Vol. 9 (2025): Proceedings of the 2025 International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability
(IC-AIMEES 2025)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ic-aimees.202516

Survey of Human-computer Interaction Based on Multimodal Fusion

Zihui Zhao*

School of Computer Science and Technology / School of Artificial Intelligence, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China

**Corresponding author: Zihui Zhao*, ORCID : 0009-0009-2771-1631*

Abstract

Multimodal fusion technology achieves information communication and exchange between human and computer by integrating different modal information such as vision, speech, and touch, which has become an important research direction in the field of human-computer interaction. This paper focuses on the four mainstream multimodal fusion methods of graph-based feature fusion, cross-modal attention technology, cross-correlation attention architecture and multimodal emotion recognition technology, compares and analyzes their technical principles, advantages, disadvantages and application scenarios, and systematically sorts out the differences in technical characteristics. By integrating multiple input methods, these methods significantly improve the user interface interaction experience, optimize the efficiency of multi-source information processing, and provide new ideas for interaction design in complex scenes. Research shows that multimodal fusion human-computer interaction technology can effectively reduce user cognitive load and improve operation efficiency, which has important application value in education, medical care, smart home and other fields. In the future, it is necessary to solve the challenges of insufficient cross-modal data alignment accuracy and high real-time requirements, and explore the deep combination of affective computing and multimodal fusion.

Keywords

multi-modal fusion, human-computer interaction, feature fusion, cross-modal attention, emotion recognition

1. Introduction

With the rapid popularization of intelligent terminal devices, human-computer interaction is changing from traditional single input mode to natural interaction integrating multi-sensory channels. Multimodal fusion technology realizes human-computer information exchange by integrating information from different modalities, which has become a core research direction (Tao et al., 2022). Its core value lies in simulating human natural perception and communication mode, and enhancing interaction reliability through multi-source information collaboration, such as combining multiple inputs to improve operation accuracy (Tao et al., 2022) in medical scenarios.

The existing review works have promoted the development of the field. Tao et al. (2022) systematically sorted out five technical directions of multimodal human-computer interaction and provided a basic framework. Lee et al.

(2025) focused on the multimodal interaction based on Electroencephalogram (EEG), and analyzed the neural network architecture and challenges of biosignal fusion. Schreiter et al. (2025) summarized the application status of multimodal interaction in interventional radiology and surgery scenarios, and emphasized the clinical value of speech-hand interaction combination. Chen et al. (2022) focus on the theoretical framework of multimodal fusion algorithms and compare the performance differences between early feature-level fusion and emerging deep learning methods. Liu et al. (2023) focused on hardware implementation and proposed a sensory-computation collaborative optimization scheme for smart home scenarios. Wang et al. (2024) and Sun et al. (2024) extended the application research in the fields of education assistance and medical health; Bao et al. (2025) and Chen Yunfang's team (2023) promote the transformation of methodology to dynamic evaluation. The former uses the technology maturity matrix to compare the performance of fusion methods, and the latter reveals the transfer logic of cross-correlation attention architecture through the technology evolution roadmap (Bao et al., 2025, Chen et al., 2023).

However, existing studies have limitations. Technical analysis focuses on a single field and lacks cross-scene adaptation comparison. The systematic evaluation of emerging methods such as graph feature fusion and cross-modal attention is insufficient. Research on the combination of affective computing and multimodal fusion is scattered, and most of them have not established a unified technical evaluation standard, and insufficient attention has been paid to the scene adaptation in vertical fields such as education and medical care and the challenges of real-time fusion in edge computing environment. In this paper, integrating the above core review found that incorporates the characteristics of figure, for the first time across the modal attention such as technical path combined with vertical scene depth, by comparing the different methods for the border, provide the comprehensive decision-making basis for technology selection, drive a multimodal interaction system to the development of intelligent, scene.

