Abstract

After decades of unrealistic predictions and expectations, robots have finally escaped from industrial workplaces and made their way into our homes, offices, museums and other public spaces. These service robots are increasingly present in our environments and many believe that it is in the area of service and domestic robotics that we will see the largest growth within the next few years. In order to realize the dream of robot assistants performing human-like tasks together with humans in a seamless fashion, we need to provide them with the fundamental capability of understanding complex, dynamic and unstructured environments. More importantly, we need to enable them the sharing of our understanding of space to permit natural cooperation. To this end, this thesis addresses the problem of building internal representations of space for artificial mobile agents populated with human spatial semantics as well as means for inferring that semantics from sensory information. More specifically, an extensible approach to place classification is introduced and used for mobile robot localization as well as categorization and extraction of spatial semantic concepts from general place appearance and geometry. The models can be incrementally adapted to the dynamic changes in the environment and employ efficient ways for cue integration, sensor fusion and confidence estimation. In addition, a system and representational approach to semantic mapping is presented. The system incorporates and integrates semantic knowledge from multiple sources such as the geometry and general appearance of places, presence of objects, topology of the environment as well as human input. A conceptual map is designed and used for modeling and reasoning about spatial concepts and their relations to spatial entities and their semantic properties. Finally, the semantic mapping algorithm is built into an integrated robotic system and shown to substantially enhance the performance of the robot on the complex task of active object search. The presented evaluations show the effectiveness of the system and its underlying components and demonstrate applicability to real-world problems in realistic human settings.

**Keywords**: spatial understanding, semantic mapping, place recognition, place categorization, mobile robotics.
Acknowledgments

There are many people who have supported, encouraged and inspired me throughout this thesis. I have worked with and was surrounded by a number of great people who contributed in many ways to this research and deserve special mention.

First and foremost, I would like to express my gratitude to Patric Jensfelt, my main advisor, for his supervision, great engagement and guidance. Thank you for supporting my ideas and all aspects of my work, for rich and inspiring discussions and great collaboration, and last but certainly not least for your friendship, understanding and great sense of humor. I owe many thanks to Barbara Caputo for giving me the opportunity to explore the exciting field of computer vision, introducing me to research in general, as well as hosting me in her group at IDIAP during my internship. Thank you for your good advices, endless enthusiasm, supervision and great help. I am also very grateful to Henrik I. Christensen for inspiring me to pursue my research in the direction of robotics as well as for his support and sharing with me his vision of AI. Many thanks go to Stefan Carlsson, Jan-Olof Eklundh and Danica Kragic for their valuable advices and encouragement during the years spent at KTH.

This thesis has been made possible with the help and cooperation of Alper Aydemir and Kristoffer Sjöö as well as Adrian Bishop, Moritz Göbelbecker, Marc Hanheide, Luo Jie, Óscar Martínez Mozos, Muhammad Muneeb Ullah, Li Xing and Hendrik Zender, and all the great people involved in the CoSy and CogX projects. Kristoffer and Alper, I will never forget our coding nights and the Turkisk peppar! Furthermore, my thanks go to our administrative staff: Jeanna Ayoubi and Friné Portal. You were always very kind, helpful and patient.

It was a pleasure to share doctoral studies and life with wonderful people like Babak, Oscar, Niklas, Mattias and Pedro among others who become close friends. Special thanks go to my dearest friend Babak for being a part of my family and a person I can always count on. And thanks for insightful discussions and excellent proof reading! Babak, Marianna, Mattias, Niklas, Oscar and Pedro thanks for sharing my great passion for climbing and inspiring me to try new things and reach new goals, especially to finish Svensk Klassiker. That was absolutely fantastic time together! Babak, Njupeskär was just the beginning!

The time at CVAP would not have been the same without: Ahmad, Alessandro, Alireza, Arnaud, Carl-Henrik, Christian, Dan, Elin, Gareth, Gert, Hedvig, Heydar, Javier, Jeannette, Jorge, Josephine, Kai, Magnus, Miroslav, Mårten, Renaud, Omid, Sagar, Simone, Yasemin, Vahid. There are many more, former and current members of CVAP that have influenced my life in positive ways. The time spent in Switzerland would not have been as amazing without my climbing and skiing buddies: Ganga, Ferran, Petr and Stan and other great people: Bogdan, Chris, Constantin, Danil, Elisa, Francesco, Ghita, Guillermo, Hamed, Kate, Niklas, Radu, Roger, Tatiana, Tristan and Vincent. Thank you all!
Special thanks go to my Polish friends: Anna J. & Marcin, Anna P. (Dziubasek 1), Grześ (Dziubasek 2), Krzyś & Gabrysia & VaNTa Junior :-), Maciej & Fredrik, Pit & Dominika and Zygi for being always next to me and deeply in my heart. When we meet it feels like we never split! Krzyś, Grześ, Zygi, you’re the best climbing mates!

Lastly, and most importantly, I want to thank my family. My parents for constant love, support and always being for me. Kocham Was drodzy Rodzice! Finally, my Marianka who gave me strength to always move forward, through all the ups and downs of the PhD time. Thank you for giving it all meaning to me and making my world so rich, beautiful and complete.

Andrzej Pronobis
Stockholm, May 2011

This work was supported by the VR project COMPLEX, the EU FP7 Project CogX and the SSF through its Centre for Autonomous Systems.
List of Papers

The thesis is based on the following papers:


In addition to papers [A]-[G], the following papers have also been produced in part by the author of the thesis:


# Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Place Classification and Semantic Mapping</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Related Work</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>Summary of the Papers</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>Discussion and Conclusions</td>
<td>37</td>
</tr>
</tbody>
</table>

---

## 1 Introduction

1. Thesis Outline | 3 |

## 2 Place Classification and Semantic Mapping

1. Problem Statement | 9 |
2. Scenario | 11 |
3. Challenges | 13 |
4. Flexible and Extensible Multi-modal Place Classification | 13 |
5. A Systems and Representational Approach to Semantic Mapping | 15 |
6. Implementation and Integration with a Robotic System | 17 |
7. Evaluation Data, Procedures and Results | 17 |

## 3 Related Work

1. Place Classification | 19 |
2. Semantic Mapping | 25 |

## 4 Summary of the Papers

1. Paper A: Single-cue Place Recognition | 31 |
2. Paper B: Incremental Learning and Knowledge Transfer | 33 |
3. Paper C: Confidence Estimation and Cue Integration | 33 |
4. Paper D: Multi-modal Place Classification for Semantic Mapping | 34 |
5. Paper E: Spatial Knowledge Representation | 34 |

---
CONTENTS

4 Discriminative Cue Integration ........................................ D10
5 Place Classification for Semantic Space Labeling .................. D13
6 Experiments with Place Classification ................................. D17
7 Experiments with Semantic Space Labeling .......................... D28
8 Conclusions ........................................................................ D32
1 Index to Multimedia Extensions ........................................ D33
References ........................................................................... D33

E Representing Spatial Knowledge in Mobile Cognitive Systems E1
1 Introduction ...................................................................... E3
2 Related Work .................................................................... E4
3 Analysis of the Problem ..................................................... E5
4 Structure of the Representation .......................................... E7
5 Instantiations ..................................................................... E11
6 Conclusions and Future Works .......................................... E13
References ........................................................................... E13

F Understanding the Real World: Combining Objects, Appearance, Geometry and Topology for Semantic Mapping F1
1 Introduction ...................................................................... F3
2 Related Work .................................................................... F5
3 Semantic Spatial Understanding ......................................... F7
4 Categorical Models of Sensory Information ............................ F10
5 The Conceptual Map .......................................................... F10
6 System Overview ............................................................... F13
7 Experimental Scenario ........................................................ F14
8 Experiments ....................................................................... F16
9 Conclusions and Future Works .......................................... F19
References ........................................................................... F19

G Exploiting Probabilistic Knowledge under Uncertain Sensing for Efficient Robot Behaviour G1
1 Introduction ...................................................................... G3
2 Related Work .................................................................... G5
3 Conceptual Map ............................................................... G5
4 Switching Continual Planner .............................................. G10
5 Experimental Evaluation .................................................... G14
6 Conclusion ......................................................................... G18
References ........................................................................... G19
Part I

Introduction
Chapter 1

Introduction

The recipient of the 1969 Turing Award and the pioneer of robotics and artificial intelligence summarized the progress of robotics in the second half of the 20th century by saying:

In the fifties, it was predicted that in 5 years robots would be everywhere.
In the sixties, it was predicted that in 10 years robots would be everywhere.
In the seventies, it was predicted that in 20 years robots would be everywhere.
In the eighties, it was predicted that in 40 years robots would be everywhere.

Marvin Minsky

Those sentences clearly illustrate the unrealistic beliefs and expectations of the robotics community which did not foresee the challenges stemming from the complexity of unstructured human environments. Challenges, which required decades of research in such fields as signal processing, statistics, machine learning and computer vision to reach the level where the developed algorithms can be applied in real-world, practical applications.

