KTH Computer Science
and Communication

Semantic Mapping with Mobile Robots

ANDRZEJ PRONOBIS

Doctoral Thesis in Computer Vision and Robotics
Stockholm, Sweden 2011

TRITA-CSC-A 2011:10
ISSN-1653-5723
ISRN-KTH/CSC/A--11/10--SE
ISBN 978-91-7501-039-7

Computer Vision and Active Perception
School of Computer Science and Communication
KTH Royal Institute of Technology
SE-100 44 Stockholm, Sweden

Tryck: AJ E-print AB

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 Stano 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:

[A] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I. Christensen. A realistic benchmark for visual indoor place recognition. *Robotics and Autonomous Systems (RAS)*, 58(1):81–96, January 2010.

[B] Andrzej Pronobis, Jie Luo, and Barbara Caputo. The more you learn, the less you store: Memory-controlled incremental SVM for visual place recognition. *Image and Vision Computing (IMAVIS), Special Issue on Online Pattern Recognition and Machine Learning Techniques for Computer-Vision: Theory and Applications*, 28(7):1080–1097, July 2010.

[C] Andrzej Pronobis and Barbara Caputo. Confidence-based cue integration for visual place recognition. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, October 2007.

[D] Andrzej Pronobis, Oscar M. Mozos, Barbara Caputo, and Patric Jensfelt. Multi-modal semantic place classification. *The International Journal of Robotics Research (IJRR), Special Issue on Robotic Vision*, 29(2-3):298–320, February 2010.

[E] Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Adrian N. Bishop, and Patric Jensfelt. Representing spatial knowledge in mobile cognitive systems. In *Proceedings of the 11th International Conference on Intelligent Autonomous Systems (IAS-11)*, Ottawa, Canada, August 2010.

[F] Andrzej Pronobis and Patric Jensfelt. Understanding the real world: combining objects, appearance, geometry and topology for semantic mapping. Technical Report TRITA-CSC-CV 2011:1 CVAP319, Kungliga Tekniska Högskolan, CVAP/CAS, May 2011. The paper contains contributions described in the following submissions:

- Andrzej Pronobis and Patric Jensfelt. Understanding the real world: combining objects, appearance, geometry and topology for semantic mapping. In *ICRA 2011 Workshop on Semantic Perception, Mapping and Exploration*, Shanghai, China, May 2011.
- Andrzej Pronobis and Patric Jensfelt. Hierarchical multi-modal place categorization. In *European Conference on Mobile Robots (ECMR'11)*, Örebro, Sweden, September 2011. (Submitted).

[G] Marc Hanheide, Charles Gretton, Richard W. Dearden, Nick A. Hawes, Jeremy L. Wyatt, Andrzej Pronobis, Alper Aydemir, Moritz Göbelbecker, and Hendrik Zender. Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, Barcelona, Spain, July 2011.

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

[1] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I. Christensen. A discriminative approach to robust visual place recognition. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, October 2006.

[2] Jie Luo, Andrzej Pronobis, and Barbara Caputo. SVM-based transfer of visual knowledge across robotic platforms. In *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS'07)*, Bielefeld, Germany, March 2007.

[3] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. Incremental learning for place recognition in dynamic environments. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, October 2007.

[4] Andrzej Pronobis, Oscar M. Mozos, and Barbara Caputo. SVM-based discriminative accumulation scheme for place recognition. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, Pasadena, CA, USA, May 2008.

[5] Muhammad Muneeb Ullah, Andrzej Pronobis, Barbara Caputo, Jie Luo, Patric Jensfelt, and Henrik I. Christensen. Towards robust place recognition for robot localization. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, Pasadena, CA, USA, May 2008.

[6] Heydar Maboudi Afkham, Alireza Tavakoli Targhi, Jan-Olof Eklundh, and Andrzej Pronobis. Joint visual vocabulary for animal classification. In *Proceedings of the 2008 19th International Conference on Pattern Recognition (ICPR'08)*, Tampa, FL, USA, December 2008.

[7] Andrzej Pronobis and Barbara Caputo. COLD: COsy Localization Database. *The International Journal of Robotics Research (IJRR)*, 28 (5):588–594, May 2009.

[8] Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Adrian N. Bishop, and Patric Jensfelt. A framework for robust cognitive spatial mapping. In *Proceedings of the 14th International Conference on Advanced Robotics (ICAR'09)*, Munich, Germany, June 2009.

[9] Andrzej Pronobis, Li Xing, and Barbara Caputo. Overview of the CLEF 2009 robot vision track. In Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth J. F. Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika, editors, *Multilingual Information Access Evaluation Vol. II Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 110–119. Springer, 2010.

[10] Andrzej Pronobis, Henrik Christensen, and Barbara Caputo. Overview of the ImageCLEF@ICPR 2010 Robot Vision track. In Devrim Ünay, Zehra Çataltepe, and Selim Aksoy, editors, *Recognizing Patterns in Signals, Speech, Images and Videos*, volume 6388 of *Lecture Notes in Computer Science*, pages 171–179. Springer Berlin / Heidelberg, 2010.

[11] Andrzej Pronobis, Marco Fornoni, Henrik I. Christensen, and Barbara Caputo. The Robot Vision track at ImageCLEF 2010. In *Working Notes for the CLEF 2010 Workshop*, Padua, Italy, 2010. ISBN 978-88-904810-0-0.

