Multi-level probabilistic uniqueness reasoning of autonomous robots based on spatial-semantic fusion

Yang, Chule

2019

Yang, C. (2019). Multi-level probabilistic uniqueness reasoning of autonomous robots based on spatial-semantic fusion. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/84521>

<https://doi.org/10.32657/10356/84521>

**MULTI-LEVEL PROBABILISTIC UNIQUENESS
REASONING OF AUTONOMOUS ROBOTS BASED
ON SPATIAL-SEMANTIC FUSION**

YANG CHULE

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

.....2019.10.8.....

Date

....

YANG CHULE

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

.....2019.10.8.....

Date

A handwritten signature in black ink, appearing to read 'Danwei Wang', written over a horizontal dotted line.

WANG DANWEI

Authorship Attribution Statement

This thesis contains material from 2 paper(s) published in the following peer-reviewed journal(s) where I was the first and/or corresponding author.

Chapter 4 is published as Chule Yang, Danwei Wang, Yijie Zeng, Yufeng Yue, P. Siritanawan, Knowledge-Based Multimodal Information Fusion for Role Recognition and Situation Assessment by Using Mobile Robot, *Information Fusion* 50, 126-138 (2019). DOI: 10.1016/j.inffus.2018.10.007.

The contributions of the co-authors are as follows:

- Prof. Wang and I discussed and established the initial project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by Dr. Siritanawan.
- Dr. Yue and Mr. Zeng assisted in data collection, data processing, and experiments.

Chapter 5 is published as Chule Yang, Yufeng Yue, Jun Zhang, Mingxing Wen, and Danwei Wang, 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 (2019). DOI: 10.1109/TMM.2018.2875513.

The contributions of the co-authors are as follows:

- Prof. Wang and I discussed and established the initial project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by Dr. Yue.
- Mr. Zhang and Mr. Wen assisted in data collection, data processing, and experiments.

2019.10.8

.....

Date

.....

YANG CHULE

Acknowledgements

First of all, I would like to express my special and great appreciation to my supervisor, Prof. Wang Danwei, for his invaluable guidance and continuous support throughout my Ph.D. study. His encouragement and praise have brought me through the difficulties and harsh times in my Ph.D. journey. This thesis would never have been possible without his brilliant ideas and excellent drive for research. He is not only my Ph.D. supervisor but also my life mentor.

I am grateful to NTU for providing me with the financial support and lab resources to carry out my research work. I wish to thank Pop, Namal, Eshan and Diluka for all the valuable discussions and the advises. To Yue Yufeng, Wang Yuanzhe, Zhang Jun, Wen Mingxing, Yosmar, Dai Bo and Fu Wenhao for their assistance and corporation in the completion of the project. To Jiang Wentao, Yang Shuai, Guo Fanghong, Fu Changhong, Zhao Wei, Cao Kun, Peng Guohao, Wu Zhenyu, Tang Xiaoyu and Zhang Haoyuan, for their friendship and support during my research. It's my privilege of working with the kind and talented fellow researchers, which has been a rewarding experience.

Furthermore, I would like to thank my parents and girlfriend. They have taught me the value of generosity, hard work, and perseverance and that, as well as their endless love, encouragement, and support, have allowed me to achieve and excel in my life.

Abstract

Advanced robots are desired to improve their reasoning ability in increasingly practical situations to quickly respond to urgent challenges such as search and rescue, public security, and autonomous driving. Advances in sensor technology and intelligent algorithms are accelerating the development of autonomous systems, allowing them to interact with dynamic environments more independently and intelligently, completing relatively complicated tasks. Nevertheless, the safety and security problems are yet to be comprehensively addressed. Currently, many efforts have been made to solve this problem by improving the unimodal perception ability of intelligent systems. Methods using surveillance cameras to collect large amounts of visual data and process them using deep networks. Various types of range scanner can sense and reconstruct the surrounding environment through precise distance measurements. However, in the above cases, there may be a lack of training data, fixed cameras cannot adequately and dynamically model the surroundings, unimodal analysis is hard to capture both appearance and depth information. Therefore, higher levels of perception and reasoning abilities are required to handle unexpected conditions and scenarios efficiently. In such cases, mobile robots should be fully aware of the dynamic environment, prioritize, and continuously analyze critical situations, share information, and collaboration under limited computing power and resources. Furthermore, considering the robots are operating alongside humans, reliable and interpretable robot behavior is highly preferred for more natural interactions.

This study aims to develop reasoning strategies for autonomous robots to improve robotic intelligence under different assumptions and conditions. The framework allows one or

several robots to address safety or security problems in a targeted manner in partially known environments, where time, dynamics, and interactions play a significant role. The objective is to identify the unique individual whose activity, in space and time, appear different from the other people in the scene. The activities considered not only the motion pattern, but also the object associated with this person. Unlike traditional abnormal detection, the uniqueness reasoning in this work does not require a predefined expected pattern. A modular analysis is also applied to decompose the entire reasoning process into submodules such as perception, analysis, and inference, to achieve algorithmic flexibility and interpretability. Probabilistic modeling is used to associate spatial relationships between semantic labels and their temporal changes. The relationships are described in either instant spatial spaces (semantic-interaction feature) or spatiotemporal spaces (spatiotemporal feature), and it provides essential information for robot navigation, situation awareness, and task planning. Unlike existing research, this is the first time that such a mission is implemented on mobile robot platforms by using the probabilistic fusion of multimodal information.

Considering how evidence information is obtained and linked in the process of reasoning, we study the problem of uniqueness reasoning at three levels. The first level is called *knowledge-based single robot uniqueness reasoning* and runs on a single robot platform. At this level, we focus on a knowledge-based approach, where knowledge is provided in terms of how motion and an object associated with a person can be used to infer his/her uniqueness, as determined by human expertise. Because such a method has advantages in dealing with structured missions and with limited training data. Qualitative rules are applied to extract human activities from observations, and then the uniqueness of each person is judged by associating the prior knowledge with the performed activities. Experiments were conducted in indoor environments with various degrees of occlusion, and the experimental results showed the rationality and accuracy of the reasoning.

However, though the knowledge-based method can successfully address specific tasks, it lacks flexibility and generality to adapt to different situations. The assumption of prior knowledge should be replaced by general analytical modeling. Thus, the second level is

analytics-based single robot uniqueness reasoning. For one thing, it applies quantitative analysis to model human activities from observations rather than using qualitative rules. For another, instead of leveraging prior knowledge, the distribution of the features characterizing a human activity is constructed and the activity of an individual is analyzed within the distribution to determine if it is unique. Experiments were conducted in both indoor and urban outdoor environments. The experimental results demonstrated the rationality, accuracy, and efficiency of the reasoning.

Moreover, since many applications require to work in an unstructured and large-scale environment as well as a long-term operation, the sensing and reasoning capabilities of a single robot are limited by occlusion, sensor failure, and narrow field of view. Robots are desired to share information and perform collaborative reasoning. Therefore, if say the first two levels are mainly using sensory data to implement the algorithm, the third level is more focused on applying the algorithm to the robot platform. The third level is *analytics-based multi-robot uniqueness reasoning*, in which a multi-robot system is adopted as the platform. By obtaining the relative pose between robots, local observations can be transformed and transmitted among the multi-robot system. Since it is a distributed system, each robot still first performs its local uniqueness reasoning in the case of long-distance and poor communication. Meanwhile, the system performs a re-judgment based on the shared global observation and then shares the judgment results back to each robot. Experiments were carried out in various unstructured environments and contained day and night situations. The experimental results showed the rationality, accuracy, and robustness of the reasoning. This is the first time that reasoning is categorized by considering the way of obtaining and linking evidence information. This study has potential applications in a variety of situations, and it can serve as a guide for further robotic motion planning and human-robot interaction.

List of Abbreviations

SS	Spatial-Semantic
SI	Semantic-Interaction
ST	Spatiotemporal
DBI	Distance-Based Inference
KBI	Knowledge-Based Inference
MIF	Multimodal Information Fusion
OEM	Object Existence Model
HAM	Human Action Model
MDP	Markov Decision Process
HMM	Hidden Markov Model
MEMM	Maximum Entropy Markov Model
MRF	Markov Random Field
OF	Optical Flow
CNN	Convolutional Neural Network
R-CNN	Region-based Convolutional Neural Network
SLAM	Simultaneous Localization And Mapping
C-SLAM	Collaborative Simultaneous Localization And Mapping
UGV	Unmanned Ground Vehicle
2D	Two-Dimensional
3D	Three-Dimensional

Contents

Acknowledgements	ii
Abstract	vi
List of Abbreviations	viii
List of Contents	xii
List of Figures	xix
List of Tables	xxii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Objective	3
1.3 Approach	3
1.4 Contributions	5
1.5 Thesis Organization	7

2	Literature Review	9
2.1	Probabilistic Reasoning	9
2.2	Abnormal Detection	10
2.3	Role Recognition	11
2.4	Multi-Robot Collaboration	12
2.5	Differences from the Related Work	12
3	Multi-Level Uniqueness Reasoning Architecture	15
3.1	Definitions and Prerequisites	17
3.1.1	Feature Taxonomy	17
3.1.2	Heterogeneous Sensors Calibration	19
3.1.3	Target Detection and Localization	21
3.1.4	Definition of Spatial-Semantic Analysis	24
3.2	Mathematical Modeling	25
3.2.1	General Uniqueness Reasoning Formulation	25
3.2.2	Uniqueness Reasoning at Different Levels	26
4	Knowledge-Based Single Robot Uniqueness Reasoning	29
4.1	Introduction	29
4.2	Main Approach	31
4.2.1	Knowledge-Based Inference (KBI)	31

4.2.2	Distance-Based Inference (DBI)	35
4.2.3	Final Uniqueness Judgment	37
4.3	Experiments	38
4.3.1	Prior Knowledge Acquisition	39
4.3.2	Raw Data Preprocessing	40
4.3.3	Experimental Description	40
4.3.4	Experimental Results	42
4.4	Conclusions	46
5	Analytics-Based Single Robot Uniqueness Reasoning	49
5.1	Introduction	49
5.2	Main Approach	51
5.2.1	Reasoning Process Decomposition	52
5.2.2	Activity Extraction	53
5.2.3	Uniqueness Judgment	54
5.3	Experiments	58
5.3.1	Effect of False Detection on Reasoning Performance	59
5.3.2	Indoor Experiments	62
5.3.3	Outdoor Experiments	66
5.4	Conclusions	69

6	Analytics-Based Multi-Robot Uniqueness Reasoning	71
6.1	Introduction	71
6.2	Main Approach	72
6.2.1	Robot Relative Pose Estimation	74
6.2.2	Local Uniqueness Reasoning	76
6.2.3	Global Uniqueness Reasoning	77
6.3	Experiments	80
6.3.1	Experiments Overview	80
6.3.2	Field Experiments	83
6.3.3	Quantitative Analysis	93
6.4	Conclusions	94
7	Conclusions and Recommendations	97
7.1	Conclusions	97
7.2	Recommendations for Further Research	99
	Author's Publications	101
	Bibliography	103

List of Figures

3.1	The whole reasoning process is decomposed into three modules, namely, perception, analysis, and inference.	16
3.2	An example of the feature taxonomy.	18
3.3	The robot platform used in our research. The Husky robot from Clearpath Robotics, the VLP-16 3D Lidar, the Hokuyo 2D range finder, the Optris PI450 thermal camera, and ZED stereo camera.	19
3.4	Diagram of the sensor frames and the two-step calibration method. (a) Sensor frames between three sensors: a 3D LiDAR $\{L\}$, a visual camera $\{C\}$ and a thermal camera $\{T\}$. (b) Calibration between 3D LiDAR and visual camera. (c) Calibration between visual and thermal camera.	20
3.5	Schematic diagram of target detection using Faster R-CNN. The input is the raw color image and the outputs are labels and regions.	22
3.6	A demonstration of sensor fusion and robot environmental perception. Multimodal information is integrated to provide a comprehensive understanding of the environment.	23

- 3.7 An example graphical model of the entire reasoning process for person j . The perception layer contains the observations of people, the feature layer contains the extracted activities of people, the decision layer is the judgment uniqueness of each person. 26
- 4.1 The overall block diagram of the knowledge-based single robot uniqueness reasoning. The proposed method starts with human and object detection to obtain their corresponding regions. Next, DBI measures the semantic-interaction feature to find the relative distance between humans and the critical object, while KBI feeds both the object existence and human action into a Bayes filter. Then, the final judgment of the uniqueness is made by the weighting of these two indicators. 30
- 4.2 Object existence is defined as whether the region of the object is intersected with the human region, and only when they intersect in the entire 3D space, then the existence is positive. S^o and S^h are the region of *object* and *human*, respectively. z is the distance to the camera and ϵ is a threshold for determining the intersection in z direction. 32
- 4.3 The motion of human is analyzed by comparing with its consecutive frame. If the difference is below a certain level, then the action is recognized as *Stationary*; otherwise, the action is recognized as *Moving*. c_t^h is the representative point for each person's region at the t th frame. . . 33
- 4.4 An example of the KBI. It is assumed that people who carry the specified object or conduct specific action become the unique individual (i^1). The bar charts show the probability of being the unique individual (i^1) and usual person (i^2) for Person 1 (Blue) and Person 2 (Orange), respectively. 34

- 4.5 An example of the DBI. It is assumed that the person who is closer to the object becomes the unique individual (i^1), others are usual people (i^2). The bar charts show the probability of being the unique individual for Person 1 (blue) and Person 2 (orange). 36
- 4.6 An illustration of the final decision process. After obtaining the inferred results from all indicators, the final uniqueness is determined through a weighted fusion of all the results. 37
- 4.7 Results of raw data preprocessing. Images on the left are raw depth images, and images on the right are filtered depth images. 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. 41
- 4.8 Results of the proposed reasoning method. Experimental data are collected from three separate indoor environments with different settings and human activities. Three bottles are used for the critical object which is different in type, size, and color. Two colors, i.e., red and green, represent the judgment result for the unique individual and usual person, respectively. 43
- 4.9 The overall uniqueness judgment result for each reasoning method. The proposed method achieved the highest accuracy in recognizing both the unique individual and the usual person. 45
- 4.10 Bias of each reasoning method. Blue and orange represent the bias that recognizes the unique individual and the usual person, respectively. It can be seen that d_{2D} and d_{3D} are more advantageous in recognizing usual persons while OF can only spot out the unique individual. The proposed method has the most balanced judgment ability for the unique individual and the usual person. 46

- 5.1 An illustration of uniqueness reasoning process. First, humans and objects are detected for obtaining the observation (region and position). Then, OEM and HAM are then built to extract human activities by performing corresponding semantic interactions and spatiotemporal analysis. Finally, the uniqueness of each person is judged by comparing the activities with the overall distribution. 50
- 5.2 The overall diagram of the analytics-based single robot uniqueness reasoning. The uniqueness of people is inferred from two observation models. OEM analyzes the overlapping area between people and objects. HAM analyzes the relative speed of people with respect to the camera. Then, the activities of each person are compared within the overall distribution in the scene. Finally, the final judgment result can be obtained by fusing all channels of information. 52
- 5.3 An example of five detected persons in the scene, and they form a fully connected undirected graphical. $x_{j,t}$ represents the single inferred result for person j either from OEM or HAM. Then, every single result is compared with each other to measure the deviation from the overall distribution for generating a complete result. 56
- 5.4 Six types of false detection and the corresponding impact on uniqueness reasoning. The green label is recognized as a unique individual. +1 and -1 are counts of corresponding reasoning results according to different false detection. 60
- 5.5 The changing trend of uniqueness reasoning *Precision* and *Recall* according to different types and degrees of false detection. 61

5.6	The three datasets are the same as in Chapter 4. But we apply the new activities extraction and uniqueness judgment approach to it. Yellow box, blue box, and green box represent some judgment results from HAM, OEM, and HAM+OEM, respectively.	63
5.7	Overall uniqueness reasoning result of each method by averaging the result of each dataset. The proposed <i>MIF</i> achieved the highest accuracy for recognizing both the unique and usual person.	65
5.8	Bias of each reasoning method. Blue and orange represent the bias that recognizes the unique individual and others, respectively. The proposed approach <i>MIF</i> has the most balanced reasoning ability.	66
5.9	Evaluation of four security-related scenarios to show the generality of the proposed algorithm, where we do not train the specific relationship but just change the detectors. We take different types and sizes of critical objects into account. Red color denote detected semantic labels, such as person, the bottle in (b), the bicycle in (c) and the car in (d). Green color denotes judged unique individuals.	67
5.10	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. . .	69
6.1	The overall schematic diagram of analytics-based multi-robot uniqueness reasoning. The blue part is the local uniqueness reasoning on each robot, and the brown part is the global uniqueness reasoning based on the shared global observation.	73
6.2	The difference between the use of raw laser points and the use of filtered laser points by a single robot SLAM. The moving person causes an inconsistent "wall" in (b).	75

6.3	A schematic diagram for human dynamic observation and motion analysis. The human motion is extracted from the spatiotemporal analysis. . . .	77
6.4	Experimental environments and platforms.	81
6.5	Overview of the final results of simultaneous collaborative mapping and reasoning in daytime rainforest. Left: multi-robot reasoning results in three scenarios. Middle: collaboratively generated 3D global map. Right: satellite image and real environment.	82
6.6	The results of multimodal environmental perception in daytime rainforest.	85
6.7	The results of multi-robot reasoning in a Daytime Rainforest environment. Points and arrows indicate the 3D position and real-time motion of the detected human. Red color represents the inferred unique person. . . .	86
6.8	Overview of the final results of simultaneous collaborative mapping and reasoning in night forest. Left: multi-robot reasoning results in three scenarios. Middle: collaboratively generated 3D global map. Right: satellite image and real environment.	87
6.9	The results of multimodal environmental perception in night forest. . . .	88
6.10	The results of multi-robot reasoning in the Night forest environment. Points and arrows indicate the 3D position and real-time motion of the detected human. Red color represents the inferred unique person.	90
6.11	Overview of the final results of simultaneous collaborative mapping and reasoning in open carpark. Left: multi-robot reasoning results in three scenarios. Middle: collaboratively generated 3D global map. Right: satellite image and real environment.	91
6.12	The results of multimodal environmental perception in open carpark. . . .	91

6.13 The results of multi-robot reasoning in an Open Carpark. Points and arrows indicate the 3D position and real-time motion of the detected human. Red color represents the inferred unique person.	92
---	----

List of Tables

4.1	Conditional Probability for Uniqueness Reasoning in Kidnapping Scenario.	39
4.2	Different reasoning methods were evaluated in six scenarios for each of the three datasets. The values are calculated from F-measure (in %) for the unique individual (U) and usual person (O). <i>Score</i> is the final weighted measurement of each method.	44
5.1	Comparison with the scope of existing works.	59
5.2	Feature detection accuracy (in %) on different labels in each dataset. . . .	64
5.3	Evaluation of uniqueness reasoning performance in three separate datasets. All the values in the table are calculated from F-measure (in %).	64
5.4	Feature detection accuracy (in %) in each scenario.	68
5.5	Evaluation of uniqueness judgment in four different scenarios (in %). Prec. = Precision, Rec. = Recall.	68
6.1	Summarize of the experiment environment.	82
6.2	Comparison of core functionality with existing approaches.	83

6.3 Human uniqueness reasoning results from the single robot and multi-robot system. (The best performance is denoted in bold.)	93
---	----

Chapter 1

Introduction

1.1 Background and Motivation

Unmanned systems are attracting more and more attention due to the increase in labor costs. Intelligent robots and autonomous vehicles have become the most popular topics in recent years. Advances in sensor technology and intelligent algorithms are accelerating the development of autonomous systems, allowing them to interact with dynamic environments more independently and intelligently, completing relatively complicated tasks.

