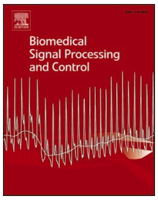
Biomedical Signal Processing and Control 85 (2023) 105052

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Multimodal emotion recognition based on audio and text by using hybrid attention networks

Shiqing Zhang [a], Yijiao Yang [a,b], Chen Chen [a], Ruixin Liu [a,b], Xin Tao [a], Wenping Guo [a], Yicheng Xu [c], Xiaoming Zhao [a,*]

[a] *Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, Zhejiang, China*
[b] *School of Science, Zhejiang University of Science & Technology, Hangzhou 310023, Zhejiang, China*
[c] *School of Information Technology Engineering, Taizhou Vocational and Technical College, Taizhou 318000, Zhejiang, China*

ARTICLE INFO

Keywords:
Multimodal emotion recognition
Deep learning
Local intra-modal attention
Cross-modal attention
Global inter-modal attention

ABSTRACT

Multimodal Emotion Recognition (MER) has recently become a popular and challenging topic. The most key challenge in MER is how to effectively fuse multimodal information. Most of prior works may not fully consider the inter-modal and intra-modal attention mechanism to jointly learn intra-modal and inter-modal emotional salient information for further improving the performance of MER. To address this problem, this paper proposes a new MER framework based on audio and text by using Hybrid Attention Networks (MER-HAN). The proposed MER-HAN combines three different attention mechanisms such as the local intra-modal attention, the cross-modal attention, and the global inter-modal attention to effectively learn intra-modal and inter-modal emotional salient features for MER. Specifically, an Audio and Text Encoder (ATE) block equipped with deep learning techniques with the local intra-modal attention mechanism is initially designed to learn high-level audio and text feature representations from the corresponding audio and text sequences, respectively. Then, a Cross-Modal Attention (CMA) block is presented to jointly capture high-level shared feature representations across audio and text modalities. Finally, a Multimodal Emotion Classification (MEC) block with the global inter-modal attention mechanism is provided to obtain final MER results. Extensive experiments conducted on two public multimodal emotional datasets, *i.e.*, IEMOCAP and MELD datasets, show the advantage of the proposed MER-HAN on MER tasks.

1. Introduction

Emotion recognition is a dynamic process that takes people's affective states as a target and plays a crucial role in human behavior analysis [1]. During the past two decades, automated emotion recognition has attracted extensive attentions due to its application to Human-Computer Interaction (HCI) [2–4]. When machines are capable of modeling and recognizing human emotions, effectively improving the efficiency of natural HCI will become a reality [5].

Most of prior emotion recognition studies concentrate on single-modal emotion recognition, such as audio/speech [6,7], facial expression [8], text [9], and so on. In despite of the achieved great progress in single-modal emotion recognition tasks, single-modal emotion recognition still has two drawbacks. First, emotion modeling from a single modality cannot accurately characterize people's affective states. This may be because human beings express their emotions in multiple different ways simultaneously, such as audio, facial expressions, text, *etc.* [10,11]. Second, single-modal data could not contain the complete emotional semantic information, since multimodal data from different modalities may have interrelated or complementary relationships on ultimate emotion recognition tasks. To tackle this issue, researchers have endeavored to extend the relatively simple single-modal emotion recognition to complicated Multimodal Emotion Recognition (MER) tasks [12–14]. MER has the advantage of multi-source knowledge related to human emotion expression compared with single-modal emotion recognition. In particular, human language is mainly composed of acoustic and textual representations [15,16]. In this case, both audio and text modalities can provide important affective information in a complementary manner to identify human emotional states. Therefore, integrating audio and text cues may have great potential for building an effective MER model, which will be investigated in this work.

[*] Corresponding author.
E-mail address: tzxyzxm@163.com (X. Zhao).

https://doi.org/10.1016/j.bspc.2023.105052
Received 1 February 2023; Received in revised form 11 May 2023; Accepted 23 May 2023
Available online 30 May 2023

However, due to the highly heterogeneous property of multimodal data, the large discrepancy between audio and text data brings about a great challenge when fusing audio and text modalities on MER tasks. Therefore, an effective strategy of multimodal data fusion integrating heterogeneous audio and text data is crucial for a basic MER system. To mitigate this issue, in early MER studies [17–19] three typical multimodal data fusion approaches, such as feature-level fusion (called Early Fusion, EF), decision-level fusion (called Late Fusion, LF), and model-level fusion, were widely employed to fuse audio and text cues, as described below.

Feature-level fusion is the most simple fusion method in which different features extracted from multimodal modalities are directly concatenated into a long feature vector as an input of the latter classifier for MER [20–22]. However, such simple feature concatenation does not consider the temporal scales, thereby ignoring the inter-modal dynamics. Moreover, feature-level fusion easily suffers from the curse of dimensionality owing to the concatenated high-dimensional feature vector. By contrast, decision-level fusion aims to model different modalities independently, and then uses certain calculation rules, such as "average", "maximum", "majority vote" and so on, to merge the obtained results of different modalities [23,24]. Nevertheless, decision-level fusion does not accurately reflect the inter-relationship across modalities. Model-level fusion aims to explicitly exploit the correlation among different modalities, and is thus a tradeoff between feature-level and decision-level fusion methods. Several representative model-level fusion methods are multiple kernel learning [21,25] and deep learning-based models [26–31]. Recently, various attention-based deep learning models, which can be roughly categorized into two groups: inter-modal attention based methods [32,33] and intra-modal attention based methods [34,35], have been applied for MER. It is noted that inter-modal attention based methods concentrate on the dynamic interaction across different modalities, and neglect the inherent connection among feature elements from single-modality. By contrast, intra-modal attention based methods stress on extracting emotional salient feature representations from single-modality, thereby ignoring the inter-modal relationship across different modalities. In this sense, the intra-modal relationship within the single-modality and the inter-modality relationship among different modalities may enhance each other to some extent. However, most of prior works may not fully consider the inter-modal and intra-modal attention mechanism to jointly learn intra-modal and inter-modal emotional salient information for potential performance improvement on MER tasks.

To address the above-mentioned problem, this paper proposes a new MER framework based on audio and text by using Hybrid Attention Networks (MER-HAN). There are two key issues to be addressed in our MER-HAN framework. One is that how to model single-modal temporal contextual information when learning salient features from audio and text sequences, respectively. The other is that how to capture complementary information in the process of fusing audio and text features. To tackle these two issues, Fig. 1 presents the framework of our proposed MER-HAN. More specifically, an Audio and Text Encoder (ATE) block equipped with deep learning techniques with the local intra-modal attention mechanism is initially designed to learn high-level audio and text feature representations from the corresponding audio and text sequences, respectively. Then, a Cross-Modal Attention (CMA) block is presented to jointly capture high-level shared feature representations for audio and text modalities. The used CMA mechanism aims to learn different emotional interaction weights across audio and text modalities, thereby capturing shared audio and text feature representations for downstream tasks. Finally, the interactively learned shared audio and text feature representations are concatenated and then fed into a Multimodal Emotion Classification (MEC) block with the global inter-modal attention mechanism so as to produce final emotion classification results. Extensive experiments are performed on two typical multimodal emotional datasets including IEMOCAP [36] and MELD [37], indicating the effectiveness of our proposed MER-HAN approach on MER tasks.

In summary, the main contributions of this paper are two-fold:

(1) This paper proposes a new MER method based on audio and text cues by using Hybrid Attention Networks (MER-HAN), in which three different attention mechanisms such as the local intra-modal attention, the cross-modal attention, and the global inter-modal attention are designed to jointly learn emotional salient features for MER.
(2) Extensive experiments on IEMOCAP and MELD datasets show that the proposed MER-HAN obtains better performance than other multimodal data fusion methods, demonstrating the advantage of the proposed MER-HAN on MER tasks.

