A Comparative Analysis of Single and Multi-Modality-Based Emotion Recognition for Disease Prevention

Han Jin

School of Artificial Intelligence and Big Data, Hefei University, Hefei, Anhui, 230000, China
100301@yzpc.edu.cn

Abstract. With the continuous growth of the economic level, disease and health problems have attracted more and more attention. Emotions and diseases are inextricably linked, which can be used as the response signal of physical diseases. Accurate emotion recognition is crucial for enhancing communication, improving customer service, and assisting mental health treatment by identifying emotional states. This paper summarizes the recent development of machine learning and deep learning-based emotion recognition technology, including fusion strategy, single-mode and multi-mode. This work aims at providing readers with a better understanding of the cutting-edge developments in this field. In addition, the possible future research and application directions are discussed, including the combination research with other recognition technologies and the application of disease prevention scenarios, and the future research directions in this field are predicted and analyzed. The purpose of this study is to provide valuable reference for researchers in the field of emotion recognition and disease prevention, so as to promote the development of emotion recognition technology in the field of disease prevention.

Keywords: Deep Learning, Machine Learning, Emotion Recognition, Disease Prevention.

1 Introduction

According to many studies, emotions are indeed closely related to a variety of diseases, especially negative emotions. When a series of human diseases occur, they are often accompanied by negative emotions. For example, people with autism often have emotional instability, severe self-laughter, irritability or irritability. And most patients with Alzheimer's disease also have irritability, paranoia or abnormal indifference. Therefore, emotion recognition has a broad application prospect in the research of disease prevention.

In the field of disease prevention, the relevant data obtained often have the characteristics of large scale and high complexity, while deep learning has the advantages of strong adaptive ability and generalization ability in this direction, which are not possessed by traditional machine learning. In the field of disease

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_35

prevention, which needs to process huge data, deep learning has been widely used in computer vision and natural language processing. Convolutional neural network, Tranformer and other technologies are used to adaptively realize the emotion recognition of different modes, so as to achieve the effect of disease prevention.

For the existing modal applications, it can be divided into single-mode and multi-mode. Among them, the single-mode emotion recognition is generally not based on video mode and signal mode. The research on video mode has a long history. Most of the researches have adopted Facial Emotion Recognition (FER) or Failure Mode and Effects Analysis (FMEA), but the accuracy based on video mode is not high. So, at present, it is more inclined to use signal mode for application research. The signal mode mainly includes Electroencephalogram (EEG) and Electrocardiogram (ECG). This mode is more accurate than video frequency for emotion monitoring, but it is difficult to obtain emotional information.

In addition, multimodal emotion recognition research has gradually become a hot spot in recent years. Due to the complexity of multi-dimensional and information types, multimodal emotion recognition often has a more profound and comprehensive understanding than single mode. In many medical fields not limited to disease prevention, multimodal also has an excellent performance. Therefore, this paper will comprehensively compare and analyze the disease prevention methods of emotion recognition based on deep learning.

2 Single Modality Emotion Recognition

2.1 Emotion Recognition Based on Video Modality

In the research methods of emotion recognition based on video modality, most of them use static pictures and dynamic video images containing emotional clues as input to capture the facial expression of patients. Many kinds of FER technology have been proposed.

From the perspective of deep learning, the early emotion recognition technology requires a large number of data sets to train the model, or, like FER, requires certain preprocessing techniques, such as face alignment, face normalization, pose normalization, etc, to align and normalize the semantic information of the face region. With the rise of multimodal technology, the research on single-mode video input methods is relatively rare, and the more meaningful method models are basically emerging at an earlier time. For example, in 2018, Zeng J etc. proposed a framework to discover their potential real tags from different tags, including: (1) training model A based on the known data set; (2) Model B is used to predict dataset to generate pseudo tags, and the two models are predicted on an unlabeled dataset at the same time; (3) All the data and their corresponding two tags are put into a network for training to generate potential real tags [1]. This method solves the uncertainty of the training model data set, and eliminates the possibility that the data set is subject to the subjective influence of the annotator to a certain extent.

