Hindawi
Scientific Programming
Volume 2021, Article ID 8354015, 11 pages
https://doi.org/10.1155/2021/8354015

Research Article

An Intention Understanding Algorithm Based on Multimodal Information Fusion

Shaosong Dou,[1,2] **Zhiquan Feng**,[1,2] **Jinglan Tian**,[1,2] **Xue Fan**,[1,2] **Ya Hou**,[1,2] **and Xin Zhang**[1,2]

[1]*School of Information Science and Engineering, University of Jinan, 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 17 June 2021; Revised 2 September 2021; Accepted 25 October 2021; Published 18 November 2021

Academic Editor: Qianchuan Zhao

This paper proposes an intention understanding algorithm (KDI) based on an elderly service robot, which combines Neural Network with a seminaive Bayesian classifier to infer user's intention. KDI algorithm uses CNN to analyze gesture and action information, and YOLOV3 is used for object detection to provide scene information. Then, we enter them into a seminaive Bayesian classifier and set key properties as super parent to enhance its contribution to an intent, realizing intention understanding based on prior knowledge. In addition, we introduce the actual distance between the users and objects and give each object a different purpose to implement intent understanding based on object-user distance. The two methods are combined to enhance the intention understanding. The main contributions of this paper are as follows: (1) an intention reasoning model (KDI) is proposed based on prior knowledge and distance, which combines Neural Network with seminaive Bayesian classifier. (2) A set of robot accompanying systems based on the robot is formed, which is applied in the elderly service scene.

1. Introduction

The current aging is a problem faced by many countries in the world. Because of the busy work of children, it is difficult to give their parents the care they need. At the same time, through the investigation of nursing home and the research on service robot by Joost and others [1], it is found that robot service is more and more recognized by the elderly, and the elderly service robot provides many services. However, the intention to understand the rate of robot service system is relatively low at present, which makes the elderly increase the interaction burden when using the service robot. As a result, based on improving the understanding of the intention of the elderly service robot system and improving the satisfaction of the elderly, we proposed a multimodal intention understanding algorithm named KDI to understand the behavior of the elderly.

In this paper, the results of single-modal identification are obtained by Neural Network. The gesture, action, and scene are obtained by Neural Network. The results of single modal are fused based on a seminaive Bayesian classifier to infer the final intention and by improving the YOLOV3 target detection model to establish a perspective matrix to obtain the actual distance between the user and the objects to enhance the user's intention.

Compared with pure Neural Network Classifier, KDI has better intention understanding because KDI greatly improves the efficiency of high-level task recognition by combining the advantages of single-modal high recognition rate of Neural Network and the advantages that Bayesian classifier is easy to adjust the results and easy to expand the multimodal information, which is more suitable for advanced identification tasks.

2. Related Work

2.1. Multimodal Fusion. In recent years, many researches have been conducted on multimodal information fusion technology, and multimodal information fusion is a technology to integrate information from different sources [2].

At present, model fusion is mainly divided into two methods, one is based on Neural Network multimodal fusion and the other is based on probability-based multimodal fusion. The multimodal fusion based on Neural Network mostly processes the multimodal information into low-level features, splices the tensors, forms a new long tensor, and trains the results [3]. During the low-level feature fusion, various features can be extracted to improve the performance of the system [4]. Reference [5] proposed to bridge the emotional gap by using a hybrid deep model, which first produces audio-visual segment features with Convolutional Neural Networks (CNN) and 3DCNN and then fuses audio-visual segment features in a Deep Belief Networks (DBN). The accuracy of recognition results is high, but the user's intention is mostly composed of multiple modal information. If adding multimodal information will introduce a large number of parameters and feature level information fusion is not easy to adjust recognition results frequently and has poor flexibility, it is not suitable for advanced intention understanding tasks. High-level feature fusion is proposed [6, 7].

There have been studies combining Neural Networks and probabilistic models for recognition tasks [8]. The two models combined exert the Neural Network to save the single-modal training cost and bring a high identification rate by fine-tuning the existing network. Probabilistic models have also been developed to easily modify training parameters and to easily augment multimodal channels and arrange and combine results to obtain personalized high-level (intention) identification results. Reference [9] used the CNN's powerful capacity of learning high-level features directly from raw data and used it to extract effective and robust action features. The HMM is used to model the statistical dependence over adjacent subactions and infer the action sequences. For example, [10] proposed an integrated probability-based decision framework for robots to infer the role of humans in a particular task. It combines Neural Network and probability model. Those methods greatly increase the flexibility of recognition task, can adjust important parameters in recognition task, and is easy to carry out incremental learning.

2.2. Scene Perception and Intention Understanding. Data from multimodal or heterogeneous sensors can provide additional scene information, which enables the system to understand objects more comprehensively and accurately [11]. So if the robot can sense the semantics of the surrounding environment, many recognition and prediction tasks can be effectively completed [6]. The robot sees the environment through sensors. And the collected sensor data are fused into the multilayer representation of spatial knowledge for semantic mapping [12, 13]. YOLOV3 not only has a certain accuracy but also maintains a high running speed after several improvements [14, 15]. YOLOV3 is favored in many tasks with high edge computing and real-time requirements. It is widely used in environment detection to provide scene information [16, 17].

Machine understanding human intention is a key problem in the field of human-computer interaction. And in order to understand the visual world, machines must recognize not only how to identify scene information but also how they interact with and upcoming interactive actions [18]. Reference [19] proposed a system to extract human, verb, and object triples in daily photos. Reference [20] proposed a new function, "active understanding of human intentions" by a robot through monitoring human behavior. Reference [21] presents a framework that allows a robot to automatically recognize and infer the action intention of a human partner based on visualization. During the collaboration, a robot with intention understanding ability can predict the successive actions that a human partner intends to perform, provide necessary assistance and support, and remind for the missing and failure actions from the human to achieve the desired task purpose.

