International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

MIUIC: A Human-Computer Collaborative Multimodal Intention-Understanding Algorithm Incorporating Comfort Analysis

Liran Zhou, Zhiquan Feng, Hongyue Wang & Qingbei Guo

To cite this article: Liran Zhou, Zhiquan Feng, Hongyue Wang & Qingbei Guo (31 Aug 2023): MIUIC: A Human-Computer Collaborative Multimodal Intention-Understanding Algorithm Incorporating Comfort Analysis, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2023.2247606

To link to this article: https://doi.org/10.1080/10447318.2023.2247606

Published online: 31 Aug 2023.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2023.2247606

Taylor & Francis
Taylor & Francis Group

MIUIC: A Human-Computer Collaborative Multimodal Intention-Understanding Algorithm Incorporating Comfort Analysis

Liran Zhou, Zhiquan Feng, Hongyue Wang, and Qingbei Guo

School of Information Science and Engineering, University of Jinan, Jinan, China

ABSTRACT

The naturalness and safety of human-computer interaction have always been primary research focuses in the field of human-computer interaction. This paper proposes a multimodal intention understanding algorithm (MIUIC), which incorporates comfort analysis, as a solution to address the issues of low intention understanding rate, weak interaction, and weak collaboration that are often observed in most massage systems. The algorithm efficiently fuses multimodal data based on objective implicit information to address the challenge of low intention understanding rates caused by non-standard expression of natural behavior. Moreover, this algorithm incorporates comfort reasoning to detect and address intentions related to security threats while providing the ability for robots to make behavioral decisions through inverse active interaction, leading to more equitable human-robot interactions. To test the validity and safety of the MIUIC algorithm, we embedded the algorithm into a mechanical arm massage system. Subsequently, 45 elderly volunteers were invited to participate in experimental tests. Finally, to verify the validity and safety of the MIUIC algorithm, we assessed the algorithm in terms of four aspects, including multimodal intention recognition rate, the ability to reduce data dispersion, the intention enhancement rate under reverse human-machine interaction, and the rate of avoiding dangerous intentions. In conclusion, the MIUIC algorithm enhances the intention understanding rate and promotes.

KEYWORDS

Human-computer collaboration; nature human-computer interaction; intention understanding; multimodal fusion

1. Introduction

The continuing growth of the global aging population has become one of the major social issues, increasing the demand for smart health services (Carros et al., 2020). Besides, the Coronavirus-19 (COVID-19) pandemic has changed the normal living style of people (Haleem et al., 2020), leading to an increase in mental health problems, including depression, anxiety, self-harm and suicide (Holmes et al., 2020; Jiménez-Pavón et al., 2020). Massage is an ancient Chinese medical practice (Naruse & Moss, 2019; Wei et al., 2017), which has been proven to have various positive effects on mental and physical health and well-being (Naruse et al., 2020; Zhang et al., 2015). The physical effects of massage include pain relief and the psychological effects include stress, anxiety and fatigue reduction (Kozak et al., 2013; Yücel et al., 2020). However, for manual massage tasks, it requires a lot of labor cost, which prompts the research on artificial intelligence technology (Li et al., 2020; Si et al., 2019). Artificial intelligence has gradually become a new driver of global value chains (Awan et al., 2022; Awan et al., 2022), avoiding repetitive operations and increasing task execution rates through technological innovation and resource allocation optimisation. Therefore, the relaxed massage robot based on artificial intelligence has become a research hotspot.

Currently, there are two common types of intelligent massage systems available in the market: the roller massage chair and the humanoid massage robotic arm. The roller massage chair is deemed safe with limited interactivity, whereas the humanoid massage robotic arm offers high flexibility but typically requires assistance from personnel or auxiliary technologies, hindering the possibility of attaining the application objectives in domestic settings (Field, 2019). To ensure safe interaction in enabling the massage robotic arm to perform relaxation massage tasks autonomously has become paramount. In addition to hardware safety protection measures such as softness design of materials (Hirzinger et al., 2001), emergency button control, and strength control (Mewes & Mauser, 2003), intention inference based on behavior perception and data is also one of the important steps to guarantee safety (Zacharaki et al., 2020).Therefore, if a human-computer collaborative robot can accurately infer the intention conveyed by human behavior, it will greatly improve the safety of human-computer interaction, and operational efficiency, thus increasing the user's sense of safety (Kleinman et al., 1970).Peterson et al. proposed that people's decisions are based on multimodal inputs (Peterson et al., 2021). However, currently, most human-computer collaborative systems predominantly depend on speech or gestures to perceive inputs. deviates from the aim of natural interaction and can be challenging

CONACT Zhiquan Feng ✉ ise_fengzq@ujn.edu.cn School of Information Science and Engineering, University of Jinan, Jinan, China

for older adults with declining expressive abilities (Lang et al., 2023). The ambiguity in input data poses a challenge for systems to comprehend users' intentions. Furthermore, collaborative systems based on multimodal data currently lack sufficient integration of the data and do not fully apply the principle of equal interaction, rendering the practical application of the system challenging. Studies have demonstrated that intention recognition methods that utilize multimodal fusion can address the issue of ambiguous or missing input information, thus effectively reducing uncertainty in intention recognition (Ernst & Banks, 2002). Additionally, safe intention comprehension and collaboration can be achieved through repeated interaction (Hajarian et al., 2019).

This paper proposes a multi-modal intention understanding algorithm that integrates comfort analysis based on the principle of equal interaction, to address the issues of low intention understanding rate and weak collaboration in current massage systems. The algorithm comprises a multimodal intention understanding algorithm and a safety human-computer collaboration strategy. The multi-modal intention understanding algorithm undertakes modal correction by exploring the implicit information in multi-modal data and resolves the problem of low intention understanding caused by input ambiguity via effective data integration. The strategy of secure human-computer collaboration employs comfort analysis and reverse-active interaction to address issues related to weak system collaboration and decision-making authority. This strategy also enhances the rate of intention understanding.

The algorithm made three main contributions: (1) constructing a safe interaction framework that prioritizes natural interaction, (2) proposing an improved evidence theory-based multi-modal intention recognition algorithm to fully exploit the objective information hidden in multimodal data, and (3) introducing a safety collaboration strategy based on comfort detection, which empowers robots to make behavior decisions, thereby increasing intention recognition rates and achieving safety collaboration to a certain extent.

This paper is organized as follows: The second section introduces the research on human-computer interaction and multimodal intention perception. The third section introduces the principle and implementation of the proposed algorithm in this paper. The fourth section describes the experimental testing and evaluation based on the effect of the algorithm in this paper. In the fifth section, a discussion and analysis are carried out and future research directions are described. Finally, the sixth section is the conclusion.

2. Related work

It is known that intention perception and human-computer collaboration are the main parts that make up algorithms when studying human-computer interaction algorithms. This section predominantly examines and discusses current formats of human-computer collaboration, together with methods for intention recognition and multimodal data fusion.

2.1. Human-Computer interaction and Human-Computer collaboration

Human-computer collaboration is a subfield of human-computer interaction research (Y. Wang & Zhang, 2017) that has gained significant attention recently and is now being widely applied across a range of research areas, including manufacturing (Papanastasiou et al., 2019) and healthcare (Pineau et al., 2003). The primary focus of human-computer collaboration is the cooperative sharing of space and task delegation between humans and robots, emphasizing intuitive and predictable interactions (Callens et al., 2020). This field encompasses both theoretical and practical research, as well as design and evaluation, that is related to human-computer interactions (Czarnowski et al., 2018; Herath et al., 2018). To enhance collaboration between humans and computers, a multitude of communication modes has been extensively researched, as evidenced by multiple studies (Berg et al., 2019; Kim et al., 2019). These communication modalities primarily involve gestures (Gadekallu et al., 2022), speech (Dinh Le et al., 2019), eye movements (Trick et al., 2019), body postures (Dutta & Zielinska, 2019), and physiological signals EMG (Peternel et al., 2016) and EEG (Tan et al., 2021). The various communication modalities available have significantly improved robots' ability to comprehend environmental states and user information more accurately and comprehensively.

Currently, the majority of human-computer collaborative systems adopt the master-servant interaction mode whereby the human provides specific instructions and the robot complies with the instructions accordingly. This model has the potential issue of "blind obedience," which can deteriorate the quality and effectiveness of human-computer interaction (Skantze, 2017). The research goal of human-computer collaboration is to achieve safe coexistence and natural interaction. This requires robots to have a minimum level of autonomy and be capable of exhibiting initiative (Gervasi et al., 2020).

Some researchers have begun exploring ways to enable active interaction, such as Sun et al. who proposed a hybrid technique that integrates passive and active haptics to develop a haptic feedback model that can be used for AR assembly tasks (Sun et al., 2019). The model enhances the user's haptic perception of the virtual object through active vibrotactile feedback when they touch the object, in addition to passive haptics that convey the object's stiffness in the virtual scene. Zheng et al. (Zheng et al., 2018) asserted that conventional animal-like robots lack the ability to interact with humans in a basic touch-based way and proposed a new type of socially assistive robot, which is active and equipped with whole-body haptic sensing. The robot responds with appropriate feedback when stimulated by touch and prompts the user to touch specific areas in a particular order during the next interaction. Flesher et al. (Flesher et al., 2021) embedded their developed brain-computer interface-controlled prosthetic system into a paralyzed. The system incorporates afferent information from muscles, skin, and joints for bidirectional interaction. The system facilitated the patient's ability to generate tactile sensations by microstimulating the cortex, which effectively supplemented visual information and resulted in significant progress in the experimental outcomes. The experimental results in the aforementioned studies demonstrate that active interaction promotes positive human-computer interaction and improves user experience. Nevertheless, most robots' active feedback content relies heavily on human instruction, and their behavioral decision-making ability is relatively weak. In contrast, Kelley et al. (Kelley et al., 2012) utilized contextual information about the form of object revelation and object state to enhance the performance of a potential intention recognition system. Robots equipped with this system actively gather the user's state information, using it for contextual reference and to prompt if further help is needed. This study led to the development of equal interaction and serves as an inspiration for this paper's call for safe and intelligent human-computer collaboration.

