Hindawi
Scientific Programming
Volume 2022, Article ID 5460766, 18 pages
https://doi.org/10.1155/2022/5460766

Research Article

MES: A Helping Elderly Escort Interactive System Based on Reverse Active Integration of Multimodal Intentions

Xujie Lang[1,2], **Zhiquan Feng**[1,2], **Xiaohui Yang**[1,2] and **Tao Xu**[1,2]

[1]*University of Jinan, Department of Information Science and Engineering, Jinan 250022, China*
[2]*Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China*

Correspondence should be addressed to Zhiquan Feng; ise_fengzq@ujn.edu.cn

Received 31 March 2022; Accepted 27 July 2022; Published 19 September 2022

Academic Editor: Daniela Briola

Lacking human-computer collaborative capability is one of the key problems commonly faced by the existing escort robots. To this end, a helping elderly escort interactive system based on reverse active integration of multimodal intentions (MES) is proposed in this paper in response to the elderly's decline in language expression ability, memory, and other abilities. This system can understand the intention of the elderly based on scene perception and three-modal information: speech, gesture, and posture. In detail, the system can extract the interactive intention from the nondeterministic multimodal data input by the elderly and evaluate the trust degree of the extracted intention. The evaluation of intention trust degree is such a process that the system autonomously judges the feasibility of the elderly's intention and corrects the wrong intention expressed by the elderly due to memory decline by reverse thinking of "find reasons based on the results"; when the intention cannot be extracted, the system will take the initiative to ask the elderly for enhanced information conducive to the intention extraction, so as to quickly and correctly extract the interactive intention of the elderly. This design aims at improving the quality of elderly care, making the interaction between the elderly and the robot more natural, improving the accuracy in intention extraction from fuzzy expression, as well as breaking the traditional "master-slave" human-computer interaction and improving the harmony of human-computer interaction. Further, the implementation principle of the system is detailed, and the system is evaluated in this paper. The evaluation experiment was conducted by a robot Pepper embedded with the system. Through experiments, it is verified that Pepper can quickly and accurately get the real intention of the elderly in the interaction with the elderly. In a challenging environment (such as the unclear expression of the elderly), it can still correctly extract the real intention of the elderly with an accuracy of 97% and can effectively avoid the wrong intention expressed by the elderly. This puts forward a valuable research path for the challenge in human-computer collaborative interaction.

1. Introduction

With the rapid development of Internet, "strong intelligence" has become a new star of concern. People hope that machines can be as intelligent as humans, which promotes the development of robots [1]. Moreover, the family planning policy in the last decade triggered a series of problems such as labor contraction and aging [2]. According to a survey [3], by 2050, China's elderly population over 60 years old will reach 430 million, and the elderly population over 65 years old will reach 320 million. Elderly caregivers may be in short supply in the future. Therefore, it has become an urgent need to let robots take care of the elderly instead of caregivers. However, there are still some thorny problems in the escort robot system at this stage. For instance, at present, few robot systems are designed according to the characteristics of the elderly; no good solution has been put forward to the problems such as the decline in robot's escort quality caused by the fuzzy expression and memory decline of the elderly [4].

At the present stage, due to the uneven level of education or the decline in expression ability with age, the elderly's expression by speech, gesture, and posture may not be as coherent and fluent as young people, so it may be difficult for a robot to understand the intention of the elderly based on single-modal information. Therefore, we hope that the robot

can integrate the multimodal information of the elderly, so as to effectively eliminate the ambiguity of expression and make the human-computer interaction develop in a more natural and harmonious direction [5]. According to the way of communication with the elderly in real life, the situation that "the intention of the elderly cannot be understood due to the unclear expression by single-modal information" may be alleviated if the robot can inversely analyze and actively ask the incomprehensible information just like a human. Moreover, in the current research studies, few robots take time and place as the parameters for evaluating the feasibility of an intention. In fact, environmental information is very important for caring the elderly. Based on environmental information, the escort robot can accurately understand "ambiguous expression," monitor wrong intention, and give a warning in time, which will greatly reduce the accident risk when the elderly is alone. For the robot, the environmental information is a general guide for extracting the real intention and may be used for eliminating the ambiguous expression, enabling the robot to get the final intention accurately.

To solve the above problems, this paper designs a helping elderly escort interactive system based on reverse active integration of multimodal intentions (MES system). Taking into account the characteristics of the elderly, such as the decline of memory and the decline of expression ability, it can correctly infer the user interaction intention, correct the wrong intention, and carry out human-computer cooperation to finally complete the intention task under the multimodal or erroneous expression of the user. The MES system uses the method of multimodal intention reverse analysis and fusion to extract the intention of the elderly, which improves the accuracy of intention recognition and the degree of interaction harmony in the process of human-computer interaction, and uses the way of active interaction to improve the fluency of human-computer interaction and reduce the interaction load of the elderly.

The rest of this article is organized as follows: the second section introduces related work; the third section introduces the general idea of the MES system; the fourth section introduces the core algorithm of the MES system in detail; the fifth section describes the testing and evaluation of the MES system; then in the sixth section, it summarizes and describes the future research direction.

2. Related Work

In recent years, robots have been changing rapidly from industrial environment to private environment with an imperfect structure and may become partners of people in daily life. Therefore, the intention understanding in human-computer interaction has become an important content of research [6]. Researchers have indeed made many outstanding contributions to the intention extraction by robot. For example, in Reference [7], Mohammed et al. manipulated a computer by recognizing the user's gesture intention. They used two position cameras to improve the computer interaction system with mouse in three dimensions and used gesture instead of computer mouse to interact with the computer, which broke the traditional mode of interface interaction and achieved a good result. However, the application environment of this improved system is limited. Only if the gesture is made in a fixed area of the two cameras can the personal computer be controlled. In Reference [8], Rafferty et al. expanded the scope of human-computer interaction and realized smart home assistance system by using sensors. The sensors were placed on furniture or routine items, and a sensor was also placed in people's pocket to realize human activity recognition through mutual sensing between sensors, so as to understand people's intention. The sensor could communicate with another sensor via Bluetooth to send the recognition result to the user's mobile phone. Over three months of transformation, the researchers created an intelligent home with sensors. Through the experiment, it was found that the correctness of intention understanding through activity recognition reached 83.3%. However, the signal receiving range of the sensor was constrained. In other words, only if the user's action was within the room where the sensor was installed could it be recognized; otherwise, the accuracy of intention understanding would be greatly reduced. Moreover, the system could only extract the intention by using sensor, and there might be a deviation in the intention recognition of similar actions of the user. In order to solve the problems in space and accuracy of sensor recognition, the intention recognition system studied by Jose and Lakshmi mainly receives and processes user's speech information [9], converts the speech into text information, extracts the speech intention, and then attaches the speech intention to each slot. Hence, a novel method for slot filling and intention prediction was proposed, where speech interaction was used to expand the scope of human-computer interaction. Through experiments, it was found that the intention understanding accuracy of this system was more than 90%. However, this system is suitable for users who can clearly express their intention in Mandarin. For the elderly with declined expression ability, the system may not give play to its real advantages, reducing the harmony of interaction. To solve this problem, Hatori further researched the speech recognition [10]. They used unconstrained oral instructions to extract intention and combined deep learning-based target detection with natural language processing technology to process unconstrained speech instructions; further, they proposed a method for robots to solve the problem of ambiguous instruction through dialogue and demonstrated that their system could effectively understand the natural instructions of human operators.

In specific circumstances, the systems mentioned in the above literature indeed can show good results. With simple interaction modality and single-modal input, such systems can give full play to their advantages only when being used by the specific group of people. However, for the elderly, it seems that the result of intention recognition by inputting a single mode cannot meet the expected requirements. Many researchers also used large-scale experiments and machine learning method to study how humans make decisions [11] and found that human made decision on the basis of multimodal input. When studying the process of human-

computer interaction, some other researchers found that the human brain did not have a single mode of operation because the brain was not composed of one thing [12]. With the continuous advancement of relevant research, human-computer interaction has been developed from a single-modal processing mode to multimodal processing mode, providing a deeper definition for "intelligence." For example, in Reference [13], researchers in the Massachusetts Institute of Technology (MIT) studied a robot that could play Jenga independently. When playing the game, the robot needed to integrate vision, touch and object behaviors, and learn the properties of Jenga building blocks by integrating a variety of network models to complete the game. Although this system really integrates multimodal information, its interaction object is not human but building blocks; but, this system indeed can integrate multiple modes and think independently. In Reference [14], Zlatintsi used online speech recognition and Kinect gesture recognition functions to assist a bathing robot to extract human intention. They proposed an automatic multimodal recognition system based on the latest signal processing technology and pattern recognition algorithm. In a noisy environment, the intention extraction accuracy of the system integrated with the two channels reached 84%, which was significantly higher than that of single-modal system. However, regarding intention evaluation, the system must work offline, which may take a long time, causing reduction in the interaction effect. In Reference [15], Zhao and Wang introduced an immersive system prototype, which integrated a three-modal (face, gesture, and speech) recognition technology to achieve human-machine interaction. The server integrated different sensor inputs in a time-sensitive way, so that the system could understand user behavior in real time. In this system, the analysis focus is speech; other modal information is the supplement of speech information. For the elderly, modal information integration should be considered from all perspectives in order to extract an intention more accurately. Considering the expression problem, Kang et al. proposed a learning-based intention detection method using first-person camera [16]. In this method, visual signals and biological EMG signals were received by cameras and wearable sensor robots, so that users could express their intentions without speech or posture movements, which effectively avoided unclear intention expression. Penaloza and Nishio proposed a more advanced wearable robot [17] that could extract users' intention via a brain-computer interface. Different from the traditional brain-computer interface, they developed a noninvasive BMI without implanting implants into human brain. Users only need to sit on a chair equipped with a mechanical arm, wear a 16-channel EEG helmet, and connect to the electrode to control the mechanical arm and realize multitask processing by using the noninvasive BMI, while using their own hands to do other tasks. This brain-computer interface breaks the traditional interaction mode, and the intention extraction is more accurate. However, the biggest problem of this wearable robot is that a data line is needed for connecting to the computer to realize real-time detection of intention, resulting in reduced range of human-computer interaction.

