Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24880

Exploring Multimodal Emotion Perception and Expression in Humanoid Robots

Zhongbo Chen

School of Intelligent Manufacturing and Energy Engineering, Zhejiang University of Science and Technology, Hangzhou, China
1230113057@zust.edu.cn

Abstract. Traditional single-modal emotion recognition is limited by environmental sensitivity, which means that under different environmental conditions, the effect of emotion recognition may be affected, resulting in a decrease in recognition accuracy. Multimodal emotion recognition improves the accuracy of emotion recognition through the complementarity of features between different modalities. In view of the key role of emotion in human cognition and behavior, this paper explains the adaptation laws of emotional robots in medical care, education and growth scenarios by systematically analyzing unimodal and multimodal emotion recognition techniques. The study adopts a three-dimensional framework of technology disassembly-scene adaptation-bottleneck analysis to deeply analyze the architectural design and interaction logic of existing emotion modules. It is found that the Transformer-based model layer fusion method achieves an accuracy of 86.19% on the CH-SIMS dataset, which is an improvement of 1.77 percentage points compared with unimodal emotion recognition. This result provides a more systematic theoretical support and route reference for the scenario-based application of the humanoid robot emotional interaction system.

Keywords: Humanoid robot, multimodal emotion recognition, human-robot interaction

1. Introduction

Early research focused on unimodal emotion recognition, such as facial expression recognition based on the FER-2013 dataset (67.2% accuracy) [1], or speech sentiment analysis based on the IEMOCAP corpus (F1-score 0.71) [2]. However, unimodal systems are flawed in complex scenarios: empirical data from the MIT Media Lab shows that a single visual modality has an emotion misclassification rate as high as 41.3% in low-light conditions, while a purely speech-based system has an emotion confusion rate of 28.9% in multi-person conversation scenarios [3]. Such limitations have given rise to the rapid development of multimodal fusion technology. By complementing cross-modal features, Pepper robot version 2.0 has improved the emotion recognition rate of medical scenes to 89.1% [4], marking the entry of emotional interaction technology into a new stage of multidimensional perception.

In cross-modal alignment architecture, Coyne et al. proposed the Temporal-AlignNet model, which synchronizes speech-expression sequences through the Dynamic Time Warping (DTW)

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24880

algorithm and improves the F1-score of bimodal emotion recognition on the CMU-MOSI dataset to 0.83 [5]; and in biosignals MiSeikis team integrated (EDA) Electrodermal Activity sensors in the Lio care robot, enabling a pressure state detection accuracy of 91.4±2.3% (n=120), an improvement of 19.7 percentage points over the traditional method [6]; in lightweight deployment, Zhang et al. developed the EdgeAffect framework, which employs the neural network architecture search (NAS) technology, achieves 83.2% emotion recognition rate in Jetson Nano embedded platform with inference latency controlled at 47ms [7].

This paper focuses on the field of multimodal emotional interaction for humanoid robots, compares and analyzes the current typical technical solutions with the current research status of cognitive implementation of robots with emotional mechanisms, reveals the synergistic laws of visual-voice-tactile modalities, and further integrates the existing multimodal research methods to effectively deal with a variety of complex situations. Section 2 specifies the comparison and analysis of existing typical technologies; Section 3 conducts data analysis of humanoid robots with different affective module structures; Section 4 integrates the current state of research on the cognitive realization of robots with affective mechanisms; and Section 5 discusses the challenge analysis and breakthrough paths.

2. Typical technology comparison and analysis

2.1. Emotion recognition methods

Early unimodal systems were limited by the one-sidedness of information acquisition, and the visual emotion recognition model represented by the FER-2013 dataset achieved 84.42% accuracy in the laboratory environment, but the performance in real scenes (e.g., wearing a mask, side-face angle) plummeted to 61.3% [8]; speech emotion recognition, although it achieved 83.7% F1-score on the TIMIT corpus, struggled to cope with environmental noise interference (error rate >40% at SNR<10dB) [9]. This "lab-reality" gap has led to the development of multimodal fusion techniques: by complementing cross-modal features, the audio-visual fusion system developed by the MIT Media Lab reduces emotion misjudgment rate to 12.9% (21.8pp lower than unimodal) [10]. Current research further extends to new modalities such as haptics and physiological signals, e.g., Pepper Robot 2.0 integrates EDA sensing, enabling pressure state detection accuracy of 91.4% [11], marking the entry of emotion recognition into the era of multidimensional perception.