In summary, current human-computer collaborative systems suffer from deficient collaboration, lack of system initiative, and ineffective multimodal intention understanding methods. This paper presents a method for intention safety analysis based on user comfort detection for human-computer collaboration in a home-based massage robot. The method enables the robot to take initiative to reverse the analysis of dangerous intentions and actively prompts the user for additional information. The proposed method emphasizes the goal of achieving equal interaction between humans and robots, enhancing the naturalness and fluency of interaction, and ensuring user safety.

2.2. Multimodal data fusion and intention recognition

The capability of machines to comprehend human intention empowers them to recognize human emotions, thereby enhancing goal achievement. The fundamental technique for recognizing intention entails selecting the appropriate output category from several categories through input information in a robotic system. This technique uses two primary implementation methods, namely traditional statistical recognition and neural network recognition.

Several statistical learning methods, including Bayesian (Khalvati et al., 2019), Random Forest (Sridhar et al., 2017), and Hidden Markov (Chang et al., 2021), can be utilized to address the intention recognition issue. Mi et al. (Mi et al., 2021) employed Bayesian probability to reconstruct an indistinct feature presentation of the target and efficaciously tackled the problem of understanding intention with incomplete data. Nonetheless, while Bayesian algorithms are suitable for smaller samples, they are prone to misclassification when the number of intentions rises. Chen et al. (L. Chen et al., 2020) classified high-dimensional emotional characteristics into distinct subclasses and employed multiple Random Forests to recognize varied emotional states in speech. Nonetheless, Random Forests are computationally expensive and time-consuming during training and exhibit poor recognition outcomes in the presence of noisier inputs. Chen et al. (Z. Chen et al., 2022) proposed an intention recognizer based on the Hidden Markov Model (HMM) for an autonomous driving system to capture the intention of other vehicles attempting to change lanes. The intention recognizer simulates the selective attention mechanism of the human visual system by incorporating the speed change of surrounding vehicles as a lane change cue to enhance its recognition performance. However, frequent HMM iterations can escalate the time consumed for recognition, and the strict adherence to Markov's assumption may decrease the algorithm's certainty when intention changes occur less frequently, leading to a lack of additional observation that provides greater accuracy (Ferguson et al., 2015). Although statistical probability-based models for intention recognition earlier mentioned have high interpretability, the requirement for initial probability choices can be susceptible to error and subjectivity.

With the advancement of technology, machine learning has demonstrated its superiority in processing multimodal data, such as speech and vision, and thereby offering new solutions for intention classification. Various artificial intelligence techniques, such as convolutional neural networks (CNNs) (Sugano et al., 2016) and recurrent neural networks (RNNs) (Singh et al., 2017), are increasingly applied in intention classification research. Multimodal data fusion, based on how the data is combined, can be categorized into two types: feature layer fusion and decision layer fusion.

Feature layer fusion is a type of multimodal data fusion that combines the pertinent features from each modality by mapping it into a high-dimensional feature vector using transformation algorithms, thereby increasing the accuracy of intention classification. An example of this is the improved single-shot multi-box detector (SSD algorithm) proposed by Bai et al. (Bai et al., 2022). In their study, they performed deep feature fusion between the target detection layer and its adjacent feature layers, resulting in improved object detection accuracy. Wang et al. (M. Wang et al., 2020) developed a bio-inspired data fusion architecture to recognize human gestures, integrating visual and somatosensory data. Their approach achieved better results than using either modality alone. The proposed model utilizes convolutional neural networks for visual processing and sparse neural networks for sensor data and visual data fusion to achieve robust recognition of a visual question-and-answer system for medical images (Sharma et al., 2021). In their study, they used ResNet-152 for image feature extraction, BERT for question feature extraction, and Multimodal Factorized Bilinear Pooling (MFB) to complete multimodal feature fusion. Additionally, they employed an attention-based multimodal deep learning model called MedFuseNet, which outperformed several state-of-the-art models.

Previous studies demonstrated the advantages of multimodal feature layer fusion. However, feature layer fusion may struggle to explain causality at the cognitive level. Additionally, adjusting network parameters becomes time-consuming and labor-intensive when the number of intentions changes. Therefore, many researchers focus on using multimodal fusion on the decision layer. This ensures differences between different attributes are highlighted, which allows researchers to use empirical and objective information for making adjustments (Jiang et al., 2020).

Ortega J et al. (Ortega et al., 2019) proposed a novel Deep Neural Network (DNN) that accurately predicts by learning and combining the representations of multiple modalities. The network encodes each modality independently using fully connected layers and then merges them into one fully connected layer with multimodal fusion completed in an end-to-end manner, achieving a high level of multimodal coordination. Dawar et al. (Dawar et al., 2019) implemented the use of a Convolutional Neural Network (CNN) to process visual input and a Long-Short Term Memory (LSTM) to process sensor input. The two output layers of these networks are multiplied by rank scores to achieve multimodal fusion results. While this approach resulted in promising fusion results, analysis of the differences and conflicts among modalities was neglected. Therefore, there is still room for improvement in the implementation of multimodal fusion.

After separately recognizing the audio, visual HD, and depth camera inputs, Rodomagoulakis et al. (Rodomagoulakis et al., 2016) utilized a decision layer weighted combination to recognize human actions by combining the scores obtained from each modality. Al-Amin et al. (2019) proposed a novel method of weighted fusion that utilizes five independent CNN models to recognize human activities through the analysis of behavioral characteristics and modal information. This approach enabled the connection between modal information and behavioral characteristics to be established, resulting in enhanced accuracy in human activity recognition. Yang et al. (Yang et al., 2019) proposed an integrated probabilistic inference approach that leverages spatial-semantic and spatial-temporal analysis to extract correlated features. They further used information entropy to fuse the results and enable the robot to infer the person's role and target in a task. The approach's efficacy lays in highlighting each modality's effects on the target in the fusion process; however, subjective weight settings in the decision-level fusion process present limitations. Analyzing the correlations and influences of modalities comprehensively can result in valuable and objective information that enhances the overall effectiveness beyond individual modalities, achieving the "$1+1>2$" effect (L. Zhang et al., 2018). Hence, it is suggested that during the fusion of modalities, objective data-reflected information should be taken into account. Objective information that is inherent in the data must be given due consideration during the modal fusion process to achieve greater objectivity.

In summary, multimodal fusion plays a pivotal role across various fields. However, most intention recognition methods overlook the objective information within the data during the fusion process. Currently, no feasible solutions exist for scenarios that involve ambiguous input or high safety requirements. Hence, this research presents a multimodal intention recognition algorithm that utilizes feature layer fusion, in conjunction with existing research methods and improved evidence theory. The integration of objective information mining, human-computer collaboration, and active interaction principles result in the proposed algorithm being highly adept at recognizing intention and avoiding hazardous situations. Additionally, the presented ideas of multimodal fusion and collaborative intention safety detection are crucial for home-based service robots. Overall, the algorithm contributes to the interpretability and robustness of algorithmic research in related fields.

3. Method

The proposed multimodal intention understanding algorithm adopts a four-part framework, as illustrated in Figure 1.

The framework consists of four components: information perception, intention recognition, intention executability analysis, and human-computer collaboration. The information perception component acquires and processes input data from each modality. The intention recognition component fuses the multimodal data to accurately extract

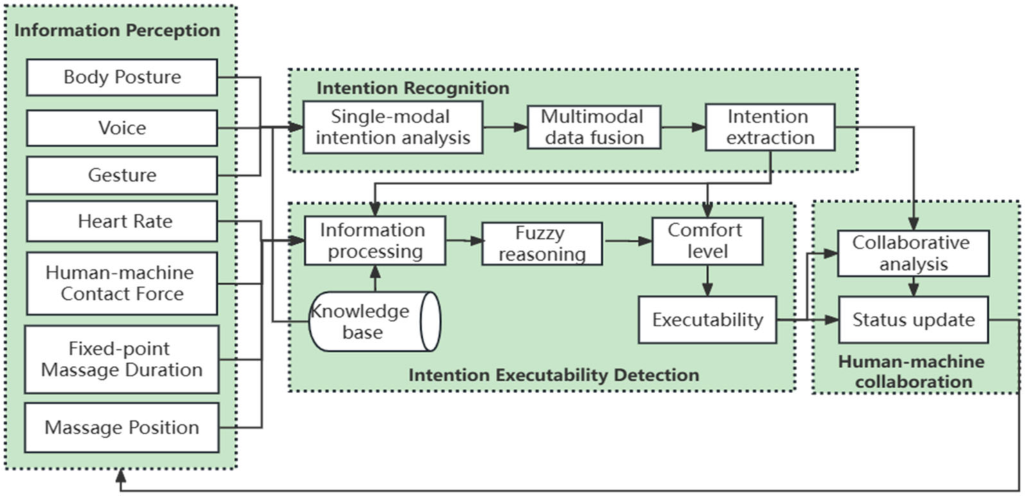

Figure 1. The Flowchart of the Overall Framework.

intention. The intention executability analysis component determines intention safety through user state detection. The human-computer collaboration component analyzes unmet intention by actively questioning and utilizing other interactive behaviors for providing information feedback. Among these four components, the intention recognition and intention executability analysis are the key elements. The following section explains the implementation details of each component.

