Chapter 1

Introduction

If some day robots are to get closer to what science fiction depicts, they will definitely require a rich perception and representation of the environment. In the past, robots used sonars as instruments for this task. Navigating with a sonar is similar to walking in a dark room trying to feel the walls and objects with the hands. Comprehensibly the first localization algorithms relied almost completely in the movements of the robot, just using the perception of the environment to correct the errors in the odometry (Elfes, 1989, 1990; Moravec, 1988). Later, ladars provided more reliable measurements, however, ladars are still too expensive for most applications, and the 2D range scanners are not able to provide enough information to qualitatively characterize places, not to mention objects. Lately, advances in computer vision along with an improvement in digital cameras have risen an increasing interest within the robotics field towards a vision-based autonomous robot.

However, extracting meaningful information from images has proven to be an arduous task. Myriad problems plague visual information, like perspective transformations introduced when the point of view changes, occlusions or motion blur, just to name a few.

Nevertheless, we humans, as vision-based autonomous navigating agents, get around these problems and manage to recognize places and objects. Numerous studies have been dedicated to understand how animals construct their mental maps of the environment and how they use them to navigate. A notable example is the work of Tolman (1948) where the author introduces the idea of a cognitive map based on ethological experiments with rodents. According to this study, rats construct mental representations of places based on the spatial relation of environment's features.

This theory gained strength when O'Keefe and Dostrovsky (1971) identified *place cells* in rodent brains. This kind of cells are neurons, mainly located in the hippocampus, that fire when the rat is located in certain places. These neurons are activated primarily by visual cues, but also by movements because they show activity even in the dark.

When published, this study had little impact on the robot navigation research

community, still in its first stages. However, time has passed and the field of robot navigation has seen enormous progress. Since then, many interesting robot navigation models taking inspiration from Tolman's theory have been proposed, such as the TOUR model by Kuipers (1978), that is designed as a psychological model of human common-sense knowledge of large-scale space; the RPLAN by Kortenkamp (1993), a model of human cognitive mapping adapted to robotics, and evaluated in an indoor scenario combining inputs from vision and sonar sensors; or the schematic maps by Freksa et al (2000) that, based on the idea of cognitive maps, distinguished between different levels of abstraction for map representation and their applications: representations closer to the geometric reality of the world are useful for local navigation and obstacle avoidance. On the other hand, more abstract or schematic representations can help in global localization and path-planning tasks.
In the beginning, these type of approaches materialized in the form of topological localization systems such as the ones described by Filliat and Meyer (2003), that make an extensive survey of internal representations of spatial layout for mobile robots with a focus on localization, or the more recent topological approach by Tapus and Siegwart (2006), that propose to use a signature which they call *fingerprint* to represent a room. This signature is constituted by a circular string that encodes the distribution of color blobs and sharp edges –extracted from omnidirectional images and a pair of 2D laser range scanners pointing in opposite directions respectively.
Recently, more ambitious approaches in the cognitive sense have been undertaken, as the one by Vasudevan et al (2007), that proposed a hierarchical probabilistic concept-oriented representation of space, constructed from objects detected in the environment and their spatial relationships. Such representation allows to endow the robot with a reasoning capacity that transcends the question "*where am I?*", typically pursued by the previous localization systems, by giving it the ability to infer the purpose or category of the room through the semantically meaningful elements or objects that can be detected.
However, if this type of approaches are to succeed, they will undoubtedly require much more advanced perception capacities than the ones typically found in a robot nowadays. Along the way to this ambitious goals in robotics research, this thesis contains the contributions described in the following section.

1.1 Contributions

The main contribution of the first part of this thesis is a signature to characterize places similar in spirit to the method proposed by Tapus and Siegwart (2006) in that it uses an omnidirectional vision sensor to perceive the environment. However, instead of relying also in range information provided by laser range scanners, the place model proposed here is purely vision based. In this approach, a place is characterized as a *constellation* of combinations of different types of visual feature regions extracted from a panoramic image that can be used as a node of a topological map graph.

