Exploration and Application Analysis of Crucial Techniques in Multimodal Emotion Recognition

Wenqi Li [1,*]

[1] School of Animal Science and Technology, Hainan University, Hainan, 570228, China
*21210502000010@hainanu.edu.cn

Abstract. Multimodal emotion recognition involves the identification of human emotional states utilizing a combination of sensory information, such as audio, video, and physiological signals. This field, straddling the domains of computer science and psychology, holds significant commercial potential in areas such as intelligent human-computer interaction and health monitoring. As a result, emotion recognition has emerged as a research hotspot in the realm of affective computing both domestically and globally. This paper offers a comprehensive overview of the multimodal sentiment analysis field, encapsulating the key technical findings to date. It further delves into pertinent experiments and application analyses pertaining to multimodal emotion recognition, contributing valuable insights to this evolving area of study. Looking ahead, the paper contemplates future trends and prospects, as well as potential research directions and strategies for the further development of multimodal emotion analysis technology. Given the rapid technological advancements and the increasing need for nuanced human-machine interactions, we posit that multimodal emotion recognition will continue to be an area of considerable focus. This paper aims to provide the foundational understanding necessary to fuel these future explorations.

Keywords: Multimodal Emotion Recognition, Key Technology of Multimodal Emotion Recognition, Multimodal Emotion Recognition Application Analysis

1 Introduction

Multimodal emotion recognition involves the utilization of diverse sensory inputs, including but not limited to audio, video, and physiological signals, to discern human emotional states. This field finds broad applications in computer science, psychology, and other disciplines, and holds potential commercial value in sectors such as intelligent human-computer interaction and health monitoring. Currently, emotion recognition has become a focal point of research in affective computing globally.Initial attempts at facial expression recognition typically relied on geometric and texture features [1]. However, with the advent of neural networks, models such as convolutional neural networks and long short-term memory networks have been effectively deployed across various modes of emotional analysis. The use of a single modality for emotional analysis has its constraints. Humans, when expressing

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_50

emotions, employ a combination of voices, content, facial expressions, and body language. Multimodalities stem from numerous heterogeneous sources [2]. As compared to single modality emotion analysis, using multiple modalities can capture emotional information more accurately.

Many researchers are dedicated to integrating multiple modalities for emotion analysis, employing techniques such as representation learning, modal alignment, and modal fusion, thereby effectively addressing the limitations of single modality emotional analysis. Current research on multimodal emotion recognition primarily centers around: external behavior expression modes, neurophysiological modes, and a combination of neurophysiological states and behavioral subconscious behaviors. Significant strides have been made in emotion recognition research in China at these levels. Key technologies largely encompass feature extraction, modal fusion, classifier design, and more [3,4].

2 Overview of Multimodal Emotion Recognition

2.1 Definition and Related Concepts of Multimodal Emotion Recognition

Multimodal emotion recognition entails the mathematical modeling of diverse human emotional representations, such as facial expressions, speech, textual language, and other modalities, with the goal of establishing the mapping relationship between feature data and human emotions. This approach goes beyond the confines of single modality analysis, enabling a comprehensive and more scientifically sound emotional analysis derived from multiple modalities [5].

Historically, emotional representation methods have been generally categorized into three models: discrete emotion model, dimensional emotion model, and other emotion models. The discrete emotion model, due to its categorical approach to emotion classification, often falls short in providing a comprehensive and accurate analysis of complex emotions [6]. With academic advancements, dimensional emotional models, with their intricate classifications, have gained prominence. One-dimensional emotional models denote the positive axis as the 'happiness' axis and the negative axis as the 'sadness' axis. Two-dimensional emotional models take the extreme and intensity of emotions as their horizontal and vertical coordinates, respectively. Three-dimensional emotion models define emotion as having three dimensions: pleasure, arousal, and dominance.

The advent of deep learning has further influenced the evolution of multimodal emotion analysis models, often merging deep learning techniques with traditional emotion analysis methodologies [7]. While deep learning is employed to analyze emotions in modalities such as speech, facial expressions, and text, references to traditional emotion analysis methods help ensure a reduction in experimental analysis errors, enhancing the overall accuracy of the emotion recognition process.