3.1. Information perception and processing

This section assesses and analyzes the input information of different modalities using diverse techniques to derive an intention extraction rate for each modality. The purpose of this section is to furnish data support for multimodal fusion in intention recognition. In the experiment, the system continuously senses user voice input data, denoted as V, and visual input, including hand gesture data Ges and body gesture change data Pos. Baidu speech recognition technology and stuttering word division are utilized for the recognition of speech information, and the resulting intention probability set of speech modality, denoted as I_{Voice}, can be obtained. Body gesture information is recognized using Kinect bone point detection technology. This enables the determination of the intention probability set of the body posture modality, denoted as I_{Pos}. Gesture information is recognized using a gesture recognition system based on CNN network classification (Ferguson et al., 2015), enabling the determination of the intention probability set of the gesture modality, denoted as I_{Ges}.

Additionally, when a user initiates an interaction, the system real-time monitors their heart rate (HR), massage intensity (P), and massage duration (T) in a fixed position. The interactive devices of this massage system mainly consist of an Xarm robotic arm, Kinect visual perception device, voice input device, pressure sensing sensor, and a computing and processing device.

3.2. Multimodal intention recognition

Accurately identifying the implicit intention behind human behavior is fundamental to the collaboration between humans and machines, constituting the ultimate goal of their interaction. Given the differences in human backgrounds and cultural levels, input ambiguity is inescapable in natural interactions between humans and robots. Consequently, fully capitalizing on the advantages of multimodality and utilizing the objective information concealed in multimodal data to diminish ambiguity is critical for successful intention recognition. This paper proposes, based on this concept, a multimodal fusion method that adopts an enhanced evidence theory. The Dempster-Shafer (D-S) evidence theory has the ability to accommodate uncertain information, thus has extensive applications in information fusion, multi-criteria decision-making, and other disciplines. Nevertheless, when evidence is highly conflicting, the information fusion outcomes derived from the D-S evidence theory will encounter a paradox, impairing its practical application. Currently, the primary approaches to addressing this issue are enhancing combination rules and correcting evidence sources. This study focuses on adopting an evidence fusion method that employs an objective correction coefficient, with evidence source correction as the starting point. In general, if a modality is supported to a greater extent by other modalities, it is reasonable to believe that the information conveyed by that modality is more credible. Consequently, in the fusion process, modalities with high support can have their weight increased to enhance their impact on the fusion outcome. Likewise, if a modality generates conflicting data, its weight should be reduced to mitigate its influence. Based on this reasoning, a novel correction coefficient is established by comprehensively assessing modalities' self-credibility, their inter-modality credibility, and the degree of falsehood. This involves discounting the basic belief assignment function in order to minimize evidence conflict. Ultimately, evidence fusion and intention extraction are implemented via the Dempster combination rule, thus enhancing the implementation process of the evidence theory.

The information perception section generates a distribution of the intention probability set, which is then combined to create the evidence set, $\boldsymbol{Inp}$. The number of effective modalities in $\boldsymbol{Inp}$ is denoted by n.

The reliability of modal information is indicated by the degree of dispersion in the intention probability set. When probability aggregates, it suggests acceptable information conveyed by the modalities. Therefore, the entropy value of the intention probability set partially captures objective implicit information. This paper calculates the trustworthiness of each modality by computing the average entropy value of n intention probability sets, which serves as the basis for the calculation. After normalization, the outcome denotes the modality's trustworthiness, $Cred_{self}(\boldsymbol{Inp})$, as shown in Equation (1). Here, $\boldsymbol{Inp}$ represents the intention probability distribution of a specific modality.

$$Cred_{self}(Inp_i) = \frac{\sum_j^{num} Inp_i(Int_j)\log_2 Inp_i(Int_j)}{\sum_i^n \left(\sum_j^{num} Inp_i(Int_j)\log_2 Inp_i(Int_j)\right)} \tag{1}$$

Here, $Inp_i(Int_j)$ is the probability of extracting the jth intention conveyed by the ith effective modality. num represents the number of intentions in the intention set, which is taken as 6 in this system. And n represents the number of effective modalities in the current intention recognition process.

The study aims to calculate the confidence of modalities since the data from multiple modalities originates from the same user intention. It is imperative that a connection exists between these sets of data. Consequently, modality trend probabilities are calculated through cosine similarity measurements. The similarity between modality data sets serves as the measurement of modality confidence, denoted by $Cred_{bet}$ as shown in Equation (2). In this formula, Inp_i and Inp_j represent the probability distribution of user intention for the i-th and j-th modalities, and $Sim()$ measures cosine similarity.

$$Cred_{bet}(Inp_i) = \frac{\sum_{\substack{j=1; \\ j \neq i}}^{num} Sim(Inp_i, Inp_j)}{n-1} \tag{2}$$

The modality similarity confidence $Cred_{bet}(Inp)$ can only reflect similarities between modality pairs. As such, it is ineffective for measuring the impact of a single modality in the global context of all modalities. To address this limitation, this study proposes using modality influence as a weight factor for global conflict. The calculation method for global conflict Con_0 is adopted from evidence theory and is shown in Equation (3).

$$Con_0 = \sum_{\cap I = \emptyset} \prod_{k=1}^{n} Inp_k(I) \tag{3}$$

Here, I represents the intention set while Inp_k represents the corresponding set of intention probabilities for modality k.

When excluding the ith modal information, the new global conflict value Con_i can be obtained. This value represents the effect of the ith modal information on the global conflict, as shown in Equation (4).

$$Con_i = \sum_{\cap I = \emptyset} \prod_{\substack{k=1; \\ k \neq i}}^{n} Inp_k(I) \tag{4}$$

Therefore, based on Equations (3) and (4), the value of spuriousness of the ith modal, $F(Inp_i)$, can be calculated, as shown in Equation (5), which can be subsequently used as a component of the correction factor.

$$F(Inp_i) = \frac{Con_0 - Con_i}{1 - Con_i} \tag{5}$$

As per the definition, if $F(Inp_i) = 0$, it indicates that the ith modal information has no effect on the overall conflict. If $F(Inp_i) = 1$, it indicates that the ith modal information is in complete conflict with other modalities, and needs to be reduced when intending to fuse. If $0 < F(Inp_i) < 1$, it indicates that the ith modal information contributes to the overall conflict, and the more significant the contribution, the greater the degree of spuriousness.

The analysis above shows that the value of self-credibility for a single modal information, the similarity between modal information, and the global falsity of modal information do not affect each other. They are all objective information implied by the modal information. In this paper, correction factors $\aleph$ for each modal information can be obtained by combining these three factors and through normalization, as shown in Equation (6).

$$\aleph(Inp_i) = \frac{Cred_{self}(Inp_i) + Cred_{bet}(Inp_i) + 1 - F(Inp_i)}{\sum_{i=1}^{n} Cred_{self}(Inp_i) + Cred_{bet}(Inp_i) + 1 - F(Inp_i)} \tag{6}$$

Finally, the system employs the correction factor $\aleph$ to adjust the set of probabilities for the intended actions of each modal information, in order to obtain the intention probability set weighted average Inp_{Ave}.

$$Inp_{Ave} = \sum_{i=1}^{n} \aleph(Inp_i) Inp_i \tag{7}$$

Applying Dempster's rule with an Inp_{Ave} combination $n-1$ times leads to a final set of intention probabilities, $\boldsymbol{I_{Fin}}$, after fusion. Once the data fusion phase completes, the system proceeds to the intention extraction phase to extract the intention element with the highest probability in $\boldsymbol{I_{Fin}}$ as the user's intention. To make the recognition process more accurate, a minimum probability threshold, th, is manually set in this paper. If the intention probability in $\boldsymbol{I_{Fin}}(Int)$ is greater than th, it confirms the user's intended action and verifies its executability. Otherwise, the system enters the information enhancement phase, where Bayesian posterior probability inverse analysis is applied to identify the intention modality with the lowest contribution, Rep_{Inp}. The user is then prompted to use that modality to re-express the intention before re-performing intention fusion.

In conclusion, this paper's proposed method leverages the objective information present in the data to make logical corrections to multimodal information and extract valuable data during the fusion process, leading to a noticeable improvement in intention recognition accuracy. Furthermore, the system incorporates reverse analysis and active interaction to adjust and strengthen unfulfilled intentions. It has the ability to "think" and "judge" to a certain extent.

3.3. Intention execution analysis

This module aims to detect the outcomes of intention recognition (Int) and ensure user safety by recognizing and avoiding hazardous situations. Assessing user comfort is an essential part of contact-based human-computer interaction (HCI) tasks, but assigning a precise value to the abstract concept of "comfort" is challenging. Therefore, this paper utilizes fuzzy logic to tackle this challenge. The system utilizes historical information and perceptual data (Int) to evaluate user comfort and determine the feasibility of a task scientifically, which serves as a foundation for decision-making to ensure safe human-computer interaction.

To start, this paper categorizes intentions in the intention set into two groups: $C_{intensity}$ for regulating intensity and $C_{position}$ for regulating position. If $Int \in C_{intensity}$, the system employs both massage intensity (P) and the user's heart rate (HR) as evaluation indices to gauge user comfort. If $Int \in C_{position}$, the system uses heart rate and a fixed position massage duration as evaluation indices. The paper uses a blend of Gaussian and rectangular Membership Functions to evaluate assessment indices and considers historical factors while fuzzifying the three assessment criteria.

