Scene Understanding for Autonomous Robots Operating in Indoor Environments

by

Alejandra Carolina Hernández Silva

A dissertation submitted by in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Electrical Engineering, Electronics and Automation

University Carlos III of Madrid

Advisor(s):

Dr. Ramón Ignacio Barber Castaño
Dr. Oscar Martínez Mozos

Tutor:

Dr. Ramón Ignacio Barber Castaño

April 2021

To my family,
especially to my loving husband Jesús.

Acknowledgements

I would like to dedicate these lines to express my gratitude to all the people who have helped me in a way or another to achieve this goal that is so important to me. First, I would like to thank my tutor and director, Dr. Ramón Barber, for introducing me to the exciting world of research, for all his support, guidance, trust, and patience over these years of my Ph.D. studies. Thank you for giving me the opportunity to work with you. I also thank Dr. Oscar Martínez, for all the advice and technical talks and for allowing me to visit his laboratory in Sweden. I would like to thank the University Carlos III of Madrid and all the Robotics Lab members I met during these years. Thank you for your support, professionalism, and friendship; It is an honor to belong to this great research group.

I would like to especially thank my lab partner Clara Gómez. Thank you for being such a great partner and friend over these years, for your support, professionalism, your dedication at work. Thank you for being my partner in conferences and research visits. Definitely, the Ph.D. would not have been the same without you.

I would like to thank Dr. Zoltan Marton and his team for giving me the opportunity to work with them at the DLR institute in Germany. I would also like to thank Prof. Robert Babuška for welcoming me to his lab at the CTU in Prague. Especially thanks to Erik Derner for being a great lab partner, for our interesting discussions, for his professionalism and friendship. I also thank the people I met at Örebro University, especially Eduardo Gutiérrez; thank you for your support and friendship.

Finally and most importantly, I would like to thank my family. Thank you for all your support, your love, and for always being there. Despite the more than 7000 km. that separate us, I am always with you, and I know that you are with me celebrating this achievement. I leave for last, the person I must thank for having embarked on this path. I want to thank my husband, Jesús. Thank you for supporting me and helping me accomplish this goal, for being my partner on this roller coaster of life, for always believing in me, and for having a word of encouragement when I felt like I could not continue. Without you, I am sure I would not have made it. For you all my love and my gratitude.

To all my sincere thanks.

Published and Submitted Content

Parts of this thesis have been published in the following papers, journals and book chapters:

- **Hernandez, A. C.**, Gomez, C., Barber, R., and Mozos, O. M. (2021). Using Miscategorization of Places to Improve Service Robotics Tasks in Indoor Environments. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). [pending notification].

  * This paper is wholly included in this thesis in Chapter 5. The material from this source is singled out with an explicit reference in the text.

- **Hernandez, A. C.**, Durner M., Gomez C., Grixa I., Teikmanis O., Marton Z.C., and Barber R. (2021). Searching for Objects in Human Living Environments based on Relevant Inferred and Mined Priors. In European Conference on Mobile Robots (ECMR). [pending notification].

  * This paper is wholly included in this thesis in Chapter 6. The material from this source is singled out with an explicit reference in the text.

- **Hernandez, A. C.**, Derner, E., Gomez, C., Barber, R., and Babuška, R. (2020). Efficient Object Search Through Probability-Based Viewpoint Selection. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 6172–6179, Las Vegas, NV, USA. DOI: 10.1109/ IROS45743. 2020.9340989.

  * This paper is wholly included in this thesis in Chapter 6. The material from this source is singled out with an explicit reference in the text.

- **Hernandez, A. C.**, Gomez, C., Derner, E., and Barber, R. (2019). Indoor scene recognition based on weighted voting schemes. In 2019 European Conference on Mobile Robots (ECMR) (pp. 1-6). IEEE. DOI: 10.1109/ECMR.2019.8870931.

  * This paper is wholly included in this thesis in Chapter 4. The material from this source is singled out with an explicit reference in the text.

- **Hernandez, A. C.**, Gomez, C., and Barber, R. (2019). MiNERVA: Toposemantic Navigation Model based on Visual Information for Indoor Enviroments. IFAC-PapersOnLine (10th IFAC Symposium on Intelligent Autonomous Vehicles), 52(8), 43-48. DOI: 10.1016/ j.ifacol.2019.08.046.

  * This paper is partially included in this thesis in Chapter 4. The material from this source is singled out with an explicit reference in the text.

- **Hernández, A. C.**, Gómez, C., Barber, R., and Mozos, O. M. (2018). Object-based probabilistic place recognition for indoor human environments. In 2018 International Conference on Control, Artificial Intelligence, Robotics and Optimization (ICCAIRO) (pp. 177-182). IEEE. DOI: 10.1109/ ICCAIRO.2018.00037.

  * This paper is wholly included in this thesis in Chapter 4. The material from this source is singled out with an explicit reference in the text.

- Barber, R., Crespo, J., Gómez, C., **Hernández, A. C.**, and Galli, M. (2018). Mobile robot navigation in indoor environments: Geometric, topological, and semantic navigation. In Applications of Mobile Robots. IntechOpen. DOI: 10.5772 /intechopen.

  * This paper is partially included in this thesis in Chapter 4. The material from this source is not singled out.

- **Hernández, A. C.**, Gómez, C., Crespo, J., and Barber, R. (2017). Adding uncertainty to an object detection system for mobile robots. In 2017 6th International Conference on Space Mission Challenges for Information Technology (SMC-IT) (pp. 7-12). IEEE. DOI: 10.1109/SMC-IT.2017.9.

  * This paper is wholly included in this thesis in Chapter 3. The material from this source is singled out with an explicit reference in the text.

- **Hernández, A. C.**, Gómez, C., Crespo, J., and Barber, R. (2016). Object detection applied to indoor environments for mobile robot navigation. Sensors, 16(8), 1180. DOI:10.3390/ s16081180.

  * This paper is partially included in this thesis in Chapter 3. The material from this source is not singled out.

In addition to the aforementioned publications, during the development of the Ph.D., we have also contributed to the following publications:

- Gomez, C., **Hernandez, A. C.**, Barber, R., and Stachniss, C. (2021). Localization Exploiting Semantic and Metric Information in Non-static Indoor Environments. IEEE International Conference on Intelligent Robots and Systems (IROS). [pending notification].

  * The material of this source is not included in this thesis.

- Derner, E., Gomez, C., **Hernandez, A. C.**, Barber, R., and Babuška, R. (2021). Change detection using weighted features for image-based localization. Robotics and Autonomous Systems, 135, 103676. DOI: 10.1016/j.robot.2020.103676.

  * The material of this source is not included in this thesis.

- Gomez, C., Marius, F., Millane, A., **Hernandez, A. C.**, Nieto, J., Barber, R., and Siegwart, R. (2020). Hybrid Topological and 3D Dense Mapping through Autonomous Exploration for Large Indoor Environments. IEEE International Conference on Robotics and Automation (ICRA), pp. 9673-9679. DOI: 10.1109 /ICRA40945. 2020.9197226.

  * The material of this source is not included in this thesis.

- Gomez, C., **Hernandez, A. C.**, Derner, E., Barber, R., and Babuška, R. (2020). Object-Based Pose Graph for Dynamic Indoor Environments. IEEE Robotics and Automation Letters (IROS+RA-L submission), vol. 5 no. 4, pp. 5401-5408. DOI: 10.1109/ LRA.2020.3007402.

  * The material of this source is not included in this thesis.

- Gomez, C., **Hernandez, A. C.**, Derner, E., and Barber, R. (2019). Semantic Localization through Propagation of Scene Information in a Hierarchical Model. IEEE European Conference on Mobile Robots (ECMR), pp. 1-6. DOI: 10.1109/ ECMR.2019. 8870972.

  * The material of this source is not included in this thesis.

- Gomez, C., **Hernandez, A. C.**, Barber, R., Moreno, L., and Mozos, O.M. (2019). Localization of Mobile Robots Incorporating Scene Information in a Hierarchical Model. IEEE International Conference on Robotic Computing (IRC), pp. 429-430. DOI: 10.1109/IRC.2019.00084.

  * The material of this source is not included in this thesis.

- Derner, E., Gomez, C., **Hernandez, A. C.**, Barber, R., and Babuška, R. (2019). Towards life-long autonomy of mobile robots through feature-based change

detection. In 2019 European Conference on Mobile Robots (ECMR) (pp. 1-6). IEEE. DOI: 10.1109/ECMR.2019.8870940.

* The material of this source is not included in this thesis.

- Gomez, C., **Hernandez, A. C.**, and Barber, R. (2019). Topological Frontier-Based Exploration and Map-Building Using Semantic Information. Sensors, vol. 19, no. 20, p. 4595. DOI: 10.3390/s19204595.

  * The material of this source is not included in this thesis.

- Gomez, C., **Hernandez, A. C.**, Moreno, L., and Barber, R. (2018). Qualitative Geometrical Uncertainty in a Topological Robotic Localization System. IEEE International Conference on Control, Artificial Intelligence, Robotics and Optimization (ICCAIRO), pp. 183-188. DOI: 10.1109/ICCAIRO.2018.00038.

  * The material of this source is not included in this thesis.

- Gomez, C., **Hernandez, A. C.**, Crespo, J., and Barber, R. (2017). A Topological Navigation System for Indoor Environments based on Perception Events. International Journal of Advanced Robotic Systems (IJARS), vol. 14, no. 1. DOI: 10.1177/ 1729881416678134.

  * The material of this source is not included in this thesis.

- Crespo, J., Gomez, C., **Hernandez, A. C.**, and Barber, R. (2017). A semantic labeling of the environment based on what people do. Sensors, vol. 17, no 2, p. 260. DOI: 10.3390/s17020260.

  * The material of this source is not included in this thesis.

- Gomez, C., **Hernandez, A. C.**, Crespo, J., and Barber, R. (2017). Uncertainty-based Localization in a Topological Robotic Navigation System. IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), pp.67-72. DOI: 10.1109/ICARSC.2017.7964054.

  * The material of this source is not included in this thesis.

Other Research Merits

During the Ph.D. studies, I made three research visits to different research laboratories, which contributed to the development of part of the work presented in this thesis. These valuable collaborations were made with the following institutions:

- Institute of Robotics and Mechatronics. German Aerospace Center (DLR), Germany. The leader of the Department for Perception and Cognition, Dr. Rudolph Triebel welcomed me for a 3-months research visit, working mainly in searching strategies. The collaboration resulted in a conference paper included in this thesis: Searching for Objects in Human Living Environments based on Relevant Inferred and Mined Priors, submitted to ECMR 2021 and which is currently under revision.

- Czech Institute of Informatics, Robotics and Cybernetics (CIIR). Czech Technical University in Prague, Czech Republic. I received an invitation from Prof. Robert Babuška to collaborate in his lab for a 2-month research visit. I continued with the work started at the DLR institute based on object searching strategies. This fruitful collaboration generated several publications, one of them included in this thesis: Efficient Object Search Through Probability-Based Viewpoint Selection published at IROS 2020.

- Center for Applied Autonomous Sensor Systems (AASS). Örebro University, Sweden. I made a third 3-month research visit thanks to the invitation of Dr. Oscar Martinez Mozos to work with his group in the Cognitive Robotic Systems Laboratory. The collaboration resulted in a conference paper included in this thesis: Using Miscategorization of Places to Improve Service Robotics Tasks in Indoor Environments, submitted to IROS 2021 and which is currently under revision.

Abstract

The idea of having robots among us is not new. Great efforts are continually made to replicate human intelligence, with the vision of having robots performing different activities, including hazardous, repetitive, and tedious tasks. Research has demonstrated that robots are good at many tasks that are hard for us, mainly in terms of precision, efficiency, and speed. However, there are some tasks that humans do without much effort that are challenging for robots. Especially robots in domestic environments are far from satisfactorily fulfilling some tasks, mainly because these environments are unstructured, cluttered, and with a variety of environmental conditions to control.

This thesis addresses the problem of scene understanding in the context of autonomous robots operating in everyday human environments. Furthermore, this thesis is developed under the HEROITEA research project that aims to develop a robot system to help elderly people in domestic environments as an assistant. Our main objective is to develop different methods that allow robots to acquire more information from the environment to progressively build knowledge that allows them to improve the performance on high-level robotic tasks. In this way, scene understanding is a broad research topic, and it is considered a complex task due to the multiple sub-tasks that are involved. In that context, in this thesis, we focus on three sub-tasks: object detection, scene recognition, and semantic segmentation of the environment.

Firstly, we implement methods to recognize objects considering real indoor environments. We applied machine learning techniques incorporating uncertainties and more modern techniques based on deep learning. Besides, apart from detecting objects, it is essential to comprehend the scene where they can occur. For this reason, we propose an approach for scene recognition that considers the influence of the detected objects in the prediction process. We demonstrate that the exiting objects and their relationships can improve the inference about the scene class. We also consider that a scene recognition model can benefit from the advantages of other models. We propose a multi-classifier model for scene recognition based on weighted voting schemes. The experiments carried out in real-world indoor environments demonstrate that the adequate combination of independent classifiers allows obtaining a more robust and precise model for scene recognition.

Moreover, to increase the understanding of a robot about its surroundings, we propose a new division of the environment based on regions to build a useful representation of the environment. Object and scene information is integrated into a probabilistic fashion generating a semantic map of the environment containing meaningful regions within each room. The proposed system has been assessed on simulated and real-world domestic scenarios, demonstrating its ability to generate consistent environment representations.

Lastly, full knowledge of the environment can enhance more complex robotic tasks; that is why in this thesis, we try to study how a complete knowledge of the environment influences the robot's performance in high-level tasks. To do so, we select an essential task, which is searching for objects. This mundane task can be considered a precondition to perform many complex robotic tasks such as fetching and carrying, manipulation, user requirements, among others. The execution of these activities by service robots needs full knowledge of the environment to perform each task efficiently. In this thesis, we propose two searching strategies that consider prior information, semantic representation of the environment, and the relationships between known objects and the type of scene. All our developments are evaluated in simulated and real-world environments, integrated with other systems, and operating in real platforms, demonstrating their feasibility to implement in real scenarios, and in some cases outperforming other approaches. We also demonstrate how our representation of the environment can boost the performance of more complex robotic tasks compared to more standard environmental representations.

Resumen

La idea de tener robots entre nosotros no es nueva. Continuamente se realizan grandes esfuerzos para replicar la inteligencia humana, con la visión de tener robots que realicen diferentes actividades, incluidas tareas peligrosas, repetitivas y tediosas. La investigación ha demostrado que los robots son buenos en muchas tareas que resultan difíciles para nosotros, principalmente en términos de precisión, eficiencia y velocidad. Sin embargo, existen tareas que los humanos realizamos sin mucho esfuerzo y que son un desafío para los robots. Especialmente, los robots en entornos domésticos están lejos de cumplir satisfactoriamente algunas tareas, principalmente porque estos entornos no son estructurados, pueden estar desordenados y cuentan con una gran variedad de condiciones ambientales que controlar.

Esta tesis aborda el problema de la comprensión de la escena en el contexto de robots autónomos que operan en entornos humanos cotidianos. Asimismo, esta tesis se desarrolla en el marco del proyecto de investigación HEROITEA que tiene como objetivo desarrollar un sistema robótico que funcione como asistente para ayudar a personas mayores en entornos domésticos. Nuestro principal objetivo es desarrollar diferentes métodos que permitan a los robots adquirir más información del entorno a fin de construir progresivamente un conocimiento que les permita mejorar su desempeño en tareas robóticas más complejas. En este sentido, la comprensión de escenas es un tema de investigación amplio, y se considera una tarea compleja debido a las múltiples subtareas involucradas. En esta tesis nos enfocamos específicamente en tres subtareas: detección de objetos, reconocimiento de escenas y etiquetado semántico del entorno.

Por un lado, implementamos métodos para el reconocimiento de objectos considerando entornos interiores reales. Aplicamos técnicas de aprendizaje automático incorporando incertidumbres y técnicas más modernas basadas en aprendizaje profundo. Además, aparte de detectar objetos, es fundamental comprender la escena donde estos se encuentran. Por esta razón, proponemos un modelo para el reconocimiento de escenas que considera la influencia de los objetos detectados en el proceso de predicción. Demostramos que los objetos existentes y sus relaciones pueden mejorar el proceso de inferencia de la categoría de la escena. También consideramos que un modelo de reconocimiento de escenas puede beneficiarse de las ventajas de otros modelos. Por ello, proponemos un multiclasificador para el reconocimiento de

escenas basado en esquemas de votación ponderados. Los experimentos llevados a cabo en entornos interiores reales demuestran que la combinación adecuada de clasificadores independientes permite obtener un modelo más robusto y preciso para el reconocimiento de escenas.

Adicionalmente, para aumentar la comprensión de un robot acerca de su entorno, proponemos una nueva división del entorno basada en regiones a fin de construir una representación útil del entorno. La información de objetos y de la escena se integra de forma probabilística generando un mapa semántico que contiene regiones significativas dentro de cada habitación. El sistema propuesto ha sido evaluado en entornos domésticos simulados y reales, demostrando su capacidad para generar representaciones consistentes del entorno.

Por otro lado, el conocimiento integral del entorno puede mejorar tareas robóticas más complejas; es por ello que en esta tesis analizamos cómo el conocimiento completo del entorno influye en el desempeño del robot en tareas de alto nivel. Para ello, seleccionamos una tarea fundamental, que es la búsqueda de objetos. Esta tarea mundana puede considerarse una condición previa para realizar diversas tareas robóticas complejas, como transportar objetos, tareas de manipulación, atender requerimientos del usuario, entre otras. La ejecución de estas actividades por parte de robots de servicio requiere un conocimiento profundo del entorno para realizar cada tarea de manera eficiente. En esta tesis proponemos dos estrategias de búsqueda de objetos que consideran información previa, la representación semántica del entorno, las relaciones entre los objetos conocidos y el tipo de escena. Todos nuestros desarrollos son evaluados en entornos simulados y reales, integrados con otros sistemas y operando en plataformas reales, demostrando su viabilidad de ser implementados en escenarios reales y, en algunos casos, superando a otros enfoques. También demostramos cómo nuestra representación del entorno puede mejorar el desempeño de tareas robóticas más complejas en comparación con representaciones del entorno más tradicionales.

Contents

List of Figures

List of Tables

1 Introduction

This thesis addresses the problem of scene understanding in the context of autonomous robots operating in everyday human environments. In this chapter, first, the motivation and an overview of the problem statement and main challenges are exposed. Second, the main objectives of this doctoral dissertation are described. Finally, the main contributions and the outline of this thesis are presented.

1.1 Motivation

The idea of having robots among us is not new. Already in science fiction books such as *I, Robot* by Isaac Asimov [2], the relationship between robots and humans was illustrated, in addition to introducing the three laws of robotics. In the classic R.U.R. (Rossum's Universal Robots) by Karel Čapek [18], a factory that made artificial people that work for humans, and the term *robot* are also introduced. Also, in media, recent movies and series show robots cohabiting with humans, assuming different tasks with a high level of intelligence.

Regarding robotics research, great efforts are continually made to replicate human intelligence. Many articles, theses, and books show the interest in introducing service robots in everyday human environments [41, 119, 167] . Some works, such as the presented by Pigini et al. [135] emphasize the need for service robots for the elderly. Their work is motivated by the aging population with the goal of covering care and social needs to improve their quality of life. Furthermore, there is research focused on robotic applications for

healthcare facilities. The work of Vänni et al. [169] establishes the need for service robots in houses and hospitals as co-workers to reduce the workload and to help patients.

If we compare the advances to date in this matter with science fiction stories, we are still far from having a general service robot with a high capacity to understand the environment, interact with humans, and able to carry out different types of activities that a user requires. So far, we have robots for specific purposes, such as in the industry, mainly in automation for car manufacturing [14, 90].

Regarding household environments, cleaning robots are one of the first service robot families to achieve commercial viability, demonstrating great results [62, 92, 152]. An example is the multiple brands of vacuum cleaning robots that we can find, such as the emblematic Roomba to newer and less expensive models. Robots in hotels and restaurants start to appear as butlers and bringing trays [76, 163]. For example, some hospitals have implemented the Savioke's Relay robot to help in transport tasks such as bringing medicines and food [22, 125]. Also, in restaurants, there are robotic platforms that work as part of the staff [81, 194]. An example is the Peanut robot (Keenon robotics) used to bring food and take away waste, with prerecorded messages to interact with the customers when the robot approaches the table. However, despite the advancements so far, getting robots to carry out all of these tasks efficiently is currently a research effort in progress.

The work developed in this thesis is framed in the HEROITEA project (Heterogeneous Social - Mobile Manipulator Robot Intelligent Teams for Elderly-People Assistance. RTI 2018-095599-B-C21). This project aims to develop a robot system to help the elderly in domestic environments as an assistant. The main activities to be carried out include developing path planning algorithms and navigation methods, developing methods for semantic understanding of the elements of the environment, and developing strategies for pre-grasping and grasping objects. This thesis tackles mainly methods to progressively improve the scene understanding of a robot so that it can improve its performance in more complex robotic tasks.

1.2 Problem Statement and Challenges

Research has demonstrated that robots are good at many tasks that are hard for us, mainly in terms of precision, efficiency, and speed. However, there are some tasks that humans do without much effort that are challenging for robots. Some examples are location in space, recognizing objects, and learning new categories to use later. Other challenges would be the performance of daily activities such as looking for something in our house, preparing a meal, among others. Let us introduce Fig. 1.1 which illustrates typical houses with different layouts and object placements. Imagine that you decide to acquire a new brand of robot for your house as an assistant. Moreover, imagine that you ask the robot to look for an object you

need to find, such as car keys or medicine. Maybe you want to prepare a robot to work in the kitchen making food or as a butler to bring you a glass of water, food, etc. The question that arises is: What would be the capabilities that a robot needs to accomplish these tasks? We have listed some of them below:

Figure 1.1: Example scenarios of household environments. Different layouts and objects placement characterize these types of dynamic environments.

- The robot has to move around the environment. It needs to explore the new environment to build and maintain a map. Furthermore, it has to be able to localize itself in it and has path planning skills to get around it.
- The robot needs to know the semantic elements present in the environment to interact with them. The robot has to know the house and the different rooms, such as where the bathroom is. What a kitchen and a living room are. Also, it is necessary the information about the type of objects, where the objects are, and where they are stored. It would also be helpful to know the place where everything usually is.
- This semantic information has to be associated with the map of the environment for later use. This implies the ability to identify objects and categorize rooms, as well as the poses of furniture and objects and the area they occupied on the map.
- Once all this information is obtained, the robot must have a strategy to follow based on the assigned task, which will define the robot's behavior. Each strategy must

adequately combine all this information to achieve the assigned goal in the most efficient way possible.

As can be seen, all these skills imply a complete knowledge of the environment to allow the robot to interact appropriately with it. In this thesis, we focus on developing methods to provide a robot with complete knowledge of the environment, especially indoor environments inhabited by humans. All of this has to do with scene understanding. The navigation and path planning skills are beyond the scope of this thesis. However, along with the chapters, we use different available algorithms, some of them developed in our laboratory, to test our approaches in real-world environments and integrate them into real robotic platforms.

Scene understanding can be considered as the ability of humans to comprehend their surroundings and the elements around them. This ability determines the degree of interaction with the environment and the activities that can be carried out. For robotics, this ability is crucial to incorporate service robots in domestic environments. Humans not only perceive the visual features of a scene but reason about the semantics and geometry of the elements in it.

In this thesis, we address the scene understanding for service robots that will operate in scenarios such as presented in Fig. 1.1. We focus on understanding the environment by analyzing the elements and the relationships between them to endow a robot with more capabilities to improve its knowledge about the environment. Likewise, this semantic information can be used to generate more meaningful representations of the environment. In this way, the robot can then use this knowledge to perform other high-level robotic tasks more efficiently.

One of the challenges is the variability of these environments; people move the objects and place them where they consider them more convenient. Furthermore, these environments are typically cluttered, occluded, and unstructured. That is one of the complexities of building general methods that work in any environment. Although these tasks can be accomplished for humans almost without a significant effort, it has not yet been shown that a robot can perform them in the same way or better without including certain constraints to work. This differs from industrial environments, where it is increasingly common to find robots performing different tasks due to the structure of the environment and the nature of the activities, which in most cases are controlled, repetitive and predefined.

1.3 Research Purpose

This thesis addresses the problem of scene understanding for service robots working in indoor human environments. This refers to developing mechanisms to provide the robot with

more capabilities to comprehend its surroundings by gathering visual information through its sensors. We focus on proposing implementations able to work in real-world scenarios with real robots and partially considering the dynamism of the environment. Furthermore, our methods must be able to integrate with other robot systems.

The main objective of this thesis is to develop and improve different methods that allow the robot to acquire more information from the environment to build knowledge that allows them to enhance the performance in more complex robotic tasks. The methods and strategies presented in this thesis have been developed to enhance the operation of autonomous robots in human indoor environments and promote a better understanding of their surroundings that allows more efficient interactions between humans and robots.

Most of this understanding is based on perception. For robots, if something is not perceived by their sensors, it does not exist. That represents a clear limitation compared with humans because we do not need to know every object to understand the functionality of the environment. Neither do we need so much information to infer other objects or perform different activities even in unknown environments. Human environments are structured and organized by humans for humans, so incorporating robots to interact with humans and act accordingly constitutes a big challenge. The robot needs to gather information about the objects, their location, knowledge about the different areas of an environment such as rooms, and how they are connected. The relationships between objects and areas can help to define the activities that can be carried out. All this implies data gathering and reasoning about it.

Scene understanding is a broad research topic. It is considered a complex task due to the multiple sub-tasks involved, such as object recognition, scene classification, semantic segmentation, pose estimation, 3D reconstruction, saliency detection, physics-based reasoning, and affordance prediction. In this thesis, we tackle three specific sub-tasks we consider essentials to reach full knowledge of the environment: object detection, scene recognition, and semantic segmentation. There are other two sub-tasks we consider transversal; that is, they are used for each of the main sub-tasks, pose estimation of objects and physics-based reasoning, specifically, support relationships of objects. In this thesis, we analyze and reason about the environment as a whole; the proposed approaches extract semantic information from visual data and combine them to acquire full knowledge of an environment.

Furthermore, this thesis covers an application that combines all available semantic information to demonstrate how a complete knowledge of the environment can improve the performance of some complex robotic tasks. To do so, we select the task of searching for objects. This task can be considered a precondition to perform many complex robotic tasks such as fetching and carrying, manipulation, user requirements, among others. The execution of these activities by service robots needs complete knowledge of the environment to perform each task efficiently. Finding objects implies reasoning from certain information

of the environment that allows the robot to infer possible locations of a target object. This process may consider prior information, semantic representation of the environment, and the relationships between known objects and the type of a scene.

According to this, we propose the following specific objectives:

- Developing and implementing methods to provide mobile robots with capabilities to recognize different object categories in real-world environments.
- Developing and implementing methods for scene recognition in real-world environments that consider uncertainties and objects' influence in the prediction process.
- Building a meaningful representation of the environment based on the available semantic information to improve some high-level tasks in service robotics.
- Developing and implementing object searching strategies making use of the available semantic information to demonstrate how semantic understanding of the environment improves the efficiency of some complex robotic tasks.

In addition to these specifics goals, scene understanding developments must be scalable, robust, generalizable, and fast to be applied in robots that operate in real-time and in real-world scenarios. Also, all the applications must be integrated with other modules such as navigation systems, path planning algorithms, interaction modules, among others.

1.4 Contributions

In this thesis, we try to provide new solutions or improvements to existing methods to the complex problem of scene understanding and the main sub-tasks considered in this work. The first contributions of this thesis are in the field of object and scene recognition. We implement object detection models capable of working in real indoor environments. Based on machine learning, the first method implements an uncertainty estimation model that considers different information sources to make the recognition process more realistic. A second method based on more modern techniques uses a convolutional neural network to detect objects in real-time. The two models are integrated into a real platform and with a navigation system guaranteeing their operation in the most realistic conditions possible. Concerning scene recognition, we propose a model that incorporates prior information to predict the scene category of a place. A model of uncertainties is proposed based on the information about the objects in the scene and the relationships between them to correct the scene recognition process's errors. Furthermore, we propose a multi-classifier

model for scene recognition considering independent base classifiers' outcomes as priors. We implement a weighted voting scheme based on genetic algorithms to combine different classifiers to improve recognition performance. We demonstrate that the combination of different techniques for independent base classifiers improves the scene recognition results. In this way, the proposed model is more general and scalable and can be adjusted to any source of information. Compensation for the errors made by each base classifier is made through the weighted fusion of its classification results, thus helping to make a better estimate of the scene perceived by the robot.

The following contribution in this thesis refers to the way to represent the environment. We propose an approach that creates a new division of the environment based on regions while maintaining confusions (miscategorizations) of places. The key idea consists of keeping the miscategorizations of smaller regions within entire rooms. Different categorization systems may result in different confusing areas due to the intrinsic nature of the perception system. Semantic labels are assigned to regions within each room by assuming that they can vary depending on contextual information. We integrate scene and object information by applying a Bayesian filter and a decision-making procedure to keep temporal coherence. Our approach has been assessed on simulated and real-world scenarios, demonstrating its ability to generate consistent representations of the environment.

The main contribution in this topic is a framework for consistent subdivision and labeling of domestic environments based on regions using different modalities. Moreover, we present a solution for merging region-based maps originated by different sources of information, in particular, visual appearance and object detection. The resulting experiments show that keeping smaller regions and their confusing labels can boost the performance of some tasks in service robotics, compared to previous methods where rooms are treated as a whole. In particular, we show an increase in the efficiency of a service robot performing an object search task. Besides, these confusions may be more intuitive for people and, therefore, improve human-robot communication.

In this thesis, we also intend to demonstrate how full knowledge of the environment can make some robotic tasks more efficient. To do so, we have selected the task of searching for objects. This mundane task can be considered a precondition for more complex robotic tasks. In several tasks, the robot first has to find an object or several objects to then decide the actions to take with them. Searching for objects is also highly relevant in showing the non-expert user that a robot can understand the world. The contributions in this topic are two novel strategies for searching for objects considering semantic information.

The first method is a global strategy that searches for unseen target objects based on the reasoning about which scenes and with which objects they co-occur. This global strategy incorporates prior semantic information to obtain an efficient search strategy for finding a specific object in an indoor environment. The method is based on a probabilistic graphical

model that can efficiently determine the most likely location of an object. Our method consists of an inference process based on a Conditional Random Field (CRF) that fuses the information about other previously detected objects, the semantic floor map, and the object-object/-room relations to build a prediction map with the most promising locations for an unseen object. The experiments demonstrate the usefulness and efficiency of our method in estimating the most probable room where an object can be located. Besides, the method has been tested in a complete dataset of apartments, which demonstrates the flexibility to apply it in different environmental conditions.

The second method is a local strategy that determines the robot's best pose based on the analysis of the candidate locations called viewpoints. The method is further refined by choosing specific places within the room to be visited. The core of the process is in the analysis and selection of the best locations. The main contribution in this work is a probabilistic analysis and selection of best viewpoints to efficiently find objects in partially known environments and the multi-objective optimization to maximize the probability of finding the target object and minimize the distance traveled. Choosing the correct viewpoint has a significant impact on the robot's performance. That is why promising robot orientations are identified as well to speed up the search further. In case the object is not found, probabilities are updated for the next best viewpoint analysis. The strategy has been assessed in simulated and real-world environments, demonstrating its validity and efficiency on the task of finding target objects.

1.5 The Thesis Path

The development of this thesis has been an incremental process of acquisition of significant information to achieve that complete understanding of the environment so necessary for service robots that operate in indoor environments. Fig. 1.2 shows the development of this thesis over the time of the Ph.D. It illustrates the evolution of the work carried out and the collaborations with other international institutions that have contributed to the successful completion of this thesis. The reader may realize that the order of the chapters of the document does not correspond chronologically with the development of the work and the publications. That is why we dedicate this section to briefly explain the evolution of the work carried out.

The Ph.D. began as a continuation of the work done in my master thesis titled *Object Detection applied to Indoor Environments for Mobile Robot Navigation*. In that work, we explored different approaches to recognize objects in indoor environments and conducted the first integration with a navigation system. Then, we tried to strengthen that work by adding the concept of uncertainty to our machine learning-based model. This work is explained in Chapter 3.3. Then, we started to think about what other kind of information

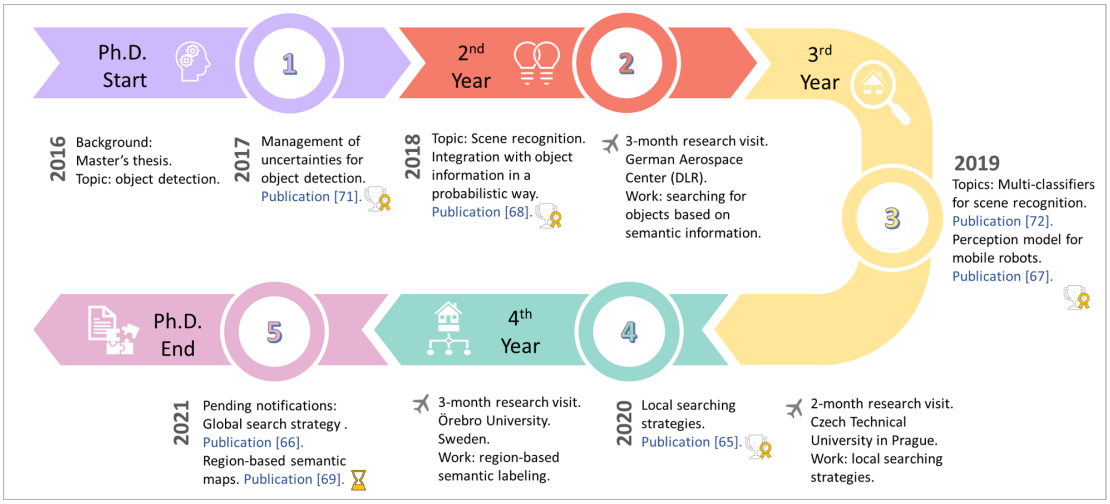


Figure 1.2: Illustration of the evolution of the work carried out in this thesis. Each color corresponds to a year of the Ph.D., showing the topics covered during this time and the publications achieved.

we could get from the environment apart from objects, and we decided to work on a scene recognition model. The proposed scene recognition model, included in Chapter 4.3, considers as priors object information in order to improve the scene recognition results.

At this point, we had the opportunity to make a first research visit to the German Aerospace Center (DLR), and the idea that came up was to work on an application for service robots that would make use of the semantic information obtained from the environment to enhance the robot's performance. We decided to work on searching for objects. The selection of this task is justified because it constitutes a precondition for more complex robotic tasks. We proposed a global searching strategy presented in Chapter 6.3 that combines different types of prior information, such as object-object and object-scene relationships, 3D object detections, and a semantic map of the environment, to obtain a probabilistic understanding of the location of unseen target objects.

Subsequently, we retook the scene recognition topic to reinforce the initial approach. We developed a multi-classifier for scene recognition based on weighted voting schemes. The idea behind this work was to take advantage of the benefits of different independent classifiers and compensating for the errors of each of them. This work is detailed in Chapter 4.4. At the same time, we moved to more modern and robust techniques to detect objects and implemented a deep learning-based approach for detecting objects in domestic environments. This implementation is presented in Chapter 3.4. Afterward, we received an invitation to do a research visit to the Czech Technical University in Prague. During this second visit, we continued with the idea started at the DLR, developing a local strategy for the

search object task. We proposed a novel method based on the analysis and selection of the best locations where a target object can be located. We were able to maximize the probability of finding the object through an optimized cost function while minimizing the distance traveled by the robot. This work is presented in Chapter 6.4.

The final work in this thesis has been developed during a third research visit to the Center for Applied Autonomous Sensor Systems (AASS) of the Örebro University in Sweden. One of the inputs to the searching methods developed is a semantic representation of the environment that is traditionally based on room labels. We decided to use the semantic information available from the environment to create a new representation based on meaningful regions. We proposed a novel method to subdivide the environment by exploiting scene and object information to infer the label of each part of the environment. This new representation was used as input of our searching strategy to demonstrate how the new representation can improve the performance of other robotic tasks. The details of this method are presented in Chapter 5.

1.6 Document Structure

In this document, each chapter addresses a specific topic of this thesis. Each one includes a review of the state of the art of the specific topic, a detailed explanation about the methods and strategies developed, the experimental evaluation, and a brief discussion of the results obtained. This document is structured as follows:

Chapter 2 - Scene Understanding Background

Chapter 2 introduces the main concepts of scene understanding. We present a revision of the main works in the literature regarding this topic that constitute the foundations of this thesis. Since scene understanding is a complex task, we review different approaches depending on their purpose and the different sub-tasks that are addressed. This chapter aims to provide the reader with a background for a better comprehension of the rest of the chapters.

Chapter 3 - Object Recognition for Mobile Robots

The understanding of the environment involves knowledge about the existing elements and their relationships. Objects represent a rich source of information for performing many robotic tasks. This chapter describes the models developed to recognize objects in indoor environments. We first review the state of the art related to object detection. Then, we present an approach based on machine learning that considers uncertainties to improve the recognition process. Subsequently, we introduce another model based on more modern and

robust methods using deep learning techniques. Both approaches are evaluated in real-world environments.

Chapter 4 - Scene Recognition for Indoor Environments

Knowing the environment implies identifying objects, but it is also necessary to recognize the scene where those objects are located and the different actions that can be carried out. This chapter presents the scene recognition models developed in this thesis. First, a probabilistic scene recognition model is proposed. The model considers the influence of the objects in the scene to modify and correct the final probability of being in a place. Second, we propose a multi-classifier for scene recognition based on weighted voting schemes taking advantage of independent classifiers. It also contributes to compensate for the errors of each of them. Two strategies are implemented, one based on accuracy and the other based on genetic algorithms.