Depending on a variety of potential applications, generally, it is highly desirable that intelligent systems can autonomously obtain useful information from the environment, think like a human, and perform high-level tasks without human participation [1, 2, 3]. Various intelligent systems have been designed to understand human behavior [4, 5] and try to interact with people in their workspaces [6, 7, 8]. Nevertheless, the safety and security problems are yet to be comprehensively addressed, especially for mobile security and rapid response. In recent decades, many efforts have been made to improve the uni-modal perception capability of intelligent systems. Methods using surveillance cameras to collect large amounts of visual data and process them using deep networks of virtually unlimited computing power [9]. Various types of range scanner sense and reconstruct the

surrounding environment through precise distance measurements [10, 11, 12]. However, in the above cases, there may be lack of training data, fixed cameras cannot adequately and dynamically model the surroundings, illusions may deceive 2D vision due to the lack of scale information, and unimodal analysis is hard to capture both appearance and depth information. Therefore, higher levels of perception and reasoning abilities are required to handle unexpected conditions and scenarios efficiently.

Recently, robotics has made exciting advances in increasingly practical situations to quickly respond to urgent challenges such as search and rescue [13], public security [14], and autonomous driving [15]. It requires the robot to comprehensively perceive the environment and understand the situation to improve the pertinence and efficiency of the task. However, workspaces often involve high complexity and uncertainty, which may lead to lethal death or injuries [16]. Among them, the inability to detect and identify risky individual or behavior in time is an essential cause of accidents. Moreover, given the limited communication bandwidth and computing power, mobile robots are unlikely to perceive, record, and process all of the surrounding information. Thus, the robot should prioritize and continually analyze unique individual or situations in the scene. Furthermore, considering the robots are operating alongside humans, reliable and interpretable robotic behavior is highly preferred for more natural interaction.

Besides, in the highly uncertain and dynamic environments, 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 robot to perform decision level reasoning. When operating in complex environments, dynamic changes of objects, potential occlusion, and varying illumination are often concerned. Reasoning strategies that rely on low-level features or single attribute are insufficient. Alternatively, integrating multimodal perceptual information from both spatial and semantic perspectives can efficiently handle uncertainties, and enrich the credibility of the reasoning.

1.2 Objective

In this research, we specifically focus on mobile security or surveillance scenarios that require targeted and dynamic monitoring of the environment as well as the ability to respond quickly. To cope with the challenges mentioned above, this study explores a new research field whose purpose is to infer unique individuals in the scene through multi-modal analysis of human activities and using robotic platforms. A unique individual is defined as the person whose activities, in space and time, appear different from the other people in the scene. The activities of an individual include not only his motion pattern but also the objects associated with him. The uniqueness is defined as the likelihood of a person being the unique individual, and the process of weighing the evidence and inferring the likelihood of a unique individual is called uniqueness reasoning. The limitations and gaps between the current and related work is discussed in Chapter 2.

Moreover, we study this uniqueness reasoning at multiple levels under different assumptions and conditions. While taking into account limited computing power and resources, the framework allows one or several robots to address safety or security problems in a targeted manner in partially known environments, where time, dynamics, and interactions play a significant role. In the end, by performing information fusion from a different data source (e.g., visual and depth data), different modalities (e.g., spatial and semantic), and different platform (e.g., single robot and multi-robot system), we show the rationality, accuracy, and robustness of the reasoning process.

1.3 Approach

The entire reasoning process is decomposed into several submodules, such as perception (observation acquisition), analysis (activity extraction), and inference (uniqueness judgment), to achieve algorithmic flexibility and interpretability. Probabilistic modeling is used to associate spatial relationships between semantic labels and their tempo-

ral changes. The relationship could either be described in instant spatial relationships (semantic-interaction analysis) or temporal spatial relationships (spatiotemporal analysis), and it provides essential information for robot navigation, situation awareness, and task planning.

Considering how evidence information is obtained and linked in the process of reasoning, we study the problem of uniqueness reasoning at three levels. The first level is called *knowledge-based single robot uniqueness reasoning* and runs on a single robot platform. At this level, we focus on knowledge-based methods in which the knowledge provided in terms of how the activities can be used to infer his/her uniqueness, as determined by human expertise who labeled the datasets as ground truth. Because such a method has advantages in dealing with structured missions and with limited training data. Based on Bayes' theorem, qualitative rules (e.g., whether to overlap and whether to move) are used to extract human activities from observations (region and position of semantic labels), and then the uniqueness of each person is judged by associating prior knowledge with extracted activities. Experiments were conducted in indoor environments with various degrees of occlusion, and the experimental results showed the rationality and accuracy of the reasoning.

However, though the knowledge-based method can successfully address specific tasks, it lacks flexibility and generality to adapt to different situations. The assumptions of prior knowledge should be replaced by general analytical modeling. We analyze the distribution of the features characterizing a human activity is constructed, and the activity of an individual is analyzed within the distribution to determine if it is unique. Thus, the second level is *analytics-based single robot uniqueness reasoning*. For one thing, it uses quantitative rules (e.g., intersection over union and relative moving speed) from observations rather than qualitative rules to model the human activities more precisely. For another, instead of using prior knowledge, the activity of each person is then compared within the overall distribution in the scene for uniqueness judgment. Experiments were conducted in both indoor and urban outdoor environments. The experimental results demonstrated the rationality, accuracy, and efficiency of the reasoning.

Moreover, since many applications require to work in an unstructured and large-scale environment as well as a long-term operation, the sensing and reasoning capabilities of a single robot are limited by occlusion, sensor failure, and narrow field of view. Robots are desired to share information and perform collaborative reasoning. Therefore, if say the first two levels are mainly using sensors to implement the algorithm, the third level is more focused on applying the algorithm to the robot platform. The third level is *analytics-based multi-robot uniqueness reasoning*, in which a multi-robot system is adopted as the platform. By obtaining the relative pose between robots, local observations can be transformed and transmitted among the multi-robot system. Since it is a distributed system, each robot still first performs its own uniqueness reasoning in the case of long-distance and poor communication. Meanwhile, the system performs a re-judgment based on the shared global observation and then shares the judgment results back to each robot for their reference to take appropriate actions. Experiments were carried out in various unstructured environments and contained day and night situations. The experimental results showed the rationality, accuracy, and robustness of the reasoning.

This is the first time that reasoning is categorized by considering the way of obtaining and linking evidence information. The study can be potentially used in a variety of situations, and it can serve as a guide for further robotic motion planning and human-robot interaction.

1.4 Contributions

In this study, main contributions are as follows.

- The uniqueness reasoning problem is extensively studied on mobile robot platforms. The overall reasoning process is decomposed into three modules. The first module is perception (observation acquisition), where heterogeneous sensor fusion is applied to detect and locate semantic labels from sensory data. The second module is analysis (activity extraction), which extracts the activity of each

individual through semantic-interaction analysis and spatiotemporal analysis based on the observations output from the perception module. The third module is inference (uniqueness judgment), which judges the likelihood of each person to be the unique individual based on the performed activities extracted from the analysis module. This is the first time that such a task is accomplished on mobile robot platforms by using the probabilistic fusion of multimodal information.

- The problem of uniqueness reasoning is studied at three levels under different assumptions and conditions, namely, *knowledge-based single robot uniqueness reasoning*, *analytics-based single robot uniqueness reasoning*, and *analytics-based multi-robot uniqueness reasoning*. This architecture demonstrates the evolution of robotic intelligence in its reasoning capabilities. At the first level, qualitative activities are extracted and then the uniqueness is judged by leveraging prior knowledge in specific tasks. At the second level, the activities are obtained by quantitative analysis and then the uniqueness is judged by comparison within the overall distribution. At the third level, the observation from each robot is shared in a multi-robot system to generate a more accurate and robust judgment result. This is the first time that reasoning is categorized by considering the way of obtaining and linking evidence information.
- A knowledge-based uniqueness reasoning approach is developed on a single robot platform. Qualitative human motion and object existence are extracted from corresponding spatiotemporal and semantic-interaction analysis. Then, based on Bayes filter, the activities are jointly modeled and associated with prior knowledge to achieve the judgment of uniqueness. Sensory data were acquired using an RGB-D camera and experimented in an indoor environment with varying degrees of occlusion to show the rationality and accuracy of the reasoning.
- An analytics-based uniqueness reasoning approach is developed on a single robot platform. The human motion and object existence is calculated based on quantitative modeling of the observation rather than using qualitative rules. Then, the distribution of the features characterizing a human activity is constructed and the

activity of an individual is analyzed within the distribution using Markov Random Field (MRF) to determine if it is unique. Sensory data were acquired using an RGB-D camera or a high-resolution camera, and experiments were performed in both indoor and outdoor environments to demonstrate the rationality, accuracy, and efficiency of reasoning.

- An analytics-based uniqueness reasoning approach is developed on a multi-robot platform. Robots build environmental maps by detecting and filtering dynamic observations, so the estimation of the relative pose between robots are more accurate by applying collaborative mapping. Next, the absolute velocity of human in the world reference frame is analyzed and local uniqueness judgment is performed based on it. Meanwhile, by applying the accurate transformation matrix, robots transform and transmit their local dynamic observations to the system to obtain a more precise dynamic global observation. Then, the activity extraction and uniqueness judgment from the shared global information is more accurate and robust. Moreover, the global judgment will also be shared back to each robot to enhance or compensate for their own belief for their reference to take the appropriate actions. Sensory data were acquired using two robots, one equipped with a calibrated visual camera and a 3D LiDAR and the other one equipped with a calibrated thermal camera and a 3D LiDAR. Experiments were conducted in a variety of unstructured environments, including day and night, to show the rationality, accuracy, and robustness of reasoning.

1.5 Thesis Organization

The remainder of this thesis is organized as follows.

- Chapter 2 gives a survey into the work published in the last decade related to uniqueness reasoning and discusses in detail the limitations and gaps between current research and the existing literature.

- Chapter 3 explains the core idea, definitions, and prerequisites of this work, including feature extraction and selection, as well as mathematical modeling of the reasoning process at each level.
- Chapter 4 details the work of uniqueness reasoning at the first level, i.e., knowledge-based single robot uniqueness reasoning. It mainly focuses on the realization of the activity extraction in multimodal analysis through qualitative rules, and applies human-labeled knowledge to achieve uniqueness judgment. Experiments are conducted in indoor environments with different scenarios and degrees of occlusion.
- Chapter 5 details the work of uniqueness reasoning at the second level, i.e., analytics-based single robot uniqueness reasoning. It focuses on the human activity extraction in multimodal analysis by using quantitative modeling. The activities of people are compared within the overall distribution to judge the uniqueness of each person. Experiments were conducted in indoor and urban outdoor environments.
- Chapter 6 details the work of uniqueness reasoning at the third level, i.e., analytics-based multi-robot uniqueness reasoning. It applies the developed algorithm to a multi-robot platform equipped with heterogeneous sensors. The observation is shared in the system to perform a more accurate and robust activity extraction. The global judgment is also shared back to enhance or compensate the belief of each robot. Experiments are conducted in various unstructured environments under occlusion, limited field of view, and varying illumination and contained day and night situations.
- Chapter 7 summaries this thesis. The contributions of this study, together with the expected future work, are stated.

Chapter 2

Literature Review

The primary objective of this study is uniqueness reasoning. Since it is a new research topic, we review the most relevant work in this chapter, such as general probabilistic reasoning, abnormal detection, and role recognition. Then, we discuss in detail the differences between this research and related works.

2.1 Probabilistic Reasoning

Probabilistic reasoning is one of the most important areas of research in an autonomous system. It is widely used in motion planning in the robotics field [17, 18, 19, 20] or in autonomous driving [21, 22, 23, 24, 25, 26, 27]. Besides, a systematic overview of computational systems that support situation awareness is provided in [28]. Recently, more related works have been investigated in the field of robotics and autonomous systems, such as autonomous driving [22, 23], intelligent transportation system [27], and disaster monitoring [29]. In [22], a situation-aware decision-making problem was modeled as a POMDP for autonomous robots to drive in the urban road safely and efficiently. [27] measured spatiotemporal trajectory similarity based on longitudinal and lateral spatiotemporal distance to deal with situation classification problems such as lane assignment and driving maneu-

ver recognition. A knowledge-based framework for human-robot collaborative context awareness in USAR missions is proposed in [29]. It used ontological representation to facilitate knowledge sharing and decision-making. However, existing works focused on either spatiotemporal measurements or ontology analysis. The proposed framework integrates these components to create a more comprehensive representation of the situation.

By fusing multimodal information, robots can expand their understanding to a decision level. In [30], a probabilistic model was proposed to infer the possible object locations by utilizing the encoded relationship between objects and the relationship between object and scene. A layered structure of the spatial knowledge representation was designed in [31] to deal with the compound and cross-modal system which is inherently uncertain and dynamic. An approach which exploits semantic knowledge and probabilistic graphical models to enhance object recognition capability of autonomous robots was presented by [32, 33]. Though a bunch of works have been done for robot searching, recognition, and motion planning, none of research in robotics field have tried to identify the unique person by jointly considering their motion and objects associated to them.

2.2 Abnormal Detection

Abnormal detection is the identification of data points, items, observations or events that do not conform to the expected pattern of a given group. Many works have been done to detect the abnormal scene by using surveillance videos [34, 35, 36]. [34] 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, [35] 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. Some other works focus on analyzing the spatiotemporal information, where people

or moving object that carries out specific actions are considered unique in the environment, and more attention needs to be paid to them. Action recognition is a relevant field [37, 38, 39, 40] and further applied to facilitate autonomous driving [41, 42, 43, 21, 17]. 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 [42]. In outdoor applications, autonomous robots need to have a general awareness of the action of critical cars [21] or people [17] to plan their navigation behaviors. 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.

2.3 Role Recognition

Role recognition is another topic to recognize the person that has critical position or function in social activities [44, 45, 46, 47, 48]. By identifying the role of human, it can help to determine their possible interactions with the environment and vice versa [49]. 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 [44] or analyzing the clothing and context of human to classify the occupation of human [45]. Some other research determines the role by considering the spatial relation between humans or between human and objects. [46] used training labels to assign the role by observing the spatiotemporal between players to predict the role labels such as “attacker” and “defender” in sports videos. Most recently, [50] 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.

2.4 Multi-Robot Collaboration

Currently, multi-robot collaboration has been studied to address the limitations of individual robots operating in large-scale environments or complex manipulations. Works use multi-robot systems to map the environment [51, 52]. They fuse the partial 3D volumetric maps generated by individual robots into a consistent global 3D volumetric map. Whether for simulation or the real-world, although they are operating in unknown environments, the tackled environment is static. Other works use multi-robot systems to manipulate objects. A team of robots is used in [53] to manipulate deformable objects while avoiding the collision and maintaining the formation. [54] proposed a decentralized algorithm that coordinates the forces of a group of robots during a cooperative manipulation task. But these works focus more on the control part rather than intelligence. Most recently, multi-robot systems are applied for object detection and tracking. [55] involves a team of cooperating micro aerial vehicles (MAVs) with onboard cameras to detect and track a person continuously, accurately and in real-time. However, existing literature focuses on developing at the perception level, but this research aims at decision level information fusion for reasoning about the unique/critical individual in the scene.

2.5 Differences from the Related Work

There are significant limitations and gaps between the existing literature and the problems we are facing in the following points:

- The state-of-the-art abnormal detection methods typically **predefine expected normal patterns**, and individuals that do not conform to the expected pattern will be identified as anomalies. But the addressed problem aims to identify individuals

whose motion patterns and objects associated with them, in space and time, appear different from the other people in the scene, which do not have the expected normal pattern as assumed in abnormal detection.

- The state-of-the-art abnormal detection methods typically focus on using **fixed views** (surveillance cameras) to identify the **anomalous scenes**. They usually apply background extraction and mostly only perform spatiotemporal analysis of the entire data sequences. However, the addressed problem requires identifying the unique individual in the environment from a mobile perspective, and expecting the performer (robot) to achieve a rapid response.
- The state-of-the-art role recognition methods aim to annotate each people's characteristics by analyzing human interactions and spatial relationships **in still images**. However, the addressed problem requires not only the analysis of instant spatial relationships, but also needs to model the full 3D environment, dynamically track and analyze human motion in spatiotemporal space.
- The state-of-the-art role recognition methods mostly use deep networks and are based on the **unlimited computing resource**. But the faced problem requires to work on mobile platforms where we need to take the limited computing power and resources into consideration.
- Both abnormal detection and role recognition focus on **unimodal analysis**, which use spatiotemporal features or semantic-interaction features separately. But the addressed problem requires simultaneous analysis of both the human motion and the object associated with him to form a comprehensive judgment.
- Existing work on multi-robot collaboration focuses on developing **robot-centric** algorithms at the **perception level**, and no literature aims to make decision-level information fusion for the reasoning of unique/critical individuals in the scene.

In summary, the uniqueness reasoning problem is somewhat similar to abnormal detection and role recognition but covers a wider range of fields. The ability to implement unique-

ness reasoning is to provide critical target information for robots to plan their behavior or movement in the environment more efficient and intelligent.

Chapter 3

Multi-Level Uniqueness Reasoning Architecture

This chapter illustrates the core idea, some specific definitions and prerequisites of the uniqueness reasoning, including the entire workflow, activity definitions, and mathematical modeling under different assumptions and conditions. It provides an overview and guideline of the entire work.

As shown in Figure 3.1, in this study, the whole reasoning process is decomposed into three modules, that is, perception, analysis, and inference.

