Proceedings of the 4th International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2026.34246

# *Emotion Perception Enhancement for Embodied Robots Using AI Interaction Technology*

**Zhaoxin Li**

*International Education School, Hebei University of Economics and Business, Shijiazhuang, China*
*lizhaoxin321@outlook.com*

***Abstract.*** Embodied intelligent robots have deeply penetrated social service fields such as medical rehabilitation, elderly care, and children's education. Its interactive naturalness and social acceptance highly depend on the accurate perception of human emotions. However, the sensor noise, real-time constraints, cultural differences, and individual specificity in the dynamic interactive environment make it difficult for the traditional non-embodied emotion computing model to migrate directly. This paper systematically summarizes how AI-based multimodal interaction technology can enhance the emotional perception of embodied robots, and analyzes it from three levels: multimodal perception technology, situational understanding model, embodied integration, and optimization. The research shows that technology evolution presents three trends. Multimodal fusion significantly improves perceptual robustness through signal complementarity; Situational modeling extends instantaneous perception to dynamic tracking; Knowledge guidance injects interpretability and common-sense constraints into data-driven models. Current technology is facing bottlenecks such as a lack of cross-cultural generalization, a lack of robustness of real scenes, and prominent privacy ethical risks. In the future, it needs to focus on the integration of brain-computer interface, the standardization of emotional operating systems, and the construction of ethical standards.

***Keywords:*** embodied intelligence, emotional perception, multimodal fusion, situational modeling, human-computer interaction

## 1. Introduction

Embodied AI robots are gradually moving from the industrial scene to human society, playing an increasingly important role in the fields of medical rehabilitation, children's education, elderly care, and public services. Unlike traditional non-embodied artificial intelligence systems (such as virtual assistants), embodied intelligent robots share space, touch contact, and dynamic collaboration with humans through physical entities. The naturalness, reliability, and social acceptance of their interaction are highly dependent on the ability to accurately interpret human emotional signals. When the robot cannot perceive the user's anxiety, resistance, or pleasure, it may lead to stiff interaction and a decline in user experience, or it may lead to security risks and trust collapse. Therefore, how to give the embodied intelligent robot real-time and context-sensitive emotional

© 2026 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: 10.54254/2755-2721/2026.34246

perception ability has become the core research direction in the field of human-computer interaction and emotional computing.

In recent years, many high-quality reviews have emerged in the field of affective computing and human-computer interaction. The application scenarios of deep learning technology of multimodal emotion recognition focus on non-embodied screen interaction or offline data sets; In the overall architecture design of a social robot, the emotion perception module is only briefly mentioned as a section. Few reviews take embodiment itself as the core analysis variable to systematically examine the unique challenges and change opportunities brought by physical entities, dynamic interaction, and environmental coupling to emotional perception technology.

Based on the above research gap, this paper aims to systematically sort out how the current interaction technology based on artificial intelligence can improve the emotional perception ability of embodied robots, and mainly analyzes its technical principle, applicable scenarios, and performance limitations.

## 2. AI technology progress of multimodal emotion perception

### 2.1. Visual modality

The visual channel is the most intuitive information source for robots to perceive human emotions. In recent years, deep learning technology has significantly improved the recognition accuracy of facial expressions, limb posture, and staring behavior.

In terms of facial expression recognition, early work mainly relied on 2D RGB images and a convolutional neural network (CNN), but the performance was seriously degraded due to head pose changes, local occlusion, and uneven illumination in real interactive scenes. 3D-CNN and graph convolution network (GCN) are introduced to improve the robustness of micro expression recognition by modeling the temporal and spatial sequence of facial key points. For example, the UGotMe system jointly proposed by Tsinghua University and Tencent proposes an active face extraction strategy for multi-person conversation scenes, which eliminates the interference of non-active individuals in the scene by dynamically screening the facial regions of active speakers, and improves the recognition accuracy by 21% under the condition of multi-target occlusion [1]. In addition, the Google DeepMind team applied the attention mechanism to the weighting of key areas, so that the model can still focus on the UN occluded face areas (such as eyes and mouth) when partially occluded, and maintain the recognition performance [2].

