Knowledge-Based Multimodal Information Fusion for Role Recognition and Situation Assessment by Using Mobile Robot

Chule Yang[a], Danwei Wang[a], Yijie Zeng[a], Yufeng Yue[a,*], Prarinya Siritanawan[b]

[a] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore*
[b] *ST Engineering - NTU Corporate Laboratory, Nanyang Technological University, Singapore*

Abstract

Decision-making is the key for autonomous systems to achieve real intelligence and autonomy. This paper presents an integrated probabilistic decision framework for a robot to infer roles that humans fulfill in specific missions. The framework also enables the assessment of the situation and necessity of interaction with the person fulfilling the target role. The target role is the person who is distinctive in movement or holds a mission-critical object, where the object is pre-specified in the corresponding mission. The proposed framework associates prior knowledge with spatial relationships between the humans and objects as well as with their temporal changes. *Distance-Based Inference (DBI)* and *Knowledge-Based Inference (KBI)* support recognition of human roles. DBI deduces the role based on the relative distance between humans and the specified objects. KBI focuses on human actions and objects existence. The role is estimated using weighted fusion scheme based on the information entropy. The situation is assessed by analyzing the action of the person fulfilling the target role and relative position of this person to the mission-related entities, where the entity is something that has a particular function in the corresponding mission. This assessment determines the robot decision on what actions it should take. A series of experiments has proofed that the proposed framework provides a reasonable assessment of the situation. Moreover, it outperforms other approaches on accuracy, efficiency, and robustness.

*Corresponding author
Email address: yyue001@e.ntu.edu.sg (Yufeng Yue)

Keywords: Decision Making, Multimodal Information Fusion, Probabilistic Inference, Role Recognition, Situation Assessment.

1. Introduction

Robotics has made exciting advancements in dealing with challenging situations in the past few years. Autonomous robots were employed in many applications, for example, assistant services, or search-and-rescue operations. Some of them can be time-consuming, and potentially life-threatening to human.

Recently, the field has shifted its attention from hardware development to intelligent agents. Robots are expected to perform high-level tasks such as visual searching, or interaction with the human. Previous studies introduced prior knowledge of environments in their systems [1, 2, 3]. This could enable robots to understand their surroundings, aware of situations, and act wisely. However, the aforementioned systems rely on the single modality for analysis and decision-making. When operating in a large scale complex environment, dynamic changes of objects and occlusions are often concerned. Single modality is clearly insufficient and prone to failure in such a situation. Alternatively, multimodality can be applied to handle uncertainties, and enrich the credibility of decisions.

Prior knowledge in this research mainly refers to the spatial relationship between objects. The relationship could be described in spatial spaces or spatio-temporal spaces. The spatial relationship provides essential information for robot navigation, scene understanding, and task planning. Considerable effort has been devoted to the usages of spatial relationships [4, 5, 6, 7]. Many approaches associated visual detection with the spatial relationship to infer the locations of objects [8, 9, 10]. However, fewer studies consider the potential relationships between human and specified objects (including their temporal changes).

This paper aims to infer the role of people and the level of the situation in a dynamic environment. By integrating both spatio-temporal and spatial-semantic features with prior knowledge, probabilistic models can be established to infer human roles. The

target person, who fulfill the target role in a specific mission, is defined as the person who is distinctive in movement or holds a mission-critical object. A mission-critical object is an object that is pre-specified in the corresponding mission. We simply use 'object' in the following content to represent the mission-critical object. Moreover, the situation is assessed into several levels based on the action of the target person and the relative position of this person to the mission-related entities. A mission-related entity is an entity that has a particular function in the corresponding mission. We simply use 'entity' in the following content to represent the mission-related entity. Then, the robot can determine what action it should take according to the inferred situation level.

Main contributions of this work are listed below:

- A probabilistic reasoning framework that can associate prior knowledge with spatial relationships between the humans and objects as well as their temporal changes to infer the implicit information in a dynamic environment.
- Distance-based inference (DBI) and knowledge-based inference (KBI), are established by spatial-semantic and spatio-temporal features with prior knowledge.
- An adaptive fusion scheme for role recognition by using the information entropy of KBI and DBI.
- Situation assessment based on the action of the target person and relative position of this person to the mission-related entities.

The rest of this paper is structured as follows. Section 2 reviews the related works of this research. Section 3 states the problems involved in this research. Section 4 details the theoretical explanation of role recognition. Section 5 introduces the situation assessment. Section 6 demonstrates the simulation of the reasoning process. Section 7 shows the experimental procedures and results. Section 8 concludes the paper and discusses future works.

2. Background and Related Works

This section will review relevant works in the field of role recognition, situation awareness, and information fusion. It is noted that object detection is beyond the scope

of this study, and we just apply the existing method to detect objects.

2.1. Role Recognition

The ability to recognize social roles in real life is crucial to robot's decision-making and task planning. It is initially a subject of Psychology and Sociology [11]. With the rapid development of robotics, role recognition can be solved by visual recognition. By identifying the role of people, robots can determine their possible interactions with the environment and vice versa. Previous research recognized the role in an image by analyzing appearance features in associate with familial relationships [12]. The clothes and their contexts were used to classify the occupation of human [13]. In conversational broadcast, [14] uses behavioral evidence from speaker turns to infer the roles played by different individuals. Some other research determines the role by considering the spatial relation between human-to-human or human-to-objects. The work in [15] observed the spatio-temporal between players to predict the "attacker" role and "defender" role in sports videos. [16] assigned roles to people in different events by tracking their actions along time series. However, most of the existing work focused on recognizing specific roles either in static images or the known scenario. This study handles abstract roles that without particular notions. We investigate spatial relationships between the humans and specified objects as well as their temporal changes to recognize the target person.

2.2. Situation Awareness

Situation awareness is divided into three steps: the perception of environmental elements concerning time or space, the comprehension of their meaning, and the projection of their future behavior [17]. The formalization of situations can be found in [18]. Besides, a systematic overview of computational systems that support situation awareness is provided in [19]. Recently, more related works have been investigated in the field of robotics and autonomous systems, such as autonomous driving [20, 21], intelligent transportation system [22], and disaster monitoring and recovery system [23]. In [20], a situation-aware decision-making problem was modeled as a POMDP for autonomous robots to drive in the urban road safely and efficiently. [22] measured spatio-

temporal trajectory similarity based on longitudinal and lateral spatio-temporal distance to deal with situation classification problems such as lane assignment and driving maneuver recognition. A knowledge-based framework for human-robot collaborative context awareness in USAR missions is proposed in [23]. It used ontological representation to facilitate knowledge sharing and decision-making. However, existing works focused on either spatio-temporal measurements or ontology analysis. The proposed framework integrates these components to create a more comprehensive representation of the situation.

2.3. Information Fusion

Information fusion is a technique to integrate information from different sources. Information fusion can be executed at low to high levels. Data from multiple modalities or heterogeneous sensors can provide additional information so that the system can understand objects more comprehensively and accurately. Traditional information fusion focuses on the integration of raw data from multiple sensors. At low-level fusion, various features can be extracted to improve the performance of the system [24]. In contrast, high-level fusion aims to integrate conceptual level information (such as knowledge) for decision-making, problem-solving, and a better understanding of situation [25]. Ontology made the most contribute to knowledge fusion [26, 27]. It provides a universal understanding of the domains that can be communicated between systems. Various works have been done by high-level knowledge fusion and context awareness [28, 29, 30, 31]. By fusing multimodal information, robots can expand their understanding to a decision level. In [8], a probabilistic model was proposed to infer the possible object locations by utilizing the encoded relationship between objects and the relationship between object and scene. A layered structure of the spatial knowledge representation was designed in [4] to deal with the compound and cross-modal system which is inherently uncertain and dynamic. An approach which exploits semantic knowledge and probabilistic graphical models to enhance object recognition capability of autonomous robots was presented by [32, 33].

However, the above research assumes that the targets are fixed, and no human activities were involved. A probabilistic world model is proposed in [34], which acquired

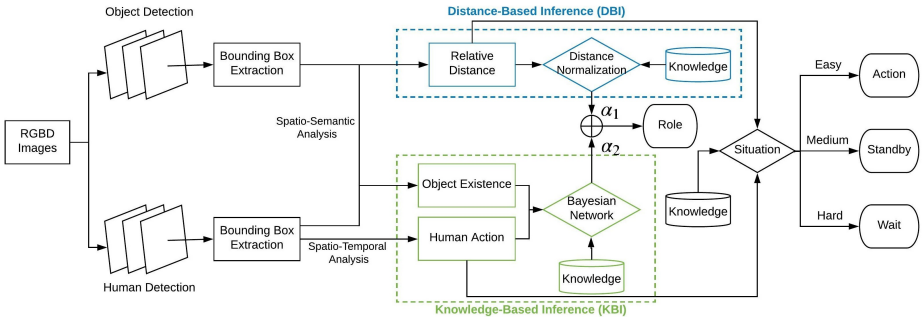


Figure 1: The block diagram for role recognition and situation assessment. The proposed method starts with human and object detection to obtain the corresponding bounding boxes. Next, DBI measures the spatial-semantic feature to find the relative distance between humans and the mission-critical object, while KBI feeds both the object existence and human action into a Bayesian network. Then, the final decision of the human role is made by the weighting of these two indicators. Once the target person is recognized, the level of the situation is assessed by integrating the action of the target person and the relative position of this person to mission-related entities in the space.

data from different sensors and integrated semantic attributes together, to detect the real persons and signs of hazardous materials in urban search and rescue (USAR). Nevertheless, their work still focused on data-level integration, rather than decision-making level. Another method has been proposed in [35], which combined the recognized human activities in the human-robot coexisting environment with the location information of objects, to infer the type of furniture. Most recently, [36] proposed a system to detect and recognize human-semantic-interaction features. The system outputs the person's bounding box, certain actions, and the target object's bounding box with labels. However, the system needs to connect to a wearable device, which is quite limited in real-world applications.