Chapter 5 - Semantic Labeling

In Chapters 3 and 4 we present several strategies to gather different types of information from the environment. In this chapter, we focus on segmenting the environment into meaningful regions considering semantic information. We propose a method to build region-based semantic maps that integrates scene and object information in a probabilistic fashion to generate different regions within a room. This information is obtained through deep learning-based classification models that are then combined by applying a Bayesian filter and a decision-making procedure that allows keeping temporal coherence. We also analyze and evaluate how our representation can enhance the robot's performance in more complex robotic tasks compared to more standard environmental representations.

Chapter 6 - Searching Strategies

In Chapter 6 we choose an essential application which is searching for objects to demonstrate how a complete knowledge of the environment can improve other robotic tasks. Finding objects can be considered an important activity and a precondition to perform other tasks, and it is highly relevant in showing the non-expert users that a robot can understand the world. We address the problem from two points of view. First, we propose a global strategy to search for target objects that have not been seen before, according to the reasoning on which scenes and which objects they co-occur with. The model combines appropriately different types of prior information and integrates them through a probabilistic graphical model. Second, we present a local search strategy that analyzes and selects the best locations within a room based on the probabilities of finding the target object in the area covered

by the sensor. Both searching strategies are designed for unknown objects and domestic environments.

Chapter 7 - Conclusions

Finally, Chapter 7 discusses the results and the contributions of this thesis. Furthermore, we suggest some research directions for future work.

2
Scene Understanding Background

In recent years, numerous advances in robotic fields such as computer vision, navigation, manipulation, artificial intelligence, to name a few, demonstrate that the robots will be between us earlier than later. Despite the promising results, some challenges remain if we talk about robot understanding. This thesis presents different methods and strategies to incrementally acquire a complete knowledge of the environment by analyzing the connection between their elements and the possibilities to use them efficiently during the execution of different robotic tasks. This chapter aims to present a revision of the main works in the literature regarding scene understanding that constitute the foundations of this thesis.

2.1 Overview

Scene understanding can be considered the ability of humans to comprehend their surroundings and the elements around them. This ability determines the degree of interaction with the environment and the activities that can be carried out. For robotics, this ability is crucial to incorporate service or domestic robots in human environments. Humans not only perceive the visual features of a scene but reason about the semantic and geometry of the elements in it.

Nasser et al. [130] define scene understanding as: "the process of analyzing a scene by considering the geometric and semantic context of its contents and the intrinsic relationships between them". Scene understanding is a complex task that implies different sub-tasks

or disciplines [20, 130, 205]. According to Nasser et al. [130], these disciplines are object recognition, scene classification, geometric reasoning, semantic segmentation, pose estimation, 3D reconstruction, saliency detection, physics-based reasoning, and affordance prediction.

Image recognition is the basis of scene understanding. The information about objects and the classification of a scene can help in this complex task. Object detection is an essential task for service robots in indoor environments [28, 197, 207] providing the category and the location of the objects and, in some cases, their shape and covered area. Usually, the outcome of an object detector is the visible parts of an object enclosed in a 2D or 3D bounding box. This task implies the pose estimation that includes the object's position and its orientation. Scene recognition includes the identification of a scene as a whole [55, 74, 139]. This task is considered very challenging, especially in indoor human environments characterized for being cluttered and unstructured, which implies occlusions, changes in the scale and viewpoint, ambiguity, and complex object interactions.

Regarding semantic segmentation, this task generates labeled pixels in an image with the semantic meaning of a category [104, 111, 193]. Depending on the application, this task can also involve labeling each pixel in a whole environment. Another important aspect is that a robot has to analyze the content of a scene and establish the relationships between scene elements. Physics reasoning [122, 180, 199] implies the inference of some dynamism from a static image. The previous patterns of motion are analyzed to predict the future events that can occur in a scene. That includes the study of the objects' stability and their physical relationships.

Since humans perceive the world in 3D, a full 3D reconstruction of the scene considering RGB-D images is desirable. The 3D reconstruction task merges the geometry of the objects, 3D shapes, and the available information of the scene to yield a robust reconstruction [40, 91]. This task has to deal with incomplete information from the environment and sensor noise, leading to pose errors.

With regards to salience detection, it is about identifying relevant parts or regions in a scene. It is based on the attention concept [15]. The relationships between objects can also be studied in terms of their functionality or affordance. That is, the set of actions that can be carried out on a specific object [43, 83]. This task provides new attributes to the objects, and this knowledge can be helpful to perform human interaction tasks in real environments.

In this thesis, we address the scene understanding for autonomous robots as a complex process that combines appropriately some of the sub-tasks described briefly above to give a robot a set of specific capabilities to understand and interact with its surroundings. In the following section, we revise the main work related to scene understanding and the sub-tasks addressed in each work according to a specific purpose.

2.2 Related Work

Scene understanding is one of the most important challenges in computer vision and robotics. It is considered a complex task due to the multiple sub-tasks that are involved, such as object recognition, scene classification, geometric reasoning, semantic segmentation, pose estimation, 3D reconstruction, saliency detection, physics-based reasoning, and affordance prediction. Scene understanding task has been addressed in different ways: through the parsing of single images [7, 20, 89, 91, 107, 176, 195, 201, 205], and considering the understanding of the whole environment where a robot moves [84, 147, 200]. In the literature, depending on the final goal pursued, the approaches emphasize some sub-tasks more than others. In most of the reviewed works, the scene understanding is oriented to object detection [89, 107, 176, 206], scene recognition [54, 55, 158, 160], 3D reconstruction [91, 147, 157, 199, 200] and semantic segmentation [7, 176, 201, 205]. Other approaches combine some of these sub-tasks such as [54, 158]. Furthermore, they may be aimed at unknown, partially known and completely known environments.

2.2.1 Scene Understanding for Single Images

Some approaches address the problem of scene understanding through the analysis of single images. Zhou et al. [205] focus on instance segmentation and semantic labeling using as input a single RGB image (Fig. 2.1). Markov Random Fields (MRF) are used to combine the different elements of the environment, inferring the semantic labels of the regions and their support relationships. Besides, Support Vector Machines (SVM) are implemented to learn the weights of the energy function of the MRF. Zheng et al. [201] introduce a method based on convolutional neural networks combined with Conditional Random Fields (CRF) to label each pixel in the image. Other works, such as the approach of Cong et al. [23] tackles the problem of scene understanding as an abstract interpretation. Their method uses the information about objects and their relations to construct a scene graph giving an abstract interpretation of each image. The nodes in the graph constitute the detected objects, and the edges are the relationships between them.

On the other hand, some works are more oriented towards object detection. The approach by Gupta et al. [54] proposes a set of algorithms to analyze RGB-D images from cluttered indoor environments. The techniques include contour detection, object detection, and semantic segmentation applied to the NYU-Depth V2 dataset [154]. The work by Wang et al. [176] reasons about semantic segmentation, as well as 3D object detection and pose estimation from a single image.

In the work by Choi et al. [20] the core of the method is the 3D representation of the objects in indoor environments. A hierarchical scene model is built combining object detections, scene recognition, and layout estimation. Fig. 2.2 shows some results of their

Figure 2.1: Instant segmentation and semantic labeling of single images presented in [205]. The method based on MRF segments images into regions and infer support relationships.

method applied to an indoor-scene-object dataset. A scene recognition method based on the Spatial Pyramid Matching (SPM) and an object detector based on the Deformable Part Model (DPM) are employed to estimate the scene composition. The image parsing is addressed as an energy maximization problem.

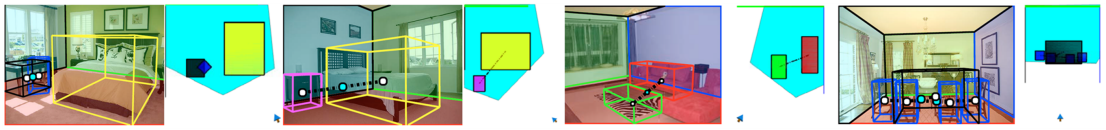


Figure 2.2: Results of the scene understanding framework proposed by [20]. The model reasons about the semantic class of the image, its spatial layout and the objects within the scene.

The approach by Zhang et al. [195] uses depth images as input of a 3D convolutional neural network that predicts the 3D location of multiple objects in a scene. The network learns the contextual information of scenes and objects. Different scene templates are defined to learn the spatial relationships between objects and global scene relations. The method generates a 3D bounding box of each object in the 2D scene (Fig. 2.3).

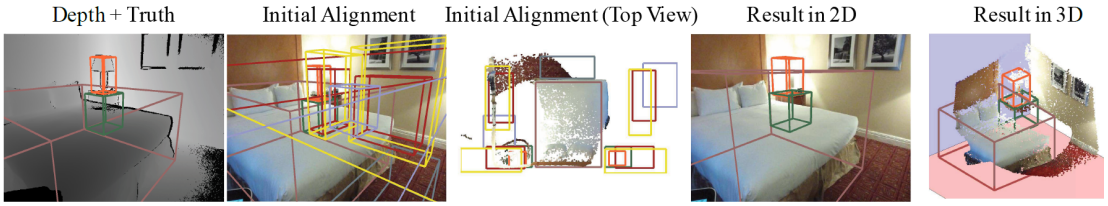


Figure 2.3: 3D object detection considering the scene context through Deep Learning architecture [195]. The method reasons the presence of objects in the image based on 3D scene templates.

Likewise, taking advantage of deep learning techniques, the approach of Kendall et al. [7] proposes the Bayesian SegNet, a semantic segmentation framework for scene understanding.

The approach consists of a deep fully convolutional neural network with a pixel-wise classification layer. The framework is applied to roads and indoor environments (Fig. 2.4).

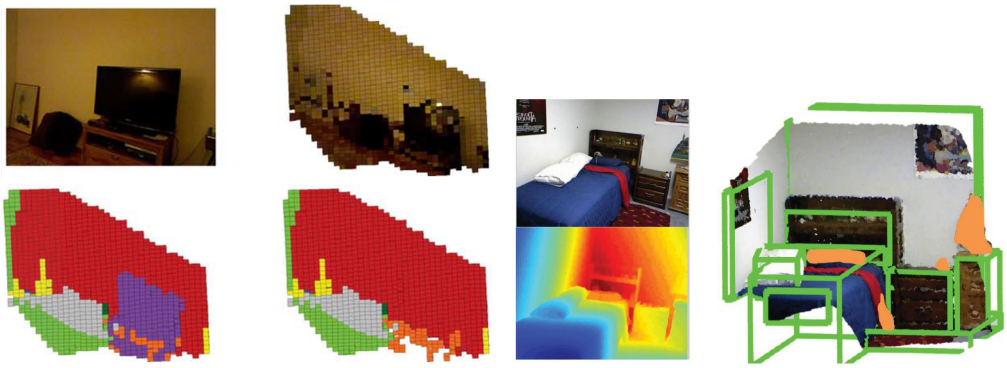
Figure 2.4: Pixel-wise segmentation through SegNet [7] applied to indoor environments from the SUN-RGB-D dataset[156].

Other approaches intend to comprehend the environment through 3D reconstruction. The approach by Kim et al. [91] proposes a method for 3D scene reconstruction and object segmentation based on a Voxel Conditional Random Fields (V-CRF). The model outputs an estimation of a dense voxel with a semantic label associated with the idea to reconstruct the structure of the scene in a 3D perspective. Some results of this approach are shown in Fig. 2.5(a). Likewise, in the approach of Khan et al. [89] layout estimation and 3D object identification through cuboids in cluttered indoor environments are jointly addressed. Fig. 2.5(b) shows some results of the method. Input RGB-D images and CRF are used to integrate geometric and local appearance features and the relationships between scene elements.

Similarly, Lin et al.[107] present a scene understanding framework based on the generation of cuboids from RGBD images. First, candidate regions are built with Constrained Parametric Min-Cut (CPMC). Then, 3D cuboids are generated for each region. Through CRF, the contextual information, appearance, and geometry are integrated to assign an object class to each cuboid and a scene label to the image.

2.2.2 Scene Understanding considering the Whole Environment

Some approaches deal with the scene understanding problem considering the whole environment. These approaches are more focused on scene recognition and semantic mapping. The approach by Zheng et al. [200] proposes a scene comprehension of unknown environments through RGB-D reconstruction and semantic segmentation. The objects are identified and segmented from RGB-D images. Then, through deep learning techniques, a



(a) (b)

Figure 2.5: (a) Scene reconstruction with the approach by [91]. An RGB image is reconstructed with V-CRF. Also, semantic labels are assign to different parts in the image, even when some objects are removed.(b) 3D Cuboid estimation (green) and identification of cluttered regions in an RGB image (orange) obtained through the method presented in [89].

semantic segmentation network is fed with 2D and 3D features to perform online semantic labeling based on voxelization. The experiments are done in simulated indoor environments as shown in Fig. 2.6.

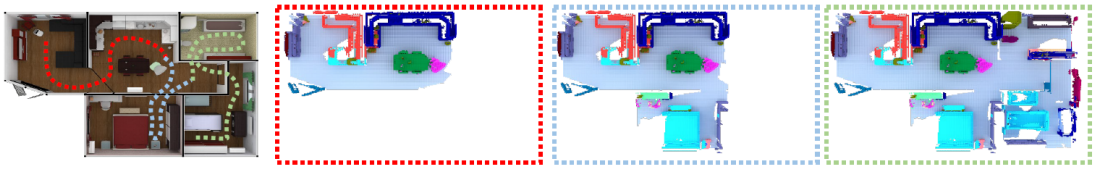


Figure 2.6: Results of applying the approach of [200] in an indoor environment. The method performs online volumetric RGBD reconstruction and semantic labeling.

In the work of Jian et al. [84] two main problems are considered, scene arrangement and 3D scene labeling. They propose a model of the environment, including object and scene relations with human poses and their interactions with objects. Through a Latent Conditional Random Field, object-object and human-object relations are merged to infer the objects' label in the scene and the scene arrangement. That implies the 3D location of each object in an RGB-D image.

Furthermore, a 3D hierarchical spatial representation of a complete environment is proposed in the approach of Rosino et al. [147]. The model includes object detection, human identification, and pose estimation. The environment is modeled as scene graphs. The graph comprises several nodes that correspond with a spatial concept as objects, places,

structures, and their respective spatial-temporal relations. The nodes are grouped into layers according to different abstraction levels of a scene.

The approach by Sunderhauf et al. [160] develops a framework for place categorization and semantic mapping through convolutional neural networks. New semantic classes are learned through the implementation of supervised one-vs-all classifiers. Prior information is embedding into the classification process through a Bayesian filter to guarantee temporal coherence. The output of the system is a map that uses the results of the scene labels to create a semantic map of the environment. Fig. 2.7(a).

Figure 2.7: (a) Semantic map of an office environment created through the approach presented in [160]. Coloured areas represents different semantic categories for the environment. (b) The resulting semantic map of the approach of [158].

The work by Sun et al. [158] addresses the scene recognition and object detection tasks jointly. A map of the environment shown in Fig. 2.7(b) is built through the SLAM algorithm creating an occupancy grid map. Each cell is labeled with a scene category. A convolutional neural network predicts in parallel the scene label and the object locations in the image. The scene is categorized based on global features while the objects are detected by region proposal-based techniques leveraging scene information.

2.3 Discussion

Scene understanding is a complex task that involves different robotic sub-tasks. As we can observe in the literature, scene understanding has been tackled depending on the purpose. In this way, different approaches exploit more tasks than others. In this thesis, we address scene understanding by properly combining some sub-tasks to provide a robot with the capabilities to know and interact with the environment. We consider essential the knowledge of objects present in the environment, the identification of the scene, the relationship between objects and scenes, the boundaries of the scene in the whole environment that helps the robot to decide how to move to perform different tasks. In the following chapters, we will propose (i) different methods to improve object and scene recognition indoor domestic environments, (ii) a method for a new division of the environment by the identification of meaningful regions of interest; we will also demonstrate how this new representation can boost other complex robotic tasks, and (iii) some strategies to improve object search tasks by combining semantic information of the environment to demonstrate how a complete knowledge of the environment can improve the performance of a high-level task.

3

Object Recognition for Mobile Robots

Full knowledge of the environment is a crucial condition for autonomous robots operating in human scenarios. This involves understanding the surrounding elements and the relationships between them. Objects are a rich source of information for performing many robotic tasks. Visual perception is a skill for mobile robots that entails identifying object categories and their position in space, among other information that can be extracted from each object. One thing to note is that vision systems have to deal with uncertainty due to sensor noise and unclear or inaccurate prior knowledge. This chapter explains the methods and algorithms implemented in this thesis to perform object recognition in human indoor environments.

3.1 Introduction

Autonomous robots need to be provided with a set of capabilities that allows them to move and interact in real environments. Among all of the skills needed, perception constitutes one of the cornerstones. The term *perception* refers to, among other things, sensory awareness. From the five different senses that humans have, vision is arguably the most important for safely moving and interacting with the world. Object detection is an important research topic in computer vision due to its wide range of applications [78]. Significant advances have been made in the past decades, especially since the work by Viola et al. [172]. In 2001, they proposed the first facial detector that could operate in real-time without restrictions.

The detection and identification of objects is a relevant task to guarantee the safe movement execution of autonomous robots, providing semantic information of a scene through understanding its visual appearance and not just the geometrical information about it that is unavailable to other sensors. Through this information, an autonomous robot should be able to learn and maintain a visual model of the environment.

The main goal of object detection is to provide methods and models able to answer these two questions: *What objects are there in an environment?* and *Where are these objects located?*. Gathering object information represents one of the most valuable and useful inputs for many robotic tasks. Then, more complex thoughts could arise, such as, *What can I do with these objects?*, *What are the relationships between them?*, *How do these objects influence the environment?*. In this chapter, we will address the two first questions. Then, in the following chapters, we will try to answer the rest of them.

In general, the techniques and methods to identify objects depend on the scope of application. On the one hand, there are methods developed for the industry, mainly based on robotic arms, that need a high accuracy to perform grasping and manipulation tasks. The environmental conditions are in most of the cases controlled, i.e., illumination conditions, organization of the environment, etc. On the other hand, when the objects have to be identified in human environments such as offices, typical houses, restaurants, the problem becomes totally different. Viewpoint variations, occlusions, illumination changes, background clutter, and sensor noise make the recognition process more challenging.

The work presented in this chapter is focused on object recognition methods applied to mobile robots operating in human indoor environments. These environments have not been modified in terms of lighting conditions, objects location, occlusions, etc. In our developments, we keep in mind that object detection must be deployed in real-world scenarios, considering the dynamism of the scene, with a robotic platform moving continuously and taking into account the noise of the sensors. Our main objective here is to provide the robot with different systems that allow it to gather as much environmental information as possible by detecting objects. That is why we start with implementations based on machine learning techniques and finish with more modern strategies based on deep learning techniques.

Initially, we present an object recognition system based on machine learning, published in [71], to identify common objects such as closets, chairs, and screens in domestic environments. The detected objects generate some events that are then used by a topological navigation system. Besides, an approach for adding uncertainty estimation to the object recognition model is presented. The idea is to have a more accurate and complete probabilistic model, allowing to improve and strengthen the navigation systems on mobile robots. Then, an object recognition model based on deep learning techniques is described. The model takes advantage of these technologies yielding a more robust and general detection

model and recognizing more objects in indoor environments.

The structure of this chapter is as follows: in Section 3.2, related work regarding object recognition for mobile robots is described. Section 3.3 presents a general overview of the object recognition system based on machine learning. The proposed uncertainty estimation model is explained in Section 3.3.1. Then, Section 3.4 explains the deep learning-based method implemented in this thesis. Finally, in Sections 3.5 and 3.6 the experimental results and the conclusions of this chapter are presented.

3.2 Related Work

Object detection forms one of the foundations for other important robotic tasks such as instant segmentation [25, 59, 60], object tracking [86, 126], autonomous driving [19, 37, 182], and scene understanding [7, 20, 107, 205], the latter being the main objective of this thesis. Recognition of objects in real-world scenarios constitutes a complex task due to the environmental conditions that must be considered. Some object detection challenges are illumination variations, different viewpoints, occlusions, scale variations, speed of detection, and object localization. In this section, we review from traditional techniques up to the latest advances in object detection.

The last two decades have been marked by great advances in terms of object detection, moving from traditional techniques to more robust applications based on deep learning [207]. Traditional object recognition techniques are based on exploiting features [3, 24, 73, 142]. Significant elements of an image (e.g., the color of the object, shape, and size) are extracted to identify the key aspects that define an object category regardless of variability in appearance. The most common features that can be extracted from images are: Scale Invariant Feature Transform (SIFT) [6], Speeded-Up Robust Features (SURF) [102], CENsus TRansform hISTogram (CENTRIST) [179], Binary Robust Independent Elementary Features (BRIEF) [17] and Oriented Fast and Rotated BRIEF (ORB) [149].

The chosen features are then learned through a classification model to determine the class to which an object belongs. In the approach of Astua et al. [3] SURF features are extracted to detect objects in indoor scenes. Ramisa et al.[142] focus on specific features, such as face detection, using different methods based on global and local descriptors. The approach by Csurka et al. [24] applies techniques based on Bag of Words during the classification process. Later, Hernandez-Lopez et al. [73] propose a method for object detection applied to indoor environments through the use of CIE-Lab color space and depth information to process the images from the Kinect sensor. In this way, a visual strategy capable of working in real-time is proposed.

Other approaches such as those presented by Felzenszwalb et al. [36] and Malisiewicz et al. [116] implement the Histogram of Oriented Gradients (HOG) descriptor for object

detection [26]. This method overcomes the scale-variance problem by allowing the detection of a variety of object categories. However, it is primarily used to detect pedestrians. As an extension of HOG detectors, the Deformable Part-based Model (DPM) has been implemented based on two main stages: root-filters to detect parts of the objects and a set of part-filters to assemble the different detected parts of the object. This method originally proposed by Felzenszwalb et al. [35] has been improved by Girshick, et al. [47] at it can work in real-world environments with larger variations.

Furthermore, some object detectors are based on classification algorithms such as Nearest Neighbor classifier (NN) [127], neural networks, and different versions of the AdaBoost algorithm [39]. The work by Meera et al.[120] applies a k-NN classifier with SIFT features to detect objects in an image dataset. Similarly, Gupta et al. [57] combine SIFT and ORB features for generic object detection. The best robust features are then integrated into a k-NN, random forest, and decision tree classifiers.

Other works employ machine learning techniques [136] for object detection. Machine learning algorithms are organized into taxonomy, taking into account the desired result. Common algorithms are based on supervised learning. In this case, a function is generated that maps the inputs to the desired outcomes. Supervised learning can be used for classification challenges: the learner needs to assimilate information to create a model capable of separating the input data into several classes. A useful technique for data classification is Support Vector Machines (SVM). The approach of Gupta et al. [56] uses RGB-D images as input to a Convolutional Neural Network (CNN) to extract rich deep features that then are fed into a linear SVM to predict objects in an image. Lim et al. [106] combine SVM with other methods to solve classification problems in 3D environments. Furthermore, the region-based system introduced by Girshick et al. [46] uses a convolutional neural network combined with a linear SVM.

More recent approaches make use of the revival of deep learning. Through CNNs, high-level feature representations of an image can be learned. The approach by Girshick et al. [45] proposes a Region-based Convolutional Neural Network (R-CNN) for object detection. The process consists of generating some object proposals that are then incorporated into a CNN model to extract the object features. Then, SVM classifiers are implemented to infer an object category within each region in the image. The work by He et al. [63] presents the Spatial Pyramid Pooling Networks (SPPNet). Other object detectors also based on region proposals are Fast-RCNN [44] and Faster-RCNN [144]. For example, the approach of Lin et al. [108] proposes the Feature Pyramid Networks (FPN) based on Faster-RCNN.

Moreover, most of these approaches have been tested on well-known datasets such as ImageNet [151], MS COCO [109], and the datasets of PASCAL VOC Challenges [32] with many categories of objects, generating very high confidence results. On the other hand, despite the accelerated development of deep learning-based object detectors, there

are still open problems to be solved [198]. The detection of small objects in cluttered environments, occlusions, scale variations of the objects, spatial correlation, and contextual modeling aspects should be considered to build more robust object detectors. Furthermore, when the methods are implemented in a real robot that moves through the environment, other additional issues pop up. For instance, the robot's movement affects the quality of the data gathered, the speed and characteristics of the terrain can cause problems during the prediction. Also, illumination changes and occlusions can make the detection task difficult. If we add to this that other systems are running at the same time, such as navigation systems, planning algorithms, sensors drivers, among others, the performance of an object recognition system can be compromised, especially in methods based on deep learning.

With all of the above in mind, our motivation in the field of object detection is to provide a real robot with a detection system that is fast, precise, and capable of operating in real-world environments. We leverage the benefits from traditional techniques and more modern strategies to provide the robot with as much information as possible. This information is useful for a complete understanding of the environment and, consequently, helps the robot to perform other robotic tasks more efficiently. Hence, our first implementation is a probabilistic object recognition model based on machine learning that includes an uncertainty model in the system. Then, we implement a more robust alternative based on deep learning techniques. It is worth mentioning that the approaches developed in this chapter will be used in the following works presented in this thesis.

3.3 Object Recognition Model based on Machine Learning

To capture information about existing objects in a given indoor environment, we have developed an object recognition model based on machine learning. The classification algorithm applied has been Support Vector Machines (SVM). A representation of the model can be seen in Fig. 3.1. The model has been integrated into a real mobile robot and is able to capture real-time images from indoor environments to locate the objects present in each frame.

The object recognition model is composed of three main stages:

- Training: the objective of this stage is to find a model for the best object classification. This offline stage includes the creation of the dataset for training, a preprocessing step by applying some techniques such as morphological operations, equalization, thresholding, and Gaussian filter. Also, we have applied some segmentation techniques based on regions of interest and region growing. The feature extraction is done from shape features and Bag of Words with SURF features. As a result, an object database and the training model are generated.

- Retrieval: during this stage, RGB-D images are captured in real-time and sent to the preprocessing step, which is similar to that applied in the training stage. Subsequently, segmentation methods and feature extraction techniques are applied to each image.
- Classification: we apply SVM as a classification algorithm, which takes into account the training data and the test matrices created from the extracted features. Besides, at this stage, uncertainties are calculated to predict the detection probability of each class of object. The uncertainty model considers the model accuracy, the SVM results after the prediction, and the empirical probability of detection for each object.

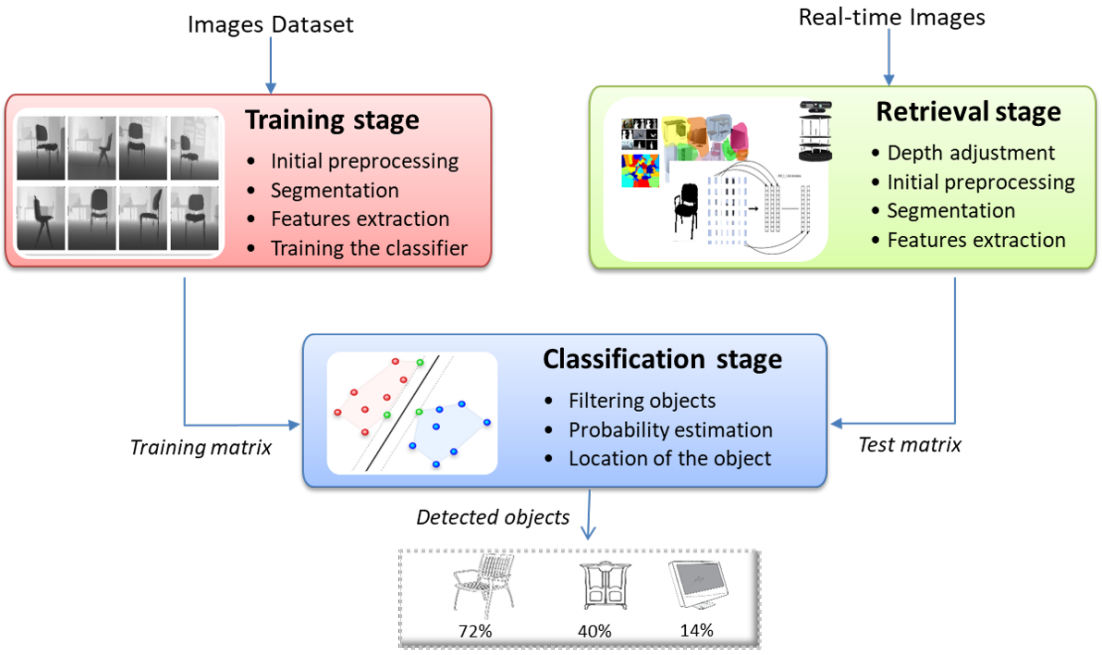


Figure 3.1: An overview of our object recognition system based on machine learning. The inputs to the model are RGB and depth images from indoor environments that are subjected to different preprocessing and feature extraction techniques to predict object categories.

Finally, our method locates the object detected in the image. To do so, the distance and angle from the object to the center of the camera are calculated. The center of mass of the detected object and its depth information is used to compute distance and angle values. The results are encapsulated and processed, so they are available to other systems, such as navigation systems and interaction modules, among others.

3.3.1 Uncertainty Model

The uncertainty in the result of a measurement, or the uncertainty of the measured value of the specific quantity subject to measurement (also called the measurand), expresses doubt about how well the result represents the true value of the measurand [11].

To estimate the uncertainties in the proposed object detection model, we have considered three factors. First, the model accuracy that has been calculated from confusion matrices in preliminary experiments in a previous work [70]. The second factor is based on the outcomes of the SVMs after the prediction process. Finally, the third factor corresponds to the empirical probability of each object as a function of distance.

Measurement Imprecision Modeling

One of the factors that affect the accuracy of an object recognition model is distance. To measure the influence of distance on the detection process, we calculate the empirical probability of each detected object [121]. This probability, also called experimental probability, can be defined as the ratio of the number of times a particular event occurs to the total number of trials. The empirical probability $p(o_i)$ is calculated as follows:

$$p(o_i) = \frac{n(o_i)}{N}, \tag{3.1}$$

where $n(o_i)$ is the number of times an object o of category i is detected, and N is the number of samples for each object. The detection frequency of each object has been determined taken into account different distances.

Then, we calculate the Pearson correlation coefficient [79] to measure the strength of the association between two variables of our uncertainty model. This coefficient is a measure of the linear dependence (correlation) between two variables A and B that allows determining how strong is the relationship between them. The correlation coefficient $\rho_{A,B}$ is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\rho_{A,B} = \frac{\sigma_{AB}}{\sigma_A \sigma_B}, \tag{3.2}$$

where A represents the distance between the center of the camera and the detected object, and B is the empirical probability $p(o_i)$ of each detected object, that is, the times an object is detected at a measured distance. In addition, σ_{AB} is the covariance and σ_A and σ_B are the standard deviations of A and B respectively.

The Pearson's coefficient can also be expressed in terms of mean and expectation as follows:

$$\rho_{A,B} = \frac{E[(A - \mu_A)(B - \mu_B)]}{\sigma_A \sigma_B}, \tag{3.3}$$

where E is the expectation and μ_A and μ_B are the mean of A and B respectively.

Once the relationship between the variables is known through the linear correlation coefficient and the regression lines have been calculated, this information can be used to predict the final probabilities of the object recognition model.

Uncertainty Estimation

As mentioned above, the uncertainty model considers three factors: the model accuracy, the results obtained from the SVMs after the prediction process, and the empirical probability of each detected object.

The model accuracy (a_{cc}) represents the proportion of the total number of predictions that are correct. This metric is based on the concept of confusion matrix [173] and is calculated as follows:

$$a_{cc} = \frac{TP + TN}{T_p}\,, \tag{3.4}$$

where TP corresponds to the true positive detections, that is, how often the model classifies an object correctly; TN represents the true negative detections, the times the model classifies correctly a negative prediction, and T_p are all detections.

Regarding the second factor, in the proposed object detection model, SVMs are used as a classification algorithm. We have implemented one SVM for each object to detect. The prediction process is based on the one-against-all approach [112]. The outcomes of each SVM are represented by numbers between 1 and N. The final result of the model denoted as p_c is obtained by applying a filter based on conditional statements and a rule based on the accuracy values. This value represents the occurrence probability of each object considering the combination of SVMs in the detection model. Then, taking into account the dependency relationships between the three factors, the uncertainty estimate can be modeled as the detection probability p_d of each object o_i as follows:

$$p_d(o_i) = \frac{a_{cc}(o_i) * p(o_i) + p_c(o_i)}{\sum_{i=1}^{n}[a_{cc}(o_i) * p(o_i) + p_c(o_i)]}. \tag{3.5}$$

Then, the final detection probability of the object recognition system (p_{sys}), which must be equal to 1 is defined by:

$$p_{sys} = \sum_{i}^{n} p_d(o_i) = \sum_{i}^{n} \frac{a_{cc}(o_i) * p(o_i) + p_c(o_i)}{\sum_{i=1}^{n}[a_{cc}(o_i) * p(o_i) + p_c(o_i)]} = 1\,. \tag{3.6}$$

Finally, the recognition system generates a ROS message [140] containing all possible perceptions and their corresponding probability of detection. The structure of the message

consists of the distance from the center of the camera to the object, the orientation angle, the name of the object category, and the probabilities of detection of each object. This data provides the robot with semantic information that can be useful for higher-level reasoning and navigation tasks.

3.4 Deep Learning-based Object Detection Model

Taking advantage of Convolutional Neural Networks (CNN), which have become leaders in efficient object detection in recent years, we have decided to implement an object detector based on deep learning techniques. For the identification of multiple objects, we have implemented a region proposal based on the CNN architecture. The model first selects regions in a single image using a proposal method and then classifies each proposal into different object classes. Specifically, we have applied the Faster-RCNN model for object detection [144]. Fig. 3.2 illustrates the model that consists of two stages: a Region Proposal Network (RPN) and a Fast-RCNN object detector [44].

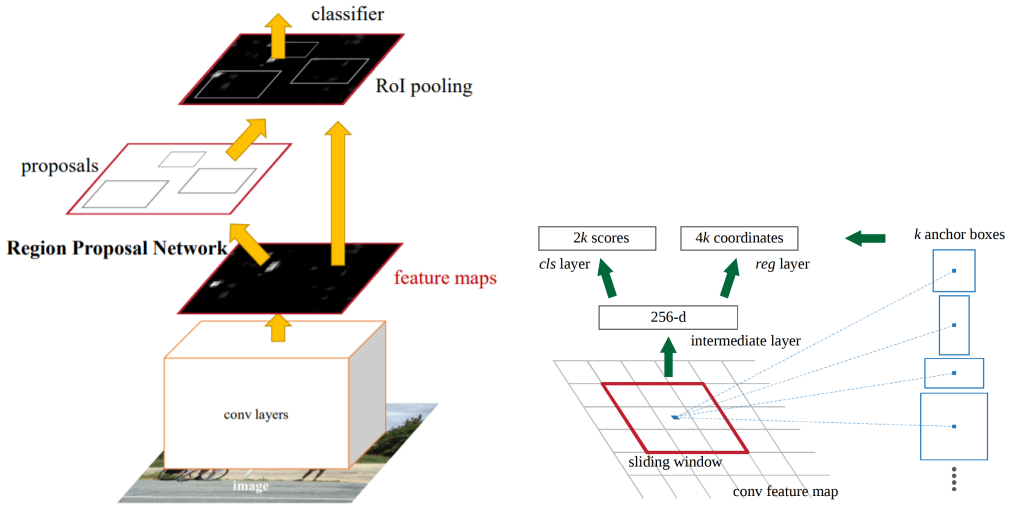


Figure 3.2: Faster R-CNN model [144]. (a) An input image is passed to the ConvNet to generate feature maps. (b) RPN architecture is then applied to obtain object proposals that are passed to a fully connected layer to predict the bounding boxes of the objects in the image.

In the first stage, feature maps are extracted from an input image using ConvNet. The resulting maps are passed through an RPN that generates region proposals in the image using the selective search (SS) method [168]. Each region is represented by a rectangular object

proposal with an objectness score. The RPN architecture shown in Fig. 3.2(b) ranks region boxes or anchors and proposes those with the highest probability of containing objects. The network slides over the convolutional feature map and fully connects to an n×n spatial window. Each sliding window generates a low-dimensional vector that feeds two sibling fully connected (FC) layers: box-regression layer (reg) and box-classification layer (cls).

The second stage receives the proposed regions with different sizes, which implies different sized CNN feature maps. Region of Interest Pooling layer (ROI) is applied to reduce the proposals to the same size. Finally, these proposed regions are analyzed by a classifier and a regressor in order to predict the bounding boxes for each object category.

The object detection model implemented in this thesis uses the feature extractor ResNet-101[64] that has been trained with the MS COCO dataset[110]. When objects are detected, the model outputs an array with information about the position of each object in the image. This array contains four numbers corresponding to the rectangle box around each object [top, left, bottom, right]. These values are the x-y coordinates at the upper-left and lower-right corners of the rectangle.

3.5 Experimental Evaluation

This section evaluates the object detection models presented in this chapter, the first based on machine learning and the second based on deep learning techniques. We are particularly interested in assessing the performance of the models in real-world environments. That is why the experiments have been carried out considering real robotic platforms and human indoor environments.

3.5.1 Evaluation of the Machine Learning-based Approach

The objective of this experiment is to demonstrate the feasibility and usefulness of incorporating uncertainties in an object detection model based on machine learning, considering different sources of information. In addition, we evaluate the performance of the method in real-time by integrating it with a navigation system.

Experimental Setup

The object detection model based on machine learning has been integrated into two mobile robots, a mobile robotic platform developed in our laboratory [49] and a TurtleBot 2 platform. Both robots are equipped with an ASUS Xtion Pro Live camera for capturing RGB-D images and a Hokuyo URG-04LX-UG01 laser scanner for navigation purposes. The environment chosen for the tests is a laboratory at the University Carlos III of Madrid. The object detection model has been developed under Robot Operating System (ROS) to

guarantee hardware and software integration. C++ programming language and OpenCV libraries [192] were used for image processing. For the tests, three common objects in indoor human environments have been selected: chairs, screens, and closets. We start by showing the results obtained during the calculation of the uncertainty model and applying the method to the chosen real-world scenario.