[12] Andrzej Pronobis and Barbara Caputo. The robot vision task. In Henning Müller, Paul Clough, Thomas Deselaers, and Barbara Caputo, editors, *ImageCLEF*, volume 32 of *The Information Retrieval Series*, pages 185–198. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15181-1.

[13] Li Xing and Andrzej Pronobis. Multi-cue discriminative place recognition. In Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth J. F. Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika, editors, *Multilingual Information Access Evaluation Vol. II Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 315–323. Springer, 2010.

[14] Andrzej Pronobis, Patric Jensfelt, Kristoffer Sjöö, Hendrik Zender, Geert-Jan M. Kruijff, Oscar M. Mozos, and Wolfram Burgard. Semantic modelling of space. In Henrik I. Christensen, Geert-Jan M. Kruijff, and Jeremy L. Wyatt, editors, *Cognitive Systems*, volume 8 of *Cognitive Systems Monographs*, pages 165–221. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-11694-0.

[15] Kristoffer Sjöö, Hendrik Zender, Patric Jensfelt, Geert-Jan M. Kruijff, Andrzej Pronobis, Nick Hawes, and Michael Brenner. The explorer system. In Henrik I. Christensen, Geert-Jan M. Kruijff, and Jeremy L. Wyatt, editors, *Cognitive Systems*, volume 8 of *Cognitive Systems Monographs*, pages 395–421. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-11694-0.

[16] Jeremy L. Wyatt, Alper Aydemir, Michael Brenner, Marc Hanheide, Nick Hawes, Patric Jensfelt, Matej Kristan, Geert-Jan M. Kruijff, Pierre Lison, Andrzej Pronobis, Kristoffer Sjöö, Alen Vrečko, Hendrik Zender, Michael Zillich, and Danijel Skočaj. Self-understanding & self-extension:a systems and representational approach. *IEEE Transactions on Autonomous Mental Development (TAMD), Special Issue on Representations and Architectures for Cognitive Systems*, 2(4):282–303, December 2010.

[17] Alper Aydemir, Kristoffer Sjöö, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, May 2011.

[18] Kristoffer Sjöö, Andrzej Pronobis, and Patric Jensfelt. Functional topological relations for qualitative spatial representation. In *Proceedings of the 15th International Conference on Advanced Robotics (ICAR'11)*, Tallinn, Estonia, June 2011.

Contents

Contents **xi**

I Introduction **1**

1 Introduction **3**
Thesis Outline 7

2 Place Classification and Semantic Mapping **9**
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 **19**
1 Place Classification 19
2 Semantic Mapping 25

4 Summary of the Papers **31**
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
6 Paper F: Combining Conceptual Knowledge, Objects, Appearance, Geometry and Topology for Semantic Mapping 35
7 Paper G: Semantic Mapping for Efficient Behaviour of Integrated Robotic Systems 35

5 Discussion and Conclusions **37**

Future Work 39

Bibliography **41**

II Included Papers **53**

A A Realistic Benchmark for Visual Indoor Place Recognition **A1**
1 Introduction A3
2 Related work A5
3 Design strategy A6
4 Data Acquisition A8
5 Baseline Evaluation A14
6 Experimental Results A20
7 Summary A29
References A30

B The More you Learn, the Less you Store: Memory-controlled Incremental SVM for Visual Place Recognition **B1**
1 Introduction B3
2 Related Work B6
3 Visual Place Recognition for Robot Localization B7
4 Memory-controlled Incremental SVM B8
5 Experimental Setup B14
6 Experiments on Support Vector Reduction B19
7 Experiments on Adaptation B20
8 Experiments on Knowledge Transfer B28
9 Summary and Conclusions B34
References B34

C Confidence-based Cue Integration for Visual Place Recognition **C1**
1 Introduction C3
2 Related Work C5
3 A Few Landmarks C5
4 Confidence Estimation C9
5 Cue Integration C14
6 Confidence-based Cue Integration C17
7 Summary and Conclusion C19
References C19

D Multi-modal Semantic Place Classification **D1**
1 Introduction D3
2 Related Work D5
3 Multi-modal Place Classification D7

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

"The eye sees only what the mind is prepared to comprehend"
Henri-Louis Bergson, 1859 –1941

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.

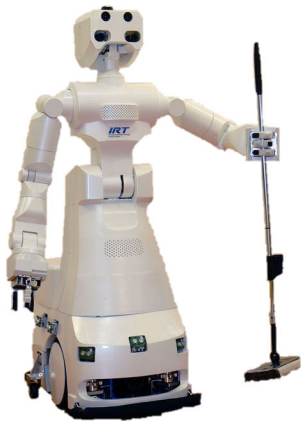
Despite the fact that we might still be far from building robots that could possess human-like intelligence, we are closer than ever to actually fulfilling the dream of ubiquitous robots. Robots have already made their way to our homes, and many believe that within the next few years, we will see a dramatic 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:

(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 [43, 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 surveillance [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. Platforms 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, 54, 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 perceive 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 foundational 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 [71, 75]. In such view, the environment can be considered an additional 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

[59, 131, 87, 99], to topological [122, 109, 56, 86], and hybrid [117, 118, 110, 52, 135, 48]. 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 [54].

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 [107], exploration [112], or navigation [65].