Despite the fact that we might still be far from building robots that could possess human-like intelligence, we are closer than ever to actually fulfilling the dream of ubiquitous robots. Robots have already made their way to our homes, and many believe that within the next few years, we will see a dramatic growth in the area of domestic and service robotics [68, 67, 44]. Our idea of robots diverges from stationary machines operating in typical industrial workplaces and starts to resemble what Karel Čapek [127], the inventor of the word robot itself, had in mind: cheap, mobile intelligent machines present in every home. Those expectations are further confirmed by the development programs implemented by the robotics industry [68, 44] as well as government agencies (23, the Korean Ubiquitous Robot Companion program) which assumes popularization of cheap service robots to the extent of one robot in every household. It is not uncommon to hear statements of the following kind:
CHAPTER 1. INTRODUCTION

(a) iRobot Roomba \[30\] (b) Pleo \[35\] (c) Aldebaran Robotics Nao \[19\] (d) Anybots \[20\] (e) Home Assistant Robot AR \[134\] (f) Care-O-Bot \[21\]

Figure 1: Examples of commercially available service and domestic robots.

The question is no longer, Will you have a robot in your home in the future? But instead, How many?

Helen Greiner, iRobot Chairman and Co-founder, 2005

These next generation robots will not only have to track their position and navigate between points in space, but reason about space and their own knowledge, plan tasks and knowledge acquisition and interact with people in a natural way.

The robots deployed in real-world human environments are mostly relatively small and simple service robots with, so far, very limited capabilities. The market is dominated by cleaning robots such as iRobot Roomba \[30\] sometimes enhanced with additional functionalities such as visual navigation \[37\] or teleoperation \[38\]. Telepresence is another quickly developing application area which require robots operating among humans \[13, 20, 29\]. Simple robots are becoming popular also in the education and entertainment sector with such examples as Nao \[19\] or Pleo \[35\].
More complex commercially available platforms are mostly found in the surveil-

lance [31] and human assistance [21, 40, 36] application areas. Those platforms are

not only capable of navigation, but are expected to autonomously interact with the

environment and communicate with the human user. At the forefront of the applied

science of service robotics, we see multiple research and prototyping platforms with

the software and embodiment designed to operate in man-made environments. Plat-

forms such as the Home Assistant Robot [134], Asimo [28] or PR2 [41] were already

shown to perform complex manipulation and navigation tasks (e.g. performing

typical household chores or even preparing pancakes). Pictures presenting some of

the commercially available service and domestic robots are shown in Figure [1]

Many recent advances in fields such as computer vision and cognitive robotics

have been driven by the goal of creating artificial cognitive systems able to perform

human-like tasks in real-world settings. Several attempts have been made to design

integrated cognitive architectures and implement them on mobile robots [24, 111,

26, 51, 27, 69, 25, 85]. Those attempts focused on creating future systems that are

more versatile than those commercially available, able to operate in unstructured

environments and still providing a sufficient level of robustness. The tasks that have

been envisioned for those future robots involve interaction with the environment

and non-expert human users.

A cornerstone for such robotic assistants is their understanding of the space they

are to be operating in. Spatial understanding is a prerequisite for such basic tasks

as navigation, obstacle avoidance, autonomous exploration or even manipulation.

While knowing the position in the world, being able to explore the environment or

find routes to known locations is a fundamental capability for a mobile agent, there

are many other tasks of the future service robots that depend on the ability to per-

ceive and understand space. These include action planning, recording and recalling

episodic memories, reasoning about spatial concepts and their relations, interacting

with objects in the environment and, finally, human-robot communication.

Spatial knowledge constitutes a fundamental component of the knowledge base

of an embodied agent operating in large scale spaces. It is considered founda-

tional to all commonsense knowledge and provides grounding for other knowledge

types [76]. Research on such problems as human augmented mapping identified

spatial knowledge as one of the major elements permitting and facilitating human

robot interaction [74, 75]. In such view, the environment can be considered an ad-

ditional communication channel which allows for disambiguation and extension of

the communicated information. Furthermore, it can be seen as a common ground

or even a “representation” shared between the agent and the human user [51].

We can identify several different types of spatial knowledge depending on the

source, point of reference, spatial scale or level of abstraction and thus different

approaches to spatial knowledge representation. Geometric aspects of space can

be represented in terms of a metric map in which the agent’s location is simply a

set of raw metric coordinates. A different representation could abstract the metric

space into a set of discrete units and focus on the spatial topology. This distinction

resulted, over the years, in a broad range of approaches spanning from purely metric
to topological, and hybrid. Recently, particularly in the case of integrated robotic systems performing more complex tasks requiring action planning in large-scale environments, topological and hybrid models are gaining popularity allowing for better scalability as well as easier access and maintenance.

Another important type of spatial knowledge stems from the semantics encoded in various observable properties of space. In the case of indoor spaces, the environment provides valuable semantic information originating from humans as designers and users. Indeed, the ability to understand the semantics of space and associate semantic terms like “corridor” or “office” with spatial locations, gives a much more intuitive idea of the position of the robot than pure metric or topological location. If further extended with such semantic concepts as room shape, size and appearance or presence of objects of certain types, the robot’s spatial knowledge representation becomes much more meaningful from the point of view of the robot’s performance on complex tasks and human interaction. Let’s take the example of a domestic gopher robot, the task of which is to find objects. Clearly, such a robot could greatly increase its performance by considering semantic types of rooms and their correlation with the location of the searched object. Moreover, such a robot should be able to communicate its internal state of knowledge using concepts known to the operator to minimize training efforts. At the same time, the semantic information can extend the capabilities of a robot in the traditional tasks of localization, exploration, or navigation.

Despite the usefulness and importance of semantic spatial knowledge, this aspect of spatial modeling has been left out by many of the previous works, mostly due to its complexity. Producing real-time solutions extracting semantic information in a robotic system is a challenging problem. In particular, realistic environments pose challenges due to their dynamic character. Indoors, the appearance of places can change due to human activity or influence of illumination. Additionally, single observations are usually not sufficiently informative and spatio-temporal information fusion is required. Finally, most of the semantics can only be discovered through visual sensing which tends to be noisy and difficult to interpret. For those reasons, many of the previous works focused on such problems as pure localization and navigation, and semantic knowledge has been included only in basic forms. At the same time, the perception of semantics is greatly enhanced by the use and integration of other information sources such as the general visual appearance, objects discovered in the environment, topological connectivity or even human actions and dialogue.

To this end, this thesis focuses on providing a robot operating in a real-world environment with a complete and efficient representation of space including semantic information. The problem is constrained to man-made environments such as homes or offices which will constitute the working space of many of the future service robots and as made by humans for humans are rich in human semantic information. The representation is meant to support such typical human-like tasks as retrieving objects, performing household chores or guiding visitors, all of which require human interaction capabilities.
More specifically, this work addresses the problem of semantic mapping, i.e. creating a representation of the environment which grounds human spatial concepts to instances of spatial entities. The problem is addressed holistically from the point of view of systems and representations, starting from the level of topological and metric maps, through place classification and building models associating concepts with sensory information, up to the level of ontologies defining more abstract concepts and their relations.

First, an extensible approach to place classification is introduced providing models that link spatial concepts to sensory information originating from multiple modalities such as vision and laser range data. The models can be incrementally adapted to the dynamic changes in the environment and provide practical measures of confidence. Second, a complete systems and representational approach is proposed to address the problem of semantic mapping. This system is capable of incorporating semantic information extracted from such sources as the geometry and general appearance of places, presence of objects, topology of the environment and/or human input. Moreover, it is able to reason about spatial concepts and infer new knowledge about the environment which cannot be directly observed. Finally, the semantic mapping algorithm is built into an integrated robotic system and shown to substantially enhance the performance of the robot on the problem of active object search.

Thesis Outline

The rest of this thesis is structured as follows.

Chapter 2 Place Classification and Semantic Mapping

Chapter 2 discusses in detail the problems addressed in this thesis, the envisioned scenario and the resulting challenges. Then, the contributions and proposed solutions are roughly divided into four groups and briefly outlined.

Chapter 3 Related Work

Chapter 3 provides an overview of related work in the areas of place classification and semantic mapping. Moreover, the approaches proposed in this thesis are placed in context and compared to other works.

Chapter 4 Summary of the Papers

Chapter 4 introduces the reader to the papers included in the second part of the thesis. First, an outline of each paper is given. Then, the contributions of the author of the thesis are summarized.
Chapter 5: Discussion

Part II concludes with a discussion of the presented solutions and lessons learned. Moreover, the directions for future research stemming from the presented work are proposed.

Part II: Included Papers

The second part of the thesis contains the included publications in the order suggested in Chapter 4. The papers provide all the details about the proposed representations, algorithms and systems.
Chapter 2

Place Classification and Semantic Mapping

1 Problem Statement

The fundamental problem considered in this thesis is that of semantic mapping. In order to provide a clear definition of the problem, a few words must be said about the spatial semantics in general as seen in this work.