The rest of this paper is organized as follows. Section 2 reviews related works, and Section 3 describes the proposed MER-HAN approach in detail. Experiments and analysis are provided in Section 4. Finally, Section 5 presents the conclusions.

2. Related work

2.1. Single-modal emotion recognition

Feature extraction and emotion classification are two key steps for single-model emotion recognition. Feature extraction is to extract feature representations related to emotion expression from the corresponding single modality. Emotion classification aims to employ a suitable classifier to learn the mapping relationship between the extracted feature representations and emotion labels so as to obtain final emotion recognition results. The conventional machine learning methods, such as Hidden Markov Models (HMM), Support Vector Machines (SVM), Bayesian Network (BN), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), decision trees, *etc.*, have been widely used as emotion classifiers for single-modal emotion recognition [38–42]. In this work, we focus on the first key step of feature extraction for single-model emotion recognition such as audio or text modality, as described below.

2.1.1. Audio emotion recognition

Affective acoustic feature extraction is very crucial for audio emotion recognition. The early-used typical acoustic features are hand-crafted acoustic Low-Level Descriptors (LLDs) [38,39,42], such as prosody features, voice quality features, spectral features, *etc*. The common prosody features contain pitch, loudness, and duration [43]. The conventional voice quality features include formants, spectral energy distribution, glottal features, voice source parameters, and so on [44]. The representative spectral features are Mel-Frequency Cepstral Coefficients (MFCCs) [45,46]. In recent years, several extensive feature sets consisting of thousands of high-level statistical parameters of LLDs, such as Interspeech-2010 [47], ComParE [48], GeMAPS and its extension called eGeMAPS [49], have exhibited good performance on audio emotion recognition tasks. Li *et al.*, [50] presented an emotional-category based feature weighting method to find the prominence of each feature under different emotions, based on the extracted Interspeech-2010, GeMAPS and eGeMAPS. Although these used hand-crafted acoustic LLDs have presented good performance on audio emotion recognition tasks, they are still low-level, thereby resulting in the big affective gap between these hand-crafted features and subjective emotions.

To bridge the above-mentioned affective gap, recently-emerged deep learning techniques [51,52] have been widely applied to audio emotion recognition due to its powerful feature learning capability. Commonly-used deep learning methods for audio emotion recognition include deep Convolutional Neural Networks (CNNs) [53], Recurrent Neural Networks (RNNs) [54], and its variants called Long Short-term Memory (LSTM) [55], *etc.* Ottl *et al.*, [56] employed different CNN models to extract deep image-based features from Emotion Recognition in the Wild

(EmotiW)-2020 group-level emotion prediction challenge data, and then fused them by early and late fusion for emotion classification. Zhang *et al.*, [57] provided a multi-scale deep convolutional LSTM model integrating multiple CNNs and LSTM for feature learning at different lengths of audio spectrograms on emotion classification tasks. Despite of the great breakthroughs achieved by deep learning techniques on audio emotion recognition tasks, the obtained performance is still limited because of the inherent drawback of the used audio modality related to emotion expression.

2.1.2. Text emotion recognition

Text emotion recognition is dedicated to automatically recognizing emotional states in textual expressions by using textual features. The commonly-used feature extraction methods in early text emotion recognition works are hand-crafted Bag-of-Words (BoW) model [58], Latent Dirichlet Allocation (LDA) [59], Latent Semantic Analysis (LSA) [60], and so on. However, these hand-crafted models have an inability of capturing high-level semantic information behind text data. To alleviate the above-mentioned issue, deep learning techniques have also been employed to improve the performance of text emotion recognition in recent years. Yang *et al.*,[61] proposed an attention-based bidirectional LSTM method by means of learning the alignment between target entities and the most significant features to improve target-dependent sentiment classification. In addition, several typical pre-trained word embedding methods, such as Word2vec [62], and GloVe [63], have also been employed for text emotion analysis. Zhou *et al.*, [64] utilized a Word2vec and stacked bidirectional LSTM models for emotion analysis in microblogs.

More recently, a variety of pre-trained language models in the field of NLP, such as Contextualized Word Vectors (CoVe) [65], Embedding from Language Models (ELMo) [66], Bidirectional Encoder Representations from Transformers (BERT) [67], have been adopted for text emotion classification. Kumar *et al.*, [68] provided a BERT based dual-channel explainable system for text emotion classification. They employed the pre-trained BERT model to extract textual features as an input of the used dual-channel networks consisting of CNNs and Bi-LSTM. Finally, the dual-channel outputs were concatenated and fed into an emotion classification module for text emotion classification. Although text emotion recognition has made great achievements in recent years, it may not be able to convey enough contextual information to distinguish human emotions when text modality is only leveraged for sentiment analysis.

2.1.3. Multimodal emotion recognition integrating audio and text

Except for the above-mentioned audio and text feature extraction, multimodal data fusion effectively integrating audio and text modalities, is another key step in a MER system. Early multimodal data fusion strategies mainly contain feature-level fusion, decision-level fusion and model-level fusion. Hazarika *et al.*, [69] presented a Self-Attentive Feature-level Fusion (SAFF) method for MER. They adopted the typical ComParE set [48] as audio features, and a CNN to extract textual features. Then, they fused audio features and textural features at feature-level for MER by performing a weighted addition using the attention score probabilities. Bayerl *et al.*, [70] extracted word-based acoustic and textual embeddings and then leveraged early and late fusion methods to detect emotion carriers defined as the segments (speech or text) in spoken narratives. Specially, they used Residual Neural Networks (ResNet) [71] for acoustic word-level feature representations, as well as the pre-trained GloVe [63] model for word-based text feature representations. Tang *et al.*, [72] presented a Hierarchical Fusion Graph Convolutional Network (HFGCN) model for MER. Specifically, they extracted the typical ComParE [48] as audio features, and adopted the pre-trained Word2vec [62] to derive text features. Then, HFGCN was used to integrate multimodal input data at model-level by means of using a two-stage graph construction method and encode the inter-modality dependencies into the conversational feature representations for MER.

However, these above-mentioned works mainly adopt early or late fusion methods for MER, thereby ignoring the correlations and interactions across audio and text modalities. To address this issue, this paper proposes a new MER-HAN framework integrating audio and text cues for emotion classification. The proposed method focuses on jointly learning intra-modal and inter-modal emotional salient information for further improving the performance of MER, as described below.

3. Proposed method

In this section, we describe the details of the proposed MER-HAN framework based on audio and text cues, as depicted in Fig. 1. Our proposed MER-HAN is mainly composed of three blocks, including an Audio and Text Encoder (ATE) block for feature extraction, a Cross-modal Attention (CMA) block for cross-modal fusion, and a Multimodal Emotion Classification (MEC) block for emotion classification, which will be elaborated below.

3.1. System overview

As shown in Fig. 1, an ATE block associated with the local intra-modal attention mechanism is initially designed to learn high-level audio and text feature representations from two input streams, respectively. Then, a CMA block associated with the cross-modal attention mechanism is presented to obtain inter-modal dynamics information. This is attributed to that the used CMA mechanism can learn different emotional interaction weights across audio and text modalities, thereby capturing shared audio and text feature representations for downstream tasks. Finally, the interactively learned shared feature representations are merged and then fed into a MEC block associated with the global inter-modal attention mechanism to conduct final MER tasks.

3.2. Audio and text encoder block for feature extraction

In this section, the designed ATE block utilizes appropriate encoders to learn audio and text feature representations from the corresponding audio and text modalities, respectively. Given the multimodal input data D, consisting of N audio-text pairs, $D = \{A_i, T_i\}_{i=1}^{N}$, where A and T separately denote audio and text modalities.

