Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

Research on Three Key Issues of Multimodal Emotion Recognition Technology in Service Robotics Context

Junze Lyu *

Department of Electronic and Information Engineering, SHENZHEN UNIVERSITY, Shenzhen, China

* Corresponding Author Email: 2024280004@mails.szu.edu.cn

Abstract. With the rapid development of artificial intelligence technology, service robots are increasingly integrating into daily life scenarios. Achieving efficient and natural human-robot interaction has become a key challenge. The ability to recognize emotions, as a core factor in enhancing the interaction experience, has attracted extensive attention from both the academic and industrial communities. This paper focuses on the emotion recognition problem of service robots, analyzes three key problems that need to be solved by the application of multimodal emotion recognition technology in service robots, lists and analyzes several advanced machine learning algorithms that can help solve the problem, evaluates and analyzes them respectively, and summarizes their common limitations as well as the possible improvement methods in the future. This research not only provides crucial theoretical support and algorithmic paths for the emotional computing of service robots, but also helps to build a more reliable human-robot collaboration relationship. It holds significant application value for promoting the real implementation of service robots in real-life scenarios.

Keywords: Service robotics; Multimodal emotional recognition; Zero-sample learning; Sensor-less emotion recognition technology; Macrolanguage model.

1. Introduction

In recent years, the development of machine learning, deep learning, and neural networks has provided the algorithmic basis for the development and improvement of the IQ and EQ of robots. With the development of living standards as well as the level of science and technology, people's requirements for service robots are getting higher and higher, not only to make them deal with chores, but also to provide a good emotional interaction experience. Accurately recognizing user emotions is the crucial first step. Emotion recognition technology is gaining momentum, and its technical approach can be divided into two categories: unimodal emotion recognition and multimodal emotion recognition. Traditional unimodal emotion recognition relies on a single data source, making it prone to interference in real-world settings. In contrast, multimodal approaches outperform unimodal ones by integrating complementary data sources, enabling more accurate and robust recognition.

The application of multimodal emotion recognition technology in service robots plays a significant role in the transformation for robotics from cold machines to empathetic human-like entities. In the context of service robots, however, the application of multimodal emotion recognition technology raises many new practical issues. For example, there is a need to improve the generalization ability of emotion recognition for service robots and reduce its dependence on large datasets. This would reduce costs while enhancing practical working ability. Additionally, it’s necessary to improve the accuracy and robustness of emotion recognition while providing users with a better interactive experience. Finally, protecting user information privacy while collecting sufficiently large, personality-rich data for model training and computation is also a critical issue. To solve these problems, it is inevitable to consider additional factors and make appropriate improvements to the current abstract modeling algorithms.

This paper explores three new, practical issues that arise from applying multimodal emotion recognition technology to service robots. It introduces several advanced model algorithms that are

currently helpful to solve these issues, discusses their common limitations, and explores potential future improvements. The goal of this study is to help researchers quickly understand the current state of development in this field and accelerate the widespread implementation of service-enabled embodied intelligence.

2. Zero-sample Learning that Improves the Generalization of Recognition

The descriptions of emotions based on dimensional emotion models can portray emotions and their changes more precisely than the descriptions based on discrete emotion models. So, the former description could improve the accuracy of machine emotion recognition. Traditional emotion recognition research faces two major challenges in achieving fine-grained emotion recognition. First, traditional multimodal emotion recognition algorithms map emotion labels to numbers (one-hot vectors), ignoring the emotion word embedding space constructed by rich and complex emotion lexicons. This leads to relatively fixed emotion classifications and could result in relatively fixed emotion categorizations. In reality, individuals often have new emotional experiences that could enrich their emotional lexicon and change the existing categorization system. In other words, the lack of an appropriate emotional structure in traditional emotion recognition research limits the accuracy and robustness of machine emotion understanding. Additionally, relying solely on large datasets for sentiment analysis poses challenges in collecting rare or novel sentiment data, and the cost of collecting and annotating substantial stimulus sentiment data increases as the scope of personal sentiment annotations grows. Second, zero-sample learning clearly has an advantage when dealing with unseen sentiment labels. However, traditional zero-sample learning efforts are limited by the ability to effectively capture the inherent discriminative properties of the labels. The data dimensions that can be fused are also extremely limited - only visual features and text embeddings can be fused. The lack of modal coverage inevitably reduces the accuracy of the output.

