applied sciences

MDPI

Article

M2ER: Multimodal Emotion Recognition Based on Multi-Party Dialogue Scenarios

Bo Zhang [1,*], Xiya Yang [1], Ge Wang [1], Ying Wang [1] and Rui Sun [2]

[1] School of Information and Communication Engineering, Communication University of China, Dingfuzhuang, Chaoyang District, Beijing 10024, China; yangxiya@cuc.edu.cn (X.Y.); wg2530280660@163.com (G.W.); yingwang@cuc.edu.cn (Y.W.)
[2] School of Computing, Newcastle University, Newcastle upon Tyne NE1 7RU, UK; r.sun5@newcastle.ac.uk
* Correspondence: zhangbo2015@cuc.edu.cn

Abstract: Researchers have recently focused on multimodal emotion recognition, but issues persist in recognizing emotions in multi-party dialogue scenarios. Most studies have only used text and audio modality, ignoring the video modality. To address this, we propose M2ER, a **m**ultimodal **e**motion **r**ecognition scheme based on **m**ulti-party dialogue scenarios. Addressing the issue of multiple faces appearing in the same frame of the video modality, M2ER introduces a method using multi-face localization for speaker recognition to eliminate the interference of non-speakers. The attention mechanism is used to fuse and classify different modalities. We conducted extensive experiments in unimodal and multimodal fusion using the multi-party dialogue dataset MELD. The results show that M2ER achieves superior emotion recognition in both text and audio modalities compared to the baseline model. The proposed method using speaker recognition in the video modality improves emotion recognition performance by 6.58% compared to the method without speaker recognition. In addition, the multimodal fusion based on the attention mechanism also outperforms the baseline fusion model.

Keywords: multimodal; emotion recognition; feature extraction; feature-level fusion; attention mechanism; speaker recognition

check for updates

Citation: Zhang, B.; Yang, X.; Wang, G.; Wang, Y.; Sun, R. M2ER: Multimodal Emotion Recognition Based on Multi-Party Dialogue Scenarios. *Appl. Sci.* **2023**, *13*, 11340. https://doi.org/10.3390/app132011340

Academic Editor: Douglas O'Shaughnessy

Received: 5 September 2023
Revised: 1 October 2023
Accepted: 11 October 2023
Published: 16 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Emotions are unique and important forms of human expression [1]. When conducting early research on emotions, Ekman [2] classified people's basic emotions according to their needs. In 1977, Picard proposed the concept of emotional computing [3], aiming to equip computers with the ability to recognize, understand, express, and adapt to human emotions. An important direction in emotional computing research is emotion recognition, which can create more intelligent and harmonious user entities for applications such as lie detection, audiovisual monitoring, online conferences, and human–computer interaction (HCI) [4].

Researchers often rely on unimodal emotion recognition [5]. Recently, significant progress has been made in the research of unimodal approaches for text, audio, and video. Particularly, facial emotion recognition (FER) technology has a wide range of applications, including HCI, emotional chat, psychological diagnosis, and other tasks [6]. AffectNet [7] is a widely recognized corpus for video modality emotion recognition. Currently, the top three models in terms of accuracy for seven-class emotion recognition on AffectNet are POSTER++ (67.49%) [8], Emotion-GCN (66.46%) [9], and EmoAffectNet (66.37%) [10]. Other studies related to FER are as follows: Bakariya et al. [11] created a real-time system that can recognize human faces, assess human emotions, and recommend music to users. Meena et al. [12] proposed a facial image sentiment analysis model based on a CNN. It is discovered that more convolution layers, a strong dropout, a large batch size, and many epochs can obtain better effects. Savchenko [13] studied lightweight convolutional neural networks (CNNs) for FER task learning and verified the effectiveness of CNNs for FER.

Meena et al. [14] utilized Inception-v3, along with additional deep features, to enhance image categorization performance. A CNN-based Inception-v3 architecture was used for emotion detection and classification. In a study by Saravanan [15], they found that CNNs are highly effective for image recognition tasks due to their ability to capture spatial features using numerous filters. They proposed a model consisting of six convolutional layers, two max-pooling layers, and two fully connected layers, which performed better than decision trees and feed-forward neural networks on the FER-2013 dataset. Li [16] used a CNN, which extracts geometric and appearance features, and LSTM, which captures temporal and contextual information on facial expressions. This CNN–LSTM architecture allows for a more comprehensive representation of facial expressions by combining spatial and temporal information. Ming et al. [17] presented a facial expression recognition method that included an attention mechanism based on a CNN and LSTM. This model was able to effectively extract information on important regions, better than general CNN–LSTM-based models. Sang [18] focused on reducing intra-class variation in facial expression depth features and introduced a dense convolutional network [19] for the FER task.

There has been an increase in the combination of transformers in various FER methods. Xue [20] was the first to use the vision transformer for FER and achieved state-of-the-art results. VTFF [21] excels in dealing with facial expression recognition tasks in the wild due to its feature fusion. Chen et al. [22] introduced CrossViT, which uses dual branches to combine image patches of different sizes to produce more reliable features. Heo et al. [23] examined the benefits of pooling layers in ViT, similar to their advantages in CNNs.

However, in real-world scenarios, the video modality often presents complex data formats. For example, multiple faces often appear in the same frame in multi-party dialogue scenarios, and the presence of non-speaking individuals' faces can interfere with the final emotion recognition. This is the reason why most of the existing research on multimodal emotion recognition in multi-party dialogues has not utilized the video modality. Challenges such as speaker recognition, significant intra-class facial expression variations, and subtle inter-class differences further highlight the room for improvement in emotion recognition. Thus, there is still considerable scope for further research and exploration in the field of emotion recognition.

It is hard to obtain accurate emotional information only through a single modality [24,25]. Compared with unimodal emotion recognition, multimodal emotion recognition can make up for the noise interference caused by the single modality and make full use of the complementary features between different modalities. Zadeh [26] conducted multimodal sentiment analysis on three modalities of text, audio, and video for the first time and released the first dataset containing text, audio, and video modalities—the YouTube dataset. Rosas [27] proposed a multimodal research dataset—Moud—and conducted sentiment analysis in sentences. Zadeh [28] constructed a large-scale multimodal dataset CMU-MOSEI. In recent years, based on the above datasets, researchers have carried out many classic multimodal emotion analysis methods based on text, audio, and video modalities. Dai [29] combined multimodal feature extraction and fusion into a model and optimized it at the same time, which improved the accuracy of emotion recognition in real-time performance. Ren [30] used the self-supervised training model to fuse the features of text, audio, and video modalities into non-standard classes and achieved better results than the baseline model.

The focus of multimodal emotion recognition lies in how to extract features and perform subsequent fusion. However, most of the current research on multimodal emotion recognition only focuses on the stage of feature fusion, neglecting the initial stage of unimodal emotional feature extraction. For example, in the case of the audio modality, most studies directly extract audio features using open-source toolkits such as Librosa and OpenSmile [31,32] and fuse them with features from other modalities. In the context of multi-party dialogues, many researchers have focused on studying the text and audio modalities while neglecting the video modality. Extracting comprehensive features from individual modalities is a prerequisite for multimodal emotion recognition. The more

comprehensive the extraction of emotional features from each modality, the better it can reflect the characteristics of emotion.