However, as far as the whole research method of emotional modality is concerned, its data set is relatively easy to collect, which is helpful for the research of early technology which is not very perfect. However, in the face of today's complex and diversified needs, the method based on video modality is too one-sided, and because the patient's own emotional expression will also be controlled by subjective consciousness, the results are less accurate than other methods.

2.2 Emotion Recognition Based on Signal Modality

For some emotional changes of patients themselves, it may be difficult to detect their external manifestations. Therefore, the method of signal modal input for emotion recognition has also become one of the research hotspots. Compared with the expression, the change of signal can more directly reflect the change of patients' emotions. It can better identify and analyze emotional behavior and explain the state. In learning human emotions, the role of physiological signals has objectivity. At present, Electroencephalogram (EEG) and Electrocardiogram (ECG) are mainly used for research.

EEG. Among the methods based on signal mode, the research of EEG signal is the most extensive. It needs to extract EEG signal through instruments, and the signal with the best characterization is called differential entropy (DE). Feature extraction plays an important role in this research method. The multi view domain adaptive representation learning method proposed in recent research breaks through the limitations of EEG research, and proposes an extended causal convolutional neural network (CADD-DCCNN) with domain discriminator based on cross attention [2]. It uses Short-time Fourier Transform (STFT) to obtain de features from EEG signals, extracts relations through convolutional neural network, and finally carries out feature fusion, which enhances the performance of features and makes EEG signals more accurate in emotion recognition. The model overview is shown in Fig. 1.

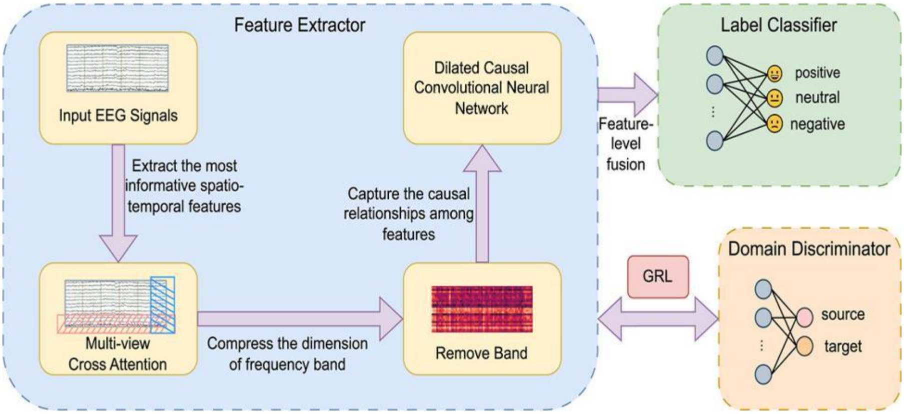
Feature Extractor
Input EEG Signals
Extract the most informative spatio-temporal features
Multi-view Cross Attention
Compress the dimension of frequency band
Remove Band
Capture the causal relationships among features
Dilated Causal Convolutional Neural Network
Feature-level fusion
Label Classifier
positive
neutral
negative
GRL
Domain Discriminator
source
target

Fig. 1. Overview of the CADD-DCCNN model [2].

ECG. ECG signal can also be used as the input of emotion recognition by mode. The same as the study of EEG signal, ECG signal also needs to be collected by instrument. The weak changes of skin potential caused by heartbeat are collected by electrode slice, and then collected by the acquisition device after passing through the amplifier to generate ECG for emotion recognition input. Therefore, emotion recognition based on ECG basically has a nine step framework. It can be summarized as follows: 1) preprocessing the signal; 2) Detection of R wave; 3) Use the window to record ECG; 4) Noise suppression; 5) Feature extraction of nonlinear analysis, especially based on time domain and frequency domain; 6) Feature standardization; 7) Floating selection of class separability feature based on pre sequence kernel; 8) Feature reduction using generalized discriminant analysis; 9) Build a classifier based on Least Squares Support Vector Machine (LS-SVM). Among them, feature extraction, feature selection and classifier are the core steps of ECG emotion recognition [3].