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 [44] 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 I 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*

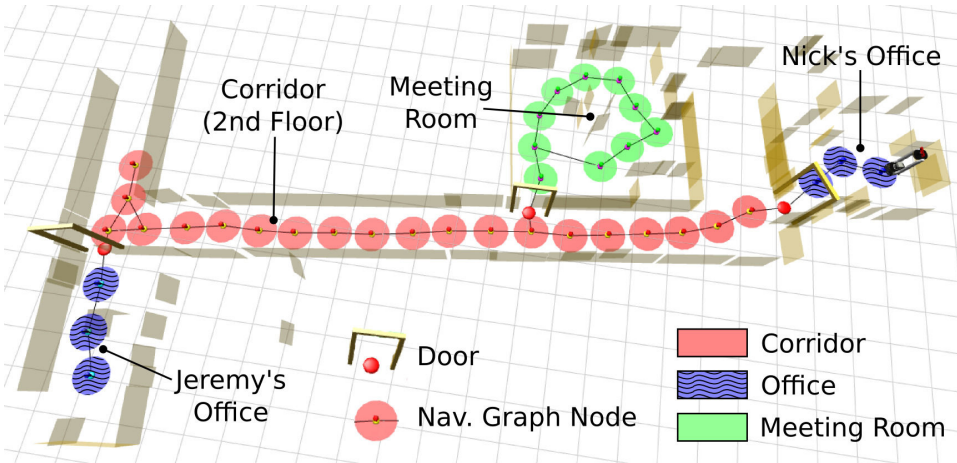


Figure 1: An example of a map augmented with semantic information.

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

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, 109, 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

(a) Office environment

(b) Home environment

Figure 2: Illustration of a typical scenario: a mobile robot platform performing semantic mapping in office and home environments.

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) [55] 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, [81]) or the Speeded Up Robust Features (SURF, [47]) local descriptors are used, combined with the bag-of-words approach [62] for place categorization. The Composed Receptive Fields Histograms (CRFH, [80]) are used as global visual features. For the laser channel the simple geometrical features proposed in [89] 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, [55]). 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 [114] with an exact method that reduces the number of support vectors needed to build the decision function without any loss in performance [61]. 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 A Systems and Representational 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

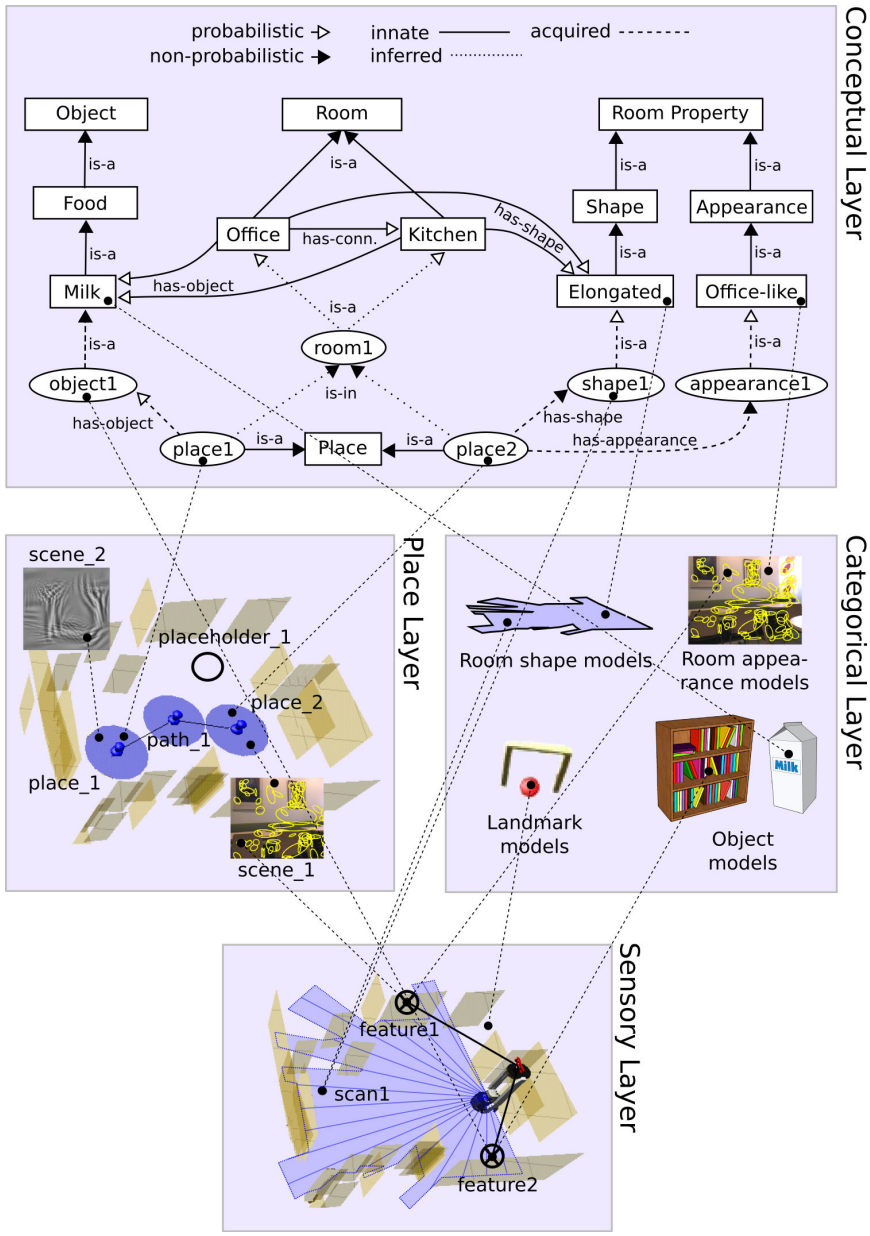


Figure 3: The layered structure of the spatial knowledge representation. The position of each layer within the representation corresponds to the level of abstraction of the spatial knowledge. The conceptual layer illustrates part of the ontology representing both instance and predefined world knowledge.