2.2 Application Scenarios for Emotion Recognition in Multimodal Contexts

Currently, the application scenarios of emotion recognition in multimodal contexts span across several sectors, including but not limited to areas like inquiry, service, and psychological diagnostics. The scientific analysis of human emotions can notably boost work efficiency in related fields, making a significant impact on industries that revolve around interaction. For instance, during interrogations, mathematical analysis of a subject's facial expressions, voice, body movements, and other modalities can be used. This approach enables the screening of valuable information disclosed by the individual, thereby achieving an optimal effect in questioning. Similarly, in the medical diagnosis of patients with psychological disorders, multimodal emotional recognition can be deployed to ascertain the patient's current emotional state, thereby enabling effective communication. In the field of human or human-computer interaction, multimodal emotion recognition holds considerable value, having the potential to significantly improve the efficiency of information exchange. This technology offers a promising future for improving interaction efficiency, whether it's between humans or between humans and machines [8].

2.3 Research Status and Challenges of Multimodal Emotion Recognition

The current research on multimodal emotion recognition often combines deep learning technology, which greatly promotes the progress of multimodal emotion recognition research, but also brings new challenges to multimodal emotion recognition research.

Firstly, compared to traditional models such as single mode, multimodal models are more effective in modeling multiple modes.As a result, the volume of the multimodal model is much larger than that of the single modal model, which often leads to problems such as slow training speed and excessive model size. Secondly, when multimodal emotion recognition research is combined with deep learning technology, the drawbacks of deep learning often also arise. Deep learning requires a large amount of training data, and its performance can be improved depending on the size of the data set. Therefore, deep learning usually requires a large amount of data as support. If a large amount of effective training cannot be carried out, it will often lead to overfitting. Moreover, deep learning cannot determine the correctness of the data, which may result in non objective results. This requires the introduction of standards for marking and scoring the data, greatly increasing the difficulty of research and development. And deep networks are too sensitive to changes in images, often localized or the changes in the background of the subject will affect their judgment of the current image [9].

At the same time, for small probability biased events, the processing ability of deep learning is significantly lower, and the processing of events is only limited to the range of delivered data, which may lead to serious biases in the research of multimodal emotion recognition.

3 Key Technologies for Multimodal Emotion Recognition

3.1 Feature Extraction Technology

The existing common feature extraction includes feature extraction for facial expressions, feature extraction for text, and feature extraction for speech.

For the feature extraction technology of facial expressions, according to the different feature representation, the FER system can be divided into two categories: FER for static images and FER for dynamic sequences. In the FER of a dynamic sequence, facial expressions exhibit two characteristics: temporality and salience. The saliency of facial expressions is often ignored in the FER of dynamic sequences, in order to solve this problem, corresponding attention modules can be added to the spatial domain subnetwork and the time domain subnetwork to improve the performance of CNN and RNN when extracting features. The facial expression recognition process includes three stages, namely face detection, feature extraction and selection, and classification. According to the characteristic representation used, it can be divided into traditional methods and deep learning-based methods.

In terms of feature extraction technology for text, we generally extract information that can express opinions and emotions fromthe text. The former is more important than extracting text that expresses opinions, while the latter is more important than extracting text that expresses emotions. In text sentiment analysis, sentiment information extraction is the most important part. The effectiveness of emotional information extraction directly affects the effectiveness of text sentiment analysis.

The extraction of emotional information is the extraction of emotional words from text. Emotional words can be divided into three types:

(1) words that only contain emotional words.

(2) Vocabulary composed of emotional words and polar orientations.

(3) Emotional words with direction and intensity.

At present, sentiment dictionary based and deep learning methods are the two main methods for text sentiment analysis.

In terms of feature extraction technology for speech. Emotional analysis based on speech. In daily life, voice communication is one of the essential ways. Voice contains rich emotional information, not only textual information, but also features that can display emotions such as pitch and rhythm. In recent years, the use of multimedia computer systems to study emotional information in speech has received increasing attention from researchers. Analyzing emotional features, judging, and simulating the speaker's emotions has become a significant research topic. In existing literature, most of the research on emotion analysis based on speech has focused on identifying some acoustic features, such as prosodic features, sound quality features, and spectral features. Currently, it is mainly divided into traditional machine learning based methods and deep learning based methods.

3.2 Modal Fusion Technology

According to existing research, there are roughly four methods of modal fusion, namely data level fusion (sensor level fusion), feature level fusion, decision level fusion, and model level fusion.