In 2024, Ritu etc. established an integrated model of u-net and artificial neural network (ANN) [4]. Its architecture is demonstrated in Fig. 2. By collecting ECG records from 12 different ECG public databases, he established a data set for short-term evaluation and another for long-term evaluation, and conducted signal preprocessing to normalize and subdivide the signal. After feature extraction, he used the method of deep learning to classify, and finally completed the architecture of the ensemble model network.

In general, the emotion recognition based on signal is more accurate than that based on video. It is less affected by the factors of the tested person, and can better reflect the patient's emotion in some aspects, but its data information is difficult to obtain. In terms of the integration with other technologies, due to the difficulty in obtaining information, the fusion technology is not as mature as other technologies.

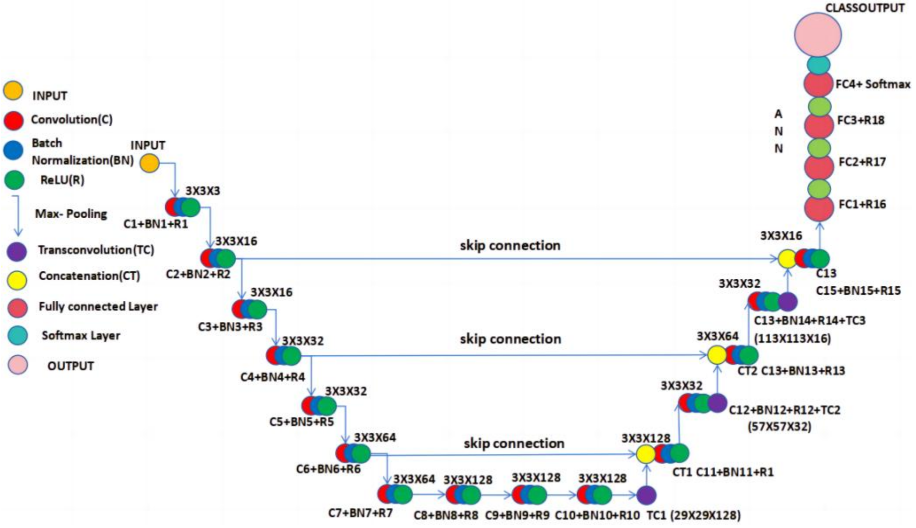

Fig. 2. Architecture of the U-Net with ANN [4]

3 Multi-modality Emotion Recognition

3.1 Integration Strategy

Due to the existence of multiple modal inputs, the fusion strategy between modes is very important in the research field of multimodal emotion recognition. In early research in the multimodal field, feature level fusion was commonly used as the fusion strategy, which was used to learn representative features from general models. Then, feature extraction was integrated through concatenation or weighted combination. In deep learning with strong representation ability, early feature fusion strategies often used recursive neural networks, convolutional neural networks, or Long Short Term Memory (LSTM) models for feature extraction in research, but these fusion strategies often had overfitting problems [5]. In the later research, it has turned to the use of model-based fusion methods. For example, LSTM model can extract deep spatio-temporal features. The model-based fusion method significantly optimizes the performance of emotion recognition tasks, which can simulate the internal dynamics of different patterns in an implicit way. However, it is undeniable that most of the current model-based methods do not consider the possibility of semantic dislocation between modes, and most of the attention-based mechanisms do not produce good results for distinguishing different modes through connection. Therefore, fusion strategy is still the main challenge for emotion analysis and emotion recognition.