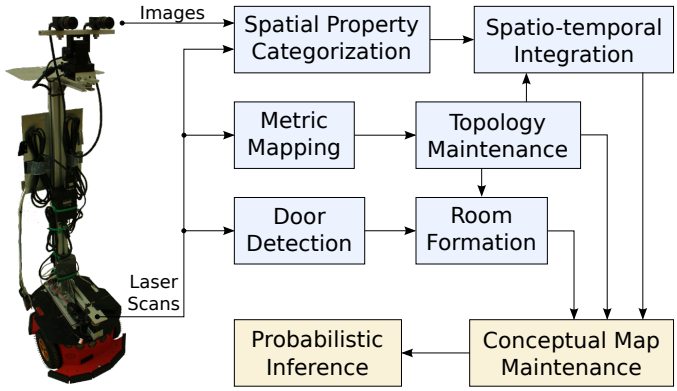


Figure 4: A coarse overview of the elements and the data flow inside the semantic mapping system.

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 sections 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

	Place recognition	Place categorization	Indoor environment	Non-omnidirectional sensor	Appearance, single-cue	Appearance, multi-cue	Geometry	Objects	Confidence estimation	Novelty detection	Spatio-temporal integration	Adaptive/Incremental
[52]	✓		✓				✓					
[122]	✓		✓		✓				✓			
[46]	✓		✓		✓							✓
[92]	✓				✓				✓			
[123], [124]	✓		✓		✓				✓			✓
[74]	✓		✓	✓	✓				✓			
[73]	✓		✓	✓	✓				✓		✓	
[98]	✓		✓	✓	✓							✓
[56], [57]	✓			✓	✓				✓	✓	✓	✓
[86]	✓			✓	✓					✓	✓	✓
[116]	✓		✓		✓		✓		✓	✓	✓	✓
[109]	✓			✓		✓			✓		✓	
[63]	✓		✓	✓		✓			✓			✓
[125]	✓		✓	✓				✓	✓		✓	✓
[106]	✓		✓	✓				✓			✓	
[121]	✓	✓	✓	✓	✓				✓		✓	
[130]		✓		✓	✓				✓			
[62]		✓		✓	✓				✓			
[104]		✓	✓	✓	✓							
[79]		✓	✓	✓	✓							
[133]	✓	✓	✓	✓	✓							
[103]		✓	✓	✓	✓	✓			✓			
[53]		✓	✓				✓		✓		✓	
[89]		✓	✓				✓					
[90]		✓	✓				✓	✓	✓		✓	
[132]		✓	✓	✓	✓				✓		✓	
[129]		✓	✓	✓				✓	✓			
[105]	✓	✓	✓	✓	✓				✓	✓	✓	✓
This work	✓	✓	✓	✓	✓	✓	✓	⋆	✓		✓	✓

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 & Košecká [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. Košecká *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

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 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

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, 63]. 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 [116, 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 better 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 adaptation 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 discriminative 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 exists a broad literature on mobile robot localization, mapping, navigation and place classification. Every such algorithm maintains a representation of spatial knowledge. However, this representation is usually specific to the particular problem

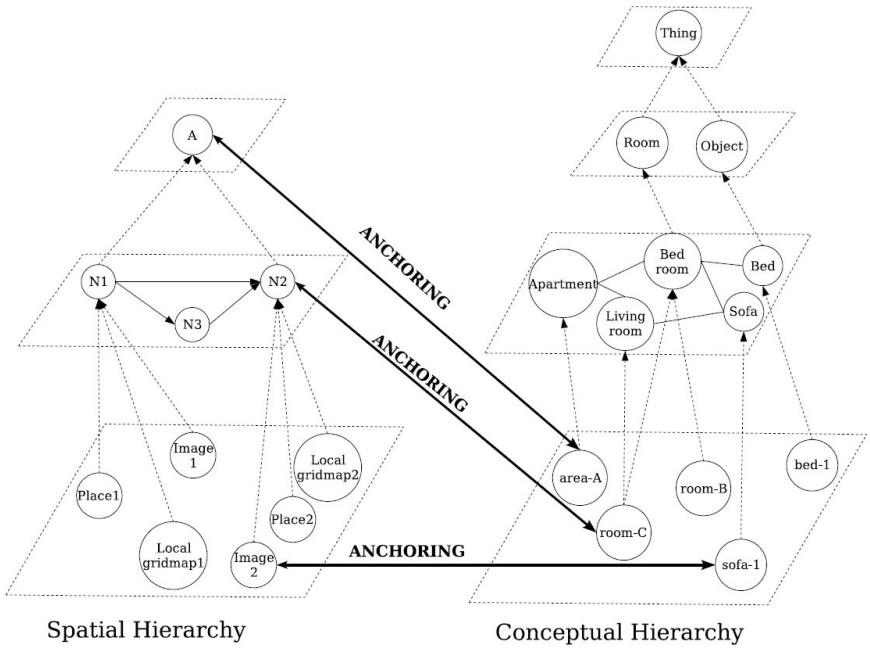


Figure 1: The spatial and semantic information hierarchies. On the left, spatial information gathered by the robot sensors. On the right, semantic information that models concepts and relations between them. Anchoring is used to establish the links between the two hierarchies (solid lines). Additional links can then be inferred by symbolic reasoning (dotted line). Reproduced from [65].

