Speech Emotion Recognition Model Based on Attention CNN Bi-GRU Fusing Visual Information

Zhangfang Hu, Lan Wang, Yuan Luo, Yanling Xia, Hang Xiao

***Abstract*—The problem of low recognition accuracy of emotion recognition models is easily caused by interference such as data redundancy and irrelevant features. In this paper, we propose a speech emotion recognition (SER) method based on an attentional convolutional neural network (CNN) bidirectional gated recurrent unit (Bi-GRU) fusing visual information. First, we pretrained the log-mel spectrograms in a ResNet-based attentional convolutional neural network (RACNN) to extract speech features. Second, the CNN-extracted facial static appearance features are fused with speech features using a deep Bi-GRU to obtain speech appearance features. A series of gated recurrent units with attention mechanisms (AGRUs) are used to extract facial geometric features. Then, the hybrid features are obtained by further combining the integrated speech appearance features with facial geometric features, and kernel linear discriminant analysis (KLDA) is used to discriminate them. Finally, the proposed method in this paper obtained accuracies of 87.92% and 89.65% on the RAVDESS and eNTERFACE'05 emotion databases, respectively. The experimental results demonstrate that the method in this paper effectively improved the accuracy and robustness of SER.**

***Index Terms*—SER, visual information, Bi-GRU, AGRUs, KLDA**

Manuscript received June 29, 2021; revised December 10, 2021. This work was supported in part by the National Natural Science Foundation Youth Fund Project (Grant No. 61703067), the Chongqing Basic Science and Frontier Technology Research Project (Grant No. Cstc2017jcyjAX0212), and the Chongqing Municipal Education Commission Science and Technology Research Project (KJ1704072).

Zhangfang Hu is a Professor of Key Laboratory of Optoelectronic Information Sensing and Transmitting Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 3565207151@qq.com).

Lan Wang is a Master Degree candidate of Key Laboratory of Optoelectronic Information Sensing and Transmitting Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (corresponding author phone: 136-576-48387; e-mail: 2268462287@qq.com).

Yuan Luo is a Professor of Key Laboratory of Optoelectronic Information Sensing and Transmitting Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 2679370283@qq.com).

Yanling Xia is a Master Degree candidate of Key Laboratory of Optoelectronic Information Sensing and Transmitting Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 1034449202@qq.com).

Hang Xiao is a Master Degree candidate of Key Laboratory of Optoelectronic Information Sensing and Transmitting Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 354612372@qq.com).

I. Introduction

EMOTION recognition (ER) is at the intersection of computational science, psychological science and cognitive science [1], [2]. Furthermore, it is an important research topic in the field of human-computer interaction (HCI) that studies the characteristics of emotional expression during interpersonal communication and designs an HCI environment with relevant feedback so that computers have the ability to recognize and understand human emotional information [3].

Voice and facial expressions are the most natural and direct ways to express emotions during interpersonal interactions [4]. In the field of speech-based ER, the 3DCNN [5], the deep-step CNN (DSCNN) [6], and feature fusion algorithms [7] have been widely used for feature extraction and learning. Despite the achievements of the relevant research work on speech-based ER technology [8], there are still problems such as a noisy speaking environment and emotion-independent factors leading to low emotion recognition accuracy. In the field of facial expression-based ER, the spatial attention CNN (SACNN), LSTM networks base on attention (ALSTMs) [9], and VGG-19 [10] have been widely used for facial emotion recognition, but interference factors such as illumination changes and facial occlusion are likely to cause face detection to fail [11], which in turn affects the discrimination of facial expressions.

With the continuous maturity of emotion recognition-related technologies, people have increasingly higher requirements for system performance such as emotion recognition accuracy, and unimodal emotion recognition has certain limitations. The use of multimodal fusion can compensate for the shortcomings of individual modalities to recognize the emotional state of the speaker more effectively. Therefore, the recognition method of multimodal fusion has gradually become a hot research topic, and a large amount of research work generally involves speech and visual information. Subhasmita et al. [12] used hidden Markov models and support vector machines to classify speech and images, respectively, which were used for emotion recognition through decision level fusion. Xu et al. [13] selected speech features using the OpenSMILE toolkit while capturing geometric features and histogram of oriented gradient (HOG) features of facial images. Cornejo et al. [14] designed a hybrid CNN to extract audio and facial features from video for concatenation, and the features were filtered by feature

selection techniques for emotion recognition. Pei et al. [15] described a model-level fusion method that uses an adaptive weighting network to incorporate auxiliary information into a multimodal emotion recognition model. Adiga et al. [16] performed comparison experiments based on different facial and speech modal features to obtain higher recognition accuracy.

