UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MATHIAS FASSINI MANTELLI

Exploiting semantic information in indoor environments

Thesis presented in partial fulfillment of the requirements for the degree of Doctor of Computer Science

Advisor: Profa. Dra. Mariana Luderitz Kolberg
Coadvisor: Prof. Dr. Renan de Queiroz Maffei

Porto Alegre
April 2022

CIP — CATALOGING-IN-PUBLICATION

Mantelli, Mathias Fassini

Exploiting semantic information in indoor environments / Mathias Fassini Mantelli. – Porto Alegre: PPGC da UFRGS, 2022.

40 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2022. Advisor: Mariana Luderitz Kolberg; Coadvisor: Renan de Queiroz Maffei.

1. Mobile robotics. 2. Object search. 3. Semantic information. 4. Robotics perception. 5. Indoor environments. I. Kolberg, Mariana Luderitz. II. Maffei, Renan de Queiroz. III. Título.

“You are seeking happiness.
Learn this lesson, once and forever,
that you will find happiness only by
helping others to find it!”
— Outwitting the Devil

ACKNOWLEDGEMENTS

Agradeço ao LaTeX por não ter vírus de macro...

ABSTRACT

Nowadays, the mobile robotics research community deals with different high-level tasks that require the robot to manipulate or interact with objects which are not in the robot's field of view or in an unexplored region of the environment. To find an object in unknown environments, the robot needs to look for it while gaining information about the environment and making decisions in real-time, known as the object search (OS) problem. The research community has proposed different approaches for dealing with the OS problem, relying on the objects' color, 3D shape, or its surroundings as visual cues to guide the search. However, this geometric information (i.e., color or size) limits the robot's perception and, consequently, the robot's performance during the search. Therefore, we proposed exploiting the advantages of semantic information inferred from two sources abundant in human-populated environments: text and dynamic agents. First, we propose an OS system for unknown indoor environments that relies on semantic information inferred from texts found in the environment, aiming to find a target door label. The use of semantic information in this scenario allowed the robot to reduce the search costs by avoiding not promising regions to contain the target door label. Our semantic planner reasons over the numbers detected from door labels to decide either to continue the search towards unknown parts of the environment or carefully search in the already known parts. Second, we present another OS system based on semantic information inferred from different objects' position over time in the environment. Composed by two modes, our system first gathers data from the objects' placement by executing its recording mode. This data is later used when the robot executes the request mode to search for the target object. Both systems were evaluated in different environments and compared against other OS approaches in simulated and real scenarios. The results support our systems' efficiency and demonstrate the improvement in the searching performance with the aid of semantic information.

Keywords: Mobile robotics. Object search. Semantic information. Robotics perception. Indoor environments.

Explorando Informações Semânticas em Ambientes Internos

RESUMO

Atualmente, a comunidade científica de robótica móvel está lidando com diferentes tarefas de alto-nível que requerem que o robô manipule ou interagir com objetos em partes não exploradas do ambiente ou que não estejam no campo de visão do robô. Para encontrar um objeto em um ambiente desconhecido, o robô precisa procurar por ele enquanto ganha informação sobre o ambiente e toma decisões em tempo-real, conhecido como o problema de busca por objetos (BPO). A comunidade de pesquisa propôs diferentes soluções para abordar o problema de BPO, se baseando na cor, tamanho ou no que existe ao redor dos objetos. Contudo, todas essas informações geométricas (como por exemplo cor ou tamanho) limita a percepção do robô e, por consequência, o seu desempenho durante a busca. Portanto, nós propomos explorar as vantagens de informações semânticas inferidas a partir de duas fontes que são abundantes em ambientes populados por humanos: textos e agentes dinâmicos. Em primeiro lugar, nós propomos um sistema para BPO para ambientes internos que se baseia em informações semânticas inferidas de textos encontrados no ambiente, o que permite que o robô reduza os custos da busca por evitar regiões não promissoras para o objeto buscado. Nosso planejador semântico raciocina sobre os números de placas de portas detectados para decidir se continua a busca em direção a regiões desconhecidas ou se realiza a busca cuidadosamente em regiões já conhecidas. Em segundo lugar, nós apresentamos outro sistema para BPO baseado em informações semânticas inferidas da diferente localização dos objectos no ambiente ao longo do tempo. Composto por dois modos, nosso sistema coleta dados do posicionamento dos objetos executando o seu modo de gravação, que depois será usado quando o robô executa o modo de busca. Ambos os sistemas foram avaliados em diferentes ambientes e comparados contra outros sistemas de BPO em simulação e ambiente real. Os resultados confirmam a eficiência dos nossos sistemas e demonstram a melhora no desempenho da busca com o auxílio das informações semânticas.

Palavras-chave: Robótica móvel. Busca por objectos. Informação semântica. Percepção robótica. Ambientes internos.

LIST OF FIGURES

Figure 2.1 OCR METHODOLOGIES. 21
Figure 2.2 OCR PAPER. 22
Figure 2.3 DEEP LEARNING SURVEY. 23
Figure 2.4 YOLO SYSTEM MODEL. 24
Figure 2.5 KDE. 25
Figure 2.6 Fundamental problems in mobile robots and their state estimation. 26
Figure 2.7 Graphical model of the fundamental mobile robotics problems. 28

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

OS	Object Search
SLAM	Simultaneous Localization and Mapping
MCL	Monte Carlo Localization
RBPF	Rao-Blackwellized Particle Filter
ROI	Region of Interest
RGB	Red, Green, Blue
RGB-D	Red, Green, Blue, Depth
GPS	Global Positioning System
BVP	Boundary Value Problem
KDE	Kernel Density Estimation
HIMM	Histogram In-Motion Mapping
DS	Door Simulator
YOLO	You Only Look Once

LIST OF SYMBOLS

$\boldsymbol{x}_t$	robot's pose at time step t. It is composed by a three dimensional vector containing x, y, which represents the position, and θ, which represents the orientation.
$\boldsymbol{m}_i$	map of the environment, represented by a list of N objects, in our case grid cells, with $1 \leq n \leq N$, in the environment along with their properties is given by the vector $\boldsymbol{m} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \cdots, \boldsymbol{m}_N)^T$.
$\boldsymbol{u}_t$	control data at instant t, and it corresponds to the change of state in the time interval $(t-1; t]$.
$\boldsymbol{z}_t$	measurement made by the robot at instant t. The vector of all of them acquired at the same instant t is $\boldsymbol{z}_t = (\boldsymbol{z}_t^1, \boldsymbol{z}_t^2, \cdots, \boldsymbol{z}_t^K)^T$.
c	a cell in the 2D grid map $\boldsymbol{m}$.
$K(\cdot)$	uniform circular kernel that computes the size of the free area covered by it.
d	Manhattan distance between two cells.
r	radius of $K(\cdot)$.
a	height of $K(\cdot)$.
$\mathbf{T}$	subset of cells that are within the area of the kernel at any moment.
c_K	centre of the kernel $K(\cdot)$.
$Q(\cdot)$	function that tests whether a cell is free.
$\Psi(\cdot)$	function that computes the free space.
$\Upsilon(\cdot)$	function that computes the map segmentation.
δ	a threshold that defines how many different sizes of free areas are considered by the segmentation function.
$\mathbf{s}$	segment from $\boldsymbol{m}$.
$\mathbf{I}$	image.
$\mathbf{L}$	list containing the recognized number from door signs.
$S(\cdot)$	function that returns the nearest segment of a cell.
$L(\cdot)$	function that returns the list of door signs from a segment.

l door number.

$\varphi(\cdot)$ probability distribution for a map cell, c, where the target object is in $\boldsymbol{m}$.

$\mathtt{S}(\cdot)$ semantic factor, the outcome of the combination of growing direction, $\varphi_g(\cdot)$, parity, $\varphi_p(\cdot)$, factors.

$\mathtt{G}(\cdot)$ geometric factor, the outcome of the combination of robot orientation, $\varphi_r(\cdot)$, door orientation, $\varphi_o(\cdot)$, and distance, $\varphi_d(\cdot)$, factors.

$\mathbf{C}$ set of candidate cells for finding the goal-door.

c_* the candidate cell most promising to bring the robot to the goal-door.

$\theta(\cdot)$ function that returns the angle of the growing direction factor.

$L^<$, $L^>$ functions that counts the amount of door labels are smaller and larger than the goal-door, respectively.

$\zeta(\cdot)$ function that measures the possibility of a segment to have door signs smaller or larger than the goal-door.

$\gamma(\cdot)$ function that measures the difference angle between the increasing angle and the Voronoi angle.

$L^{\neq}$, $L^=$ functions that counts the amount of door labels have different or equal parity than the goal-door, respectively.

w_p threshold used to control the minimum amount of detected door signs.

$H_r[\cdot]$ the door orientation histogram.

λ_r a threshold used to define how many most recent robot's orientations are saved.

$D(\cdot,\cdot)$ function that counts the number of cells between two other specific cells.

$d_{\ll}$ smallest distance between two cells.

$\mathbf{H}$ 2D grid heat map.

$\mathbf{O}$ set of detected objects.

o detected object, composed by its position within $\mathbf{H}$, its category, the hour it has been detected, and the robot's position during the detection, i.e., $o = (p^o, c^o, h^o, r^o)$.

$W(\cdot,\cdot)$ function that measures the weight of the time difference.

$K^i(\cdot)$ inversed uniform circular kernel that computes the size of the free area covered by it.