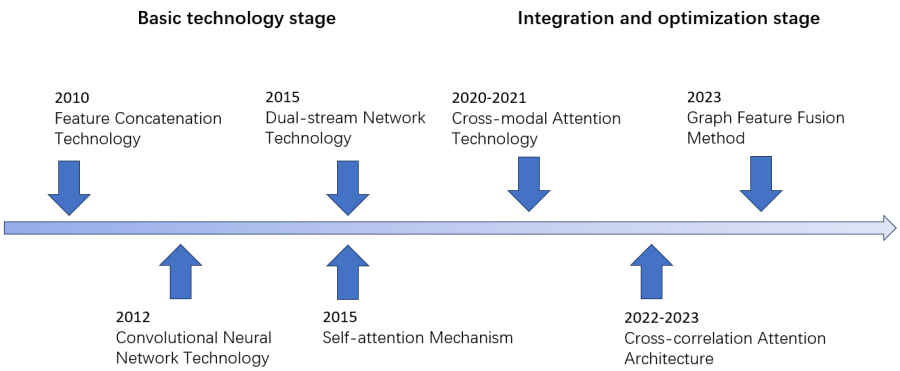
2. Overview of Multimodal Fusion Technology

2.1 Development of Multimodal Fusion Technology

Multimodal fusion technology has experienced the evolution from single mode to multimodal collaboration. Early interactive systems rely on a single input such as keyboard and mouse, which cannot simulate natural human communication. With the advancement of sensor technology and the improvement of computing power, researchers have begun to explore the integration of multi-modal information such as vision, speech, and touch, giving rise to the prototype of multi-modal fusion technology (Tao et al., 2022).

As shown in Figure 1, the evolution of multimodal fusion technology can be divided into three stages. The first stage (2000-2010) is dominated by simple modal superposition, such as speech recognition systems adding cameras to capture facial expressions, but lacking deep interaction between modalities. The second stage (2010-2020) focuses on collaborative optimization between modalities, and improves system performance through feature-level fusion, such as combining voice commands and gesture control in intelligent assistants. The third stage (2020 - present) emphasizes the dynamic adaptive fusion, in the deep learning model cross modal semantic alignment and real-time decision-making (Sun et al., 2024).

Figure 1: Timeline of human-computer interaction technology development based on multi-modal feature fusion



Multimodal fusion technology is inseparable from the development of human-computer interaction

requirements. The diversity of application scenarios makes the limitations of a single modality prominent. For example, relying on voice commands alone in noisy environments is error-prone, and the system robustness is significantly improved after introducing visual information. This demand-driven technological progress makes multimodal fusion gradually become the core means to realize natural and efficient human-computer interaction (Sun et al., 2024).

At present, multimodal fusion technology faces new opportunities and challenges. The development of artificial intelligence provides stronger tools for cross-modal representation learning, but the efficient alignment and real-time processing of different modal data is still a key issue (Wang et al., 2024). In the future, with the popularization of edge computing and 5G technology, multi-modal fusion is expected to achieve breakthroughs in a wider range of scenarios, bringing more possibilities for human-computer interaction.

3. Key Technologies of Multimodal Fusion Human-Computer Interaction

3.1 Feature Fusion of Graphs

Graph-based feature fusion achieves information complementarity by constructing an inter-modal correlation graph, which represents different modal data as nodes in the graph, and describes the potential correlation between modalities by edges, which effectively integrates multi-source information. In the field of human-computer interaction, this technology has attracted attention because of its ability to intuitively model complex interaction relationships (Deng et al., 2025).

Table 1: Overview of the principles and characteristics of the key technologies of multimodal fusion human-computer interaction

	Principle	Features
Feature fusion of graphs	The inter-modal correlation graph is constructed, and the data of different modalities are represented as nodes in the graph. The potential correlation between modalities is depicted by edges, and the cross-modal feature propagation and aggregation are realized by graph neural network	1) The graph structure is flexible to adapt to multi-modal combination scenarios; 2) Support multi-hop reasoning to capture deep semantic associations; 3) It is robust to modality missing, but faces the challenge of modality heterogeneity
Cross-modal attention techniques	By simulating the human selective attention mechanism, the correlation weights between the features of each modality are automatically learned by the attention mechanism, and the key interaction information is accurately extracted	1) Dynamic weight allocation ADAPTS to scene changes; It supports end-to-end training and reduces feature engineering. 2) attention visualization improves interpretability, 3) The high computational complexity affects the real-time performance, and there is a mode imbalance problem
Cross-correlation attention architecture	The cross-correlation operation is used to capture the dependence of different modal features in the spatio-temporal dimension, and the cross-modal feature map similarity matrix is calculated to establish the dynamic correlation	1) Local feature comparison can capture subtle cross-modal correlations; 2) dynamic weight assignment ADAPTS to complex scenes; 3) strong tolerance to modal asynchrony
Multimodal emotion recognition technology	Multi-modal information such as speech, facial expression, and physiological signals are integrated to determine the user's emotional state through feature extraction and fusion strategies	1) Multi-source information complementarity improves recognition robustness; 2) Cross-modal correlation captures the delicate emotional dimension; 3) real-time feedback to support dynamic interactive adjustment

