Probabilistic reasoning for unique role recognition based on the fusion of semantic-interaction and spatio-temporal features

Yang, Chule; Yue, Yufeng; Zhang, Jun; Wen, Mingxing; Wang, Danwei

2019

Yang, C., Yue, Y., Zhang, J., Wen, M., & Wang, D. (2019). Probabilistic reasoning for unique role recognition based on the fusion of semantic-interaction and spatio-temporal features. IEEE Transactions on Multimedia, 21(5), 1195-1208. doi:10.1109/TMM.2018.2875513

https://hdl.handle.net/10356/106302

https://doi.org/10.1109/TMM.2018.2875513;
https://doi.org/210580

Probabilistic Reasoning for Unique Role Recognition Based on the Fusion of Semantic-Interaction and Spatio-Temporal Features

Chule Yang, Yufeng Yue, Jun Zhang, Mingxing Wen and Danwei Wang, *Senior Member, IEEE*

***Abstract*—This paper deals with the problem of recognizing the unique role in dynamic environments. Different from social roles, the unique role refers to those who are unusual in their carrying items or movements in the scene. In this paper, we propose a hierarchical probabilistic reasoning method that relates spatial relationships between interested objects and humans with their temporal changes to recognize the unique individual. Two observation models, *Object Existence Model (OEM)* and *Human Action Model (HAM)*, are established to support role inference by analyzing the corresponding semantic-interaction features and spatio-temporal features. Then, OEM and HAM results of each person are compared with the overall distribution in the scene, respectively. Finally, we can determine the role through the fusion of two observation models. Experiments are conducted in both indoor and outdoor environments concerning different setting, degrees of clutter and occlusions. The results show that the proposed method can adapt to a variety of scenarios and outperforms other methods on accuracy and robustness, moreover, exhibits stable performance even in complex scenes.**

***Index Terms*—Probabilistic Inference, Multimodal Information Fusion, Decision Making, Unique Role Recognition.**

I. Introduction

As a result of the growth in workforce costs, unmanned systems are addressing increasing attention. Depending on a variety of potential applications, it is highly desirable that an intelligent system can autonomously obtain information from the environment and perform high-level tasks without human participation [1]–[3]. Various intelligent systems have been designed to understand human behavior [4], [5] and try to interact with people in their workspaces [6]–[8].

Recently, autonomous systems have been employed to handle the problem of intelligent decision-making and rapid response, e.g., assistant services or search-and-rescue operations. It requires the autonomous system to respond fast in order to improve the operational efficiency of the mission. Besides, in these highly uncertain and dynamic environments, potential targets may only have an abstract description or an ambiguous appearance. Thus, the ability that can infer implicit information from limited knowledge and perception is beneficial for the autonomous system to perform decision level inference. When operating in complex environments, dynamic changes of objects and potential occlusion are often concerned. Decision strategies that rely on low-level features or single attribute are insufficient. Alternatively, integrating information from both spatial and semantic perspectives can handle uncertainties, and enrich the credibility of decisions.

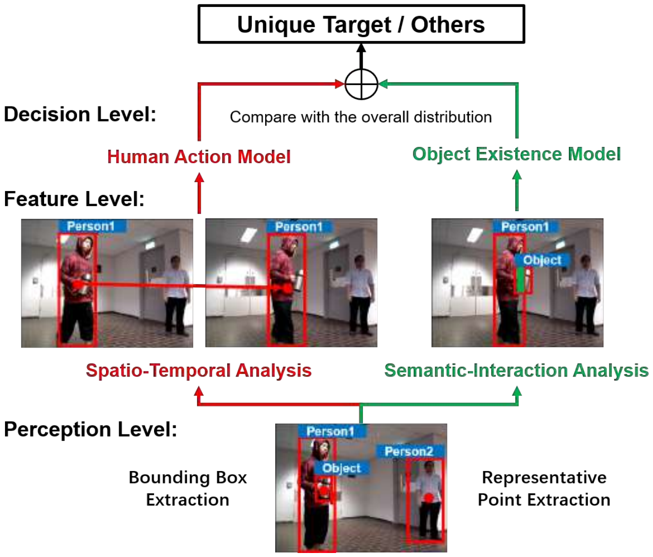


Fig. 1. An illustration of unique role recognition. First, it detects humans and objects of interest for obtaining their corresponding bounding boxes. Next, OEM and HAM are established by analyzing the semantic-interaction and spatio-temporal features, respectively. Then, the inferred result of each individual is compared with the overall distribution. Finally, the unique target can be determined by fusing the results from the two observation models.

By knowing the potential role of individuals, the system can spot out the targeted person and achieve a more comprehensive understanding of the environment. The ability to recognize the unique target is critical for many cases, such as discovering children drinking, looking for people with a rough description. It is initially a subject of Psychology and Sociology [9]. With the development of image processing and machine learning techniques, the perception capability of intelligent systems is drastically increasing. For example, the state-of-the-art visual detection method Region-based Convolutional Neural Network (R-CNN) [10] has performed well on detecting individual objects. However, the deep learning-based method has two main drawbacks when dealing with our problems. 1) It is time-consuming and resource-intensive to train a massive network that encodes multiple relationships among numerous semantic objects. 2) It lacks scalability and adaptability. Objects and actions can vary according to different scenarios or require-

Manuscript received February 1, 2018; revised April 21, 2018, and August 5, 2018; accepted September 24, 2018. This work was partially supported by the ST Engineering - NTU Corporate Laboratory through the NRF Corporate Laboratory@University Scheme. (*Corresponding author: Yufeng Yue*)

The authors are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: {yang0438, yyue001, jzhang061, mingxing001}@e.ntu.edu.sg, edwwang@ntu.edu.sg).

ments. Deep learning-based approaches need to retrain the relationship whenever objects change. Thus, this study separates the decision-making process from the object detection problem. We only exploit the bounding boxes that generated by detection algorithms to make inferences and decisions in any scenarios.

We need to clarify the difference between this study and several related studies. First, this study differs from scene understanding [11] in that it aims to quickly recognize the target in the scene instead of describing the entire scene. Besides, we do not train specific relationships between semantic objects, which may result in untimely responses in real situations. Second, this study is different from the anomaly detection through surveillance video [12] because it uses dynamic vision. The purpose of this study is to recognize the distinctive person who is different from others, not those with abnormal behavior. Thus, we do not have a standard or abnormal pattern as a reference. Finally, this study differs from traditional role recognition [13] in that it does not have a specific notion of role. The roles defined in this research are classified according to different human-object interactions and human actions, rather than specific appearance.

Therefore, this research aims to develop a flexible reasoning strategy that can associate spatial relationships between interested objects with their temporal changes to recognize the target individual who does a specific activity or holds a specified item. The spatial relationship could be described in spatial spaces or spatio-temporal spaces. As an extended version of previous work [14], the focus of this paper is to build observation models by dynamically analyzing semantic-interaction and spatio-temporal features in the whole 3D space. Then, after comparing the observation result of an individual with the overall distribution in the scene, a fusion scheme is applied to integrate all the available information channels to determine the final role. An illustration is shown in Fig. 1.

Main contributions of this work are listed below:

- A probabilistic reasoning scheme is proposed that relates spatial relationships and temporal changes between humans and semantic objects to infer roles of humans in dynamic and unknown environments without training particular relationships.
- Two observation models, object existence model (OEM) and human action model (HAM), are established by analyzing semantic-interaction and spatio-temporal features.
- A hierarchical graphical model that integrates multimodal information into a comprehensive decision scheme by fusing of multiple observation models.

The rest of this paper is structured as follows. Section II reviews the work related to this paper. Section III describes the problem and gives an overview of the proposed method. Section IV explains the features we use in this study. Section V details the theoretical basis of the proposed reasoning method. Section VI shows the experimental procedures and results. Section VII concludes the paper and discusses future works.

II. Related Works

Identifying the unique or desired entity in the scene is a complex problem that requires simultaneous consideration of multimodal information and high-level knowledge from both spatial and semantic perspectives. This section will review works that are closely related to our research in four different fields: role recognition, action recognition, multimodal information fusion and other relevant probabilistic reasoning applications.

A. Role Recognition

Role recognition is mainly a research problem in social activities [13], [15]–[18]. By identifying the role of human, it can help to determine their possible interactions with the environment and vice versa [19]. Some research recognizes the role by analyzing appearance features in a single image, such as training with familial social relationship labels to find the relation between pairs of people [15] or analyzing the clothing and context of human to classify the occupation of human [13]. Some other research determines the role by considering the spatial relation between humans or between human and objects. [16] used training labels to assign the role by observing the spatio-temporal between players to predict the role labels such as "attacker" and "defender" in sports videos. Most recently, [20] proposed work which detected and recognized human-semantic-interaction features to output a triplet which showed the person's box, certain actions, and the target object's box as well as its category. However, existing works focused on recognizing the specific appearance or relationship either on a single image or known scenarios, which is insufficient in the continuously changing and highly uncertain environment. This study handles the abstract role that does not have a particular notion. We simultaneously investigate spatial relationships between semantic objects and their temporal changes to infer the role.