2.3. Summary. To sum up, we greatly improve the efficiency of high-level task recognition by combining the advantages of Neural Network and the advantages of the probabilistic model. We use scene information as one of the pieces of multimodal information. In the intention understanding task, we add the system active response process, which greatly improves the intelligence of the system.

3. Materials and Methods

The algorithm is mainly divided into two parts: intention reasoning based on prior knowledge and intention reasoning based on distance. Finally, the two reasoning probabilities are combined. The detailed process is shown in Figure 1.

Based on the reasoning of prior knowledge, we used Neural Network to identify the single-modal information to obtain the category label, such as gesture recognition label (h_i) and action recognition label (a_i). At the same time, we obtain the target detection information obtained through YOLOV3 and input the above information into the semi-naive Bayesian classifier to obtain the intended result as P^K.

In distance-based intention reasoning, we give intention to objects in the scene. At present, we only give each object one intention; that is, intention and object belong to one-to-one correspondence. In our understanding, the user is close to an object and is likely to interact with it. Therefore, we assume that the closer the user to the object, the greater the probability of interaction as P^D.

Finally, we combined the above two intention results to obtain the final intention result.

Next, we will introduce the process of intention reasoning based on Bayesian and distance and the basis of intention classification after intention fusion.

3.1. Prior Knowledge-Based Intention Reasoning

3.1.1. Acquisition of Gestures, Actions, and Target Detection. The method to get gestures is the gesture recognition model proposed by [22]. Their experimental results show that the

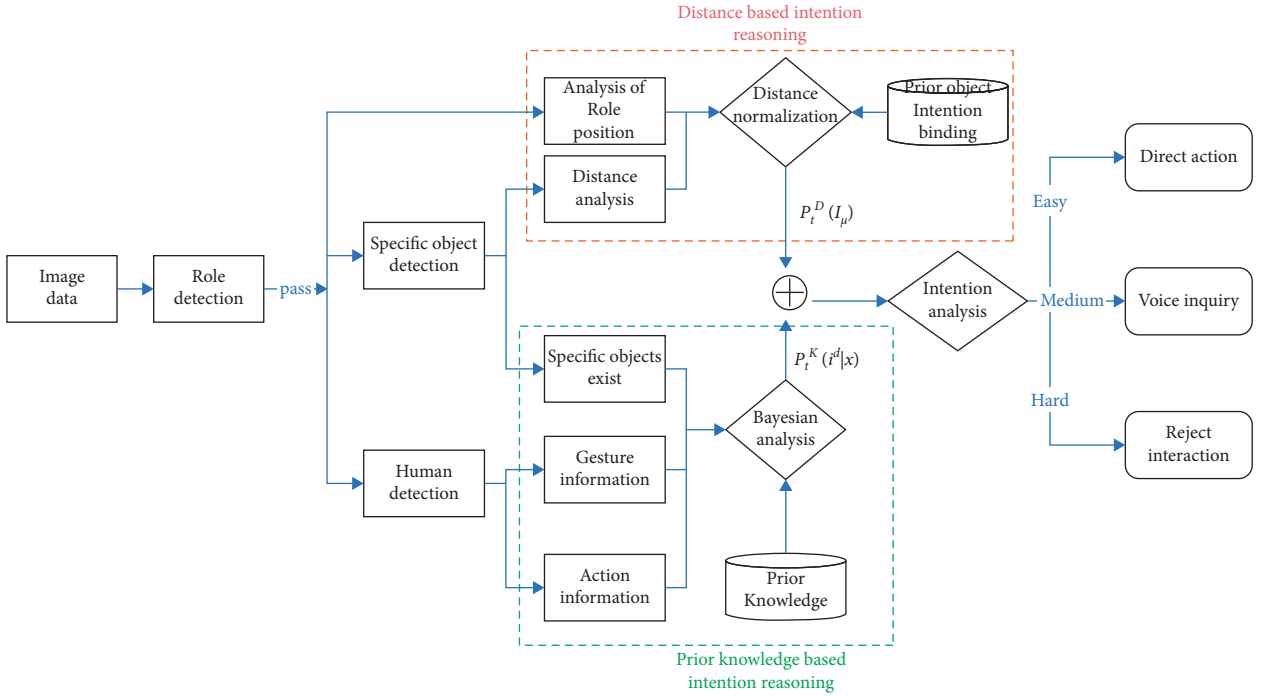


Figure 1: KDI intention understanding algorithm framework.

method can handle complex gesture interactions with a recognition rate of 97.7%. We used their model and data set for Kinect-based gesture recognition using five gestures (grab, fist, five fingers open, three fingers open, and extend index finger) from its classification results. Using Kinect-based skeleton information detection, we preset five human actions (walk, bend over, reach out, lie down, and sit down) through the relative distance and angular relationship between the three-dimensional skeletal points and set corresponding thresholds [23]. The test accuracy was 84%.

We use the YOLOV3 network to train the object recognition network for real-time object detection. The data set for Neural Network training consists of 330000 pictures. There are 80 kinds of pictures, which cover almost all indoor daily necessities and contain many object information to take care of the daily living of the elderly living alone.

3.1.2. Intentional Probability Formula Based on Seminaive Bayesian. In the actual daily life scene, one of the many pieces of information that affect the understanding of intention always has a feature dependence on intention. For example, the existence of water cups around the user is particularly important for the probability determination of drinking intention. That means if the water cup is near the user, the user's intention to drink water will be significantly greater. The reason why we set super parent is to increase the importance of a key attribute. Property dependencies are shown in Figure 2.