Spatial Semantics

In the view taken by this thesis, semantic information is expressed by the relations between spatial entities and a set of predefined concepts. These concepts are meant to be meaningful for humans and therefore are transferred to the robot either by direct interaction with a human user or by analyzing available common-sense knowledge databases such as Open Mind [32], ConceptNet [22], OpenCyc [34], or WordNet [42]. Recently, Internet search engines (e.g. Google Image Search), social networks (e.g. Facebook) and image repositories (e.g. Flicker) became a valuable source of common-sense knowledge obtained directly from user generated content.

An important concept employed by humans in indoor spaces is that of a room which can be loosely defined as a bounded area in the environment. Rooms tend to share similar functionality as well as many other spatial properties. In most cases, rooms are naturally categorized based on their functionality and can be described in terms of discrete concepts such as “a kitchen” or “an office”. Rooms can also be associated with other concepts describing their spatial properties. The experiments presented in this thesis employ such properties as the shape of a room (e.g. square or elongated), the size of a room (e.g. small or large, compared to other typical rooms) or the general appearance of a room (e.g. corridor-like or office-like appearance). However, more fine grained semantic descriptions are often desired. Those can be associated with objects or landmarks in the environment. One important landmark which facilitates segmentation of continuous space is a
door. Indeed, in the case of indoor environments, rooms are usually separated by doors or other narrow openings.

**Semantic Mapping**

Given the view on semantics presented above, *semantic mapping* can be defined as a process of building a representation of the environment which associates spatial concepts with spatial entities. The outcome of semantic mapping should ideally be a complete and efficient representation of the environment visited by the agent. The efficiency in this case is defined by the performance of the agent on certain typical tasks and the representation should not be decoupled from the tasks and treated in isolation. Such representation should not only contain the semantic information, but should explicitly represent the spatial entities to which the semantics is tied. Additionally, it is assumed that the robot starts without any prior knowledge that comes from actual observations of the environment in which it is to be operating. Instead, it is equipped with a set of categorical and conceptual models acquired either in other environments or from databases. An example of a map augmented with semantic information is shown in Figure 1. In this work, the problem is expanded by stating that knowledge should not only be derived directly from the immediate sensory information but should also be inferred based on the whole body of knowledge available. A typical example would be prediction of categories of objects that might be present in yet unexplored rooms connected topologically to a room for which evidence is available.

There are multiple sources of semantic information that the agent can exploit. Semantic spatial knowledge can be provided directly by the user, for instance through a situated dialogue. The topology of the environment itself can be a
2. SCENARIO

valuable cue for discovering semantic categories of rooms. A good example is a corridor which is likely to be connecting many other rooms. We also mentioned objects, landmarks and spatial properties of areas such as shape, size, or general appearance. Perception of these requires robust models of sensory information such as the visual models of object categories or models describing various shapes and appearances of spatial regions. The object detection, recognition and categorization problem is vastly researched in the computer vision community [101, 58, 66] and multiple approaches to modeling object categories, each having different limitations, are available both theoretically and as software implementations [33, 88]. This thesis is not concerned about building object or landmark models. Instead, attention is given to the problem of designing models of geometry and appearance of spatial regions for the purpose of place classification.

Place Classification

Place classification, can be characterized as a pattern recognition problem of assigning a region in an environment to one of predefined classes based on multi-modal sensory input and a set of models. In order to support the scenarios considered in this thesis, we assume a supervised case (either by a human or an independent sub-system). First, the models are built from a collection of labeled data samples acquired in places belonging to the modeled classes. The models store intrinsic visual and geometric properties of the classes. Then, the algorithm is presented with data samples acquired in one of the same places or in a novel place belonging to one of the modeled classes, possibly under different conditions. The goal is to classify correctly as much of the sensory data samples as possible.

Place classification can further be subdivided into place recognition and place categorization depending on the scenario. We talk about place recognition if the models are tested on the sensory data collected in the same environment in which the models were trained. Place recognition is mostly used as a solution for topological localization [122, 103, 52, 56] or together with traditional localization and mapping algorithms for initialization (e.g. in case of the kidnapped robot problem) [108] and loop closing [93, 82]. This is different from the problem of place categorization where the task is to classify test data captured in a novel, previously unseen place. In this case, the algorithms have to tackle additional challenges resulting from the within-category variability. Place categorization models will be employed as sub-components providing shape, size and appearance information about places to the semantic mapping system. However, this thesis proposes and evaluates a model which can be applied in a much broader context and to both place categorization and recognition.

2 Scenario

This section gives an overview of the general scenario for which the proposed algorithms were designed and in which they were evaluated. The primary assumption
is that the environment in which the robot operates is unstructured and does not contain any artificial markers or beacons. As the primary interest is the human semantics, the environments were constrained to indoor spaces, such as offices or homes, which are typical for the interaction between humans and robots \[135\]. In order to provide natural, real-world conditions, humans could be present and performing typical actions during the experiments.

The considered scenario assumes a mobile robot platform performing typical human assistance tasks. The platform is assumed to be equipped with a standard set of robotic sensors, in particular a monocular camera and optimally a laser range scanner. The fetch-and-carry task is used as a concrete application example. In this case, the robot is sent to find objects in a large-scale indoor space, often without any previous knowledge about that concrete part of the environment. Imagine the case where a mobile courier robot is tasked with finding and fetching an object on a 15-room office floor. It is unreasonable to assume that such a robot will receive timely updates on the exact locations of every relevant object. At the same time, it would be very inefficient to require the robot to scan the entire environment in search for the object. In such case, semantic information indicating the functionality of spatial regions and typical locations of objects belonging to certain categories (e.g. plates are often found in kitchens) could be very valuable and could greatly improve the performance of the robot.

From the point of view of the semantic mapping system, there is one more important element of the scenario. In the application mentioned above, the semantic mapping or place classification sub-subsystem is integrated into a larger robotic architecture. Therefore, the requirements and properties of other sub-systems should influence the design of the spatial understanding component.
3 Challenges

The considered scenario results in several challenges that semantic mapping and place classification systems have to tackle. First, the proposed solutions must co-exist with other components in an integrated system on a robot platform and work in real-time. This is a strong constraint on the computational complexity of the algorithms and their memory consumption. Additionally, the algorithms must deal with uncertain perceptions and this uncertainty must be modeled and presented to other components of an integrated robotic system such as a decision theoretic planner.

Other challenges stem from the characteristics of the environment. Real-world indoor environments are usually dynamic and their appearance changes over time. For example, the appearance is affected by illumination changes. For a visual sensor, the same room might look different during the day, during sunny weather, under direct natural illumination, and at night with only artificial light turned on. The perception of the environment is also influenced by short term (presence of people) and long term (furniture moved around, objects being relocated etc.) human activities. The models of sensory information must be robust to those variations and, in case of categorization, must be able to generalize across multiple instances of places belonging to the same category. Additionally, many indoor places cannot be uniquely characterized by their geometry, or even general appearance, and integration of multiple types of information is required. As a result, most approaches that work well for outdoor environments will perform poorly when applied indoors [103].

Another set of challenges arises due to the properties of the sensors employed. The fact that the sensors have a limited field of view requires the algorithms to internally integrate information and deal with frequent occlusions. Moreover, viewpoint variations cause the sensors to capture different aspects of the same place. Many viewpoints in separation do not contain discriminative information (e.g. when the robot is looking towards the wall) and the information that the robot gathers is not evenly spread across the viewpoints.

The fact that so many different parameters influence the performance of a semantic mapping or place classification system is another challenge itself, especially burdensome at the design stage. As the results depend on the choice of training and test input data, which in real environments would change over time, it is hard to measure the influence of the different parameters on the overall performance of the system. There is a need for realistic benchmarks and databases which would allow for precise analysis and simplification of the experimental process.

4 Flexible and Extensible Multi-modal Place Classification

This thesis contributes a method for multi-modal place classification. The method effectively utilizes information from different robotic sensors by fusing multiple visual cues and laser range data in order to combine the stability of geometrical
solutions with the versatility and richness of vision. The method relies on discriminative Support Vector Machine (SVM) \(^5^5\) models of place classes known for their superior generalization abilities. The models are built from different types of both global and local visual features as well as a set of geometrical cues extracted from range data. For the vision channel, either the Scale-invariant Feature Transform (SIFT, \(^8^1\)) or the Speeded Up Robust Features (SURF, \(^4^7\)) local descriptors are used, combined with the bag-of-words approach \(^6^2\) for place categorization. The Composed Receptive Fields Histograms (CRFH, \(^8^0\)) are used as global visual features. For the laser channel the simple geometrical features proposed in \(^8^9\) are applied. The resulting algorithm is capable of real-time and robust place recognition as well as categorization and was evaluated for both problems. It is robust to different types of natural variations that occur for indoor environments due to changing illumination and configuration of furniture and small objects.