These types of features are designed to be resistant to viewpoint and illumination changes and, in consequence, the proposed signature is also resistant to some extent to these problems. Furthermore, as the signature is composed from many individual features, it can tolerate some degree of dynamic changes in the environment while still being capable of recognizing the place. In order to test the proposed method, we have performed localization tests in various sequences of panoramas taken in different rooms of various buildings.
Since analyzing the whole map can become a time consuming task, we propose and evaluate a fast re-ranking method based in the *bag of features* approach to speed-up this step. Finally, we show that the presented signature is notably resistant even in the case of using a conventional perspective camera to perform localization.
In order to allow the robot to move between the nodes of the topological map, the second contribution of this first part is a biologically inspired inexpensive visual homing method based on the Average Landmark Vector (ALV). This homing method is able to determine the direction home by comparing the distributions of landmarks corresponding to the *home* with the current omnidirectional images, but without having to explicitly put them in correspondence. Typically, artificial landmarks have been used in experiments with the ALV. However, the method presented in this work combines the ALV with the feature regions employed earlier for global localization, thus complementing the localization method. First, a theoretical study is performed to evaluate the applicability of the proposed method in the domain of the local features and, next, it is evaluated in real world experiments showing promising results.

Even though the localization method proposed in the first part of this thesis is able to reliably model and recognize places, still few semantic knowledge about the world is available to the robot to reason with. As mentioned earlier, Vasudevan et al (2007) proposed a powerful space representation constructed from semantically rich elements of the environment. However, in order for this model to be applicable, a fast and robust object recognition or classification method is indispensable.
Indeed, not only localization would benefit from having a robust, generalistic and easily trainable object recognition system. Also other fields such as robot manipulation, human robot interaction and, in general, any discipline that addresses a practical use of robotics in a not highly structured environment would benefit from such a method. On the other side, computer vision is obtaining impressive results with recent object recognition and classification methods, but we are aware of little effort on porting it to the robotics domain. Therefore, a lightweight object perception method which allows robots to interact with the environment in a human cognitive level is still lacking.
In order to help reduce a bit this gap, in the second part of the thesis the main contribution is the evaluation of two successful state of the art object recognition methods – the SIFT object recognition method from Lowe (2004), and the Nister and Stewenius (2006) Vocabulary Tree – on a realistic mobile robotics sce-

nario, that includes many of the typical problems that will be encountered when roboticists try to use these methods on practical matters. Both methods have several properties that make them attractive for the problem of mobile robotics: the SIFT object recognition method detects object hypothesis location up to an affine transformation and has a low ratio of false positives; the Vocabulary Tree is a *bag of features* type method that was designed with the objective of being fast and scalable. Furthermore, it is suitable for types of objects that may confuse the SIFT method because of few texture or repetitive patterns. Additionally, and more importantly, several modifications and improvements of the original methods are proposed in this thesis in order to adapt them to the domain of mobile robotics.
The selected algorithms are evaluated under different perspectives, as for example:

- Detection: Does the method have the ability to easily and accurately detect where in the image is located the detected object? In most situations, large portions of the image are occupied by background texture that introduce unwanted information which may confuse the object recognition method.
- Classification: A highly desirable capability for an object detection method is to be able to generalize and recognize previously unseen instances of a particular class, is this achievable by the method?
- Occlusions: Usually a clear shot of the object to recognize will not be available to the robot. An object recognition method must be able to deal with only partial information of the object.
- Image quality: Since we are interested in mobile robots, motion blur needs to be taken into account.
- Scale: Does the method recognize the objects over a wide range of scales?
- Texture: Objects with a rich texture are typically easier to recognize than those only defined by its shape and color. However, both types of objects are equally important and we want to evaluate the behavior of each method in front of them.
- Repetitive patterns: Some objects, such as a chessboard, present repetitive patterns which cause problems in methods that have a data association stage.
- Training set resolution: Large images generate more features at different scales that are undoubtedly useful for object recognition. However, if training images have a resolution much higher than test images descriptor distributions may become too different.
- Input features: Most modern object recognition methods work with local features instead of raw image pixels. There are two reasons for this: in the first place, concentrating on the informative parts of the image the size of

the redundant input data is significantly reduced and, on the second place, the method is insensitive to small pixel intensity variations due to noise in the pixels, as well as small changes in point of view, scale or illumination. We evaluate several state of the art visual feature detectors.

- Run-Time: One of the most important limitations of the scenario we are considering is the computation time. We want to measure the frame-rate at which comparable implementations of each method can work.