3.2.1. Audio encoder

For audio modality, we initially extract 40-dimensional MFCCs from raw audio signals as the input of audio encoder, since MFCCs are a typical kind of spectral features which are extracted in terms of the characteristics of human audition, making them much more suited to audio emotion recognition [46]. Each utterance is divided into different frames with a 25 ms width by using a Hamming window. The frame stride is set to 10 ms. Then, we design an audio encoder to obtain the discriminative audio feature representations, as described below.

Suppose the extracted MFCC features are denoted as $A = \{a_1, ..., a_i, ..., a_n\} \in \mathrm{R}^{d_A \times n}$, in which d_A is the dimension of MFCC features, n is the number of audio frames. As shown in Fig. 1, the audio encoder contains two components: a two-layer Bi-LSTM network and a local intra-attention network. These two components have different effects in the encoding procedure. The used two-layer Bi-LSTM network encodes the contextual information by iteratively updating its hidden states for capturing the temporal dependencies within the extracted MFCC features. The output h^A of Bi-LSTM network for the audio encoder is defined as:

$$h^A = \text{Bi-LSTM}(A). \tag{1}$$

The used local intra-attention network aims to focus on learning emotional salient features of audio sequences by attending to important parts in input data with a weighting scheme. To this end, the Multi-head

Self-attention (MHSA) mechanism [73] is adopted. It aims to model internal dependencies between data elements at different positions in the learned feature representations, thereby enhancing the attention weights of important feature parts in input data. The adopted MHSA aggregates useful information from the output h^A of the used Bi-LSTM network to obtain the discriminating feature representations of audio signals as the whole output of the audio encoder, denoted as H^A. More specially, the used MHSA mechanism utilizes linear projections to calculate query (Q), key (K), and value (V) with the corresponding dimension d_q, d_k and d_v, respectively. The attention matrix of its output is computed as follows:

$$\mathrm{Att}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2}$$

MHSA enables the model to jointly concentrate on feature information from different representation subspaces at various locations. In particular, self-attention carries out different linear projections of queries, keys and values for t times. Then, the attention function is executed in parallel to generate output values with the corresponding dimension of $\frac{d_v}{t}$. Finally, these values are concatenated and projected again to produce final feature representations. The process of MHSA is expressed as:

$$\text{Multi-Head}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_t)W^O, \tag{3}$$

where the used i-th Scaled Dot-Product Attention (SDPA) module is represented by $\mathrm{head}_i = \mathrm{Att}(QW_i^Q,\ KW_i^K,\ VW_i^V)$, and $W_i^Q \in R^{d_k/h}$, $W_i^K \in R^{d_k/h}$, $W_i^V \in R^{d_v/h}$ are the weight matrices of parallel attention, separately. W^O is the weight matrix with the dimension d_o in the linear output function.

3.2.2. Text encoder

For text modality, we use the BERT model [67] to extract text feature representations. BERT can learns left and right contextual information through the Transformer module [73] to obtain a deep contextual semantic embedding of a particular word. Specially, we fine-tune the pre-trained BERT model to achieve a 768-dimensional text feature vector h^T, which can be expressed as

$$h^T = \mathrm{BERT}(p_1,\ldots,p_i,\ldots,p_m), \tag{4}$$

where p_i is the i-th word, and m is the number of words.

After extracting text feature representations h^T with contextual information, similar to the audio encoder, we feed it into the used local intra-attention network consisting of the MHSA mechanism to further learn emotional salient features in the text sequences as the output of the text encoder, denoted as H^T.

3.3. Cross-modal attention block for cross-modal fusion

In the Cross-modal Attention (CMA) block, we design a cross-modal attention strategy to learn cross-modal interaction information so as to effectively capture shared feature representations across audio and text modalities, as described below.

Firstly, we project the extracted audio and text feature representations into the same space by a 1D-CNN operation:

$$\widetilde{H}\{A,T\} = \mathrm{Conv1D}\left(H_{\{A,T\}}, k_{\{A,T\}}\right), \tag{5}$$

where $k_{\{A,T\}}$ is the size of the convolution kernel of 1D-CNN for audio and text modalities $\{A,T\}$. Here, the used 1D-CNN has two effects. First, 1D-CNN aims to capture local contextual information from the extracted audio or text feature representations. Second, 1D-CNN can map the extracted audio and text feature representations into the same subspace with a dimension d. In this case, the dot product operations between audio and text feature representations can be performed in the next step.

To explore the interactive dynamics information across unaligned audio and text modalities, the cross-modal attention mechanism is designed to capture cross-modal interactions. Here, when passing information from audio modality to text modality, such interaction is denoted as "$A\rightarrow T$", otherwise "$T\rightarrow A$". The audio and text feature representations projected by 1D-CNN are denoted as $\widetilde{H}_A$ and $\widetilde{H}_T$, respectively.

To learn the interaction information across audio and text modalities, we initially transform the extracted feature representations from these two modalities into the query (Q), key (K), and value (V) by using the following linear projections:

$$Q_l = W_l^Q\widetilde{H}_l, \tag{6}$$

$$K_l = W_l^K\widetilde{H}_l, \tag{7}$$

$$V_l = W_l^V\widetilde{H}_l, \tag{8}$$

where $Q_l, K_l, V_l \in R^{S_l\times d}$ separately denote the query, key and value of the corresponding feature representations $\widetilde{H}_l$for input modality l ($l \in \{A,T\}$). $W_l^Q, W_l^K, W_l^V \in R^{d\times d}$ separately represent the corresponding weight matrices, and d is feature dimension. Then, we compute the dot product of the query and key of audio and text modalities, and leverage the Softmax function to scale and normalize the computed results row-wisely for producing the attention weights. Finally, we aggregate each feature representation by means of multiplying the corresponding weights to obtain the interaction information across two modalities.

3.3.1. Cross-modal interaction from audio to text

To learn the interaction information from audio modality (A) to text modality (T), we employ h heads for cross-modal attention with the query (Q_T), key (K_A), and value (V_A). In this case, we can perform cross-modal interaction for $A\rightarrow T$, thereby interactively learning high-level textual feature representations with the aid of the extracted audio feature representations. This process is denoted as follows:

$$\mathrm{Att}_{A\rightarrow T}\left(\widetilde{H}_A,\widetilde{H}_T\right) = \mathrm{softmax}\left(\frac{Q_TK_A^T}{\sqrt{d_k}}\right)V_A. \tag{9}$$

Then, the obtained results of h heads are concatenated and projected as follows:

$$M_{A\rightarrow T}\left(\widetilde{H}_A,\widetilde{H}_T\right) = \mathrm{Concat}(\mathrm{Att}_{A\rightarrow T}(1),\cdots,\mathrm{Att}_{A\rightarrow T}(i))W, \tag{10}$$

where $\mathrm{Att}_{A\rightarrow T}(i)$ means the i-th ($i \in [1,h]$) cross-modal attention, and W is the weight matrix. Finally, we apply residual connections and layer normalization to the encoding features $\widetilde{H}_A$ and the output is expressed as:

$$CM_{A\rightarrow T} = \mathrm{LayerNorm}\left(\widetilde{H}_A + M_{A\rightarrow T}\left(\widetilde{H}_A,\widetilde{H}_T\right)\right), \tag{11}$$

where $CM_{A\rightarrow T}$ denotes the output feature representations of cross-modal interaction for $A\rightarrow T$. In this case, $CM_{A\rightarrow T}$ can interactively learn the complementary information between audio and text modalities, and thus capture cross-modal dynamics information across two modalities.

3.3.2. Cross-modal interaction from text to audio

Similarly, we use h heads for cross-modal attention to learn interaction information from text modality to audio modality. This process is denoted as follows:

$$\mathrm{Att}_{T\rightarrow A}\left(\widetilde{H}_A,\widetilde{H}_T\right) = \mathrm{softmax}\left(\frac{Q_AK_T^T}{\sqrt{d_k}}\right)V_T. \tag{12}$$

Then, we concatenate and project the achieved results of h heads as:

Table 1
Recognition performance (%) of baseline and our methods.