Several extensions of the models are proposed that increase the robustness in different situations. First, a confidence estimation algorithm is contributed which provides a practical measure of confidence of the decision of the place classification algorithm. The method is based on the distance of the test sample from the SVM hyperplane and the average distance of each training class. Through experiments, it is shown to increase robustness and reliability as well as efficiency in case of multi-cue classification. Second, an algorithm that integrates various cues and modalities is proposed which is based on the principle of high-level discriminative accumulation. For each cue, a discriminative SVM classifier is trained which outputs a set of scores encoding confidence of the decision. Integration is then achieved by either accumulating the scores linearly or feeding them to a Support Vector Machine (SVM, \(^5^5\)). Such an approach allows to optimally combine cues, even obtained using different types of models, with a complex, possibly non-linear function. Finally, in order to tackle the challenges arising on a mobile platform which might observe the environment from many, often non-informative viewpoints, an algorithm is provided performing spatio-temporal integration of evidence.

The thesis presents extensions of the models allowing for incremental learning and adaptation. A SVM-based incremental method is designed which performs like the batch algorithm while maintaining bounded complexity of the models, the last one being an important feature for real-time robotic systems. The approach is based on a combination of an approximate technique for incremental SVM \(^1^1^4^3\) with an exact method that reduces the number of support vectors needed to build the decision function without any loss in performance \(^6^1\). The algorithm is applied in two scenarios: adaptation in presence of dynamic changes and transfer of knowledge between autonomous agents. In the first scenario, the resulting system is able to maintain performance of the models despite dynamic changes. In the second scenario, we consider the case when a robot, proficient in solving the place recognition task within a known environment, transfers its visual knowledge to another robotic platform with different characteristics. In this case, the incremental algorithm allows the receiver of the information to gradually adapt the transferred representation to its own sensing.
5. AN APPROACH TO SEMANTIC MAPPING

Aiming toward the goal of building a complete semantic mapping system, this work first analyzes the problem of representing the whole body of spatial knowledge. As a result, a structure of a layered spatial knowledge representation is proposed which takes into account assumptions and requirements imposed by the considered scenario and possible interactions between the representation and other components of a robotic system.

The structure of the representation is shown in Figure 3. It consists of four layers corresponding to different levels of abstraction, from low-level sensory input to high-level conceptual symbols. The lowest level of the representation is the sensory layer which maintains an accurate representation of the robot’s immediate environment. Above this are the place and categorical layers. The place layer discretizes continuous space into a finite number of places, plus paths between them. As a result, the place layer represents the topology of the environment. The categorical layer contains categorical models of the robot’s sensory information such as object models or place classification models. On top of this, the conceptual layer creates a unified representation relating sensed instance knowledge to general conceptual knowledge.

The conceptual knowledge constitutes a crucial part of the representation. It includes taxonomy of human-compatible spatial concepts which are linked to the sensed instances of these concepts drawn from lower layers. It is the conceptual layer which contains the information that kitchens commonly contain cereal boxes and have certain general appearance and allows the robot to infer that the cornflakes box in front of the robot makes it more likely that the current room is a kitchen. The conceptual layer is described in terms of a probabilistic ontology defining spatial concepts and linking those concepts to instances of spatial entities (see Figure 3). Based on this design, a probabilistic graphical chain graph model is proposed as a representation for performing inferences on the knowledge represented in the conceptual layer. This results in an efficient approach to probabilistic modeling and reasoning about conceptual knowledge.

Based on the principles included into the design of the representation, a complete semantic mapping system is built which maintains it. An overview of the components of the system is presented in Figure 4. The system incorporates the conceptual reasoner and the place categorization sub-system as well as components building representations of other aspects of spatial knowledge such as a SLAM algorithm and object/landmark recognizers. It performs segmentation of space into rooms based on detected doorways and narrow openings. Moreover, the system implements a hierarchical structure decoupling the categorical models of sensory information from the conceptual reasoning by introducing an intermediate level of the so called properties of space. Those properties can represent the general appearance of a room, its geometrical attributes such as shape or size or object
Figure 3: The layered structure of the spatial knowledge representation. The position of each layer within the representation corresponds to the level of abstraction of the spatial knowledge. The conceptual layer illustrates part of the ontology representing both instance and predefined world knowledge.
presence. The universal character of the properties permits integration of semantic information obtained from multiple sources such as topology, general appearance and geometry, object information, human input, and potentially, human actions.

6 Implementation and Integration with a Robotic System

The semantic mapping system was implemented in a cognitive robotic software architecture (CAST [70]). This facilitates integration with other components of a robotic system and permits analysis of performance and usefulness of the semantic mapping system on real tasks. This thesis presents a system in which the semantic mapping is used together with active exploration and view planning components as well as a switching planner. The planner automatically switches between using decision-theoretic and classical AI planning procedures in order to create a system capable of autonomous active visual search for objects in a large-scale environment. In order to show the importance of semantic information for solving complex tasks, the performance of the system employing the semantic mapping component is compared to a simplified version which does not have access to semantic information.

7 Evaluation Data, Procedures and Results

In order to evaluate and analyze various properties of the proposed solutions thoroughly and in realistic settings, several datasets were collected. The datasets were designed to capture the input that semantic mapping and place classification systems would receive when running on a mobile robot platform. The datasets were collected in multiple office and home environments. Based on the datasets, several benchmarks were proposed and released to the robotics and computer vision com-
munities. Those benchmarks served as a basis for evaluation of performance and analysis of properties of the proposed methods.

The first benchmark was proposed based on two different databases: the INDECS (INDoor Environment under Changing conditionS) database and the IDOL (Image Database for rObot Localization) database. In case of INDECS, images of an office environment were captured from a fixed set of points using a standard camera mounted on a tripod. The resolution of the images is high; this makes this database suitable for context-based object recognition. The IDOL database, instead, consists of image sequences recorded using two mobile robot platforms equipped with perspective cameras, and thus is well suited for experiments with robot localization. The databases represent a different approach to the problem and can be used to analyze different properties of a place recognition system. The acquisition was performed under several different illumination settings and over a significant span of time. Six months after the acquisition of the IDOL database, an extension referred to as IDOL2 was acquired. Together, IDOL and IDOL2 capture significant long-term variations that occur in indoor environments and were used for evaluating the adaptive place classification models.

In order to evaluate the systems in larger environments and permit experiments with categorization, another database, COsy Localization Database (COLD), was acquired in three different office environments across Europe. In each environment, the acquisition was performed in several rooms of different functionality and short-term dynamic changes caused by illumination were captured. Unfortunately, the images were taken with low quality cameras. In order to increase the number of categories and the image quality, the database was extended with a large dataset COLD-Stockholm. The new dataset captures appearance and geometry of almost 50 rooms belonging to different semantic categories. This dataset was used during the offline categorization experiments and to train the appearance and geometry models of the semantic mapping system. Besides those databases, smaller datasets were created for the purpose of specific experiments in both home and office environments.
Chapter 3

Related Work

This section provides an overview of the related work in the area of place classification and semantic mapping. Place classification is a vastly researched topic in the computer vision and robotics community, usually considered as an independent problem and employed in a variety of applications. In computer vision the problem is often referred to as scene classification. Despite the fact that, in this work, place classification is ultimately used as an intermediate step towards semantic mapping, the proposed models also have much wider potential applications, often experimentally demonstrated. Therefore, the work on place classification will first be analyzed followed by the more general area of semantic mapping.

1 Place Classification

As previously mentioned, place classification can be divided into place recognition and place categorization and several of the proposed approaches were used for both problems. However, many of them, particularly in robotics, were focused on place recognition and its typical application - topological localization. Table 1 compares some of the approaches discussed below and maps them to keywords representing properties of place classification algorithms.

Place Recognition

Even in the early days, due to its richness, vision was considered a solution for the problem of place recognition. Already back in 1994, Kortenkamp & Weymouth [72] proposed an approach to topological localization using vision as one of the sensors and the concept of vision-based maps has been explored much earlier [49, 50]. Still, some of the later approaches relied only on geometrical cues and laser range data. Brunskill et al. [52] used a method based on simple geometrical features previously proposed for place categorization [89] in the context of topological localization. In this work, place recognition models were used to select one of the submaps which were earlier identified by decomposing a map into separate segments using
CHAPTER 3. RELATED WORK

<table>
<thead>
<tr>
<th></th>
<th>Place recognition</th>
<th>Place categorization</th>
<th>Indoor environment</th>
<th>Non-omnidirectional sensor</th>
<th>Appearance, single-cue</th>
<th>Appearance, multi-cue</th>
<th>Geometry</th>
<th>Objects</th>
<th>Confidence estimation</th>
<th>Novelty detection</th>
<th>Spatio-temporal integration</th>
<th>Adaptive/Incremental</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[52]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[122]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[46]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[92]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[123], [124]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[114]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[73]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[98]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[56], [57]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[88]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[116]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[109]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[63]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[125]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[106]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[121]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[130]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[62]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[104]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[119]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[133]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[103]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[53]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[89]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[90]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[132]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[129]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[105]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: Properties of the discussed place classification approaches. The first part of the table lists place recognition methods. The second part focuses on approaches applied to place categorization. Finally, the properties of the proposed approach are listed for comparison. (*) Objects are introduced to the semantic mapping system and integrated with place categorization models on the conceptual level.
spectral clustering. The model proposed in this thesis, was also evaluated for place recognition based on similar features.