Researchers believe that human-computer interaction should take the interactive relationship into consideration. Some studies have shown that at this stage, most robot research studies still focus on the "master-slave" interaction; the equality of the interactive relationship directly affects the quality and results of interaction [18]. In most of the existing research studies on human-computer interaction, "computer" is in passive position, while "human" takes the absolute initiative; the service robot absolutely listens to human instructions without own thinking. The effect of such human-computer interaction cannot meet people's expectations of "intelligence," and the interactive experience is poor. That is, upon receiving human instruction, the robot will do the task step by step regardless of the correctness, so that the robot seems stiff and clumsy when interacting with people. In this regard, Sun et al. [19] studied the use of active and passive haptic forms in AR interaction; in addition that people could use passive haptic form to change the shape of an object in the virtual scene, the machine could actively give tactile feedback to people when people touch an object in the virtual scene, so that people can know about the stiffness of the object in the virtual scene. In Reference [20], Flesher et al. applied this method to the brain-computer interface. The system they developed is embedded into the body of a paralyzed person, integrating the incoming information of muscles, skin, and joints. And, this interaction is two-way. The system can use the input touch feeling to supplement human vision, so as to achieve the natural interaction model based on human-computer integration. However, this system is only a preliminary attempt of equal interaction; it still relies on human instructions to actively send feedback. Meanwhile, Zheng et al. adopted a haptic perception-based active social assistant robot for patients with Alzheimer's disease [21]. They thought that the existing animal-like robots were usually passive, so the robot they designed used the whole-body sense of touch to sense the input information of people. When user touches the robot, the robot will give corresponding feedback and actively ask the user to touch the positions of the robot in a certain order when the user touches it next time. However, this interaction between robot and user is only based on haptic single-modal input, and the function is only for rehabilitation training. Moreover, the human-computer interaction is not strong enough, with limitation in usage scenes. Therefore, Kelley et al. [22] made a breakthrough over the previous two studies in terms of the "initiative" of robot system. Context information in the form of object revelation and object state was used to improve the performance of potential intention recognition system. Robots equipped with this system can actively obtain user state and treat it as the context information, such as opening a book, closing a book, and opening a computer. It can also predict the possible state of the user after this state and actively ask the user whether he needs further help. This way has basically realized equal participation in human-computer interaction, but the context information of this system is static. In an unfamiliar environment, the interaction may lack flexibility and robustness.

As can be discovered from the above research studies, there are still many deficiencies in the escort robot at this

Table 1: Comparison between related work and MES.

Work	Multimodal identification	Unconstrained command	Unconstrained recognition space	Equality of interactive participation	Scalability	Recognition effect (reaches more than 90%)
[7]	✗	✗	✗	✗	✓	✓
[8]	✗	✓	✓	✗	✗	✗
[9]	✗	✗	✓	✗	✓	✓
[10]	✗	✓	✓	✗	✓	✓
[13]	✓	—	✓	—	✗	✓
[14]	✓	✓	✓	✗	✓	✗
[15]	✓	✓	✗	✗	✓	✗
[16, 17]	✓	✓	✗	✗	✗	✓
[19]	✗	✗	✗	✓	✗	✓
[20]	✓	—	✗	✓	✗	✓
[21]	✗	✗	✗	✓	✗	✓
[22]	✓	✗	✓	✓	✓	✓
MES	✓	✓	✓	✓	✓	✓

stage, and it is rarely designed according to the characteristics of the elderly. In order to solve the problem of interaction between the elderly and the escort robot, the escort robot system must have a set of cognitive process based on human interaction with the outside world and an interaction model built for the common characteristics of the elderly [23]. The research group specially went to families and nursing homes to research the elderly's demand and know about the required aspects of coordination in daily life of the elderly. The results showed that the elderly had decline in expression ability and memory; their gestures and body language in interaction often expressed important information [24]. Therefore, each mode of information should be comprehensively synthesized in the process of intention extraction, so that the robot can better understand the elderly from many aspects. Moreover, experimental support has been obtained in previous work of the research group [25, 26], and the proposed method is also consistent with the human-human interaction. However, poor interaction and weak human-computer coordination ability are still one of the challenges faced by the elderly escort robot system. Therefore, considering various factors, the MES system was constructed in this research. As proved in previous research studies, the environment plays a positive role in the robot's intention extraction [27]. Hence, this system used four-modal information (speech, gesture, posture, and environment) as the system input to extract and integrate the intention of the four-channel information. In the aspect of reverse and active interaction, the trust degree of the extracted intention was evaluated in this research, taking the modal information entropy, scene perception information, and historical information as the parameters of intention trust. For the intention with trust degree substandard, the system can reversely seek the trust degree of single-modal information, take the initiative to ask the user to re-input enhancement information against the modal information with substandard trust degree, extract the intention again, and re-evaluate the trust degree to repeatedly refine the user's real intention. This method not only helps enhance the interactive sense and improve the accuracy in intention extraction but also helps reduce the probability of potential risk of the elderly living alone and improve the escort quality. The system adopts environmental analysis and perception, the cognitive model uses multimodal fusion algorithm and reverse understanding thinking, and adopts the form of robot active interaction, which breaks the restriction of environment use and gets rid of the problem of unequal interaction in the traditional human-computer interaction.

In the context of caring for the elderly, in order to more intuitively highlight the advantages of the MES system, we systematically compared the related work with the MES system, and the results are shown in Table 1.

We cannot deny that related work can perform very well in specific scenarios, but this paper mainly focuses on the elderly care, so we only judge whether the system is suitable for the elderly care. In the table, the performance of the system is judged from six aspects: whether the system can receive multimodal input, whether the user input instructions are constrained, whether the recognition space is constrained, whether the system participates equally in the interaction with the user, whether the system is easy to expand, and the recognition effect. It is not difficult to see that the MES has more prominent advantages than other related work in the elderly care environment.

The system adopts environmental analysis and perception method, multimodal fusion algorithm and reverse understanding thinking, and robot active interaction form, breaking the restriction in usage environment and getting rid of unequal interaction problem in traditional human-computer interaction.

3. General Idea of the MES System

The system interacts by reverse active integration of multimodal intentions. The general idea is to extract and integrate an intention from the multimodal information obtained by the system and evaluate the trust degree of the intention. If the trust degree of the intention is substandard, the system will reversely check each mode of information, actively ask the user to re-enter the unqualified single-modal information, and integrate it with the original standard information till the user's real intention is extracted. The general idea of the MES system is as shown in the figure:

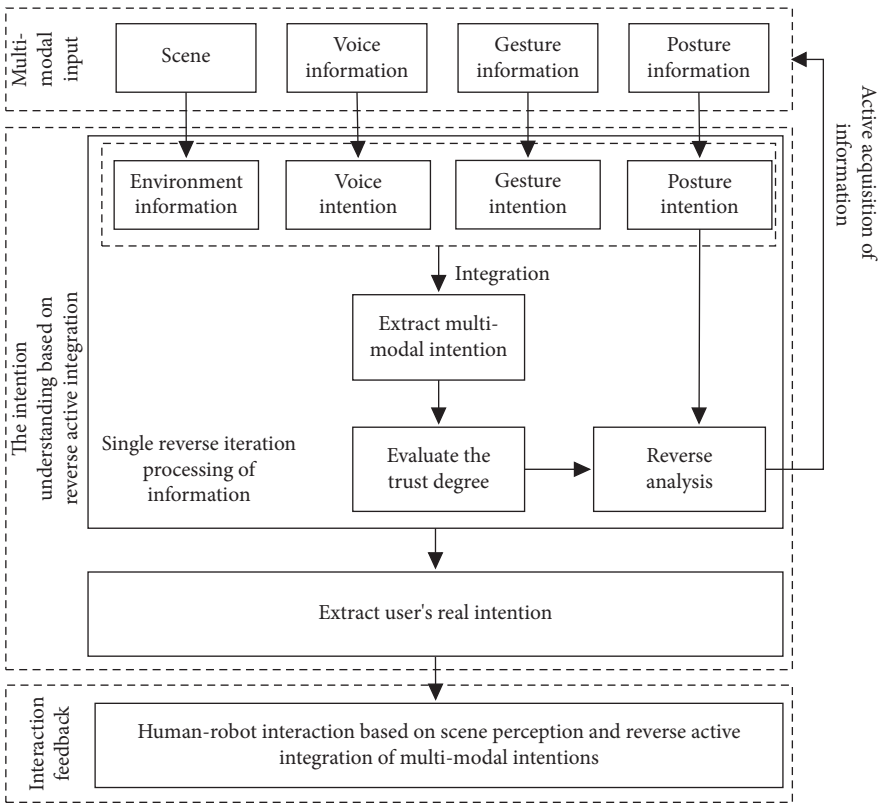

Figure 1: General idea of the MES system.

The general idea of the MES system can be divided into three parts: the multimodal input, the intention understanding based on reverse active integration, and the interaction feedback. The core of the system is the intention understanding based on reverse active integration. The operation process of the system is shown in Figure 1. First, the system needs multimodal input. The system receives the user's speech information through the microphone, obtains the gesture information and posture information through Kinect2.0 [28], and uses the ordinary camera of Pepper robot to obtain the scene. In the intention understanding based on reverse active integration, firstly, the keyword matching algorithm is used to get the speech intention, and the gesture intention and posture intention are obtained through gesture recognition and posture recognition. The YOLOv3 [32] target detection algorithm is used to detect the semantic information in the scene, so as to confirm the interaction places at this time, such as the living room, kitchen, and bedroom. The MES system also obtains the current time information and takes the interactive place and time information as environmental information. Then, the MES system fuses the speech intention, gesture intention, posture intention, and environmental information and extracts the fusion intention. The extracted fusion intention needs to be evaluated by trust degree. The trust degree evaluation mechanism will analyze the trust degree of intention, which refers to the degree of extracting the correct intention. In other words, the higher the probability that the extracted intention is correct, the greater the trust, and vice versa. Trust degree evaluation is a process to acquire the trust degree of an intention (refer to subsection 4.1.3 for details). If the trust degree does not reach the specified threshold, the system will reverse analyze the trust degree of each modal intention according to the current noncompliance results. For the modal information whose trust degree is not up to the standard, the robot actively asks the user for this modal information and continues to use this information as the enhanced information as the system input, fuses with the previously qualified modal information again, extracts the intention, and it is executed repeatedly until the executable intention whose trust degree meets the requirements is condensed. Finally, in the interactive feedback stage, the robot carries out human-computer cooperation according to the extracted intention.