The core of graph feature fusion is to construct a graph structure that reflects the semantic relationship between modalities. Taking smart home as an example, when the user uses voice commands and gestures to control the lights at the same time, the system uses voice and gesture feature vectors as graph nodes, and constructs edge connections through semantic similarity calculation. This representation preserves the characteristics of each modality, and cross-modal feature propagation and aggregation are realized through

Graph Neural Network (GNN). Rui Yao et al. pointed out that "the cross-modal feature interaction fusion module based on channel-spatial attention can realize the complementary fusion of different modal features" (Yao et al., 2025). This mechanism is applicable in graph feature fusion, and the information transmission strength between modalities can be dynamically adjusted through the attention weight.

As shown in Table 1, the advantages of graph feature fusion are reflected in three aspects: first, the flexibility of graph structure can adapt to different modal combination scenarios, such as doctors in medical assistance systems annotate medical images by voice description combined with 3D gestures, and the graph topology can be dynamically adjusted with interaction (Yu et al., 2025). Second, the graph-based representation naturally supports multi-hop reasoning and captures deep semantic associations. For example, in educational applications, students 'gesture trajectories and voice explanations are associated with knowledge graph concept nodes through multi-layer graph convolutional networks. Third, it is robust to modal loss. When a certain modality (e.g. speech) is invalid due to environmental noise, the system can still infer the missing information through the graph structure.

However, it faces the challenge of modality heterogeneity. As stated by Deng et al., "Modality heterogeneity and expression inconsistency pose challenges for effective feature fusion" (Deng et al., 2025). Specifically, the sampling frequency, feature dimension, and semantic granularity of different modalities are different, and direct construction of association graphs may lead to information loss. The current research adopts two solutions: mapping heterogeneous features into a unified space through a modality alignment network, such as the "gesture rotation mapping" method proposed by Yu Xinyi's (2025) team; Namely, an adaptive graph learning mechanism is designed to dynamically optimize the node connection weights.

In the practice of human-computer interaction, graph-based feature fusion technology has been successfully applied in many scenarios. In the virtual reality training system, by integrating eye tracking, gesture operation and voice feedback data, a three-dimensional interaction graph is constructed, which significantly improves the accuracy of operation guidance. In the field of intelligent customer service, a more natural emotional interaction experience is achieved by establishing a node correlation graph between user's voice emotion and facial micro-expression. In the future, with the deep integration of graph neural network and adaptive learning technology, this method is expected to further break through the cross-modal semantic gap and give stronger situational understanding ability to human-computer interaction systems.

3.2 Cross-Modal Attention Technology

Cross-modal attention technology simulates the human selective attention mechanism, realizes the dynamic capture and weighted fusion of correlation features between different modalities, and has significant advantages in scenarios that require real-time response and multi-source information collaborative processing. The core of the cross-modal attention technology is to automatically learn the correlation weights between the features of each modality and accurately extract the key interaction information (Luo et al., 2025).

The technical implementation consists of three components: the feature encoder converts the original input into a high-dimensional feature vector; The attention computing module quantifies the correlation strength between modalities through a learnable weight matrix. The fusion output layer integrates the weighted features into a unified semantic representation. Zhixin Luo et al. pointed out that "the cross-modal attention mechanism can effectively perform weighted fusion of audio and video features, so that the model can better capture the interaction between audio and video modalities" .(Qu and Xu, 2025) This mechanism is prominent in complex tasks such as emotion recognition, which improves the accuracy of the system's understanding of the user's intention.