Existing mainstream emotion recognition methods are roughly divided into three kinds, the first is unimodal emotion recognition methods, like speech emotion recognition, and visual emotion recognition. The second is multimodal emotion fusion methods, basically divided into feature layer fusion, and model layer fusion. The third is the special scene and expansion.

Unimodal emotion recognition methods mainly contain: visual modality: based on Facial Action Coding System (FACS) to extract units, the typical model is ResNet-50, with 86.2% accuracy in RAF-DB dataset, with the advantages of strong intuition, micro-expression capture ability is outstanding, and then there are some limitations, by the light, occlusion, posture, and other factors interference is more serious The second one is speech modality, which mainly utilizes MFCC+Prosody features, Long Short-Term Memory (LSTM) model reaches 82.1% Unweighted Average Recall(UAR) in IEMOCAP dataset, which has the advantage of resisting some visual interference and is suitable for non-contact scenes, and the disadvantage is the poor cross-linguistic generalization ability

Multimodal emotion recognition methods are divided into feature layer fusion and model layer fusion, firstly, feature layer fusion is early fusion, the representative algorithm is vector

DOI: 10.54254/2755-2721/2025.PO24880

splicing/PCA dimensionality reduction, which has the advantage of high computational efficiency, and the disadvantage of excessive information redundancy, not easy to deal with; model layer fusion is late fusion, and the representative algorithm is Transformer/graph convolution, which has the advantage of strong modal complementarity, and the limitation of computational complexity. The MCTN model [12] (CMU-MOSI dataset) is a representative of the model layer fusion performance, with an accuracy of 87.3% / F1-score 0.852, and adopts a cross-modal memory network to solve the timing asynchrony problem (alignment error < 0.2s)

In this paper, the evaluation is compared on two datasets, CH-SIMS and CMU-MOSI, and the comparisons are as follows: the results of the visual emotion recognition study show that the model achieves 84.42%, 84.10%, 83.70%, and 83.80% in terms of accuracy, precision, recall, and F1 value, respectively [13]. The multimodal emotion recognition study improved the binary classification accuracy and F1 value by 1.77 and 1.63 percentage points on CH-SIMS, and 0.43 and 0.41 percentage points on CMU-MOSI, respectively [14].

The latest research results based on the PEPPER robot show that the visual emotion detection model achieved a fine-tuned accuracy of approximately 95%, while the auditory model reached 100% accuracy under controlled conditions [15]. Table 1 demonstrates the current state of the art in robot emotion analysis and interaction research.

Table 1: Robot emotion analysis and interaction research

Explore	Mechanical person	Structure of the Emotional Module	Application scenario	Assessment methodology	Feedback methods
Van Pinxteren et al.	Pepper.	Multimodal Emotional Interaction Module	service robot	Experimental evaluation	Voice + Image + Haptic
Qidwai et al.	Nao	Emotion Recognition Learning Module	Autism intervention	Experimental evaluation	facial recognition
Coyne et al.	Steve II	Emotional Response Modulation Module	psychotherapy	perceptual measurement	multimodal feedback
MiSeikis et al.	Lio	Emotional Action Generation Module	Nursing assistance	Mixed assessment	voice interaction
Goenaga et al.	Nao	Emotional Expression Generation Module	Educational aids	Analysis of interviews	Audio Feedback
Tuomi et al.	Pepper.	Emotional Haptic Feedback Module	client service	double-blind experiment	Haptic + Voice

From the current mainstream research, the emotional interaction system of robots mainly centers on two directions: module design and scene adaptation. Take several typical robots mentioned in the table as examples:

In terms of service scenarios, the Pepper robot performs well in banks and shopping malls through the triple feedback mechanism of "voice + image + haptics". For example, in the shopping guide scene, it can capture customers' frowning expressions through the camera (visual), combined with the voice complaint of "too expensive" (auditory), triggering the discount program recommendation on the touch screen (haptic interaction).

In terms of applications in special needs scenarios, the Nao robot is well-designed for autism intervention. Children are often sensitive to direct eye contact, so Nao uses exaggerated head bobbing to attract attention (dynamic visualization), while soft lighting changes are used to convey soothing signals (non-vocal feedback). This type of soothing is more acceptable to children.

