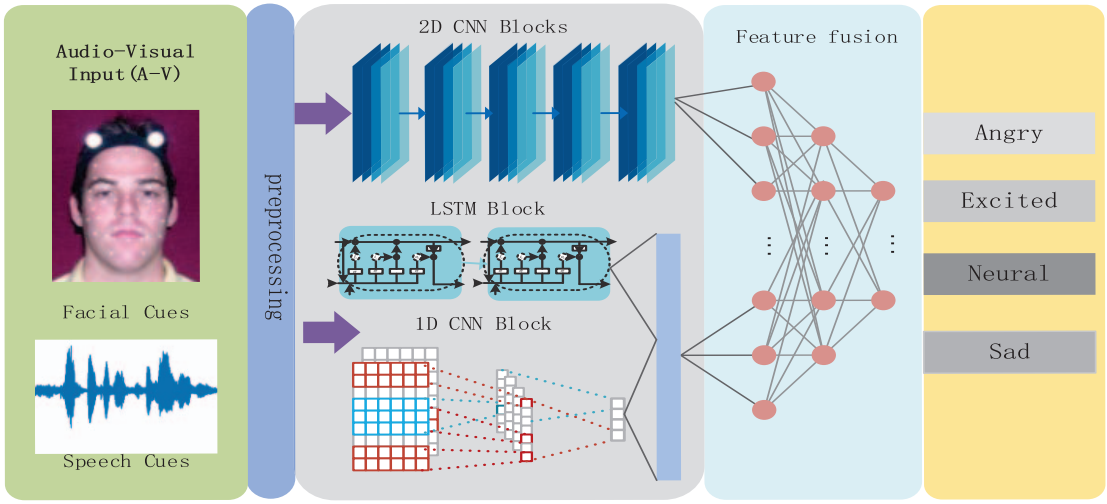

Fig. 1 The structure of multimodal modal

A. *Data Preprocessing*

The key issue of speech feature extraction is to select the most judgmental speech feature from the original speech signal.

we use pyAudioAnalysis [2] open-source python library to extract low-level audio features. The 34-dimensional features are then calculated using the feature set shown in Table 1. Each utterance uses clipping and padding to make a fixed length of 100 frames, so the input is a matrix of 34x100.

TABLE I. EXTRACTED LOW-LEVEL AUDIO FEATURES

Feature Group	Number
Zero Crossing Rate	1
Energy	1
Entropy of Energy	1
Spectral Centroid	1
Spectral Spread	1
Spectral Energy	1
Spectral Flux	1
Spectral Roll off	1
MFCCs	13
Chroma Vector	12
Chroma Deviation	1

For the facial expression data, we use 46 three-dimensional Mocap trajectories. The 46 total three-dimension facial markers allow for 165-dimensional facial landmark representation.

B. *High-level Feature Learning of Speech and facial expression*

Convolutional neural network (CNN) has strong adaptability and is good at mining local features of data, extracting global training features and classification. The basic components of a CNN include a convolutional layer, a pooling layer (also called a sampling layer) and a fully connected layer. The convolutional layer performs local perception and realizes parameter sharing, the pooling layer is used for feature dimensionality reduction, compressing the number of parameters, and the fully connected layer integrates local feature information. CNN network can reduce the variation of input frequency and capture local information, but does not need to consider global features and context.

Long short-term memory (LSTM) is an excellent Recurrent Neural Network (RNN). It inherits the characteristics of most RNN models, and solves the vanishing gradient problem caused by the gradual reduction of gradient back propagation process. LSTM is very suitable for dealing with the problems highly related to time series. Current research on NLP shows that such networks are often used to extract speech contextual information.

We use a CNN block for audio high-level feature extraction, the CNN blocks were composed of four 1D CNN layers, a MaxPooling1D layer and a Dropout layer that can prevent model overfitting. The output is connected to a Dense layer. We set the size of the convolution kernel of CNN to 3, the window of the Maxpooling1D layer to 2, and use ReLU as the activation function. At the same time, we use two Bi-LSTM layers, two Dropout layers, and a Dense layer to extract the contextual features. Finally, a Concatenate layer is used to fuse high-level features and contextual features, the output as speech emotion features. Combined CNN and LSTM, we can learn the deep features including the local information and the long-term contextual dependencies.

We borrowed the VGG16 model idea and used small-scale kernel convolution instead of large-scale kernel convolution, and designed a multiple small-scale kernel convolution model to extract facial expression features. The model consists of five CNN blocks, each CNN block includes two 2D CNNs with a convolution kernel of 3, a Batch Normalization (BN) layer, and a Dropout layer. We use two 2D CNNs with a convolution kernel of 3 replace a 2D CNN with a convolution kernel of 5.

C. *Feature Fusion*

The purpose of feature fusion is to combine the advantages of different modalities, so as to achieve the purpose of complementary disadvantages. There are four main current integration strategies.

Feature level fusion [3] is also called early fusion (EF), which is the most common and direct method, in which multiple independent modalities are fused into a single feature vector, and then input into machine learning classifier. But, feature-level fusion cannot model complex relationships.