B. Action Recognition

Action recognition is another important work which can reveal the potential significance of entities in 3D space. By analyzing the spatio-temporal information, the people or moving object that carries out specific action is unique in the environment, and more attention needs to be paid to them. Many works have been done on action recognition [21]–[24] and further applied to facilitate autonomous driving [25]–[29] and intelligent surveillance systems [30]–[32].

As robots are becoming more common in our life, some service robots are designed to learn the appropriate way to approach a customer by dynamically observing the human actions [26]. In outdoor applications, autonomous robots need to have a general awareness of the action of critical cars [28] or people [29] to plan their navigation behaviors. Besides, action recognition is critical to detect abnormal scene in the environment. [30] presented an approach to detect the abandoned luggage in surveillance videos. They extracted static foreground regions concerning temporal transition information, then identified the abandoned objects by analyzing the back-traced trajectories of luggage owners. Also based on the foreground object detection, [31] focused on detecting the frequent or infrequent motions in the scene. They used an unsupervised method to infer the background from a subset of frames and compare them with other frames to generate an accurate background subtraction. The off-the-shelf action recognition methods focus on trajectory analysis of a single person or a group of people. However, the actions studied in this study are more involved in analyzing the relative direction and relative speed of all tested people in the scene.

C. Multimodal Information Fusion

Information fusion is a technology to integrate data and information from heterogeneous sources, and it can be handled at multiple levels. Traditional information fusion concentrates on low-level fusion which fused the raw data from multiple sensors. This fusion scheme takes advantage of the characteristics of different modalities to enhance the data authenticity and availability [33]. Data from multiple modalities or heterogeneous sensors can provide us with more information and enable us to understand the environment more comprehensively and accurately. For example, in mobile robot applications, both the RGB-D camera and 3D LiDAR are used to build maps of unknown environments, and then the acquired information from each modality are fused into a consistent global map for precise 3D reconstruction [34], [35]. Another form of sensor fusion is data registration, which is an alignment problem of finding the true transformation from the local frame of each sensor into a common frame. It is of critical importance to the successful deployment of fusion systems in practice. Building a map of the environment is one of the important tasks for SLAM and robotic perception. The method proposed in [36], [37] generated dense 3D models of environments with both appearance and temperature information by combing a thermal infrared camera with a range sensor. As another example, Kinect was used as the mapping tool in the indoor environment because it can detect rich features from RGB-D images. However, if we still want to use it in outdoor, the problem of short measurement distance and sensitivity to light should be solved firstly [38]. Though these methods can obtain richer data from the environment, they are not able to interpret the information.

More recently, the focus has changed to knowledge fusion, which is considered as high-level fusion. It aims to integrate information from the conceptual level into some common knowledge that could be used for decision-making and problem-solving or could provide a better understanding of the situation [39]. Ontology is considered to be the greatest contribution to knowledge fusion [40], [41]. It provides sharing and common understanding of some domains that can be communicated either across multiple information or knowledge sources. Various works have been done in this high-level knowledge fusion and further pushed for context awareness [42]–[45]. By fusing of the multimodal information, the intelligent system can push the understanding into a decision level. In the view of this research, semantic knowledge is illustrated by the relationships between spatial entities and a group of specific concepts. These concepts are intuitive and meaningful for humans and therefore transferred to the system to establish the fundamental or practical knowledge.

D. Probabilistic Reasoning Applications

Some other research also used probabilistic reasoning, such as place categorization and object search. The inference could be made by analyzing the presence of visual features or reasoning of the spatial knowledge. In [46], a probabilistic model was proposed to infer the possible object locations utilizing the encoded relationship between objects and the relationship between object and scene. A layered structure of the spatial knowledge representation was designed in [47] 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 [48], [49]. However, the above research assumes that the targets are fixed and no human activities involved. A probabilistic world model is proposed in [50] which acquired 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, it was still focused on data-level integration, rather than lifting it to the decision-making level. Another method has been proposed in [51], which combined the recognized human activities in the human-robot coexisting environment with the location of objects to infer the type of furniture. Nonetheless, it needed to be connected to a wearable device, which was quite limited in real-world applications.

Thus, this research is more concentrated on using the multimodal information to discover the unique or targeted individual in the entire 3D space over time, rather than on the specific relationship between people in a single image or particular instant. We intend to develop a probabilistic approach to reason the role by simultaneously observing the semantic-interaction feature and measuring the spatio-temporal changes in the full 3D environment.

III. Problem Statement

In this research, the proposed method aims to infer roles of people and recognize the target individual in dynamic and unknown environments. Mathematical problems are formulated into three parts, namely, feature extraction from analyzing spatial relationships and temporal changes between the bounding boxes of semantic objects; model building for each observed feature by using Hidden Markov Model; probabilistic reasoning over the scene based on Markov Random Field. This method infers through multiple forms of information that are scalable and flexible for various situations. Two types of observation models are established, that is, OEM and HAM. OEM analyzes the semantic-interaction feature, which considers the relative spatial position between humans and other semantic objects. Meanwhile, HAM analyzes the spatio-temporal feature, which is determined by spatial variation of humans in time series. Roles are categorized as binary representations, the unique target ($\mathcal{I}_1$) and others ($\mathcal{I}_2$), $I \in \{\mathcal{I}_1, \mathcal{I}_2\}$.

Specifically, we investigate specified objects as the interacted semantics in this research. When the autonomous system is sent into an unknown environment, first, it needs to detect people and objects for acquiring the bounding box

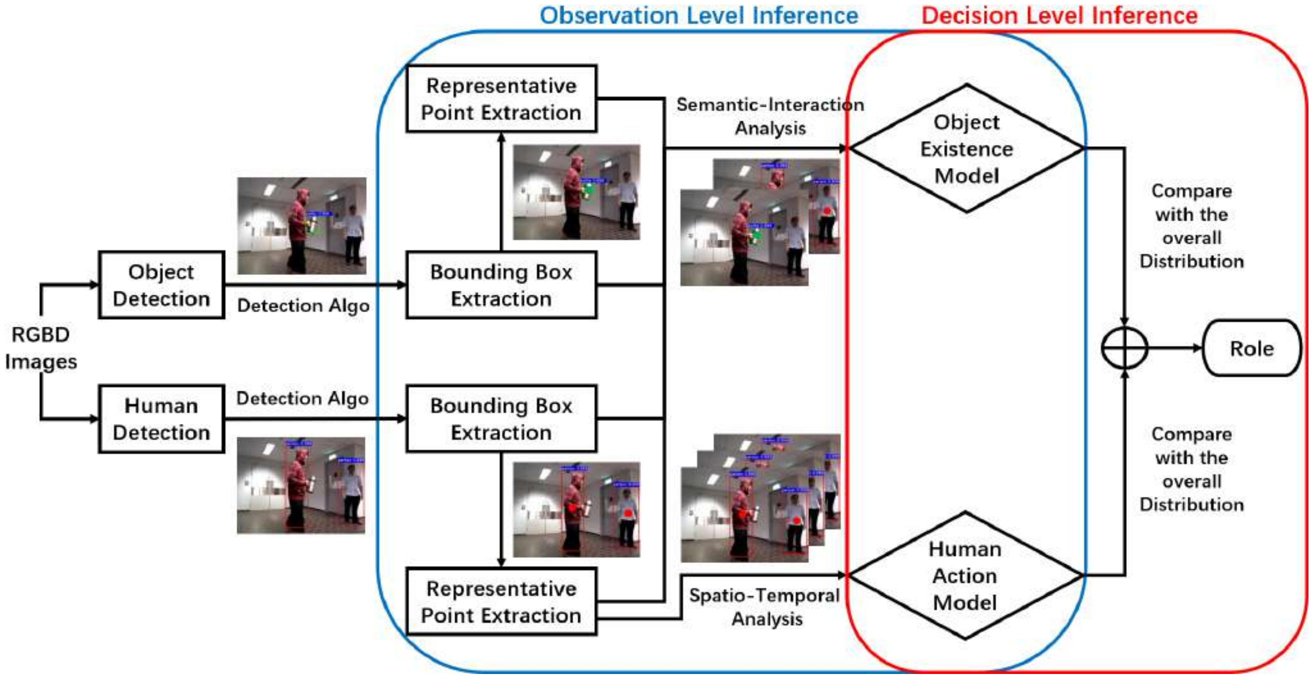


Fig. 2. The diagram of the overall role recognition. Roles of human are inferred from two observations. On the one hand, OEM analyzes spatial relationships between people and specified objects. On the other hand, HAM analyzes temporal changes in human actions and their positions in the space. Then, compare the observation of each individual with the overall distribution in the scene. Finally, the final decision result can be obtained by fusing all channels of information.

of them. Next, the system is going to extract the 2D/3D representative points of people and objects from the obtained bounding boxes. Then, OEM measures the object existence (semantic-interaction feature), while the HAM analyzes the performed human action (spatio-temporal feature). After that, comparing each individual's OEM and HAM results with the overall distribution in the scene, respectively. Finally, the target can be recognized by integrating all observation models into a comprehensive decision scheme. The overall procedures of the proposed method are described in Fig. 2. It can be applied to many tasks, such as delivering items to people with specific attachments, identifying car thieves or recognizing children drinking in public, etc. The detailed theoretical basis is explained in the following parts, including the construction of the probabilistic model, how those channels are integrated, and how final decisions are made.