There are three main methods of multimodal fusion: data-level fusion, feature-level fusion, and decision-level fusion. The specific process of feature-level fusion is illustrated in Figure 1. Feature-level fusion can fully leverage the advantages of each modality, effectively integrate information from different modalities, and consider the correlation between various data in different modalities. However, if the feature-level fusion is achieved by directly concatenating the feature vectors, it will result in high-dimensional vectors, leading to problems such as the curse of dimensionality.

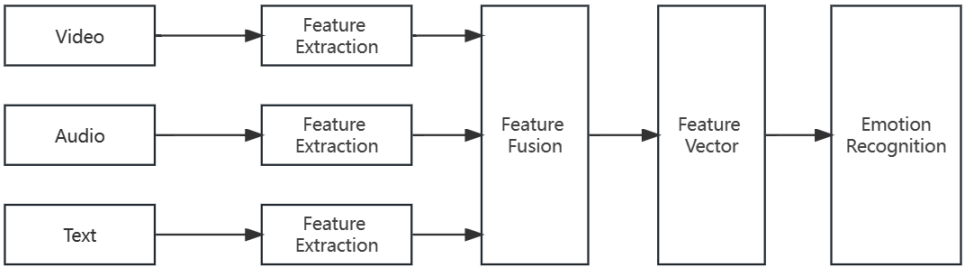


Figure 1. The specific process of feature-level fusion, which involves extracting emotional features from individual modalities, combining the obtained feature vectors in a specific way, and finally using an emotion classifier to recognize the fused features.

Recently, many research works have focused on attention-based fusion and its variants, such as self-attention, multi-head attention, and transformers [33]. The attention-based fusion integrates the advantages of early fusion and late fusion and compensates for their shortcomings [34]. The attention mechanism is a specialized structure that can be embedded in the framework of machine learning models. By employing the attention mechanism, the problem of information overload can be addressed. Furthermore, the attention mechanism can provide an effective resource allocation scheme in neural networks [35]. As the number of model parameters increases in deep neural networks, the model generally becomes more expressive and capable of storing a greater amount of information. However, the increasing number of parameters also demands significant computational resources during model training, making it challenging. By incorporating the attention mechanism into neural networks, it becomes possible to identify which data in the input sequence contributes more significantly to the task at hand [36]. Consequently, more limited attention can be allocated to the most valuable portions of information, while reducing attention or disregarding irrelevant information, thus efficiently utilizing computational resources [37]. Hu [38] proposed the Multimodal Dynamic Fusion Network (MM-DFN) to recognize emotions by fully understanding multimodal conversational context. Wang et al. [34] proposed a cross-attention asymmetric fusion module, which utilized information matrices of the acoustic and visual modality as weights to strengthen the text modality.

Based on the above situation, we propose M2ER that optimizes the key steps of multimodal emotion recognition in multi-party dialogue scenarios. We mainly focus on how to fully utilize video modalities. The contributions of M2ER are summarized as follows:

- We constructed suitable feature extraction models for text, audio, and video modalities. Addressing the challenge of multiple faces appearing in a single frame in the video modality, we propose a method using multi-face localization for speaker recognition, thus extracting features from facial expression sequences of the identified speaker.
- For the multimodal fusion model, we adopted the feature-level fusion approach utilizing a multimodal fusion model based on the attention mechanism. The extracted unimodal emotional features are combined using cross-modal attention to capture the intermodal interactions. Furthermore, the attention mechanism employed determines

the contribution of each modality to the final emotion classification, enabling the fusion with different weights.

- We conducted experiments on the Multimodal Emotion Lines Dataset (MELD) [39] using both unimodal and multimodal fusion methods and further evaluated the scalability of our models on the MEISD dataset [40]. The extensive experiments show that our unimodal feature-based emotion recognition model of M2ER outperforms the baseline models. The multimodal fusion model achieves higher recognition accuracy compared to the unimodal emotion recognition systems. Moreover, our fusion model of M2ER exhibits superior performance in multimodal emotion recognition tasks compared to directly concatenated models.

The remaining parts of the paper are structured as follows: Section 2 presents the detailed design of the proposed M2ER, including the extraction of unimodal features and the multimodal feature fusion model. In Section 3, we outline the experiments conducted on unimodal and multimodal emotion recognition separately and verify the scalability of the models. Furthermore, we discuss the advantages of our work as well as the limitations and future work in Section 4; Finally, Section 5 concludes the work of this paper.

2. Detailed Design

Figure 2 illustrates the overview framework of M2ER, which includes the extraction of emotional features from text, audio, and video modalities, as well as the multimodal fusion classification framework adopted in our work based on the attention mechanism. We will introduce the detailed scheme of the unimodal extraction model in Section 2.1 and fusion model information in Section 2.2.

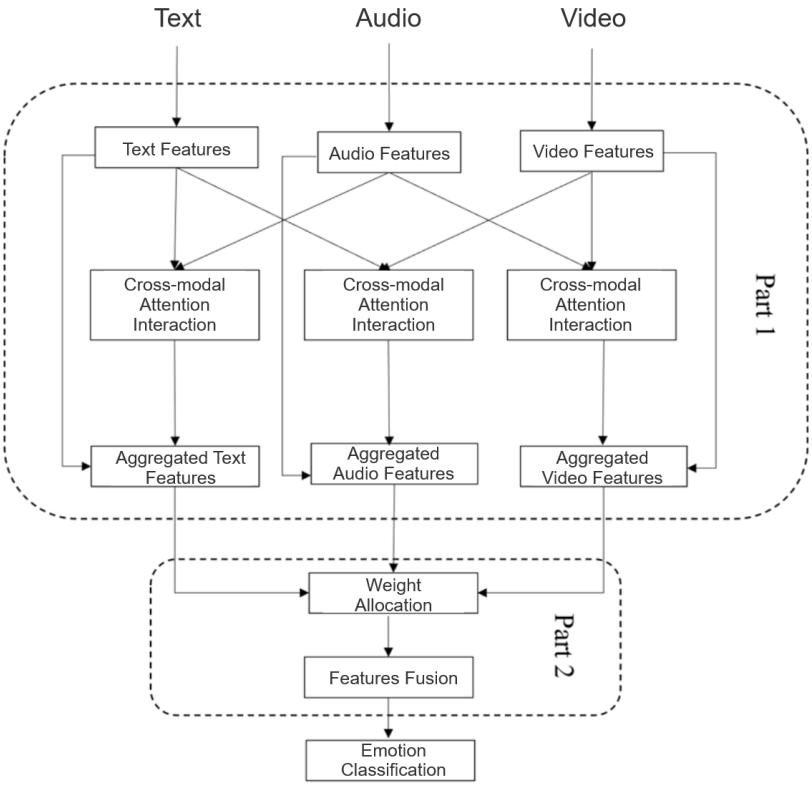


Figure 2. The framework of M2ER. Detailed information will be introduced in the following Sections 2.1 and 2.2.

2.1. The Unimodal Extraction Model of M2ER

We adopted a feature-level fusion method for multimodal emotion recognition, so we need to perform feature extraction for each modality in the first step. The detail of feature

extraction models of M2ER for the three modalities (including text, audio, and video) are described in Sections 2.1.1–2.1.3.

