Research Article

Lulu Nie*

A robot electronic device for multimodal emotional recognition of expressions

https://doi.org/10.1515/pjbr-2022-0127
received May 23, 2023; accepted April 8, 2024

Abstract: This study addresses the challenge of low recognition rates in emotion recognition systems, attributed to the vulnerability of sound data to ambient noise. To overcome this limitation, we propose a novel approach that leverages emotional information from diverse modalities. Our method integrates speech and facial expressions through advanced feature layer fusion and decision layer fusion strategies. Unlike traditional fusion algorithms, our proposed multimodal emotion recognition algorithm incorporates a dual fusion process at both the feature layer and the decision layer. This dual fusion not only preserves the distinctive characteristics of emotional information across modalities but also maintains inter-modal correlations. To evaluate the effectiveness of our approach, experiments were conducted using the eNTERFACE'05 multimodal emotion database. The results demonstrate a remarkable recognition accuracy of 89.3%, surpassing the highest recognition rate of 83.92% achieved by the current state-of-the-art kernel space feature fusion method. Our algorithm exhibits a significant improvement of 5.38% in recognition accuracy. By combining emotional data from speech and facial expressions using a data fusion methodology, our study demonstrates a significant improvement of 5.38% in recognition accuracy, contributing to the progress of multimodal emotion recognition systems.

Keywords: multimodal emotion recognition, speech emotion recognition, facial expression recognition, convolutional recurrent neural network

*** Corresponding author: Lulu Nie,** Department of Information Engineering, Zhengzhou Institute of Technology and Business, Zhengzhou, Henan 451400, China, e-mail: lulunie6@163.com

1 Introduction

Emotion is a collective term for a series of subjective cognitive experiences – a psychological and physiological state that combines various feelings and behaviors. Factors such as mood, personality, and hormones [1] interact with emotion. Different emotions have different guiding effects on daily behavior, and everything people do has different emotional expressions. Understanding the emotions of users when using products can greatly improve service quality and effectiveness. In recent years, with the increasing development of science and technology, the efficiency and accuracy of emotion recognition have greatly improved. Emotional recognition systems can help customer service personnel understand the emotional status of customers in advance, improve service quality, and thereby improve service efficiency and satisfaction. However, improving the recognition accuracy of emotion recognition in a single mode is challenging and often requires the use of a high-performance GPU. This limitation restricts the application scope of the system and makes it difficult to achieve the desired effect in real-life scenarios. To achieve the desired effect in real-life scenarios, we need to reduce the usage conditions of the system, design a multimodal emotion recognition system that combines software and hardware, accelerate the calculation process in parallel, improve the recognition accuracy of the system, and enable its use on a wider range of devices. With the advent of the artificial intelligence era, the intersection between computers, robots, and human life continues to expand.

A variety of intelligent devices bring people many conveniences in daily life and work and further promote the construction of smart cities. Emotion recognition is a multidisciplinary research topic, mainly including physiological psychology, cognitive science, information science, and other disciplines. Researchers actively explore methods to enable intelligent devices such as computers to recognize and understand emotional information conveyed through human voice and expression, thereby facilitating enhanced emotional communication between people and computers. Human beings generally use speech and facial expressions to convey emotional information. The existing idiom "observing

speech and facial expressions" means paying attention to listening to others' words and seeing their expressions; subsequently, the brain understands the other person's current state, changes the way they speak, and enables them to communicate harmoniously [2,3]. Although single-modal emotion recognition technology is becoming increasingly mature, there are some unavoidable drawbacks to the emotional feature information of single-modality, such as poor quality of emotional features in databases and single-data information. Moreover, single-modal emotion recognition technology has limited applicability and cannot be fully employed for all research objects. The commonly extracted features in speech emotion recognition (SER) are prosodic features, sound quality features, and spectral features [4]. The intonation, intensity, and duration of human speech vary, and these unique prosodic features construct beautiful and pleasant sounds. The commonly used prosodic features include fundamental frequency, duration, energy, etc.

The characteristics of sound quality mainly lie in the characteristics of language spectrum and timbre, which depend on the form of the sound waves spoken. Spectral features show how signals behave in the frequency domain. Some common spectral feature parameters are Mel Frequency Cepstral Coefficients, Linear Predictive Cepstral Coefficients, and others. Although SER is relatively simple to analyze from a time domain perspective, the spectrum reflects the most important perceptual characteristics in speech signals [5,6]. With the continuous maturity of Internet of Things technology and the continuous development of sensor devices and deep learning algorithms, the robot service scene has gradually covered the field of personal homes. So, for family service robot-related field research by researchers from all walks of life, emotion is the embodiment of human communication, and emotional recognition is the basis of emotional calculation. The user expects the family robot to observe the communication object and then understand the emotions expressed by each other, and based on understanding the user's emotional state, combined with the uncertainty of dynamic family environment information, make an appropriate response. So, the service robot has like human "observation" ability and can be in the process of man-machine or all interaction to observe and understand the user's emotional state. At the same time, in the uncertainty of a dynamic family environment, the robot, according to the user's emotional state, can take reasonable action, becoming an important direction of service robot technology development.

For image data, the most widely extracted features are directional gradient histograms (HOG), local binary patterns (LBPs), and Haar-like features. The fields of use for these three features vary widely in facial expression recognition systems. Haar-like features describe the pixel brightness transformation information of an image within a local range; therefore, it mainly detects the front of the face; the HOG feature describes the shape edge gradient information corresponding to the local range, so it has more advantages in detecting the side of a face in pedestrian recognition; LBP features represent the texture information corresponding to the image in a local range. Although the extraction speed is faster than Haar-like features, the accuracy of the extraction will decrease. Therefore, this article chooses to extract Haar-like features for the extraction of facial expression features. Figure 1 shows the method for facial expression recognition. To address this research question, Zheng et al. proposed a multimodal emotion recognition method based on speech, text, and action [7]. A deep wave off-field push improved wave physics model (DWE-WPM) is designed for SER. A custom feature extraction scheme reconstructed the waveform and injected it into DWE-WPM to simulate the information mining process of LSTM. In text emotion recognition, the deformation model of the multi-attention mechanism is used to identify the text emotion combination. Groups extracted the sequential features of facial expressions and hand movements in motor emotion recognition. We designed the four-channel joint model by combining it with a bidirectional three-layer