In this paper, we propose an ER method that fuses speech and visual features. First, the original speech signal is preprocessed to generate log-mel spectrograms, which are used as the input of the RACNN speech network for pretraining; and the weight parameters of the network are transferred to the subsequent learning process, thereby obtaining better weight initialization results and reducing the probability of overfitting. Second, regarding the problem that face detection is prone to failure, through facial frame- and landmark-based facial feature extraction methods, the static appearance features and facial geometric features are extracted using CNN and AGRUs, respectively. Then, the rich facial features are used to improve the low recognition accuracy of speech features. The extracted speech features and static appearance features are fused by deep Bi-GRU to obtain speech appearance features. Finally, the facial geometric features and speech appearance features are further fused to obtain better hybrid features. KLDA is used to filter irrelevant features for emotion recognition after feature selection to reduce the influence of emotion irrelevant factors and improve the ER accuracy.

The remainder of this paper is structured as follows: Section II gives the proposed method of fusing speech and visual information for ER. Section III shows the experimental results of our method on relevant databases and completes the analysis of the experimental results, and section IV concludes the work of this paper.

II. Proposed Method

The SER model based on attention CNN Bi-GRU fusing visual information that we proposed, which consists of both speech and face components, is shown in Fig. 1. First, we designed and pretrained the RACNN using 3D log-mel spectrograms extracted from the speech databases (part (a) in Fig. 1). Then, we used the CNN to extract static appearance features from all frames of the faces and used the well-designed AGRUs to extract facial geometric features from key frames. Finally, the speech features extracted by the pretrained RACNN and the static appearance features extracted by the CNN from the multimodal databases were sent into the Bi-GRU fusion network to obtain the speech appearance features, and then the facial geometric features and speech appearance features were fused to generate hybrid features, which were used for emotion recognition after dimensionality reduction by KLDA (part (b) in Fig. 1).

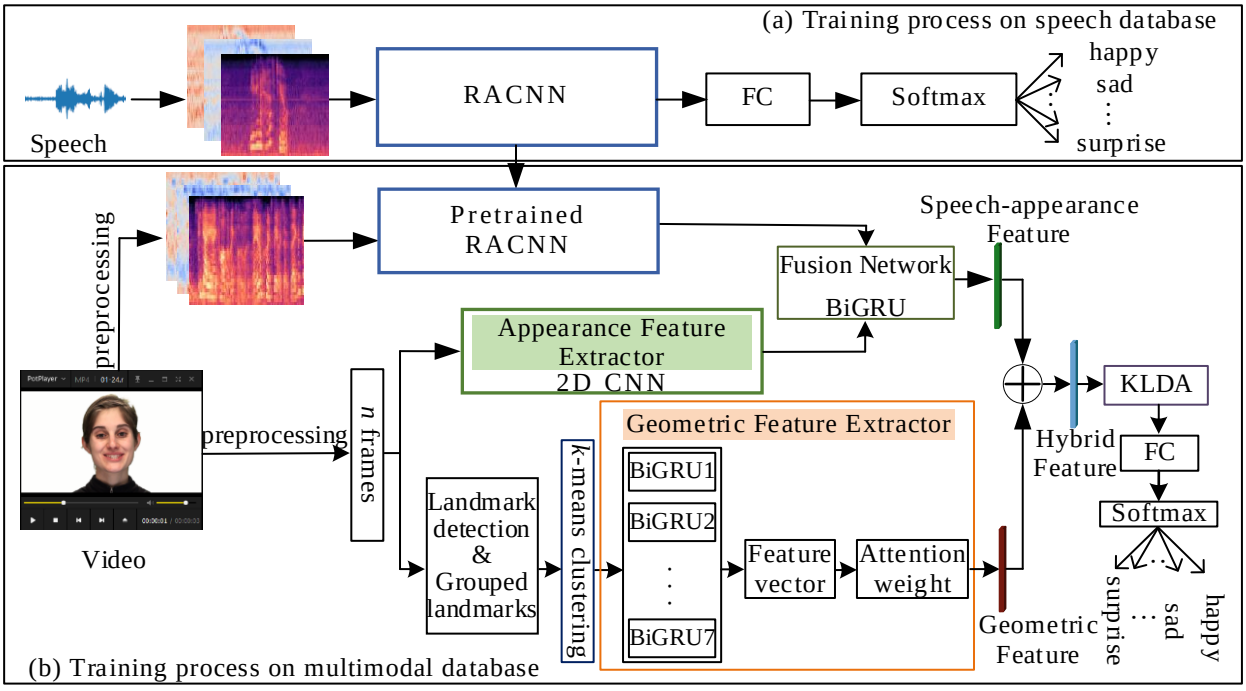

Fig. 1. Overall structure of our proposed SER model

A. Speech Network

1) Preprocessing

There are many problems in speech features, such as severe feature redundancy and many irrelevant features, so it is necessary to preprocess the original speech signals. The log-mel spectrum is a method that can reflect changes in emotion [17], so it can be used as the input of the network to effectively improve the speech emotion recognition accuracy. The process of preprocessing the speech signal to extract 3D log-mel spectrograms is shown in Fig. 2. We superimposed the log-mel spectrum and its deltas and delta-deltas to obtain a three-channel colour picture with a horizontal length related to the signal duration and vertical length related to the filter bank.