One of the main contributions of this research is to explore a new mode of interaction based on multimodal integration. The characteristic of this mode is that user can input multiple modes such as speech and gesture in parallel, and then the user's intention can be perceived by the MES proposed in this paper. However, if the input is unreliable, the integrated result is certainly unreliable. For this reason, both the trust degree of the integrated result and the reliability of each mode of input were evaluated in the research. In case of mode with low reliability, the system can actively prompt the user to re-enter the mode of information. After obtaining a new input of a certain mode, the system will further make integrated calculation with the other existing modes of input. It can be seen that the reverse input is single-modal, but the input of integrated calculation is multimodal.

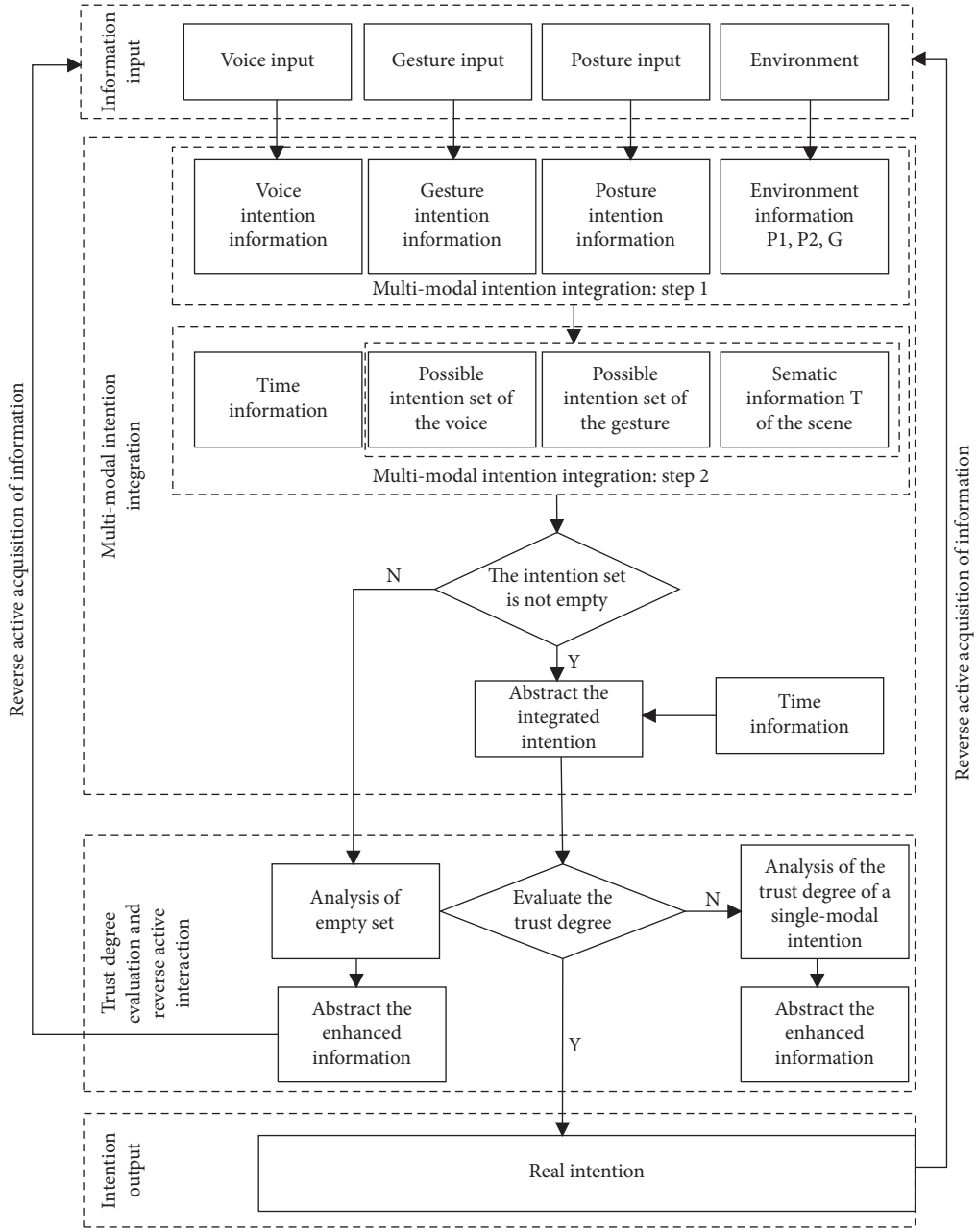

Figure 2: Multimodal intention extraction based on the reverse active integration algorithm.

4. Algorithm Design and Analysis

This section mainly introduces the implementation principle of the MES. The MES refers to such a model that infers user's interactive intention according to the multimodal information of speech, gesture, posture, and scene, evaluates the trust degree of the inferred intention, then actively asks the user to re-enter the modal information substandard, if any, and repeat the steps till the trust degree reaches the set threshold, and finally extracts the intention up-to-standard.

4.1. Reverse Active Integration Algorithm for Multimodal Intention Extraction. The specific implementation process of multimodal intention extraction based on the reverse active integration algorithm is presented in Figure 2.

4.1.1. Information Input. The input multimodal information is represented by the symbol $A_j\,(j = 1, 2, 3, 4)$, where A_1 represents speech information, A_2 represents gesture information, A_3 represents posture information, and A_4 represents environment information. First, the user's speech information A_1, gesture information A_2, posture information A_3, and the environment information A_4 actively obtained by the system are used as the multimodal information input of the system. The system infers the speech intention set iA_1, gesture intention set iA_2, and posture intention set

iA_3 from A_1, A_2, and . . . Respectively. Here, the intention set i_{Aj} is extracted from the modal information A_j because the same modal information may express different intentions. For example, if the user's gesture information A_2 is to stretch out his thumb and little finger at the same time, he may express "number 6" or "phoning" intention. Then, the gesture intention set $i_{A2} = \{''\text{number6}'', ''(\text{phoning})''\}$. In this way, all possible intentions are extracted to generate an intention set, which effectively improves the accuracy in intention extraction of the system.

In terms of multimodal information processing, we use Baidu AI speech [29] for speech modal information processing, which uses a deep learning algorithm similar to neural network to replace the previous recognition module. The reason why Baidu AI speech is adopted is that we consider that most elderly people speak Mandarin non-standard or only use dialects for communication. Baidu AI speech can recognize multiple dialects, with better robustness and a very fast recognition speed. Gesture modal recognition adopts the deep gesture recognition model of our laboratory [30], and in order to improve the robustness of the system, we adopt the method that multiple gestures correspond to the same semantics [31]. When the system obtains enhanced information, the user can change a way of expression to better let the robot understand the intention, and the advantage of this method [31] is that this method uses depth information when recognizing gestures, which makes the gesture recognition effect not affected by the angle of camera recognition and environmental brightness, and it can still work effectively in a complex environment such as home, with an accuracy rate of 99.89%. To obtain environmental information, first of all, use Pepper's own camera to collect photos from different directions of the interaction scene and use the YOLOv3 target detection algorithm [32] to obtain the semantic information of the collected environmental photos. According to the semantic information, determine the interaction location, such as the place where the semantic information includes "sofa, TV, coffee table, and TV cabinet," probably the living room and the place where semantic information includes "bed, bedside table, table lamp, and wardrobe" is the bedroom. Compared with other target detection algorithms, the biggest feature of YOLOv3 is its fast recognition speed, which can improve the efficiency of human-computer interaction. Moreover, we draw lessons from the method of using scenarios to calculate user intentions in Reference [22] and also integrate scenario information into the intention understanding of the system, making full use of scenario information to contribute to the extraction of system intentions. So, we integrate the semantic information identified here into intention understanding. Posture recognition adopts our research results [33]. Compared with other recognition algorithms, this algorithm has a fast recognition speed and is less affected by ambient light, with an accuracy of 95% and high reliability.

4.1.2. Multimodal Intention Integration. The speech modal intention set i_{A1}, gesture modal intention set i_{A2}, environmental information, and posture information i_{A3} were integrated. Among them, the posture information was used to judge whether an intention could be executed in this posture, for instance, people could not drink water when lying down and could not do exercise when sitting down. This information was further used to exclude the impossible intention expressed by speech and gesture in this state and obtain the constrained speech modal intention set I_{A1} and gesture modal intention set I_{A2}. Then, the system got $P1_{y_n}^{L_i}$ of I_{A1} and I_{A2} according to the interaction place (Schedule 2), where L_i represents the place, y_n represents the intention, and $P1_{y_n}^{L_i}$ represents the probability of y_n in L_i. According to the time information, the system got $P2_{y_n}^{t}$ (Schedule 3), where t represents the current time, y_n indicates the intention, and $P2_{y_n}^{t}$ indicates the probability of y_n at time t. The results shown in Schedules 2 and 3 were obtained by a survey on the daily life of 40 elderly people. It was found that the elderly input the same modal information, but might want to express different intentions in different interaction environments, and the probability of occurring the same intention in different interaction places was also different. Hence, the intention-place probability was expressed as $P1_{y_n}^{L_i}$ (Schedule 2). Further, an investigation and analysis were also made on the relationship between time and intention. The result revealed that different time might also affect the probability of intention expression. For example, taking medicine is most likely to occur in the morning, noon, and evening, while the probability of taking medicine in the late night or early morning is small. According to this relationship, the time-intention probability was expressed as $P2_{y_n}^{t}$ (Schedule 3).

$P1_{y_n}^{L_i} * P2_{y_n}^{t}$ was taken as the comprehensive probability $P(y_n)$ of intention y_n, and I_{A1} and I_{A2} were ranked according to the values of $P(y_n)$. This was because the place and time both had effect on the accuracy of intention extraction. The most likely 3 (possibly less than 3) intentions were extracted from I_{A1} and I_{A2}, respectively, to generate modal possible intention sets I_{A1}' and I_{A2}'. Then, the two possible intention sets were intersected to get the integrated intention set I_{A1A2}'.