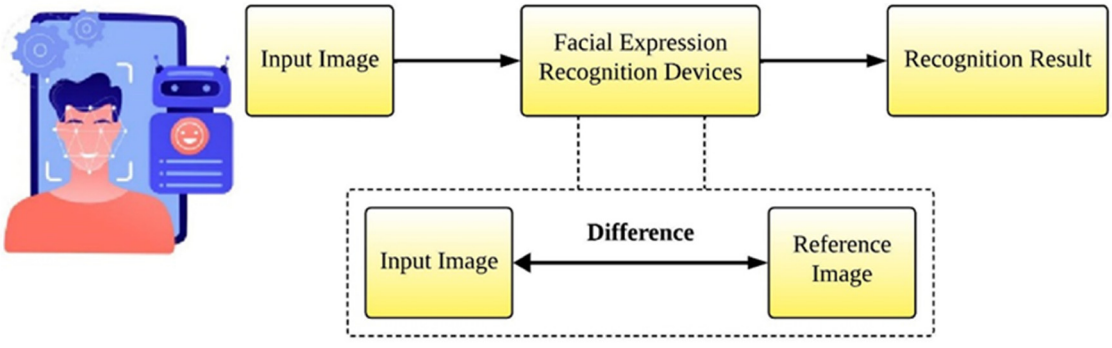

Figure 1: Methodology for recognizing facial expressions.

LSTM model with an attention mechanism. Experimental results show that the proposed method has high recognition accuracy in the multimodal regime. Xie et al. studied robust approaches for multimodal emotion recognition during conversation [8]. Researchers structured and fine-tuned three separate audio, video, and text models on MELD, using a transformer-based cross-mode fusion technique with blackmail network architecture. The proposed multimodal network architecture can achieve up to 65% accuracy, greatly exceeding any unimodal model. We provide multiple evaluation techniques to demonstrate the robustness of our model, which can even outperform the state-of-the-art models on MELD. To improve customer service, human–computer interaction, and the creation of intelligent gadgets, emotion recognition technology is essential. Our motivation comes from the aim to greatly increase recognition accuracy and broaden the application of emotion recognition technologies to real-life scenarios, as we are aware of the limitations of single-modal emotion detection systems, especially in noisy surroundings. Our goal is to close the emotional gap between humans and artificial intelligence by leveraging voice and facial expression data in a revolutionary multimodal way, leading to more natural and intelligent human–machine interactions. Our research offers a novel dual-fusion approach that combines speech and facial expressions at the feature and decision layers. In contrast to conventional fusion algorithms, our method maintains important inter-modal correlations as well as unique features of emotional information across modalities. The key innovation is our redesigned fusion algorithm that gives emotional features the right weights, overcoming the shortcomings of previous techniques that found it difficult to decide how much weight to give each feature. The fusion algorithm in our method makes it stand out because it can recognize and use different kinds of emotional information from different modes. This could make multimodal emotion identification systems more accurate.

2 Related work

The complexity of human emotions and the challenges of combining information from different modes of communication make it harder to create multimodal emotion recognition systems that are both accurate and useful. Prevalent approaches frequently encounter difficulties in adjusting to practical situations, which compromises the overall dependability and efficacy of such systems. The field of emotion recognition has made significant strides forward, with an emphasis on improving precision across a range of applications. Nigam et al. [9] introduced a multimodal approach to emotion recognition that incorporates speech, text, and action. An enhanced wave physics model based on deep-wave off-field push facilitated their successful implementation. Wang et al. [10] looked into reliable ways to recognize emotions across multiple modes of communication by using a transformer-based cross-mode fusion method on separate text, video, and audio models. Researchers have used VGG-Face model fine-tuning in the area of recognizing single-modal emotions, as shown in the study [11], which created a complete system for teaching a strong expression recognition model. The literature thoroughly discusses feature extraction techniques, including Haar-like features for facial expression recognition [12]. Furthermore, Wang et al. [13] investigated the domain of affective computing by presenting a multimodal deep-learning methodology designed to identify emotions in videos. The model utilized a blend of auditory and visual components, demonstrating encouraging outcomes by encapsulating intricate emotional states. In the interim, Cannata et al. [14] made a significant contribution to the field of facial expression analysis by introducing the OpenFace toolkit, which offered an all-encompassing infrastructure for the extraction of facial features and classification of expressions. Certain researchers, including Akçay and Oğuz [15], have placed significant emphasis on the value of dynamic acoustic features in the context of SER within the domain of feature extraction. They center their research on utilizing dynamic features, including cadence and tempo, to enhance the classification capabilities of emotion recognition models.

Fang et al. [16] investigated the incorporation of physiological signals, which provide a holistic comprehension of emotions by correlating physiological reactions with facial expressions. Zhang et al. [17] were trailblazers in the field of emotion recognition by introducing convolutional recurrent neural networks to analyze facial expressions and utterances simultaneously. The model exhibited the ability to capture temporal dependencies in multimodal data, indicating encouraging outcomes in the identification of subtle affective states. Zhong et al. [18] investigated the amalgamation of deep learning methodologies to identify emotions in both speech and face, with a particular focus on the significance of end-to-end learning in encoding intricate emotional characteristics. Hossain et al. [19] made a substantial contribution to the field of real-world applications through the introduction of the AFEW dataset, which serves as a benchmark for real-world facial expression recognition. By addressing the constraints associated with controlled environments, this dataset proves to be an invaluable asset when assessing the resilience of emotion recognition systems across a wide range of environments.

Additionally, Tripathi et al. [20] contributed to the discipline by introducing a context-aware methodology for emotion recognition that incorporated contextual data to improve the precision of emotion forecasting. The wide range of methodologies employed demonstrates the interdisciplinary character of emotion recognition, which extends to physiological signal analysis, computer vision, and affective computing. However, developing a cohesive and resilient structure that effectively incorporates data from diverse modalities, taking into consideration the intrinsic intricacy of human emotions in practical situations, remains an ongoing obstacle despite considerable advancements achieved thus far. To confront these ongoing obstacles, our study employs a holistic methodology by combining speech and facial expression data via a dual fusion procedure. By capitalizing on developments in feature extraction, such as the refinement of the VGG-Face model and implementing an enhanced fusion algorithm, our objective is to augment the resilience and precision of multimodal emotion recognition. A rigorous evaluation of the proposed methodology using the eNTERFACE'05 database demonstrates its effectiveness, outperforming current state-of-the-art approaches.

