ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/320058025

Knowledge-Based Role Recognition by Using Human-Object Interaction and Spatio-Temporal Analysis

Conference Paper · December 2017

DOI: 10.1109/ROBIO.2017.8324411

CITATIONS
5

READS
324

6 authors, including:

Chule Yang
Nanyang Technological University
21 PUBLICATIONS **75** CITATIONS
SEE PROFILE

Yijie Zeng
Nanyang Technological University
2 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Yufeng Yue
Nanyang Technological University
30 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Prarinya Siritanawan
Japan Advanced Institute of Science and Technology
23 PUBLICATIONS **57** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Relative Dynamics and Control for Spacecraft Formation View project

Unique Person Recognition View project

All content following this page was uploaded by Chule Yang on 10 December 2018.

The user has requested enhancement of the downloaded file.

Proceedings of the 2017 IEEE
International Conference on Robotics and Biomimetics
December 5-8, 2017, Macau SAR, China

Knowledge-Based Role Recognition by Using Human-Object Interaction and Spatio-Temporal Analysis

Chule Yang[1], Yijie Zeng[1], Yufeng Yue[1], Prarinya Siritanawan[2], Jun Zhang[1] and Danwei Wang[1]

Abstract— **Role recognition is a key problem when dealing with the unspecified human target whose description is limited, or appearance is ambiguous. Moreover, the ability to recognize the role of human can help to spot out the exceptional person in the scene. In this paper, a knowledge-based inference approach is proposed to categorize human roles as a binary representation of the targeted person and others by using the object-interaction feature and spatio-temporal feature. The method can associate spatial observations with prior knowledge and efficiently infer the role. An intelligent system equipped with an RGB-D sensor is employed to detect the individual and designated objects. Then, a probabilistic model of the existence of objects and human action is built based on prior knowledge. Finally, the system can determine the role through a Bayesian inference network. Experiments are conducted in multiple environments concerning different setups and degrees of clutter. The results show that the proposed method outperforms other methods regarding accuracy and robustness, moreover, exhibits a stable performance even in complex scenes.**

I. Introduction

In recent years, with the rapid development of computer technologies, various intelligent systems have been designed to learn the human behavior and try to interact with people in their workspace. Moreover, smart and autonomous systems are highly anticipated operating in the increasingly complex environment and understand human context. However, in highly uncertain and dynamic environments, the targeted person may only have an abstract description or ambiguous appearance. Thus, the ability that can infer implicit information from the limited perception is beneficial for robots to perform high-level task planning. On the other hand, due to the abstractness, dynamic changes and potential occlusion in the environment, recognition methods that rely on low-level features or single attribute are inefficient. In order to solve these problems, we should equip the intelligent system with practical knowledge and integrate multimodal information to tolerate the uncertainty, understand the environmental context comprehensively and reasoning based on it.

The ability to recognize the role of a person is critical in many tasks such as social activities and surveillance video. By knowing the function of individuals, the system can have a more comprehensive understanding of the human context and identify the critical or abnormal person in the environment. It is originally a subject of Psychology and Sociology [1]. With the development of image processing and machine learning techniques, the perception capability of intelligent systems is drastically increasing both from hardware [2] and software perspective [3]. For example, the state-of-the-art visual detection method R-CNN [4] has performed well on detecting individual objects. However, due to the possible loss of structural information, the deep learning-based approach is hard to debug and maintain. Consequently, by leveraging the rich features extracted from vision-based methods and combining with other relevant information, different modules could be constructed to solve the problem in a decomposable manner.

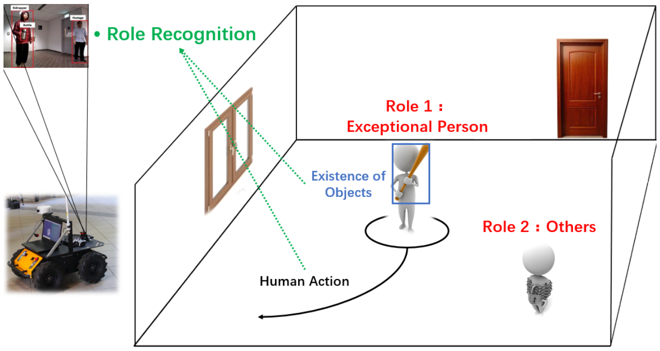


Fig. 1. An illustration of role recognition. In order to distinguish the targeted person from others, the inference is based on the analysis of movement behavior together with carry-on objects.