Decision level fusion, also known as late fusion (LF), aims to combine several unimodal emotion recognition results by algebraic combination rules. However, decision-level fusion does not capture the interrelationships between

the different modes because they are considered to be independent.

Score-level fusion is a special form of decision-level fusion, has been implemented by combining the individual classification scores. These classification scores indicate the likelihood that the sample will fall into different categories.

Model-level fusion has the advantages of feature-level fusion and decision-level fusion in order to obtain multi-modal joint feature representation. Its implementation depends largely on the fusion model used. In this paper, we use deep neural networks (DNNs) as the fusion model.

III. Experiments and Results

A. Dataset

IEMOCAP is an emotion database recorded by the University of Southern California. It contains about 12 hours of audiovisual data, namely video, audio and voice text, facial expressions, which is 10 actors (5 females and 5 male actors) in the lines or impromptu scenes, leading to emotional expression. After that, each dialogue is manually divided into single sentences, and each sentence is labeled by at least three taggers.

In order to balance the data of different emotion categories, happy and excite were combined into excite category. Finally, four kinds of emotion recognition databases are composed of excite, angry, sad and neutral.

B. Evaluation Matrices

In machine learning, confusion matrix is an error matrix, which is often used to visually evaluate the performance of supervised learning algorithms. This paper uses confusion matrix as the evaluation index of the algorithm model.

C. Experiment Settings and Results

For speech emotion recognition, we use two models in the experiment. The first model (model SER_1) contains two LSTM layers with 256 units and a recurrent dropout of 0.2. Model SER_2 adds four 1D CNNs with a kernel of 3 to extract speech timing information based on SER_1. The first model of facial emotion recognition (model FER_1) consists of five convolutional blocks, each convolutional block includes a 2D CNN with a convolution kernel of 5, a BN layer and a dropout layer. In model FER_2, we use two 2D CNNs with a convolution kernel of 3 to replace the 2D CNN with a convolution kernel of 5 in model FER_1.

For the multi-modal, model MER_1 uses model SER_1 to extract speech emotion features and FER_2 to extract facial emotion features, their outputs are fused by DNNs and use SoftMax to classify. Model MER_2 uses SER_2 to extract speech emotion features and model FER_2 to extract facial emotion features and combines the output of speech and facial expression model the same as MER_1.

We divide the IEMOCAP dataset into three parts: training set, validation set and test set. Our final experimental data and comparative data are shown in Table 2. the confusion matrixes of model SER_2, FER_2, MER_2 are showing in Fig.2, Fig.3 and Fig.4. We use model SER_1, FER_1 and MER_1 as the baseline.

Compared with the baseline model, CNN-LSTM model has better recognition performance, which fully demonstrates the effectiveness of the model proposed in SER. SER_2 shows the superiority of the LSTM network. Model FER_2 uses small convolution kernel instead of large convolution kernel, which has 134752 less parameters than model FER_1 on the premise of ensuring recognition performance.

Model MER_2 uses the structures in Model SER_2 and Model FER_2 to extract speech emotion features and facial expression features， then uses DNNs to fuse emotion features, and finally uses the SoftMax function to classify to get the best recognition effect.

We also compared with others' methods on the IEMOCAP dataset. For the single model emotion recognition, our model SER_2 and FER_2 perform more accurately than [4]. [7] employed a weighted SVM method to explore the use of prototypicality information in audio-visual emotion recognition and achieve the best recognition rate of 67.59%. [6] have presented a CHfusion (context-aware hierarchical) multimodal fusion strategy, its A+V modality combination performs accuracy of 69.5% on IEMOCAP dataset.

We combined CNN and LSTM to extract speech global and context high-level features and employed multiple small-scale kernel convolution blocks to extract facial expression high-level features, then DNNs are used for multimodal fusion. Finally, we get the best accuracy of 70.24%. Compared with others' methods on the same dataset[4-7], our model shows higher recognition performance which demonstrates that the model proposed in this paper is effective.

IV. Conclusion and Future Work

In this paper, a multimodal emotion recognition model based on speech and facial expression is proposed, which uses CNN and LSTM to learn global and context high-level speech emotion features, and design multiple small-scale kernel convolution blocks to extract facial expression features.