- The Perception module is used to process raw sensor data from the camera and LiDAR for target detection and positioning. It generates the regions and labels of the detected targets and calculates their coordinates. We refer to the target region, label, and coordinates together as an observation of this target.
- Analysis module takes the observations from the perception module as input to extract the activity of each person in the scene. As defined in Section 1.2, the activities of an individual include not only the motion pattern but also the objects associated with this person. By performing semantic-interaction analysis, the relative position,

Figure 3.1: The whole reasoning process is decomposed into three modules, namely, perception, analysis, and inference.

distance, or overlapping between people and objects can be obtained. By applying spatiotemporal analysis, the movement frequency, relative or absolute velocity of people can be obtained.

- Inference module takes the extracted activities from the analysis module as input to judge the likelihood of each person to be the unique individual. Also as defined in Section 1.2, the person whose activity, in space and time, appear different from the other people in the scene is judged as the unique individual. The extracted activities can be either associated with specific human-labeled prior knowledge, or compared within the overall distribution in the scene to judge if it is unique.

The uniqueness is defined as the likelihood of a being the unique individual, and the entire process of weighing the evidence and inferring the likelihood is called uniqueness reasoning. By applying this modular structure, the output of each module can be evaluated at any time for easier debugging and maintenance. On the other hand, it allows flexible changes to the module to suit various tasks.

3.1 Definitions and Prerequisites

Before autonomous robots can judge the uniqueness of people, they first need to extract the relevant features from the environment as the evidence and understand how the evidence will link and influence the reasoning results over time. Therefore, feature extraction and selection are one of the prerequisites for performing uniqueness reasoning. In this section, we introduce the feature taxonomy, the approaches of target detection and localization, and the definition of spatial-semantic analysis.

3.1.1 Feature Taxonomy

Features are the most foundational elements of autonomous system execution perception, analysis, decision-making, and planning, and it is considered as the basis of intelligence. However, the features can be in various forms. In this study, we categorize the features into four levels, i.e., geometric feature, regional feature, semantic feature, and spatial-semantic feature. An example of the feature taxonomy is demonstrated in Figure 3.2.

- Geometric features are features of objects constructed by a set of geometric elements like points, lines, curves or surfaces. These features can be the corner, edge, blobs, ridges, salient points image texture and so on. It is the most fundamental and low-level feature in the feature taxonomy.
- Regional features are features that segment a scene into different regions by the contrast of information modalities, such as color, depth, thermal and so on. The regions can show the boundaries of different objects, but they do not contain any information about their class labels.
- Semantic features are upgraded versions of regional features. It can not only describe regional boundaries but also divide scenes according to different semantic classes. In other words, the regional feature may be able to distinguish different objects, but it does not know what they are, but the semantic feature can tell both.

(a) Geometric Feature

(b) Regional Feature

(c) Semantic Feature

(d) Spatial-Semantic Feature

Figure 3.2: An example of the feature taxonomy.

- Spatial-semantic features are the most high-level features in the feature taxonomy. It is not for a single element but expresses the spatial relationships between different semantic entities as well as their temporal changes. The relationship could be

Figure 3.3: The robot platform used in our research. The Husky robot from Clearpath Robotics, the VLP-16 3D Lidar, the Hokuyo 2D range finder, the Optris PI450 thermal camera, and ZED stereo camera.

described in either spatial spaces or spatiotemporal spaces.

High-level features are aggregates of many lower features. Being able to extract high-level features is the key to improve the intelligence and autonomy of robots as well as facilitate further interaction and collaboration with the human.

3.1.2 Heterogeneous Sensors Calibration

Future robots are desired to achieve a comprehensive and long-term perception. Thus heterogeneous sensors are critical because they provide complementary data characteristics. 3D LiDAR can record 360° environmental views with accurate depth information, which is essential for large-scale positioning and mapping in unstructured environments. How-

Figure 3.4: Diagram of the sensor frames and the two-step calibration method. (a) Sensor frames between three sensors: a 3D LiDAR $\{L\}$, a visual camera $\{C\}$ and a thermal camera $\{T\}$. (b) Calibration between 3D LiDAR and visual camera. (c) Calibration between visual and thermal camera.

ever, due to the particularity of the sensor, 3D LiDAR can only provide sparse geometric information, which makes detecting objects extremely challenging. At the same time, with advances in computer vision technology, visual and thermal cameras can compensate for the power of 3D LiDAR by accurately identifying and tracking dynamic objects during the day and night. Therefore, we need precise calibration to establish precise geometric relationships (rotation and translation) between the sensor frames and then fuse these data deeply. The robot platform is shown in Figure 3.3.

Generally, the geometric relationship between multiple sensors mounted on a robot is unknown. In this study, to obtain the 6 DOF extrinsic parameters among three sensors (a 3D LiDAR, a visual and a thermal camera), a two-step method is developed [56], as shown in Figure 3.4. The extrinsic parameter between a visual camera and a thermal camera is denoted as T_c^t and the extrinsic parameter between a visual camera and a 3D LiDAR is denoted as T_l^c .

In the first step, T_l^c is obtained by using a calibration toolbox [57] (Figure 3.4(b)). Six checkerboards are used, and they should cover the entire field of view for the toolbox to work properly. In particular, the top, bottom, left, right, and four corners of the field of view should be filled by the checkerboard.

In the second step, the thermal camera similarly provides spatial thermal images just like grayscale images from a visual camera. Thus, to obtain the extrinsic parameters T_c^t , the problem can be solved via standard stereo calibration solution as shown in Figure 3.4(c). The solution mainly follows Zhang's method [58] and OpenCV functions. A checkerboard with 4×11 asymmetric circular holes is employed.

At last, T_l^t can be calculated by multiplying T_c^t and T_l^c . Once the transformation parameters are obtained, the three sensors can be fused together for further applications.

3.1.3 Target Detection and Localization

3.1.3.1 Target Detection

Since the main idea of uniqueness reasoning is to analyze the real-time spatial relationship between specified semantic entities (e.g., objects and humans), it is a preliminary step to obtain the region and class label of them. To this end, any detection algorithm can be selected to generate the region of the detected entity. In order to verify the reliability of the proposed reasoning method, a detection algorithm with high precision and stability is preferred. Consequently, deep learning techniques are considered to be the most appropriate approach, and the Faster R-CNN [59] 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 target 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, and the obtained information is then used as the input of the reasoning process. The standard procedures are shown in Figure 3.5.

Figure 3.5: Schematic diagram of target detection using Faster R-CNN. The input is the raw color image and the outputs are labels and regions.

3.1.3.2 Target Localization

After extracting the class labels and regions of each entity, it is necessary to obtain their 3D position in the environment for distance measurement and temporal tracking. In this study, we use heterogeneous sensor fusion to calculate the precise 3D position of each entity. By calibrating the 3D LiDAR and the camera, an accurate 3D point cloud can be projected onto the image pixels, and a demonstration is shown in Figure 3.6. Thus, once the area of each entity is obtained from the target detection, its 3D position can be calculated as follows.

Definition 1 (Collection of Points). Let $S_{x,t}, S_{y,t}, S_{z,t}$ denote the collection of all points in the corresponding x, y, z axis within the bounding box at the timestamp t .

Definition 2 (Representative Coordinate). Let x_t, y_t, z_t denote the representative coordinate in the corresponding x, y, z axis at the timestamp t , which is obtained by taking the median of the collection of points in the corresponding axis.

$$x_t = \text{Median}(S_{x,t}), \quad y_t = \text{Median}(S_{y,t}), \quad z_t = \text{Median}(S_{z,t}) \quad (3.1)$$

Definition 3 (Representative Position). Let c_t denotes the representative position of the corresponding bounding box at timestamp t . 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

(a) The result of data fusion and dynamics separation by using 3D LiDAR and visual camera.

(b) The result of data fusion and dynamics separation by using 3D LiDAR and thermal camera.

Figure 3.6: A demonstration of sensor fusion and robot environmental perception. Multimodal information is integrated to provide a comprehensive understanding of the environment.

factor according to the camera specification.

$$c_t = (X_t, Y_t, Z_t) = (\delta x_t, \delta y_t, z_t), \quad (3.2)$$

3.1.4 Definition of Spatial-Semantic Analysis

In this study, the spatial-semantic analysis is a process of extracting human activity from multimodal information. It consists mainly of two sub-process, namely semantic-interaction analysis and spatiotemporal analysis.

3.1.4.1 Semantic-Interaction Analysis

The activity extracted from semantic-interaction analysis is generally denoted as Θ^{SI} , which represents the instant spatial relationship between the people h and a critical object o in the scene at the same timestamp t . It can be expressed as a function of the degree of overlap between the regions S of semantic labels, or the distance between the representative position c of semantic labels. The function varies depending on the chosen evaluation criteria, such as the relative position, relative distance, or overlapping.

$$\Theta^{SI} = f(c_t^h, c_t^o, S_t^h, S_t^o) \quad (3.3)$$

Specifically, interaction between human and specific object is critical in determining the human uniqueness. The unique individual in the scene may wear particular clothes or hold specific items. For example, “Chef” wearing “Tall Hat” and “Soldier” holding a “Gun” are representative examples.

3.1.4.2 Spatiotemporal Analysis

The activity extracted from spatiotemporal analysis is generally denoted as Θ^{ST} , which represents the spatial relationship of people according to temporal changes. It can be expressed as a function of the change in the representative position c of people under continuous timestamps. The function varies depending on the chosen evaluation criteria,

such as the moving frequency, relative or absolute velocity, and so forth.

$$\Theta^{ST} = f(c_t^h, c_{t-1}^h, \dots, c_{t-n}^h) \quad (3.4)$$

Specifically, the temporal changes of human motion is another critical feature to infer the human uniqueness. The unique individual may perform particular actions. For example, the “Patrol” always moves frequently in the environment and “Waitress” usually bows to the guest.

3.2 Mathematical Modeling

In this study, the uniqueness of each person is judged based on their performed activities, and the activities are extracted from observations. The uniqueness reasoning problem is studied at multiple levels under different assumptions and conditions.

3.2.1 General Uniqueness Reasoning Formulation

After acquiring sensory data, perception module detects the boundary area S of each label (superscript h for human and o for critical objects), and calculate their representative positions c . Let g^h and g^o define the corresponding observations (c and S) for all people and critical objects in the scene, correspondingly. Then, the analysis module performs semantic-interaction analysis and spatiotemporal analysis to extract the activities for each person, denoted as Θ^{SI} and Θ^{ST} . Finally, inference module either applies domain knowledge or neighborhood comparison to judge the uniqueness I of each person j from their activities. An example graphical model is illustrated in Figure 3.7, and the general process can be decomposed as below:

$$P_j(I_t | \Theta_t^{SI}, \Theta_t^{ST}, g_t^o, g_{t-1:t}^h) = \underbrace{P_j(I_t | \Theta_t^{SI}, \Theta_t^{ST})}_{\text{Uniqueness Judgment}} \cdot \underbrace{P_j(\Theta_t^{SI}, \Theta_t^{ST} | g_t^o, g_{t-1:t}^h)}_{\text{Activity Extraction}} \quad (3.5)$$

Figure 3.7: An example graphical model of the entire reasoning process for person j . The perception layer contains the observations of people, the feature layer contains the extracted activities of people, the decision layer is the judgment uniqueness of each person.

- Activity extraction is the process of modeling the performed activities of people from the observation. In the following contents, according to different assumptions and conditions, this process can be achieved either by taken qualitative rule (e.g., whether to overlap and whether to move) or quantitative analysis (e.g., actual intersection over union and relative velocity).
- Uniqueness judgment is the process of judging the likelihood of each person being the unique individual based on their activities. In the following contents, according to different assumptions and conditions, this process can be achieved by either applying human-labeled transition parameters or comparing within the overall distribution in the scene.

3.2.2 Uniqueness Reasoning at Different Levels

- For knowledge-based single robot uniqueness reasoning, the assumption is that it works in a specific domain where prior knowledge can be provided by human ex-

pertise in terms of how motion and an object associated with a person can be used to judge his uniqueness. The correspondence between observations and activities $P_j(\Theta_t^{SI}, \Theta_t^{ST} | g_t^o, g_{t-1:t}^h)$ is obtained through qualitative rules (e.g., whether to overlap and whether to move). The correspondence between activities and the uniqueness $P_j(I_t | \Theta_t^{SI}, \Theta_t^{ST})$ is obtained from the expertise experience such as case studies and surveys, and modeled through Bayes filter. The overall process can be decomposed as in (3.6), and this content will be detailed in Chapter 4.

$$\begin{aligned}
 P_j(I_t | \Theta_t^{SI}, \Theta_t^{ST}, g_t^o, g_{t-1:t}^h) &= \underbrace{P_j(I_t | \Theta_t^{SI}, \Theta_t^{ST})}_{\text{Uniqueness Judgment}} \cdot \underbrace{P_j(\Theta_t^{SI}, \Theta_t^{ST} | g_t^o, g_{t-1:t}^h)}_{\text{Activity Extraction}} \\
 &\propto \underbrace{P_j(I_t | \Theta_t^{SI})}_{\text{Ontology Model}} \cdot \underbrace{P_j(\Theta_t^{SI} | g_t^o, g_t^h)}_{\text{Semantic-Interaction}} \cdot \underbrace{P_j(I_t | \Theta_t^{ST})}_{\text{Motion Model}} \cdot \underbrace{P_j(\Theta_t^{ST} | g_{t-1:t}^h)}_{\text{Spatiotemporal}} \quad (3.6)
 \end{aligned}$$

- For analytics-based single robot uniqueness reasoning, we assume no prior knowledge (human labeling information) available. The correspondence between observations and activities $P_j(\Theta_t^{SI}, \Theta_t^{ST} | g_t^o, g_{t-1:t}^h)$ is modeled into two observation model, that is, object existence model (OEM) and human action model (HAM), through quantitative analysis of the observation (e.g., actual intersection over union and the relative velocity). Then, the distribution of the features characterizing a human activity is constructed in to a decision model, and the activity of an individual is analyzed within the distribution by using MRF ϕ . It also takes the previous judgment into account. It is more versatile and easier to apply to different scenarios. The overall process can be decomposed as in (3.7), and this content will be detailed in Chapter 5.

$$\begin{aligned}
 P_j(I_t | \Theta_t^{SI}, \Theta_t^{ST}, g_t^o, g_{t-1:t}^h) &= \underbrace{P_j(I_t | \Theta_t^{SI}, \Theta_t^{ST})}_{\text{Uniqueness Judgment}} \cdot \underbrace{P_j(\Theta_t^{SI}, \Theta_t^{ST} | g_t^o, g_{t-1:t}^h)}_{\text{Activity Extraction}} \\
 &\propto \underbrace{(\underbrace{\phi_j(\Theta_t^{SI} | g_t^o, g_t^h)}_{\text{OEM}} + \underbrace{\phi_j(\Theta_t^{ST} | g_{t,t-1}^h)}_{\text{HAM}})}_{\text{Decision Model}} \cdot \sum_I \underbrace{P_j(I_t | I_{t-1})}_{\text{Decision Transition}} \underbrace{P_j(I_{t-1} | \Theta_{1:t-1}^{ST}, \Theta_{1:t-1}^{SI})}_{\text{Previous Belief}} \quad (3.7)
 \end{aligned}$$

- For analytics-based multi-robot uniqueness reasoning, it deals with the problem in large-scale environments as well as a long-term operation. In such a case, it is still tricky for a single robot to address occlusion, limited field of view, and changing lighting condition. In this work, we focus more on the spatiotemporal analysis over the semantic-interaction analysis because of the large-scale and unstructured environments. Robots analyze the motion of people in the world reference frame rather than the local reference frame in previous chapters. As a distributed system, each robot first performs uniqueness reasoning based on their local observation. Meanwhile, robots also shared their local observations directly to each other to generate a global observation, and then the uniqueness is judged based on the global observation. For single robot uniqueness reasoning, it can be referred to as Chapter 5. For multi-robot uniqueness reasoning, we use superscript r to indicate the local robot frame and use superscript R to indicate it is in the global robot frame. This, let $T_t^{r,R}$ denotes the transformation between neighborhood robots, which is obtained from map fusion [60]. g_t^r denotes the observation of robot r , while g_t^R denotes the global observation. I_t^r is the uniqueness judged by single robot r , while I_t^R is the uniqueness judged by the multi-robot system. By sharing the information among robots, it can help robots enhance their belief or compensate for individual failures. This content will be detailed in Chapter 6.

$$\begin{aligned}
& P_j(I_t^R | \Theta_t^{STR}, g_{t-1:t}^R, T_{t-1:t}^{r,R}, I_t^r, \Theta_t^{STr}, g_{t-1:t}^r) \\
& \propto \underbrace{P_j(I_t^R | \Theta_t^{STR}, g_{t-1:t}^R)}_{\text{Global Uniqueness Reasoning}} \cdot \underbrace{P_j(g_{t-1:t}^R | g_{t-1:t}^r, T_{t-1:t}^{r,R})}_{\text{Global Dynamic Observation}} \cdot \underbrace{P_j(I_t^r | \Theta_t^{STr}, g_{t-1:t}^r)}_{\text{Local Uniqueness Reasoning}} \quad (3.8)
\end{aligned}$$

Chapter 4

Knowledge-Based Single Robot Uniqueness Reasoning

This chapter focuses on using the knowledge-based approach to deal with a specific security scenario (motivated by kidnapping scenario). But not limited to this case, the proposed method can apply to other scenarios with similar needs. This work has been published on [61, 62].

4.1 Introduction

This work presents an integrated probabilistic reasoning strategy for a robot to infer the uniqueness that humans fulfill in a specific scenario (motivated kidnapping scenario). Based on the application background, the unique individual in this chapter is defined as the person who is holding a critical object or frequently moving in the environment. A critical object is usually pre-specified in the corresponding scenario. We use ‘object’ in the following content to represent the critical object. The knowledge is provided in terms of how the activities can be used to infer the uniqueness of people, as determined by human expertise who labeled the datasets as ground truth. Based on Bayes’ theorem,

Figure 4.1: The overall block diagram of the knowledge-based single robot uniqueness reasoning. The proposed method starts with human and object detection to obtain their corresponding regions. Next, DBI measures the semantic-interaction feature to find the relative distance between humans and the critical object, while KBI feeds both the object existence and human action into a Bayes filter. Then, the final judgment of the uniqueness is made by the weighting of these two indicators.

qualitative rules are used to extract human action (moving frequency) and associated objects (object existence) from observations. Then the prior knowledge is applied to judge the uniqueness of each person based on their performed activities.