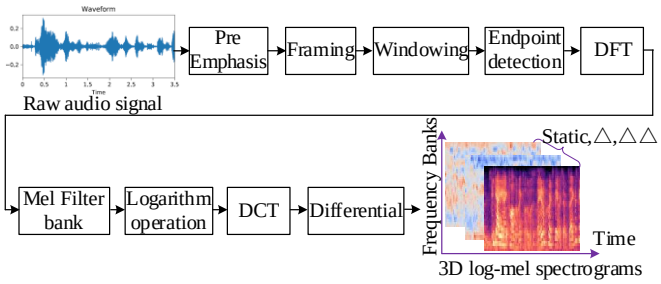

Fig. 2. Flowchart of preprocessing to obtain 3D log-mel spectrograms

2) RACNN

We designed a 3D RACNN, which consists of convolution blocks and attention modules, to extract the deep-level features of 3D log-mel spectrograms (i.e., the static, deltas and delta-deltas) from speech databases. A convolutional block is made of convolutional layers, group normalization (GN), and rectified linear units (ReLUs) for feature acquisition. In addition, drawing on the relevant experience of the convolutional block attention module (CBAM) [18], channel and spatial attention modules were designed. The two attention modules assist the RACNN in capturing the refined spatial and channel features. The Architecture of the proposed 3D RACNN is shown in Fig. 3.

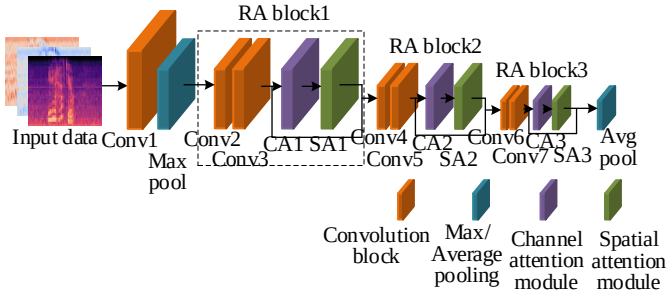

Fig. 3. Architecture of the proposed 3D RACNN

The RACNN designed in this paper is a CNN-based three-dimensional network model in which not only the convolutional attention module is introduced but also the idea of residual networks is introduced, and three residual attention blocks (RA) are designed to learn deep-level features sequentially using the jump connection technique.

We resize the log-mel spectrograms to 224 × 224 × 3 as the input to the RACNN, and the parameters of each network layer are shown in Table 1. The network uses the max pooling layer to downsample the feature maps and retains the salient features in the prominent parts. In this, each RA block has two convolutional blocks as the first step, and then the channel and spatial attention modules are concatenated sequentially. Channel and spatial attention are performed for the features obtained from the convolutional blocks. Finally, an average pooling layer is applied to characterize the global features.

TABLE I
THE NETWORK PARAMETERS OF THE PROPOSED 3D RACNN

Layers	Input-Dimension	Output-Dimension	Kernel	Stride
Conv1	9×224×224×3	7×112×112×32	3×2×2	1×2×2
Max pool	7×112×112×32	7×56×56×32	1×2×2	1×2×2
Conv2	7×56×56×32	6×56×56×32	2×1×1	1
Conv3	6×56×56×32	5×56×56×32	2×1×1	1
Conv4	5×56×56×32	5×28×28×96	1×2×2	1×2×2
Conv5	5×28×28×96	4×28×28×96	2×1×1	1
Conv6	4×28×28×96	4×28×28×288	1×2×2	1×2×2
Conv7	4×28×28×288	3×14×14×288	2×1×1	1
Avg pool	3×14×14×288	3×7×7×288	1×2×2	1×2×2

B. Visual Network

Before facial feature extraction, image preprocessing and face detection are conducted. There are two independent branches for feature extraction. One uses a CNN to extract static appearance features and uses EfficientNet to learn the spatial features in all frames. The other uses AGRUs to extract facial geometric features, which can effectively extract temporal features based on facial landmarks from key frames by learning facial morphological variations and the dynamic evolution of expressions. The face part uses CNN-GRU networks to extract local-holistic, geometric-appearance and temporal-spatial features, which enriches the expression representation of facial features.

1) Face Detection

In this paper, the face detection module uses OpenFace2.0 [19] to locate and crop faces from complex backgrounds. By using a CNN-based face detector and facial landmark detection algorithm, the module is able to handle faces in non-frontal, occluded, and low-illumination conditions, which improves the facial landmark detection accuracy, as well as the face detection accuracy. We extracted face-related frames and facial landmark points via OpenFace2.0 from the video, and the specific distribution of the 68 facial landmarks of the face is shown in Fig. 4.