The early adopters of vision for topological localization in robotics relied mainly on omnidirectional sensors. Ulrich & Nourbakhsh [122] proposed an appearance-based method which relied on color histograms extracted from omnidirectional images and a nearest neighbor image retrieval system. In constrast, in [46], Artač et al. implemented an incremental eigenspace model for representing the panoramic images captured at different locations in order to allow for incremental learning and adaptation without the need to store all the input data. Later approaches employing omni-directional sensing focused on scalability and large-scale environments; however, also preferring outdoor settings. Murillo & Košeká [92] presented an algorithm using a global descriptor computed for portions of panoramic images and a similarity measure for image matching. The method was tested on a large scale outdoor Street View dataset. Finally, in [123, 124] an incremental spectral clustering algorithm was applied to segment continuous space into topological nodes and local feature matching was used for localization. These clusters are defined by appearance and the aim is to support localization rather than human robot interaction. The clusters therefore have no obvious semantic meaning. The work focused on robustness to seasonal changes in mixed large-scale indoor/outdoor environment.

Many solutions relied on perspective vision being a popular and easily available sensor. Košeká et al. [74] proposed models of places built by segmenting temporally adjacent views based on a global appearance-based similarity measure and using the resulting segments for qualitative topological localization. In later work [73], local scale-invariant keypoints were used instead and spatial relationships between locations were modeled using Hidden Markov Models (HMM). In [98], the experimental setup presented in this thesis was used to evaluate place recognition models built using online learning extension of Support Vector Machines (SVM) in order to adapt to long-term appearance variations. As in case of the methods using omnidirectional vision, several recent works focused on scalability in large outdoor environments. Cummins & Newman [56, 57] proposed a probabilistic appearance-based framework for SLAM evaluated on paths up to 1000km length. At the same time, Milford & Wyeth [86] mapped a suburb with a SLAM system inspired by computational models of the rodent hippocampus.

Most of the above mentioned approaches only one modality was used for the recognition of places. However, several authors observed that robustness and efficiency of the recognition system can be improved by combining information provided by vision and geometrical sensors. Kortenkamp & Weymouth [72] combined vision with sonar sensing for topological localization. Also, Tapus & Siegwart [116] combined omnidirectional vision with features extracted from laser range data to build rotationally invariant descriptors, called fingerprints of places, identifying the topological locations. Those were then used for incremental topological mapping. In a similar spirit, several authors integrated multiple types of visual features in order to increase performance. In [109], the use of global and local visual features was motivated by the studies of human visual capabilities and a biologically-inspired
vision system was built for computing the gist of a scene and salient local regions. Those, in turn, were integrated into a Monte-Carlo localization system evaluated in an outdoor environment. At the same time, Filliat [63] proposed an algorithm for global localization in an indoor environment based on bag-of-words representation of scale-invariant local key-points and texture and color information which is able to incrementally learn the appearance of the environment based on interaction with a human.

A different take to the place recognition problem was offered by Vasudevan et al. [125] and Ranganathan & Dellaert [106]. In both cases objects detected in the environment were used as cues for place recognition. In [106], a constellation object model is extended to 3D and built in a coordinate frame local to the place. The observed constellation was then matched to the place models for place recognition. In [125], hierarchical probabilistic representation of space is proposed that is composed of places which are connected to each other through doors and are represented by local probabilistic object graphs. In contrast to [106], each object was first independently detected and used to update the hypothesis about the current location. In both cases, the object models were learned in a supervised manner.

Place Categorization

The problem of place categorization based on visual information was first addressed in the computer vision community. In this case, the research focused mainly on the problem of classifying single images captured in indoor or outdoor environments (scene classification). At the same time, robotics researchers initially employed the 2D laser range sensor being much more robust to variations occurring in the environment and much easier to handle computationally in real time.

In computer vision one of the first works to address the problem of place categorization was by Torralba et al. [121, 120] which employed an image representation called the gist of the scene [97], which is a vector of principal components of outputs of a bank of spatially organized filters applied to the image. The approach was tested in the context of both recognition and categorization, both indoors and outdoors, and used a HMM to fuse information over time and space. One of the key insights in that work is that the context is very important for recognition and categorization of both places and objects and that these processes are intimately connected. In [130], the problem of grouping images into semantic categories is addressed. It is pointed out that many natural scenes are ambiguous and the performance of the system is often quite subjective. They argue that typicality is a key measure to use in achieving meaningful categorizations. Each cue used in the categorization should be assigned a typicality measure to express the uncertainty in the categorization. The system is evaluated in natural outdoor scenes. In [62] another method is presented for categorization of outdoors scenes based on the distribution of codewords in each scene category obtained by clustering local interest point descriptors. A similar approach was used by Quelhas et al. [104] which also relies on the bag-of-words representation and studies analogies between scene classification
1. PLACE CLASSIFICATION

Based on visual words and text documents classification. In [62] a Bayesian hierarchical model was employed, while [104] used SVM for performing classification. Lazebnik et al. [79] extends the bag-of-words paradigm by introducing a spatial pyramid encoding approximate global geometric correspondence between local features. The approach is evaluated on the Caltech-101 database and the work reports increased performance compared to the orderless approach. In [133] a new global image descriptor, PACT, is presented and shown to give superior results on the datasets used in [121, 62] when combined with an SVM classifier. Finally, Quattoni & Torralba [103], extend the previous work in [121] by combining the global gist descriptor with local features. The method is evaluated on a large database of 67 indoor scene categories.

In robotics, the early systems for place categorization relied on omnidirectional laser range data for extracting simple semantic descriptions. In their work, Buschka & Saffiotti [53] partitioned grid maps of indoor environments into two different classes of open spaces, i.e. rooms and corridors. The division of the open spaces was done incrementally on local submaps. Mozos et al. [89] applied boosting to create a classifier based on a set of geometrical features extracted from range data to classify different places in indoor environments into rooms, corridors and doorways. A similar idea was used in [119] to describe regions from laser readings. In [90], the work by Mozos et al. was extended to also incorporate visual information in the form of object detections. Furthermore, this work also added a HMM on top of the point-wise classifications to incorporate information about the connectivity of space and make use of information such as offices are typically connected to corridors. Viswanathan et al. [129] adopted a purely object-based approach and performed automated learning of object-place relations and visual object models from the online LabelMe database. In [132] the work from [133] is extended with a new image descriptor, CENTRIS, and a focus on visual place categorization in indoor environment for robotics. A Bayesian filtering scheme is added on top of the frame based categorization to increase robustness and give a more smooth category estimate. Recently, Ranganathan [105] addressed the problem of place categorization in a different and novel way. The problem was cast in a fully probabilistic framework which operates on sequences of images rather than individual images. The method uses change point detection to detect abrupt changes in the statistical properties of the data. A Rao-Blackwellized particle filter implementation is presented for the Bayesian change point detection. All information deemed to belong to the same segment can then be used to estimate the category for that segment using a bag-of-words technique.

Properties

Table 1 compares the place classification approaches in terms of their key properties. The first important difference is the problem to which the approach was applied i.e. recognition or categorization. Despite that several of the methods are capable of performing both, many of the place recognition approaches are specifically designed
CHAPTER 3. RELATED WORK

for topological localization and utilize a set of techniques and heuristics useful only in this scenario. In particular, methods focusing on large-scale datasets such as [86, 56] specialize towards localization in order to achieve high efficiency and scalability. Compared to those, the model presented in this work is much less scalable when applied to the topological mapping problem; however, it can be directly applied to place categorization and learn human spatial concepts. Another important scenario-related distinction results from the type of the environment for which the problem is designed. It is not obvious that a method performing well in an outdoor environment will perform equally well indoors [103]. The model presented here was evaluated indoors according to the primary scenario outlined in the previous chapter.

The approaches differ mostly with respect to the way the environment is perceived, and thus the sensory modalities employed and the method used to extract characteristic features of the scene. Purely geometric solutions based mostly on laser range data have proven to be successful for certain tasks [53, 52, 89]. Yet, the inability to capture many aspects of complex realistic environments leads to the problem of perceptual aliasing [77] and greatly limits the usefulness of purely geometrical methods. This inspired many researchers to turn towards vision which nowadays is tractable in real-time applications. The available methods employed either perspective or omnidirectional cameras. One of the requirements in this work was to use non-omnidirectional sensors which are commonly used on service robots and require being robust to partial observations and occlusions which will occur if the robot is deployed among humans.

