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

Ph.D. Thesis Proposal

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

Porto Alegre
October 2021

"If I have seen farther than others,

it is because I stood on the shoulders of giants."

— Sir Isaac Newton

ABSTRACT

Este documento é um exemplo de como formatar documentos para o Instituto de Informática da UFRGS usando as classes LaTeX disponibilizadas pelo UTUG. Ao mesmo tempo, pode servir de consulta para comandos mais genéricos. *O texto do resumo não deve conter mais do que 500 palavras.*

Keywords: Formatação eletrônica de documentos. Padronização de documentos. Instituto de Informática da UFRGS. LaTeX. ABNT. UFRGS.

Using LaTeX to Prepare Documents at II/UFRGS

RESUMO

This document is an example on how to prepare documents at II/UFRGS using the LaTeX classes provided by the UTUG. At the same time, it may serve as a guide for general-purpose commands. *The text in the abstract should not contain more than 500 words.*

Palavras-chave: Electronic formatting of documents. Instituto de Informática da UFRGS. LaTeX. ABNT. UFRGS.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

OS	Object Search
SLAM	Simultaneous Localization and Mapping

CONTENTS

1 INTRODUCTION

The first decades of research in Mobile Robotics, from the beginning until 2004, handled the challenges of connecting efficiency and data association. They introduced probabilistic formulations to path planning, exploration, simultaneous localization and mapping (SLAM), and many other areas. Some of the approaches from these areas are still popular nowadays, such as RaoBlackwellised Particle Filters and Extended Kalman Filters. The majority of them were based on ultrasonic or lidar sensors, as they were the most popular and robust sensors at the time. Consequently, the outcome maps were mostly 2D grid ones, in which the cells represented the free, occupied (obstacles), and unknown regions (CESAR et al., 2016).

After building a solid foundation for many problems with probabilistic approaches, the research community took a forward step. They concentrated on improving the properties of the already proposed and new approaches like observability, convergence, and consistency (CESAR et al., 2016). Simultaneously in this period (2004-2015), visual sensors have been in the spotlight as an alternative to gain information about the environment. Their considerable improvement in data quality and variety (e.g., depth images, point clouds, stereo images) aided their increasing employment. In fact, building 2D and 3D maps from the environment with a visual sensor resulted in a new term, Visual SLAM (SALAS-MORENO, 2014).

Mobile robotics have enjoyed formidable advantages in performing tasks that only expect robots to navigate free spaces and avoid obstacles. Moving items from point A to point B or vacuuming free spaces are examples of robotics tasks with satisfactory solutions. However, the same level of success does not apply so far to many other high-level tasks that robots are supposed to dealing with nowadays. Since mobile robotics shifted its focus from factory floors and assembly lines to everyday living spaces, robots are demanded to perform human-like tasks in different scenarios that are not necessarily as strict and organized as the industrial world (AYDEMIR, 2012). Relying only on purely geometric maps and having limited perceptions that do not allow going beyond basic geometric representations does not allow the robot to obtain a high-level understanding of the environment. This might be the reason for robots not prosper as much in high-level tasks. The robots are deprived of processing the environmental data to infer or estimate extra valuable knowledge useful in various tasks. We then claim that when building autonomous robots, the high-level information inferred from the sensor readings (semantic information) has to be heavily exploited to complement the robot's perception.

The association of semantic information (or concepts) to geometric entities in the map is called semantic mapping, one of the newest topics the researchers have explored. It enhances the robot's autonomy and robustness in many ways, besides facilitating some high-level tasks (CESAR et al., 2016). Fig. 1.1 is an image from Zoox's autonomous car, and it illustrates the advantage of using semantic information in robotics tasks. If the car maps this scene with its geometric perception, it will have the information that four obstacles are in its front, and two are closer than the other two. With a different perception based on semantic information, the car associates the concept of "person" to the three people and "vehicle" to the truck within the scene. Most importantly, it estimates that one person is distracted using his phone, and another is holding a stop sign. Combining the detection of a walking person and a phone allows the car to estimate the semantic information that this person is likely distracted. Hence, the car should drive itself even more carefully. This whole process is natural for human drivers, but the same can not be said about robots.