Therefore, we choose the existence of a special object (μ) which is relatively important as the super parent to estimate independently.

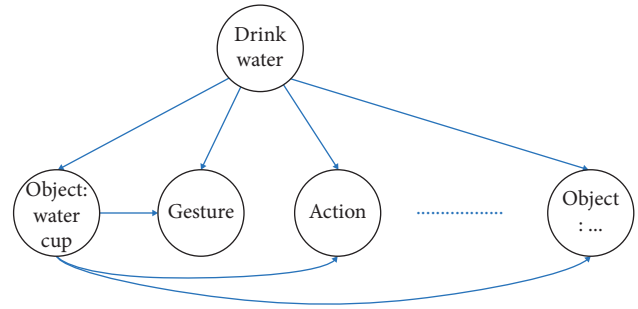


Figure 2: Attribute dependencies in seminaive Bayesian.

We take h_{t-1}, a_{t-1} and h_t, a_t for $t-1$ and t moment gesture and action features, respectively. μe is whether a specific object μ exists at time t and its value range is $\mu e = \{\text{cup}_1, \text{cup}_0, \text{chair}_1, \text{chair}_0, \ldots\}$. For example, cup_1 is the existence of cup and cup_0 is the nonexistence of cup. x_j is the property value and its value range is $x_j \in \{h_{t-1}, h_t, a_{t-1}, a_t, \mu e\}$. I is the intention and its value range is $I \in (i_1, i_2, \ldots, i_n)$. x' is the super parent and its value range is $x' \in \{\text{cup}_1, \text{chair}_1, \ldots\}$, which means that, under different intentions, we take μe different attribute values of object existence as super parent classes. $p(I, x')$ is the prior probability of I. $p(x_j|I, x')$ is the conditional probability of I. $P_t(I|x)$ is the posterior probability of I at time t as formula (1).

$$P_t(I|x) \propto P(I, x') \prod_{j=1}^{d} P_t\left(x_j|I, x'\right). \tag{1}$$

We use the SPODE model of seminaive Bayes to estimate the prior probability and conditional probability as formulas

(2) and (3). In formulas (2) and (3), D represents the complete data set. Let N denote the number of possible classes in data set D. N_i is the number of possible values of the i attribute. $D_{I,x'}$ is an aggregate whose intention category is I and whose values on the μe attribute is x'. D_{I,x',x_j} is an aggregate whose intention category is I and whose values on the μe and j attributes are x' and x_j. We used data sets in Table 1 to train the prior probability and conditional probability.

$$P(I, x') = \frac{|D_{I,x'}| + 1}{|D| + N \times N_{x'}}. \tag{2}$$

$$P(x_j | I, x') = \frac{|D_{I,x',x_j}| + 1}{|D_{I,x'}| + N_j}. \tag{3}$$

Next, the prior probability and conditional probability are obtained according to formulas (2) and (3). Then, the test data are input into the classifier (1) in real time. Finally, the posterior probability of classifying the test data at time t into each intention label is calculated as $P_t(I|x)$, $I \in (i_1, i_2, \ldots, i_n)$.

By formula (4), we normalize the probability of each intention obtained by seminaive Bayesian classifier (1) and obtain the proportion of each intention I in the total intention. $P_t^K(I|x)$ is the proportion of each intention in all intentions, $I \in (i_1, i_2, \ldots, i_n)$.

$$P_t^K(I|x) = \frac{P_t(I|x)}{\sum_t P_t(I|x)}. \tag{4}$$

3.2. Intentional Reasoning Based on Distance Calculation of Actual Distance

3.2.1. Four-Point Perspective Method. At present, most of the cameras used for target detection are ordinary RGB cameras. In the general target detection pictures, the distance between people and objects will be perspectively distorted. Therefore, we need to get the actual distance through perspective transformation.

In order to get the actual distance between the user and the object, we use the perspective matrix, calibrate the camera, and place it 2.5 m away from the ground. We define the perspective matrix (5). $\begin{bmatrix} x \\ y \end{bmatrix}$ and $\begin{bmatrix} x' \\ y' \end{bmatrix}$ are the coordinates of the source image and the coordinates of the target image, respectively.

$$\begin{bmatrix} x'' \\ y'' \\ \omega \end{bmatrix} = \omega \cdot \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = M \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} m0 & m1 & m2 \\ m3 & m4 & m5 \\ m6 & m7 & m8 = 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{5}$$

There are eight unknowns. We take four points to solve the matrix from the coordinates of the source image. By establishing the perspective change matrix and improving the YOLOV3, we get the mapping relationship of the image as follows. For example, the distance between the chairs in the upper left corner and the upper right corner is equal to the distance between the chairs in the lower-left corner and the lower right corner. After perspective transformation, the distance is basically the same as shown in Figure 3.

Table 1: Partial prior data set.

#	h_{t-1}	h_t	a_{t-1}	a_t	Cup	Thermos	Chair	...	Intent
1	h_1	h_3	a_1	a_4	cup_1	thermos_1	chair_1		i_3
2	h_4	h_3	a_3	a_4	cup_1	thermos_0	chair_0	...	i_3
3	h_1	h_2	a_5	a_3	cup_1	thermos_1	chair_1		i_3
4	h_1	h_3	a_3	a_3	cup_1	thermos_1	chair_0		i_3
						...			
60	h_1	h_3	a_2	a_2	cup_0	thermos_0	chair_1	...	i_2
						...			

After the actual test, we find that the best result is to select the midpoint at the bottom of the bounding box to calculate the distance between objects. The error is within 15 cm to ensure the accuracy of the estimated probability.

