frontiers in Psychology

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/fpsyg.2022.857891

Enterprise Strategic Management From the Perspective of Business Ecosystem Construction Based on Multimodal Emotion Recognition

Wei Bi[1]*, Yongzhen Xie[1], Zheng Dong[2] and Hongshen Li[3]

[1] School of Management, Shandong University, Jinan, China, [2] School of Business Administration, Dongbei University of Finance and Economics, Dalian, China, [3] School of Economics and Management, Shandong Youth University of Political Science, Jinan, China

OPEN ACCESS

Edited by:
Deepak Kumar Jain,
Chongqing University of Posts
and Telecommunications, China

Reviewed by:
Zhiyong Jiang,
Guilin University of Aerospace
Technology, China
Huicheng Liu,
North China Electric Power University,
China

***Correspondence:**
Wei Bi
201990000050@sdu.edu.cn

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Psychology

Received: 19 January 2022
Accepted: 31 January 2022
Published: 03 March 2022

Citation:
Bi W, Xie Y, Dong Z and Li H
(2022) Enterprise Strategic
Management From the Perspective
of Business Ecosystem Construction
Based on Multimodal Emotion
Recognition.
Front. Psychol. 13:857891.
doi: 10.3389/fpsyg.2022.857891

Emotion recognition (ER) is an important part of building an intelligent human-computer interaction system and plays an important role in human-computer interaction. Often, people express their feelings through a variety of symbols, such as words and facial expressions. A business ecosystem is an economic community based on interacting organizations and individuals. Over time, they develop their capabilities and roles together and tend to develop themselves in the direction of one or more central enterprises. This paper aims to study a multimodal ER method based on attention mechanism. It analyzes the current emotional state of consumers and the development direction of enterprises through multi-modal ER of human emotions and analysis of market trends, so as to provide the most appropriate response or plan. This paper firstly describes the related methods of multimodal ER and deep learning in detail, and briefly outlines the meaning of enterprise strategy in the business ecosystem. Then, two datasets, CMU-MOSI and CMU-MOSEI, are selected to design the scheme for multimodal ER based on self-attention mechanism. Through the comparative analysis of the accuracy of single-modal and multi-modal ER, the self-attention mechanism is applied in the experiment. The experimental results show that the average recognition accuracy of happy under multimodal ER reaches 91.5%.

Keywords: multimodal emotion recognition, deep learning, attention mechanism, business ecosystem, enterprise strategic management

INTRODUCTION

Ecosystems have been the hottest topic in theory and practice in the last decade. As an important branch of the ecosystem field, the business ecosystem with unique theoretical attributes has gradually attracted the attention of the people and actively carried out theoretical exploration. Strategic emerging industries are based on major technological breakthroughs and major development needs, and represent the direction of technological innovation as a direction of industrial development. In recent years, with the development of AI, the application of deep learning and the continuous update of computer science, more intelligent terminals have appeared in people's field of vision, bringing people convenient and considerate services. In the process of human development and communication, an important part of it is emotional expression. Emotion recognition (ER) combines speech processing, psychology, standard recognition and video image processing, it can be applied to various fields such as education, transportation and medical care. And emotions can be expressed in different ways. Since single ER often has shortcomings such as insufficient information utilization and low recognition accuracy, more and more researchers have begun to pay attention to multimodal ER.

In this information age, ER has broad prospects for development. It can collect multiple emotional information of people in different backgrounds with the help of computers. Based on this, it can respond appropriately by analyzing people's emotional states and guessing people's inner emotions. In some specific occasions, it can show the effect of killing two birds with one stone. Multimodal ER is complemented by the pooling of information between different modalities, thereby improving the final recognition rate. In the business ecosystem, ER technology can initially control the corporate strategy, while improving the efficiency of the company, it also reduces the cost of human and material resources. At the same time, it can also provide reference for other business ecosystems and provide ideas for enterprises to choose value models in the new business environment.

The innovations of this paper are: (1) The fusion method of multimodal ER is analyzed. Among them, the fusion method based on feature layer can effectively utilize the information between different modes. However, the fusion method adopted by the directly cascaded feature layers is only to splicing the output emotion feature vectors of each modality. In this paper, the attention mechanism is effectively introduced into the process of feature layer fusion according to the research needs, which ensures the accuracy and rationality of the multimodal ER results. (2) A multimodal ER based on self-attention mechanism and neural network is proposed. It uses the self-attention mechanism to analyze the text data with the most abundant and accurate information for the three modal data of the user's text, image and sound.

RELATED WORK

In recent years, with the continuous development of artificial intelligence technology, people hope that computers can serve people with emotion, and ER is one of the key technologies. Rao KP studied the proposed computer vision system that would automatically analyze videos of learners to identify students' comprehension levels from their facial expressions and head movements. In the first stage, facial features are extracted to generate feature vectors using Accelerated Robust Features (SURF). In the second stage, head motion (nodding, shaking, or still) is detected based on nostril tracking. However, his proposed system cannot give feedback in real time (Rao et al., 2019). Wei proposed to use a multimodal strategy to extract emotional features from facial expression images. The basic idea is to combine low-level empirical features and high-level self-learning features into a multimodal feature. The two-dimensional coordinates of facial key points are extracted as low-level empirical features, and high-level self-learning features are extracted by convolutional neural network (CNN). However, there is a lot of redundant data in the implementation of this technology (Wei et al., 2020). Saha T thoroughly investigated the role of emotion in the automatic recognition of DA in task-independent dialog in a multimodal framework (especially audio and text). DL-based multi-task networks have been developed for DAC and ER, incorporating attention to facilitate the fusion of different modalities. An open-source, benchmarked ER multimodal dataset, IEMOCAP, has been manually annotated for its corresponding DA. However, the consumption cost of this network will be relatively large, and the practicability needs to be improved (Saha et al., 2020). Zhao S first briefly introduces widely used emotion representation models and emotion modalities, then he summarizes existing emotion annotation strategies and corresponding computational tasks, describing the main challenges in MER. Furthermore, he proposes some representative methods on representation learning for each emotion modality, feature fusion for different emotion modalities, MER classifier optimization, and MER domain adaptation. However, these methods are not based on reality and lack operability (Zhao et al., 2021). Wang X proposed a two-level attention mechanism with a two-stage multi-task learning (2Att-2Mt) framework for facial emotion estimation only on static images. First, features of corresponding regions (location-level features) are automatically extracted and enhanced by a one-level attention mechanism. A bidirectional recurrent neural network (Bi-RNN) with self-attention (two-level attention) is then used to adaptively make full use of the relational features (hierarchical features) of different layers. However, the shortcoming of this study is the lack of specific case analysis (Wang et al., 2018). The bimodal ER system proposed by Manisha S consists of three main parts, such as audio processing, video processing and data fusion, for detecting human emotions. The fusion of visual information and audio data obtained from two different sources improves the ER rate by providing complementary data. The proposed method aims to classify 7 basic emotions from input videos. He takes audio and image frames from video input to predict a person's final emotion, using an audio-visual dataset that is particularly suitable for studying multimodal emotion expression and perception. But he did not propose specific measures (Manisha et al., 2021). Yoo G This study proposes a system that can identify human emotional states from biological signals, aiming to improve the interaction between humans and computers to achieve effective human-machine capable of intelligent interaction. The proposed method is able to identify six emotional states, such as joy, happiness, fear, anger, despair, and sadness, which are widely used for ER purposes. However, the accuracy of this experiment is not enough and the performance is low (Yoo et al., 2018). Kim J H proposed a multi-level fusion method to combine visual information and physiological signals for ER. For visual information, he proposes serial fusion of two-stage features to enhance the representation of facial expressions in video sequences. He proposes to combine neural aggregation networks with CNN feature maps to strengthen important emotional frameworks. For physiological signals, he proposed a parallel fusion scheme to widen the labeling range of EEG signals. However, the experiments in the article lack a lot of data (Kim and Lee, 2021).