Therefore, this research aims to develop a scalable inference strategy, which can exploit object-interaction features and spatio-temporal features to recognize a particular role and distinguish from others in the environment. The focus of this research is on inferring the human role by observing their carry-on objects and movement behaviors in the environment.

Main contributions of this work are listed below:

- A probabilistic inference approach is proposed that can associate spatial observations with semantic labels to identify the specific role in the environment.
- An algorithm is developed to categorize human roles as a binary representation of the targeted person and others by analyzing object-interaction and spatio-temporal feature and associating them with prior knowledge.
- Experiments are conducted under different setups and degrees of clutter, which showed that our method outperforms others regarding accuracy and robustness and exhibits a stable performance even in complex scenes.

*This research was partially supported by the ST Engineering - NTU Corporate Lab through the NRF corporate lab@university scheme.

[1]Chule Yang, Yijie Zeng, Yufeng Yue, un Zhang and Danwei Wang are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore {yang0438,yzeng004,yyue001, jzhang061,edwwang}@ntu.edu.sg

[2]Prarinya Siritanawan is with ST Engineering - NTU Corporate Laboratory, Nanyang Technological University, Singapore prarinya@ntu.edu.sg

978-1-5386-3741-8/17/$31.00 © 2017 IEEE

The rest of this paper is structured as follows. Section II reviews relevant works in role recognition and action recognition. Section III details the theoretical basis of the proposed knowledge-based recognition method. Section IV shows the experimental procedures and results. Section V concludes the paper and discusses future works.

II. Background and Related Works

Role recognition is a research problem in social activities. By understanding the role of human, it can help to determine their possible interactions with the environment and vice versa [5]. The research [6] used appearance features to recognize the relationship between pairs of people by training with familial social relationship labels. In [7], it classified the occupation of human by analyzing the clothing and context in human images. [8] predicted the role labels such as "attacker" and "defender" in sports videos for group activity recognition. They used training labels to assign the role by the spatio-temporal interaction between players. Most recently, [9] proposed work which detected and recognized human-object interactions to output a triplet which shows the person's box, certain actions, and the target object's box as well as its category. However, recognizing the relationship only on a single image or a particular instant may be insufficient in the continuously changing and highly uncertain environment. Moreover, models based on deep learning are usually computationally expensive because it takes a long time to learn parameters and there is a loss of structural information during the process.

Action recognition is another important feature used to infer the role. By analyzing the spatio-temporal information, many works have been done to recognize the human action [10]–[13]. [11] built the relationships of interest points by generating visual words from local motion and appearance features and then formed the corresponding neighbor features. A hierarchical structure is introduced in [10] to model the spatio-temporal contextual information of interest points detected by SIFT. In their work, point level, intratrajectory and intertrajectory contextual information are exploited. Besides, action recognition is critical to detect abnormal scene in the environment. [14], [15] presented an approach to identify 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, [16] was 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.

Thus, this research not only considers the interaction between human and objects but also speculates on the role of people in the entire 3D space over time. We intend to develop a knowledge-based approach to generate an intuitive and dynamic inference of human role by simultaneously observing the human-object interaction and analyzing the spatio-temporal information in the 3D space.

