Abstract—With the fast development of artificial intelligence and short videos, emotion recognition has become one of the most important research topics in human-computer interaction. At present, most emotion recognition methods still stay in a single modality. However, in daily life, human beings will usually disguise their real emotions, which leads to the problem that the low accuracy of single modal emotion recognition. Moreover, it is not easy to distinguish similar emotions. Therefore, we propose a new approach denoted as ICANet to achieve multimodal short video emotion recognition by employing three different modalities of audio, video, and optical flow, making up for the lack of a single modality and then improving the accuracy of emotion recognition in short videos. ICANet has a better accuracy of 80.77% on the IEMOCAP benchmark. The cross-modal fusion method of short video emotion recognition established in this paper can effectively improve the accuracy of emotion recognition in human-computer interaction scenarios.

Index Terms—Multimodal Deep Learning, Human-Computer Interaction, Emotion Recognition, Attention Mechanism.

People’s life always accompanies emotion, which is an essential part of the dynamic mechanism in people’s psychological activities. The “pure cognition” research on human emotion recognition has long been the focus of many scholars. With the continuous improvement of technology, emotion recognition methods are gradually applied to various fields, such as healthcare, psychology, etc. For example, based on physiological signals main deploy emotion recognition methods. The emotion recognition methods and non-physiological signals-based emotion recognition methods are unsuitable for practical application. Emotion recognition based on non-physiological signals mainly employs non-physiological signals such as EEG, EMG, and ECG to predict. These signals are relatively easy to collect and have the capability of getting more objective results, it is relatively difficult for us to collect these physiological signals. Furthermore, it lacks reasonable evaluation standards, which is expected to find more potential applications in the future.

Currently, the research on emotion recognition in short videos is mainly divided into physiological signals-based emotion recognition and non-physiological signals-based emotion recognition. Emotion recognition methods mainly include methods based on physiological signals and methods based on non-physiological signals. The former mainly includes EEG, EMG, ECG-based emotion recognition methods. The latter mainly refers to methods based on non-physiological signals such as facial expression, voice, and other multimodal inputs. The multimodal fusion methods will effectively improve the accuracy of emotion recognition in short video emotion recognition established in this paper can effectively improve the accuracy of emotion recognition in human-computer interaction scenarios.

ICANet: A Method of Short Video Emotion Recognition Driven by Multimodal Data

The overall structure of ICANet is divided into three parts: data preprocessing module, multimodal feature extraction module, and fusion classification module. The overall illustration of ICANet. It consists of the Data Preprocessing Module, the Multimodal Feature Extraction Module, and the Fusion Classification Module, shown in Fig. 1.

In this paper, we select the videos, optical flow and LFCC spectrograms as the inputs. The three feature extraction networks, I3D (RGB/FLOW) and CA-VGG16 are deployed to model the three different modalities, namely audio, video, and optical flow. The corresponding to the three modalities are then extracted. The specific feature extraction results are transferred to the SoftMax classifier, and the final scores are obtained. We will then finetune and simulate human behaviors. Emotion recognition techniques have many application scenarios, such as fatigue status detection and prediction of fatigue status [1]. Therefore, we propose a new approach denoted as ICANet to achieve multimodal short video emotion recognition by employing three different modalities of audio, video, and optical flow, making up for the lack of a single modality and then improving the accuracy of emotion recognition in short videos.
A. Data Preprocessing Module

1) LFCC spectrogram

In this paper, the audio signal is preprocessed by generating the LFCC spectrograms, which is a three-dimensional spectrum representing the voice frequency graph changing with time. Since the spectrogram deploys a two-dimensional plane to express the three-dimensional information, the energy value is expressed by the depth of the colors. Firstly, this paper preprocesses the Wav files of the audio section in the IEMOCAP dataset, using the scipy voice processing tool and python_speech_features library to read the speech information and extract the LFCC speech features, and finally converts the ordinary Wav speech signals into the LFCC spectrograms.

2) Optical flow extraction

Optical Flow is the apparent motion mode of two consecutive interframe images caused by the motion of an object or a camera. In this paper, we deploy the interface of the optical flow estimation algorithm provided in the OpenCV, which includes the sparse optical flow estimation algorithm cv2.calcOpticalFlowPyrlk() and dense flow estimation algorithm cv2.calcOpticalFlowFarneback(). Specifically, we deploy the sparse flow estimation algorithm, which is denoted as the Lucas Kanade algorithm [4].

B. Multimodal Feature Extraction Module

1) RGB and flow feature extraction module

In this paper, the I3D [5] is deployed to extract the specific features of the characters in the videos, and the model training method of separating the video stream and the optical flow stream is adopted. I3D (two-stream inflated 3D CNN) is a video action recognition model proposed by Google DeepMind in 2017, using Inception-V1 as the backbone network. Its specific network structure is shown in Fig. 2. In Inception-V1, the stride of the first convolution transformation layer is 2. In addition to the parallel maximum pooling layer, there are four maximum pooling layers with a stride of 2×2 and 7×7. The detailed structure of the initial space sub-module “Inc.” is shown in Fig. 3.

Fig. 2. The overall network structure of Inflated Inception-V1. Here, “Rec. Fields” represents the receptive fields for the specific feature tensors of feature extraction networks in the decision-level feature fusion module.

Fig. 3. The overall illustration of the initial space sub-module “Inc.”. The strides of convolution and pooling operators are 1 were not specified.

This paper adopts the form of separate model training. One I3D network training takes RGB (video) stream as input, and the other takes FLOW (optical flow) stream as input, carrying optimized and smooth stream information. In this paper, the two I3D networks are first trained, and the pre-trained weights are then loaded during model validation. The models are trained on the IEMOCAP multimodal dataset, and all the video frames have been adjusted to the appropriate size, number, and channels before being input into the networks. Seventy-nine video frames are extracted at uniform time intervals, and the entire video content is captured. Finally, the pre-trained weights of RGB and FLOW modalities are obtained.