3 Methods

In this section, we methodically describe the suggested method for multimodal emotion recognition. It includes gathering data, preprocessing methods for speech and facial expression data, and feature extraction using transfer learning with the VGG-Face model. In this section, we describe the dual fusion technique at the feature and decision layers in detail, utilizing multi-kernel learning and an advanced fusion algorithm. The section also explains the utilization of the eNTERFACE'05 Multimodal Emotion Database for evaluation, along with the architecture of the ACNN-LSTM model for classification. The proposed methodology functions without interruption by employing a methodical and groundbreaking procedure to identify emotions across multiple modalities which is depicted in Figure 2. The process begins with the procurement of speech data and facial expression images of superior quality, which guarantees a wide range of emotional expressions. The preprocessing module thoroughly cleanses the data by implementing signal processing techniques to reduce noise in speech and facial detection and alignment, extracting consistent features from facial expressions. During the phase of feature extraction, the VGG-Face model performs fine-tuning to extract complex facial features. Concurrently, the analysis of speech data includes extracting prosodic and spectral features. At both the feature and decision layers, the strategy distinguishes itself through its dual fusion approach.

Multi-kernel learning enables the integration of facial and speech features while maintaining their distinct attributes and inter-modal correlations. A sophisticated fusion algorithm improves the precision of decisions by taking into account the weights assigned to affective features. The architecture utilized by the classification model is ACNN-LSTM, which combines attention mechanisms with convolutional recurrent neural networks. Using a softmax activation function, the output layer generates final recognition results by classifying various emotional states. Utilizing random partitioning for both training and testing, the assessment demonstrates the approach's resilience by utilizing the eNTERFACE'05 Multimodal Emotion Database. Comparative analysis shows that the approach is better than current methods, and its success opens the door for more research and improvements in the field of multimodal emotion recognition.

3.1 Expression recognition based on VGG-face model fine-tuning structure

Deep learning is a machine learning method based on neural networks that has emerged with the boom in big data and computing. However, researchers are unable to obtain sufficiently effective datasets, so they cannot train deeper convolutional neural networks from scratch. In response to the problem of poor training results for complex networks with a small number of samples, the author proposes an expression recognition based on VGG-Face model fine-tuning, which first trains a large expression dataset and, next, uses the trained model as a training task weight or feature extractor [21].

3.1.1 Framework of the algorithm

In addition to training the final classifier, the fine-tuning model can also use backpropagation algorithms to adjust the parameters of the network layer according to its needs. The fine-tuning object can be all network layers or specified layers, and for the emotion of insufficient data storage during the fine-tuning process, to prevent overfitting, only fine-tune high-level parameters. With the increase in convolution depth, the process of extracting parameters will be very different. For image data, low-level CNN features have universality and can be considered edge detectors or color block detectors. As the number of convolutional layers increases, the number of feature blocks containing information from the expression database can also increase.

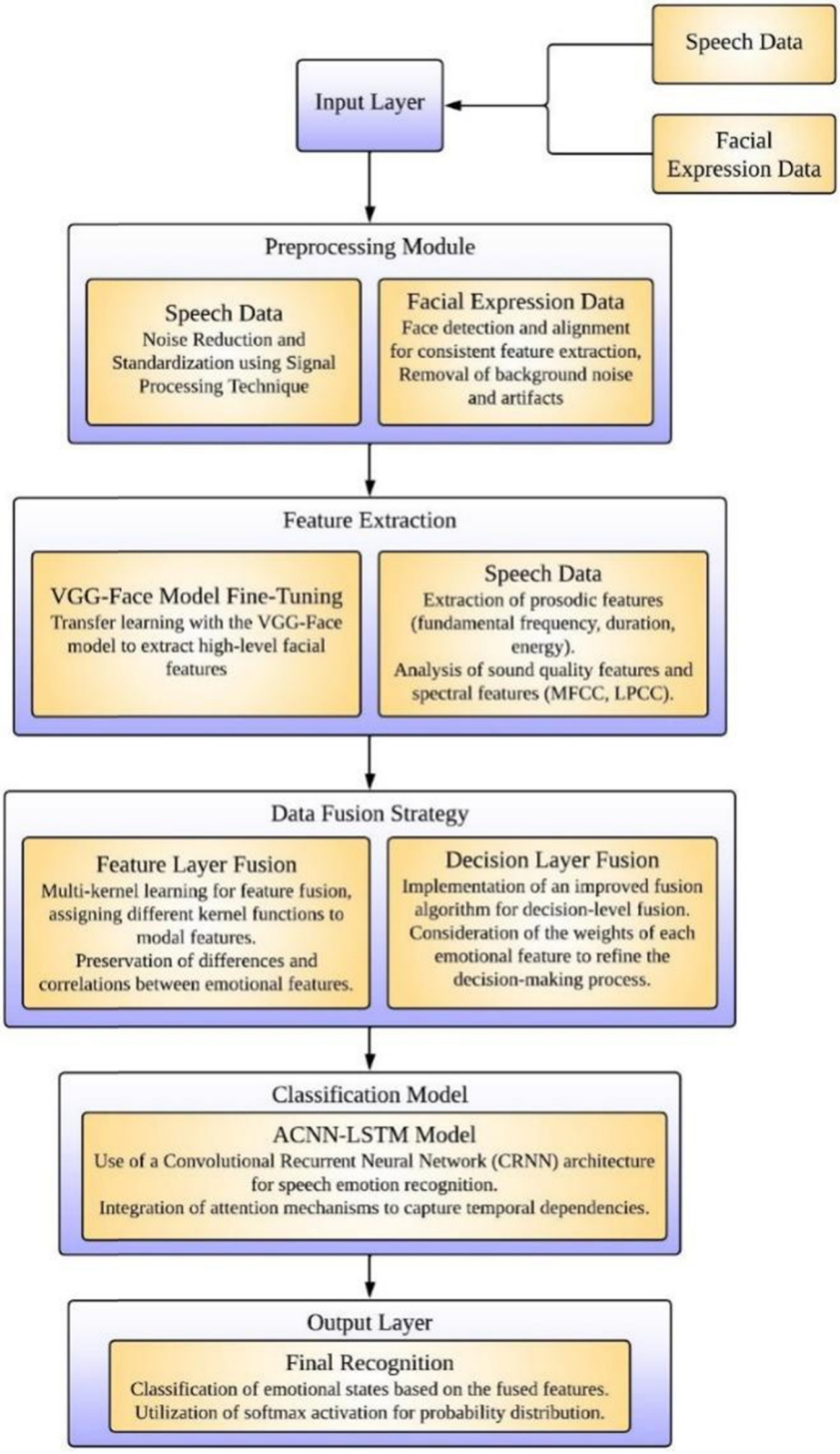

Figure 2: Proposed approach for multimodal emotion recognition.

The recognition results of facial expression samples were obtained, and a large number of facial expression images in the database were selected as the training subjects to fine-tune the VGG-Face model, adjust the parameters, and train the optimal model [22].