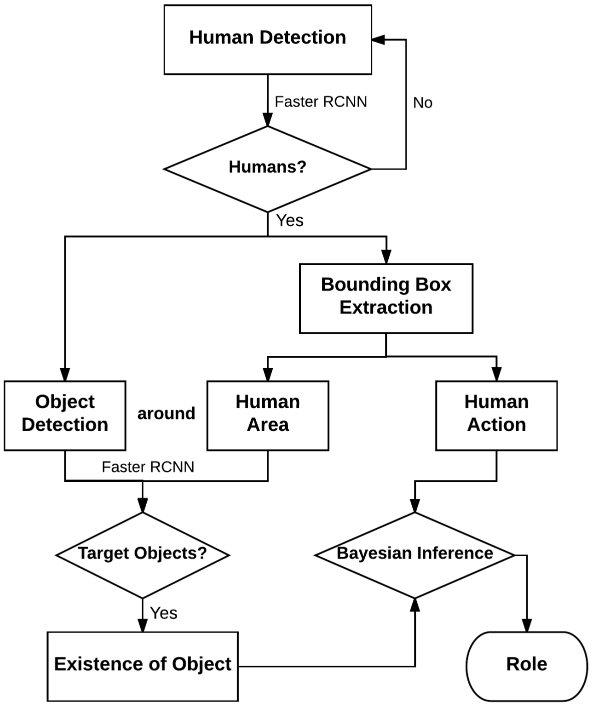


Fig. 2. The diagram of knowledge-based role recognition. Faster R-CNN is applied to detect human and designated objects in the scene. Then, the role of human is inferred through a Bayesian network by concerning human action and the existence of objects simultaneously.

III. Knowledge-Based Role Recognition

The proposed knowledge-based role recognition method is to dynamically reason the role of human in uncertain environment. This method is scalable and it makes the inference through a Bayesian scheme by considering object-interaction and spatio-temporal features simultaneously. Human roles are considered as a binary representation, namely, the specific/exceptional role and others, which are denoted as I_1 and I_2, respectively. $I \in \{I_1, I_2\}$.

A. Problem Statement

In many scenarios, the appearance of the target could be ambiguous or lack of description. Thus, in order to recognize the role of human and spot out the exceptional person in the scene, the analysis of human activities and their carry-on items is a more reliable way. After detecting people in the environment, the intelligent system will investigate the area around them and identify whether there exists the designated object. Then, the system needs to locate the 3D position of both humans and objects, and simultaneously analyze the action of people. Then the system can distinguish the specific role from others through a Bayesian inference strategy. The detailed theoretical basis is illustrated in the following parts, including the construction of the probabilistic model, how those channels are integrated, and how decisions are made.

B. High-Level Feature Extraction

*1) **Object Interaction Feature** Θ^{OI}:* The interaction between human and certain object is critical in determining the role. A specific role may require people to wear

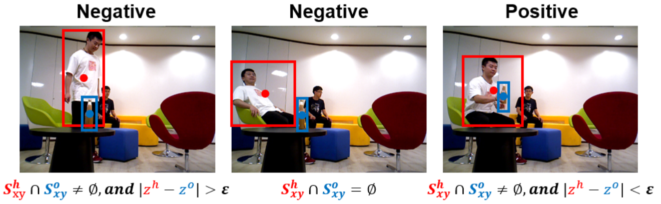

Negative
Negative
Positive
$S^h_{xy} \cap S^o_{xy} \neq \emptyset$, and $|z^h - z^o| > \varepsilon$
$S^h_{xy} \cap S^o_{xy} = \emptyset$
$S^h_{xy} \cap S^o_{xy} \neq \emptyset$, and $|z^h - z^o| < \varepsilon$

Fig. 3. The existence of an object is defined as whether the 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.

particular clothes or using particular items. For example, "Cook" wearing "Tall Hat" and "Soldier" holding a "Gun" are representative examples.

a) Existence of Objects: 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, O is represented as designated objects and it is assumed that all objects are independent. The existence of certain object is either positive or negative, which are denoted as o_1 and o_2, respectively. $O \in \{o_1, o_2\}$. The probabilistic model is represented as $P(O_{j,t}|I)$, which gives the probability of the jth person carrying the designated object O at time t when knowing its role I. If there are multiple objects, it can be represented as $P(O_{n,j,t}|I)$.