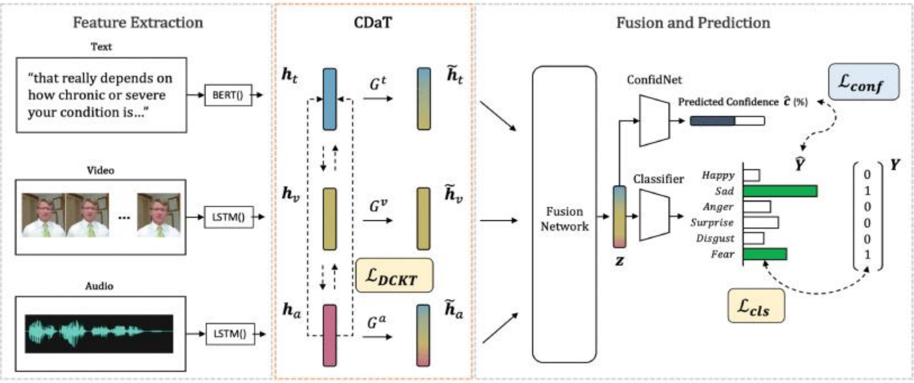

Fig. 3. Dynamic Transfer Learning Process [6]

In order to solve this problem, a representation learning method called Cross Modal Dynamic Transfer Learning (CDAT) has been proposed recently [6]. Its biggest advantage is the dislocation between the modes of scientific research dynamic adjustment. The learning method mainly includes two stages: (1) dislocation modal detection and (2) modal knowledge transfer. First, the distortion mode filter (MMF) stage is introduced, in which the additional network is trained to evaluate the modal reliability at the example level. Including the proportion adjustment of unrelated and other high reliability modes. In order to make these settings dynamic in each instance, the model combines the study of the difference loss between the knowledge transfer probability (PKT) and the features extracted from each module encoder. Finally, experiments on two datasets show that it achieves consistent performance gains compared with the most advanced fusion model. The learning process is shown in Fig. 3.

3.2 Typical Work and Cutting-edge Technology

In the research process of multimodal emotion analysis, a large number of excellent and classic models have emerged, and they all put forward more efficient model methods or have good improvements on the basis of the original methods. Table 1 demonstrates the classic multimodal emotion analysis methods.

Table 1. Classic Multimodal Emotion Recognition Methods.

Year	Feature Representation	Classifier	Fusion Strategy
2011	Acoustic-prosodic Semantic labels [7]	Multi-Disciplinary Team(MDT), MaxEnt, Base classifiers	Decision-leve l
2013	Mel-Frequency Cepstral Coe fficients (MFCC), Facial movement,	Bidirectional Long Short-Term Memory (BiLSTM), Support Vector Machine (SVM)	Hybrid-level
2015	handcrafted, Convolutional Neural Net-work (CNN), Principal Component Analysis (PCA), Confidence-Based Feature Selection (CFS)	Math Kernel Library (MKL)	Feature-level, Decision-leve l
2017	Convolutional Coding with Spatial Transformer, Networks(C3D+), Deep Belief Network (DBN) [8]	Score-level Fusion	Model-based
2019	2D CNN, 3D CNN [9],	ELM-based fusion, SVM	Model-based
2020	Self-attention, Attention-driven CNN, Temporal-Differential Neura l Network (T-DNN) [10]	Fully Connected (FC)	Feature-level

In recent years, with the rapid development of artificial intelligence and deep learning, more and more needs to be completed by multimodal emotion analysis. Among them, graph based multimodal technology has made rapid progress. A method of multimodal emotion analysis by capturing high-level semantic correlation through graph has been proposed, hierarchical graph contrastive learning Hierarchical Graph Contrastive Learning (HGraph-CL) [11]. First, three modal forms of T, A and V are given, and modal specific encoders are used to further extract emotion related information, so as to create a capsule network, realize the coordination and cooperation between low-level capsules and high-level capsules, use dynamic routing to build nodes from the output sequence, design a learnable adjacency matrix to realize edge construction based on self-concern, and then convert the adjacency matrix into multimodal. Figure, after layered comparative learning, can recognize and predict emotions. According to the experimental data, it performs better on a variety of data sets than other existing models, and performs well on the task of emotion recognition. The main contribution is to break through the limitations of HGraph-CL method in graph construction, avoid the neglect of high-level language relevance in the traditional graph based multimodal identification method, and reduce the information redundancy. And better capture implicit and long-distance correlation.