2.1.1. Text Modality Preprocessing and Feature Extraction

We used the Embeddings from Language Model (ELMo) [41] pre-trained model to obtain dynamic word vector features for the text modality. The core of ELMo lies in utilizing a bidirectional Long Short-Term Memory (LSTM) [42] recurrent neural network structure for feature extraction. During training, ELMo leverages the entire input text and considers both forward and backward input sequence information simultaneously to obtain more comprehensive text emotional features. We also adopted BERT [43] to extract semantic information at the sentence level. BERT can be used to extract text emotional features, where the proximity of words in the feature vector space reflects their semantic similarity [44]. The BERT model utilizes the transformer as a feature extractor. When processing a task, BERT first transforms the input text to obtain BERT input representation. Then, the transformer encoder performs computations on the input, then the computed results serve as the input for the next transformer encoder. This process is repeated, resulting in the representation of the entire text.

In conclusion, in the feature extraction part of text modality, we utilized the pre-trained model of ELMo and BERT to obtain text emotional features from the word-level and semantic-level perspectives, respectively. Finally, the extracted features from both parts were combined to obtain complete text emotional features. The process of text modality feature extraction is illustrated in Figure 3.

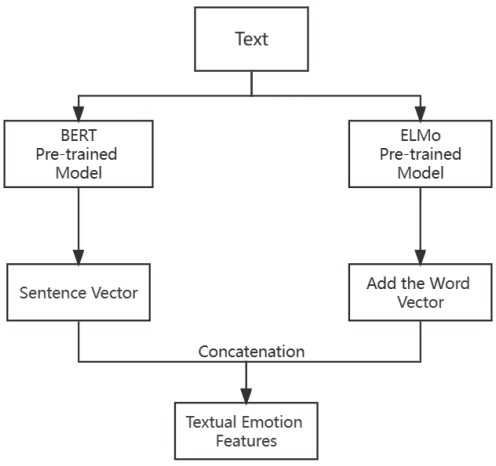


Figure 3. Text feature extraction model.

2.1.2. Audio Modality Preprocessing and Feature Extraction

The representation of audio signals is quite diverse, and the way audio signals are described greatly impacts the performance of subsequent feature extraction and emotion recognition. The purpose of the preprocessing is to transform audio signals with different quality into signals with smooth and uniform representative characteristics, which is convenient for the subsequent feature extraction. Preprocessing includes pre-emphasis, framing, and windowing. The next step is to process the data by transforming the raw audio into spectrograms, which contain both temporal and frequency domain information. These spectrograms are fed into a pre-trained model. Due to its excellent performance in audio emotion recognition, the pre-trained DenseNet [19] network model was selected for extracting emotional features from the spectrograms. The overall steps for audio emotion feature extraction are illustrated in Figure 4.

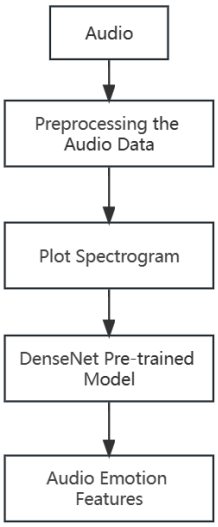


Figure 4. Audio feature extraction model.

Spectrograms visualize audio signals, and they can be regarded as color images in terms of their representation. By using the two-dimensional image to describe the three-dimensional information of time, frequency, and energy, the differences between different audio data can more effectively captured. Moreover, spectrograms are two-dimensional, colorful images, making them suitable for feature extraction using CNNs. The spectrograms corresponding to the seven emotions are shown in Figure 5.

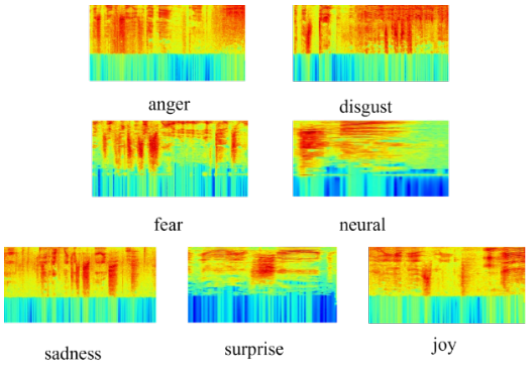


Figure 5. Spectrogram of seven emotions. The horizontal axis of the spectrogram represents the temporal information of the audio signal, while the vertical axis represents the frequency of the audio signal. The two-dimensional coordinates in the spectrogram represent the frequency of the audio at a specific moment, and the intensity of the coordinates also reflects the energy of the audio. The darker the color in the spectrogram, the higher the energy.

Finally, each spectrogram corresponding to each short-time frame is input into DenseNet to extract emotional features from the spectrogram. The detailed architecture of the DenseNet used for feature extraction is shown in Figure 6.

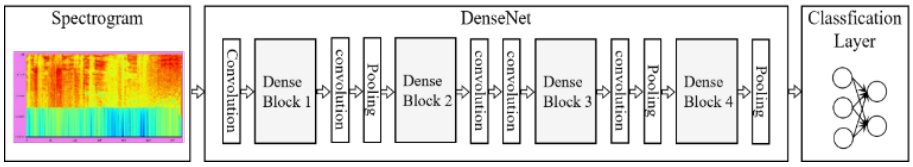


Figure 6. Structure of DenseNet in our method.

2.1.3. Video Modality Preprocessing and Feature Extraction

Capturing the emotional features of facial expressions from speakers accurately is a key challenge in implementing video-based emotion analysis. In real-life applications, the analysis is typically focused on the emotions of the subject (usually human) in a video, and human emotions tend to change slowly over time. Therefore, it is not necessary to analyze every frame in the video when extracting emotional features. Sampling frames from the video and analyzing those samples is sufficient. However, there are usually multiple faces present in the same frame in the case of multi-party dialogue scenarios. The facial expressions of unrelated persons can interfere with the analysis of the speaker's emotions. Therefore, the challenge in analyzing facial expressions in multi-party dialogue scenarios is how to isolate the facial expressions of the speaker.

We selected MELD, which is a widely used multi-party dialogue dataset. In our study, we first read all the sample data from the dataset. For all the video data, we sampled every fifth frame and applied multi-face localization to locate all the faces in the sampled frames, as shown in Figure 7. There are three persons: *Rachel*, *Monica*, and *Phoebe* with distinct facial expressions, i.e., neutral and anger. The facial expressions of the non-speakers (*Rachel* and *Monica*) in the frame can affect the emotion recognition of the real speaker (*Phoebe*). Therefore, it is necessary to exclude the faces of unrelated persons from the frames. Then, we extracted facial expression images of all the faces in the sampled frames.

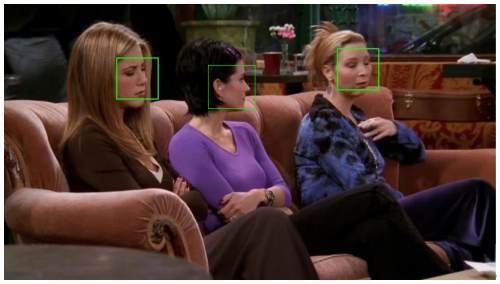


Figure 7. An example of multiple face detection technology locating all faces.

The facial expression image obtained in the previous step is for everyone in the picture, including both the speaker and the non-speaker. We use the speaker recognition method to extract the facial expression sequence of the speaker in the video. The length of the video modality in MELD is set to correspond to a single sentence in the text modality. The text modality also provides speaker annotations for each sentence in the dialogue. We can determine who is speaking in the video by loading the labels from the text modality. We applied speaker recognition to filter out the facial images of the speaker for each video segment. Figure 8 illustrates the changes in a speaker's facial expressions in a specific sequence of video segments.