IV. High-Level Feature Extraction

Besides explicit labels and spatial distance between people and objects, 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 high-level features that serve as the evidence for role inference. It can be assumed that people carrying the specified object and performing a particular action are recognized as the unique target ($\mathcal{I}_1$) in the scene; otherwise, it will be recognized as others ($\mathcal{I}_2$).

A. Human and Object Detection

Since the main idea of role recognition is to analyze the spatio-temporal changes of human position and the real-time distance between the specified object and humans. Therefore, it is a preliminary step to obtain the 3D location of the specified object and involved persons. To achieve this, any detection algorithm can be chosen to generate a bounding box of each person and object. In order to verify the reliability of the proposed decision algorithm, a detection algorithm with high precision and stability is preferred.

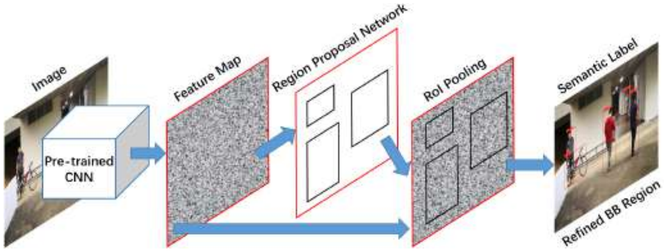


Fig. 3. Schematic diagram of Faster R-CNN. The input is the raw color image and the outputs are corresponding labels and bounding boxes.

Consequently, deep learning technique is considered to be the most appropriate approach, and the Faster R-CNN [52] is selected as the detection method of this study. Overall, Faster R-CNN has a significant increase in speed over its earlier versions, and its accuracy has reached a very high level. It is worth mentioning that although the future model can improve the detection speed, few models can significantly exceed the Faster R-CNN in performance. In other words, Faster R-CNN may not be the easiest and fastest way for object detection, but its performance is still considered as leading at the moment. This detection algorithm is applied to each frame for generating the corresponding labels and regions of the specified object and human, the obtained information is then used as the input of decision algorithm. The standard

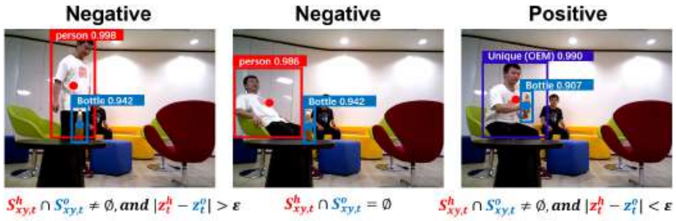

Negative
Negative
Positive
person 0.998
Bottle 0.942
person 0.986
Bottle 0.942
Unique (OEM) 0.990
Bottle 0.907
$S^h_{xy,t} \cap S^o_{xy,t} \neq \emptyset, and\ |z^h_t - z^o_t| > \varepsilon$
$S^h_{xy,t} \cap S^o_{xy,t} = \emptyset$
$S^h_{xy,t} \cap S^o_{xy,t} \neq \emptyset, and\ |z^h_t - z^o_t| < \varepsilon$

Fig. 4. The existence of an object is defined as whether the object intersects with people in the entire 3D space. $S^o_{xy,t}$ and $S^h_{xy,t}$ are the collections of all points in xy plane in the object and human's bounding boxes, respectively. z^o_t and z^h_t are the depth of the corresponding object and human. ε is the threshold that determines the level of intersection in the z direction.

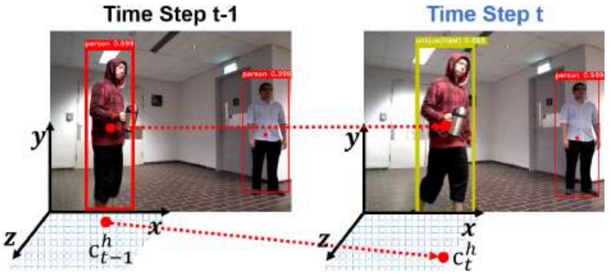


Fig. 5. Human action is described as a vector from the representative point of the previous frame (c^h_{t-1}) to it of the current frame (c^h_t).

procedures are shown in Fig. 3.

B. Semantic-Interaction Feature Θ^{SI}

The interaction between human and certain semantic object is critical in determining the role. A particular role may require people to wear particular clothes or using particular items. For example, "Chef" wearing "Tall Hat" and "Soldier" holding a "Gun" are representative examples.

Many objects have their special functions, and people who hold such objects can be inferred as the certain role and may conduct predictable activities. Therefore, the existence of related items is one of the critical factors for role recognition. In this research, it is assumed that all objects are independent and the object existence is either positive or negative, which are denoted as e_1 and e_2, respectively, $E \in \{e_1, e_2\}$. If there are multiple objects, $E \triangleq \{E_i\}_{i=1:m}$.

Definition 1 (Collection of Points): Let $S^k_{m,t}$ denote the collection of all points in the $m \in \{x, y, z, xy, xz, yz\}$ axis in the detected bounding box of the entity $k \in \{h, o\}$ at the time step t, where 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.

Definition 2 (Representative Coordinate): Let n^k_t denote the representative coordinate in the $n \in \{x, y, z\}$ axis at the time step t, which is obtained by taking the median of the collection of points in the corresponding axis.

$$n^k_t = Median(S^k_{n,t}) \tag{1}$$

More specifically, the traditional definition of object existence is for the entire scene, which gives 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. ε is a threshold for determining whether there is an intersection in the depth direction. Only when bounding boxes of the human and object intersect in the 3D space, the existence is assigned positive, as illustrated in Fig. 4 and (2).

$$E_t = \begin{cases} e_1(Positive), & \text{if } S^h_{xy,t} \cap S^o_{xy,t} \neq \emptyset \text{ and } |z^h_t - z^o_t| \leq \varepsilon \\ e_2(Negative), & \text{Otherwise} \end{cases} \tag{2}$$

C. Spatio-Temporal Feature Θ^{ST}

Human action is another critical feature to infer the role. A particular role may require people to perform particular actions. For example, the "Patrol" always moves frequently in the environment and "Waitress" usually bows to the guest.

Definition 3 (Representative Point): Let c_t denotes the 3D coordinates of the point at the tth time step.

$$c_t = (X_t, Y_t, Z_t) = (\delta x_t, \delta y_t, z_t), \tag{3}$$

where X_t, Y_t, Z_t denote the coordinates in real space, and x_t, y_t, z_t denote the representative coordinates in pixel, and δ is the scale factor according to the camera specification.

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 $A \in \{a_1, \cdots, a_n\}$.

More specifically, the action of people at each time step is analyzed by comparing with its previous position. The 3D position of a person at each time step is represented by a representative point. By measuring the 3D coordinate difference between the consecutive frames, the human action can be deduced, as illustrated in Fig. 5. f is introduced as a mapping function that computes specific actions according to different scenarios or requirements. It can be of many forms, such as absolute values, exponential functions and so forth.

$$A_t = f(c^h_t - c^h_{t-1}) \tag{4}$$

V. Role Inference from Semantic-Spatial Fusion

The objective is to estimate a distribution over a role about what the person could be by giving the human action and object existence as input. This problem is formulated as a Bayesian filter [53], where the hidden variable $I \in \mathcal{I}$, is the role of a person that currently inferred. The robot observes the person's features, G, including semantic-interaction and spatio-temporal features, and estimates a distribution over the current role at each time step:

$$P_j(I_t|G_{1:t}) \tag{5}$$

To estimate this distribution, we alternately perform a measurement update and a time update. The measurement update constructs a decision model from two observation models to update each belief of role, whereas the time update updates the belief of the role given previous information.

$$\begin{aligned} P_j(I_t|G_{1:t}) &= \frac{P_j(G_t|I_t)\cdot P_j(I_t|G_{1:t-1})}{P_j(G_t|G_{1:t-1})} \\ &\propto \underbrace{P_j(G_t|I_t)}_{\text{Decision Model}} \cdot \underbrace{P_j(I_t|G_{1:t-1})}_{\text{Previous Information}} \end{aligned} \tag{6}$$

The measurement update contains all the channels of information by fusing all available observations from potential m semantic-interaction features and n spatio-temporal features.

$$\begin{aligned} P_j(G_t|I_t) &= P_j(\Theta^{SI}_{1,t},\cdots,\Theta^{SI}_{m,t},\Theta^{ST}_{1,t},\cdots,\Theta^{ST}_{n,t}|I_t) \\ &\propto \prod_{h=1}^{m} \underbrace{P_j(\Theta^{SI}_{h,t}|I_t)}_{\text{Semantic-Interaction}} \cdot \prod_{k=1}^{n} \underbrace{P_j(\Theta^{ST}_{k,t}|I_t)}_{\text{Spatio-Temporal}} \end{aligned} \tag{7}$$

The time update contains the transition probability from the previous role to the current role and the previous belief:

$$P_j(I_t|G_{1:t-1}) = \sum_{I_{t-1}\in\mathcal{I}} \underbrace{P_j(I_t|I_{t-1})}_{\text{Decision Transition}} \cdot \underbrace{P_j(I_{t-1}|G_{1:t-1})}_{\text{Previous Belief}}, \tag{8}$$

where the current role is dependent either on the previous role or the previous measurement.