CONTENTS

1 INTRODUCTION....14
1.1 Scope....15
1.2 Objective....17
1.3 Contributions of this Thesis....18
1.4 Outline....18
2 THEORETICAL BACKGROUND....20
2.1 Perception....20
2.1.1 Text Localization and Recognition....20
2.1.2 You Only Look Once (YOLO)....22
2.2 Kernel Density Estimation on Images....25
2.3 The Basics of Mobile Robotics....25
2.4 Object search problem formulation....30
REFERENCES....32
APPENDIX A — RESUMO EXPANDIDO....37
A.1 Hipótese e objetivos....38

1 INTRODUCTION

Robots can be grouped into different classes depending on their function and the workplace they are designed for (KUMAR et al., 2005; ROBOTICS, 2012; HAIDEGGER et al., 2013). Among all the classes of robots, there are two major ones, called *industrial* and *service* (UN, 2003; ALMEIDA; FONG, 2011; CHIANG; TRIMI, 2020; GARCIA-HARO et al., 2020; CHEN; BARNES, 2021). Industrial robots (IRs) are, according to the Robotic Industries Association (RIA), automatically controlled, reprogrammable, multi-purpose manipulators programmable in three or more axes (RIA, 2012). They can be mobile or fixed in place platforms for industrial automation applications. Traditionally, such robots are designed for factories, where they are deployed for different application categories such as palletizing (MOURA; SILVA, 2018), painting (ASADI; LI; CHEN, 2018), welding (YAO et al., 2019), assembly (KYRARINI et al., 2019), and general handling tasks (WILLIGENBURG; HOL; HENTEN, 2004; HÄGELE et al., 2016). To meet the requirements for this set of applications, IRs have a wide variety of designs regarding payload capacity, workspace volume, and the number of robot axes (HÄGELE et al., 2016). On the other hand, service robots (SRs) are robots that work semi or completely autonomously to perform useful services for the well-being of humans and equipment, excluding manufacturing operations (ROBOTICS, 2012). The SRs come in all different designs, as they may or may not be equipped with an arm structure, and even though most of them are mobile, they can also be fixed in place (GARCIA-HARO et al., 2020). The International Federation of Robotics (IFR) divides SRs into two subclasses based on their usability: *professional* and *personal/domestic* service robots (LITZENBERGER, 2018). Some examples of the professional service robots (PSRs) are defence robots (MARTINIC, 2014), farmer-assistants (VAKILIAN; MASSAH, 2017), medical (ABUBAKAR et al., 2020), and logistic (THAMRONGAPHICHARTKUL et al., 2020). Examples of domestic service robots (DSRs) include but are not limited to vacuum cleaners (FORLIZZI; DISALVO, 2006), lawn-mowers (BORINATO, 2017), food and beverage waiters (WAN et al., 2020), and elderly assistants (HERSH, 2015). The market for SRs has been regularly rising, and it is no surprise that there is an expectation that it will grow even further in the next few years (ALMEIDA; FONG, 2011; CHIANG; TRIMI, 2020). The decreasing cost of hardware components (processors, motor drivers, and sensors), the increasing energy density and lower cost of batteries, and the current pandemic situation drive this expansion (CHIANG; TRIMI, 2020).

1.1 Scope

According to the Population Division (PD) of the United Nations, in 2015, there were 901 million people aged 60 or over, representing 12% of the global population (DIVISION, 2015). Besides, the PD projects that by 2030, the number of older adults in the world will reach 1.4 billion and 2.1 billion by 2050. Several policies to tackle the problems of population ageing have been proposed by several countries, including, for example, facilities for the elderly (LIN; CHEN, 2018; SEDDIGH et al., 2020). However, placing elderly people in facilities for retirement or even in nursing homes may cause some problems, such as physically, emotionally, and psychologically dependencies (THEURER et al., 2015). Additionally, some elderly do not voluntarily stay at nursing homes, preferring to spend their remaining years at their home where they have a more positive self-image than those who live in the nursing homes (KOK; BERDEN; SADIRAJ, 2015; LIN; CHEN, 2018). The increasing number of older people living at home supports the need for DSRs to automate processes and tasks that may be tedious, inconvenient, or even challenging for older people (PAULIUS; SUN, 2019; TORRESEN; KURAZUME; PRESTES, 2020). In general, these sorts of robots can contribute to practical tasks for humans as robot assistants or robot companions, such as watching older adults concerning emergencies, reminding them to take their medicines, and searching, picking, and placing objects (SPRUTE et al., 2017; TORRESEN et al., 2018; PAULIUS; SUN, 2019).

Additionally, while some of the main motivations for deploying SRs have been elderly assistance and productivity improvement, the current COVID-19 pandemic has brought a more critical purpose for them (CHIANG; TRIMI, 2020). PSRs can be deployed to perform a series of applications to provide contactless services, ensuring humans can practice social distancing (SEIDITA et al., 2021). Besides disinfecting indoor environments (MANTELLI et al., 2022), PSRs also have the potential to support the hospitality industry (ROSETE et al., 2020), and deliver medications and food (LEE et al., 2009; YANG et al., 2020). The use of SRs in logistic applications is relevant during such unusual scenarios. Some national organizations from the United States identified logistics as one of the broad areas where robotics can make a difference during outbreaks (SEIDITA et al., 2021).

In many example applications that we listed above, it is likely that SRs have to perform some searching tasks. Simple examples would be DSRs searching and picking ob-

jects for elderly with mobility restrictions, and PSRs delivering packages to a specific spot in an unknown environment. Similar to humans in the context of object searching tasks, SRs should also not rely on the assumption that the object (or regions of interest) they are searching for is already within their field of view (FoV) (SJöö; AYDEMIR; JENSFELT, 2012). Hence, they have to find the target object in large-scale environments based on primarily their visual sensors, known as object search (OS) problem (AYDEMIR et al., 2013). However, how does an SR find the target object that is not initially within its FoV? One way to address this problem is to make the SR perform a brute-force OS, in which it visits the whole environment following a predefined search route. Even though this strategy seems a straightforward solution, it does not efficiently solve the problem (RASOULI et al., 2020). As long as the target object is within the environment, the SR will eventually find it. However, the searching process may be time-consuming due to the long distances travelled by the robot. Another more efficient solution is to consider a search strategy that incorporates information from both the environment and the target object, to improve the searching performance. For example, such information could be the shape of the room for the environment (AYDEMIR et al., 2011), and the colour or category/class for the target object (RASOULI et al., 2020). The search strategy is one of the most critical parts of an OS approach, as it directly impacts the efficiency of an OS system (AYDEMIR et al., 2013). Therefore, it must be robust and effective regardless the environment the SR is performing the search.

The research community has proposed valuable works related to the OS problem (EKVALL; KRAGIC; JENSFELT, 2007; SJöö et al., 2009; SJöö; AYDEMIR; JENSFELT, 2012; AYDEMIR et al., 2013; RASOULI et al., 2020). The problem is proven to be NP-Complete (TSOTSOS, 1992; YE; TSOTSOS, 2001), which means that the optimal search solution can be computed by approximation (SJöö; AYDEMIR; JENSFELT, 2012), minimizing the search cost as much as possible. In the case of SR performing OS tasks, such approximation could be computed with the aid of strong cues provided by the semantics of both the environment and other objects in the SR's surroundings (SJöö; AYDEMIR; JENSFELT, 2012). Semantics can be regarded as the high-level information inferred (or "perceived") from the environment, including but not limited to names and categories of different objects, rooms and locations (VASUDEVAN et al., 2007; SJöö; AYDEMIR; JENSFELT, 2012; LIU et al., 2016). Similarly, semantic maps encode not just the geometric and topological description of the environment but also its semantic interpretation, providing a friendly way for robots to communicate with humans (LIU et

al., 2016). Then, when the SR processes its sensor readings to infer further knowledge about its surroundings, it increases the level of abstraction of the environment over time (BARBER et al., 2018). The use of both semantic information and map in robotic applications enhances the robot's autonomy and robustness in many ways, besides facilitating some challenging tasks (CESAR et al., 2016).

1.2 Objective

This thesis aims to exploit the organization of both the environment and its objects to infer semantic search cues to address the OS problem.

We claim that in general our society is not randomly organized, and there are several patterns and rules we follow everyday. It is no surprise that humans can improve their efficiency while performing daily tasks, like OS, if the environment is merely logically organized. Thus, they can save energy and time during such tasks. For example, most cities have their own rules for numbering the properties, although there is no unique and global rule for that. Then, the habitants can study it to understand the numbering pattern. Thus, they can estimate where a particular unknown building is in the city, even if they have never been there. On the contrary, when there are no written rules to specify the organization of the environment, humans can understand them just by observing the environment for a while. Therefore, we are interested in making the SRs take advantage of such available organizations to improve their performance in the OS problem.

We consider that semantic information could be inferred from the environment, and it could be used to help SRs in search tasks. Such semantic information is helpful to OS systems because it could be used as high-level search cues. Then, with a semantic-based OS system, the SR would not need to search the whole environment to find the target object. The human reasoning process relies on several sorts of high-level search cues during searches. In our daily life, we read signs, symbols, and labels to evaluate which direction we should go to find a specific room in an unknown environment. Another example would be someone who first checks whether the family's car is at home, to then search for the car key. In particular, we focus on the positional semantic information inferred from the organization of the environment, like the labels and signs in the first example or the acknowledgement that other people may move objects.