Compared with traditional methods, cross-modal attention technology has roughly three characteristics: dynamic weight allocation ADAPTS to different scenarios, such as increasing the weight of gesture modality when the environment is noisy in smart home, and enhancing the weight of speech modality when the light is insufficient. It supports end-to-end training to avoid traditional tedious feature engineering, and automatically discovers cross-modal associations at different abstraction levels through cascaded multi-head attention layers. Attention visualization supports system interpretability, and developers analyze the key feature areas focused by the model through heat maps.

In practical application scenarios, this technology has shown remarkable results. For example, in the field

of education, intelligent tutoring systems can more accurately assess students 'learning status by synchronously analyzing multimodal information such as students 'voice responses, facial micro-expressions and gesture pointing, and provide strong support for personalized teaching. Qu and Xu (2025) show that "using cross-modal attention mechanism to explicitly construct correlations between modalities" can effectively improve the performance of emotional interaction systems. These cases verify its practical value in complex interaction scenarios.

Current challenges include high computational complexity and modal imbalance. The former is caused by global feature comparison, which affects the real-time performance of mobile devices. The latter shows that the weak mode is easily masked by the dominant mode. Researchers have proposed improvement schemes such as hierarchical attention and lightweight attention. Future directions include developing efficient attention computing paradigms to reduce resource consumption, exploring adversarial training to balance modal contributions, and introducing meta-learning mechanisms to make the system quickly adapt to new scenarios and user groups (Qu and Xu, 2025, Zhao et al., 2025).

3.3 Cross-Correlation Attention Architecture

Cross-correlation attention architecture realizes multi-modal fusion by modeling the dynamic interaction between modalities. It was originally applied to object tracking in computer vision, and has been transferred to human-computer interaction scenarios in recent years (Chen et al., 2023). The combination of cross-correlation attention architecture with human-computer interaction is to analyze the user's multi-channel input signals in real time to achieve natural interactive experience.

The technical principle of cross-correlation attention architecture is to calculate the similarity matrix of cross-modal feature maps to establish dynamic correlation. Taking the intelligent driving system as an example, when the vehicle is driving at a complex intersection, the system will simultaneously capture the dynamic trajectory of the driver's voice command and hand steering operation. Through the cross-correlation calculation of the voice spectrum features and the steering wheel sequence motion features, the attention heat map reflecting the degree of coordination between commands and operations is generated. This heat map can determine the consistency of the driver's intention and operation in real time. If there is a delay or conflict between the voice command and the hand movement, the system can trigger the reminder mechanism in time to improve driving safety. The chained frame processing method proposed by Chen Yunfang's team shows that "taking two consecutive frames of pictures as input, the target association problem is transformed into the problem of regression of two frames of detection boxes" (Chen et al., 2023), which is suitable for processing cross-modal temporal alignment such as voice-gesture. The specific implementation process is as follows: firstly, the local features of each modality are extracted through the convolutional network, and then the cross-correlation matrix between the feature maps is calculated. Finally, the attention weights are generated by softmax normalization. As shown in Table 1, the technical principle of the proposed architecture determines its advantage in fine-grained association capture. As shown in Table 1, these features make it uniquely valuable in real-time interaction scenarios, but there are also problems of computational efficiency and modal balance.

In human-computer interaction scenarios, cross-correlation attention architecture shows unique advantages. Its local feature comparison mechanism can capture subtle cross-modal correlations. For example, in the virtual piano teaching scene, the accuracy of performance movements can be accurately judged by analyzing the spatio-temporal consistency between the user's finger position (visual modality) and the key sound (auditory modality). The dynamic weight distribution feature enables it to adapt to complex interaction scenarios. Zhao et al. (2025) pointed out that the traditional cross-modal attention "is easy to ignore some fine-grained spatio-temporal information", while the cross-correlation attention can retain richer interaction details through pixel-level feature comparison. At the same time, the architecture has strong tolerance to modal asynchrony. When there is a short delay between voice command and gesture input, the correlation between them can still be maintained through a sequential sliding window.

The current challenges of this technology are mainly focused on computational efficiency and modal span. The computational overhead caused by cross-correlation operation may affect the real-time performance in resource-constrained scenarios such as mobile devices. For the combination of visual-tactile modalities with large differences, special feature mapping methods need to be designed to ensure the fusion effect. To solve these problems, researchers have proposed optimization schemes such as hierarchical cross correlation and

lightweight feature extraction. Future research can further optimize the computing module to reduce resource consumption, explore cross-modal contrastive learning to improve feature compatibility, and combine the meta-learning mechanism to make the system quickly adapt to new interaction scenarios, so as to further expand its application potential in intelligent education, medical assistance and other fields (Bao et al., 2025, Chen et al., 2023).