2) Audio feature extraction module

This paper first converts audio files into the LFCC spectrograms, and then evolves them into an image classification task. We deploy the independently improved VGG16 denoted as CA-VGG16. We mainly introduce the Coordinate Attention Module (CA module) to improve the performance of VGG16, adding the CA Module after the first, fourth, and fifth convolution transformation blocks of VGG16 to improve the model performance. The overall network structure of CA-VGG16 is shown in Fig. 4.
In the CNNs, the full dataset and finally form an emotion recognition dataset. We filter the unbalanced data in the original recognition dataset is deployed to conduct model training and validation. We conclude the highest accuracy of 58.96%, which is a practical result.

### A. IEMOCAP Dataset Introduction

The IEMOCAP [9] multimodal emotion recognition dataset is composed of four emotion categories: happy, sad, neutral, and anger, with a total number of 5531. The distribution for four categories is shown in Table I.

### B. Evaluation Indicator

The evaluation indicator chosen in experiments is accuracy (ACC). The formula of ACC is shown in Eq. 5:

\[
ACC = \frac{TP + TN}{TP + TN + FP + FN}
\]

Where TP is the true positive sample detected as a correct; TN represents a negative sample detected as correct; FP represents the positive sample detected as a negative sample; FN represents the negative sample detected as a positive sample.

### C. Ablation Studies

We conduct four sets of ablation experiments for the models of the proposed improvement based on SE Attention and CBAM Attention. The models of these four ablation experiments are original VGG16, SE-VGG16, CA-VGG16-3, and CA-VGG16-5, respectively. The specific ablation results are shown in Fig. 5.

| Happy | Sad | Neutral | Anger | Total |
|-------|-----|---------|-------|-------|
| 143   | 245 | 1708    | 1103  | 5531  |
| 305   | 384 | 320     | 362   | 1241  |
| 286   | 278 | 327     | 240   | 1031  |
| 197   | 194 | 197     | 170   | 1151  |
| 1151  | 1085| 1151    | 1151  | 5531  |

### C. Fusion Classification Module

The ICANet proposed in this paper adopts the decision fusion classification module. From Eq. 1 to Eq. 4, where W and H represent the width and height, respectively. P, K, and S represent the filter size, convolution kernel size, and stride, respectively.

### Convolution Transformation Layer

\[
W_{n+1} = (W_n + 2 + P - K)/S + 1
\]

\[
H_{n+1} = (H_n + 2 + P - K)/S + 1
\]

### Pooling Layer

\[
W_{n+1} = (W_n - K)/S + 1
\]

\[
H_{n+1} = (H_n - K)/S + 1
\]

### Full Connection Layer

The full connection layer utilizes the SoftMax layer for classification. It deploys three full connection layers at the end, and the last one is responsible for the task of “Classifier”, which classifies the feature tensors accurately.

### Table I. The distribution for four categories

| Session1 | Session2 | Session3 | Session4 |
|----------|----------|----------|----------|
| Happy    | Sad      | Neutral  | Anger    |
| 1905     | 366      | 1698     | 2097     |
| 277      | 327      | 197      | 170      |
| 305      | 384      | 320      | 362      |
| 286      | 278      | 327      | 240      |
| 197      | 194      | 197      | 170      |
| 1151     | 1085     | 1151     | 1151     |

### Experimental Results

The experimental results show that CA-VGG16-3 performs better than CA-VGG16. However, the ablation studies also attempt to add the CA modules after each second, fourth, and fifth blocks of VGG16, denoted as CA-VGG16-3. At the same time, we also attempt to add the CA modules after each second, fourth, and fifth blocks of VGG16, denoted as CA-VGG16-5. The specific ablation results are shown in Fig. 5.
D. Comparison with other SOTA Methods

To further verify the effectiveness of the emotion recognition method ICANet proposed in this paper, we compare ICANet with the current mainstream emotion recognition methods. The results of the comparison are shown in Table II. It can be seen from the abovementioned table that the accuracy of the new emotion recognition method ICANet is significantly higher than the above-mentioned five mainstream emotion recognition methods. Further, the accuracy of ICANet in short videos driven by multimodal data can reach 80.77%, exceeding the existing CNN-based state-of-the-art methods by 15.89%.

| Method          | Modality       | ACC(%) |
|-----------------|----------------|--------|
| C3D             | RGB            | 53.37  |
| I3D             | RGB            | 53.04  |
| ON THE IEMOCAP MULTIMODAL DATASET. | FLOW | 61.33 |
| ResNet50+LSTM   | RGB            | 60.12  |
| 1D Music CNN    | RGB            | 64.88  |
| Simam           | RGB:FLOW:Audio| 77.50  |
| Ours            | RGB:FLOW:Audio| 80.77  |

Therefore, we select 4:2:4 as the optimal weight ratio when RGB:FLOW:Audio = 4:2:4, we can get the best model performance to a certain extent. The reason for the above phenomenon is that in this approach, the overall network cannot achieve better results. On the contrary, moderately increasing the weight of RGB stream will not achieve the optimal performance. According to the experimental results, the ratios of RGB, FLOW, and Audio are 1:1:1, 3:2:5, 4:3:3, respectively, the ACC are respectively 77.50%, 80.77%, 85.52%.

To explore the best recognition effect of ICANet, this paper gives different modalities different weight ratios for multimodal fusion to obtain the optimal weight ratio. When the weight ratio is considered to further improve the performance of emotion recognition performance. In future work, more modes can be considered.