The evaluation index for "massage intensity," denoted as P, has a theoretical range of [0, $P_{\max}$]. Here, $P_{\max}$ signifies the maximum massage intensity that has been set manually. Historical information associated with the index reveals that it can be classified into five different fuzzy linguistic values: very small, small, medium, large, and larger. Corresponding fuzzy sets for these linguistic values are defined as VS_P, S_P, M_P, CL_P and L_P, respectively.

The evaluation index for "duration of massage in the same position," denoted as T, has a theoretical range of [0, $T_{\max}$]. Here, $T_{\max}$ denotes the maximum massage time that has been set manually. Historical information associated with the index reveals that it can be classified into five different fuzzy linguistic values: very short, short, medium, longer, and long. Corresponding fuzzy sets for these linguistic values are defined as VS_T, S_T, M_T, LL_T and L_T, respectively.

The evaluation index, HR, measures the user's heart-rate, and theoretically, it has a domain of ($HR_{\min}$, $HR_{\max}$), where $HR_{\min}$ is the artificially set minimum heart-rate limit and $HR_{\max}$ is the artificially set maximum heart-rate limit. The index has been partitioned into three fuzzy linguistic values of low, medium and high, and their corresponding fuzzy sets are S_{HR}, M_{HR}, and F_{HR}.

The output linguistic variable of the system is the comfort level, C_{omf}, and its theoretical domain is $\left[0, \quad 1\right]$. The fuzzy subsets are very low VL_C, low L_C, medium M_C, high H_C and very high VH_C.

The setting conditions for fuzzy rules are as follows:

IF P **is** $P_{\boldsymbol{i}}$ **AND** HR **is** $HR_{\boldsymbol{i}}$ **THEN** C_{omf} **is** C_i;

IF T **is** T_i **AND** HR **is** $HR_{\boldsymbol{i}}$ **THEN** C_{omf} **is** C_i. The number of fuzzy rules is obtained by the Cartesian product of fuzzy linguistic values for all input indicators. Therefore, the number of fuzzy rules for both types of intention reasoning systems is 15, as show Table 1.

For convenience in computation, this paper chooses the "maximum-minimum" composite rule to complete the calculation of the fuzzy entailment relationship. Let the fuzzy output of the comfort degree in the original state (before the execution intention Int) obtained by the system inference be denoted as μ'_{Comf-o}, and then use Equation (8) to perform defuzzification to obtain the precise output value of the comfort degree, μ_{Comf-o}.

$$\mu_{Comf-o} = \frac{\sum x_i \cdot \mu'_{Comf-o}[i]}{\sum \mu'_{Comf-o}[i]} \tag{8}$$

Table 1. Fuzzy inference rule base.

Input/Output **No.**	***P***	***HR***	$\boldsymbol{C_{omf}}$
1	VS_P	S_{HR}	M_C
2	VS_P	M_{HR}	VL_C
3	VS_P	L_{HR}	VL_C
4	S_P	S_{HR}	H_C
⋮	⋮	⋮	⋮
Input/Output No.	T	HR	C_{omf}
1	VS_T	S_{HR}	M_C
2	VS_T	M_{HR}	M_C
3	VS_T	L_{HR}	VL_C
4	S_T	S_{HR}	H_C
⋮	⋮	⋮	⋮

Here, x_i represents the center value of the membership degree function of the ith linguistic value in the output fuzzy linguistic set. Similarly, the exact numerical value of user comfort in the new state (after the execution of intention Int) can be obtained as μ_{Comf}.

The impact of intention Int on user comfort can be used as one of the criteria for judging the degree of execution, E_{Int}, which is calculated according to Equation (9) and ranges from (th −1, 2).

$$E_{Int} = (\mu_{Comf} - \mu_{Comf-o}) + \boldsymbol{I_{Fin}}(Int) + S \tag{9}$$

The safety factor, S, has only two values, 0 and −1. If $Int \in C_{power}$ and the massage intensity of the new state exceeds the limit, S is set to −1; otherwise, S is set to 0. Similarly, if $Int \in C_{position}$ and the massage location of the new state is in a sensitive area, S is set to −1; otherwise, S is set to 0. These criteria ensure that the massage device operates within safe limits and avoids potential harm to the user.

The following paragraph describes the security assessment process for intention Int. It first considers whether the execution threshold E_{Int} is greater than th_e. If it is, intention Int can be executed independently, but if it is not, the intention is deemed dangerous and passes through the collaborative module. The procedure then employs Bayesian posterior probability reverse analysis to determine the maximum negative impact factor. If the factor is found to be $(\mu_{Comf} - \mu_{Comf-o})$, the system will actively interrogate the user about the intention. If it is S, the system will reject the execution of intention Int. The minimum executable threshold is set artificially as th_e. The procedure for processing unqualified intention in human-computer collaboration is as follows: use Bayesian posterior probability reverse analysis to determine the modal Rep_{Inp} that contributes the least to the intention Int and require the user to re-express the intention actively. Doing so promotes information enhancement.

The intention executable degree detection module is capable of detecting potentially harmful intentions and thereby improving the safety of the interaction process. This module works in conjunction with the collaborative mode of reverse analysis and active interaction, which imbues the robot with the ability to "think and judge" to some extent and enhances the overall fluency of the interaction process.

3.4. MIUIC algorithm

The MIUIC algorithm consists of three primary steps: (1) determining and analyzing the effective input modality of the user and corresponding intention probability set; (2) utilizing objective information conveyed by multimodality, including unimodal information entropy, inter-modal similarity, and modal falsity, to correct and fuse multimodal data for comprehensive intention extraction; and (3) executing an analysis of the extracted intention based on user comfort. The algorithm is presented below, building upon the aforementioned discussion.

Algorithm: A Human-Computer Collaborative Multimodal Intention Understanding Algorithm (MIUIC) Incorporating Comfort Analysis

```
Input: The intensity of massage $P$; massage duration $T$; user's heart rate $HR$; user's voice $V$; user's posture $Pos$; user's gesture $Ges$.

Output: Safe intention $Final_{Int}$ that can be finally enforced

Step1. Information perception and processing:
    Determining the overall set of probabilities of intention $\boldsymbol{Inp} \leftarrow \{ V \rightarrow I_{Voice}; Pos \rightarrow I_{Pos}; Ges \rightarrow I_{Ges} \}$;
Step2. Intention fusion and intention recognition:
    IF length($\boldsymbol{Inp}$)==1:
    $Final_{Int} = Int = \max(\boldsymbol{Inp}(.))$, return $Final_{Int}$;
    Else:
        Modal self-confidence: $Cred_{self}(\boldsymbol{Inp}) \leftarrow$ formula (1);
        Modal Mutual Confidence: $Cred_{con}(\boldsymbol{Inp}) \leftarrow$ formula (3);
        Modal falsity: $F(Inp_i) \leftarrow$ formula (6);
        Modal correction coefficient: $\aleph(Inp_i) \leftarrow$ formula (7);
        Fusion of Multi-modal Data:
        $Inp_{Ave} = \sum_{i=1}^{n} \aleph(Inp_i) Inp_i$;
        $\boldsymbol{I_{Fin}} =$ Dempster $(Inp_{Ave}, n-1)$;
        $Int = argmax(\boldsymbol{I_{Fin}})$,
        IF $Int > th$ : return $Int$;
        Else: Reverse analysis:
        $Rep_{Inp} = \text{argmin}_i \; P(Inp_i \mid Int)$;
            Active interaction; Return to step 1
Step3. Judgment of Executability:
    $E_{Int} \leftarrow$ formula (1);
    IF $E_{Int} \geq th_e$ :
        $Final_{Int} = Int$, Return $Final_{Int}$.
    Else: Reverse analysis; Active interaction; Return to step 1
    End.
```

The algorithm's primary features are as follows: (1) the effective fusion of data and intention extraction through the utilization of a range of objective information implied by multimodality, which enhances the evidence theory; (2) the application of fuzzy mathematics knowledge to facilitate interactive intention safety analysis; (3) elevated rates of intention recognition and avoidance of hazardous intention resulting from behaviors such as active interaction and reverse analysis.

4. Experiment and analysis

To evaluate the proposed multimodal intention understanding model's effectiveness, this paper implements the system into the Xarm robotic arm for interaction testing.

4.1. Experimental flow and design

The experimental hardware consisted of a Win10 laptop with an i7-10850H CPU and an RTX2070S graphics card, a six-axis Xarm robotic arm, Kinect2.0, Realsense, and a Xiaomi wristband. The Xarm robotic arm was affixed to the desktop, with the Kinect2.0 positioned 1.5 meters from the user. Realsense was attached to the end-effector of the robot arm, located 5 cm away from the human machine contact surface. The software code was written in Python 3.7.

For this study, we recruited 45 participants, composed of 25 females and 20 males, by posting the recruitment information with an age limit of 55 to 70 years old and without any other additional requirements. It is important to note two aspects of the volunteer-based experiment in this study: (1) all participants signed an informed consent form before the research was conducted and were aware of the nature and purpose of the study, (2) all participants voluntarily participated in the experiment and received no compensation. Moreover, all participants completed the test efficiently following our guidelines and without experiencing any adverse reactions.