Uncertainty model results

The proposed uncertainty model is based on three factors. First, the empirical probability of each selected object was determined considering different distances from the center of the camera to each object. In total, 275 samples for closets, 300 for chairs, and 325 for screens were obtained. Table 3.1 shows the results of the empirical probability calculation.

Table 3.1: Empirical probability for the three common objects in indoor environments.

Distance (m)	$p(o_{closet})$	$p(o_{chair})$	$p(o_{screen})$
0.25	0.01	0.01	0.01
0.50	0.01	0.99	0.01
0.75	0.01	0.96	0.01
1.00	0.76	0.96	0.72
1.25	0.80	0.92	0.72
1.50	0.96	0.92	0.68
1.75	0.92	0.72	0.60
2.00	0.80	0.52	0.56
2.25	0.72	0.04	0.48
2.50	0.68	0.01	0.16
2.75	0.64	0.01	0.04
3.00	0.64	0.01	0.04
3.25	0.01	0.01	0.01
3.50	0.64	0.01	0.01
3.75	0.42	0.01	0.01
4.00	0.32	0.01	0.01
4.25	0.16	0.01	0.01
4.50	0.16	0.01	0.01
4.75	0.10	0.01	0.01
5.00	0.08	0.01	0.01

Then, we have obtained the following correlation coefficients: -0.929 for closets, -0.849 for chairs, and -0.944 for screens. As we can see, all coefficients are less than 0, which means a negative correlation. The results indicate a strong negative correlation between the two variables called inverse relation: when one increases, the other decreases in constant proportion. Fig. 3.3 shows the graphical representation of the relationship between the two variables, the distance, and the empirical probability of each object. The scatter diagrams show the inverse relationship between variables. Once the relationship is determined, a linear regression can be employed to obtain best-fit curves.

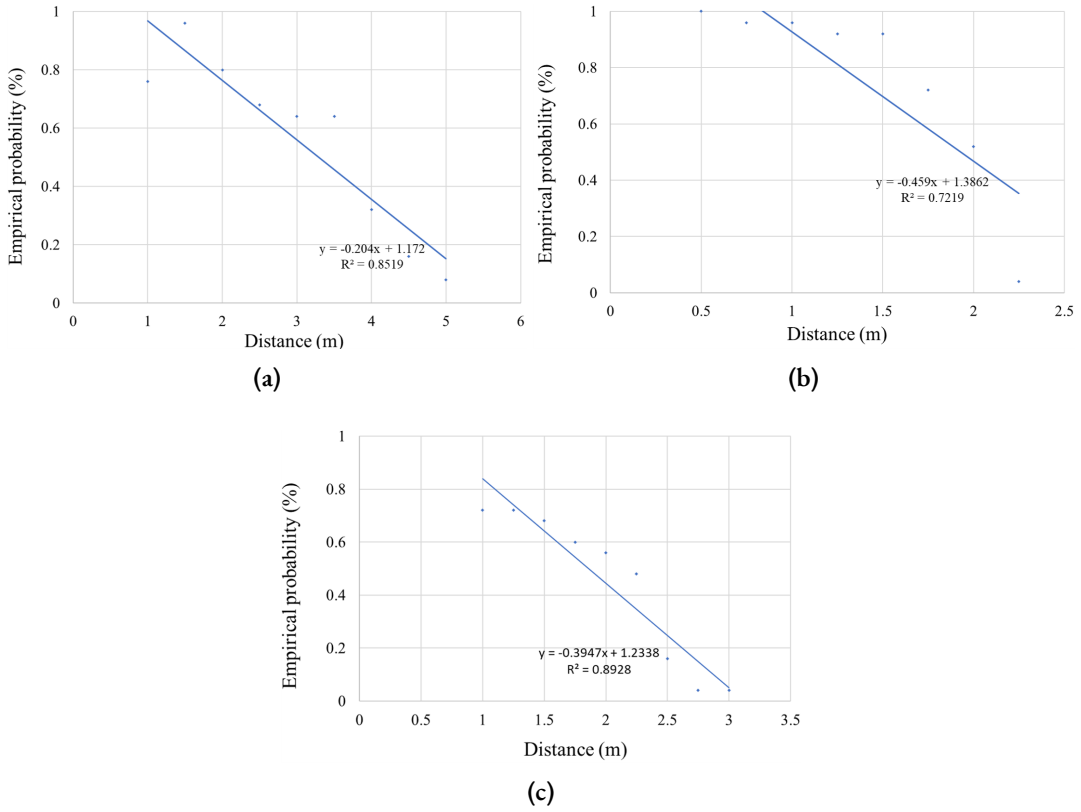


Figure 3.3: Scatter diagrams for the selected objects: (a) closets, (b) chairs and (c) screens. The negative correlation indicates a high dependence between the two variables: distance and empirical probability.

Regarding the second factor, Table 3.2 shows the model accuracy obtained after the detection process. The accuracy results are 78.07% for closets, 71.45% for chairs, and 65.47% for screens with the lowest value.

For the third factor, the occurrence probability of each object, we implement one SVM

Table 3.2: Model accuracy results of the object recognition system.

Evaluations	Closets	Chairs	Screens
Model accuracy (a_{cc})	0.781	0.715	0.655

for each chosen object. Each SVM generates different results after the prediction process. Table 3.3 shows the combinations of the outputs of the SVMs used in this approach. Then, the final occurrence probability for each object is shown in Table 3.4.

Table 3.3: Results of the SVMs after the prediction process.

Case	1	2	3	4	5	6	7	8
SVM_1	1	1	1	1	0	0	0	0
SVM_2	1	1	0	0	1	1	0	0
SVM_3	1	0	1	0	1	0	1	0
Result	Closet	Closet	Closet	Closet	Chair	Chair	Screen	-

Table 3.4: Occurrence probability for each chosen object based on the SVM results.

Objects	p_c
Closets	0.5
Chairs	0.25
Screens	0.125
Undefined	0.125

Implementation in Real-World Environments

The main goal of the following experiments is to demonstrate the usefulness and feasibility of our object recognition model in real-world scenarios. All experiments have been performed in real-time. The first part of the experiment consists of detecting the selected objects: chairs, screens, and closets using a mobile platform. In this case, we used a TurtleBot 2 that moves through the environment by teleoperation while the detection system runs in real-time. Fig. 3.4 shows the results of the detection model in a laboratory environment.

Figure 3.4: Results of the object detection model applying uncertainties. The experiment has been carried out in a laboratory at the University Carlos III of Madrid. The label of the detected object is displayed in each image frame.

For each detected object, the predictions about the probability that the object belongs to each category are shown in red (closets), green (chairs), and yellow (screens). The final decision about the object is based on the object category with the highest detection probability. In each image frame, the definitive label for the object appears in the center of the object.

A quantitative evaluation of the method has been carried out from confusion matrices, considering the following basic terms: the true positives predictions (*TP*), the true negative predictions (*TN*), false positive predictions (*FP*), and false negative predictions (*FN*). Then, we select the following metrics:

- Model accuracy: the percentage of correctly classified objects.

$$\text{Model accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.7}$$

- Misclassification rate: also called error rate, indicates how often the model performs incorrect classifications.

$$\text{Misclassification} = \frac{FP + FN}{TP + TN + FP + FN}. \tag{3.8}$$

- Sensitivity: also known as the true positive rate, is calculated as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \tag{3.9}$$

- Specificity: also called true negative rate is computed as:

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{3.10}$$

Table 3.5 shows the results of the evaluation of the proposed model during the detection of the three object categories: chairs, screens, and closets.

Table 3.5: Evaluation of the proposed object detection model based on machine learning.

Evaluations	Closets	Chairs	Screens
Model Accuracy	0.854	0.731	0.766
Misclassification	0.146	0.269	0.234
Sensibility	0.708	0.619%	0.636
Specificity	0.100	0.842	0.895

The results show that the highest model accuracy is for closets, with 85.42% followed by screens with 76.56% and chairs with 73.06%. The highest misclassification rate is for chairs with 26.94% and screens with 23.44%. Sometimes, if the robot perceives the chair back, the system is unable to detect it. In the case of closets, when the robot is more than three meters away from the object, it cannot be detected. The experiments carried out show the feasibility and usefulness of implementing an object detection system that incorporates uncertainty.

The second part of the experiment consists of integrating the proposed detection model with the topological navigation system developed by Gomez et al. [50] to evaluate the real-time functioning of both systems together. The navigation system is categorized as a topological representation based on movements; this means that the relations between nodes have no geometrical meaning. In this integration, the information perceived by the detection model is processed, and some objects in the room are detected. Navigation is structured according to the position of these objects. The navigation system receives the detected object's information: name, distance, orientation angle, and detection probability. If this

information corresponds to the desired one for the event, the robot moves towards the object. For this test, the mobile platform TurtleBot 2 has been used. Fig. 3.5 shows the results of this experiment. As we can observe, the integration process has been completed successfully, and the navigation objectives have been achieved.

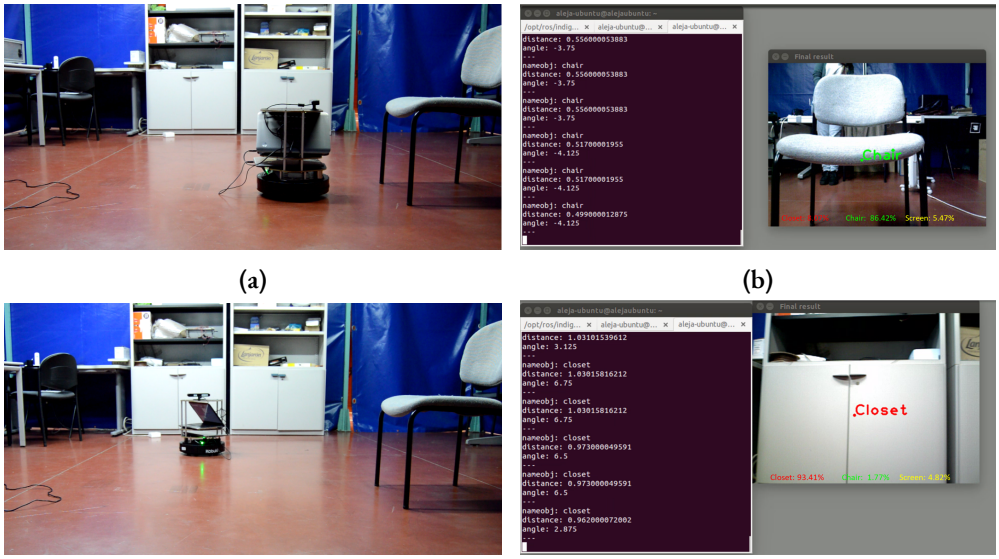


Figure 3.5: Results of the integration of both systems into a TurtleBot 2 platform. (a) The robot in front of the first target object; (b) the point of view of the robot; (c) the robot in front of the second target object; (d) the view of the robot during the detection of an object of category closet.

3.5.2 Evaluation of the Deep Learning-based Approach

This test aims to evaluate the performance of the object detection model based on deep learning in real-world environments. This method has already been tested on a wide variety of datasets, including PASCAL VOC Challenge [33] and Microsoft COCO [110] with tremendous results in static images. Conversely, we focus on evaluating the method performance while a real mobile platform moves through a home environment considering real environmental conditions.

Experimental Setup

The experiments have been conducted on a TurtleBot 2, equipped with a RealSense Dense camera D435i to capture RGB images. The real-world environments selected correspond

to two typical houses of 45 m^2 and 75 m^2 respectively. We have implemented the open-source framework TensorFlow Object Detection API [77] which has been developed to build, train and implement object detection models. The object detection model based on Faster- RCNN has been adapted to interact with other systems through the middleware ROS. In this way, the detector is prepared to be integrated with the rest of the algorithms and methods developed in this thesis. The model runs in real-time on an Intel Core i7-8750H CPU with an NVIDIA GeForce GTX 1060 GPU. For robot movement, we implemented the Adaptive Monte-Carlo Localization (AMCL) algorithm.

Implementation in Real-World Environments

The TensorFlow Object Detection API allows using different pre-trained models. Table 3.6 shows some models trained with the COCO data set. The COCO mAP column represents the accuracy of the model. Higher numbers mean better accuracy for the model. However, in these models, an increase in speed generates a decrease in accuracy.

Table 3.6: Detection models pre-trained on the COCO dataset. The speed model indicates the running time in milliseconds (ms).

Architecture	Feature extractor	Speed (ms)	COCO mAP
SSD	mobilenet v1	30	21
SSD	inception v2	42	24
R-FCN	Resnet-101	92	30
Faster R-CNN	Resnet-101	106	32
Faster R-CNN	Inception-Resnet v2	620	37

Considering the results above, we have selected the model architecture Faster R-CNN with the feature extractor Resnet-101, which allows us to identify 80 different object categories. The test consists in identifying the existing objects present in a real home environment. To do so, the model evaluates each RGB image frame captured by the sensor while the robot moves through the environment. Fig. 3.6 shows some results during the detection process in real-time.

For this experiment, a total of 540 images were considered. Each RGB image frame of $640x480$ pixels is provided to the model that outputs a list with the detected objects, the coordinates of the bounding box that contains each object, and a score that indicates the detection confidence of the underlying object category. This value goes from 0 to 1, where the closer the value to 1, the more confident the model is. It is worth mentioning that a cut-off threshold can be set to discard detection results depending on the application.

Figure 3.6: Results of object detection model based on deep learning techniques. Each detected object is enclosed in a bounding box with the detection confidence of the underlying object category.

To measure the model's effectiveness, we adopted the following performance metrics: precision, recall, and F1-score. Table 3.7 shows the results of the evaluation of the 20 objects detected in the house environment.

- Recall: is a ratio that allows us to know how much the detection model correctly predicted. It indicates how well the model works at choosing the correct elements. The higher the value, the better is the result. It is calculated as:

$$\text{Recall} = \frac{TP}{TP + FP}. \tag{3.11}$$

- Precision: indicates from all the classes the model correctly predicted , how many are actually positive. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FN}. \tag{3.12}$$

- F1-score: is a measure that combines precision and recall in a single value by calculating the harmonic mean of precision and recall. It is a way to find the balance between these metrics and measure the accuracy of the model. It is obtained by:

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \tag{3.13}$$

Table 3.7: Evaluation of the deep learning-based object detection model in real-world environments.

Objects	Precision	Recall	F1-score
couch	0.793	1.000	0.885
tv	0.686	0.857	0.762
book	1.000	0.667	0.800
vase	0.458	1.000	0.629
chair	0.735	0.947	0.828
bottle	0.833	0.833	0.833
teddy bear	1.000	0.714	0.833
laptop	0.444	0.571	0.500
remote	0.500	0.375	0.429
dining table	0.700	0.538	0.609
keyboard	0.667	0.857	0.750
bed	0.852	0.920	0.885
cup	0.645	1.000	0.784
mouse	1.000	0.714	0.833
sink	0.600	0.450	0.514
refrigerator	0.682	0.882	0.769
backpack	1.000	0.333	0.500
microwave	0.818	1.000	0.900
oven	0.714	0.385	0.500
toilet	0.838	1.000	0.912
Avg. model	0.748	0.752	0.723

As can be seen, the model's precision on average exceeds 70% for most of the object categories detected. However, for some classes such as sink, vase, laptop, and remote, the

detection is a more challenging process with a precision below 60%. During the test's execution, our deep learning-based model demonstrated a good combination of speed and accuracy so required in real-world environments. The object detection model originally developed to operate with static images has been adapted to work online in real indoor environments while the robot is moving. Furthermore, the model has been integrated into a ROS node to deliver the detection results as a ROS message available to be used by the other systems developed in this thesis.

3.6 Discussion

This chapter presented two approaches for object detection, one based on machine learning techniques and the other based on deep learning strategies. The first approach implements an uncertainty estimation model to an object detection system based on machine learning. The proposed system was implemented into a real mobile robot and the tests were conducted in real-time. The main idea behind the proposed method was gathering information that represents the real world, considering that the data is not perfect and is never exactly as measured. In this way, the system allowed getting the information of the object category, the object's distance from the center of the sensor, the object's angle with respect to the center of the sensor, and the detection confidence of each object.

Second, we implemented an object detection model based on deep learning techniques to strengthen the object detection process. The model allows the detection of more objects, in this case, 80 object categories, and most importantly, it can work in real-world scenarios while the robot is moving. All this information is essential to provide a robot with enough capabilities to understand its surroundings, starting with the objects around it. The introduction of deep learning in object detection approaches has demonstrated better results than using machine learning techniques because of their robustness and capability to detect more objects in the environment. The increasing advances in the hardware imply that these approaches guarantee some crucial conditions when we work in real scenarios: the speed and the operation in real-time.

4

Scene Recognition in Indoor Environments

The ability to know where a robot is has been a relevant topic in robotics. Given an image, it is not only important to identify the objects in the scene but also the environment where the objects are and the different actions that can occur. Assigning a semantic label to a scene, i.e., kitchen, bedroom, bathroom, implies knowing the objects around and the relationship between them. In this chapter, the scene recognition models developed in this thesis are presented. First, a probabilistic scene recognition model based on the relationship between the objects is proposed. Second, a multi-classifier model to take advantage of different base scene classifiers is explained.

4.1 Introduction

Autonomous robots moving and performing different tasks in human environments is one of the most challenging problems in robotics [114]. Full knowledge of the environment is necessary to guarantee the robot's safe operation and success during the execution of a task. Vision plays a relevant role in collecting perceptual information from the environment, so the quality of the data gathered can improve robots' performance in their most important tasks such as navigation, scene recognition, manipulation, among others. A scene recognition process basically consists of a label assignation to an area of the environment considering some visual features and the spatial arrangement of the elements present in each place. These labels have semantic meaning to humans and determine the different actions in the place and

the human-robot interaction possibilities. In this thesis, a scene is defined as a part of the environment (a place) such as a living room, a bathroom, an office, among others occupied by different kinds of objects where real life occurs [67]. Fig. 4.1 shows a representation of the concept.

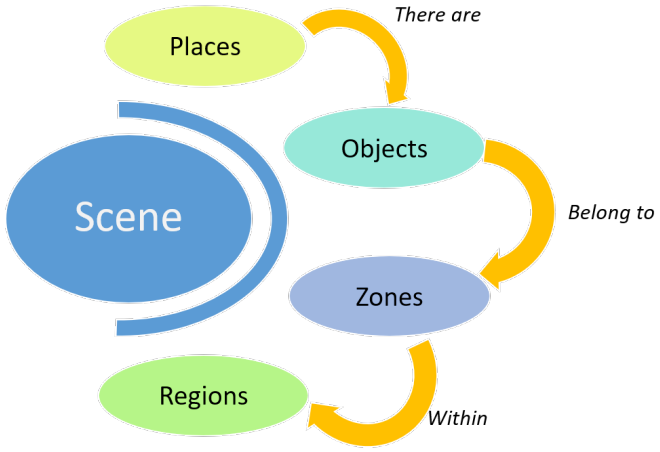


Figure 4.1: Scene concept illustration. A scene is a place occupied by objects. The objects in turn belong to a specific zone within a region. A region is an area of the environment and a place is an abstraction of it.

The categorization of a given scene according to specified categories implies analyzing geometric and semantic information and the scene elements, in this case, objects and their relationships, to develop a robust and efficient scene recognition model. Several environmental conditions, such as illumination, variability, viewpoint, ambiguity, and scale changes, can make the classification process easier or more challenging and affect the process of identifying a scene. This analysis process increases the robot's understanding of the scene, its objects, and the relationships between them. As an example, considers these two scenes: a dining room and a meeting room. In both environments, there are chairs and tables. This makes object identification not enough to distinguish the scenes. However, the relationships and arrangement of objects are vital to improve the inference about the scene category. That is why scene recognition is a more challenging problem than other image recognition tasks.

This chapter addresses the process of identifying an area in an indoor environment. Also, we intend to answer these general questions: *Where am I? What actions can I do? What are the elements with which I can interact?*. The robot has to know about: the place where it is, the elements (objects) present, and their location. Besides, the category of the existing objects and their relationships can also influence the scene categorization. To deal with this, first, a probabilistic scene categorization model is proposed. The model considers

object information to generate an uncertainty model that allows a representation of the environment as close as possible to the real world. The modeling of uncertainty includes the sensor's noise, environmental conditions, and the system's limitations and constraints. The uncertainties of the scene categorization model itself and the uncertainties from an object recognition system are incorporated. Lastly, the relationships between objects and scenes adjust the final probability result about the place where the robot is.

Moreover, we present a multi-classifier model for indoor scene recognition based on weighted voting schemes. A multi-classifier, also known as an ensemble classifier or combined classifier, aims to build a more robust model by combining the outcomes of independents base classifiers. We study two weighted voting schemes: based on the accuracy of independent classifiers and based on genetic algorithms. The weights are computed for each independent classifier to obtain a final estimation of the place the robot is perceiving according to its visual information. Genetic algorithms calculate the optimal weights assigned to individual classifiers in the multi-classifier to obtain satisfactory results. The motivation, in this case, is to profit from the advantages that different techniques can offer and merge them into a multi-classifier to obtain a more robust model for scene recognition.

The remain of the chapter is structured as follows: Section 4.2 presents the related work regarding scene recognition applied to indoor environments. Section 4.3 explains the details of the proposed probabilistic scene recognition model. Section 4.4 presents the proposed multi-classifier for scene recognition. An experimental evaluation for the two proposed methods is explained in Section 4.5. Finally, the conclusions of the chapter are presented in Section 4.6.

4.2 Related Work

Scene recognition is a field that has awakened great interest in recent decades due to the large amount of information that it gives and its influence on the interaction of robots with human beings. The knowledge about the place where a robot is can determine the possible actions to be carried out in that area. This topic still represents a big challenge, and many works intend to solve it. Some works are based on handcrafted image features algorithms [55], learning features [190] and contextual information [74]. Other approaches apply more complex learning methods such as convolutional neural networks [195]. In this section, we analyze different approaches to address the scene recognition problem.

According to feature-based algorithms, Gupta et al. [55] propose using generic and specific characteristics to address scene categorization. They consider different features: geocentric pose, geometric features, pixel depth, size and shape, and superpixels' appearance. The approach by Zhou et al. [203] apply dense extraction techniques to obtain SIFT features for all the images (grayscale and color). Then, it is implemented a bag of features approach to

generate a training model. The approach by Jianchao et al. [82] proposes a variation of the Spatial Pyramid matching method for scene recognition. For this, SIFT feature sparse codes are used as local appearance descriptors. Nicosevici et al. [131] propose a model based on visual vocabularies to determine similarities between scenes. Through this, re-visited regions are detected to reduce the errors during robot navigation and mapping. Also, the approach by Xie et al. [183] implements a framework that expands the traditional Bag of Words model [24]. They combine texture and edge-based local features, creating geometric visual phrases to model spatial context. The approaches by Khan et al. [88] and Banerji et al. [10] propose scene recognition models based on visual features such as color, texture, shape, and geometric information. This type of strategy is effective but can be easily affected by external factors such as illumination, environmental conditions, and robot movements. Despite the good results that some of these approaches report, the tests have been made with standard datasets, not evaluating them in real-world environments and with real robots in motion.

Other works incorporate object information to improve the scene recognition rate. The work by Herranz et al. [74] combines scene and object information through convolutional neural networks. A single CNN is used as a feature extractor. The network is trained using different datasets of scenes and objects considering different scale ranges. In the approach by Espinace et al. [31] common objects are associated with a class of scene considering contextual information. Then, a search strategy is implemented to look for meaningful objects. The objects found are then combined through a Naive Bayes approximation to infer the scene category. Also, the approach by Kollar et al. [93] incorporates information of the environment, including the relations between objects and scenes, to propose a probabilistic model to predict the location of novel objects. They assume classifiers and object detectors are always correct, so the classifiers' errors are not considered. In the approach by Sünderhauf et al. [160] a place categorization model is defined as a probabilistic estimation problem. This estimation is then used to boost the results of the object detectors.

More recent works try to solve the scene recognition problem through convolutional neural networks employing large image datasets to train the models. The work of Zhu et al. [204] combines RGB and depth features in a multi-modal fusion framework for scene recognition. The approach by Zhang et al. [195] proposes a CNN based on 3D information to obtain a representation of the environment. The training set consists of 3D scene templates, including furniture arrangement. In this way, context information is incorporated into the model. The approach by Nascimento et al. [129] proposes a method that combines global and local features through CNNs. The system is evaluated on standard datasets for scene recognition, reporting improvements compared to other approaches. Likewise, the work of Herranz et al. [74] combines object and place datasets for scene recognition, considering how the scale of the input patches affects the final results.

The work in this chapter uses object and scene information to improve the robot's

understanding abilities. It is considered the systems are not perfect, and for this reason, the model includes the errors of each sensor used. In this work, a probabilistic scene recognition model that considers the information of different elements of the environment is proposed. The essential part of the work is the proper and accurate modeling of the environment, which has to be as close as possible to reality. This model incorporates elements that influence to a greater or lesser degree the process of recognizing the scene. In this way, we develop a robust recognition model that allows robots to have a better knowledge of the environment to perform essential tasks such as navigation, manipulation, among others.

So far, the approaches presented corresponds to single classifiers. These approaches have to lead to decent and accurate results in some situations. However, we are still far from having a general scene classifier that is valid for all types of environments and situations. For this reason, we also propose to combine the outputs of independents classifiers to overcome the limitations and constraints of each of them and take advantage of their strengths. The purpose of a multi-classifier, also called ensemble classifier or combined classifier, is to combine heterogeneous or homogeneous independent classifiers to generate a final decision about a problem [178]. Multi-classifiers have been extensively used to improve the classification outcomes in face recognition [128, 141], texture recognition [187], emotion identification [12], and object detection [116]. However, few approaches have been proposed for the scene recognition problem.

The approach by Yan et al. [184] proposes a hierarchical SVM strategy to deal with the categorization of rare scenes. In standard datasets, a scene is considered rare when the negative class is much larger than the positive class. Each SVM is trained independently, then the outcomes of each of them are aggregated with another SVM. The approach by Gu et al. [52] proposes an ensemble of parallel deep rule-based classifiers. They use the features obtained from a pre-trained DCNN to train each classifier separately. The approach winner-takes-all is applied to the decision layer considering each classifier's result and the confidence scores. The approach by Bai et al. [8] uses a single CNN classifier. The classification process is carried out independently based on the features of different layers of the CNN. Then, the layers' outputs are combined to predict the scene category. They proposed a soft combination of the ensemble of layers of the classifier using static and dynamic weights computed through genetic algorithms. Similarly, the work of Guo et al. [53] implements a Local Convolutional Supervision (LCS) layer. This layer keeps the importance of fine-grained and detailed information in the image, improving the final outcome of the model. Unlike the approaches mentioned above, we proposed a multi-classifier that applies dynamic weighted voting schemes to fuse the outcomes of independent classifiers. Each base classifier extracts particular features from the environment to predict the scene that a robot is perceiving. We study two schemes: one based on each classifier's performance rate and another based on genetic algorithms.

4.3 Object-based Probabilistic Scene Recognition Model

We propose a scene recognition model based on objects' information and their relationships to allow real-time interaction with the human environments [68]. The model is based on machine learning, implementing Support Vector Machine (SVM) as a classification algorithm. The model is divided into two stages depending on how processes are executed (offline and online). Fig. 4.2 shows the schematic for the proposed scene recognition model. The offline stage includes all the processes that the robot performs before the final prediction process. This contributes to guarantee efficiency in the execution times and the optimal consumption of the robot's resources. The online stage consists of all the processes executed in real-time during the robot's navigation, allowing an immediate interaction between the robot and the environment.

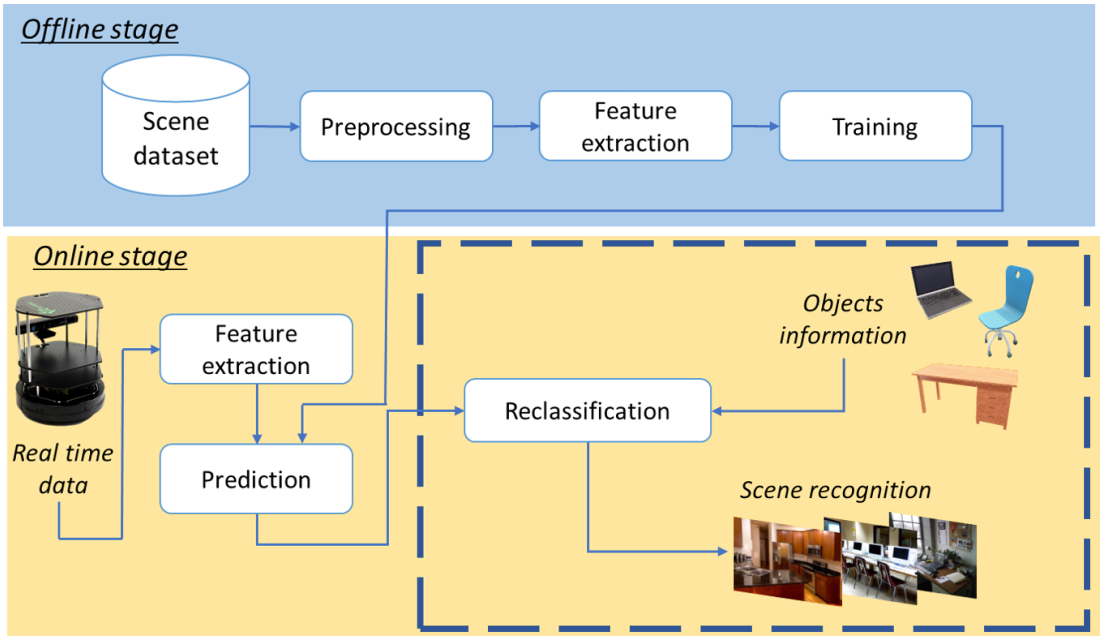


Figure 4.2: Probabilistic scene recognition model. The reclassification step includes the information of the scene and the results of the object recognition system to perform the estimation of the uncertainty in order to refine the final decision of the place that the robot is perceiving through its sensors.

The Probabilistic Scene Recognition Model comprises five steps: preprocessing, feature extraction, training, prediction, and reclassification. The final result of the model consists

of the semantic label of the place that the robot perceives and the detection confidence that represents the degree of uncertainty of the result. The detail of each step is explained below:

- *Preprocessing*. In this first step, the initial data is selected and prepared to create the training set. The data selection is built through two available scene datasets. First, the Kyushu University Indoor Semantic Place Dataset [124] that contains RGB and depth images for five scene categories: corridor, kitchen, laboratory, study room, and office. Second, the SUN397 Scene benchmark [181] with 397 categories of outdoor and indoor scenes. Besides, in this step, image resizing and color conversion are performed.

- *Feature extraction*. To obtain the features of each image, the well-known Bag of Words technique has been implemented. To do so, the scale and rotation invariant descriptor SURF (Speeded Up Robust Features) is implemented [102]. The sets of visual features are clustered using a vocabulary or codebook of visual words. We use K-means clustering [155] to build the visual vocabulary. Likewise, this step is applied to extract the features to the training dataset in the offline stage as well as to the retrieval images on the online stage.

- *Training*. Through the features extracted in the previous step, the training model is generated. Additionally, the general parameters of the selected classifier are defined. The features are grouped into matrices to find the rules for the best classification of the scenes. The output of this step is the model of the scenes to be detected, which will be used to train the classifier.

- *Prediction*. Due to its proven effectiveness in this type of problem, SVM as a classification algorithm has been implemented. Considering this, the general parameters for the selected classifier are defined. The result of this step is the detection confidence of the scene perceived by the robot.

- *Reclassification*. In this last step, the final scene detection confidence is adjusted due to the influence of the objects in the scene that can modify the final probability of being in a place. To get the information from the environment, the object detection system presented in Chapter 3.3 is used. The system operates in real-time, giving the object category and its respective object detection confidence. The delivered information also includes the distance and the angle to the center of the camera.

  Once the first estimation of the place and the surrounding objects' information are obtained, the reclassification process begins. First, a model of rules based on learning is created to determine how often an object is located in a specific place. With this information, we apply the Bayes' theorem to establish a relationship between objects

and the scene that improves the final results. This information is finally available as input for other essential tasks such as navigation, manipulation, human-robot, and robot environment interaction.

4.3.1 Uncertainties Management

Proper management of uncertainty is a crucial factor in obtaining final results closer to reality. Our uncertainty model considers the sensors' errors, the methods used, and the different objects present in the environment to generate accurate and precise results. In this work, we model the uncertainties through the objects' occurrence probabilities in the scene and the initial scene detection confidence calculated by the scene recognition model.

Probability of Occurrence of Objects

The probability that a robot is in a place can be improved by incorporating the relationships between objects and scenes in the environment. To do this, we calculate the probabilities of occurrence to determine how often an object tends to be in a specific scene. The SUN397 Scene Recognition Benchmark [181] is used as the scene dataset. This dataset contains 397 categories of scenes, with at least 100 images per class and 108,754 images in total. For this work, we chose images of typical houses and university buildings. Specifically, we selected six categories: laboratory, garage, classroom, bathroom, kitchen, and bedroom. With the object recognition system, we detect the objects in each scene. Then, the frequency distribution of the number of times that an object o of class s appears in a scene ξ of type i has been calculated as follows:

$$p(o_s|\xi_i) = \frac{n_{s,i}}{\sum n_{s,k}}, \tag{4.1}$$

where $n_{s,i}$ refers to the number of times an object appears in a scene and $n_{s,k}$ represents the sample size. Fig. 4.3 shows the frequency distribution of the appearance of a certain type of objects in four scene categories selected for this work. On the y-axis, the selected object classes are listed, and the x-axis, corresponds to the occurrence count depending on the scene.

It can be observed that some objects have a greater probability of occurrence than others. For example, a bed in a bedroom (96%) or a sink in a bathroom (72%), as one might expect in everyday human environments. Through this information, the occurrence probabilities of 17 common objects in indoor environments are calculated. Table 4.1 shows the normalized results for each object in each of the selected scene categories.

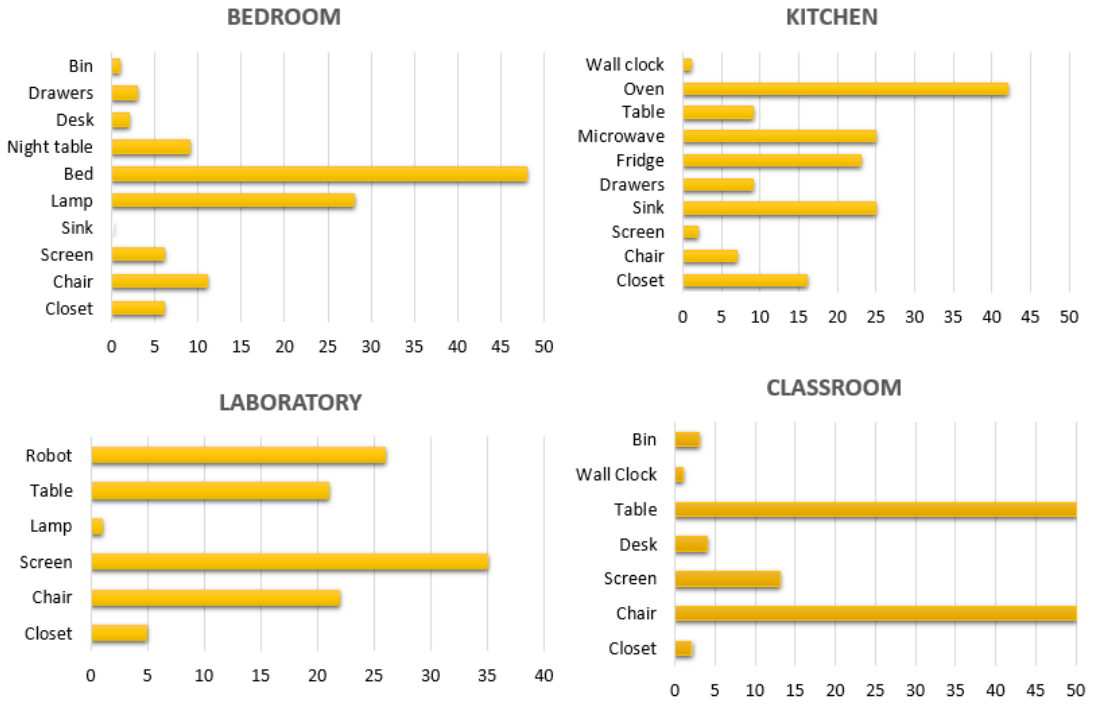


Figure 4.3: Frequency distribution of the appearance of certain common objects in four indoor environments: laboratory, classroom, kitchen, and bedroom.

Incorporating Prior Information to the Scene Recognition Model

This section details how to incorporate objects' information and their relationships in the proposed scene recognition model. This information is used to generate updated probabilities giving a more precise measurement of a potential result. We propose to combine prior information in the form of occurrence probabilities calculated in the previous section and the detection results obtained from the object recognition system to determine the final decision about the probability that a robot is in a specific place.

To solve this problem, the Bayes' theorem has been applied to calculate the posterior probability. The posterior probability of an uncertain proposition can be defined as the conditional probability that is assigned after new or additional evidence is uncovered. Formally, given a class of scene i, the probability that a scene ξ_i is the underlying scene given that an object o_s is present in the scene is:

$$p(\xi_i|o_s) = \frac{p(o_s|\xi_i) * p(\xi_i)}{p(o_s)}, \tag{4.2}$$

Table 4.1: Occurrence probability of 17 objects in the six scene categories selected for this work.

	bathroom	bedroom	kitchen	office	classroom	laboratory
closet	0.075	0.055	0.099	0.039	0.016	0.044
chair	0.025	0.018	0.043	0.284	0.398	0.192
screen	0.006	0.055	0.012	0.265	0.104	0.306
toilet	0.365	0.005	0.003	0.003	0.004	0.004
sink	0.453	0.005	0.154	0.003	0.004	0.004
lamp	0.006	0.258	0.003	0.045	0.004	0.004
bed	0.006	0.442	0.003	0.003	0.004	0.004
night table	0.006	0.083	0.003	0.003	0.004	0.004
desk	0.006	0.018	0.003	0.239	0.032	0.061
drawers	0.006	0.028	0.056	0.058	0.004	0.004
fridge	0.006	0.005	0.142	0.003	0.004	0.004
microwave	0.006	0.005	0.154	0.003	0.004	0.004
table	0.006	0.005	0.056	0.013	0.398	0.122
oven	0.006	0.005	0.259	0.003	0.004	0.004
wall clock	0.006	0.005	0.003	0.006	0.008	0.004
bin	0.006	0.005	0.003	0.026	0.004	0.004
robot	0.006	0.005	0.003	0.003	0.004	0.227

where $p(o_s)$ is the aleatory probability of finding an object o_s in a scene, which can be calculated through,

$$p(o_s) = p(o_s|\xi_i) * p(\xi_i) + p(o_s|\xi_i)' * p(\xi_i)'. \tag{4.3}$$

The initial scene detection confidence $p(\xi_i)$ is obtained from the scene recognition model during the classification phase.