Different types of cues were used to represent visual information. Landmark-based techniques make use of either artificial or natural landmarks in order to extract information about a place. In [83, 113] information signs are used as a source of spatial information. [125, 106, 90, 129] rely on objects detected in the environment. In those cases object models are trained beforehand in a supervised fashion. Visually distinctive image regions were also used as landmarks [109]. Other solutions employed mainly local image features such as SIFT [81, 45, 73, 62, 98, 104, 105], SURF [47, 91, 63, 123, 124, 56, 57], also using the bag-of-words approach [62, 63, 64, 104, 56, 57, 105], or other representation based on information extracted from local patches [115, 62, 79, 64]. Global features are also commonly used for place recognition. Torralba et al. [121, 120, 103] used the gist of the scene. Similar approach has been adopted by others [109, 92]. Other approaches use color histograms [122], gradient orientation histograms [74], eigenspace representation of images [46], Composed Receptive Field Histograms (CRFH) [80, 98], representations obtained using the Census Transform (CT) [133, 132] or a scanline intensity profile [86].

Several works combined vision with geometrical sensors [110, 90]. Others, used a combination of global and local visual features to increase performance and robustness [109, 63, 103]. The place classification approach presented in this thesis seems to be unique in that it integrates multiple visual cues with geometrical information extracted from laser range data, only when it is likely to increase performance,
which thus also improves efficiency. The cue integration technique fuses cues on a
high-level after discriminative classification which has been shown to achieve bet-
ter performances than probabilistic approaches [95]. Moreover, object information
is also used in the final semantic mapping system and is integrated with place
categorization models on the conceptual level.

An important property of a place classification system is the ability to estimate
the confidence of its own decision. Therefore, many systems provide some practical
measure of confidence. In most cases, this is based on the similarity of a query
image to the training images and implemented in terms of nearest neighbor models
or other non-parametric approaches [122, 92, 74, 73, 63, 123] while other methods
use probabilistic models [56, 57, 105]. The advantage of generative probabilistic
models is that it is possible to estimate how novel the observation is i.e. how likely
it is that the observation does not belong to any of the place classes available during
training. In the context of place categorization, the only work that implements
that functionality is [105]. In order to provide good generalization, especially in
presence of the large within-category variability, in this thesis a discriminative SVM
classifier is used. SVMs do not provide an out-of-the-box solution for the confidence
estimation problem. Therefore, a practical method based on the distance to the
SVM hyperplane is designed and when applied yielding good results.

In many of the works, especially when non-omnidirectional sensors are used, the
authors observed that the ability to fuse observations over time and space is crucial
for robust operation. Several works applied techniques known from the metric
localization domain e.g. particle filters [109, 105] or other Bayesian filters [132].
Others employed graphical models such as HMM [121, 73, 90]. In this work, we
use a two step approach. First a technique performing evidence accumulation over
time and space is used for evidence gathering inside places. Then, in the semantic
mapping system, information is fused across places and combined with typical room
connectivity information by a chain graph [78], i.e. a probabilistic graphical model.

Finally, several methods applied to place recognition and topological mapping
build their representations in an incremental fashion and allow updating and adap-
tation of the place models [46, 123, 124, 98, 56, 57, 86, 116, 63, 125]. In case of
place categorization, this feature is not common and only [105] provides a way to
build the representation online. This work shows how the presented discrimina-
tive model can be extended to allow for incremental learning and adaptation to
long-term environment variations.

2 Semantic Mapping

The semantic mapping problem has only recently received significant attention and
several systems were proposed within the last 5 years. As shown above, there ex-
ists a broad literature on mobile robot localization, mapping, navigation and place
classification. Every such algorithm maintains a representation of spatial knowl-
dge. However, this representation is usually specific to the particular problem
and designed to be efficient within the single mapping system detached from any other interacting components. Other, more general concepts, such as the Spatial Semantic Hierarchy [76] concentrate on lower levels of spatial knowledge abstraction and do not support higher-level conceptualization or representation of categorical information.

One of the first systems that was able to build a representation from both spatial and semantic perspective was proposed by Galindo et al. [65]. In their system, two hierarchies are maintained, spatial and semantic which are interrelated through the concept of anchoring (see Figure 1). The spatial hierarchy contains simple sensory data like camera images or local grid maps as well as the topology of the environment. The conceptual hierarchy represents concepts and their relations modeled by employing standard AI languages. This permits the robot to do inferences about symbols e.g. infer the room category based on detected objects as well as the presence of typical objects based on room category. However, the representation does not contain the uncertainties about the instances. Objects are the only source of semantic information in the system and the semantic hierarchy is built manually.
2. SEMANTIC MAPPING

In order to identify discrete spatial entities, a grid map segmentation algorithm is used to detect open spaces. Finally, an AI planner is used together with the representation to actively resolve ambiguities in the room categorization.

Zender et al. [135] proposed a system that is similar in spirit but with some extensions. The authors design a representation composed of layers representing maps at different levels of abstraction: metric, navigation, topological, and conceptual divided into two groups: mapping (first 3 layers) and reasoning (last layer). In that sense, their approach is similar to [65] and the spatial and semantic hierarchies. The conceptual layer contains an innate conceptual ontology that defines categories for rooms and objects and how they are related. Also, the information extracted from sensors and given through situated dialogue is represented as instances of concepts. The conceptual knowledge is encoded in an OWL-DL ontology and a description-logic reasoner is used to infer new knowledge about the world that is neither given verbally nor actively perceived. As in case of [65], uncertainty is not represented at the conceptual level and the ontology is provided manually. What is new in this work is the inclusion of place classification models by Mozos et al. [89] for the purpose of distinguishing rooms from corridors. Additionally, door detection is used in order to segment space into rooms. The system is integrated in a mobile robot endowed with laser and vision sensors for place and object recognition. The system also incorporates a linguistic framework that supports the map acquisition process. Overview of the system presented in Figure 2.
CHAPTER 3. RELATED WORK

Vasudevan & Siegwart [126] focus again on purely object-based semantic mapping, but make their representation of the world fully probabilistic. The approach is based on a generative model of place categories based on a Naive Bayesian Classifier. These objects detected in the environment are grouped into spatial and semantic abstractions. The robot then uses the semantic groups as concepts and assigns them to places identified by spatial object groups. The concepts arise during the training process and are extracted from input training data. During testing, the detected objects are used to segment space. The approach is feed-forward i.e. the object information is used to classify places; however, knowledge about place categories is not used to infer presence of objects.

In [84], Meger et al. focus on the autonomous detection and perception of objects and augmenting spatial metric maps with object information. Their system is much more advanced, compared to the previously described, in terms of the vision subsystem. The system uses a peripheral-foveal vision and an attention system combining bottom-up visual saliency with structure from stereo. This is integrated with FastSLAM for localization and mapping. The object models are trained on image data collected by submitting text-based queries to internet image search engines. The system is capable of autonomous exploration and object search and was demonstrated during the Semantic Robot Vision Challenge [39]. The work by Viswanathan et al. [128] can be seen as an attempt to provide similar functionality as in [84] in an more robust and autonomous way. The paper presents a semantic mapping system able to annotate places with semantic labels based on the object information. For this purpose, a Bayesian model of place categories is built based on object occurrence frequencies for various semantic place categories learned from an online annotated database. Then, these models together with spatial information are used to cluster the space into discrete places. The resulting representation is used in to infer typical object locations and perform an informed search for objects.

Another approach for augmenting spatial maps with object-based semantic information was proposed by Nüchter & Hertzberg [96]. In contrast to all the previously described approaches, the objects were detected from a 3D representation of the world built using a 6D SLAM algorithm from laser range data. The system first analyzes the obtained point-cloud map and identifies coarse scene features such as walls or floors. Then objects are detected by a trained classifier and projected back on to the map. The resulting representation is meant to be visualized for human inspection.

A completely different approach was taken by Nieto-Granda et al. [94]. The aim of this work is to assign semantic labels obtained from human augmented mapping directly to the metric space. This is a different approach than that of Topp & Christensen [119] in case of which the space is segmented into regions. The semantic layer is a multivariate probability distribution on the coordinates of our metric map. This multivariate distribution is modeled as a Gaussian model and each of the Gaussians is based on the robot’s sensor data when it was provided a label by a human guide. The semantic information can then be expanded to cover the entire metric map.
Finally, there is a number of works devoted to semantic mapping of outdoor environments. Since, many of the approaches can also be relevant in case of indoor spaces, they are reviewed below for completeness. Posner et al. [102] presents a system for augmenting the representation of outdoor space with semantic labels. A supervised learning scheme is employed to train a set of classifiers to respond to common scene attributes given a mixture of geometric and visual scene information obtained using a 3D laser scanner and a camera. A set of SVM classifiers is used, each specialized to detect a certain type of semantic attribute like pavement, tarmac, bush. The SVM models are trained on hand-labeled data. A similar problem was approached by Douillard et al. [60] with a focus on objects. The authors tried to classify objects in urban environments based on laser and vision data and used the classification results to augment metric maps. The novelty in this case results from the applied technique. The system extracts visual features from color images and shape features from 2D laser scans. From those, a probabilistic model exploiting spatial and temporal dependencies is created based on Conditional Random Fields (CRF) which can be trained from partially labeled data. Finally, Persson et al. [100] describe a method for automatic classification of outdoor scenes captured with omnidirectional vision into two classes: nature or buildings. The classification is performed using AdaBoost and the results are used to annotate a grid map of the environment with the semantic information.