MULTIMODAL EMOTION RECOGNITION AND DEEP LEARNING RELATED METHODS

Multimodal Feature Fusion Method

Although multimodal ER can well overcome the single shortcomings of single-modal ER, it is difficult to process and combine information in different ways. The traditional multimodal information fusion method framework includes data layer fusion, feature layer fusion and decision layer fusion, and the flow charts are shown in **Figures 1–3**, respectively (Min et al., 2018).

Data layer fusion refers to the direct fusion of the collected raw data layers, and the raw data of various emotional information is prepared in advance for data synthesis and analysis. Data layer fusion generally adopts a centralized fusion system for fusion processing.

Feature layer fusion is a fusion that belongs to the middle level. It first extracts effective features from the collected raw information, and then comprehensively analyzes and processes the feature information. The advantage of feature layer fusion is that considerable information compression can be achieved, which is beneficial to the real-time processing of the system. And because the extracted features are directly related to decision analysis, the fusion result can provide the feature information required for decision analysis.

Decision layer fusion is to perform preprocessing and feature processing on the single-modal information one by one, and then obtain the respective classification results through the classifier, and then fuse the respective classification results according to a certain form to obtain the final sentiment classification result. Common decision fusion criteria include: maximum, minimum, sum, average, voting, and product, etc. The fusion of the decision layer is simpler than the fusion of the data layer and the feature layer, because the dimensions of the classification results of each modal are consistent.

Comparing the three recognition frameworks for ER, they have their own advantages. However, practical problems must be considered in practical work in order to choose the best fusion method.

Deep Learning Related Technologies

In multimodal ER, deep neural networks are widely used. Commonly used ones include: long short-term memory network (LSTM), CNN, and attention mechanism. In previous studies,

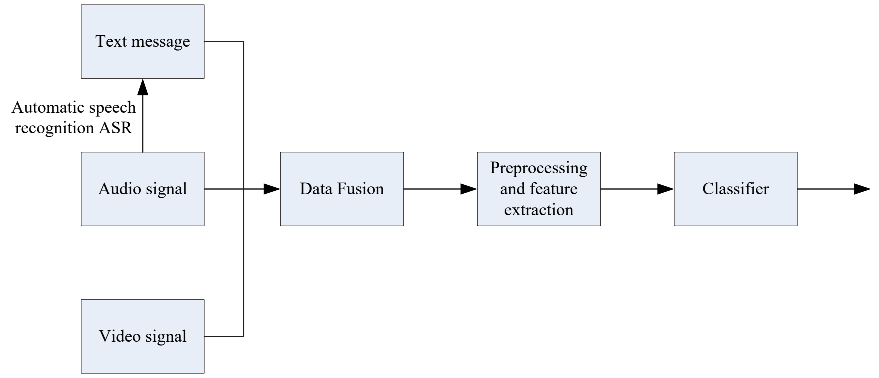
Text message
Automatic speech recognition ASR
Audio signal
Video signal
Data Fusion
Preprocessing and feature extraction
Classifier

FIGURE 1 | Data layer fusion flow chart.

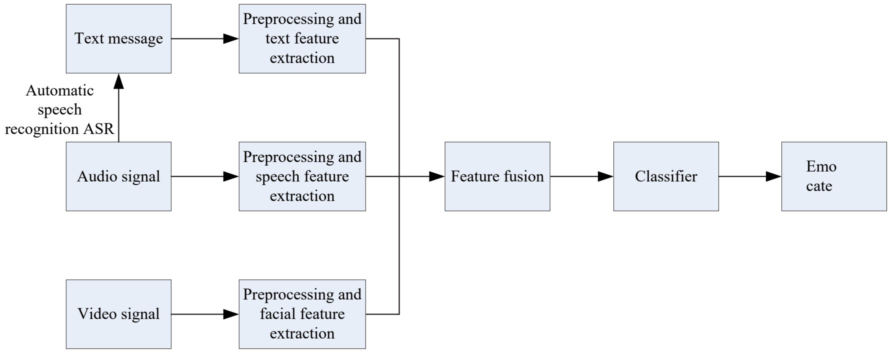
Text message
Automatic speech recognition ASR
Audio signal
Video signal
Preprocessing and text feature extraction
Preprocessing and speech feature extraction
Preprocessing and facial feature extraction
Feature fusion
Classifier
Emo cate

FIGURE 2 | Feature layer fusion flow chart.

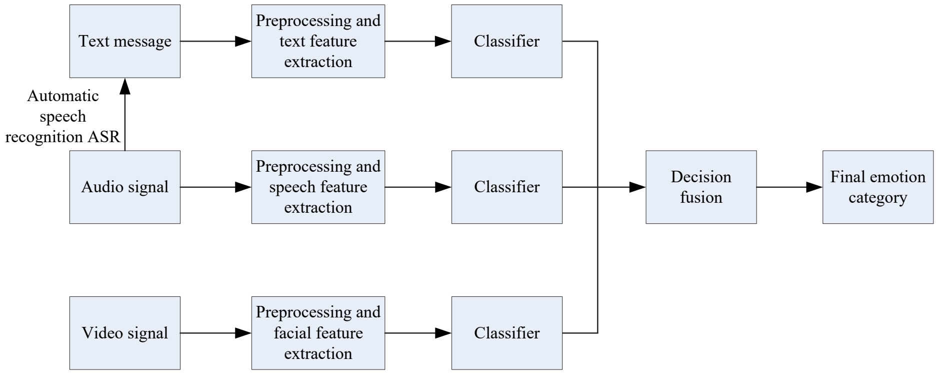
Text message
Preprocessing and text feature extraction
Classifier
Automatic speech recognition ASR
Audio signal
Preprocessing and speech feature extraction
Classifier
Decision fusion
Final emotion category
Video signal
Preprocessing and facial feature extraction
Classifier

FIGURE 3 | Flow chart of decision-making level fusion.