1.3 Contributions of this Thesis

This thesis presents results (MANTELLI et al., 2021; MANTELLI et al., 2022) showing that semantic information inferred from the organization of the environment can help SRs in the OS problem. Specially, we show that the use of organizational semantic information as search cues in the search strategy of OS systems can make the SR save resources by not visiting the whole environment. Besides, we show that the proper use of semantic information can improve the SRs' perception to perform high-level tasks, bringing them closer to humans.

The first contribution of this thesis is an OS system that seeks to find a specific room based on the organization of door labels within the environment (MANTELLI et al., 2021). Although humans heavily rely on texts, characteres, and symbols for accomplishing several tasks, the use of characteres as a data source is not very popular in robotics. In this work, we have argued that characteres have a great potential for providing search clues and are often found in man-made environments. This idea came from human behaviour when searching for someone's office in an unknown building. More specifically, we are interested in how the characteres from door labels in corridors are used to estimate whether a corridor is promising for finding the target office. The search strategy relies on the patterns of door labels in indoor scenarios, and it reasons over them to estimate which corridor is more promising for achieving the goal.

Another contribution is a long-term semantic system that searches for a target object in dynamic unknown environments (MANTELLI et al., 2022). It assumes that some objects within the environment are not always static, and people move them around over time. In this way, its goal is to incorporate a person's routine and habits into the search strategy and then make search estimations. This work aimed to model the semantic information of how objects are organized over time within an environment. Then, it uses this information to avoid making the SRs search for the target object in not promising regions. This idea came from observing how the objects are placed over time and that every person has their own singularities in terms of object placement.

1.4 Outline

The outline of this thesis is as follows. First, in Chapter 2, we introduce a background on the main problems in mobile robotics, as well as the most popular approaches

that deal with each problem. Besides, it also presents the general concepts of the OS problem, which is used throughout this thesis along with the basic concepts of mobile robotics. In Chapter **??**, we provide our first semantic OS system that is based on text as the main source of semantic information. Next, in Chapter **??**, we discuss our second semantic OS system, which is the one that aims to understand how the objects within the environment are moved through a period of time, to make predictions about their future positions. Lastly, in Chapter **??**, we conclude this thesis proposal by discussing the current contributions and draw the future directions for this PhD work.

2 THEORETICAL BACKGROUND

In the previous chapter, we have argued that multiple robotic tasks would benefit from exploiting the semantic information inferred from spatial and temporal organization of the environments that surrounds the robot. We have chosen the object search (OS) problem to explore this idea, which aims to estimate a target object's location in a large unknown environment, usually with a camera attached to a mobile robot. We believe investigating this problem can enlarge our understanding regarding the benefits of employing semantic information to expand the robot's perception.

This chapter presents a theoretical background detailing techniques used throughout this thesis. The OS problem requires the robot to map the unknown environment and to estimate its position simultaneously. SLAM systems fulfill these requirements, as it computes the state estimation and builds an environment representation. Hence, we address the basic concepts of such systems and mobile robotics in general, from the individual localization and mapping problems to how they are combined into the SLAM systems. Besides, we cover the generic and central formulation of OS problems, which is the basis for the works presented in Chapters **??** and **??**.

2.1 Perception

2.1.1 Text Localization and Recognition

Text can be embedded into documents or scenes as a means of communicating information, and it is considered one of the most expressive means of communications (YE; DOERMANN, 2014). The process of Optical Character Recognition (OCR) aims to detect and recognize text in printed materials, images or videos, to then convert it into a digitized form so that machines can manipulate the digital text (YE; DOERMANN, 2014; ISLAM; ISLAM; NOOR, 2017). The OCR process has been in the spotlight for several years (ISLAM; ISLAM; NOOR, 2017). The motivation for such attention from the research community is that text provides meaningful information to be used in many applications. Besides, OCR is a complex problem due to the variety of languages, fonts, and styles in which text can be written, along with the complex rules of languages (ISLAM; ISLAM; NOOR, 2017).

There are two methodologies commonly used in OCR systems, integrated and

stepwise (YE; DOERMANN, 2014). The Integrated methodology recognizes words when the detection procedures share information with character classification and relies on joint optimization strategies, as shown in Fig. 2.1a. On the other hand, Stepwise methodologies have separated detection and recognition modules. Besides, it uses a feed-forward pipeline to detect, segment, and recognize text regions, also shown in Fig. 2.1b (YE; DOERMANN, 2014). Lastly, it relies on a feedback procedure from text recognition to text detection to reduce false detections.

(a) Integrated Methodology (b) Stepwise Methodology

Figure 2.1 – OCR METHODOLOGIES by (YE; DOERMANN, 2014).

The latter methodology usually employs a coarse-to-fine strategy, which is done in four steps: localization, verification, segmentation, and recognition (YE; DOERMANN, 2014). The goal of the first step, localization, is to classify components and groups into candidate text regions coarsely. In the second step, verification, such regions are further classified into text or non-text regions. The third step, segmentation, separates the characters in a way that exclusive, accurate outlines of image blocks remain for the recognition step. Lastly, the recognition step converts image blocks into characters. It is also possible that some stepwise methodologies ignore the verification and/or segmentation steps. Further, another adaptation is the inclusion of additional steps to perform text enhancement and/or rectification (YE; DOERMANN, 2014).

One of the leading methods in scene text detection is based on detecting characters, such as the one proposed in Neumann and Matas (2012) (ZHANG et al., 2016). It is an end-to-end real-time scene text localization and recognition method based on the stepwise methodology (NEUMANN; MATAS, 2012; YE; DOERMANN, 2014). Neumann and Matas addressed the character detection problem as an efficient sequential selection from the set of Extremal Regions (ERs) to achieve real-time performance. Such ER is a character detector that analyses image regions whose outer boundary pixels have higher values than the region itself. It is robust to blur, illumination, colour, and texture variation. Additionally, it also handles low-contrast text (NEUMANN; MATAS, 2012). The pixels within each ER's bounding box are described by a class of region descriptors that serve as features for the character classification.

In a given grayscale image, Figure 2.2b, the authors do the text localization by estimating the probability of each ER being a character using a set of features, Figure 2.2c.

The verification step is done by selecting only the ERs from the previous step with locally maximal probability (of being a character), Figure 2.2d. Since the verification step aims to eliminate not promising character candidates, the authors further classify the high-likely ERs with more computationally expensive features to improve the character classification. Then, they group the verified ERs into words and select the most probable character segmentation, Figure 2.2e. The grouping is computed with a highly efficient exhaustive search with feedback loops. To get the text detected, Figure 2.2f, the average run time of the proposed method on a 800×600 image is 0.3s on a standard PC (NEUMANN; MATAS, 2012). The method proposed by Neumann and Matas achieved state-of-the-art text localization results when evaluated in two public datasets. Besides, they were the first ones to report results for the end-to-end text recognition on the ICDAR 2011 Robust Reading competition dataset (NEUMANN; MATAS, 2012)

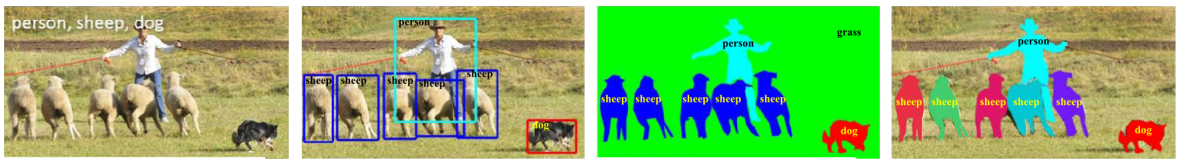
(a) (b) (c) (d) (e) (f)

Figure 2.2 – OCR PAPER (NEUMANN; MATAS, 2012).

2.1.2 You Only Look Once (YOLO)

Generic object detection problem is defined as, given an image, determining whether or not there are instances of objects from predefined categories. For the present instances, it is also necessary to return their spatial location and extent (LIU et al., 2020). This problem is also known as generic object category detection, object class detection, or even object category detection (LIU et al., 2020). It focuses on detecting a broad range of natural categories instead of specific object category detection where only a narrower predefined category of interest may be present, such as faces, pedestrians, or cars. Nowadays, the research community is more interested in detecting highly structured objects, like cars, bicycles, and airplanes, and articulated objects, such as humans and pets, rather than unstructured scenes (e.g., sky, grass, and cloud) (LIU et al., 2020). Regarding the spatial location and extend of an object within an image, like the detected objects in Figure 2.3a, it can be defined coarsely using a bounding box (EVERINGHAM et al., 2010; REDMON et al., 2016), which is a rectangle tightly bounding the object, Figure 2.3b, a precise pixel-wise segmentation maks (ZHANG et al., 2013), Figure 2.3c, or a closed

boundary (RUSSELL et al., 2008; LIN et al., 2014), Figure 2.3d (LIU et al., 2020).



(a) Object classification (b) Generic object detection (Bounding box) (c) Semantic segmentation (d) Object instance segmentation

Figure 2.3 – DEEP LEARNING SURVEY (LIU et al., 2020).

Recently, deep learning-based techniques have been proposed to deal with the object detection problem (REDMON et al., 2016). Deep learning has been used to solve many other challenging tasks, in areas such as image classification (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; HE et al., 2016) and language modeling (GONG et al., 2018; DAI et al., 2019)(HE; ZHAO; CHU, 2021). Deep learning techniques have arisen as powerful techniques for learning feature representations automatically from data (LIU et al., 2020). The success of deep learning in general in these areas is partly due to the rapidly developing computational resources, such as powerful graphical cards, the availability of big training data, and the effectiveness of deep learning to extract hidden representations from images, texts, and videos (WU et al., 2020).