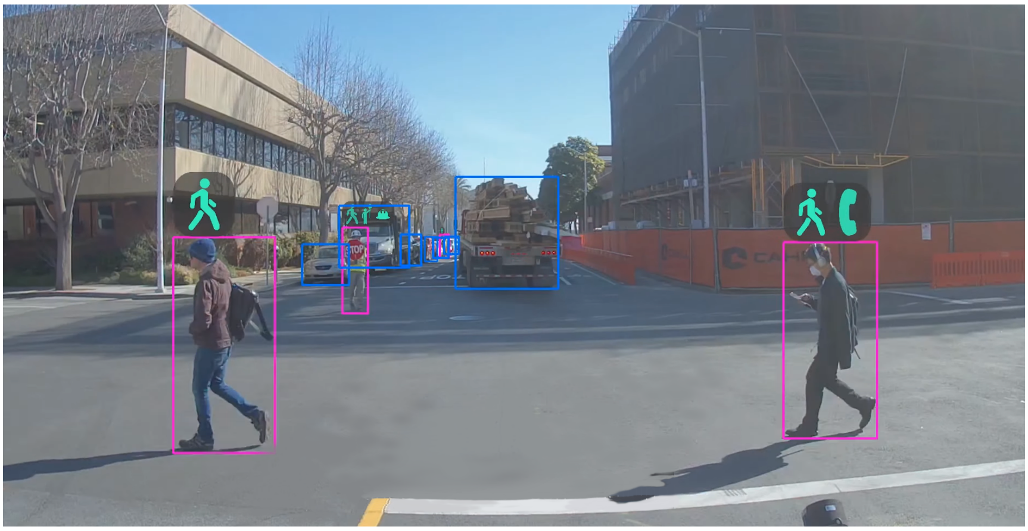

Figure 1.1 – plane.

As semantic information is more like a specific knowledge for each task and inferred from the robot's surroundings than a particular type of data from a sensor reading, several questions must be answered before using it in a robotic task. We see the following as noteworthy challenges:

- Deciding on what type of semantic information is possible to infer and associate to the robot's surroundings that is relevant to the task;
- How to perform the inferring or estimate the semantic information;

- How to use semantic information to improve the robot's performance in a particular task.

Since semantic information is relatively new in the literature, the first point is frequently discussed in the context of geometric information. Briefly, for the context of a robotic task, what information is not explicitly in the environment but could be inferred or estimated to improve the robot's performance? This demands a deep understanding of the task and the general environment characteristics where the robot operates. An inspiration for answering this point is to consider how humans reason under the same circumstance and solve such a task and how we process the environment's information to accomplish the task efficiently.

Second, depending on the needed semantic information, it may be necessary to use methods based on machine learning to estimate it. For example, training a deep learning model for estimating terrain traversability for an outdoor ground robot may provide a suitable result. However, besides the training requirement, the solution's quality depends on the training data, and this approach does not scale well. Probabilistic-based estimations appear as a second option, as it does not require a large set of data for training, and accepts a wide range of different models.

The third and last point, the proper use of the inferred semantic information in the robot's system, is crucial for successful task completion. As the robot gains more information from the environment, it is important to keep updating the estimations, and it is even better if the estimations become more robust over time.

The exploitation of semantic information in robotics is an idea that has recently gained attention from researchers, and thus, most of the challenging problems are still unsolved. A simple way of pushing the limits further and exploring these problems is to study the advantages of semantic information in different areas. In this thesis, we have chosen a task with a high difficulty level that can benefit from semantic information: object search (OS) in indoor and unknown environments, a yet unsolved problem in robotics.

In OS tasks, the robot's goal is to find a target object in the environment with a visual sensor. Usually, the environment is unknown to the robot, and the data it uses for searching are gathered with its own sensors. Since we are complementing the robot's perception with semantic information models for different OS tasks, the extra knowledge from the environment inferred by the robot has to aid the OS searching by reducing the search space. The robot plans a search strategy that estimates the most promising regions to contain the target object. This thesis exploits the improvements in OS tasks by the use of semantic information inferred from two different data sources disregarded by the research community: text and dynamic obstacles.

1.1 Outline

The outline of this thesis is as follows.

1.2 Contributions

Parts of this thesis have been previously published or submitted as journal articles. The following publications are the results of research carried during this PhD:

1. MANTELLI, M. et al. Temporal object search system based on heat maps. *Journal of intelligent & robotic systems* (in review), Springer, v. 101, n. 2, p. 1–23, 2021.
   Summary and Individual contribution: This paper is on how to search a target object in unknown and dynamic environments efficiently. As opposed to other OS works that consider the objects' position static and ignore the human-object interaction, the idea presented in this work is to incorporate a person's routine and habits in the search strategy. This work aimed to model the semantic information of how objects are moved over time within an environment and use the inferred information to reduce the searching space. 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. The contribution of the author of this thesis is in modeling the semantic information as part of the search strategy and in building a heat map with the inferred data.
2. 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.
   Summary and Individual contribution: This paper is on how to find a target door label based on text analysis. Although humans heavily rely on texts for accomplishing several tasks, text as a data source is not very popular in robotics. In this work, we have argued that texts have a great potential for providing search clues and are often found in man-made environments. This idea came from the human behavior

when searching for a someone's office in an unknown building, and how the door labels are used for estimating whether the current 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.