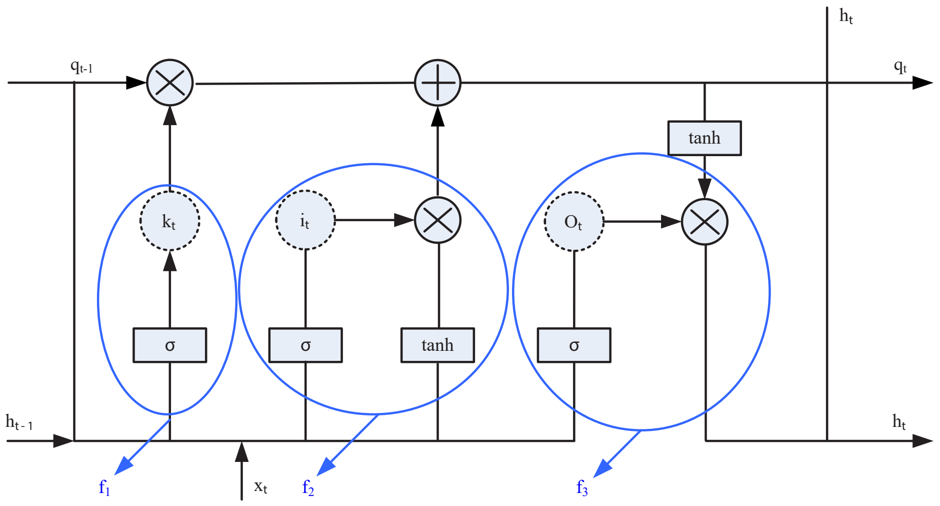

FIGURE 4 | Long short-term memory network (LSTM) cell structure.

these kinds of networks have achieved good recognition results. In this section, they will be briefly introduced.

(1) Long and short-term memory network.

The Recurrent Neural Network (RNN) network model was first proposed in the 1980s, and now this network model has been widely used in various research fields (Zheng et al., 2019). Because the RNN can make good use of past information, it can be used as an important reference for current decision-making. Therefore, compared to other traditional models, this model can better utilize timing information that other models cannot. But at the same time, a new long-term dependency problem arises, which prevents the RNN from learning information that is relevant for a longer time (Rossi and Ruocco, 2019; Barabanschikov and Suvorova, 2020). In order to better solve the long-term dependence problem generated by RNNs, in 1997, some scholars proposed a new neural network structure–LSTM. The LSTM network is a special network structure with three "gate" structures. **Figure 4** is a structural block diagram of the LSTM network.

The long-short-term memory network relies on the "gate" structure to effectively control the transmission information in the input network, thereby affecting the state of each moment in the neural network. Gate is a structure that combines a sigmoid neural network and a point multiplication function. There are three Gates in LSTM that work together, namely: forget gate, input gate and output gate. The key to solving the long-term dependency problem lies in the "forgetting gate" and the "input gate," which allow the LSTM network to retain long-term memory.

In short, the forget gate can discard useless information, so that the neural network can forget the previous invalid information. The current input xt, the state qt-1 at the previous moment, and the output ht-1 at the previous moment will jointly decide which information needs to be forgotten by the forgetting gate. After discarding invalid relevant information, it is necessary to supplement the network with appropriate information. At this point, the forget gate comes into play. Similarly, according to xt, qt-1, ht-1 will jointly decide which information needs to be supplemented by the input gate into the current state qt. The function of the input gate is to make a valid output. According to the latest state qt updated by the network, the output ht-1 at the previous moment, and the current input xt, the output gate will generate the output ht at this moment. Through the "three-gate structure," LSTM can effectively control the transmission state, discard unimportant information, and retain useful long-term memory. The specific work flow is described as follows:

In the LSTM network, the first step is to filter out the information that is not needed in the LSTM and discard it.

$$f_1 = \phi 1(W_1 \times x^*) \tag{1}$$

Where $x^* \overset{\Delta}{=} [x, s^{(old)}]$ means that the current input sample is connected to downstream time channel $s^{(old)}$.

The second step is to supplement some new information in the LSTM. The sigmoid layer of the input gate f2 determines the information that needs to be updated, and the tanh layer is used to generate the updated alternative content.

$$f_2 = \phi 1(W_2 \times x^*) \bullet \varphi 2(W_3 \times x^*) \tag{2}$$

The third step is to output information. The output gate f3 uses the sigmoid layer to determine the state of the output unit, which is processed first and then multiplied by the tanh layer, and finally output.

$$f_3 = s^{(new)} = \phi 2(h^{(new)}) \bullet \varphi 1(W_4 \times x^*) \tag{3}$$

Among them, W1∼W4 represent the weight matrix corresponding to each gate.

The LSTM network relies on its unique three-gate structure to effectively solve the long-term dependency problem of the neural network. It is suitable for the modeling of speech time signals and text signals that are closely related to time.

Convolutional Neural Network

Convolutional Neural Networks is one of the representative algorithms of deep learning. CNNs are widely used in speech-related tasks such as image classification, object detection, and speech synthesis. In the speech ER task, researchers also use CNNs to extract emotion-related speech features for ER. The basic structure includes convolutional layers, pooling layers, and fully connected layers, as shown in **Figure 5**.

Convolutional neural networks can automatically extract some advanced features, usually using backpropagation algorithm for parameter training and learning. The convolutional layer is the feature extraction unit of the CNN. The convolutional layer is composed of a convolutional kernel filter. The convolutional kernel parameters determine the convolutional layer to extract specific features. The parameters of a convolutional layer include: the number of input channels, the number of output channels, the size of the convolution kernel and the stride. The convolution kernel responds to local input features rather than the entire input feature map, and the local area where the response is obtained is the receptive field. In a feature map of a convolutional layer, the weights of the convolutional kernels are shared, which can detect specific patterns in different regions and greatly reduce the network parameters. The pooling layer often adopts mean pooling or max pooling to reduce noise and extract more robust features (Li et al., 2018).