Methods	IEMOCAP			MELD		
	WAR	UAR	F1-score	WAR	UAR	wF1-score
Audio	55.52	56.98	56.06	45.56	20.19	40.50
Text	68.57	69.54	68.96	58.51	31.38	54.88
EF (baseline)	69.62	70.23	69.71	61.23	33.52	56.96
LF (baseline)	70.26	71.66	70.06	61.15	35.73	58.46
Ours	**73.33**	**74.20**	**73.66**	**62.87**	**37.91**	**60.22**

Table 2
Performance (%) comparisons of existing methods on IEMOCAP dataset.

Methods	Year	WAR	UAR
SAFF [69]	2018	72.10	71.90
Att-alignment [79]	2019	70.40	69.50
CMDM-fully [74]	2020	70.30	68.60
CMDM-semi [74]	2020	72.60	72.10
IA-MMTF [75]	2022	71.96	72.49
Ours	2023	**73.33**	**74.20**

Table 3
Performance (%) comparisons of existing methods on MELD dataset.

Methods	Year	wF1-score
DialogueRNN [77]	2019	55.90
ConGCN [78]	2019	59.40
CMDM-fully [74]	2020	56.10
CMDM-semi [74]	2020	57.10
AGHMN [76]	2020	58.10
IA-MMTF [75]	2022	48.96
HFGCN [72]	2022	59.71
Ours	2023	**60.22**

$$M_{T\to A}\left(\widetilde{H}_A,\widetilde{H}_T\right) = \mathrm{Concat}\left(\mathrm{Att}_{T\to A}(1),\cdots,\mathrm{Att}_{T\to A}(i)\right)W, \tag{13}$$

where $\mathrm{Att}_{T\to A}(i)$ represents the i-th $(i\in[1,h])$ cross-modal attention. Finally, the residual connections and layer normalization are applied to the encoding features $\widetilde{H}_T$ and the output is defined as:

$$CM_{T\to A} = \mathrm{LayerNorm}\left(\widetilde{H}_T + M_{T\to A}\left(\widetilde{H}_A,\widetilde{H}_T\right)\right), \tag{14}$$

where $CM_{T\to A}$ is the output feature representations of cross-modal interaction for $T\to A$.

3.4. *Multimodal emotion classification block for emotion recognition*

We obtain the shared feature representations by means of directly concatenating two cross-modal learned feature representations, *i.e.*, $I^{fusion} = [CM_{A\to T}, CM_{T\to A}]$. Then, the designed MEC block employs the global inter-modal attention mechanism to further enhance the shared feature representations that contribute most to MER. A residual connection is also used to reduce information loss after performing the global inter-modal attention operation, as described below.

Given a trainable parameter $U\in R^{2d}$, the weights and outputs are calculated by:

$$\alpha_i = e^{UI_i^{fusion}} \Big/ \sum_{i=1} e^{UI_i^{fusion}} \tag{15}$$

In this way, we can obtain the final shared feature representations P across audio and text modalities, as expressed as:

$$P = \sum_{i=1} \alpha_i I_i^{fusion}. \tag{16}$$

Then, the fused audio-text feature representations P_j for j-th sample are fed into a linear FC layer and a Softmax output layer to produce the final prediction vector y_j^{pred}. Then, the categorical cross-entropy loss is adopted as the objective function to minimize the negative log-likelihood, as expressed as:

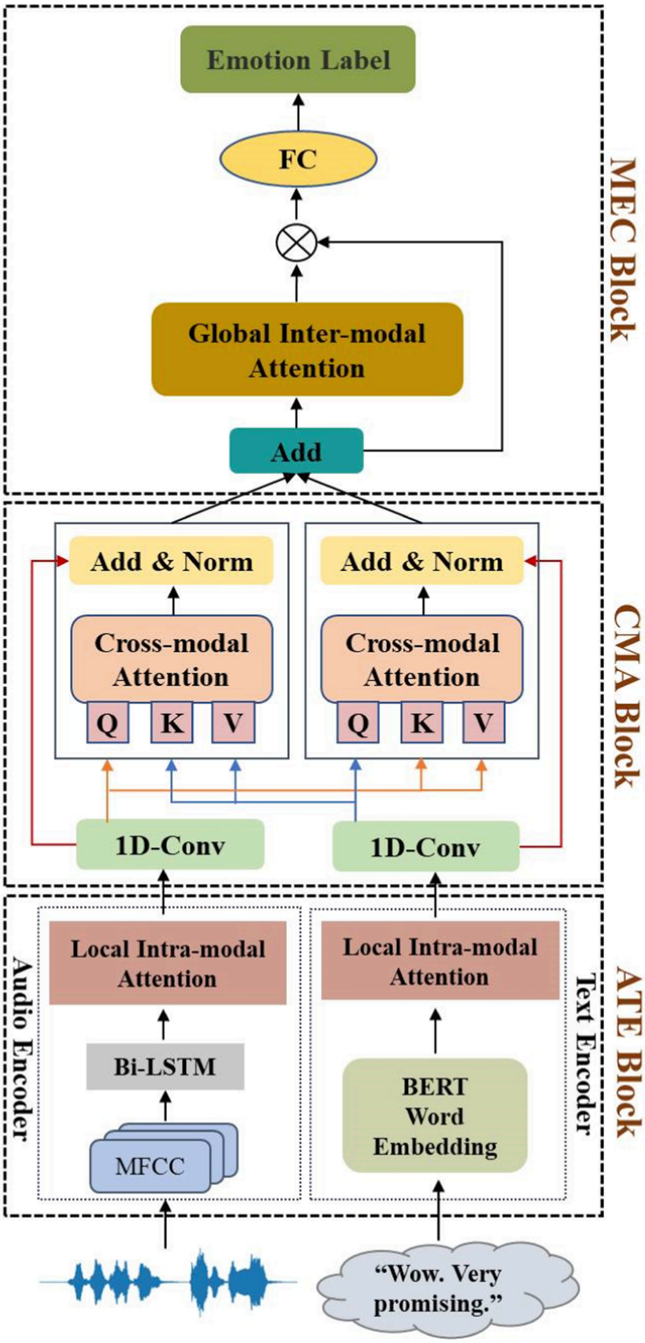


Fig. 1. The framework of our proposed MER-HAN.

$$y_j^{pred} = \mathrm{softmax}(W_0P_j + b_0) \tag{17}$$

$$\mathrm{loss} = -\sum_j y_j \log y_j^{pred} \tag{18}$$

where $y_j (j\in[1,2,\ldots,C])$ means the predicted emotion labels, C is the total number of emotional categories, W_0 represents the weight vectors, and b_0 is the bias.

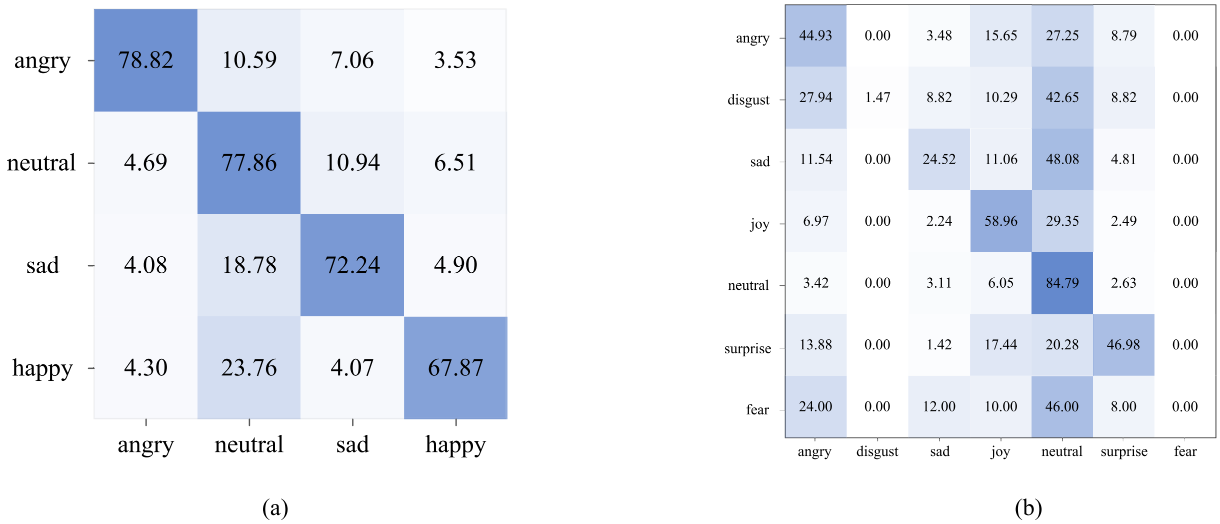


Fig. 2. Confusion matrices of recognition results obtained with MER-HAN when MER-HAN performs best on two datasets: (a) IEMOCAP, (b) MELD.