A popular deep learning approach for object detection that has been improved since its first version is called YOLO (REDMON et al., 2016). The name is an acronym and is short for "*You Only Look Once*", which partially explains the general idea of the approach. The YOLO's authors have entirely abandoned the initial object detection paradigm of "*proposal detection + verification*". Instead, they followed an entirely different idea: to apply a single neural network to the whole image. This idea made YOLO the first one-stage object detector in the deep learning era (where the name comes from) (ZOU et al., 2019). Thanks to this unified approach, YOLO detects objects extremely fast, processing images in real-time at 45 frames per second (FPS) (REDMON et al., 2016). To compare, the best object detector before YOLO, called Faster RCNN, processes images at 5 7 FPS (LIU et al., 2020).

Redmon et al. (REDMON et al., 2016) approached object detection as a regression problem, straight from image pixels to bounding box coordinates and class probabilities (LIU et al., 2020). Each image is evaluated once in their system by a single neural network that predicts bounding boxes and class probabilities. Due to this single network detection pipeline, YOLO can be optimized end-to-end directly on detection performance. In more detail, YOLO divides an image into an $U \times U$ grid, and each grid cell predicts

B bounding boxes and confidence scores for those boxes. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts. In the second image of Figure 2.4, the score of the bounding boxes is represented by their thickness. Hence, the thicker the bounding box, the higher the confidence that there is an object in that location. Here is important to mention that at this point, YOLO does not know which objects there are in the image. It just knows whether exist any and their locations. To find out the object classes within the image, YOLO predicts C conditional class probabilities for each grid cell. These probabilities are conditioned on the grid cell containing an object. It means that YOLO is predicting that if there is an object in a cell, that object is an instance of the class with the highest probability. YOLO only predicts a set of class probabilities per grid cell regardless of the number of bounding boxes B for each grid cell. The third image of Figure 2.4 illustrates this prediction. Finally, YOLO multiplies the conditional class probabilities and the individual bounding box confidence predictions at the test time, which results in the class-specific confidence scores for each box. Hence, these scores encode both the probability of that class appearing in the bounding box and how well the predicted box fits the object, represented by the fourth image of Figure 2.4. Through a non-max suppression process, YOLO discards low-value bounding boxes and duplicated detections, resulting then in the final detections, illustrated by the fifth image of Figure 2.4.

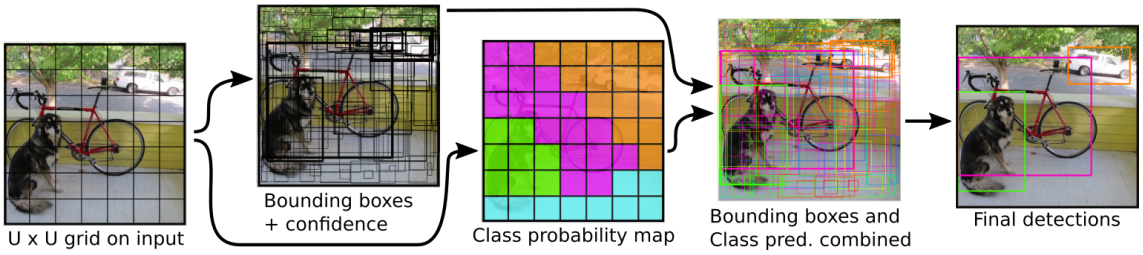


Figure 2.4 – YOLO SYSTEM MODEL. Adapted from (REDMON et al., 2016).

Unlike sliding window and region proposal-based techniques, YOLO reasons globally about the image when making predictions, processing the entire image during training and test time. Hence, it implicitly encodes contextual information about classes and their appearance. Besides, YOLO learns generalizable representations of objects. In one of its experiments, YOLO was trained on natural images and tested on artwork, outperforming top detection methods proposed by the research community (REDMON et al., 2016). However, despite its significant improvement in detection speed, YOLO suffers from a slight drop in localization accuracy compared with two-stage detectors, especially for some small objects (ZOU et al., 2019).

2.2 Kernel Density Estimation on Images

TO DO

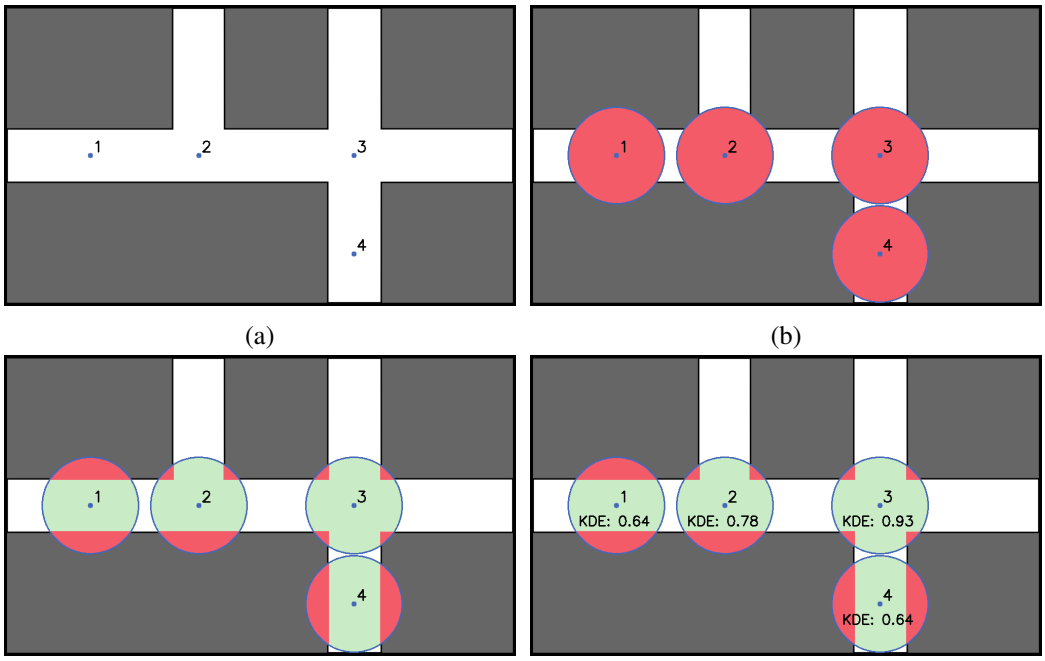


Figure 2.5 – KDE.

2.3 The Basics of Mobile Robotics

Mobile robots perform several tasks that require them to be aware of their positions in the environment and obstacles' positions to avoid collisions. In most realistic scenarios where the robots are deployed, such information is not directly available. Hence, the robots have to estimate it with their sensors, which provide noisy and partial data from the environment (THRUN; BURGARD; FOX, 2006).

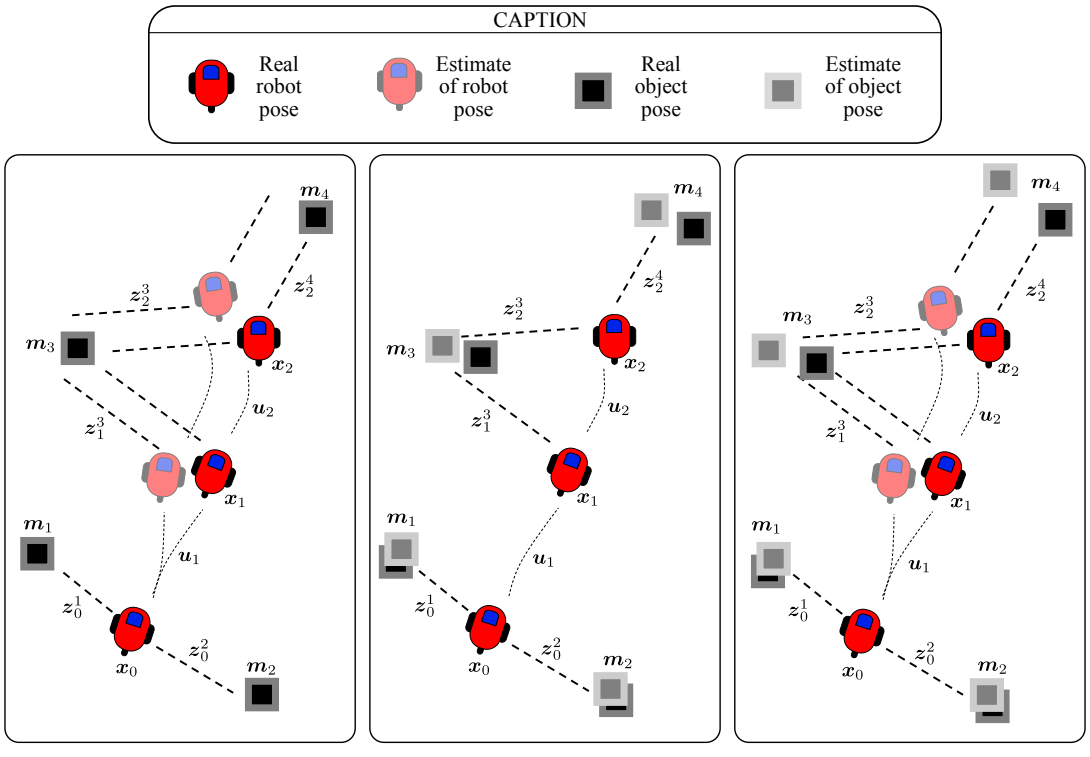
The state estimation in mobile robotics can be summarized in four variables:

- $\boldsymbol{x}_t$: robot's pose at time step t. It is composed by a three dimensional vector containing $(x, y, \theta)^T$, in which x, y represent the position and θ the orientation. A sequence of robot's poses from time step 0 to time step t is defined as $\boldsymbol{x}_{0:t} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_t\}$.
- $\boldsymbol{m}_i$: object i's position in the environment. A list of N objects, with $1 \leq n \leq N$, in the environment along with their properties is given by the vector $\boldsymbol{m} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \cdots, \boldsymbol{m}_N)^T$.
- $\boldsymbol{u}_t$: control data at instant t, and it corresponds to the change of state in the time

interval $(t-1;t]$. The sequence of control data that takes the robot from the initial position to $\boldsymbol{x}_t$ is denoted by $\boldsymbol{u}_{1:t} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_t\}$.