This research considers role recognition, situation awareness and information fusion as a whole. We associate prior knowledge with spatial analysis between humans and objects as well as their temporal changes to infer the implicit information, i.e., roles of human and levels of the situation. The overall block diagram of the proposed framework is shown in Figure 1.

3. Role Recognition

Besides explicit labels and spatial distance between people and object, there are still many attributes that can be used as human inputs, such as body temperature, motion and so forth. In this research, the object existence and human action are chosen as the features that serve as the evidence for role inference. The proposed role recognition method is an association of two indicators, DBI and KBI. DBI decides by measuring the relative spatial distance between humans and the object (if applicable) and then obtains the normalized probability. KBI decides from a Bayesian network which integrates human action and object existence with the prior knowledge. For the final role recognition, it is a weighted fusion of the above two indicated results by using the information entropy. From the knowledge, in a specific mission, it can be assumed that people holding a particular object or performing a specific action can be recognized as the target person (i_1) in the scene; otherwise, it will be recognized as others (i_2), $I \in \{i_1, i_2\}$.

3.1. Distance-Based Inference (DBI)

This indicator is a global probability estimation, which is used to estimate who is more likely to be the target one among a group of people.

3.1.1. Role Inference from Relative Distance Measurement

After obtaining the bounding boxes of humans and the object in the scene, the relative distance between the object and each person is compared by calculating the mean value of the coordinates in the corresponding bounding box. Let $d_{j,t}$ denotes the distance between the object and the jth person at time t, D_t is the sum of distances from the object to each person at time t. The relative distance can be either 2D or 3D. The probability of a person being the target is inversely proportional to the distance to the object. The probability of being the target for the jth person is presented in Eq.(2), where the superscript $\mathcal{D}$ means probability from the DBI. An example is illustrated in Figure 2.

$$D_t = \sum_j d_{j,t} \tag{1}$$

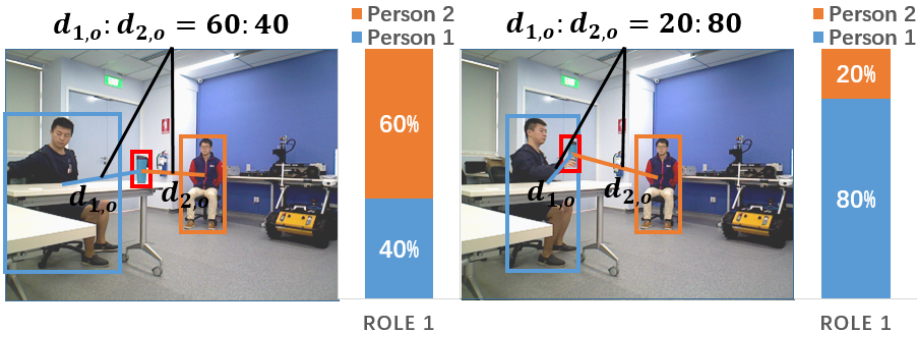


Figure 2: An example of the DBI. It is assumed that the person who is closer to the object becomes the target (i_1). The bar charts show the probability of being the target for Person 1 (blue) and Person 2 (orange).

$$P_j^{\mathcal{D}}(i_{1,t}) = \frac{D_t - d_{j,t}}{D_t}, \qquad P_j^{\mathcal{D}}(i_{2,t}) = \frac{d_{j,t}}{D_t} \tag{2}$$

3.1.2. Confidence of DBI Result

$H_j^{\mathcal{D}}(I_t)$ is the information entropy of the role for the jth person from DBI at time t, which shows the confidence of DBI result. The equation is illustrated as below:

$$H_j^{\mathcal{D}}(I_t) = -\sum_I P_j^{\mathcal{D}}(I_t) \log P_j^{\mathcal{D}}(I_t) \tag{3}$$

3.2. Knowledge-Based Inference (KBI)

This indicator is a local probability estimation, which is used to independently estimate the likelihood of which role a person could be. Prior knowledge can be predefined and directly applied to the robot. In this research, the object existence and human action are selected as the additional features to facilitate the inference. According to knowledge, it can be known that the specific role of people is related to his movements and attachments.

3.2.1. Relationship between Role and Spatial-Semantic Feature Θ^{SS}

In this paper, spatial-semantic feature considers the relative spatial position between humans and other mission-critical objects. The interaction between humans and

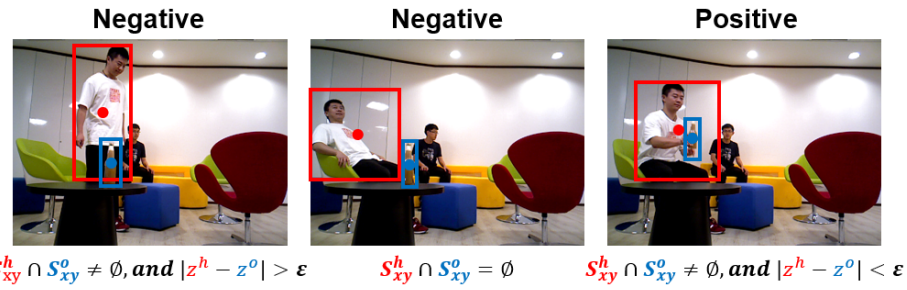


Figure 3: Object existence is defined as whether the specified object exists in the human region. Only when the bounding box of human and object intersect in the whole 3D space, then the existence is positive. S^o_{xy} and S^h_{xy} are collections of all points in xy plane within the $object$'s and $human$'s bounding box, respectively. z is the distance to the camera and ϵ is a threshold for determining the intersection in z direction.

these specified objects is significant in determining the role. A specific role may require people to wear particular clothes or using specific items. For example, "Cook" wearing "Tall Hat" and "Soldier" holding a "Gun" are representative examples.

Objects Existence. In each particular mission, there are critical objects that require particular attention. People carrying such objects can be inferred as the particular role, and their likely future activities may be predicted. Therefore, the existence of such items is critical to recognize the role. E is represented as object existence. It is assumed that all objects are independent and the existence is either positive or negative, which can be denoted as e_1 and e_2, respectively, $E \in \{e_1, e_2\}$. The probabilistic model is represented as $P_j(E_t|I_t)$, which gives the probability of the specified object E attached to the jth person at time t when knowing its role I. If there are multiple objects, the relationship between the role and the existence of the nth object is denoted as $P_j(E_{n,t}|I_t)$.

More specifically, the traditional definition of the object existence is for the entire scene, which will give positive as long as the object is detected in the image. However, in this research, the object existence is defined as whether the object exists in the human region. Let S represents the collection of all points in the detected bounding box. h and o denote $human$ and $object$, respectively. For example, S^h_{xy} means a collection of all points in xy plane within the bounding box of the detected human. z_t is the coordinate of the center point in z axis, which is obtained by calculating the medium

of depth value from all points, $z_t = Median(S_{z,t})$. ϵ is a threshold for determining whether there is an intersection in the depth direction. Only when the bounding box of human and object intersect in the 3D space, then the existence is assigned positive, as illustrated in Fig. 3 and Eq. (4) .

$$E_t = \begin{cases} e_1(Positive) & \text{if } S^h_{xy,t} \cap S^o_{xy,t} \neq \emptyset \; and \; |z^h_t - z^o_t| \leq \epsilon \\ e_2(Negative) & \text{Otherwise} \end{cases} \tag{4}$$

3.2.2. Relationship between Role and Spatio-Temporal Feature Θ^{ST}

In this paper, spatio-temporal feature is determined by spatial variation of humans in time series. Human action is another critical feature to indicate the role. A specific role may require people to perform particular actions. For example, the "Patrol" always moves frequently in their environment and "Waitress" usually bows to the guest.

Human Action. Human actions can reveal their different functions when they are performing certain tasks. Recognizing human actions can help to understand their roles in the environment. Human action is represented as A and it could be of many kinds, such as standing, sitting, lying and so forth, which can be denoted as $A \in \{a_1, a_2, \cdots, a_n\}$. In this research, mainly two actions are considered, $Moving$ and $Stationary$, which are denoted as a_1 and a_2, respectively. The probabilistic model $P_j(A_t|I_t)$ represents the likelihood of taking the action A of the jth person at time t when the role is I.

