ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

Research on user intent perception and experience optimization in multimodal intelligent interaction scenarios

Ziyou Ma
Washington, USA

Abstract: To address the issues of low accuracy in intent recognition and poor user experience in single-modal interaction in complex scenarios, this paper conducts research on user intent perception and experience optimization in multi-modal intelligent interaction scenarios. The research integrates voice, gesture, facial recognition, and environmental perception technologies, builds a feature fusion model based on Convolutional Neural Network (CNN) + Transformer, designs a hierarchical intent understanding mechanism and interaction execution optimization strategy, and realizes precise perception and efficient response of user intentions. Experiments are selected to verify in three typical scenarios: smart home control, immersive entertainment, and personalized services. The results show that this research scheme can achieve an intent recognition accuracy of over 85%, and the system response delay is controlled within 800ms. This effectively improves the naturalness and intelligence level of multi-modal interaction, providing technical references for the design and optimization of intelligent terminal interaction systems.
Keywords: multimodal intelligent interaction; intention perception; feature fusion; experience optimization; human-computer interaction

1. Introduction

The deep integration of artificial intelligence and Internet of Things technology has driven smart terminals to evolve from single-function devices to multi-scenario interaction hubs. Multimodal intelligent interaction has become the core development direction in the field of human-computer interaction. Compared with traditional single-modal interaction methods such as remote controllers and touch screens, multimodal interaction integrates various input methods including voice, gestures, and vision, which better conforms to human natural communication habits. However, in complex real-world scenarios, it still faces many problems: the recognition rate of voice interaction significantly decreases in noisy and multi-person environments, gesture recognition is easily interfered by environmental factors such as lighting and occlusion, and existing systems mostly remain at the simple superposition of single-modal technologies, lacking a unified feature fusion architecture, and are unable to accurately perceive the user's deep intentions. At the same time, most interaction systems ignore the correlation between environmental information and user behavior, and cannot dynamically adjust response strategies according to the scene, resulting in insufficient coherence and adaptability of the interaction experience.

Currently, domestic and foreign manufacturers have been conducting research on multimodal interaction technologies. Companies like Google and Amazon have launched intelligent interaction devices that integrate visual and audio features. Several domestic brands have also incorporated multimodal perception functions into their smart terminals. However, the existing solutions have not addressed the core issues of insufficient cross-modal feature fusion and shallow understanding of intentions. In this context, conducting research on user intention perception and experience optimization in multimodal intelligent interaction scenarios, building an integrated multimodal perception fusion architecture, and designing precise intention understanding and decision-making mechanisms have significant practical significance for enhancing the interaction capabilities of smart terminals and optimizing user experiences.

2. Multimodal intelligent interaction system design

2.1. Overall system architecture

To achieve precise perception of user intentions and optimization of interaction experience, this research

ISSN: 2424-8460(O)
2251-2608(P)

designs a hierarchical multi-modal intelligent interaction architecture. The overall structure is divided into the input layer, the perception layer, and the output layer. Each layer collaborates to complete the entire process tasks of multi-source information collection, feature fusion, intention perception, and interaction execution.

The input layer serves as the information collection terminal, equipped with an RGB-D camera, a 6-speaker circular array, and temperature/humidity/light sensors, enabling the simultaneous collection of user behavior information and environmental scene information, providing comprehensive data support for subsequent intention perception. The perception layer is the core processing unit of the system, completing feature extraction and cross-modal fusion of multimodal data based on the CNN+Transformer feature fusion model. It combines the Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) to achieve precise recognition of facial expressions and behavioral characteristics, and uses the semantic-level intention understanding algorithm to complete in-depth analysis of user behavior. The output layer adopts the dual-mode communication protocol of Wi-Fi 6 and Bluetooth 5.2, coordinating with the functional modules of the intelligent terminal to convert the user intentions parsed by the perception layer into executable instructions. At the same time, it enhances the operational smoothness and accuracy through the interaction execution optimization module.

2.2. Core technology system

2.2.1. Multimodal perception fusion technology

Multimodal perception fusion is the core for accurately perceiving user intentions. Its main goal is to integrate and collaboratively process heterogeneous information from different sensors, eliminating the limitations of single-modal data. The RGB-D images, voice signals, and environmental parameters collected at the input layer are first preprocessed through the feature extraction module: the visual modality uses an improved CNN to extract spatial features of gestures and faces, combined with the LBP algorithm to achieve facial micro-expression recognition; the speech modality uses the deep Fourier transform to extract spectral features, and through the time-delay neural network to capture the temporal information of speech.