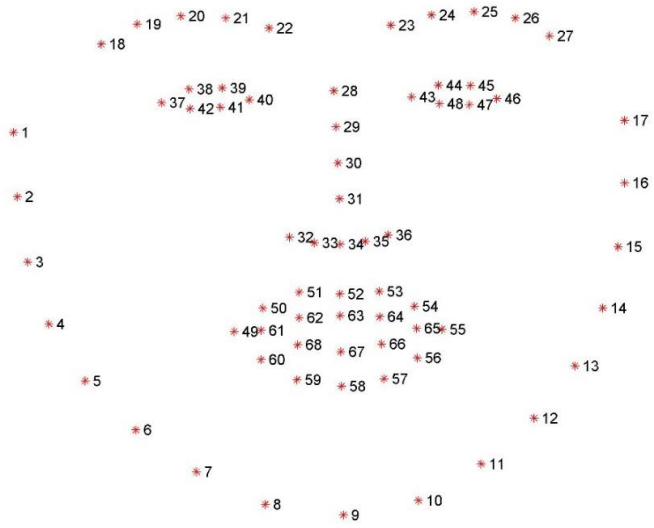

Fig. 4. 68 facial landmarks

Using a set of keyframes to represent each video enables more efficient feature learning, and we performed clustering separately for each video clip extracted via OpenFace 2.0 in terms of frames. OpenFace 2.0 provides a fast, simple and accurate methodology to preserve face videos to address possible effects in which clustering-based approaches are highly sensitive to the noise and motion of the face [20]. In this paper, we use a k-means clustering algorithm [21] based on radial basis functions (RBFs) [22], which uses RBFs instead of the Euclidean distance to model the nonlinear human visual perception section, helping to find and calculate the similarity between frames.

2) Static Appearance Features

The specific CNN structure used in this paper for static appearance feature extraction network is based on EfficientNet [23], which uses the same convolutional structure but removes the final fully connected (FC) layer. The network uses the mobile inverted bottleneck convolution (MBConv) and replaces the pooling layer in the CNN model with a global average pooling layer.

3) Facial Geometric Features

The facial geometric features used for expression analysis are based on locating the landmark points and determining the relative position relation of left and right eyes, left and right eyebrows, nose, mouth, and chin [24]. OpenFace 2.0 can directly align faces and then detect landmarks. As different facial regions contribute unequally to expression recognition, in order to learn features with better representativeness from key facial regions, we use the AGRUs network to learn local-holistic geometric features. Specifically, the AGRUs model consists of seven GRU subnetworks and an attention mechanism. The facial landmark points are divided into seven groups according to different facial positions as inputs for the corresponding seven GRU subnetworks to obtain the relative geometric position dependencies. Then, all the features learned separately from the seven facial regions are connected to obtain the holistic geometric features of the entire face. Finally, the weight vectors are learned through the attention mechanism, and the weights are adaptively readjusted to estimate the importance of different landmark regions to assist in extracting more discriminative features.

C. Fusion Model

In the process of fusing different modal features, the visual information is used as auxiliary clues to strengthen the information of the speech modality by selectively learning the relative importance of different modalities. For the joint learning of speech and static appearance features, a highly nonlinear fusion of audio-visual emotion features is performed by using the deep Bi-GRU to obtain speech appearance features. For the facial geometric features, we fuse them with the speech appearance features via weighted summation according to the weighted average algorithm to obtain the hybrid features. First, the integrated hybrid features are subjected to feature dimensionality reduction by using KLDA to filter out the most discriminative features. Then, the features are further sent to the FC layer. Finally, the emotion recognition performed by softmax layer. In this paper, by jointly combining the attention-based CNN and Bi-GRU, hybrid features are obtained to express emotions more effectively, and the structures of the proposed fusion model are shown in Fig. 5.

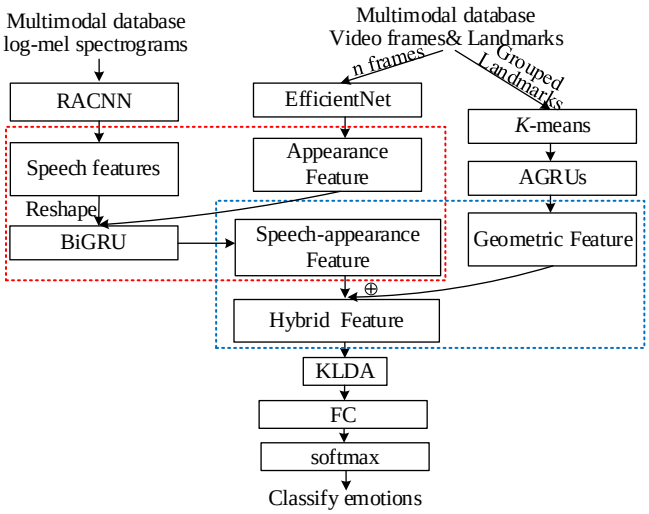

Fig. 5. Architecture of the proposed fusion model

We reshaped the feature maps obtained in the RACNN into two-dimensional data with a size of 147 × 288 as the input to a deep Bi-GRU fusion network. As a simplified and modified form of recurrent neural network (RNN), a GRU is capable of capturing long-term dependencies for sequences of arbitrary length. A GRU uses hidden states for information transfer and contains two gates: an update gate and a reset gate. The update gate determines the information in the previous hidden state that needs to be retained and memorized. The reset gate forgets part of the previous hidden state that is unimportant for the current moment and is used to calculate the current hidden state. Because contextual information should also be considered in emotion recognition, a deep bidirectional GRU, which consists of two independent hidden layers, one forward pass and another backward, and calculates the joint output built on their hidden state, is used in this paper.