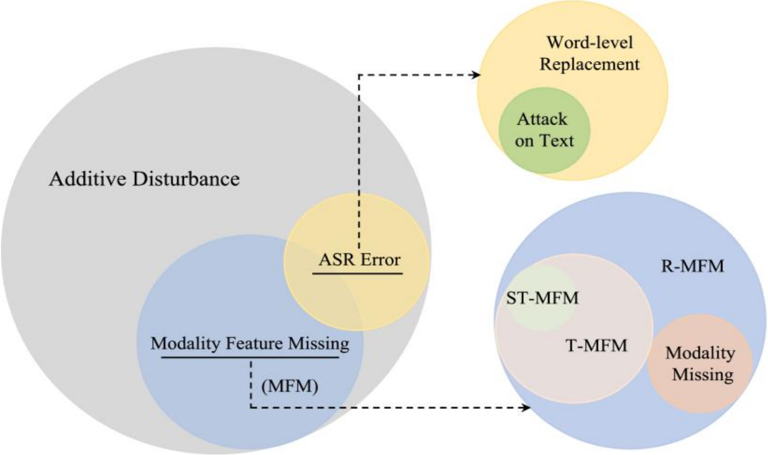

Fig. 4. Indication of Defect Classification [12]

In addition to the multimodal emotion recognition method based on graph method, the analysis of characteristic noise by emotion recognition model has become one of the most critical research topics in the field of multimodal emotion analysis. For emotion analysis, data defects are almost inevitable. Here is the classification of different defects, as shown in Fig. 4.

For multimodal emotion recognition, different regions in the picture correspond to different features in the input syntax. Coarse grained image and feature alignment will introduce noise, while the framework for multimodal emotion analysis of feature noise (META-NA) can solve this problem [13]. It first constructs the source noise task, designs the test case count K, samples the training case count range, and uses Bernoulli distribution and uniform distribution for sampling. Finally, it injects the feature noise into all training and test cases, so as to achieve the purpose of obtaining the source task. It uses the backbone network and auxiliary feature adaptive module based on post fusion.

In order to reduce the negative impact of feature noise on the single peak representation of learning, the meta noise adaptation strategy is used for nested optimization. From the later analysis of the results, it can be seen that all the indicators of META-NA in model comparison are due to all the existing baseline methods. The proposed method greatly reduces the possibility of inaccurate recognition due to the lack of information. Compared with the previous traditional baseline method, it can still accurately carry out emotion analysis in the absence of voice, and puts forward solutions for other methods in noise processing.

3.3 Summary

There are still many fields to be explored in multimodal emotion recognition. The research methods proposed in this section have solved the key problems in traditional fields. From the perspective of fusion strategy, dynamic transfer learning has a good solution to the dislocation problem between modes, but for the whole research field, the vast majority of models still lack the ability of autonomous learning, and the over fitting problem is still the top priority of fusion strategy. The two methods proposed later have good improvements on the multimodal graph method and baseline method, which help multimodal emotion recognition improve the accuracy and accuracy. They have a good performance on the experimental data set, and have inspiration and breakthrough on the future research.

4 Discussion

4.1 Modal Discussion

The discussion of single-mode and multi-mode has always been a long-standing topic. From the point of view of this article, for single-mode, the degree of refinement of mode is higher than that of single-mode. For example, the accuracy of processing text, voice or image alone is often higher than that of multi-mode. It has good specificity and mature technology, has a wide range of practical applications, and does not need to face the problem of integration during operation, which is more convenient and faster. However, due to the lack of modes in dealing with complex problems, it cannot achieve high accuracy, and cannot meet the needs of multi-level. In this regard, intersection multimodal analysis has certain disadvantages.

For multimodal, it can handle the complex situation of multimodal input, and the comprehensive judgment of emotion is more comprehensive than that of single mode. It also has a good performance in the intersection with a variety of fields, and has more possibilities. However, at present, multimodal technology is not fully mature, and issues such as integration still need to be explored and improved.