Table 2 compares the properties of the discussed semantic mapping approaches. Out of the above mentioned methods designed for semantic mapping of indoor environments, none uses topology of the environment as a source of semantic information. Furthermore, those only two that use general appearance of places as semantic information, only do so for outdoor settings. This is surprising given the large body of work on appearance-based place categorization. Two methods, [135] and [96] make use of geometric place information extracted from laser range sensors, and only [135] applies a previously developed place classification technique for this purpose. In [135], semantic cues can be obtained by a situated dialogue with a user and [94] build maps augmented with semantic symbols purely from human input. Almost every method is focused primarily on using objects for extracting spatial semantics [65, 135, 126, 84, 128, 96]. Objects clearly carry a lot of semantic information; however, they are also sparse and reliable object categorization in real-world environments is still a major open challenge. At the same time, valuable semantic cues are also encoded in geometry, general appearance and topology and robust methods for extracting that information have been proposed, including the approach presented in this thesis. The inability to fuse together all the sources of information is likely a result of the different character of the different inputs. In this work, we present a system able to combine all the aforementioned sources of semantic information: general appearance and geometry of places, object information, topological structure and human input. This is made possible by creating a hierarchical, property-based system in which all sources of information contribute to various properties of space which are then fused seamlessly on the conceptual level.
CHAPTER 3. RELATED WORK

Table 2: Properties of the discussed semantic mapping systems compared to the proposed approach.

<table>
<thead>
<tr>
<th></th>
<th>Indoor environment</th>
<th>Place appearance cues</th>
<th>Place geometry cues</th>
<th>Object information</th>
<th>Topology information</th>
<th>Human input</th>
<th>Segmentation</th>
<th>Conceptual map / Ontology</th>
<th>Uncertain concepts</th>
<th>Inferring properties</th>
<th>Active component</th>
<th>Concepts built automatically</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>135</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>126</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>84</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>128</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>96</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>94</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>102</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>60</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>100</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

The conceptual map in our system is also a unique feature. The most comprehensive relevant representations has been proposed in [65] and [135]. Both approaches encode an ontology of an indoor environment. However, those ontologies are built manually and use traditional AI reasoning techniques which are unable to incorporate uncertainty that is inherently connected with semantic information obtained through robot sensors in realistic environments. In contrast, we implement a probabilistic ontology and a probabilistic inference engine incorporating uncertainty in definitions of concepts and their links to instances of spatial entities. Moreover, the values of all properties for which direct evidence is not available can be inferred based on all the available semantic information. Additionally, as in case of [126] and [128] the concept definitions are built automatically from online databases and floor plans obtained from robotics datasets. Finally, the semantic mapping is combined with AI planning components resulting in a system able to actively search for objects in a similar fashion to [128].
Chapter 4

Summary of the Papers

In this section the included papers are summarized and briefly discussed. First, an outline of each paper is presented followed by an overview of the contributions of the author of the thesis. Additionally, Table 4 groups the included papers as well as the other papers co-authored by the author according to keywords relevant to the problem considered in this work. The relation of each paper to other works, including works by the author of the thesis, as well as the impact on the respective fields are discussed inside the papers.

1 Paper A: Single-cue Place Recognition

1.1 Outline of the Paper

This paper presents two carefully designed and annotated image databases augmented with an experimental procedure and extensive baseline evaluation. The databases were gathered in an uncontrolled indoor office environment using two mobile robots and a standard camera. The acquisition spanned across a time range of several months and different illumination and weather conditions. Thus, the databases are very well suited for evaluating the robustness of algorithms with respect to a broad range of variations, often occurring in real-world settings. We thoroughly assessed the databases with a purely appearance-based place recognition method based on Support Vector Machines and two types of rich visual features (global and local).

1.2 Contribution by the Author

Acquired the databases used for evaluating the visual place recognition algorithms in the paper. Designed a benchmark for visual place recognition. Built a visual place recognition system based on global and local visual features. Performed the evaluation of the system on two different databases.
### Table 1: The papers co-authored by the author of the thesis grouped according to keywords relevant to the problem considered in this work.

<table>
<thead>
<tr>
<th></th>
<th>Databases and benchmarks</th>
<th>Place recognition</th>
<th>Knowledge transfer</th>
<th>Incremental learning</th>
<th>Confidence estimation</th>
<th>Cue integration</th>
<th>Sensor fusion</th>
<th>Place categorization</th>
<th>Knowledge representation</th>
<th>Semantic mapping</th>
<th>Conceptual map</th>
<th>Integrated cognitive systems</th>
<th>Active semantic perception</th>
</tr>
</thead>
<tbody>
<tr>
<td>[A]</td>
<td>✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[B]</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[C]</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[D]</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[E]</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[F]</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[G]</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The papers co-authored by the author of the thesis grouped according to keywords relevant to the problem considered in this work.
2 Paper B: Incremental Learning and Knowledge Transfer

2.1 Outline of the Paper

This paper presents an SVM-based algorithm, capable of learning representations incrementally while maintaining memory requirements. We combine an incremental extension of SVMs with a method reducing the number of support vectors needed to build the decision function without any loss in performance introducing a parameter which permits a user-set trade-off between performance and memory. The resulting algorithm is able to achieve the same recognition results as the original incremental method while reducing the memory growth. Our method is especially suited for autonomous systems in realistic settings. We present experiments on two common scenarios in this domain: adaptation in presence of dynamic changes and transfer of knowledge between two different autonomous agents, focusing in both cases on the problem of visual place recognition applied to mobile robot topological localization.

2.2 Contribution by the Author

Acquired the database used for experiments in the paper. Designed and implemented the memory-controlled incremental SVM algorithm. Performed a part of the evaluation of the algorithm on the place classification databases. Helped with the design and implementation of the knowledge transfer algorithm.

3 Paper C: Confidence Estimation and Cue Integration

3.1 Outline of the Paper

This paper presents a recognition algorithm able to measure its own level of confidence and, in case of uncertainty, to seek for extra information so to increase its own knowledge and ultimately achieve better performance. We focus on the visual place recognition problem for topological localization, and we take an SVM approach. We propose a new method for measuring the confidence level of the classification output, based on the distance of a test image to the average distance of training vectors. This method is combined with a discriminative accumulation scheme for cue integration. We show with extensive experiments that the resulting algorithm achieves better performances for two visual cues than the classic single cue SVM on the same task, while minimising the computational load. More important, our method provides a reliable measure of the level of confidence of the decision.

3.2 Contribution by the Author

Researched several approaches to confidence information extraction for Support Vector Machines. Combined the confidence estimation approaches with discrimi-
native cue integration and performed an evaluation of the resulting algorithm in the context of visual place recognition.

4 Paper D: Multi-modal Place Classification for Semantic Mapping

4.1 Outline of the Paper

In this paper we present a multi-modal place classification system that allows a mobile robot to identify places and recognize semantic categories in an indoor environment. The system effectively utilizes information from different robotic sensors by fusing multiple visual cues with laser range data. This is achieved using a high-level cue integration scheme based on a Support Vector Machine that learns how to optimally combine and weight each cue. Our multi-modal place classification approach can be used to obtain a real-time semantic space labeling system which integrates information over time and space. We perform an extensive experimental evaluation of the method for two different platforms and environments, on a realistic off-line database and in a live experiment on an autonomous robot.

4.2 Contribution by the Author

Designed and implemented the multi-modal place classification system. Researched various cue integration techniques and proposed a modified discriminative cue accumulation scheme. Performed an extensive experimental evaluation of the place classification system. Built a semantic mapping system based on the place classification models. Finally, evaluated the system in real-time on a mobile robot platform.

5 Paper E: Spatial Knowledge Representation

5.1 Outline of the Paper

In this paper, we carefully analyze the problem and design a spatial knowledge representation for a cognitive mobile system. Our representation is layered and represents knowledge at different levels of abstraction. It deals with complex, cross-modal, spatial knowledge that is inherently uncertain and dynamic. Furthermore, it incorporates discrete symbols that facilitate communication with the user and components of a cognitive system. We present the structure of the representation and propose concrete instantiations.

5.2 Contribution by the Author

Specified the principles behind the spatial knowledge representation. Designed the general theoretical structure of the representation.

6.1 Outline of the Paper

In this paper, we present a multi-layered semantic mapping algorithm able to combine information about the existence of objects in the environment with knowledge about the topology and semantic properties of space such as room size, shape and general appearance. We use it to infer semantic categories of rooms and predict existence of objects and values of other spatial properties. We perform extensive experiments offline and online on a mobile robot demonstrating the efficiency and usefulness of our system.