Prior to conducting the experiment, we gathered and evaluated data from all participants regarding the relationship between the intensity and duration of massages and their comfort levels, to establish a fuzzy inference rule base. Participants were acquainted with the operation of the system and the motion path of the robotic arm by watching the experimental video recording. Following the completion of all preparations, each participant was requested to interact with the system ten times during the massage process. Initially, the user naturally expressed their intention within the range of the system's vision, and the system summarized the intention based on the input data and assessed its feasibility. If the intention obtained does not meet the execution standards, the system engages in a human-computer collaborative analysis for more detailed information or prompts the user to re-express their intention. Throughout the experiment, the system avoided voice prompts to prevent interfering with the user's interaction with the system. In addition, to enhance the safety of the research system, an emergency control button was incorporated. The experimenter holds the controller throughout the whole procedure and can press the emergency control button in case of any discomfort.

While conducting the experiment, we noticed that the accuracy of intention extraction is significantly influenced by the uncertainty of input information. Unclear expression is anticipated during natural human-computer interaction, particularly among older individuals. For a more detailed evaluation of the effectiveness of the multimodal fusion algorithm in identifying fuzzy data and reducing blurs, we partitioned input information based on the degree of blurriness and created a formula for reducing blurs. The steps of fuzzy data classification include averaging the n sets of intention probabilities in the ***Inp*** list to obtain an average set of intention probabilities. We then utilized the entropy value of the average set of intention probabilities to determine the degree of fuzziness for input data D.

Table 2. Table of Intention.

I_1	I_2	I_3	I_4	I_5	I_6
Increase Massage Intensity	Reduce Massage Intensity	Move up	Move down	move to left	move to right

The paper categorizes multimodal intention sets based on entropy values at probability boundaries of 0.9, 0.7, and 0.5. Specifically, these entropy values are 0.7012, 1.5779, and 2.161, respectively. Data with $\mathrm{D} \leq 0.7012$ is categorized as clear data and is able to enter the intention extraction stage directly. Data with $\mathrm{D} > 0.7012$ is categorized as unclear intention sets, indicating that the user has provided an ambiguous or unclear expression. If $0.7012 < \mathrm{D} < 1.5779$, the unclear data is categorized as a weak dispersal intention probability set (W_D). If $1.5779 < \mathrm{D} < 2.161$, it is recorded as a medium dispersal intention probability set (M_D). If $\mathrm{D} > 2.161$, it is recorded as a strong dispersal intention probability set (S_D). In addition, to showcase the benefits of the multimodal intention understanding algorithm, we perform a comparative study by varying the input modality. Specifically, we consider four input modes: (1) speech only (V-SM); (2) gesture only (H-SM); (3) speech and gesture combined (DM); and (4) multimodal intention understanding (MM). The first three modes can be obtained by removing respective modal information from the multimodal mode.

After tallying the expression of the six types of intentions expressed by the volunteers, listed in Table 2, we assessed the algorithm performance based on five metrics.

The advantages and recognition effects of the multimodal fusion algorithm (before entering the human-computer collaborative analysis module) were evaluated by calculating the intention recognition rate (RR); $\mathrm{RR} = \mathrm{R\_count}_c/\mathrm{T\_count}_c{*}100\%$ for different input modes, considering different levels of data ambiguity, where c represents various levels of ambiguity. Here, $\mathrm{R\_count}_c$ represents the number of times the system correctly extracted intentions from the c-class input data, and $\mathrm{T\_count}_c$ represents the total number of times that the system extracted the c-class input data, during the experiment. Thus, the intention recognition rate RR can most directly reflect the effectiveness of the multimodal recognition algorithm.

1. Intention Recognition Rate (RR) before entering the human-computer collaborative analysis module: $\mathrm{RR} = \mathrm{R\_count}_c/\mathrm{T\_count}_c{*}100\%$. Here, c represents input data with different degrees of dispersion. $\mathrm{R\_count}_c$ represents the number of times the system correctly extracted intentions from the c-class input data, and $\mathrm{T\_count}_c$ represents the total number of times that the system extracted the c-class input data, during the experiment. Thus, RR can directly reflect the effectiveness of the multimodal recognition algorithm.
2. Dispersion Reduction Capability (DR) before entering the human-computer collaborative analysis module:

$$DR = \begin{cases} \dfrac{\mathrm{D} - D(\boldsymbol{I_{Fin}})}{\mathrm{D} - 0.7012}{*}100\%, & D(\boldsymbol{I_{Fin}}) \geq 0.7012 \\ 1, & D(\boldsymbol{I_{Fin}}) < 0.7012 \end{cases} \quad (10)$$

Here, the variable D represents the degree of dispersion in the probability set of multimodal intention. Meanwhile, $D(\boldsymbol{I_{Fin}})$ corresponds to the entropy value of the probability distribution of the intention set after performing multimodal information fusion. The effectiveness of the multimodal fusion algorithm can be accurately and precisely reflected through the numerical representation of BR.

3. Intention Correction/Enhancement Rate (CR): $\mathrm{CR} = \mathrm{R\_count}_{\mathrm{un}}/\mathrm{T\_count}_{\mathrm{un}}{*}100\%$. Here, $\mathrm{R\_count}_{\mathrm{un}}$ indicates the number of unmet intentions converted to standards following a human-computer collaborative analysis, and $\mathrm{T\_count}_{\mathrm{un}}$ represents the total number of unmet intention recognition cases during the experiment. CR is a critical performance indicator that provides crucial information regarding the safety collaborative strategy's efficacy, thereby enabling potential areas for improvement.
4. Hazardous Intention Avoidance Rate (AR):$\mathrm{AR} = \mathrm{DF\_count}/\mathrm{DT\_count}{*}100\%$, where DF_count represents the total number of times the system recognized hazardous intention during the experiment and DT_count represents the total number of times hazardous intention appeared during the experiment. The effectiveness of the collaborative intention safety assurance method can be reflected numerically through AR.
5. Subjective Evaluation by Users: We used the System Usability Scale (SUS) (Vlachogianni & Tselios, 2022) to evaluate the usability of the M-algorithm based massage system in a way that was convenient for volunteers due to its wide availability and lack of time constraints. In order to obtain more data, we also had volunteers test the home-based massage chair and rate its usability using the same SUS questionnaire.

4.2. Experimental results and analysis

4.2.1. Intention recognition rate

During the course of the experiment, we observed 358 instances of dispersal probability distribution in the multimodal intention set caused by non-standard input behavior. These instances had an average information entropy higher than 0.7012 and were categorized as weak dispersion, moderate dispersion, or strong dispersion with 141, 132, and 85 instances respectively. The analysis of intention recognition rates with varying input modes allowed us to discern the correlation between input mode and the probability distribution of the input intention set. We also ascertained the intention recognition rate (RR), illustrated in Figure 2.

From the Figure 2, it is evident that the dispersion of input data significantly affects the accuracy of intention recognition. As the data dispersion increases, the accuracy of intention recognition decreases. However, when the data has the same degree of dispersion, multimodal information can effectively complement the shortcomings of unimodal information and achieve a higher accuracy of intention recognition.

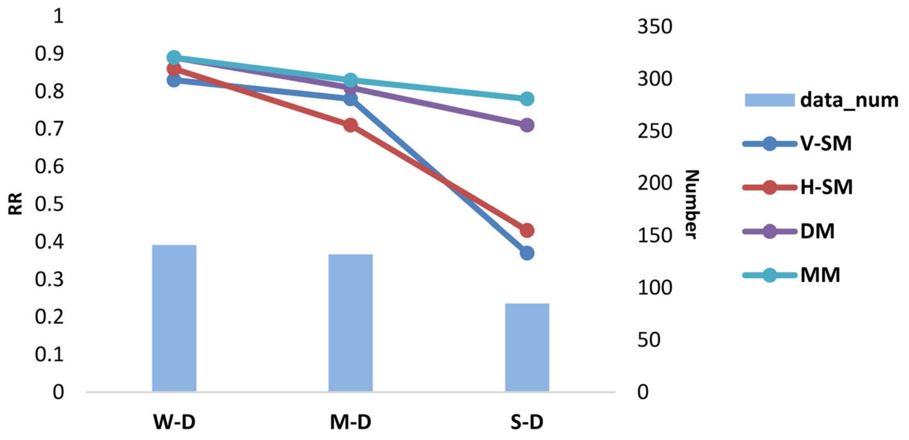

Figure 2. The plot of intention recognition accuracy versus input pattern and input dispersion.

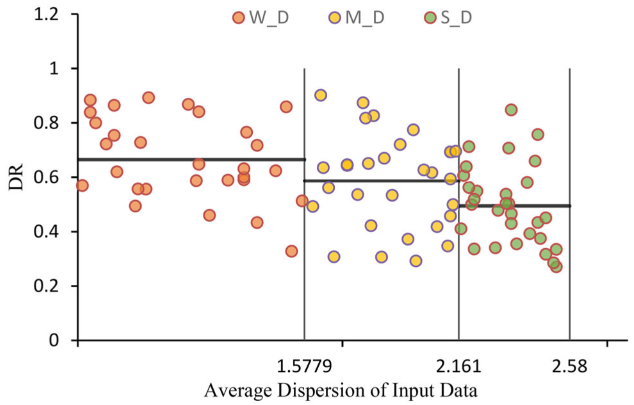

Figure 3. Statistical graph of dispersion reduction capability of MIPIC algorithm for various input data.

