2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)/2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud (EdgeCom)

Deep Learning Emotion Recognition Method

Weidong Xiao
Schoole of Software
Engineering Xiamen
University of Technology
Xiamen 361024, China
xiaoweidong@xmut.edu.cn

Wenjin Tan
School of Computer Science and
Engineering,Hunan University
of Science and Technology
Xiangtan 411201,China
13370639594@163.com

Naixue Xiong
School of Computer Science and
Engineering,Hunan University
of Science and Technology
Xiangtan 411201,China
dnxiong@126.com

Ce Yang
School of Computer Science and
Engineering,Hunan University
of Science and Technology
Xiangtan 411201,China
554734744@qq.com

Lin Chen
School of Computer Science and
Engineering,Hunan University
of Science and Technology
Xiangtan 411201,China
lynn070021@163.com

Rui Xie
School of Computer Science and
Engineering,Hunan University
of Science and Technology
Xiangtan 411201,China
xierui54@mail.hnust.edu.cn

***Abstract*—Emotion recognition refers to the process of actively analyzing human emotions through computer technology, and it has become an important part of modern society. Traditional emotion recognition is mainly based on a single information source, such as text, speech, video, etc., from which emotional features are extracted for classification or regression to recognize human emotions. With the development of artificial intelligence technology, multimodal emotion recognition is gradually becoming widely used. It combines two or more types of information, such as text, speech, and visual information, in different ways to analyze emotions. Multimodal emotion recognition is far superior to a single modality in understanding emotions. This article mainly analyzes the technology of emotion analysis. Firstly, we introduce the basic concepts and research status of emotion recognition. Then, we introduce the main types of emotion recognition and describe various methods used in the process in detail. Finally, we discuss the challenges and future developments of emotion recognition.**

Keywords*—multimodal, emotion, feature extraction, emotion fusion, emotion classification.

I. INTRODUCTION

Emotion recognition technology utilizes various technical means to identify emotions expressed in different media, such as text, audio, and images, and has broad applications in fields such as social media analysis, mental health, and intelligent customer service. It has become a hot research topic in recent years. In the business field, emotion recognition can automatically capture the correlation between different emotional features, helping companies understand consumer needs and feedback and increase sales and customer satisfaction [1]. In the medical field, emotion recognition can help doctors better understand the emotions and psychological conditions of patients, thus providing more personalized and effective treatment plans[2]. In the education field, emotion recognition can effectively analyze students' learning status and emotions, thus improving their academic performance[3]. In the multimedia field, emotion recognition can be applied to product satisfaction, film and television quality reviews, and news event analysis, among other areas, to improve the user experience. [4].

Identifying different emotional states in humans is a rather specific task. Carroll Izard and his research team considered the complexity and diversity of emotions and refined them into eight categories: joy, adoration, empathy, anger, fear, surprise, disgust, and sadness[5]. In the exploration of emotion analysis, most research is based on a single mode, which cannot fully capture the diversity and complexity of emotional expressions. Shu et al. reviewed physiological signal methods for emotion recognition, including measuring physiological indicators such as heart rate, skin conductance, and muscle activity and inferring emotional states by analyzing these indicators[6]. Zou et al. found that by using multiple CNNs and fusing their results in the classifier, the accuracy of facial expression recognition was improved [7].

To overcome the limitations of a single modality, in recent years, research on emotion analysis has begun to combine two or more modalities to achieve cross-modal emotion recognition. Multimodal emotion recognition can use multiple types of information obtained from different sensors, such as sound, facial expressions, physiological signals, etc., to obtain data from different perspectives to compensate for missing information and integrate and analyze this information to improve the accuracy of emotion recognition. Figure 1 shows the framework of emotion analysis.