- $\boldsymbol{z}_t^i$: the i-th measurement made by the robot at instant t. The vector of all of them acquired at the same instant t is $\boldsymbol{z}_t = (\boldsymbol{z}_t^1, \boldsymbol{z}_t^2, \cdots, \boldsymbol{z}_t^K)^T$, whereas $\boldsymbol{z}_{1:t} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_t\}$ expresses the history of all observations.

After defining the four variables that are the basic foundation for state estimation in mobile robotics, it is worthing to explain their role in different estimation problems. The set of controls $\boldsymbol{u}_{1:t}$ and measurements $\boldsymbol{z}_{1:t}$ are always known since the robot's sensors provide them. Inertial measurement units and wheel encoders are examples of sensors that provide control data, whereas lidars, sonars, and cameras measure the environment. The other two variables, robot's pose, $\boldsymbol{x}_{0:t}$, and environmental map, $\boldsymbol{m}$, are not necessarily known. Depending on the estimation problem, it is necessary to estimate different variables, like the three examples depicted in Figure 2.6. In *Localization*, Figure 2.6a, the map is known in advance, and hence, only the robot's pose is estimated. The opposite happens in *Mapping*, Figure 2.6b, as the map is built based on the known robot's pose. Lastly, in *SLAM*, Figure 2.6c, which combines the two previous problems, none of them is given a priori, and therefore, both are estimated simultaneously.



(a) Localization (b) Mapping (c) SLAM

Figure 2.6 – Fundamental problems in mobile robots and their state estimation. It is estimated: (a) robot's pose, (b) map, (c) both of them simultaneously. Extracted from (MAFFEI, 2017).

Localization is the most basic perceptual problem in robotics. It aims to determine the robot's pose relative to a given map of the environment. Localization can also be seen as a problem of coordinate transformation, in which it is established a correspondence between the map coordinate system and the robot's local coordinate system (THRUN; BURGARD; FOX, 2006). There are multiple localization problems, and they are not equal in terms of their difficulty level. One characteristic that divides this problem into local and global localization is the awareness of the robot's initial pose. The former assumes that the initial robot's pose is known. Therefore, the problem becomes a sort of position tracking in which the noise in the measurements is adjusted in robot motion, commonly by a Gaussian distribution. On the other hand, the latter is unaware of the initial pose, making it perform the localization globally (where the name comes from) in the map. The global localization has a higher difficulty level than the local one, but one of its variations is even more challenging, called the kidnapped robot problem. It addresses the problem of a localized robot being teleported to some other location in that the robot might believe it knows where it is while it does not. Although a robot is rarely kidnapped in practice, recovering from localization failures is essential for autonomous robots.

The formulation of the global localization problem is presented in Figure 2.7, which depicts a few iterations of the robot's pose estimation and how the variables are used. The map $\boldsymbol{m}$ is already known, whereas the $\boldsymbol{x}_{0:t}$ must be estimated based on the controls $\boldsymbol{u}_{1:t}$ and the measurements $\boldsymbol{z}_{1:t}$. For the case of local localization, the $\boldsymbol{x}_0$ is known and hence, does not need to be estimated. Markov localization is a probabilistic algorithm that addresses all the localization problems mentioned earlier. It applies the Bayes filter, $p(\boldsymbol{x}_t \mid \boldsymbol{u}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{m})$, to transform a probabilistic belief at time $t-1$ into a belief at time t.

Many other localization algorithms implement Markov localization in mobile robotics. Three of them have been in the spotlight for a long time and are prevalent in this field: Kalman filter, grid-based filter, and particle filter. The former filters predicts in linear dynamics and measurement functions (LEONARD; DURRANT-WHYTE, 1991), whereas the grid-based filter approximates the estimations by decomposing the state space into finitely many regions of the grid map (BURGARD et al., 1998). The key idea of the latter, particle filter, is to represent the estimation by a set of random state samples, called particles, drawn from the previous estimation. It can represent a much broader space of distribution, in contrast to the Kalman filter that is more strict to Gaussians (DELLAERT et al., 1999). The particle filter implementation for mobile robotics is also known as

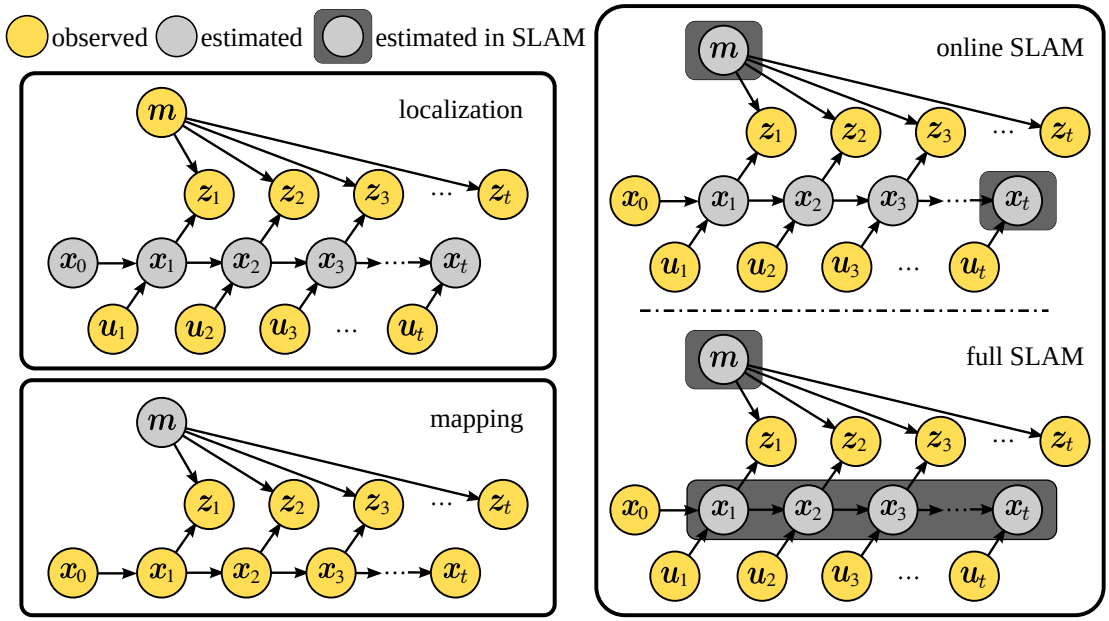


Figure 2.7 – Graphical model of the fundamental mobile robotics problems: localization, mapping, and two SLAM variations. Adapted from (MAFFEI, 2017).

Monte Carlo Localization (MCL), widely used in many different robotics applications for multiple robot types (DELLAERT et al., 1999).

Mapping, for the case of the robot's poses are known, is the problem of generating consistent maps from noisy and imprecise measurement data (THRUN; BURGARD; FOX, 2006). The estimated belief of the map, $p(\boldsymbol{m} \mid \boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t})$, considers the set of all measurements up to time t, $\boldsymbol{z}_{1:t}$, along with the robot's path defined by its history of all poses, $\boldsymbol{x}_{1:t}$, as shown in Figure 2.7. Comparing the graphical models of the localization and mapping problems in Figure 2.7, one can say that they are opposite each other in terms of which variable each estimates. This thought makes sense, since whereas the former relies on $\boldsymbol{m}$ to estimate $\boldsymbol{x}_{0:t}$, the latter relies on $\boldsymbol{x}_{0:t}$ to estimate $\boldsymbol{m}$. It is important to mention that the controls $\boldsymbol{u}_{1:t}$ play no role in this context, as the path is already known. Besides, the robot's initial pose $\boldsymbol{x}_0$ is omitted from the map estimation because no measures are taken when the robot is at that pose.

Similar to the localization problem that groups multiple localization types, the mapping problem also represents a general idea implemented by different map types. The feature-based maps represent the cartesian location of features, which are distinct objects in the physical world, extracted from the measurements, such as images from visual sensors or a vector of distances from a 2D lidar (SALAS-MORENO et al., 2013; ENGEL; SCHÖPS; CREMERS, 2014; MUR-ARTAL; MONTIEL; TARDOS, 2015). The advantage of such a map type is the reduction of computational complexity, as the feature space

has a lower dimension than the raw measurement. For example, the eight 3D edges of a boudingbox encircling a car are computationally cheaper to process than a point cloud from a 3D lidar. Another map type within the mapping problem is called location-based. It represents in each map component $\boldsymbol{m}_i$ the regions from the environment, regardless of whether they contain objects. This way, any location in the world has a label on the map, not only features. Occupancy grid maps are often considered the most popular location-based map (THRUN; BURGARD; FOX, 2006). They discretize the environment into small portions called grid cells, which store information about the area it covers. In general, this information in each cell is a single value representing the probability that an obstacle occupies this cell. The size of the cells defines the map resolution, which brings a tradeoff between the level of details and the demand for memory resources.