Data level fusion, also known as sensor level fusion. Data level fusion is the direct combination of the most primitive and unprocessed data collected by various sensors to construct a new set of data. At present, the methods of data level fusion processing include numerical processing and parameter estimation. Specifically, linear and nonlinear estimation and statistical operation methods are used to calculate and process data from multiple data sources. Its advantage is that it can effectively preserve the data information on various modal sensors, avoid information loss, and maintain information integrity.But its drawbacks are also obvious, as the data is fused at the feature level in the original state, which involves constructing multiple modal data into corresponding modal features and then concatenating them into a feature set that integrates various modal features. At the feature level, the commonly used fusion strategy is to cascade all modal feature data after feature extraction into feature vectors, which are then fed into an emotion classifier [10].

Decision level fusion is to identify the credibility of each modality, and then coordinate and make joint decisions. Compared to feature level fusion, decision level fusion is easier to perform, but the key is to explore the importance of each modality in emotion recognition.

Model level fusion does not rely on the architecture of the three fusion levels mentioned above. The key to decision level fusion is to identify the credibility of different modalities in the decision-making stage, but model level fusion does not need to focus on exploring the importance of each modality. Instead, appropriate models need to be established based on modal characteristics to jointly learn related information. Feature level fusion mainly involves constructing feature sets or mixed feature spaces, and then sending them to classification models for classification decision-making. Model level fusion can input different modal features into different model structures for further feature extraction. Overall, the biggest advantage of model level fusion over decision level fusion and feature level fusion is the flexibility to choose the fusion location.

3.3 Classifier Design Techniques

The classification problem is to classify objects into a certain category based on their observed values. The specific steps are:

(1) Establish the training set in the feature space, and know the category to which a point in the training set belongs.

(2) Seek a certain discriminant function or discriminant criterion from these conditions, and design a discriminant function model.

(4) Determine the parameters in the model according to the samples in the training set, and obtain a perfect discriminant function mode.

Use a perfect discriminant function or discriminant criterion to determine which class the point of each unknown category should belong to.

In statistical pattern recognition, the main issue discussed is not the accuracy of decision-making, but the probability of accuracy of decision-making. The "best"optimization emphasized in pattern recognition is aimed at a certain design principle, which becomes the criterion.This criterion includes:

Minimum error rate criterion: based on the principle of reducing classification errors.

Minimum risk criterion: Introduce the concept of risk loss and assign different weights to minimize the total risk.

Nearest neighbor criterion: Distinguish based on the principle of clustering characteristics of similar objects in space.

Fisher's Rule: Seeking the Best Direction of a Straight Line and How to Realize the Transformation of Projection in the Best Direction.

Perception criterion: The perception criterion function minimizes the sum of distances between misclassified samples and the interface.

3.4 Basic methods

(1)Template matching method:

Usually, the nearest neighbor principle is used, which is the simplest classification method, but its drawbacks are high computational complexity and storage capacity.

(2) Discriminant function method:

Classification method based on probability statistics.It often depends on the relevant knowledge of the previous statistical distribution.The most classic BAYES classifier uses the prior probability and the class conditional probability density function to calculate the posterior probability, so as to design the discriminant function and decision surface.

Geometric classification.Independent of the knowledge of conditional probability density, the feature space is decomposed into subspaces corresponding to different categories through geometric methods.

4 Application Analysis of 4 Multimodal Emotion Recognition

4.1 Dataset Introduction

At present, the research of multimodal emotion recognition is generally based on a large amount of data, and for this reason, a large number of data sets related to modality are established. The most ubiquitous datasets today are the two-mode Pacific dataset and the trimodal dataset. Among them, most of the former sources and various review websites, its mode is generally a combination of text image language, such as YELP website, a famous American merchant review website, its business scope covers shopping centers, restaurants, shopping centers, hotels and other merchants around the rating project, users can score merchants in the Yelp website,

submit reviews, exchange shopping experience, etc. Its counterpart, Yelp Open Dataset, is a subset of Yelp business, reviews and user data, which can study the emotions of users during evaluation, and then improve the service model of merchants. Most of the latter comes from video websites, and the modalities generally include text, video, and audio data, including YouTube-8M, for example, this dataset contains 8,000,000 YouTube video links, and these video sets are video-level labeled as 4800 KnowledgeGraph entities [4]. The following table 1 are several common datasets:

Table 1. Multimodal correlation datasets.

	name	Modal	
Bimodal sentiment dataset	CMU-MOSI	Images, text	https://www.amir-zadeh.com/dataset
	Yelp	Images, text	https://www.yelp.com/dataset/challenge
Three-modal sentiment dataset	MELD	Text, video, audio	https://affective-meld. Github.io/
	IEMOCAP	Text, video, audio	http://sail.usc. edu/iemocap/
	YouTube-8M	Text, video, audio	https://arxiv.org/abs/1609.08675