I_{A1A2}' contains three different intentions at most. Then, calculate the time interval between the current time t_{now} and the last time of intention in I_{A1A2}' as time information because the frequency of different intentions is different, and under the same intention, the longer the time difference from the last intention, the more likely the user is to express the intention, we established the Set Time Difference Probability, as shown in Schedule 4. This Schedule 4 shows the difference between the interval of different intentions and the standard time difference t^{y_n} and the probability of intention occurrence. Among them, $T^{\wedge}$ is the average value of the time difference between the two occurrences of an intention after investigation, which is taken as the standard time difference of intention in this paper. Based on this, the most likely intention expressed by the user at this place and time was analyzed and extracted as the integrated intention; this process is presented by equation (1), where the φ function is for extracting the corresponding intention of the obtained value, t_{now} is the current time, $t_{before}^{y_n}$ is the last occurrence time of t^{y_n}, t^{y_n} is the time difference-intention

setting, and set $t_{now} - t_{before}^{y_n} - t^{y_n}$ to T^{y_n}. Hence, the integrated intention p_s was expressed as follows:

$$p_s = \varphi\left[\max\left(T^{y_n}\right)\right], y_n \in I_{A1A2}'. \quad (1)$$

For example, if I_{A1A2}' contains the intention of "drinking water" and "drinking tea," first calculate the result of the interval of "drinking tea" intention, that is, the intention expressed by the user at t_{now} is "drinking tea," and the last time when the intention expressed by the user at "drinking tea" is $t_{\text{before}}^{y_n}$, we can calculate the time interval $t_{\text{now}} - t_{\text{before}}^{y_n}$ of the user expressing the intention of "drinking tea" twice. t^{y_n} is the standard time difference of two "tea drinking" intentions obtained through investigation (see Schedule 4). By comparing the time interval of "tea drinking" with the standard time difference of "tea drinking," the result $T^{\text{DrinkTea}} = t_{\text{now}} - t_{\text{before}}^{\text{DrinkTea}} - t^{\text{DrinkTea}}$ of "tea drinking" intention interval is obtained. Similarly, calculate the "drinking water" intention interval result $T^{\text{DrinkWater}} = t_{\text{now}} - t_{\text{before}}^{\text{DrinkWater}} - t^{\text{DrinkWater}}$ and select the intention with the largest interval result as the fusion intention p_s. If I_{A1A2}' is an empty set, different treatment will be made for different situations, and then the system will reversely and actively obtain the enhanced information.

4.1.3. Trust Degree Evaluation and Reverse Active Interaction. After extracting the fusion intention p_s, the system needs to further confirm the extracted fusion intention p_s, so this section proposes a trust evaluation and reverse active interaction method to prevent the user from expressing the wrong intention due to forgetfulness and check whether the system makes mistakes in intention fusion stage, so as to improve the fault tolerance rate of the system itself. For the case of successfully extracting the intention, that is, after successfully extracting the fusion intention p_s, we need to get the comprehensive probability $P(y_n)$ of the intention set I_{A_j} corresponding to the different modal information of p_s. If the sum of the $P(y_n)$ of I_{A_j} is not less than 1, we need to normalize all intention probabilities in this mode, making $\sum P'(y_n) = 1$. If $P'(y_n) = P(y_n)$ when the sum of the $P(y_n)$ of I_{A_j} is less than 1, the information entropy H_{A_j} of this mode can be calculated by equation (2) based on the comprehensive probability $P'(y_n)$ of the intention set. The larger the information entropy is, the higher the information divergence is. In order to facilitate calculation and better represent the information quality, the system adopts $(2 - H_{A_j}) * P'(y_n)/\sum P'(y_n)$ as the data quality of modal information. The smaller the information entropy is, the clearer the expression of modal information is. For the convenience of calculation, we use $(2 - H_{A_j})$. $P'(y_n)/\sum P'(y_n)$ represents the proportion of useful information in all information of this modal expression. The purpose of this is to prevent false high information quality in this modal information. A large number of studies have found that the divergence of information plays a vital role in the accuracy of information [34]. Next, it is needed to obtain the recognition rate of single-modal information, such as the success recognition rate of speech information α_1. α_1 refers to the accuracy of speech mode in the current environment, which is the average of the accuracy obtained after multiple speech recognition tests in different scenarios. See Schedule 5 for the corresponding relationship between α_1 and location. In different environments, α_1 is different, so that the noise degree of the environment should be taken into comprehensive consideration. For this reason, experiments were conducted at different places. The corresponding relationship between α_1 and the place is as presented in Schedule 5. The recognition of gesture information is sensitive to environmental information. In other words, different brightness or clutter degree of the environment may affect the recognition rate of gesture, namely, each gesture in each environment has different recognition rate α_2. α_2 refers to the accuracy rate of gesture mode in the current environment, which is the average value of the accuracy rate obtained by multiple gesture recognition in different scenes. See Schedule 5 for the corresponding relationship between α_2 and location. Further, it is also needed to make clear the correctness of the intention extracted from different modes of information. The probability of correct intention extraction was represented by information quality. Based on the above parameters, equation (3) was inferred out for the trust degree evaluation of intention:

$$H_{A_j} = -\sum_{n=1}^{\text{length}\left(I_{A_j}\right)} P'(y_n) log_2\left(P'(y_n)\right), j = 1, 2, \quad (2)$$

$$p_e = E(p_s) = \left[\left(2 - H_{A_1}\right) * \frac{P'(y_n)}{\sum P'(y_n)} * \alpha_1 + \left(2 - H_{A_2}\right) * \frac{P'(y_n)}{\sum P'(y_n)} * \alpha_2\right] * P(y_n). \quad (3)$$

In the equation, p_e is the trust degree of the integrated intention p_s and E is the trust degree evaluation function. The trust degree of the integrated intention depends on the modal data quality H_{A_j}, the modal recognition rate α_j, and the probability of correct extraction of the intention $P(y_n)$ jointly. Finally, the probability of intention occurrence at certain time interval $P3_{y_n}$ (Schedule 4) was used to judge the correctness of the intention expressed by the user. For example, if the user's speech or gesture modal expression is clearer, the smaller H_{A_j} is, the larger $(2 - H_{A_j})$ is. If in the intention set expressed by speech or gesture, the greater the comprehensive probability of an intention in this mode accounts for the sum of the comprehensive probabilities of all intentions in this modal intention set, the greater is $P'(y_n)/\sum P'(y_n)$; if the credibility of gesture or voice is higher, α_j is larger; A is positively correlated with B, C, and D, indicating that the clearer the modal information, the greater the sum of intention synthesis probability and intention set synthesis probability, and the higher the reliability of modal identification, the greater the trust degree *Pe*, and vice versa. *Pe* is positively correlated with $(2 - H_{A_j})$, $P'(y_n)/\sum P'(y_n)$ and α_j, indicating that the clearer the modal information is, the greater the sum of intention synthesis probability and intention set synthesis probability,

and the higher the reliability of modal identification, the greater the trust degree p_e, and vice versa.

After the intention trust degree p is obtained, it is first compared with the set threshold to determine whether the intention can be executed when the intention trust degree is p, that is, whether the execution of this intention will pose a threat to the user's health or quality of life. The u mentioned below refers to the trust degree threshold of the integrated intention; u_{p1} is the trust degree threshold of $P1_{y_n}^{L_i}$; u_{p2} is the trust degree threshold of $P2_{y_n}^{t}$; u_{p3} is the trust degree threshold of $P3_{y_n}$; and u_{ei} is the trust degree threshold of mode i. Here, the mentioned thresholds are expressed in average value because when the trust degree is greater than the average value, the probability of obtaining real intention tends to be credible. u_{p1} refers to the average occurrence probability of y_n at all places, u_{p2} refers to the average occurrence probability of y_n at all times, and u_{p3} refers to the shortest standard occurrence time interval of y_n. In the process of trust degree evaluation, if $P3$ was 1, u_{p3} was also 1 when calculating the threshold. If $P3$ was not 1, for the sake of safety, 6 h was used as the shortest standard interval probability; the maximum threshold of the modal entropy was defined as the average entropy when the $P'(y_n)$ of the obtained intention was 0.5 and that of the other intentions was the residual remaining probability if there were n intentions in the I_{A_j}, or the entropy when the $P'(y_n)$ of the obtained intention was 0.5 if there was one intention in the I_{A_j}. u_{ej} is the product that 2 subtracts the maximum entropy of the mode and then multiplies the average correctness of mode j (it is expressed in the same way as the data quality of the modal information A_j); and u was calculated as per the said threshold in the same algorithm as p_e.

If $p_e \geq u$, it indicated that the real intention was successfully extracted; if $p_e < u$, the system will analyze the cause step by step. As we mentioned above, the clarity of modal information, the proportion of intention comprehensive probability to the sum of intention set comprehensive probability, and the reliability of modal identification are all related to the size of trust p_e. In the case of $p_e < u$, the system will gradually analyze the above relevant parameters and actively find the reasons for improvement. So when $p_e < u$, the system reversely analyzes the single-mode information, namely, the robot would trace the substandard intention back to each single mode of intention branch before intention integration and calculate the trust degrees e_1 and e_2 of each modal branch of this intention as per the following equation:

$$e_j = \left(2 - H_{A_j}\right) * \frac{P'(y_n)}{\sum P'(y_n)} * \alpha_j,\ j = 1, 2. \tag{4}$$

If $e_j < u_{ej}$, this situation is generally caused by the unclear modal information input by the user. In view of this situation, the system will actively ask the user to re-input the corresponding modal information as enhanced information and then the system would repeatedly make the multimodal integration and trust degree evaluation based on the enhanced information and the previously up-to-standard modal information till $e_j \geq u_{ej}$. If $e_j \geq u_{ej}$, but $p_e < u$, This situation shows that there is no problem with the information entered by the user, but that the time factor or place factor restricts the trust degree of the intention. At this time, we should discuss it in two cases: (1) if $P1_{y_n}^{L_i} < u_{p1}$, the system would prompt the user that the intention was not feasible in the present scene, let p_e be 0 and end the interaction task; (2) if $P1_{y_n}^{L_i} \geq u_{p1}$ and $P2_{y_n}^{t} \leq u_{p2}$, this shows that the time factor leads to low trust degree of intention. The system will actively ask the user whether to adhere to the current intention. If the user adheres to the current intention, p_e will be directly set to u. If the user abandons the current intention, p_e will be directly set to 0, and the user's intention will be asked again. Until $p_e \geq u$, the real intention p_f is obtained.