2 THEORETICAL BACKGROUND

In the previous chapter, we have argued that multiple robotic tasks would benefit from exploiting the semantic information inferred from everyday 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 expand our understanding regarding the benefits of employing semantic information to improve 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, and hence, we address the basic concepts of such systems and other basic concepts in mobile robotics. Besides, we cover the generic and central formulation of OS problems, which is the basis for the works presented in Chapters X and Y [TO DO].

2.1 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 (CITE PROB. ROBOTICS).

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 sensor provides them. Inertial measurement units and wheel encoders are 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 Fig. [REF THE FIG]. In *Localization*, the map is known in advance, and hence, only the robot's pose is estimated. The opposite happens in *Mapping*, as the map is built based on the robot's pose. Lastly, in *SLAM*, which combines the two previous problems, none of them is given a priori, and therefore, both are estimated simultaneously.

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. (CITE PROB. ROBOTICS). There are multiple localization problems, and not each of them is equally difficult. 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 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 [REF THE FIG], which depicts a few iterations of the robot's pose estimation and how the variables are used. The map $\boldsymbol{m} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \boldsymbol{m}_3, \boldsymbol{m}_4)^T$ 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 and predicts in linear dynamics and measurement functions (CITE KALMAN), whereas the grid-based filter approximates the estimations by decomposing the state space into finitely many regions of the grid map (CITE GRID). 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 (CITE PARTICLE). The particle filter implementation for mobile robotics is also known as Monte Carlo Localization (MCL), widely used in many different robotics applications for multiple robot types.

Mapping, for the case of the robot's poses are known, is the problem of generating consistent maps from noisy and imprecise measurement data (CITE PROB ROB). 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 Fig [REF FIG]. Comparing the graphical models of the localization and mapping problems, [REF FIGS], 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 (CITE EXAMPLES images from visual sensors or a vector of distances from a 2D lidar.). 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 (CITE PROB ROB). 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 (CITE PROB ROBOT). 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 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 [REF FIG]. 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 SLAM, which focus on estimating only the posterior over the current robot's pose $\boldsymbol{x}_t$ and the map $\boldsymbol{m}$, $p(\boldsymbol{x}_t, \boldsymbol{m} \mid \boldsymbol{z}_{1:t}, \boldsymbol{u}_{1:t})$, and the full SLAM, which 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 (CITE MAFFEI THESIS). The online

SLAM problem that is based on particle filter is known as Rao-Blackwellized particle filter (RBPF) (CITE LOTS OF PAPERS). 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 (CITE GMAPPING). 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.

2.2 OS problem formulation

The OS problem discussed in this chapter aims to find an efficient strategy for localizing a target object in a large unknown indoor environment. Since our works in this thesis are based on a 2D grid map, the search strategies 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 traveled distance by the robot during the search because the longer is 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.

The OS problem discussed in this chapter aims to find an efficient strategy for localizing 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 $\Psi(c)$ 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 $\Psi(c)$, it is possible to address the OS problem in three different ways:

- **$\boldsymbol{m}$ and $\Psi(c)$ are known**: the problem becomes a sensor placement, aiming to reduce the search cost by moving the robot straight to the cell c.
- **$\boldsymbol{m}$ 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 $\Psi(c)$ 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 ad the exploration vs. exploitation problem.

In this thesis, both the second and third points are considered, addressed individually in different works, in chapters [REF CHAPTERS]. 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

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

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

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

APPENDIX A — RESUMO EXPANDIDO

Resolução 02/2021 – Redação de Teses e Dissertações em Inglês Dissertações de Mestrado e Teses de Doutorado do PPGC, bem como outros trabalhos escritos tais como Proposta de Tese e PEP, poderão ser redigidas em inglês desde que contenham um título e resumo expandido redigidos em português. O resumo expandido deve conter no mínimo duas páginas inteiras, deve aparecer como apêndice e deve conter as principais contribuições e resultados do trabalho.