Based on Eq. (4.2), to include the influence of several objects in the prediction process, we implement the Bayes' extended theorem to combine multiple conditions with independent ancestors. This way, the information about not only one object at a time but several objects at the same time has been incorporated. The probability of being in a scene ξ_i given several objects at the same time $\cap_s o_s$ is shown in Eq. (4.4).

$$p(\xi_i| \cap_s o_s) = \frac{p(\cap_s o_s|\xi_i) * p(\xi_i)}{p(\cap_s o_s)}, \tag{4.4}$$

where the probability of detecting several object in a scene, $p(\cap_s o_s|\xi_i)$, is defined as:

$$p(\cap_s o_s|\xi_i) = \prod_s p(o_s|\xi_i). \tag{4.5}$$

These probabilities have been obtained through occurrence matrices that indicate how often the objects appear in a specific scene. Finally, $p(\cap_s o_s)$ is the probability to find several specific objects in any scene category.

4.4 Multi-classifier Model for Scene Recognition

In the previous section, a probabilistic scene recognition model that considers some prior information to obtain a final scene detection confidence more precise and accurate has been presented. However, we consider that the model proposed in this section can benefit from the advantages that other models offer. The combination of several classifiers can lead to a meaningful improvement in the recognition rate and the overall system performance. Each of these classifiers can be designed using several techniques, input features, and different information sources. Each base classifier generates an output with the prediction about the class to which a scene belongs. Then, these outputs must be appropriately combined in order to obtain a final prediction.

We propose a multi-classifier model for scene recognition based on parallel topology. Fig. 4.4 shows a general overview of the proposed model [72]. Each base classifier receives RGB and depth images. They also perform the scene classification independently according to their feature extraction methods and classification algorithms. The outputs of the classifiers are the probabilities of the classes to which the scene belongs. Finally, the individual results are fused through a weighted voting scheme.

The proposed model is capable of working with an arbitrary number of base classifiers. For this work, we have selected two base classifiers. The first base classifier consists of a scene recognition system based on machine learning that generates as output the score values for each scene category. The second base classifier corresponds to the object-based probabilistic scene recognition model presented in the previous section. Each base classifier generates a ranked list of the classes with their scene detection confidences.

To ensure an adequate combination of the base classifiers' results, we study two weighted voting schemes. In the first voting scheme, the weights are assigned to each classification output considering the individual performance of each scene recognition model. Each model's accuracy is used to determine the final result of the scene category perceived by the robot. The second scheme applies genetic algorithms to determine the optimal weights that have to be applied to each base classifier model's output. Finally, both voting schemes are compared to determine the best solution for the scene recognition problem.

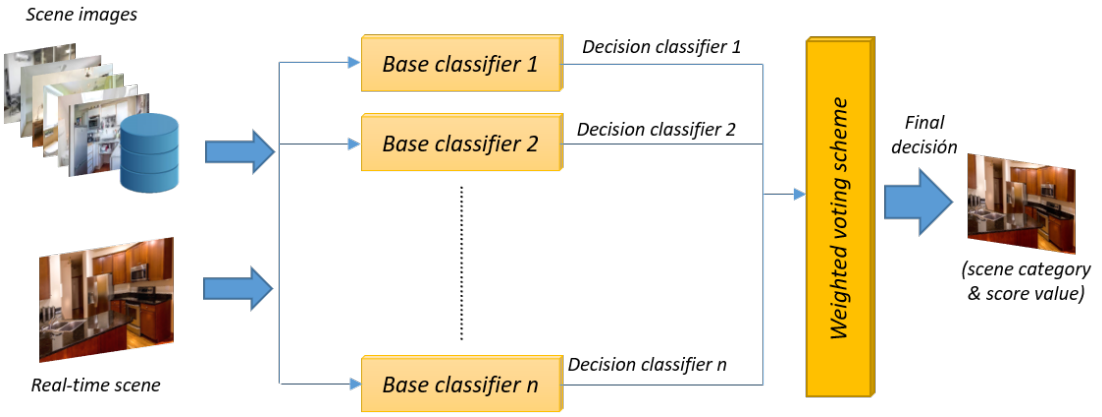


Figure 4.4: General overview of the proposed multi-classifier model for scene recognition. The final decision of the place where the robot is calculated based on weighted voting schemes.

4.4.1 Base Scene Classifiers

To build the multi-classifier model, we select two scene recognition models as base classifiers. First, a scene recognition system based on machine learning is employed. The model makes use of images of different indoor environments to generate a training model. Color and grayscale images are used as input data. The model consists of two stages, the training and classification stage. For feature extraction, the Bag of Features (BoF) technique combined with SURF descriptors is employed. The classification method is based on neural networks. When deployed in a real-world environment, the robot identifies the current scene according to the trained model, and the result is a list of the trained classes with their respective score values. Figure 4.5 shows a general diagram of the first base classifier.

As a second base classifier, the probabilistic scene recognition model explained in Section 4.3 has been used.

4.4.2 Weighted Voting Schemes

Properly selecting the weights is a crucial task to have a more robust multi-classifier. In this subsection, we implement two weighted voting schemes to solve the scene recognition problem.

Weighted Voting Scheme by Accuracy

This voting scheme assigns a weight to each base classifier model. The weights are based on the assumption that classifier models with high recognition performance are more reliable

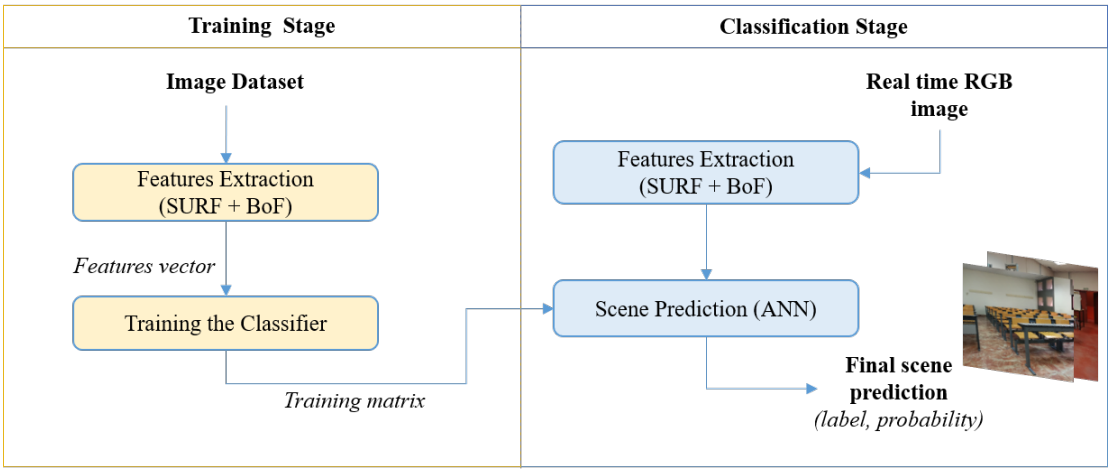


Figure 4.5: A general representation of the scene recognition model based on BoF and SURF descriptors. The model uses multilayer perceptron as classification method.

than classifiers with lower performance rates. In this work, the weight of each base classifier is represented by the accuracy of each model. The accuracy is a measure to evaluate the performance of a classification model that represents the proportion of the correct predictions that is determined as in Eq. (3.7) of Chapter 3.

In the process, each scene class probability of each independent classifier is multiplied by its respective weight. Then, the weights of the classifiers are added for each class of scene. The final prediction is the scene class with the highest sum of weights. Eq. (4.6) shows the formulation of this voting scheme:

$$\hat{y} = \underset{i}{argmax} \sum_{j=1}^{m} \omega_j p_{ij}, \tag{4.6}$$

where $\hat{y}$ is the final outcome after the combination process, ω_j is the weight of each base independent classifier j associated with the accuracy of the model. Finally, p_{ij} is the prediction result of the base classifier j for the scene class i.

Weighted Voting Scheme through Genetic Algorithms

In this section, we propose a second voting scheme for the base classifiers combination. The method considers a set of weights as free parameters in order to find a combination of values that produces the best result for the complete system. For this work, we use genetic algorithms due to their robustness and good performance demonstrated in many other complex problems [96]. Genetic algorithms (GAs) [75] can be defined as a subset of the

computer science branch called Evolutionary Computing. GAs are search-based algorithms inspired by the processes of natural and genetic selection. Fig. 4.6 illustrates the process.

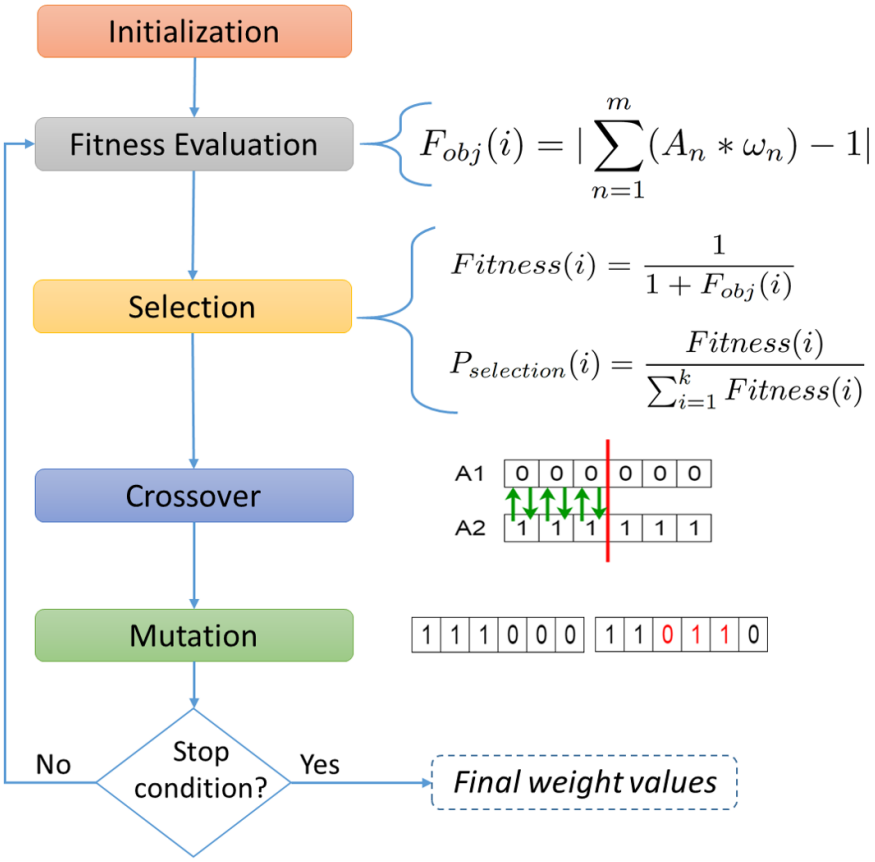


Figure 4.6: Implementation of genetic algorithms for scene recognition. The process consists of five phases: initialization, fitness evaluation, selection, crossover and mutation.

The process starts with a group or population of possible solutions to a given optimization problem. Each candidate solution is called a chromosome, and initially, each of them is created randomly. It is necessary to define a fitness function to evaluate the quality of the solution. Each chromosome is assigned a fitness value, and only the fittest individuals, that is, the chromosomes that constitute a better solution to the optimization problem, are allowed to "reproduce". During reproduction, new individuals are created from the random changes (mutation) and fusion of two chromosomes (crossover). This way, better individuals or solutions are obtained over the generations (evolving) until a stop criterion is reached.

The phases of a GA are: initialization, fitness function definition, selection, crossover, and mutation. The details of each phase with respect to our method are described below.

- *Initialization*. Each chromosome is represented by an array of weights that are real numbers between 0 and 1. The number of elements of the array is equal to the number of base classifier models (m). All the values of the chromosomes are generated randomly. For this work, the population size (N) is set to 50.

- *Fitness Function* The fitness function or objective function $F_{obj}(i)$ is defined as the accuracy of each base classifier model (A_n) combined with the weights (ω_n) corresponding to each chromosome i. Through this function, shown in Eq. (4.7) a fitness score for each chromosome is calculated to determine if an individual is able to compete with others.

$$F_{obj}(i) = |\sum_{n=1}^{m}(A_n * \omega_n) - 1|. \tag{4.7}$$

- *Selection*. In this step the idea is to select the fittest individuals. The fitness of each chromosome is calculated according to:

$$Fitness(i) = \frac{1}{1 + F_{obj}(i)}. \tag{4.8}$$

  The selection probability is calculated considering the fitness score of each chromosome, Eq. (4.9). Lower values of $Fitness(i)$ lead to lower selection probabilities.

$$P_{selection}(i) = \frac{Fitness(i)}{\sum_{i=1}^{k} Fitness(i)}. \tag{4.9}$$

- *Crossover*. This is the most important step in a genetic algorithm. During the crossover, for each pair of parents that must be mated, a random crossover point is chosen within the genes. This way, a new population is created by fusing the information in the chromosomes.

- *Mutation*. This step involves replacing a gene in a random position with a new value. This process guarantees the diversity of the population and avoids early convergence. Once the mutation process is done, the GA's iteration has finished, yielding a new generation. With this, the objective function is evaluated. If the result of the objective function decreases, it means that a better solution was found compared with the previous chromosome. Finally, the best chromosome represents the final result of the weights that will be applied for the fusion of the base classifiers.

4.5 Experimental Evaluation

The main goal in these experiments is to evaluate the robot's performance during a scene recognition task in real-time. Also, through the experiments, we will demonstrate the usefulness of the inclusion of object information, their object and scene relationships, and the calculation of uncertainties in a scene recognition model.

4.5.1 Experimental Setup

The models developed in this chapter are prepared to operate in real-world environments by real mobile robots. For the experiments, a Turtlebot 2 robotic platform and a self-designed robot built in our laboratory [49] have been used. The mobile platforms are equipped with an ASUS Xtion Pro Live camera as an RGB-D sensor to recognize objects and scenes. Also, to demonstrate the usefulness of our proposed models in the real world, two environments have been selected: a university building and a typical house. In all the experiments, the robots were teleoperated to move through the environment. However, the different modules of the proposed probabilistic scene recognition model and the multi-classifier model operate autonomously.

4.5.2 Evaluation of the Probabilistic Scene Recognition Model

The first experiment consists of evaluating the performance of the object-based probabilistic scene recognition model in real-world environments. Through confusion matrices as testing methodology, different measurements have been calculated to analyze the proposed model. True positive rate (sensitivity), true negative rate (specificity), accuracy, and misclassification are the selected measures. Table 4.2 shows the results after several scene detections.

The first environment, a university building, reports an accuracy between 73% and 76% and an average misclassification of 25.23%. In the house environment, the accuracy is higher than in the previous environment, with an average of 85.41%. In the same way, the misclassification rate is low, with an average of 14.58%, being the lowest measurement for the bedroom with 10.94%. The results show that our model has a better average accuracy of 80.10% compared to that reported in the approach of Nascimento et al. [129] of 71.08% using the same dataset. This accuracy improvement demonstrates the validity of our approach, taking into account that our experiments are performed in real-time and in real-world environments.

A second experiment is conducted to evaluate the uncertainty model proposed in this chapter. The model considers the object's information and their relationship with the environment and between them. During this test, the robot is moved by teleoperation capturing real-time images that contain objects in different scenes. The scenes to recognize

Table 4.2: Evaluations of the probabilistic scene recognition model in the two selected environments: a university building and a typical house.

Evaluations	Accuracy	Misclassification	Sensitivity	Specificity
Laboratory	0.749	0.252	0.529	0.968
Garage	0.761	0.239	0.733	0.788
Classroom	0.734	0.266	0.813	0.656
University Avg.	0.748	0.252	0.692	0.804
Kitchen	0.813	0.188	0.750	0.875
Bedroom	0.891	0.109	0.813	0.969
Bathroom	0.859	0.141	0.813	0.906
House Avg.	0.854	0.146	0.792	0.917
Total Avg.	0.801	0.199	0.727	0.860

are: bathroom, kitchen, bedroom, classroom, laboratory, and garage. Fig. 4.7 and Fig. 4.8 show the main results.

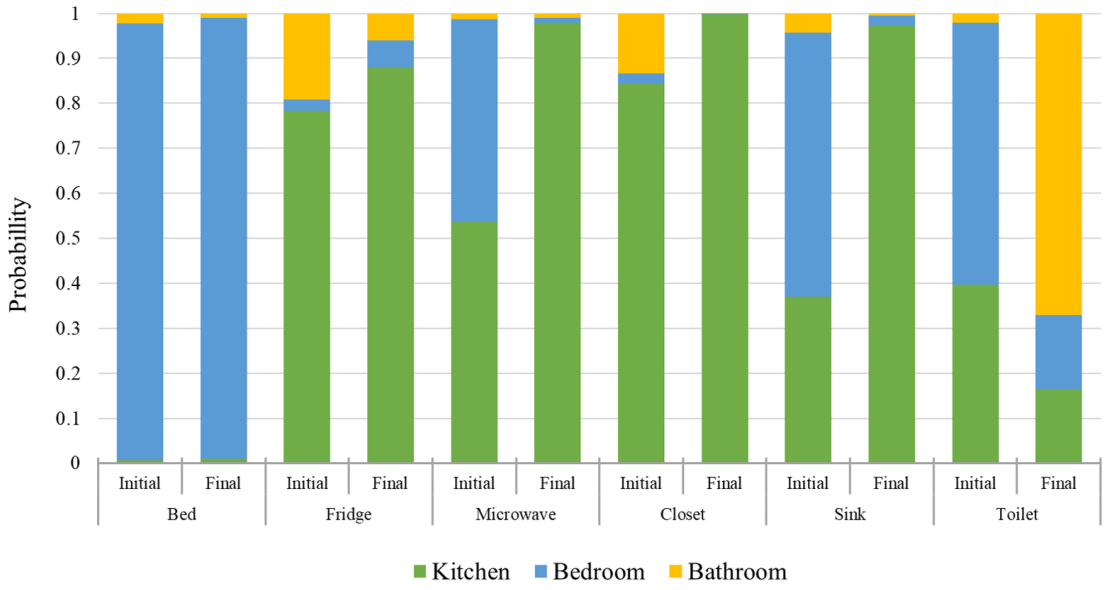


Figure 4.7: The uncertainty model influence in the results of the probabilistic scene recognition model for the environment: a typical house.

In the bar charts, the y-axis shows the recognized scene categories, while the x-axis shows the objects detected in each place. Two columns are displayed for each detected object class.

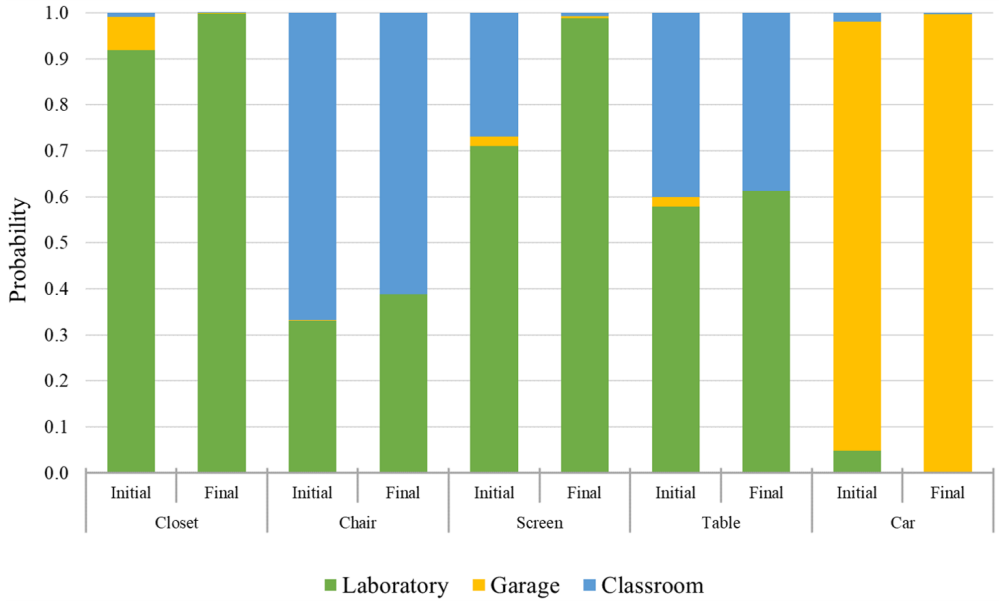


Figure 4.8: The uncertainty model influence in the results of the probabilistic scene recognition model for the environment: a university building.

The first column on the left, labeled "Initial", shows the scene detection confidence for each scene category without any object information. The second column, labeled "Final", shows the final scene detection confidence for each scene category after the reclassification process has run.

In most cases, detected objects enhance scene detection confidences. When a microwave is detected in the experiments, the scene category kitchen increases its confidence from 53.6% to 97.9%. On the contrary, in other situations, the probabilities decrease. When detecting a sink, the probability of being in a bedroom decreases from 58.69% to 2.1%. We can also observe how initially erroneous classifications are corrected. For example, in one of the cases, the recognition system gave these results: bedroom 58.26% and bathroom 39.66%. After the reclassification process, the probabilities were modified by decreasing the bedroom probability to 16.44% and increasing the probability of being in a bathroom to 67.1%. Through the conducted experiments, the usefulness of the probabilistic scene recognition model presented in this chapter has been demonstrated. Also, it was proved how the system's performance improves through the uncertainty model proposed in this work. We have also validated how the model enhances using the objects' information and their relationship with the environment in which they are located.

4.5.3 Evaluation of the Multi-classifier Model

The experiments below evaluate the weighted voting schemes based on accuracy and genetic algorithms proposed in this chapter for the scene recognition task.

Evaluation of the Weighted Voting Scheme by Accuracy

In this experiment, the robot moves by teleoperation capturing images of the different scene categories, while the base classifiers execute independently at the same time. We selected two human indoor environments: a typical house and a university building. Each base classifier gives its scene detection confidence about the place that the robot is perceiving. The multi-classifier model receives this information and merges it using as weights the accuracy of each model obtained during its respective training phase. These accuracy values are 78.5% for the first base classifier model and 88.4% for the second base classifier model. Table 4.3 shows the final results of the base classifier models and the result of the proposed multi-classifier model using a weighted voting scheme based on accuracy.

Table 4.3: Evaluations of the multi-classifier model compared to the results of the base independent classifiers based on accuracy in two environments: a typical house and a university building.

Environments	Recognition rate (1)	Recognition rate (2)	Final recognition rate
Laboratory	0.725	0.799	0.876
Garage	0.639	0.746	0.788
Classroom	0.720	0.723	0.748
University Avg.	0.695	0.756	0.804
Kitchen	0.788	0.892	0.915
Bedroom	0.729	0.789	0.799
Bathroom	0.632	0.761	0.789
Living room	0.670	0.652	0.695
House Avg.	0.705	0.774	0.799
Total Avg.	0.699	0.765	**0.802**

The first two columns show the scene recognition rate of each base classifier model, 69.98% and 76.48%, respectively. The last column shows the scene recognition rate of the proposed multi-classifier model (80.17%) and how the scene recognition task improves after merging the base classifier models considering the accuracy as weights.

Evaluation of the Weighted Voting Scheme through Genetic Algorithms

The second experiment was carried out under the same conditions as the previous one. The robot moves through the two selected environments, a university and a typical house, classifying the scene that it perceives. The voting scheme strategy used in this experiment is based on genetic algorithms. Table 4.4 shows the configuration of the main parameters used for the generation of the weights through genetic algorithms.

Table 4.4: Main parameters used in the implementation of genetic algorithms for scene recognition.

Parameters	Values
Population size	50
Nº iterations	100
Crossover rate	0.8
Mutation rate	0.1

Fig. 4.9 shows the prediction results after applying the weighted voting scheme based on genetic algorithms in different scene categories.

Figure 4.9: Execution of the multi-classifier model based on weighted voting scheme through GA in a typical house and in a laboratory of a university. The information of each base classifier model are considered to obtain the final result.

After the weights are generated, they are multiplied by the individual results of each base classifier model. In Table 4.5 the results of the application of our multi-classifier model based on GAs are presented.

The results show that the implementation of the two weighted voting schemes proposed in this work increases the recognition rate in comparison with each of the base classifier

Table 4.5: Evaluations of the multi-classifier model based on genetic algorithms.

Environments	Recognition rate (1)	Recognition rate (2)	Final recognition rate
Laboratory	0.719	0.779	0.877
Garage	0.694	0.811	0.864
Classroom	0.709	0.754	0.795
University Avg.	0.707	0.781	0.845
Kitchen	0.829	0.918	0.959
Bedroom	0.738	0.782	0.856
Bathroom	0.574	0.729	0.754
Living room	0.666	0.679	0.731
House Avg.	0.702	0.777	0.825
Total Avg.	0.705	0.779	**0.835**

models. The voting scheme based on accuracy achieved good results with an 80.17% of average recognition rate. However, a weighted voting scheme based on genetic algorithms produces the best result for the scene recognition task with an average recognition rate of 83.52%. The experiments show that the proper combination of independent classifiers can compensate for the errors during the classification process and generate much more robust scene recognition models.

4.6 Discussion

This chapter proposed an object-based probabilistic scene recognition model considering uncertainties and a multi-classifier model for scene recognition based on weighted voting schemes. Regarding the probabilistic scene recognition model, the experiments have demonstrated the helpfulness of the proposed model when information of the objects is incorporated. Likewise, it has been proved how the relationship between objects and scenes can influence the final decision of where the robot is based on what it perceives. With the uncertainty model proposed in this work, we have strengthened the classification process incorporating perceptual information.

In classical approaches, classification algorithms, feature extraction methods, and image processing techniques are the core of the process. However, in this approach, the essential part is the proper and accurate modeling of the environment, which has to be as close as possible to reality. This model must incorporate all the elements that intervene in a greater and lesser degree in the scene recognition process. In this way, we have obtained a robust

recognition model that allows robots to understand the environment better to perform essential tasks such as navigation, manipulation, and interaction with humans.

On the other hand, we proposed a multi-classifier model for scene recognition based on weighted voting schemes. In this work, two strategies were considered. In the first strategy, the weights assigned to each base classifier were equal to the respective accuracy of each model. On the other hand, in the second scheme, a genetic algorithm was implemented for weight optimization. In this process, accuracy was used to create the objective function. The experiments showed that the recognition rate of the proposed multi-classifier model increases with respect to the recognition rate of individual classifiers. Furthermore, we have implemented two variants of voting schemes that achieve both good results. However, the weighted voting scheme based on genetic algorithms works better than the simple voting scheme by accuracy. The adequate combination of independent classifiers through genetic algorithms allows obtaining a more robust and precise model for scene recognition, taking advantage of the benefits of each classifier and compensating for the errors of each of them.

5
Region-based Semantic Labeling

Intelligent robots need to understand the world to interact significantly with it. This includes knowledge about the objects and where they are located. Likewise, the ability to generate meaningful representations of the environment aids a service robot to fulfill a variety of tasks in domestic environments and interacting with humans. In this chapter, we address the problem of segmenting the environment into meaningful regions considering semantic information. We propose a region-based semantic segmentation method, which integrates scene and object information in a probabilistic way to generate different regions in the environment. We also demonstrate how our representation can boost the performance of more complex robotic tasks compared to more standard environmental representations.

5.1 Introduction

To operate in human environments, autonomous robots need to understand their surroundings at a human-like level in a way that not only allows them to navigate and identify elements but also interact with them and make decisions. Service robots need to be able to collect and interpret semantic information from the environment to build useful representations of it that allow them to know where they are, what the place is like, and also the actions they can take. This knowledge can help robots to more efficiently perform complex high-level tasks. Likewise, it can contribute to decision-making processes and facilitate human-robot interaction when performing tasks requested by its users.

Recent events in the world have made us change our way of thinking about our environments due to the increase of *from home* activities. We have turned our houses into work areas, some parts are now classrooms and offices, and what is more, teaching activities designed for campus-based facilities are conducted through a laptop on the kitchen table. Therefore, could the table and the chair in my kitchen also be called my *office*? All this leads us to rethink the representation of our environments and how to transfer it to a service robot that operates in domestic scenarios.

Traditionally, previous methods for constructing semantic maps aim to assign a semantic label to each complete room of a given environment [48, 138, 160, 186]. In this way, a map is assigned one semantic label per room. These labels are usually associated with the utility of the room, its physical layout, or the way humans understand the space. Great advances have been reported extracting semantic information to generate maps, but always keeping in mind the physical boundaries. However, these methods might fail, for instance, in a studio apartment where such boundaries are not so clear. Also, object placement that changes constantly, and the increasing multi-functionality of domestic environments can affect the semantic labeling process.

In contrast, in this work, we assign the semantic labels to regions within each room by assuming that they can vary depending on contextual information. For example, a living room can be at the same time a bedroom, if a sofa is used to sleep, or an office if a laptop is placed on the dining table. We present an approach that creates a subdivision of the environment into regions by maintaining the confusions (miscategorizations) which are due to the appearance or to the distribution of objects inside. Fig. 5.1 presents the main idea behind our method.

We maintain smaller regions that emerge automatically inside full rooms. Some of these regions will be assigned a semantic label that actually corresponds to the room, but some others will be assigned labels that do not correspond and then create confusing areas. Instead of applying further filters to try to correct these confusions, like in [123, 166], we keep these areas assuming that confusions in the perception system are due to similarities. We merge several semantic maps built from different sensing modalities, each of which generates different confusing regions.

In this chapter, we implement, on the one hand, a categorization system based on visual appearance, and on the other hand, a categorization system based on objects detected in the scene. Each method generates a semantic map with different categorized regions within each room (Fig. 5.1). The shape, size, and category of each region depend on the results of the specific classifier used. Finally, we present a method to merge the resulting maps.

The main contribution of this work is a framework for consistent subdivision and labeling of domestic environments based on regions using different modalities [69]. Moreover, we present a solution for merging region-based maps originated by different sources of

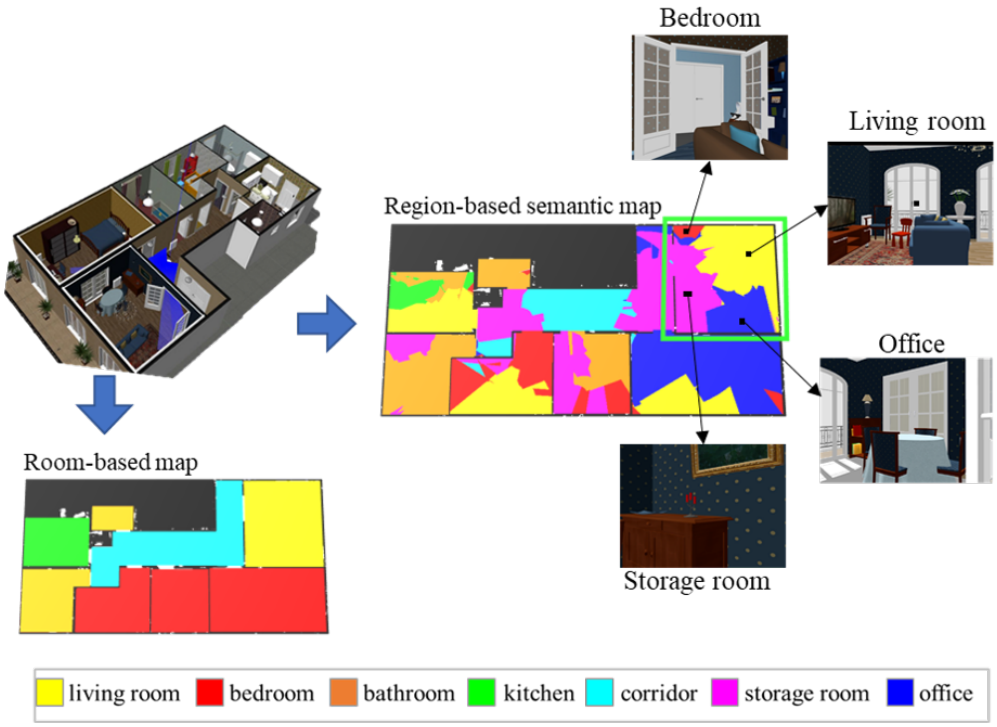


Figure 5.1: A 2D semantic map built by our proposed approach. Right: Different parts of a living room are categorized into different regions due to visual similarities. Bottom-left: the ground truth of the environment based on the physical boundaries of each room.

information, in particular, visual appearance and object detection. The resulting experiments show that keeping smaller regions and their confusing labels can improve some domestic tasks in service robotics. In particular, we show an increase in the efficiency of a service robot performing an object search task. In addition, these confusions may be more intuitive for people and, therefore, improve human-robot communication.

The remainder of this chapter is organized as follows. The related work is discussed in Section 5.2. Then, Section 5.3 describes our proposed method. The results of the conducted experiments are shown in Section 5.4. Finally, in Section 5.5 the conclusions of the chapter are presented.

5.2 Related Work

The problem of constructing semantic maps has been widely explored in robotics. It can be divided according to the type of source of information (laser scans, scene, objects, user utterances, activity recognition, among others) and its application in image frames or in a

complete environment. Works such as [9, 159, 161, 175] deal with semantic mapping by exploiting only scene information in the whole environment.

The approach of Sünderhauf et al. [161] proposes a place categorization and semantic mapping system based on a Bayesian filter framework. This allows to incorporate prior scene information and guarantee temporal coherence during the classification outcomes. Besides, Wang et al. [175] apply Bayesian update for spatial and temporal consistency, considering the scene recognition problem as a probabilistic estimation. Sun et al. [159] propose a convolutional neural network for object detection and scene classification simultaneously. Although they detect and segment the objects in the environment, a semantic map is built only upon scene information. The approach of Balaska et al. [9] semantically defines the environment through graph-based segmentation. SURF features are extracted from each image frame and similarity is studied to define semantic clusters or regions through the Louvain Community Detection Algorithm (LCDA) [13].

On the other hand, some works leverage object information to create representations of the environment [115, 138, 165]. The approach of Trevor et al. [165] extracts planar surfaces such as shelves, counters and tables from point clouds that are then integrated into a semantic map. In a more recent work by Luo et al. [115], objects are detected through a CNN and associated with a topological node. Then, the nodes are grouped into rooms by identifying regions of interest and applying morphological operations and distance transformations based on the metric map. In the work by Qi et al. [138] an object semantic grid map is built by merging the occupancy grid map and object point clouds. Common sense about the relations between objects and rooms is used to segment a small environment and get the room labels.

The combination of different sources of information has also been exploited to generate consistent semantic maps. In the approach of Kostavelis et al. [94], a place recognition method based on appearance-based histograms and an object detection based on a saliency attentional model are implemented. Both methods generate inference results about the scene label. Then, in the final decision about the scene label, the place recognition results always prevailed, and when the results diverge, the image frames are discarded. Later, in the work by Ruiz et al. [150] a Conditional Random Field (CRF) is employed to exploit contextual relations between objects and rooms, and also to measure uncertainties related to the categorization process. Similarly, in the approach of Brucker et al. [16] a pre-segmentation of the environment based on geometric primitives is assumed. Scene and object information is combined through a CRF to obtain a semantic label for each room in different simulated environments.

Other works deal with other sources of information from the interaction with humans or other devices. The approach of Rosa et al. [146] infers semantic labels for each room by exploiting user activities. The map is updated considering the relationships between human

activities and room types. A people detector and a wearable device are implemented to obtain information about the activity that a person is performing to update the final map. Moreover, in the work by Katsumata et al. [87] user utterances are used to create a bag of words to associate it with a room category.

In all the above approaches, semantic labels are assigned according to the physical boundaries of the environment. It is assumed that all cells belonging to the same room should have the same scene label. So, for instance, by averaging the values in the cells, a unique category is obtained for each room. However, real-world environments can be categorized not only according to a physical division based on walls and doorways but also by taking into account existing objects, their location, and the uses of space given by humans. That is why in our work, we focus on the subdivision of the environment into parts that we call regions. Object and scene information is properly integrated to increase scene understanding required for service robots moving in human environments, resulting in a more efficient performance of complex robotic tasks.

Few works attempt to represent the environment as small regions. In the work by Fowler et al. [38] scene human motion and actions are used to train a CNN to segment a 3D environment that defines certain regions such as walkable, sittable, usable, and structure. In the same way, the work by Roy et al. [148] develops a CNN-based approach to predict regions based on five affordance types. In this case, the segmentation has been carried out at the image level. In these works, regions are predefined and trained previously. In our work, we try to minimize human intervention; that is why only object and scene information is used. Region labels are determined by the scene classifier, and the arrangement of objects in the environment helps to define these regions, making our approach more generalizable and scalable for different human environments.

5.3 Proposed Model for Segmenting the Environment

The construction of our region-based semantic map is made up of three key steps:

- Generation of a semantic map based on the classification of RGB data according to the global structure of the scene.
- Histogram analysis of the existing objects in each image frame to infer a scene label to build an object-based semantic map.
- Integration of the scene-based and object-based maps to build a grip map with the final semantic regions.

In this work, we assume a mobile robot that initially explores the environment to build a 2D occupancy grid map using a Simultaneous Localization and Mapping (SLAM) algorithm

such as gmapping. The robot is also able to localize and navigate the environment using a localization algorithm such as amcl/particle filter pose estimation. As it builds the map, RGB data is gathered through its camera sensor to then, use as an input of each step of our method. Thus, the input information corresponds to RGB images with their corresponding pose information (x, y, θ) where x and y are the 2D coordinates of the robot on the map and θ is the robot's orientation. Each part of our approach will be described in more detail below.