Then, the p_f was analyzed to see whether $P3_{p_f}$ reached the standard or not. If $P3_{p_f} < u_{p3}$, the intention expressed by user might be wrong and the system would re-ask the user to confirm whether to execute this intention or not; if not, the system would end this task; if yes or $P3_{p_f} \geq u_{p3}$, the system would execute p_f.

We summarize the trust evaluation and reverse active interaction process of the system as formula (5), where S represents the system's active feedback of information, which belongs to the robot's feedback action in the above three situations.

$$S = \left(p_e,\ u, u_{p1}, u_{p2}, u_e, P3_{y_n}\right). \tag{5}$$

If the real intention was not successfully extracted (i.e., I'_{A1A2} was empty), "$I'_{A1A2} = \varnothing$" might be caused by the following cases: first, I_{A1}' or I_{A2}' caused empty $I'_{A1} \cap I'_{A2}$ due to the empty i_{A3}; second, I_{A1}' or I_{A2}' was empty, but the empty $I'_{A1} \cap I'_{A2}$ was not caused by the empty i_{A3}; and third, I_{A1}' and I_{A2}' were both not empty, but $I'_{A1} \cap I'_{A2}$ was empty.

The cause of the first case might be that the user could not complete the expressed intention in this state. At this time, the system would remind that the user intention could not be conducted in this state and actively ask the user to change the state. The cause of the second case might be that the system failed to successfully extract the information of a mode. In this case, the system would actively ask the user to re-enter the modal information with empty intention set and then re-extract the integrated intention p_s. The cause of the third case might be that the user expressed inconsistent speech intention and gesture intention in the input, or the system made serious mistake in the extraction of single-modal intention. In this case, to ensure the correctness of the extracted intention, the system would ask the user to re-enter the speech information and gesture information and then integrate and extract the intention p_s again.

Table 2 shows a more intuitive description of the entire process of the algorithm.

4.2. Algorithm Analysis. Based on the algorithm proposed in this paper, a behavior recognition model for interaction between the elderly and robot was established taking into account the declined language expression ability and memory, the multimodal intention expression manner and other features of the elderly as well as the time and scene of human-computer interaction. This model solves the

Table 2: Multimodal intention integration algorithm.

The algorithm for reverse active integration of multimodal intentions
While(1):
While(1):
Modal information $A_j\,(j = 1, 2, 3, 4) \longrightarrow i_{A1}, i_{A2}, i_{A3}, (L_i, t, G)$
$i_{A1}, i_{A2} \longrightarrow^{i_{A3}} I_{A1}, I_{A2} \longrightarrow^{L_i,t} I^{'}_{A1}, I^{'}_{A2} \longrightarrow^{G} I^{'}_{A1A2}$ //extract speech and gesture intentions from the speech and gesture information as per the posture information and then integrate the two-modal intentions based on the environment information
IF: $I^{'}_{A1A2} != \varnothing$: $p_s = \varphi[\max(t_{now} - t^{y_n}_{before} - t^{y_n})]$, $y_n \in I^{'}_{A1A2}$; break; //if the integrated intention set is not empty, obtain the most possible intention p_s based on the time difference-intention probability
ELSE: analyze the cause of empty set, actively obtain enhanced information to cover the original A_j, and continue; //the system obtains the enhanced information as per the cause of empty set of the integrated intention
$H_{A_j} = -\sum_{n=1}^{\text{length}(I_{A_j})} P'(y_n) log_2(P'(y_n))$, $L_i \longrightarrow \alpha_j$ (check the corresponding table); //calculate the entropy of single-modal information by using the improved method for calculation of information entropy
$p_e = E(p_s) = [(2 - H_{A_1}) * (P'(y_n)/\sum P'(y_n)) * \alpha_1 + (2 - H_{A_2}) * (P'(y_n)/\sum P'(y_n)) * \alpha_2] * P(y_n)$
$S = (p_e, u, u_{p1}, u_{p2}, u_e, , P3_{y_n})$ //obtain the system's feedback action S on the basis of the integrated intention, the time difference, and the trust degree threshold if($S == \varnothing$): $p_e = 0$; break;
Else: execute S; continue

troublesome problems commonly encountered by elderly escort. Among the existing intention understanding models with good effect, the slot-gated intention understanding model [35], multimodal "human-computer integrated" cooperation system [27], Bayesian context-based intention understanding model [22], and the aforementioned smart home assistance system [8] all adopt single-modal intention information input and have good effect in specific scenes, but they cannot serve the elderly well. For example, the Bayesian context-based intention understanding model [22] mainly adopts the contextual information of user's action. According to Bayesian probability, the model can calculate the user's next possible intention and proactively ask the user whether he needs help or not so that the robot can know about the user's possible intention in advance, avoiding the time to wait for intention understanding of the robot in interaction. However, the contextual information used in this method is static and the smooth interaction in this method depends on specific action order. If dynamic contextual information is used, it will take longer time to predict the intention. Moreover, this method cannot avoid wrong intention. In contrast, in the method proposed in this paper, the environment information is used to obtain the feasibility parameter of user's intention, which breaks the limit to a static context. Furthermore, the YOLOv3 used for environmental perception takes both time and accuracy into account to the extent that the scene is recognized accurately with the least time; and the environmental perception contributes greatly to the accurate extraction of intention [37]. This way not only improves the accuracy in intention extraction but also avoids wrong intention expressed by user. Besides, the way of real intention extraction based on reverse active integration of multimodal information and the repeated trust degree evaluation mainly solves the problem that the elderly's real intention cannot be normally extracted due to vague expression. Meanwhile, making clear the feasibility of the intention by interaction helps improve the smoothness in human-computer interaction, as well as user's experience and the accuracy in intention extraction.

5. Experimental Results and Analysis

To verify the effectiveness of MES, interaction experiments were made on this system in combination with a Pepper robot.

5.1. Experimental Setting. The experiment was conducted on an HP Pavilion Gaming Laptop 15 equipped with a 2.20 GHz Intel Core i7 CPU, 1920 p× 1080p display, and Intel(R) UHD Graphics 630 GPU. The experimental robot was a humanoid smart robot Pepper developed by SoftBank Robotics, with algorithm body programmed in Python language and communication realized by Python 3.7, Python 2.7, and Visual Studio 2015. Pepper was connected to the computer to execute all functions of this system. The connection diagram of system equipment is shown in Figure 3.

5.2. Feasibility Study. During the interaction, this system mainly realized the scene perception, the reverse active interaction, the extraction of integrated intention, and the avoidance of wrong intention and guided the robot to complete the corresponding escort tasks. As shown in Figure 4, the scene perception function of this system improves the accuracy and speed of intention extraction, the reverse active interaction of the robot makes human-computer interaction more harmonious, and the intention integration function and trust degree evaluation mechanism contribute to more accurate extraction of the intention by robot. According to the intention extracted in this way, the robot can successfully complete the elderly escort work.

5.3. Experimental Design and Results. In this research, the experimental subjects were 10 users (average age: 75 years old) including 4 males and 6 females who lived in high-rise residential buildings in urban communities alone during the day time (living together with their children who however needed to work during the day time) or in long-term (not living together with their children). The experiments were mainly conducted in the simulated living room and

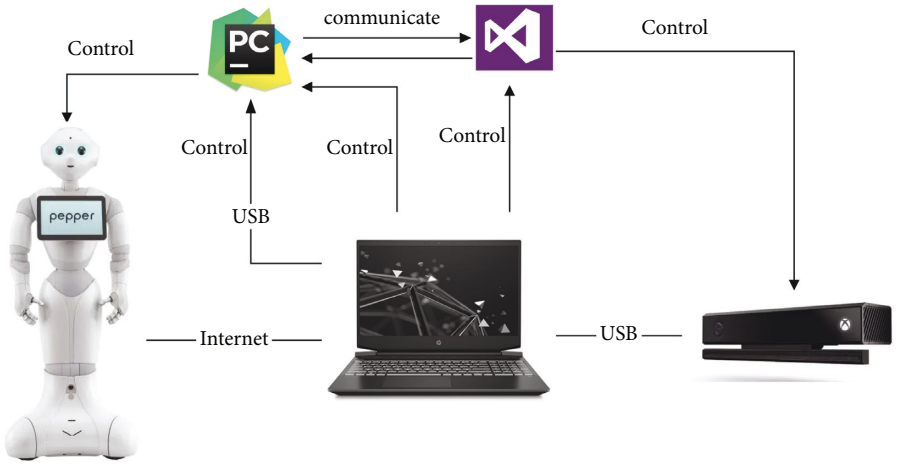

Figure 3: MES system equipment connection diagram.

Figure 4: Robot's cooperation with the elderly.

bedroom of a laboratory. Each user was tested at each place 5 times. Hence, we obtained 100 ($5 \times 2 \times 10$) groups of interaction data finally. All experiments are not completed in the same time period on the same day.

5.3.1. Experimental Protocol. In the experiment of this paper, we have some specific regulations in the design of interactive tasks, which need to be followed by each user. We list them as experimental protocols, which need to be read in detail before the user carries out the experiment, so that the experiment can be carried out smoothly. The experimental agreement is as follows:

(1) Wake up the MES system: each user uses the speech command "start" to wake up the MES system when initiating an interactive task.

(2) Scene recognition: after waking up the MES system, the user can use the speech command "scene recognition" within 3 seconds to let the robot determine its environment. The Pepper robot will rotate at a uniform speed for one week to obtain the surrounding environment information. This process takes about 5 seconds. If the user does not use the voice command "scene recognition," the scene information of the last interaction will be used, so this process is not time-consuming.