The net input to the convolutional layer is calculated as follows:

$$Z^p = W^p \otimes X + b^p = \sum_{d=1}^{D} W^{p,d} \otimes X^d + b^p \tag{4}$$

Where W represents the convolution kernel, X represents the feature map, and b represents the bias term.

After the network input is operated by a non-linear activation function, the feature output map Y^p is obtained.

$$Y^p = f(Z^P) \tag{5}$$

Where f(·) is the non-linear activation function.

Suppose the input of the convolutional layer is m, and the output of the l-th feature map of the convolutional layer is n^l. The formula is expressed as:

$$n_{ij}^{l} = \sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C} \omega_{abc}^{l}{}^{m}(ki+a)(kj+b) + \beta_{ij}^{l} \tag{6}$$

Among them, ω^l represents the convolution kernel, and its size is A × B × C, A and B are the spatial dimensions, respectively, and C is the number of input channels. β^l is the bias term, and k is the step size of the convolution kernel.

The expression for mean pooling is:

$$z_{ij} = \frac{1}{AB}\sum_{a}\sum_{b} n(ki+a)(ki+b) \tag{7}$$

The expression for max pooling is:

$$z_{ij} = \max\nolimits_{a,b} n(ki+a)(kj+b) \tag{8}$$

Where A and B are the size of the pooling layer, and k is the stride of the pooling layer.

The activation layer provides non-linearity to the model. In general, the convolutional layer is followed by the activation layer. The normal activation functions include ReLU activation function, Sigmoid function and tanh function, etc. The formula expressions are as follows:

$$R(x) = \max(0, x) \tag{9}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{10}$$

$$T(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{11}$$

Typical CNNs, such as AlexNet, VGGNet, and ResNet, etc., due to the existence of fully connected layers in their structure, the input must be a fixed-size image (No author Listed, 2017; Cai et al., 2019). The difference is that AlexNet uses a large-size convolution kernel design, and there is a grouping

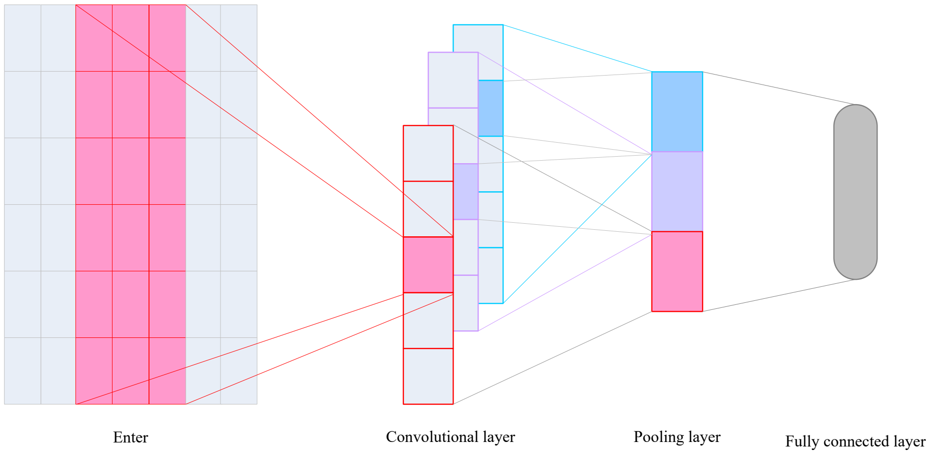

FIGURE 5 | Convolutional neural network (CNN) basic network structure diagram.

operation in the convolution layer. All convolutional layers of AlexNet are followed by ReLU-type activation functions, and the two convolutional layers starting after ReLU use local response regularization. The calculation of the local response normalization is shown in formula (12).

$$b^i_{x,y} = a^i_{x,y} l(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2)_\beta \tag{12}$$

In formula (12), $a^i_{x,y}$ represents the activation value after passing through the ReLU function after computing the response with the convolution kernel i at position (x,y). $b^i_{x,y}$ is the output of local response regularization, and the total number of convolution kernels in this convolutional layer is N. The summation formula here is performed on the feature maps of n adjacent kernel functions at the same position, in order to suppress the output of those convolution kernels with larger output values. The constants k, n, α and β here are hyperparameters determined from the validation set. In AlexNet, $k = 2$, $n = 5$, $\alpha = 10e\text{-}4$, $\beta = 0.75$.

(3) Attention mechanism.

The attention mechanism first appeared in the field of images, just as people pay more attention to the parts of things when they observe things. The principle of the attention mechanism is to assign high numerical weights to information units that have a great influence on the results, that is, to ensure that important information is not covered by irrelevant information through training parameters. Usually models that process sequence data, such as RNN and LSTM, use max pooling layers or average pooling to extract contextual information. Doing so often brings some losses, and using the max pooling layer may get misjudged information. The operation of average pooling often blurs the target and weakens the useful information.

Take the attention mechanism in machine translation as an example to illustrate in detail. The input sample sequence is sent to the encoder in time sequence, and the hidden state of the encoder is extracted at the last moment to initialize the hidden state of the decoder. The output sequences are then generated one after the other in time series. This encoder-decoder structure achieves better results than purely statistical based methods. But this RNN-based structure still has two serious flaws. First, RNNs are forgettable, which means that after thousands of temporal states, old information is forgotten. Second, there is no explicit word segmentation during decoding, so the system's attention is spread over the entire input. To address these two issues, researchers introduced an attention mechanism into machine translation (Sarvestani and Boostani, 2017). They keep the RNN encoder, and during decoding, for each temporal state j, they compute an attention weight α_{ji} for each hidden state vector h^{in}_i of the encoder, and finally get a content vector r_j:

$$e_{ji} = a(h^{in}_j, h^{out}_j) \tag{13}$$

$$\alpha_{ij} = \frac{e_{ij}}{\sum_i e_{ij}} \tag{14}$$

$$r_j = \sum_i \alpha_{ji} h^{in}_i \tag{15}$$

Here r_j is the weighted average of the input sequence elements, which is the encoded expression of the entire sentence for the current element h^{out}_j. $a(h^{in}_i, h^{out}_j)$ is the segmentation function, which is used to measure the similarity between two vectors. Then, r_j combines the current hidden state vector h_j and the output vector y_{j-1} of the previous step to generate the current moment output y_i4:

$$y_i = f_y(h^{out}_j, y_{j-1}, r_j) \tag{16}$$

$$h^{out}_{j+1} = f_h(h^{out}_j, y_j) \tag{17}$$

Among them, f_y and f_h represent the output layer and hidden layer functions of the RNN. This process is repeated until the end of the output sequence. By introducing this attention mechanism, the weak memory of RNN can be compensated. Because we compute the attention weight for each element in the input sequence, encoding vector r_j is not affected by the length of the input sequence. On the other hand, in this process, different attention weights are assigned according to the importance of different positions of the input sequence to the current decoding, and the soft localization of the input words can be achieved through this weight. This attention-based machine translation model has achieved great success.