4.2 Application Analysis

Based on the understanding of multimodal emotion recognition technology, this paper believes that its possible future research directions and schemes are as follows:

- Cross-modal emotion recognition based on deep learning: Develop more accurate and efficient neural network models to achieve the fusion of audio, video, physiological signals and other information.
- Complementarity of facial expression recognition and speech analysis: By combining the recognition results of facial expressions and speech, the accuracy and robustness of emotion classification are improved.
- Cross-cultural emotion recognition: Considering the influence of cultural backgrounds of different countries and regions on emotions, research on cross-cultural emotion recognition can be carried out in the future, and cross-cooperation with the field of psychology can be strengthened.
- Physiological signal emotion recognition: use physiological data (such as galvanic skin response, brain waves, etc.) as input for emotion recognition to achieve more objective and reliable emotion classification.

5 Future Work Prospects

Nowadays, multimodal emotion recognition technology has gradually been combined with algorithms such as deep learning, becoming an interdisciplinary field. Deep learning technology has become the mainstream of multimodal emotion recognition research. By establishing large-scale neural network models, it can learn features from a large amount of data and achieve more accurate emotion classification. The cross modal fusion technology that integrates different sensory information to improve the accuracy and robustness of emotion recognition is one of the current research focuses. Some of these methods include neural network based cross modal fusion and collaborative filtering based fusion. At the same time, multimodal emotion recognition needs to combine relevant knowledge of psychology in order to better understand the essence of human emotions. Future research can focus more on the psychological mechanisms of human emotions and the differences in emotional expression in different cultural backgrounds.

6 Conclusion

This article presents an encompassing review of the application scenarios, the current state of research, and the challenges inherent in emotion recognition within the field of multimodal emotion analysis. It provides a summary of pivotal technical research achievements to date, and organizes the prevalent technologies used in multimodal emotion analysis into a coherent taxonomy. A series of relevant experiments have been conducted on multimodal emotion recognition, and these findings have been applied to perform a comprehensive analysis. The scope of this work extends beyond simply outlining the current state of the field, but also offers a forward-looking perspective. The closing section of the article forecasts the future trajectory of multimodal emotion analysis technology, identifying the emergent trends and future prospects of multimodal emotion recognition. This includes possible research pathways and promising solutions, giving a speculative glimpse into the advancements that may shape the field in the coming years.

References

1. Wang, Y., Jiang, B., & Liu, J. (2019). A survey of multimodal emotion recognition research. Information Fusion, 52, 222-237.
2. Baltrusaitis, T., Robinson, P., & Morency, L. P. (2018). Multimodal emotion analysis in the wild. Image and Vision Computing, 68, 101-115.
3. Karray, F., & Saleh, J. A. (2018). Multimodal Emotion Recognition Using Deep Neural Networks: A Review. IEEE Transactions on Affective Computing, 9(3), 377-401.
4. Song, X., Lu, D., & Li, H. (2020). Multimodal Emotion Recognition Based on Deep Learning: A Survey. IEEE Transactions on Cognitive and Developmental Systems, 12(3), 567-579.

5. Yin, M., Duan, J., Zhang, Y., Zhang, S., & Yan, Q. (2019). Multimodal emotion recognition based on deep neural network. IEEE Access, 7, 7093-7103.
6. Li, Y., Mao, Y., Chen, L., & Lin, C. Y. (2018). A survey of multimodal emotion recognition using physiological signals. Information Fusion, 45, 99-113.
7. Zhao, Y., Wu, X., Wang, L., Zhang, J., & Chen, Y. (2020). A Survey on Multimodal Emotion Recognition Methods and Applications. Frontiers in Psychology, 11, 335.
8. Guan, J., Wang, D., & Hong, X. (2017). Multimodal emotion recognition using dynamic fuzzy support vector machine. IEEE Transactions on Cybernetics, 48(5), 1393-1404.
9. Liu, Y., Wang, Y., & Zhang, J. (2019). Convolutional neural network based multimodal emotion recognition applied to video games. Neurocomputing, 323, 70-77.
10. Wang, S., Li, Y., & Li, Q. (2017). Multimodal emotion recognition using sparse autoencoder neural network. IEEE Transactions on Industrial Electronics, 64(8), 6486-6495.