3.1.2 Design and theoretical analysis of algorithms

This section is based on the VGG-Face fine-tuning model to train facial expressions, modifying parameters to train a fast and efficient model. The specific steps of the expression recognition model are the same as the training process for speech-emotion recognition. The specific training steps of the model are to divide the samples in the expression database into training sets (train_data) and testing sets (test_data) and obtain the training set label (train_class), testing set label (test_class), and category (class), respectively [23]; import pre-trained models; train the expression samples in the pre-trained model, adjust the parameters to fine-tune the model, and change the category of the output layer to five categories; conduct training; and call the function to calculate the mean of the training set, save the updated mean, and then call the function to iteratively train the expression samples. After training all facial expression samples, the author saved the trained network model and then convolved, pooled, and fully connected the facial expression dataset; conducted testing; and inputted the test set data into the VGG-Face fine-tuning model for testing and obtaining emotional recognition rates for each category [24,25].

3.2 Improvement of multimodal emotion recognition fusion algorithm

Based on previous research on single-modal SER methods, this section uses useful information from several information sources for multimodal emotion recognition methods because single-modal emotion information has some flaws and can't fully express human emotions. Feature fusion algorithms use the complementary emotional features of different modalities, but they can't figure out how much weight to give to each emotional feature. This means that they can't show how different emotional features differ in recognizing emotions. Although decision-level fusion algorithms consider the differences between different modal information, they ignore the correlation between information, unable to meet the requirements of this section of the experiment. An improved fusion algorithm is proposed for multimodal emotion recognition [26,27].

3.2.1 Framework for improving fusion algorithms

At present, the commonly used information fusion strategies in multimodal emotion recognition research include feature layer fusion and decision layer fusion. Feature layer fusion combines multiple modal features to form a new fused feature, which is then classified into a training model to obtain recognition results. Decision layer fusion is based on the results of individual recognition of each modality and then makes decision judgments based on rules to obtain the final recognition result. However, the feature layer fusion algorithm cannot reflect the differences in different modal features, and the decision layer fusion algorithm cannot reflect the correlation of different modal information. To fix these problems, the author makes the traditional fusion algorithm better by first combining the extracted speech and facial features to get the expected results. They then make a 3-Dlog Mels map of the speech features and put it in the trained ACNN-LSTM model to get results for SER. Then, the two results for decision-making are fused to obtain the final recognition result [28].

3.2.2 Design and theoretical analysis of improved fusion algorithms

First, the extracted speech and facial features are fused. Because speech and facial features are not the same, the types and parameters of kernel functions used will not be the same. To get around this, the author uses the feature layer fusion method of multi-core learning, giving different kernel functions to the extracted modal features and getting the weight coefficients of each kernel function during the learning process. The multimodal emotional features can be divided into speech features, facial features, and serial features obtained by combining speech and facial features. We use different kernel functions to map each feature and obtain the weight coefficients of the kernel functions corresponding to each feature type through learning. Not only does it preserve the differences between different modalities of emotional information, but it also preserves the correlation between emotional information. Sample space S is taken as an example.

The feature space S includes speech feature S_{Speech}, facial feature S_{face}, and serial feature S_{all}, and can be $S_{\text{all}} = [S_{\text{Speech}}, S_{\text{face}}]$. Using different types of kernel functions for multi-kernel learning mapping of speech, facial expressions, and serial features, represented by F_s, F_f, and F_a, the kernel functions used for speech, facial expressions, and serial features correspond to weight coefficients represented by d_s, d, and d_a, respectively. According to the multi-

core definition expression, equation (1) is expressed as follows:

$$\begin{cases} F(x,x) = \sum_{t=1}^{n_1} d_{st} F_{st}(x,x) + \sum_{t=1}^{n_2} d_{ft} F_{ft}(x,x) \\ \quad + \sum_{t=1}^{n_1} d_{at} F_{at}(x,x) \\ d_{st}, d_{ft}, d_{at} \geq 0, \sum_{t=1}^{n_1} d_{st} + \sum_{t=1}^{n_2} d_{ft} + \sum_{t=1}^{n_3} d_{at} = 1 \\ n_1 + n_2 + n_3 = N. \end{cases} \tag{1}$$

In the equation, N is the total number of kernel functions used, while n_1, n_2, and n_3 represent the number of kernel functions used on speech features, facial features, and serial features, respectively. The author placed the fused features into a classifier for classification, obtained the classification results, and then made decision fusion with the results of single modal SER. Decision layer fusion is an emotion classifier constructed for a single modality in multimodality, and the emotion classification results are obtained by using each classifier, and then fused according to the fusion strategy of the decision-making layer to obtain the final classification result. In the research on multimodal decision layer fusion, common fusion rules include mean rule, product rule, maximum rule, minimum rule, and sum rule. The decision-level fusion strategy is to use n classifiers to classify a test sample x under m classification conditions and obtain a posterior probability set $\{p_{ij}(x), i = 1,2, \ldots, n, j = 1,2, \ldots, m\}$, where i is the classifier and j is the category. The author's multimodal emotion recognition requires two classifiers to fuse according to rules to obtain a new probability set $\{q_j(x), j = 1,2, \ldots, m\}$ and finally select the largest new_$q_j(x)$ as the final recognition result; the author mainly adopts summation rules and fusion for multimodal emotion recognition of speech and facial expressions, which should meet the following equation:

$$\begin{cases} q_j(x) = \sum_{i=1}^{2} p_j(x) \\ \text{new\_}q_j(x) = \dfrac{q_j(x)}{\sum_j q_j(x)}. \end{cases} \tag{2}$$

4 Results and analysis

4.1 Multimodal emotion database and experimental environment

To evaluate the performance of the model, the author conducted multimodal emotion recognition experiments on the eNTERFACE'05 multimodal emotion database, the database was collected from 34 males and 8 females from 14 different countries, including six basic emotions: anger, disgust, fear, happiness, sadness, and surprise; the author used five sentences to express each emotion [29]. The author selected all the data in the database as the experimental subjects, extracted corresponding speech samples from the video file, and used screenshot software to evenly intercept facial expressions in the video sequence. The author also selected the image that best expresses this emotion as the expression sample, obtained 1,260 speech and expression samples each, and removed invalid information from the edge of the sample, so that each image has 256 × 256 pixels. An image of six basic emotions of an object is presented. This experiment mainly focuses on processing speech signals and images, so Tensorflow and Keras were selected as deep learning frameworks. The eNTERFACE'05 database introduced above was used for testing on Windows 7 (×64), Python 2.7, and AMD18-4500M [30].