Table 4
Recognition performance (%) in an ablation study on IEMOCAP dataset.

Methods	WAR	UAR	F1-score
Audio without local intra-modality attention	50.60	52.18	51.05
Audio without local intra-modality attention	50.60	52.18	51.05
Audio with local intra-modality attention	55.52	56.98	56.06
Text without local intra-modality attention	67.43	67.72	66.61
Text with local intra-modality attention	68.57	69.54	68.96
MER-HAN without CMA block	70.34	71.87	70.56
MER-HAN without global inter-modality attention	71.21	71.43	71.31
MER-HAN	**73.33**	**74.20**	**73.66**

Table 5
Recognition performance (%) in an ablation study on MELD dataset.

Methods	WAR	UAR	wF1-score
udio without local intra-modality attention	44.06	18.71	38.43
Audio without local intra-modality attention	44.06	18.71	38.43
Audio with local intra-modality attention	45.56	20.19	40.50
Text without local intra-modality attention	57.62	27.16	50.49
Text with local intra-modality attention	58.51	31.38	54.88
MER-HAN without CMA block	59.08	35.68	56.53
MER-HAN without global inter-modality attention	60.46	37.15	58.30
MER-HAN	**62.87**	**37.91**	**60.22**

4. Experiment study

In this section, we employ two typical public datasets such as IEMOCAP [36] and MELD [37] to conduct MER experiments so as to evaluate the effectiveness of our proposed MER-HAN method on MER tasks, as described below.

4.1. Database

IEMOCAP: The IEMOCAP dataset [36] consists of five sessions performed by 10 unique speakers. Each session contains the dialogs of two speakers, *i.e.*, a male and a female. Each conversation in dialogs includes video, audio, motion capture recordings, and text transcriptions. The language in dialogs in all videos is English. In this work, we utilize audio and text transcriptions to implement MER tasks. Following in [69,74,75], we conduct speaker-independent MER experiments. In particular, the first four sessions are used as the training set, while the last session is adopted as the testing set. The IEMOCAP dataset is categorized into four common emotion categories, namely angry (1,103), sad (1,084), happy (1,636), and neutral (1,708). The happy data is a combination of happy and excitement.

MELD: the MELD dataset [37] is a multi-party emotional conversational database consisting of video, audio and text transcriptions. Each conversation in dialogs is collected from the Friends TV series with multiple speakers. This dataset contains 1,433 dialogs in total from 407 speakers. It is divided into fixed-size three parts with a certain number of dialogs: the training set (1,039), the validation set (114), and the testing set (280). Seven basic emotion categories, including angry, disgust, sad, joy, neutral, surprise and fear, are included. Following in [37,76], we present the same experiment settings.

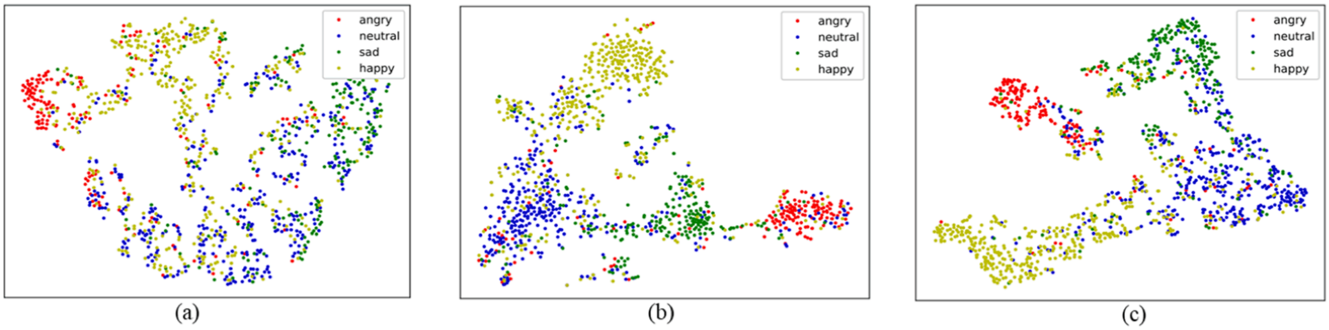


Fig. 3. Feature visualization with t-SNE on IEMOCAP dataset: (a) audio encoder, (b) text encoder, (c) MER-HAN. Different colors represent different emotions.

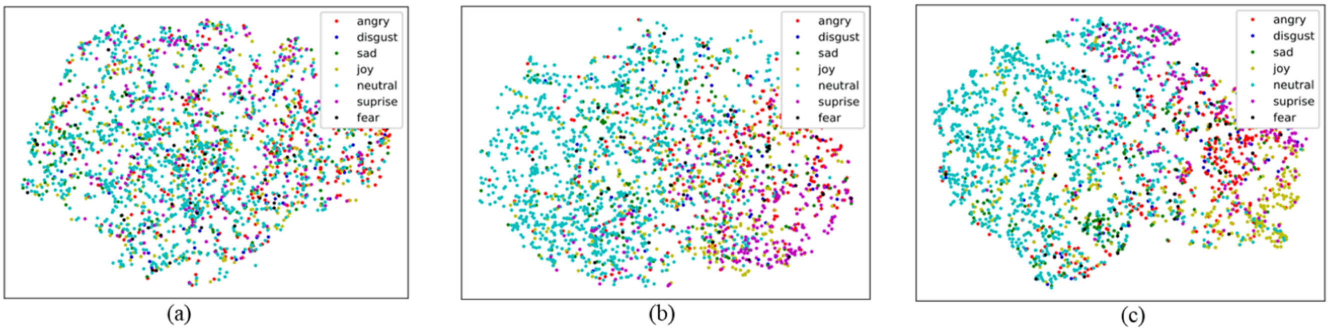


Fig. 4. Feature visualization with t-SNE on MELD dataset: (a) audio encoder, (b) text encoder, (c) MER-HAN. Different colors represent different emotions.

4.2. Experimental setup

4.2.1. Evaluation metrics

We adopt typical Weighted Average Recall (WAR), Unweighted Average Recall (UAR) and F1-score to evaluate the performances of all used methods on MER tasks. Once we separately obtain "Recall" and "Precision" based on the confusion matrices, we can easily figure out WAR, UAR and F1-score. More specially, WAR represents the weighted average accuracy of different emotion categories where the weights are proportional to the number of samples in each category, which is expressed as:

$$\mathrm{WAR} = \frac{\sum_{j=1}^{C} N_j * \mathrm{Recall}_j}{\sum_{j=1}^{C} N_j} \tag{18}$$

where C is the emotion category, and N_j is the sample number for the j-th class.