Figure 8. The changes in the speaker's facial expression. For each video segment, the number of facial expression images obtained through the previous steps is different. To ensure a consistent frame count for each video segment, a trimming and padding process is performed.

Firstly, the average number of facial expression frames obtained is calculated for each video segment in the dataset after the previous steps. During the experiment, we chose to retain a sequence of 30 frames for each video segment. For segments with fewer than 30 frames, zero-padding is used to fill the remaining frames, while for segments with more than 30 frames the sequence is trimmed to 30 frames. After that, the 30-frame facial expression sequence represents the entire video segment, and it can be directly input into the facial feature extraction model.

In the stage of feature extraction, we input the facial expression sequences of the speaker obtained from the previous steps into a pre-trained model VGG16 [45] to extract emotional features from each frame. Since facial expressions are slow-changing sequences, we also adopted LSTM to capture temporal context information through multiple rounds of training, thereby obtaining richer and more comprehensive emotional features.

The specific process for extracting emotional features from the video modality is shown in Figure 9. The preprocessing primarily involves using the OpenCV library to read video frame data. Then, our speaker recognition method of M2ER is applied to filter out the facial expressions of the speaker in the video. As a result, the complete video samples are processed into facial expression sequence images with dimensions of (30, 3, 224, 224). The feature extraction stage utilizes a combination of VGG16 and LSTM. The output from the fully connected layers is used as the emotional feature vector of the video modality.



Figure 9. Video feature extraction model.

2.2. Multimodal Emotion Recognition Based on Attention Mechanism

Building an effective multimodal fusion model is a crucial step in multimodal sentiment recognition. We adopted a feature-level fusion approach to combine the emotional features extracted from the text, audio, and video modalities obtained by the aforementioned models. Current research on multimodal sentiment recognition often relies on extracting a large number of features to identify emotion. However, directly concatenating these features can lead to the curse of dimensionality, and there is no distinction in their importance, which may result in the overshadowing of relatively significant features. Furthermore, there are often correlations among the features from the text, audio, and video modalities. Additionally, it is observable that people express emotion differently in real-life scenarios, but existing multimodal fusion models often overlook this phenomenon. The attention mechanism can be used in neural networks to achieve more effective resource allocation. Based on these issues, M2ER explores an attention-based multimodal fusion model.

Our fusion model consists of three main parts: (shown in Figure 2).

(1) Cross-Modal Attention Interaction—*Part 1*: This module utilizes cross-modal attention to capture the intermodal relationships and obtain the feature representation of the interaction between different modalities.
(2) Multimodal Attention Fusion—*Part 2*: This module employs the attention mechanism to determine the importance of each modality in the final fusion classification. It obtains the weight distribution of each modality's features in the fusion process and performs the fusion accordingly.
(3) Finally, the fused multimodal features are passed through the softmax classification layer for emotion recognition.

2.2.1. Cross-Modal Attention Interaction

Because multimodal emotion recognition often involves a large number of features, determining the importance of these features and capturing the relationships among multimodal emotional features are key issues. In our fusion model, we incorporate the emotional features extracted from the text, audio, and video modalities. This is achieved by utilizing cross-modal attention to facilitate the interaction among different modalities. The input to the Cross-Modal Attention Interaction—*Part 1* is the emotional features of the text, audio, and video modalities, represented as M^T, M^A, and M^V, respectively. The specific architecture of *Part 1* is shown in Figure 10.

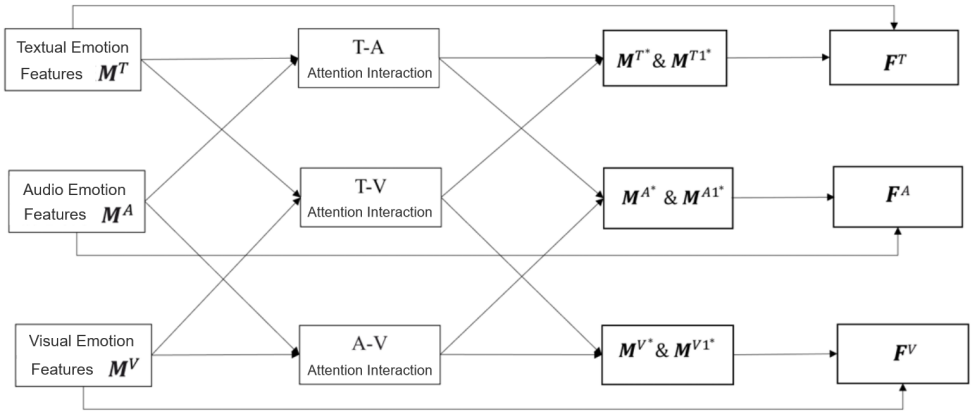


Figure 10. Cross-modal attention interaction architecture.

Taking the example of inputting the text and audio modality into the Cross-Modal Attention Interaction module, M^{T^*} represents the feature representation with interaction obtained from the text through this module. The calculation formula for M^{T^*} is shown in Equations (1)–(3). Similarly, M^{A^*} represents the feature representation with interaction obtained from the audio modality through the T-A attention module. The calculation formula for M^{A^*} is shown in Equations (4)–(6).

$$H^{TA} = M^T M^{A^T}, \tag{1}$$

$$\alpha^{TA} = softmax(H^{TA}, \tag{2}$$

$$M^{T^*} = (\alpha^{TA} M^T) * M^T, \tag{3}$$

$$H^{AT} = M^A M^{T^T}, \tag{4}$$

$$\alpha^{AT} = softmax(H^{AT}, \tag{5}$$

$$M^{A^*} = (\alpha^{AT} M^A) * M^A, \tag{6}$$

where H^{TA} and H^{AT} represent the cross-modal interaction information between the text and audio modalities. α^{TA} and α^{AT} represent the scores obtained for the text and audio modalities in the cross-modal attention interaction. By applying the soft attention mechanism to the emotional features of the input text and audio modalities in *Part 1* and multiplying M^T, M^A with the corresponding elements of their respective matrices, we obtain the feature representation with interaction for the text and audio modalities M^{T^*}, M^{A^*}.

Part 1 is divided into three main parts: text–audio attention interaction (*T-A*), text–video attention interaction (*T-V*), and audio–video attention interaction (A-V).

(1) *T-A*: The emotional features M^T, M^A of the input text and audio modalities in *Part 1* are used to obtain the interaction representation between text and audio M^{T^*}, M^{A^*} through cross-modal attention;
(2) *T-V*: The emotional features M^T, M^V of the input text and video modalities in *Part 1* are used to obtain the interaction representation between text and video, M^{T1^*}, M^{V^*} through cross-modal attention;
(3) *A-V*: The emotional features M^A, M^V of the input audio and video modalities in *Part 1* are used to obtain the interaction representation between audio and video M^{A1^*}, M^{V1^*} through cross-modal attention.

Finally, we obtained the interaction feature representations of the text modality: M^{T^*}, M^{T1^*}; the interaction feature representations of the audio modality: M^{A^*}, M^{A1^*}; and the interaction feature representations of the video modality: M^{V^*}, M^{V1^*}. These representations are concatenated with the respective emotional features of each modality using

fully connected layers to obtain the complete representation of the text, audio, and video modalities' emotional features. The calculation formulas for this process are shown in Equations (7)–(9).

$$F^T = \tanh(W^T[M^T \oplus M^{T^*} \oplus M^{T1^*}] + b^T), \quad (7)$$

$$F^A = \tanh(W^A[M^A \oplus M^{A^*} \oplus M^{A1^*}] + b^A), \quad (8)$$

$$F^V = \tanh(W^V[M^V \oplus M^{V^*} \oplus M^{V1^*}] + b^V), \quad (9)$$

W^T, W^A, W^V, b^T, b^A, b^V are the parameters to be learned, $\oplus$ denotes the concatenation operation. By performing the concatenation operation, we obtain the final text emotional features F_i^T, audio emotional features F_i^A, and video emotional features F_i^V for *Part 1*.