III. Experimental Results And Analysis

A. Database

1) Interactive Emotional Dyadic Motion Capture database

IEMOCAP [25] is an English dyadic sessions multimodal database containing five sessions with a total of 10 actors performing improvised or scripted scenes. The database is annotated in two ways: using nine emotion categories (e.g., anger, happy, sad, etc.) for discrete annotations of speech segments and continuous annotations of emotions from three emotion attribute dimensions (i.e., valence, activation, and dominance).

2) Ryerson Audio-Visual Database of Emotional Speech and Song

RAVDESS [26] is a multimodal database of facial and voice expressions recorded by 24 professional actors. Each actor performed eight emotions, including happiness, sadness, anger, fear, surprise, disgust, calm, and neutrality. There are a total of 1440 audio files and 1012 song files, and we only used the 1440 audio files to classify the emotions.

3) eNTERFACE'05

The eNTERFACE'05 [27] audio-visual emotion database contains audio and video data recording six basic emotional states performed by 42 subjects. All data in the database were selected to obtain 1260 speech samples and expression samples each, and invalid information at the edges of the samples was removed. In purpose to ensure the convenience of face detection and tracking, a dark gray pure background was used.

B. Experimental Settings

The hardware conditions for the experiments in this paper include an Intel Xeon E5 CPU and an NVIDIA 2080Ti GPU. The system environment is Ubuntu 16.04 LTS, and the experimental tool is MATLAB 2016b. To avoid over-fitting, we chose the early stopping method in the experiments and used the cross-entropy error function as the training objective function to train the network by minimizing the cross-entropy loss; furthermore, the Adam algorithm was adopted for optimization. The learning rate of the network was 0.001, the batch size=10, and the number of epochs=100. For each experiment, we divided the data at an 8:2 ratio in which 80% was used for training and 20% for testing.

C. Experimental Results

To test the performance of the improved RACNN speech network and the speech emotion recognition model fusing visual information, experiments are conducted on the IEMOCAP, RAVDESS and eNTERFACE'05 databases, and the experimental results are studied in comparison with those in other literature.

1) RACNN Speech Network Performance Testing

In this paper, the proposed RACNN speech network is trained on the IEMOCAP and RAVDESS databases. For experiments on the IEMOCAP database, we only consider four discrete emotions (i.e., anger, happy, sad, and neutral) to remain consistent with the majority of the existing literature on speech-based emotion recognition. Referring to the model evaluation criteria used in a large number of studies, we evaluate each model using unweighted accuracy (UA) to obtain reliable results. The comparison results are shown in Table 2.

TABLE II
The Results of Comparing the Proposed ACNN with Other SER Models in the Literature Using IEMOCAP and RAVDESS Databases

Literature	Method	Database (UA%)	
		IEMOCAP	RAVDESS
Refer [28]	3D log-mels-ACRNN	64.74	-
Refer [18]	3D log-mels-ADRNN	69.32	-
Refer [29]	Log-mels-ResNet	-	73.26
Refer [30]	Log-mels-VACNN	-	74.31
Refer [31]	Spectrum-lightweight CNN	77.01	-
Refer [32]	Spectrogram-SAM	78.01	80
Proposed	**3D Log-mels-RACNN**	**79.58**	**81.76**

The above table shows that 3D log-mel spectrograms can retain valid emotional information and reduce the influence of emotion-independent factors. Furthermore, the experimental results of the RACNN speech network in this paper show that the attention-based 3D CNN model improved by drawing on the idea of residual network exhibiting a better spatial-temporal feature learning ability and effectively improves the emotion recognition accuracy.

2) Test Results and Performance Analysis of Speech Emotion Recognition Network Fusing Visual Information

To validate the performance of the network proposed in this paper, we trained it on the RAVDESS and eNTERFACE'05 databases. To avoid starting from an empty network, which leads to a time-costly learning phase, we trained the weights of EfficientNet with initial values coming from a pretrained face classification network in the ImageNet database.

Table 3 lists the accuracy of the RACNN with or without pretraining and the visual information with or without introductions for different cases. The initial values of the RACNN weights are derived from the speech network that has been trained on the speech databases in the previous section. If the RACNN network is trained from scratch on the database, the only difference is the weight initialization, i.e., using random initialization for each layer.

TABLE III
Comparison of the Accuracy Using the Proposed RACNN, RACNN + Visual, and Pretrained RACNN + Visual

Database	RACNN	RACNN + Visual	Pretrained RACNN + Visual
RAVDESS	81.76	86.8	87.92
eNTERFACE'05	-	87.53	89.65

The above table demonstrates that the pretrained speech emotion recognition model fusing visual information has better performance than the model fully trained from scratch. Pretraining provides better initialization of the weights for the speech emotion recognition network fusing visual information. This is particularly important in the field of ER, where a pretrained speech model helps to avoid the overfitting phenomena and thus improves the performance of the model.