3.2.2. Distance-Based Probabilistic Calculations. Generally speaking, when the user is close to an object, a large probability is intended to operate it. For example, when the user is close to a chair, the user has a large probability of interacting with the chair, such as sitting or moving chair, as shown in Figure 4.

In formula (6), D_t is the sum of the distance from the detected special objects as $\mu \in (\text{cup}, \text{chair}, \ldots)$ to the user at time t. And $d_{\mu,t}$ is the distance from μ to the user at time t. For example, $d_{\text{cup},t}$ is the actual distance from the cup to the user at the t time.

$$D_t = \sum_{\mu} d_{\mu,t}. \tag{6}$$

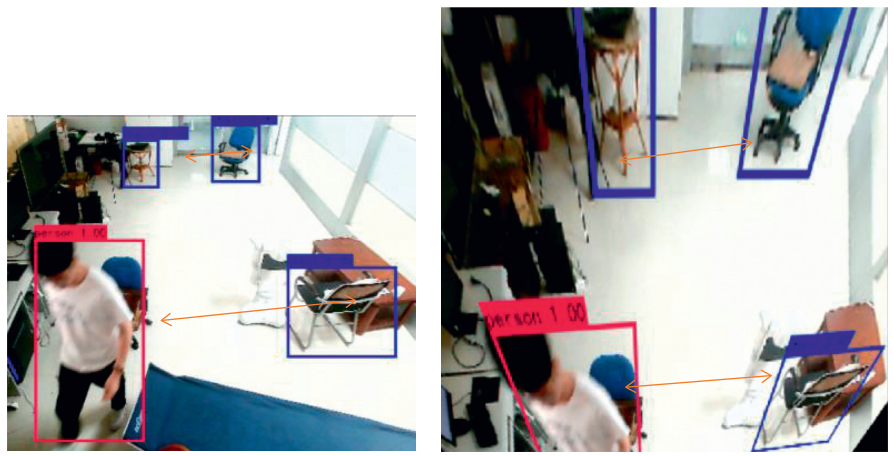
At present, we only bind one intention to one object, and we can bind multiple intentions in future work. Intention and object belong to one-to-one correspondence as $I_\mu = I \in (i_1, i_2, \ldots, i_n)$ (e.g., if $\mu = \text{cup}$, I = i_3(pour the water)). $P^D(I_\mu)$ means the probability of intention I_μ between the user and object μ.

After normalization with formula (7), the probability $P_t^D(I_\mu)$ that the target will interact with the user is the difficulty of interaction.

$$P_t^D(I_\mu) = \frac{D_t - d_{\mu,t}}{D_t}. \tag{7}$$

3.3. Intentional Fusion Based on Multimodal Information

3.3.1. Intentional Reasoning Formula Based on Multimodal Fusion. By combining the probability of each intention probability obtained from knowledge reasoning and distance reasoning, algorithm (8) complements and corrects the two intention recognition results. Finally, the prediction probability $P_t(I)$ of each intention at time t is obtained.



(a) (b)

Figure 3: Distance relationship after perspective transformation. (a) After. (b) Before.

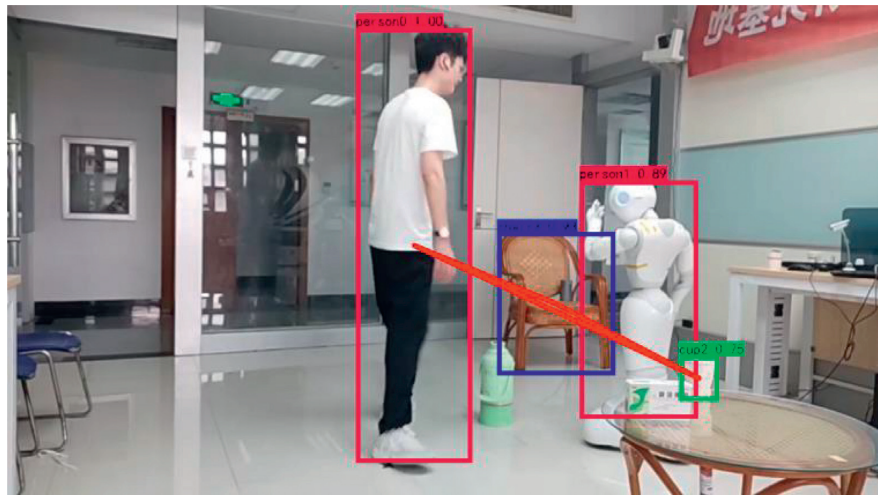


Figure 4: Distance map ($d_{\text{chair},t} < d_{\text{cup},t}$, more likely to interact with a chair).

$$P_t(I) = \frac{P_t^D(I_\mu) \times P_t^K(I|x)}{\sum_I P_t^D(I_\mu) \times P_t^K(I|x)}. \tag{8}$$

3.3.2. KDI Intention Understanding Algorithm Flow. Next, according to several statistical experiments, we determine the thresholds ε_1 and ε_2. Test with the threshold range of 0.1~0.9. For example, when we take the threshold ε_1 as 0.9, although the user's intention is obvious, it still fails to meet the system standard; that is, the threshold setting is unreasonable. After 200 statistical experiments, the accuracy of system feedback is 95% when $\varepsilon_1 = 0.7$ and $\varepsilon_2 = 0.3$. We reduce the error of the threshold to about 5%.

When $\max(P_t(I)) > \varepsilon_1$, the level of intention in the current situation is considered to be simple and the interaction can be initiated actively. For example, when all the features of the mobile chair are satisfied at time t (gesture is grab and action is bow, the chair exists, and the distance is closest), the result is simple, and the robot takes the initiative to help move the chair.