2.2.2. Multimodal Attention Fusion

People often express emotions in different ways in reality. Some people prefer to express their emotions through various facial expressions while others through different tones of voice. Based on this phenomenon, it can be inferred that different modalities of emotional features contribute differently to the final emotion classification. Therefore, in our fusion model, an attention mechanism was adopted to determine the importance of each modality in the final classification. Specifically, the attention mechanism is used to allocate attention weights to the emotional features F^T, F^A, and F^V obtained in *Part 1*. Finally, these weighted features are summed to obtain the fused emotional feature, denoted as F^*. The calculation process is illustrated in Equations (10)–(12):

$$H_X = \tanh(W_{att}^X F^X + b_{att}^X), \quad (10)$$

$$\beta_X = softmax(H_X), \quad (11)$$

$$F^* = \sum_X F^X \beta_X^T, \quad (12)$$

where X represents the modality, which can be text, video, or audio. H_X represents the hidden unit state, W_{att}^X represents the weights, and b_{att}^X represents the biases. Equation (11) is used to normalize the weight vector. The resulting F^* is then fed into the fully connected layer and the softmax classification layer for emotion classification.

3. Evaluation

3.1. Dataset Introduction

This study primarily adopted MELD for experiments. MELD is a multimodal dataset based on dialogues which is widely used for emotion recognition. The dataset consists of over 1400 dialogues, which contain more than 13,000 utterances. Due to the presence of multiple speakers in the same scenario, multi-party dialogues are more challenging than binary dialogues.

For each dialogue segment in MELD, researchers have annotated the corresponding emotion category for each utterance. Table 1 presents the emotion distribution in MELD. Table 2 provides several key statistical data of the dataset. By analyzing the emotion distribution in the training, validation, and test sets, it can be observed that the emotion distribution in the dataset is uneven. The majority of emotions are neutral, while the categories of fear and disgust have fewer instances. So, we conducted further experiments on the MEISD dataset with a more balanced emotional distribution to verify the scalability of our model.

Table 1. The emotion distribution in MELD.

Emotion	Training Set	Test Set	Validation Set
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

Table 2. The detailed distribution of MELD. In the training, validation, and test sets, the average utterance length is almost the same.

MELD Statisic	Training Set	Test Set	Validation Set
No. of modalities	{a, v, t}	{a, v, t}	{a, v, t}
No. of unique words	10643	2384	4361
Avg./Max utterance length	8.0/69	7.9/37	8.2/45
No. of dialogues	1039	114	280
No. of dialogues dyadic MELD	2560	270	577
No. of utterances	9989	1109	2610
No. of speakers	260	47	100
Avg. No. of utterances per dialogue	9.6	9.7	9.3
Avg. No. of emotions per dialogue	3.3	3.3	3.2
Avg./Max No. of speakers per dialogue	2.7/9	3.0/8	2.6/8
No. of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59 s	3.59 s	3.58 s

3.2. Experimental Setting

In our work, the experiments primarily utilized Python with PyTorch. Table 3 displays the hardware configuration used during the experiments. Python 3.7 with PyTorch 1.12.1 was installed on the PC via Anaconda.

Table 3. The server hardware configuration information.

Graphics Card	Server	RAM
NVIDIA GeForce RTX 3090 Ti	AMD Ryzen 9 5950X 16-Core Processor 3.4 GHz	32 G

Our experiment used the cross-entropy loss function and optimized the model parameters using the Adam optimizer [46] with the learning rate of 0.001. To prevent overfitting, We applied the dropout rate of 0.2. The model was trained for 100 epochs with the batch size of 64, which we found to be the most effective.

The experiment was mainly divided into three processes: training, validation, and testing. The model was trained on the training set of MELD, and the validation set was used to observe the training progress of the model and adjust relevant parameters based on the actual training process. Finally, the trained model was used to predict the results on the test set.

3.3. Performance Evaluation

Precision, Recall, and F1 Score are the main key performance indicators used to compare the performance of various models or algorithms [47]. Precision is the ability of the classifier not to label as positive a sample that is negative, and Recall is the ability of the classifier to find all the positive samples. The F1 Score can be interpreted as a weighted harmonic mean of the Precision and Recall. All of them were computed for the proposed model and other baseline models. In our experiment, the micro-F1 Score was used as the evaluation metric.

For binary classification evaluation metrics, the calculation formula for F1 Score is shown in Equations (13)–(15), as follows:

$$Recall = \frac{TP}{TP + FN}, \tag{13}$$

$$Precision = \frac{TP}{TP + FP}, \tag{14}$$

$$F1 = 2 * Recall * \frac{Precision}{Recall + Precision}, \tag{15}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

Multiclass evaluation metrics are derived from binary classification evaluation metrics. The micro-F1 Score takes into account the issue of class imbalance. This approach calculates the global Precision and Recall directly based on individual samples. The calculation formulas are shown in Equations (16)–(18), as follows:

$$Precision_{micro} = \frac{\sum_{i=1}^{L} TP}{\sum_{i=1}^{L} TP + \sum_{i=1}^{L} FP}, \tag{16}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{L} TP}{\sum_{i=1}^{L} TP + \sum_{i=1}^{L} FN}, \tag{17}$$

$$micro - F1 = \frac{2 \cdot Precision_{micro} \cdot Recall_{micro}}{Precision_{micro} + Recall_{micro}}, \tag{18}$$

3.4. Results of Unimodal Experiments

This section primarily outlines the related experiments conducted on MELD, comparing our proposed method with commonly used baseline models for emotion recognition. Specifically, we compare the Text-CNN (text modality only) [48], bcLSTM [49], and DialogueRNN [50] models with our proposed model.

In these experiments, due to the imbalanced distribution of emotions within MELD, we utilized micro-F1 and weighted-average F1 (w-avg F1) as evaluation metrics. Table 4 presents the results for the seven emotion categories on the test set.

Table 4. Scores for unimodal emotion classification on the test set. To facilitate a clearer comparative analysis, this table was transformed into a bar chart as shown in Figure 11.

Model		Emotion							
		Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**	**w-avg F1**
text-CNN	text	34.49	8.22	3.74	49.39	74.88	21.05	45.45	55.02
bcLSTM	text	42.06	21.69	7.75	54.31	71.63	26.92	48.15	56.44
	audio	25.85	6.06	2.9	15.74	61.86	14.71	19.34	39.08
DialogueRNN	text	40.59	2.04	8.93	50.27	75.75	24.19	49.38	57.03
	audio	35.18	5.13	5.56	13.17	65.57	14.01	20.47	41.79
M2ER	text	40.45	15.24	5.55	53.31	77.57	37.85	52.42	60.05
	audio	31.51	9.02	5.25	29.08	64.84	13.06	20.13	43.39
	video	24.43	6.62	4.54	22.89	63.68	20.07	29.44	42.73

It can be observed that the emotion recognition performance in the text modality is generally better than that in the audio and video modalities. The text modality achieved a w-avg F1 of 60.05%, which is an improvement of 9.14%, 6.4%, and 5.3% compared to the Text-CNN, bcLSTM, and DialogueRNN models, respectively. The audio modality achieved a w-avg F1 of 43.39%, which is an improvement of 10.03% and 3.8% compared to the bcLSTM and DialogueRNN models, respectively. These results demonstrate the effectiveness of the text and audio modality feature extraction models, surpassing the performance of popular baseline models.