From the obtained results, conclusions on the methods viability for the mobile robots domain are extracted and some ways to improve them are suggested. The final aim of this work is to develop or adapt an object recognition method that is suitable to be incorporated in a mobile robot and used in common indoor environments.
However, as it was found that none of the evaluated object recognition methods completely fulfilled the requirements of a robotics application, we proposed, as a last contribution, a Reinforcement Learning based approach to select on-line which object recognition schema should be used in a given image. The Reinforcement Learning approach is based on low level features computed directly from the image pixels, such as mean gray-level value or image entropy. It was evaluated in a challenging dataset and found to have very good performance. Another possible use of this method –although not directly addressed on this work because of time constrains– could be speeding up the Nister and Stewenius Vocabulary Tree by quickly discarding irrelevant areas of an image.

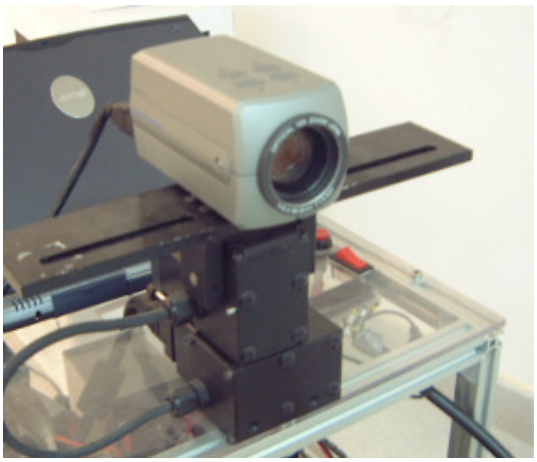
1.2 Robot

All experimentation carried on through this thesis have been done with the help of a real robot. Figure 1.1 shows pictures of it with the single-camera and the stereo head modes. The robot is a Pioneer 2AT, and is equipped with a Directed Perception PTU 46-70 pan tilt unit. On top of the pan tilt unit one or two Sony DFW-VL500 cameras are mounted. The cameras have a resolution of 640×480 pixels. The CPU of the robot is an Acer Travelmate C110 laptop (Intel Pentium M 1000MHz, 799 MHz, 760 MB RAM) placed on top of the robot and running Microsoft Windows XP. For the object recognition experiments, the platform that supports the cameras and the laptop was raised in order to gain a more human-like perspective and be able to see objects on top of the tables.

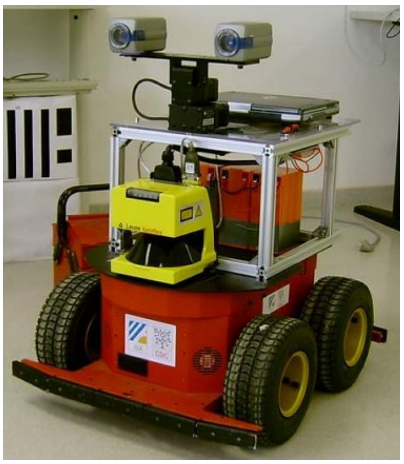
1.3 Publications

From the work carried on while pursuing this thesis, several publications have been derived:

- A. Ramisa, A. Tapus, R. Lopez de Mantaras, R. Toledo; "Mobile Robot Localization using Panoramic Vision and Combination of Local Feature



(a)



(b)

Figure 1.1: (a) The Pioneer 2AT robot used in the experiments described in Chapters 3 and 4. (b) The stereo setup has been used in Chapters 5 and 6 of the thesis.

Region Detectors", In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation*, Pasadena, California, May 19-23, 2008, pp. 538-543.

- R. Bianchi, A. Ramisa, R. Lopez de Mantaras; "Learning to select Object Recognition Methods for Autonomous Mobile Robots", In *Proceedings of the 18th European Conference on Artificial Intelligence*, Patras, Greece, July 21-25, 2008, pp. 927-928.

- R. Bianchi, A. Ramisa, R. Lopez de Mantaras; "Automatic Selection of Object Recognition Methods using Reinforcement Learning", In *Recent Advances in Machine Learning (dedicated to the memory of Prof. Ryszard S. Michalski)*. Springer Studies in Computational Inteligence. To appear.

- A. Ramisa, S. Vasudevan, D. Scharamuzza, R. Lopez de Mantaras, R. Siegwart; "A Tale of Two Object Recognition Methods for Mobile Robots", In *Proceedings of the 6th International Conference on Computer Vision Systems, Lecture Notes in Computer Science 5008*, Santorini, Greece, May 12-15, 2008, pp. 353-362.