Knowledge-Based Inference (KBI) is used as the primary indicator, and the guidance mathematical formulation is as (3.6). KBI focuses on the activities related to human action and the object existence. In this chapter, *Distance-Based Inference (DBI)* serves as an aid to the primary indicator. DBI deduces the uniqueness based on the relative distance between people and the object. The final uniqueness is estimated using a weighted fusion scheme based on the information entropy. Experiments were conducted in indoor environments with various degrees of occlusion, and the experimental results showed the rationality and accuracy of the reasoning.

4.2 Main Approach

In this work, The uniqueness are judged as binary representations, the unique individual (i^1) and usual person (i^2), $I \in \{i^1, i^2\}$. The overall schematic diagram is shown in Figure 4.1, and the main approach contains three processes, which are listed below.

- KBI is performed, which takes qualitative human actions (whether to move) as the spatiotemporal feature and qualitative object existence (whether to overlap) as the semantic-interaction feature. It will be detailed in Section 4.2.1.
- An DBI is performed to support the knowledge-based approach by taking distance measure between involved people and the critical object. It will be detailed in Section 4.2.2.
- A weighted fusion scheme is developed by using the information entropy to integrate the result of both KBI and DBI to generate the final uniqueness judgment. It will be detailed in Section 4.2.3.

4.2.1 Knowledge-Based Inference (KBI)

This indicator is a local probability estimation, which is used to independently estimate the likelihood of whether the person could be the unique individual. Prior knowledge is provided and directly applied to the Bayes Filter. In this research, the object existence and human action are regarded as the activities of people. The person who is holding the critical object and frequently moving in the scene has a higher probability of being judged as a unique individual.

Figure 4.2: Object existence is defined as whether the region of the object is intersected with the human region, and only when they intersect in the entire 3D space, then the existence is positive. S^o and S^h are the region of *object* and *human*, respectively. z is the distance to the camera and ϵ is a threshold for determining the intersection in z direction.

4.2.1.1 Activity Extraction

Semantic-Interaction Analysis - Object Existence

In this chapter, the semantic-interaction analysis considers the relative spatial position between people and a critical object. In different scenarios, there could have different critical objects that require particular attention. People carrying such objects can be judged as the unique individual, and their likely future activities may be predicted. Therefore, the existence of such items in the human body is critical to recognize the unique individual.

Let E denote object existence, which represents the Θ^{SI} feature in this chapter. It is assumed that all objects are independent and the existence is either positive or negative, which can be denoted as e_1 and e_2 , respectively, $E \in \{e_1, e_2\}$. The $p_j(I_t | \Theta_t^{SI})$ shown in (3.6) can be rewritten as $p_j(I_t | E_t)$. If there are multiple objects, the relationship between the uniqueness and the existence of the n th object is denoted as $p_j(I_t | E_{n,t})$.

More specifically, the traditional definition of the object existence is for the entire scene, which will give positive as long as the object is detected in the image. However, in this research, object existence is defined as whether the object exists in the human region. Let ϵ is a threshold for determining whether there is an intersection in the depth direction.

Figure 4.3: The motion of human is analyzed by comparing with its consecutive frame. If the difference is below a certain level, then the action is recognized as *Stationary*; otherwise, the action is recognized as *Moving*. c_t^h is the representative point for each person's region at the t th frame.

Only when regions of human and object intersect in the 3D space, then the existence is assigned positive, as illustrated in Figure 4.2 and (4.1).

$$E_t = \begin{cases} e_1(Positive) & \text{if } S_t^h \cap S_t^o \neq \emptyset \text{ and } |z_t^h - z_t^o| \leq \epsilon \\ e_2(Negative) & \text{Otherwise} \end{cases} \quad (4.1)$$

Spatiotemporal Analysis - Human Action

In this chapter, the spatiotemporal analysis is determined by the spatial variation of people in time series. Human action can reveal their different functions when they are performing certain tasks. Recognizing human action can help to understand their uniqueness in the environment. Let A represent human action, and it could be of many kinds, such as standing, sitting, lying and so forth, which can be denoted as $A \in \{a_1, a_2, \dots, a_n\}$. In this chapter, mainly two qualitative actions are considered, *Moving* and *Stationary*, which are denoted as a_1 and a_2 , respectively. $p_j(I_t | \Theta_t^{ST})$ shown in (3.6) can be rewritten as $p_j(I_t | A_t)$.

Figure 4.4: An example of the KBI. It is assumed that people who carry the specified object or conduct specific action become the unique individual (i^1). The bar charts show the probability of being the unique individual (i^1) and usual person (i^2) for Person 1 (Blue) and Person 2 (Orange), respectively.

More specifically, the action of human at each decision frame is analyzed by comparing with its previous frame. The human action is recognized as *Moving* if the difference exceeded certain level; otherwise, the action is recognized as *stationary*. As illustrated in Figure 4.3 and (4.2), the 3D position of human at each frame is represented by a single point. Let σ denotes the threshold for determining the motion between the consecutive frames.

$$A_t = \begin{cases} a_1(Moving) & \text{if } \|\mathbf{c}_t^h - \mathbf{c}_{t-1}^h\| \geq \sigma \\ a_2(Stationary) & \text{if } \|\mathbf{c}_t^h - \mathbf{c}_{t-1}^h\| < \sigma \end{cases} \quad (4.2)$$

4.2.1.2 Uniqueness Judgment from Bayes Filter

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

$$p_j(I_t | A_t, E_{1,t}, \dots, E_{n,t}) \propto p_j(A_t, E_{1,t}, \dots, E_{n,t} | I_t) = \underbrace{p_j(A_t | I_t)}_{\text{Human Action}} \cdot \prod_{k=1}^n \underbrace{p_j(E_{k,t} | I_t)}_{\text{Object Existence}} \quad (4.3)$$

Initially, the probability of becoming a unique individual or usual person are set equal.

The last equation is derived based on the conditional independence between the two channels of information.

An illustration is shown in Figure 4.4. Before the final decision, the normalization of the probability is performed first. The probability of uniqueness for the j th person is illustrated as below, where the superscript $\mathcal{K}$ means probability from KBI.

$$p_j^{\mathcal{K}}(i_t^1) = \frac{p_j(i_t^1)}{p_j(i_t^1) + p_j(i_t^2)}, \quad p_j^{\mathcal{K}}(i_t^2) = 1 - p_j^{\mathcal{K}}(i_t^1) \quad (4.4)$$

4.2.1.3 Confidence of KBI Result

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

$$H_j^{\mathcal{K}}(I_t) = - \sum_I p_j^{\mathcal{K}}(I_t) \log p_j^{\mathcal{K}}(I_t) \quad (4.5)$$

4.2.2 Distance-Based Inference (DBI)

This indicator serves as an aid to the primary indicator. Because by applying the qualitative rules, in the case where the critical object existence in the scene but does not overlap with any people, KBI can not assign reasonable belief to each people from the semantic-interaction analysis. It is a global probability estimation, which is used to estimate who is more likely to be the unique individual among a group of people based on the relative distance to the critical object.

4.2.2.1 Uniqueness Judgment from Relative Distance Measurement

After obtaining the regions of humans and the object in the scene, we first calculate the representative coordinate for each semantic label, which is the mean value of all coordinates in the corresponding region. The distance between the object and each person is

Figure 4.5: An example of the DBI. It is assumed that the person who is closer to the object becomes the unique individual (i^1), others are usual people (i^2). The bar charts show the probability of being the unique individual for Person 1 (blue) and Person 2 (orange).

actually the norm of the 3D vector between the representative coordinates. The ratio is calculated by comparing the norm of the 3D vector. Let $d_{j,t}$ denotes the distance between the object and the j th person at time t , D_t is the sum of distances from the object to each person at time t . The probability of a person being the unique individual is inversely proportional to the distance to the object. The probability of being the unique individual for the j th person is presented in (4.7), where the superscript $\mathcal{D}$ means probability from the DBI. An example is illustrated in Figure 4.5.

$$D_t = \sum_j d_{j,t} \quad (4.6)$$

$$p_j^{\mathcal{D}}(i_t^1) = \frac{D_t - d_{j,t}}{D_t}, \quad p_j^{\mathcal{D}}(i_t^2) = \frac{d_{j,t}}{D_t} \quad (4.7)$$

4.2.2.2 Confidence of DBI Result

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

Figure 4.6: An illustration of the final decision process. After obtaining the inferred results from all indicators, the final uniqueness is determined through a weighted fusion of all the results.

below:

$$H_j^{\mathcal{D}}(I_t) = - \sum p_j^{\mathcal{D}}(I_t) \log p_j^{\mathcal{D}}(I_t) \quad (4.8)$$

4.2.3 Final Uniqueness Judgment

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

$$\alpha_1 = \frac{H_j^{\mathcal{K}}(I_t)}{H_j^{\mathcal{D}}(I_t) + H_j^{\mathcal{K}}(I_t)} \quad \alpha_2 = \frac{H_j^{\mathcal{D}}(I_t)}{H_j^{\mathcal{D}}(I_t) + H_j^{\mathcal{K}}(I_t)} \quad (4.9)$$

The less the entropy, the higher the confidence of the judgment results. Combining the results of KBI and DBI, we can determine the final uniqueness of the j th person as below:

$$p_j(I_t) = \alpha_1 p_j^{\mathcal{D}}(I_t) + \alpha_2 p_j^{\mathcal{K}}(I_t) \quad (4.10)$$

$$I_{j,t} = \arg \max_{I \in \{i^1, i^2\}} (p_j(I_t)) \quad (4.11)$$

Algorithm 1 Uniqueness Reasoning**Input:** $g_{t,t-1}^h$ and g_t^o **Output:** Bounding box of the unique person

```

1: function JUDGMENT(Uniqueness)
2:   1. Knowledge-Based Inference (KBI)
3:   Extract  $A_t$  and  $E_t$ 
4:   Feed  $A_t$  and  $E_t$  into Bayes Filter
5:   Probability of being the unique individual( $i^1$ )  $\leftarrow p_j^K(i_t^1)$ 
6:   Calculate confidence  $H_j^K$ 
7:
8:   2. Distance-Based Inference (DBI)
9:   Measure the relative distance between  $S_t^h$  and  $S_t^o$ 
10:  Probability of being the unique individual( $i^1$ )  $\leftarrow p_j^D(i_t^1)$ 
11:  Calculate confidence  $H_j^D$ 
12:
13:  3. Final Decision
14:  Calculate weights  $\alpha_1, \alpha_2$  based on  $H_j^K$  and  $H_j^D$ 
15:  Weighted fusion  $\leftarrow p_j(I_t)$ 
16:  return Determined uniqueness  $I$ 
17: end function

```

Figure 4.6 illustrates the final reasoning process. The overall procedure of uniqueness reasoning is described in Algorithm 1.

4.3 Experiments

According to the research background, the motivation of this experiment is inspired by a security case. The robot police need to identify the victims and terrorists in the environment autonomously, so they need to identify their carrying object (gun) and motion pattern (moving frequency). Therefore, we use several water bottles to indicate critical objects and analyze human motion in some indoor environments.

The sensor used throughout the experiment is the ASUS Xtion Pro Live camera. Faster R-CNN [59] was employed to detect humans and the object in the scene. Since object detection is not the focus of this study, we use the existing network parameters which are trained on VOC 2007 [63] and VOC 2012 [64] dataset. Three separate datasets were

Table 4.1: Conditional Probability for Uniqueness Reasoning in Kidnapping Scenario.

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

		Uniqueness (I)	
		Unique (i^1)	Usual (i^2)
Action (A)	Moving (a_1)	0.54	0.39
	Stationary (a_2)	0.46	0.61

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

		Uniqueness (I)	
		Unique (i^1)	Usual (i^2)
Existence (E)	Positive (e_1)	0.56	0.01
	Negative (e_2)	0.44	0.99

collected from different indoor environments.

4.3.1 Prior Knowledge Acquisition

Prior knowledge is an important part to infer the uniqueness. All parameters are obtained from the case study and surveys. By consulting experts (100 police officers) in the field and knowing of a similar scenario, prior knowledge was obtained in advance. On the one hand, it could be inferred that the unique individual might carry some potential weapons and perform more movements in the environment. On the other hand, prior knowledge about determining the appropriate interaction time and the corresponding actions could be leveraged. Action-wise, when the rescuer (robot in this case) wants to conduct rescue operations, it is best to participate when the unique individual is in a low alert state. Usually, the unique individual is vigilant when they are walking around and lower their alertness when they stay still. Position-wise, it should be when the unique individual is near the windows or in an open space and not close to any usual person because they do not want to cause any accidental damage to the usual person.

To acquire prior knowledge, a survey has been sent out to 100 ordinary people and the 50 experts in the field. This survey is a questionnaire, by asking how they would behave if they were in this kind of scenario, the statistical distribution of the knowledge could be

learned. For example, we asked ‘Assuming you are a unique individual, are you going to keep stationary or moving around?’, in the end, 54% of the answers chose that they would move around, so the number ‘0.54’ was assigned in (a) of Table 4.1. All the numbers are acquired similarly, and we weighted 0.7 for the experts’ answers whereas 0.3 for the ordinary people’ answers.

4.3.2 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 a 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 Figure 4.7. The actual depth value does not change after the preprocessing, the color difference is just because we used a different color map to increase the contrast for better visualization, and no information is lost during this process.

4.3.3 Experimental Description

To evaluate the quality of the proposed reasoning method, a series of scenarios are investigated in this experiment. As shown in Figure 4.8. In each dataset, two persons perform a series of activities in the scene. Three different bottles, regarding shape, size, and color, are chosen as the critical object. For each dataset, we collect approximately 2 minutes of video and convert to 30 frames per second. In each dataset, we divide it into six most common scenarios based on the ground truth motion of the unique individual, motion of the usual person, and the existence of the critical object.

Figure 4.7: Results of raw data preprocessing. Images on the left are raw depth images, and images on the right are filtered depth images. 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.

Let U , O , and E represent for *Unique* (unique individual), *Others* (usual persons) and *Existence*, respectively. m and s denote the human action *moving* and *stationary*. p and n represent for the object existence *positive* and *negative*. Six most common scenarios are evaluated based on the ground truth human action and object existence as shown in Table 4.2. For example, the scenario ' $U_m O_s E_n$ ' denotes that the unique individual is moving while the usual person stays stationary and the bottle does not exist in the scene.

Besides, to get a complete and fair evaluation result, scenario weight γ is introduced to describe the possibility of occurrence of each scenario. The more scenarios there are, the higher the weight assigned, and the less the scenario, the lower the weight assigned. The weight for each scenario is calculated based on Table 4.1, and the calculated values are shown in Table 4.2. For each scenario, weight γ is calculated as follows and then normalized into $[0,1]$:

$$U_m O_m E_p \rightarrow \gamma_1 = p(a_1|i^1)p(a_1|i^2)p(e_1|i^1), U_m O_m E_n \rightarrow \gamma_2 = p(a_1|i^1)p(a_1|i^2)p(e_2|i^1),$$

$$U_m O_s E_p \rightarrow \gamma_3 = p(a_1|i^1)p(a_2|i^2)p(e_1|i^1), U_m O_s E_n \rightarrow \gamma_4 = p(a_1|i^1)p(a_2|i^2)p(e_2|i^1),$$

$$U_s O_s E_p \rightarrow \gamma_5 = p(a_2|i^1)p(a_2|i^2)p(e_1|i^1), U_s O_s E_n \rightarrow \gamma_6 = p(a_2|i^1)p(a_2|i^2)p(e_2|i^1)$$

In this experiment, the input features of the autonomous reasoning system are extracted from three different data sources (i.e., RGB image, depth image, and RGBD image), and four reasoning methods were evaluated. First, by analyzing RGB images, people and the specified object were detected using Faster R-CNN [59], and the judgment is made by measuring the 2D distance (d_{2D}) between them. Second, by processing depth images, people and the specified object were detected using template matching [65], and the judgment is made by measuring the 3D distance (d_{3D}) between them. For these reasoning methods from distance-based measurement, the unique individual is identified when the object is detected to be intersected with this person. Another reasoning method is through optical flow analysis (*OF*) [66], where the judgment is made by measuring the spatial changes between successive time steps, and the person with distinctive movement is identified as the unique individual. The last method use only *KBI* from analyzing RGBD images, which is the submodule of the proposed method. The proposed approach also works on RGBD images and decides by jointly considering *KBI* and *DBI*.

Thus, the desired unique individual is the person who is distinctive in movement (more frequently) and holds the critical object (bottle).

4.3.4 Experimental Results

For each dataset, all the frames were split into six scenarios for evaluation. Judgment accuracy for both the unique individual and the usual person in each examined method was obtained by taking the F-measure [67]. *Score* is the final measurement for each method by multiplying the F-measure result with the corresponding scenario weight.

Results of each dataset were shown in Table 4.2 and the average performance of the overall reasoning result was revealed in Figure 4.9. For judgments based on the analysis of 2D images (d_{2D}), it can only identify the unique individual (the person with the bottle)

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

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

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

Figure 4.8: Results of the proposed reasoning method. Experimental data are collected from three separate indoor environments with different settings and human activities. Three bottles are used for the critical object which is different in type, size, and color. Two colors, i.e., red and green, represent the judgment result for the unique individual and usual person, respectively.

Table 4.2: Different reasoning methods were evaluated in six scenarios for each of the three datasets. The values are calculated from F-measure (in %) for the unique individual (U) and usual person (O). *Score* is the final weighted measurement of each method.