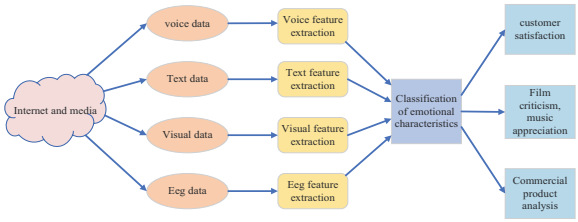

Fig. 1. Framework for Emotion Recognition

II. TYPES OF EMOTION ANALYSIS

In order to obtain accurate and effective emotional features, researchers have begun to explore various fields based on visual information, textual information, speech information, etc. These modalities can be used separately for emotion recognition or combined together to perform multimodal fusion for emotional analysis.

A. Emotion Analysis Based on Visual

Research on visual-based emotion analysis refers to the analysis and processing of images or videos to obtain emotional information about the people in the images or videos. In recent years, with the advancement of visual processing technology, visual emotion analysis has become a popular research field focused on solving the problem of predicting emotions in images. Hossain et al. [8] provided an overview of the history and development of visual emotion analysis and introduced various visual emotion datasets and evaluation metrics, such as SentiBank, MIR Flickr, and NRC Emotional Intensity, which provide a more comprehensive

979-8-3503-1246-1/23/$31.00 ©2023 IEEE
DOI 10.1109/CSCloud-EdgeCom58631.2023.00067

2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)/2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud (EdgeCom) | 979-8-3503-1246-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSCloud-EdgeCom58631.2023.00067

background and related knowledge for understanding visual emotion analysis. Meanwhile, Ko et al. [9] briefly reviewed research on visual emotion analysis, and the mixed CNN-LSTM method proposed can outperform the CNN method using time averaging for aggregation. In visual recognition, facial recognition is a typical method, and facial expressions are crucial for better understanding the emotions of people of different ages. However, the development of this feature has been hindered due to the lack of strongly labeled data for model training. To improve this situation, Huang et al. [10] proposed a facial deformation-based data augmentation method in the literature, using a technique called "morphable model" for feature point matching and deformation of two facial images to generate a new facial image, thereby improving the performance of emotion intensity recognition.

B. *Emotion Analysis Based on Speech*

Research on speech-based emotion analysis uses a technology that uses speech signals to identify and analyze the emotional state of a speaker. Speech contains a lot of semantic information, such as intonation, speech rate, volume, pauses, etc. Compared with traditional image-based emotion analysis, speech-based emotion analysis has the advantage of more directly reflecting a person's emotional state. When a person makes a sound, the vocal tract filters the sound. [11] Vocal tract features can be well described in the frequency domain, and the most widely used spectral feature in automatic speech recognition is Mel-frequency cepstral coefficients (MFCC). Segmentation of speech into different segments can be used to obtain MFCC and speech features for emotion recognition [12]. However, in traditional speech emotion recognition, the efficiency of the algorithm heavily depends on the quality of handcrafted acoustic features, which can lead to problems such as a lack of high-quality data and insufficient model accuracy. Shchetinin et al. [13] analyzed some traditional emotion recognition methods and their limitations and introduced methods and techniques for using deep learning models to recognize human emotions in speech, including commonly used models such as CNN, long short-term memory networks (LSTM), and bidirectional recurrent neural networks (BRNN), which demonstrate the advantages of deep learning technology in speech emotion recognition. Meanwhile, Kang et al. [14] introduced how to use deep learning models for modeling and representation learning of speech data, including preprocessing, feature extraction, and selection.

C. *Emotion analysis based on text.*

Emotion recognition based on textual data is a widely used technique for emotion classification. Identifying human emotions from written texts and user conversations can help people better understand and analyze users. [15]. Saffar et al. [16] introduced the progress and applications of text emotion detection in the health field and summarized the current technologies and algorithms used for text emotion detection, including dictionary-based methods, machine learning-based methods, and deep learning-based methods. Analyzing emotions from text is mainly based on emotional vocabulary [17]. If contextual dependencies in dialogues are not considered, general text classification methods can be used to solve emotion classification problems in dialogue text. However, because dialogues themselves have many elements, such as short text length and dynamic contextual information, emotional recognition of utterances is not simply equivalent to emotional recognition of individual text sentences but requires comprehensive consideration of relevant information [18]. To enhance the emotional association of context, Shu et al. [19] used a combination of multiple classifiers, such as dictionary-based methods, naive Bayes classifiers, support vector machines, and decision trees, to generate the final classification results from the predicted results.