When $\varepsilon_1 < \max(P_t(I)) < \varepsilon_2$, the level of intention I occurs in the current situation is medium and the user can be asked tentatively. When $\max(P_t(I)) < \varepsilon_2$, it is difficult to determine the level of intention in the current situation and then refuse to interact at this t time.

Figure 5 is the interactive diagram of the algorithm. In the following figure, we take getting the user's intention as an example to move the table. Firstly, the robot models the scene information and learns that there are people, table, thermos, and cup. Next, it judges the user's gestures and actions and combines with the location information of the objects and the user. The final result is that the judgment intention level is simple, and the robot actively helps to move the table.

3.3.3. KDI Algorithm Analysis. KDI algorithm can process multichannel information in parallel and alleviate the multimodal conflict problem. The acquired action, gesture, and objects information are processed in real time, and the final probability is calculated based on the fusion algorithm (1). In

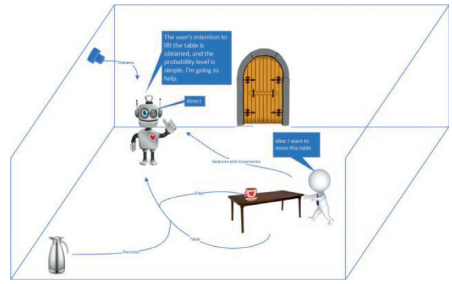


Figure 5: A schematic diagram of human-robot interaction (moving the table).

Input: A frame image at time t.
Output: The max probability of classification of intentions $\max(P_t(I))$.
Start:
User's information detection (from Kinect).
(1) If the user is found in the scene, then
(2) wake up the system and find the user's information.
(3) If user's action (a_{t-1}, a_t) and gesture (h_{t-1}, h_t) are not found, then
(4) continue to find the user's information (1).
(5) end if.
(6) If a_{t-1}, a_t, h_{t-1}, and h_t are found, then
(7) record h_{t-1}, h_t, a_{t-1}, and a_t at the t-1 and t moment.
(8) end if.
Scene detection (from base station camera).
(9) If the special objects (μ) are detected in the prior knowledge base, then
(10) record as $\mu e = \{\text{cup}_1, \text{cup}_0, \text{chair}_1, \text{chair}_0, \ldots\}$ and the actual distance between user and μ in t moment as $d_{\mu,t}$.
(11) end if.
Knowledge reasoning.
(12) Put the data into (1) $\leftarrow$ h_{t-1}, h_t, a_{t-1}, a_t, andμe to get the results of each intention.
(13) If attribute information (x_j) is lost, then
(14) determine the conditional probability of missing attribute that is 1 as $p(x_j|I, x') = 1$ and then
(15) obtain ${P_t}^K(I|x)$ after probability normalization with (4) $\leftarrow P_t(I|x)$.
(16) end if.
Distance reasoning.
(17) We use (7) $\leftarrow$, $d_{\mu,t}$ to get ${P_t}^D(I_\mu)$.
Get the final intention.
(18) Fuse by (8) $\leftarrow$, ${P_t}^D(I_\mu)$, ${P_t}^K(I|x)$ to get the final result $P_t(I)$, $I \in (i_1, i_2, \ldots, i_n)$.
(19) max = 0, for I = i_1:i_n *do*
if max $P_t(I)$, then max = $P_t(I)$
end if.
end for.
(20) $\max(P_t(I))$, $\leftarrow$ max.
(21) end if.
End

Algorithm 1: Intentional understanding algorithm based on knowledge and distance.

our experiment, there is a probability that a channel recognition is not successful, such as being affected by illumination. If attribute information (x_j) is lost at the current time, we determine that the conditional probability of missing attribute is 1 ($p(x_j|I, x') = 1$) and continue the intention calculation. It achieves the purpose that the overall probability will not be affected even if the information of a certain channel cannot be obtained. However, when all the information (motion and gesture) of the user is lost, we assume that the Kinect is blocked and fails to recognize and retrieve the user's information.

4. Results and Discussion

The host processor selected in the experiment is Intel (R) Core (TMi7-9750 CPU). In the 64-bit Windows 10 system, the target detection network model is based on YOLOV3 and calibrates the RGB camera 2.5 m away from the ground to collect the target detection image. We use Kinect as pepper's eyes to take images, use CNN to analyze gesture information, and use bone point information to analyze action information (see 3.1.1). The development language is C++ and python, and the development platform is Microsoft

Visual Studio and PyCharm. The microphone of the pepper robot is used for voice channel information.

4.1. Experimental Process

4.1.1. The Process of Knowledge-Based Reasoning. For the settings, gestures, and actions with their attribute values, as shown in Table 2. Gesture feature has attributes h_1, h_2, h_3, h_4, and h_5, and action feature has attributes a_1, a_2, a_3, a_4, and a_5. The existence attributes of objects are cup (binding intention i_3), chair (binding intention i_2), and thermos (no binding intention). The intention is classified as i_1, i_2, i_3, i_4, and i_5.

Table 1 is part of the data set of this experiment. The data set of this classifier determines the on-site action demonstration of 10 elderly people (five women and five men). In the demonstration process, we informed them of the intention range and provided different interactive scenes (whether there are special objects), and then the user makes gesture and action feedback under his own selected intention. From the test results, we screened 100 test data for each intention, and a total of 500 test data formed a prior data set. We used data sets to set prior probability and conditional probability by (2) and (3).

4.2. Examples and Analysis of Experiments. Table 3 is a predesigned human-computer interaction behavior for us. We consulted our interviewees and got the most desired action for the robot under each intention as our robot feedback. It is based on our actual research results.