5.3.1 Scene-based Map Building

To classify each image individually, we first apply a scene recognition model based on deep learning using the VGG16 architecture trained on the Places365 database [202]. As we are focused on domestic environments, the network's output layer has been limited to seven scene categories: kitchen, bedroom, bathroom, office, living room, storage room, and corridor. The scene classification results are then used to build a 2D semantic map of the environment. Each image is assigned a scene probability estimate and associated with a robot position, so to create the map, the scene probability is propagated into cells within the field of view (FOV) of the robot. Each cell k in the map is assigned a vector $\hat{v}_k = (v_0, v_1, v_2, ...v_n)$ which indicates the probability that cell k belongs to a certain scene category i. We maintain one map layer per semantic category; therefore, seven map layers are created as illustrated in Fig. 5.2.

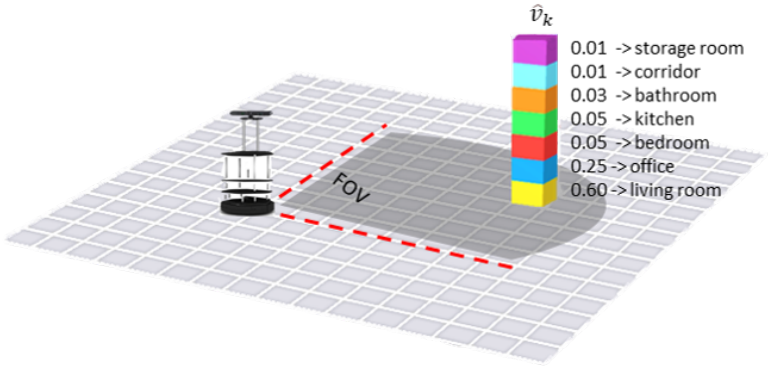


Figure 5.2: Semantic map structure. Each cell contains a vector with the probabilities of belonging to different scene categories (denoted by different colors). All cells within the FOV of the robot's sensor are updated with the scene classification results.

To update each cell, we apply the occupancy grid mapping algorithm [29]. This way, it is possible to maintain spatial and temporal coherence, as well as correct false classifications and incorporate the observations as prior knowledge. The probability $p(v_k^i \mid I_{1:t})$, i.e., that cell k has the label i given a set of input images is denoted as:

$$p(v_k^i \mid I_{1:t}) = [1 + \frac{1 - p(v_k^i \mid I_t)}{p(v_k^i \mid I_t)} \frac{1 - p(v_k^i \mid I_{1:t-1})}{p(v_k^i \mid I_{1:t-1})} \frac{p(v_k^i)}{1 - p(v_k^i)}]^{-1}, \tag{5.1}$$

where $p(v_k^i \mid I_t)$ denotes the probability that cell k is assigned the label i given a current image I_t when the cell k falls inside the field of view of the robot when capturing the image as shown in Fig. 5.2. The previous estimate is represented by $p(v_k^i \mid I_{1:t-1})$ and $p(v_k^i)$ is a prior probability. For practical purposes the implementation is done by using the log-odds notation [29], Eq. (5.1) can be expressed as:

$$L(v_k^i \mid I_{1:t}) = L(v_k^i \mid I_{1:t-1}) + L(v_k^i \mid I_t), \tag{5.2}$$

with

$$L(A) = \log[\frac{p(A)}{1 - p(A)}], \tag{5.3}$$

where A is the probability of an event happening. Finally, the label i with the highest probability $v_k^{i^*}$ is assigned to each cell:

$$v_k^{i^*} = \underset{i^*}{\operatorname{argmax}}\, v_k^i \forall i. \tag{5.4}$$

By grouping cells with the same category, different regions are formed within each room.

5.3.2 Object-based Map Building

Scene recognition implicitly considers surrounding information. However, certain objects can provide relevant information to define some parts of the environment [105, 134]. This section of our proposed method relies on the detection of objects in order to infer semantic information about the place that the robot perceives through its sensors. To exploit local features and relationships between objects, we build an object-based semantic map by comparing the object histogram obtained from each image frame with the object reference histograms created for each scene category considered in this work.

Histograms Generation

We have built seven object reference histograms for the seven previously chosen scene categories. These histograms have been created upon the NYU depth V2 dataset [154]. A total of 5000 RGB images for each scene category have been selected from the dataset to which an object detector based on Resnet-101 [64] trained on the COCO dataset [109] has been applied. The detector outputs are the object class, the detector confidence, and

the bounding box of the object in the image. The detector that originally identifies 80 object categories has been limited to detect 54 object categories that are more common in domestic environments. Then, the appearance frequency of objects in each image in the dataset is calculated to obtain the seven reference histograms that will be used later for comparisons. This is a way of representing the objects' distribution and their interaction with the environment based on the category of the scene in which they are located. Fig. 5.3 shows the object reference histograms for two scene categories: (a) a living room and (b) a kitchen.

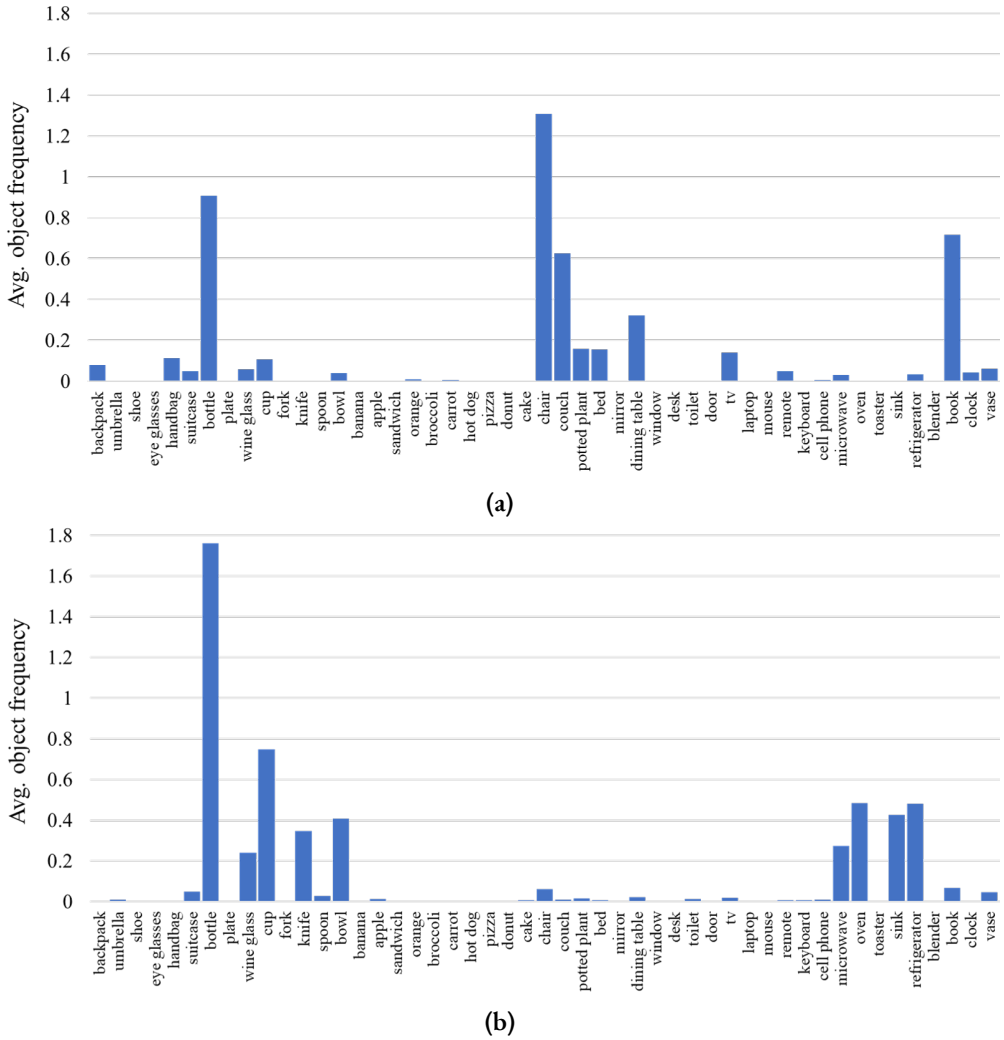


Figure 5.3: Object reference histograms. (a) histogram of the scene category: living room. (b) histogram for a scene category: kitchen.

As it can be observed, the histograms are considerably different, suggesting the presence of variant patterns in the scene categories. Depending on the type, some scenes may be richer in information than others, and this is the fact that we employ to infer the scene category in an image frame based on the encountered objects.

Histograms Comparison

Considering the environment that we want to semantically subdivide, each image frame obtained from the camera sensor is sent to the object detector. Then, the corresponding histogram, from now on called test histogram, is built by counting the number of times the objects appear in the image. The comparison between histograms (the seven reference histograms with the test histogram, which have been previously normalized) is made through the Kullback Leibler (KL) divergence calculation [98]. This measure (also called relative entropy) allows us to know how different a probability distribution is from another. We use histogram comparison to determine which histogram the image is most similar to, in order to deduce the scene category to which it belongs.

Let us consider two probability distributions. H_i represents the object reference histogram for a scene category i, while H_t represents the probability distribution of objects in each image frame t of the environment under study. Both probability distributions are defined in the same probability space X. Then, The KL divergence can be defined as:

$$D_{kl}(H_t \parallel H_i) = \sum_{x \varepsilon X} H_t(x) \ln \frac{H_t(x)}{H_i(x)} , \tag{5.5}$$

where x corresponds to the possible outcomes of each distribution, that in this case are the object classes coming from the detector. A KL divergence of 0 means that the two distributions are identical. The lower the KL values, the better we have matched the model with our approximation. Each test histogram has been assessed against the seven object reference histograms. Therefore, seven KL divergence values are generated per image frame. The minimum KL value will represent the best matching and in consequence, it can be inferred that the image is more probable to belong to the scene category associated with the respective reference histogram.

With the resulting scene categories, a semantic grid map is created. Each cell within the field of view of the camera sensor is assigned the final scene category of the respective image frame. If no objects are detected in an image frame, the associated cells are marked as unlabeled. Finally, to update the map and overcome the overlapping the occupancy grid mapping algorithm has been applied as in Section 5.3.1.

5.3.3 Integration of Semantic Features

In this section, we present an approach to integrate region-based maps that come from different categorization methods. In particular, we integrate the region-based maps from the scene classifier (Section 5.3.1) together with the map obtained by using object histograms (Section 5.3.2).

Let $\mathcal{M}^s$ be the scene-based map built in Section 5.3.1 and $\mathcal{M}^o$ the object-based map built in Section 5.3.2. The final map $\mathcal{M}^f$ is created by applying the following decision-making procedure in each cell:

$$\mathcal{M}^f(k) = \begin{cases} \mathcal{M}^s(k) & \text{if } \mathcal{M}^s(k) = \mathcal{M}^o(k) \\ & \text{or } \mathcal{M}^o(k) = \textit{unlabeled} \\ P(v_k^i \mid \cap_s O_s) & \text{if } \mathcal{M}^s(k) \neq \mathcal{M}^o(k)\,. \end{cases} \tag{5.6}$$

In case the most likely scene categories are identical for cell k in both maps, or when no objects are detected, the method updates the cell k in the final map with the scene category of the scene-based map and proceeds to evaluate the next cell. Otherwise, if the scene categories in cell k of the two maps diverge, the Bayes' extended theorem is applied to determine the definitive scene category for the final map:

$$P(v_k^i \mid \cap_s o_s) = \frac{p(\cap_s o_s \mid v_k^i) p(v_k^i)}{p(\cap_s o_s)}\,, \tag{5.7}$$

with

$$P(\cap_s o_s) = \prod_s P(o_s \mid v_k^i)\,, \tag{5.8}$$

where $p(v_k^i \mid \cap_s o_s)$ is the probability estimation that cell k has the label i given a set of objects $\cap_s o_s$ which have been recognized in the respective image frame and $p(\cap_s o_s \mid v_k^i)$ corresponds to the object-scene occurrences obtained from [65]. In addition, $p(v_k^i)$ is the current estimate on the scene category given by the scene-based map. Finally, $p(\cap_s o_s)$ is calculated by multiplying the detection confidences obtained from the object detector.

Through this, the scene category with the highest probability in each cell is selected to get a final semantic grid map with meaningful regions of interest. This representation of the environment can increase the scene understanding of the robot as well as improve its performance in more complex robotic tasks.

5.4 Experimental Evaluation

The main objective in this chapter is the development of a method for the division of the environment into meaningful regions by maintaining the confusions (miscategorizations)

which are due to the appearance or to the distribution of objects inside. The experiments demonstrate the feasibility of assigning a semantic label to areas of the environment to build semantic regions in both simulated and real-world scenarios. Furthermore, we prove that our semantic representation can boost the efficiency of a robot during the execution of more complex robotic tasks.

5.4.1 Experimental Setup

The experiments were conducted in simulated and real-world scenarios. We select two home-like environments obtained from the SweetHome 3D [137] software. This software is an open source 3D interior design software application. It allows us to create houses with different layouts and exporting each model in an appropriate format for robotic applications. The real-world environments consist of two houses of $45m^2$ and $75m^2$, respectively. We have used a TurtleBot 2 platform equipped with a Hokuyo URG-04LX-UG01 laser scanner to create the geometric map of the environment and for navigation, and a RealSense D435i camera to capture RGB images from each environment.

5.4.2 Semantic Segmentation Results

The first experiment shows the ability of the proposed method to assign a semantic label to parts of the environment. We demonstrate the feasibility of applying it in simulated and real-world scenarios. We are focused on showing how different sources of information affect the region's creation during the construction of a semantic map, considering only scene information, only object information and the combination of both. The chosen environments (two simulated and two from the real world) are typical houses composed of different room types such as living room, bedroom, bathroom, kitchen, and corridor. In addition, all environments contain objects that are commonly found in domestic scenarios.

Initially, the robot builds a map of the environment beforehand by executing the gmapping ROS package. As the robot moves, it captures images of the surroundings with the camera's sensor. These RGB images are later used to build the semantic maps. Fig. 5.4 shows the resulting semantic maps built by our proposed method. Given the four environments shown in Fig. 5.4(a) and Fig. 5.4(b) , the resulting semantic maps that include only scene information are shown in Fig. 5.4(c), the object-based maps are shown in Fig. 5.4(d) and the resulting maps combining scene and object information are shown in Fig. 5.4(e).

It can be seen that depending on the semantic information, different regions are formed in each environment. The arbitrary form of each region depends on the grouped cells of the same category and the shape of the environment itself. In this work, we focus on assigning a semantic label to parts of the environment to generate meaningful regions of interest within

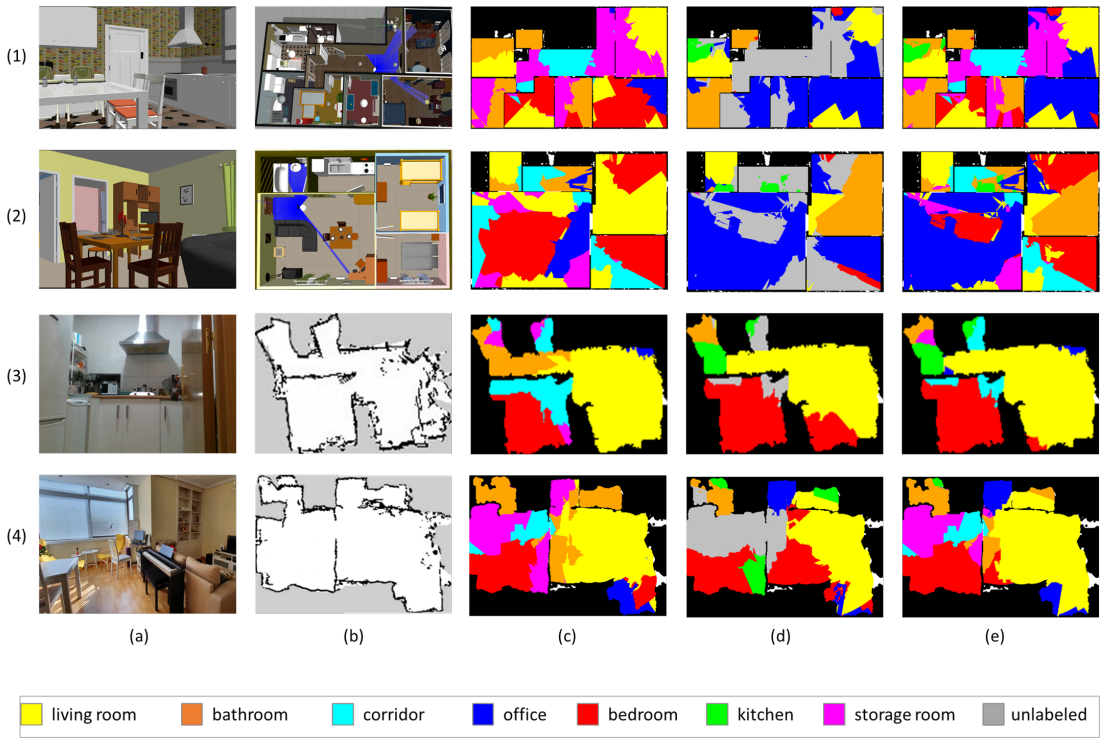


Figure 5.4: Resulting semantic maps. Rows (1) and (2) correspond to simulated houses. The last two rows (3) and (4) represent real houses. Columns (a) and (b) correspond to example images and maps of each environment. Each color on the map represents a different region category. Column (c) shows the semantic maps generated considering only scene information. In column (d) are the semantic maps built through only object information. Column (e) illustrates the resulting maps of combining both sources of information.

each room. That is why we do not apply any procedure to unify the regions, unlike [48, 138, 160, 186], to obtain a unique semantic label for each room.

Seven region categories are defined represented by different colors. These regions correspond to the scene categories defined for the scene classification model used in this work. When only scene information is used, a semantic label is assigned to each cell of the map and different regions are generated in each room. In a second case, when only object information is considered, notice that some areas in column Fig. 5.4(d) are gray; this is because the object detector has recognized no objects; therefore, cells that belong to these areas are marked as unlabeled. Then, when the fusion procedure is applied (Fig. 5.4(e)), the method takes the best of the scene-based and object-based maps to create a final map much rich in regions and therefore in semantic information.

We cannot compare our region's generation with a ground truth because it is not possible to define the limits of our regions such as in traditional room segmentation approaches where the physical room boundaries determine the size of a region and indicates if it is correct or not. That is why we have presented the resulting maps step by step to show the influence of each source of information in the semantic segmentation process. In the next experiment, we demonstrate the efficiency of our approach by its deployment in a more complex robotic task.

5.4.3 Performance in the Object Search Task

The second experiment is presented to prove that the representation of the environment obtained by the proposed method is capable of increasing the robot's efficiency when it performs more complex robotic tasks. To do so, we have selected the task of searching for objects. The robot is asked to search for an object that is not known in advance.

We compare the results with our work on [65], where the full rooms are assigned a single scene category. The searching strategy uses as input a 2D floor plan that includes the semantic room labels and occurrence statistics about objects and scenes to determine the most probable room where the target object can be found. The method selects the most likely room, then the robot moves to a random position within it that has the highest probability of finding the target object. If the object is not found, the robot moves to the next best position. The complete description of the search process is presented in Chapter 6.4.

To test the approach presented in this chapter, the search method uses our semantic maps based on regions as input. In addition, instead of selecting the most likely room, the method has been adapted to select the most likely region in which the object can be found. To make comparisons it has been considered the following metrics:

- number of viewpoints: count the locations the robot has visited until it finds the target object.
- number of segments: the segments are the rotations that the robot has to make in each viewpoint. This has to do with the horizontal field of view of the camera sensor.
- percentage of the covered area: represents the total area of the environment that the robot has explored until it finds the target object.
- distance traveled: total distance traveled by the robot until it finds the object.
- rooms visited : number of rooms visited by the robot until it finds the object.

For this experiment, six objects that can be found in domestic environments have been selected: chair, cup, bottle, TV, laptop, and bowl. These objects have not been seen in

advance by the robot and have been placed in the environment randomly, so the robot does not know their location beforehand. The complete approach of [65] has been applied as a baseline in a simulated environment and in a real-world one used in the previous experiment. Then, the search method was executed considering as input the different semantic maps generated by the method: the scene-based map, the object-based map and the resulting map when both sources of information are merged. The method was executed 18 times for each environment and method. An ablation study has been conducted achieving the results shown in Table 5.1.

Table 5.1: Ablation study : searching for objects using different semantic maps.

Environment	Method	Avg. viewpoints visited	Avg. segments visited	Avg. covered area (%)	Avg. distance traveled (m)	Avg. rooms visited
simulated	baseline [65]	5.481	21.759	25.042	23.898	2.889
	scene-based	4.241	14.370	11.678	25.934	2.917
	object-based	3.407	12.111	14.039	22.546	2.537
	fusion	**2.833**	**10.407**	**11.059**	**18.583**	**2.110**
real-world	baseline [65]	3.370	11.667	37.139	9.647	2.000
	scene-based	3.722	11.370	29.351	14.703	1.741
	object-based	3.981	10.019	24.540	12.466	1.907
	fusion	**2.315**	**6.241**	**22.790**	**7.506**	**1.333**

The results show that using the region-based semantic map that combines scene and object information (fusion) offers better performance in complex tasks such as searching for objects. In the simulated environment, the robot visited on average 2.83 viewpoints which represent 48.31% less than the baseline (5.48 viewpoints) that uses a standard semantic map divided into rooms as input. The number of segments and the area covered by the robot are also lower than the baseline. The covered area is significantly less than in the baseline, on average only 11.06% of the environment has been explored with our approach, which compared to the 25.04% resulting in the baseline, is equivalent to 55.84% less area.

Similar behavior is observed in the real-world environment. The viewpoints visited by the robot are on average 2.31 and the area covered until the robot finds the object is on average 22.79%. This represents 38,64% less area covered. Likewise, the distance traveled by the robot is on average 7.5 meters applying our environment representation, which indicates 22.2% less than when using the semantic map by rooms (baseline). Finally, the number of rooms visited by the robot in simulated environments is on average 2.11 rooms, and for real-world environments 1.33 rooms. Both results are lower compared to the results of the baseline

method. Fig. 5.5 shows the average number of rooms visited by the robot while searching for various objects in the simulated environment.

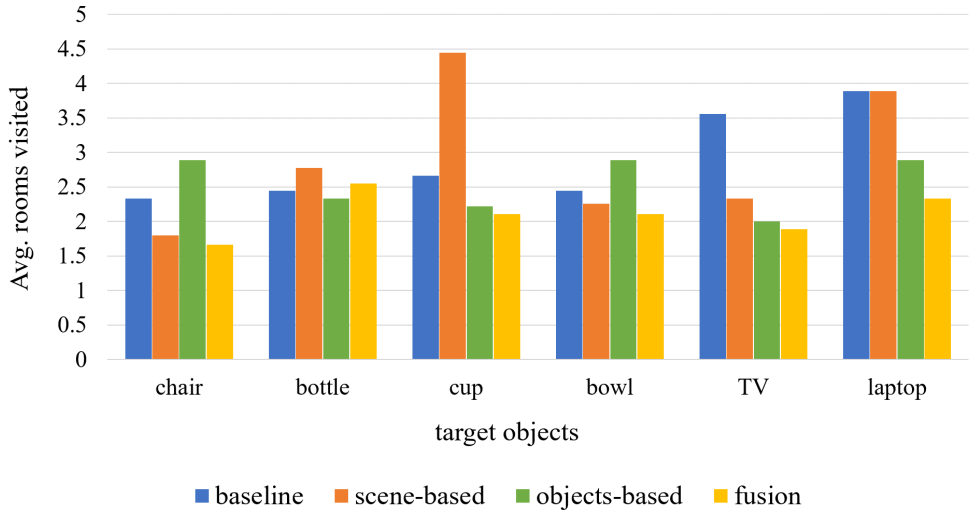
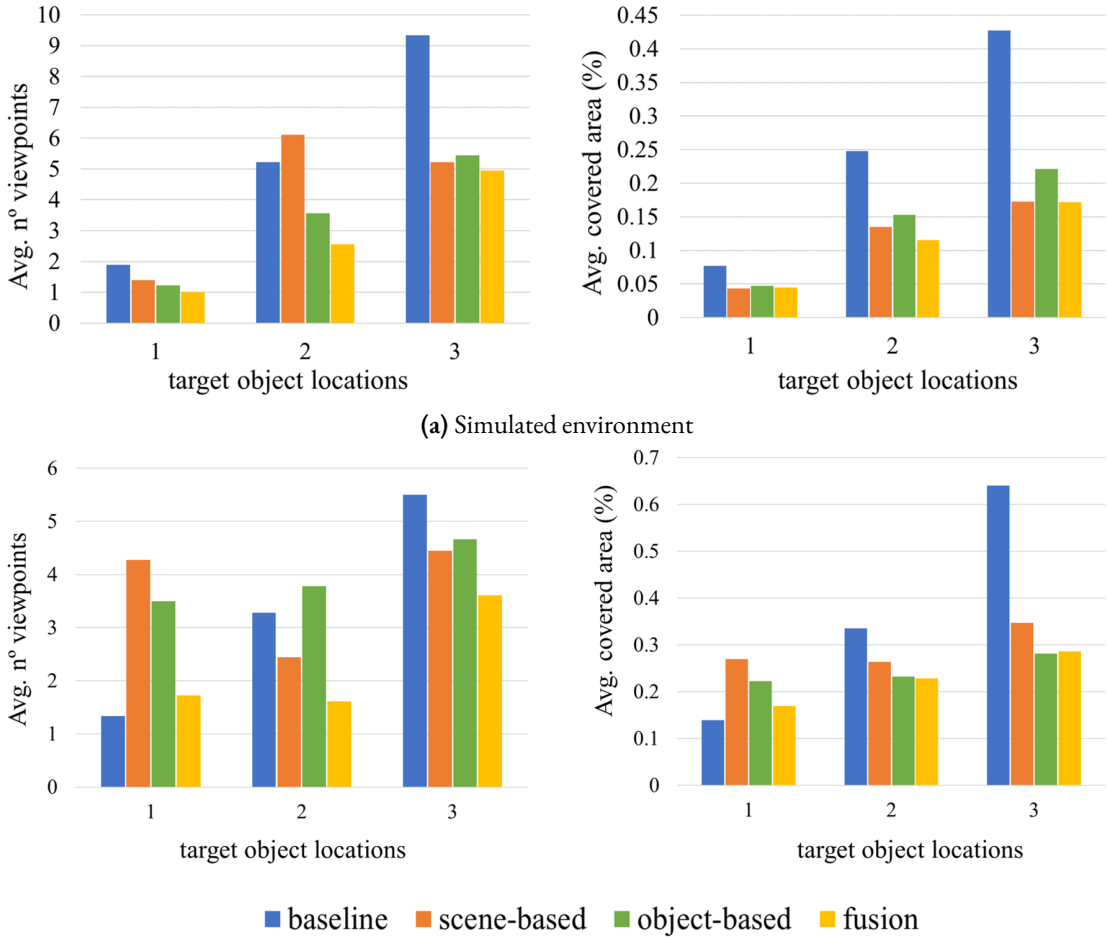


Figure 5.5: Average rooms visited by the robot during the searching process for each object in the simulated environment. Our approach (yellow) outperforms the baseline method (blue) during the object search process.

The evaluation shows that the map resulting from the integration of semantic information contributes to a better performance in the task of searching for objects, reducing the number of rooms visited and the area explored by the robot until it finds the object.

To increase the complexity of the search process and to further stress the method, the locations of the objects were changed three times, considering the most probable rooms where each object could be located (object-scene occurrences [65]). Firstly, the objects were placed in the most likely location, this is the easiest option. Then, in the subsequent runs, the objects were moved to the second most likely place that is less probable than the first. The third option represents a place with an even lower probability of finding the object. Fig. 5.6 shows the average final area covered by the robot and the number of viewpoints visited while it finds the objects in three defined locations for the two selected environments.

As the object is located in a less probable place, applying the baseline, the covered area increases, since the robot has to visit more viewpoints moving from room to room until it finds the object. The less likely the place, the greater the effort to find the object. However, when the object-based map and the scene-based map are combined to generate the final semantic map, the search strategy keeps the covered area at a low level. That is, even though the location of the object is not in the most likely room, the robot does not have to explore



(a) Simulated environment

(b) Real-world environment

Figure 5.6: Object search results considering different locations for each target object in a simulated and real-world environment. On the x-axis "1" represents the most obvious place where the object can be found. "2" is the second alternative to look for the object. And "3" is the third and the least likely place to find the object.

each room completely. Because our region-based semantic map contains different categories of regions within the same room, the robot only explores the region of interest within it. In other cases, the robot does not even have to move to another room to explore a different region while searching for objects.

In summary, our evaluations suggest that our method provides competitive results to be applied in more complex robotic tasks. We have demonstrated that our proposed method is capable of generating a useful semantic map based on regions for domestic environments,

and also, that our semantic representation of the environment can aid improve the robot's performance in tasks such as searching for objects.

5.5 Discussion

In this chapter, we have presented a new approach for dividing indoor environments into meaningful semantic regions. The key idea consists of keeping the miscategorizations of smaller regions within entire rooms. Different categorization systems may result in different confusing areas due to the intrinsic nature of the perception system. Therefore, we have also presented a framework for integrating region-based semantic maps built from different sources of information: one based on visual appearance and another based on the distribution of detected objects. Our approach has been implemented in simulated and real-world scenarios. The experimental results have proved the feasibility of applying our method to create a more information-rich representation of the environment based on regions. The results have also shown greater efficiency in the task of searching for objects in indoor environments compared to previous methods where rooms are treated as a whole.

6

Searching Strategies for Mobile Robots

This thesis focuses on understanding the environment by analyzing the elements and the relationships between them. We aim to endow robots with more capabilities to build knowledge that allows them to improve their performance in more complex robotic tasks. This thesis covers an application that combines all available information on the environment to demonstrate how a complete knowledge of the environment can improve the performance of other tasks. To do so, we select the task of searching for objects. This process considers prior information, semantic representation of the environment, and the relationships between known objects and the type of a scene. This chapter proposes a global and a local search strategy to find unseen objects in human indoor environments.

6.1 Introduction

The ability to search for objects can be considered an essential task and a precondition for carrying out other robotic tasks. Finding objects is highly relevant in showing the non-expert end-user that a robot can understand the world. Robots operating alongside humans are permanently confronted with changing environments. Besides modeling the robot's surroundings, a major challenge is that the locations of task-relevant objects are changing within a dynamic environment. Despite the efforts to equip robots with complex perception systems, modeling and updating the robot's world representation remains open research questions. Object search in human environments has drawn a lot of research attention

recently. Many service robotics activities, such as fetch-and-carry, require the robot to autonomously search for an object in a dynamic environment. Generally, a map of the environment can include some objects such as cabinets and tables. In contrast, the location of other objects such as cups, chairs, or laptops is often unknown and more easily changed.

In this chapter, we focus on a robot's ability to find possible locations of movable objects in human indoor environments. The search problem is tackled from two points of view: globally and locally. The global strategy aims to find the most probable room where the target object can be located. Complementary, a local search strategy is carried out within the most probable room to find the target object in a more precise location.

Given a query to search for a target object, several simple strategies can be applied, e.g., random or room-by-room exploration. However, such brute-force strategies do not consider multiple information channels that we as humans would use, which makes them inefficient and highly influenced by the robot's starting position, as well as the dimensions and complexity of the environment. For the global search strategy, we propose a multi-cue search method for objects in domestic environments that can appropriately combine different types of prior knowledge, such as semantic information and object-object/-scene relations. The goal is to obtain a probabilistic understanding of target objects' location that results in a more efficient search strategy. The core of the method is a Conditional Random Field (CRF) [101] and its ability to encode known relations between different observations.

Regarding the local search strategy, we propose a novel method for object search guided by co-occurrence probabilities that based on the most probable room, the search is refined by choosing specific places within the room to be visited so that the distance traveled is minimized. The most important part of the search process is analyzing and selecting the best locations called *viewpoints*. Therefore, a method that generates viewpoints from which the target object is likely to be seen is employed. Note that a viewpoint itself does not specify the robot orientation. However, promising robot orientations are identified as well to further speed up the search. If the object is not found, the probabilities are updated and the process is repeated to choose the new best viewpoint. Finally, the decision of the best viewpoint considers the maximum probability of finding the object with the minimum distance traveled.

The main contributions of this work are: first, the development of two novel object search methods combining semantic information about the environment with prior co-occurrence probabilities. Second, a global search method based on a probabilistic graphical model that can efficiently determine a set of the most likely rooms where a queried unseen target object can be found by exploiting co-occurrence statistics mined from online datasets. Third, we conduct a qualitative and quantitative evaluation on the Bosch Semantic Interpretation Challenge dataset [16] obtaining better performance concerning a baseline method. Besides, we integrate the global search method with a topological navigation

system [51] for performing the optimized search in order to show the benefits and feasibility of implementing our approach in mobile robots. Regarding the local strategy, the main contributions are: first, a probabilistic analysis and selection of best viewpoints to find objects in partially known environments efficiently. Second, a multi-objective optimization to maximize the probability of finding the target object and minimize the distance traveled. Finally, the last contribution is the implementation of our local object search strategy on a mobile robot and its evaluation in simulated and real-world environments.

The remainder of this chapter is organized as follows. Section 6.2 discusses the related work in the field of searching for objects and viewpoint selection. In Section 6.3, the proposed global search method is explained in detail. Section 6.4 describes the local search method and explains its individual parts in detail. Experimental results are discussed in Section 6.5. Finally, Section 6.6 summarizes the conclusions of the chapter.

6.2 Related Work

The interaction with objects has become crucial to build a semantic representation of the space [170]. Research with a focus on object search strategies has appeared in the literature since the 70s. Searching for objects can be classified into direct and indirect methods. Direct methods focus on finding the target object. In contrast, indirect methods first look for an intermediate object that is supposed to be related to the object that is actually being searched for. Also, it is assumed that this intermediate object is easier to find and that the target object is close to it.

Early works like [42, 177] propose to extend the search for one or several intermediate objects that commonly have a relation with the actual target object. More recently, the approach by Kunze et al. [100] proposes an indirect search method considering the distance and directional relations between objects. In their method, some relations such as in front of and left to are defined to calculate the probability of finding the object. Additionally, the method finds supporting surfaces such as tables, and after that, the target object is searched. The selection of viewpoints with a high probability of seeing the target object is based on the number of voxels of a supporting plane.

The approach by Elfring et al. [30] uses object-object relations to find the target object via a chain of intermediate objects. In the works by Chumtong et al. [21] and Aydemir et al. [5], the search process is based on object-spatial locations and relations between objects are modeled to obtain the probability distribution of the target object in a room. In the approach of Zhang et al. [196], the search process starts by looking for environmental objects (defined as objects with less movability) and then looking for the dynamic target object. The work by Loncomilla et al. [113] proposes a Bayesian-based method for searching through secondary objects. Even though this strategy can reduce the computational costs as well as the search

area, sometimes the spatial relationship between both target and intermediate object is not strong or does not exist. Furthermore, the difficulty of accurate object detection is still valid.

Other approaches address the problem of searching for objects through direct strategies. In the approach of Aydemir et al. [4], a search for specific objects is performed directly in large and unknown environments. Relations between objects and rooms are defined to build a probabilistic model of the environment, revealing the promising areas where the target object can be located. As exploration is a part of the problem presented, a cost function is used to decide where to explore and search simultaneously. The approach by Joho et al. [85] proposes a reactive search strategy for unknown environments that predicts promising directions for the robot based on decision trees. In the works by Shubina et al. [153] and Izquierdo-Cordoba et al. [80], an unknown environment is also considered. The proposed methods only consider as prior information the assumption that the target object is located on one of the tables of the environment. Also, in the approach by Shubina et al. [153], multiple cost functions are evaluated to reduce not only the distance traveled but also the number of actions performed by the robot. In the approach by Wang et al. [174], the object search problem is modeled as a Partially Observable Markov Decision Process (POMDP). Only object-room relations are used and are encoded in a belief map.

Veiga et al. [171] proposes a method using a semantic map with information about objects and their relations to obtain uncertainty estimates about the object location in a domestic environment. However, their experiments consider only the kitchen area of a house. In the work by Nie et al. [132] a method to search for objects in a simulated cluttered environment is presented. The model is based on the Markov-chain Monte Carlo method and includes the co-occurrence knowledge about which objects tend to be close to one another. The approach by Kunze et al. [99] incorporates ontological concepts and relations for reasoning about the locations of object classes. Experiments show increasing efficiency in the object search process by minimizing the movement cost and the number of processed images. In the approach by Toris et al. [164] a model based solely on temporal relationships between objects and search locations is proposed.

Other approaches, e.g., [143, 189], implement search models based not only on spatial and temporal information, such as attention models. In the work by Rasouli et al. [143], a search model based on visual attention is proposed. Visual salience stimuli are extracted to identify promising locations where the target object might be.

With the rise of deep learning techniques, some works apply more complex methods to find objects. In the approaches of Ye et al.[188, 189], the search problem is addressed through a deep reinforcement learning model that learns action policies to reach the target object. In [188] the target object has been previously seen and the semantic segmentation and the depth information are computed in each robot observation. Druon et al. [27] propose a visual navigation method based on previously detected objects' context information to calculate

the similarity to the target object. The main drawback of these methods is that they are time-consuming and demanding and sometimes fail due to the agent getting stuck in the same place repeating the same actions. Confusions between objects with similar semantics can also appear.