As a complementary channel, body language analysis has attracted increasing attention. Openpose, mediapipe, and other open-source tools have realized real-time human skeleton extraction. The research team of Munich University of Technology has further constructed the posture emotion correlation model to infer the emotional state from clues such as gait, gesture, and body orientation. The team found that when individuals are in an anxious state, they tend to contract their limbs and increase the frequency of gestures; The relaxed state is accompanied by an open posture and soothing limb movement [3]. In a noisy environment or long-distance interaction, when facial details cannot be clearly captured, limb posture becomes a key alternative information source.

### 2.2. Auditory modality

Speech carries rich emotional information, including paralinguistic features such as tone, rhythm, and intensity, as well as non-semantic sounds such as laughter, sighs, and breath.

DOI: 10.54254/2755-2721/2026.34246

Speech emotion recognition has experienced a paradigm shift from manual features (MFCC, prosodic features) to end-to-end depth models. The emergence of self-supervised pre-training models such as wav2vec2 and Hubert, proposed by the meta AI team, makes it possible to directly learn emotional representations from the original speech waveform, significantly improving the generalization ability across corpora [4]. The Human-Computer Interaction Laboratory of Tsinghua University found that personalized fine-tuning combined with voiceprint analysis can improve the recognition accuracy of specific users. In the real robot interaction scene, speech emotion recognition faces the challenges of background noise and multiple people speaking at the same time, which needs to be preprocessed by combining sound source separation and noise reduction technology.

Paralinguistic feature analysis expands the boundary of traditional speech emotion recognition. The intensity and duration of laughter can reflect the degree of pleasure, while sighing indicates fatigue or depression. The research team of Nanyang Technological University has developed a special sound event detection module, which combines these non-semantic signals with voice content to successfully identify the acoustic signs of early depression tendency in the elderly care scene [5]. For example, individuals with depressive tendencies tend to show slower speech speed, increased pauses, and flat intonation.

Multi-speaker separation is the core challenge of real conversation scenes. Ugotme system uses active sound source localization and beamforming technology to separate the target speaker's voice under the condition that multiple people speak at the same time, providing clean audio input for subsequent emotional analysis. This technology can still maintain stable sound source tracking performance in the process of robot dynamic motion [1].

### 2.3. Tactile and other modes

Embodied interaction gives the robot a unique tactile perception channel, which is completely missing in non-embodied systems.

Tactile pressure mapping technology through a flexible piezoresistive sensor array, real-time capture of handshake strength, hug posture, shoulder position, and other physical contact signals. The Intelligent Robot Laboratory of Osaka University in Japan found that the duration and change pattern of grip strength can effectively distinguish between friendly greetings and nervous resistance. A short, strong handshake is often accompanied by anxiety, while a mild and lasting contact conveys trust [6]. The hugvie project of the laboratory integrates tactile sensors into the pillow robot, and infers the emotional state and intimacy of users through the distribution of hugging pressure [7].

As a new research direction, temperature sensing monitors the changes in user skin temperature through infrared thermal imaging or a contact temperature sensor. The level of emotional arousal is closely related to peripheral blood flow, and the temperature fluctuation of specific areas below the state of anxiety or excitement (such as the nose tip and cheek) can be used as an implicit emotional indicator. In the elderly care scene, abnormal temperature changes may indicate physical discomfort or emotional fluctuations.

Physiological signal fusion will push interaction to a deeper level. Near infrared spectral imaging (fNIRS) can non invasively monitor the blood oxygen level of the prefrontal cortex, reflecting cognitive load and emotional regulation; Heart rate variability (HRV) and skin electrical response (EDA) collected by wearable devices provide indicators of autonomic nervous system activity. The research team of Munich University of Technology proposed to integrate the low-power physiological sensor into the robot arm contact surface, and synchronously collect the user's

DOI: 10.54254/2755-2721/2026.34246

physiological signals during the handshake process to realize the implicit perception of emotion [8]. This non-invasive sensing method avoids the burden of additional wearable devices and improves the nature of interaction.