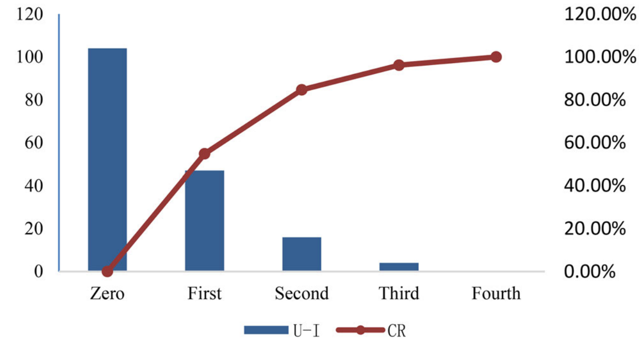

Figure 4. The correlation between active interaction and the rate of intention enhancement. The horizontal axis denotes the number of active interactions, while the primary-vertical axis represents the total number of unfulfilled intentions, the secondary-vertical axis describes the intention correction/enhancement rate.

4.2.2. Analysis of the blur reduction ability of multimodal fusion algorithm

Further, to better detect the precise effects of our multimodal fusion algorithm on atypical inputs, we extracted 30 sets of data from each of the different dispersed multimodal intention probability sets to form three test groups. By analyzing the changes in the dispersion degree of the intention probability sets before and after multimodal fusion, we generated the results depicted in Figure 3.

As depicted in the figure, multi-modal data fusion effectively reduces dispersion degree by an average of 0.6646, 0.5869, and 0.4947 for the weakly, moderately, and highly dispersed test group data, respectively. These results indicate that the multi-modal fusion algorithm presented in this paper effectively reduces the dispersion degree of the probability of intention and promotes a concentrated distribution. The data above demonstrates that the multi-modal fusion algorithm presented in this paper is capable of reducing the dispersion degree of multi-modal data, enhancing the concentration distribution of the probability of intention, and improving the recognition rate of intention.

4.2.3. Active interaction effect of human-computer collaboration and hazardous intention avoidance effect

In order to evaluate the impact of active analysis and interaction on the multimodal intention understanding algorithm, we collected statistical data regarding collaborative behavior between humans and machines in response to unfulfilled intentions. Among the 358 instances of intention identification, 104 fell short of the fulfillment threshold ($\boldsymbol{I}_{\boldsymbol{Fin}}(Int) < th$), henceforth referred to as unfulfilled intentions and denoted as U-I.

Through Bayesian analysis and active interaction, the system acquires more accurate information regarding intentions. In this paper, we investigate the correlation between the number of human-computer interactions and the frequency of unmet intentions, as well as the correlation with the intention correction/enhancement rate (CR). Figure 4 depicts these relationships, where the horizontal axis represents the number of reverse analyses, the primary vertical axis represents the frequency of unmet intentions, and the secondary vertical axis represents the intention correction/enhancement rate. The figure demonstrates that as the number of reverse human-computer interactions increases, the frequency of unmet intentions decreases, and the intention correction/enhancement rate increases proportionally. The third human-computer interaction marks a notable threshold, after which both the intention extraction and intention correction/enhancement rates reach a high level. Therefore, human-computer interaction contributes to the naturalness of interaction, enables enhanced intention information acquisition, and improves the rate of intention recognition.

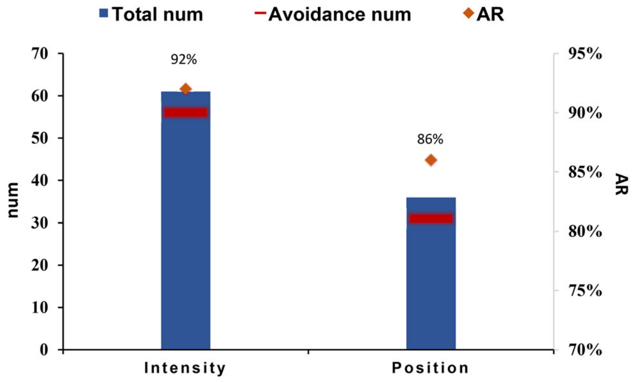

Figure 5. Statistical chart of hazard intention avoidance.

The collaborative safety strategy between humans and machines, with a proactive reverse function, can not only enhance the identification of intentions that violate safety standards, but can also distinguish and prevent dangerous intentions that compromise safety or cause discomfort. To evaluate the effectiveness of this function, we recorded the system's ability to prevent and rectify dangerous intentions during the experiment, as illustrated in Figure 5. Dangerous intentions are defined as attempts to increase the intensity level beyond the limits, or to change the massage position to an unsafe area, which were voiced spontaneously by volunteers 97 times during the experiment, including 61 times of adjusting the intensity level and 36 times of adjusting the massage position. Based on the data in the figure, the rate of successful prevention for dangerous intensity level adjustments was 92%, while the rate for dangerous massage position adjustments was 86%. As a result, the safety human-computer collaborative strategy employed in this study provides an effective means of preventing hazardous intentions to achieve a safe interaction.

4.2.3. User experience analysis

At the end of the experiment, we asked volunteers to complete a SUS questionnaire with 10 questions, rating satisfaction on a scale of 1–5 for the massage system implemented in this paper and a home massage chair. See Table 3 for the questionnaire contents. Questions 1–4 explored user's use of the system, questions 5–6 investigated user's psychological safety using the system, and questions 7–10 assessed user's cognitive load while using the system. Figure 6 summarizes the questionnaire statistics.

The statistical results demonstrate that the human-computer collaborative massage system described in this paper provides a more positive experience for elderly users. This is due to the system's increased customization options, making it more adaptable to user needs and therefore easier to use. Additionally, the system's natural interaction reduces the user's mental effort since it does not require memorization of fixed expression patterns. However, the size of the robot arm and its self-navigation feature may cause some users to experience nervous pressure. It has been found that the implementation of comfort-detection technology and human-computer collaboration can gradually reduce this pressure.

Table 3. SUS questionnaire.

Question	Description (low score → high score)
Q1	Whether you are willing to use this massage system (unwilling → willing)
Q2	Whether the function of this massage system is simple (difficult → simple)
Q3	Whether you need to ask for help to complete the massage (required → not required)
Q4	Whether this massage system is worth promoting (no → yes)
Q5	Whether you feel psychologically uncomfortable when using the system (yes →no)
Q6	Whether you feel physically uncomfortable when using the system (yes →no)
Q7	Confident or frustrated in the process of using it (frustrated → confident)
Q8	Whether the interaction process of this massage system is simple (difficult → simple)
Q9	How do you feel about self-performance using this system (not good → good)
Q10	Does it require a lot of learning before using this massage system (required → not required)

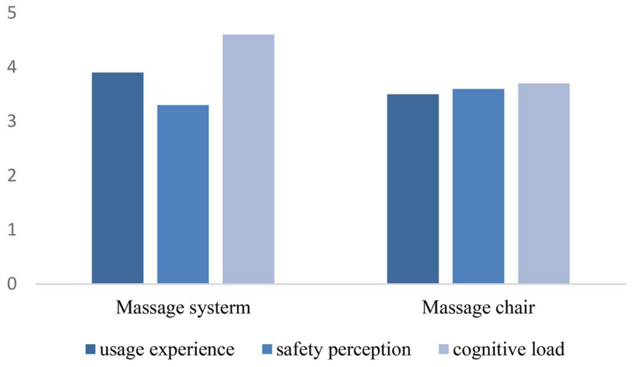

Figure 6. Natural pointing massage positioning vs. 3D mannequin massage positioning NASA survey results.

The experiments conducted in this study demonstrate that the algorithm for multimodal intention recognition proposed in this paper achieves good intention recognition even with fuzzy input data (expressions), which solves the problem of limited range of expressions and promotes natural human-machine interaction for elderly users. Additionally, the module for intention feasibility assessment added to the system enables effective identification and prevention of unsafe intentions, ensuring the safety of users. In conclusion, this study proposes a safe and natural interaction-based algorithm for multimodal intention recognition that can be applied to home-based intelligent massage systems for the elderly population.

5. Discussion

In the context of human-computer interaction, natural expression and the age of the target audience may lead to inconsistent communication styles, making it challenging for the system to recognize the user's intentions. To promote safe interaction, this study proposes a multi-modal fusion-based intention recognition algorithm and a reverse-active collaborative strategy based on comfort reasoning for intelligent massage tasks. Experimental results demonstrate that optimizing multi-modal data's hidden objective information by implementing data correction and fusion can decrease input data's dispersion and enhance the system's intention recognition rate. Moreover, the reverse-active interactive strategy can naturally boost the intention recognition rate and provide safety. As a result, the proposed MIUIC algorithm highlights (1) the facility with which users can express themselves without specific constraints, (2) how multi-modal data's potential objective information can achieve efficient data fusion, and (3) how during the intention recognition process, intention feasibility tests based on comfort reasoning enable robots to make decisions and steer clear of harmful circumstances resulting from "blind obedience."

The MIUIC algorithm has achieved high levels of multimodal intention recognition and dangerous intention avoidance. However, the limited number of modalities used in this study presented challenges in obtaining data from each modality. Furthermore, the system recognized only a limited range of intentions. For speech recognition, the algorithm relied mainly on keyword-based speech recognition technology, which has inherent limitations. To mitigate this, we recommend integrating phonetic characteristics and visual lip shape information as supplementary information to increase human-computer interaction and improve speech recognition rates. Further experimentation and improvement are necessary to overcome these limitations (Bastanfard et al., 2009).

Additionally, human safety is critical for tasks that involve close contact between humans and machines. The part localization and depth perception techniques used in this paper rely mainly on skeletal point recognition techniques, detecting hand gestures and depth cameras, which may experience issues such as inaccurate recognition of skeletal points and depth detection. To augment these techniques, we can incorporate more modal information, such as speech emotion recognition (Keshtiari et al., 2015), expression recognition (Kollias, 2022), and sensors to analyze human intentions.