3.4 Multi-modal emotion Recognition Technology

Multimodal emotion recognition technology integrates multi-modal information such as voice, facial expression, and physiological signals to accurately determine the user's emotional state, which is the key to improve the emotional ability of human-computer interaction. Its core is to use the complementarity of different modal data to overcome the limitations of a single recognition method in complex scenes. Tao Jianhua and other scholars pointed out that "multimodal fusion emotion recognition research is increasingly paid attention by researchers to fully exploit the complementarity of different modal data" (Sun et al., 2025), which makes it of outstanding value in fields such as education assistance and mental health monitoring.

The technical implementation includes three steps: feature extraction. Special methods are designed for different modalities, such as convolutional network to analyze facial expression images and time-frequency analysis to extract speech emotional parameters. The fusion strategy is divided into early fusion (feature level) and late fusion (decision level). Qu and Xu (2025) show that "the core part of multi-modal sentiment analysis is multi-modal representation learning and fusion, which encodes and integrates multi-modal representations to understand the emotion behind the original data". For example, the remote psychological counseling system analyzes the frequency of voice trembling (auditory), the change of eyebrow wrinkles (visual) and galvanic skin response (tactile) of the consultant, and outputs a comprehensive emotional score through hierarchical fusion.

The advantages of multimodal emotion recognition are mainly reflected in three aspects. The complementarity of multi-source information can improve the robustness of recognition. For example, when the user wears a mask and loses facial information, the system can judge the emotional state through the comprehensive analysis of speech prosody and body movements. The cross-modal correlation mechanism can capture more delicate emotional dimensions, and can effectively distinguish easily confusing emotions such as anxiety and anger by combining features such as speech pause duration and finger pressure strength. The real-time feedback feature supports dynamic interactive adjustment, and the educational robot can automatically slow down the speaking speed of the explanation according to the confused expression of the students, which is a typical case. Deng et al. (2025) emphasize that "multimodal emotion recognition is essential for understanding human emotions from multiple sources such as speech, text, and video", which makes it irreplaceable in personalized interaction scenarios.

This technology has achieved remarkable results in practical applications. In the field of intervention for children with autism, multimodal emotion recognition system can capture the abnormal body movements and vocal characteristics of children, and provide comprehensive evaluation basis for therapists. These practices verify the viewpoint of Tao Jianhua's team, that is, "Multimodal emotion understanding and interaction technology aims to fully model multi-dimensional information from audio, video and physiological signals to achieve more accurate emotion understanding" (Sun et al., 2025).

The current challenges in this field mainly focus on the accuracy and computational efficiency of cross-modal alignment. There is time asynchrony in different modalities of emotional expression, such as facial expression changes often lag behind the burst of speech emotion, which requires the system to have a dynamic alignment mechanism. At the same time, the real-time processing requirements of mobile devices put forward higher requirements for lightweight models. In recent years, the dual-channel attention fusion method represented by DBSQFusio further optimizes the modal imbalance problem through adaptive weight allocation, which provides new ideas for the development of multimodal emotion recognition technolog (Liu et al., 2025)y.

In summary, graph feature fusion, cross-modal attention technology, cross-correlation attention architecture and multimodal emotion recognition technology have their own characteristics in multimodal fusion human-computer interaction, and the specific comparison is shown in Table 1.

4. Application Expansion of Multimodal Fusion in Specific Fields

4.1 Field of Computer-Aided Design (CAD)

In computer aided design (CAD), multimodal natural interaction technology has significantly improved the design efficiency. Niu et al. (2022) proposed a multi-modal interaction framework that integrates gesture, speech, and eye tracking, allowing designers to directly manipulate 3D models through natural movements, reducing the dependence on traditional keyboards and mice. The framework constructs a modal correlation graph by feature fusion of graphs, and maps hand posture, voice commands, and gaze point information into a unified design space to achieve a "what you want is what you get" interaction experience.