More specifically, the traditional definition of the existence of objects is for the entire scene, which will give positive as long as the object is detected in the image. However, in this research, the existence of objects 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_f is the coordinate of the center point in z axis, which is obtained by calculating the medium of depth value from all points, $z_f = Median(S_{z,f})$. 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. (1) .

$$O_{j,t} = \begin{cases} o_1(Positive) & \text{if } S^h_{xy,f} \cap S^o_{xy,f} \neq \emptyset \text{ and } |z^h_f - z^o_f| \leq \varepsilon \\ o_2(Negative) & \text{Otherwise} \end{cases} \tag{1}$$

*2) **Spatio-Temporal Feature** Θ^{ST}:* 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.

a) 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 $A \in \{a_1, a_2, \cdots, a_n\}$. The probabilistic model is represented as $P(A|I)$, which gives the probability of taking action A when it is known as the role I. In this research, mainly two actions are considered, namely, *Moving* and *Stationary*, which are denoted as a_1 and a_2, respectively. The probabilistic model $P(A_{j,t}|I)$ represents the likelihood of taking the action A of the jth person at time t when the role is I.

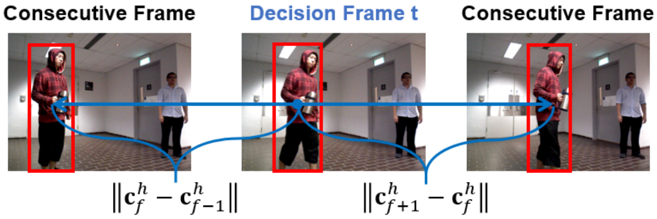

Consecutive Frame
Decision Frame t
Consecutive Frame
$\|\mathbf{c}^h_f - \mathbf{c}^h_{f-1}\|$
$\|\mathbf{c}^h_{f+1} - \mathbf{c}^h_f\|$

Fig. 4. At each decision frame, the action of human is analyzed by comparing with its neighbors. If both comparison results show that the person is staying stationary, then the action is recognized as *Stationary*; otherwise, the action is recognized as *Moving*. $\mathbf{c}^h_f$ is the representative point for each *human* bounding box at the fth frame.

More specifically, the action of human at each decision frame is analyzed by comparing with its neighbors. By comparing the 3D position difference in the space between the consecutive frames, the human action is moving if the difference exceeded certain level; otherwise, the action is stationary. Only if both comparison results show that the person is staying stationary, then the action at that decision frame is confirmed as *Stationary*; otherwise, the action is recognized as *Moving*. As illustrated in Fig. 4 and Eq. (2), the 3D position of human at each frame is represented by a single point. Let $\mathbf{c_f}$ denotes the 3D coordinates of the point at the fth frame, $\mathbf{c_f} = (X_f, Y_f, Z_f) = (\delta x_f, \delta y_f, z_f)$. Where X_f, Y_f, Z_f denotes the coordinate in real space, x_f, y_f, z_f 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 $x_f = Median(S_{x,f})$, $y_f = Median(S_{y,f})$. σ is the threshold for determining the motion between the consecutive frames.