Regarding viewpoint generation and selection, in the work by McGreavy et al. [118], a set of poses reachable by the robot are generated. Then, a visibility analysis of the object candidate from each viewpoint is performed. The analysis considers the occlusions and the visible object features to define the best viewpoint. In other works, such as [4, 80, 100, 113], the viewpoints are randomly generated in the reachable space. Every viewpoint is computed until a covered area threshold is reached. In [113], the selection of the best pose is based on the maximum probability of finding the target object. In the approach by Kunze et al.[100], the full area covered by each viewpoint is evaluated until the object is found. In [80], the target object is expected to be placed over a flat surface and the selection of the best viewpoint considers the highest portion of the room covered by the viewpoint.

The methods mentioned above have certain drawbacks. The purpose of indirect strategies is to reduce the search space in order to decrease the computational complexity. However, searching for the intermediate object can sometimes be as difficult as searching for the target object itself. Besides, relations between intermediate and target objects are not always available. Approaches based on deep learning techniques require highly time-consuming training. Computational complexity also increases if the object detector needs to be continuously running.

In most of the approaches, a fixed number of viewpoints is generated in advance. Also, the pose selection strategies take into account only a single objective, e.g., the highest visibility of the object or the minimum time or minimum distance traveled.

In our models, RGB-D cameras to capture the environment in a more realistic way are used. The model contains information about objects and scenes, considering the relations between them as well as their influence on the task of finding new unseen objects. Furthermore, we propose a direct search method to infer the best viewpoints from which the target object could be seen. New positions of the viewpoints are generated in case the object is not found at the first attempt. Two criteria, the probability of finding the target object and the distance traveled, are optimized simultaneously through a cost function. The method is computationally lightweight, and it can be run on a low-cost mobile robot.

6.3 Global Strategy for the Object Search Task

In this section, a detailed description of the global strategy to search for unseen objects in indoor environments is presented.

6.3.1 Global Search Strategy Overview

Our approach is based on the assumptions that objects mainly occur in specific environments (e.g., a pan in a kitchen) and that objects often co-occur with other objects. Fig. 6.1 shows an overview of the proposed method [66]. As it can be seen, the core element is the fully connected Conditional Random Field (CRF) [101], which fuses the information about some previously detected objects, a semantic floor map, object-object/room relations to calculate the probability of finding the target object within the corresponding grid-cell.

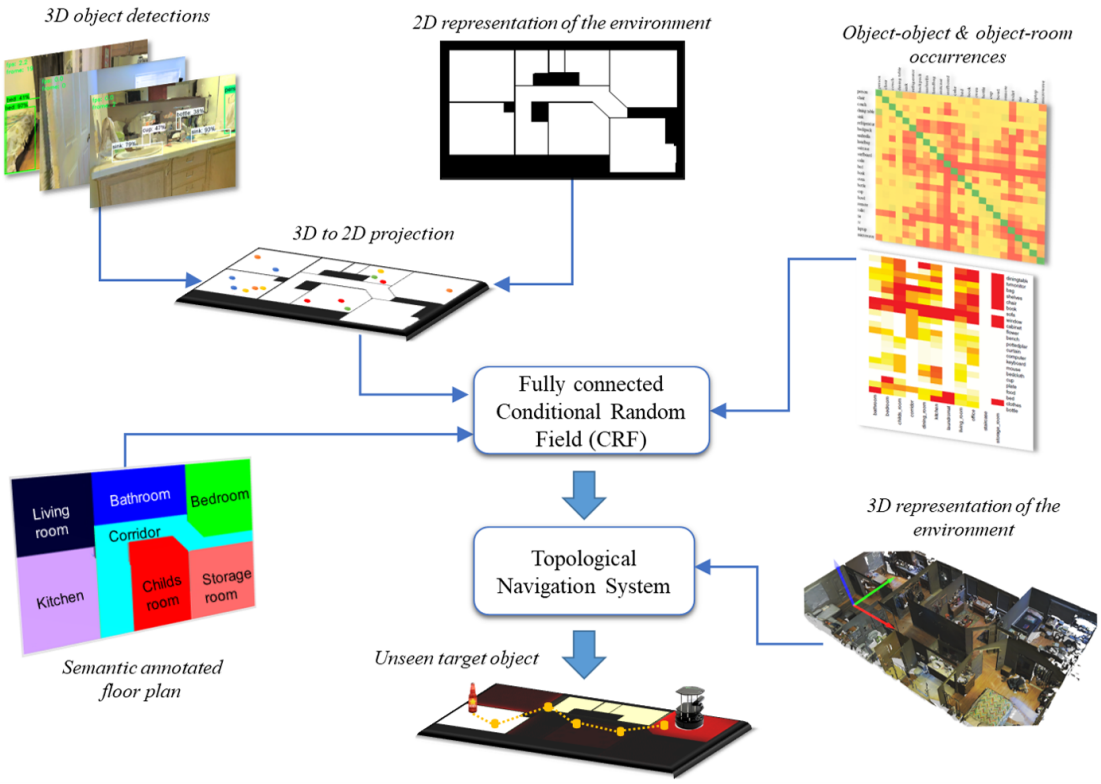


Figure 6.1: The proposed object search method. A 2D representation of the environment, 3D object detections and the semantic information of the environment are combined properly through a fully connected Conditional Random Field to obtain the most probable room where a target object is located.

Initially, the method assumes as input information: first, a 2D representation of the environment which is generated previously by the robot as shown in [1]. Second, we feed a room-wise semantic annotated floor plan, as proposed in [16], in combination with object-object and object-room co-occurrences into the CRF. The co-occurrences have been built

based on the NYU-Depth V2 dataset [154] and the COCO dataset [109].

Third, the 3D information of some previously detected objects is considered to feed our CRF. A Faster R-CNN object detector [145] is trained on a large amount of different object classes included in the COCO dataset. Thereby, given the robot's location and the corresponding depth image of the detector's RGB input image, the detected objects are projected from 3D (by taking the median depth in the bounding box) into the 2D floor-map of the environment. Finally, the resulting object-annotated floor-map is further input of the CRF. Then, during the inference process, the CRF outputs a heat map visualizing the most probable locations (rooms) for the unseen target object. The method prioritizes the high-probability locations during the search. Next, the information about the most probable room is incorporated into a topological navigation system to generate an optimal search plan through the environment to reach the desired room.

6.3.2 Building Object and Room Associations

Associating objects of daily use with certain categories of places facilitates the search for an object in a specific room context among other tasks [58, 191]. In the same way, some objects tend to be near or far from others. We include both concepts by exploiting object-room and object-object occurrences. While for the object-room relations the statistics presented in [16] based on NYU-Depth V2 dataset are applied, the object-object co-occurrences are built based on the COCO dataset [109].

The object categories that appear among our detections have been identified, and the probabilities that they occur close to other objects from this subset are computed. Working with co-occurrences generally implies the use of similarity measures to normalize the data. The associations between objects have been computed through the Jaccard Similarity Index, which allows comparing the similarity, dissimilarity, and distance of members for two sets [133]. Fig. 6.2 shows the object-object co-occurrences of the object categories selected for this work.

The Jaccard coefficient $J(\mathcal{A}_k, \mathcal{A}_l) \in [0,1]$ is the ratio between the intersection and the union of the two sets $\mathcal{A}_k$ and $\mathcal{A}_l$:

$$J(\mathcal{A}_k, \mathcal{A}_l) = \frac{|\mathcal{A}_k \cap \mathcal{A}_l|}{|\mathcal{A}_k \cup \mathcal{A}_l|} = \frac{|\mathcal{A}_k \cap \mathcal{A}_l|}{|\mathcal{A}_k| + |\mathcal{A}_l| - |\mathcal{A}_k \cap \mathcal{A}_l|}, \tag{6.1}$$

where $\mathcal{A}_k$ and $\mathcal{A}_l$ depict subsets of the set of training images $\mathcal{I}$ in which object k respectively l is detected. The higher the value of $J(\mathcal{A}_k, \mathcal{A}_l)$, the greater is the probability of the two objects k and l occurring close to each other. Through this, helpful semantic cues are obtained that are subsequently incorporated as inputs into our search method.

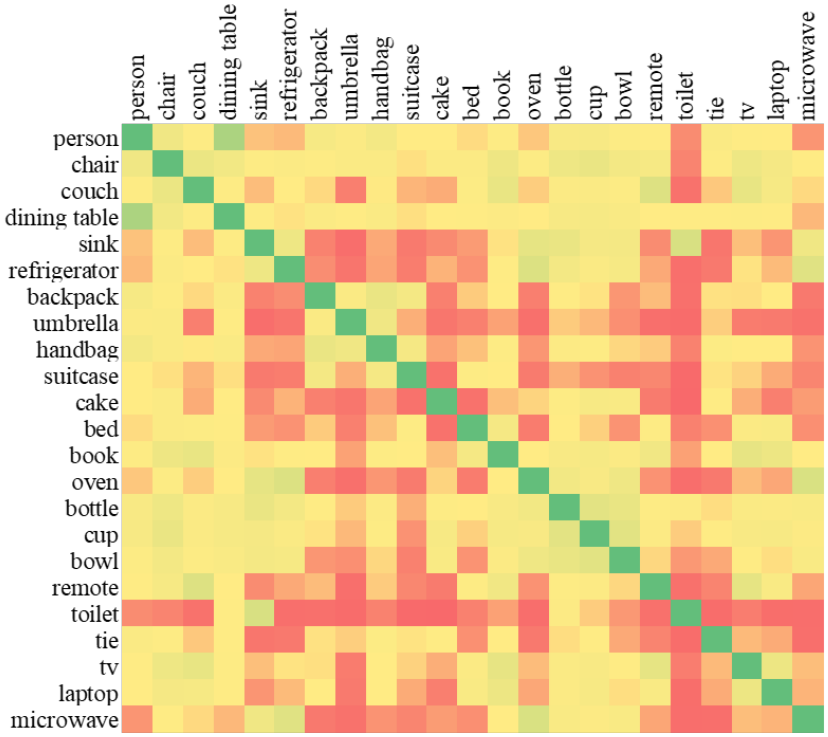


Figure 6.2: Object-object co-occurrences of the object categories detected in our dataset. Red colour represents the lowest value and green color the highest value. All remaining values get a colour based on the probability.

6.3.3 Modeling the Method to Search for Objects with CRF

In this section, we present a detailed explanation of how to fuse the different input data to obtain a robust estimate of the location of an unseen target object. Graphical models provide a natural way to represent the dependence of some variables with others which makes them suitable for this use case. A CRF [162] is a discriminative undirected probabilistic graphical model that considers known relationships (contexts) between observations to construct consistent predictions. Here, the model generates a pixel-wise prediction in the floor-map about the probability of the target object's location.

Consider a set of random variables $X = \{X_1, \ldots, X_N\}$ where X_i corresponds to pixel x_i of the 2D geometric floor-map $G \in \mathbb{R}^{mxn}$, obtained by [1]. The Gibbs energy function that characterizes a fully connected CRF to obtain the final probability of finding a target object in pixel x_i is:

$$E(X|G) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j)\,, \tag{6.2}$$

where i and j range from 1 to N. The pixel-wise unary potential is depicted as $\psi_u(x_i)$, while the pairwise potential $\psi_p(x_i, x_j)$ is modeled as mixtures of Gaussian kernels.

The *unary potential* is calculated independently for each pixel in G by fusing the object detector output and the semantic room annotation. Given a target object o_τ, the softmax output for a previously detected object o_s in pixel x_i is defined as $p(o_s)$. Hence the unary potential is:

$$\psi_u(x_i) = \begin{cases} p(o_s) * J(o_\tau, o_s) & if\, p(o_s) > 0 \\ J(o_\tau, r_m) & otherwise \end{cases}, \tag{6.3}$$

where $J(o_\tau, o_s)$ is the co-occurrence probability between the objects o_τ and o_s, and $J(o_\tau, r_m)$ is the occurrence probability of finding the target object o_τ in the room r_m.

The *pairwise potentials* represent the relationships between all pair of pixels, that in our model have the form:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} \omega^{(m)} k^{(m)}(f_i, f_j)\,, \tag{6.4}$$

where, similar to [95], $\mu(x_i, x_j)$ is a label compatibility function and $\omega^{(m)}$ a linear combination weight. The term $k^{(m)}(f_i, f_j)$ defines a Gaussian kernel where f_i and f_j are feature vectors for pixels x_i and x_j in an arbitrary feature space. We implement two kernels which are defined as:

$$\underbrace{\omega^{(1)} exp(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2})}_{\text{appearance kernel}} + \underbrace{\omega^{(2)} exp(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2})}_{\text{smoothness kernel}}\,. \tag{6.5}$$

These kernels have been defined in terms of the color vectors I_i and I_j and positions p_i and p_j. The parameters θ_α, θ_β and θ_γ control the weighting within a kernel and have to be set experimentally. The appearance kernel is a color-dependent term where the features are composed of the pixel location and the RGB pixel values. The second kernel, smoothness kernel, removes local isolated outliers within the map. The scalars $\omega^{(1)}$ and $\omega^{(2)}$ define the weights which must also be adjusted. In this way, the inference process is applied in the whole environment to merge all the potentials and thus obtain a 2D floor-map with the final probabilities in each pixel for the object sought.

6.4 Local Search Strategy in a Room Scale

Once the most probable room is determined, the next step should be to get closer to the target object. To do that, reducing the search space inside the room and minimizing the distance traveled by the robot are the main requirements to satisfy.

6.4.1 Local Search Strategy Overview

A schematic overview of the method is shown in Fig. 6.3. There are three inputs of the algorithm. The first one is a 2D floor plan previously built by the robot that includes the room layouts and the information about the position and class of the mapped objects. The second one is a probabilistic representation of the object–room and object–object co-occurrences. Finally, the method is given the class of the target object to be found, e.g., a cup. Even though the method can be used to search also for mapped objects, we focus mainly on the more challenging case when the location of the object sought is not known in advance.

When the search process starts, the semantic information included in the map and the prior co-occurrence probabilities are fused to create an initial probability map. Note that for a more intuitive understanding, top-view maps of the environment, as well as the probability maps calculated by the method, can be represented as images, where a pixel corresponds to a real-world square with a fixed size (e.g., $0.1 \times 0.1\,\text{m}$). If the target room to search for the object is not explicitly specified by the user, we start in the room with the highest probability of containing the target object (Section 6.3). The probability map of the room, which can be visualized as a heat map, encodes the promising areas of the room where the target object is likely to be present.

In the second step, a set of random candidate viewpoints is generated with the goal of maximizing the room coverage. Then, in the third step, the viewpoints are analyzed, taking into account the visibility model of the camera. It provides information about the minimum and maximum distances for the perception.

As a result, the probability of finding the target object in the area covered by each viewpoint is calculated. The area covered by each viewpoint is afterward divided into a given number of segments, according to the horizontal field of view (FOV) of the camera. The probability of the target object is in each segment of the viewpoint is calculated. The fourth step consists of selecting the best viewpoint and, in turn, the best segment. The best viewpoint selection maximizes the probability of finding the target object while minimizing the distance traveled by the robot. To achieve this, we have designed a utility function and automatically tuned its parameters.

Finally, the best segment is sent to the robot navigation system. When the robot reaches the desired pose, the object detector based on deep learning described in Chapter 3.4 is executed. If the object is found, the process ends. If not, the robot moves to the next best segment and

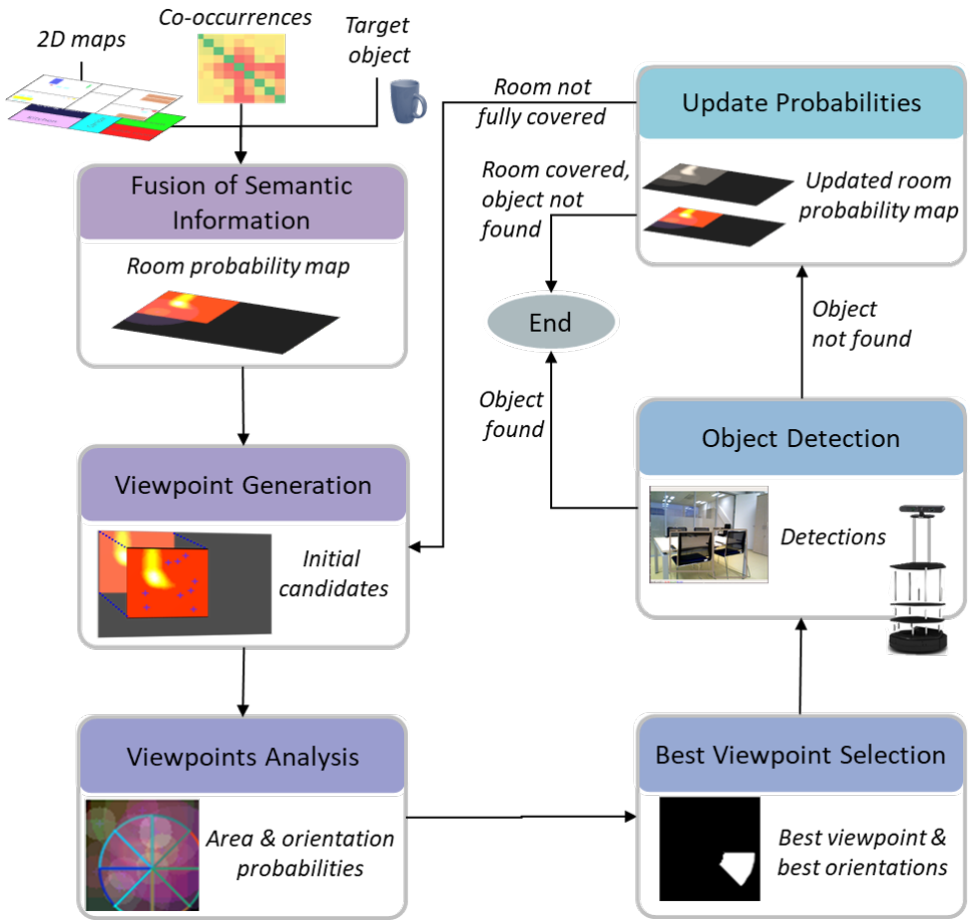


Figure 6.3: The local approach for searching for objects. First, the semantic information and prior knowledge are fused. Then, an exhaustive analysis of the generated candidate viewpoints is performed to obtain an optimal strategy to search for the target object. If the object is not found, the probabilities are updated and a set of new candidate viewpoints is generated.

attempts to detect the object again. If there are no more candidate segments available and the target object has not been found yet, the probabilities in the explored area are updated, and then the process returns to the second step to determine a new best viewpoint.

Also, the covered area is calculated and used as a termination condition. When the room is considered covered, the process ends. This may indicate that the target object is not in that room and the search process can start again in other rooms.

6.4.2 Initial Probabilities Assignment

An efficient object search approach is fundamental to reduce the search space and thus minimize the length of the robot trajectory, decrease power consumption and computational costs. A possible way to achieve this is by using prior knowledge about the environment [93].

Prior information available to the robot has several sources: a 2D floor plan of the environment, including the type l_j of each room r_j (e.g., a kitchen), and the information about the previously seen objects o_s in the room. The latter includes the object class c_s, the detection confidence $p(o_s)$ given by the object detector, and the space occupied by the object in the map.

This space is represented as a function $g(o_s, P_{x,y}) \rightarrow \{0,1\}$, which is equal to one if the object o_s is present in the pixel $P_{x,y}$ of the top-view floor plan and zero otherwise.

The last component of the prior knowledge are the co-occurrence probabilities. We use two publicly available datasets to obtain the co-occurrence probability values. The probabilities for the object–room co-occurrences, denoted $p(o_s|r_j)$, have been extracted from the NYU-Depth V2 dataset [16]. Object–object co-occurrences $p(o_s|o_{s'})$ for all pairs of object classes c_s, $c_{s'}$ were calculated using the COCO dataset [109]as is shown in Section 6.3.2.

Using the co-occurrence probabilities between the room type and the target object class, the most likely room r_j^* is selected:

$$r_j^* = \underset{j \in \{1,\dots,m\}}{\operatorname{argmax}} p(o_\tau|r_j)\,, \tag{6.6}$$

where m is the number of rooms in the environment. The robot will explore the room r_j^* first.

Next, we generate the room probability map, which identifies the most likely areas where the target object o_τ can be located. The probability of finding the target object in each pixel of the room is defined as:

$$p(o_\tau|P_{x,y}) = \begin{cases} \sum_{s=1}^{n} f(o_s, P_{x,y}) p(o_\tau|o_s) & \text{if } \sum_{s=1}^{n} f(o_s, P_{x,y}) > 0\,, \\ p(o_\tau|r_j^*) & \text{otherwise}\,, \end{cases} \tag{6.7}$$

where $f(o_s|P_{x,y}) = g(o_s, P_{x,y})p(o_s)$. The number of all mapped objects in the environment is n. The values of $p(o_\tau|P_{x,y})$ are normalized. Then, a Gaussian filter is applied around each detected object. As pixels are further from the detected object, they receive a gradually decreasing probability. Fig. 6.4 shows the room probability map for four target objects visualized as a heat map. The brighter the area, the more likely it is to encounter the target object.

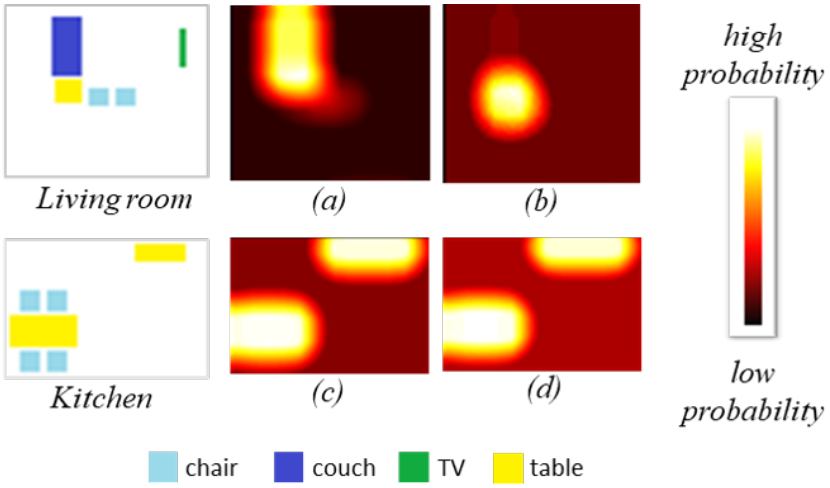


Figure 6.4: Room probability maps of four target objects: (a) *laptop*, (b) *cup*, (c) *bowl*, and (d) *bottle*. Darker areas represent lower probabilities, whereas lighter areas indicate promising zones where the target object can be located.

Once the initial room probability map is calculated, the next step is the generation of a set of candidate locations on the map, and the selection of the best one in order to execute a detector that locates the target object.

6.4.3 Candidate Viewpoints Generation

Viewpoint selection is one of the most important steps in the search strategy. Choosing the right viewpoint has a significant impact on the robot's performance. In this step, an initial set of N candidate viewpoints is generated in the room r_j^*. Each candidate viewpoint $V_{x,y}$ represents a position in the map. This process considers the hardware limitations of the camera: the minimum and maximum distance d_{min} and d_{max}, from which the sensor can perceive the objects, and the horizontal FOV of the sensor, see Fig. 6.5.

To cover the room with candidate viewpoints, we follow an iterative process. The process starts by placing an initial viewpoint at a random position in the free space $\mathcal{F}(r_j^*)$ of the room r_j^*. The free space is defined as:

$$\mathcal{F}(r_j) = \left\{ P_{x,y} \;\middle|\; \sum_{s=1}^{n} g(o_s, P_{x,y}) = 0 \right\} . \tag{6.8}$$

Next viewpoints are randomly generated in the free space while keeping a minimum Euclidean distance d_{min} from the previously generated viewpoints.

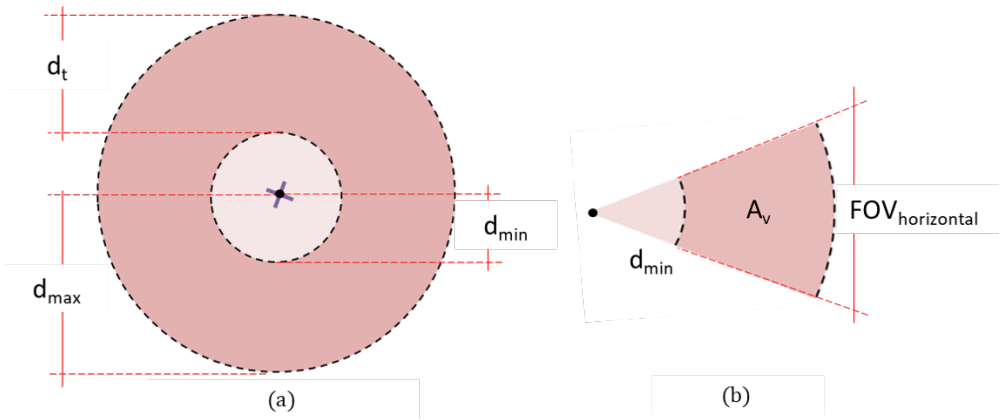


Figure 6.5: Visibility model of the camera. In (a), d_{min} and d_{max} represent the limits of the coverage area of each viewpoint. In (b), A_v denotes the search area covered by a particular segment of a viewpoint.

Each candidate viewpoint $V_{x,y}$ covers a set of pixels in r_j. The union of all pixels inside the detection zone determined by d_t, see Fig. 6.5, represents the area $\mathcal{A}(V_{x,y})$ covered by the viewpoint $V_{x,y}$. The total room area $\mathcal{R}(r_j)$ is defined as the union of all pixels $P_{x,y}$ belonging to that room. The coverage $C(r_j)$ for the room r_j is then calculated as:

$$C(r_j) = \frac{\left|\bigcup_{V_{x,y}} \mathcal{A}(V_{x,y})\right|}{\left|\mathcal{R}(r_j)\right|}. \tag{6.9}$$

The coverage $C(r_j)$ is updated after the insertion of each viewpoint. A threshold ρ determines when the room is assumed to be sufficiently covered. The viewpoint generation process terminates once $C(r_j) > \rho$.

6.4.4 Viewpoints Analysis

Each candidate viewpoint is analyzed to determine the most promising location from which the target object can be seen. First, the probability of finding the target object inside the area covered by each viewpoint is calculated:

$$p(o_\tau | V_{x,y}) = \frac{\sum_{P_{x,y} \in \mathcal{A}(V_{x,y})} p(o_\tau | P_{x,y})}{\left|\mathcal{A}(V_{x,y})\right|}. \tag{6.10}$$

This measure represents how good a certain viewpoint $V_{x,y}$ is for the search task. Next, we analyze the segments of each viewpoint. According to the visibility model of the camera, the

area covered by a viewpoint $V_{x,y}$ can be divided into Q segments $\theta_{x,y,q}$. In this work, $Q = 8$ segments have been used, given the $\text{FOV}_{\text{horizontal}}$ of the camera 58°. Note that the segments are partially overlapping, which is desirable for improved performance as the chances of detecting objects that are in the limit between two segments are increased.

The probability of finding the target object o_τ in a given segment $\theta_{x,y,q}$ of a viewpoint $V_{x,y}$ is defined as:

$$p(o_\tau|\theta_{x,y,q}) = \frac{\sum_{P_{x,y}\in\mathcal{S}(\theta_{x,y,q})} p(o_\tau|P_{x,y})}{\left|\mathcal{S}(\theta_{x,y,q})\right|} . \tag{6.11}$$

Analogously to viewpoint probability calculation Eq. (6.10), $\mathcal{S}(\theta_{x,y,q})$ denotes the set of pixels that belong to the segment $\theta_{x,y,q}$.

6.4.5 Best Viewpoint Selection

Using the information available at this point, the best viewpoint can be selected. To maximize the probability of finding the target object while minimizing the distance traveled, we adapt the utility function $U_{x,y}$ from [153]. It is calculated for a given viewpoint $V_{x,y}$ as follows:

$$U_{x,y} = p(o_\tau|V_{x,y}) \left(1 + \frac{\beta}{d(V_{x,y})}\right) , \tag{6.12}$$

where $d(V_{x,y})$ is the distance between the current (starting) pose of the robot and the viewpoint $V_{x,y}$.

In [153], the parameter β is set to 1 to relax the requirement of distance minimization. In our work, we have decided to apply the Levenberg-Marquardt algorithm [103, 117] in order to find the optimal β value. The algorithm was provided with a set of 100 measured data points. To obtain this data, the search method has been applied with different β values and looking for different o_τ. Using the ground truth location of the target object, the objective function is minimizing the distance traveled until the target object is found.

The best viewpoint $V^*_{x,y}$ is selected as follows:

$$V^*_{x,y} = \underset{V_{x,y}}{\text{argmax}}\ U_{x,y} . \tag{6.13}$$

The viewpoint $V^*_{x,y}$ specifies the most likely position in the map from which the robot is expected to be able to detect the object sought. To specify also the orientation of the robot, a set of best segments is determined, given an empirically chosen threshold σ:

$$\theta^*_{x,y,q} \in \left\{\theta_{x,y,q} \;\middle|\; p(o_\tau|\theta_{x,y,q}) > \sigma\right\} . \tag{6.14}$$

This way, the segments with lower probabilities are discarded, making the search process more efficient. The robot visits the best segments, as for mobile robots it is typically faster to

turn on the spot. As soon as the best segment is determined, it is sent to the robot navigation system. Once the robot reaches it, an object detector is run to identify the objects present in the segment.

6.4.6 Probability Map Update

In case the processing of the best viewpoint has terminated and the target object has not been found, the room probability map and the percentage of the room's covered area should be updated for a new iteration. This update allows for deciding whether to continue the search process in the current room or whether other actions, such as searching in another room, should be taken. If the object detector reports that o_τ is not found for a given viewpoint, the next viewpoint generation considers the areas already explored by the robot and the detection results in order to avoid looking at these parts again. In some cases, areas of the viewpoints may overlap. Because of this, the probability of finding the target object in the pixels belonging to an already explored viewpoint has to be reduced. We introduce a discrete time variable t and we define how $p(o_\tau|P_{x,y})_t$ decreases in a new iteration:

$$p(o_\tau|P_{x,y})_{t+1} = 0.5\, p(o_\tau|P_{x,y})_t\,. \tag{6.15}$$

Through Eq. (6.15) the room probability map is updated in each pixel. As a termination condition, the covered area of the room $C(r_j)$ is updated according to Eq. (6.9). In this work, the threshold ρ has been empirically set to 0.85 to declare when the room has been completely covered. It is not always possible to cover 100% of the environment, in particular in cases when the free space $\mathcal{F}(r_j)$ in the room r_j is limited by many obstacles. If the threshold is not reached, the search process begins a new iteration. This time, the new set of viewpoints is generated taking into account the current pose of the robot and considering the updated probabilities after an unsuccessful detection.

6.5 Experimental Evaluation

To evaluate the validity and efficiency of our approaches, the proposed search strategies have been tested in both simulated and real-world environments. In this section, details about the experiments and a discussion about the results are presented.

6.5.1 Evaluation of the Global Search Strategy

Experimental Setup

To evaluate the performance of the global search model, all the experiments have been conducted using the Bosch Semantic Interpretation Challenge dataset [16]. This dataset

consists of 10 apartments from real homes that contain for each a 3D mesh, rendered mesh views from various viewpoints, and a 2D projected ground truth floor plan with annotated room types: bedroom, bathroom, living room, dining room, storage room, kitchen, office, laundromat, child room, and corridors. As target objects, we selected seven objects, which normally occur in a typical house: chair, bed, bottle, cup, bowl, tv, and book. Since this dataset does not contain object annotations and, to our knowledge, no other datasets for the task of object search exist, the GT object locations are generated by applying a Faster R-CNN object detector and map onto the 2D floor plan. Fig. 6.6 shows qualitative results of the detector. As it can be seen, these are unstructured environments with many objects, sometimes superimposed, which makes the task for the detector more difficult.

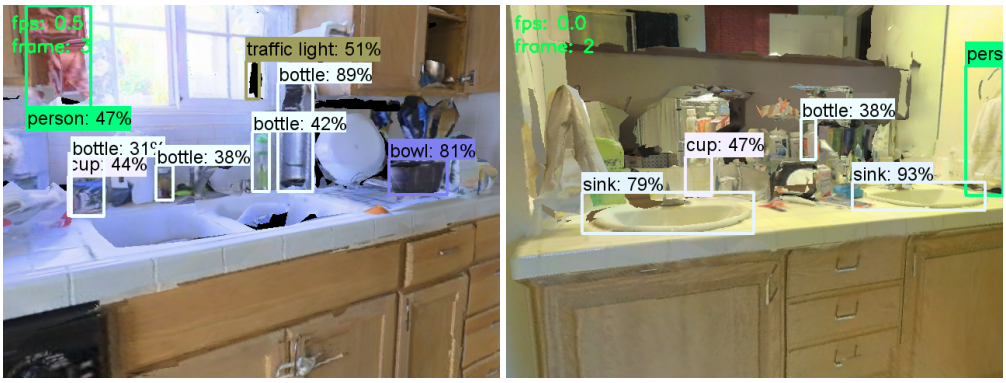


Figure 6.6: Faster R-CNN detections in the Bosch Semantic Interpretation Challenge dataset. The system gives the probability score and the label of the detected object.

Besides, we can not place new objects into the apartments of the dataset and scan them again, we simulate this process by removing an object from the previous detections and then ask to start searching for it. With this, our search method receives a modified map with some missing objects guaranteeing that the target object was not seen before. This way, we have an unbiased GT location for the object, and all the information is equivalent to it being placed there after the initial scan. In the experiments we use the library presented in [95] to implement our CRF based model. Also, the method has been integrated with a topological navigation system to obtain the best path to reach the most probable room.

Parameters Adjustment

In our experiments, the parameters: $\omega^{(1)}$, $\omega^{(2)}$, θ_α, θ_β and θ_γ of our CRF, Eq. (6.5) are set experimentally. Due to a good train-test split is not possible, as the dataset is very small, to adjust these parameters we have performed a grid-search study. We define $\omega^{(1)}$ as a parameter

between 0 and 1 and $\omega^{(2)} = 1 - \omega^{(1)}$ and the parameter $\theta = \theta_\alpha = \theta_\beta = \theta_\gamma$. For each value of $\omega^{(1)}$ and θ, we evaluate our CRF based method in all the apartments and target objects. Fig. 6.7 shows a heat map of testing our CRF based method with different parameter values. The darker the color the better are the results predicting the most probable room. For a wide range of values for $\omega^{(1)}$ and θ the method obtains good results.

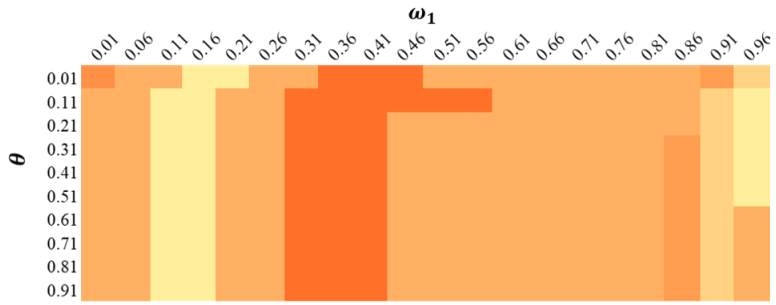


Figure 6.7: Heat map of the parameters used by the CRF based method. The heat map values correspond to the mean order of the GT room.

This study shows that the CRF based method is robust to the choice of the relative weight of the pairwise potentials, with a wide range of values performing well on all apartments, where $\omega^{(1)} \in [0.31, 0.41]$ would describe the range of values that work best for our method. The standard deviations have a very limited effect because we are ranking complete rooms; thus, the variance is not too sensitive. This suggests that the parameters can be intuitively selected with good generalization to unseen situations.

Quantitative CRF results

Fig. 6.8 shows an example of the proposed object search method. Three different objects (bottle, cup, and bowl), which were not seen in advance, are searched in three apartments (1, 2, and 5) of the Bosch dataset. First, the 3D detections of some previously detected objects are projected in the 2D floor plan (a). Based on the scene labeled maps (b), an inpainted process is applied to each room in order to eliminate the walls (c). Since walls do not contain valuable information, their unary potential is 0 which negatively affects the final outcome. Due to the fully connected characteristic of the CRF, the 0 potential would radiate into the room which would result in higher probabilities in the center of the room than closer to the walls. Then, the unary potentials are calculated (d), fusing the object-object/-room probabilities and the information about the previously detected objects. Through this, an initial probability map is obtained for the target object in each room. After that, the pairwise potentials are computed, resulting in a final heat map for the target object location (e). The

GT room used for comparisons can be seen in (f). Lighter colors on the heat map represent more likely locations for the target object. The darker areas represent lower probabilities to find the target object there.

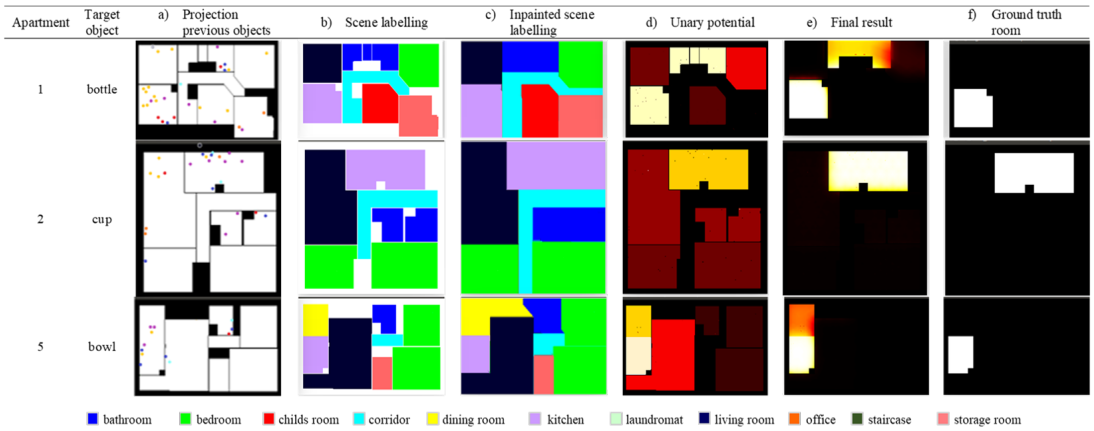


Figure 6.8: The proposed search method applied to three apartments searching for three unseen objects. a) shows the projection of some previously detected objects, b) corresponds with the scene labeling of the apartment, c) is the inpainted scene, d) the unary potential heat map, e) the final CRF heat map and (f) the ground truth created for comparisons.