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.

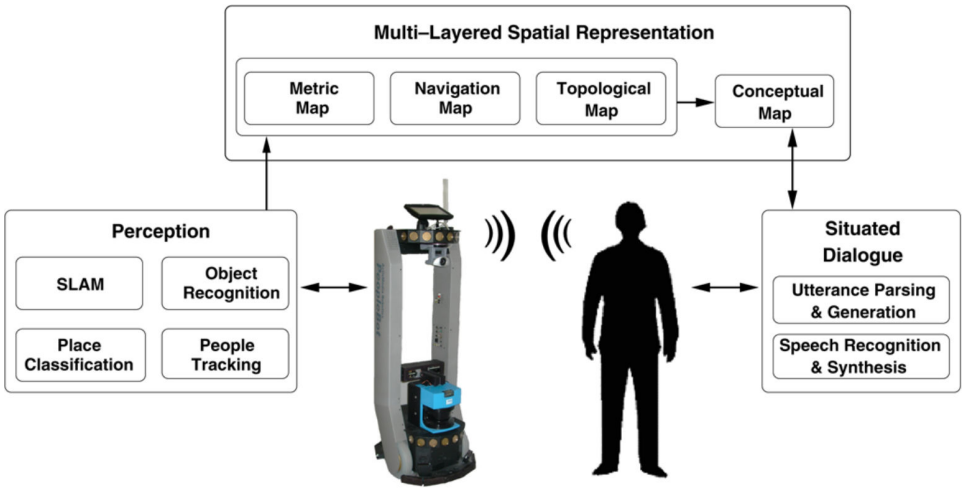


Figure 2: Overview of the components of the system presented in [135] as well as the four layers of the conceptual spatial representation divided into two groups: mapping (metric, navigation, topological) and reasoning (conceptual). Reproduced from [135].

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.

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.

	Indoor environment	Place appearance cues	Place geometry cues	Object information	Topology information	Human input	Segmentation	Conceptual map / Ontology	Uncertain concepts	Inferring properties	Active component	Concepts built automatically
[65]	✓			✓			✓	✓		✓	✓	
[135]	✓		✓	✓		✓	✓	✓		✓		
[126]	✓			✓			✓		✓			✓
[84]	✓			✓							✓	
[128]	✓			✓			✓		✓	✓	✓	✓
[96]	✓		✓	✓								
[94]	✓					✓	✓					
[102]		✓	✓									
[60]				✓								
[100]		✓										
This work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

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

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 1 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.

	Databases and benchmarks	Place recognition	Knowledge transfer	Incremental learning	Confidence estimation	Cue integration	Sensor fusion	Place categorization	Knowledge representation	Semantic mapping	Conceptual map	Integrated cognitive systems	Active semantic perception
[A]	✓	✓											
[B]		✓	✓	✓									
[C]		✓			✓	✓							
[D]		✓				✓	✓	✓		✓			
[E]									✓				
[F]	✓					✓	✓	✓	✓	✓	✓		
[G]										✓	✓	✓	✓
[1]	✓	✓											
[2]		✓	✓										
[3]	✓	✓		✓									
[4]	✓	✓				✓	✓						
[5]	✓	✓						✓					
[6]													
[7]	✓												
[8]									✓				
[9]	✓												
[10]	✓												
[11]	✓												
[12]	✓												
[13]		✓			✓	✓							
[14]		✓				✓	✓	✓	✓	✓	✓		
[15]												✓	
[16]									✓			✓	✓
[17]													✓
[18]									✓		✓		

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 Paper F: Combining Conceptual Knowledge, Objects, Appearance, Geometry and Topology for Semantic Mapping

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 Paper G: Semantic Mapping for Efficient Behaviour of Integrated Robotic Systems

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 [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

[1] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I. Christensen. A discriminative approach to robust visual place recognition. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, October 2006.

[2] Jie Luo, Andrzej Pronobis, and Barbara Caputo. SVM-based transfer of visual knowledge across robotic platforms. In *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS'07)*, Bielefeld, Germany, March 2007.

[3] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. Incremental learning for place recognition in dynamic environments. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, October 2007.

[4] Andrzej Pronobis, Oscar M. Mozos, and Barbara Caputo. SVM-based discriminative accumulation scheme for place recognition. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, Pasadena, CA, USA, May 2008.

[5] Muhammad Muneeb Ullah, Andrzej Pronobis, Barbara Caputo, Jie Luo, Patric Jensfelt, and Henrik I. Christensen. Towards robust place recognition for robot localization. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, Pasadena, CA, USA, May 2008.

[6] Heydar Maboudi Afkham, Alireza Tavakoli Targhi, Jan-Olof Eklundh, and Andrzej Pronobis. Joint visual vocabulary for animal classification. In *Proceedings of the 2008 19th International Conference on Pattern Recognition (ICPR'08)*, Tampa, FL, USA, December 2008.

[7] Andrzej Pronobis and Barbara Caputo. COLD: COsy Localization Database. *The International Journal of Robotics Research (IJRR)*, 28(5):588–594, May 2009.