UAR is the average accuracy of different emotion categories, and defined by:

$$\mathrm{UAR} = \frac{1}{C}\sum_{j=1}^{C} \mathrm{Recall}_j \tag{19}$$

F1-score is a classification metric that represents a harmonic average of precision and recall, and defined by:

$$\mathrm{F1\ -\ score} = \frac{2}{\frac{1}{\mathrm{Precision}} + \frac{1}{\mathrm{Recall}}} \tag{20}$$

When the number of each category in the dataset is definitely unbalanced, the weighted average F1-score (abbreviated as wF1-score), which is computed by means of averaging the support-weighted mean per-class F1-scores, is usually adopted. The wF1-score aims to consider the varying degrees of importance of each category in the dataset.

4.2.2. Implementation details

We implement all used models in the PyTorch framework configured by a NVIDIA GPU 3090 with a 24 GB memory. For audio and text modalities, the processed sequence length is uniformly set to the maximum of all samples, and the shorter sequence length is padded with zero values. In the ATE block, the number of MHSA is set to 4. In the CMA block, the number of cross-modal attention block is set to 1, and the number of MHSA is set to 4. The neuron number of Bi-LSTM in the audio encoder is 256. For training deep learning models, we use an Adam optimizer with a learning rate of 0.0001 and the cross-entropy loss function for MER. For the single modality, the batch size is 32 and the epoch number is 64.

4.3. Results and analysis

Baseline methods: For multimodal data fusion, two typical methods including Early Fusion (EF) and Late Fusion (LF) are employed as baseline methods for a comparison. For EF, we directly concatenate the outputs of audio encoder and text encoder, and then feed them into a FC layer for emotion classification. For LF, we first employ the outputs of audio encoder and text encoder as the input of a FC layer to individually obtain the corresponding audio emotion classification results y_A and text emotion classification results y_T. Then, we use a weighted approach to fuse these two modal classification results, as defined as

$$y_{LF} = \mu_A y_A + \mu_T y_T, \tag{21}$$

where y_{LF} denotes the final predicted emotion labels, and the weight parameters $\mu_A + \mu_F = 1$. In this work, we use $\mu_A = 0.4$, and $\mu_F = 0.6$ for its best performance.

Table 1 shows the recognition performance of different methods. It is noted that on MELD dataset, we report the performance of wF1-score except for WAR and UAR, as done in [37,76]. From the results in Table 1, we have the following two observations:

(1) The proposed MER-HAN separately achieves the highest WAR of 73.33 %, UAR of 74.20 % and F1-score of 73.66 % on IEMOCAP dataset, and WAR of 62.87 %, UAR of 37.91 %, wF1-score of 60.22 % on MELD dataset. In comparison with the single audio and text modality, integrating audio and text cues can clearly achieve higher performance. In particular, our MER-HAN obtains the WAR performance with a 17.51 % and 4.76 % improvement on IEMOCAP dataset over the single audio and text modality, respectively. Likewise, our MER-HAN yields the wF1-score performance with a 19.72 % and 5.34 % improvement on MELD dataset over the single audio and text modality, respectively. This indicates that combining multiple modalities yields better performance than the single modality for emotion recognition.

(2) Compared with the baseline methods, our MER-HAN makes an improvement of about 3 % over EF and LF in terms of WAR, UAR and F1-score on IEMOCAP dataset, and about 2 % over EF and LF in terms of WAR, UAR, and wF1-score on MELD dataset. This demonstrates the validity of the proposed MER-HAN on MER tasks. This is because the proposed MER-HAN employs the hybrid cross-modal attention strategy to learn cross-modal interaction information so as to effectively capture the shared feature representations across audio and text modalities.

Comparing methods: To validate the effectiveness of our proposed MER-HAN method, we compare our method with several previous works which has the same settings as ours, as depicted below.

SAFF: Hazarika *et al.*, [69] developed a Self-Attentive Feature-level Fusion (SAFF) method for MER.

DialogueRNN: Majumder *et al.*,[77] adopted three GRUs with an attention layer to model preceding emotion status of two speakers and global contextual clues in conversations.

Cmdm: Liang *et al.*, [74] presented a MER method of semi-supervised learning model based on Cross-Modality Distribution Matching (CMDM), in which abundant unlabeled data were utilized to promote the MER performance. They found that semi-supervised CMDM (CMDM-semi) outperformed fully-supervised CMDM (CMDM-fully).

ConGCN: Zhang *et al.*, [78] provided a MER method based on Graph-based Convolutional Neural Network towards Conversations, called ConGCN, in which both context-sensitive and speaker-sensitive clues were captured to detect emotion.

Att-alignment: Xu *et al.*,[79] employed the attention mechanism to capture the alignment clues between audio frames and text words, resulting in more accurate shared feature representations for MER.

Aghmn: Jiao *et al.*, [76] presented an Attention Gated Hierarchical Memory Network (AGHMN) model to obtain utterance-level features and the contextual clues for real-time MER tasks.

IA-MMTF: Guo *et al.*, [75] developed an Implicitly Aligned Multimodal Transformer Fusion (IA-MMTF) method for MER. The IA-MMTF method learned the complementarity between audio and text modalities with their inter-modality guidance for each other.

HFGCN: Tang *et al.*, [72] provided a MER method based on Hierarchical Fusion Graph Convolutional Network (HFGCN), which aimed to learn more discriminative shared feature representations by means of considering the inter-modality dependencies.

Tables 2 and 3 separately present the performance comparisons of different methods on IEMOCAP and MELD datasets. More specially, on IEMOCAP dataset WAR and UAR are usually adopted for a comparison, whereas on MELD dataset wF1-score is usually employed for a comparison. From the results in Tables 2 and 3, we can see that compared with other MER methods, the proposed MER-HAN exhibits better performance on two datasets in terms of WAR, UAR and wF1-score. This shows the superiority of our proposed method to other comparing methods. Specially, our MER-HAN is capable of capturing the intra-modal and inter-modal emotional information simultaneously between audio and text modalities.

To show the recognition performance of each emotion, Fig. 2 presents the confusion matrices of recognition results obtained with our MER-HAN when MER-HAN performs best on IEMOCAP and MELD datasets. As depicted in Fig. 2 (a), on IEMOCAP dataset MER-HAN yields an accuracy of 78.82 % for angry, 77.86 % for neutral, 72.24 % for sad, 67.87 % for happy, respectively. Likewise, it can be seen from Fig. 2(b) that on MELD dataset MER-HAN gives an accuracy of 44.93 % for angry, 24.52 % for sad, 58.96 % for joy, 84.79 % for neural, and 46.98 % for surprise, respectively. Nevertheless, MER-HAN identifies disgust and fear with the worst accuracy of less than 2 % on MELD dataset. This shows that disgust and fear on MELD dataset are much more difficult to be classified than other emotions, since they are easily confused with neutral.

4.4. Ablation study

It is noted that our proposed MER-HAN adopts hybrid attention networks including the local intra-modality attention in the ATE block, the cross-modal attention in the CMA block, and the global inter-modality attention in the MEC block. To verify the effectiveness of each attention block, we conduct ablation studies. Tables 4 and 5 individually show the recognition results in an ablation study on IEMOCAP and MELD datasets.

