Research Progress and Prospects of Human-Computer Interaction Based on Multimodal Affective Computing

Bote Chen
School of Computer Science and Technology, Qingdao University, Qingdao, China

Keywords: Multimodal Affective Computing, Human-Computer Interaction, Emotion Recognition, Emotion Recognition, Data Fusion.

Abstract: In the era of intelligence, human-computer interaction is developing towards naturalness, emotion, and personalization, and emotional computing has become a research hotspot. It can identify, understand, and respond to the user's emotional state to enhance the experience. However, single-modal emotional data has limitations and it is difficult to fully reflect emotions. Multimodal emotional computing integrates data such as voice, expression, and physiological signals, significantly improving the accuracy of emotion recognition, and is an important direction for human-computer interaction. This paper systematically sorts out its research progress and discusses it from three aspects: theory, technology, and application. First, the characteristics of multimodal emotional data and the necessity of fusion are explained, and then combined with deep learning, the core technologies such as feature level, decision level, and hybrid fusion are analyzed; then its application cases in scenarios such as intelligent assistants, virtual reality, and intelligent driving are discussed. Finally, the current challenges such as data synchronization, modality loss, noise interference, and model generalization are summarized, and suggestions such as optimizing modal fusion algorithms, expanding data sets, and improving model generalization capabilities are proposed for future research. This paper provides a reference for relevant researchers to promote the development of multimodal emotional computing technology and help the intelligent upgrade of human-computer interaction systems.

1 INTRODUCTION

In the era of intelligence, human-computer interaction (HCI) is undergoing a profound transformation from functional to emotional, natural, and personalized. Traditional human-computer interaction systems mainly rely on physical input devices such as keyboards and mice, while modern interaction methods focus more on natural interaction modes such as voice, gestures, and expressions. Affective computing, as the core technology of this transformation, aims to achieve recognition, understanding, and response to human emotional states through the cross-integration of computer science, psychology, and cognitive science.

Rosalind Picard first proposed the concept of affective computing in 1997, pointing out that "emotion is an integral part of intelligence", a view that laid a theoretical foundation for subsequent research (Plutchik, 2001). However, single-modal affective data (such as recognizing emotions only through voice or facial expressions) has obvious limitations. For example, voice emotion recognition is easily affected by environmental noise and language differences; facial expression recognition may lead to misjudgment due to lighting conditions, occlusion or cultural differences; and text-based sentiment analysis is difficult to capture non-verbal information. These limitations make single-modal emotion recognition systems often unstable in practical applications. Multimodal affective computing can capture the user's emotional state more comprehensively and accurately by integrating multiple data sources such as voice, expression, and text, thereby significantly improving the robustness and user experience of human-computer interaction systems.

The research on multimodal affective computing not only has important theoretical value, but also shows broad prospects in practical applications. From a theoretical perspective, it promotes the deep integration of affective computing and multimodal information processing, and provides new solutions to cutting-edge issues such as affective modeling and

Chen, B.
Research Progress and Prospects of Human-Computer Interaction Based on Multimodal Affective Computing.
DOI: 10.5220/0014306900004718

In *Proceedings of the 2nd International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2025)*, pages 23-27
ISBN: 978-989-758-792-4

cross-modal representation learning. From an application perspective, multimodal affective computing has been widely used in intelligent assistants (such as Siri, Alexa), virtual reality (VR)/augmented reality (AR), intelligent driving, distance education, mental health monitoring and other fields. For example, in an intelligent driving system, by real-time monitoring of the driver's facial expressions, voice intonation and physiological signals, the system can promptly warn of fatigue or anger, thereby reducing the risk of accidents.

In recent years, the rapid development of deep learning technology has injected new vitality into multimodal sentiment computing. Models such as convolutional neural networks (CNNs) have performed well in sentiment feature extraction and cross-modal fusion (Zhao et al., 2019, Li et al., 2021). For example, CNNs can efficiently extract local features in image and speech emotion recognition. Nevertheless, multimodal sentiment computing still faces many challenges, such as data synchronization and alignment, modality loss and noise interference, model generalization ability, and computational efficiency and real-time performance.