Regarding future research directions, inspired by the literature (Hudec et al., 2021), this study will explore ways to enable the affiliation function to possess generalization capabilities based on the clarification of the fuzzy information about the behavior. This is crucial for massage robots in terms of human-computer interaction and is necessary to ensure user safety. Additionally, this study will extend the concept of human-computer collaboration to enhance the machine's intelligence and decision-making capabilities to achieve more flexible massage functions to address massage safety issues.

6. Conclusion

This paper describes the MIUIC algorithm, which combines multi-modal data analysis and human-computer interaction for a better understanding of user intention. It specifically addresses the lack of safety measures in existing intelligent massage systems. The MIUIC algorithm utilizes multi-modal data to infer user intention, effectively reducing the dispersion of input data and improving recognition rate. To ensure user safety, the algorithm employs reverse active interaction strategy by detecting user comfort level and effectively avoiding dangerous intentions. Experimental results show that the MIUIC algorithm recognizes user intention effectively, improves the collaborative relationship between the user and the robot, and promotes the development of home-style massage systems.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This paper is supported by the Independent Innovation Team Project of Jinan City [No. 2019GXRC013].

References

Al-Amin, M., Tao, W., Doell, D., Lingard, R., Yin, Z., Leu, M. C., & Qin, R. (2019). Action recognition in manufacturing assembly using multimodal sensor fusion. *Procedia Manufacturing*, *39*, 158–167. https://doi.org/10.1016/j.promfg.2020.01.288

Awan, U., Gölgeci, I., Makhmadshoev, D., & Mishra, N. (2022). Industry 4.0 and circular economy in an era of global value chains: What have we learned and what is still to be explored? *Journal of Cleaner Production*, *371*, 133621. https://doi.org/10.1016/j.jclepro.2022.133621

Awan, U., Sroufe, R., & Bozan, K. (2022). Designing value chains for industry 4.0 and a circular economy: A review of the literature. *Sustainability*, *14*(12), 7084. https://doi.org/10.3390/su14127084

Bai, D., Sun, Y., Tao, B., Tong, X., Xu, M., Jiang, G., Chen, B., Cao, Y., Sun, N., & Li, Z. (2022). Improved single shot multibox detector target detection method based on deep feature fusion. *Concurrency and Computation: Practice and Experience*, *34*(4), e6614. https://doi.org/10.1002/cpe.6614

Bastanfard, A., Aghaahmadi, M., Kelishami, A. A., Fazel, M., & Moghadam, M. (2009). *Persian viseme classification for developing visual speech training application* [Paper presentation]. Advances in Multimedia Information Processing-PCM 2009: 10th Pacific Rim Conference on Multimedia, Thailand, December 15–18, 2009 Proceedings 10.

Berg, J., Lottermoser, A., Richter, C., & Reinhart, G. (2019). Human-Robot-Interaction for mobile industrial robot teams. *Procedia CIRP*, *79*, 614–619. https://doi.org/10.1016/j.procir.2019.02.080

Callens, T., Van der Have, T., Van Rossom, S., De Schutter, J., & Aertbeliën, E. (2020). A framework for recognition and prediction of human motions in human-robot collaboration using probabilistic motion models. *IEEE Robotics and Automation Letters*, *5*(4), 5151–5158. https://doi.org/10.1109/LRA.2020.3005892

Carros, F., Meurer, J., Löffler, D., Unbehaun, D., Matthies, S., Koch, I., Wieching, R., Randall, D., Hassenzahl, M., Wulf, V. (2020). Exploring human-robot interaction with the elderly: Results from a ten-week case study in a care home. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.*

Chang, L. J., Jolly, E., Cheong, J. H., Rapuano, K. M., Greenstein, N., Chen, P.-H A., & Manning, J. R. (2021). Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Science Advances*, *7*(17), eabf7129. https://doi.org/10.1126/sciadv.abf7129

Chen, L., Su, W., Feng, Y., Wu, M., She, J., & Hirota, K. (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences*, *509*, 150–163. https://doi.org/10.1016/j.ins.2019.09.005

Chen, Z., Shen, D., Yu, F., Ren, Y., Xu, X., & Zhang, H. (2022). *Aerial target recognition method based on improved sensor credibility* [Paper presentation]. Sixth International Conference on Electromechanical Control Technology and Transportation (ICECTT 2021). https://doi.org/10.1117/12.2623890

Czarnowski, J., Dąbrowski, A., Maciaś, M., Główka, J., & Wrona, J. (2018). Technology gaps in human-machine interfaces for autonomous construction robots. *Automation in Construction*, *94*, 179–190. https://doi.org/10.1016/j.autcon.2018.06.014

Dawar, N., Ostadabbas, S., & Kehtarnavaz, N. (2019). Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. *IEEE Sensors Letters*, *3*(1), 1–4. https://doi.org/10.1109/LSENS.2018.2878572

Dinh Le, T.-S., An, J., Huang, Y., Vo, Q., Boonruangkan, J., Tran, T., Kim, S.-W., Sun, G., & Kim, Y.-J. (2019). Ultrasensitive anti-interference voice recognition by bio-inspired skin-attachable self-cleaning acoustic sensors. *ACS Nano*, *13*(11), 13293–13303. https://doi.org/10.1021/acsnano.9b06354

Dutta, V., & Zielinska, T. (2019). Predicting human actions taking into account object affordances. *Journal of Intelligent & Robotic Systems*, *93*(3–4), 745–761. https://doi.org/10.1007/s10846-018-0815-7

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Ferguson, S., Luders, B., Grande, R. C., & How, J. P. (2015). *Real-time predictive modeling and robust avoidance of pedestrians with uncertain, changing intentions* [Paper presentation]. Algorithmic Foundations of Robotics XI: Selected Contributions of the Eleventh International Workshop on the Algorithmic Foundations of Robotics.

Field, T. (2019). Pediatric massage therapy research: A narrative review. *Children*, *6*(6), 78. https://doi.org/10.3390/children6060078

Flesher, S. N., Downey, J. E., Weiss, J. M., Hughes, C. L., Herrera, A. J., Tyler-Kabara, E. C., Boninger, M. L., Collinger, J. L., & Gaunt, R. A. (2021). A brain-computer interface that evokes tactile sensations improves robotic arm control. *Science (New York, N.Y.)*, *372*(6544), 831–836. https://doi.org/10.1126/science.abd0380

Gadekallu, T. R., Srivastava, G., Liyanage, M., Iyapparaja, M., Chowdhary, C. L., Koppu, S., & Maddikunta, P. K. R. (2022). Hand gesture recognition based on a Harris hawks optimized convolution neural network. *Computers and Electrical Engineering*, *100*, 107836. https://doi.org/10.1016/j.compeleceng.2022.107836

Gervasi, R., Mastrogiacomo, L., & Franceschini, F. (2020). A conceptual framework to evaluate human-robot collaboration. *The International Journal of Advanced Manufacturing Technology*, *108*(3), 841–865. https://doi.org/10.1007/s00170-020-05363-1

Hajarian, M., Bastanfard, A., Mohammadzadeh, J., & Khalilian, M. (2019). SNEFL: Social network explicit fuzzy like dataset and its application for Incel detection. *Multimedia Tools and Applications*, *78*(23), 33457–33486. https://doi.org/10.1007/s11042-019-08057-3

Haleem, A., Javaid, M., & Vaishya, R. (2020). Effects of COVID-19 pandemic in daily life. *Current Medicine Research and Practice*, *10*(2), 78–79. https://doi.org/10.1016/j.cmrp.2020.03.011

Herath, D. C., Jochum, E., & Vlachos, E. (2018). An experimental study of embodied interaction and human perception of social presence for interactive robots in public settings. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(4), 1096–1105. https://doi.org/10.1109/TCDS.2017.2787196

Hirzinger, G., Albu-Schaffer, A., Hahnle, M., Schaefer, I., & Sporer, N. (2001). *On a new generation of torque controlled light-weight robots* [Paper presentation]. Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164). https://doi.org/10.1109/ROBOT.2001.933136

Holmes, E. A., O'Connor, R. C., Perry, V. H., Tracey, I., Wessely, S., Arseneault, L., Ballard, C., Christensen, H., Cohen Silver, R., Everall, I., Ford, T., John, A., Kabir, T., King, K., Madan, I., Michie, S., Przybylski, A. K., Shafran, R., Sweeney, A., … Bullmore, E. (2020). Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science. *The Lancet. Psychiatry*, *7*(6), 547–560. https://doi.org/10.1016/S2215-0366(20)30168-1

Hudec, M., Mináriková, E., Mesiar, R., Saranti, A., & Holzinger, A. (2021). Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. *Knowledge-Based Systems*, *220*, 106916. https://doi.org/10.1016/j.knosys.2021.106916

Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, *53*, 209–221. https://doi.org/10.1016/j.inffus.2019.06.019

Jiménez-Pavón, D., Carbonell-Baeza, A., & Lavie, C. J. (2020). Physical exercise as therapy to fight against the mental and physical consequences of COVID-19 quarantine: Special focus in older people. *Progress in Cardiovascular Diseases*, *63*(3), 386–388. https://doi.org/10.1016/j.pcad.2020.03.009

Kelley, R., Tavakkoli, A., King, C., Ambardekar, A., Nicolescu, M., & Nicolescu, M. (2012). Context-based bayesian intention recognition. *IEEE Transactions on Autonomous Mental Development*, *4*(3), 215–225. https://doi.org/10.1109/TAMD.2012.2211871

Keshtiari, N., Kuhlmann, M., Eslami, M., & Klann-Delius, G. (2015). Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD). *Behavior Research Methods*, *47*(1), 275–294. https://doi.org/10.3758/s13428-014-0467-x

Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., Sestito, M., Dreher, J.-C., & Rao, R. P. (2019). Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances*, *5*(11), eaax8783. https://doi.org/10.1126/sciadv.aax8783

Kim, D., Goyal, A., Newell, A., Lee, S., Deng, J., & Kamat, V. R. (2019). Semantic relation detection between construction entities to support safe human-robot collaboration in construction (*Computing in Civil Engineering 2019*. In *Data, Sensing, and Analytics* (pp. 265–272). American Society of Civil Engineers.