6.2 Contribution by the Author

Acquired the COLD-Stockholm database being used for the experimental evaluation of the approach. Implemented and tested the categorical models of sensory information providing basis for the spatial properties. Designed the spatial knowledge representation, the ontology behind the conceptual map and its chain-graph inference model. Designed and implemented the property-based semantic mapping system. Finally, performed online experimental evaluation of the system on a mobile robot.


7.1 Outline of the Paper

In this work we present a robot system that combines common-sense knowledge about the structure of the world with probabilistic modeling of the uncertainty and demonstrate improvements in efficiency and reliability. Our first contribution is a probabilistic relational model integrating common-sense knowledge about the world in general, with observations of a particular environment. Our second contribution is a switching planning system which is able to plan on the large problems posed by that model, by automatically switching between decision-theoretic and classical procedures. We evaluate our system on object search tasks in two different real-world indoor environments. By reasoning about the trade-offs between possible courses of action with different informational effects, and exploiting the cues and general structures of those environments, our robot is able to consistently demonstrate efficient and reliable goal-directed behavior.
7.2 Contribution by the Author

Designed and implemented the conceptual map and the semantic mapping algorithm. Integrated the semantic mapping algorithm with other components of the cognitive system. Finally, performed online experiments in the office environment evaluating properties of the integrated system.
Chapter 5

Discussion and Conclusions

This thesis explored the problem of enabling mobile robots with the ability to understand human environments. Several methods have been proposed for extraction of semantic information from robotic sensors, modeling spatial concepts and finally building semantic maps. The methods were experimentally evaluated on realistic offline datasets as well as in real-time on mobile robot platforms. Those evaluations showed that semantic spatial understanding is within our grasp and is getting ready to be deployed outside research environments. Moreover, several important scientific questions have been posed and addressed in the course of this work.

Firstly, it was shown that useful models of place instances and place categories can be constructed from the general appearance of the environment as well as its geometry measured through laser range sensors. Those models can be made robust to most typical variations that occur in indoor environments such as illumination changes and variations caused by human intervention. In case of place recognition, the models were shown to perform topological localization with high precision, although in relatively small environments compared to the more recent techniques developed for large-scale outdoor spaces \cite{56,57,86}. In the context of place categorization, it was shown that assuming a certain level of within-category variability that occurs within a single multi-storey building, the methods can be robust and provide important spatial semantic concepts.

Secondly, confidence measures for the place classification models have been proposed and thoroughly evaluated. It was shown that confidence measures have important practical value for increasing robustness of the system as well as its efficiency in case of multi-cue models. Indeed, in many real-world applications it is more desirable to refrain from action because of a self-recognized lack of confidence, rather than take a hard decision which might result in a costly error. When combined with a cue integration scheme, confidence estimation can be used to decide about acquisition and processing of additional cues only if it is required to improve the confidence of the system. This results in an improved efficiency without compromising the overall performance.

Furthermore, this thesis studied the problem of cue integration and sensor fusion
and proposed a method for cue accumulation from multiple sensors applicable to the place classification model. Through extensive experiments, it was shown that robustness of the system can be increased if multiple cues extracted from the same modality (in this case global and local visual features) are integrated. Also, larger performance gain can be obtained if the cues come from different sensors having different characteristics. In this work, the most robust solution was obtained when rich visual cues were combined with illumination invariant geometrical features extracted from laser range data.

As another way of solving the long-term dynamic variations problem, this thesis advocated the use of incremental and adaptive systems. An incremental extension to the SVM discriminative classifier was proposed and applied to the place recognition problem. The method was shown to achieve recognition performances statistically equivalent to those of the batch algorithm, while obtaining a substantial memory reduction. Moreover, in case a limit is set on the size of the model, the method tends to forget the oldest information making it suitable for adaptation to changing conditions. It was experimentally validated that an adaptive place recognition model can greatly improve its performance by tracking the dynamic changes. The algorithm was also applied to the problem of knowledge transfer between two robotic platforms. In this case, the incremental learning algorithm allowed the receiver of the information to gradually adapt the model to its own sensing.

In order to perform experimental evaluations of the proposed solutions in controlled, yet realistic, settings, several databases were collected including: INDECS, IDOL, IDOL2, COLD and its recent extension COLD-Stockholm. The offline evaluations were then compared to online experiments. The robot achieved comparable performance to that obtained offline. This suggests that the proposed databases and benchmarks based on them are indeed realistic. At the same time, using databases permitted thorough analysis of properties of the methods and their fair comparison.

This thesis expressed a belief that objects play an important role in understanding of space, as does spatial topology. However, as shown by the review of related works, no principled method previously existed for fusing different sources of semantic information such as objects, general appearance and geometry into one comprehensive representation. The property-based paradigm proposed in this work provided a seamless way of integrating objects with place categorization. Moreover, it was shown that the topology itself can be a strong cue for room categorization, especially in case of such rooms as a corridor which is likely to be connecting other rooms. Combination of all sources of knowledge inside the conceptual inference framework resulted in a reliable place categorization technique.

Another advantage with the property based system is that it permitted training of the concept definitions independently from the models of sensory information. As a result, it became possible to train the system with data from common sense knowledge databases or crawling the internet for information about typical topologies and objects-room relations. The experiments showed that those can be valuable source of conceptual information, and through the abstraction provided by place classification and object models, useful in practice in realistic environments.
One of the important characteristics of the presented method is the ability to represent the conceptual knowledge in a probabilistic framework. This turned out to be particularly important in integrated systems. When the semantic mapping system was integrated with a planning and execution monitoring component, the uncertainty presented to the planner allowed for much more efficient behavior. For example, the planning component could trade the room exploration cost with the likelihood of finding a particular object in a specific room. Finally, the importance of semantic knowledge for behavior planning was shown by the experiments with the active visual object search system. The system was run with and without the possibility to use the results of semantic mapping and the time required for finding the object was measured. It became clear that the search behavior becomes much more efficient if the objects are searched in their canonical positions inferred by the semantic mapping system.

As final words, it is important to say that despite this thesis being concerned with the use of semantic mapping system on mobile robot platforms, there are multiple other applications which could benefit from the availability of semantic information. Those include wearable devices in contexts such as assistance of elderly and disabled people. Such devices could provide information about the presence of objects, the typical actions that should be performed and could monitor the behavior of a person by comparing it to the typical behavior. Moreover, in the era of ubiquitous mobile devices equipped with substantial computational units and multiple sensors, we can think about the presented system running in our pockets and extending our experience of localization services or social networks. Surely, the future will bring many new exciting scenarios and applications for artificial intelligent systems understanding and exploiting spatial semantics.

**Future Work**

The presented work can be extended in many directions and several possible directions for future research are outlined briefly below.

**3D sensing** Recently, cheap RGB-D sensors became broadly available providing depth information fused with the visual input. This inspired many researchers to introduce 3D information into their approaches. The place classification technique presented in this work could benefit from introducing depth information and integrating it with the appearance models. Moreover, the geometrical information, so far provided by expensive laser range sensors could instead be computed from RGB-D sensing.

**Online learning of place models** An incremental adaptive model of places was presented in this work. However, optimally, the models should be updated not in batches, but online and in real time. At the same time, the complexity of the models must be controlled. The future work will address this issue.
Novelty detection and learning of novel concepts The probabilistic generative model of the conceptual information opens new possibilities in terms of automatic detection and learning of novel concepts. In future, the approach should be extended with the ability to identify gaps in spatial and semantic knowledge and perform learning of new concepts.

Using properties for space segmentation Currently, the doors and narrow openings detected in the environment are used as the only cue for segmentation of space into rooms. However, the conceptual map already assigns spatial properties to distinct places in the environment identified in an unsupervised fashion. This semantic information associated with places should be used for more informative and robust room segmentation and detected doorways should be fused with other spatial properties.

Life-long learning and autonomy Finally, the future work will investigate the use of the system on a mobile platform operating uninterruptedly over long periods of time. This will create opportunities for the system to update its representation gradually to changing conditions in an unsupervised or semi-supervised fashion. This should provide a setting to study many problems currently addressed by extensive offline training or generalization abilities of the learning algorithms. Moreover, the concept definitions currently generated based on common sense knowledge databases could be used only for bootstrapping and the robot could update or extend those definitions based on its own experience. Lastly, many properties of space related to its functionality only become apparent after long-term observations. A robot operating with humans in an indoor environment could learn to link actions to room categories and objects. Life-long learning is a complex problem which together with opportunities brings many challenges. Identifying and tackling those challenges is one of the most exciting directions for the future work.
Bibliography


[16] Jeremy L. Wyatt, Alper Aydemir, Michael Brenner, Marc Hanheide, Nick Hawes, Patric Jensfelt, Matej Kristan, Geert-Jan M. Kruijff, Pierre Lison,


44

BIBLIOGRAPHY


[34] OpenCyc. URL http://opencyc.org/


[38] Samsung VC-RL87W. URL http://www.samsung.com/


[40] TWENDY-ONE. URL http://www.twendyone.com/


[72] David M. Kortenkamp and Terry Weymouth. Topological mapping for mobile robots using a combination of sonar and vision sensing. In *Proceedings of


