Key Issues and Application Analysis of Multimodal Emotion Recognition

Ruiyang Lin *

Department of Fujian Normal University, Fuzhou, 350000, China

* Corresponding Author Email: 136132023008@student.fjnu.edu.cn

Abstract. This article mainly explores the key issues of multimodal emotion recognition and analyzes some applications. It discusses the key issues from two core problems: data quality and coverage, semantic understanding and modal association. In the issue of data quality coverage, it is pointed out whether the data is good enough and comprehensive enough. In semantic understanding and modal association, it is pointed out whether the semantics are clear and whether they can be coordinated with other modalities. It also combines two core issues to discover the problems that occur in multimodal scenarios. At the application level, research on learning emotion recognition for the education field is listed. By real-time judgment of learning concentration and emotional state, it provides support for personalized teaching. The design and application of multimodal emotion recognition technology in security systems, the role of multimodal emotion recognition in security, MemoCMT: Multimodal emotion recognition is achieved through feature fusion based on cross-modal converters. By using an innovative feature fusion strategy, the three application aspects of multimodal emotion features are effectively integrated.

Keywords: Multimodal emotion recognition, data quality and coverage, semantic understanding and modal association, Application.

1. Introduction

Emotions play a significant role in human life. By understanding people's emotions, one can grasp their inner feelings, which provides certain assistance for communication and interaction among them. It is difficult to know people's emotional changes based on experience and observation, but multimodal emotion recognition technology can achieve this. Multimodal emotion recognition refers to a technology that can more comprehensively and accurately infer human emotional states by integrating multiple modal data such as text, voice, images, and physiological signals. This technology can be widely used in fields such as education, finance and healthcare to help people solve problems. Multimodal emotion recognition can sense users' emotions and achieve personalized emotional services. Further expand the boundaries of emotional perception and provide support in areas such as mental health monitoring, intelligent healthcare, and education, thereby enhancing the quality of social services. This article analyzes the key issues and applications of multimodal emotion recognition to promote the future development of multimodal emotion recognition technology, better achieve human-computer interaction, and enable multimodal emotion recognition to be more widely used.

2. Key Issues in Multimodal Emotion Recognition

2.1. Multimodal and Unimodal-

Multimodality is composed of multiple single modalities. Compared with multimodality, single modality has simpler technology, requires less data, and has lower computational costs. The advantage of multimodality lies in its more comprehensive information, strong robustness, and the ability to be applied to more complex scenarios and tasks. The key issues of multimodal emotion recognition are mainly analyzed from two parts: macroscopic carriers and microscopic modalities.

2.2. Page Numbers

From the perspective of macro carriers, the main problems lie at the dataset level. There are mainly three issues: insufficient data volume, insufficiently wide application, and unbalanced data distribution. From the perspective of insufficient data volume, it can be seen from the heat map of publications in the field of multimodal emotion recognition grouped by country that there are actually quite a few publications in this field [1]. Most regions have corresponding publications, but the dataset collection is very small, resulting in an insufficient amount of available data. From the perspective of unbalanced data distribution, the number of neutral emotion samples in daily conversations is far greater than that of stronger emotions like joy and sadness. Therefore, the model tends to focus on identifying common emotions, while the data of emotion categories is uneven, thus ignoring niche but crucial emotions. In terms of limited application, the current multimodal emotion recognition is mainly used in laboratory scenarios and has not been widely implemented in actual complex scenarios, such as social platforms and the entire process of intelligent customer service. Moreover, no machines have appeared in real environments. The practicality is limited due to the diversity of data. Overall, the main issue is whether the data is good enough and comprehensive enough. The core issue is the problem of data quality and coverage.