## 3. How multimodal technology can enhance emotional perception

### 3.1. Collaborative enhancement of multimodal signals

Single-mode often has incomplete information or noise interference. Multimodal fusion can significantly improve the perceptual robustness through signal complementarity and cross-validation. Visual modality provides high-precision facial expression recognition in good illumination, but its performance decreases in occlusion or long-distance conditions. At this time, auditory modality (speech emotion, paralinguistic features) can be used as an effective supplement to maintain perceptual continuity. Tactile modality provides unique emotional clues in physical interaction -- handshake strength, hug posture -- this information can not be obtained through vision or hearing, and constitutes the "exclusive channel" for emotional perception in embodied scenes.

The fusion architecture evolves from early/late fusion to intermediate fusion based on a transformer, and dynamically aligns heterogeneous features through cross-modal attention. The ECRP framework jointly proposed by Beijing University of Technology and Huawei Noah Ark Laboratory uses the confidence weighting strategy to reduce the visual weight and improve the auditory weight when the vision is blocked; When the background noise is too large, the opposite is true. This confidence weighting strategy can make the F1 score of emotion classification reach 92.1% in the real interaction scene, which is 17.3% higher than that of the unimodal visual model [9]. Cross-modal completion further strengthens the synergy effect. When visual occlusion occurs, the generated model can predict the facial expression features according to speech and realize the self-repair of emotion inference.

### 3.2. From instantaneous perception to dynamic tracking

Multimodal technology expands emotion perception from "single point identification" to "dynamic tracking", and understands the evolution process of emotion by introducing time dimension and context information.

Temporal modeling captures the dynamics of emotions. The sequence model encodes the previous rounds of conversations into implicit states as the context of the current emotion inference, which can improve the classification accuracy by 8%-12% [10]. The emotional state transition graph models the temporal constraints between emotional categories as a graph structure, and the consistency of constraint inference - "pleasure" will not directly jump to "Rage", and the predicted consistency can be improved by 28% [11].

Context fusion reduces scene ambiguity. The threatening meaning of a frowning expression when holding a knife in the kitchen is much higher than that in the living room scene. The multimodal system combines object recognition, human posture, and facial expression through a scene graph neural network, which can effectively reduce the misjudgment caused by the environment.

Personalized modeling solves the problem of expression differences. The meta-learning framework (such as MAML) can quickly adapt to new users with a small number of samples (1-5 interactions), and the F1 score is 14.6% higher than the transfer learning baseline [12]. The long-term memory mechanism maintains users' emotional files (emotional baseline, trigger events, typical

DOI: 10.54254/2755-2721/2026.34246

reactions), continuously updates users' portraits within a few weeks of the interaction cycle, and protects privacy under the federal learning framework [8].

### 3.3. From data-driven to interpretable reasoning

A pure data-driven model lacks explicability and is easy to violate common sense. Knowledge-guided multimodal technology combines symbolic knowledge with a neural network to achieve deeper emotional understanding.

The formalization of emotional psychology theory transforms classical theories (such as the OCC model) into symbolic rules. OCC model defines emotion as a function of the relationship between event consequences and individual goals - when events promote the achievement of goals, they produce pleasure, and when they hinder, they produce depression [13]. These rules, as logical constraints, guide the neural network to focus on key causal variables and provide interpretable reasoning paths.

Knowledge map embedding encodes emotional common sense (such as "loss of loved ones leads to sadness" and "reward leads to joy") into vectors, which are jointly learned with visual and speech features. In the recognition of long tail emotions (such as guilt, pride, and contempt) with sparse samples, knowledge guidance significantly improves the recognition performance. Cross-cultural modeling introduces cultural dimension theory (individualism and collectivism) as knowledge constraints, and guides the model to give differentiated weights to the emotional cues of different cultural groups.

Interpretability enhancement is the key to human-computer trust. The symbolic reasoning module built by the Emotional Computing Laboratory of Peking University, when the model predicts that the user is in the "anxiety" state, the symbolic reasoning module can trace back to the key evidence in the input characteristics (such as voice trembling, frequent blinking, and clenching hands) to provide users or developers with understandable explanations [14]. This ability of "knowing what is and why" makes multimodal technology move from simple pattern recognition to real emotional understanding.

## 4. Embodied integration and system optimization challenges

Deploying the multimodal emotion perception model to the physical robot platform faces the engineering challenges that non-embodied systems do not need to deal with.