D. *Multimodal Fusion-based Emotion Analysis.*

The emotion analysis methods described above are all based on considering a single modality, while multimodal emotion analysis is a method of combining multiple sensory modalities (such as audio, video, text, and images) to perform emotion analysis. The multimodal emotion recognition framework is shown in Figure 2. It can utilize the complementarity between different modalities to improve the accuracy and robustness of emotion recognition [20]. In recent years, research on multimodal emotion recognition has received increasing attention. In the study by Tan et al. [21], a multimodal emotion analysis method based on facial expressions and EEG was proposed. Useful features were extracted from visual and brain information, and the extracted features were fused. Support vector machines (SVM) and multilayer perceptrons (MLP) were then used to classify the fused features based on emotion. In order to obtain more comprehensive emotion features, Huang et al. [22] proposed a multimodal data-based emotion analysis method that aims to form a multidimensional feature vector by extracting features from different modalities of data, such as audio, video, and text, and comprehensively considering the emotional state from multiple perspectives.

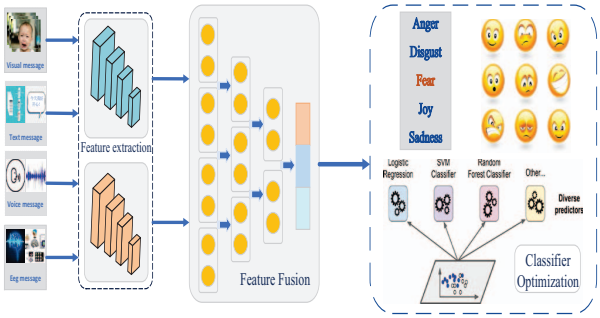

Fig. 2. Framework of Multi-Modal Emotion Recognition

III. Emotion Recognition Feature Processing Techniques Research.

The feature processing of emotion recognition plays a crucial role in the effectiveness and reliability of emotion analysis methods. Sleeman et al. [23] reviewed the development and related research of emotion classification and divided the feature processing of emotion recognition into two main processes: data preprocessing and feature extraction. In this section, we will discuss the relevant techniques and methods for feature processing in emotion recognition based on these two processes.

A. *Preprocessing process*

In emotion recognition, the preprocessing step is a crucial part that can perform various operations on input data such as cleaning, normalization, dimensionality reduction, and data augmentation to improve the accuracy, efficiency, generalization ability, and robustness of the algorithms. The preprocessing step can ensure the stability and reliability of the input data quality while also enhancing the correlation and

consistency among different modalities, thereby improving the recognition performance of the algorithms. Therefore, a good preprocessing process can have a direct impact on the results of emotion analysis. In this section, we will introduce different feature processing operations for various modalities in the following four aspects:

a) Visual Information Preprocessing

The visual modality for emotion recognition typically refers to facial images or videos. To extract features from visual information, facial information needs to be obtained first. Nguyen et al. [24] used a method based on the Haar cascade classifier for face detection and aligned the face by detecting facial landmarks in the image. Chen et al. [25] used a pixel-based method to extract image features of facial expressions and a geometry-based method to capture the dynamic features of facial expressions, which preprocessed the data. However, when training deep neural networks, a large amount of data is usually needed to train an accurate model. Zhu et al. [26] used a generative adversarial network called Conditional Wasserstein GAN, which can generate synthetic images based on the input emotion label, to expand the size of the training dataset and increase the diversity of the data.

b) Preprocessing of Audio Information.

Speech signals are high-dimensional and unstructured signals that contain a large amount of information. Song et al. [27] performed preprocessing operations such as pre-emphasis, framing, windowing, and short-time Fourier transform on the raw speech signals to improve the accuracy of cross-corpus emotion recognition. Meanwhile, Deng et al. [28] also performed a series of preprocessing operations on speech signals in speech emotion recognition, including speech endpoint detection, silence removal, speech segmentation, and speech length normalization, to improve model robustness. In order to reduce the dimensionality and redundant information of speech information, Daneshgar et al. [29] optimized the backpropagation algorithm to automatically select and learn the most discriminative features, thereby reducing the dimensionality and enhancing the discriminability of emotion features.

c) Text information preprocessing.