Figure 6 shows a group of data at a certain time t and we split it into multiple single modals as shown in Figure 7. They illustrated the operation process and intention recognition process of the algorithm by giving examples of each attribute value. The actual distance between the user with the chair and the water cup at time t measured by the base station camera is $d_{\text{chair},t} = 3.1$ and $d_{\text{cup},t} = 17$, respectively.

Table 4 shows the values of each attribute captured by the algorithm at a certain time t and input them into the classifier. Based on the prior and conditional probabilities, the probability of each intention is calculated by (1). At the same time, the distance probability between each special object and the user is calculated by (7). Finally, the intention probability after fusion is obtained by (8).

For those intentions that can be inferred without binding objects, we make their distance probability P^D as 0.5 by default. That is, i_1 means that shaking hands with robot does not need to bind objects so that $P_t^D(i_1) = 0.5$.

After comparing the fusion probability, the probability of intention i_2 is the largest. Next, the robot will classify the user's intention according to the intention level classification method. Compared with the threshold value, that is, $P_t(i_2) > \varepsilon_1$, the intention level is easy, and the robot walks to move the chair.

4.3. Comparison Test and Analysis. Three groups of experiments were designed as control experiments. In the identification layer, the acquisition of single-modal information is based on the CNN model proposed in [22] and the LSTM model proposed in [24] as a control experiment. In the intention fusion layer, the seminaive Bayesian classifier without distance reasoning and the deep belief network proposed in [25] are used for control experiments. In the control experiments, the same prior data set was used, and 100 experiments were carried out with each model. We summarized the data in Table 5. We only care about the single-modal accuracy, the understanding rate of the final intention, and the time spent for the classifier. And we evaluate the understanding rate of the intent classifiers by building four confusion matrices in Figure 8.

It can be seen from the above table that the KDI model proposed in this paper is superior to the other three models in intention understanding rate. Moreover, the proposed distance reasoning optimizes the intention understanding rate in the experiment. Compared with the deep learning fusion model, this algorithm is more extensible, and new intention classification can be added by establishing its prior knowledge.

4.4. User Experience. In this section, 30 elderly people (15 men and 15 women) were invited to participate in the user experience. The experience results show that the following four systems can roughly understand and interact with the user's intention. We recorded the user experience scale in detail by using a questionnaire Figure 9. We use system convenience, system helpfulness, system user load, and system accuracy to conduct a satisfaction survey. The score is 1~10 points (1 is the worst and 10 is the best). Finally, the user satisfaction questionnaire is statistically analyzed to obtain the satisfaction chart, as shown in Figure 10.

5. Results and Discussion

This paper proposes a multimodal intention understanding algorithm (KDI), which achieves 91% intention understanding rate through experiments. Compared with other intention understanding algorithms, it has the characteristics of high efficiency, high accuracy, and easy to realize incremental learning. And in the practical application scenario, through the user experience interview of the elderly, we get a higher user experience evaluation. The low user experience load of the algorithm model proves the practicability and convenience of the algorithm.

There are still many short weaknesses in our work, as follows:

(1) In persona recognition, only one person can be the protagonist. When multiple users appear in the scene, the system selects a "zero" user by default for system operation.

(2) The system has a limitation of camera occlusion. When there is an obstruction between the user and Kinect, the user information capture is incomplete. In the future, it is considered to obtain the user's information through wearable devices.

Table 2: Knowledge base.

Gestures		Actions		Intention		Object	Binding intention
h_1	Grab	a_1	Walk	i_1	Shake hands	Cup	i_3
h_2	Fist	a_2	Bend over	i_2	Move object	Chair	i_2
h_3	Five fingers open	a_3	Reach out	i_3	Pour the water	Thermos	Null
h_4	Three fingers open	a_4	Lie down	i_4	Draw the curtain		
h_5	Extend index finger	a_5	Sit down	i_5	Disease alarm		

Table 3: Human-computer interaction behavior.

A pepper robot is used to predict the interaction of intention:
Intention i_1: the robot approaches the user and raises its arm to shake hands.
Intention i_2: the robot moves to help the old man move heavy objects.
Intention i_3: the robot moves to the thermos, picks up the thermos, and moves to the user.
Intention i_4: the robot moves to draw curtains for the user.
Intention i_5: the robot calls its children for sudden anomalies.

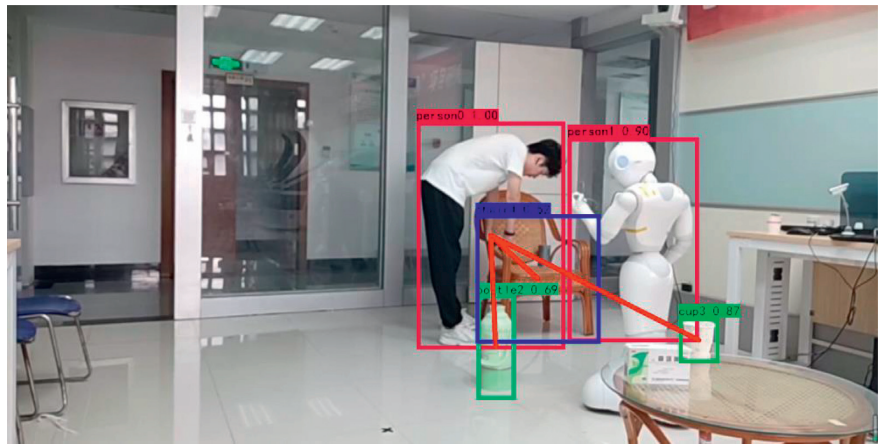
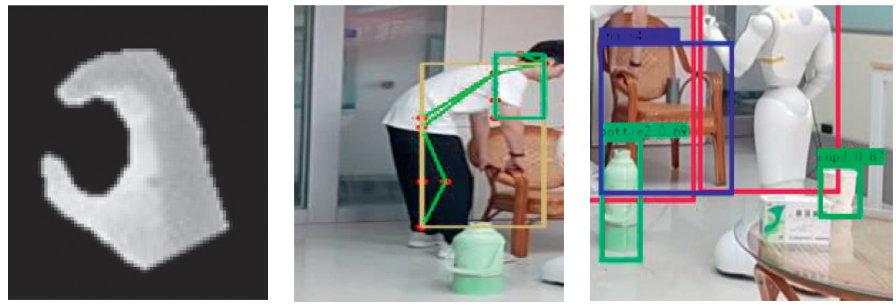


Figure 6: Practical scene diagram of experiment. The robot captures the user's gesture as h1 and action as a2, detects the chair, thermos, and cup, and draws their actual distance from the user.