The basic form of the attention mechanism is summarized as:

$$e_i = a(u, \upsilon_i) \tag{18}$$

$$\alpha_i = \frac{e_i}{\sum_i e_i} \tag{19}$$

$$c = \sum_i \alpha_i \upsilon_i \tag{20}$$

Formula (18) calculates the attention weights, (19) normalizes the weights, and (20) encodes them according to the attention weights. First, u is a vector related to the task, which is matched according to the segmentation function $a(u, \upsilon_i)$ and the i-th element υ_i in the sequence, and then the output weight scalar e_i indicates the degree of matching.

Strategic Implications of Business Ecosystem Enterprises

American strategist first proposed the concept of "business ecosystem" (Liu et al., 2021). He defines a business ecosystem as an economic union based on the interaction of organisms and individuals. The definition was then further refined. It is considered to be a global network system, a real physical system, composed of business context and WWW-related abiotic factors, a virtual habitat system for this real system (Aung et al., 2017). Others emphasize the structure and interconnection of business ecosystems, that is, interconnected, dynamic businesses that organize consumers' customers into business ecosystems.

MULTIMODAL EMOTION RECOGNITION EXPERIMENT AND ANALYSIS BASED ON SELF-ATTENTION MECHANISM

Experimental Dataset

Multimodal ER is an emerging field of research. Human communication and emotional expression are inherently multimodal. From a resource point of view, we really need large-scale datasets to deeply study multimodal ER (Alam et al., 2017). In this experiment, CMU-MOSI and CMU-MOSEI are the datasets we selected, and the data are summarized in **Tables 1**, **2**.

The first dataset is a video dataset MOSI7I (Multimodal Opinion-Level Sentiment Intensity) of 93 people commenting on certain topics in English, the videos are full of emotions. Each segment of the video was scored by 5 annotators based on the sentiment expressed by the segment, ranging from –3 for extremely negative sentiment to +3 for extremely positive sentiment. In this paper, the five annotation scores are averaged, and this average result is used as the sentiment polarity. That is, only two categories of positive emotions and negative emotions are considered. The videos of the first 62 people in the dataset are used as the training set, and the videos of the remaining 31 people are used as the test set. The training and test sets have 1,447 and 752 video clips, respectively.

TABLE 1 | Statistical overview of the CMU-MOSI dataset.

Total number of video clip samples	2,199
Different speakers	89
Total number of videos	93
The average number of video clips cut by the video	23.2
Average length of video clip samples	4.2 s
The total number of words in the video clip sample	26,295
The total number of unique words in the video clip sample	3,107

TABLE 2 | Statistical overview of the CMU-MOSEI dataset.

Total number of video clips	23,453
Total number of videos	3,228
Number of different speakers	1,000
Number of different topics	250
The number of video clips in a video	7.3
Average duration of each video clip	7.28 s
The total number of words in the video clip sample	447,143
The total number of unique words in the video clip sample	23,026

Recognition Design Under the Self-Attention Mechanism

To demonstrate the strengths of the models presented in this chapter, these two datasets are compared with classic models that have been used to solve multimodal sentiment analysis problems very well. There are three models that need to be compared: Simple LSTM (Simple Long Short-Term Memory Neural Network), which is proposed by the paper in ACL2017 and can represent a class of algorithm models that complete multimodal fusion through learning; Hierarchical RNN (RNN with hierarchical structure), the idea of this method is to splicing vectors of multiple modalities to complete multi-modal sentiment analysis; CAT-LSTM (Contextual Attention-Based LSTM), which is a multimodal fusion model based on attention mechanism (Sirai et al., 2017).

Figure 6 shows the results of running all methods when only unimodal data is input. The self-attention-based unimodal model proposed in this chapter achieves better results than other basic models. This result suggests that self-attention mechanisms can be more effective than LSTM-based models, such as Simple LSTM. Because the self-attention mechanism can obtain more contextual information between video clips.

The self-attention mechanism can be applied to unimodal sentiment analysis. In order to further illustrate that this method can improve the accuracy of sentiment analysis by extracting contextual information between video clips in a single modality, this chapter conducts a single-modal sentiment analysis experiment to evaluate the model proposed in this chapter.

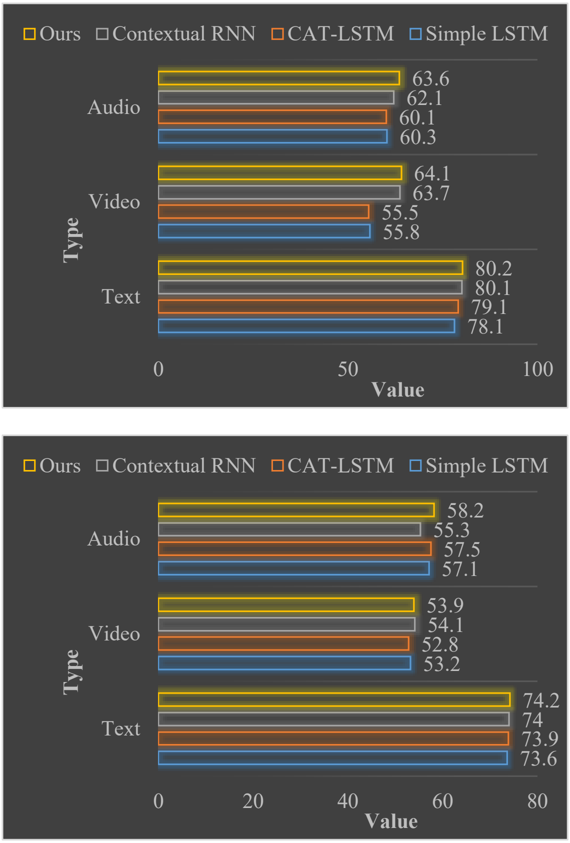

FIGURE 6 | The accuracy of single-modal emotion analysis in MOSI and MOSEI, respectively.