Preprocessing is a crucial step in emotion recognition, and for text data, it involves feature extraction and dimensionality reduction. Methods such as principal component analysis and linear discriminant analysis can be used to reduce the dimensionality of text features, thereby improving the accuracy and efficiency of emotion recognition. Li et al. [30] proposed a deep learning-based emotion recognition system that uses self-supervised learning, feature selection, and dimensionality reduction techniques to optimize the feature extraction and preprocessing processes to improve model performance and efficiency. In addition, since natural language text contains various noises and errors, the text data also needs to be cleaned while reducing dimensionality. Goyal et al. [31] reduced noise and unnecessary information in the text by removing stop words and transforming words into their basic forms to reduce variations and repetitions in language.

B. Feature Extraction

The core task of emotion recognition is to extract useful features from different modalities of input signals to help identify different emotional states. Feature extraction is an important component of emotion recognition research, which aims to extract discriminative features from raw data and use them as inputs for emotion classifiers. This section will introduce various modality-specific feature extraction methods.

a) Visual feature extraction

Visual feature extraction can be divided into local feature extraction and global feature extraction. Local features refer to a certain part of the image, such as the SIFT and SURF algorithms, which are classic local feature extraction algorithms. Agarwal et al. [32] proposed an emotion recognition method based on local feature extraction, which uses multiple local feature descriptors and visual word techniques to model and represent emotion-related information. Global features can obtain more comprehensive feature information. Ding et al. [33] proposed a joint local and global feature extraction method based on deep learning, which uses a convolutional neural network to extract global features and then extracts local features at each convolutional layer to achieve visual emotion recognition.

b) Speech Feature Extraction

Pitch, intensity, MFCC, spectral centroid, and other audio features play important roles in emotion analysis. Among them, MFCC has good human auditory perception performance and stability, and is one of the commonly used features in speech emotion recognition. Fayek et al. [34] conducted pre-emphasis and framing processing on the audio signal in the process of studying speech emotion recognition, and then used MFCC to extract 13-dimensional features for each frame. These features were treated as input vectors and input into a DNN for training and classification. In addition, there are also some emerging feature extraction methods, such as deep neural network, and feature extraction methods based on dictionary learning and sparse coding. Lu et al. [35]used CNN as a feature extractor. By using a deep convolutional neural network composed of multiple convolutional layers and pooling layers to perform convolution and pooling operations on the spectrogram of the speech signal, high-level abstract features of the speech signal were extracted, thereby achieving recognition of speech emotion.

c) Text Feature Extraction

In the research on emotion recognition in text information, the feature extraction method based on the bag-of-words model is widely used. This method treats text as a bag of words and represents each text document as a vector of word counts, thereby extracting text features. Miller et al. [36] used the bag-of-words model to remove non-alphabetic characters such as punctuation marks and numbers from the text in text emotion analysis, divided each text into words, and then used the text-word matrix as the feature representation of the text. In recent years, a series of deep learning-based text feature extraction methods have been proposed, such as word embedding models, convolutional neural networks, and recurrent neural networks. Shen et al. [37] proposed a hierarchical attention network that uses word embedding to extract features from documents for emotion analysis. The hierarchical attention network can capture the relationships between words in the document and assign higher weights to more important words.

IV. Research on Multimodal Emotion Fusion Technology.

During the research of emotion recognition, it is often necessary to combine practical applications and fuse various pieces of information from different sensors. In multimodal emotion recognition, commonly used modal information includes speech, images, biometrics, etc. Table 1 analyzes the commonly used multimodal datasets. According to research in the field of multimodal emotion recognition, multimodal fusion technology can be divided into the following three categories: feature fusion, decision fusion, and model fusion.