TABLE II. The comparison of the experimental results on IEMOCAP dataset, where A stands audio, V stands visual, WA stands weighted accuracy

Model	Modality	Accuracy (%)				
		Angry	*Excite*	*Neutral*	*Sad*	*WA*
Poria et al[4]	A					57.1
SER_1		62	43	40	88	56.01
SER_2		42	53	63	75	**58.97**
Poria et al[4]	V					53.3
FER_1		52	69	60	52	58.16
FER_2		56	71	45	79	**60.19**
Poria et al[4]	A+V					67.4
Zadeh et al[5]						68.4
Poria et al[6]						69.5
Yelin Kim et al[7]		77.23	68.44	46.89	77.80	67.59
MER_1		71	70	67	67	67.21
MER_2		**75**	**66**	**58**	**88**	**70.24**

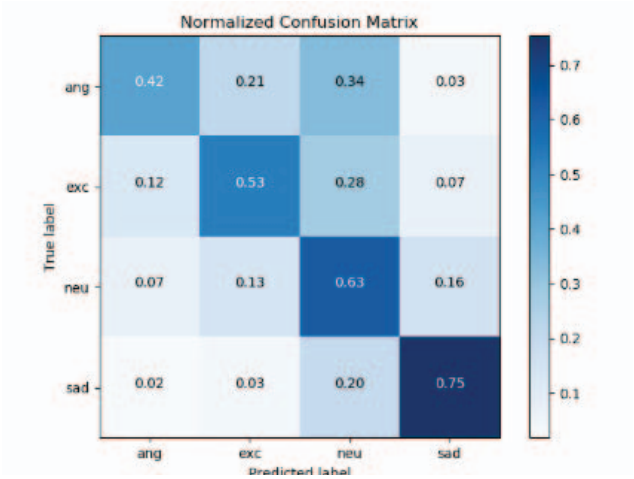

Fig. 2 the confusion matrixes of model SER_2

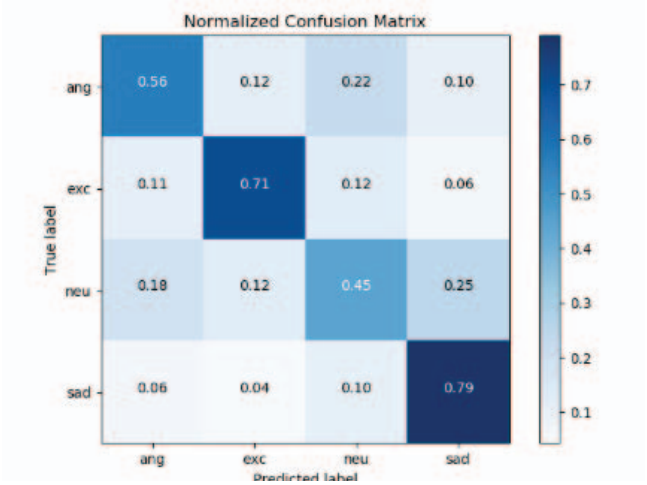

Fig. 3 the confusion matrixes of model FER_2

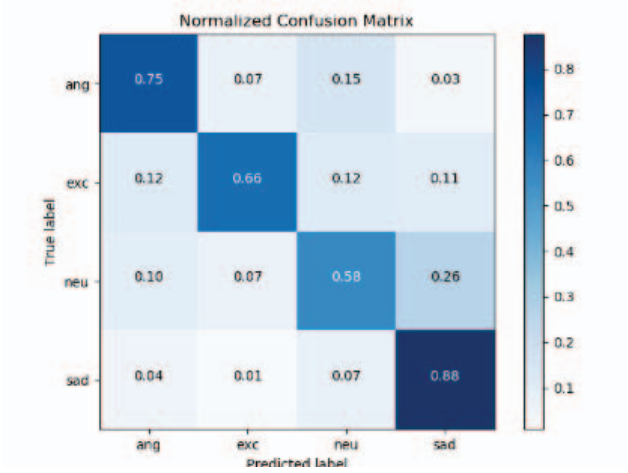

Fig. 4 the confusion matrixes of model MER_2

Finally, the DNNs are used for multimodal feature fusion. Compared with unimodal, the recognition performance has been greatly improved by fusing speech and facial expression information; we also compared with other multimodal methods, our method also has a good improvement. In addition, using multiple small-scale kernel convolution instead of large-scale kernel convolution can not only guarantee the recognition rate but can also reduce the training parameters.

In future work, we plan to introduce modalities such as text and gestures into the multimodal model, while exploring more effective feature extraction methods, and exploring more effective multimodal fusion methods. While improving the model recognition rate, we will also consider using the model in a virtual environment for human-computer interaction experiments.

REFERENCES

[1] C. Busso, M. Bulut, C. Lee, et al. “IEMOCAP: interactive emotional dyadic motion capture database,” Language Resources & Evaluation., vol.42, no. 4, 2008.

[2] G. Theodoros, P. Gianni “PyAudioAnalysis: An open-source python library for audio signal analysis,” PLOSONE., vol, 10, no, 12, December 2015.

[3] J. Cai et al., “Feature-Level and Model-Level Audiovisual Fusion for Emotion Recognition in the Wild,” 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2019, pp. 443-448.

[4] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, “Convolutional MKL based multimodal emotion recognition and sentiment analysis,” 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016, pp. 439-448.

[5] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017 pp. 1114–1125.

[6] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, “Multimodal sentiment analysis using hierarchical fusion with context modeling,” Knowledge Based Systems., vol, 161, pp.124-133, December. 2018.

[7] Y. Kim and E. M. Provost, “Leveraging inter-rater agreement for audio-visual emotion recognition,” 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, 2015, pp. 553-559.