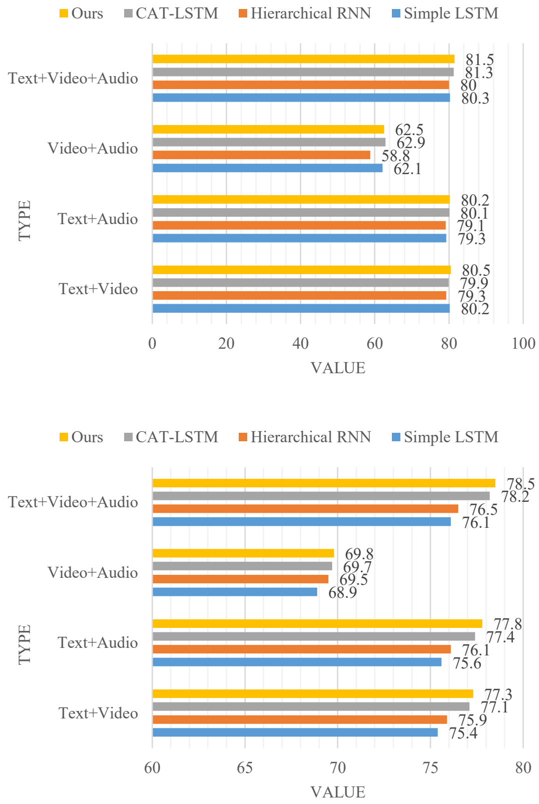

FIGURE 7 | The accuracy of multimodal sentiment analysis in MOSI and MOSEI, respectively.

Since the algorithm structure provided by Hierarchical RNN and Simple LSTM in single-modal sentiment analysis is exactly the same, the results of Hierarchical RNN are no longer written into the table. This chapter selects the Contextual RNN proposed in the paper published on EMNLP 2018 as another basic model for comparison. For this model, it is only necessary to apply the self-attention mechanism and Bi-GRU to three modalities of text, image and audio. The goal of the unimodal sentiment analysis experiment is to observe the effect of applying the self-attention mechanism to sentiment analysis.

The results of the multimodal sentiment analysis comparison experiment are shown in **Figure 7**. We tested the sentiment analysis accuracy of all methods under three modalities and two modalities.

The three-modal sentiment analysis and two-modal sentiment analysis of the method proposed in this chapter outperform the three classical models in **Figure 7** on both datasets. It shows that the model proposed in this chapter can more effectively utilize multimodal information in multimodal sentiment analysis problems. It is especially worth noting that the accuracy of this model is higher than that of Simple LSTM (Zhou et al., 2021). This shows that learning the connections between video clips with self-attention can improve the performance of the model in multimodal sentiment analysis. Noting that the accuracy of this model is higher than all methods except CAT-LSTM when only the fusion experiments are performed on image modalities and audio modalities. The main reason for this phenomenon is because the self-attention mechanism is only applied to text modalities, but not to images and audios. If the text modality is not considered at all, the variant of this model will adopt the idea of two-modal fusion based on self-attention, and the performance of the variant model will match or even exceed the performance of CAT-LSTM. In the presence of text modalities, this model needs to prioritize text. The vectors in the attention matrices generated by the three modalities are only concatenated with the vectors corresponding to the same video clip in the text modalities, and then used for sentiment classification.

Multimodal Emotion Recognition Scheme

The single-modal and bi-modal ER results of the MOSI dataset are shown in **Figure 8**.

As can be seen from **Figure 8**, the accuracy of single-modal ER is generally low, especially in language and video. The accuracy of the binary classification is only 56.0 and 54.9%, while the text can reach 74.1%. In addition, for the text mode, the effect of using the Text CNN model for ER is not as good as that of LSTM+FC,

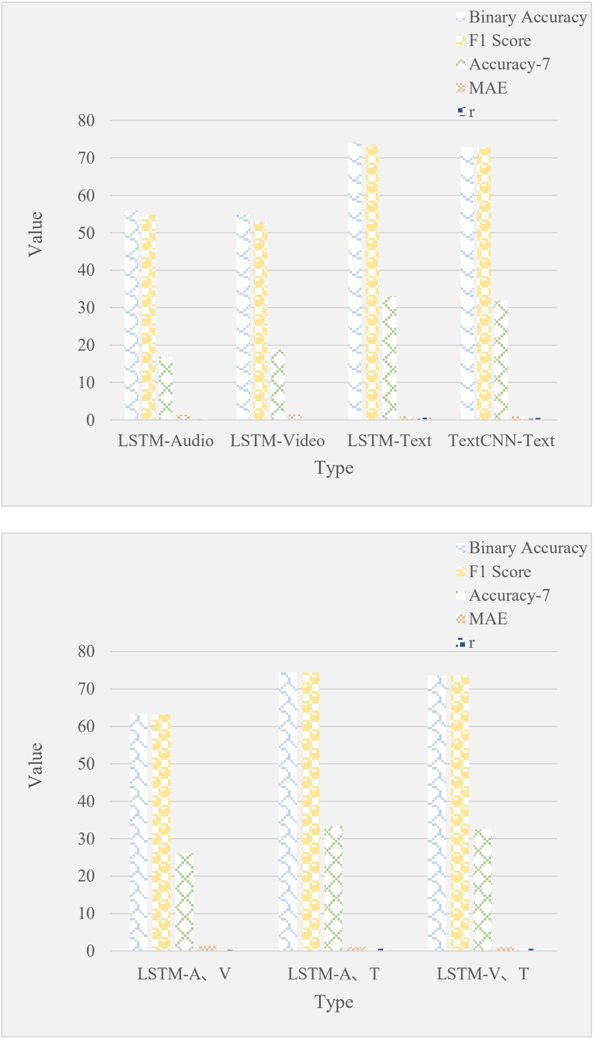

FIGURE 8 | Single-modal and bi-modal emotion recognition (ER) results in MOSI dataset.

and the Binary Accuracy is only 73.0%. From here, it can be analyzed that text as the main axis mode has a greater impact on ER, and compared to Text CNN, LSTM is also more suitable for our model. In the dual-modal ER results, LSTM-A and V are the models that use LSTM for cascade and FC processing using two modalities of speech and video. LSTM-A and T are two modes of speech and text, LSTM-V and T are two modes of video and text, and their models are the same as above. We will see that the BA of LSTM-A, V is 63.4%, the BA of LSTM-A, T is 74.5%, and the accuracy of LSTM-V, T is 73.6%. It can be clearly seen from the figure that the effect of dual-modal ER is higher than that of single-modal ER, and it can also be seen that the recognition results of related texts in dual-modal ER are relatively improved (Elleuch and Wali, 2019; Kim et al., 2021b; Liu and Fu, 2021).