Bhati et al. proposed the first generalized zero-sample sentiment classifier (GZS-ConvNet) [1]. GZS-ConvNet utilizes high-quality convolutional kernel weights to extract sentiment features and then classifies sentiment through a fully connected layer and Softmax. Bhati et al. designed a two-stage training strategy. In the first phase, Reextracted is frozen, and only the classifier parameters are trained. This significantly reduces cross-dataset accuracy. In the second phase, the entire network is unfrozen, and all parameters, including the convolutional kernel weights, are fine-tuned. This results in a significant improvement in cross-dataset generalization ability, which plays a fundamental role in improving the ability to recognize emotions.

Fan Qi et al. combined professional psychological knowledge with a priori emotional knowledge to improve emotional structure. They constructed a dynamic emotional graph space by strengthening semantic connections between contextually related concepts using a simple spectrogram convolutional network. This network can adaptively analyze connections between new and existing emotional words, greatly improving the model's generalization ability [2]. Additionally, Fan Qi et al.'s study converts a zero-sample multimodal emotion recognition task into a transformer-based cross-modal learning problem. This conversion is achieved through an adversarial modal de-entanglement module, a cross-modal alignment module, and an emotion-guided group decoding scheme. This scheme utilizes a hybrid multimodal synergistic attentional mechanism that retains diverse information and the ability to discriminate against other data points [2]. And this model addresses the migration of machine-recognized emotions from the categories observed during training to the novel and unseen categories observed during testing. This addresses the generalized zero-sample learning problem in emotion recognition. It’s a fact that Fan Qi et al. have made a unique innovation in zero-sample learning. However, this algorithm is currently only tested in a laboratory environment, so its performance in real-world service robot scenarios is unknown. With updates and iterations to their algorithm and the integration of more dimensional information in the future, it is believed that service robots' recognition generalization ability will improve.

3. Sensor-less Emotion Recognition Technology

Human emotional signals can be categorized into two types. One type is based on extrinsic physical signals. And the other type is based on intrinsic physiological signals. One significant feature of intrinsic physiological signals is their authenticity, which makes it difficult to hide or intentionally alter the real physiological signals of a recognized object. This greatly enhances the accuracy and robustness of the algorithm. Commonly utilized physiological signals include electroencephalography (EEG), galvanic skin response (GSR), electrocardiography (ECG), and eye tracking (ET) [3] [4].

3.1. Contact Emotion Recognition

Zhu, Enguo, et al. integrated physiological signals, behavioral data, and stimulus source signals to enhance the robustness of machine emotion recognition [5]. They proposed a new emotion recognition model, the EMFNN framework, which is based on the time-series processing of EEG data and the direct processing of event-related potential data through a CNN module. First, to obtain the EEG feature set, the study extracted EEG features using filtering and differential entropy methods. Then, they utilized the extracted EEG features using feature-level fusion. A convolutional neural network and long-short-term memory network were applied to process the EEG feature set. And the EEG feature data output was obtained through decision-level fusion. Finally, decision-level fusion was used to recognize emotions by processing the fused data through a deep neural network and outputting the results [5].

Zhuozheng et al. used the complementary nature of EEG and ECG signals to extract time-domain, frequency-domain, and nonlinear features. They optimized the performance of sentiment analysis by filtering the top nine discriminative features through the random forest approach [6].

This fusion algorithm excels in emotion recognition accuracy and robustness, but its shortcomings are also obvious. First, stimulus data is difficult to obtain and expensive to collect, easily resulting in a limited dataset. Second, this technology belongs to Contact-Based Emotion Recognition, which is not suitable for real-life scenarios involving service robots. The future of emotion recognition technology should focus on convenience, practicality and robustness, like the contactless emotion recognition technology.

References are cited in the text just by square brackets [1]. (If square brackets are not available, slashes may be used instead, e.g. /2/.) Two or more references at a time may be put in one set of brackets [3, 4]. The references are to be numbered in the order in which they are cited in the text and are to be listed at the end of the contribution under a heading References, see our example below.