The performance visualization results of the proposed speech emotion recognition model fusing visual information are shown in Fig. 6 and Fig. 7, which show the accuracy and the loss value for training and testing on both the RAVDESS and

eNTERFACE'05 databases.

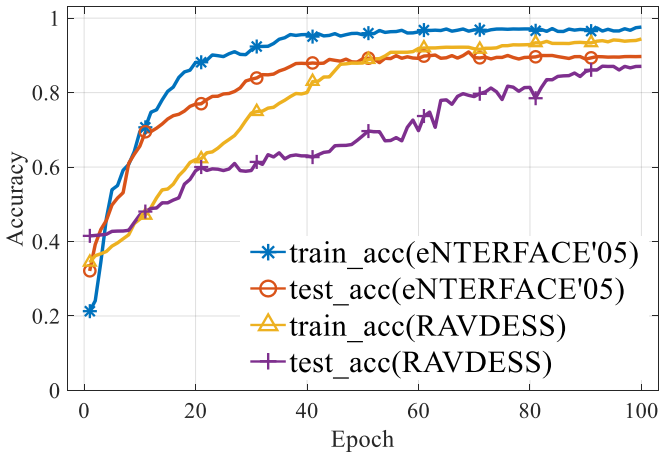

Fig. 6. Model training and testing performance

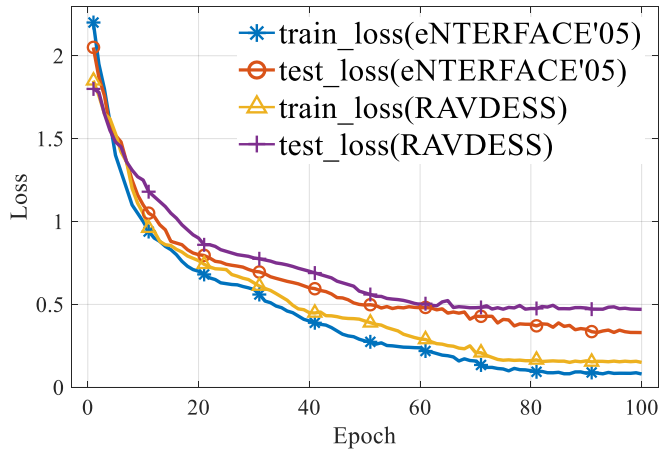

Fig. 7. Model training and testing loss

The analysis of the above figure shows that on the RAVDESS database, a recognition accuracy of nearly 0.95 with a loss value of 0.15 can be achieved on the training set, and a recognition accuracy of 0.8792 with a loss value decreasing to 0.47 is obtained on the testing set. On the eNTERFACE'05 database, the model recognition accuracy basically tends to be stable after approximately 40 epochs. The model recognition accuracy can reach nearly 0.98 and the loss value can be reduced to 0.08 on the training set, and a recognition accuracy of 0.8965 and a loss value decreasing to 0.33 was obtained on the testing set. In summary, the method in this paper can obtain high accuracy and low loss values during training and testing, which demonstrates the feasibility and effectiveness of the method.

To further study and show the accuracy between categories and the confusion with each other, we give the confusion matrix for each database. Fig. 8 and Fig. 9 show the confusion matrices for RAVDESS and eNTERFACE'05, respectively.