A. Decision Level Inference

1) Decision Model: The decision model calculates the probability of the respective feature given the role. Each feature is a set of human actions and the object existence, $\langle A, E\rangle$, where:

- A represents the performed action.
- E represents the object existence.

Thus, a decision model for a single person j is formed as:

$$P_j(G_t|I_t) = P_j(A_t, E_t|I_t) \tag{9}$$

We factor the features assuming that each modality is independent of the other given the state. Namely, it is assumed that if the real role is known, the probabilities of this person performed action and the object existence are independent:

$$P_j(G_t|I_t) = P_j(A_t|I_t)\cdot P_j(E_t|I_t) \tag{10}$$

By taking the logarithm on both sides, the processed result after the logarithmic function is proportional to the original. Thus, the relationship can be further decomposed as:

$$\begin{aligned} P_j(G_t|I_t) &\propto \log(P_j(G_t|I_t)) \\ &= \log(P_j(A_t|I_t)\cdot P_j(E_t|I_t)) \\ &= \log(P_j(A_t|I_t)) + \log(P_j(E_t|I_t)) \\ &= L_j(A_t|I_t) + L_j(E_t|I_t), \end{aligned} \tag{11}$$

where L denotes the function after taking the logarithm.

2) Decision Transition: Since the environment is unknown, it is assumed that the role of a person is unlikely to change in a short period of time. At each time step, it has a relatively larger probability, q_1, of transitioning to the same role. Otherwise, the transition probability is uniformly assigned based on the total number of roles, $|\mathcal{I}|$.

$$P_j(I_t|I_{t-1}) = \begin{cases} q_1, & \text{if } I_t = I_{t-1} \\ \frac{1-q_1}{|\mathcal{I}|-1}, & \text{otherwise} \end{cases} \tag{12}$$

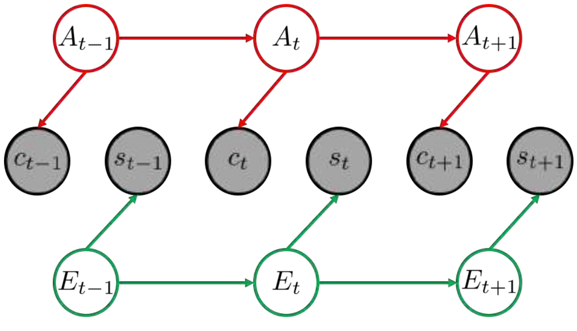


Fig. 6. Graphical models of the two proposed observation models for a single person. Hidden variables are white, and observed variables are gray. The red and green circles are two types of hidden variables from the corresponding OEM and HAM. c and s is the corresponding person's position and the object's bounding box.

B. Observation Level Inference

In this research, the interacted semantics is taken as specified objects. The observation model calculates the probability of the observation given corresponding features. Each observation is a set of object's bounding box and person's position, $\langle s, c\rangle$, where:

- s denotes the bounding box area of the object.
- c denotes the representative point of the detected person.

Since there are two types of features for determining the role of people, two observation models are established for the corresponding object existence and human actions, namely, *Object Existence Model (OEM)* and *Human Action Model (HAM)*. The graphical model for achieving this is illustrated in Fig. 6.

1) Object Existence Model (OEM): OEM contains the observation of the object existence and the transition probability between consecutive states.

a) Object Existence Observation: The object existence is modeled as the intersection area between a person and the specified object at each time step. First, the area of the bounding box of both the person j ($S^h_{j,t}$) and the specified object (S^o_t) are extracted. Let $P_j(s_t|E_t)$ define the probability of the person j carrying the specified object at time t, which is set to be proportional to the area of the intersection. As illustrated in (13), it limits the probability from 0 to 1.

$$P_j(s_t|E_t) \propto \frac{S^h_{j,t}\cap S^o_t}{S^o_t} \tag{13}$$

b) Object Existence Transition: It is assumed that the object existence of a person is likely to be continuous. At each time step, it has a relatively large probability, q_3, of transitioning to the same existence. Otherwise, the transition probability is uniformly assigned based on the total number of existences ($|\mathcal{E}|$).

$$P_j(E_t|E_{t-1}) = \begin{cases} q_3, & \text{if } E_t = E_{t-1} \\ \frac{1-q_3}{|\mathcal{E}|-1}, & \text{otherwise} \end{cases} \tag{14}$$

2) Human Action Model (HAM): HAM contains the observation of human action and the transition probability between consecutive states.

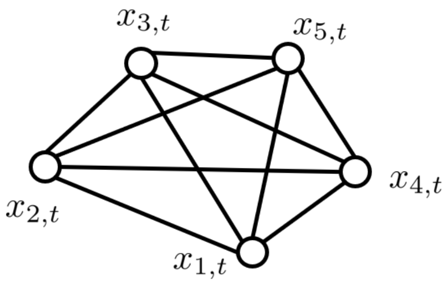


Fig. 7. An example of five detected persons in the scene, which forms a fully connected undirected graphical model. $x_{j,t}$ represents the single inferred result for person j either from OEM or HAM. Then, each single result is compared with each other to measure the deviation from the overall distribution for generating a more comprehensive result.

a) Human Action Observation: Human action is modeled as a positional difference between two consecutive time steps. First, the representative point $c^h_{j,t}$ is extracted for person j at time step t. Next, the movement vector between the two consecutive time steps is calculated. Then, let $P_j(c_t|A_t)$ define the probability of the person j taking a specific action, which is set to be proportional to the corresponding function f of the movement vector. The f function is set to take the absolute value in this research. As illustrated in (15), it limits the probability from 0 to 1.

$$P_j(c_t|A_t) \propto \frac{1}{1+e^{f(c^h_{j,t}-c^h_{j,t-1})}} \tag{15}$$

b) Human Action Transition: It is assumed that the human action feature is likely to be continuous. At each time step, it has a relatively large probability, q_2, of transitioning to the same action. Otherwise, the transition probability is uniformly assigned based on the total number of actions ($|\mathcal{A}|$).

$$P_j(A_t|A_{t-1}) = \begin{cases} q_2, & \text{if } A_t = A_{t-1} \\ \frac{1-q_2}{|\mathcal{A}|-1}, & \text{otherwise} \end{cases} \tag{16}$$

3) Comparison with Overall Distribution: The distance between the same semantic category can reveal the role to a certain extent. For example, the wolf king is always in front of other wolves, or a team leader is usually standing in front of the team. Since these semantic labels belong to the same category, it is not intuitive to recognize the unique role from the appearance-based approaches, but we can get some hints from the spatial distance measurements. Other than the physical spatial distance between subjects, some semantics can also be converted into distance forms such as age, height, and color. Unique or critical entities in the scene may differ from others in these standards.

In this research, final inference results of OEM and HAM are calculated separately by dynamically measuring the degree of deviation of a single inferred result from the overall probability distribution. It is formed as an undirected graphical model, as an example shown in Fig. 7. Let $\mathcal{V} = \{1, 2, \cdots, n\}$ define a set of n nodes (persons) in the scene; $\mathcal{L}$ is a set of undirected edges (l, j), which is linking pairs of nodes (persons). The neighbors of a node (person) j are defined as:

$$\Gamma(j) = \{l \in \mathcal{V} | (l, j) \in \mathcal{L}\}$$

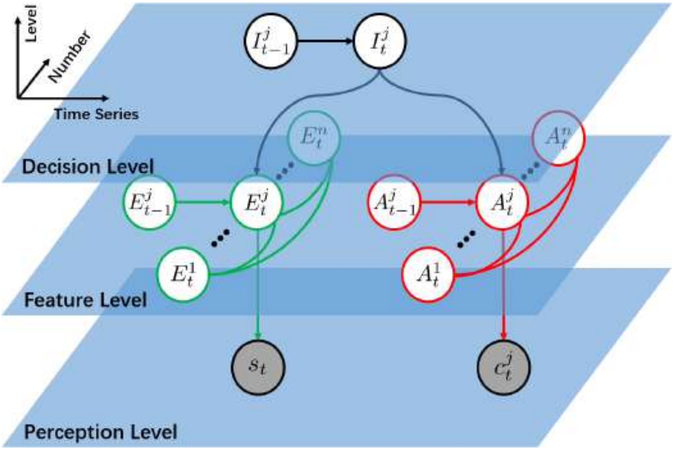


Fig. 8. A hierarchical graphical model of the role recognition for person j at the time t. The inferred result for each individual from either OEM or HAM is interlinked with other persons in the scene. After measuring the degree of the derivation of an individual from the overall distribution, the final role of each person is inferred by fusing the result from both observation models.