The binary classification results report on the Confusion Matrix dataset for the test set on the MOSI dataset is shown in **Table 3**.

TABLE 3 | MOSI test set Confusion Matrix.

Emotion category	Negtive	Positive	Total
Negtive	306	65	375
Positive	90	225	311
Total	396	290	686

At the same time, we also performed 6-class multimodal ER on the MOSEI dataset, and **Figure 9** is the recognition result.

It can be seen that the effect of multimodal ER is satisfactory. Among them, happy has the highest average recognition accuracy, reaching 91.5%, and surprise is the lowest, with an accuracy of only 55.2%. It can be seen that the recognition accuracy is generally more accurate than the original recognition, and the effect of multimodal ER is remarkable.

DISCUSSION

This paper is devoted to the study of multi-modal ER of deep learning, and three modalities of text, speech and video are selected. Firstly, a comprehensive analysis and overview of the background, significance, and related work of this paper are carried out, and the significance of this topic to enterprise strategic management from the perspective of business ecosystem is explained. Then, the multimodal feature fusion method and the important application of deep learning neural network in ER are introduced. Then, this paper uses deep learning neural network to identify single-modal and multi-modal emotional features. Finally, the multi-modal experimental results of this paper are described. This paper introduces the ER process that introduces the attention mechanism, and designs the analysis process of the experiment to compare the single-modal and multi-modal recognition results of this model (Khalaf et al., 2021; Kim et al., 2021a; Surendran et al., 2022).

In the business ecosystem, corporate strategic management faces a huge, non-centralized, borderless business environment, such as the network economy. When making strategic decisions for a company, it cannot just focus on a particular aspect of internal resources and closely related external companies. It cannot be guided solely by strategic management theory derived from this business environment. In the new era of entrepreneurial ecological environment, the practical orientation of traditional strategic management theory has been lost (Weisman et al., 2017). The perspective of enterprise management strategy should be "showing emotion from the scene," and the focus of management strategy should be "like and love." The strategic decision of an enterprise in a business ecosystem should be based on its value and connectivity characteristics, and according to which it forms the business ecosystem in which the enterprise is located. In this paper, a multimodal neural network model based on self-attention mechanism is established, and the review videos uploaded by users are taken as the research object, and sentiment analysis is performed on different

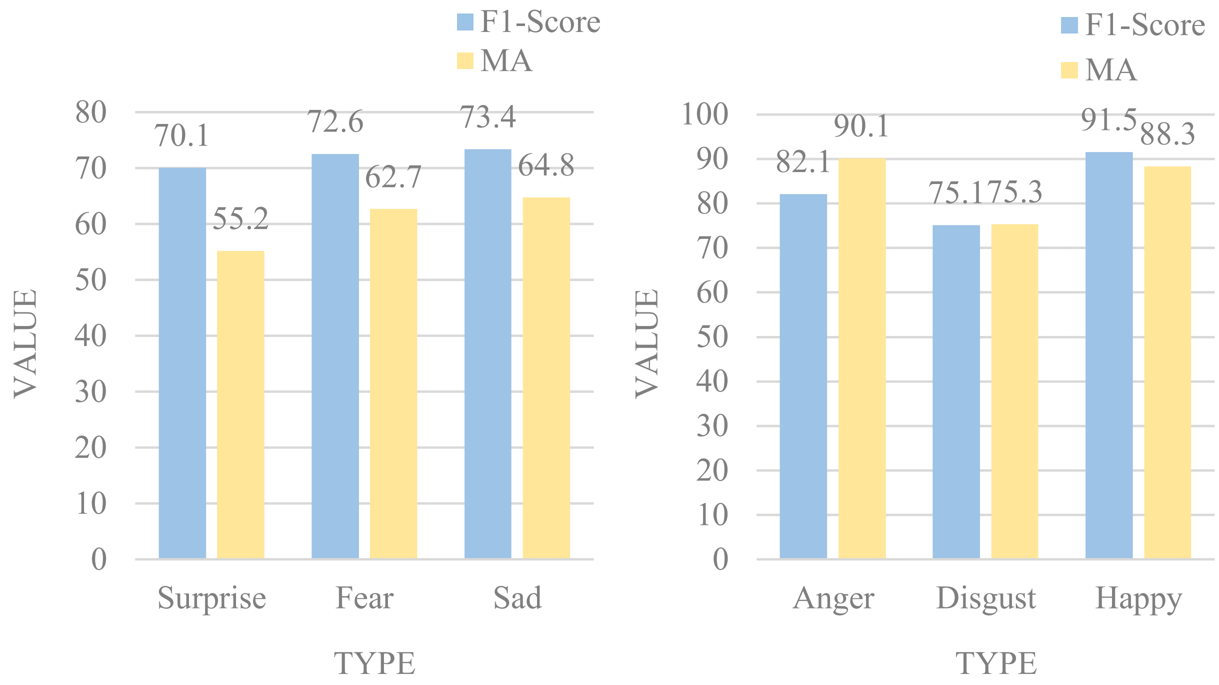

FIGURE 9 | Emotion recognition (ER) results on the MOSEI dataset.

parts of the videos. The experimental results are verified by comparing the experimental data of single-modal ER and multi-modal ER. On the basis of experiments, the analysis shows that the recognition of single modality is lower than the effect of multimodal ER, which proves that introducing additional methods will significantly improve the accuracy of ER.

CONCLUSION

The vigorous development of science and technology has brought more development of artificial intelligence, and human-human interaction has turned to human-computer interaction. Therefore, people's emotional requirements for artificial intelligence are getting higher and higher, and they want them to have the same emotions as people and can give more spiritual souls. Emotions are also a novel topic in both strategic and entrepreneurial studies. According to the concept and structure of the business ecosystem, the connection characteristics between the core enterprise and the complementary alternative enterprises in the business ecosystem will have an important impact on the strategic operation performance of the enterprise and the performance of the entire business ecosystem. The multimodal ER is based on the enterprise strategic management under the business ecosystem. It can make more accurate judgments on human emotions and better understand people's psychological needs by reading human gestures, tone, facial expressions and other aspects of conveying emotions, so as to give correct and timely responses. In real life, there are many ways to express human emotions. When we talk, we make random expressions that match the words, the context, and often some body language. Therefore, the expressions of human emotions are diverse, and only a single modality will have a higher error rate, and the emotional judgment made by a single modality is far less than multimodal ER.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

WB: writing. YX: editing. ZD and HL: data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Shandong Social Science Planning and Research Project (No. 20CPYJ29).