More specifically, the action of human at each decision frame is analyzed by comparing with its previous frame. The human action is recognized as $Moving$ if the difference exceeded certain level; otherwise, the action is recognized as $stationary$. As illustrated in Fig. 4 and Eq. (5), the 3D position of human at each frame is represented by a single point. Let $\mathbf{c_t}$ denotes the 3D coordinates of the point at the tth frame, $\mathbf{c_t} = (X_t, Y_t, Z_t) = (\delta x_t, \delta y_t, z_t)$. Where X_t, Y_t, Z_t denotes the coordinate in real space, x_t, y_t, z_t denotes the coordinate in pixel, and δ is the scale factor according to the camera. The specific value is obtained by calculating the median of the coordinates of all points within the aforementioned bounding box, which is denoted as $x_t = Median(S_{x,t}),\ y_t = Median(S_{y,t})$. σ is the threshold for determining the

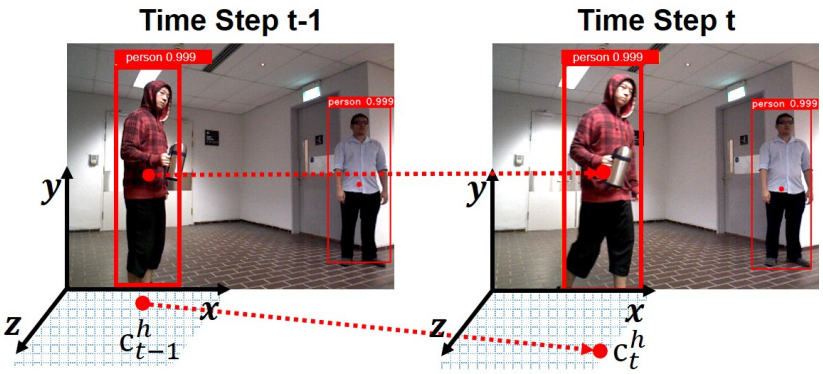


Figure 4: At each decision frame, the action of human is analyzed by comparing with its previous frame. If the difference is below certain level, then the action is recognized as $Stationary$; otherwise, the action is recognized as $Moving$. $\mathbf{c}_t^h$ is the representative point for each person's bounding box at the tth frame.

motion between the consecutive frames.

$$A_t = \begin{cases} a_1(Moving) & \text{if } ||\mathbf{c}_t - \mathbf{c}_{t-1}|| \geq \sigma \\ a_2(Stationary) & \text{if } ||\mathbf{c}_t - \mathbf{c}_{t-1}|| < \sigma \end{cases} \tag{5}$$

3.2.3. Role Inference from Bayesian Network

If there are multiple types of Θ^{SS} or Θ^{ST} features, let m and n denote the number of corresponding features. On the basis of the Bayes rule, the posterior probability of the role is illustrated as below:

$$P_j(\Theta^{SS}_{1,t}, \cdots, \Theta^{SS}_{m,t}, \Theta^{ST}_{1,t}, \cdots, \Theta^{ST}_{n,t}|I_t) \propto \prod_{h=1}^{m} \underbrace{P_j(\Theta^{SS}_{h,t}|I_t)}_{\text{Spatial-Semantic}} \cdot \prod_{k=1}^{n} \underbrace{P_j(\Theta^{ST}_{k,t}|I_t)}_{\text{Spatio-Temporal}} \tag{6}$$

In this research, the role of the jth person can be inferred by fusing of two channels, i.e., human action and object existence. The posterior probability of the role is illustrated as below:

$$\begin{aligned} P_j(I_t|A_t, E_{1,t}, \cdots, E_{n,t}) \propto & P_j(A_t, E_{1,t}, \cdots, E_{n,t}|I_t) P_j(I_{t-1}) \\ = & \underbrace{P_j(A_t|I_t)}_{\text{Human Action}} \cdot \prod_{k=1}^{n} \underbrace{P_j(E_{k,t}|I_t)}_{\text{Existence of Objects}} \cdot P_j(I_{t-1}) \end{aligned} \tag{7}$$

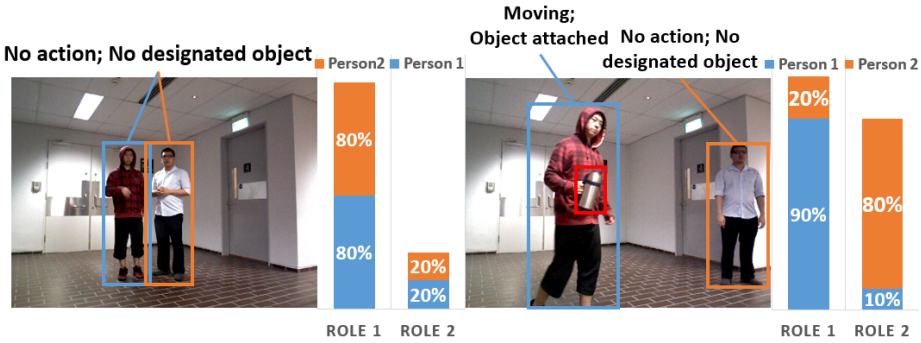


Figure 5: An example of the KBI. It is assumed that people who carry the specified object or conduct specific action becomes the target (i_1). The bar charts show the probability of being the target (i_1) and others (i_2) for Person 1 (Blue) and Person 2 (Orange), respectively.

Initially, the probability of becoming each role is set equal. The last equation is de-
185 rived based on the conditional independence between the two channels of information.

An illustration is shown in Figure 5. Before the final decision, normalization of the probability is performed first. The probability of role type for the jth person is illustrated as below, where the superscript $\mathcal{K}$ means probability from KBI.

$$P_j^{\mathcal{K}}(i_{1,t}) = \frac{P_j(i_{1,t})}{P_j(i_{1,t}) + P_j(i_{2,t})}, \quad P_j^{\mathcal{K}}(i_{2,t}) = 1 - P_j^{\mathcal{K}}(i_{1,t}) \tag{8}$$

3.2.4. Confidence of KBI Result

$H_j^{\mathcal{K}}(I_t)$ is the information entropy of the role for the jth person from KBI at time t. It shows the confidence of KBI result and the equation is illustrated as below:

$$H_j^{\mathcal{K}}(I_t) = -\sum_{I} P_j^{\mathcal{K}}(I_t) \log P_j^{\mathcal{K}}(I_t) \tag{9}$$

3.3. Final Role Recognition

To make the final decision, the results from both DBI and KBI should be combined. Let α_1 and α_2 denote the weight factor for the DBI and KBI, respectively. The weight

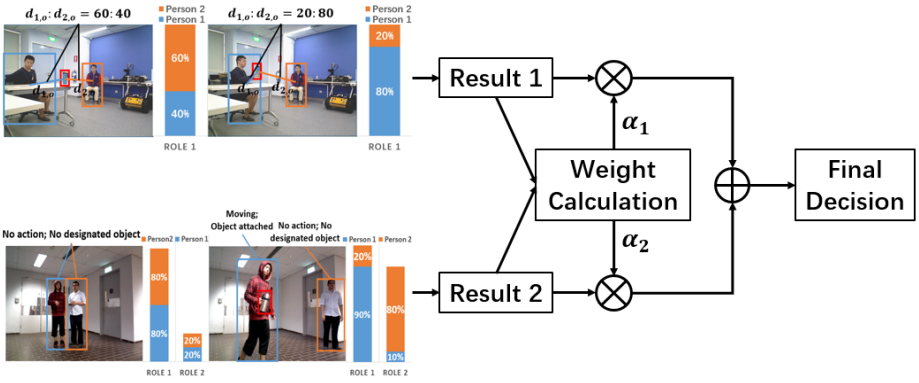


Figure 6: An illustration of the final decision process. After obtaining the inferred results from all indicators, the final role is determined through a weighted fusion of all the results.

factors are defined as below:

$$\alpha_1 = \frac{H_j^{\mathcal{K}}(I_t)}{H_j^{\mathcal{D}}(I_t) + H_j^{\mathcal{K}}(I_t)} \qquad \alpha_2 = \frac{H_j^{\mathcal{D}}(I_t)}{H_j^{\mathcal{D}}(I_t) + H_j^{\mathcal{K}}(I_t)} \tag{10}$$

The less the entropy, the higher the confidence of the inference results. Combining the results of DBI and KBI, we can determine the final role of the jth person as below:

$$P_j(I_t) = \alpha_1 P_j^{\mathcal{D}}(I_t) + \alpha_2 P_j^{\mathcal{K}}(I_t) \tag{11}$$

$$I_{j,t} = \underset{I \in \{i_1, i_2\}}{\arg\max}(P_j(I_t)) \tag{12}$$

Figure 6 illustrates the final decision process. The overall procedure of role recognition is described in Algorithm 1.