TABLE I. Common Multimodal Datasets

Data set name	language	Modal type	Source
YouTube	English	Speech, Text, Video	stratou@ict.usc.edu
MELD	English	Speech, Text, Video	https: //affective-meld.github.io/
MOUD	English	Speech, Text, Video	http://web.eecs.umich.edu/~mihalcea/downloads.html
CMU-MOSI	English	Face, Speech, Text	https://www.amir-zadeh.com/datasets
SEED	Chinese	EEG	http: //bcmi.sjtu.edu.cn/~seed/
Yelp	English	Text, Video	https: //www.yelp.com/dataset/challenge
IEMOCAP	English	Speech, Text, Video	https://sail.usc.edu/iemocap/
DEAP	English	Face, EEG, GSR	https: //www.eecs.qmul.ac.uk/mmv/datasets/deap/
SEMAINE	English	Speech, Video	http: //semaine-db.eu

A. Feature Fusion

Feature fusion is the process of combining features from different sensors to generate a more comprehensive and accurate emotional feature vector [38]. In feature fusion, data from different modalities is merged together and pre-processed in the feature extraction stage, as shown in Figure 3. In this way, redundancy and noise can be reduced, and the expressiveness and robustness of emotional features can be improved.

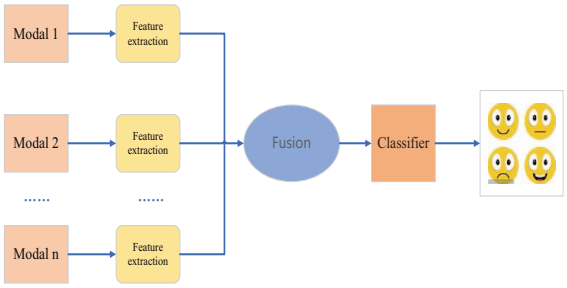

Fig. 3. Feature Fusion Model

Hazarika et al. [39] proposed a multimodal emotion recognition method based on feature-level fusion and context modeling. This method uses features from three modalities: speech, text, and image, and fuses them through feature-level fusion. Zhang et al. [40] proposed a feature fusion method based on deep canonical correlation analysis (DCCA), which uses the DCCA algorithm to map features from different modalities into a common low-dimensional space so that the feature representations from different modalities have the maximum correlation in this space.

B. Decision Fusion

Decision fusion is a relatively simple and direct fusion technique that combines emotion recognition results from multiple sensors to obtain the final emotion recognition result, as shown in Figure 4. In decision fusion, the outputs from different sensors are fed into a decision layer for aggregation, which is typically achieved using methods such as voting, weighted averaging, or logistic regression.

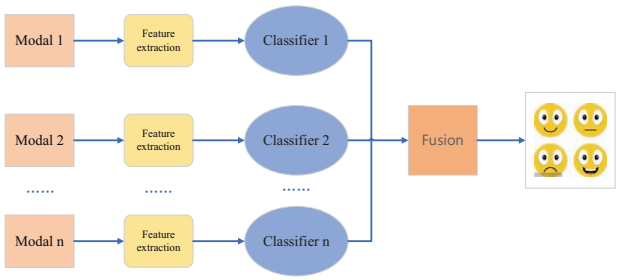

Fig. 4. Decision Fusion Model

Zhang et al. [41] implemented multimodal emotion recognition using decision-level fusion. They input the feature vectors of audio, image, and text modalities into separate classifiers for classification. In the study [42], four different physiological signal modalities were used to represent emotional information: electrocardiogram (ECG), electrodermal activity (EDA), electromyography (EMG), and EEG. The authors classified ECG, EDA, EMG, and EEG using four independent classifiers and combined the classification results of different modalities using a voting strategy as the final emotion recognition result.