(3) Main task interaction: after completing (1) or (2), the user can input the intention he wants to express within 3 seconds. If the user does not send any speech command within 3 seconds to the robot, the robot will actively ask the user what help he needs. The user must express his intention within 3 seconds after the robot asks, that is, input three-modal information within 3 seconds after the robot asks. The robot analyzes the intention and calculates its trust degree according to the information input by the user. For cases that do not meet the trust degree requirements, the robot will actively interact with the user and ask for enhanced information. The user needs to complete the input of enhanced information within 3 seconds according to the requirements of the robot. Until the extracted intention trust meets the requirements, the robot will perform the corresponding task and the interaction ends.

(4) Terminate interaction: during the interaction between the user and the robot, the user can use the speech command "end" to terminate this interaction at any time.

5.3.2. Experimental Design. In this experiment, the system is embedded in the Pepper robot. The task of the robot in the experiment is to correctly identify the environmental information and the intention of the user input, evaluate and reverse analyze the fused intention, actively correct the user's wrong intention, and help the user complete the task. In the process of interaction, we hope that the robot can correctly and actively extract the user's expressed and feasible intention and help the user complete the task corresponding to

Figure 5: Photos of the furnishings in the simulated living room (parts 1 and 2) and bedroom (parts 3 and 4).

the intention. Therefore, this experiment mainly focused on the following aspects: (1) the accuracy in multimodal intention recognition, (2) the effect of the mechanism for evaluating the trust degree of the intention, (3) the participation equality in the human-computer interaction, and (4) the rate of wrong intention avoidance. In each interaction experiment, we will record whether the intention recognition is correct, the evaluation results of intention trust, the number of active interactions initiated by users and robots, and whether there is wrong intention avoidance.

As presented in Figure 5, in the laboratory, there are two simulated scenes (i.e., living room and bedroom) of the elderly as at home. The simulated brightness of the laboratory is provided by natural light and fluorescent lamps. The simulated ambient noise volume is the daily noise volume (≤60 db). There are TV, remote control, sofa, chair, tea table, cups, tea canister, and green plants in the simulated living room; there are beds, bedside tables, cups, green plants, and medicine bottles in the simulated bedroom. The scene recognition algorithm proposed in this paper can not only be extended to specific places where the elderly live, such as kitchens and study, but also can be used in many fields, such as urban nursing homes, social services, and so on. For the convenience of the experiment, the intention that the system can recognize in this experiment is set as follows: drinking water, taking medicine, drinking tea, eating fruit, eating health products, watching TV, reading books, and playing phone. Of course, this system can also add or delete recognizable intentions at will as per the needs of the elderly, so as to better serve the elderly. During the experiment, the user could interact with the robot by various combinations of speech, gesture, and posture, and the robot would give corresponding feedback based on the acquired information.

10 users need to complete five interactive experiments in the living room and bedroom simulated in the laboratory. After initiating the interaction, each user selectively commands the robot to perceive the scene. The user inputs the information of speech and gesture at will. The robot actively obtains the posture information and environmental information and extracts the intention and evaluates the trust degree according to the input information. According to the evaluation results, the robot acquires enhanced information reversely and actively to further improve the accuracy of intention and interaction effect. The robot will not execute the intention task until the intention with the required degree of trust is extracted.

5.3.3. Experimental Result. According to the above experimental design, we obtained 100 experimental data. Among the 100 interactive experiments, we randomly selected 3 experiments and took these 3 experiments as examples to further experience the functions of the MES system, as shown in Tables 3–5. According to these three experiments, we found that when the user interacts with the robot, the robot can point out the wrong intention that the user may express according to the environmental information, so as to achieve the effect of avoiding mistakes. In the same experiment, the user can change the intention that he wants to express, which well reflects the intelligence of the robot. Moreover, the number of interactions between users and Pepper in the same interaction task is uncertain, which is caused by the trust evaluation mechanism, while some systems [36] only have one interaction process: the user sends out a command, and the robot starts to execute the task after obtaining the intention. When the robot fails to obtain the intention, it will not react. At this time, the user needs to reissue the command. Although such a system reduces the consumption of interaction time, it increases the interaction load, and the accuracy of the extracted intention is low. It does not judge the correctness of the intention, and the user experience is not good.

5.3.4. Interaction Design. After the elderly starts an interaction, the robot will perceive the scene and ask the user needs. Then, the user inputs the speech and gesture information. The robot extracts the intention from the input information, evaluates the trust degree of the intention, and then selectively, reversely, and actively obtains enhanced information as per the evaluation result to further improve the intention extraction accuracy and interaction effect. In the interaction experiments, the users were asked to interact with the robot at a specified place and time. Tables 2–4 show the interactive dialogue between user and the robot in two experiments randomly selected from the 100 experiments. As can be discovered from the three tables, during the user-robot interaction, the robot can point out possible wrong intention expressed by the user based on the environment information to avoid error. Moreover, in the same experiment, user can convert the intention, which well reflects the intelligence of the robot.

In the same interaction task, the number of interactions between user and Pepper is uncertain due to the trust degree evaluation mechanism. Some systems [36] only perform one interaction: namely, after receiving the instruction of user, the robot starts to obtain the intention and execute it; if no intention is obtained, the robot will not react; at this time, the user needs to give the instruction again. Although such a system reduces the interaction time, the interaction load is

Table 3: Interaction task 1 (task no. 31).

Interaction experiment (place: living room; time: 4:00 pm; user no: 3)			
Interaction no.	Integrated intention	Start	
1:	Drinking tea	User:	I want to drink tea (posture: sitting on the chair; gesture: making a fist and shaking it)
		Pepper:	(Reversely and actively ask) I am not sure about your intention because the gesture information is substandard. Please re-enter the gesture information
2:	Drinking tea	User:	(Gesture: acting like grasping tea)
		Pepper:	(Extract the intention of drinking tea and reversely and actively ask) the evening sleeping quality may be affected if you drink tea in the afternoon. Are you sure you want to drink tea now?
3:	Drinking tea	User:	Yes, now
		Pepper:	(Execute the task of "taking a cup, tea box, and thermos" as no articles relevant to drinking tea were found in front of the elderly in previous target detection)
		End	

Table 4: Interaction task 2 (task no. 39).

Interaction experiment (place: living room; time: 5:00 pm; user no: 3)			
Interaction no.	Integrated intention	Start	
1:	Empty	User:	I want to drink something (gesture: acting like grasping and shaking hand in front of the mouth; posture: standing)
		Pepper:	(Reversely and actively say) do not hurry, please sit down
2:	Drinking water	User:	(Posture: sitting on the sofa)
		Pepper:	(Extract the intention of drinking water and ask) do you want to drink water? (as cups and tea box were found in front of the elderly in the early target detection)
3:	Drinking tea	User:	I prefer to drink tea. (Gesture: acting like grasping tea)
		Pepper:	(Extract the intention of drinking tea, and reversely and actively ask) the evening sleeping quality may be affected if you drink tea in the afternoon. Are you sure you want to drink tea now?
4:	Drinking tea	User:	Now
		Pepper:	(Evaluate the trust degree; if $pe < u$, reversely and actively ask) you have just made tea, are not you? Do you want to make tea again?
		User:	Oh, yes. I forget it. Don't make tea again
		End	

Table 5: Interaction task 3 (task no. 78).

Interaction experiment (place: bedroom; time: 10:00 am; user no: 7)			
Interaction no.	Integrated intention	Start	
1:	Taking medicine	User:	I feel a little uncomfortable (posture: sitting on the bed; gesture: acting like grasping a medicine bottle)
		Pepper:	(Reversely and actively ask) it's not the time to take medicine. Are you sure you want to take medicine now?
2:	Watching TV	User:	(Posture: sitting on bed) no, I want to watch TV (gesture: swinging the thumb up and down)
		Pepper:	As extracted by the system, you want to watch TV. (Reversely and actively say) there is no TV in the bedroom, please go to the living room to watch TV
		End	

increased, the accuracy of the extracted intentions is low, the correctness of the extracted intention is not judged, and the user experience is also not good.

Table 6 displays the intentions and their trust degrees extracted from the above three interaction tasks. In the table, N represents the task number; IN represents the number of interactions (interaction no.); Ex represents whether to execute or not; Av represents whether wrong intention is avoided; under the item of integrated intention, tea means drinking tea, wat means drinking water, med means taking medicine, and TV means watching TV; and the trust degree marked red means that this trust degree fails to reach the threshold. As can be seen from the table, in the same interaction task, the intention may change with the increase in number of interactions; and with the increase in the number of interactions, the trust degree of the same intention may be increased, and the feasibility threshold may also become 0, which depends on whether the enhanced information

Table 6: Change in integrated intention and the trust degree of the intention.

N	IN	p_s	p_e	u	e_1	u_{e1}	e_2	u_{e2}	$P1^{Li}_{yn}$	u_{p1}	$P2^{t}_{yn}$	u_{p2}	$P3_{yn}$	u_{p3}	Ex	Av
31	1	tea	0.146	0.268	1.519	1.45	0.099	0.162	-	-	-	-	-	-	No	Yes
	2	tea	0.287	0.477	1.519	1.395	1.67	1.485	0.9	0.6	0.1	0.276	-	-	No	
	3	tea	0.477	0.477	-	-	-	-	-	-	-	-	1	1	Yes	
39	1	ø	-	-	-	-	-	-	-	-	-	-	-	-	No	Yes
	2	wat	0	1.294	-	-	-	-	-	-	-	-	-	-	No	
	3	tea	0.287	0.477	1.519	1.395	1.67	1.485	0.9	0.6	0.1	0.276	-	-	No	
	4	tea	0.477	0.477	-	-	-	-	-	-	-	-	0.1	0.6	No	
78	1	med	0.83	0.82	-	-	-	-	-	-	-	-	0.3	0.5	No	
	2	TV	0.197	0.518	1.62	1.395	1.67	1.485	0.1	0.5	-	-	-	-	No	

extracted by the robot is conducive to improving the trust degree or not. In the three interaction tasks, the feasible intentions extracted were accurate and wrong intentions expressed due to poor memory were avoided, which should attribute to the trust degree evaluation mechanism. For example, in the first interaction of task 31, the total trust degree p_e failed to meet the requirements. In response, the system reversely analyzed e_1 and e_2 and found that the trust degree of e_2 was substandard, so the second interaction was performed. After user re-entered the enhanced information, the system found that p_e was still not up to the standard. By reversed analysis, it was found that the trust degrees of the modal information were up to the standard; by analyzing $P1^{L_i}_{y_n}$ and $P2^{t}_{y_n}$, it was discovered that the trust degree of $P1^{L_i}_{y_n}$ was too low. Hence, the third interaction was conducted. The robot actively asked the user whether to continue the task or not; the answer was yes, so the system got the trust degree of the intention set as 0.477 and executed the intention. In the three interaction experiments, the parameter of trust degree well reflects the feature that the robot repeatedly makes reverse analysis and actively asks enhanced information, which contributes to improving the accuracy and trust degree of the extracted intention and effectively avoids wrong intention expressed by the user. The data in the table intuitively reflect the flexibility and reliability of the system; and the trust degree evaluation mechanism makes the intention extraction more accurate, the interaction more harmonious, and improves the intelligence of the elderly escort robot.