In the development of multimodal sentiment analysis, scholars have summarized existing technologies from different perspectives. Author PENG X J et al. summarized existing technologies through sentiment analysis based on visual information, voice information, text information and brain information (Peng, 2018). Author Li D et al. introduced and differentiated sentiment recognition, opinion mining and sentiment analysis in detail, and classified and summarized the three modal technologies of text, voice and vision used in sentiment analysis (Soleymani et al., 2017). Author HUDDAR M G et al. discussed the existing single-modal sentiment analysis technology, and then summarized the multimodal sentiment analysis in recent years. While summarizing the literature, they also pointed out the method of modal fusion (Huddar et al., 2019). Author GAO J et al. summarized the existing modal fusion algorithms from the perspective of deep learning (Gao et al., 2020).

This paper focuses on the fusion and calculation of multimodal emotional data, and explores its theoretical basis, technical methods, application scenarios and challenges. First, starting from the characteristics of multimodal emotional data, the necessity and core tasks of multimodal data fusion are explained. Secondly, combined with deep learning methods, its application in multimodal emotional computing is discussed, and practical cases in scenarios such as intelligent assistants, virtual reality, and intelligent driving are analyzed. This paper aims to provide reference for researchers in related fields by comprehensively reviewing multimodal emotional data computing, summarizing its theories, technologies and applications, and analyzing existing technologies, so as to promote the development of multimodal emotional computing technology and help the intelligent upgrade of human-computer interaction systems.

2 MULTIMODAL AFFECTIVE COMPUTING

2.1 Basic Theory of Multimodal Affective Computing

Affective computing aims to identify, understand, simulate and respond to human emotional states through technical means. It mainly includes three core tasks: emotion recognition, emotion generation, and emotion feedback. Single-modal emotion data often has limitations and cannot fully reflect the user's emotional state. For example, speech emotion recognition may be interfered by environmental noise, while facial expression recognition may be affected by lighting conditions. At this time, multimodal emotion recognition solves this problem well. Multimodal emotion data refers to emotion-related information collected through multiple perceptual channels (such as speech, facial expressions, physiological signals, text, etc.). Each modality provides a unique way of expressing emotions and can capture the user's emotional state from different angles (Wu et al., 2014). Through multimodal data fusion, the shortcomings of a single modality can be made up (Zhang et al., 2020).

2.2 Current Challenges of Multimodal Affective Computing

Although multimodal data fusion has significant advantages, the effect of modal fusion will directly affect the accuracy of the results (Wu et al., 1999). In addition, it still faces many challenges in practical applications: data synchronization and alignment: data of different modalities may have different timestamps and sampling rates. How to achieve data synchronization and alignment is an important technical challenge; modality missing and noise interference: in practical applications, data of some modalities may be missing or interfered by noise; model generalization ability: multimodal emotion

computing models need to have good generalization ability and be able to adapt to different users and scenarios. However, due to individual differences and environmental changes, the generalization ability of the model is often insufficient; computational complexity and real-time requirements: multimodal data fusion involves the collaborative analysis of multiple modalities, and the algorithm complexity is relatively high.

2.3 Detailed Introduction to Convolutional Neural Algorithms in Deep Learning

In recent years, deep learning methods have made significant progress in emotion recognition tasks, among which convolutional neural networks (CNNs) are particularly famous. CNN is a deep learning model specifically designed to process grid-structured data (such as images, speech, and video). Its core idea is to effectively extract hierarchical features of input data through local perception, weight sharing, and pooling operations, thereby reducing the complexity and computational complexity of the model and improving the accuracy of the overall decision-making results (Krizhevsky et al., 2017). Convolutional neural networks have been widely used in the field of deep learning by extracting features through convolution operations, relying on the two major characteristics of local connection and weight sharing (Tan et al., 2022). On this basis, graph convolutional neural networks came into being and developed rapidly. After multiple stages of evolution, many variants with their own characteristics have been derived. The first generation of graph convolutional neural networks is the spectral convolutional neural network (Bruna et al., 2013). It was the first to perform convolution operations on graph data. The model is based on spectral graph theory and graph signal processing (Spielman, 2012, Sandryhaila & Moura, 2013). According to the convolution theorem, the graph convolution operation is implemented in the spectral domain to calculate the features of each node (Shuman et al., 2013).

From the perspective of structural composition, convolutional neural networks are mainly composed of convolutional layers, pooling layers, fully connected layers, and output layers. Among them, the convolutional layer is composed of multiple feature maps, and its main function is to extract local feature information of the input through convolution operations. Unlike traditional neural networks, where each layer of neurons is connected to all neurons in the previous layer, convolutional neural networks have local connection characteristics, and neurons are only connected to some neurons in the previous layer (Hao, 2020). Convolutional layers are connected through convolution kernels, which are shared weight matrices. The main function of the pooling layer is to merge the semantically similar features extracted from the convolutional layer, aiming to extract the most useful feature information while reducing the dimension of the features (Chen, 2020). Commonly used pooling methods are max-pooling and average-pooling. The fully connected layer is located before the output layer, and its main function is to map the extracted distributed features to the sample label space and integrate the local information with category distinction.