The real-time implementation requires the model to complete the perception-reasoning closed-loop within a hundred milliseconds. Model lightweight is the key path. Lightweight backbone networks such as mobilenet and efficientnet can reduce the model volume to 1/10 while maintaining more than 80% performance [15]. TensorRT quantization and kernel fusion technology will further improve the reasoning speed by 2-3 times. In the edge cloud collaborative architecture, the basic emotion recognition is completed in the robot ontology, the complex emotion analysis is uploaded to the cloud, and the end-to-end response delay can be controlled within 34.8 milliseconds [9].

Physical constraint response includes sensor noise suppression, occlusion processing, and energy consumption optimization. The visual jitter caused by robot motion needs IMU to assist image stabilization; The baseline drift of tactile sensor needs regular automatic calibration; Active face extraction strategy is required to maintain identity Association for visual occlusion in multi person conversation scenes [1]; Neuromorphological sensors such as event cameras can reduce power consumption by two orders of magnitude, making it possible for all-weather operation [16].

DOI: 10.54254/2755-2721/2026.34246

Perception-action closed loop realizes emotion-driven behavior adaptation. The reinforcement learning framework encodes the real-time emotion recognition results and task context into a state space, and learns to select the optimal response strategy under different emotional states [9]. Reduce task complexity and provide encouragement when users are detected as tired; Slow down and keep a safe distance when users are anxious. Safety constraints must be embedded in the closed loop - when expressing comfort, ensure that the contact strength is within the safe range, and avoid physical injury caused by emotional response [17].

## 5. Discussion

### 5.1. Critical analysis of technological limitations

Mainstream facial expression datasets such as AffectNet and RAF-DB mainly use Western samples. According to Fan et al., in the AffectNet dataset, Europeans make up approximately 67.3%, Asians only 6.1%, and Africans only 7.4%. This sample imbalance leads to a significant decrease in cross-racial emotion recognition performance. Li et al. found through cross-race validation of FaceReader and DeepFace systems that the accuracy of negative emotion recognition on Eastern face datasets was significantly lower than that on Western faces [18]. In addition, there are significant differences between the deliberate expressions collected in the laboratory and the natural expressions of real interactions, and the sparse long tail emotion samples further constrain the model's generalization ability.

In the real environment, changes in lighting lead to a decrease in facial recognition accuracy, background noise increases the error rate of speech emotion recognition, and the sensor pollution introduced by the robot's own motion lacks an effective response. The performance generally deteriorates when transferring across scenes.

The sampling rate and transmission delay of multimodal signals vary, and existing alignment methods are still unstable under severe motion or network fluctuations. Kimura et al. pointed out that different sensors (vision, hearing, touch, etc.) usually operate at different sampling rates, resulting in time dislocation; Network latency, hardware failures, or processing bottlenecks further introduce time offset, and alignment errors may lead to poorer fusion performance than single modality [19].

Emotional expression has cultural specificity, with East Asian cultures tending to suppress the outward manifestation of negative emotions while Western cultures are more open. The "Handbook of Human Emotional Neuroscience" published by the University of Cambridge specifically explores the neural evidence of cultural variation in emotions, pointing out that there are systematic differences between East Asian and Western cultures in emotional experiences, expression norms, and neural responses. The current model lacks perception of this and is prone to misjudging implicit nonverbal cues such as silence and gaze in high-context cultures.

### 5.2. Ethics and social impact

Physiological signals involve the deepest psychological states, and traditional anonymization is difficult to defend against re-identification attacks. Federated learning and differential privacy have a fundamental conflict between protecting data and maintaining accuracy, and current regulations lack implementation details for robot scenarios. Business scenarios can utilize emotional weaknesses for targeted marketing, while political propaganda can manipulate public opinion. When robots perceive emotions to cater to their preferences, users' autonomy is eroded, which may

DOI: 10.54254/2755-2721/2026.34246

constitute spiritual exploitation. Excessive dependence on robots among the elderly may lead to increased social isolation, long-term interaction among children may affect the development of real interpersonal cognition, and the impact of human-machine emotional bonds on the quality of emotional connections among individuals needs to be tracked over the long term.