5.3.5. User Comment. Finally, the participants were asked to score their interaction experience. NASA TLX was used to make an overall cognitive evaluation after the participants fully used the entire system. In this research, 100 experiments were conducted, and each of them was a whole experience of the entire system. Upon finishing each experiment, the user was asked to score; then, the average of the 100 scores was treated as the user comment, with result

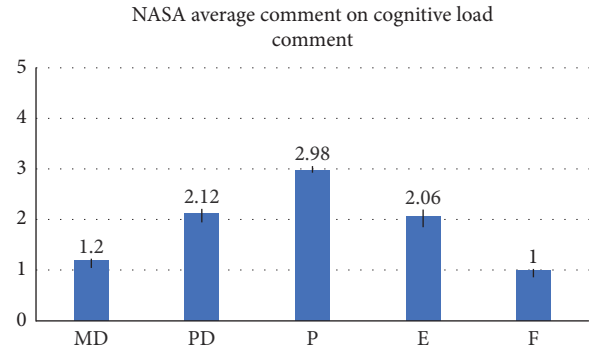

Figure 6: Average scores of the interaction effect.

as shown in Figure 6 and standard deviation as presented in Table 7.

User comment indicators include mental demand (MD), physical demand (PD), operational performance (P), effort (E), and frustration (F). Among them, MD describes the user's performance memory load, PD describes the user's effort made in the operation, P describes the smoothness in the user's operation (the smoother the operation is, the less the user's operation burden is), E describes whether the user feels easy to operate or not, and F describes the negative degree produced in the user's operation. The NASA evaluation indicators adopted a 5-point system. Each indicator was divided into 5 levels; the score within 0-1 point indicates a tiny cognitive load, within 1-2 points indicates a small cognitive load, within 2-3 points indicates a general cognitive load, within 3-4 points indicates a large cognitive load, and within 4-5 points indicates a quite large cognitive load.

5.4. Evaluation Criteria. The accuracy in intention extraction, the participation equality in human-computer interaction, the rate of wrong intention avoidance, and the user's subjective comment are used as the evaluation criteria of the system. The accuracy in intention extraction refers to the

Table 7: Score variance in user's interaction effect.

	MD	PD	P	E	F
Mean ± standard deviation ($\overline{x} \pm \sigma$)	1.20 ± 0.52	2.12 ± 0.6	2.98 ± 0.36	2.06 ± 1.03	0.87 ± 0.48

Table 8: Performance evaluation on the MES.

Performance evaluation on the MES				
Evaluation criteria	Accuracy in intention extraction	Participation equality in human-computer interaction	Rate of wrong intention avoidance	User's subjective comment
Experimental result	97.2%	12.6	100%	16

ratio of the number of intentions correctly extracted by the system to the number of all intentions expressed by the user in the experiment. The participation equality in human-computer interaction is represented by ten times of the active interaction ratio. The active interaction ratio refers to the ratio of the number of robot's active interactions to the number of human's active interactions; the closer the ratio is to 10, the more equal the participation in the human-computer interaction is; if the ratio is greater than 10, it indicates that the robot is more active; if the ratio is smaller than 10, it reveals that the user is more active. The rate of wrong intention avoidance refers to the rate that the wrong intention expressed by user is successfully avoided by the robot. The user's subjective comment is obtained by reducing the total score of NASA comment load (25 points) with the sum of the average values of such five items of interaction effect as scored by the user; the larger the value is, the better the comment is.

5.5. Result Analysis

5.5.1. Result Evaluation. The above evaluation criteria were used to evaluate the experimental results. In the 100 experiments, the 10 participants expressed intention 145 times in total (multiple intentions might be expressed in one experiment), and the system successfully extracted the participants' intention 141 times, with an accuracy of 97.2%; totally, 205 interactions were conducted actively by users and 260 interactions by the robot; hence, the participation equality in human-computer interaction was 12.6. Among the 145 intention expressions of the 10 participants, totally 36 wrong intentions were expressed and successfully avoided by the system, reaching a wrong intention avoidance rate of 100%. According to the above table of user comment, the user's subjective evaluation score was 16. The performance evaluation results are listed in Table 8.

5.5.2. Comparative Experiment. In order to reflect the advantages of this system in performance, a comparison was made between this system and four human-computer interaction-based intention understanding models selected as per the above evaluation criteria (i.e., multimodal "human-computer integrated" cooperation system [27], the slot-gated intention understanding model [35], Bayesian context-based intention understanding model [22], and the smart home assistance system [8]); then, the performance of this system was comprehensively analyzed from three perspectives: the accuracy in intention extraction, the participation equality in human-computer interaction, and the rate of wrong intention avoidance.

In the multimodal "human-computer integrated" cooperation system [27], user and the service robot can naturally communicate and retrieve information from the cooperation interface in multiple modes (e.g., head gestures and line of sight) in a manner of interactive dialogue. In this way, the service robot can fully understand human intentions and thus well collaborate to complete the corresponding task. The slot-gated intention understanding model [35] can extract key demonstrative verbs containing large core task information based on relationship analysis, construct such a function as combining key verbs with the contextual information, and provide a new dual-slot-gated mechanism for intention understanding. Based on Bayesian context-based intention understanding model [22], a framework was proposed in this research to improve the performance of an underlying intention recognition system by using the application of the object and the contextual information of the object's state form; this system used the digraphs extracted automatically from a large natural language corpus to represent the object and its application so as to understand the intention of user. The smart home assistance system [8] uses sensors to identify user's activity and further determine the user's intention. In this research, the said four systems were reproduced as per the methods stated in the corresponding literature in the same operating system environment as the MES and entered with information based on the features of the elderly; then, the four systems were tested 30 times, with the same intention as tested on the MES; after that, the same evaluation indicators and methods as the MES's were taken to evaluate the experimental results of the four systems, as shown in Table 9. Figure 7 can more intuitively compare the performance of various systems.

As discovered from the above comparative experiments, in case of obtaining vague input information, the four systems cannot extract the intention as accurately as the MES and do not have outstanding performance in avoiding wrong intention. In terms of interactive equality, the MES

Table 9: Comparative experiments on the accuracy in intention extraction, the participation equality in human-computer interaction, and the rate of wrong intention avoidance.

Method/evaluation criteria	Accuracy in intention extraction (%)	Participation equality in human-computer interaction	Rate of wrong intention avoidance (%)	User's subjective comment
Human-computer integrated [27]	82.5	5	0	12
Slot-gated [35]	87.0	5	0	14
Bayesian [22]	79.1	7	33.3	15
Smart home [8]	90.0	0	0	16
MES	97.0	12.6	100	16

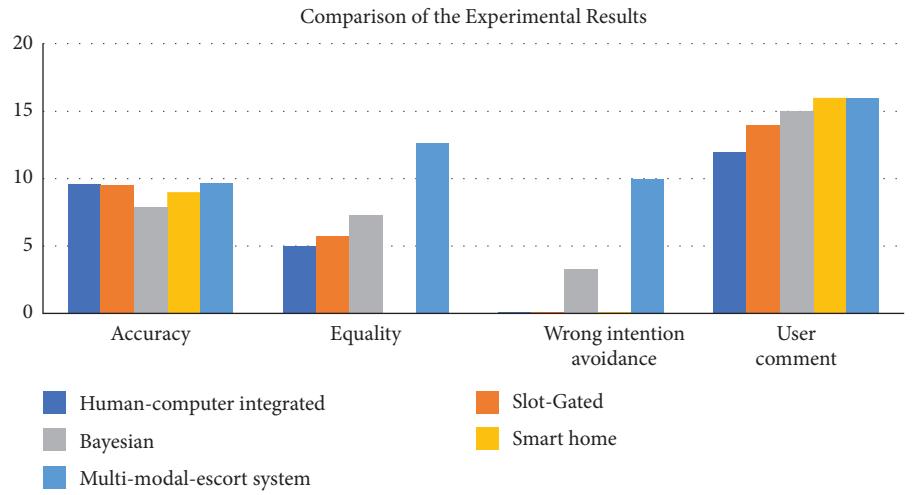

Figure 7: Comparison of the experimental results (bar graph).

Table 10: Intention extraction results under fuzzy expressions.

System	Human-computer integrated [27]	Slot-gated [35]	Bayesian [22]	Smart home [8]	MES
Consistent	30	22	26	16	44
Inconsistent	20	28	24	34	6
Correctness	60%	44%	52%	32%	88%

performs the best; the other systems also take interactive equality into consideration, but human is more active in the whole interaction. In regard to user comment, the elderly give low score to the model with input limitation; the elderly prefer to express their intentions habitually so as to reduce the interactive load.

In order to further highlight the robustness of this research work in intention extraction from fuzzy expressions, a special test was made on the correctness of intention extracted from fuzzy expression. In detail, some ambiguous expressions and 10 participants were selected; each participant was asked to express five intentions in the same form and manner as ambiguous under the 5 systems. Hence, each system experienced 50 interactions and obtained corresponding experimental data. Then, a comparison was made on the real intention of user and the intention extracted by the system; if the real intention was consistent with the extracted intention, it indicated that the intention extracted by the system was correct; otherwise, it indicated that the extracted intention was wrong. The results are as illustrated in Table 10.