[8] Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Adrian N. Bishop, and Patric Jensfelt. A framework for robust cognitive spatial mapping. In *Proceedings of the 14th International Conference on Advanced Robotics (ICAR'09)*, Munich, Germany, June 2009.

[9] Andrzej Pronobis, Li Xing, and Barbara Caputo. Overview of the CLEF 2009 robot vision track. In Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth J. F. Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika, editors, *Multilingual Information Access Evaluation Vol. II Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 110–119. Springer, 2010.

[10] Andrzej Pronobis, Henrik Christensen, and Barbara Caputo. Overview of the ImageCLEF@ICPR 2010 Robot Vision track. In Devrim Ünay, Zehra Çataltepe, and Selim Aksoy, editors, *Recognizing Patterns in Signals, Speech, Images and Videos*, volume 6388 of *Lecture Notes in Computer Science*, pages 171–179. Springer Berlin / Heidelberg, 2010.

[11] Andrzej Pronobis, Marco Fornoni, Henrik I. Christensen, and Barbara Caputo. The Robot Vision track at ImageCLEF 2010. In *Working Notes for the CLEF 2010 Workshop*, Padua, Italy, 2010. ISBN 978-88-904810-0-0.

[12] Andrzej Pronobis and Barbara Caputo. The robot vision task. In Henning Müller, Paul Clough, Thomas Deselaers, and Barbara Caputo, editors, *ImageCLEF*, volume 32 of *The Information Retrieval Series*, pages 185–198. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15181-1.

[13] Li Xing and Andrzej Pronobis. Multi-cue discriminative place recognition. In Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth J. F. Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika, editors, *Multilingual Information Access Evaluation Vol. II Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 315–323. Springer, 2010.

[14] Andrzej Pronobis, Patric Jensfelt, Kristoffer Sjöö, Hendrik Zender, Geert-Jan M. Kruijff, Oscar M. Mozos, and Wolfram Burgard. Semantic modelling of space. In Henrik I. Christensen, Geert-Jan M. Kruijff, and Jeremy L. Wyatt, editors, *Cognitive Systems*, volume 8 of *Cognitive Systems Monographs*, pages 165–221. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-11694-0.

[15] Kristoffer Sjöö, Hendrik Zender, Patric Jensfelt, Geert-Jan M. Kruijff, Andrzej Pronobis, Nick Hawes, and Michael Brenner. The explorer system. In Henrik I. Christensen, Geert-Jan M. Kruijff, and Jeremy L. Wyatt, editors, *Cognitive Systems*, volume 8 of *Cognitive Systems Monographs*, pages 395–421. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-11694-0.

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

Andrzej Pronobis, Kristoffer Sjöö, Alen Vrečko, Hendrik Zender, Michael Zillich, and Danijel Skočaj. Self-understanding & self-extension:a systems and representational approach. *IEEE Transactions on Autonomous Mental Development (TAMD), Special Issue on Representations and Architectures for Cognitive Systems*, 2(4):282–303, December 2010.

[17] Alper Aydemir, Kristoffer Sjöö, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, May 2011.

[18] Kristoffer Sjöö, Andrzej Pronobis, and Patric Jensfelt. Functional topological relations for qualitative spatial representation. In *Proceedings of the 15th International Conference on Advanced Robotics (ICAR'11)*, Tallinn, Estonia, June 2011.

[19] Aldebaran Robotics Nao. URL `http://www.aldebaran-robotics.com/`.

[20] Anybots. URL `http://anybots.com/`.

[21] Care-O-Bot. URL `http://www.care-o-bot.de/`.

[22] ConceptNet. URL `http://csc.media.mit.edu/conceptnet`.

[23] DESIRE (Deutsche Servicerobotik Initiative). URL `http://www.service-robotik-initiative.de/`.

[24] EU FP6 Integrated Project COGNIRON: The Cognitive Robot Companion. URL `http://www.cogniron.org`.

[25] EU FP6 Integrated Project RobotCub. URL `http://www.robotcub.org/`.

[26] EU FP6 IST Cognitive Systems Integrated Project CoSy: Cognitive Systems for Cognitive Assistants. URL `http://www.cognitivesystems.org/`.

[27] EU FP7 ICT Cognitive Systems Large-Scale Integrating Project CogX: Cognitive Systems that Self-Understand and Self-Extend. URL `http://cogx.eu/`.

[28] Honda Asimo. URL `http://asimo.honda.com/`.

[29] iRobot AVA. URL `http://www.irobot.com/`.

[30] iRobot Roomba. URL `http://www.irobot.com/`.

[31] MobileRobots PatrolBot. URL `http://www.mobilerobots.com/`.

[32] Open Mind Common Sense. URL `http://openmind.media.mit.edu/`.

[33] OpenCV (Open Source Computer Vision) Library. URL `http://opencv.willowgarage.com/`.

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

[35] Pleo. URL `http://www.pleoworld.com/`.

[36] RoboDynamics Luna. URL `http://robodynamics.com/`.

[37] Samsung Navibot. URL `http://www.samsung.com/`.

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

[39] The Semantic Robot Vision Challenge. URL `http://www.cs.cmu.edu/~srvc/`.

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

[41] Willow Garage PR2. URL `http://www.willowgarage.com/`.

[42] WordNet. URL `http://wordnet.princeton.edu/`.

[43] WowWee Rovio. URL `http://www.wowwee.com/`.