Lastly, Simultaneous localization and mapping (SLAM), also known as Concurrent Mapping and Localization, is undoubtedly the most fundamental and challenging problem in robotics (THRUN; BURGARD; FOX, 2006). SLAM problems appear in scenarios where the environmental map is unavailable and the robot is unaware of its pose. In contrast to the other two problems presented earlier, which have to estimate either the map $\boldsymbol{m}$ or $\boldsymbol{x}_{1:t}$, in SLAM problems, the robot has to perform the estimation of both variables at the same time, as shown in Figure 2.7. Since the robot does not know its pose and there is no map, the pose $\boldsymbol{x}_0$ is assumed, by convention, to be $(0, 0, 0)^T$. The high difficulty level of SLAM comes from the double dependency of localization and mapping: to estimate the pose, the robot needs a map from the environment, whereas to estimate the map, the robot need to know its pose.

The SLAM problem is divided into two forms based on what is estimated: online and full SLAMs. The former focus on estimating only the posterior over the current robot's pose $\boldsymbol{x}_t$ and the map $\boldsymbol{m}$, $p(\boldsymbol{x}_t, \boldsymbol{m} \mid \boldsymbol{z}_{1:t}, \boldsymbol{u}_{1:t})$. The full SLAM computes the same estimation, but with the entire robot's trajectory $\boldsymbol{x}_{1:t}$ along with the map $\boldsymbol{m}$, $p(\boldsymbol{x}_{1:t}, \boldsymbol{m} \mid \boldsymbol{z}_{1:t}, \boldsymbol{u}_{1:t})$.

The majority of the algorithms for the online SLAM problem are incremental, i.e., the idea is to estimate the posterior probability on the current robot state and map as the robot moves, discarding past measurements and controls once they have been processed. The Kalman and particle filters are also used in this context, besides the localization one as previously discussed. The Extended Kalman Filter is the basis of one of the earliest online SLAM approaches, linearizing motion and observation models, which usually are nonlinear, to perform the online SLAM estimations (MAFFEI, 2017). An online

SLAM problem that is based on particle filter is known as Rao-Blackwellized particle filter (RBPF) (MURPHY et al., 1999; DOUCET et al., 2000; GRISETTIYZ; STACHNISS; BURGARD, 2005; GRISETTI; STACHNISS; BURGARD, 2007). In RBPF, each particle carries an individual grid map of the environment, representing a hypothesis of the robot's trajectory. The number of particles is directly related to the map quality since the higher this number, the broader is the hypotheses variety. However, there is a cost associated with each particle, and hence, it is not practical to increase the number of particles until the estimated map matches the physical world.

The algorithms for the full SLAM problem calculate a posterior over the entire path, which solves an issue in the online SLAM problem. Discarding the previous states after estimating the current one, also known as Markov assumption, implies that the possible poor estimations in the past are not adjustable. In contrast, the full SLAM problems backpropagate to the previous estimations the error reduction computed in the current state calculation. GraphSLAM captures the essence of the full SLAM problem, since it calculates a solution for the offline problem over $\boldsymbol{x}_{1:t}$ and $\boldsymbol{z}_{1:t}$ in $\boldsymbol{m}$. Despite the advantage of improving previous state estimations, full SLAM algorithms are computationally heavy due to the optimization of nonlinear quadratic constraints.

Explaining the fundamental problems of mobile robotics, from the simplest localization to the more complex SLAM problems, helps to understand why the OS works for unknown environments depend on a SLAM system. Since our works presented in the following chapters are designed for similar conditions (large and unknown environments), we opted to rely on GMapping (GRISETTI; STACHNISS; BURGARD, 2007). It is an online SLAM algorithm based on RBPF that provides a 2D grid map, and each cell contains a value that means whether the region it represents is unknown (to the SLAM system), occupied (obstacle), or free.

2.4 Object search problem formulation

The OS problem relies on an efficient strategy for finding a target object in a large unknown indoor environment. Since our works presented in this thesis are based on a 2D grid map, the search strategies from these works reason over the map $\boldsymbol{m}$, and they decide what cell c is currently more promising to localize the target object while minimizing the total cost. We define cost as the distance traveled by the robot during the search, as the longer the robot's path, the higher is the amount of resources (battery and time) it spends.

The robot is equipped with a 2D lidar to build the grid map and a camera used to gather visual cues for semantic information estimation. Both sensors are fixed to the robot's body, and hence, we consider only the movements performed by a ground mobile robot.

Additionally, let $\varphi(c)$ be the probability distribution for a map cell, c, where the target object is in $\boldsymbol{m}$. Depending on the level of a priori knowledge of $\boldsymbol{m}$ and $\varphi(c)$, it is possible to address the OS problem in three different ways:

- $\boldsymbol{m}$ **and** $\varphi(c)$ **are known**: the problem becomes a sensor placement, aiming to reduce the search cost by moving the robot straight to the cell c.
- **only** $\boldsymbol{m}$ **is known**: in case the map is available a priori (or acquired through a separate mapping step), the mobile robot should either rely on a generic probability distribution or move through the environment to gather information. The inspection performed by the robot is to get information about the objects and update the probability distribution.
- $\boldsymbol{m}$ **and** $\varphi(c)$ **are unknown**: the robot needs to map the environment with the aid of a SLAM system, at the same time that it collects information to compute the probability distribution. Since the robot performs OS in an unknown environment, it has to tradeoff between expanding the mapped area and executing sensing actions to search for the target object carefully. This scenario is also known as the exploration vs. exploitation problem.

In this thesis, both the second and third points are considered, addressed individually in different works, in Chapters **??** and **??**, respectively. In general, each of these works has a semantic search strategy, i.e., it incorporates semantic information into the estimations to improve the performance. However, it is important to mention that these semantic search strategies consider common-sense knowledge, which is not environment-specific, and integrate high-level human concepts. In the context of this thesis, common-sense knowledge encodes semantic information inferred from text signs and objects' placement over a while. Such information is valuable for our works because it reduces the search space and improves the search for a human-like performance.

REFERENCES

ABUBAKAR, S. et al. Arna, a service robot for nursing assistance: System overview and user acceptability. p. 1408–1414, 2020.

ALMEIDA, A. T. de; FONG, J. Domestic service robots [tc spotlight]. **IEEE robotics & automation magazine**, IEEE, v. 18, n. 3, p. 18–20, 2011.

ASADI, E.; LI, B.; CHEN, I.-M. Pictobot: A cooperative painting robot for interior finishing of industrial developments. **IEEE Robotics Automation Magazine**, v. 25, n. 2, p. 82–94, 2018.

AYDEMIR, A. **Exploiting structure in man-made environments**. Thesis (PhD) — KTH Royal Institute of Technology, 2012.

AYDEMIR, A. et al. Active visual object search in unknown environments using uncertain semantics. In: **Transactions on Robotics**. [S.l.]: IEEE, 2013. v. 29, n. 4, p. 986–1002.

AYDEMIR, A. et al. Object search guided by semantic spatial knowledge. In: **The RSS**. [S.l.: s.n.], 2011. v. 11.

BARBER, R. et al. Mobile robot navigation in indoor environments: Geometric, topological, and semantic navigation. In: . [S.l.]: IntechOpen, 2018.

BORINATO, G. **Auto mowing system**. [S.l.]: Google Patents, 2017. US Patent 9,820,433.

BURGARD, W. et al. Integrating global position estimation and position tracking for mobile robots: the dynamic markov localization approach. v. 2, p. 730–735, 1998.

CESAR, C. et al. Simultaneous localization and mapping: Present future and the robust-perception age. **arXiv preprint arXiv: 1606.05830**, 2016.

CHEN, J. Y.; BARNES, M. J. Human–robot interaction. **Handbook of human factors and ergonomics**, Wiley Online Library, p. 1121–1142, 2021.

CHIANG, A.-H.; TRIMI, S. Impacts of service robots on service quality. **Service Business**, Springer, v. 14, n. 3, p. 439–459, 2020.

DAI, Z. et al. Transformer-xl: Attentive language models beyond a fixed-length context. **arXiv preprint arXiv:1901.02860**, 2019.

DELLAERT, F. et al. Monte carlo localization for mobile robots. v. 2, p. 1322–1328, 1999.

DIVISION, P. **World Population Prospects**. New York, NY: Department of Economic and Social Affairs, United Nations, 2015.

DOUCET, A. et al. Rao-blackwellised particle filtering for dynamic bayesian networks. **Conference on Uncertainty in Artificial Intelligence**, 2000.

EKVALL, S.; KRAGIC, D.; JENSFELT, P. Object detection and mapping for service robot tasks. In: **Robotica**. [S.l.: s.n.], 2007. v. 25, n. 2, p. 175–187.

ENGEL, J.; SCHÖPS, T.; CREMERS, D. Lsd-slam: Large-scale direct monocular slam. p. 834–849, 2014.

EVERINGHAM, M. et al. The pascal visual object classes (voc) challenge. **International journal of computer vision**, Springer, v. 88, n. 2, p. 303–338, 2010.

FORLIZZI, J.; DISALVO, C. Service robots in the domestic environment: a study of the roomba vacuum in the home. p. 258–265, 2006.

GARCIA-HARO, J. M. et al. Service robots in catering applications: A review and future challenges. **Electronics**, MDPI, v. 10, n. 1, p. 47, 2020.

GONG, C. et al. Frage: Frequency-agnostic word representation. **Advances in neural information processing systems**, v. 31, 2018.