2.3. Microscopic Modes

From the perspective of micro-modality, the main issue lies in the text information level. At the text information level, there are mainly three problems, which can be further divided into strong ambiguity of characters' emotions, explainability of modality, and the absence of modality. Text serves as the foundation for emotional expression, but there is a problem of strong concealment of characters' emotions. For instance, "fine, thanks" may seem neutral on the surface, but in reality, it might imply a sense of perfunctoriness and helplessness. It is difficult for models to accurately capture the deep meaning that characters want to convey. Different people express themselves in many different ways. Meanwhile, the text modal also interacts with other modalities, thereby increasing the difficulty of recognition and making it hard for machines to interact with people. The speaker's dependence and emotions in semantics. For instance, the heterogeneity problem of multimodal data can be initially alleviated by mapping the features of different modalities to a shared semantic space using a shared encoder. Then, a shared convolutional network is employed in the shared semantic space to further learn the shared semantic information between each modality, thereby eliminating the gap between multimodal features [2]. However, most models do not fully consider the speaker dependency relationship in semantics and emotional expression. The semantics and emotions conveyed by different speakers on the same topic can show discrepancies in the dependency list. Ignoring these dependencies may result in extracting noisy and misaligned unimodal features from the speaker's turn. Secondly, the modal differences bring difficulties in effectively fusing these noisy unimodal features. Without explicitly processing feature differences, simply merging noise features will hinder the subsequent fusion and alignment of the hidden layer [3]. In terms of modal interpretability, it is difficult for models to combine multiple modalities to judge people's emotions, such as combining text information and micro-expressions to determine why the emotion expressed when combined is this one. Is it the textual information that dominates people's emotions more, or is it the micro-expressions that better illustrate people's emotions. Some studies have conducted reinforcement learning by using deepseek R1 and introduced the RLVR model for training. The RLVR training process aims to optimize the emotion recognition task of humanni-0.5b using multimodal inputs including video and audio data [4]. Compared with the previous old models, this model can significantly enhance the ability of emotion recognition, as well as improve reasoning and comprehension capabilities in emotion recognition tasks. Finally, in our practical applications, it is inevitable that multimodal data will be missing. Transmission problems caused by poor network conditions prevent images from being transmitted and other situations occur. The main issue is whether the semantics are clear and whether they can be coordinated with other modalities. The core problem lies in semantic understanding and modal association.

2.4. Multimodal Scenarios

Multimodal scenarios require the integration of more than two information modalities, which can break through the limitations of a single modality and enable machines to better understand the thoughts and intentions generated by people. This can make our information more comprehensive and the interaction between humans and machines more self-warming. However, the current situation is complex and diverse in form, so choosing an appropriate multimodal feature fusion strategy is a problem. From multiple information and high and low-frequency information, the effect of multimodal conversation emotion recognition is further improved [5]. It is necessary to combine the above two. The combination of macroscopic carriers and microscopic modes will also lead to the problem of difficult cross-modal alignment. The advancement of multimodal human-computer interaction can solve this problem. Multimodal human-computer interaction is an advanced human-computer interaction method that allows users to interact with computers or intelligent devices through multiple different modalities, such as voice, hand posture, touch, etc [6]. When fusing data from different modalities, a person's cultural background, age and gender will all affect the experiment and its impact on research. Physiological signals are limited. Human responses to different stimuli or situations vary in intensity when using the same emotional response system [7]. Due to the differences in the characteristics of various modalities and individual differences, it is very difficult to precisely match the corresponding emotional expression content in terms of time and semantic dimensions. It is extremely challenging to identify the key nodes of emotions, and these nodes are hard to synchronize, which leads to the model's inability to determine the interactive relationship of emotions.

3. Application Analysis of Multimodal Emotion Recognition

3.1. Research on Learning Emotion Recognition in the Field of Education

3.1.1 Emotion Recognition task

Two indicators, A and F1, are commonly used to evaluate the performance of models. A represents the proportion of correctly classified samples among the total samples, which is divided into overall A (measuring the overall accuracy rate of all samples) and average A (the average accuracy rate of each emotion category, the unweighted average is directly taken as the average value). However, when the number of category samples is unbalanced, A cannot accurately reflect the performance of the classifier in different categories. While F1, as the harmonic average of the precision P and the recall R, can more comprehensively demonstrate the recognition effect of the classifier on positive and negative samples, which can make up for this deficiency. The calculation formula for A is

$$A(y, \hat{y}) = (1/N) \sum_{1}^{N} 1(\hat{y}_i = y_i) \tag{1}$$

Here, N represents the number of samples, y^ represents the prediction result, y represents
The true values. The calculation formulas for P and R are