[44] Robotic visions to 2020 and beyond. The strategic research agenda for robotics in Europe. Technical report, The European Robotics Technology Platform (EUROP), July 2009.

[45] Henrik Andreasson, André Treptow, and Tom Duckett. Localization for mobile robots using panoramic vision, local features and particle filter. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, 2005.

[46] Matej Artač, Matjaž Jogan, and Aleš Leonardis. Mobile robot localization using an incremental eigenspace model. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation (ICRA'02)*, pages 1025–1030, May 2002. ISBN 0-7803-7272-7.

[47] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, June 2008.

[48] Patrick Beeson, Joseph Modayil, and Benjamin Kuipers. Factoring the mapping problem: mobile robot map-building in the hybrid spatial semantic hierarchy. *The International Journal of Robotics Research (IJRR)*, 29(4):428–459, April 2010.

[49] Rodney A. Brooks. Aspects of Mobile Robot Visual Map Making. In *Proceedings of the 2nd International Symposium on Robotics Reseach*, pages 369–375, Kyoto, Japan, 1984.

[50] Rodney A. Brooks. Visual Map Making for a Mobile Robot. In *Proceedings of the 1985 IEEE International Conference on Robotics and Automation (ICRA'85)*, pages 824–829, 1985.

[51] Rodney A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[52] Emma Brunskill, Thomas Kollar, and Nicholas Roy. Topological mapping using spectral clustering and classification. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 3491–3496, San Diego, CA, USA, October 2007.

[53] Pär Buschka and Alessandro Saffiotti. A virtual sensor for room detection. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'02)*, Lausanne, Switzerland, 2002.

[54] Henrik I. Christensen, Geert-Jan M. Kruijff, and Jeremy L. Wyatt, editors. *Cognitive Systems*, volume 8 of *Cognitive Systems Monographs*. Springer, 2010. ISBN 978-3-642-11693-3.

[55] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. ISBN 0521780195.

[56] Mark Cummins and Paul M. Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research (IJRR)*, 27(6):647–665, 2008.

[57] Mark Cummins and Paul M. Newman. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Proceedings of Robotics: Science and Systems (RSS'09)*, 2009.

[58] Kostas Daniilidis and Jan-Olof Eklundh. *3-D vision and recognition*, pages 543–562. Springer, 2008. ISBN 978-3-540-23957-4.

[59] Gamini Dissanayake, Paul M. Newman, Steven Clark, Hugh F. Durrant-Whyte, and M. Csorba. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241, June 2001.

[60] Bertrand Douillard, Dieter Fox, and Fabio Ramos. Classification and semantic mapping of urban environments. *The International Journal of Robotics Research (IJRR)*, 30(1):5–32, 2010.

[61] Tom Downs, Kevin E. Gates, and Annette Masters. Exact simplification of support vector solutions. *Journal of Machine Learning Research (JMLR)*, 2 (2):293–297, May 2002.

[62] Li Fei-Fei and Pietro Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005.

[63] David Filliat. A visual bag of words method for interactive qualitative localization and mapping. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[64] Friedrich Fraundorfer, Christopher Engels, and David Nistér. Topological mapping, localization and navigation using image collections. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[65] Cipriano Galindo, Alessandro Saffiotti, Silvia Coradeschi, Pär Buschka, Juan Antonio Fernández-Madrigal, and Javier González. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, 2005. ISBN 0-7803-8912-3.

[66] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6): 712–722, June 2010.

[67] Bill Gates. A robot in every home. *Scientific American*, 296(1):58–65, January 2007.

[68] Rodolphe Gelin and Henrik I. Christensen. Building the European Robotics Platform - EUROP, Sectoral Report on Service Robotics. Technical Report May 2005, May 2005.

[69] Nick A. Hawes, Marc Hanheide, Jack Hargreaves, Ben Page, and Hendrik Zender. Home alone: Autonomous extension and correction of spatial representations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, 2011.

[70] Nick A. Hawes and Jeremy L. Wyatt. Engineering intelligent information-processing systems with CAST. *Advanced Engineering Informatics*, 24(1): 27–39, January 2010.

[71] Helge Huettenrauch, Kerstin Eklundh, Anders Green, and Elin Anna Topp. Investigating spatial relationships in human-robot interaction. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, pages 5052–5059, Beijing, China, October 2006. ISBN 1-4244-0258-1.

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

the 12th National Conference on Artificial Intelligence (AAAI'94), Seattle, Washington, USA, 1994.

[73] Jana Košecká, Fayin Li, and Xialong Yang. Global localization and relative positioning based on scale-invariant keypoints. *Robotics and Autonomous Systems*, 52:27–38, 2005.

[74] Jana Košecká, Liang Zhou, Philip Barber, and Zoran Duric. Qualitative image based localization in indoors environments. In *Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, WI, USA, 2003.

[75] Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems, Special Issue on Human and Robot Interactive Communication*, 4(1):125–138, 2007.

[76] Benjamin Kuipers. The Spatial Semantic Hierarchy. *Artificial Intelligence*, 119(1-2):191–233, 2000.

[77] Benjamin Kuipers and Patrick Beeson. Bootstrap learning for place recognition. *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI'02)*, pages 174–180, 2002.

[78] Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal Of The Royal Statistical Society Series B*, 64(3):321–348, 2002.

[79] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2169–2178, 2006. ISBN 0-7695-2597-0.

[80] Oskar Linde and Tony Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proceedings of the 2004 15th International Conference on Pattern Recognition (ICPR'04)*, 2004.