C. Model Fusion

Model fusion is a more advanced and complex fusion technique that involves inputting emotional data from different sensors into multiple models for emotion recognition, as compared to decision fusion and feature fusion. Abdullah et al. [43] proposed a deep multimodal fusion method that first extracts features from speech signals using deep learning methods and then uses a convolutional neural network to extract features from electrocardiogram signals. The outputs of the two models are then fused to obtain the final emotion recognition result. Priyasad et al. [44] proposed a multimodal emotion recognition method based on model fusion. Firstly, suitable feature extraction methods and classifiers were selected for each modality, and then the different models were fused through a weighted sum to obtain the final emotion recognition results. The experimental results demonstrate that, compared to single-modality and decision-level fusion methods, this approach achieves better recognition

performance. Fang et al. [45] used CNN for visual feature extraction, performed physiological feature dimensionality reduction using PCA, and combined SVM for classification. Liu et al. [46] proposed a hierarchical fusion method based on multi-attention mechanisms that enhances modality features layer by layer and selectively fuses multimodal information to achieve more accurate emotion classification.

V. Research on Multimodal Emotion Classification Techniques

After fusing the features from multiple modalities, a more comprehensive and accurate feature vector that expresses emotional information can be obtained. These feature vectors can be used as input for emotion classification. Emotion classification is a crucial step in multimodal emotion recognition tasks and is the core of the entire task. Through emotion classification, we can categorize the input emotional information into different emotional categories, such as happy, sad, angry, etc., in order to better understand and respond to human emotions.

A. Machine learning methods.

Traditional machine learning methods typically use linear or nonlinear classifiers, such as SVM, the K-nearest neighbor algorithm (KNN), decision trees, etc. Tang et al. [47] proposed a recognition method based on SVM classification. The paper used video, speech, and text data to identify emotions, and after fusing the three types of features, SVM was used for classification. Zhang et al. [48] analyzed a multimodal emotion recognition method based on audio and video data. The authors used KNN and SVM classifiers for emotion classification and compared the performance of the classifiers. The experimental results showed that the performance of the KNN classifier was slightly better than that of the SVM classifier. Liu et al. [49] proposed a decision tree-based model that takes physiological signals and facial expression features as inputs. The decision tree was trained to extract key features from the input features and associate them with emotion categories.

B. Deep learning-based methods.

Deep learning methods are widely used in multimodal emotion recognition, and their main advantage is that they can automatically learn and fuse. In deep learning methods, common structures include DNN, CNN, RNN, LSTM, etc. These methods can input the features from multiple modalities into different branches of the network for processing and fuse and classify the features in the upper layers of the network. Among them, Majumder et al. [50] proposed a hierarchical fusion method for multimodal emotion recognition based on deep neural networks. The authors used multiple deep neural networks to fuse audio, text, and video modalities and finally achieved the multimodal emotion recognition task through hierarchical fusion of these networks. Tzirakis et al. [51] proposed an end-to-end multimodal emotion recognition method that used a deep neural network based on CNN and LSTM to automatically learn emotional representations from inputs from multiple sensors. Compared with other deep learning methods, using an end-to-end training method allows the model to automatically learn from the raw input data to the final emotion classification without the need for manual feature extraction, which has obvious advantages.

VI. Conclusion

This article comprehensively explores the latest research progress and methods in emotion recognition. In the introduction section, the paper elaborates on the applications of emotion recognition in various fields and the advantages of multimodal emotion recognition. In the second part, the paper details several different types of emotion recognition methods, including those based on visual, speech, text, and other modalities, and lists relevant literature. In the third, fourth, and fifth parts, the paper focuses on several important stages in the emotion recognition process, including feature processing, multimodal emotion fusion, and emotion classification. However, it is worth noting that existing emotion recognition research still faces some challenges, such as a lack of publicly available datasets and mutual interference between different perception channels. Therefore, future research should focus on addressing these challenges and further exploring various application scenarios and methods in the field of multimodal emotion recognition.

References

[1] Sun X, Zhang C, Li G, et al. Detecting users' anomalous emotion using social media for business intelligence[J]. Journal of Computational Science, 2018, 25: 193-200.

[2] Y Gao, X Xiang, et al., Human action monitoring for healthcare based on deep learning, IEEE Access 6, 52277-52285, 2018.

[3] Wang W, Xu K, Niu H, et al. Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation[J]. Complexity, 2020, 2020: 1-9.

[4] Andalibi N, Buss J. The human in emotion recognition on social media: Attitudes, outcomes, risks[C]//Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020: 1-16.