(a) (b) (c)

Figure 7: Single-channel data at certain time t: (a) hand h1; (b) action a2; (c) object detection: cups, chairs, and thermos.

Table 4: At the time t, the value of each attribute and probability.

Property	h_{t-1}	h_t	a_{t-1}	a_t	Cup (i_3)	Chair (i_2)	Thermos
Value	h_1	h_1	a_2	a_3	cup_1	chair_1	thermos_1
P^K of probability	${P_t}^K(i_1)$	${P_t}^K(i_2)$	${P_t}^K(i_3)$	${P_t}^K(i_4)$	${P_t}^K(i_5)$		
Value	0.0060	0.9414	0.000681	0.0164	0.0109		
Distance	$d_{\text{cup},t}$	$d_{\text{chair},t}$					
Value	17	3.1					
P^D of probability	${P_t}^D(i_1)$	${P_t}^D(i_2)$	${P_t}^D(i_3)$	${P_t}^D(i_4)$	${P_t}^D(i_5)$		
Value	0.5	0.7360	0.1342	0.5	0.5		
Intention after fusion	$P_t(i_1)$	$P_t(i_2)$	$P_t(i_3)$	$P_t(i_4)$	$P_t(i_5)$		
Probability	0.0042277	0.9764077	0.0001288	0.0115556	0.0076803		

Table 5: Comparative analysis of comparative tests.

	CNN + SNB (without distance)	**CNN + SNB (KDI)**	CNN + DBN (with distance)	LSTM + DBN (with distance)
Single-modal accuracy	0.77	**0.77**	0.77	0.82
Intentional comprehension rate	0.81	**0.90**	0.84	0.88
Time spent (second)	27	**27**	83	116

The bold values highlight the advantages of the model results in this paper.

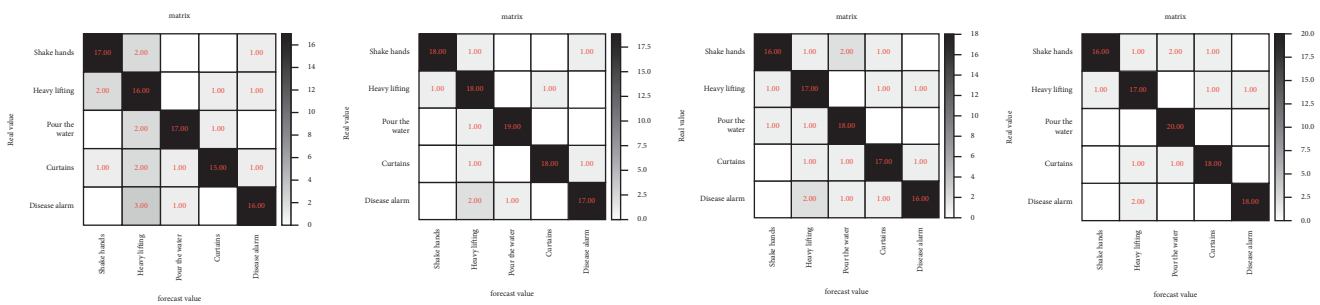


Figure 8: Comparison of the intention comprehension rate of the confusion matrix. The calculated understanding rate of their intention is 0.81 for CNN + SNB (without distance), 0.90 for (KDI), 0.84 for CNN + DBN, and 0.88 for LSTM + SNB. They demonstrated that KDI has higher intent understanding. (a) CNN + SNB (without distance). (b) CNN + SNB (KDI). (c) CNN + DBN. (d) LSTM + SNB.

User satisfaction survey (score 1-10)

1. Scoring system convenience
(the user will respond to the identification speed of each channel and the response speed of the system)

System 1 (CNN + Nb) ( ) system 2 (KDI) ( )

System 3 (CNN + DBN) ( ) system 4 (LSTM + DBN) ( )

2. Score system helpfulness
(users to feedback whether the system helps users meet their current intention)

System 1 (CNN + Nb) ( ) system 2 (KDI) ( )

System 3 (CNN + DBN) ( ) system 4 (LSTM + DBN) ( )

3. System user load scoring
(the intellectual and physical load to be carried out by the user when using system (the higher the score, the smaller the load))

System 1 (CNN + Nb) ( ) system 2 (KDI) ( )

System 3 (CNN + DBN) ( ) system 4 (LSTM + DBN) ( )

4. Scoring system accuracy
(whether the system correctly recognizes the user's current intention)

System 1 (CNN + Nb) ( ) system 2 (KDI) ( )

System 3 (CNN + DBN) ( ) system 4 (LSTM + DBN) ( )

Figure 9: Use satisfaction questionnaire.