GRISETTI, G.; STACHNISS, C.; BURGARD, W. Improved techniques for grid mapping with rao-blackwellized particle filters. **IEEE transactions on Robotics**, IEEE, v. 23, n. 1, p. 34–46, 2007.

GRISETTIYZ, G.; STACHNISS, C.; BURGARD, W. Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling. p. 2432–2437, 2005.

HÄGELE, M. et al. Industrial robotics. In: ____. **Springer Handbook of Robotics**. Cham: Springer International Publishing, 2016. p. 1385–1422.

HAIDEGGER, T. et al. Applied ontologies and standards for service robots. **Robotics and Autonomous Systems**, Elsevier, v. 61, n. 11, p. 1215–1223, 2013.

HE, K. et al. Deep residual learning for image recognition. p. 770–778, 2016.

HE, X.; ZHAO, K.; CHU, X. Automl: A survey of the state-of-the-art. **Knowledge-Based Systems**, Elsevier, v. 212, p. 106622, 2021.

HERSH, M. Overcoming barriers and increasing independence–service robots for elderly and disabled people. **International Journal of Advanced Robotic Systems**, SAGE Publications Sage UK: London, England, v. 12, n. 8, p. 114, 2015.

ISLAM, N.; ISLAM, Z.; NOOR, N. A survey on optical character recognition system. **arXiv preprint arXiv:1710.05703**, 2017.

KOK, L.; BERDEN, C.; SADIRAJ, K. Costs and benefits of home care for the elderly versus residential care: a comparison using propensity scores. **The European journal of health economics**, Springer, v. 16, n. 2, p. 119–131, 2015.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.

KUMAR, V. et al. Industrial, personal, and service robots. **DRAFT REPORT**, v. 41, 2005.

KYRARINI, M. et al. Robot learning of industrial assembly task via human demonstrations. **Autonomous Robots**, Springer, v. 43, n. 1, p. 239–257, 2019.

LEE, M. K. et al. The snackbot: documenting the design of a robot for long-term human-robot interaction. In: **Proceedings of the 4th ACM/IEEE international conference on Human robot interaction**. [S.l.: s.n.], 2009. p. 7–14.

LEONARD, J. J.; DURRANT-WHYTE, H. F. Mobile robot localization by tracking geometric beacons. **IEEE Transactions on robotics and Automation**, v. 7, n. 3, p. 376–382, 1991.

LIN, T.-Y. et al. Microsoft coco: Common objects in context. p. 740–755, 2014.

LIN, X.; CHEN, T. A qualitative approach for the elderly's needs in service robots design. p. 67–72, 2018.

LITZENBERGER, G. Service robots. **International Federation of Robotics**, 2018. Accessed: 2022-03-23. Available from Internet: <https://ifr.org/service-robots>.

LIU, L. et al. Deep learning for generic object detection: A survey. **International journal of computer vision**, Springer, v. 128, n. 2, p. 261–318, 2020.

LIU, Q. et al. Extracting semantic information from visual data: A survey. **Robotics**, Multidisciplinary Digital Publishing Institute, v. 5, n. 1, p. 8, 2016.

MAFFEI, R. d. Q. **Translating sensor measurements into texts for localization and mapping with mobile robots.** Thesis (PhD) — Federal University of Rio Grande do Sul, 2017.

MANTELLI, M. et al. Semantic temporal object search system based on heat maps. **Journal of intelligent & robotic systems**, Springer, v. 101, n. 2, p. 1–23, 2022.

MANTELLI, M. et al. Semantic active visual search system based on text information for large and unknown environments. **Journal of intelligent & robotic systems**, Springer, v. 101, n. 2, p. 1–23, 2021.

MANTELLI, M. F. et al. Autonomous environment disinfection based on dynamic uv-c irradiation map. **IEEE Robotics and Automation Letters**, IEEE, 2022.

MARTINIC, G. The proliferation, diversity and utility of ground-based robotic technologies. **Canadian Military Journal**, v. 14, n. 4, p. 52, 2014.

MOURA, F. M.; SILVA, M. F. Application for automatic programming of palletizing robots. p. 48–53, 2018.

MUR-ARTAL, R.; MONTIEL, J. M. M.; TARDOS, J. D. Orb-slam: a versatile and accurate monocular slam system. **IEEE transactions on robotics**, IEEE, v. 31, n. 5, p. 1147–1163, 2015.

MURPHY, K. P. et al. Bayesian map learning in dynamic environments. p. 1015–1021, 1999.

NEUMANN, L.; MATAS, J. Real-time scene text localization and recognition. In: **Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2012.

PAULIUS, D.; SUN, Y. A survey of knowledge representation in service robotics. In: . [S.l.]: Elsevier, 2019. v. 118, p. 13–30.

RASOULI, A. et al. Attention-based active visual search for mobile robots. In: **Autonomous Robots**. [S.l.]: Springer, 2020. v. 44, n. 2, p. 131–146.

REDMON, J. et al. You only look once: Unified, real-time object detection. p. 779–788, 2016.

RIA, Robotic Industries Association. **ANSI/RIA R15.06-2012**. American National Standards Institute, Inc., 2012. Accessed: 2022-03-23. Available from Internet: <https://webstore.ansi.org/preview-pages/RIA/preview_ANSI+RIA+R15-06-2012.pdf>.

ROBOTICS, I. . **ISO 8373:2012 Robots and robotic devices — Vocabulary**. International Organization for Standardization, 2012. Accessed: 2022-03-23. Available from Internet: <https://www.iso.org/standard/55890.html>.

ROSETE, A. et al. Service robots in the hospitality industry: An exploratory literature review. p. 174–186, 2020.

RUSSELL, B. C. et al. Labelme: a database and web-based tool for image annotation. **International journal of computer vision**, Springer, v. 77, n. 1, p. 157–173, 2008.

SALAS-MORENO, R. F. **Dense Semantic SLAM.** Thesis (PhD) — Imperial College London, 2014.

SALAS-MORENO, R. F. et al. Slam++: Simultaneous localisation and mapping at the level of objects. p. 1352–1359, 2013.

SEDDIGH, M. et al. A comparative study of perceived social support and depression among elderly members of senior day centers, elderly residents in nursing homes, and elderly living at home. **Iranian Journal of Nursing and Midwifery Research**, Wolters Kluwer–Medknow Publications, v. 25, n. 2, p. 160, 2020.

SEIDITA, V. et al. Robots as intelligent assistants to face covid-19 pandemic. **Briefings in Bioinformatics**, Oxford University Press, v. 22, n. 2, p. 823–831, 2021.

SJöö, K.; AYDEMIR, A.; JENSFELT, P. Topological spatial relations for active visual search. In: **Robotics and Autonomous Systems**. [S.l.: s.n.], 2012. v. 60, n. 9. ISSN 0921-8890.

SJöö, K. et al. Object search and localization for an indoor mobile robot. In: **Journal of Computing and Information Technology**. [S.l.]: SRCE-University Computing Centre, 2009. v. 17, n. 1.

SPRUTE, D. et al. Ambient assisted robot object search. In: SPRINGER. **International Conference on Smart Homes and Health Telematics**. [S.l.], 2017. p. 112–123.

THAMRONGAPHICHARTKUL, K. et al. A framework of iot platform for autonomous mobile robot in hospital logistics applications. p. 1–6, 2020.

THEURER, K. et al. The need for a social revolution in residential care. **Journal of aging studies**, Elsevier, v. 35, p. 201–210, 2015.

THRUN, S.; BURGARD, W.; FOX, D. **Probabilistic Robotics**. [S.l.]: Massachusetts Institute of Technology, 2006.

TORRESEN, J.; KURAZUME, R.; PRESTES, E. Special issue on elderly care robotics – technology and ethics. In: . [S.l.]: Springer Nature BV, 2020. v. 98, n. 1, p. 3–4.

TORRESEN, J. et al. Robot companions for older people – ethical concerns. In: . [S.l.: s.n.], 2018. p. 53.

TSOTSOS, J. K. On the relative complexity of active vs. passive visual search. In: **International journal of computer vision**. [S.l.]: Springer, 1992. v. 7, n. 2, p. 127–141.

UN. United nations and the international federation of robotics. **Proceedings of the World Robotics**, 2003.

VAKILIAN, K. A.; MASSAH, J. A farmer-assistant robot for nitrogen fertilizing management of greenhouse crops. **Computers and electronics in agriculture**, Elsevier, v. 139, p. 153–163, 2017.

VASUDEVAN, S. et al. Cognitive maps for mobile robots—an object based approach. **Robotics and Autonomous Systems**, Elsevier, v. 55, n. 5, p. 359–371, 2007.

WAN, A. Y. S. et al. Waiter robots conveying drinks. **Technologies**, Multidisciplinary Digital Publishing Institute, v. 8, n. 3, p. 44, 2020.

WILLIGENBURG, L. V.; HOL, C.; HENTEN, E. V. On-line near minimum-time path planning and control of an industrial robot for picking fruits. **Computers and electronics in agriculture**, Elsevier, v. 44, n. 3, p. 223–237, 2004.

WU, Z. et al. A comprehensive survey on graph neural networks. **IEEE transactions on neural networks and learning systems**, IEEE, v. 32, n. 1, p. 4–24, 2020.

YANG, G.-Z. et al. **Combating COVID-19—The role of robotics in managing public health and infectious diseases**. [S.l.]: American Association for the Advancement of Science, 2020. eabb5589 p.

YAO, P. et al. Light-weight topological optimization for upper arm of an industrial welding robot. **Metals**, v. 9, n. 9, 2019.