Let $x_{j,t} \in \{P_j(s_t|E_t), P_j(c_t|A_t)\}$ define the random variable (inferred result from either OEM or HAM) associated with the node (person) j at the time step t. $\mathcal{C} = \Gamma(x_{j,t})$ is defined as a set of cliques (fully connected subsets) of nodes (persons) to $x_{j,t}$. Then, ϕ_j is described as the degree of derivation of $x_{j,t}$ from the overall distribution in the scene, which can be formed as:

$$\phi_j(x_{j,t}) = \frac{1}{Z} \cdot \prod_{c \in \mathcal{C}} x_{c,t},$$

where Z is a partition function for normalization:

$$Z = \sum_{x_{j,t}} \prod_{c \in \mathcal{C}} x_{c,t}$$

C. Final Role Inference Result

After finishing building the two observation models, these two feature states can be used as evidence to determine the final role. Similar to (6), the final role inference results for each person can be obtained as follows.

1) OEM Results: The inference result from OEM is proportional to the degree of derivation of the existence from the overall distribution and the previous existence.

$$L_j(E_t|I_t) \propto \phi_j(P_j(s_t|E_t)) \cdot P_j(E_t|s_{1:t-1}), \tag{17}$$

where the time update is denoted as:

$$P_j(E_t|s_{1:t-1}) = \sum_{E_{t-1} \in \mathcal{E}} P_j(E_t|E_{t-1}) \cdot P_j(E_{t-1}|s_{1:t-1}) \tag{18}$$

2) HAM Results: The inference result from HAM is proportional to the degree of derivation of human action from the overall distribution and the previous action.

$$L_j(A_t|I_t) \propto \phi_j(P_j(c_t|A_t)) \cdot P_j(A_t|c_{1:t-1}), \tag{19}$$

where the time update is denoted as:

$$P_j(A_t|c_{1:t-1}) = \sum_{A_{t-1} \in \mathcal{A}} P_j(A_t|A_{t-1}) \cdot P_j(A_{t-1}|c_{1:t-1}) \tag{20}$$

TABLE I
COMPARISON WITH THE SCOPE OF EXISTING WORKS.

Works	Abstract Appearance	Spatial Relationship	Dynamic Environment
[15]	√	√	×
[13]	×	×	×
[16]	√	×	√
[18]	√	×	√
[20]	√	√	×
Proposed	√	√	√

3) Models Fusion: As illustrated in Fig. 8, by acquiring the information from all channels, the entire recognition process forms a hierarchical graphical model. Thus, the role of each person can be finally inferred as follows:

$$P_j(I_t|E_{1:t},A_{1:t}) \propto \underbrace{(\underbrace{L_j(E_t|I_t)}_{\text{OEM}} + \underbrace{L_j(A_t|I_t)}_{\text{HAM}})}_{\text{Decision Model}} \cdot \sum_{I_{t-1}\in\mathcal{I}} \underbrace{P_j(I_t|I_{t-1})}_{\text{Decision Transition}} \underbrace{P_j(I_{t-1}|A_{1:t-1},E_{1:t-1})}_{\text{Previous Belief}} \tag{21}$$

The probability that each person becomes the target or others is normalized after each inference, and the following inference is based on the previous normalized value.

VI. EXPERIMENTS

Both indoor (Section VI-D) and outdoor (Section VI-E) environments are investigated. These experiments aim to recognize the unique target ($\mathcal{I}_1$) in the scene. In indoor experiments, the algorithm is evaluated under various degrees of clutter and occlusions. In outdoor experiments, the algorithm is tested with different sizes of specified objects to assess the performance. Since most of the existing works are pre-trained with role-specific features or events, in this experiment, we compared their core ideas rather than specific methods. Besides, the scope of problems that studied in the related works is compared in TABLE I. It shows whether the method can handle such situations: the abstract appearance means that people do not have specific clothing; spatial relationships signify that whether the method analyzes relative spatial position between different semantic objects; dynamic environments indicate whether the method is to process a single image or a video stream. Besides, the parameters and observations of the probabilistic model can be easily modified according to different cases. The imaging equipment used in this experiment was an ASUS Xtion Pro Live RGB-D camera for indoor experiments and a monocular camera for outdoor experiments.

A. Raw Data Preprocessing

Due to the distortion or misalignment of the camera and illumination changes, the raw data acquired from the sensor might be damaged or missed, so preprocessing is needed to improve the quality of the input. In this experiment, the fault data were mainly caused by two problems, i.e., missing data due to no reflection on the glass and distortion on the edge of the RGB-D camera. To solve this problem, a 3×3 median filter was implemented to smooth the image with the size of 640×480. Because most of the fault data in this experiment existed around the edge, so the filter was set to start from the central point and stretch out in four directions. Processing results are presented in Fig. 9.

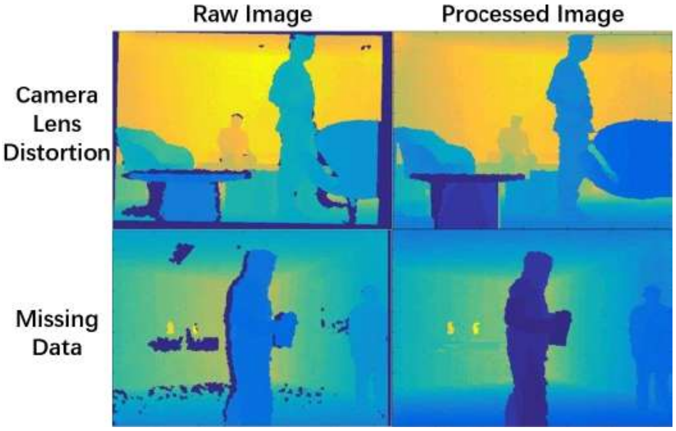


Fig. 9. Results of raw data preprocessing. The two images on the left are the raw depth images, and the two on the right are the images after filtering. The problem with the above case is caused by the distortion of the camera lens, whereas the below one is missing data due to no reflection on the glass.

B. Bounding Box Extraction

There are diverse methods available for object detection. The focus of this research is on the decision-making strategy that based on the analysis of obtained bounding boxes. Therefore, in this experiment, we directly employed the existing Faster R-CNN network parameters that have been trained on VOC 2007 [54] and VOC 2012 [55] datasets to detect humans and objects for acquiring the corresponding bounding boxes. Then, the coordinated depth value for each pixel within the bounding box is extracted from the R-GBD camera.

C. Effect of False Detection on Decision Performance

Since the decision process is based on perception results, the fault detections of objects and humans may affect decision making to varying degrees. Therefore, how false detection affects decision results is a critical part that we need to investigate. As shown in Fig. 10, six different types of false detection of both human and object may result in the false inference of the unique target. Let fp^h and fn^h denote the false positive and false negative detection of human, and fp^o and fn^o denote the false positive and false negative detection of the object. There might also have different conditions under certain type, thus we use subscript 1 and 2 to distinguish them. TP, TN, FP and FN denote the corresponding true positive, true negative, false positive and false negative recognition of the unique target. $Precision$ and $Recall$ are the main metrics for recognition, which is defined as below:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

The quantitative impact of false detection on role decision performance is demonstrated in Fig. 11. This figure displays

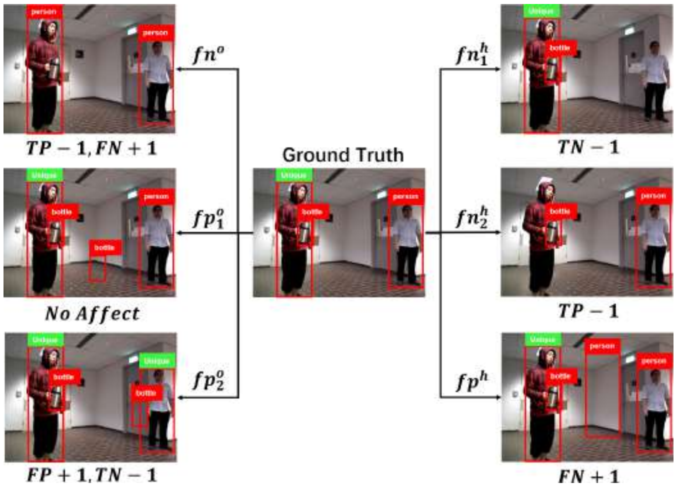


Fig. 10. Different types of false detection and the corresponding impact on role decisions. Six types of false detections are investigated, and the green label is recognized as the unique target. $+1$ and -1 are counts of decision results according to different false detections.

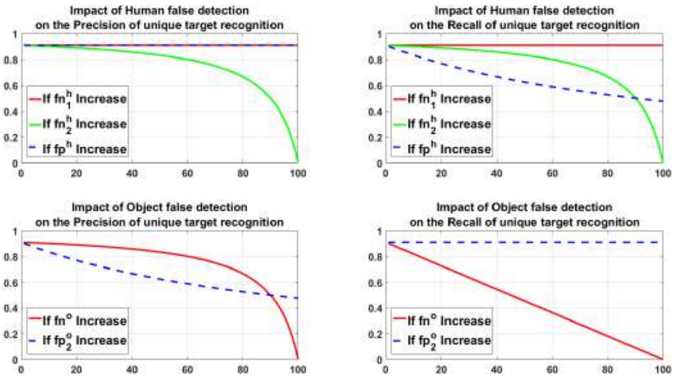


Fig. 11. The changing trend of unique role recognition $Precision$ and $Recall$ according to different types and degrees of false detection.

the changing trend of unique role recognition according to different degrees and types of false detection. For example, the top left figure in Fig. 11 shows that fn_1^h and fp^h do not affect the recognition $Precision$, but the $Precision$ will drop if fn_2^h happens. fn^o and fn_2^h have a relatively larger impact on the performance.