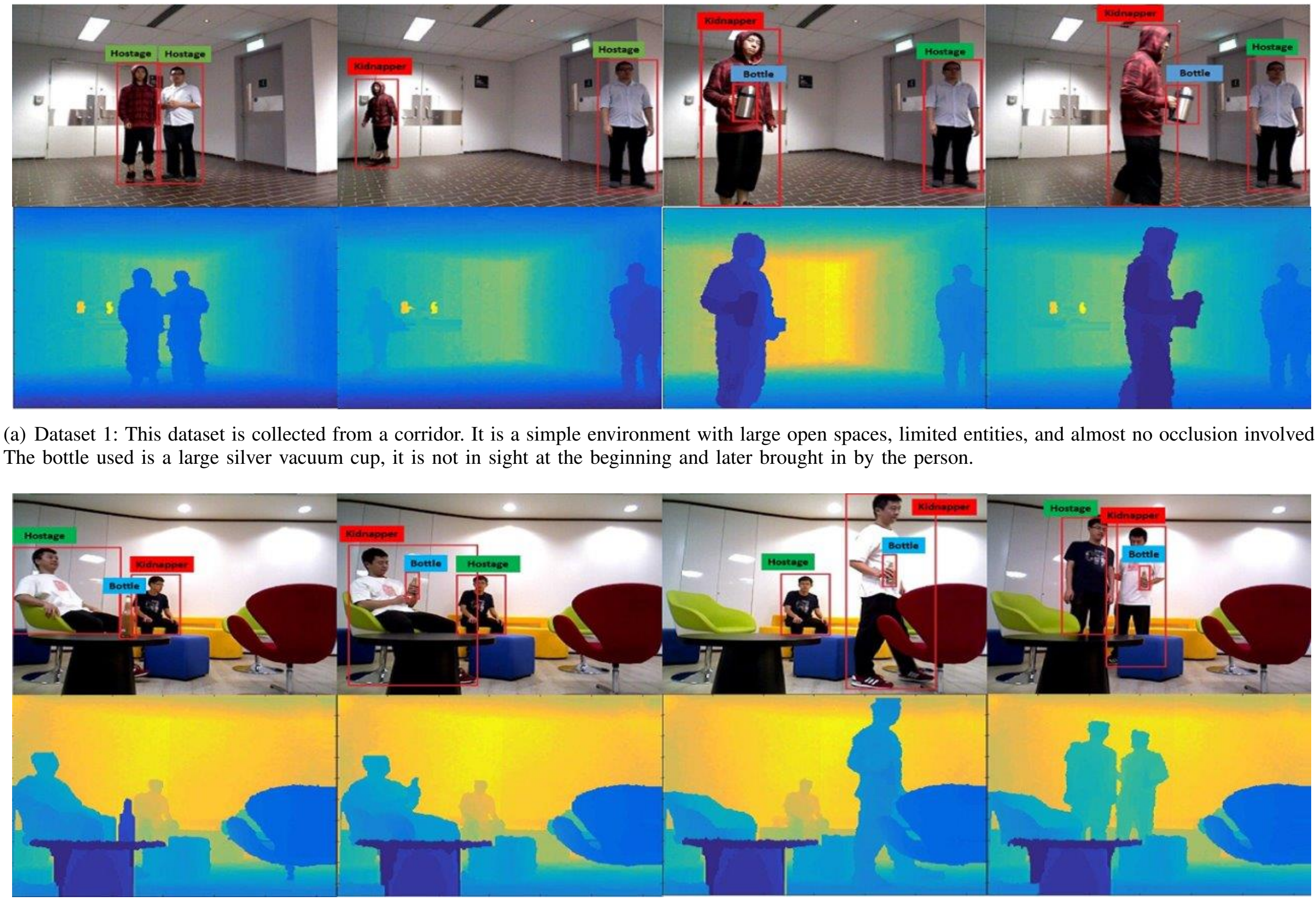
$$A_{j,t} = \begin{cases} a_1(Moving) & \text{if } ||\mathbf{c}_f - \mathbf{c}_{f-1}|| \geq \sigma \text{ or } ||\mathbf{c}_{f+1} - \mathbf{c}_f|| \geq \sigma \\ a_2(Stationary) & \text{if } ||\mathbf{c}_f - \mathbf{c}_{f-1}|| \leq \sigma \text{ and } ||\mathbf{c}_{f+1} - \mathbf{c}_f|| \leq \sigma \end{cases} \tag{2}$$

C. Role Inference from Bayesian Network

By acquiring all channels of information, the role of the jth person can be inferred by fusing of all those available object-interaction features and spatio-temporal features.

$$P_{j,t}(I|\Theta^{OI}_1, \cdots, \Theta^{OI}_m, \Theta^{ST}_1, \cdots, \Theta^{ST}_n) = \prod_{i=1}^{m} \underbrace{P_{j,t}(I|\Theta^{OI}_i)}_{\text{Object-Interaction}} \prod_{k=1}^{n} \underbrace{P_{j,t}(I|\Theta^{ST}_k)}_{\text{Spatio-Temporal}} \tag{3}$$

Specifically, in this research, based on the Bayes rule, the posterior probability of the role is illustrated as below:



(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 body blocks from the objects as well as each other. The bottle used is a small brown beer bottle.