4. Situation Assessment

Once the target person is recognized in the scene, the robot needs to start to look into the situation and infer the appropriate interaction with this person. Therefore, the level of the situation is defined as the degree of the appropriateness of dynamic interaction. Different actions and positions of the target person can affect the results

Algorithm 1 Role Recognition

Input: $S^h_{xy,t}$ and $S^o_{xy,t}$
Output: Role of People

function DECISION($Role$)
    1. Distance-Based Inference (DBI)
    Measure the distance between $S^h_{xy,t}$ and $S^o_{xy,t}$
    Probability of being the Target$(i_1) \leftarrow P^{\mathcal{D}}_j(i_{1,t})$
    Calculate confidence $H^{\mathcal{D}}_j$

    2. Knowledge-Based Inference (KBI)
    Analyze A_t and E_t
    Feed A_t and E_t into Bayesian Network
    Probability of being the Target$(i_1) \leftarrow P^{\mathcal{K}}_j(i_{1,t})$
    Calculate confidence $H^{\mathcal{K}}_j$

    3. Final Decision
    Calculate weights α_1, α_2 based on $H^{\mathcal{D}}_j$ and $H^{\mathcal{K}}_j$
    Weighted fusion $\leftarrow P_j(I_t)$
    return Determined role I
end function

of the possible human-robot interaction. By dynamically analyzing the movement and position of the target person in the space, the robot can infer the level of the situation. Moreover, it can determine what action it should take based on the current situation level. In this research, L is expressed as the level of the situation, which illustrates the difficulty of the possible interaction operations. The level can be divided as $L \in \{l_1, l_2...l_n\}$. The action and position are specific for the target person and denoted as 'target action' and 'target position', respectively.

4.1. Relationship between Situation Level and Target Action

The action of the target person is a major factor in changing the environment and affecting the situation. Being able to understand the action of the target person and respond appropriately based on this is critical for a more efficient human-robot interaction. According to knowledge, it can be known that the situation level is closely related to the movement of the target person and in the space. If the target person moves too fast or too frequently, the robot may have difficulty to interact with this target. However, if the target person remains stationary or moves slowly in the scene, then it could

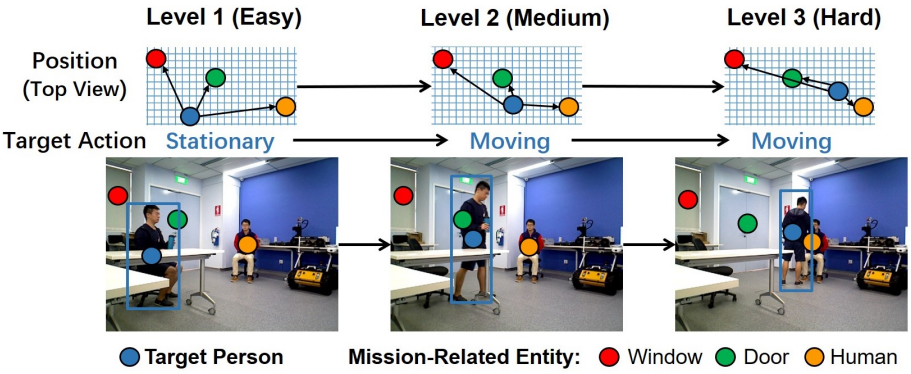


Figure 7: Illustration of situation level inference. By dynamically observing the action of the target person and the relative position of this person to several mission-related entities, the situation can be assessed into different levels.

be much easier for the robot to locate or track the position of this target. $P(A_t^T|L_t)$ gives the probability of the performed action A of the target at time t when knowing the level L.

4.2. Relationship between Situation Level and Target Position

Another key criterion that robots need to measure is the spatial distance between the target person and multiple mission-related entities. Certain entity may have a particular semantic function in the corresponding mission. For example, the door and window are considered passable for humans in the room, but the wall or desk is not. The dynamical distance between the target person and those entities may affect the current situation level. According to the knowledge, the severity of the situation can be either proportional or inversely proportional to the distance between the target person and those entities. Thus, the closest entity to the target person is taken as an important matter. C_t^T is denoted as the closest entity to the target person at time t, and $P(C_t^T|L_t)$ gives the probability of the closest entity C at time t when knowing the level L.

4.3. Inference for the Situation Level

In this research, the situation is assessed into three levels, which is denoted as $L \in \{l_1, l_2, l_3\}$. For instance, when the level $Hard\ to\ operate$ is given, the probability

Algorithm 2 Situation Assessment

Input: $Target$ from Algorithm 1
Output: Situation level and robot decision

function ASSESSMENT($Level\ of\ the\ Situation$)
    Analyze A_t^T and C_t^T
    Feed A_t^T and C_t^T into Bayesian network
    Calculate $P(L_t|A_t^T, C_t^T)$
    if Level of the Situation ← Hard **then**
        Robot Decision ← Wait
    else if Level of the Situation ← Easy **then**
        Robot Decision ← Action
    else
        Robot Decision ← Standby
    end if
end function

of the target is $Moving$ is much higher than that of $Stationary$. When the level is predicted as $Hard$, the robot should stay still and keep observing; when the level is predicted as $Medium$, the robot need to standby and prepare for operation; once the level is predicted as $Easy$, the robot will take the corresponding actions.

$$l_1 = Easy \quad l_2 = Medium \quad l_3 = Hard$$

With the above two channels of information, i.e., action of the target and the closest entity, the level of the situation can be inferred. Specifically, based on the Bayes rule, the posterior probability of the level at time t is formulated as following:

$$\begin{aligned} P(L_t|A_t^T, C_t^T) \propto & P(A_t^T, C_t^T|L_t) \cdot P(L_t) \\ = & \underbrace{P(A_t^T|L_t)}_{\text{Target Action}} \cdot \underbrace{P_t(C_t^T|L_t)}_{\text{Closest Entity}} \cdot P(L_t) \end{aligned} \tag{13}$$

The last equation is derived based on the conditional independence between the two channels of information. The level of the situation at time t (Figure 7) can be determined with respect to the maximum a posterior probability criterion.

$$L_t = \underset{L \in \{l_1, l_2, l_3\}}{\arg\max} \ (P(L_t)) \tag{14}$$

The procedure of situation assessment is described in Algorithm 2.

5. Simulation Study

The performance of the proposed probabilistic model was evaluated through simulation. In this simulation, two persons (Person 1 and Person 2) were set to conduct a series of actions in the environment, and a virtual mission-critical object is placed at black star. It is simulated in an indoor environment, and the person is supposed to be recognized as the target if he is moving frequently or holding the particular object. The situation is assessed by considering the action and position of the target in the space. The performed walking route and standing position of the two persons are demonstrated in Figure 8. Nine types of scenes are considered and simulated as follows, where the scene is defined as the different combinations of individual action and spatial position. Each scenario is set to consume two time steps, where the time step is defined as the period in which people are staying in the same state.

$$s_1 = (T)Walk\ toward\ Window, (O)Walk\ rightward$$
$$s_2 = (T)Stand\ beside\ Window,\ (O)Stand$$
$$s_3 = (T)Walk\ toward\ Robot,\ (O)Stand$$
$$s_4 = (T)Stand\ beside\ Robot,\ (O)Stand$$
$$s_5 = (T)Walk\ toward\ O,\ (O)Stand$$
$$s_6 = (T)Stand\ beside\ O\ pickup\ Weapon, (O)Stand$$
$$s_7 = (T\&O)Walk\ toward\ Door$$
$$s_8 = (T)Walk\ toward\ Window,\ (O)Stand$$
$$s_9 = (T)Stand\ beside\ Window,\ (O)Stand$$

T and O represent for $Target$ and $Others$, respectively. The considered actions are walking (a_1) and standing (a_1) whereas the object existence are considered either positive (e_1) or negative (e_2). Arrows indicate walking trajectories. Depending on the action of the target person and the object existence, the sequence of these scenes can

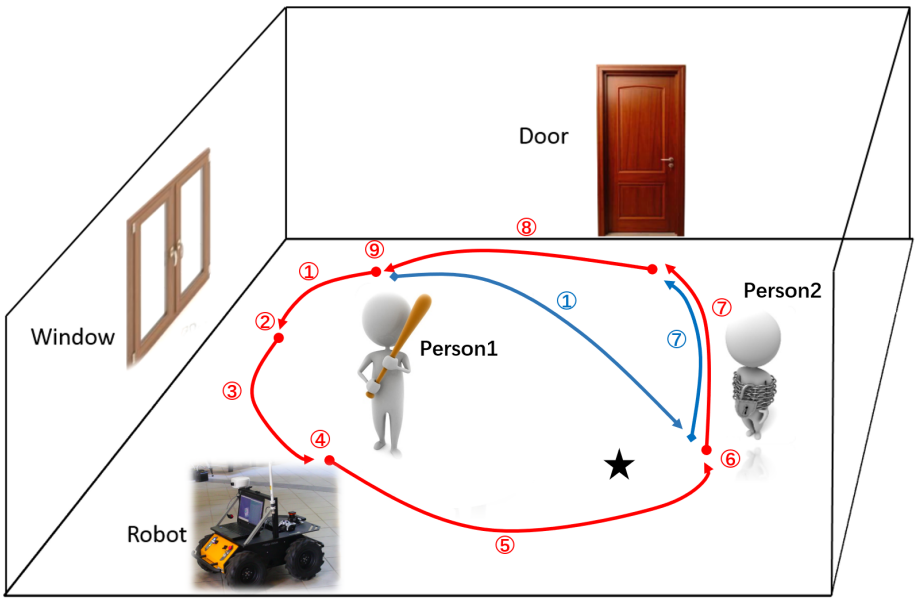


Figure 8: Simulated scenarios in a dynamic indoor environment. The red arrow represents the movement of Person 1 whereas the blue arrow indicates the movement of Person 2. The star represents the place where the object is placed.

be subdivided into:

$$a_1(e_2) \rightarrow a_2(e_2) \rightarrow a_1(e_2) \rightarrow a_2(e_2) \rightarrow a_1(e_2)$$
$$\rightarrow a_2(e_1) \rightarrow a_1(e_1) \rightarrow a_1(e_1) \rightarrow a_2(e_1)$$

5.1. Simulation Results for Role Recognition

Figure 9 displays the probability of being the target for both persons based on each time step. The prior knowledge of the role is initialized to be equal for two persons and the entropy is the largest at the beginning. The probability is updated according to the performed actions and distance between the particular object and human. After a period of observation, it was inferred that Person 1 was the target at the time step 12 whereas Person 2 was the others.