The fusion of information from different modalities is the core issue of multimodal sentiment computing, which integrates the information extracted from different single modalities into a multimodal feature (Zhang et al., 2020). At present, the modal fusion methods are mainly divided into three types: feature-level fusion, decision-level fusion and hybrid fusion. The following is an introduction and explanation of additive fusion in feature-level fusion.

Additive fusion is a direct and simple multimodal feature fusion method. Its basic idea is to add pixels from feature maps (or feature vectors) from different modalities at corresponding positions to obtain a comprehensive feature map (or feature vector).

First, emotional features are extracted from the data of each modality. Since the features of different modalities may have different dimensions and spatial resolutions, these features need to be aligned before additive fusion. The aligned feature maps (or feature vectors) are added pixels at corresponding positions to obtain a comprehensive feature map (or feature vector). This comprehensive feature map (or feature vector) contains emotional information from different modalities. The fused features can be used for subsequent emotion recognition or analysis tasks. For example, the fused features can be input into a classifier for emotion classification or used for emotion intensity assessment.

Additive fusion is an effective multimodal emotion fusion method, especially when fast implementation and simple application are required. The performance and accuracy of multimodal emotion fusion can be further improved by combining other more complex fusion methods or algorithms.

2.4 Current Applications of Multimodal Affective Computing

Multimodal affective computing has a wide range of application scenarios in human-computer interaction: Emotional interaction in smart assistants (such as Siri): Smart assistants can identify users' emotional states by analyzing multimodal data such as their voice, facial expressions, and text, thereby providing more personalized and considerate services. For example, when a user shows anxiety, smart assistants can provide comforting words or suggestions; Emotional feedback in virtual reality (VR) and augmented reality (AR): In VR and AR environments, multimodal affective computing can enhance users' immersion. For example, by analyzing users' facial expressions and physiological signals, the system can adjust the emotional feedback of virtual characters in real time, allowing users to experience a more realistic interactive experience; Driver emotion monitoring in intelligent driving systems: Intelligent driving systems monitor drivers' emotional states in real time by analyzing their facial expressions, voice, and physiological signals. When drivers show fatigue or anger, the system can issue warnings or take appropriate safety measures.

3 DISCUSSION ON MULTIMODAL AFFECTIVE COMPUTING

This paper summarizes the challenges and development trends of multimodal sentiment analysis by understanding the current status of research. In multimodal sentiment analysis, data collection under the condition of people expressing problems naturally is one of the main problems. There are few data sets and most of them are composed of three modalities: vision, text and speech, lacking modal data such as posture and brain waves. There are certain differences between different data. Due to different collection conditions and the subjectivity of annotations, data deviation and inconsistent annotations are very common in different data sets. There is a certain correlation between the features extracted from different modalities. How to effectively use the correlation between modalities to improve the accuracy of sentiment analysis is one of the future research directions. When conducting multimodal sentiment analysis, too many modalities will increase the complexity of the fusion algorithm, and too few modalities will affect the accuracy of the results. When fusion of modalities, the optimal weight allocation of different modalities in different environments is one of the important factors affecting the results of sentiment analysis. How to give a larger weight to the modality that has the greatest impact on the analysis results is one of the key directions of modal fusion in the future.

In multimodal sentiment analysis technology, some literature only improves the accuracy of single-modal feature extraction, but ignores the contextual information in the sequence, resulting in insufficient feature mining of different modalities. With the deepening of research, researchers have introduced RNN, LSTM, GRU and other networks to extract contextual information and improve the accuracy of sentiment analysis, but it is easy to lose information when processing long sequences. At this stage, it is possible to consider using multi-level GRU to encode contextual information to solve the problem of long-term dependence, so as to obtain more comprehensive sentiment information. In terms of modal fusion technology, researchers use multi-layer fusion methods to perform modal fusion. This method can improve the quality of single-modal feature vectors and achieve better results when the data is large, but it may cause overfitting problems in small samples. Since the attention mechanism plays an important role in finding the optimal weight in modal fusion, the tensor can project the features of all modalities into the same space to obtain a joint representation space, which is easy to calculate the interaction between modalities. In recent years, the mainstream modal fusion methods are attention-based methods and tensor-based methods.