Fig. 5. Experimental data are collected from two separate indoor environments with different settings and human actions. Two bottles are used for the designated object which is different in type, size, and color. Three colors, i.e., red, green and blue, represent the recognition result for the kidnapper, hostage, and bottle, respectively.

$$\begin{aligned} &P_{j,t}(I|A_{j,t},O_{1,j,t},\cdots,O_{n,j,t})\\ \propto &P_{j,t}(A_{j,t},O_{1,j,t},\cdots,O_{n,j,t}|I)P_{j,t-1}(I)\\ \propto &\underbrace{P_{j,t}(A_{j,t}|I)}_{\text{Human Action}}\cdot\prod_{i=1}^{n}\underbrace{P_{j,t}(O_{i,j,t}|I)}_{\text{Existence of Objects}}\cdot P_{j,t-1}(I) \end{aligned} \tag{4}$$

Initially, the probability of becoming each role is set as equal, i.e., $P_{j,0}(I_1) = P_{j,0}(I_2)$. The last equation is derived based on the conditional independence between different channels of information. Then, the role is determined by taking the maximum a posterior possibility.

$$I_{j,t} = \underset{I\in\{I_1,I_2\}}{\arg\max}(P_{j,t}(I)) \tag{5}$$

IV. Experiments

The parameters of the probabilistic model could be modified according to different cases. To simplify the procedures, an indoor kidnapping scenario is simulated in this experiment to recognize the role of people, namely, kidnapper (I_1) and hostage (I_2). Due to limitations, two different bottles are used as the designated object, which are different in type, size, and color. And two datasets were collected from different indoor environments regarding different settings and degrees of clutter, as shown in Fig. 5.

A. Prior Knowledge Acquisition

All parameters are obtained from case study and surveys. By consulting experts (100 policeman) in the field and knowing of the corresponding scenario, practical knowledge could be obtained in advance. Thus, it could be inferred that the kidnapper might carry some potential weapons and perform more movements in the environment. TABLE I (a) illustrates the relationship between roles and human action $P(A|I)$. Similarly, the relationship between roles and the existence of objects $P(O|I)$ is shown in TABLE I (b).

B. Human and Object Detection

In this research, an ASUS Xtion Pro Live RGB-D camera is used as the imaging equipment. Faster R-CNN [17] is employed to detect humans and objects in the environment. Since detection is not the focus of this paper, we just use the existing network model which is trained on VOC 2007

TABLE I

CONDITIONAL PROBABILITY FOR ROLE RECOGNITION IN KIDNAPPING SCENARIO.

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

		Role (I)	
		Kidnapper (I_1)	Hostage (I_2)
Action	Moving (a_1)	0.54	0.39
(A)	Stationary (a_2)	0.46	0.61

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

		Role (I)	
		Kidnapper (I_1)	Hostage (I_2)
Object	Positive (o_1)	0.56	0.01
(O)	Negative (o_2)	0.44	0.99

TABLE II

EVALUATION OF DIFFERENT DECISION METHODS ON THE SIX SCENARIOS IN TWO SEPARATE DATASETS. K AND H REPRESENT RESULTS FOR THE CORRESPONDING KIDNAPPER AND HOSTAGE. THE VALUES IN THE TABLE ARE CALCULATED BY F-MEASURE. *Score* IS THE FINAL WEIGHTED MEASUREMENT OF EACH METHOD. d_{2D}, d_{3D}, *IS*, and *KBI* STAND FOR DECISION FROM 2D DISTANCE, 3D DISTANCE, IMAGE STREAM, AND KNOWLEDGE-BASED INFERENCE, RESPECTIVELY.