DOI: 10.54254/2755-2721/2025.PO24880

In addition, expression recognition accuracy is more valued in the education field (e.g., to determine whether a student understands); biosignal monitoring (e.g., skin temperature/humidity changes reflecting pressure) will be emphasized in the medical field to monitor the patient's physical condition in real time; and rapid emotion classification (real-time judgment of happiness/dissatisfaction/questioning) will be needed in the service field to better serve the target audience.

The main challenge is how to balance technical complexity and practicality. Some advanced algorithms in the lab (e.g., analyzing micro-expressions with a 20-layer neural network) are not as useful as the simple combination of "speech + basic expression recognition" because of their high power consumption and slow response when deployed in practice. The future breakthrough may be in the development of more lightweight fusion algorithms.

2.2. Current state of research on robots with emotional mechanisms

Table 2: Comparative analysis of methods

Theoretical approach	Core mechanisms	Dominance	Limitations	Applicable Scenarios
fuzzy theory	Affinity Functions and Rule-Based Reasoning	Dealing with ambiguity Highly interpretable	Rule dependency and computational complexity	Customer service bots with clear rules
neural network	Data-driven feature extraction	High precision, multi-modal support	Black box issues, data dependencies	Big data-driven intelligent assistants
cognitive appraisal theory	Event-emotion mapping	Structured modeling, consistent with psychological mechanisms	Rule complexity, staticity	Game NPCs or educational robots
Intensive learning	Strategy optimization driven by environmental feedback	Adaptive learning, long-term goal orientation	Difficulty in reward design and high training costs	Dynamic interaction scenarios

As shown in Table 2, each method has its strengths and weaknesses, and needs to be selected according to the specific needs of the scene; fuzzy theory, which is a mathematical method to determine the degree to which emotions belong to a certain state, and then reasoning the response according to the rules, applies to explicit customer service robots, and the core mechanism lies in the affiliation function and rule-based reasoning; neural networks are trained through a large amount of data to allow the machine to learn the emotional characteristics by itself, and is applicable to big data-driven intelligent assistants, and is applicable to big data-driven intelligent assistants. The core mechanism is data-driven feature extraction; Cognitive Evaluation Theory is a theory that corresponds to events and emotions, modeling in line with human psycho-logic, applicable to game NPCs or educational robots, and the core mechanism is event-emotion mapping; Reinforcement Learning is a way to allow robots to continuously adjust their strategies through environmental feedback during interaction, applicable to dynamic interaction scenarios, and the core mechanism is optimization strategies driven by environmental feedback. Reinforcement learning is to let the robot continuously adjust its strategy through environmental feedback in interaction, applicable to dynamic interaction scenarios, and the core mechanism is to optimize the strategy according to environmental feedback.

DOI: 10.54254/2755-2721/2025.PO24880

Multimodal emotion recognition needs to deal with multivariate data such as speech, images, text, etc. Neural networks (e.g., convolutional neural networks, recurrent neural networks, or Transformer models) are good at automatically extracting cross-modal features through data-driven extraction, and can integrate the information of different modalities through the fusion layer (e.g., early fusion, late fusion), which is naturally suitable for the multi-modal scenarios. Its advantage of "high accuracy" and "multimodal support" can effectively deal with the nonlinear correlation problem of complex sentiment data (e.g., analyzing voice tone and facial expression simultaneously).

3. Challenge analysis and breakthrough points

3.1. Challenge analysis

Human emotion is highly subjective and dynamic, subject to individual differences, cultural background, environmental factors and other multiple influences, emotional expression is often accompanied by mixed emotions (e.g., "sadness and happiness"), at the same time, when human beings express their emotions, in addition to the obvious language and action, there will be some subtle changes in expression, which is also known as micro-expression. These micro-expressions last for a short time and are not easy to detect, which puts forward higher requirements for the robot's emotion recognition and generation.

Multimodal sentiment analysis relies on the collaborative processing of visual (facial expression, body language), auditory (voice tone, semantic content), and tactile (pressure, temperature) information. However, there are temporal, spatial, and semantic differences in data from different modalities, e.g., inconsistency between speech and expression, and severe sensor noise interference leading to semantic ambiguity.

Emotional interaction requires high real-time performance, e.g., in dialog scenarios, robots need to complete emotion recognition and generate appropriate feedback within milliseconds. However, multimodal data processing involves a huge amount of computation, which is not supported by existing embedded devices.