However, bounding box-based detection methods have the inherent defects that they are difficult to accurately describe the boundaries of objects. The inaccurate boundaries may affect the 3D position measurements or the intersection area estimation between humans and the specified object, which could lead to erroneous interaction/relationship identification. This would affect the decision performance to some extent when encountering occlusions, illusions, and changing view perspectives. But this problem can be improved if there is a detection method that can provide more accurate boundary information. The focus of this research is to develop a comprehensive probabilistic decision framework that is independent of the detection algorithms. Hence, the selection of the state-of-art detection approaches is flexible in this framework.

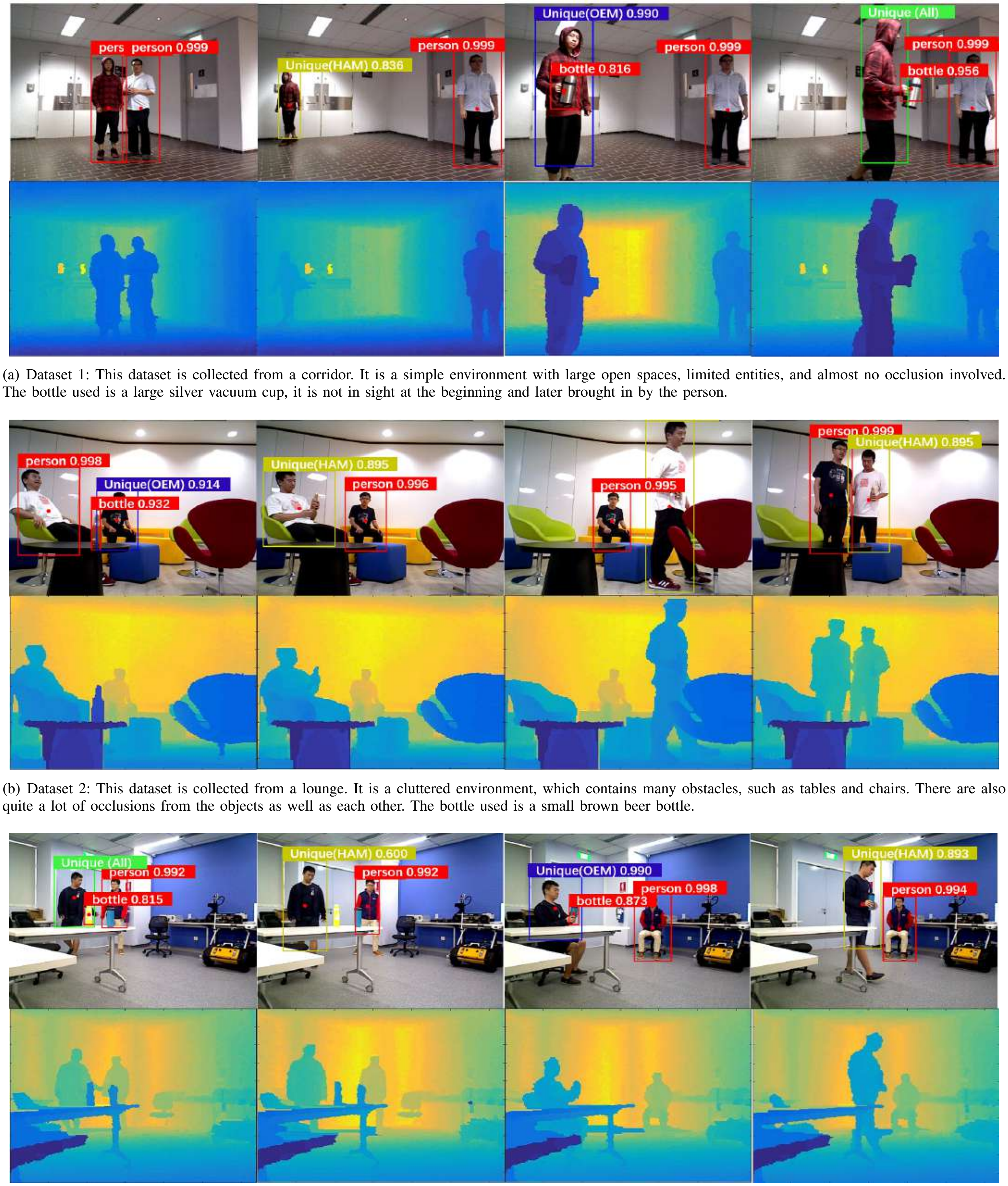
D. Indoor Experiments

In indoor experiments, three separate datasets were collected from different indoor environments regarding different settings and degrees of clutter, as shown in Fig. 12, to comprehensively evaluate the proposed approach concerning human actions and the existence of the specified object. The unique target is the person who moves frequently and carrying on the specified item. Due to limitations, three different bottles are used as the specified object, which is different in type, size, and color.

1) Indoor Experimental Description: In this experiment, the input features of decision algorithms were extracted from three different data sources (i.e., RGB images, depth images and RGB-D images), and four main decision methods were evaluated. First, by analyzing RGB images, the decision is made by measuring 2D distance (d_{2D}) between the bounding boxes [15], [20]. Second, by processing depth images, people and the specified object were detected using template matching [56], and the decision is made by measuring 3D distance (d_{3D}) between them. For these decision methods from distance-based measurement, the person is identified as the unique target only if the object is detected to be intersected with him (her). Another main decision method is through spatio-temporal analysis (*STA*) [16], [18], [57], where the decision is made by measuring the spatial changes between successive time steps, and the person with distinctive moving speed is identified as the unique target. The last one is Knowledge-Based Inference (*KBI*) from analyzing RGB-D images. The decision is made through a Bayesian network based on the prior knowledge [14]. The proposed approach, Multimodal Information Fusion (*MIF*), also works on RGB-D images and decides by jointly considering OEM and HAM.

2) Indoor Experimental Results: Detection results on different modalities for each dataset are shown in TABLE II. Detecting humans on color images was robust, but it was challenging to identify the small object. However, it was tough to discover the target on depth images by using template matching. The human detection result on depth images in dataset 2 was quite low because it contained high illumination in the background, which has a significant impact on the depth image generation.

Recognition results of each examined decision method were obtained by taking the F-measure, and numerical values are shown in TABLE III. The average performance of the overall recognition result was revealed in Fig. 13. As a result, by fusing the information from different modalities, the proposed approach *MIF* can take advantage of each observation to compensate for the failure of a single model to generate robust and reasonable recognition performance. Comparing with other methods in each dataset, it produced the highest accuracy among all methods for recognizing both roles. Even in some complex scenes where there were many occlusions and illusions, the performance remained stable. For decisions based on analyzing 2D distance from RGB images (d_{2D}), it can only recognize the unique target when the specified object is successfully detected, and the performance can be affected by illusions and variant viewing angles. For decisions based on analyzing depth images (d_{3D}), it is hard to detect a person when encountering a variety of postures and occlusions. Be-



(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 lounge. It is a cluttered environment, which contains many obstacles, such as tables and chairs. There are also quite a lot of occlusions from the objects as well as each other. The bottle used is a small brown beer bottle.

(c) Dataset 3: This dataset is collected from a laboratory. It is a typical work area, where many desks are presented. More human actions are performed in this environment, and there are many occlusions from the objects as well as each other. The bottle used is a blue water cup.

Fig. 12. The experimental data were collected from three separate indoor environments with different settings and degrees of clutter. The three bottles used as the specified object which was different in type, size, and color. The four colors of bounding boxes, red represents the detection result yellow and blue represent results from HAM, OEM; green represents the final result by comprehensively considering both observation models.

TABLE II
FEATURE DETECTION ACCURACY (IN %) ON DIFFERENT MODALITIES IN EACH DATASET.

Input Data	Dataset 1		Dataset 2		Dataset 3	
	Object	Person	Object	Person	Object	Person
RGB Image	17.22	96.18	51.13	96.62	5.39	96.27
Depth Image	0.00	37.70	0.00	9.61	0.00	17.93

TABLE III
EVALUATION OF ROLE RECOGNITION PERFORMANCE IN THREE SEPARATE DATASETS. ALL THE VALUES IN THE TABLE ARE CALCULATED FROM F-MEASURE (IN %). d_{2D}, d_{3D}, *STA*, *KBI* and *MIF* REPRESENT FOR DECISION-MAKING FROM 2D DISTANCE, 3D DISTANCE, SPATIO-TEMPORAL ANALYSIS, KNOWLEDGE-BASED INFERENCE AND MULTIMODAL INFORMATION FUSION, RESPECTIVELY.