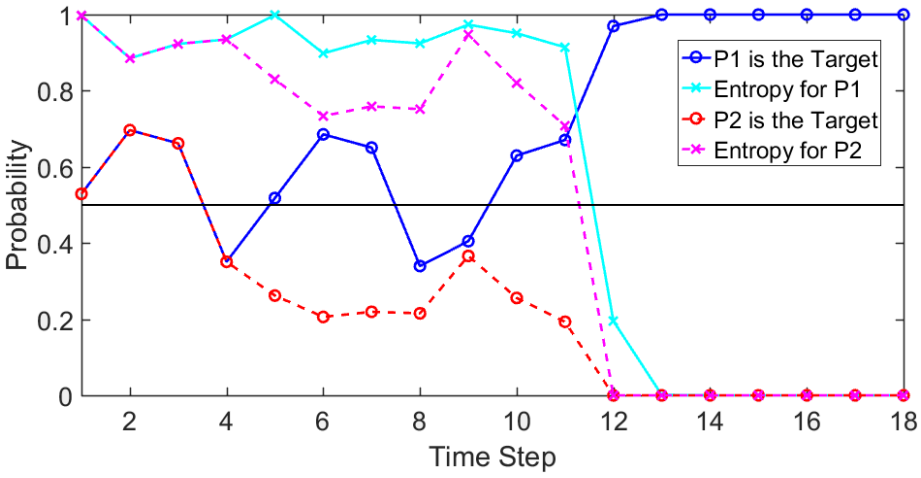


Figure 9: Probability update process of recognizing as the target. It is inferred that Person 1 is the target at the time step 12.

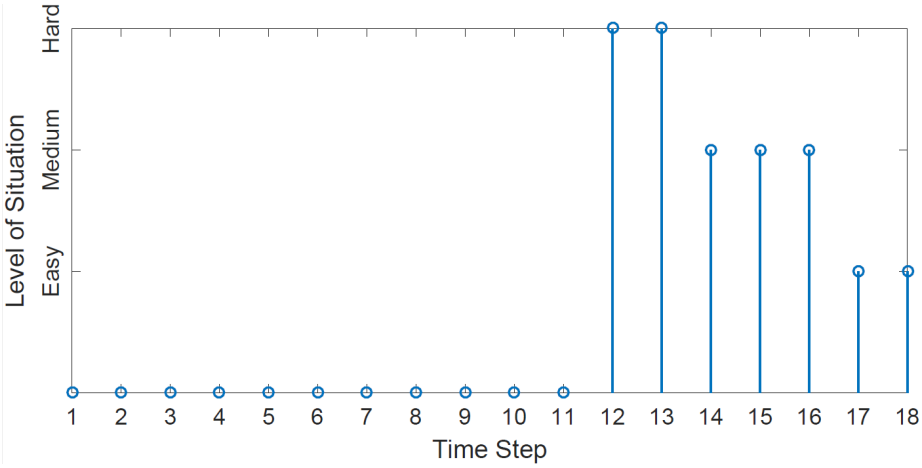


Figure 10: The situation level inference starts after recognizing the target at time step 12. Time steps 12 and 13 are considered to be hard to operate so that the robot will wait; time steps 14 to 16 are considered as the medium level to operate so that the robot will standby; time steps 17 and 18 are inferred as the proper time to take action.

5.2. *Simulation Results for Situation Assessment*

After recognizing the target at time step 12, the robot reasons the level of the situation by integrating the current human action and spatial distance between the target and several entities. The inference results are shown in Figure 10. It predicted that

Table 1: Conditional Probability for Role Recognition in Kidnapping Scenario.

(a) Model $P(A|I)$

		Role (I)	
		Target (i_1)	Others (i_2)
Action (A)	Moving (a_1)	0.54	0.39
	Stationary (a_2)	0.46	0.61

(b) Model $P(E|I)$

		Role (I)	
		Target (i_1)	Others (i_2)
Existence (O)	Positive (e_1)	0.56	0.01
	Negative (e_2)	0.44	0.99

time step 12 or 13 was a wrong time to get involved and time step 17 or 18 might be an excellent chance to take action.

6. Experiments

Based on our research background, an indoor scenario is tested to recognize the target person and assess the situation in this experiment. The sensor used throughout the experiment is the ASUS Xtion Pro Live camera. Faster R-CNN [37] was employed to detect humans and the object in the scene. Since object detection is not the focus of this paper, we just use the existing network parameters which are trained on VOC 2007 [38] and VOC 2012 [39] dataset. Three separate datasets were collected from different indoor environments, as shown in Figure 11. In each dataset, two persons perform a series of actions in the scene. Three different bottles, regarding shape, size and color, are used as the specified object.

6.1. Prior Knowledge Acquisition

Domain knowledge is an important part to infer the role. All parameters are obtained from case study and surveys. By consulting experts (100 policemen) in the field and knowing of the corresponding scenario, prior knowledge could be obtained in advance. On the one hand, it could be inferred that the target might carry some potential weapons and perform more movements in the environment. On the other hand,

Table 2: Conditional Probability for Situation Level Inference in Kidnapping Scenario.

(a) Model $P(A|L)$

		Situation Level (L)		
		Easy (l_1)	Medium (l_2)	Hard (l_3)
Action (A)	Moving (a_1)	0.19	0.40	0.61
	Stationary (a_2)	0.81	0.60	0.39

(b) Model $P(C|L)$

		Situation Level (L)		
		Easy (l_1)	Medium (l_2)	Hard (l_3)
Entity (E)	Closest to Others (c_1)	0.09	0.33	0.57
	Closest to Robot (c_2)	0.26	0.26	0.26
	Closest to Window (c_3)	0.39	0.22	0.05
	Closest to Door (c_4)	0.26	0.19	0.12

prior knowledge about determining the appropriate interaction time and corresponding actions could be leveraged. Action-wise, when the rescuer wants to conduct rescue operations, it is best to participate when the target is in a low alert state. Usually, the targets are vigilant when they are walking around and lower their alertness when they stay still. Position-wise, it should be when the targets are near the windows or in an open space and not close to any others because they do not want to cause any accidental damage to the others.

To represent the prior knowledge in a quantitative way, a survey has been sent out to both ordinary people and the experts in the field. This survey is a questionnaire, by asking how they would behave if they were in this kind of scenario, the statistical distribution of the knowledge could be learned. For example, we asked 'Assuming you are a target, are you going to keep stationary or moving around?', in the end, 54% of the answers chose that they would move around, so the number '0.54' was assigned in (a) of Table 1. Similarly, we asked 'Assuming you are a policeman trying to rescue the others, when is the proper time to conduct the operation?'. As shown in Table 2, 81% of the answers chose that they would act when the target is not moving, and 57% chose not to act if the target is too close to the others. All the numbers are acquired in a similar way, and we weighted 0.7 for the experts' answers whereas 0.3 for the ordinary

people' answers.

6.2. Experimental Description

To evaluate the quality of the proposed reasoning method, a series of scenarios are investigated in this experiment. Let T, O, and E represent for $Target$, $Others$ and $Existence$, respectively. m and s denote the human actions $moving$ and $stationary$. p and n represent for the object existence $positive$ and $negative$. For example, the scenario '$T_m O_s E_n$' denotes that the target person is moving while others stays stationary and the specified object is not existed in the scene. Statistically, these six scenarios can cover almost all the possible situations in real cases.