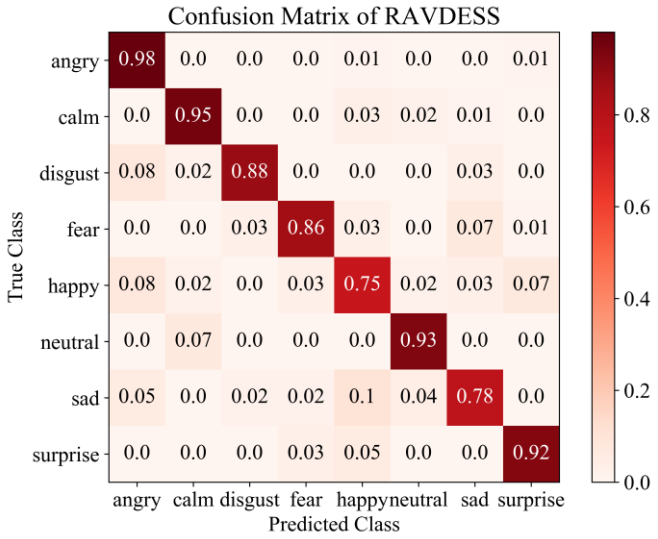

Fig. 8. The confusion matrix for the RAVDESS database

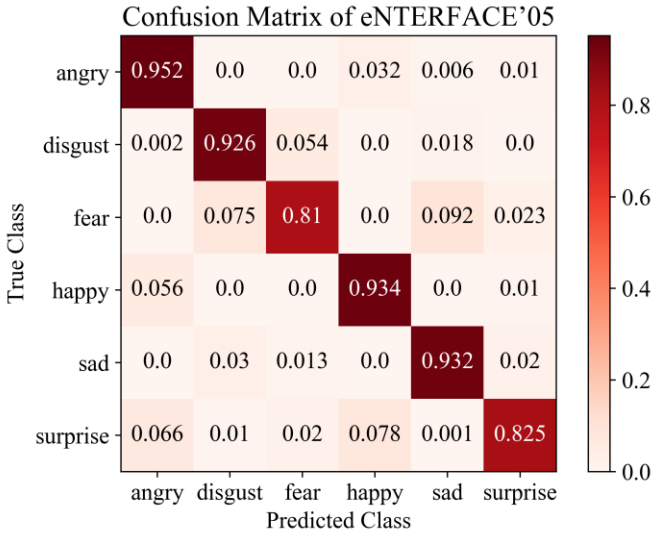

Fig. 9. The confusion matrix for the eNTERFACE'05 database

The confusion matrix shows the recognition accuracy and confusion rate for each class in the corresponding row. The horizontal axis indicates the predicted labels, and the vertical axis indicates the true labels. The confusion matrix of RAVDESS has high recognition accuracy for “anger”, “calm”, “neutral” and “surprise” and low recognition accuracy for “happy” and “sad”, below 0.80. In the confusion matrix of eNTERFACE'05, the recognition accuracy of all emotions is high. The recognition accuracies for “anger”, “disgust”, “happy” and “sad” are above 0.90; and the recognition accuracies for “fear” and “surprise” are relatively low but also above 0.80.

3) Comparison of the Network Architecture

To further demonstrate the performance of the SER network based on the attention CNN Bi-GRU fusing visual information proposed in this paper, we compared the experimental results with the results of other approaches, as displayed in Table 4.

TABLE IV
THE COMPARATIVE ACCURACY OF THE PROPOSED MODEL AND RELATED WORKS ON ENTERFACE'05 AND RAVDESS

Literature	Method	Database (Accuracy%)	
		eNTERFACE '05	RAVDESS
Refer [33]	ResNet-BiLSTM	-	77.02
Refer [32]	SAM	-	80
Refer [14]	Facial-hybrid DCNN	83.33	-
Refer [30]	VACNN-BOVW	-	83.33
Refer [34]	Facial-CNN-DBN	85.69	-
Refer [35]	Video-CNN-DBN	85.97	-
Ours	**Facial-RACNN-AGRUs**	**89.65**	**87.92**

Compared with the methods in literatures [37] and [32], the recognition accuracy of the method in this paper is 3.68 percent and 4.59 percent better on the eNTERFACE'05 and RAVDESS databases, respectively. In this paper, the attention model and residual network improve the 3DCNN speech network, and the speech network is pretrained to improve the generalization ability of cross-database SER research. Furthermore, a deep Bi-GRU fusion network is used for speech and visual feature fusion to achieve cross-modal modelling. The experimental results show that the method in this paper has excellent performance on different databases, which proves that the SER model designed in this paper has good generalization ability. In addition, it is demonstrated that the introduced visual information helps to improve the emotion recognition accuracy and makes the model have better robustness.

IV. CONCLUSION

In this paper, we proposed a SER method based on attention CNN Bi-GRU fusing visual information. By fusing speech and visual features, the proposed method can solve the problem of low emotion recognition accuracy, which is difficult to overcome using unimodal information. First, we designed and pretrained an RACNN speech network for deep-level feature extraction of 3D log-mel spectrograms. Then, static appearance features and facial geometry features are extracted by a modified CNN and AGRUs visual network. Finally, a Bi-GRU network is used to fuse the speech features with the static appearance features and then fuse them with the facial geometry features to obtain the hybrid features. Experiments tested the improved RACNN speech model on the IEMOCAP and RAVDESS databases, and the results demonstrated that the 3D log-mel spectrograms of the speech signals can effectively recognize the emotional state of the speaker and help improve the robustness of the model. Furthermore, the pretrained RACNN network also greatly reduces the overfitting problem that may be caused by the high-dimensional network. In addition, an emotion recognition network fusing speech and visual information was tested on the RAVDESS and eNTERFACE'05 audio-visual emotion databases, and it was demonstrated that the introduction of visual information could achieve better recognition results. Compared with other methods of the same type, the method proposed in this paper can effectively raise the accuracy and robustness of SER.

REFERENCES

[1] J. O'Dwyer, "Speech, Head, and Eye-based Cues for Continuous Affect Prediction," 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, UK, 2019, pp. 16-20.

[2] H. J. Zhao, Y. Ning, and W. Ruchuan, "Coarse-to-Fine Speech Emotion Recognition Based on Multi-Task Learning," Journal of Signal Processing Systems 93, 2021, pp. 299-308.

[3] A. Jaratrotkamjorn and A. Choksuriwong, "Bimodal Emotion Recognition using Deep Belief Network," 2019 23rd International Computer Science and Engineering Conference (ICSEC), Phuket, Thailand, 2019, pp. 103-109.

[4] S. Adiga, D. Vaishnavi, S. Saxena and S. Tripathi, "Multimodal Emotion Recognition for Human Robot Interaction," 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI), Stockholm, Sweden, 2020, pp. 197-203.