Method			*Scenario*$_1$		*Scenario*$_2$		*Scenario*$_3$		*Scenario*$_4$		*Scenario*$_5$		*Scenario*$_6$		*Score*	
			$K_mH_mO_n$		$K_sH_sO_n$		$K_mH_sO_n$		$K_sH_sO_p$		$K_mH_sO_p$		$K_mH_mO_p$			
			$\gamma_1 = 0.11$		$\gamma_2 = 0.15$		$\gamma_3 = 0.18$		$\gamma_4 = 0.19$		$\gamma_5 = 0.23$		$\gamma_6 = 0.14$			
Decision	Input		K	H	K	H	K	H	K	H	K	H	K	H	K	H
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}	Point Clouds		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
IS	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
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}	Point Clouds		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
IS	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**

and VOC 2012 dataset. The detection algorithm generates a bounding box for each detected human and object. Then, the coordinated depth value for each pixel can be directly extracted from the camera.

C. Raw Data Preprocessing

Due to the lens distortion and misalignment of the camera, 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 and lens distortion on the edge of the camera. To solve this problem, a 3×3 median filter was implemented to smooth the data. According to the distribution of fault data, the filter was set to start from the central point and stretch out in four directions. Processing results were presented in Fig. 6.

D. Experimental Description

Multiple scenarios are examined in this experiment to evaluate the proposed method comprehensively concerning different human actions and the existence of objects. Let K, H, and O represent for *Kidnapper*, *Hostage* and *Object*, respectively. m and s denote *moving* and *stationary*. p and n represent for *positive* and *negative*. For example, the scenario '$K_mH_sO_n$' means that the kidnapper is moving while the hostage is stationary and no designated object is carried. These six scenarios can cover almost all possible situations in real cases. Weights γ are introduced to describe the possibility of occurrence of each scenario. The weight for each scenario is calculated on the basis of TABLE I, and the values are shown in TABLE II. As there are six scenarios considered in this experiment, $\gamma_i \in \{\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6\}$.

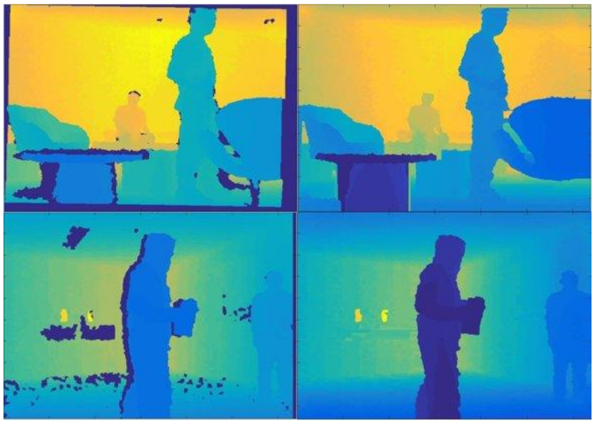


Fig. 6. The two images on the left are raw depth images, and the two on the right are images after filtering. The above is caused by the camera lens distortion whereas the below due to missing data.

In this experiment, input features of decision systems are extracted from three different perspectives (i.e., RGB image, Point Clouds and RGB-D image). By analyzing RGB images and point clouds, the decision is made by measuring 2D distance (d_{2D}) and 3D distance (d_{3D}) between the detected person and objects, respectively. For decisions from distance measurement, the person is identified as the kidnapper only if the object is detected to be located in the human area. Another one is to analyze the Image Stream (*IS*), the decision is made by measuring the difference between successive images, and the moving person is identified as the kidnapper. The proposed Knowledge-Based Inference (*KBI*) method operates on the RGB-D images, and the decision is made through a Bayesian network.