[5] Kwon O W, Chan K, Hao J, et al. Emotion recognition by speech signals[C]//Eighth European conference on speech communication and technology. 2003.

[6] Shu L, Xie J, Yang M, et al. A review of emotion recognition using physiological signals[J]. Sensors, 2018, 18(7): 2074.

[7] Zou J, Cao X, Zhang S, et al. A facial expression recognition based on improved convolutional neural network[C]//2019 IEEE International Conference of Intelligent Applied Systems on Engineering (ICIASE). IEEE, 2019: 301-304.

[8] Hossain M S, Muhammad G. Emotion recognition using deep learning approach from audio–visual emotional big data[J]. Information Fusion, 2019, 49: 69-78.

[9] Ko B C. A brief review of facial emotion recognition based on visual information[J]. sensors, 2018, 18(2): 401.

[10] Huang T R, Hsu S M, Fu L C. Data augmentation via face morphing for recognizing intensities of facial emotions[J]. IEEE Transactions on Affective Computing, 2021.

[11] Fahad M S, Deepak A, Pradhan G, et al. DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features[J]. Circuits, Systems, and Signal Processing, 2021, 40: 466-489.

[12] Uddin M Z, Nilsson E G. Emotion recognition using speech and neural structured learning to facilitate edge intelligence[J]. Engineering Applications of Artificial Intelligence, 2020, 94: 103775.

[13] Shchetinin E Y. Recognition of emotions in human speech with deep learning models[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1703(1): 012036.

[14] L Kang, RS Chen, et al., Selecting hyper-parameters of Gaussian process regression based on non-inertial particle swarm optimization in Internet of Things, IEEE Access 7, 59504-59513, 2019.

[15] Halim Z, Waqar M, Tahir M. A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email[J]. Knowledge-based systems, 2020, 208: 106443.

[16] Saffar A H, Mann T K, Ofoghi B. Textual emotion detection in health: Advances and applications[J]. Journal of Biomedical Informatics, 2022: 104258.

[17] Z Wang, T Li, et al., A novel dynamic network data replication scheme based on historical access record and proactive deletion, The Journal of Supercomputing 62 (1), 227-250, 2012.

[18] Chen X, Li S. A Review of Dialogue Emotion Recognition. Computer Engineering and Applications, 2023, 59(03): 33-48.

[19] L Shu, Y. Zhang, et al., Context-aware cross-layer optimized video streaming in wireless multimedia sensor networks, The Journal of Supercomputing 54 (1), 94-121, 2010.

[20] Zhang C, Yang Z, He X, et al. Multimodal intelligence: Representation learning, information fusion, and applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478-493.

[21] Tan Y, Sun Z, Duan F, et al. A multimodal emotion recognition method based on facial expressions and electroencephalography[J]. Biomedical Signal Processing and Control, 2021, 70: 103029.

[22] Huang J, Tao J, Liu B, et al. Multimodal transformer fusion for continuous emotion recognition[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 3507-3511.

[23] Sleeman IV W C, Kapoor R, Ghosh P. Multimodal classification: Current landscape, taxonomy and future directions[J]. ACM Computing Surveys, 2022, 55(7): 1-31.

[24] Nguyen H D, Kim S H, Lee G S, et al. Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks[J]. IEEE Transactions on Affective Computing, 2019, 13(1): 226-237.

[25] Chen J, Chen Z, Chi Z, et al. Facial expression recognition in video with multiple feature fusion[J]. IEEE Transactions on Affective Computing, 2016, 9(1): 38-50.

[26] Zhu X, Liu Y, Li J, et al. Emotion classification with data augmentation using generative adversarial networks[C]//Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22. Springer International Publishing, 2018: 349-360.

[27] Song P, Zheng W, Ou S, et al. Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization[J]. Speech Communication, 2016, 83: 34-41.

[28] Y Deng, H Hu, N Xiong, W Xiong, L Liu. A general hybrid model for chaos robust synchronization and degradation reduction. Information Sciences 305, 146-164, 2015.

[29] Daneshfar F, Kabudian S J. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm[J]. Multimedia Tools and Applications, 2020, 79: 1261-1289.