REFERENCES

Alam, F., Danieli, M., and Riccardi, G. (2017). Annotating and modeling empathy in spoken conversations. *Comput. Speech Lang.* 50, 40–61. doi: 10.1016/j.csl.2017.12.003

Aung, M. S. H., Kaltwang, S., Romera-Paredes, B., Martinez, B., Singh, A., Cella, M., et al. (2017). The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset. *IEEE Trans. Affect. Comput.* 99, 1–1. doi: 10.1109/TAFFC.2015.2462830

Barabanschikov, V. A., and Suvorova, E. V. (2020). Human emotional state assessment based on a video portrayal. *Exp. Psychol.* 13, 4–24. doi: 10.1037/pas0000595

Cai, L., Hu, Y., Dong, J., and Zhou, S. (2019). Audio-textual emotion recognition based on improved neural networks. *Math. Probl. Eng.* 2019, 1–9. doi: 10.1155/2019/2593036

Elleuch, H., and Wali, A. (2019). Unwearable multi-modal gestures recognition system for interaction with mobile devices in unexpected situations. *IIUM Eng. J.* 20, 142–162. doi: 10.31436/iiumej.v20i2.1000

Khalaf, O. I., Tavera Romero, C. A., Jaisudhan Pazhani, A. A., and Vinuja, G. (2021). VLSI implementation of a high-performance nonlinear image scaling algorithm. *J. Healthcare Eng.* 2021:6297856. doi: 10.1155/2021/6297856

Kim, J. H., and Lee, S. P. (2021). Multi-modal emotion recognition using speech features and text embedding. *Transac. Korean Inst. Electric. Eng.* 70, 108–113.

Kim, T.-Y., Kim, S.-H., and Ko, H. (2021a). Design and implementation of BCI-based intelligent upper limb rehabilitation robot system. *ACM Transac. Internet Technol.* 21, 1–17.

Kim, T.-Y., Ko, H., Kim, S.-H., and Kim, H.-D. (2021b). Modeling of recommendation system based on emotional information and collaborative filtering. *Sensors* 21:1997. doi: 10.3390/s21061997

Li, X., Lu, G., Yan, J., Li, H., Zhang, Z., Sun, N., et al. (2018). Incomplete cholesky decomposition based kernel cross modal factor analysis for audiovisual continuous dimensional emotion recognition. *KSII Trans. Internet Inf. Syst.* 13, 810–831.

Liu, S., Zhang, M., Fang, M., Zhai, J., Hou, K., Hung, C.-C., et al. (2021). Speech emotion recognition based on transfer learning from the FaceNet framework. *J. Acoust. Soc. Am.* 149, 1338–1345. doi: 10.1121/10.0003530

Liu, Y., and Fu, G. (2021). Emotion recognition by deeply learned multi-channel textual and EEG features. *Future Gen. Comput. Syst.* 119, 1–6. doi: 10.1016/j.future.2021.01.010

Manisha, S., Nafisa, H. S., Gopal, N., and Roshni, A. P. (2021). Bimodal emotion recognition using machine learning. *Int. J. Eng. Adv. Technol.* 10, 189–194.

Min, C., Hao, Y., Kai, L., Yuan, Z., and Hu, L. (2018). Label-less learning for traffic control in an edge network. *IEEE Network* 32, 8–14. doi: 10.1109/mnet.2018.1800110

No author Listed. (2017). "Functional architecture of visual emotion recognition ability: a latent variable approach": correction to Lewis, Lefevre, and Young (2016). *J. Exp. Psychol. Gen.* 146, 1085–1085. doi: 10.1037/xge0000160

Rao, K. P., Chandra, M., and Rao, S. (2019). Assessment of students' comprehension using multi-modal emotion recognition in E-learning environments. *J. Adv. Res. Dyn. Control Syst.* 10, 767–773.

Rossi, S., and Ruocco, M. (2019). Better alone than in bad company: effects of incoherent non-verbal emotional cues for a humanoid robot. *Interaction Stud.* 20, 487–508. doi: 10.1075/is.18066.ros

Saha, T., Gupta, D., Saha, S., and Bhattacharyya, P. (2020). Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cogn. Comput.* 11, 1–13.

Sarvestani, R. R., and Boostani, R. (2017). FF-SKPCCA: kernel probabilistic canonical correlation analysis. *Appl. Intell.* 46, 438–454. doi: 10.1007/s10489-016-0823-x

Sirai, E., Aran, L. R., and Wong, F. A. (2017). Review of methods in speech and facial expressions recognition for human-computer interaction. *J. Comput. Theor. Nanosci.* 23, 10236–10240. doi: 10.1166/asl.2017.10426

Surendran, R., Khalaf, O. I., and Andres, C. (2022). Deep learning based intelligent industrial fault diagnosis model. *CMC-Comput. Mater. Con.* 70, 6323–6338. doi: 10.32604/cmc.2022.021716

Wang, X., Peng, M., Pan, L., Hu, M., Jin, C., Ren, F., et al. (2018). Two-level attention with two-stage multi-task learning for facial emotion recognition. *J. Vis. Commun. Image Representation* 62, 217–225.

Wei, W., Jia, Q., Feng, Y., Chen, G., and Chu, M. (2020). Multi-modal facial expression feature based on deep-neural networks. *J. Multimodal User Interfaces* 14, 17–23. doi: 10.1007/s12193-019-00308-9

Weisman, O., Chetouani, M., Saint-Georges, C., Bourvis, N., Delaherche, E., Zagoory-Sharon, O., et al. (2017). Dynamics of non-verbal vocalizations and hormones during father-infant interaction. *IEEE Trans. Affect. Comput.* 7, 337–345.

Yoo, G., Seo, S., Hong, S., and Kim, H. (2018). Emotion extraction based on multi bio-signal using back-propagation neural network. *Multimed. Tools Appl.* 77, 4925–4937. doi: 10.1007/s11042-016-4213-5

Zhao, S., Jia, G., Yang, J., Ding, G., and Keutzer, K. (2021). Emotion recognition from multiple modalities: fundamentals and methodologies. *IEEE Signal Process. Mag.* 38, 59–73. doi: 10.1109/msp.2021.3106895

Zheng, W. L., Liu, W., Lu, Y., Lu, B.-L., and Cichocki, A. (2019). EmotionMeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zhou, B., Wan, J., Liang, Y., and Guo, G. (2021). Adaptive cross-fusion learning for multi-modal gesture recognition. *Virtual Reality Intell. Hardw.* 3, 235–247. doi: 10.1016/j.vrih.2021.05.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.