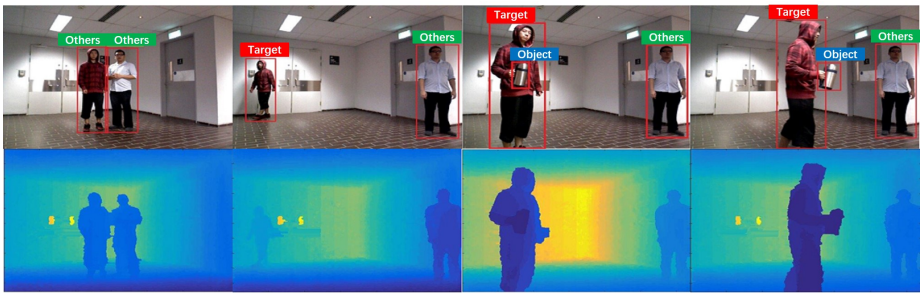
Besides, to get a comprehensive and fair evaluation result, scenario weight γ is introduced to describe the possibility of occurrence of each scenario. The more scenarios there are, the higher the weight assigned, and the less the scenario, the lower the weight assigned. The weight for each scenario is calculated on the basis of Table 1, and the values are shown in Table 3. For each scenario, weight γ is calculated as following:

$$\gamma = P(a_x|i_1)P(a_y|i_2)P(e_z|i_1) \quad x, y, z \in \{1, 2\} \tag{15}$$

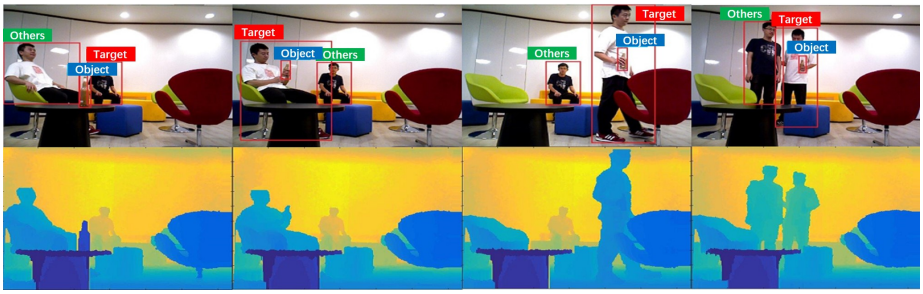
As there are six scenarios considered in this experiment, $\gamma_n \in \{\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6\}$. The final weight γ_n for $Scenario_n$ was then normalized by:

$$\gamma_n = \frac{\gamma}{\sum_{k=1}^{6} \gamma_k} \tag{16}$$

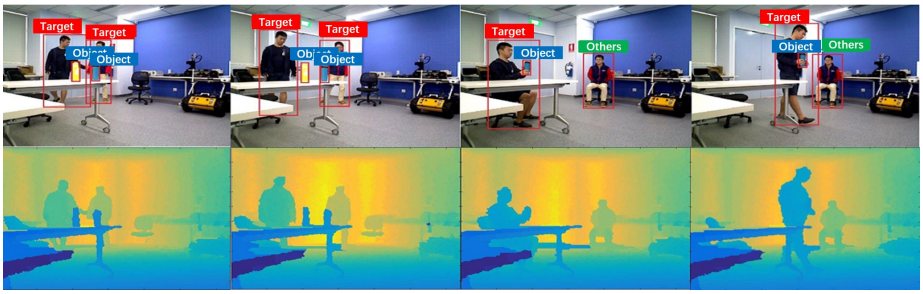
In this experiment, the input features of the decision system are extracted from three different data sources (i.e., RGB image, depth image and RGBD image), and four decision method were evaluated. First, by analyzing RGB images, people and the specified object were detected using Faster R-CNN [37], and the decision is made by measuring 2D distance (d_{2D}) between them. Second, by processing depth images, people and the specified object were detected using template matching [40], and the decision is made by measuring 3D distance (d_{3D}) between them. For these decision methods from distance-based measurement, the target is identified when the object



(a) Dataset 1: This dataset is collected from a corridor. It is a simple environment with large open spaces, limited entities, and almost no occlusion involved. The bottle used is a large silver vacuum cup, it is not in sight at the beginning and later brought in by the person.



(b) Dataset 2: This dataset is collected from a laboratory. It is a typical work area, where many desks are presented. There are many body blocks by the objects. The bottle used is a blue water cup, which is originally placed on the table and later picked up by the person.



(c) Dataset 3: This dataset is collected from a lounge. It is a cluttered environment, which contains many obstacles, such as tables and chairs. More human actions are performed in this environment, and there are many body blocks from the objects as well as each other. The bottle used is a small brown beer bottle.

Figure 11: Recognition results by the proposed method. Experimental data are collected from three separate indoor environments with different settings and human actions. Three bottles are used for the mission-critical object which is different in type, size, and color. Three colors, i.e., red, green and blue, represent the recognition result for the target, others, and specified object, respectively.

is detected to be intersected with this person. Another decision method is through optical flow analysis (*OF*) [41], where the decision is made by measuring the spatial changes between successive time steps, and the person with distinctive movement is identified as the target. The last method use only Knowledge-Based Inference (*KBI*) from analyzing RGBD images, which is the submodule of the proposed method. The proposed approach also works on RGBD images and decides by jointly considering DBI and KBI.

For the situation assessment, Dataset 3 is used as an example to assess the situation level. On the basis of the assessment result, the operator (robot) aims to interact with the person who fulfills the target role. In this experiment, four mission-related entities are taken into consideration, namely, others, the robot, the window and the door. Others are the people do not fulfill the target role; the robot is the operator; the door or window is the entity that is passable for humans. The detection of windows and doors is beyond the scope of this paper. We manually label the boundaries of windows and doors in the environment.

With respect to the position of the target, four types of positions can be allocated, $C \in \{c_1, c_2, c_3, c_4\}$, which is illustrated as follows:

$$c_1 = Closest\ to\ Others \qquad c_2 = Closest\ to\ Robot$$
$$c_3 = Closest\ to\ Window \qquad c_4 = Closest\ to\ Door$$

Thus, the desired target role is the person who is distinctive in movement (moves faster or more frequently) and holds the mission-critical object (bottle). Once the target role is identified, the robot can assess the situation level with respect to the movement of the target person and the position of this person to the above mentioned four mission-related entities. After the robot completes the assessment of the situation, it could decide what action should be taken to interact with the target person.

6.3. Experimental Results

For each dataset, all the frames were split into six scenarios for recognition. Recognition accuracy for both target and others in each examined method was obtained by

Table 3: Evaluation of different decision methods on the six scenarios in three separate datasets. T and O represent recognition accuracy for the corresponding Target and Others. The values in the table are calculated from F-measure. $Score$ is the final weighted measurement of each method. d_{2D}, d_{3D}, OF, KBI and $Proposed$ represent for decision from 2D distance, 3D distance, Optical Flow, KBI and DBI+KBI, respectively.

Method			$Scenario_1$		$Scenario_2$		$Scenario_3$		$Scenario_4$		$Scenario_5$		$Scenario_6$		$Score$	
			$T_mO_mE_n$		$T_sO_sE_n$		$T_mO_sE_n$		$T_sO_sE_p$		$T_mO_sE_p$		$T_mO_mE_p$			
			$\gamma_1 = 0.11$		$\gamma_2 = 0.15$		$\gamma_3 = 0.18$		$\gamma_4 = 0.19$		$\gamma_5 = 0.23$		$\gamma_6 = 0.14$			
Decision	Input		T	O	T	O	T	O	T	O	T	O	T	O	T	O
d_{2D}	RGB Image	Dataset 1	0.0	63.8	1.3	64.8	3.0	66.2	50.9	60.0	0.0	70.1	9.9	65.7	11.8	**65.4**
d_{3D}	Depth Image		0.0	46.6	0.0	46.2	0.0	60.3	0.0	65.1	0.0	64.3	0.0	33.7	0.0	54.8
OF	RGB Image		28.1	0.0	6.5	0.0	33.6	0.0	15.7	0.0	55.1	0.0	19.8	0.0	28.5	0.0
KBI	RGBD Image		40.0	57.1	25.0	25.0	57.1	25.0	73.7	44.4	57.1	40.0	54.6	44.4	**53.2**	38.4
$Proposed$	RGBD Image		40.0	58.1	13.3	43.5	57.1	45.0	73.7	44.4	57.6	40.0	54.6	44.4	51.6	58.2
d_{2D}	RGB Image	Dataset 2	25.7	34.2	0.0	0.0	0.0	6.1	11.1	27.9	0.0	7.7	2.0	0.0	5.2	11.9
d_{3D}	Depth Image		0.0	40.2	0.0	45.8	0.0	29.3	0.0	3.0	0.0	0.0	0.0	0.0	0.0	17.5
OF	RGB Image		41.3	0.0	33.7	0.0	49.5	0.0	16.7	0.0	41.5	0.0	26.7	0.0	35.0	0.0
KBI	RGBD Image		58.8	36.4	50.0	50.0	20.0	22.2	57.1	40.0	60.0	33.3	60.0	33.3	50.6	35.4
$Proposed$	RGBD Image		63.6	42.9	54.6	44.4	26.1	29.1	58.8	36.4	62.1	36.7	62.1	37.1	**54.0**	**37.2**
d_{2D}	RGB Image	Dataset 3	0.0	64.3	0.0	66.7	0.0	66.2	0.0	67.0	28.0	57.3	60.0	51.8	14.8	**62.2**
d_{3D}	Depth Image		0.0	25.8	0.0	23.7	0.0	39.1	0.0	51.4	0.0	12.4	3.2	30.9	0.0	30.4
OF	RGB Image		6.7	0.0	0.0	0.0	30.5	0.0	11.5	0.0	66.1	0.0	13.4	0.0	25.5	0.0
KBI	RGBD Image		40.0	57.2	0.0	66.7	58.8	36.4	50.0	50.0	67.1	40.0	61.5	28.6	48.5	45.5
$Proposed$	RGBD Image		50.0	50.0	57.2	40.0	61.5	48.6	61.6	54.6	57.1	40.0	59.5	41.1	**58.3**	45.6