4.2 Simulation experiments

We extracted an equal number of voice and emoticon samples from the eNTERFACE'05 databases to maintain a balance between the two types of data. We trained a total of 1,260 voice samples and 1,260 emoticon samples on the network model, with 80% of the samples used for training and the remaining samples used for testing. Tables 1 and 2 show the SER results and expression recognition results for six emotional types: anger, disgust, fear, happiness, sadness, and surprise, using the experimental parameters consistent with the above experiments. It contains 42 subjects from 14 different nationalities (81% of the respondents were male, the remaining 19% were female, 31% wore glasses, and 17% were female).

We conducted the experiments in English. We instructed each subject to listen to six consecutive short stories, each triggering a specific emotion. They then responded to each condition, and two human experts judged whether the response had clearly expressed emotion. If this is true, the database adds the samples. Processing the video sequence is done at a frame rate of 25 frames per second using the 72Q 576 Microsoft AVl format. The pixel aspect ratio is D1 LRV PAL (1.067e). To ensure easy porting, we compress the video using the DivX 5.0.5 codec. The unexpectedly high-frequency sampling rate is 48,999 using the uncompressed-sound 16-bit format. Finally, the database contains a total of 1,166 video sequences. Of the 1,166 video sequences, 264 involved female recordings (23%), and 902 involved male recordings (77%).

Table 1: Unimodal speech and emotion recognition results (average)

Emotions	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Anger	0.68	0.15	0.03	0.11	0.01	0.02
Hate	0.02	0.86	0.02	0.07	0	0.03
Fear	0.11	0.21	0.55	0.04	0.01	0.08
Happy	0.02	0.18	0.05	0.69	0	0.06
Sadness	0.03	0.10	0.03	0.02	0.78	0.04
Surprised	0.03	0.18	0.04	0.07	0.01	0.67
Average recognition rate	70.5%					

Table 2: Unimodal expression recognition results (average)

Emotions	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Anger	0.56	0.29	0.07	0.02	0.02	0.04
Hate	0.01	0.90	0.03	0.01	0.02	0.03
Fear	0	0.04	0.89	0.03	0.01	0.03
Happy	0	0.06	0.04	0.85	0.03	0.02
Sadness	0.01	0.03	0.05	0.04	0.86	0.01
Surprised	0.02	0.06	0.14	0.02	0.01	0.75
Average recognition rate	80.17%					

The analysis in Table 1 reveals a recognition rate of 70.5% for these six single-modal speech emotions. Among them, the correct recognition rates for disgust and sadness are higher, at 86 and 78%, respectively. However, the recognition rate of fear is less than 60% – only 55% – which can be considered anger and disgust. The quality of multimodal emotion databases has a certain impact on emotion recognition research. The analysis in Table 2 reveals that these six single-modal expressions have a recognition rate of 80.17%. Among them, the recognition rates for disgust, fear, happiness, and sadness are relatively high, all above 85%, but the recognition rate for anger is low, only 56%, and it is considered disgusted. Analyzing the emotions of anger and disgust, anger is characterized by furrowed brows, a gaze, and nasal expansion, while disgust is characterized by furrowed brows, narrowed eyes, and a wrinkled nose.

Establishing a high-quality multimodal emotional database is crucial due to the low recognition rate caused by the inaccurate facial expression images captured in the video. Table 3 shows the results of multimodal emotion recognition for speech and facial expressions based on feature fusion. Table 3 shows that the average recognition rate for multimodal emotion recognition of speech and expression based on feature fusion is 83.67%. Table 4 shows the emotional recognition results obtained through decision fusion based on summation rules. According to Table 4, we can see that the average recognition rate of emotion recognition by decision fusion based on sum rules is 87.3%. Table 5 shows the results of emotion recognition based on the improved fusion algorithm. Table 5 shows that the average recognition rate of emotion recognition based on the improved fusion algorithm is 89.3%.

Table 3: Emotion recognition results based on feature fusion (average)

Emotions	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Anger	0.74	0.15	0.02	0.04	0.03	0.02
Hate	0.01	0.87	0.03	0.05	0.02	0.02
Fear	0.02	0.08	0.85	0.02	0.02	0.01
Happy	0.03	0.01	0.03	0.87	0.04	0.02
Sadness	0	0.03	0.04	0.01	0.89	0.03
Surprised	0.02	0.01	0.07	0.06	0.04	0.8
Average recognition rate	83.67%					

Table 4: Emotion recognition results based on decision layer fusion (average)

Emotions	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Anger	0.74	0.15	0.02	0.04	0.03	0.02
Hate	0.01	0.87	0.03	0.05	0.02	0.02
Fear	0.02	0.08	0.85	0.02	0.02	0.01
Happy	0.03	0.01	0.03	0.87	0.04	0.02
Sadness	0	0.03	0.04	0.01	0.89	0.03
Surprised	0.02	0.01	0.07	0.06	0.04	0.8
Average recognition rate	87.3%					

Table 5: Emotion recognition results based on improved fusion algorithm (average)

Emotions	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Anger	0.82	0.02	0.05	0.04	0.06	0.01
Hate	0.01	0.94	0	0.03	0.01	0.01
Fear	0.02	0.01	0.89	0.03	0.03	0.02
Happy	0	0.02	0.01	0.92	0.02	0.03
Sadness	0	0	0.03	0.02	0.94	0.01
Surprised	0.01	0.03	0.02	0.05	0.04	0.85
Average recognition rate	89.3%					

The analysis of the above table shows that the average recognition rate of the six emotions based on feature fusion is 83.67%, while the average recognition rate of the decision layer fusion based on summation rules is 87.3%. Compared to the results of single-modal emotion recognition, the improved fusion algorithm in multimodal emotion recognition achieves an average recognition rate of 89.3%. The author's improved algorithm achieved a higher recognition rate compared to the algorithm based on feature fusion and decision layer fusion, resulting in an improvement in the recognition rate of each emotion. Table 6 displays the results of single-modal and multimodal emotion recognition. To compare the recognition results of emotion recognition methods, different identification methods were plotted into line plots, as shown in Figure 3 (horizontal axis from 1 to 7: anger, disgust, fear, joy, sadness, surprise, and average recognition rate) [31].

From the analysis in Figure 3, it can be seen that the emotion recognition based on eNTERFACE'05 audio and video multimodal emotion database has an accuracy of 70.5% for single modal SER, 80.17% for single modal expression recognition, and 83.67% for multimodal emotion recognition based on feature fusion, the accuracy of multimodal emotion recognition based on summation rule decision layer fusion is 87.3%, while the accuracy of multimodal emotion recognition based on improved fusion algorithm can reach 89.3%, which is significantly improved compared to other multimodal emotion recognition accuracy rates of 60.09 and 83.92%, this indicates that this method has certain improvements in emotion recognition, and experimental results show that the performance of multimodal emotion recognition combining speech and facial expressions is better than that of single modal emotion recognition [32].