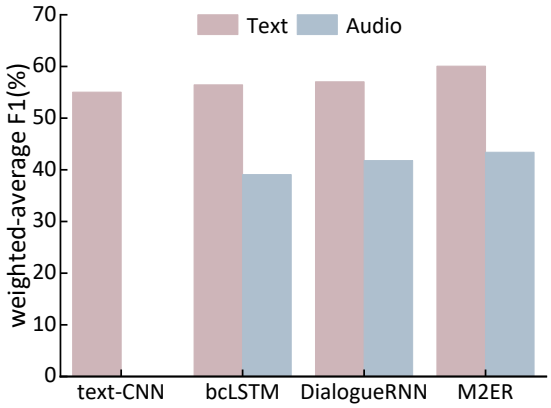


Figure 11. Performance results of different models of text and audio modalities on MELD. The figure shows the comparison of the w-avg F1 values for different models in text and audio modality emotion recognition on MELD.

3.5. Results of Multimodal Experiments

Relevant experiments were conducted on the training, validation, and test sets of MELD. The experimental results are shown in Table 5. This model utilizes cross-modal attention interaction to capture correlated information between modalities, obtaining feature representations with interactive effects. By using the attention mechanism, it determines the importance of each modality in the final fusion classification and combines the multimodal information.

Table 5. Scores for the seven emotion classifications on the test set. *"text + audio + video"* represents our attention-based multimodal fusion model. This table was transformed into a bar chart, as shown in Figure 12.

Model		**Emotion**							
		Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**	**w-avg F1**
bcLSTM	text + audio	43.39	23.66	9.38	54.48	76.67	24.34	51.04	59.25
DialogueRNN	text + audio	43.65	7.89	11.68	54.40	77.44	34.59	52.51	60.25
M2ER	text	40.45	15.24	5.55	53.31	77.57	37.85	52.42	60.05
	audio	31.51	9.02	5.25	29.08	64.84	13.06	20.13	43.39
	video	24.43	6.62	4.54	24.89	63.68	20.07	29.44	42.73
	text + audio + video	43.75	16.93	9.62	56.63	80.11	41.67	54.14	62.83

Comparing the results in Figure 12 shows that our three-modality fusion emotion recognition model achieved a w-avg F1 score of 62.83%, outperforming the individual modalities of text, audio, and video in emotion recognition. Compared to unimodal data, multimodal data can capture more diverse emotional features. Multimodal fusion can also compensate for the limitations of individual modalities. Furthermore, our multimodal fusion model shows improved fusion performance compared to several baseline models. It was found that the recognition results for disgust and fear were not satisfactory. To address this, we further conducted a five-class emotion recognition experiment on MELD, excluding the less frequent emotions of fear and disgust. The results of this experiment are shown in Table 6.

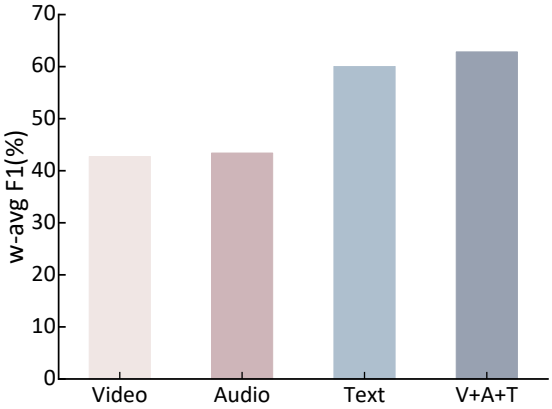


Figure 12. Results of unimodal and multimodal emotion classification of our model on the test set. They present a comparison of the w-avg F1 values for both unimodal and multimodal emotion recognition.

Table 6. Scores for the five emotion classifications on the test set.

Model		Emotion					
		Anger	**Joy**	**Neutral**	**Sadness**	**Surprise**	**w-avg F1**
bcLSTM	text + audio	45.9	52.2	77.9	11.2	49.9	60.6
DialogueRNN	text	41.7	53.7	77.8	21.2	47.7	60.8
	audio	34.1	18.8	66.2	16	16.6	44.3
	text + audio	48.2	53.2	77.7	20.3	48.5	61.6
M2ER	text	42.1	53.2	78.6	35.9	52.3	62.9
	audio	32.5	31.0	66.1	13.6	23.2	46.7
	video	29.2	26.0	65.7	19.5	29.6	46.3
	text + audio + video	45.6	55.1	79.4	36.3	53.9	64.3

By comparing the results in Tables 5 and 6, it can be observed that after excluding two less frequent emotions, the five-class emotion recognition performance significantly improved compared to the seven-class classification because there are very few samples for the emotions of fear and disgust in the training set. Additionally, distinguishing between anger, disgust, and fear is challenging as the differences between these emotions are subtle. This explains why the recognition results for disgust and fear were relatively poor in the seven-class emotion recognition experiment. Furthermore, the performance of the text modality in emotion recognition remained generally superior to that of the audio and video modalities in the five-class emotion recognition experiment, and the multimodal emotion recognition outperformed the single modality.

3.6. Ablation Experiments

We conducted ablation experiments to validate the effectiveness of different components designed in our multimodal feature fusion model by removing specific modules in the multimodal fusion part.

Our fusion model mainly consists of two parts: *Part 1* utilizes cross-modal attention to capture the interaction between modalities; *Part 2* utilizes the attention mechanism to determine the importance of each modality for the final classification and fuses the multimodal information. We conducted comparative experiments between direct concatenation and the fusion mechanism we adopted, as shown in Table 7.

Comparing the experimental results in Figure 13, *Fusion 2* improved the performance by 1.09% compared to *Fusion 1*. Compared with *Fusion 1* and *Fusion 2*, *Fusion 3* improved the performance by 3.4% and 2.3%, respectively, as *Fusion 2* and *Fusion 3* utilize the correlated information between modalities and effectively allocate importance weights to each modality. The ablation experiments confirmed the effectiveness of each module in the fusion model in our work—*Part 1* and *Part 2*.

Table 7. Results of ablation experiments for multimodal fusion emotion recognition. *Fusion 1* represents direct concatenation; *Fusion 2* represents the first part of the fusion model, *Part 1*, which considers only the interaction between modalities and within each modality; *Fusion 3* represents the complete fusion model, *Part 1* + *Part 2*. This table has been transformed into a bar chart, as shown in Figure 13.