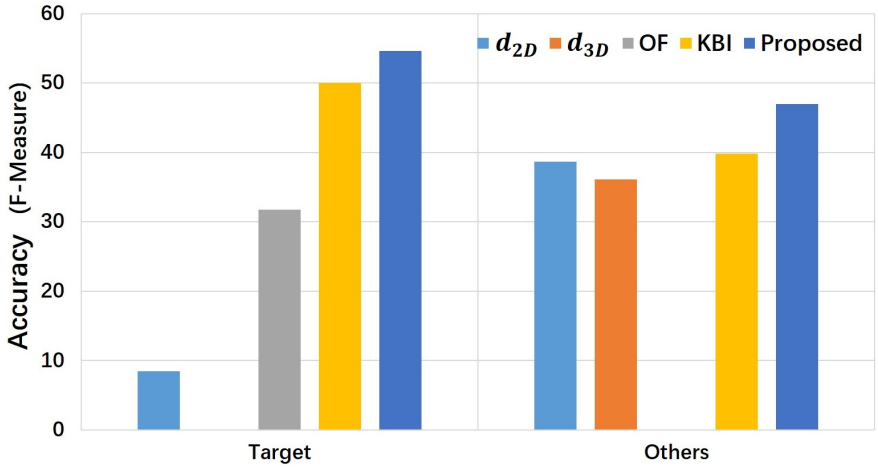


Figure 12: The overall role recognition result for each decision method. The proposed method achieved the highest accuracy in recognizing both target and others.

taking the F-measure [42]. $Score$ is the final measurement for each method by multiplying the F-measure result with the corresponding scenario weight.

6.3.1. Role Recognition Results

Recognition results of each dataset were shown in Table 3 and the average performance of the overall recognition result was revealed in Figure 12. For decisions based on the analysis of 2D images (d_{2D}), it can only identify the target (the person with the bottle) when the bottle is successfully detected, moreover, the performance can be affected due to occlusion and various viewing angles. For decisions based on analyzing depth images (d_{3D}), it is difficult to detect a person when encountering a variety of postures and occlusion. Besides, since a small object detection from only depth data is very challenging, the robot cannot recognize the target because of failing to detect the bottles. For these decision methods using distance measurements, they tended to recognize all detected persons as others due to the limitation of detection. For decisions based on optical flow analysis (OF), it can only recognize the target (the person with distinctive movement) in the presence of a prominent moving speed, but it might fail if people are stationary or moving at similar speeds. Consequently, by combining both DBI and KBI, the proposed method well compensates for the limitation of failing to

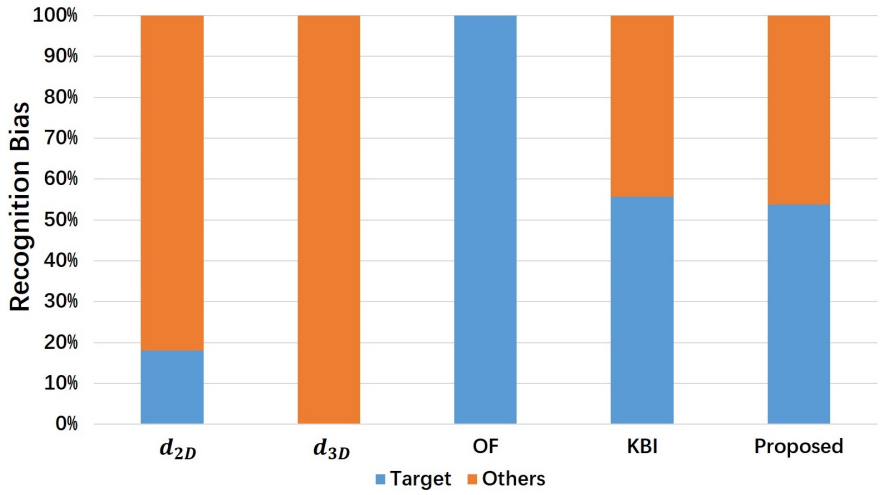


Figure 13: Recognition bias of each decision method. Blue and orange represent the bias that recognizes the target and others, respectively. It can be seen that d_{2D} and d_{3D} are more advantageous in recognizing others while OF can only spot out the target. The proposed method has the most balanced recognition ability for both roles.

detect the single attribute by the ability to fuse information from different perspectives. Compared with other methods in each dataset, it produced the highest accuracy among all methods for target recognition and a competitive result for others recognition. Even in some complex scenes, the performance remained stable and outperformed its sub-module (KBI).

The recognition preference of each method was shown in Figure 13. It revealed that the proposed method has the most balanced recognition ability for both roles. The two decision methods, i.e., d_{2D} and d_{3D}, tend to recognize the people as others since they are hard to detect the bottle. For decisions based on OF, it can recognize the target, but unable to distinguish between target and others.

6.3.2. Situation Assessment Results

For predicting the level of the situation, Dataset 3 is used for evaluation. The target action, closest entity, and the corresponding inferred situation level are displayed in Figure 14. The trajectory of the target, individual decision points and determined easy level points are demonstrated in Figure 15. The 3D map of the environment, the location of the window and door, and two representative easy points are shown in Figure

Figure 14: The inferred result of the level of the situation in Dataset 3. The figure above shows the target action and the closest entity to the target at each decision step. The figure below shows inferred levels of the current situation. It indicates that the situation level is easy at decision steps 7, 9, 11, 32, 34.

16. As a result, it successfully assessed the level and inferred several opportunities for the robot to respond. According to the inferred result, it is appropriate timing for the mobile robot to take action when the target person is in relatively free space, or the target remains stationary.

7. Conclusions and Future Work

In this paper, a comprehensive probabilistic reasoning approach is proposed. The approach enables a mobile robot to recognize the target person based on the identification of the role fulfilled by humans in a mission. The idea behind the approach is to find associations between the prior knowledge and spatial relationships between humans and mission-critical objects as well as the temporal changes concerning humans and mission-related entities. Spatial-semantic and spatio-temporal analysis are used to

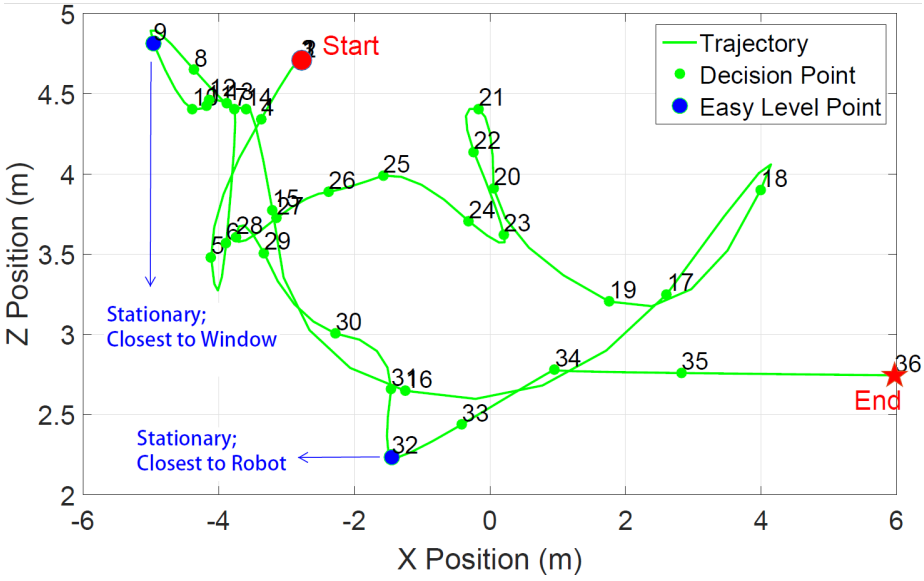


Figure 15: The green curve is the trajectory of the target in Dataset 3. The red dot is the starting point, and the red star is the endpoint. According to Figure 14, green dots are all decision points, and blue dots are determined as easy level.

extract corresponding features. Different inference results are fused into an integrated decision by using the information entropy. As a result, the proposed method can well compensate for the limitations that hard to detect a single attribute and shows a balanced and stable role recognition capability. Besides, it generates reasonable inference results of the situation level and determines the suitable action for interaction with the target person.

The problems of high-level decision support systems are of great significance to the future development of intelligent robots. The operating environment of the robot will be in a highly uncertain, dynamic and complex situation. Efficient human-robot interaction should be performed in a seamless manner to promote the quality of human life and enrich their social experience. In the future work, it is desirable for the robot to identify more human behaviors and be able to relate them to the corresponding scenes. Besides, heterogeneous sensors (e.g., Lidar, thermal camera, etc.) should also be employed to obtain more types of data, thereby generating richer information to enhance the analytical ability and understanding of the situation.