4 CONCLUSION

This paper reviews the research progress in the field of multimodal emotion computing and analyzes it from three dimensions: theoretical basis, technical methods, and application practice. In terms of theoretical basis, the characteristics of multimodal emotion data and the necessity of fusion are explained, and the limitations of a single modality and the advantages of multimodal complementarity are pointed out. At the technical method level, multimodal fusion strategies based on deep learning are discussed, including feature-level fusion, decision-level fusion, and hybrid fusion, and the application mechanism of convolutional neural networks in multimodal emotion recognition is analyzed. In the application practice part, the practical value of multimodal emotion computing technology

is verified through typical scenarios such as smart assistants, VR/AR, and smart driving. The study also pays special attention to the core challenges currently faced, such as data synchronization, modality loss, and noise interference.

Research has found that multimodal emotion computing significantly improves the accuracy and robustness of emotion recognition by integrating data such as speech, facial expressions, and physiological signals. Deep learning models have shown great potential in feature extraction and cross-modal association modeling. However, existing technologies still have obvious limitations.

Future research on multimodal affective computing should focus on the following directions: building richer multimodal datasets; developing lightweight fusion algorithms; exploring cross-modal self-supervised learning; and promoting cross-innovation between affective computing and cutting-edge technologies. With breakthroughs in key technologies, multimodal affective computing is expected to be deeply applied in areas such as smart healthcare and emotional AI assistants. This study provides a systematic reference for related fields, but the technology is still in a rapid development stage and requires continuous collaborative innovation between academia and industry.

REFERENCES

Bruna, J., Zaremba, W., Szlam, A., et al.: 'Spectral networks and locally connected networks on graphs.' arxiv preprint arxiv:1312.6203, 2013

Chen, J.: 'Research on multi-label classification of unstructured medical text based on multi-channel convolutional neural network.' Southwest Jiaotong University, 2020

Gao, J., Li, P., Chen, Z., et al.: 'A survey on deep learning for multimodal data fusion.' Neural Computation, 2020, 32 (5): 829-864

Hao, X.: 'Research on video description algorithm based on deep learning sequence model.' Beijing University of Posts and Telecommunications, 2020

Huddar, M., Sannakki, S., Rajpurohit, V.: 'A survey of computational approaches and challenges in multimodal sentiment analysis.' International Journal of Computer Sciences and Engineering, 2019, 7(1): 876-883

Krizhevsky, A., Sutskever, I., Hinton, G E.: 'ImageNet classification with deep convolutional neural networks.' Communications of the ACM, 2017, 60(6): 84-90

Li, D., Chai, B., Wang, Z., et al.: 'EEG emotion recognition based on 3-D feature representation and dilated fully convolutional networks.' IEEE Transactions on Cognitive and Developmental Systems, 2021, 13(4): 885-897

Peng, X.: 'Multi-modal affective computing: a comprehensive survey.' Journal of Hengyang Normal University, 2018, 39(3): 31-36

Plutchik, R.: 'The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice.' American Scientist, 2001, 89(4): 344-350

Sandryhaila, A., Moura, J.: 'Discrete signal processing on graphs.' IEEE transactions on signal processing, 2013, 61(7): 1644-1656

Shuman, D., Narang, S., Frossard, P., et al.: 'The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains.' IEEE signal processing magazine, 2013, 30(3): 83-98

Soleymani, M., Garcia, D., Jou, B., et al.: 'A survey of multimodal sentiment analysis.' Image and Vision Computing, 2017, 65: 3-14

Spielman, D.: 'Spectral graph theory.' Combinatorial scientific computing, 2012, 18: 18

Tan, Y., Wang, J., Zhang C.: 'A review of text classification methods based on graph convolutional neural networks.' Computer Science, 2022, 49(08): 205-216

Wu, C., Lin, J., Wei, W.: 'Survey on audiovisual emotion recognition: databases,features, and data fusion strategies.' APSIPA Transactions on Signal and Information Processing, 2014, 3: e12

Wu, L., Oviatt S., Cohen, P.: 'Multimodal integration— a statistical view.' IEEE Transactions on Multimedia, 1999, 1(4): 334-341

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

Zhao, J., Mao, X., Chen, L.: 'Speech emotion recognition using deep 1D & 2D CNN LSTM networks.' Biomedical signal processing and control, 2019, 47: 312-323