4.2 Text Classification and Sentiment Analysis

The breakthrough in the application of multimodal technology in the field of text is reflected in cross-modal semantic alignment. The multi-module fusion network proposed by Yu et al. (2022) associates text features with additional information such as images and audios through the attention mechanism, which significantly improves the accuracy of multi-label text classification. In sentiment analysis tasks, RAFT (Robust Adversarial Fusion Transformer) enhances modal robustness through adversarial training, which can still maintain high sentiment classification accuracy even when part of the modal data is missing, and provides a reliable tool for public opinion analysis of social media (Wang et al., 2025a).

5. Final Remarks

Multimodal fusion technology significantly improves the naturalness and efficiency of interaction by integrating multiple sensory information. This paper reviews the four methods of graph-based feature fusion, cross-modal attention technology, cross-correlation attention architecture and multimodal emotion recognition technology, and analyzes their principles, advantages and disadvantages, and application scenarios. These technologies have significant effects in reducing cognitive load and improving experience, and have been applied to education assistance, telemedicine, smart home and other fields. However, these technologies still face the challenges of insufficient cross-modal data alignment accuracy and high real-time response requirements, which need to be further optimized.

Future research is expected to be carried out in three directions: in the aspect of multimodal representation learning, more efficient algorithms are developed to solve the modal heterogeneity, such as improving the adaptive learning ability of graph neural networks to improve cross-modal semantic alignment; Affective computing is deeply combined with multimodal fusion, the emotion recognition model that fuses physiological signals and behavioral features captures the mental state more accurately, and explores a lightweight framework for specific scenarios (such as depression screening). For complex scene technology adaptation, for dynamic environments such as smart home and automatic driving, the fusion mechanism with environmental perception ability is studied, so that the system can automatically adjust the strategy.

From the perspective of technology evolution, multimodal interaction is leap ingfrom "human-computer collaboration" to "human-intelligence collaboration" (Wang et al., 2025b). By introducing cognitive science theory and adaptive learning, the system will more accurately understand the user's intention and realize truly natural intelligent interaction. At the application level, multimodal technology will penetrate into a wider range of fields, such as the education field combining eye tracking and voice analysis to realize personalized learning recommendation. In the field of medical rehabilitation, multimodal biosignal fusion will be used to improve the accuracy of remote diagnosis and treatment. With the popularity of edge computing devices, real-time fusion under resource-constrained conditions will become the key to engineering, promoting multimodal technology from the laboratory to practical application.

References

Bao, Y., Zhao, X., Zhang, P., Qi, Y. and Li, H., (2025). HIAN: A hybrid interactive attention network for multimodal sarcasm detection. *Pattern Recognition,* vol. 164, p. 111535.

Chen, X., Li, Y. and Wang, Z., (2022). A comparative study of early and deep learning-based feature fusion for multimodal interaction. *Journal of Intelligent Systems,* vol. 31**,** no. 2, pp. 189-205.

Chen, Y. F., Lu, Y. Y. and Zhou, X., (2023). Multi-object tracking algorithm based on cross-correlation attention and chain frame processing. *Computer Science***,** no. 1, pp. 231-237.

Deng, Y., Li, C. and Gu, Y., (2025a). Graph-based multimodal fusion for emotion recognition: A review. *Pattern Recognition Letters,* vol. 189, pp. 45-53.

Deng, Y., Li, C., Gu, Y., Zhang, H., Liu, L., Lin, H., Wang, S. and Mo, H., (2025b). Graph Convolution-Based Decoupling and Consistency-Driven Fusion for Multimodal Emotion Recognition. *Electronics,* vol. 14**,** no. 15, p. 3047.

Lee, H.-T., Shim, M., Liu, X., Cheon, H.-R., Kim, S.-G., Han, C.-H. and Hwang, H.-J., (2025). A review of hybrid EEG-based multimodal human–computer interfaces using deep learning: applications, advances, and challenges. *Biomedical Engineering Letters*, pp. 1-32.

Liu, J., Zhang, H. and Li, S., (2023). Sensor-computing collaborative optimization for smart home multimodal interaction. *IEEE Internet of Things Journal,* vol. 10**,** no. 5, pp. 4210-4223.

Liu, S., Shao, F., He, X., Xue, J., Zhang, H. and Liu, Q., (2025). DBSQFusion: a multimodal image fusion method based on dual-channel attention. *Complex & Intelligent Systems,* vol. 11**,** no. 10, p. 421.