The preprocessed multimodal feature vectors are sent to the Transformer fusion module. Through the self-attention mechanism, the weighted fusion of multi-source features is achieved. During the fusion process, the weights are dynamically adjusted according to the confidence threshold of each modality (CT = 0.65). The data of modalities with a confidence lower than this threshold will be temporarily suppressed to avoid the interference of invalid information on the fusion result. To solve the problem of temporal alignment between modalities, the timestamp synchronization mechanism and the unified mapping strategy in the coordinate space are introduced to ensure the consistency of different modal data in the temporal and spatial dimensions, thereby improving the overall perception accuracy.

2.2.2. Hierarchical intention understanding and decision-making mechanism

Intention understanding and decision-making are crucial for converting multimodal perception information into executable instructions. This study designs a hierarchical intention understanding mechanism to achieve layered parsing of user's instruction-level and demand-level intentions. Firstly, the fused semantic vectors are identified at the instruction level. Common control instructions such as "increase volume" and "switch channel" are matched through a predefined action template library. The template matching threshold is set at 0.72. If the threshold is lower than this value, the deep semantic analysis will be initiated.

In the demand-level reasoning stage, scene state variables are introduced as context constraints, and the response strategy is dynamically adjusted based on the operating mode of the intelligent terminal. The deep user demand is parsed through the context-enhanced intent classifier. The decision engine generates specific operation instructions according to the priority rules of the instructions, and at the same time, a context update mechanism with a 3-second time window sliding is designed to achieve semantic ambiguity resolution and state consistency maintenance during scene switching, ensuring the responsiveness while avoiding misjudgment.

2.2.3. Interactive execution optimization technology

Interactive execution optimization is a crucial link that connects intention perception with actual interaction. Its core objective is to enhance the smoothness and accuracy of operations. The optimization strategies are designed from three dimensions: interface interaction, handling of multi-person scenarios, and multi-device coordination. In the interface interaction stage, a gesture confidence threshold (GCT = 0.78) is set. Gestures with a confidence level below this threshold are judged as invalid actions to prevent accidental triggering. In multi-person usage scenarios, a spatial position weight function is introduced:

$$W(d)=e^{-kd} \quad (1)$$

(k = 0.15, d being the horizontal distance between the user and the center axis of the terminal), evaluate the

ISSN: 2424-8460(O)
2251-2608(P)

operation priorities of each user and solve the problem of multiple instructions conflicts.

The multi-device linkage process design is based on a timestamp-synchronized task scheduling mechanism to ensure the temporal consistency of instructions from different terminals; at the same time, a context state locking mechanism and a 1.5-second operation rollback window are set up to effectively solve the problems of signal interference and accidental operations, and enhance the stability and continuity of interaction and experience.

3. Experimental verification and result analysis

3.1. Experimental scenario and parameter settings

To verify the effectiveness of the multimodal intent perception and experience optimization scheme, a control experiment was designed for three typical application scenarios: smart home control, immersive entertainment, and personalized services. The experiment was conducted in a standard living room environment of 5m×4m, with RGB-D cameras and a 6-speaker circular array deployed. The core parameters were initialized as follows: CT = 0.65, GCT = 0.78, distance attenuation coefficient k = 0.15, and time window length of 3 seconds.

Recruit 15 test users, conduct 20 repetitions for each scenario, and accumulate a total of 300 sets of test data. Control the environmental light intensity at 300-500 lx, keep the distance between the user and the terminal at 2.5m, and maintain the room temperature at (22 ± 2) °C. Eliminate the influence of environmental variables on the perception accuracy. Set evaluation indicators and preset goals: the accuracy rate of intention perception ≥ 85%, the average response time ≤ 800 ms, the hit rate of the confidence threshold ≥ 75%, the success rate of fusion ≥ 80%, and the system stability ≥ 90%. Use descriptive statistical analysis and one-way analysis of variance to verify the significant differences between scenarios.

3.2. Experimental results and analysis

The experimental results show that all the evaluation indicators in the three scenarios have reached the preset targets, verifying the technical feasibility of the multimodal perception fusion architecture and the experience optimization strategy. The performance test results of each scenario are shown in **Table 1** and **Figure 1**.

Table 1. Performance test results of the multimodal intelligent interaction system in various scenarios.