YE, Q.; DOERMANN, D. Text detection and recognition in imagery: A survey. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 37, n. 7, p. 1480–1500, 2014.

YE, Y.; TSOTSOS, J. K. A complexity-level analysis of the sensor planning task for object search. In: **Computational Intelligence**. [S.l.: s.n.], 2001. v. 17, n. 4, p. 605–620.

ZHANG, X. et al. Object class detection: A survey. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 46, n. 1, p. 1–53, 2013.

ZHANG, Z. et al. Multi-oriented text detection with fully convolutional networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 4159–4167.

ZOU, Z. et al. Object detection in 20 years: A survey. **arXiv preprint arXiv:1905.05055**, 2019.

APPENDIX A — RESUMO EXPANDIDO

As primeiras décadas de pesquisa sobre robótica móvel, a partir do início de 2004, lidaram com os desafios de conectar eficiência e associação de dados. Eles introduziram formulações probabilísticas para planejamento de caminhos, exploração, localização e mapeamento simultâneos, e muitas outras áreas. Algumas das abordagens para estas areas são populares ainda hoje, tal qual o Filtro de Partículas RaoBlackwellised e o Filtro Extendido de Kalman. A maioria deles foram baseados em sensores de laser ou ultrassônicos, já que esse sensores eram os mais populares na época. Consequentemente, os mapas resultantes eram na maioria grades 2D, no qual as células representavam regiões livres, ocupadas (obstáculos), ou desconhecidas (CESAR et al., 2016).

Após construir uma fundação sólida para muitos problemas robóticos com abordagens probabilísticas, a comunidade científica deu um passo a frente. Eles concentraram em melhorar as propriedades das abordagens já propostas e das novas também, tais como obersabilidade, convergências, e consistência (CESAR et al., 2016). Simultaneamente neste período (2004-2015), sensores visuais estiveram nos centros das atenções como uma alternativa para ganhar informação sobre o ambiente. A melhora considerável deles na qualidade e variedade (por exemplo imagens de profundidade, nuvem de pontos e images estéreo) dos dados ajudaram no aumento do seu uso. De fato, construir mapas 2D e 3D do ambiente com sensores visuais resultou em um novo termo, SLAM Visual (SALAS-MORENO, 2014).

A robótica móvel aproveitou formidáveis vantagens ao realizar tarefas que requerem apenas que os robôs naveguem por ambientes livres e desvie de obstáculos. Mover itens do ponto A para o ponto B ou aspirar espaços livres são exemplos de tarefas robóticas com soluções satisfatórias. Contudo, o mesmo nível de sucesso não se aplica até agora para muitas outras tarefas de alto-nível que os robôs supostamente têm que abordar hoje em dia. Desde que a robótica móvel mudou o seu foco de chão de fábrica e linhas de produção para espaços povoado diariamente por pessoas, os robôs são requeridos a desempenhar tarefas tais quais humanos em diferentes cenários que não são necessariamente tão controlados e organizados como o mundo industrial (AYDEMIR, 2012). Se basear apenas em mapas puramente geométricos e ter percepção limitada que não permitem ir além de representações geométricas básicas não permite o robô a obter uma compreensão de alto-nível do ambiente. Isso pode ser a razão para robôs não properarem tanto em tarefas de alto-nível. Os robôs são impedidos de processar os dados do ambiente

para inferir ou estimar conhecimentos extras e valiosos para várias tarefas.

A.1 Hipótese e objetivos

Como mencionado anteriormente, tarefas robóticas de alto-nível demandam que os robôs leiam o ambiente de uma forma similar aos humanos, que raciocinam sobre várias características do ambiente ou dos objetos e não apenas sobre o tamanho das áreas livre ou ocupada. A percepção geométrica do robô, ou seja, leituras de sensores puros que geram mapas métricos padrões, não são descritivos e informativos o suficiente para fornecer tais melhorias que são necessárias pelas tarefas de alto-nível. Esta limitação não significa que mapas geométricos são inúteis ou irrelevantes atualmente. Pelo contrário, eles ainda são significantes para a segurança da navegação do robô ou pelo planejamento de caminhos. Contudo, há uma demanda para complementar e extender a percepção geométrica do robô com conhecimento significativo do ambiente. Nós, então, reivindicamos que informações de alto-nível inferidas a partir das leituras dos sensores, também chamada de informação semântica, deve ser explorada para complementar a percepção do robô quando construir robôs autônomos.

A associação de informação semântica (ou conceitos) a entidades geométricas no mapa é chamado mapeamento semântico, um dos mais novos tópicos que os pesquisadores têm explorado. Ele realça a autonomia e a robustês do robô em diversas formas, além de facilitar algumas tarefas de alto-nível (CESAR et al., 2016). Carros autônomos são um bom exemplo de uma aplicação robótica que demanda uma compreensão do ambiente parecida com uma feita por humanos. A Figura **??** ilustra uma nuvem de pontos do carro autônomo da Waymo enquanto se dirigia por uma rua. A nuvem de pontos crua não diferencia os obstáculos ao redor do carro, uma vez que ela apenas indica aonde eles estão e quais regiões são livres. Se o carro se baseaia apenas nessa nuvem de pontos, ele pode até se mover pela rua desviando dos obstáculos, mas ele está longe de se comportar como um motorista adequado que segue todas as regras de trânsito. A sua percepção geométrica não especifica, por exemplo, quais obstáculos são estáticos ou dinâmicos, o que é vital para a segurança de todos. Contudo, com o auxílio da informação semântica, a percepção do carro vai além de apenas detectar objetos tal como o semáforo, carros e pessoas, Figura **??**. O carro entende qual semáforo está aberto ou fechado pela cor de sua luz, onde os outros carros estão indo, e o tipo dos carros (normal ou de polícia), Figura **??**. Motoristas humanos naturalmente e rapidamente compreendem a cena na Figura **??**, mas

o mesmo não pode ser dito sobre os robôs.

Nossa hipótese é que a informação semântica e mapas semânticos preenchem o vazio para melhorar a robótica móvel no sentido de tarefas de alto-nível. Nossa idéia é que o ambiente fornece mais informações inferidas pelo sistema do robô do que simples leituras de snesores. Como informações semânticas são muito mais um tipo de conhecimento muito específico para cada terafa e inferidas a partir do que há ao redor do robô do que um tipo particular de dado das leituras dos sensores, várias questões precisam ser respondidas antes de usá-las em tarefas robóticas. Nós discutir os seguintes tópicos:

- Decidir qual tipo de informação semântica é possível inferir e associar ao que há ao redor do robô e que é relevante para a tarefa
- Como realizar a inferência ou estimação da informação semântica
- Como usar a informação semântica para melhorar a performance do robô em uma tarefa partícular.

Uma vez que informação semântica é relativamente nova na literatura, o primeiro ponto é frequentemente discutido no contexto de informação geométrica. Resumidamente, para o contexto de uma terafa robótica, qual informação não está explicitamente disponível no ambiente, mas poderia ser inferida ou estimada para melhorar a performance do robô? Isso demanda uma profunda compreensão da tarefa e das características gerais do ambiente onde o robô opera. Uma inspiração para responder este ponto é considerar como humanos pensam e raciocinam sob as mesmas circunstâncias e resolvem tal tarefa, e como eles processam as informações do ambiente para realizar a tarefa com eficiência.

Dependendo da informação semântica disponível, pode ser necessário usar métodos baseados em aprendizagem de máquina para estimá-la. Por exemplo, treinar um modelo de aprendizagem profunda para estimar a traversabilidade de terrenos para um robô terrestre para ambientes externos pode fornecer um resultado adequado. Contudo, apesar da necessidade do treinamento, a qualidade da solução depende dos dados de treinamento, e esta abordagem não escala muito bem. Estimativas baseadas em probabilidade aparecem como uma segunda opção, já que elas não requerem um grande conjunto de dados para treinamento, e aceita uma larga variedade de modelos.

O terceiro e último ponto, que é o uso apropriado da informação semântica inferida no sistema do robô, é crucial para completar com sucesso uma tarefa. Conforme o robô ganha mais informação do ambiente, é importante manter as estimativas atualizadas, e é

ainda melhor se as estimativas se tornam mais robustas ao longo do tempo.

A exploração de informação semântica na robótica é uma ideia que recentemente tem ganhado atenção dos pesquisadores, e então, a maioria dos desafios estão ainda não resolvidos. Uma forma simples de avançar os limites e investigar este problemas é estudar as vantagens da informação semântica em diferentes áreas. Nesta proposta de tese, nós escolhemos a tarefa com um alto nível de dificuldade que pode ter benefícios com a informação semântica: busca por objetos (BPO) em ambientes internos desconhecidos, um problema ainda não resolvido na robótica.

Em tarefas de BPO, o objetivo do robô é encontrar um objeto específico no ambiente por meio de um sensor visual. Normalmente, o ambiente é desconhecido para o robô, e os dados que ele usa para a busca são coletados com seus próprios sensores. Uma vez que nós estamos complementando a percepção do robô com modelos de informação semântica pra diferentes tarefas de BPO, o conhecimento extra do ambiente inferido pelo robô tem que auxiliar a busca por reduzir o espaço de busca. O robô planeja a estratégia de busca que estima as regiões mais promissoras que contém o objeto específico. Esta proposta de tese explora os avanços em tarefas de BPO pelo uso de informação semântica inferida a partir de duas fontes de dados ignorados pela comunidade científica: textos e obstáculos dinâmicos.