## 6. Conclusion

This review systematically outlines the evolution path of AI-based multimodal interaction technology to enhance embodied robots' emotional perception. At the perceptual level, multimodal fusion significantly improves perceptual robustness through complementary and cross-validation of visual, auditory, tactile, and physiological signals. Transformer-based cross-modal alignment methods have achieved excellent performance in sentiment classification, and dynamic confidence weighted fusion and cross-modal complementation techniques further enhance the synergistic effect. At the level of understanding, situational modeling extends perception from single-point recognition to dynamic tracking. Temporal models integrate dialogue history, emotional state transition graph constraints to infer coherence, and environmental context fusion to reduce scene ambiguity. Meta-learning frameworks and long-term memory mechanisms effectively solve the challenges of cold start and personalized adaptation. At the inference level, knowledge guidance injects interpretability and common sense constraints into data-driven models by formalizing OCC sentiment theory, embedding sentiment knowledge graphs, and introducing cultural dimension theory, significantly improving the performance of long tail sentiment recognition. At the deployment level, model lightweighting, edge cloud collaborative architecture, IMU assisted image stabilization, event cameras, and other neural morphological sensors, as well as perception-action closed loops embedded with security constraints, together constitute a systematic solution from algorithm to physical platform implementation.

In response to the current challenges faced by technology, such as insufficient cross-cultural generalization ability, lack of robustness in real-world scenarios, challenges in cross modal asynchrony, and privacy ethical risks, it is necessary to construct a multimodal dataset covering multiple cultures in the future, develop domain generalization and non aligned fusion technologies, explore deep integration solutions between federated learning and differential privacy, and promote the establishment of ethical standards for embodied emotion computing. Overall, the current technology is in a critical transition period from laboratory prototypes to real-world deployment. Although significant progress has been made in multimodal fusion, situational modeling, and knowledge guidance, bottlenecks such as data bias, environmental generalization, and cultural adaptability have not been fundamentally resolved. Looking ahead to the future, only by synergistically promoting algorithm innovation, hardware breakthroughs, and ethical standards can it truly realize the vision of human-computer interaction from "emotional perception" to "empathetic understanding".

## References

[1] Zhang, Y., et al. (2023). UGotMe: An embodied system for active speaker detection and emotion recognition in dynamic environments. IEEE Transactions on Robotics.

[2] Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective body expression perception and recognition: A survey. IEEE Transactions on Affective Computing.

[3] Ruhland, K., et al. (2015). A review of eye gaze in virtual agents, social robotics, and HCI. ACM Transactions on Interactive Intelligent Systems.

DOI: 10.54254/2755-2721/2026.34246

[4] Baevski, A., et al. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (NeurIPS).
[5] Cummins, N., et al. (2015). A review of depression and suicide risk assessment using speech analysis. Speech Communication.
[6] Fukuda, H., et al. (2012). Huggable communication medium encourages people to touch and be touched. Journal of Robotics and Mechatronics.
[7] Goulart, C., et al. (2019). Emotion analysis in children through facial thermal imaging. IEEE Access.
[8] Tsai, Y. H. H., et al. (2019). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).
[9] Zhang, C., et al. (2022). Cross-modal generation for missing modality compensation in emotion recognition. IEEE Transactions on Affective Computing.
[10] Hu, J., et al. (2023). Graph-based emotion evolution tracking in long-term human-robot interaction. IEEE Transactions on Affective Computing.
[11] Yang, S., et al. (2021). Context-aware emotion recognition in the wild using scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[12] Jackson, J. C., et al. (2019). Emotion semantics show both cultural variation and universal structure. Science.
[13] Ortony, A., et al. (1988). The cognitive structure of emotions. Cambridge University Press.
[14] Liu, B. (2022). Explainable emotion recognition via neural-symbolic integration. ACM Transactions on Interactive Intelligent Systems.
[15] Howard, A. G., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications.
[16] Gallego, G., et al. (2022). Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[17] Wang, H., et al. (2024). ELA: An emotion-language-action large model for embodied affective interaction.
[18] Li, Y. T., Yeh, S. L., & Huang, T. R. (2023). The cross-race effect in automatic facial expression recognition violates measurement invariance. Frontiers in Psychology, 14, 1201145.
[19] Kimura, T., Li, X., Hanna, O., et al. (2025). InfoMAE: Pair-efficient cross-modal alignment for multimodal time-series sensing signals. In Proceedings of the ACM Web Conference 2025 (WWW‘25).