Application scenarios	Response accuracy rate/%	Average response time/ms	Confidence threshold hit rate/%	Integration success rate/%	System stability/%
Smart home control	87.5	732	78.3	82.7	92.1
Immersive Entertainment	89.2	675	81.6	85.4	94.3
Personalized service	85.8	789	76.9	80.2	90.5
Pre-set goal	≥85	≤800	≥75	≥80	≥90

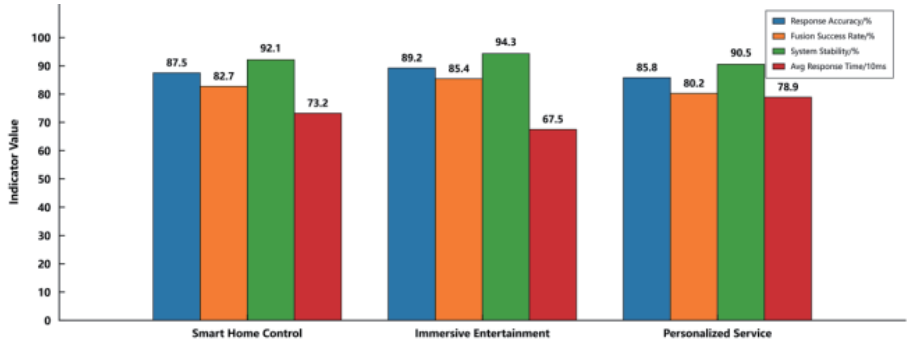

Figure 1. Comparison of performance indicators for various scenarios of the multimodal intelligent interaction system.

The intent perception accuracy rate of the smart home control scenario is 87.5%, with an average response time of 732ms. The excellent performance of this scenario is attributed to the high matching efficiency of the predefined action template library, which can quickly parse standardized home control instructions. The immersive entertainment scenario performs the best, with an intent perception accuracy rate of 89.2% and an average response time of 675ms. This is because the user's gesture actions in this scenario are relatively standardized, and the CNN + Transformer model can effectively recognize standardized operation instructions.

ISSN: 2424-8460(O)
2251-2608(P)

The design of a 3-second time window and a 1.5-second operation rollback window significantly improves the response speed and stability.

The accuracy rate of intent perception in the personalized service scenario is 85.8%, and the average response time is 789ms. The main reason for the slightly lower indicators in this scenario is that personalized services involve complex semantic understanding, and higher requirements are placed on the fusion of multimodal features and the deep analysis of intentions. This reflects that the intent perception in complex scenarios still needs to further optimize the model's processing capabilities.

Overall, the dual-mode communication protocol of Wi-Fi 6 and Bluetooth 5.2 ensures the low-latency response of the system. The multi-modal perception fusion architecture effectively improves the accuracy of intent recognition. The hierarchical intent understanding and interaction execution optimization strategies significantly enhance the user's interaction experience and solve many pain points of single-modal interaction.

4. Conclusion

This research focuses on the perception of user intentions and the optimization of user experience in multi-modal intelligent interaction scenarios. A hierarchical multi-modal intelligent interaction architecture was constructed, and a core technical system for multi-modal perception fusion, hierarchical intention understanding, and interaction execution optimization was designed. Through empirical experiments, the effectiveness and practicality of the scheme were verified. The research results show that the CNN + Transformer feature fusion model can effectively integrate multi-source heterogeneous information, the hierarchical intention understanding mechanism can enhance the ability to interpret users' deep needs, and a series of interaction execution optimization strategies can effectively solve problems such as multiple-person conflicts and accidental triggering, significantly improving the fluency and stability of the interaction.

References

[1] Yujia L , Huijun D , Chao Z ,et al.User acceptance-based interaction design and teaching application of intelligent in-vehicle virtual robots[J].Experimental Technology & Management, 2025, 42(2).DOI:10.16791/j.cnki.sjg.2025.02.028.

[2] Ji Y , Zhang C .Research on the Interaction Design of Sense of Control in Intelligent Driving Scenarios[C]// International Conference on Human-Computer Interaction.Springer, Cham, 2025.DOI:10.1007/978-3-031-92692-1_5.

[3] Haipeng C , Li Y , Jing Y ,et al.Research on User Experience Adaptive Algorithm in Intelligent Connected Vehicle Interaction[J].2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM), 2024:340-344.DOI:10.1109/ispcem64498.2024.00063.

[4] Zhang Y , Zheng Y , Yin C ,et al.Exploration on Intelligent Interaction Design and User Experience of VR Platform Based on Intelligent Perception and Deep Learning[J].Proceedings of the First International Conference on Science, Engineering and Technology Practices for Sustainable Development, ICSETPSD 2023, 17th-18th November 2023, Coimbatore, Tamilnadu, India, 2024.DOI:10.4108/eai.17-11-2023.2342820.