Method		Scenario ₁												Scenario ₂												Scenario ₃												Scenario ₄												Scenario ₅												Scenario ₆												Score	
		$U_m O_m E_p$						$U_m O_m E_n$						$U_m O_s E_p$						$U_m O_s E_n$						$U_s O_s E_p$						$U_s O_s E_n$																																											
		$\gamma_1 = 0.14$						$\gamma_2 = 0.11$						$\gamma_3 = 0.23$						$\gamma_4 = 0.18$						$\gamma_5 = 0.19$						$\gamma_6 = 0.15$																																											
Decision		Input																																																																									
d_{2D}		RGB Image		9.9		65.7		0.0		63.8		0.0		70.1		3.0		66.2		50.9		60.0		1.3		64.8		11.8		65.4																																													
d_{3D}		Depth Image		0.0		33.7		0.0		46.6		0.0		64.3		0.0		60.3		0.0		65.1		0.0		46.2		0.0		54.8																																													
OF		RGB Image		19.8		0.0		28.1		0.0		55.1		0.0		33.6		0.0		15.7		0.0		6.5		0.0		28.5		0.0																																													
KBI		RGBD Image		54.6		44.4		40.0		57.1		40.0		57.1		25.0		73.7		44.4		25.0		25.0		53.2		38.4		51.6																																													
$Proposed$		RGBD Image		54.6		44.4		40.0		58.1		57.6		40.0		57.1		45.0		73.7		44.4		13.3		43.5		51.6		58.2																																													
d_{2D}		RGB Image		2.0		0.0		25.7		34.2		0.0		7.7		0.0		6.1		11.1		27.9		0.0		0.0		5.2		11.9																																													
d_{3D}		Depth Image		0.0		0.0		0.0		40.2		0.0		0.0		0.0		29.3		0.0		3.0		0.0		45.8		0.0		17.5																																													
OF		RGB Image		26.7		0.0		41.3		0.0		41.5		0.0		49.5		0.0		16.7		0.0		33.7		0.0		35.0		0.0																																													
KBI		RGBD Image		60.0		33.3		58.8		36.4		60.0		33.3		20.0		22.2		57.1		40.0		50.0		50.0		50.6		35.4																																													
$Proposed$		RGBD Image		62.1		37.1		63.6		42.9		62.1		36.7		26.1		29.1		58.8		36.4		54.6		44.4		54.0		37.2																																													
d_{2D}		RGB Image		60.0		51.8		0.0		64.3		28.0		57.3		0.0		66.2		0.0		67.0		0.0		66.7		14.8		62.2																																													
d_{3D}		Depth Image		3.2		30.9		0.0		25.8		0.0		12.4		0.0		39.1		0.0		51.4		0.0		23.7		0.0		30.4																																													
OF		RGB Image		13.4		0.0		6.7		0.0		66.1		0.0		30.5		0.0		11.5		0.0		0.0		0.0		25.5		0.0																																													
KBI		RGBD Image		61.5		28.6		40.0		57.2		67.1		40.0		58.8		36.4		50.0		50.0		0.0		66.7		48.5		45.5																																													
$Proposed$		RGBD Image		59.5		41.1		50.0		50.0		57.1		40.0		61.5		48.6		61.6		54.6		57.2		40.0		58.3		45.6

Figure 4.9: The overall uniqueness judgment result for each reasoning method. The proposed method achieved the highest accuracy in recognizing both the unique individual and the usual person.

when the bottle is successfully detected. Moreover, the performance can be affected due to occlusion and various viewing angles. For judgments based on analyzing depth images (d_{3D}), it is difficult to detect a person when encountering a variety of postures and occlusion. Besides, since a small object detection from only depth data is very challenging, the robot cannot recognize the unique individual because of failing to detect the bottles. For these reasoning methods using distance measurements, they tended to recognize all detected persons as usual persons due to the limitation of detection. For judgments based on OF , it can only recognize the unique individual (the person with distinctive movement) in the presence of a prominent moving speed, but it might fail if people are stationary or moving at similar speeds. Consequently, by combining both KBI and DBI, the proposed method well compensates for the limitation of failing to detect the single attribute by the ability to fuse information from different perspectives. Compared with other methods in each dataset, it produced the highest accuracy among all methods for recognizing the unique individual and a competitive result for recognizing the usual person. Even in some complex scenes, the performance remained stable and outperformed its submodule *KBI*.

The bias of each reasoning method was shown in Figure 4.10. It revealed that the pro-

Figure 4.10: Bias of each reasoning method. Blue and orange represent the bias that recognizes the unique individual and the usual person, respectively. It can be seen that d_{2D} and d_{3D} are more advantageous in recognizing usual persons while OF can only spot out the unique individual. The proposed method has the most balanced judgment ability for the unique individual and the usual person.

posed method has the most balanced reasoning ability for both the unique individual and the usual person. The two reasoning methods, i.e., d_{2D} and d_{3D} , tend to recognize the people as usual persons since they are hard to detect the bottle. For judgments based on OF , it can recognize the unique individual, but unable to distinguish between this person and the usual person.

4.4 Conclusions

In this chapter, a comprehensive probabilistic reasoning approach is proposed. The method enables a mobile robot to recognize the unique individual by using probabilistic reasoning from both spatial and semantic information. The idea behind the approach is to find associations between the prior knowledge and empirical qualitative spatial relationships between humans and critical objects as well as their temporal changes. Semantic-interaction

and spatiotemporal analysis have been applied to extract corresponding features. Different judgment results are fused into an integrated reasoning decision by using the information entropy. As a result, the proposed method can well compensate for the limitations that hard to detect a single attribute and shows a balanced and stable uniqueness reasoning capability.

Chapter 5

Analytics-Based Single Robot Uniqueness Reasoning

The previous chapter introduced a knowledge-based reasoning approach that adequately addressed the uniqueness reasoning problem in a specific scenario. However, it lacks flexibility and generality to adapt to different situations. Therefore, this chapter aims to use quantitative models to quantify measurements and observations to analyze the uncertainty in the reasoning process and thereby improve its applicability. This work has been published on [68].

5.1 Introduction

This work deals with the problem of recognizing the unique individual by using analytical models. In this work, we propose a hierarchical probabilistic reasoning method that relates quantitative spatial relationships between critical objects and humans with their temporal changes to recognize the unique individual. The guidance mathematical formulation is as (3.7). Two observation models, *Object Existence Model (OEM)* and *Human Action Model (HAM)*, are established to support uniqueness reasoning by analyz-

Figure 5.1: An illustration of uniqueness reasoning process. First, humans and objects are detected for obtaining the observation (region and position). Then, OEM and HAM are then built to extract human activities by performing corresponding semantic interactions and spatiotemporal analysis. Finally, the uniqueness of each person is judged by comparing the activities with the overall distribution.

ing the corresponding semantic-interaction features and spatiotemporal features. Then, OEM and HAM results of each person are compared with the overall distribution in the scene, respectively. Finally, we can determine the uniqueness 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, exhibiting stable performance even in complex scenes. An illustration is shown in Figure 5.1.

5.2 Main Approach

Specifically, when the autonomous system is sent into an unknown environment, first, it needs to detect people and objects for acquiring regions of them. Next, the system is going to extract the representative positions 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 (spatiotemporal feature). After that, we compare each individual's OEM and HAM results with the overall distribution in the scene, respectively. Finally, the unique individual can be recognized by integrating all observation models into a comprehensive reasoning scheme. 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, and so forth. The individual theoretical basis is explained in the following parts, including the construction of the probabilistic model, how those channels are integrated, and how final judgments are made. The overall schematic diagram is shown in Figure 5.2, and the main processes are listed as below.

- The overall mathematical modeling and decomposition is discussed in Section 5.2.1.
- Two observation models are constructed to model human activities from the observation, namely, object existence model (OEM) and human action model (HAM). OEM performs semantic-interaction analysis that measures the overlapping area between people and critical objects. Meanwhile, HAM performs spatiotemporal analysis that measure the relative moving speed of people with respect to the camera. It is detailed in Section 5.2.2
- The uniqueness of each person is judged by comparing their performed activities within the overall distributions in the scene, which is illustrated in Section 5.2.3.

Figure 5.2: The overall diagram of the analytics-based single robot uniqueness reasoning. The uniqueness of people is inferred from two observation models. OEM analyzes the overlapping area between people and objects. HAM analyzes the relative speed of people with respect to the camera. Then, the activities of each person are compared within the overall distribution in the scene. Finally, the final judgment result can be obtained by fusing all channels of information.

5.2.1 Reasoning Process Decomposition

The objective is to estimate a distribution over the uniqueness about what the person could be by giving the human action and object existence as input. This problem is formulated as a Bayes filter. The robot observes the activities of each person, including the semantic-interaction feature Θ^{SI} and spatiotemporal feature Θ^{ST} , and estimates a distribution over the current uniqueness at each timestamp:

$$P_j(I_t | \Theta_{1:t}^{SI}, \Theta_{1:t}^{ST}) \quad (5.1)$$

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 the uniqueness, whereas the time update updates the belief of the

uniqueness given previous information.

$$\begin{aligned}
 P_j(I_t | \Theta_{1:t}^{SI}, \Theta_{1:t}^{ST}) &= \frac{P_j(\Theta_t^{SI}, \Theta_t^{ST} | I_t) \cdot P_j(I_t | \Theta_{1:t-1}^{SI}, \Theta_{1:t-1}^{ST})}{P_j(\Theta_t^{SI}, \Theta_t^{ST} | \Theta_{1:t-1}^{SI}, \Theta_{1:t-1}^{ST})} \\
 &\propto \underbrace{P_j(\Theta_t^{SI}, \Theta_t^{ST} | I_t)}_{\text{Decision Model}} \cdot \underbrace{P_j(I_t | \Theta_{1:t-1}^{SI}, \Theta_{1:t-1}^{ST})}_{\text{Previous Information}}
 \end{aligned} \tag{5.2}$$

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

$$P_j(\Theta_t^{SI}, \Theta_t^{ST} | I_t) = P_j(A_t, E_t | I_t) \tag{5.3}$$

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

$$P_j(I_t | \Theta_{1:t-1}^{SI}, \Theta_{1:t-1}^{ST}) = \sum \underbrace{P_j(I_t | I_{t-1})}_{\text{Decision Transition}} \cdot \underbrace{P_j(I_{t-1} | \Theta_{1:t-1}^{SI}, \Theta_{1:t-1}^{ST})}_{\text{Previous Belief}}, \tag{5.4}$$

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

5.2.2 Activity Extraction

The general types of the activity in this chapter is similar to Section 4.2.1.1, but the way to extract them are different from the previous chapter. Since there are two types of activities for determining the uniqueness of human, two observation models are established for the corresponding object existence and human actions, namely, *Object Existence Model (OEM)* and *Human Action Model (HAM)*. The observation model calculates the probability of the observation given corresponding activities. Each observation is a set of regions and positions of the human and the object, $\langle S, c \rangle$.

5.2.2.1 Object Existence Model (OEM)

Let E denotes the object existence in this chapter. The object existence is modeled as the intersection area between a person and the object over the union of their area at each timestamp. First, the region of the person j ($S_{j,t}^h$) and the object (S_t^o) are extracted. Let $P_j(S_t|E_t)$ define the probability of the person j carrying the critical object at time t , which is set to be proportional to the area of the intersection. As illustrated in (5.5), it limits the probability from 0 to 1.

$$P_j(S_t|E_t) \propto \frac{S_{j,t}^h \cap S_t^o}{S_{j,t}^h \cup S_t^o} \quad (5.5)$$

5.2.2.2 Human Action Model (HAM)

Let A denotes the object existence in this chapter. Human action is modeled as a positional difference between consecutive timestamps. First, the representative point $c_{j,t}^h$ is extracted for person j at timestamp t . Next, the movement vector between the two consecutive timestamps 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 study. As illustrated in (5.6), it limits the probability from 0 to 1.

$$P_j(c_t|A_t) \propto \frac{1}{1 + e^{f(c_{j,t}^h - c_{j,t-1}^h)}} \quad (5.6)$$

5.2.3 Uniqueness Judgment

5.2.3.1 Decision Model

The decision model calculates the probability of the respective activities given the uniqueness. Each activity is a set of human actions and the object existence, $\langle A, E \rangle$.

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

$$P_j(\Theta_t^{SI}, \Theta_t^{ST} | I_t) = P_j(A_t | I_t) \cdot P_j(E_t | I_t) \quad (5.7)$$

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

To simplify the calculate process, we take the logarithm on both sides to turn the multiplication into addition while keeping its trend unchanged. the processed result after the logarithmic function is proportional to the original. Thus, the relationship can be further decomposed as:

$$\begin{aligned} P_j(\Theta_t^{SI}, \Theta_t^{ST} | I_t) &\propto \log(P_j(\Theta_t^{SI}, \Theta_t^{ST} | I_t)) = \log(P_j(A_t | I_t) \cdot P_j(E_t | I_t)) \\ &= \log(P_j(A_t | I_t)) + \log(P_j(E_t | I_t)) \quad (5.8) \\ &= L_j(A_t | I_t) + L_j(E_t | I_t), \end{aligned}$$

where L denotes the function after taking the logarithm.

5.2.3.2 Decision Transition

We assumed that the uniqueness of a person is unlikely to change in a short period of time. Thus, at each timestamp, there is a relatively larger probability, q , of transitioning to the same uniqueness. Otherwise, the transition probability is uniformly assigned based on the total types of uniqueness, $|\mathcal{I}|$. In this work, there are only two types of uniqueness, namely, unique individual and usual people, thus $|\mathcal{I}| = 2$. q is set as 0.7 in this work.

$$P_j(I_t | I_{t-1}) = \begin{cases} q, & \text{if } I_t = I_{t-1} \\ \frac{1-q}{|\mathcal{I}|-1}, & \text{otherwise} \end{cases} \quad (5.9)$$

Figure 5.3: An example of five detected persons in the scene, and they form a fully connected undirected graphical. $x_{j,t}$ represents the single inferred result for person j either from OEM or HAM. Then, every single result is compared with each other to measure the deviation from the overall distribution for generating a complete result.

5.2.3.3 Comparison within the Overall Distribution

The distance between the same semantic category can reveal the uniqueness 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 individual 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 study, the final judgment 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 Figure 5.3. Let $\mathcal{V} = \{1, 2, \dots, n\}$ 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:

$$\Gamma(j) = \{l \in \mathcal{V} | (l, j) \in \mathcal{L}\} \quad (5.10)$$

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 timestamp 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:

$$\phi_j(x_{j,t}) = \frac{1}{Z} \cdot \prod_{c \in \mathcal{C}} x_{c,t}, \quad (5.11)$$

where Z is a partition function for normalization:

$$Z = \sum_{x_{j,t}} \prod_{c \in \mathcal{C}} x_{c,t} \quad (5.12)$$

5.2.3.4 Final Judgment

After finishing building the two observation models, these two feature states can be used as evidence to determine the final uniqueness. By referring to (5.11), the final uniqueness reasoning results for each person can be obtained as follows.

OEM Judgment

The judgment result from OEM is proportional to the degree of derivation of the existence from the overall distribution.

$$L_j(E_t|I_t) \propto \phi_j(P_j(S_t|E_t)) \quad (5.13)$$

HAM Judgment

The judgment result from HAM is proportional to the degree of derivation of human action from the overall distribution and the previous action.

$$L_j(A_t|I_t) \propto \phi_j(P_j(c_t|A_t)) \quad (5.14)$$

Belief Fusion

By acquiring the information from all channels, the entire reasoning process forms a hierarchical graphical model. Updating from (5.2), the uniqueness of each person can be finally inferred as follows:

$$\begin{aligned} P_j(I_t|E_{1:t}, A_{1:t}) &\propto \underbrace{(\phi_j(P_j(S_t|E_t)))}_{\text{OEM}} + \underbrace{\phi_j(P_j(c_t|A_t))}_{\text{HAM}} \\ &\quad \underbrace{\hspace{10em}}_{\text{Decision Model}} \\ &\cdot \sum \underbrace{P_j(I_t|I_{t-1})}_{\text{Decision Transition}} \underbrace{P_j(I_{t-1}|A_{1:t-1}, E_{1:t-1})}_{\text{Previous Belief}} \end{aligned} \quad (5.15)$$

The probability that each person becomes the unique individual or the usual person is normalized after each judgment, and the following judgment is based on the previous normalized value.

5.3 Experiments

This experiment attempts to remove the assumptions of working in specific domains and focus on a wider range of potential applications to show the generality of the proposed algorithm. Both indoor (Section 5.3.2) and outdoor (Section 5.3.3) environments are investigated to recognize the unique individual in the scene. In indoor experiments, the algorithm is evaluated under various degrees of clutter and occlusions. In outdoor ex-

Table 5.1: Comparison with the scope of existing works.

Works	Abstract Appearance	Spatial Relationship	Dynamic Environment
[44]	✓	✓	×
[45]	×	×	×
[46]	✓	×	✓
[48]	✓	×	✓
[50]	✓	✓	×
Proposed	✓	✓	✓

periments, the algorithm is tested with different sizes of critical 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 5.1. 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 for indoor experiments was an ASUS Xtion Pro Live RGB-D camera. But because the RGB-D camera is sensitive to the light, thus the depth estimation is inaccurate in outdoor. Therefore, a monocular camera with higher resolution is used instead in outdoor experiments. Then, the relative distance is calculated based on the pixel distance rather than the 3D distance, but the principle is the same.

5.3.1 Effect of False Detection on Reasoning Performance

Since the reasoning process is based on perception results, the fault detections of objects and humans may affect reasoning performance to varying degrees. Therefore, how false detection affects reasoning results is a critical part that we need to investigate. As shown in Figure 5.4, six different types of false detection of both human and object may result

Figure 5.4: Six types of false detection and the corresponding impact on uniqueness reasoning. The green label is recognized as a unique individual. +1 and -1 are counts of corresponding reasoning results according to different false detection.

in the false judgment of the unique individual. 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 a particular 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 judgment of the unique individual. *Precision* and *Recall* are the primary metrics for judgment, which is defined as below:

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

The quantitative impact of false detection on uniqueness reasoning performance is demonstrated in Figure 5.5. This figure displays the changing trend of uniqueness reasoning

Figure 5.5: The changing trend of uniqueness reasoning *Precision* and *Recall* according to different types and degrees of false detection.

according to different degrees and types of false detection. For example, the top left figure in Figure 5.5 shows that fn_1^h and fp^h do not affect the judgment *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, region-based detection methods have the inherent defects that they are difficult to describe the boundaries of objects accurately. The inaccurate boundaries may affect the 3D position measurements or the intersection area estimation between humans and the critical object, which could lead to false interaction/relationship identification. This would affect the reasoning performance to some extent when encountering occlusions, illusions, and changing view perspectives. However, this problem can be improved if there is a detection method that can provide more accurate boundary information. The focus of this study is to develop a comprehensive probabilistic reasoning framework that is independent of the detection algorithms. Hence, the selection of state-of-art detection approaches is flexible in this framework.

5.3.2 Indoor Experiments

For indoor experiments, we use the same datasets as in Chapter 4 but applied the new activity extraction and uniqueness judgment algorithms to show the superiority of the improved approach. As shown in Figure 5.6, to comprehensively evaluate the proposed approach concerning human actions and the object existence. The unique individual here is the person who moves frequently and carrying a critical object. Due to limitations, three different water bottles are chosen as the critical object in this experiment, which is different in type, size, and color.