Kleinman, D. L., Baron, S., & Levison, W. (1970). An optimal control model of human response part I: Theory and validation. *Automatica*, *6*(3), 357–369. https://doi.org/10.1016/0005-1098(70)90051-8

Kollias, D. (2022). Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Kozak, L., Vig, E., Simons, C., Eugenio, E., Collinge, W., & Chapko, M. (2013). A feasibility study of caregiver-provided massage as supportive care for Veterans with cancer. *The Journal of Supportive Oncology*, *11*(3), 133–143. https://doi.org/10.12788/j.suponc.0008

Lang, X., Feng, Z., Yang, X., & Xu, T. (2023). HMMCF: A human-computer collaboration algorithm based on multimodal intention of reverse active fusion. *International Journal of Human-Computer Studies*, *169*, 102916. https://doi.org/10.1016/j.ijhcs.2022.102916

Li, C., Fahmy, A., Li, S., & Sienz, J. (2020). An enhanced robot massage system in smart homes using force sensing and a dynamic movement primitive. *Frontiers in Neurorobotics*, *14*, 30. https://doi.org/10.3389/fnbot.2020.00030

Mewes, D., & Mauser, F. (2003). Safeguarding crushing points by limitation of forces. *International Journal of Occupational Safety and Ergonomics*, *9*(2), 177–191. https://doi.org/10.1080/10803548.2003.11076562

Mi, Q., Wang, C., Camerer, C. F., & Zhu, L. (2021). Reading between the lines: Listener's vmPFC simulates speaker cooperative choices in communication games. *Science Advances*, *7*(10), eabe6276. https://doi.org/10.1126/sciadv.abe6276

Naruse, S. M., Cornelissen, P. L., & Moss, M. (2020). 'To give is better than to receive?'Couples massage significantly benefits both partners' wellbeing. *Journal of Health Psychology*, *25*(10-11), 1576–1586. https://doi.org/10.1177/1359105318763502

Naruse, S. M., & Moss, M. (2019). Positive Massage for Couples' Wellbeing and Relationships: The Bridge between Positive Psychology and Massage. *Health*, *11*(12), 1609–1624. https://doi.org/10.4236/health.2019.1112122

Ortega, J. D., Senoussaoui, M., Granger, E., Pedersoli, M., Cardinal, P., & Koerich, A. L. (2019). Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196*

Papanastasiou, S., Kousi, N., Karagiannis, P., Gkournelos, C., Papavasileiou, A., Dimoulas, K., Baris, K., Koukas, S., Michalos, G., & Makris, S. (2019). Towards seamless human robot collaboration: Integrating multimodal interaction. *The International Journal of Advanced Manufacturing Technology*, *105*(9), 3881–3897. https://doi.org/10.1007/s00170-019-03790-3

Peternel, L., Tsagarakis, N., Caldwell, D., & Ajoudani, A. (2016). *Adaptation of robot physical behaviour to human fatigue in human-robot co-manipulation* [Paper presentation]. 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). https://doi.org/10.1109/HUMANOIDS.2016.7803320

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science (New York, N.Y.)*, *372*(6547), 1209–1214. https://doi.org/10.1126/science.abe2629

Pineau, J., Montemerlo, M., Pollack, M., Roy, N., & Thrun, S. (2003). Towards robotic assistants in nursing homes: Challenges and results. *Robotics and Autonomous Systems*, *42*(3-4), 271–281. https://doi.org/10.1016/S0921-8890(02)00381-0

Rodomagoulakis, I., Kardaris, N., Pitsikalis, V., Mavroudi, E., Katsamanis, A., Tsiami, A., & Maragos, P. (2016). *Multimodal human action recognition in assistive human-robot interaction* [Paper presentation]. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP.2016.7472168

Sharma, D., Purushotham, S., & Reddy, C. K. (2021). MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, *11*(1), 19826. https://doi.org/10.1038/s41598-021-98390-1

Si, W., Srivastava, G., Zhang, Y., & Jiang, L. (2019). Green internet of things application of a medical massage robot with system interruption. *IEEE Access. 7*, 127066–127077. https://doi.org/10.1109/ACCESS.2019.2939502

Singh, D., Merdivan, E., Hanke, S., Kropf, J., Geist, M., & Holzinger, A. (2017). *Convolutional and recurrent neural networks for activity recognition in smart environment* [Paper presentation]. Towards Integrative Machine Learning and Knowledge Extraction: BIRS Workshop, Banff, AB, Canada, July 24–26, Revised Selected Papers.

Skantze, G. (2017). Predicting and regulating participation equality in human-robot conversations: Effects of age and gender. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.*

Sridhar, S., Markussen, A., Oul€asvirta, A., Theobalt, C., Boring, S. (2017). Watchsense: On-and above-skin input sensing through a wearable depth sensor. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.*

Sugano, Y., Zhang, X., Bulling, A. (2016). Aggregaze: Collective estimation of audience attention on public displays. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology.*

Sun, M., He, W., Zhang, L., & Wang, P. (2019). *Smart haproxy: A novel vibrotactile feedback prototype combining passive and active haptic in AR interaction* [Paper presentation]. 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). https://doi.org/10.1109/ISMAR-Adjunct.2019.00026

Tan, Y., Sun, Z., Duan, F., Solé-Casals, J., & Caiafa, C. F. (2021). A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical Signal Processing and Control*, *70*, 103029. https://doi.org/10.1016/j.bspc.2021.103029

Trick, S., Koert, D., Peters, J., & Rothkopf, C. A. (2019). *Multimodal uncertainty reduction for intention recognition in human-robot interaction* [Paper presentation]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems *(IROS)*. https://doi.org/10.1109/IROS40897.2019.8968171

Vlachogianni, P., & Tselios, N. (2022). Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. *Journal of Research on Technology in Education*, *54*(3), 392–409. https://doi.org/10.1080/15391523.2020.1867938

Wang, M., Yan, Z., Wang, T., Cai, P., Gao, S., Zeng, Y., Wan, C., Wang, H., Pan, L., Yu, J., Pan, S., He, K., Lu, J., & Chen, X. (2020). Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors. *Nature Electronics*, *3*(9), 563–570. https://doi.org/10.1038/s41928-020-0422-z

Wang, Y., & Zhang, F. (2017). *Trends in control and decision-making for human-robot collaboration systems.* Springer.

Wei, X., Wang, S., Li, L., & Zhu, L. (2017). Clinical evidence of Chinese massage therapy (Tui Na) for cervical radiculopathy: A systematic review and meta-analysis. *Evidence-Based Complementary and Alternative Medicine*, *2017*, 1–10. https://doi.org/10.1155/2017/9519285

Yang, C., Wang, D., Zeng, Y., Yue, Y., & Siritanawan, P. (2019). Knowledge-based multimodal information fusion for role recognition and situation assessment by using mobile robot. *Information Fusion*, *50*, 126–138. https://doi.org/10.1016/j.inffus.2018.10.007

Yücel, Ş. Ç., Arslan, G. G., & Bagci, H. (2020). Effects of hand massage and therapeutic touch on comfort and anxiety living in a nursing home in Turkey: A randomized controlled trial. *Journal of Religion and Health*, *59*(1), 351–364. https://doi.org/10.1007/s10943-019-00813-x

Zacharaki, A., Kostavelis, I., Gasteratos, A., & Dokas, I. (2020). Safety bounds in human robot interaction: A survey. *Safety Science*, *127*, 104667. https://doi.org/10.1016/j.ssci.2020.104667

Zhang, L., Xie, Y., Xidao, L., & Zhang, X. (2018). *Multi-source heterogeneous data fusion* [Paper presentation]. 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD). https://doi.org/10.1109/ICAIBD.2018.8396165

Zhang, Y., Tang, S., Chen, G., & Liu, Y. (2015). Chinese massage combined with core stability exercises for nonspecific low back pain: A randomized controlled trial. *Complementary Therapies in Medicine*, *23*(1), 1–6. https://doi.org/10.1016/j.ctim.2014.12.005

Zheng, Z. K., Zhu, J., Fan, J., & Sarkar, N. (2018). *Design and system validation of rassle: A novel active socially assistive robot for elderly with dementia* [Paper presentation]. 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). https://doi.org/10.1109/ROMAN.2018.8525819

About the authors

Liran Zhou is a graduate student at the Department of Computer Science and Technology, University of Jinan. Her research interests lie in human-computer interaction and collaboration in elderly care.

Zhiquan Feng is a professor of Computer Science and Technology at University of Jinan. His work explores human-machine interaction and collaboration issues in topics such as smart education, elderly robots, and robotic arms.

Hongyue Wang is a graduate student at the Department of Computer Science and Technology, University of Jinan. His research interests lie in human-computer interaction, virtual reality and artificial intelligence research in smart education.

Qingbei Guo is an associate professor of Computer Science and Technology at University of Jinan. His research at intersection of pat tern recognition and computer vision focuses especially on human computer collaboration.