Table 1. Examples of Senseless Emotion Recognition Techniques [7-11]

Emotional recognition algorithms	Data dimensions	Emotional recognition algorithms	Data dimensions
Graphic multimodal sentiment analysis	Images and text	Graphic multimodal sentiment analysis	Images and text
Art Multimodal sentiment analysis	The expression of emotion within an image and its beauty	Art Multimodal sentiment analysis	The expression of emotion within an image and its beauty
Face multimodal sentiment analysis	Facial expressions and micro-expressions of the human face	Face multimodal sentiment analysis	Facial expressions and micro-expressions of the human face

In reality, non-sensory sensors can collect a lot of valuable information, so contactless emotion recognition technology has a variety of data sources (see Table 1).

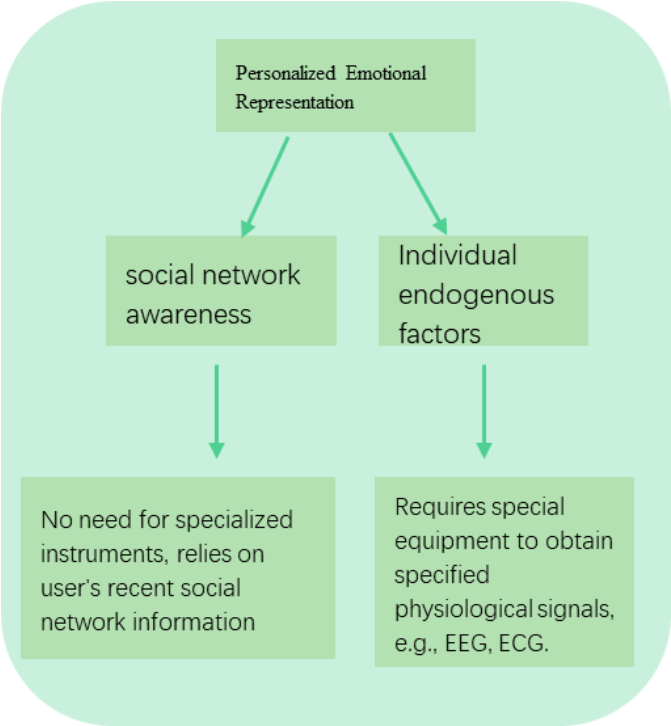

Figure 1. Personalized Emotional Representation [11]

As shown in Figure 1, the emotion recognition data sources are a class of data that can be collected without contact. This includes information about an individual's social network and personality traits. Due to the specific structure of these data sources, they are not typically included in multimodal data, such as speech and language.

3.2. Contactless Emotion Recognition

Among the recent innovations, Yuqing et al. proposed the MESCA model to compute the semantic consistency of emotions among text, audio, and video modalities. This model solves the problem of excessive computational overhead caused by the multi-branch training structure through structural re-parameterization. It also mitigates the redundancy problem in traditional pairwise interaction methods and significantly improves computational efficiency, merged stimulus source data with user physiological signals and proposed a new multimodal emotion recognition fusion algorithm, the Emotion Multimodal Fusion Neural Network (E-MFNN) [12][13]. They had experimentally verified its effectiveness. SUN Ying et al. observed that functional paralanguage contains a significant amount of information about emotions conveyed through speech. Thus, they adopted an integrated learning approach using the functional paralanguage proportion coefficient (FPPC) attention mechanism and adaptive entropy weight decision fusion to improve the system's overall emotion recognition rate [14]. These studies offer new insights into non-contact emotion recognition.

4. Sentiment Recognition in a Large Language Model

Understanding emotions is the first step, and it is expected that service robots will also have certain output capabilities. Large language models like ChatGPT and GPT-4 have demonstrated remarkable zero/few-shot learning, in-context learning (ICL), chain-of-thought, and other impressive capabilities, garnering significant attention.

Generalized macromodels (e.g., GPT-4 and LLaMA) perform well in language comprehension and generation tasks, but struggle with those requiring specialized knowledge, such as sentiment analysis and micro-expression recognition [15] [16]. For this reason, researchers have created high-quality datasets (e.g., EMER Coarse and EMER Real-Time) and targeted model optimization strategies.