To evaluate the influence of the potentials in our method, Fig. 6.9 shows the results of the ranking of the GT room for each target object in each apartment, by applying only unary potentials and incorporating the pairwise potentials into the search process.

The y-axis represents the frequency with which the GT room is identified as the most likely room to find the target object. The x-axis corresponds to the ranking of the GT room. In addition, Table 6.1 shows the results of evaluating our method using only unary potentials and when incorporating our CRF based method in the 10 apartments of the Bosch Semantic dataset.

As it can be seen, our CRF based approach generates better results, with the highest percentage of times that the GT room is detected in the first place. 65% of the times the most likely room corresponds to the GT room when our CRF based search method is applied. In addition, 75% of the times the GT room is classified as the first or second option to look for the object. On the other hand, 25% of the times, the GT room is classified as the first option to find the objects when the model only considers the unary potential information. Furthermore, the average of the final probability in the most likely room is 89.23% considering only unary potentials in Eq. (6.5), which increases to 97.16% when pairwise potentials are included in our CRF based method.

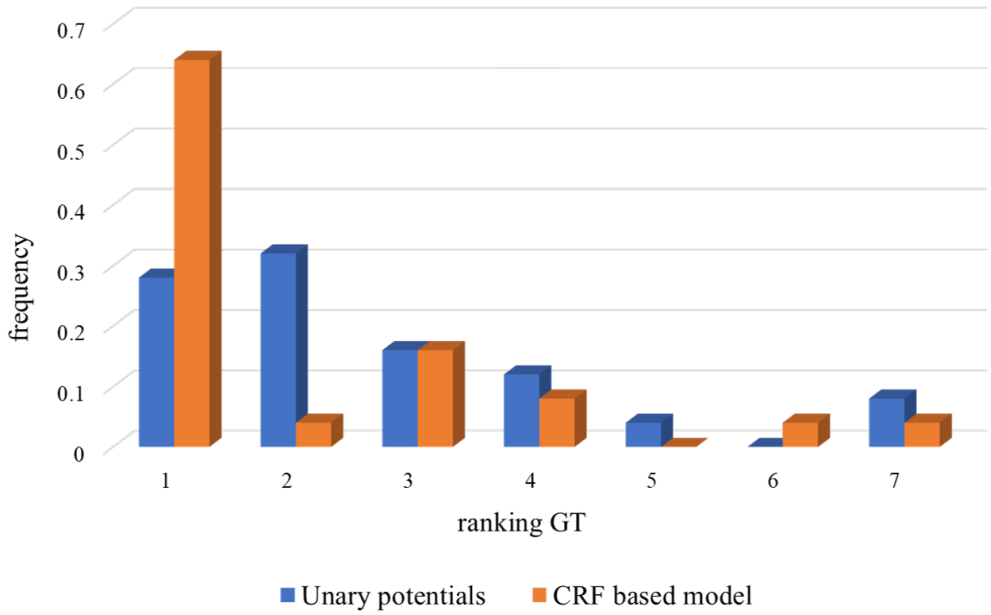


Figure 6.9: Histograms of the ranking of the GT room for each target object in each apartment. The GT room occupies the first position many more times when using our CRF approach compared to applying only unary potentials.

Table 6.1: Percentages of hits of the Ground Truth room for the 10 apartments used during the experiments.

Apartment	Unary potentials	CRFbased method
1	0.33	0.67
2	0.17	0.67
3	0.17	0.50
4	0.0	0.67
5	0.17	0.67
6	0.20	0.80
7	0.50	0.83
8	0.40	0.60
9	0.43	0.43
10	0.14	0.57
Avg.	0.25	**0.65**

Integration with a Topological Navigation System

In order to show the applicability and to evaluate the efficiency of our proposed method, we integrate it with the topological navigation system presented in [51]. The test has been carried out in the apartment 1 of the Bosch dataset using the Gazebo simulator. We have selected a differential drive robot (20*cm*x30*cm*), equipped with a Hokuyo URG-04LX-UG01 laser to map the environment. Three target objects were selected: a bottle, a bed and a chair. Initially, the robot is placed on the starting point to begin executing the search task. The most likely room information generated by our search method feeds the navigation system that calculates the optimal path to reach it. The path planning process is based on the Dijkstra algorithm. The decision about where to go is based on the maximum utility, that is, the robot chooses the room with the highest probability given by our CRF based approach. In the case of ambiguities, meaning several rooms with the same probability, the model considers the minimum path. Table 6.2 compares the search task by applying only unary potentials and incorporating pairwise potentials to find the target object. The distance traveled, the time spent on the search task and the number of steps through which the robot has to pass until it reaches the goal are calculated.

Table 6.2: Results of the search for objects through the topological navigation system.

Method	Unary potentials			CRF based method		
Target object	distance (m)	time (s)	steps	distance (m)	time (s)	steps
bottle	14.81	12.21	2	3.39	3.17	1
bed	27.55	15.64	3	12.20	8.06	1
chair	23.53	13.63	3	10.01	8.20	1

The results show that our CRF based method requires less distance and invests less time going to the room where the target object is located compared to using only unary potential information. Regarding the number of steps, in most cases using the estimates of the CRF method, the room with the highest probability corresponds to the correct room. Hence, the robot tends to make only one step to reach the goal. In other cases, when the correct room is not the most likely, the planner directs to the first room and then re-plans to the next most likely room and so on until it reaches the GT room. The results from the experiments carried out in this work demonstrate the effectiveness, robustness and validity of our approach to the task of searching for an object when information of the environment is appropriately considered.

Comparison with a Baseline Method

To obtain fair comparison results, the methods have to be evaluated in the same conditions, target objects and object poses. To the best of the authors' knowledge, there is no dataset available for comparison of object search methods in everyday human environments. To overcome this issue, we have designed a baseline method to compare with our CRF model using the Bosch Semantic dataset. While in the CRF, cues from several methods and information channels are merged, one could also think of a CNN learning implicitly a mapping from visual cues in an image to the probability of finding the target object within this camera view. Based on this idea, we implement a baseline method that consists of four main steps (Fig. 6.10).

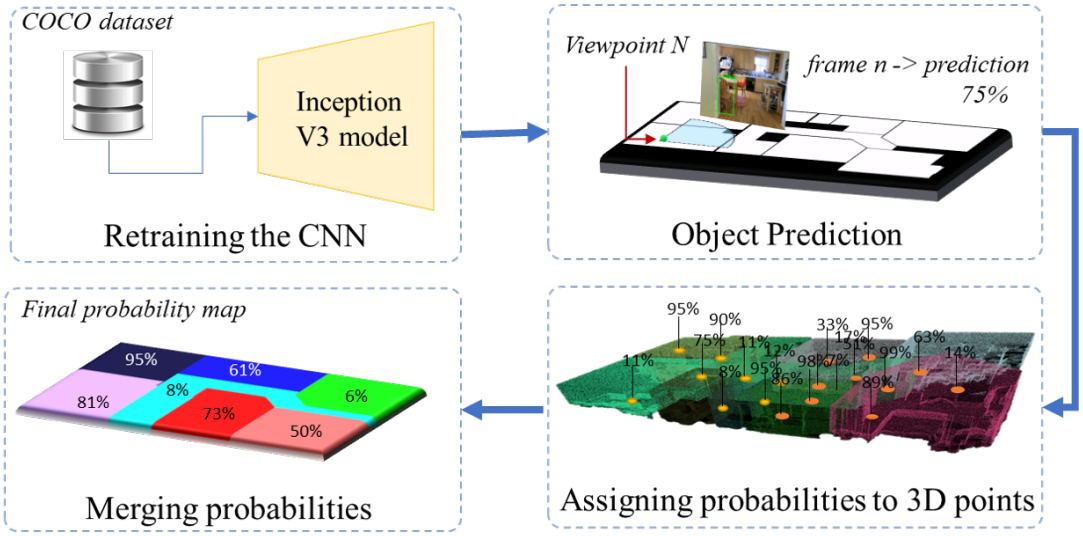


Figure 6.10: Baseline method. Using transfer learning a CNN is retrained to predict the most probable location of a target object in each of the 10 apartments of the Bosh dataset.

First, the fully connected layer of a pre-trained Inception V3 model is retrained on the COCO dataset. The dataset contains images with and without the target objects, leading to a binary classification problem. This way, the CNN implicitly learns the object-object/-room relations. Second, the CNN is applied to image frames within apartments of the Bosch dataset. As a result, the probability of the target object being in each image is obtained. Third, the predicted output for each frame is associated with the 3D point cloud according to the viewpoint. In the end, each seen 3D point of the cloud has a probability of the target object. Then, the probability of finding the target object in each room is obtained by merging the probabilities of the points that belong to each room. To do that, we apply a majority voting to

obtain the maximum probability found in each room. Through this, the highest probability of each room associated with the target object is obtained.

Fig. 6.11 shows the results of comparing the baseline method with our CRF based approach in the estimation of the most likely room where the unseen target objects can be located. This method has been executed on the same dataset and looking for the same target objects as described in section 6.5.1. As it can be seen, our CRF approach outperforms the baseline method, with the highest percentage of times that the GT room is detected in the first place.

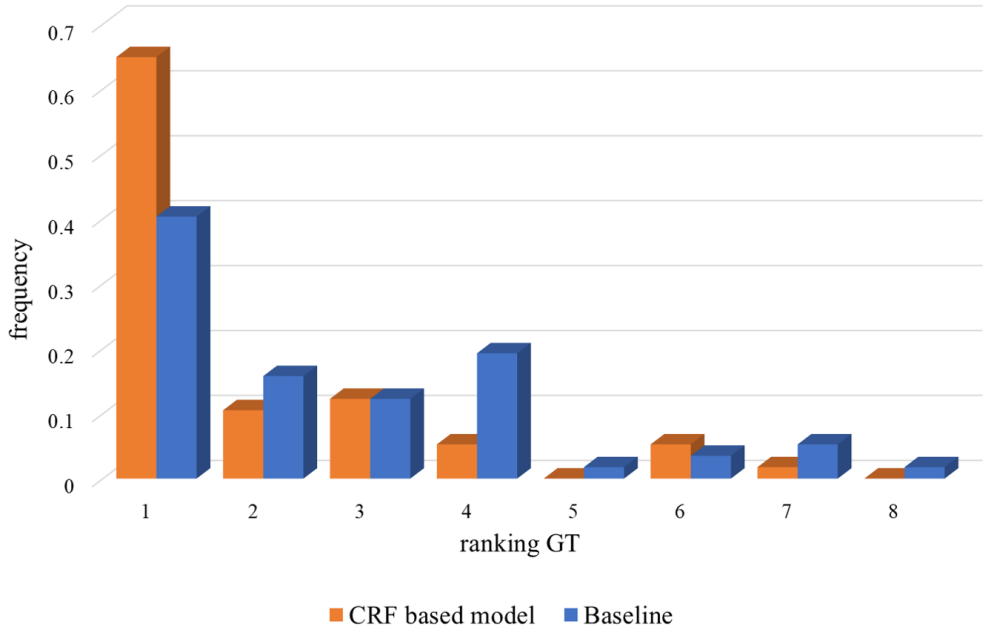


Figure 6.11: Comparison of the ranking of the GT room for each target object in each apartment using our CRF approach and the baseline method.

Table 6.3 shows the results of evaluating both approaches in the 10 apartments of the dataset. With the baseline method, only 42% of the times the most probable room corresponds to the GT room compared to 65% when the CRF based model is applied. Moreover, 75% of the times the GT room is classified as the first or second option applying our CRF based method compared to the 55% when the baseline method is applied.

In addition, the average of the final probability in the most likely room is 94.83% applying the baseline method and 97.16% building the estimates with our CRF based search method. Applying the baseline approach, in some cases the model predicts objects that are not present in the real detections. This is due to the method's characteristic training the network on

objects related to the context where ambiguities might occur.

Table 6.3: Percentages of hits of the Ground Truth room for the 10 apartments applying our CRF method and the baseline.

Apartment	CRF based method	Baseline
1	0.67	0.05
2	0.67	0.50
3	0.50	0.33
4	0.67	0.67
5	0.67	0.33
6	0.80	0.80
7	0.83	0.67
8	0.60	0.40
9	0.43	0.43
10	0.57	0.14
Avg.	**0.65**	0.42

6.5.2 Evaluation of the Local Search Strategy

Experimental Setup

We have selected the widely-used mobile robotic platform TurtleBot 2 to perform the real-world experiments. It is equipped with an ASUS Xtion Pro Live camera and a Hokuyo URG-04LX-UG01 laser scanner. For object detection, we have implemented the object recognition framework based on deep learning presented in Chapter 3. The model architecture is based on ResNet-101[64] and it has been trained with the COCO dataset [109]. The components of the framework are integrated through the middleware ROS. To build a map of the environment, the ROS gmapping package has been used. For path planning, the Adaptive Monte-Carlo Localization (AMCL) algorithm has been implemented.

Evaluation in Simulated Environments

We set up a simulated environment inspired by real-world homes that includes common objects and different types of rooms. The 12 m $\times$ 8 m environment was created using Gazebo simulator (Fig. 6.12) and consists of a typical house with 6 different rooms: a bedroom, a

child's room, a corridor, a bathroom, a living room, and a kitchen. We have tested the method with four target objects: a laptop, a cup, a bowl, and a TV.

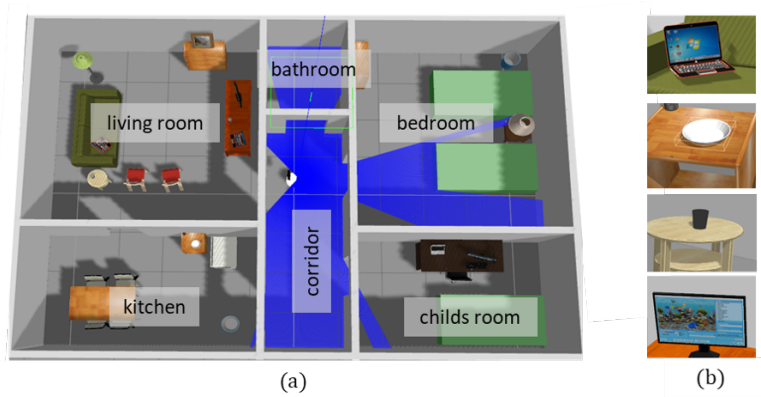


Figure 6.12: The simulated home environment used in the experiments. (a) A home environment with six rooms. (b) The target objects selected for the experiments.

We explain the working of the method on an example of searching for a cup. Fig. 6.13 illustrates the steps to select the best viewpoint during the search process. At first, the robot has determined the room that most likely contains the object sought and is standing at its entrance. The prior knowledge about the object–object co-occurrences and the information about the mapped objects are fused into a room probability map. The map provides the initial prediction of the promising areas where the target object can be located. In Fig. 6.13(a), the lighter areas are the most probable locations for the target object. In this example, the cup is on a small dining table in the living room. Then, a set of random viewpoints is iteratively generated considering the distance between the viewpoints and the covered area of the room (b). After that, the viewpoint analysis begins. The probability of finding the target object in the entire viewpoint area is calculated (c). Similarly, the probabilities of finding the target object in each segment of the viewpoint are determined (d).

Through the utility function maximization, the best viewpoint is chosen. Next, the set of best segments is determined. Then, the best viewpoint is sent to the robot navigation system. The robot moves to the given viewpoint and adjusts its orientation to visit the best segment. If the object is not found, the probabilities in the explored area are updated. Fig. 6.14 illustrates the sub-tasks of the robot during the search process. In (a) and (b), the target object, the best viewpoint and the best candidate segments are shown. In (c), a representation of the covered area of the room after the processing of the first best segment can be observed. In this case, the covered area was 10.3%. In (d), the results of the object detector are shown.

Table 6.4 shows the results of the proposed method. The starting robot pose is at the

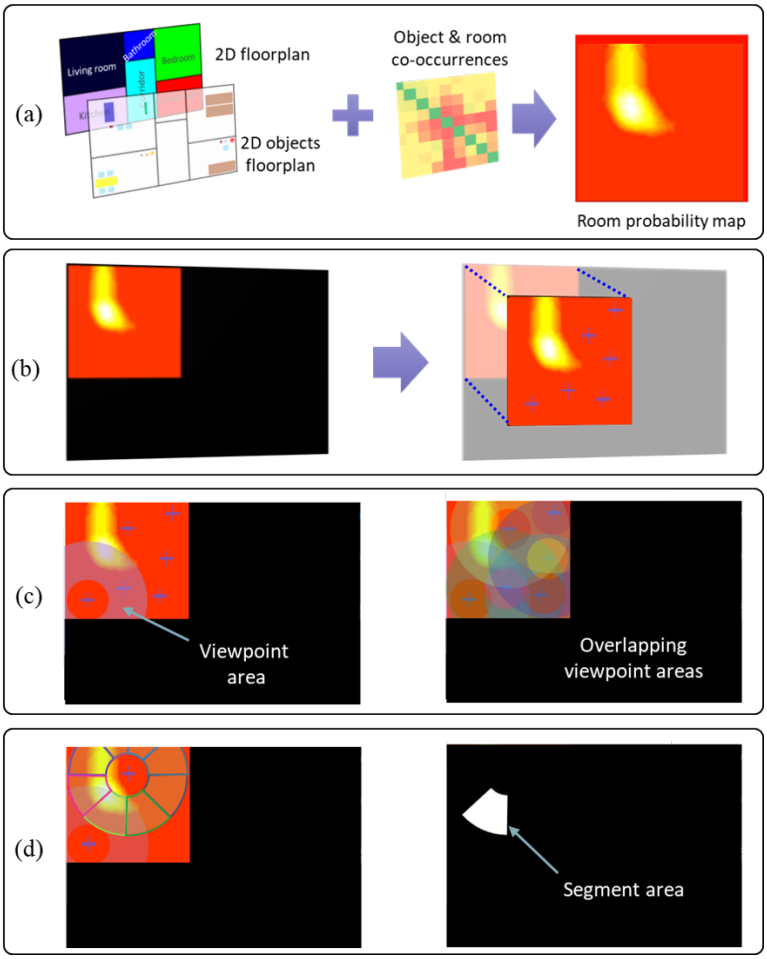


Figure 6.13: The proposed search strategy operating in a home simulated environment. The most probable room is the living room. (a) The prior information is fused to generate a room probability map. (b) The random viewpoints are generated inside the room. In (c) the analysis of the all viewpoint areas is conducted. Finally, (d) shows the best candidate segment.

entrance to the room. We calculate the covered area of the room, the time spent and the total distance traveled by the robot until it finds the object. Total viewpoints visited and the number of segments explored during the search process are also counted. We have repeated the search for each of the four target objects three times. The results show that the method limits the search area through an analysis of the viewpoints and only the most promising areas are considering for searching. On average, the search process takes 157.87 seconds and the covered area of the room is 19%.

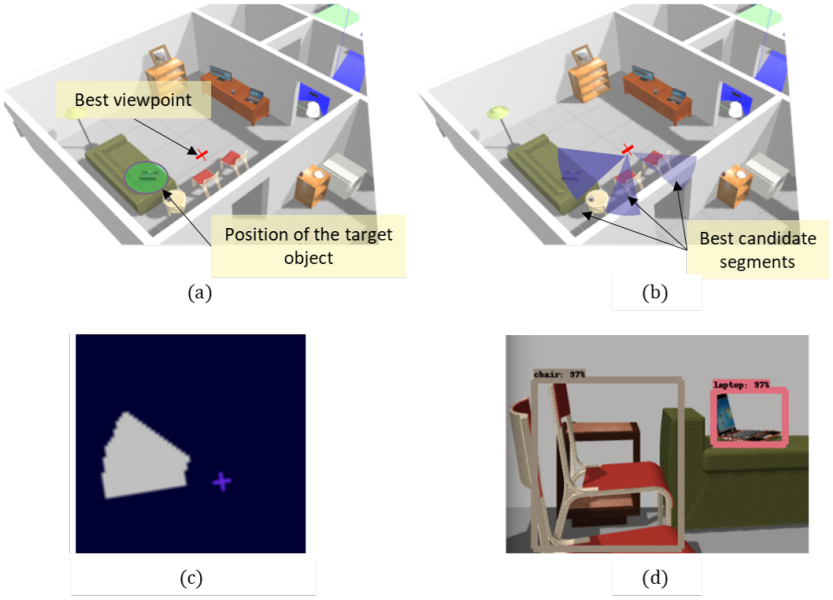


Figure 6.14: Illustration of the sub-tasks in the search process. In (a), (b) the best viewpoint and the best candidate segments are chosen; (c) the covered area of the room after exploring the best segment; (d) object detections.

Table 6.4: Evaluation of the proposed strategy during the search for four target objects in a simulated home environment.

Target object	Time (s)	Distance traveled (m)	Covered area	Total viewpoints	Total segments
laptop	120.66	1.27	0.21	1	2
laptop	63.43	1.18	0.10	1	1
laptop	82.20	1.11	0.10	1	1
cup	270.50	2.44	0.30	2	3
cup	245.50	2.04	0.16	1	2
cup	451.22	2.75	0.29	3	3
tv	63.01	0.56	0.19	1	1
tv	91.55	0.75	0.16	1	1
tv	118.24	1.24	0.21	2	2
bowl	152.59	0.92	0.19	1	2
bowl	135.27	1.02	0.10	1	1
bowl	186.42	1.94	0.25	2	2
Avg.	157.87	1.44	0.19	1.42	1.75

Evaluation in Real-World Environments

The experiments were carried out in a living room of 14 m × 4 m. The room was first mapped and the information about some objects in the room such as tables, chairs, and sofas was therefore available to the robot. Fig. 6.15 shows the environment and the target objects selected for the experiments. The robot has to search for two different objects: a cup and a laptop, each of them placed at two different locations in each execution.

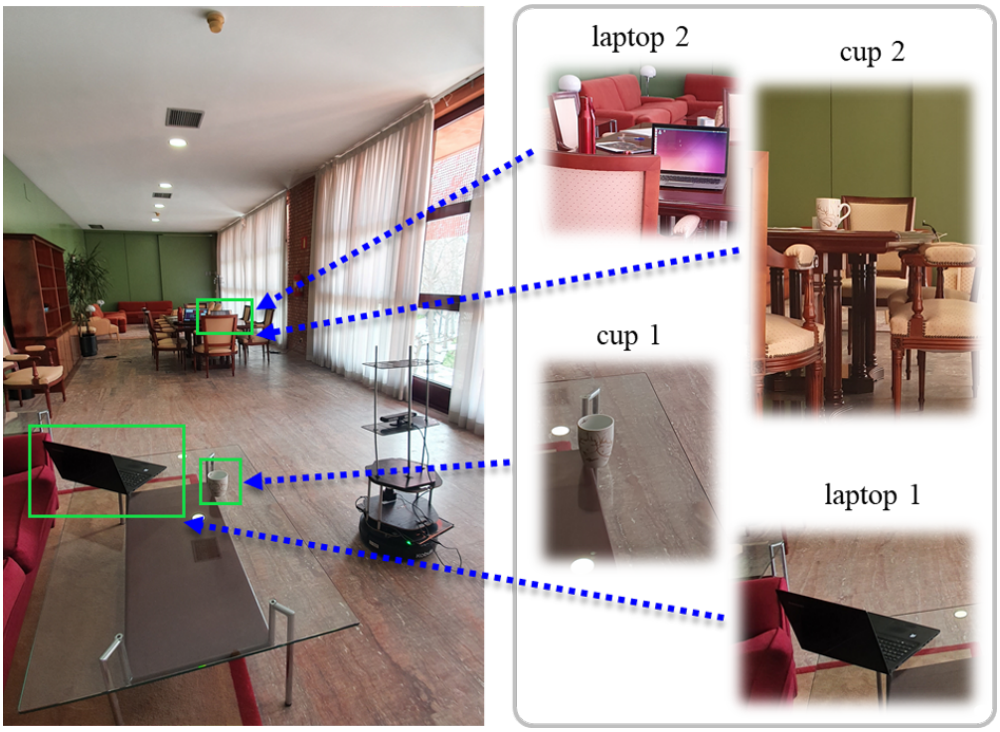


Figure 6.15: The real environment used in our experiments. The robot is asked to find a cup and a laptop in different locations of the living room.

In Fig. 6.16, the search for a cup is illustrated: (a) shows the map of the environment, (b) shows the room probability map for the target object, and (c) the covered area of the room.

Additionally, Table 6.5 shows the results of evaluating the proposed search strategy in the real-world environment. In some cases, the detector fails to recognize the object, although the target object is within the field of view of the camera. This forces the robot to explore another segment or a new viewpoint from which the object detector is able to identify the object. Despite this, the results demonstrate the feasibility and the efficiency of the proposed method for the task of searching for objects in real scenarios.

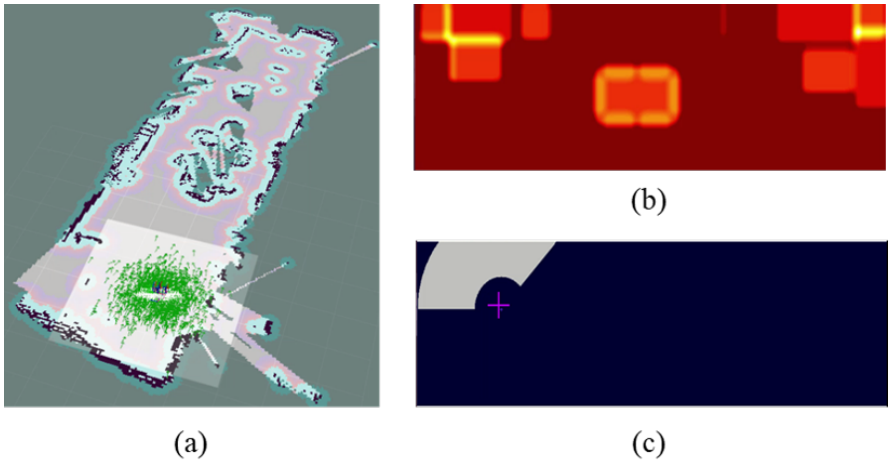


Figure 6.16: Execution of proposed search strategy. (a) The map of the environment and the execution of the path planning. (b) The room probability map built for the target object cup. (c) The covered area of the room (gray) after the evaluation of the best viewpoint.

Table 6.5: Evaluation of the proposed local search strategy in a real-world environment.

Target object (o_τ)	Time (s)	Distance traveled (m)	Covered area	Total viewpoints	Total segments
cup 1	275.34	0.95	0.18	2	2
cup 2	140.71	1.94	0.19	1	2
laptop 1	139.11	0.36	0.10	1	1
laptop 2	383.44	1.05	0.20	2	3
Avg.	234.65	1.08	0.17	1.5	2

Comparison with Other Approaches

Quantitative Comparison: to the best of the authors' knowledge, there is no dataset publicly available that would be specifically designed for comparison of object search methods in human-inhabited environments. To obtain valid results in a comparison, the compared methods have to be evaluated under the same conditions of the environment, target objects, and the position of the objects. To overcome this issue, we have built a baseline approach to object search based on random viewpoint selection. The method has been executed 12 times in the same simulated environment and looking for the same target objects as in Fig. 6.12.

This method does not take into account any prior information to select the best viewpoint. Table 6.6 shows the results of the baseline method compared to our proposed search strategy.

Table 6.6: Comparison between the proposed search strategy and the baseline search method.

Search method	Avg. time (s)	Avg. distance (m)	Avg. coverage	Avg. viewpoints	Avg. segments
random	395.31	4.41	0.28	2.16	9.09
our approach	157.87	1.44	0.19	1.42	1.75

The results show that the random strategy is substantially less efficient than the proposed method. On average, finding an object through the random selection of viewpoints takes 395.31 seconds and the average distance traveled is 4.41 meters. On the contrary, with our search strategy, the task takes an average of 157.87 seconds with a distance traveled of 1.44 meters. Our approach allows determining the best viewpoint and the best segments to search for an object in a more efficient way. The number of viewpoints explored and the covered area are also lower than those obtained with the random method. An appropriate viewpoint selection reduces the search space and the robot trajectory, yielding better results in less time.

Qualitative Comparison: in [100] and [80], the authors associate the target object with supporting planes to reduce the search space. The search space is limited to objects that can be found, for example, on tables. If the objects do not have associated supporting planes or if the tables are not identified, then the target object can not be found. In our work, we incorporate object–room and object–object co-occurrences to generate promising areas where a target object can be, regardless of whether they are in supporting planes or, for instance, on a chair or a sofa. In [196], an intermediate object with a strong relationship with the target object is found first to guide the search process. The problem is that sometimes detecting the intermediate objects can be as challenging as identifying the target object. In our approach, the robot searches for the object directly, based on a probability map of the room. Our efforts are focused on an analysis of the viewpoints to determine the most promising poses. Besides, the room probability map is updated after visiting a viewpoint if the object is not found, which influences the selection of the new best viewpoint.

As for the viewpoint generation, in [100] the number of random viewpoints is set to 20. Similarly, in [80], the fixed number of viewpoints is 1000, assuming that the area of the rooms is always the same. In our method, we build a set of random viewpoints, in which the size of the set depends on the covered area of the room. Therefore, our method ensures that the room is completely covered and that the amount of points is sufficient to carry out the search task. Another contribution of our work is the optimization of the utility function.

Our utility function is adapted from [153], where the authors set the β value to 1 to relax the distance minimization. In contrast to that, we use an optimization algorithm to find the β value to satisfy not only the maximum probability of finding the target object but also the minimum distance traveled until the object is found.

6.6 Discussion

In this chapter, we have demonstrated how full knowledge of the environment can improve more complex robotic tasks. We selected the task of searching for objects as it is considered a precondition for performing many robotic tasks and is highly relevant in showing the non-expert end-user that a robot can understand the world. With this work, we follow the idea of incorporating prior semantic information to obtain efficient search strategies to find target objects that have not been seen before in human indoor environments. First, we proposed a CRF based method to consider additional cues which influence the robot's understanding of its environment, namely the object and context relations as well as semantic information. The presented experiments demonstrate the usefulness and efficiency of our method to search for objects based on fully connected conditional random fields compared to other approaches, such as the use of the unary potentials and methods based only on current deep learning techniques. In addition, the global strategy has been tested in different real apartments, which demonstrates the flexibility to apply it in different environmental conditions.

Then, we have presented an efficient search strategy to find unseen objects on a room scale. Prior knowledge in the form of object-object and object-room co-occurrences has been employed to build a room probability map with the most promising locations of the target object. Our method is based on the selection of the best viewpoint through an analysis of the probabilities of finding the target object in the area covered by the viewpoint. In addition, the best orientations within the viewpoint are determined to further speed up the search. An optimized cost function is used to maximize the probability of finding the target object while minimizing the distance traveled. The method was evaluated in simulated and real-world environments, demonstrating its validity and efficiency on the task of finding unseen objects. Finally, a quantitative and qualitative comparison has been presented, showing the advantages of the proposed local search strategy.

7 Conclusions and Future Work

In this thesis, we have proposed several methods that allow robots to acquire more information from the environment to progressively build knowledge that helps them perform better in more complex robotic tasks. This chapter summarizes the conclusions and main contributions of each chapter. We also propose some lines derived from this thesis for future research.

7.1 Conclusions

Understanding the environment is a crucial skill for a service robot operating in human environments. In this thesis, we consider indoor environments, which means that the robot has to inhabit environments made for humans by humans. These environments are characterized as unstructured, cluttered, with high variability, different layouts, non-controlled environmental conditions, such as lighting conditions, and different types of objects. A robot has to be able to interpret information about the environment in order to interact with it. In this dissertation, we tackle the problem of scene understanding for service robots. The robot has to identify the elements present in the environment, the relationships between them, understand their meaning and uses in order to define the actions that it can perform later. This knowledge is progressively built by adding information about the environment through visual perception. The main contribution of this thesis aims to develop and implement different methods and techniques to provide the robot with capabilities for

a better understanding of the environment. Furthermore, this knowledge can be used to enhance the robot's performance in high-level robotic tasks.

Throughout the chapters of this thesis, we presented methods to obtain meaningful information about the environment, mainly from the existing objects and the scene's global features. We used the information gathered to create representations of the environment useful for other robotic tasks. Likewise, we developed applications where the knowledge of the environment is vital to perform other tasks efficiently. The work presented in Chapter 3 focused on object recognition methods applied to mobile robots operating in human indoor environments. These environments were not modified in terms of lighting conditions, object location, and occlusions. In our developments, we keep in mind that object detection must be deployed in real human environments, considering the dynamism of the scene, with a robotic platform moving continuously, and considering the noise of the sensors. Our primary motivation was to provide the robot with different systems to gather as much environmental information as possible by detecting objects. That is why we start with implementations based on machine learning techniques and finish with more modern strategies based on deep learning techniques.

Based on machine learning, the first method implemented an uncertainty model to improve the detection confidence of different objects in real-world environments. The model considers that the data is not perfect and is never exactly as measured. In this way, the system allows getting the information of each object's category and detection confidence, as well as the distance and angle from the center of the sensor to each detected object. Through the experiments, we demonstrated the feasibility and usefulness of implementing the proposed system into a real mobile robot, performing the detection in real-time.

Then, we implemented an object detection model based on the Fast-RCNN architecture, taking advantage of deep learning techniques. The model allows identifying more objects, in this case, 80 object categories, and was adapted to work in real-time. The model outputs the object class, the detection confidence, and the bounding box's coordinates of the detected object. Through the experiments conducted in real-time, with a mobile platform in domestic environments, we demonstrated the feasibility of employing this method in real-world scenarios. The introduction of deep learning in object detection approaches has demonstrated better results than using machine learning techniques because of their robustness and capability to detect more objects in the environment. The increasing advances in the hardware imply these approaches guarantee one of the crucial conditions when we work in real scenarios: the speed and the operation in real-time.

In Chapter 4, we addressed the process of identifying a scene in an indoor environment. Given an image, it is not only important to identify the objects in the scene but also the environment where the objects are and the different actions that can occur. Assigning a semantic label to a scene, i.e., kitchen, bedroom, bathroom, implies knowing the objects

around and the relationships between them. First, we proposed a probabilistic scene recognition model based on uncertainties. The model considers object information to improve the scene recognition results. All the experiments were performed in real-time with real platforms demonstrating the helpfulness of the scene recognition model when information of the objects is incorporated. Likewise, it has been proved how the relationship between objects and scenes can influence the final decision of where the robot is based on what it perceives. With the uncertainty model proposed in this work, we have strengthened the classification process incorporating perceptual information. In this approach, the essential part is the proper and accurate modeling of the environment, which has to be as close as possible to reality.

Second, we proposed a multi-classifier for scene recognition based on weighted voting schemes. We explored two voting strategies. In the first strategy, the weights assigned to each base classifier were equal to the respective accuracy of each model. On the other hand, in the second scheme, a genetic algorithm was implemented for weight optimization. In this process, accuracy was used to create the objective function. The experiments demonstrated that the recognition rate using the proposed multi-classifier is better than using individual classifiers. In this work, we have implemented two variants of voting schemes that achieve both excellent results. However, the weighted voting scheme based on genetic algorithms works better than the simple voting scheme by accuracy. The adequate combination of independent classifiers through genetic algorithms allows obtaining a more robust and precise model for scene recognition, taking advantage of the benefits of each classifier and compensating for their errors.

Next, in Chapter 5, we introduced a novel method to subdivide the environment into meaningful regions endowing the robot with semantic understanding about its surroundings. Our approach exploits semantic information to infer the label of each part of the environment. Traditional approaches address this problem by segmenting the environment into room-delimited regions that are consistent with a label defined by human knowledge about environments, e.g., office. However, smaller areas inside a room can be used in different ways. For example, the table and the chair in a kitchen may become a temporal office, and thus the semantic category may change to office in that specific place. This may create a confusing area inside the kitchen. To address this situation, we presented an approach that creates a new division of the environment based on regions while maintaining confusions (miscategorizations) of places. We maintain smaller regions that emerge automatically inside full rooms. Some of these regions will be assigned a semantic label that actually corresponds to the room. However, some others will be assigned labels that do not correspond and then create confusing areas. Instead of applying additional filters to try to correct these confusions, like in [123, 166], we keep these areas assuming that confusions in the perception system are due to similarities. We applied a Bayesian filter and

a decision-making procedure to properly merge object and scene information, guaranteeing temporal coherence and correcting misclassifications. The method has been implemented and evaluated in simulated and real-world environments. The experimental results proved the feasibility of applying our method to create a more information-rich representation of the environment based on regions. Moreover, we proved that the representation of the environment obtained by our proposed method is capable of increasing the robot's efficiency when it performs more complex robotic tasks. To do so, we selected the task of searching for objects. The robot is asked to search for an object that is not known in advance. The search method uses our semantic maps based on regions as input. In addition, instead of selecting the most likely room, the method has been adapted to select the most likely region in which the object can be found. We compared our results with the approach presented in [65]. Our evaluations suggested that our method provides competitive results to be applied in more complex robotic tasks. We have demonstrated that our proposed method is capable of generating a useful semantic map based on regions for domestic environments and also, that our semantic representation of the environment can aid improve the robot's performance in tasks such as searching for objects.

Later, in Chapter 6, we demonstrated how the knowledge of the environment can improve more complex robotic tasks. In this chapter, we selected the task of searching for objects. Finding objects can be considered a precondition for performing many robotic tasks and is highly relevant in showing the non-expert end-user that a robot can understand the world. The search problem is tackled from two points of view: globally and locally. The global strategy aims to find the most probable room where the target object can be located. Complementary, a local search strategy is carried out within the most probable room in order to find the target object in a more precise location. For the global search strategy, we proposed a multi-cue search method for objects in human indoor environments. The method appropriately combines different types of prior knowledge, such as semantic information and object-object/-scene relations, to obtain a probabilistic understanding of the location of target objects, which results in a more efficient search strategy. The core of the method is a Conditional Random Field (CRF) and its ability to encode known relations between different observations. The experiments demonstrated the usefulness and efficiency of our method to search for objects based on fully CRF compared to other approaches, such as using the unary potentials and methods based only on current deep learning techniques. The global strategy was tested on a complete dataset built from 10 different real apartments, demonstrating the flexibility to apply it in different environmental conditions.