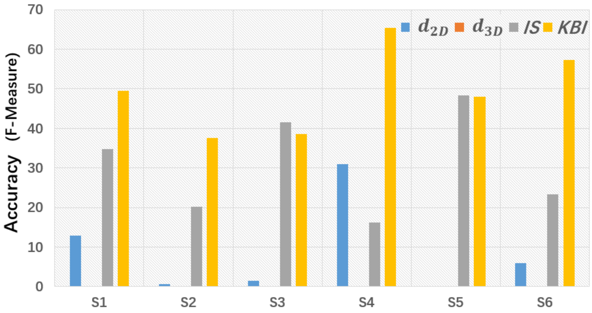


Fig. 7. Average recognition results for the exceptional person (kidnapper) in each scenario. According to F-Measure, the proposed method (*KBI*) achieved the highest accuracy among all decision methods.

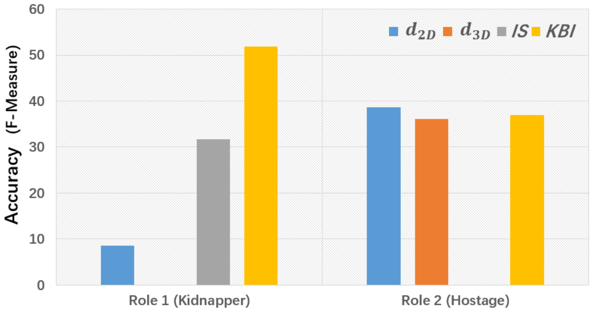


Fig. 8. Overall role recognition result for each decision method. The proposed method (*KBI*) achieved the highest accuracy for kidnapper recognition and competitive result for hostage recognition.

E. Experimental Results

Results of each examined method were obtained by taking the F-measure. *Score* is the final measurement for each method by multiplying the F-measure result of each scenario with its corresponding weight. Numerical results for each dataset were shown in TABLE II. The performance for recognizing the specific role and overall recognition result were revealed in Fig. 7 and Fig. 8.

The proposed method (*KBI*) well compensates for the limitations of failing to detect the single attribute by fusing information from different modalities. Comparing with other methods in each dataset, it produces the highest accuracy among all methods for kidnapper recognition and a competitive result for hostage recognition. Even in some complex scenes, the performance remains robust. For decisions based on analyzing RGB images (d_{2D}), it can only identify the kidnapper when the specified object is successfully detected, and the performance can be affected by illusions and variant viewing angles. For decisions based on analyzing point clouds (d_{3D}), it is hard to detect a person when encountering a variety of postures and occlusions. Besides, as detection of the small object is still very challenging, they can not recognize the kidnapper because of failing to detect the bottle. For decisions based on Image Stream (*IS*), it can only recognize the moving person, thus it will fail if there is a lack of movements in the scenes or multiple persons are involved.

V. Conclusions

In this paper, it is proposed to combine spatial observations with prior knowledge to recognize a specific role and distinguish him from others in indoor environments. The probabilistic model established in this research can integrate the human action and the existence of objects to give a robust role recognition. As a result, the proposed method well compensated for the limitation of failing to detect the single attribute, which produced the highest accuracy among all methods for kidnapper recognition and a competitive result for hostage recognition. Moreover, it achieved robust performance even in some complex scenes.

Since the role recognition problem is of significant uncertainty, in the future work, it is desired that the system could deal with more persons and objects as well as recognize more human behaviors. Besides, if we want to apply it on mobile robots, the ability to determine their actions based on the visual recognition result is highly anticipated.

References

[1] A. Sapru, "Automatic social role recognition and its application in structuring multiparty interactions," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2015.

[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] C. Yang, D. Wang, and P. Siritanawan, "Organ-based facial verification using thermal camera," in *Multimedia (ISM), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 321–324.

[4] 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.

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

[6] 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.

[7] 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.

[8] T. Lan and L. Sigal, "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.

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

[10] J. Sun, X. Wu, and S. Yan, "Hierarchical spatio-temporal context modeling for action recognition," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2004–2011.

[11] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2046–2053.

[12] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[13] 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.

[14] F. Lv, X. Song, and V. K. Singh, "Left luggage detection using bayesian inference," in *Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*. Citeseer, 2006, pp. 83–90.

[15] K. Lin, S.-C. Chen, and C.-S. Chen, "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.

[16] 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.

[17] 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.

View publication stats