$$P = (1/M) \sum_{1}^{M} (TP_i / (TP_i + FP_i)) \tag{2}$$

$$R = (1/M) \sum_{1}^{M} (TP_i / (TP_i + FN_i)) \tag{3}$$

Among them, M represents the number of categories, TPi represents the number of true positives (TP) of the i-th category, FPi represents the number of false positives of the i-th category, and FNi represents the number of false negatives (FN) of the i-th category. The calculation formula for F1 is [8]

$$F_1 = (2 \times P \times R)/(P + R) \tag{4}$$

3.1.2 Learn emotion recognition methods

Learning emotion recognition methods are mainly divided into four types: learning emotion recognition methods based on speech information, learning emotion recognition methods based on text information, learning emotion recognition methods based on physiological information, and multimodal learning emotion recognition methods. An end-to-end speech emotion recognition architecture was designed on speech, stacking multiple Transformer layers to enhance the learning ability of aggregated global features. Experimental results on the IEMOCAP database show that, compared with the traditional method EMO-DB (German speech dataset), This method enhances emotion recognition by 20%. From the perspective of text, the learning emotion automatic recognition pipeline framework method has developed to the combination of deep learning and topic modeling. Here is the algorithm for learning emotion recognition evaluation indicators. Common data modalities such as visual information and gestures can all be evaluated. From a physiological perspective, in terms of the technical framework of thermal imaging and visual imaging, the application of physiological signal acquisition usually requires professional equipment, which is very costly. Physiological signal acquisition typically requires professional equipment [8]. Finally, they are combined for multimodal integration. The information and data contained in different modalities will vary. There are still many limitations in the in-depth research of learning emotion recognition at the current specific field stage, which also leaves room for future academic and practical exploration.

3.2. Design and Application of Multimodal Emotion Recognition Technology in Security Systems

The design and application of Multimodal emotion Recognition Technology in Security Systems: The intelligent security system based on multimodal emotion recognition technology mainly consists of five modules: personnel emotion recognition module, environmental status recognition module, data transmission module and communication terminal module. The personnel status recognition module mainly monitors through camera radar, captures the movements of personnel within the visible range, acquires information such as their speed, distance, and azimuth Angle, and transmits it to the communication terminal module via the data transmission module. The remote computer control platform analyzes and recognizes it, makes predictions, and displays the personnel status recognition results: safe or warning [9]. Feature acquisition for humans is mainly designed based on four features. Acoustic features are auditory attributes extracted from audio signals, reflecting the physical and emotional characteristics of sound. Grayscale features are statistical attributes of pixel brightness in an image. Geometric features focus on the shape and spatial structure of objects in images and videos. Texture features represent the repetitive grayscale patterns and spatial distribution in an image.

3.3. MemoCMT: Multi-modal Emotion recognition Based on Feature Fusion of Cross-modal Converters

The MemoCMT architecture and the updated CMT that utilizes cross-attention effectively integrate the mechanisms of audio and text representations extracted by SER and TER. Use the cross-attention module, which is designed to extract key emotional features from audio and text cues. Therefore, CMT has created a better insight feature that enables classifiers to have more important information about the predicted classes. To further improve the extracted emotional features, various aggregation techniques were attempted before the fused features were passed to the classifier. This method helps to select the appropriate reduction features created by the cross-attention mechanism of the classifier in terms of dimensions [10]. The core lies in that the cross-modal attention mechanism is based on the principle of the classical self-attention mechanism. Self-attention enables the model to weigh the importance of different elements in a single input sequence and determine how much attention to pay to each part when dealing with other parts of the same sequence. In multimodal contexts, this concept is extended to facilitate interaction (among different modalities) [11]. Cross-modal converters, as the core model of multimodal processing, have a profound impact on the field of emotion recognition. Firstly, by virtue of its attention mechanism, it can not only precisely align the temporal and semantic nodes of various modalities such as text, speech, and images, but also dynamically associate the emotional features of different modalities, thereby effectively alleviating the problem of difficult cross-modal alignment. Furthermore, in response to the issue of unbalanced data distribution, this model can also deeply explore the multimodal features of niche emotions. Moreover, in the face of the absence of modalities, it can "imagine" key information through existing modalities, significantly enhancing the robustness of the model. At the semantic understanding level, multi-layer attention can iteratively parse implicit emotions and enhance the depth of recognition. Meanwhile, the visualization of attention weights also enhances the interpretability of the model.