Figure 4 presents the results of the experimental analysis of the proposed approach for several motions. The percentage values are also presented in this experimental analysis in Table 7.

Table 6: Accuracy of emotional recognition by different methods

Method	Average recognition rate (%)
Single-modal SER results	70.5
Single-modal facial expression recognition results	80.17
Emotional recognition results based on feature fusion	83.67
Emotional recognition results based on decision layer fusion	87.3
Emotion recognition results based on improved fusion algorithm	89.3

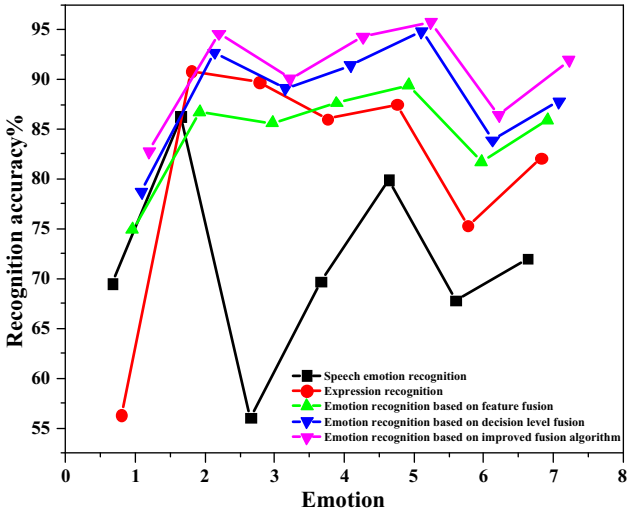

Figure 3: Average recognition rates of different methods.

The experimental analysis results demonstrate the effectiveness and resilience of the proposed emotion recognition approach. The single-modal speech method demonstrated an average recognition rate of 77.4% when applied to a range of emotions, whereas the single-modal expression method achieved 84.3%. The fusion of features improved recognition by 87.5%, and the fusion of decisions increased it to 90.8%. The novel algorithm, Improved Fusion Algorithm, demonstrated the maximum level of accuracy, attaining a noteworthy average recognition rate of 92.8%. The present comparative analysis demonstrates that the proposed algorithm surpasses both single-modal approaches and conventional fusion methods, highlighting the substantial progress made in the field of multimodal emotion recognition. The outcomes indicate the potential for real-world implementations across various domains, underscoring the proposed method's superiority in capturing intricate emotional manifestations. Table 8 presents the comparative analysis of the proposed approach with existing studies [7,11,13]. The comparative analysis reveals that our proposed approach outperforms existing studies across multiple critical parameters. With a remarkable recognition rate of 92.8%, our approach, combining speech and facial expressions, surpasses Zheng et al.'s [7] method by 9.3% points and Zhang et al. [11] by a substantial margin of 27.8% (Figure 5).

The advanced fusion algorithm implemented showcases the effectiveness of our dual-layer fusion strategy, contributing to this superior performance. Notably, our method excels in utilizing a multimodal dataset (eNTERFACE'05), ensuring a robust evaluation environment comparable to other studies. Moreover, the fusion strategy employed in our approach, namely the Improved Fusion Algorithm, demonstrates its effectiveness against diverse strategies like wave off-field push, cross-mode fusion, and kernel fusion, contributing to a significant accuracy boost. These results affirm the innovation and efficacy of our proposed approach, positioning it as a standout solution in the field of multimodal emotion recognition.

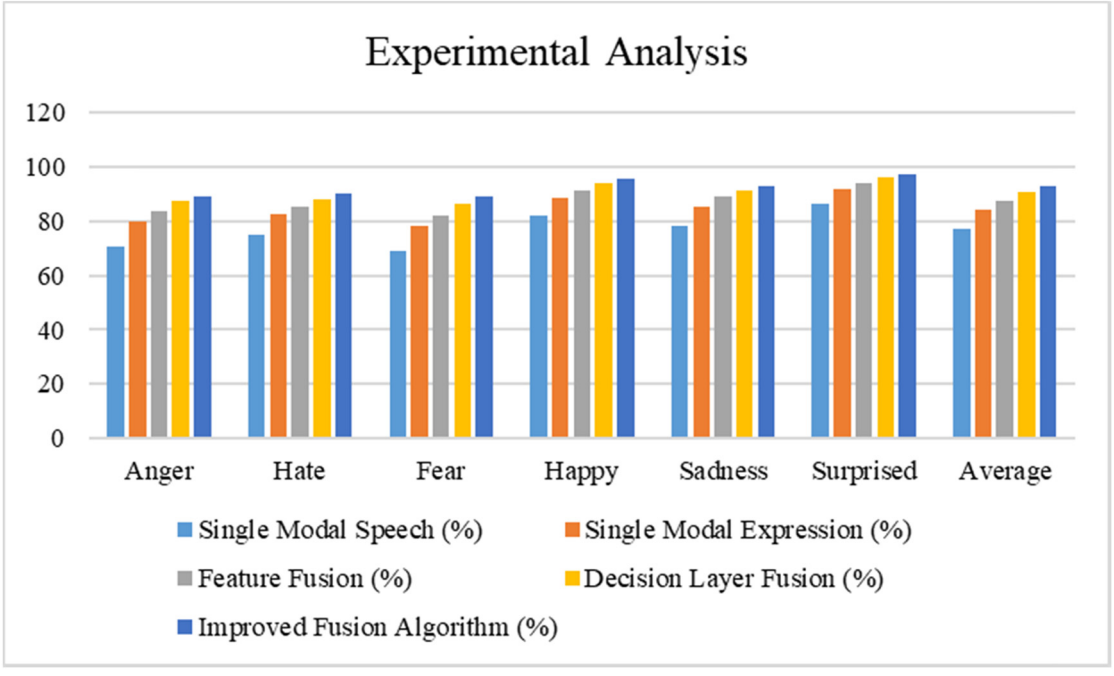

Figure 4: Experimental analysis of the proposed approach.