1) *Effects of local intra-modal attention:* When using the local intra-modality attention for feature learning of audio and text modality, audio encoding features and text encoding features are individually input into a linear FC layer to obtain the single-modal emotion recognition results. By contrast, in the absence of the local intra-modality attention, audio and text encoding features are fed into a linear FC layer for emotion classification. The results in Table 4 on IEMOCAP dataset show that after integrating the local intra-attention, audio emotion classification results are improved by a WAR of 4.92 %, a UAR of 4.8 %, and a F1-score of 5.01 %, respectively. Text emotion classification results on IEMOCAP dataset are increased by a WAR of 1.14 %, a UAR of 1.82 %, and a F1-score of 2.35 %, respectively. Similarly, the results in Table 5 on MELD dataset demonstrate that compared with the obtained results without the local intra-modality attention, audio emotion classification results with the local intra-modality attention make an improvement by a WAR of 1.5 %, a UAR of 1.48 %, and a wF1-score of 2.07 %, respectively. Text emotion classification results with the local intra-modality attention promote the performance by a WAR of 0.89 %, a UAR of 4.22 %, and a wF1-score of 4.39 %, respectively. This indicates that the local intra-modality attention can help the ATE block learn more salient emotional features for emotion recognition.
2) *Effects of CMA block:* To explore the effects of the used CMA block, we compare the different recognition results obtained with or without the CMA block. When implementing our method without the CMA block, we directly concatenate audio encoding features and text encoding features in feature-level as an input of a linear classification network. As shown in Table 4, in comparison with the reported results without the CMA block, the obtained results by MER-HAN equipped with the CMA block on IEMOCAP dataset are increased by a WAR of 2.99 %, a UAR of 2.33 %, and a F1-score of 3.10 %, respectively. Likewise, it can be seen from Table 5 that the achieved results by MER-HAN equipped with the CMA block on MELD dataset are improved by a WAR of 3.79 %, a UAR of 2.23 %, and a wF1-score of 3.69 %, respectively. This shows that integrating the CMA Block in our method is beneficial to improve the performance of our model.
3) *Effects of global inter-modal attention:* In the absence of the global inter-modality attention in the MEC block, the shared feature representations across audio and text modalities are fed directly into a linear FC layer for emotion classification. The results in Table 4 on IEMOCAP dataset indicate that compared with our designed MER-HAN, in the absence of the global inter-modality attention, there is a performance decrease of 2.12 %, 2.77 % and 2.35 % on WAR, UAR and F1-score metrics, respectively. Likewise, from the results in Table 5 on MELD dataset, we can see that compared with the used MER-HAN without the global inter-modality attention, our designed MER-HAN makes an improvement by a WAR of 2.41 %, a UAR of 0.76 %, and a wF1-score of 1.92 %, respectively. This demonstrates that the used global inter-modality attention focusing on the salient parts of the shared feature representations across audio and text modalities can effectively promote the performance of MER.

4.5. Feature visualization

In order to intuitively present the advantage of our proposed MER-HAN, we use the t-SNE [80] method to visualize the learned emotional features from the single audio encoder, text encoder and the proposed MER-HAN.

Figs. 3 and 4 shows the results of feature visualization for each emotion on IEMOCAP and MELD datasets. From the visualization results in Figs. 3 and 4, it can be clearly observed that the emotional label distributions of the proposed MER-HAN on IEMOCAP and MELD datasets are more discriminative compared with the single audio and text modality, since the proposed MER-HAN produce better clustering effects for each emotion than the single audio and text modality, in which different emotion categories are separated as far as possible. For instance, on IEMOCAP dataset the emotion label distributions of the single audio and text modality show that four emotion categories are easily confused for each other. By contrast, in the MER-HAN distribution, various emotions can be much better separated. This finding demonstrates that the proposed MER-HAN method can make full use of hybrid attention networks to jointly capture high-level shared feature representations across audio and text modalities for MER.

5. Conclusion

In this paper, we propose a new MER-HAN framework for emotion classification based on audio and text cues. The main purpose of the proposed MER-HAN approach is to integrate audio and text modalities by using hybrid attention networks, which aims to jointly learn discriminative shared feature representations across audio and text modalities for emotion classification. Experiment results on the popular IEMOCAP and MELD datasets demonstrate the advantage of the proposed MER-HAN method compared with other used methods. In our future work, we will combine much more modalities, such as visual and physiological signals such as electroencephalography (EEG) [81], to further improve the performance of automatic MER.

CRediT authorship contribution statement

Shiqing Zhang: Conceptualization, Writing - original draft. **Yijiao Yang:** Software. **Chen Chen:** Software. **Ruixin Liu:** Software. **Xin Tao:** Software. **Wenping Guo:** Software. **Yicheng Xu:** Software. **Xiaoming Zhao:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC), and Zhejiang Provincial National Natural Science Foundation of China under Grant No. 62276180, 61976149, LZ20F020002, and LQ21F020002, and the Humanities and Social Science Project of the Chinese Ministry of Education under Grant No. 20YJAZH033.

References

[1] N. Perveen, D. Roy, K.M. Chalavadi, Facial expression recognition in videos using dynamic kernels, IEEE Trans. Image Process. 29 (2020) 8316–8325.

[2] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, M.R. Wrobel, Emotion recognition and its applications, in: Human-Computer Systems Interaction: Backgrounds and Applications, Vol. 3, Springer, 2014, pp. 51–62.

[3] M.K. Chowdary, T.N. Nguyen, D.J. Hemanth, Deep learning-based facial emotion recognition for human–computer interaction applications, Neural Comput. Appl. (2021) 1–18.

[4] D. Wang, X. Zhao, Affective video recommender systems: a survey, Front. Neurosci. 16 (2022), 984404.

[5] P. Sarkar, A. Etemad, Self-supervised ECG representation learning for emotion recognition, IEEE Trans. Affect. Comput. 13 (2022) 1541–1554.

[6] S. Parthasarathy, C. Busso, Semi-supervised speech emotion recognition with ladder networks, IEEE/ACM Trans. Audio, Speech, Lang. Process. 28 (2020) 2697–2709.

[7] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, B. Schuller, Attention-enhanced connectionist temporal classification for discrete speech emotion recognition, Interspeech (2019) 206–210.

[8] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio–visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 28 (2017) 3030–3043.

[9] J. Deng, F. Ren, A survey of textual emotion recognition and its challenges, IEEE Trans. Affect. Comput. (2021), https://doi.org/10.1109/TAFFC.2021.3053275.

[10] G. Castellano, L. Kessous, G. Caridakis, Emotion recognition through multiple modalities: face, body gesture, speech, in: Affect and Emotion in Human-Computer Interaction, Springer, Berlin, Heidelberg, 2008, pp. 92–103.

[11] S. Yoon, S. Dey, H. Lee, K. Jung, Attentive modality hopping mechanism for speech emotion recognition, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Barcelona, Spain, 2020, pp. 3362–3366.

[12] Y. Qian, Z. Chen, S. Wang, Audio-visual deep neural network for robust person verification, IEEE/ACM Trans. Audio, Speech, Lang. Process. 29 (2021) 1079–1092.

[13] J. Huang, J. Tao, B. Liu, Z. Lian, M. Niu, Multimodal transformer fusion for continuous emotion recognition, in: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 3507–3511.

[14] M. Sharafi, M. Yazdchi, R. Rasti, F. Nasimi, A novel spatio-temporal convolutional neural framework for multimodal emotion recognition, Biomed. Signal Process. Control 78 (2022), 103970.

[15] S. Yoon, S. Byun, S. Dey, K. Jung, Speech emotion recognition using multi-hop attention mechanism, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Brighton, UK, 2019, pp. 2822–2826.

[16] L. Gao, H. Li, Z. Liu, Z. Liu, L. Wan, W. Feng, RNN-transducer based Chinese sign language recognition, Neurocomputing 434 (2021) 45–54.

[17] N.J. Shoumy, L.-M. Ang, K.P. Seng, D.M. Rahaman, T. Zia, Multimodal big data affective analytics: a comprehensive survey using text, audio, visual and physiological signals, J. Netw. Comput. Appl. 149 (2020), 102447.

[18] S.K. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Comput. Surv. 47 (2015) 1–36.

[19] S. Poria, H. Peng, A. Hussain, N. Howard, E. Cambria, Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis, Neurocomputing 261 (2017) 217–230.

[20] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Signal Process. 11 (2017) 1301–1309.

[21] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: 2016 IEEE 16th international conference on data mining (ICDM), IEEE, Barcelona, Spain, 2016, pp. 439–448.