5.3.2.1 Indoor Experimental Description

In this experiment, the input features of reasoning algorithms were extracted from three different data sources (i.e., RGB images, depth images, and RGB-D images), and four main reasoning methods were evaluated. First, by analyzing RGB images, the judgment is made by measuring the 2D distance (d_{2D}) between the detected bounding boxes [44, 50]. Second, by processing depth images, people and the critical object were detected using template matching [65], and the judgment is made by measuring the 3D distance (d_{3D}) between them. For these reasoning methods from distance-based measurement, the person is identified as the unique individual only if the object is detected to be intersected with him. Another primary reasoning method is through spatiotemporal analysis (STA) [46, 48, 66], where the judgment is made by measuring the spatial changes between successive timestamps, and the person with distinctive moving speed is identified as the unique individual. The last one is knowledge-based inference (KBI) from analyzing RGB-D images. The judgment is made through a Bayes filter based on the prior knowledge [61]. The proposed approach, multimodal information fusion (MIF), also works on RGB-D images and decides by jointly considering OEM and HAM.

(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.

Figure 5.6: The three datasets are the same as in Chapter 4. But we apply the new activities extraction and uniqueness judgment approach to it. Yellow box, blue box, and green box represent some judgment results from HAM, OEM, and HAM+OEM, respectively.

Table 5.2: Feature detection accuracy (in %) on different labels 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 5.3: Evaluation of uniqueness reasoning performance in three separate datasets. All the values in the table are calculated from F-measure (in %).

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
<i>STA</i>	RGB Image	44.10	78.11	50.70	79.37	62.12	81.32
<i>KBI</i>	RGBD Image	67.40	71.31	64.70	74.66	70.31	67.82
<i>MIF</i>	RGBD Image	93.32	94.96	83.22	90.44	85.41	90.38

5.3.2.2 Indoor Experimental Results

Detection results on different modalities for each dataset are shown in Table 5.2. Detecting humans on color images was robust, but it was challenging to identify the small object. However, it was tough to discover the unique individual 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.

Results of each examined reasoning method were obtained by taking the F-measure, and numerical values are shown in Table 5.3. The average performance of the overall judgment result was revealed in Figure 5.7. 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 reasoning performance. Comparing with other methods in each dataset, it produced the highest accuracy among all methods for recognizing both the unique and usual person. Even in some complex scenes where there were many occlusions and illusions, the performance remained stable. For judgments based on analyzing 2D distance from RGB images (d_{2D}),

Figure 5.7: Overall uniqueness reasoning result of each method by averaging the result of each dataset. The proposed *MIF* achieved the highest accuracy for recognizing both the unique and usual person.

it can only recognize the unique individual when the critical object is successfully detected, and the performance can be affected by illusions and variant viewing angles. For judgments based on analyzing depth images (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 unique individual because of failing to detect the bottle. For these reasoning methods using distance measurements, they tended to recognize all detected human as others due to the limitation of detection. Thus, they achieved relatively high precision although their recall was low. For judgments based on the spatiotemporal analysis (*STA*), they were able to pick out the unique individual 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 bias of each reasoning method is shown in Figure 5.8. It revealed that the proposed method *MIF* had the most balanced reasoning ability for recognizing both the unique and usual person. d_{2D} and d_{3D} tended to recognize the people as others since they were hard to detect the critical object.

Figure 5.8: Bias of each reasoning method. Blue and orange represent the bias that recognizes the unique individual and others, respectively. The proposed approach *MIF* has the most balanced reasoning ability.

5.3.3 Outdoor Experiments

In outdoor experiments, another four scenarios were assessed with respect to different potential applications to show the generality of the proposed algorithm. We do not need to train any specific relationship for each scenario, but just need to change the detectors we need, which shows the flexibility and adaptability of the proposed approach. As shown in Figure 5.9, these four scenarios were collected from public places to recognize people interacting with critical objects or having a distinctive motion in the scene. For OEM, it aims to identify the one who is holding or very close to a critical object. For HAM, it is supposed to recognize the one that is moving at a distinctive motion pattern compared to the other people in the scene.

5.3.3.1 Outdoor Experimental Description

Four different scenarios were collected from the public streets, college campus and parking lots, respectively. In scenario 1, we show the proposed algorithm has the adaptability

(a) People falling in public.

(b) People drinking in public.

(c) People holding the bicycle.

(d) People near the car.

Figure 5.9: Evaluation of four security-related scenarios to show the generality of the proposed algorithm, where we do not train the specific relationship but just change the detectors. We take different types and sizes of critical objects into account. Red color denote detected semantic labels, such as person, the bottle in (b), the bicycle in (c) and the car in (d). Green color denotes judged unique individuals.

to applications that no critical objects involved. The other three scenarios take different critical objects into consideration, a bottle is used to recognize people drinking in public in scenario 2, a bicycle is used to recognize people carrying (stealing) the bicycle in scenario 3; a car is used to recognize people behind (damaging) the car in scenario 4. The experiment first evaluates two observation models separately and then combines them for a more comprehensive evaluation.

5.3.3.2 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 the judgment was relatively higher. The results of human and object detection are shown in Table 5.4.

Table 5.4: Feature detection accuracy (in %) in each scenario.

Input Data	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	Object	Person	Object	Person	Object	Person	Object	Person
RGB Image	-	92.11	24.75	99.01	93.59	99.90	99.99	98.97

Table 5.5: Evaluation of uniqueness judgment in four different scenarios (in %). Prec. = Precision, Rec. = Recall.

Decision	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Time ($10^{-4}s$)
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	
OEM	-	-	74.63	27.03	87.22	88.90	97.05	91.86	2.00
HAM	86.92	78.14	63.10	53.00	78.96	76.12	27.40	14.91	6.00
OEM+HAM	86.92	78.14	92.17	81.82	88.85	89.14	98.21	99.79	6.80

Detecting larger objects like bicycles and cars is very accurate. However, because of the possible occlusions and various poses, identifying small objects such as bottles is still challenging.

As shown in Table 5.5, the result validated that the proposed method could tolerate single model failure and fuse them into a comprehensive reasoning scheme to achieve higher accuracy. Decision time refers to the time from the analysis of the observation to the completion of the judgment. It showed that the reasoning 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 in Table 5.4, the contribution rate of each observation model is displayed in Figure 5.10. It illustrated that the proposed method *MIF* (OEM+HAM) relied more on HAM when the size of the critical 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 critical object became larger, so the proposed method *MIF* would dependent more on OEM.

Figure 5.10: 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.

5.4 Conclusions

In this chapter, a probabilistic reasoning approach was proposed that enables the intelligent system to recognize the unique individual in the scene. The idea behind the approach is to associate spatial relationships between humans and the critical object with the temporal changes concerning the positions of human in dynamic environments. Semantic-interaction and spatiotemporal features were used to form corresponding observation models, namely, OEM and HAM. Then, a hierarchical probabilistic model was established based on the mutual 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 reasoning decision. We also evaluated the effects of false detection on uniqueness reasoning 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 individual in the scene. Moreover, it achieved a timely and robust performance even in some complex scenes.

Chapter 6

Analytics-Based Multi-Robot Uniqueness Reasoning

We explained in detail in the previous chapters how to realize the uniqueness reasoning from a single robot perspective. However, many of the security applications also require robots to work in an unstructured and large-scale environment as well as a long-term operation, like search-and-rescue operation or mobile patrol. Moreover, single robot operation is still tricky to practically deal with problems such as occlusion, sensor failure, limited field of view, and varying illumination. Therefore, this chapter aims to apply the developed algorithm to robot platforms and share information in the multi-robot system to improve the accuracy and robustness of the reasoning process.

6.1 Introduction

The collaborative operation of the multi-robot system has garnered significant attention [69], which has been gradually employed to cope with tough missions, such as visual tracking [70], collaborative 3D mapping [71], intelligent reasoning [72], and rapid response [73]. However, the systematic framework for multi-robot reasoning is not studied

in depth, especially using mobile robots in dynamic, large-scale, and unstructured environments.

But by introducing a multi-robot system, several new problems will also be introduced. The first new problem is the need to estimate the relative pose between the robots. Only when the relative pose between the robots is obtained can the robots share their dynamic observations and information. The second new problem is to track people from multiple views. Only by achieving that, the robot can generate a global dynamic observation to compensate the local perception failure. In this chapter, since robots are operating in large-scale and unstructured environments, we focus more on the activities related to spatiotemporal feature (human motion) rather than the semantic-interaction features.

For the overall reasoning process, the system first applies heterogeneous sensor fusion to perceive dynamic people and separate them from the static environment (please refer to Section 3.1.2). Then, the collaborative mapping is performed to generate the relative transformation between robots. Next, each robot performs uniqueness reasoning based on its local dynamic observations. Meanwhile, each robot also shares its local dynamic observation directly with the system to generate a global dynamic observation. After that, the system extracts the human activity from the global observation and re-judge uniqueness based on it. Finally, the judgment from the system will be shared back to each robot for their reference to take the appropriate action. In the experiment, three cases were investigated, namely Daytime Rainforest, Night Forest, and Open Carpark. Each scenario involves varying degrees of dynamic human movements. The results show the accuracy, robustness, and versatility of uniqueness reasoning using a multi-robot platform.

6.2 Main Approach

This section develops a flexible distributed multi-robot reasoning strategy for a multi-robot system to identify unique individuals in the environment. The overall schematic diagram is shown in Figure 6.1, and the main approach contains three processes, which

Figure 6.1: The overall schematic diagram of analytics-based multi-robot uniqueness reasoning. The blue part is the local uniqueness reasoning on each robot, and the brown part is the global uniqueness reasoning based on the shared global observation.

are listed below.

- **Robot Relative Pose Estimation.** Robots should first estimate their relative pose and obtain the real-time transformation matrix before the robots can share their dynamic observations and information. Since these are not the main contributions of this work, we briefly introduce them in Section 6.2.1.
- **Local Uniqueness Reasoning.** Each robot perceives the environment from their onboard sensors, extracts the human motion and performs the uniqueness reasoning based on their local dynamic observation. It will be discussed in Section 6.2.2.
- **Global Uniqueness Reasoning.** It perform a two-stage information sharing, namely *forward information sharing* and *backward information fusion*. For forward information sharing, robots share their local dynamic observations directly with their neighborhood to generate a more precise global dynamic observation. Then the system re-judges uniqueness of each person based on this global dynamic observation. For backward information sharing, the judgment from the system will be

shared back to each robot for their reference to take appropriate actions. It will be discussed in Section 6.2.3.

6.2.1 Robot Relative Pose Estimation

In order to perform multi-robot reasoning using multiple robots, it is first necessary to obtain the real-time transformation between the robots before the robots can share their dynamic observations and information with each other. The transformation between neighborhood robots includes the translation and orientation relationship. We use collaborative mapping and map fusion to obtain the transformation matrix. The collaborative mapping problem has been addressed in different aspects, such as estimating the relative transformation between the robots [74], accounting for the uncertainty of the relative localization [75], integrating the information from neighboring robots [11], managing the communication [76], updating maps in the global perspective [77]. However, the methods mentioned above assume that the environment is entirely static, and dynamic objects such as people are ignored. Besides, most of them directly apply dense sensor data registration to estimate relative transformation without providing comprehensive analysis and modeling of the integrated map fusion framework. In this section, we take the dynamic people into account and filter them out to generate more precise maps for a more accurate transformation matrix.

6.2.1.1 Single Robot Mapping

A single SLAM problem is to build a map m_t^r and estimate the robot location x_t^r at each time step according to the dynamic observation g_t^r and control commands u_t^r (see (6.1)). The single robot SLAM problem has been well studied, and many efficient approaches have been developed. Any SLAM method can be applied here for building the foundation of single robot mapping and reasoning. Figure 6.2 shows the mapping results using different 3D points, and the red box highlights their differences. It can be seen that by filtering

Figure 6.2: The difference between the use of raw laser points and the use of filtered laser points by a single robot SLAM. The moving person causes an inconsistent "wall" in (b).

out the moving people, the accuracy and consistency of the map have been improved.

$$p(m_t^r, x_{1:t}^r | g_{1:t}^r, u_{1:t}^r) = \underbrace{p(x_{1:t}^r | g_{1:t}^r, u_{1:t}^r)}_{\text{Localization}} \cdot \underbrace{p(m_t^r | x_{1:t}^r, g_{1:t}^r)}_{\text{Mapping}} \quad (6.1)$$

6.2.1.2 Probabilistic Map Matching

For robot r , it receives partial maps sequentially, which means the relative transformation is a serial processing upon receiving of new partial maps. Let T^{r,r_n} defines the relative pose between individual robots r and r_n . R is a set of neighboring robots in communication with robot r . $T^{r,R}$ is a set of relative pose between robot r and all neighboring robots R . Since the relative transformation T^{r,r_n} is also pairwise independent of each pairs of robots, the posterior of relative transformation $p(T^{r,R} | m_t^r, m_t^R)$ can be factorized into a product of the posterior of relative transformation between each pair, as shown in (6.2):

$$p(T^{r,R} | m_t^r, m_t^R) = \prod_{r_n \in R} p(T^{r,r_n} | m_t^r, m_t^{r_n}) \quad (6.2)$$

where $T^{r,r_n} \in SE(3)$ is a rigid transformation with rotation $R^{r,r_n} \in SO(3)$ and translation $T^{r,r_n} \in R^3$. Instead of simply considering T^{r,r_n} as a fixed unknown parameter, we model it as a 6D random variable. Then, T^{r,r_n} can be modeled as a multivariate Gaussian distribution and solved by MAP estimation, where $T^{r,r_n} \sim (\hat{T}_{MAP}, \hat{T}_{\Sigma_{MAP}})$.

Since robot relative pose estimation is just a prerequisite for our work, not the main contribution, thus for the detailed explanation, please refer to [60].

6.2.2 Local Uniqueness Reasoning

Local uniqueness reasoning is first performed on each single robot. The principle is similar to Chapter 5, but since we have introduced multiple robots and changes to different environments, there are several points needs to be clarified.

- The dynamic observation of human g_t^r is in the world reference frame, which is obtained in a robot map by jointly considering the pose of the robot coordinate x_t^r (obtained from SLAM [78]) and the local coordinate of the detected person b_t^r .
- Since robots are operating in large-scale environments, we focus more on the activities related to spatiotemporal feature over the semantic-interaction features.
- Since heterogeneous sensor fusion is applied, we are able to calculate the absolute human moving speed and direction rather than the relative velocity in the previous chapters.

6.2.2.1 Activity Extraction

For local uniqueness reasoning, we take human motion A as the spatiotemporal feature Θ^{ST} . Then, by applying the law of total probability and chain rule, the local uniqueness reasoning is decomposed as follows:

$$p_j(I_t^r | A_t^r, g_{t-1:t}^r) = \underbrace{p_j(I_t^r | A_t^r)}_{\text{Uniqueness Judgment}} \cdot \underbrace{p_j(A_t^r | g_{t-1:t}^r)}_{\text{Motion Extraction}} \cdot \underbrace{p_j(g_{t-1:t}^r | b_{t-1:t}^r, x_{t-1:t}^r)}_{\text{Dynamic Observation}} \quad (6.3)$$

The human dynamic observation g^r is obtained by adding the dynamic perception local coordinate b^r to the corresponding robot coordinates x^r . An illustration is shown in Figure

Figure 6.3: A schematic diagram for human dynamic observation and motion analysis. The human motion is extracted from the spatiotemporal analysis.

6.3. The motion of a person considered here is mainly his moving speed and direction. Two motion models are established for the corresponding moving speed and direction of people.

Let $A_{s,t}$ defines the moving speed of the person j in a single robot map at time t . ΔT is the time interval, the speed can be extracted as:

$$A_{s,t} = \left\| \frac{g_t - g_{t-1}}{\Delta T} \right\| \quad (6.4)$$

Let $A_{d,t}$ defines the moving direction of the person j in a single robot map at time t . ΔT is the time interval, the direction can be extracted as:

$$A_{d,t} = \cos\left(\frac{\frac{g_t - g_{t-1}}{\Delta T}}{\left\| \frac{g_t - g_{t-1}}{\Delta T} \right\|}, \frac{b_t}{\|b_t\|}\right) \quad (6.5)$$

Then, for the following uniqueness judgment of each person, we follow the same logic as in Chapter 5, thus we skip this content in this chapter.

6.2.3 Global Uniqueness Reasoning

For multi-robot uniqueness reasoning, it either performs a *forward information sharing* and a *backward information sharing*.

- Forward Information Sharing is defined as information sharing from a single robot to the multi-robot system, which can solve the problem of occlusion and false detection. Multiple robots may observe different people or view the same person from various viewpoints. It is beneficial to integrate perceptual information or local dynamic observations from each robot to compensate or enhance each other. By incorporating local dynamic observations, each person's activity can be estimated more accurately. Besides, the uncertainty of reasoning can be reduced from such information sharing due to the uncertainty of detection and tracking.
- Backward Information Sharing is defined as information sharing from the multi-robot system to each robot, which solves the problem of the limited field of view and sensor failure. On the one hand, the robot who is not able or fails to perform local uniqueness reasoning can benefit from this information sharing. On the other hand, the robot can correct or enhance its judgment of the unique individual in the scene for allocating computing resources more targeted and efficient.

6.2.3.1 Forward Information Sharing

After obtaining the transformation matrix $T^{r,R}$ from Section 6.2.1, the dynamic observations is transformed into a common frame. But because the person's label is different in the local map of each robot, thus after dynamic observation transformation, the same person need to be identified from different labels. In this chapter, since we already have a relatively precise positions of each person, we applied KNN matching after the transformation to match the same person in the global frame. Since human tracking is beyond the scope of this chapter, we do not talk it into detail. Once the global dynamic observation is obtained, the multi-robot system can judge the uniqueness of each person by using the global dynamic observation g_t^R . The guidance mathematical formulation refers to (3.8).

Thus, the multi-robot system first updates the labels and position of each person concerning the global dynamic observation. The single robot dynamic observation g_t^r is transformed to coordinate frame of neighborhood robot R , which is denoted as g_t^R . Then,

the uniqueness of each person is re-judged using a similar approach as in single robot reasoning but feeding the global observation, i.e., global moving speed and direction.

$$\begin{aligned}
 p_j(I_t^R | A_t^R, g_{t-1:t}^R, g_{t-1:t}^r, T_{t-1:t}^{r,R}) \\
 \propto \underbrace{p_j(I_t^R | A_{s,t}^R, A_{d,t}^R)}_{\text{Uniqueness Re-Judgment}} \cdot \underbrace{p_j(A_{s,t}^R, A_{d,t}^R | g_{t-1:t}^R)}_{\text{Global Motion Extraction}} \cdot \underbrace{p_j(g_{t-1:t}^R | g_{t-1:t}^r, T_{t-1:t}^{r,R})}_{\text{Global Dynamic Observation}} \quad (6.6)
 \end{aligned}$$

Then, for the following global uniqueness judgment of each person, we follow the same logic as in Chapter 5, thus we skip this content in this chapter.