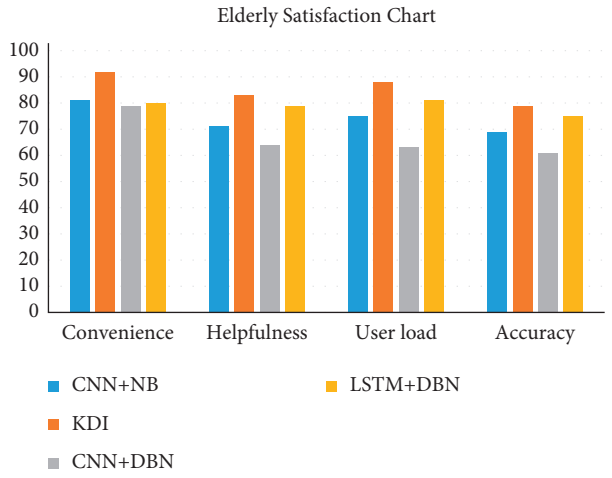


Figure 10: Use satisfaction chart for older persons.

(3) At present, the algorithm has less prior knowledge and user's intention and only gives a single intention to a single object. In the future, multiple intentions will be considered for a single object.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by the Independent Innovation Team Project of Jinan City (no. 2019GXRC013).

References

[1] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: a review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.

[2] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and p," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[3] S. Kumari, U. Kowsalya, R. Preethi, R. Theepa, J. Paulraj, and S. J. Anusuya, "Audio-visual emotion recognition using 3DCNN and DBN techniques," *International Journal of Advance Research, Ideas and Innovations in Technology*, 2018.

[4] M. Wang, Z. Yan, T. Wang et al., "Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors," *Nature Electronics*, vol. 3, no. 9, 2020.

[5] S. Zhang, S. Zhang, T. Huang, G. Wen, and T. Qi, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[6] S. Lauro and b Jesús García, "Context-based information fusion: a survey and discussion - ScienceDirect," *Information Fusion*, vol. 25, pp. 16–31, 2015.

[7] N. Da War, S. Osta Da Bbas, and N. Kehtarnavaz, "Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition," *Electronics Letters*, vol. 3, no. 1, 2018.

[8] A. Savran, H. Cao, A. Nenkova, and V. Ragini, "Temporal bayesian fusion for affect sensing: combining video, audio, and lexical modalities," *IEEE Transactions on Cybernetics*, vol. 45, no. 9, pp. 1927–1941, 2014.

[9] J. Lei, G. Li, J. Zhang, Q. Guo, and D. Tu, "Continuous action segmentation and recognition using hybrid convolutional neural network-hidden Markov model model," *IET Computer Vision*, vol. 10, no. 6, pp. 537–544, 2016.

[10] Y. Chule, W. Danwei, Z. Yijie, Y. Yufeng, and S. Prarinya, "Knowledge-based multimodal information fusion for role recognition and situation assessment by using mobile robot," *Information Fusion*, vol. 50, pp. 126–138, 2019.

[11] M. Shoaib, S. Bosch, H. Scholten, J. M. H. Paul, and D. I. Ozlem, "Towards detection of bad habits by fusing smartphone and smartwatch sensors," in *Proceedings of the IEEE International Conference on Pervasive Computing & Communication Workshops*, March 2015.

[12] S. Rosa, A. Patanè, C. X. Lu, and T. Niki, "Semantic place understanding for human–robot coexistence—toward intelligent workplaces," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 2, pp. 160–170, 2018.

[13] A. Aydemir, A. Pronobis, M. Gobelbecker, and P. Jensfelt, "Active visual object search in unknown environments using uncertain semantics," *IEEE Transactions on Robotics*, vol. 29, no. 4, pp. 986–1002, 2013.

[14] J. Redmon, S. Divvala, R. Girshick, G. Ross, and F. Ali, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, San Juan, PR, USA, May 2016.

[15] J. Redmon and A. Farhadi, *Yolov3: An Incremental improvement*, arXiv preprint arXiv:1804.02767, 2018.

[16] W. Y. Hsu and W. Y. Lin, "Ratio-and-scale-aware YOLO for pedestrian detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 934–947, 2020.

[17] Q. C. Mao, H. M. Sun, Y. B. Liu, and S. J. Rui, "Mini-YOLOv3: real-time object detector for embedded applications," *IEEE Access*, vol. 99, p. 1, 2019.

[18] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: visual semantic role labeling for image

understanding[C]//computer vision & pattern recognition," *IEEE*, pp. 5534–5542, 2016.

[19] G. Gkioxari, R. Girshick, P. Dollar, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE*, Salt Lake City, UT, USA, June 2018.

[20] T. Sato, Y. Nishida, J. Ichikawa, Y. Hatamura, and H. Miczoguchi, "Active understanding of human intention by a robot through monitoring of human behavior," *Intelligent Robots and Systems*, vol. 13, pp. 349–372, 1995.

[21] H. I. Lin, X. A. Nguyen, and W. K. Chen, "Active intention inference for robot-human collaboration," *International Journal of Computational Methods and Experimental Measurements*, vol. 6, p. 12, 2017.

[22] X. Guo, Z. Feng, K. Sun, H. Liu, W. Xie, and J. Bi, "Research on unified recognition model and algorithm for multi-modal gestures," *The Journal of China Universities of Posts and Telecommunications*, vol. 26, no. 2, pp. 30–42, 2019.

[23] L. Xie, H. J. Liao, and Y. B. Yang, "Recognition and application research of kinect-based gesture," *Gesture Recognition*, vol. 5, pp. 258–260, 2013.

[24] Z. Liang, G. Zhu, P. Shen, S. Juan, A. S. Syed, and B. Mohammed, "Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition," in *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017.

[25] M. M. Hassan, M. Alam, M. Z. Uddin, H. Shamsul, A. Ahmad, and F. Giancarlo, "Human emotion recognition using deep belief network architecture," *Information Fusion*, vol. 51, pp. 10–18, 2018.