Decision Method	Input Data	Dataset 1		Dataset 2		Dataset 3	
		Unique	Others	Unique	Others	Unique	Others
d_{2D}	RGB Image	27.14	70.34	26.19	78.46	10.59	67.93
d_{3D}	Depth Image	0.00	54.76	0.00	17.52	0.00	30.41
STA	RGB Image	44.10	78.11	50.70	79.37	62.12	81.32
KBI	RGBD Image	67.40	71.31	64.70	74.66	70.31	67.82
MIF	RGBD Image	**93.32**	**94.96**	**83.22**	**90.44**	**85.41**	**90.38**

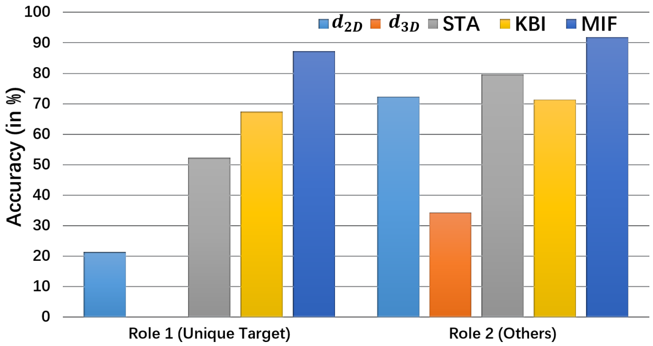


Fig. 13. Overall role recognition result of each decision method by averaging the result of each dataset. The proposed *MIF* achieved the highest accuracy for both unique target and others recognition.

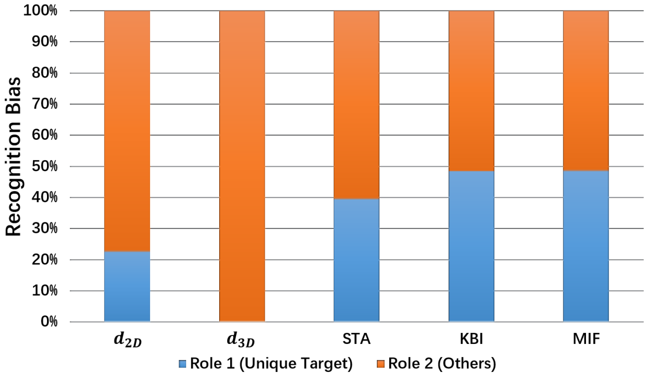


Fig. 14. Recognition bias of each decision method. Blue and orange represent the bias that recognizes the unique target and others, respectively. The proposed approach *MIF* has the most balanced recognition ability.

sides, as detection of the small object is still very challenging, they can not recognize the unique target because of failing to detect the bottle. For these decision methods using distance measurements, they tended to recognize all detected human as others due to the limitation of detection. Thus, they achieved a high precision although their recall was low. For decisions based on the spatio-temporal analysis (*STA*), they were able to pick out the unique person with distinctive moving speed, but it might fail if there was a lack of movements in the scenes or even moving speed. The knowledge-based approach (*KBI*) is viable, but it requires prior knowledge of the specific domain, which is not very flexible for adapting to other situations.

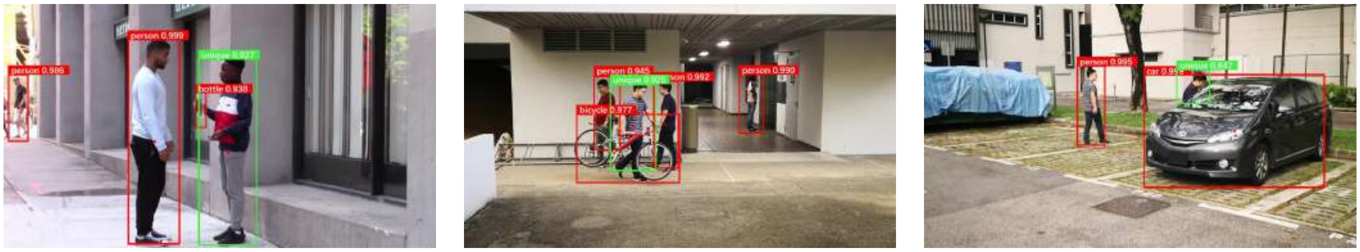
The recognition bias of each method is shown in Fig. 14. It revealed that the proposed method *MIF* had the most balanced recognition ability for both roles. d_{2D} and d_{3D} tended to recognize the people as others since they were hard to detect the specified object.

E. Outdoor Experiments

In this outdoor experiment, another three outdoor scenarios were assessed to recognize the unique target who is different from others concerning the object existence and their performed actions. As shown in Fig. 15, these three scenarios were collected from public places to recognize the people interacting with specified objects or performed distinctive movements. For OEM, it aims to identify the one who is holding or very close to the specified object. For HAM, it is supposed to recognize the one that is moving at a distinctive speed compared to rest of the people in the scene.

1) Outdoor Experimental Description: Three different scenarios were collected from the street, college campus and parking lots, respectively. Each scenario has an specified object, that is, the bottle used to recognize people drinking in public in scenario 1; the bicycle used to recognize people carrying (stealing) the bicycle in scenario 2; the car used to recognize people behind (damaging) the car in scenario 3. The experiment first evaluates two observation models separately and then combines them together for another evaluation.

2) Outdoor Experimental Results: Different from the indoor environment, the outdoor environment had ample space that did not have a lot of occlusions and visual illusions. Therefore, the accuracy of recognition was relatively higher.



(a) People drinking in public.

(b) People carrying (stealing) the bicycle.

(c) People behind (damaging) the car.

Fig. 15. Three different outdoor scenarios for unique role recognition, concerning different sizes of the specified object. The red bounding box is the detected feature, such as persons, the bottle in (a), the bicycle in (b) and the car in (c). The green bounding box is the recognized unique target in the scene.

TABLE IV
FEATURE DETECTION ACCURACY (IN %) IN EACH SCENARIO.

Input Data	Scenario 1		Scenario 2		Scenario 3	
	Object	Person	Object	Person	Object	Person
RGB Image	24.75	99.01	93.59	99.90	99.99	98.97

TABLE V
EVALUATION OF UNIQUE ROLE RECOGNITION IN THREE DIFFERENT SCENARIOS (IN %).

Decision Method	Input Data	Scenario 1		Scenario 2		Scenario 3		Decision Time
		Precision	Recall	Precision	Recall	Precision	Recall	$(10^{-4}s)$
d_{2D} (OEM)	RGB Image	74.63	27.03	87.22	88.90	97.05	91.86	2.00
STA (HAM)	RGB Image	63.10	53.00	78.96	76.12	27.40	14.91	6.00
MIF (OEM+HAM)	RGB Image	**92.17**	**81.82**	**88.85**	**89.14**	**98.21**	**99.79**	6.80

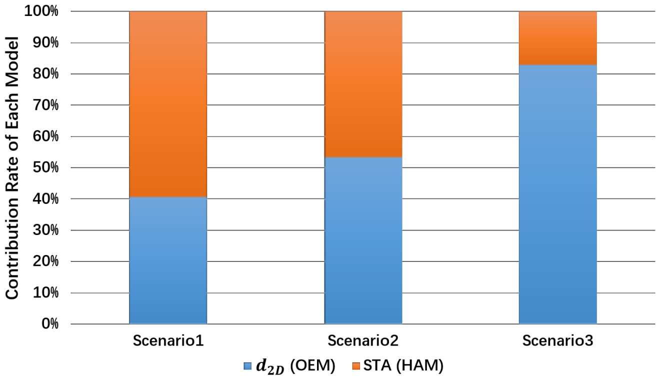


Fig. 16. Contribution rate of each observation model. It showed that the larger the size of the object, the higher the contribution of the OEM. Conversely, the smaller the size of the object, the higher the contribution of the HAM.

The results of human and object detection are shown in TABLE IV. Detecting larger objects like bicycles and cars is very accurate. However, because of the possible occlusions, changing lighting conditions and various poses, identifying small objects such as bottles is still challenging.

As shown in TABLE V, the result validated that the proposed method could tolerate single model failure and fuse them into a comprehensive decision scheme to achieve higher accuracy. Decision time refers to the time from the analysis of the bounding box to the completion of the decision. It showed that the decision time on a regular Intel i7 CPU is speedy enough to handle real-time tasks. Since the OEM and HAM work in parallel, the total time spent is more than a single model, but less than the sum of their time. Taking F-measurement of each method and averaging results of all three scenarios, then, by comparing with the detection result in each scenario IV, the contribution rate of each observation model is displayed in Fig. 16. It illustrated that the proposed method *MIF* relied more on HAM when the size of the specified object was small since OEM had difficulty detecting small objects. In contrast, the detection of these objects was more accurate as the size of the specified object became larger, so the proposed method *MIF* would dependent more on OEM.

VII. CONCLUSIONS

In this paper, a probabilistic reasoning approach was proposed that enables the intelligent system to recognize the unique person in the scene. The idea behind the approach is to associate spatial relationships between humans and the specified object with the temporal changes concerning human positions in dynamic environments. Semantic-interaction and spatio-temporal features were used to form corresponding observation models, namely, OEM and HAM. Then, a hierarchical probabilistic model was established based on the joint relationship of both observation models. By investigating the relationship between single observation and the overall distribution, the proposed method (*MIF*) fused different inferred results into an integrated decision. We also evaluated the effects of false detection on role decision performance. In summary, the proposed method well compensated for the limitation of single model failure and produced the highest accuracy among all methods of recognizing the unique person in the scene. Moreover, it achieved a timely and robust performance even in some complex scenes.