4.2 Outlook

In the direction of disease prevention, future work could consider combining behavior recognition with multimodal emotion recognition in future research. Through the idea of multimodal emotion recognition, future work transforms behavior recognition through cross modal transformation, analyze it as a mode, and give it a special emotion quantitative index. After the analysis, according to the suggestions of professional doctors, set the corresponding emotional indicators for a variety of diseases, output and control the gate logic through weighting, voting and other methods, and use data enhancement technology, including rotation, scaling, clipping and other operations, to increase the diversity of data and the robustness of the model, so as to complete a set of system for judging diseases through daily emotions and behaviors. Through the identification of emotions and behaviors, the goal of early detection of diseases can be achieved, and the occurrence and development of related diseases can be better prevented.

5 Conclusion

This paper discusses the current research status of both single modality and multimodality, summarizes the early research methods of single modality and recent emerging technologies of both single modality and multimodality, classifies and analyzes the current research methods of multimodal emotion recognition in various fields from different angles and methods. Moreover, it conducts a multi-level comparison between single modality, multimodality, and the earlier and recent stages of multimodal research, multi-modality and multimodal early and recent, and make a future outlook on the possible technical extension of multimodal technology and the possibility of combining with other fields. It is expected that in the future field of emotion recognition, this work can further improve methodology, make contributions to disease prevention, and realize the integration of technology in more fields.

References

1. Zeng, J., Shan, S., & Chen, X.: Facial expression recognition with inconsistently annotated datasets. In Proceedings of the European conference on computer vision, pp. 222-237. Springer, Munich (2018).
2. Li, C., Bian, N., Zhao, Z., Wang, H., & Schuller, B. W.: Multi-view domain-adaptive representation learning for EEG-based emotion recognition. Information Fusion, **104**, 102156 (2024).
3. Dessai, A., & Virani, H.: Multimodal and Multidomain Feature Fusion for Emotion Classification Based on Electrocardiogram and Galvanic Skin Response Signals. Sci, **6**(1), 10 (2024).
4. Begum, R. N. A., Sharma, A., & Singh, G. K.: An Ensemble Model of DL for ECG-based Human Identification. IEEE Transactions on Instrumentation and Measurement, **73**, 1-15 (2024).

5. Zhang, F., Li, X. C., Lim, C. P., Hua, Q., Dong, C. R., & Zhai, J. H.: Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. Information Fusion, **88**, 296-304 (2022).
6. Hong, S., Kang, H., & Cho, H.: Cross-Modal Dynamic Transfer Learning for Multimodal Emotion Recognition. IEEE Access, **12**, 14324-14333 (2024).
7. Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., ... & Zhang, W.: A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, **83**, 19-52 (2022).
8. Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q.: Learning affective features with a hybrid deep model for audio–visual emotion recognition. IEEE transactions on circuits and systems for video technology, **28**(10), 3030-3043 (2017).
9. Zhang, E., Xue, B., Cao, F., Duan, J., Lin, G., & Lei, Y.: Fusion of 2D CNN and 3D DenseNet for dynamic gesture recognition. Electronics, **8**(12), 1511 (2019).
10. Priyasad, D., Fernando, T., Denman, S., Sridharan, S., & Fookes, C.: Attention driven fusion for multi-modal emotion recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3227-3231. IEEE, Virtual (2020).
11. Qian, F., Han, J., Guan, Y., Song, W., & He, Y.: Capturing High-level Semantic Correlations via Graph for Multimodal Sentiment Analysis. IEEE Signal Processing Letters, **31**, 561-565 (2024).
12. Yuan, Z., Liu, Y., Xu, H., & Gao, K.: Noise imitation based adversarial training for robust multimodal sentiment analysis. IEEE Transactions on Multimedia. 26, 529-539 (2023).
13. Yuan, Z., Zhang, B., Xu, H., & Gao, K.: Meta Noise Adaption Framework for Multimodal Sentiment Analysis With Feature Noise. IEEE Transactions on Multimedia. 26, 7265-7277 (2024).