Model			Emotion							
			Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**	**w-avg F1**
M2ER	Fusion 1	text + audio + video	41.81	11.63	7.10	54.37	77.84	40.09	53.41	60.74
	Fusion 2	text + audio + video	41.99	15.83	7.90	55.07	78.52	40.39	53.86	61.40
	Fusion 3	text + audio + video	43.75	16.93	9.62	56.63	80.11	41.67	54.14	62.83

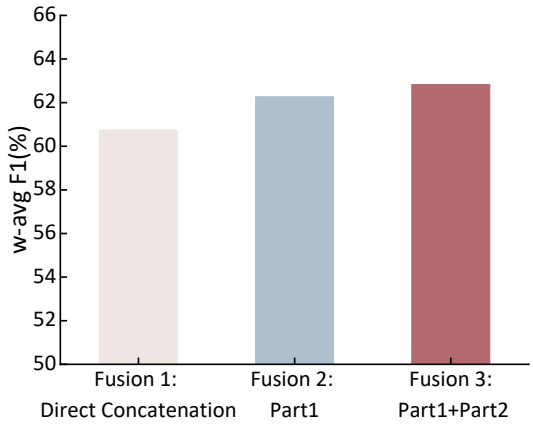


Figure 13. Experimental results of multimodal emotion recognition ablation. They present the comparison of the w-avg F1 scores for different variants of the multimodal fusion model on MELD.

Since the baseline models on MELD did not utilize the video modality, and it was found that most dialogue emotion recognition studies based on MELD also did not utilize the video modality through research, A comparative analysis was performed between the video modality emotion recognition models without speaker recognition and the models utilizing it in order to validate the effectiveness of our proposed speaker recognition method. The results are presented in Table 8.

Table 8. Results of the ablation experiments in the video modality. *video′* represents the method where our method was not used. *video* represents our proposed method of using speaker recognition.

Model	Emotion							
	Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**	**w-avg F1**
video′	22.64	5.41	3.32	23.07	60.34	14.15	24.16	40.05
video	24.43	6.62	4.54	22.89	63.68	20.07	29.44	42.73

By comparing the data in Table 8, it can be observed that our model achieved a w-avg F1 score of 42.73% for emotion classification in the video modality. Furthermore, by comparing the data of *video′* and *video*, it is evident that our proposed speaker recognition method improved the emotion recognition performance in the video modality by 6.58%. The comparison in Table 8 indicates that the method effectively enhances the efficiency of extracting emotional features in multi-party dialogue scenarios, highlighting the role of the video modality in emotion recognition during multi-party dialogues.

3.7. Model Scalability Verification

To validate the scalability of our multimodal emotion recognition model, we conducted emotion recognition experiments on the MEISD dataset, which is also a multi-party dialogue dataset. Additionally, we performed emotion recognition tests on some real-world

data using the fusion model and presented the test results to visualize the generalization performance of the model.

We further conducted multi-class emotion recognition experiments on the MEISD dataset, using the micro-F1 score as the evaluation metric. Since the distribution of each emotion label in the training, validation, and test sets of the MEISD dataset is relatively balanced, the prediction performance for each emotion label is almost the same. Therefore, we collected the overall w-avg F1 score for comparison and presentation in the experiments. The results are shown in Table 9.

Table 9. Scores for emotion classification on the MEISD dataset. This table has been transformed into a bar chart, as shown in Figure 14.

Model		w-avg F1
text-CNN	text	54.18
bcLSTM	text	57.05
	audio	41.17
	video	39.45
	text + audio + video	59.32
DialogueRNN	text	58.73
	audio	41.52
	video	40.87
	text + audio + video	60.57
M2ER	text	61.10
	audio	43.49
	video	42.91
	text + audio + video	62.97

Figure 14 shows that our multimodal emotion recognition approach also performs well on the MEISD dataset. The w-avg F1 score for text modality emotion recognition is 61.10%, for audio modality is 43.49%, and for the video modality is 42.91%. The performance of individual modalities in emotion recognition surpasses that of classical baseline models. The fusion model achieves a w-avg F1 score of 62.97%, outperforming the individual modality recognition results. These experimental results demonstrate the effectiveness and scalability of our multimodal emotion recognition model in multi-party dialogue scenarios.

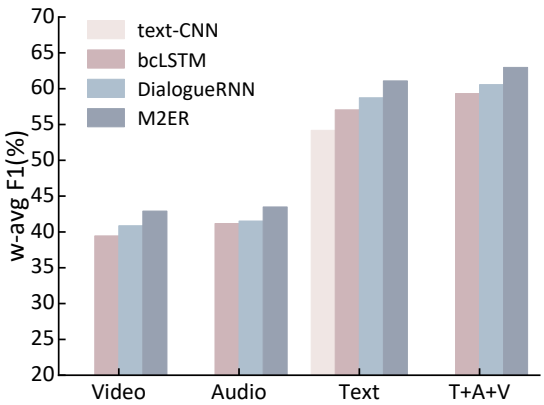


Figure 14. Emotional classification results of the MEISD. They present the comparison of w-avg F1 scores for unimodal and multimodal emotion recognition of various models on the MEISD dataset.

To better illustrate the scalability of the model, we tested it on some actual examples, as shown in Table 10. From the table, it can be observed that in the case of Example A, the facial image of the speaker obtained from the video modality shows an upward curvature of the mouth, indicating a smiling expression. Additionally, the speaker's voice has a high pitch and a cheerful speaking rate. The text also expresses a positive emotion, leading to

a predicted emotion of positive. In the case of Example B, the facial image of the speaker obtained from the video modality shows a furrowed brow, and the speaker's speech is slow and filled with a tone of sadness. The text modality also exhibits negative sentiment, resulting in a predicted emotion of negative, which aligns with the authentic label. In the case of Example C, the facial expression of the speaker obtained from the video modality is relatively neutral without a clear emotional color, but its text and audio modalities have evident negative sentiment, so the final prediction result is also negative, which is consistent with the real label. Through the analysis and presentation above examples, it is evident that our multimodal emotion recognition model based on multi-party dialogue scenarios can effectively identify the speaker and successfully fuse information from the text, audio, and video modalities.

Table 10. Examples of multimodal emotion recognition. *T* represents the dialogue text, *V* represents the speaker's visual information, and *A* represents the audio information in the video.

Example	Speaker	T	V	A	Authentic Emotion	Predicting Emotion
A	Phoebe	Ohh! I'm gonna be on the news.		high pitch, cheerful tone	Positive	Positive
B	Monica	So, I hear you, you hate me!		slow pace, downcast tone	Negative	Negative
C	Ross	Look! I did not feel like dancing. Okay?		downcast tone, high pitch	Positive	Positive

4. Discussion

4.1. Strengths

M2ER has solved the problem of speaker recognition when multiple faces appear in the same video frame, which enables utilization of the video modality for emotion recognition. When encountering multiple people in the same scene, M2ER eliminates the interference of other people by recognizing the speaker and using the facial expressions of the speaker in the video and the changes during video playback.

To incorporate the multimodal fusion model into our approach, we employed a feature-level fusion that relies on the attention mechanism. By utilizing cross-modal attention, we combined the extracted unimodal emotional features to effectively capture intermodal interactions.

4.2. Limitations and Future Work

In real-world scenarios, a variety of factors can cause modality absence, such as the faces in the video not appearing within the range of the camera at some moments. The datasets we used are also affected by modality absence, which definitely affects the accuracy of the modality. In future work, we will focus on addressing the issue of modality absence, which may enhance M2ER.

The proposed method involves multiple components, including face detection, face recognition, and attention-based fusion; while these components enhance the effectiveness of the model, they also introduce complexity that undoubtedly increases the difficulty of implementation for others.

Another limitation is that only two datasets were utilized for the experiment, without further testing the generalization of the model. Due to the potential impact of different data sources on performance, it is necessary to explore how well the proposed method generalizes to other datasets due to the potential impact of diverse data sources on performance in the future.