Regarding the local search strategy, we proposed a novel method guided by co-occurrence probabilities based on the most probable room. The search is refined by choosing specific places inside the room to be visited to minimize the distance traveled. The core of the search process is in the analysis and selection of the best locations called *viewpoints*. Our method

is based on selecting the best viewpoint by analyzing the probabilities of finding the target object in the area covered by the viewpoint. Also, the best orientations within the viewpoint are determined to further speed up the search. An optimized cost function is proposed to maximize the probability of finding the target object while minimizing the distance traveled. We conducted several experiments in simulated and real-world environments. The results showed that the robot successfully finds the target object in the environment while covering only a small portion of the search space. The real-world experiments with the TurtleBot 2 mobile robot validate the proposed approach and demonstrate the good performance of the method also in real-world scenarios.

To summarize, through the work developed in this thesis, we contributed by increasing the robot's knowledge about its environment. We have proposed methods and strategies that extract essential information from indoor environments with the idea of helping service robots to better comprehend and interpret their surroundings. Furthermore, we have proved through our experiments that the appropriate combination of semantic information can improve the robot's performance in some robotic tasks. In this way, through this thesis, we have tried to get closer to that ideal service robot for general purposes.

7.2 Future Perspectives

Many challenges still remain before having autonomous robots cohabiting in domestic environments and helping us with everyday housework. This thesis sets out several research lines that we believe are necessary to solve in further research, which have been divided into five sections and described below:

Scene understanding

As we mentioned in Chapter 1 and Chapter 2, scene understanding is a complex task that involves multiple sub-tasks. Each sub-task by itself is a research topic. This thesis mainly addressed three essential sub-tasks, scene recognition, object detection, semantic segmentation of the environment, and two other sub-tasks that we consider transversal for this thesis, such as object pose estimation and physics-based reasoning, specifically, support relationships of objects. The methods presented in this thesis try to combine all these sub-tasks to acquire full knowledge of the environment to improve robot performance in some robotic tasks. However, we consider it is essential to widen the robot understanding by exploiting and combining other sub-tasks such as 3D reconstruction, saliency detection, and affordance prediction. A 3D representation of the environment can help get more accurate information to perform other tasks such as grasping objects. It is also desirable for the safe navigation of robots. On the other hand, it can be helpful to find relevant objects or attractive

specific regions of the environment for further processing. Thereby, another interesting line is the incorporation of affordances to objects and regions of the environment. The robot could know the object category, its position in the environment, and also its different functional roles. Likewise, the affordance concept would improve the knowledge about the surrounding regions, having not only the semantic label, such as office or kitchen area, but also other functionalities related to the actions that can be carried out, such as sittable, walkable, lyable, among others. The process of having full knowledge of the environment is incremental. Therefore, the information of the environment obtained from these tasks could be combined to improve understanding and contribute to more efficient robot performance.

Object recognition

Despite the advances reported in the literature in object detection, there is still no generic object detection model for all purposes capable of identifying any object category. Our machine learning-based model only detects a few categories that we compensate for by implementing a deep learning-based approach. Although the method based on deep learning can detect 80 object categories, many objects cannot be detected. That is why one of the challenges is the capability of the model to learn new object instances. The systems developed from traditional or more modern techniques work with pre-trained models with a large amount of data for a specific set of objects. Works such as the proposed by Hariharan et al. [61] and Yan et al. [185] can serve as an inspiration to develop more robust systems that will be able to learn new object categories without a strong previous training process. Their approaches are based on low-shot visual learning, where novel object categories are learned based on a small amount of training data.

Another interesting improvement could be incorporating uncertainties into the deep learning-based approaches. As it has been presented by Kraus et al. [97] despite the significant progress made over the last years with modern deep learning techniques, the deep learning-based object detection models do not incorporate uncertainties about how certain a model is in its predictions. To overcome this, authors incorporate a Bayesian Neural Network into a YOLOv3 [34] framework for object detection to estimate the uncertainties. A similar approach could be applied to the deep learning-based approach presented in Chapter 3.

Scene recognition

Scene recognition is a more complex topic than object detection because its variability makes it difficult to assign a semantic label to a scene. To strengthen the methods proposed in this thesis, one suggestion could be to improve the ensemble classifier proposed in Chapter 4, adding other independent classifiers based on deep learning. It might also be interesting to study other voting strategies to evaluate the robustness and performance of multi-classifier

models. Additionally, even though in Chapter 6 we implemented a scene recognition model based on deep learning, that method does not consider objects information. It would be necessary to incorporate information about the objects present in the scene into the CNN to strengthen the scene recognition model.

Semantic labeling

In Chapter 5, our model for subdividing the environment works offline. One idea would be to extend the proposed method to allow the generation of semantic regions online. A way to go is to build the regions while the robot is moving, and when the robot finishes the exploration, save the map for future uses. Besides, it is necessary to consider that the variability of human environments and the fact that humans are constantly changing the objects' placement can affect the definition of the regions. A more robust model should update the regions according to the changes, reassigning the semantic regions in the maps each time. This would be useful to maintain such semantic maps for long-term robots operation. Other ideas are to implement our semantic representation in other robotic tasks to evaluate performance and usefulness. It would also be interesting to explore different techniques to merge semantic information and integrate other information types from the environment.

Searching Strategies

Searching for objects is a mundane task that is the basis for other more complex robotic tasks. Some promising directions could be to improve the method for working in long-term operation. Saving the places where the objects have been found can serve to update the object-scene occurrences and strengthen the search process itself. Another interesting point is to enhance the cost function of the topological navigation model to optimize the calculation of the best path. Likewise, the CRF based strategy presented in Chapter 6.3.1 does not consider the influence of the walls on the method. It would be interesting to study the influence of the walls in the formulation of the CRF, as in real life, some objects can be located near the walls or even on the wall. Additionally, the incorporation of another type of semantic information could be studied in order to improve the search strategies proposed in this thesis.

References

[1] Ambrus, R., Claici, S., & Wendt, A. (2017). Automatic Room Segmentation from Unstructured 3-D Data of Indoor Environments. *IEEE Robotics and Automation Letters*, 2(2), 749–756.

[2] Asimov, I. (1950). *I, robot*. Fawcett Publications.

[3] Astua, C., Barber, R., Crespo, J., & Jardon, A. (2014). Object detection techniques applied on mobile robot semantic navigation. *Sensors (Switzerland)*, 14(4), 6734–6757.

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

[5] Aydemir, A., Sjöö, K., Folkesson, J., Pronobis, A., & Jensfelt, P. (2011). Search in the real world: Active visual object search based on spatial relations. *Proceedings - IEEE International Conference on Robotics and Automation*, (pp. 2818–2824).

[6] Azad, P., Asfour, T., & Dillmann, R. (2009). Combining Harris interest points and the SIFT descriptor for fast scale-invariant object recognition. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, (pp. 4275–4280).

[7] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481–2495.

[8] Bai, S. & Tang, H. (2018). Softly combining an ensemble of classifiers learned from a single convolutional neural network for scene categorization. *Applied Soft Computing Journal*, 67(March), 183–196.

[9] Balaska, V., Bampis, L., Boudourides, M., & Gasteratos, A. (2020). Unsupervised semantic clustering and localization for mobile robotics tasks. *Robotics and Autonomous Systems*, (pp. 103567).

[10] Banerji, S., Sinha, A., & Liu, C. (2013). New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. *Neurocomputing*, 117, 173–185.

[11] Bartiromo, R. & De Vincenzi, M. (2016). *Electrical measurements in the laboratory practice*. Springer.

[12] Bejani, M., Gharavian, D., & Charkari, N. M. (2014). Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. *Neural Computing and Applications*, 24(2), 399–412.

[13] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

[14] Brogårdh, T. (2007). Present and future robot control development—an industrial perspective. *Annual Reviews in Control*, 31(1), 69–79.

[15] Bruce, N. D. & Tsotsos, J. K. (2009). Saliency, attention and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 1–24.

[16] Brucker, M., Durner, M., Ambruş, R., Márton, Z. C., Wendt, A., Jensfelt, P., Arras, K. O., & Triebel, R. (2018). Semantic labeling of indoor environments from 3d rgb maps. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1871–1878).: IEEE.

[17] Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010). BRIEF: Binary robust independent elementary features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6314 LNCS(PART 4), 778–792.

[18] Capek, K. (1920). *RUR (Rossum's universal robots)*. Doubleday, Page.

[19] Chabot, F., Chaouch, M., Rabarisoa, J., Teuliere, C., & Chateau, T. (2017). Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2040–2049).

[20] Choi, W., Chao, Y. W., Pantofaru, C., & Savarese, S. (2013). Understanding indoor scenes using 3D geometric phrases. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (c), 33–40.

[21] Chumtong, P., Mae, Y., Ohara, K., Takubo, T., & Arai, T. (2014). Object search using object co-occurrence relations derived from web content mining. *Intelligent Service Robotics*, 7(1), 1–13.

[22] Chung, M. J.-Y. & Cakmak, M. (2018). "how was your stay?": Exploring the use of robots for gathering customer feedback in the hospitality industry. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 947–954).: IEEE.

[23] Cong, Y., Ackermann, H., Liao, W., Yang, M. Y., & Rosenhahn, B. (2020). Nodis: Neural ordinary differential scene understanding. In *Lecture Notes in Computer Science: Computer Vision-ECCV 2020* (pp. 636–653). Springer.

[24] Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1 (pp. 1–2).: Prague.

[25] Dai, J., He, K., & Sun, J. (2016). Instance-Aware Semantic Segmentation via Multi-task Network Cascades. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 3150–3158.

[26] Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1 (pp. 886–893).: IEEE.

[27] Druon, R., Yoshiyasu, Y., Kanezaki, A., & Watt, A. (2020). Visual object search by learning spatial context. *IEEE Robotics and Automation Letters*, 5(2), 1279–1286.

[28] Ekvall, S., Kragic, D., & Jensfelt, P. (2007). Object detection and mapping for service robot tasks. *Robotica*, 25(2), 175–187.

[29] Elfes, A. (1989). Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6), 46–57.

[30] Elfring, J., Jansen, S., van de Molengraft, R., & Steinbuch, M. (2013). Active object search exploiting probabilistic object–object relations. In *Robot Soccer World Cup* (pp. 13–24).: Springer.

[31] Espinace, P., Kollar, T., Soto, A., & Roy, N. (2010). Indoor scene recognition through object detection. *Proceedings - IEEE International Conference on Robotics and Automation*, (pp. 1406–1413).

[32] Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1), 98–136.

[33] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2), 303–338.

[34] Farhadi, A. & Redmon, J. (2018). Yolov3: An incremental improvement. *Computer Vision and Pattern Recognition, cite as*.

[35] Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1–8).: IEEE.

[36] Felzenszwalb, P. F., Girshick, R. B., & McAllester, D. (2010). Cascade object detection with deformable part models. In *2010 IEEE Computer society conference on computer vision and pattern recognition* (pp. 2241–2248).: IEEE.

[37] Feng, D., Haase-Schuetz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., & Dietmayer, K. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*.

[38] Fowler, S., Kim, H., & Hilton, A. (2018). Human-centric scene understanding from single view 360 video. In *2018 International Conference on 3D Vision (3DV)* (pp. 334–342).: IEEE.

[39] Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780), 1612.

[40] Fu, K., Peng, J., He, Q., & Zhang, H. (2021). Single image 3d object reconstruction based on deep learning: A review. *Multimedia Tools and Applications*, 80(1), 463–498.

[41] Garcia-Haro, J. M., Oña, E. D., Hernandez-Vicen, J., Martinez, S., & Balaguer, C. (2021). Service robots in catering applications: A review and future challenges. *Electronics*, 10(1), 47.

[42] Garvey, T. D. (1976). Perceptual strategies for purposive vision.

[43] Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology Press.

[44] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

[45] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1), 142–158.

[46] Girshick, R., Donahue, J., Darrell, T., Malik, J., Berkeley, U. C., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1, 5000.

[47] Girshick, R. B. (2012). *From rigid templates to grammars: Object detection with structured models*. Citeseer.

[48] Goeddel, R. & Olson, E. (2016). Learning semantic place labels from occupancy grids using cnns. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 3999–4004).: IEEE.

[49] Gomez, C., Hernandez, A. C., Crespo, J., & Barber, R. (2015). a Ros-Based Middle-Cost Robotic Platform With High-Performance. *Iceri2015: 8Th International Conference of Education, Research and Innovation*, (November), 8031–8039.

[50] Gomez, C., Hernandez, A. C., Crespo, J., & Barber, R. (2016). A topological navigation system for indoor environments based on perception events. *International Journal of Advanced Robotic Systems*, 14(1), 1729881416678134.

[51] Gomez, C., Hernández, A. C., Crespo, J., & Barber, R. (2017). Uncertainty-based localization in a topological robot navigation system. In *2017 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)* (pp. 67–72).: IEEE.

[52] Gu, X., Angelov, P. P., Zhang, C., & Atkinson, P. M. (2018). A Massively Parallel Deep Rule-Based Ensemble Classifier for Remote Sensing Scenes. *IEEE Geoscience and Remote Sensing Letters*, 15(3), 345–349.

[53] Guo, S., Huang, W., Wang, L., & Qiao, Y. (2017). Locally supervised deep hybrid model for scene recognition. *IEEE Transactions on Image Processing*, 26(2), 808–820.

[54] Gupta, S., Arbeláez, P., Girshick, R., & Malik, J. (2015). Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation. *International Journal of Computer Vision*, 112(2), 133–149.

[55] Gupta, S., Arbelaez, P., & Malik, J. (2013). Perceptual organization and recognition of indoor scenes from RGB-D images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (pp. 564–571).

[56] Gupta, S., Girshick, R., Arbeláez, P., & Malik, J. (2014). Learning rich features from RGB-D images for object detection and segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8695 LNCS(PART 7), 345–360.

[57] Gupta, S., Kumar, M., & Garg, A. (2019). Improved object recognition results using SIFT and ORB feature detector. *Multimedia Tools and Applications*, 78(23), 34157–34171.

[58] Hanheide, M., Gretton, C., Dearden, R., Hawes, N., Wyatt, J., Pronobis, A., Aydemir, A., Göbelbecker, M., & Zender, H. (2011). Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour. *IJCAI International Joint Conference on Artificial Intelligence*, (pp. 2442–2449).

[59] Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2014). Simultaneous detection and segmentation. In *European Conference on Computer Vision* (pp. 297–312).: Springer.

[60] Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015, 447–456.

[61] Hariharan, B. & Girshick, R. (2017). Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3018–3027).

[62] Hasan, K. M. & Reza, K. J. (2014). Path planning algorithm development for autonomous vacuum cleaner robots. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)* (pp. 1–6).: IEEE.

[63] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9), 1904–1916.

[64] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

[65] Hernandez, A. C., Derner, E., Gomez, C., Barber, R., & Babuška, R. (2020). Efficient object search through probability-based viewpoint selection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 6172–6179).

[66] Hernandez, A. C., Durner, M., Gomez, C., Grixa, I., Teikmanis, O., Marton, Z.-C., & Barber, R. (2021a). Searching for objects in human living environments based on relevant inferred and mined priors. In *2021 European Conference on Mobile Robots (ECMR)*.

[67] Hernandez, A. C., Gomez, C., & Barber, R. (2019a). Minerva: Toposemantic navigation model based on visual information for indoor enviroments. *IFAC-PapersOnLine*, 52(8), 43–48.

[68] Hernandez, A. C., Gomez, C., Barber, R., & Martinez Mozos, O. (2018). Object-Based Probabilistic Place Recognition for Indoor Human Environments. *Proceedings - 2018 International Conference on Control, Artificial Intelligence, Robotics and Optimization, ICCAIRO 2018*, (pp. 177–182).

[69] Hernandez, A. C., Gomez, C., Barber, R., & Mozos, O. M. (2021b). Using miscategorization of places to improve service robotics tasks in indoor environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[70] Hernández, A. C., Gómez, C., Crespo, J., & Barber, R. (2016). Object detection applied to indoor environments for mobile robot navigation. *Sensors (Switzerland)*, 16(8).

[71] Hernandez, A. C., Gomez, C., Crespo, J., & Barber, R. (2017). Adding uncertainty to an object detection system for mobile robots. In *Proceedings - 6th IEEE International Conference on Space Mission Challenges for Information Technology, SMC-IT 2017*, volume 2017-Decem (pp. 7–12).

[72] Hernandez, A. C., Gomez, C., Derner, E., & Barber, R. (2019b). Indoor scene recognition based on weighted voting schemes. In *2019 European Conference on Mobile Robots (ECMR)* (pp. 1–6).: IEEE.

[73] Hernández-López, J.-J., Quintanilla-Olvera, A.-L., López-Ramírez, J.-L., Rangel-Butanda, F.-J., Ibarra-Manzano, M.-A., & Almanza-Ojeda, D.-L. (2012). Detecting objects using color and depth segmentation with Kinect sensor. *Procedia Technology*, 3, 196–204.

[74] Herranz, L., Jiang, S., & Li, X. (2016). Scene recognition with CNNs: Objects, scales and dataset bias. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem (pp. 571–579).

[75] Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

[76] Huang, G.-S. & Lu, Y.-J. (2017). To build a smart unmanned restaurant with multi-mobile robots. In *2017 International Automatic Control Conference (CACS)* (pp. 1–6).: IEEE.

[77] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., & Guadarrama, S. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7310–7311).

[78] Huang, T. S. (1997). Computer Vision: Evolution and Promise. *Report*.

[79] Huber, P. J. (2004). *Robust statistics*, volume 523. John Wiley & Sons.

[80] Izquierdo-Cordova, R., Morales, E. F., Sucar, L. E., & Murrieta-Cid, R. (2017). Searching objects in known environments: Empowering simple heuristic strategies. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9776 LNAI, 380–391.

[81] Jang, H.-W. & Lee, S.-B. (2020). Serving robots: Management and applications for restaurant business sustainability. *Sustainability*, 12(10), 3998.

[82] Jianchao Yang, Kai Yu, Yihong Gong, & Huang, T. (2010). Linear spatial pyramid matching using sparse coding for image classification. (pp. 1794–1801).

[83] Jiang, Y., Koppula, H., & Saxena, A. (2013). Hallucinated humans as the hidden context for labeling 3D scenes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (pp. 2993–3000).

[84] Jiang, Y., Koppula, H. S., & Saxena, A. (2016). Modeling 3D Environments through Hidden Human Context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10), 2040–2053.

[85] Joho, D., Senk, M., & Burgard, W. (2011). Learning search heuristics for finding objects in structured environments. *Robotics and Autonomous Systems*, 59(5), 319–328.

[86] Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., & Wang, X. (2017). T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10), 2896–2907.

[87] Katsumata, Y., Taniguchi, A., Hagiwara, Y., & Taniguchi, T. (2019). Semantic mapping based on spatial concepts for grounding words related to places in daily environments. *Frontiers in Robotics and AI*, 6, 31.

[88] Khan, S. H., Bennamoun, M., Sohel, F., & Togneri, R. (2014). Geometry driven semantic labeling of indoor scenes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1), 679–694.

[89] Khan, S. H., He, X., Bannamoun, M., Sohel, F., & Togneri, R. (2015). Separating objects and clutter in indoor scenes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015, 4603–4611.

[90] Khan, Z. H., Khalid, A., & Iqbal, J. (2018). Towards realizing robotic potential in future intelligent food manufacturing systems. *Innovative food science & emerging technologies*, 48, 11–24.

[91] Kim, B. S., Kohli, P., & Savarese, S. (2013). 3D scene understanding by voxel-CRF. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 1425–1432).

[92] Kim, J., Mishra, A. K., Limosani, R., Scafuro, M., Cauli, N., Santos-Victor, J., Mazzolai, B., & Cavallo, F. (2019). Control strategies for cleaning robots in domestic applications: A comprehensive review. *International Journal of Advanced Robotic Systems*, 16(4), 1729881419857432.

[93] Kollar, T. & Roy, N. (2011). Utilizing object-object and object-scene context when planning to find things. In *English*: IEEE.

[94] Kostavelis, I. & Gasteratos, A. (2017). Semantic maps from multiple visual cues. *Expert Systems with Applications*, 68, 45–57.

[95] Krähenbühl, P. & Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems* (pp. 109–117).

[96] Kramer, O. (2017). *Genetic algorithm essentials*, volume 679. Springer.

[97] Kraus, F. & Dietmayer, K. (2019). Uncertainty estimation in one-stage object detection. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 53–60).: IEEE.

[98] Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79–86.

[99] Kunze, L., Beetz, M., Saito, M., Azuma, H., Okada, K., & Inaba, M. (2012). Searching objects in large-scale indoor environments: A decision-theoretic approach. *Proceedings - IEEE International Conference on Robotics and Automation*, (pp. 4385–4390).

[100] Kunze, L., Doreswamy, K. K., & Hawes, N. (2014). Using Qualitative Spatial Relations for indirect object search. *Proceedings - IEEE International Conference on Robotics and Automation*, (pp. 163–168).

[101] Lafferty, J. & Mccallum, A. (2014). Conditional Random Fields. *Computer Vision*, 2001(June), 146–146.

[102] Leonardis, A., Bischof, H., Pinz, A., Bay, H., Tuytelaars, T., & Van Gool, L. (2006). *Computer Vision – ECCV 2006 SURF: Speeded Up Robust Features*, volume 3951. Springer.

[103] Levenberg, K. & Arsenal, F. (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics*, 1(278), 536–538.

[104] Li, Y., Qi, H., Dai, J., Ji, X., & Wei, Y. (2017). Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2359–2367).

[105] Li, Y., Zhang, J., Cheng, Y., Huang, K., & Tan, T. (2018). Df 2 net: Discriminative feature learning and fusion network for rgb-d indoor scene classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[106] Lim, J. J., Pirsiavash, H., & Torralba, A. (2013). Parsing IKEA objects: Fine pose estimation. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 2992–2999).

[107] Lin, D., Fidler, S., & Urtasun, R. (2013). Holistic scene understanding for 3D object detection with RGBD cameras. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 1417–1424).

[108] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

[109] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014a). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).: Springer.

[110] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014b). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5), 740–755.

[111] Liu, X., Deng, Z., & Yang, Y. (2019). Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 52(2), 1089–1106.

[112] Liu, Y. & Zheng, Y. F. (2005). One-against-all multi-class svm classification using reliability measures. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2 (pp. 849–854).: IEEE.

[113] Loncomilla, P., Ruiz-del Solar, J., & Saavedra A, M. (2018). A Bayesian based Methodology for Indirect Object Search. *Journal of Intelligent and Robotic Systems: Theory and Applications*, 90(1-2), 45–63.

[114] Lowry, S., Sunderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., & Milford, M. J. (2016). Visual Place Recognition: A Survey. *IEEE Transactions on Robotics*, 32(1), 1–19.

[115] Luo, R. C. & Chiou, M. (2018). convolutional neural networks for intelligent service robotics. *IEEE Access*, 6, 61287–61294.

[116] Malisiewicz, T., Gupta, A., & Efros, A. A. (2011). Ensemble of exemplar-svms for object detection and beyond. In *2011 International conference on computer vision* (pp. 89–96).: IEEE.

[117] Marquardt, D. W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2), 431–441.

[118] McGreavy, C., Kunze, L., & Hawes, N. (2017). Next best view planning for object recognition in mobile robotics. *CEUR Workshop Proceedings*, 1782(2).

[119] McLeay, F., Osburg, V. S., Yoganathan, V., & Patterson, A. (2021). Replaced by a robot: Service implications in the age of the machine. *Journal of Service Research*, 24(1), 104–121.

[120] MeeraM, K. & ShajeeMohanB, S. (2016). Object recognition in images. *2016 International Conference on Information Science (ICIS)*, (pp. 126–130).

[121] Mendenhall, W., Sincich, T., & Boudreau, N. S. (1996). *A second course in statistics: regression analysis*, volume 5. Prentice Hall Upper Saddle River, NJ.

[122] Mottaghi, R., Bagherinezhad, H., Rastegari, M., & Farhadi, A. (2016). Newtonian Image Understanding: Unfolding the Dynamics of Objects in Static Images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 3521–3529.

[123] Mozos, O. M. & Burgard, W. (2006). Supervised learning of topological maps using semantic information extracted from range data. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2772–2777).: IEEE.

[124] Mozos, O. M., Mizutani, H., Jung, H., Kurazume, R., & Hasegawa, T. (2013). Categorization of indoor places by combining local binary pattern histograms of range and reflectance data from laser range finders. *Advanced Robotics*, 27(18), 1455–1464.

[125] Mucchiani, C., Sharma, S., Johnson, M., Sefcik, J., Vivio, N., Huang, J., Cacchione, P., Johnson, M., Rai, R., & Canoso, A. (2017). Evaluating older adults' interaction with a mobile assistive robot. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 840–847).: IEEE.

[126] Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., & Ghanem, B. (2018). Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 300–317).

[127] Murty, M. N. & Devi, V. S. (2011). Nearest neighbour based classifiers. In *Pattern Recognition* (pp. 48–85). Springer.

[128] Nanni, L., Lumini, A., & Brahnam, S. (2017). Ensemble of texture descriptors for face recognition obtained by varying feature transforms and preprocessing approaches. *Applied Soft Computing Journal*, 61(August), 8–16.

[129] Nascimento, G., Laranjeira, C., Braz, V., Lacerda, A., & Nascimento, E. R. (2018). A robust indoor scene recognition method based on sparse representation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10657 LNCS (pp. 408–415).

[130] Naseer, M., Khan, S., & Porikli, F. (2018). Indoor Scene Understanding in 2.5/3D for Autonomous Agents: A Survey. *IEEE Access*, 7, 1859–1887.

[131] Nicosevici, T. & Garcia, R. (2012). Automatic visual bag-of-words for online robot navigation and mapping. *IEEE Transactions on Robotics*, 28(4), 886–898.

[132] Nie, X., Wong, L. L., & Kaelbling, L. P. (2016). Searching for physical objects in partially known environments. *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June, 5403–5410.

[133] Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. *Lecture Notes in Engineering and Computer Science*, 2202(March), 380–384.

[134] Pereira, R., Gonçalves, N., Garrote, L., Barros, T., Lopes, A., & Nunes, U. J. (2020). Deep-learning based global and semantic feature fusion for indoor scene classification. In *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)* (pp. 67–73).: IEEE.

[135] Pigini, L., Facal, D., Blasi, L., & Andrich, R. (2012). Service robots in elderly care at home: Users' needs and perceptions as a basis for concept development. *Technology and Disability*, 24(4), 303–311.

[136] Pontil, M. & Verri, A. (1998). Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6), 637–646.

[137] Puybaret, E. (2021-03-26). Sweet Home 3D. http://www.sweethome3d.com/.

[138] Qi, X., Wang, W., Yuan, M., Wang, Y., Li, M., Xue, L., & Sun, Y. (2020). Building semantic grid maps for domestic robot navigation. *International Journal of Advanced Robotic Systems*, 17(1), 1729881419900066.

[139] Quattoni, A. & Torralba, A. (2009). Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 413–420).: IEEE.

[140] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., & Ng, A. Y. (2009). Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3 (pp.5).: Kobe, Japan.

[141] Raafat, H. M., Tolba, A. S., & Aly, A. M. (2011). A novel training weighted ensemble (twe) with application to face recognition. *Applied Soft Computing*, 11(4), 3608–3617.

[142] Ramisa, A., Vasudevan, S., Scaramuzza, D., De Mántaras, R. L., & Siegwart, R. (2008). A tale of two object recognition methods for mobile robots. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5008 LNCS, 353–362.

[143] Rasouli, A., Lanillos, P., Cheng, G., & Tsotsos, J. K. (2020). Attention-based active visual search for mobile robots. *Autonomous Robots*, 44(2), 131–146.

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

[145] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6), 1137–1149.

[146] Rosa, S., Patane, A., Lu, C. X., & Trigoni, N. (2018). Semantic place understanding for human–robot coexistence—toward intelligent workplaces. *IEEE Transactions on Human-Machine Systems*, 49(2), 160–170.

[147] Rosinol, A., Gupta, A., Abate, M., Shi, J., & Carlone, L. (2020). 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans.

[148] Roy, A. & Todorovic, S. (2016). A multi-scale cnn for affordance segmentation in rgb images. In *European conference on computer vision* (pp. 186–201).: Springer.

[149] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 2564–2571).

[150] Ruiz-Sarmiento, J.-R., Galindo, C., & Gonzalez-Jimenez, J. (2017). Building multiversal semantic maps for mobile robot operation. *Knowledge-Based Systems*, 119, 257–272.

[151] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.

[152] Samarakoon, S. B. P., Muthugala, M. V. J., Le, A. V., & Elara, M. R. (2020). Htetro-infi: A reconfigurable floor cleaning robot with infinite morphologies. *IEEE Access*, 8, 69816–69828.

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

[154] Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *European conference on computer vision* (pp. 746–760).: Springer.

[155] Sivic, J. & Zisserman, A. (2009). Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4), 591–606.

[156] Song, S., Lichtenberg, S. P., & Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 567–576).

[157] Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., & Funkhouser, T. (2017). Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1746–1754).

[158] Sun, H., Meng, Z., Tao, P. Y., & Ang, M. H. (2018a). Scene Recognition and Object Detection in a Unified Convolutional Neural Network on a Mobile Manipulator. *Proceedings - IEEE International Conference on Robotics and Automation*, (pp. 5875–5881).

[159] Sun, H., Meng, Z., Tao, P. Y., & Ang, M. H. (2018b). Scene recognition and object detection in a unified convolutional neural network on a mobile manipulator. In *2018 IEEE international conference on robotics and automation (ICRA)* (pp. 1–5).: IEEE.

[160] Sunderhauf, N., Dayoub, F., McMahon, S., Talbot, B., Schulz, R., Corke, P., Wyeth, G., Upcroft, B., & Milford, M. (2016). Place categorization and semantic mapping

on a mobile robot. *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June, 5729–5736.

[161] Sünderhauf, N., Dayoub, F., McMahon, S., Talbot, B., Schulz, R., Corke, P., Wyeth, G., Upcroft, B., & Milford, M. (2016). Place categorization and semantic mapping on a mobile robot. In *2016 IEEE international conference on robotics and automation (ICRA)* (pp. 5729–5736).: IEEE.

[162] Sutton, C. & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4), 267–373.

[163] Thanh, V. N., Vinh, D. P., & Nghi, N. T. (2019). Restaurant serving robot with double line sensors following approach. In *2019 IEEE International Conference on Mechatronics and Automation (ICMA)* (pp. 235–239).: IEEE.

[164] Toris, R. & Chernova, S. (2017). Temporal persistence modeling for object search. In *2017 IEEE international conference on robotics and automation (ICRA)* (pp. 3215–3222).: IEEE.

[165] Trevor, A. J., Rogers, J. G., Nieto-Granda, C., & Christensen, H. I. (2010). : Georgia Institute of Technology.

[166] Triebel, R., Schmidt, R., Mozos, O. M., & Burgard, W. (2007). Instance-based amn classification for improved object recognition in 2d and 3d laser range data. In *Proceedings of the 20th international joint conference on Artifical intelligence* (pp. 2225–2230).

[167] Tussyadiah, I. P. & Park, S. (2018). Consumer evaluation of hotel service robots. In *Information and communication technologies in tourism 2018* (pp. 308–320). Springer.

[168] Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, 104(2), 154–171.

[169] Vänni, K. J. & Salin, S. E. (2017). A need for service robots among health care professionals in hospitals and housing services. In *International Conference on Social Robotics* (pp. 178–187).: Springer.

[170] Vasudevan, S. & Siegwart, R. (2008). Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robotics and Autonomous Systems*, 56(6), 522–537.

[171] Veiga, T. S., Miraldo, P., Ventura, R., & Lima, P. U. (2016). Efficient object search for mobile robots in dynamic environments: Semantic map as an input for the decision maker. *IEEE International Conference on Intelligent Robots and Systems*, 2016-November, 2745–2750.

[172] Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1(July 2014).

[173] Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *MAICS*, 710, 120–127.

[174] Wang, C., Cheng, J., Wang, J., Li, X., & Meng, M. Q.-H. (2018a). Efficient object search with belief road map using mobile robot. *IEEE Robotics and Automation Letters*, 3(4), 3081–3088.

[175] Wang, P., Cheng, J., & Feng, W. (2018b). An approach for construct semantic map with scene classification and object semantic segmentation. In *2018 IEEE International Conference on Real-time Computing and Robotics (RCAR)* (pp. 270–275).: IEEE.

[176] Wang, S., Fidler, S., & Urtasun, R. (2015). Holistic 3D scene understanding from a single geo-tagged image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015, 3964–3972.

[177] Wixson, L. E. & Ballard, D. H. (1994). Using intermediate objects to improve the efficiency of visual search. *International Journal of Computer Vision*, 12(2-3), 209–230.

[178] Woźniak, M., Graña, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16(1), 3–17.

[179] Wu, J. & Rehg, J. M. (2011). CENTRIST: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8), 1489–1501.

[180] Wu, J., Yildirim, I., Lim, J. J., Freeman, W. T., & Tenenbaum, J. B. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in Neural Information Processing Systems*, 2015-January, 127–135.

[181] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3485–3492).

[182] Xiaozhi, C., Huimin, M., Ji, W., Bo, L., & Tian, X. (2017). Multi-View 3D Object Detection Network for Autonomous Driving | Spotlight 4-2B - YouTube. *ComputerVisionFoundation Videos*, (pp. 1907–1915).

[183] Xie, L., Tian, Q., Wang, M., & Zhang, B. (2014). Spatial pooling of heterogeneous features for image classification. *IEEE Transactions on Image Processing*, 23(5), 1994–2008.

[184] Yan, R., Liu, Y., Jin, R., & Hauptmann, A. (2003). On predicting rare classes with SVM ensembles in scene classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 3(2), 21–24.

[185] Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., & Lin, L. (2019). Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 9577–9586).

[186] Yang, B., Rosa, S., Markham, A., Trigoni, N., & Wen, H. (2018). Dense 3D object reconstruction from a single depth view. *Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*.

[187] Yang, F., Xia, G. S., Liu, G., Zhang, L., & Huang, X. (2016). Dynamic texture recognition by aggregating spatial and temporal features via ensemble SVMs. *Neurocomputing*, 173, 1310–1321.

[188] Ye, X., Lin, Z., Lee, J.-Y., Zhang, J., Zheng, S., & Yang, Y. (2019). Gaple: Generalizable approaching policy learning for robotic object searching in indoor environment. *IEEE Robotics and Automation Letters*, 4(4), 4003–4010.

[189] Ye, X., Lin, Z., Li, H., Zheng, S., & Yang, Y. (2018). Active Object Perceiver: Recognition-Guided Policy Learning for Object Searching on Mobile Robots. *IEEE International Conference on Intelligent Robots and Systems*, (pp. 6857–6863).

[190] Yin, H., Jiao, X., Chai, Y., & Fang, B. (2015). Scene classification based on single-layer sae and svm. *Expert Systems with Applications*, 42(7), 3368–3380.

[191] Young, J., Basile, V., Suchi, M., Kunze, L., Hawes, N., Vincze, M., & Caputo, B. (2017). Making Sense of Indoor Spaces Using Semantic Web Mining and Situated

Robot Perception. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10577 LNCS, 299–313.

[192] Zelinsky, A. (2009). *Learning OpenCV—Computer Vision with the OpenCV Library (Bradski, G.R. et al.; 2008)[On the Shelf]*, volume 16.

[193] Zhang, H., Zhang, H., Wang, C., & Xie, J. (2019a). Co-occurrent features in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 548–557).

[194] Zhang, J., Ou, Y., Jiang, G., & Zhou, Y. (2016). An approach to restaurant service robot slam. In *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (pp. 2122–2127).: IEEE.

[195] Zhang, Y., Bai, M., Kohli, P., Izadi, S., & Xiao, J. (2017a). DeepContext: Context-Encoding Neural Pathways for 3D Holistic Scene Understanding. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October, 1201–1210.

[196] Zhang, Y., Tian, G., Lu, J., Zhang, M., & Zhang, S. (2019b). Efficient dynamic object search in home environment by mobile robot: A priori knowledge-based approach. *IEEE Transactions on Vehicular Technology*, 68(10), 9466–9477.

[197] Zhang, Y., Wang, H., & Xu, F. (2017b). Object detection and recognition of intelligent service robot based on deep learning. In *2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)* (pp. 171–176).: IEEE.

[198] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object Detection with Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), 3212–3232.

[199] Zheng, B., Zhao, Y., Yu, J. C., Ikeuchi, K., & Zhu, S. C. (2013). Beyond point clouds: Scene understanding by reasoning geometry and physics. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (pp. 3127–3134).

[200] Zheng, L., Zhu, C., Zhang, J., Zhao, H., Huang, H., Niessner, M., & Xu, K. (2019). Active Scene Understanding via Online Semantic Reconstruction. *Computer Graphics Forum*, 38(7), 103–114.

[201] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., & Torr, P. H. (2015). Conditional random fields as recurrent neural networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 International Conference on Computer Vision, ICCV 2015, 1529–1537.

[202] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6), 1452–1464.

[203] Zhou, L., Zhou, Z., & Hu, D. (2013). Scene classification using a multi-resolution bag-of-features model. *Pattern Recognition*, 46(1), 424–433.

[204] Zhu, H., Weibel, J. B., & Lu, S. (2016). Discriminative Multi-modal Feature Fusion for RGBD Indoor Scene Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 2969–2976.

[205] Zhuo, W., Salzmann, M., He, X., & Liu, M. (2017). Indoor scene parsing with instance segmentation, semantic labeling and support relationship inference. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January, 6269–6275.

[206] Zia, M. Z., Stark, M., & Schindler, K. (2015). Towards Scene Understanding with Detailed 3D Object Representations. *International Journal of Computer Vision*, 112(2), 188–203.

[207] Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object Detection in 20 Years: A Survey. (pp. 1–39).