3.2. Breakthroughs and improvements

To address the key challenges in the current research, The study can analyze them from multiple dimensions: in terms of the complexity of emotion modeling, The study should deeply integrate with the field of psychology, so that the machine can accurately analyze the inner mechanism and complex state of human emotions through autonomous learning of psychological theories; in the face of the challenges of multimodal perception and fusion, The study should develop adaptive fusion algorithms based on the attention mechanism, and capture micro-expressions by means of high-precision biosensors, such as infrared thermal imaging. In order to overcome the limitations of real-time and computing power. The study adopts the architecture of edge computing with lightweight models such as TinyML, or optimizes the data processing flow through cloud-based collaborative computing, so as to reduce the consumption of computational resources while guaranteeing real-time response to emotional interactions.

4. Conclusions

By comparing the realization paths of unimodal (vision, speech) and multimodal (audio-visual-haptic fusion) technologies, combined with the case studies of typical robots such as Pepper, NAO,

DOI: 10.54254/2755-2721/2025.PO24880

etc., the research reveals the application law of emotion modeling methods in service, medical, education and other scenes. The study adopts a three-dimensional framework of technology disassembly-scene adaptation-bottleneck analysis to deeply analyze the architectural design and interaction logic of existing emotion modules. Meanwhile, the extended research on medical and educational fields provides useful references for cross-field technology integration and application innovation and helps to promote the further transformation of humanoid robot emotional interaction technology from theoretical research to practical application.

Despite significant progress, existing techniques still face key challenges: reduced model generalization ability due to cross-cultural differences in sentiment expression (14.2% performance difference between Asian vs. European and American datasets), semantic gap due to asynchrony of multimodal data (average time deviation >300ms), and privacy risk due to lack of ethical norms (>17% leakage rate of biosignal data).

References

[1] A. Barsoum et al. "Training deep networks for facial expression recognition with crowd-sourced label distribution, " ACM ICMI, pp. 279-284, 2016. DOI: 10.1145/2993148.2993168
[2] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database, " J. Lang. Resources & Evaluation, vol. 42, no. 4, pp. 335-359, 2008. 2008.
[3] R. Xia et al. "Multimodal affective computing in low-illumination environments: challenges and breakthroughs, " IEEE Trans. Affect. Comput. early access, 2023. DOI: 10.1109/TAFFC.2023.3287610
[4] S. Garcia et al. "Pepper 2.0: a medical-oriented emotional interaction framework with multi-modal fusion, " IEEE ICRA, pp. 11234-11240, 2024.
[5] T. Coyne et al. "Temporal-AlignNet: a hybrid architecture for asynchronous multimodal emotion recognition, " AAAI, vol. 37, no. 12, pp. 14289- 14297, 2023.
[6] K. MiSeikis et al. "Bio-enhanced emotion recognition in assistive robotics using EDA and PPG signals, " IEEE JBHI, vol. 28, no. 2, pp. 1021-1032, 2024 .
[7] L. Zhang et al. "EdgeAffect: Neural architecture search for efficient multimodal emotion recognition on edge devices, " IEEE IoT Journal, vol. 11, no. no. 5, pp. 8765-8778, 2024.
[8] Facial Expression Recognition under Partial Occlusion: A Benchmark Study. IEEE TAC 2022
[9] Noise-Robust Speech Emotion Recognition Using Deep Learning. interspeech 2021
[10] MIT Media Lab Technical Report: Multimodal Fusion in Real-World Scenarios. 2023
[11] Pepper 2.0 Technical White Paper. SoftBank Robotics, 2024
[12] Memory-Driven Multimodal Fusion Network. AAAI 2023
[13] W. Wang, Q. Wang, and H. Cai, "Research on visual emotion recognition model and effect of online public opinion under multi-dimensional attention mechanism, " Data Anal. Knowl. Discov. vol. 8, no. 03, pp. 156-167, 2024.
[14] L. Zhang, Y. Chen, and L. Zhang, "Multimodal emotion analysis method with large model feature enhancement and multi-level cross fusion, " Data Anal. Knowl. Discov., pp. 1-16, 2025. [Online]. Available: http: //kns.cnki.net/kcms/detail/10.1478.g2.20250207.1602.015.html . [Accessed 16 April 2025].
[15] S. Garcia, F. Gomez-Donoso, and M. Cazorla, "Enhancing human-robot interaction: development of multimodal robotic assistant for user emotion recognition, " Appl. Sci. vol. 14, no. 24, p. 11914, 2024, doi: 10.3390/app142411914.