Table 2. Examples of emotion recognition models based on large language models

Models	Emotion-LLaMA	EmoLLM	ExpLLM
Core functionality	Multimodal Emotion Recognition and Reasoning	Mental Health Support	Experience-driven self-directed learning
Technological base	Improved LLaMA + multimodal coding	Multi-model fusion + command fine-tuning	Experience Pool + Natural Language Retrieval
Input modal	Audio, Visual, Text	text-based	Text (expandable to images)

As shown in Table 2, Emotion-LLaMA model is an emotion recognition model based on large language models, which was proposed by Cheng et al. This model introduces an emotion-specific encoder and fuses audio, visual, and textual information. It also fine-tunes the large model. Another model is the EmoLLM, which was proposed by Yang et al. in 2024, which is based on a systematic evaluation of the generalized large model's ability to comprehend emotions. This model constructs a benign closed loop of "evaluation-optimization" to promote the development of multimodal emotion recognition [17] [18]. The EmoLLM model is based on a systematic evaluation of the general model's ability to understand emotions and builds a closed loop of "evaluation-optimization" to promote the development of multimodal emotion recognition. For the problem of recognizing facial microexpressions, Lan et al. proposed the ExpLLM model. This model analyzes the causes of emotions from facial action units and uses the chain thinking mechanism to generate detailed explanations of emotions. The ExpLLM model performs outstandingly in the microexpression recognition task.

Zixing Zhang et al. extensively compared the performance of the ExpLLM model with that of other state-of-the-art (SOTA) models. They demonstrated that LLMs can achieve comparable or better performance in emotion recognition tasks. LLMs can capture more diverse patterns, linguistic cues, and emotion-related contextual information through large amounts of training data[19]. The excellent generalization ability and interpretability not only significantly improve performance but also expand the range of application scenarios.

The generalized large model's strong visual emotion understanding and language generation capabilities are a big help in accelerating the development of truly intelligent service robots. Expanding the data dimensions for fusion analysis in the future will further improve the effectiveness of multimodal emotion recognition in service robot applications.

5. Personalized Information Representation and Dynamic Computing

The problem of missing context is a common issue in most studies. Individualized information, such as personal social networks and personality traits, is an excellent choice for solving this problem because it is very rich in data. Individual characteristics, such as gender, age, cultural background, and personality traits, as well as differences in social networks, interaction patterns, and personal interests, can affect the accuracy of sentiment recognition computations. Rui et al. propose a model for predicting individual sentiments that combines user interests and social influence. In offline training, a user interest graph and a social influence graph are constructed by this model. A probabilistic graph model and a SampleRank algorithm are then used to learn personalized weights, α and β, for each user based on historical data. In online prediction, the user interest graph is used with a Bayesian model to fuse textual and image information to predict endogenous sentiment. The social influence graph considers the strength and timeliness of influence. The average friend sentiment is weighted to predict the influence of friends on user u's sentiment. Finally, the learned personalized weight parameters are combined with the prediction to obtain individual sentiment[20]. This approach considers personal factors and the social environment and can capture the complexity and dynamics of individual emotions more comprehensively.

Zhao et al. proposed a long short-term memory (LSTM)-based model that fuses facial expressions, speech, and text information. This model uses an attention mechanism to dynamically assign weights to different modal information [21]. Mollahosseini et al. , proposed a deep learning–based approach to map facial expressions into the valence-arousal-dominance (VAD) space, capturing subtle emotional changes through a model trained using the AffectNet dataset [22]. It performs well in continuous emotion recognition tasks.

6. Challenges and Prospects

As affective computing becomes more widely adopted, privacy and data security issues must be addressed. To this end, differential privacy techniques introduce appropriate noise into the data. Blockchain technology guarantees transparent, non-tamperable data circulation through decentralization. Federated learning techniques enable local model training without uploading data to a central location [11]. And it is worth mentioning that the cost of federated learning technology is not so high, and it can be utilized to its full advantage in relatively private scenarios, such as those in household robot environments. The local attributes of home robot environments provide convenient opportunities for collecting large amounts of personalized information, while fully guaranteeing user privacy. This greatly facilitates collecting personalized information representations and performing dynamic computations.

Multimodal emotion recognition algorithms have limited data dimensions and can only fuse three to four-dimensional data sources for analysis and computation. This is also one reason why various algorithms have different strengths and weaknesses. Additionally, personalized information is a type of data source with rich characteristics, but due to the specificity of its data structure, it is difficult to fuse with other data dimensions. If people can expand the data dimensions, optimize the algorithms, or introduce a more efficient auxiliary reference mechanism for personalized information in the future, multimodal emotion recognition technology will certainly make further breakthroughs.