[30] Li Y, Gao Y, Chen B, et al. Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(5): 3190-3202.

[31] Goyal R, Chaudhry N, Singh M. Personalized Emotion Detection from Text using Machine Learning[C]//2022 3rd International Conference on Computing, Analytics and Networks (ICAN). IEEE, 2022: 1-6.

[32] Agarwal S, Mukherjee D P. Facial expression recognition through adaptive learning of local motion descriptor[J]. Multimedia Tools and Applications, 2017, 76: 1073-1099.

[33] Ding L, Tian Y, Fan H, et al. Joint coding of local and global deep features in videos for visual search[J]. IEEE Transactions on Image Processing, 2020, 29: 3734-3749.

[34] Fayek H M, Lech M, Cavedon L. Towards real-time speech emotion recognition using deep neural networks[C]//2015 9th international conference on signal processing and communication systems (ICSPCS). IEEE, 2015: 1-5.

[35] Lu C, Zong Y, Zheng W, et al. Domain invariant feature learning for speaker-independent speech emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2217-2230.

[36] Miller R E, Strickland C, Fogerty D. Multimodal recognition of interrupted speech: Benefit from text and visual speech cues[J]. The Journal of the Acoustical Society of America, 2018, 144(3): 1800-1800.

[37] X Shen, B Yi, H Liu, W Zhang, Z Zhang, S Liu, N Xiong, Deep variational matrix factorization with knowledge embedding for recommendation system,IEEE Transactions on Knowledge and Data Engineering 33 (5), 1906-1918, 2019.

[38] R Wan, N Xiong, Q Hu, H Wang, J Shang, Similarity-aware data aggregation using fuzzy c-means approach for wireless sensor networks, EURASIP Journal on Wireless Communications and Networking 2019, 1-11, 2019.

[39] Hazarika D, Gorantla S, Poria S, et al. Self-attentive feature-level fusion for multimodal emotion detection[C]//2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2018: 196-201.

[40] Zhang K, Li Y, Wang J, et al. Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis[J]. IEEE Signal Processing Letters, 2021, 28: 1898-1902.

[41] Zhang F, Li X C, Lim C P, et al. Deep emotional arousal network for multimodal sentiment analysis and emotion recognition[J]. Information Fusion, 2022, 88: 296-304.

[42] Fu Z, Zhang B, He X, et al. Emotion recognition based on multi-modal physiological signals and transfer learning[J]. Frontiers in Neuroscience, 2022, 16.

[43] Abdullah S M S A, Ameen S Y A, Sadeeq M A M, et al. Multimodal emotion recognition using deep learning[J]. Journal of Applied Science and Technology Trends, 2021, 2(02): 52-58.

[44] Priyasad D, Fernando T, Denman S, et al. Attention driven fusion for multi-modal emotion recognition[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 3227-3231.

[45] Fang Y, Rong R, Huang J. Hierarchical fusion of visual and physiological signals for emotion recognition[J]. Multidimensional Systems and Signal Processing, 2021, 32: 1103-1121.

[46] Liu X, Xu Z, Huang K. Multimodal Emotion Recognition Based on Cascaded Multichannel and Hierarchical Fusion[J]. Computational Intelligence and Neuroscience, 2023, 2023.

[47] Tang H, Liu W, Zheng W L, et al. Multimodal emotion recognition using deep neural networks[C]//Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24. Springer International Publishing, 2017: 811-819.

[48] Zhang J, Yin Z, Chen P, et al. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review[J]. Information Fusion, 2020, 59: 103-126.

[49] Liu, Zhen-Tao, et al. "Speech emotion recognition based on feature selection and extreme learning machine decision tree." Neurocomputing 273 (2018): 271-280.

[50] Majumder N, Hazarika D, Gelbukh A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-based systems, 2018, 161: 124-133.

[51] Tzirakis P, Trigeorgis G, Nicolaou M A, et al. End-to-end multimodal emotion recognition using deep neural networks[J]. IEEE Journal of selected topics in signal processing, 2017, 11(8): 1301-1309.