In Table 10, "consistent" means that the intention expressed by the user is consistent with the intention extracted by the system; "inconsistent" means that the intention expressed by the user is inconsistent with the intention extracted by the system; and "correctness" means the ratio of the number of "consistent" results to the total number of results. The results in Table 10 were visualized as shown in Figure 8 to make it clearer.

According to the results of robustness experiments on the 5 systems and the 50 groups of experimental data obtained by each system in the same expression mode of participants, the MES performs more outstanding in processing fuzzy expression of user's intention; other systems can also process fuzzy intention expression but without that good effect, hence cannot be directly used by the elderly with declined intention expression ability. Above all, the MES proposed in this paper has better robustness than the other systems.

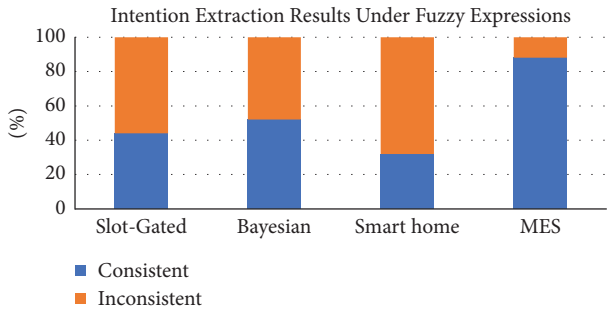

Figure 8: Visualization of intention extraction results under fuzzy expressions.

5.6. Limitations. This research also has some limitations. First, the human-computer cooperation still has low efficiency and sometimes needs multiple reverse input interactions between human and robot in order to have the robot understand the real intention of the user. Second, facial expression and other input modes of face emotion have not been taken into consideration and integrated into the system.

6. Conclusion

In this paper, the MES, a robot interactive system based on reverse active integration of multimodal intentions and scene perception, is proposed, making up for the key shortcomings of the existing helping elderly escort robots. The systems taking the declined expression ability and memory and other features of the elderly into consideration are rarely proposed in the existing research studies; and few researchers attach importance to the interaction environment. In contrast, the system proposed in this paper can perceive the scene before extracting the user's intention, which greatly improves the accuracy and success rate of the intention extraction. Then, the system can reversely and actively obtain multimodal information, integrate and extract the intention over multiple iterations, and further evaluate the trust degree of the extracted intention. This way solves a series of problems caused by the declined expression ability and memory and other features of the elderly. This system breaks the absolute instruction-based interaction mode in traditional human-computer interaction, giving robot the initiative. In general, the MES has encouraging research value and application scenes.

Data Availability

The data needed for this experiment can be obtained in the attached catalogue of the paper.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This paper was supported by Jinan Independent Innovation Team Project (No. 2019GXRC013).

Supplementary Materials

In the research of this subject, we need to investigate and obtain the experimental parameters. All the investigation and experimental parameters are legally obtained and have been unified by the investigators. The experimental parameters obtained from the investigation can be consulted from Schedule part. . (*Supplementary Materials*)

References

[1] Y. Tao and T. Wang, "Thoughts and suggestions on the research status and development trend of intelligent robot," *High tech communication*, vol. 29, no. 2, pp. 149–163, 2019.

[2] F. Yang, *Research on Contemporary Chinese Family Planning History doctoral Dissertation*, Zhejiang University, Hangzhou, 2004.

[3] D. Peng, "Centennial development trend of population aging in China," *Population Research*, no. 06, pp. 92–95, 2005.

[4] T. Guo, "A basic study on capacity and reaction time of visual working memory for elderly memory training," in *Proceedings of the 2019 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 600–604, Tianjin, China, 2019.

[5] G. X. Jing and W. Z. Liang, "Research on multi-modal interactive system of smart home for the elderly[J]," *Computer Science*, vol. 38, no. 11, pp. 216–219, 2011.

[6] G. Z. Yang, P. Dario, and D. Kragic, "Social robotics-Trust, learning, and social interaction," *Science Robotics*, vol. 3, no. 21, Article ID eaau8839, 2018.

[7] S. Mohammed, S. Shajideen, and V. H. Preetha, "Hand gestures - virtual mouse for human computer interaction," in *Proceedings of the 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 543–546, Tirunelveli, India, 2018.

[8] J. Rafferty, C. D. Nugent, J. Liu, and L. Chen, "From activity recognition to intention recognition for assisted living within smart homes," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 3, pp. 368–379, 2017.

[9] K. J. Jose and K. S. Lakshmi, "Joint slot filling and intent prediction for natural language understanding in frames dataset," in *Proceedings of the 2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 179–181, Coimbatore, India, 2018.

[10] J. Hatori, "Interactively picking real-world objects with unconstrained spoken language instructions," in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3774–3781, Brisbane, Australia, 2018.

[11] J. C. Peterson, D. Bourgin, T. L. Agrawal, D. D. Reichman, and M. Griffiths, "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, vol. 372, no. 6547, pp. 1209–1214, 2021.

[12] M. Gary and D. Ernest, "Artificial intelligence from the perspective of human brain [J]," *Journal of Computer Science*, vol. 02, no. 017, pp. 74–79, 2021.

[13] N. Fazeli, M. Oller, J. Wu, Z. Wu, J. B. Tenenbaum, and A Rodriguez, "See, feel, act: hierarchical learning for complex manipulation skills with multisensory fusion," *Science Robotics*, vol. 4, no. 26, Article ID eaav3123, 2019.

[14] A. Zlatintsi, "Multimodal signal processing and learning aspects of human-computer interaction for an assistive bathing robot," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3171–3175, Calgary, Canada, 2018.

[15] R. Zhao and K. Wang, "An immersive system with multimodal human-computer interaction," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 517–524, Xi'an, China, 2018.

[16] D. Kang, B. B. Kim, K. B. Choi et al., "Eyes are faster than hands: a soft wearable robot learns user intention from the egocentric view," *Science Robotics*, vol. 4, no. 26, Article ID eaav2949, 2019.

[17] C. I. Penaloza and S. Nishio, "BMI control of a third arm for multitasking," *Science Robotics*, vol. 3, no. 20, p. 1228, 2018.

[18] S. Gabriel, "Predicting and regulating participation equality in human-robot conversations: effects of age and gender," in *Proceedings of the 2017 12th ACM/IEEE International Conference on Human-computer interaction*, pp. 196–204, Vienna, Austria, 2017.

[19] M. Sun, W. He, L Zhang, and P. Wang, "Smart haproxy: a novel vibrotactile feedback prototype combining passive and active haptic in AR interaction," in *Proceedings of the 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 42–46, Beijing, China, 2019.

[20] S. N. Flesher, J. E. Downey, J. M. Weiss et al., "A brain-computer interface that evokes tactile sensations improves robotic arm control," *Science*, vol. 372, no. 6544, pp. 831–836, 2021.

[21] Z. K. Zheng, J. Zhu, J. Fan, and N. Sarkar, "Design and system validation of rassle: a novel active socially assistive robot for elderly with dementia," in *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 1–4, Nanjing, China, 2018.

[22] R. Tavakkoli, A. King, C. Ambardekar, A. Nicolescu, M. Nicolescu, and M. Nicolescu, "Context-based bayesian intent recognition," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 3, pp. 215–225, 2012.

[23] N. Li, G. Jin, F. Tian, G. Dai, and H. Wang, "ICOM~(DT): a dynamic task - oriented interactive computing model," *Journal of Software*, vol. 30, no. 10, pp. 2927–2941, 2019.

[24] G. Xu, L. Tao, and Y. Shi, "Human computer interaction in pervasive computing mode [J]," *Journal of Computer Science*, vol. 30, no. 007, pp. 1041–1053, 2007.

[25] Y Hou, Z. Feng, and T. Xu, "Decision making of mobile robot based on multimodal fusion," in *Proceedings of the 2020 the 6th International Conference on Computing and Data Engineering (ICCDE 2020)*, pp. 243–246, Association for Computing Machinery, New York, NY, USA, 2020.

[26] X. Qiu, Z. Yang, X. Tian, and J. Tian, "Multimodal fusion of speech and gesture recognition based on deep learning," *Journal of Physics: Conference Series*, vol. 1453, no. 1, Article ID 012092, 2020.

[27] K. Khalvati and S. A. Park, "Modeling other minds: Bayesian inference explains human choices in group decision-making [J]," *Science Advances*, vol. 5, no. 11, Article ID eaax8783, 2019.

[28] X. Zhou, *Research on Gesture Recognition Algorithm for Interactive Teaching interface[D]*, University of Jinan, Jinan, 2018.

[29] https://cloud.baidu.com/doc/SPEECH/index.html.

[30] Z. Feng, Bo Yang, T. Xu, H. Tang, and Na Lv, "Direct manipulation 3D human-computer interaction paradigm based on natural gesture tracking," *Journal of Computer Science*, vol. 37, no. 06, pp. 1309–1323, 2014.

[31] Z. Feng, "Research on flexible mapping interaction algorithm for multiple gestures corresponding to the same semantics," *Journal of Electronics*, vol. 47, no. 08, pp. 1612–1617, 2019.

[32] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, 2018.

[33] L XuJie, F ZhiQuan, and Y XiaoHui, "Research On Human-Robot Natural Interaction Algorithm Based On Body Potential Perceptions," in *Proceedings of the 2020 the 6th International Conference on Computing and Data Engineering (ICCDE 2020). New York, United States*, pp. 260–264, 2020.

[34] A. Kang, "Collaborative filtering algorithm based on trust and information entropy," in *Proceedings of the 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pp. 262–266, Bangkok, Thailand, 2018.

[35] S. Zhang, J. Jiang, Z. He, X. Zhao, and J. Fang, "A novel slot-gated model combined with a key verb context feature for task request understanding by service robots," *IEEE Access*, vol. 7, pp. 105937–105947, 2019.

[36] Q. Zhao, D. Tu, and S. Xu, "Natural human-computer interaction for elderly and disabled healthcare application[C]," in *Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine*, IEEE, 2015.

[37] A. Anderson, M. S. Homdee, N. Alam et al., "BESI: behavioral and environmental sensing and intervention for dementia caregiver empowerment-phases 1 and 2," *American Journal of Alzheimer's Disease and Other Dementias*, vol. 35, no. 3, p. 15, 2020.