7. Conclusion

This paper first points out that the application of multimodal emotion recognition technology to service robots requires solving three key problems: improving the generalization ability of emotion recognition, optimizing the interaction experience, providing more accurate and personalized emotion recognition. Then it gives examples of several advanced model algorithms that can help solve these problems and evaluate and analyze them, to summarize their common limitations and possible future improvements.

This paper provides an overview of the research on multimodal emotion recognition technology for service robots. On the one hand, it allows relevant researchers to quickly grasp the development status of this technology. On the other hand, it helps them clearly understand several real-world problems for the application of multimodal emotional recognition in service robotics and refer to advanced modeling algorithms to make improvements and innovations, thus pushing the emotion recognition and interaction capabilities of service robots forward.

References

[1] Bhati V S, Tiwari N, Chawla M. A generalized zero-shot deep learning classifier for emotion recognition using facial expression images. IEEE Access, 2025, 13: 18687 - 18700.

[2] Qi F, Zhang H Z, Yang X, et al. A versatile multimodal learning framework for zero-shot emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34 (7): 5728 - 5741.

[3] Gandhi A, Adhvaryu K, Poria S, et al. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 2023, 91: 424 - 444.

[4] Sedehi J F, Dabanloo N J, Maghooli K, et al. Multimodal insights into granger causality connectivity: Integrating physiological signals and gated eye-tracking data for emotion recognition using convolutional neural network. Heliyon, 2024, 10 (16): e36411.

[5] Guo Z E, Yang M Q, Lin L, et al. E-MFNN: an emotion-multimodal fusion neural network framework for emotion recognition. PeerJ Computer Science, 2024, 10: e1977.

[6] Wang Z, Wang Y H. Emotion recognition based on multimodal physiological electrical signals. Frontiers in Neuroscience, 2025, 19: 1512799.

[7] Sun Y, Zhou Y, Zhang X. Speech emotion recognition fusing functional paralanguage proportion coefficient. Journal of Northeastern University (Natural Science), 2024, 45 (1): 40 - 48.

[8] Zhu T, Li L, Yang J, et al. Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia, 2023, 25: 3375 - 3385.

[9] Wen C S, Jia G L, Yang J F. DIP: dual incongruity perceiving network for sarcasm detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 2540 - 2550.

[10] Ghandeharioun A, McDuff D, Czerwinski M, et al. EMMA: an emotion-aware wellbeing chatbot. In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction, 2019: 1 - 7.

[11] Zhao S C, Feng Y F, Zhang Z C, et al. Research advancements on emotionally and intellectually integrated digital humans and robotics. Journal of Image and Graphics, 2025, 30 (6): 2139 - 2160.

[12] Zhao S, Jia Z, Chen H, et al. PDANet: polarity-consistent deep attention network for fine-grained visual emotion regression. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019: 192 - 201.

[13] Zhang Y, Xie D, Luo D, et al. Modality emotion semantic correlation analysis for multimodal emotion recognition. Computers and Electrical Engineering, 2025, 126: 110467.

[14] Guo Z, Yang M, Lin L, et al. E-MFNN: an emotion-multimodal fusion neural network framework for emotion recognition. PeerJ Computer Science, 2024, 10: e1977.

[15] Lian Z, Sun L, Sun H, et al. GPT-4V with emotion: a zero-shot benchmark for generalized emotion recognition. Information Fusion, 2024, 108: 102367.

[16] Chen Y, Wang H, Yan S, et al. Emotion Queen: a benchmark for evaluating empathy of large language models. [EB/OL], 2025.

[17] Cheng Z, Cheng Z Q, He J Y, et al. Emotion-LLaMA: multimodal emotion recognition and reasoning with instruction tuning. 2025.

[18] Qu Y, Mang Y, Du B. EmoLLM: multimodal emotional understanding meets large language models. 2025.

[19] Zhang Z, Peng L, Pang T, et al. Refashioning emotion recognition modeling: the advent of generalized large models. IEEE Transactions on Computational Social Systems, 2024, 11 (5): 6690 - 6704.

[20] Rui T, Cui P, Zhu W. Joint user-interest and social influence emotion prediction for individuals. Neurocomputing, 2017, 230: 66 - 76.

[21] Zhao Z P, Zheng Y, Zhang Z X, et al. Exploring spatio-temporal representations by integrating attention based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition. In: Proceedings of Interspeech 2018, 2018: 272 - 276.

[22] Molla Hosseini A, Hasani B, Mahoor M H. AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 2019, 10 (1): 18 - 31.