Table 7: Experimental analysis of proposed approach in (% age)

Motion	Single modal speech (%)	Single modal expression (%)	Feature fusion (%)	Decision layer fusion (%)	Improved fusion algorithm (%)
Anger	70.5	80.17	83.67	87.3	89.3
Hate	75.2	82.45	85.1	88.2	90.1
Fear	68.9	78.3	82	86.5	88.9
Happy	82	88.7	91.2	93.8	95.5
Sadness	78.3	85.6	88.9	91.5	93.2
Surprised	86.5	92.1	94.3	96	97.2
Average	77.4	84.3	87.5	90.8	92.8

5 Discussion

In this section, we elaborate on the potential of our multimodal emotion recognition system to transform robot behavior in the real world. Adding emotional intelligence to robots could change everything, especially when it comes to improving relationships between humans and robots. The system is good at reading both spoken and unspoken emotional cues, which leads to more natural and understanding exchanges that improve the way people talk to each other. Our system is also adaptive, which means that robots can change how they help and support

Table 8: Comparative analysis of proposed approach with existing studies [7,11,13]

Parameters	Proposed approach	Zheng et al. [7]	Zhang et al. [11]	Wang et al. [13] (Kernel Fusion)
Recognition rate (%)	92.8	83.5	65	83.92
Modalities utilized	Speech, Facial Exp.	Speech, Text, Action	Audio, Video, Text	Speech, Facial Exp.
Fusion strategy	Improved Fusion Alg.	Wave Off-Field Push	Cross-Mode Fusion	Kernel Fusion
Database used	eNTERFACE'05	MELD	eNTERFACE'05	eNTERFACE'05
Average experiment accuracy (%)	0.928	0.835	0.65	0.8392

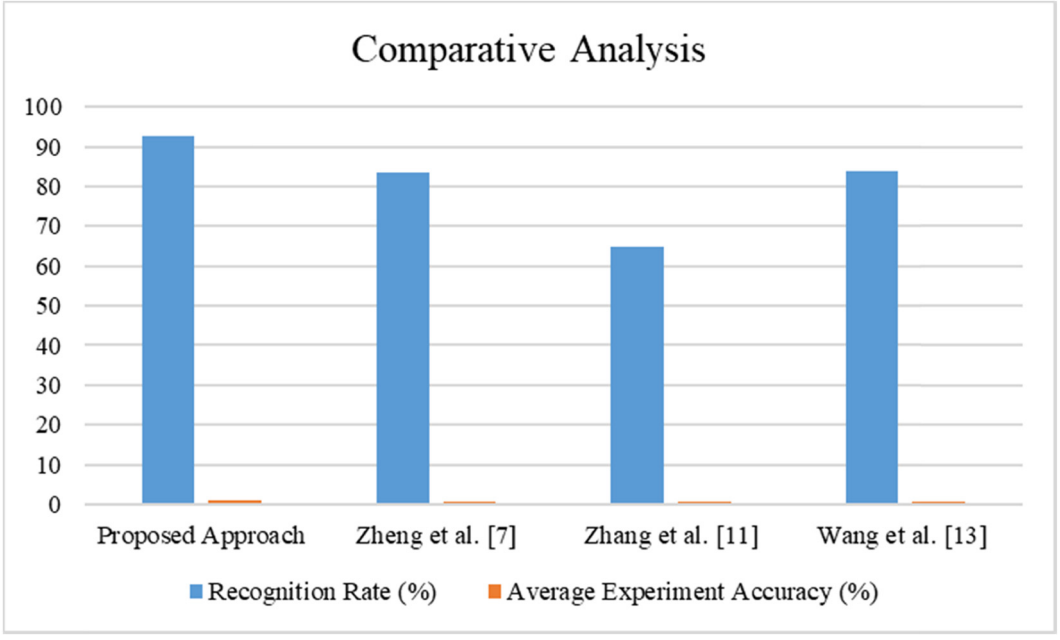

Figure 5: Comparative analysis of the proposed approach.

people based on the feelings they detect. This customized method improves the level of care in places like healthcare settings. Our method helps robots become more emotionally intelligent, which makes them better team members who can work together and understand each other. This emotional intelligence makes relationships more peaceful and helps people accept robots as helpful team members.

This has effects on socially aware navigation and interaction, where robots using our approach can move through crowded areas with a greater awareness of how people are feeling. This skill makes sure that people treat each other with respect and care, and it deals with social limits in public places. Even with these improvements, though, moral concerns are still the most important thing. We stress how important it is to have strong privacy protections to keep user consent and sensitive emotional data safe. In the end, our multimodal mood recognition system is useful in real life as well as being a step forward in technology. This changes how people and robots communicate, making robots more understanding, responsive, and socially aware. The effects we talked about show how possible it is to make robots understand and respond to emotions. This is a big step toward making robot friends that are smart and sensitive to our feelings.

The practical integration of our algorithm into robot behavior involves navigating various challenges for seamless real-world implementation. In a family setting, the algorithm's effectiveness may face hurdles such as ambient noise and dynamic lighting conditions. Incorporating noise cancellation techniques and adaptive algorithms ensures robust performance, addressing challenges such as ambient noise and dynamic lighting conditions. Real-time responsiveness becomes crucial, prompting the need for efficient processing and immediate feedback. Handling incorrect identifications demands a thoughtful approach, potentially involving user feedback loops to enhance the algorithm's learning over time. Considerations extend to user-centric design principles, ethical implications, and the establishment of human–robot trust. The adaptability of the algorithm to diverse human expressions and emotions should be a focal point, enhancing its applicability in dynamic, real-world scenarios. As robots increasingly become part of daily life, our algorithm's successful integration hinges on a comprehensive understanding of the nuanced challenges and considerations inherent in human–robot interactions.

6 Conclusion

This research has delved deeply into critical aspects including preprocessing, feature extraction, fusion strategies, and emotion recognition classification for both speech signals and expression images. The suggested method for detecting emotions is based on a dual fusion strategy at both the feature layer and the decision layer. It takes advantage of the fact that different sources of emotional information work well together. Through extensive experimentation on the eNTERFACE'05 audio and video multimodal emotion database, we achieved a commendable recognition rate of 89.3%. Comparative analysis revealed a substantial improvement in multimodal emotion recognition over single-mode recognition. The applications of emotion recognition have found widespread use in diverse fields such as remote education, customer service, and safe driving, attesting to the practical significance of this research. While the advent of convolutional neural networks has introduced novel approaches to the study of emotion recognition, challenges persist, particularly in the realm of multimodal emotion recognition. Notably, there is a pressing need for the establishment of a comprehensive and standardized multimodal emotion database. The study opens avenues for future research in several directions. First, further exploration and refinement of fusion strategies can contribute to enhancing the robustness and generalizability of multimodal emotion recognition systems. Additionally, the incorporation of advanced machine learning techniques, including deep learning architectures, could yield improvements in accuracy and efficiency. Moreover, addressing the existing challenges in establishing a comprehensive and standardized multimodal emotion database is imperative.

Funding information: This study did not receive any funding in any form.

Author contributions: The author has made important personal contributions to this manuscript. Lulu Nie: writing and performing experiments; data analysis and performing experiments; article review and intellectual concept of the article.