5. Conclusions

In this paper, our work primarily focuses on the research of multimodal emotion recognition in multi-party dialogue scenarios. We propose a novel approach using multi-face localization for speaker recognition in the video modality, thus enhancing the efficiency of utilizing the video modality in the field of multimodal emotion recognition. In the multimodal fusion part, we explore a multimodal feature fusion model based on attention mechanism to address dimension explosion and poor correlation in the directly concatenated fusion model. We conducted seven-class emotion experiments, five-class emotion experiments, and scalability experiments. The results validate the effectiveness of M2ER.

Author Contributions: Conceptualization, G.W.; methodology, B.Z. and X.Y.; validation, B.Z., X.Y. and G.W.; formal analysis, Y.W.; investigation, Y.W.; resources, X.Y.; data curation, G.W. and B.Z.; writing—original draft preparation, X.Y.; writing—review and editing, B.Z. and R.S.; visualization, Y.W; supervision, B.Z. and R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
2. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [CrossRef] [PubMed]
3. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
4. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
5. Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; Tian, Q. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 3030–3043. [CrossRef]
6. Perveen, N.; Roy, D.; Chalavadi, K.M. Facial expression recognition in videos using dynamic kernels. *IEEE Trans. Image Process.* **2020**, *29*, 8316–8325. [CrossRef] [PubMed]
7. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
8. Mao, J.; Xu, R.; Yin, X.; Chang, Y.; Nie, B.; Huang, A. Poster v2: A simpler and stronger facial expression recognition network. *arXiv* **2023**, arXiv:2301.12149.
9. Panagiotis, A.; Filntisis, P.P.; Maragos, P. Exploiting emotional dependencies with graph convolutional networks for facial expression recognition. *arXiv* **2021**, arXiv:2106.03487.
10. Ryumina, E.; Dresvyanskiy, D.; Karpov, A. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing* **2022**, *514*, 435–450. [CrossRef]
11. Bakariya, B.; Singh, A.; Singh, H.; Raju, P.; Rajpoot, R.; Mohbey, K.K. Facial emotion recognition and music recommendation system using cnn-based deep learning techniques. In *Evolving Systems*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 1–18.
12. Meena, G.; Mohbey, K.K.; Indian, A.; Khan, M.Z.; Kumar, S. Identifying emotions from facial expressions using a deep convolutional neural network-based approach. In *Multimedia Tools and Applications*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 1–22.
13. Savchenko, A.V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In Proceedings of the 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 16–18 September 2021; pp. 119–124.
14. Meena, G.; Mohbey, K.K.; Kumar, S. Sentiment analysis on images using convolutional neural networks based inception-v3 transfer learning approach. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100174. [CrossRef]
15. Mehendale, N. Facial emotion recognition using convolutional neural networks (ferc). *SN Appl. Sci.* **2020**, *2*, 446. [CrossRef]
16. Li, T.H.S.; Kuo, P.H.; Tsai, T.N.; Luan, P.C. Cnn and lstm based facial expression analysis model for a humanoid robot. *IEEE Access* **2019**, *7*, 93998–94011. [CrossRef]

17. Ming, Y.; Qian, H.; Guangyuan, L. Cnn-lstm facial expression recognition method fused with two-layer attention mechanism. *Comput. Intell. Neurosci.* **2022**, *2022*, 7450637. [CrossRef] [PubMed]
18. Sang, D.V.; Ha, P.T. Discriminative deep feature learning for facial emotion recognition. In Proceedings of the 2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR), Ho Chi Minh City, Vietnam, 5–6 April 2018; pp. 1–6.
19. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
20. Xue, F.; Wang, Q.; Guo, G. Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3601–3610.
21. Ma, F.; Sun, B.; Li, S. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1236–1248. [CrossRef]
22. Chen, C.-F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.
23. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S.J. Rethinking spatial dimensions of vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11936–11945.
24. Nguyen, D.; Nguyen, K.; Sridharan, S.; Ghasemi, A.; Dean, D.; Fookes, C. Deep spatio-temporal features for multimodal emotion recognition. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 1215–1223.
25. Guanghui, C.; Xiaoping, Z. Multi-modal emotion recognition by fusing correlation features of speech-visual. *IEEE Signal Process. Lett.* **2021**, *28*, 533–537. [CrossRef]
26. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, http://arxiv.org/abs/1606.06259.
27. Zadeh, A.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018. Available online: https://api.semanticscholar.org/CorpusID:51868869 (accessed on 15 July 2018).
28. Dai, W.; Cahyawijaya, S.; Liu, Z.; Fung, P. Multimodal end-to-end sparse model for emotion recognition. *arXiv* **2021**, arXiv:2103.09666.
29. Ren, M.; Huang, X.; Shi, X.; Nie, W. Interactive multimodal attention network for emotion recognition in conversation. *IEEE Signal Process. Lett.* **2021**, *28*, 1046–1050. [CrossRef]
30. Khare, A.; Parthasarathy, S.; Sundaram, S. Self-supervised learning with cross-modal transformers for emotion recognition. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 381–388.
31. Lv, F.; Chen, X.; Huang, Y.; Duan, L.; Lin, G. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2554–2562.
32. Xie, B.; Sidulova, M.; Park, C.H. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. *Sensors* **2021**, *21*, 4913. [CrossRef]
33. Zhang, L.; Liu, C.; Jia, N. Uni2mul: A conformer-based multimodal emotion classification model by considering unimodal expression differences with multi-task learning. *Appl. Sci.* **2023**, *13*, 9910. [CrossRef]
34. Wang, H.; Yang, M.; Li, Z.; Liu, Z.; Hu, J.; Fu, Z.; Liu, F. Scanet: Improving multimodal representation and fusion with sparse-and cross-attention for multimodal sentiment analysis. *Comput. Animat. Virtual Worlds* **2022**, *33*, e2090. [CrossRef]
35. Ma, H.; Wang, J.; Qian, L.; Lin, H. Han-regru: Hierarchical attention network with residual gated recurrent unit for emotion recognition in conversation. *Neural Comput. Appl.* **2021**, *33*, 2685–2703. [CrossRef]
36. Jiao, W.; Lyu, M.R.; King, I. Real-time emotion recognition via attention gated hierarchical memory network. *arXiv* **2019**, http://arxiv.org/abs/1911.09075.
37. Xing, S.; Mai, S.; Hu, H. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1426–1439. [CrossRef]
38. Hu, D.; Hou, X.; Wei, L.; Jiang, L.; Mo, Y. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7037–7041.
39. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.
40. Firdaus, M.; Chauhan, H.; Ekbal, A.; Bhattacharyya, P. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 4441–4453. Available online: https://aclanthology.org/2020.coling-main.393 (accessed on 8 December 2020).
41. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
42. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

43. Munikar, M.; Shakya, S.; Shrestha, A. Fine-grained sentiment classification using bert. In Proceedings of the 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 5 November 2019; Volume 1, pp. 1–5.
44. Jiang, D.; He, J. Tree framework with bert word embedding for the recognition of Chinese implicit discourse relations. *IEEE Access* **2020**, *8*, 162004–162011. [CrossRef]
45. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
46. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
47. Sharma, S.; Rana, V.; Kumar, V. Deep learning based semantic personalized recommendation system. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100028. [CrossRef]
48. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882.
49. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.-P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 873–883. Available online: https://aclanthology.org/P17-1081 (accessed on 30 July 2017).
50. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *arXiv* **2018**, arXiv:1811.00405.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.