[22] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, I. Marsic, Multimodal affective analysis using hierarchical attention strategy with word-level alignment, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, NIH Public Access, Athens, Greece, 2018, pp. 2225.

[23] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, X. Chen, Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild, in: Proceedings of the 16th International Conference on multimodal interaction, Istanbul, Turkey, 2014, pp. 494–501.

[24] S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 112–118.

[25] J. Chen, Z. Chen, Z. Chi, H. Fu, Emotion recognition in the wild with feature fusion and multiple kernel learning, in: Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul Turkey, 2014, pp. 508–513.

[26] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio–visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 28 (2018) 3030–3043.

[27] A.I. Middya, B. Nag, S. Roy, Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities, Knowl.-Based Syst. 244 (2022), 108580.

[28] Y. Fu, S. Okada, L. Wang, L. Guo, Y. Song, J. Liu, J. Dang, Context-and knowledge-aware graph convolutional network for multimodal emotion recognition, IEEE Multimedia 29 (2022) 91–100.

[29] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, NIH Public Access, 2019, pp. 6558.

[30] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[31] P.P. Liang, Z. Liu, A.B. Zadeh, L.-P. Morency, Multimodal language analysis with recurrent multistage fusion, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 150–161.

[32] H. Chen, D. Jiang, H. Sahli, Transformer encoder with multi-modal multi-head attention for continuous affect recognition, IEEE Trans. Multimedia 23 (2021) 4171–4183.

[33] D.S. Chauhan, M.S. Akhtar, A. Ekbal, P. Bhattacharyya, Context-aware interactive attention for multi-modal sentiment and emotion analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5647-5657.

[34] J. Tang, D. Liu, X. Jin, Y. Peng, Q. Zhao, Y. Ding, W. Kong, BAFN: bi-direction attention based fusion network for multimodal sentiment analysis, IEEE Trans. Circuits Syst. Video Technol. 33 (2023) 1966–1978.

[35] M.G. Huddar, S.S. Sannakki, V.S. Rajpurohit, Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM, Multimed. Tools Appl. 80 (2021) 13059–13076.

[36] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.

[37] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: a multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 527–536.

[38] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recogn. 44 (2011) 572–587.

[39] C.-N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artif. Intell. Rev. 43 (2015) 155–177.

[40] N. Alswaidan, M.E.B. Menai, A survey of state-of-the-art approaches for emotion recognition in text, Knowl. Inf. Syst. 62 (2020) 2937–2987.

[41] F.A. Acheampong, H. Nunoo-Mensah, W. Chen, Transformer models for text-based emotion detection: a review of BERT-based approaches, Artif. Intell. Rev. 54 (2021) 5789–5829.

[42] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, B.W. Schuller, Survey of deep representation learning for speech emotion recognition, IEEE Trans. Affect. Comput. (2021), https://doi.org/10.1109/TAFFC.2021.3114365.

[43] L. Ten Bosch, Emotions, speech and the ASR framework, Speech Comm. 40 (2003) 213–225.

[44] J. Sundberg, S. Patel, E. Bjorkner, K.R. Scherer, Interdependencies among voice source parameters in emotional speech, IEEE Trans. Affect. Comput. 2 (2011) 162–174.

[45] Y. Sun, G. Wen, J. Wang, Weighted spectral features based on local Hu moments for speech emotion recognition, Biomed. Signal Process. Control 18 (2015) 80–90.

[46] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using Fourier parameters, IEEE Trans. Affect. Comput. 6 (2015) 69–75.

[47] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C.A. Müller, S.S. Narayanan, The INTERSPEECH 2010 paralinguistic challenge, INTERSPEECH Makuhari, Chiba, Japan, 2010, pp. 2794–2797.

[48] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, INTERSPEECH-2013Lyon, France, 2013, pp. 148–152.

[49] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2016) 190–202.

[50] D. Li, Y. Zhou, Z. Wang, D. Gao, Exploiting the potentialities of features for speech emotion recognition, Inf. Sci. 548 (2021) 328–343.

[51] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[52] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.

[53] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, Lake Tahoe, Nevada, United States, 2012, pp. 1097–1105.

[54] J.L. Elman, Finding structure in time, Cognit. Sci. 14 (1990) 179–211.

[55] S. Hochreiter, J.J.N.C. Schmidhuber, Long short-term memory, 9 (1997) 1735–1780.

[56] S. Ottl, S. Amiriparian, M. Gerczuk, V. Karas, B. Schuller, Group-level speech emotion recognition utilising deep spectrum features, in: Proceedings of the 2020 International Conference on Multimodal Interaction, ACM, Utrecht, the Netherlands, 2020, pp. 821–826.

[57] S. Zhang, X. Zhao, Q. Tian, Spontaneous speech emotion recognition using multiscale deep convolutional LSTM, IEEE Trans. Affect. Comput. 13 (2022) 680–688.

[58] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. (CSUR) 34 (2002) 1–47.

[59] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[60] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by latent semantic analysis, J. Assoc. Inf. Sci. Technol. 41 (1990) 391–407.

[61] M. Yang, W. Tu, J. Wang, F. Xu, X. Chen, Attention based LSTM for target dependent sentiment classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2017, pp. 5013–5014.

[62] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, ACM, Lake Tahoe, Nevada, USA, 2013, pp. 3111–3119.

[63] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543.

[64] J. Zhou, Y. Lu, H.-N. Dai, H. Wang, H. Xiao, Sentiment analysis of Chinese microblog based on stacked bidirectional LSTM, IEEE Access 7 (2019) 38856–38866.

[65] Y.C. Tan, L.E. Celis, Assessing social and intersectional biases in contextualized word representations, in: Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 2019, pp. 1–12.

[66] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237.

[67] J.D.M.-W.C. Kenton, L.K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[68] P. Kumar, B. Raman, A BERT based dual-channel explainable text emotion recognition system, Neural Netw. 150 (2022) 392–407.

[69] D. Hazarika, S. Gorantla, S. Poria, R. Zimmermann, Self-attentive feature-level fusion for multimodal emotion detection, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, Miami, FL, USA, 2018, pp. 196–201.

[70] S.P. Bayerl, A. Tammewar, K. Riedhammer, G. Riccardi, Detecting emotion carriers by combining acoustic and lexical representations, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, Cartagena, Colombia, 2021, pp. 31–38.

[71] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, Nevada, USA, 2016, pp. 770–778.

[72] S. Tang, Z. Luo, G. Nan, J. Baba, Y. Yoshikawa, H. Ishiguro, Fusion with hierarchical graphs for multimodal emotion recognition, in: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE, Chiang Mai, Thailand, 2022, pp. 1288–1296.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, ACM, Long Beach, CA, USA, 2017, pp. 6000–6010.

[74] J. Liang, R. Li, Q. Jin, Semi-supervised multi-modal emotion recognition with cross-modal distribution matching, in: Proceedings of the 28th ACM International Conference on Multimedia, Seattle WA, USA 2020, pp. 2852–2861.

[75] L. Guo, L. Wang, J. Dang, Y. Fu, J. Liu, S. Ding, Emotion recognition with multimodal transformer fusion framework based on acoustic and lexical information, IEEE Multimedia 29 (2022) 94–103.

[76] W. Jiao, M. Lyu, I. King, Real-time emotion recognition via attention gated hierarchical memory network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 8002–8009.

[77] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 6818–6825.

[78] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, G. Zhou, Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)Macao, China, 2019, pp. 5415–5421.

[79] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, X. Li, Learning alignment for multimodal emotion recognition from speech, in: Proc. Interspeech 2019, Graz, Austria, 2019, pp. 3569–3573.

[80] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[81] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, C.F. Caiafa, A multimodal emotion recognition method based on facial expressions and electroencephalography, Biomed. Signal Process. Control 70 (2021), 103029.