Conflict of interest: The author declares no competing interests.

Data availability statement: The data used to support the findings of this study are available from the corresponding author upon request.

References

[1] S. Lee, D. K. Han, and H. Ko, "Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification," *IEEE Access*, vol. 9, pp. 94557–94572, 2021.

[2] L. Li and M. Yang, "A context-aware gated recurrent units with self-attention for emotion recognition," *Journal of Physics: Conference Series*, vol. 1880, no. 1, 2021, p. 012026.
[3] S. Veni, R. Anand, D. Mohan, and E. Paul, "Feature fusion in multimodal emotion recognition system for enhancement of human-machine interaction," *IOP Conference Series: Materials Science and Engineering*, vol. 1084, no. 1, 2021, p. 012004.
[4] M. Ren, X. Huang, X. Shi, and W. Nie, "Interactive multimodal attention network for emotion recognition in conversation," *IEEE Signal. Process. Lett.*, vol. 28, pp. 1046–1050, 2021.
[5] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition using a hierarchical fusion convolutional neural network," *IEEE Access*, vol. 9, pp. 7943–7951, 2021.
[6] T. Mittal, A. Bera, and D. Manocha, "Multimodal and context-awareemotion perception model with multiplicative fusion," *IEEE Multimed.*, vol. 28, no. 2, pp. 67–75, 2021.
[7] C. Zheng, C. Wang, and N. Jia, "Emotion recognition model based on multimodal decision fusion," *Journal of Physics: Conference Series*, vol. 1873, no. 1, 2021, 012092.
[8] B. Xie, M. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, p. 4913, 2021.
[9] N. Nigam, Y. Zhang, P. Chen, G. Wolfe, T. Pillsbury, N. M. Wereley, et al., "Adaptive control and actuation system development for biomimetic morphing," in *24th AIAA/AHS Adaptive Structures Conference*, 2016, p. 1084.
[10] J. Wang, S. Wang, M. Lin, Z. Xu, and W. Guo, "Learning speaker-independent multimodal representation for sentiment analysis," *Inf. Sci.*, vol. 628, pp. 208–225, 2023.
[11] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert. Syst. Appl.*, vol. 237, p. 121692, 2023.
[12] I. Gangopadhyay, A. Chatterjee, and I. Das, "Face detection and expression recognition using Haar cascade classifier and Fisherface algorithm," in *Recent Trends in Signal and Image Processing: Proceedings of ISSIP 2018*, Singapore, Springer, 2019, pp. 1–11.
[13] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, et al., "A systematic review on affective computing: Emotion models, databases, and recent advances," *Inf. Fusion.*, vol. 83, pp. 19–52, 2022.
[14] D. Cannata, S. Redfern, and D. O'Hora, "OpenFaceR: Developing an R Package for the convenient analysis of OpenFace facial information," in *PSYCHOBIT*, 2020, October.
[15] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun*, vol. 116, pp. 56–76, 2020.
[16] Y. Fang, R. Rong, and J. Huang, "Hierarchical fusion of visual and physiological signals for emotion recognition," *Multidimension. Syst. Signal. Process.*, vol. 32, pp. 1103–1121, 2021.
[17] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, and G. Dai, "Pre-trained deep convolution neural network model with attention for speech emotion recognition," *Front. Physiol.*, vol. 12, p. 643202, 2021.
[18] X. Zhong, Y. Gu, Y. Luo, X. Zeng, and G. Liu, "Bi-hemisphere asymmetric attention network: recognizing emotion from EEG signals based on the transformer," *Appl. Intell.*, vol. 53, no. 12, pp. 15278–15294, 2023.
[19] S. Hossain, S. Umer, V. Asari, and R. K. Rout, "A unified framework of deep learning-based facial expression recognition system for diversified applications," *Appl. Sci.*, vol. 11, no. 19, p. 9174, 2021.
[20] A. Tripathi, T. S. Ashwin, and R. M. R. Guddeti, "EmoWare: A context-aware framework for personalized video recommendation using affective video sequences," *IEEE Access*, vol. 7, pp. 51185–51200, 2019.
[21] T. Chen, H. Yin, X. Yuan, Y. Gu, and X. Sun, "Emotion recognition based on fusion of long short-term memory networks and svms," *Digital Signal. Process.*, vol. 117, no. 1, p. 103153, 2021.
[22] X. Wang, Y. Chang, V. Sugumaran, X. Luo, and H. Zhang, "Implicit emotion relationship mining based on optimal and majority synthesis from multimodal data prediction," *IEEE Multimedia*, vol. 28, no. 2, pp. 96–105, 2021.
[23] Y. Zhang, G. Zhao, Y. Shu, Y. Ge, and X. Sun, "Cped: a chinese positive emotion database for emotion elicitation and analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1417–1430, 2021.
[24] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomed. Signal. Process. Control.*, vol. 70, no. 9212, p. 103029, 2021.
[25] Y. Zhu, Y. F. Tang, and Y. X. Wu, "Design of character recognition robot based on fpga," *Journal of Physics: Conference Series*, vol. 1748, no. 4, 2021, p. 042036.
[26] H. W. Chuah and J. Yu, "The future of service: the power of emotion in human-robot interaction," *J. Retail. Consum. Serv.*, vol. 61, no. 3, p. 102551, 2021.
[27] W. Sheng, M. Pham, H. M. Do, Z. Su, and A. J. Bishop, "Negative emotion management using a smart shirt and a robot assistant," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 4040–4047, 2021.
[28] H. Dong, W. Song, B. Luan, and G. Li, "Autonomous recognition technology of carrier robot on various terrain environment," *Proc. Inst. Mech. Eng. Part. D: J. Automob. Eng.*, vol. 235, no. 9, pp. 2568–2584, 2021.
[29] Q. Wang and D. Li, "Multimodal soft jumping robot with self-decision ability," *Smart Mater. Struct.*, vol. 30, no. 8, p. 085038, 2021.
[30] L. Lu, H. Wang, B. Reily, and H. Zhang, "Robust real-time group activity recognition of robot teams," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2052–2059, 2021.
[31] S. Liu, G. Tian, Y. Zhang, and P. Duan, "Scene recognition mechanism for service robot adapting various families: a cnn-based approach using multi-type cameras," *IEEE Trans. Multimed.*, vol. 24, pp. 2392–2406, 2021.
[32] Z. Sun, X. Guo, X. Zhang, J. Han, and J. Hou, "Research on robot target recognition based on deep learning," *Journal of Physics: Conference Series*, vol. 1948, no. 1, 2021, p. 012056.