6.2.3.2 Backward Information Sharing

After the system makes a global judgment on the uniqueness of each person, it will share the result back to each robot for their reference. By jointly consider the local judgment $p(I_t^r)$ and the global judgment $p(I_t^R)$, each robot will obtain a final judgment $p(I_t^F)$ and be able to take appropriate actions based on it in the environment. For every single robot, there will be three most likely cases when obtaining this backward information sharing. For case 1, if the multi-robot system has a clear judgment (belief exceeds a certain threshold of σ), but the individual robot judges relatively vague or fails, this individual robot will adopt the judgment from the system, and stops to take appropriate actions to this unique individual. For case 2, if the individual robot has a clear judgment, but the multi-robot system judges relatively vague, this individual robot will rely on its own judgment, and stop to standby for potential interactions. For case 3, if neither the multi-robot system nor the individual robot can provide a clear judgment, this individual robot will maintain its original motion.

$$p(I_t^F | I_t^r, I_t^R) = \begin{cases} \max(p(I_t^r), p(I_t^R)), & \text{if } (I_t^R \geq \sigma \text{ and } I_t^r < \sigma) \text{ or } (I_t^r \geq \sigma \text{ and } I_t^R < \sigma) \\ p(I_t^r), & \text{otherwise} \end{cases} \quad (6.7)$$

Fuzzy laws are illustrated in (6.7), where $\sigma \in [0, 1]$ is set as the threshold that indicates the sensitivity of the algorithm. Generally, a higher value of σ indicates that the sensitivity of the algorithm is lower, which may lose some unique individual recognition (true-positive) but can reduce false alarm (false-positive).

6.3 Experiments

This experiment is motivated by the cases that mobile robot patrols are required to work in large-scale and unstructured environments to identify intruders. It aims to demonstrate the performance of the proposed simultaneous collaborative mapping and reasoning (SCMR), and both qualitative and quantitative results are presented. Qualitative results of collaborative mapping are shown in each scenario. The validation of the performance of the proposed SCMR by presenting extensive results in three challenging scenarios, including changing illumination, being highly dynamic, and also enabling day and night operation with high robustness and accuracy. Moreover, we demonstrate the superiority and robustness of multi-robot reasoning by sharing dynamic observations from single robots.

6.3.1 Experiments Overview

6.3.1.1 Experiment Setup

Experiments are conducted by operating two robots in three different scenarios in a university campus (see Table 6.1). All algorithms are executed on the ROS platform for multimodal environmental perception, collaborative static mapping, and collaborative dynamic reasoning. Two robot platforms (Husky) are equipped with the Intel Core i7-6700HQ CPU @ 2.60GHz CPU and the Nvidia GeForce GTX 1050 @ 4GB RAM GPU. Each robot is equipped with a 3D Velodyne LiDAR, thermal camera or visual camera. 3D LiDAR provides 16 laser beams @ 10Hz, visual camera produces RGB images with resolution: 672×376 @ 30Hz, and thermal camera produces thermal images with resolution:

Figure 6.4: Experimental environments and platforms.

$382 \times 288 @ 23\text{Hz}$. The communication between robots is established by long-range Wi-Fi with limited bandwidth.

Due to the limited computational resources onboard, all incoming data are running at maximum 6Hz. In all cases, YOLOv3 [79] is used for moving people detection. Then, the static and dynamic observations are separated. LOAM [80] is applied for single robot pose estimation, and Octomap [81] is utilized for 3D occupancy mapping at a resolution of 0.1m. The experiment environment settings are:

- **Daytime Rainforest:** two Husky robots equipped with Velodyne 3D LiDAR and visual camera in a dense rainforest under a messy background and changing illumination.
- **Night Forest:** one Husky robots equipped with visual camera and Velodyne 3D Li-

Figure 6.5: Overview of the final results of simultaneous collaborative mapping and reasoning in daytime rainforest. Left: multi-robot reasoning results in three scenarios. Middle: collaboratively generated 3D global map. Right: satellite image and real environment.

Table 6.1: Summarize of the experiment environment.

Dataset	Robot 1 Sensors	Robot 2 Sensors	GPS-Denied	Dynamic
Daytime Rainforest	Vision, 3D LiDAR	Vision, 3D LiDAR	✓	✓
Night Forest	Thermal, 3D LiDAR	Vision, 3D LiDAR	✓	✓
Open Carpark	Vision, 3D LiDAR	Vision, 3D LiDAR	✓	✓

DAR, while the other one equipped with thermal camera and Velodyne 3D LiDAR in a forest at night.

- **Open Carpark:** two Husky robots equipped with Velodyne 3D LiDAR and visual camera in an open carpark, which is a highly dynamic environment.

Figure 6.4 and Table 6.1 summarize the settings and the main characteristics of each experiment.

6.3.1.2 Comparison Baseline

Most of the existing works either focus on map fusion in a structured environment or reasoning with a fixed camera. In this experiment, we compared the core contributions

Table 6.2: Comparison of core functionality with existing approaches.

Algorithms	SCMR	Multimodal Perception	Collaborative Mapping	Collaborative Reasoning	Distributed Framework
[11]	×	×	✓	×	×
[82]	×	×	✓	×	×
[83]	×	✓	×	×	✓
[50]	×	✓	×	×	×
Ours	✓	✓	✓	✓	✓

of our method with them, as summarized in Table 6.2. It can be seen that they are unable to achieve multi-robot reasoning, and it is more challenging to perform simultaneous collaborative mapping and reasoning. [11] achieves map fusion in a 2D indoor environment, and [82] carries out 3D map fusion by directly apply ICP in an off-line setting. [83] provides a distributed dynamic perception framework, however, did not consider collaborative actions. [50] detects multimodal information to recognize the action of the people in images. However, it is not feasible to be performed with a robotic system and therefore can not locate the person in the environment.

As no article has addressed the SCMR problem, we demonstrate extensive qualitative results of the proposed algorithm. As no work is available for multi-robot reasoning of human uniqueness by using a multi-robot system, we compare the results of multi-robot uniqueness reasoning with the results of a single robot uniqueness reasoning.

6.3.2 Field Experiments

Extensive unstructured scenarios are investigated, including Daytime Rainforest (Section 6.3.2.1), Night Forest (Section 6.3.2.2) and Open Carpark (Section 6.3.2.3).

6.3.2.1 Daytime Rainforest

As shown in Figure 6.5, the robots operate under a dense forest canopy, which is a fully dynamic and unstructured 3D environment that contains trees, slopes and moving people.

The mission is to collaboratively map out the full 3D environment and simultaneously estimate the uniqueness of people. It can be used for missions that require rapid deployment to an unknown forest environment. The two robots started moving from a nearby place, while robot **1** (red trajectory) turned left and went uphill and robot **2** (orange trajectory) turned right and went downhill. People moved into the environment in different directions and speeds.

As shown in Figure 6.5, the overall results of collaborative mapping and reasoning are provided. It can be found from the satellite image that it is a dense rainforest, and it is difficult to see the ground from the top view. Real environmental images also verify the complexity of the environment from a human perspective. The experiment is carried out immediately after the rainstorm, and the robot is operated in muddy terrain. The middle image shows the final merged global map at the end of the operation. The global map reconstructs the entire unstructured environment in detail with high precision. It is worth noting that while dynamic objects are moving around in the environment, we are still able to build an accurate 3D static map based on multimodal context-aware frameworks. The foundation behind the global map is to accurately estimate the relative pose between the robots, which is also the basis for multi-robot reasoning. The left three images show the multi-robot reasoning results on the global map. In the image, the dots and arrows show the 3D position and motion of each person. The red dot denotes the recognized unique person from comprehensive human motion analysis.

Multimodal Perception Results

We first present the performance of multimodal environmental perception. In this challenging environment, moving people will cause additional difficulty for robots to perceive their surroundings. Alternatively, multi-robot reasoning does not need to process the entire sensor dynamic observations. The purpose of this part is to separate the static point cloud and simultaneously detect and track moving people.

Firstly, the extrinsic parameters of the rotation matrix and translation vector between 3D

Figure 6.6: The results of multimodal environmental perception in daytime rainforest.

LiDAR and visual camera are calculated. Then, YOLOv3 is applied to detect the moving people in the visual camera and to perform target tracking using a Kalman filtering based approach. It operates at low-resolution images at a 6 Hz calculation rate to ensure efficiency. Figure 6.6(a) shows the detected people with labeled bounding box. By projecting the image to the 3D LiDAR, we can visualize the 3D color point cloud of the people in Figure 6.6(b). Based on the results of Figure 6.6(a) and Figure 6.6(b), we can get filtered static point cloud in Figure 6.6(c), where each moving person is represented with a labeled 2D bounding box. The magnified image shows the existence of three detected people in the point cloud.

(a) Uniqueness reasoning when detection fails.

(b) Uniqueness reasoning under occlusion.

(c) Uniqueness reasoning concerning the motion of people in the scene.

Figure 6.7: The results of multi-robot reasoning in a Daytime Rainforest environment. Points and arrows indicate the 3D position and real-time motion of the detected human. Red color represents the inferred unique person.

Multi-Robot Reasoning Results

Then, we present the results of the collaborative uniqueness reasoning of the humans in the Daytime Rainforest, as shown in Figure 6.7. Due to the limitations of visual cameras

Figure 6.8: Overview of the final results of simultaneous collaborative mapping and reasoning in night forest. Left: multi-robot reasoning results in three scenarios. Middle: collaboratively generated 3D global map. Right: satellite image and real environment.

and 3D LiDAR, each sensor can only obtain one form of information, namely a human bounding box or an actual 3D distance. Therefore, each person's 3D position can be obtained by fusing data from two sensors. Also, multiple robots have different angles of view. By combining dynamic observations from all neighborhood robots, human motion can be analyzed globally and more accurately. As shown in Figure 6.7(a), robot 1 fails to detect the person, but robot 2 detects the person. Thus, multi-robot reasoning takes the dynamic observation and judgment result of robot 2. In Figure 6.7(b), due to the occlusion, the true unique person cannot be seen from robot 2, but robot 1 successfully observes all the persons. The multi-robot reasoning fuses the dynamic observation from two robots and generates the final result. In Figure 6.7(c), both robots judge the person with different movement behavior from other people to be a unique person. The multi-robot reasoning confirms the judgment results of each robot.

6.3.2.2 Night Forest

The forest in this experiment is more sparse than the environment of the previous experiment, and the primary emphasis is that this experiment is conducted on a night. As presented in Figure 6.8, comprehensive results concerning collaborative mapping and

Figure 6.9: The results of multimodal environmental perception in night forest.

reasoning has been provided. Here, we demonstrate the full-day operational capabilities of our system in this challenging environment. As summarized in Section 6.3.1.1, both robots are equipped with 3D LiDAR. In this setting, only robot 1 is equipped with the thermal camera, and robot 2 carries a normal visual camera. The algorithm produces a highly accurate map and performs uniqueness detection at night. The collaboration framework enables robot 2 to maintain the same information as robot 1, and therefore also gains nighttime awareness ability. This capability is critical in the real world because the robot team will carry heterogeneous sensors for different tasks.

Multimodal Perception Results

It can be directly observed that the visual camera does not see anything on a night. Different from the RGB image produced by the visual camera, the thermal camera detects

radiation in the long-infrared range of the electromagnetic spectrum and produces images of that radiation, called thermal image. In thermal image, it records only temperature information for each pixel instead of color information. A special color palette is defined where different colors are mapped to different temperature values. In this case, the brighter color means the higher temperature, such as the three persons detected. The calibration method between a thermal camera and 3D LiDAR has been explained in Section 3.1.2. The thermal images then input into YOLOv3 for human detection. Figure 6.9(a) shows detected people with labels. Figure 6.9(b) is acquired by projecting the thermal image to 3D laser points. The different colors in the 3D point cloud represent different temperatures, so we call it the "temperature point cloud." Finally, Figure 6.9(c) presents filtered static point cloud and human states.

Multi-Robot Reasoning Results

We also show the results of uniqueness reasoning in the night forest environment. As shown in Figure 6.10, the robot 1 equips a thermal camera, while the robot 2 equips with a visual camera. Humans can only be detected from thermal images, and the 3D position of each person is obtained by projecting 3D point clouds onto the thermal frame. Figure 6.10(a) and Figure 6.10(b) show the results of the corresponding reasoning at different directions and speeds of movement. People that move in a different pattern from others is recognized to be the unique person. Since human detection is not available in the visual image, and thus, multi-robot reasoning relies on the result of robot 1.

6.3.2.3 Large Scale Open Carpark

The open car park depicted in Figure 6.11 is different from the first two purely unstructured forest environment experiments. It is a structured and unstructured hybrid environment. As presented in Figure 6.11, robots started from a parking lot, passed through the underpass of the office building, and then moved to the road between the building and

(a) The uniqueness reasoning under different moving direction.

(b) The uniqueness reasoning under different moving speed.

Figure 6.10: The results of multi-robot reasoning in the Night forest environment. Points and arrows indicate the 3D position and real-time motion of the detected human. Red color represents the inferred unique person.

the unstructured slope. The unique person is also denoted in red at the left part of Figure 6.11. Here we show the capability of the system to perform tasks in a complex urban environment without GPS.

Multimodal Perception Results

In this experiment, both robots are equipped with a visual camera and 3D LiDAR. The difference is that in this case, we have more walking people. Besides, the surrounding is a mixed environment. As can be seen in Figure 6.12(a), the left part is an office building,

Figure 6.11: Overview of the final results of simultaneous collaborative mapping and reasoning in open carpark. Left: multi-robot reasoning results in three scenarios. Middle: collaboratively generated 3D global map. Right: satellite image and real environment.

Figure 6.12: The results of multimodal environmental perception in open carpark.

(a) Uniqueness reasoning under different angles of view.

(b) Uniqueness reasoning under occlusion.

(c) Uniqueness reasoning concerning the motion of people under varying perspectives.

Figure 6.13: The results of multi-robot reasoning in an Open Carpark. Points and arrows indicate the 3D position and real-time motion of the detected human. Red color represents the inferred unique person.

while the right hand is the unstructured environment with trees. Figure 6.12 shows five detected walking people and the filtered static point cloud. This will serve as input for simultaneous collaborative mapping and reasoning.

Table 6.3: Human uniqueness reasoning results from the single robot and multi-robot system. (The best performance is denoted in bold.)

Experiments	Robot 1 Reasoning		Robot 2 Reasoning		Multi-Robot Reasoning	
	Precision	Recall	Precision	Recall	Precision	Recall
Daytime Rainforest	85.04	82.76	66.97	53.28	86.24	88.85
Night Forest	89.47	88.54	-	-	89.47	88.54
Open Carpark	59.40	54.86	57.65	52.64	64.88	61.57

Multi-Robot Reasoning Results

There are more dynamic moving objects in this experiment, and people's movements are more chaotic than the forest environment, making robot recognition more difficult. As shown in Figure 6.13, many people move in the environment with various patterns. Furthermore, due to the large scale of the environment, each robot may observe only a part of the environment that may result in entirely different judgment results. In Figure 6.13(a), since the field of view is limited, robot 1 cannot recognize the unique person, but robot 2 successfully recognizes the unique person. After the fusion, the robotic system can finally identify unique people. In Figure 6.13(b), occlusion makes it difficult to see everyone in the scene, but the robot judged the uniqueness of people from the current dynamic observation. Therefore, the robot 1 incorrectly judges the unique person because some persons are occluded. After sharing the information, the multi-robot system makes the right judgment. In Figure 6.13(c), each robot observes a part of the scene and infers the uniqueness of people in its field of view. However, after integrating the dynamic observations of all the robots, a truly unique person is discovered.

6.3.3 Quantitative Analysis

As stated before, there is no relevant work to directly address multi-robot reasoning using mobile robots, let alone SCMR. To show the quantitative results in respect of the accuracy and efficiency of SCMR, we compared intermediate results of our algorithm.

6.3.3.1 Result of Uniqueness Reasoning

Results of collaborative uniqueness reasoning are demonstrated in Table 6.3. It shows the overall precision and recall of each experiment. The experiments in the forest environment (Daytime Rainforest and Night Forest) involved fewer people and with less complexity of movement behavior. Thus, the overall accuracy is relatively higher in forest environments than carpark environment. In the Daytime Rainforest experiment, most people move in the field of view of the robot 1, and the robot 2 can not see many people. Therefore, the robot 2 cannot accurately recognize the unique person in the environment. Meanwhile, some people only appear in the field of view of the robot 2 and the robot 1 cannot see it. However, under the multi-robot reasoning, the two robots can exchange information with each other, so the accuracy of the reasoning is improved. In the Night Forest experiment, robot 1 equips with a thermal camera while robot 2 equips with a visual camera. Thus, humans can only be observed from thermal images. The multi-robot reasoning can compensate for the limitation of visual blindness and takes advantage of the thermal information. Therefore, although Robot 2 (with a visual camera) cannot contribute to reasoning, Robot 1 (with thermal camera) can share its dynamic observations and judgment results to Robot 2 to help it understand the environment and provides information for its future path planning. This is also seen as a kind of collaboration. In Open Carpark experiment, many humans are moving in the environment with messy moving behavior. Thus, each robot can infer the unique person in their own field of view, but it is tough to discover a unique person in the global view. The result shows that multi-robot reasoning can somehow link the significantly distinctive person from both robots' view and improve the overall recognition accuracy.

6.4 Conclusions

This chapter presents a method for simultaneous collaborative mapping and reasoning (SCMR) in GPS-denied, dynamic, and unstructured environments. We have designed a

new framework to provide theoretical formulas and system implementations for collaborative static mapping and dynamic reasoning, showing accurate and robust performance in day and night operations. Each robot in the group first executes the separation of static and dynamic observations based on heterogeneous sensor fusion and human detection. Explicitly environmental mapping is addressed by adopting map fusion. It estimates the relative pose between robots and generates an accurate transformation matrix between them. Once the transformation is obtained, the robots can share their dynamic observations and local judgment results with each other to perform global uniqueness reasoning of people. The method is evaluated during day and night in a real-world unstructured environment, which exhibits improved precision, efficiency and robustness compared to existing methods. In summary, the proposed collaboration system provides a new perspective for sensing and adapting to dynamic and unstructured environments, which compensates for the limitations of individual robot perception, mapping, and reasoning.

Chapter 7

Conclusions and Recommendations

This chapter concludes this thesis by summarizing the main contributions and suggesting directions for future research.

7.1 Conclusions

In this thesis, we have extensively studied the uniqueness reasoning problem at three levels under different assumptions and conditions in the context of security or surveillance scenarios. It enabled one or two robots to identify the unique individual whose motion pattern and an object associated with him, in space and time, appear different from the other people in the scene.