4. Summary

Multimodal emotion recognition, by integrating multi-dimensional information, provides significant assistance for people to understand their emotions. This technology breaks through the limitations of single-mode and can more clearly understand people's emotional changes. However, it also faces challenges such as data quality, modal correlation, and insufficient application. The high cost and strong subjectivity of annotation in terms of data quality led to the scarcity of high-quality datasets. In modal association, problems such as diverse dynamic emotional changes and high difficulty in cross-modal alignment have emerged. From the application perspective, there are currently few applications in people's living environments in society. They are mostly confined to controlled environments and mainly appear in laboratories, with few being implemented in real scenarios. In the future, it is necessary to promote it from the laboratory to real scenarios, enhance the human-computer interaction experience, and the lightweight and real-time performance of the model will become the core support for its implementation. At the same time, interdisciplinary integration will also be deepened, leveraging the power of external fields to solve existing problems, enabling multimodal emotion recognition to be applied in a wider range of scenarios.

References

[1] Sepideh Kalateh, Luis a. Estrada-jimenez, Sanaz Nikghadam-hojjati, et al. A systematic review on multimodal emotionrecognition: building blocks, current state, applications, and challenges. Centre of Technology and Systems (CTS-UNINOVA),2829-516 Caparica, Portugal Associated Laboratory on Intelligent Systems (LASI), 2829-516 Caparica, Portugal Department of Electrical Engineering, NOVA School of Science and Technology, NOVA University of Lisbon, 1099-085 Lisbon, Portugal, 2024.

[2] Xiaofei Zhu, Chenyao Li, Xu Chen, et al. Multimodal conversation sentiment recognition based on modal de-heterogeneity and adaptive fusion. Chongqing University of Technology, 2025.

[3] Changzeng Fu, Fengkui Qian, Kaifeng Su, et al. HiMul-LGG: A hierarchical decision fusion-based local–global graph neural network for multimodal emotion recognition in conversation. Hebei Key Laboratory of Marine Perception Network and Data Processing,2025.

[4] Wei Ai1, Fuchen Zhang1, Yuntao Shou1, et al. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. college of computer and mathematics, central south university of forestry and technology, 410004, china. college of computer science and electronic engineering, hunan university, 410082, China Department of Computer Science, State University of New York, 12561, USA, 2010, 30(1): 158-160,2025.

[5] Jiaxing Zhao, Xihan Wei, Liefeng Bo. R1-Omni: explainable omni-multimodal emotion recognition with reinforcement learning. Tongyi Lab, Alibaba Group, 2025.

[6] Xiaofei Luo. A review of multimodal human-computer interaction evaluation. Hangzhou Guyun Business Consulting Co., LTD, 2025.

[7] Manju Priya Arthanarisamy Ramaswamy, Suja Palaniswamy. Multimodal emotion recognition: A comprehensive review, trends, and challenges. Multimodal emotion recognition: A comprehensive review, trends, and challenges, 2024.

[8] Yumei Tan, Shuxiang Song, Haiying Xia. A review of learning emotion recognition research in the field of education. key laboratory of integrated circuits and microsystems, Guangxi Normal University, Guangxi Higher Education Institutions, 2025.

[9] Qingqing Li. A design and application of multimodal emotion recognition technology in security systems. Jiangsu Fushite Electrical Technology Co, LTD,2025.

[10] Mustaqeem Khan, Phuong-NamTran, Nhat Truong Pham, et al. MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. Scientific Reports, 2025, 15;5473.

[11] Karthik Parvathinathan, Sudarshana Karkala, Sazzad Hossain, et al. Multimodal emotion recognition from text and audio using cross-attention fusion. Federal University Oye Ekiti, 2025.