Luo, Z. X., Tang, R. and Li, L., (2025). Audio-visual emotion recognition based on multi-scale KAN convolution and cross-modal attention. *Computer Technology and Development***,** no. 7, pp. 100-107.

Niu, H., Van Leeuwen, C., Hao, J., Wang, G. and Lachmann, T., (2022). Multimodal natural human–computer interfaces for computer-aided design: A review paper. *Applied sciences,* vol. 12**,** no. 13, p. 6510.

Peng, H., Shi, N. and Wang, G., (2023). Remote sensing traffic scene retrieval based on learning control algorithm for robot multimodal sensing information fusion and human-machine interaction and collaboration. *Frontiers in neurorobotics,* vol. 17, p. 1267231.

Qu, H. C. and Xu, B., (2025). Multimodal sentiment analysis based on adaptive graph learning weights. *Journal of Intelligent Systems***,** no. 2, pp. 516-528.

Schreiter, J., Heinrich, F., Hatscher, B., Schott, D. and Hansen, C., (2025). Multimodal human–computer interaction in interventional radiology and surgery: a systematic literature review. *International Journal of Computer Assisted Radiology and Surgery,* vol. 20**,** no. 4, pp. 807-816.

Sun, B., Jia, L. and Cui, Y., (2024a). Dynamic multimodal fusion: A survey. *Neurocomputing,* vol. 562, p. 126890.

Sun, B., Jia, L., Cui, Y., Wang, N. and Jiang, T., (2025). Conv-Enhanced Transformer and Robust Optimization Network for robust multimodal sentiment analysis. *Neurocomputing,* vol. 634, p. 129842.

Sun, B., Jiang, T., Jia, L. and Cui, Y. M., (2024b). Multimodal sentiment analysis based on cross-modal joint-encoding. *Computer Engineering and Applications,* vol. 60**,** no. 18, pp. 208-216.

Tao, J. H., Wu, Y. C., Yu, C., Weng, D. D., Li, G. J., Han, T., Wang, Y. T. and Liu, B., (2022). A survey on multi-modal human-computer interaction. *Journal of Image and Graphics,* vol. 27**,** no. 06, pp. 1956-1987.

Wang, L., Zhang, Y. and Zhao, J., (2024). Multimodal interaction in education: A case study on intelligent tutoring systems. *Educational Technology Research,* vol. 36**,** no. 1, pp. 56-2.

Wang, R., Xu, D., Cascone, L., Wang, Y., Chen, H., Zheng, J. and Zhu, X., (2025a). Raft: robust adversarial fusion transformer for multimodal sentiment analysis. *Array*, p. 100445.

Wang, Z. Y., Tian, D., Dong, Y., Qiao, N. and Shan, G. I., (2025b). Multimodal interaction: From human-computer collaboration to human-intelligence collaboration. *Frontiers of Data & Computing,* vol. 7**,** no. 3, pp. 81-93.

Yao, R., Wang, K., Guo, H. F., Hu, W. T. and Tian, X. R., (2025). Infrared and visible image fusion based on cross-modal feature interaction and multi-scale reconstruction. *Infrared and Laser Engineering,* vol. 54**,** no. 08, pp. 269-280.

Yu, X., Li, Z., Wu, J. and Liu, M., (2022). Multi-module Fusion Relevance Attention Network for Multi-label Text Classification. *Engineering Letters,* vol. 30**,** no. 4, p. 1237.

Yu, X. Y., Zhang, X., Xu, C. J. and Ou, L. L., (2025). Human-robot interaction method and system design by fusing human perception and multimodal gestures. *Chinese High Technology Letters,* vol. 35**,** no. 02, pp. 183-197.

Zhao, Q., Guo, B., Liu, Y. B., Sun, Z., Wang, H. and Chen, M. Q., (2025). Generation of enrich semantic video dialogue based on hierarchical visual attention. *Computer Science,* vol. 52**,** no. 01, pp. 315-322.

Zhu, C., Yi, B. and Luo, L., (2024). Base on contextual phrases with cross-correlation attention for aspect-level sentiment analysis. *Expert Systems with Applications,* vol. 241, p. 122683.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access