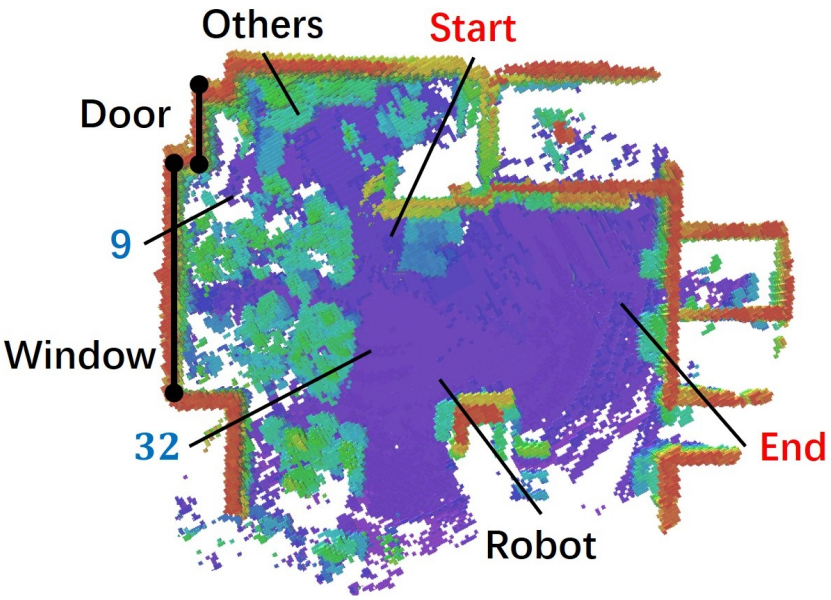


Figure 16: 3D Map of the environment in Dataset 3. Red cubes represent high altitudes while purple cubes indicate low altitudes. Numbers are corresponding locations as shown in Figure 15.

Acknowledgement

The research was partially supported by the ST Engineering - NTU Corporate Lab through the NRF corporate lab@university scheme. We also would like to thank all the participants involved in our experiment for their corporation and patience.

References

[1] K. Shubina, J. K. Tsotsos, Visual search for an object in a 3d environment using a mobile robot, Computer Vision and Image Understanding 114 (5) (2010) 535–547.

[2] A. Aydemir, A. Pronobis, M. Göbelbecker, P. Jensfelt, Active visual object search in unknown environments using uncertain semantics, IEEE Transactions on Robotics 29 (4) (2013) 986–1002.

[3] M. Lorbach, S. Höfer, O. Brock, Prior-assisted propagation of spatial informa-

tion for object search, in: Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on, IEEE, 2014, pp. 2904–2909.

[4] A. Pronobis, K. Sjöö, A. Aydemir, A. N. Bishop, P. Jensfelt, Representing spatial knowledge in mobile cognitive systems, in: 11th International Conference on Intelligent Autonomous Systems (IAS-11), Ottawa, Canada, 2010.

[5] K. Sjöö, A. Aydemir, P. Jensfelt, Topological spatial relations for active visual search, Robotics and Autonomous Systems 60 (9) (2012) 1093–1107.

[6] A. X. Chang, M. Savva, C. D. Manning, Learning spatial knowledge for text to 3d scene generation., in: EMNLP, 2014, pp. 2028–2038.

[7] L. Kunze, C. Burbridge, M. Alberti, A. Thippur, J. Folkesson, P. Jensfelt, N. Hawes, Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding, in: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2014, pp. 2910–2915.

[8] T. Kollar, N. Roy, Utilizing object-object and object-scene context when planning to find things, in: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on, IEEE, 2009, pp. 2168–2173.

[9] A. Pronobis, P. Jensfelt, Large-scale semantic mapping and reasoning with heterogeneous modalities, in: Robotics and Automation (ICRA), 2012 IEEE International Conference on, IEEE, 2012, pp. 3515–3522.

[10] X. Nie, L. L. Wong, L. P. Kaelbling, Searching for physical objects in partially known environments, in: Robotics and Automation (ICRA), 2016 IEEE International Conference on, IEEE, 2016, pp. 5403–5410.

[11] A. Sapru, Automatic social role recognition and its application in structuring multiparty interactions, Ph.D. thesis, École Polytechnique Fédérale de Lausanne (2015).

[12] G. Wang, A. Gallagher, J. Luo, D. Forsyth, Seeing people in social context: Recognizing people and social relationships, Computer Vision–ECCV 2010 (2010) 169–182.

[13] Z. Song, M. Wang, X.-s. Hua, S. Yan, Predicting occupation via human clothing and contexts, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 1084–1091.

[14] H. Salamin, A. Vinciarelli, Automatic role recognition in multiparty conversations: An approach based on turn organization, prosody, and conditional random fields, IEEE Transactions on Multimedia 14 (2) (2012) 338–345.

[15] T. Lan, L. Sigal, G. Mori, Social roles in hierarchical models for human activity recognition, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 1354–1361.

[16] V. Ramanathan, B. Yao, L. Fei-Fei, Social role discovery in human events, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2475–2482.

[17] T. Ziemke, K. E. Schaefer, M. Endsley, Situation awareness in human-machine interactive systems (2017).

[18] M. M. Kokar, C. J. Matheus, K. Baclawski, Ontology-based situation awareness, Information fusion 10 (1) (2009) 83–98.

[19] V. Loia, G. D'Aniello, A. Gaeta, F. Orciuoli, Enforcing situation awareness with granular computing: a systematic overview and new perspectives, Granular Computing 1 (2) (2016) 127–143.

[20] W. Liu, S.-W. Kim, S. Pendleton, M. H. Ang, Situation-aware decision making for autonomous driving on urban road using online pomdp, in: Intelligent Vehicles Symposium (IV), 2015 IEEE, IEEE, 2015, pp. 1126–1133.

[21] W. Schwarting, J. Alonso-Mora, D. Rus, Planning and decision-making for autonomous vehicles, Annual Review of Control, Robotics, and Autonomous Systems (0).

[22] F. Damerow, S. Klingelschmitt, J. Eggert, Spatio-temporal trajectory similarity and its application to predicting lack of interaction in traffic situations, in: Intel-

ligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on, IEEE, 2016, pp. 2512–2519.

[23] R. Chandra, R. P. Rocha, Knowledge-based framework for human-robots collaborative context awareness in usar missions, in: Autonomous Robot Systems and Competitions (ICARSC), 2016 International Conference on, IEEE, 2016, pp. 335–340.

[24] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Information Fusion 14 (1) (2013) 28–44.

[25] L. Snidaro, J. García, J. Llinas, Context-based information fusion: a survey and discussion, Information Fusion 25 (2015) 16–31.

[26] E. G. Little, G. L. Rogova, Designing ontologies for higher level fusion, Information Fusion 10 (1) (2009) 70–82.

[27] R. Dapoigny, P. Barlatier, Formal foundations for situation awareness based on dependent type theory, Information Fusion 14 (1) (2013) 87–107.

[28] G. Cagalaban, S. Kim, Context-aware service framework for decision-support applications using ontology-based modeling, in: Pacific Rim Knowledge Acquisition Workshop, Springer, 2010, pp. 103–110.

[29] A. Smirnov, T. Levashova, N. Shilov, Patterns for context-based knowledge fusion in decision support systems, Information Fusion 21 (2015) 114–129.

[30] H. Aloulou, M. Mokhtari, T. Tiberghien, R. Endelin, J. Biswas, Uncertainty handling in semantic reasoning for accurate context understanding, Knowledge-Based Systems 77 (2015) 16–28.

[31] U. Alegre, J. C. Augusto, T. Clark, Engineering context-aware systems and applications: A survey, Journal of Systems and Software 117 (2016) 55–83.

[32] J.-R. Ruiz-Sarmiento, C. Galindo, J. Gonzalez-Jimenez, Scene object recognition for mobile robots through semantic knowledge and probabilistic graphical models, Expert Systems with Applications 42 (22) (2015) 8805–8816.

[33] J.-R. Ruiz-Sarmiento, C. Galindo, J. Gonzalez-Jimenez, Exploiting semantic knowledge for robot object recognition, Knowledge-Based Systems 86 (2015) 131–142.

[34] J. Meyer, P. Schnitzspan, S. Kohlbrecher, K. Petersen, M. Andriluka, O. Schwahn, U. Klingauf, S. Roth, B. Schiele, O. von Stryk, A semantic world model for urban search and rescue based on heterogeneous sensors, in: RoboCup 2010: Robot Soccer World Cup XIV, Springer, 2010, pp. 180–193.

[35] W. Sheng, J. Du, Q. Cheng, G. Li, C. Zhu, M. Liu, G. Xu, Robot semantic mapping through human activity recognition: A wearable sensing and computing approach, Robotics and Autonomous Systems 68 (2015) 47–58.

[36] G. Gkioxari, R. Girshick, P. Dollár, K. He, Detecting and recognizing human-object interactions, arXiv preprint arXiv:1704.07333.

[37] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[40] L. Xia, C.-C. Chen, J. K. Aggarwal, Human detection using depth information by kinect, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, IEEE, 2011, pp. 15–22.

[41] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, M. Chen, Spatio-temporal analysis for human action detection and recognition in uncontrolled environments, International Journal of Multimedia Data Engineering and Management (IJMDEM) 6 (1) (2015) 1–18.

[42] D. M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.