[5] N. Hajarolasvadi, H. Demirel, "3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms," Entropy, 2019, 21(5), 479.

[6] Mustaqeem, S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," Sensors, 2019, 20(1), 183.

[7] L. Huang, D. Jing, D. Zhou, Q. Zhang, "Speech Emotion Recognition Based on Three-Channel Feature Fusion of CNN and BiLSTM," 2020 the 4th International Conference on Innovation in Artificial Intelligence, Xiamen, China, ICIAI, no. 7, pp. 52–58, 2020.

[8] Y. Yu, Y.-J. Kim, "Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database," Electronics, 2020, 9(5), 713.

[9] C. Liu, K. Hirota, J. Ma, Z. Jia and Y. Dai, "Facial Expression Recognition Using Hybrid Features of Pixel and Geometry," in IEEE Access, vol. 9, pp. 18876-18889, 2021.

[10] N. Hajarolasvadi, E. Bashirov, and H. Demirel, "Video-based person-dependent and person-independent facial emotion recognition," Signal, Image and Video processing (SIViP), 2021.

[11] N. Zhou, R. Liang and W. Shi, "A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection," in IEEE Access, vol. 9, pp. 5573-5584, 2021.

[12] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," 2016 IEEE Students' Technology Symposium (TechSym), Kharagpur, India, 2016, pp. 7-12.

[13] F. Xu and Z. Wang, "Emotion Recognition Research Based on Integration of Facial Expression and Voice," 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 2018, pp. 1-6.

[14] J. Cornejo, H. Pedrini, "Bimodal Emotion Recognition Based on Audio and Facial Parts Using Deep Convolutional Neural Networks," 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, 2019, pp. 111-117.

[15] E. Pei, D. Jiang, H. Sahli, "An efficient model-level fusion approach for continuous affect recognition from audiovisual signals," Neurocomputing, vol. 376, pp. 42-53, 2019.

[16] S. Adiga, D. V. Vaishnavi, S. Saxena, and S. Tripathi, "Multimodal Emotion Recognition for Human Robot Interaction," 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2020.

[17] H. Meng, T. Yan, F. Yuan and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms with Deep Learning Network," in IEEE Access, vol. 7, pp. 125868-125881, 2019.

[18] S. Woo, J. Park, JY. Lee, I.S. Kweon, "CBAM: Convolutional Block Attention Module," European Conference on Computer Vision - ECCV 2018, Springer, Cham, vol. 11211, pp. 3-19.

[19] T. Baltrusaitis, A. Zadeh, Y. C. Lim and L. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 2018, pp. 59-66.

[20] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera and G. Anbarjafari, "Audio-Visual Emotion Recognition in Video Clips," in IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 60-75, 1 Jan.-March 2019.

[21] M. Capó, A. Pérez, and J. A. Lozano, "An efficient approximation to the K-means clustering for massive data," Knowledge-Based Systems, vol. 117, pp. 56-69, Feb. 2017.

[22] S. S. Chouhan, A. Kaul, U. P. Singh, and S. Jain, "Bacterial foraging optimization based radial basis function neural network (BRBFNN) for

identification and classification of plant leaf diseases: An automatic approach towards plant pathology," in IEEE Access, vol. 6, pp. 8852-8863, 2018.
[23] M. Tan, and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019.
[24] K. Dobs, L. Isik, D. Pantazis, and N. Kanwisher, "How face perception unfolds over time," Nature Communications, Mar 19, 2019, 10(1), 1258.
[25] C. Busso, M. Bulut, CC. Lee, et al. "IEMOCAP: interactive emotional dyadic motion capture database," Language Resources and Evaluation, 42.4(2008), pp. 335-359.
[26] S. R. Livingstone, F. A. Russo, and N. Joseph, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLOS ONE, 13.5(2018), pp. e0196391-e0196391.
[27] O. Martin, I. Kotsia, B. Macq and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 2006, pp. 8-8.
[28] M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, Oct. 2018.
[29] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.
[30] M. Seo, M. Kim, "Fusing Visual Attention CNN and Bag of Visual Words for Cross-Corpus Speech Emotion Recognition," Sensors, 2020, 20, 5559.
[31] T. Anvarjon, Mustaqeem, S. Kwon, "Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features," Sensors, 2020, 20, 5212.
[32] Mustaqeem, S. Kwon, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," Applied Soft Computing, vol. 102, pp. 1568-4946, 2021.
[33] Mustaqeem, M. Sajjad and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," in IEEE Access, vol. 8, pp. 79861-79875, 2020.
[34] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, A. Koši, "Audio-visual emotion fusion (AVEF): a deep efficient weighted approach," Information Fusion, 2019, vol. 46, pp. 184-192.
[35] S. Zhang, S. Zhang, T. Huang, W. Gao and Q. Tian, "Learning Affective Features with a Hybrid Deep Model for Audio–Visual Emotion Recognition," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 3030-3043, Oct. 2018.