Firstly, at each level of reasoning, different sensors were used to perceive the environment. In Chapter 4, an ASUS Xtion PRO LIVE camera was used. It has point clouds registered with RGB information, which can produce a dense 3D representation of the environment. However, it can be easily affected by changing lighting condition, the FoV is narrow, and the detection range is short. Thus, in Chapter 5, we added a high-resolution camera to obtain a larger FoV and more extended detection range. In Chapter 6, we calibrated a camera with 3D LiDAR (VLP-16) and also added a thermal camera (Optris PI 450) into

the robot system to handle the environmental perceptional at dark conditions. It enables the robot system to perform long-term operations, including day-and-night, bright-and-dark switching. Moreover, the distance measurement is more accurate from LiDAR than RGBD camera.

Secondly, at each level of reasoning, different approaches to activity extraction were developed to extract the spatiotemporal and semantic-interaction features from observations. In Chapter 4, the activities were extracted by applying qualitative rules, such as relative position, whether to overlap, whether to move and so forth. In Chapter 5, instead of qualitative rules, quantitative analysis is applied to extract the features, such as actual distance, the overlapping area over the union, relative velocity, and so forth. Furthermore, in Chapter 6, dynamic observations are shared in the system. The activity is extracted in both local frame and global frame based on the corresponding local and global observation.

Thirdly, at each level of reasoning, different uniqueness judgment methods that linked extracted activities with the uniqueness were developed. In Chapter 4, we associated the activities with prior knowledge, which was provided by human expertise. In Chapter 5, we used neighborhood comparison to analyze the activity of an individual within the distribution to determine if it is unique. In Chapter 6, the final uniqueness is judged by jointly considered the local judgment result and the global judgment result.

Finally, at each level of reasoning, different experimental conditions were tested with increasing difficulty. In Chapter 4, the experiment was done in indoor environments with different degrees of occlusion. In Chapter 5, the experiments were conducted in both indoor and urban outdoor environments and with more scenarios involved. In Chapter 6, the tests were carried out in unstructured outdoor environments with more dynamic targets and even night conditions.

7.2 Recommendations for Further Research

Since we have successfully found unique individuals in the scene, we can proceed to the next step from two perspectives. On the one hand, we can focus on the perceptual level, which predicts the future movements and intentions of unique individuals. On the other hand, we can try to design an appropriate robot motion strategy to achieve dynamic interaction.

For target intention, it is a primary step for the robot to perform situation/risk assessment in the environment, and thereby achieve a more comprehensive environmental awareness. Firstly, the robot should be able to keep tracking of the target and analyze its past positions, velocities, and trajectory and project them into its future states. Once the robot obtains possible future states of the target, the robot can accordingly associate them with its own future states to assess the situation/risk.

For dynamic interactions, it is a critical step for the robot to achieve higher intelligence and seamlessly get into our daily life. The robot should be able to make decisions for its motion planning based on the assessed situation. Possible objectives of decision-making can be in motion, such as acceleration, steering, or it can also be in behavior, such as following, Intercepting, alarming, and so forth.

Author's Publications

Journal Papers:

1. **Chule Yang**, Danwei Wang, Yijie Zeng, Yufeng Yue, P. Siritanawan, "Knowledge-Based Multimodal Information Fusion for Role Recognition and Situation Assessment by Using Mobile Robot," *Information Fusion*, vol.50, pp.126-138, 2019. (IF:10.716)
2. **Chule Yang**, Yufeng Yue, Jun Zhang, Mingxing Wen, and Danwei Wang, "Probabilistic Reasoning for Unique Role Recognition Based on the Fusion of Semantic-Interaction and Spatio-Temporal Features," *IEEE Transactions on Multimedia*, vol.21, no.5, pp.1195-1208, May 2019. (IF:5.452)
3. Yufeng Yue, P.G.C.N. Senarathne, **Chule Yang***, Jun Zhang, Mingxing Wen, and Danwei Wang, "Hierarchical Probabilistic Fusion Framework for Matching and Merging of 3D Occupancy Maps," *IEEE Sensors Journal*, vol.18, no.21, pp.8933-8949, Nov 2018. (Corresponding Author, IF:3.076)
4. Yufeng Yue, **Chule Yang**, Yuanzhe Wang, Jun Zhang, Mingxing Wen, and Danwei Wang, "A General Probabilistic Framework for Multi-Robot 3D Mapping Using Heterogeneous Sensors," *IEEE Systems Journal*, 2019. (IF: 4.463)
5. Yufeng Yue, **Chule Yang***, Jun Zhang, Mingxing Wen, and Danwei Wang, "SCMR: Simultaneous Collaborative Mapping and Reasoning in GPS-Denied and Dynamic

Unstructured Environments,” *International Journal of Robotics Research*, 2019.
(Corresponding Author, Under Review, IF: 6.134)

Conference Papers:

1. **Chule Yang**, Yufeng Yue, Jun Zhang, Mingxing Wen, and Danwei Wang, “Semantic-Spatial Reasoning for Unique Role Recognition in Dynamic,” *IEEE International Conference on Robotics and Automation (ICRA) workshop*, 2018.
2. Jun Zhang, Yufeng Yue, **Chule Yang**, Mingxing Wen, and Danwei Wang, “A Two-step Method for Extrinsic Calibration between a Sparse 3D Lidar and a Thermal Camera,” *IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2018.
3. Yufeng Yue, P.G.C.N. Senarathne, **Chule Yang**, Jun Zhang, Mingxing Wen and Danwei Wang, “Probabilistic Fusion Framework for Collaborative Robots 3D Mapping,” *21st International Conference on Information Fusion (ICIF)*, 2018.
4. **Chule Yang**, Yijie Zeng, Yufeng Yue, Jun Zhang and Danwei Wang, “Knowledge-Based Role Recognition by Using Human-Object Interaction and Spatio-Temporal Analysis,” *IEEE Conference on Robotics and Biomimetics (ROBIO)*, 2017. **Finalist of Best Student Paper Award**
5. Yufeng Yue, Danwei Wang, P.G.C.N. Senarathne and **Chule Yang**, “A General Framework for Distributed Multi-Robot 3-D Map Fusion,” *IEEE International Conference on Robotics and Automation (ICRA) workshop*, 2017.
6. Yufeng Yue, Danwei Wang, P.G.C.N. Senarathne and **Chule Yang**, “Robust Submap-Based Probabilistic Inconsistency Detection for Multi-Robot Mapping,” *European Conference on Mobile Robotics (ECMR)*, 2017.
7. **Chule Yang**, Danwei Wang, P. Siritanawan, “Organ-Based Facial Verification Using Thermal Camera,” *IEEE International Symposium on Multimedia (ISM)*, 2016.

Bibliography

- [1] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, “Vision-based state estimation for autonomous rotorcraft mavs in complex environments,” in *2013 IEEE International Conference on Robotics and Automation (ICRA)*, 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 *2015 International Conference on Unmanned Aircraft Systems (ICUAS)*, 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 *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 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] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémont, “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.
- [10] S. Vidas, P. Moghadam, and M. Bosse, “3d thermal mapping of building interiors using an rgb-d and thermal camera,” in *2013 IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 2311–2318.
- [11] S. Saeedi, L. Paull, M. Trentini, M. Seto, and H. Li, “Group mapping: A topological approach to map merging for multiple robots,” *IEEE Robotics Automation Magazine*, vol. 21, no. 2, pp. 60–72, June 2014.
- [12] G. Grisetti, C. Stachniss, and W. Burgard, “Improved Techniques for Grid Mapping With Rao-Blackwellized Particle Filters,” *IEEE Transactions on Robotics*, vol. 23, no. 1, pp. 34–46, Feb 2007.
- [13] X. Chen, H. Zhang, H. Lu, J. Xiao, Q. Qiu, and Y. Li, “Robust slam system based on monocular vision and lidar for robotic urban search and rescue,” in *2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*, 2017, pp. 41–47.
- [14] S. Hoshino, S. Ugajin, and T. Ishiwata, “Patrolling robot based on bayesian learning for multiple intruders,” in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 603–609.

- [15] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, “A survey of motion planning and control techniques for self-driving urban vehicles,” *IEEE Transactions on intelligent vehicles*, vol. 1, no. 1, pp. 33–55, 2016.
- [16] T. B. LEE, “The hype around driverless cars came crashing down in 2018,” <https://arstechnica.com/cars/2018/12/uber-tesla-and-waymo-all-struggled-with-self-driving-in-2018/>, 2018.
- [17] D. Mehta, G. Ferrer, and E. Olson, “Autonomous navigation in dynamic social environments using multi-policy decision making,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1190–1197.
- [18] A. Vemula, K. Muelling, and J. Oh, “Modeling cooperative navigation in dense human crowds,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1685–1692.
- [19] M. Pfeiffer, M. Schaeuble, J. Nieto, R. Siegwart, and C. Cadena, “From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1527–1533.
- [20] H. Shi, Z. Lin, S. Zhang, X. Li, and K.-S. Hwang, “An adaptive decision-making method with fuzzy bayesian reinforcement learning for robot soccer,” *Information Sciences*, vol. 436, pp. 268–281, 2018.
- [21] A. G. Cunningham, E. Galceran, R. M. Eustice, and E. Olson, “Mpdpm: Multipolicy decision-making in dynamic, uncertain environments for autonomous driving,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1670–1677.
- [22] W. Liu, S. Kim, S. Pendleton, and M. H. Ang, “Situation-aware decision making for autonomous driving on urban road using online pomdp,” in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 1126–1133.

- [23] W. Schwarting, J. Alonso-Mora, and D. Rus, “Planning and decision-making for autonomous vehicles,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, pp. 187–210, 2018.
- [24] C. Hubmann, M. Becker, D. Althoff, D. Lenz, and C. Stiller, “Decision making for autonomous driving considering interaction and uncertain prediction of surrounding vehicles,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1671–1678.
- [25] C. Dong, J. M. Dolan, and B. Litkouhi, “Intention estimation for ramp merging control in autonomous driving,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1584–1589.
- [26] C. Hubmann, J. Schulz, M. Becker, D. Althoff, and C. Stiller, “Automated driving in uncertain environments: Planning with interaction and uncertain maneuver prediction,” *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 1, pp. 5–17, March 2018.
- [27] F. Damerow, S. Klingelschmitt, and J. Eggert, “Spatio-temporal trajectory similarity and its application to predicting lack of interaction in traffic situations,” in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016, pp. 2512–2519.
- [28] V. Loia, G. D’Aniello, A. Gaeta, and F. Orciuoli, “Enforcing situation awareness with granular computing: a systematic overview and new perspectives,” *Granular Computing*, vol. 1, no. 2, pp. 127–143, 2016.
- [29] R. Chandra and R. P. Rocha, “Knowledge-based framework for human-robots collaborative context awareness in usar missions,” in *2016 International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 2016, pp. 335–340.
- [30] T. Kollar and N. Roy, “Utilizing object-object and object-scene context when planning to find things,” in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 2168–2173.

- [31] 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)*, 2010.
- [32] 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.
- [33] —, “Exploiting semantic knowledge for robot object recognition,” *Knowledge-Based Systems*, vol. 86, pp. 131–142, 2015.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] M. Jain, H. Jegou, and P. Bouthemy, “Better exploiting motion for better action recognition,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2555–2562.
- [38] 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.
- [39] —, “Action recognition by dense trajectories,” in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3169–3176.

- [40] 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.
- [41] 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.
- [42] O. A. I. Ramírez, H. Khambhaita, R. Chatila, M. Chetouani, and R. Alami, “Robots learning how and where to approach people,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 347–353.
- [43] C. Dondrup and M. Hanheide, “Qualitative constraints for human-aware robot navigation using velocity costmaps,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 586–592.
- [44] G. Wang, A. Gallagher, J. Luo, and D. Forsyth, “Seeing people in social context: Recognizing people and social relationships,” in *2010 European conference on computer vision*. Springer, 2010, pp. 169–182.
- [45] Z. Song, M. Wang, X.-s. Hua, and S. Yan, “Predicting occupation via human clothing and contexts,” in *2011 IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1084–1091.
- [46] T. Lan, L. Sigal, and G. Mori, “Social roles in hierarchical models for human activity recognition,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1354–1361.
- [47] 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.
- [48] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Chun Zhu, “Joint inference of groups, events and human roles in aerial videos,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4576–4584.

- [49] B. J. Biddle, “Recent developments in role theory,” *Annual review of sociology*, vol. 12, no. 1, pp. 67–92, 1986.
- [50] G. Gkioxari, R. Girshick, P. Dollr, and K. He, “Detecting and recognizing human-object interactions,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.
- [51] Y. Yue, D. Wang, P. Senarathne, and D. Moratuwage, “A hybrid probabilistic and point set registration approach for fusion of 3d occupancy grid maps,” in *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on.* IEEE, 2016, pp. 001 975–001 980.
- [52] Y. Yue, D. Wang, P. Senarathne, and C. Yang, “Robust submap-based probabilistic inconsistency detection for multi-robot mapping,” in *Mobile Robots (ECMR), 2017 European Conference on.* IEEE, 2017, pp. 1–6.
- [53] J. Alonso-Mora, R. Knepper, R. Siegwart, and D. Rus, “Local motion planning for collaborative multi-robot manipulation of deformable objects,” in *Robotics and Automation (ICRA), 2015 IEEE International Conference on.* IEEE, 2015, pp. 5495–5502.
- [54] Z. Wang and M. Schwager, “Kinematic multi-robot manipulation with no communication using force feedback,” in *Robotics and Automation (ICRA), 2016 IEEE International Conference on.* IEEE, 2016, pp. 427–432.
- [55] E. Price, G. Lawless, H. H. Bühlhoff, M. Black, and A. Ahmad, “Deep neural network-based cooperative visual tracking through multiple micro aerial vehicles,” *arXiv preprint arXiv:1802.01346*, 2018.
- [56] J. Zhang, P. Siritanawan, Y. Yue, C. Yang, M. Wen, and D. Wang, “A two-step method for extrinsic calibration between a sparse 3d lidar and a thermal camera,” in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2018, pp. 1039–1044.

- [57] A. Geiger, F. Moosmann, mer Car, and B. Schuster, "Automatic camera and range sensor calibration using a single shot," in *2012 International Conference on Robotics and Automation (ICRA)*, 2012.
- [58] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, 2000.
- [59] 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.
- [60] Y. Yue, C. Yang, Y. Wang, P. C. N. Senarathne, J. Zhang, M. Wen, and D. Wang, "A multilevel fusion system for multirobot 3-d mapping using heterogeneous sensors," *IEEE Systems Journal*, 2019.
- [61] C. Yang, Y. Zeng, Y. Yue, P. Siritanawan, Z. Jun, and D. Wang, "Knowledge-based role recognition by using human-object interaction and spatio-temporal analysis," in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2017.
- [62] C. Yang, D. Wang, Y. Zeng, Y. Yue, and P. Siritanawan, "Knowledge-based multimodal information fusion for role recognition and situation assessment by using mobile robot," *Information Fusion*, vol. 50, pp. 126–138, 2019.
- [63] 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>.
- [64] —, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [65] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011, pp. 15–22.

- [66] 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.
- [67] D. M. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” 2011.
- [68] C. Yang, Y. Yue, J. Zhang, M. Wen, and D. Wang, “Probabilistic reasoning for unique role recognition based on the fusion of semantic-interaction and spatio-temporal features,” *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1195–1208, May 2019.
- [69] S. Saeedi, M. Trentini, M. Seto, and H. Li, “Multiple-Robot Simultaneous Localization and Mapping: A Review,” *Journal of Field Robotics*, vol. 33, no. 1, pp. 3–46, 2016.
- [70] C. Fu, A. Carrio, M. A. Olivares-Mendez, R. Suarez-Fernandez, and P. Campoy, “Robust real-time vision-based aircraft tracking from unmanned aerial vehicles,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 5441–5446.
- [71] Y. Yue, P. G. C. N. Senarathne, C. Yang, J. Zhang, M. Wen, and D. Wang, “Probabilistic fusion framework for collaborative robots 3d mapping,” in *2018 21st International Conference on Information Fusion (FUSION)*, 2018, pp. 488–491.
- [72] S. Omidshafiei, S.-Y. Liu, M. Everett, B. T. Lopez, C. Amato, M. Liu, J. P. How, and J. Vian, “Semantic-level decentralized multi-robot decision-making using probabilistic macro-observations,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 871–878.
- [73] D. Scaramuzza, M. C. Achtelik *et al.*, “Vision-Controlled Micro Flying Robots: From System Design to Autonomous Navigation and Mapping in GPS-Denied En-

- vironments,” *IEEE Robotics Automation Magazine*, vol. 21, no. 3, pp. 26–40, Sept 2014.
- [74] Saeedi Sajad and Paull Liam and Trentini Michael , “Map merging for multiple robots using Hough peak matching,” *Robotics and Autonomous Systems*, vol. 62, no. 10, pp. 1408–1424, 2014.
- [75] Y. Yue, D. Wang, P. G. C. N. Sena, and C. Yang, “Robust submap-based probabilistic inconsistency detection for multi-robot mapping,” in *2017 European Conference on Mobile Robots (ECMR)*, Sept 2017, pp. 1–6.
- [76] P. C. Niedfeldt, A. Speranzon, and A. Surana, “Distributed map fusion with sporadic updates for large domains,” in *2015 IEEE International Conference on Robotics and Automation*, May 2015, pp. 2806–2813.
- [77] Y. Yue, D. Wang, P. Senarathne, and D. Moratuwage, “A hybrid probabilistic and point set registration approach for fusion of 3D occupancy grid maps,” in *2016 IEEE International Conference on Systems, Man, and Cybernetics*, 2016, pp. 1975–1980.
- [78] Y. Yue, P. G. C. N. Senarathne, C. Yang, J. Zhang, M. Wen, and D. Wang, “Hierarchical probabilistic fusion framework for matching and merging of 3-d occupancy maps,” *IEEE Sensors Journal*, vol. 18, no. 21, pp. 8933–8949, Nov 2018.
- [79] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” 2018.
- [80] J. Zhang and S. Singh, “Loam: Lidar odometry and mapping in real-time,” in *Robotics: Science and Systems Conference*, July 2014.
- [81] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, “OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees,” *Autonomous Robots*, 2013.
- [82] J. Jessup, S. N. Givigi, and A. Beaulieu, “Robust and efficient multirobot 3-d mapping merging with octree-based occupancy grids,” *IEEE Systems Journal*, vol. 11, no. 3, pp. 1723–1732, 2017.

- [83] R. Hadidi, J. Cao, M. Woodward, M. S. Ryoo, and H. Kim, “Distributed perception by collaborative robots,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3709–3716, Oct 2018.