- A. Ribes, A. Ramisa, R. Toledo, R. Lopez de Mantaras; "Object-based Place Recognition for Mobile Robots Using Panoramas", In *Proceedings of the 11th International Conference of the ACIA, Frontiers in Artificial Intelligence and Applications, Vol. 184. IOS Press*, Sant Marti d'Empuries, Girona, October 22-24, 2008, pp. 388-397.

- A. Ramisa, R. Lopez de Mantaras, D. Aldavert, R. Toledo; "Comparing Combinations of Feature Regions for Panoramic VSLAM", In *Proceedings of the 4th International Conference on Informatics in Control, Automation and Robotics*, Angers, France, May 2007.

- M. Vinyals , A. Ramisa, R. Toledo; "An Evaluation of an Object Recognition Schema Using Multiple Region Detectors", In *Proceedings of the 10th International Conference of the ACIA. Frontiers in Artificial Intelligence and Applications, Vol. 163. IOS Press*, Sant Julia de Lria, Andorra, October 2007, pp. 213-222.

- A. Goldhoorn, A. Ramisa, R. Lopez de Mantaras, R. Toledo; "Using the Average Landmark Vector Method for Robot Homing", In *Proceedings of the 10th International Conference of the ACIA. Frontiers in Artificial Intelligence and Applications, Vol. 163. IOS Press*, Sant Julia de Lria, Andorra, October 2007, pp. 331-338.

- D. Aldavert, A. Ramisa, R. Toledo; "Wide Baseline Stereo Matching Using Voting Schemas", In *1st CVC Research and Development Workshop*, October 2006.

- A. Ramisa, D. Aldavert, R. Toledo; "A Panorama Based Localization System", In *1st CVC Research and Development Workshop*, October 2006.

Besides, three more papers have been submitted for publication:

- A. Ramisa, A. Tapus, D. Aldavert, R. Toledo, R. Lopez de Mantaras; "Robust Vision-Based Localization using Combinations of Local Feature Regions Detectors".
- A. Goldhorn, A. Ramisa, D. Aldavert, R. Toledo, R. Lopez de Mantaras; "Combining Invariant Features and the ALV Homing Method for Autonomous Robot Navigation based on Panoramas".
- D. Aldavert, A. Ramisa, R. Toledo, R. Lopez de Mantaras; "Visual Registration Method for a Low Cost Robot".

1.4 Outline of the Thesis

This thesis contains eight chapters that can be grouped in two parts of closely related content: chapters three and four describe an approach to visual-based indoor global localization without any semantic capability, while chapters five to seven address the issue of object recognition, the main difficulty if a semantically enhanced approach is to be attempted. Chapters two and eight present the related work and preliminaries, and the conclusions and future work respectively. Next is a brief outline of the thesis starting at chapter two.

Chapter 2: Related Work and Preliminaries

In this chapter, we review literature related to both of the robot localization and object recognition fields. Approaches to global localization with similarities to the one proposed are discussed, and interesting methods for object recognition that have some characteristics relevant for robotic applications are presented. Finally, some preliminaries on the type of visual features employed through all this work are reviewed.

Chapter 3: Global Localization Method

In this chapter our proposed topological indoor localization system is presented and evaluated in a dataset of panorama sequences from various buildings. Additionally, a re-ranking of the map nodes to speed up the search of the current location is proposed. Also experiments with conventional perspective images instead of panoramas are performed.

Chapter 4: Appearance-Based Homing with the Average Landmark Vector

In order to travel between the topological map nodes proposed in the previous chapter, here we experiment with the ALV homing method using image features as the ones employed for localization. Simulation experiments, as well as real-world ones, are done to verify the robustness of the method.

Chapter 5: SIFT Object Recognition Method

In this chapter the SIFT object recognition method, as well as the proposed modifications, are described and evaluated. First we review the effects of varying the different parameters of the algorithm and, next, the most successful configurations are further evaluated on the whole test data.

Chapter 6: Vocabulary Tree Method

Similarly to the previous chapter, here different choices of parameters for the Vocabulary Tree method, and the proposed adaptations to detect objects in unsegmented images, are compared. The best performing combination is again evaluated on the whole test data.

Chapter 7: Object Recognition Method Selection with Reinforcement Learning

In our experiments, we found that both methods have advantages and drawbacks and therefore, in this chapter, we propose a Reinforcement Learning approach for selecting the best performing method for a given input image.

Chapter 8: Conclusions and Future Work

Finally, in this chapter, the conclusions of the thesis and future research directions are presented.