Autonomous decision-making is of great significance for the future development of intelligent systems. The operating environment of future intelligent systems will have a high degree of uncertainty and complexity. Since the reliable decision process depends on comprehensive and precise perception results. Efficient human-computer interaction should be carried out in a seamless manner to improve the quality of human life and enrich the social experience. In future work, it is desired that more accurate detection algorithms can be utilized; more human behaviors can be identified; the behaviors can be associated with corresponding scenes. Besides, heterogeneous sensors should be employed to obtain more types of data that can generate richer information to promote the analytical ability and understanding of the situation.

References

[1] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, "Vision-based state estimation for autonomous rotorcraft mavs in complex environments," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1758–1764.

[2] C. Fu, A. Carrio, and P. Campoy, "Efficient visual odometry and mapping for unmanned aerial vehicle using arm-based stereo vision pre-processing system," in *Unmanned Aircraft Systems (ICUAS), 2015 International Conference on*. IEEE, 2015, pp. 957–962.

[3] S. Minaeian, J. Liu, and Y.-J. Son, "Vision-based target detection and localization via a team of cooperative uav and ugvs," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 7, pp. 1005–1016, 2016.

[4] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 23, no. 11, pp. 1993–2008, 2013.

[5] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[6] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[7] J. H. Oh, A. Suppé, F. Duvallet, A. Boularias, L. E. Navarro-Serment, M. Hebert, A. Stentz, J. Vinokurov, O. J. Romero, C. Lebiere *et al.*, "Toward mobile robots reasoning like humans." in *AAAI*, 2015, pp. 1371–1379.

[8] K. Charalampous, I. Kostavelis, and A. Gasteratos, "Recent trends in social aware robot navigation: A survey," *Robotics and Autonomous Systems*, 2017.

[9] A. Sapru and H. Bourlard, "Automatic recognition of emergent social roles in small group interactions," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 746–760, 2015.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[11] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, 2015.

[12] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 683–695, 2017.

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

[14] C. Yang, Y. Zeng, Y. Yue, P. Siritanawan, J. Zhang, and D. Wang, "Knowledge-based role recognition by using human-object interaction and spatio-temporal analysis," in *Robotics and Biomimetics (ROBIO), 2017 IEEE International Conference on*. IEEE, 2017.

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

[16] T. Lan, L. Sigal, and 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.

[17] H. Salamin and A. Vinciarelli, "Automatic role recognition in multiparty conversations: An approach based on turn organization, prosody, and conditional random fields," *IEEE Transactions on Multimedia*, vol. 14, no. 2, pp. 338–345, 2012.

[18] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Chun Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4576–4584.

[19] B. J. Biddle, "Recent developments in role theory," *Annual review of sociology*, vol. 12, no. 1, pp. 67–92, 1986.

[20] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," *arXiv preprint arXiv:1704.07333*, 2017.

[21] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2555–2562.

[22] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[23] ——, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[24] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[25] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726–1743, 2013.

[26] O. A. I. Ramírez, H. Khambhaita, R. Chatila, M. Chetouani, and R. Alami, "Robots learning how and where to approach people," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 347–353.

[27] C. Dondrup and M. Hanheide, "Qualitative constraints for human-aware robot navigation using velocity costmaps," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 586–592.

[28] A. G. Cunningham, E. Galceran, R. M. Eustice, and E. Olson, "Mpdm: Multipolicy decision-making in dynamic, uncertain environments for autonomous driving," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1670–1677.

[29] D. Mehta, G. Ferrer, and E. Olson, "Autonomous navigation in dynamic social environments using multi-policy decision making," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 1190–1197.

[30] K. Lin, S.-C. Chen, C.-S. Chen, D.-T. Lin, and Y.-P. Hung, "Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1359–1370, 2015.

[31] Y. Lin, Y. Tong, Y. Cao, Y. Zhou, and S. Wang, "Visual-attention-based background modeling for detecting infrequently moving objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1208–1221, 2017.

[32] J. Huang, G. Li, N. Li, R. Wang, and W. Wang, "A violence detection approach based on spatio-temporal hypergraph transition," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2017, pp. 218–229.

[33] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[34] Y. Yue, P. Senarathne, C. Yang, J. Zhang, M. Wen, and D. Wang, "Hierarchical probabilistic fusion framework for matching and merging of 3d occupancy maps," *IEEE Sensors Journal*, 2018.

[35] ——, "Probabilistic fusion framework for collaborative robots 3d mapping," in *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 2018, pp. 488–491.

[36] S. Vidas, P. Moghadam, and M. Bosse, "3d thermal mapping of building interiors using an rgb-d and thermal camera," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2311–2318.

[37] S. Vidas and P. Moghadam, "Heatwave: A handheld 3d thermography system for energy auditing," *Energy and Buildings*, vol. 66, pp. 445–460, 2013.

[38] A. Geiger, F. Moosmann, Ö. Car, and B. Schuster, "Automatic camera and range sensor calibration using a single shot," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3936–3943.

[39] L. Snidaro, J. García, and J. Llinas, "Context-based information fusion: a survey and discussion," *Information Fusion*, vol. 25, pp. 16–31, 2015.

[40] E. G. Little and G. L. Rogova, "Designing ontologies for higher level fusion," *Information Fusion*, vol. 10, no. 1, pp. 70–82, 2009.

[41] R. Dapoigny and P. Barlatier, "Formal foundations for situation awareness based on dependent type theory," *Information Fusion*, vol. 14, no. 1, pp. 87–107, 2013.

[42] G. Cagalaban and 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.

[43] A. Smirnov, T. Levashova, and N. Shilov, "Patterns for context-based knowledge fusion in decision support systems," *Information Fusion*, vol. 21, pp. 114–129, 2015.

[44] H. Aloulou, M. Mokhtari, T. Tiberghien, R. Endelin, and J. Biswas, "Uncertainty handling in semantic reasoning for accurate context understanding," *Knowledge-Based Systems*, vol. 77, pp. 16–28, 2015.

[45] U. Alegre, J. C. Augusto, and T. Clark, "Engineering context-aware systems and applications: A survey," *Journal of Systems and Software*, vol. 117, pp. 55–83, 2016.

[46] T. Kollar and 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.

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

[48] J.-R. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez-Jimenez, "Scene object recognition for mobile robots through semantic knowledge and probabilistic graphical models," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8805–8816, 2015.

[49] ——, "Exploiting semantic knowledge for robot object recognition," *Knowledge-Based Systems*, vol. 86, pp. 131–142, 2015.

[50] J. Meyer, P. Schnitzspan, S. Kohlbrecher, K. Petersen, M. Andriluka, O. Schwahn, U. Klingauf, S. Roth, B. Schiele, and 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, 2011, pp. 180–193.

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

[52] S. Ren, K. He, R. Girshick, and 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.

[53] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*. MIT press, 2005.

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

[55] ——, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[56] L. Xia, C.-C. Chen, and 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.

[57] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, and M. Chen, "Spatio-temporal analysis for human action detection and recognition in uncontrolled environments," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 6, no. 1, pp. 1–18, 2015.

Chule Yang received his B.Eng. degree in Electrical Engineering from Wuhan University, Wuhan, China, in 2014. He received his M.Sc. degree in Computer Control and Automation from Nanyang Technological University, Singapore, in 2015. He is currently a Ph.D. student with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include multimodal perception, probabilistic reasoning and intelligent decision making for autonomous system.

Yufeng Yue received his B.Eng. degree in automation from Beijing Institute of Technology, Beijing, China in 2014. He is currently a Ph.D. student with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include collaborative mapping, multi-robot coordination, multi-robot information fusion and reasoning.

Jun Zhang received his B.Eng. degree in mechanical engineering from Huazhong University of Science and Technology (HUST) in 2012, and the M.Eng. degree in mechatronics from HUST, Wuhan, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological Univrsity. His research focuses on place recognition and localization, and multi-sensor calibration.

Mingxing Wen received his B.Eng. degree of Automation and M.Sc. degree of Computer Control and Automation from Beijing Institute of Technology, Beijing, China in 2015 and Nanyang Technological University, Singapore in 2016 respectively. After working as a Research Associate in NTU, he is currently pursuing the Ph.D. in School of Electrical and Electronic Engineering, NTU. His research focuses on the robot learning, and reinforcement learning applications in autonomous system.

Danwei Wang is leading the autonomous mobile robotics research group. He received his Ph.D and MSE degrees from the University of Michigan, Ann Arbor in 1989 and 1984, respectively. He received his B.E degree from the South China University of Technology, China in 1982. Since 1989, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Currently, he is professor and the co-director of the ST Engineering NTU Corporate Laboratory. He is a senator in NTU Academics Council. He has served as general chairman, technical chairman and various positions in international conferences, such as ICARCV and IROS conferences. He is an associate editor for the International Journal of Humanoid Robotics and served as an associate editor of Conference Editorial Board, IEEE Control Systems Society from 1998 to 2005. He was a recipient of Alexander von Humboldt fellowship, Germany. His research interests include robotics, control theory and applications.