[81] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[82] Martin Magnusson, Henrik Andreasson, Andreas Nüchter, and Achim J. Lilienthal. Appearance-based loop detection from 3D laser data using the normal distributions transform. *Journal of Field Robotics*, 26(11-12):892–914, 2009.

[83] Mario Mata, Jose M. Armingol, Arturo de la Escalera, and Miguel Angel Salichs. Using learned visual landmarks for intelligent topological navigation of mobile robots. In *Proceedings of the 2003 IEEE International Conference on Robotics and Automation (ICRA'03)*, 2003.

[84] David Meger, Per-Erik Forssen, Kevin Lai, Scott Helmer, Sancho McCann, Tristram Southey, Matthew Baumann, James J. Little, and David G. Lowe. Curious George: An attentive semantic robot. *Robotics and Autonomous Systems (RAS)*, 56(6):503–511, 2008.

[85] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. The iCub humanoid robot: an open platform for research in embodied cognition. In Raj Madhavan and Elena Messina, editors, *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems (PerMIS)*, NIST Special Publication 1090, pages 50–56, 2008.

[86] Michael J. Milford and Gordon F. Wyeth. Mapping a suburb with a single camera using a biologically inspired SLAM system. *IEEE Transactions on Robotics*, 24(5):1038–1053, October 2008.

[87] Michael Montemerlo and Sebastian Thrun. *FastSLAM: A Scalable Method for the Simultaneous Localization and Mapping Problem in Robotics*. Springer Tracts in Advanced Robotics. Springer, 2007.

[88] Thomas Mörwald, Johann Prankl, Andreas Richtsfeld, Michael Zillich, and Markus Vincze. BLORT - The blocks world robotic vision toolbox. In *Proceedings of the ICRA Workshop Best Practice in 3D Perception and Modeling for Mobile Manipulation*, 2010.

[89] Oscar Martinez Mozos, Cyrill Stachniss, and Wolfram Burgard. Supervised learning of places from range data using AdaBoost. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, 2005.

[90] Oscar Martinez Mozos, Rudolph Triebel, Patric Jensfelt, Axel Rottmann, and Wolfram Burgard. Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems (RAS)*, 55(5): 391–402, 2007.

[91] Ana Cris Murillo, José Jesús Guerrero, and Carlos Sagüés. SURF features for efficient robot localization with omnidirectional images. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[92] Ana Cris Murillo and Jana Košecká. Experiments in place recognition using gist panoramas. In *The 9th ICCV Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras*, Kyoto, Japan, 2009.

[93] Paul M. Newman and Kin Ho. SLAM-loop closing with visually salient features. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, pages 635–642, 2005. ISBN 0-7803-8914-X.

[94] Carlos Nieto-Granda, John G. Rogers, Alexander J. B. Trevor, and Henrik I. Christensen. Semantic map partitioning in indoor environments using regional analysis. In *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'10)*, pages 1451–1456, Taipei, Taiwan, October 2010.

[95] Maria-Elena Nilsback and Barbara Caputo. Cue integration through discriminative accumulation. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, DC, USA, 2004.

[96] Andreas Nüchter and Joachim Hertzberg. Towards semantic maps for mobile robots. *Robotics and Autonomous Systems (RAS)*, 56(11):915–926, 2008.

[97] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155, 2006.

[98] Francesco Orabona, Claudio Castellini, Barbara Caputo, Jie Luo, and Giulio Sandini. Indoor place recognition using online independent Support Vector Machines. In *Proceedings of the British Machine Vision Conference (BMVC'07)*, Warwick, UK, 2007.

[99] Lina María Paz, Patric Jensfelt, Juan D. Tardós, and José Neira. EKF SLAM updates in O(n) with Divide and Conquer SLAM. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[100] Martin Persson, Tom Duckett, Christoffer Valgren, and Achim J. Lilienthal. Probabilistic semantic mapping with a virtual sensor for building/nature detection. In *2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 236–242, Jacksonville, Florida, USA, June 2007. IEEE. ISBN 1-4244-0789-3.

[101] Axel Pinz. Object categorization. *Foundations and Trends in Computer Graphics and Vision*, 1(4):255–353, 2006.

[102] Ingmar Posner, Derik Schroeter, and Paul M. Newman. Describing composite urban workspaces. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[103] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, Miami Beach, FL, USA, 2009.

[104] Pedro Quelhas, Florent Monay, Jean-Marc Odobez, Daniel Gatica-Perez, and Tinne Tuytelaars. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(9):1575–89, September 2007.

[105] Ananth Ranganathan. PLISS: Detecting and labeling places using online change-point detection. In *Proceedings of Robotics: Science and Systems (RSS'10)*, Zaragoza, Spain, June 2010.

[106] Ananth Ranganathan and Frank Dellaert. Semantic modeling of places using objects. In *Proceedings of Robotics: Science and Systems (RSS'07)*, 2007.

[107] Axel Rottmann, Oscar Martinez Mozos, Cyrill Stachniss, and Wolfram Burgard. Semantic place classification of indoor environments with mobile robots using boosting. In *Proceedings of the 20nd National Conference on Artificial Intelligence (AAAI'05)*, pages 1306–1311, Pittsburgh, Pennsylvania, USA, July 2005. ISBN 1-57735-236-x.

[108] Stephen Se, David G. Lowe, and James J. Little. Vision-based global localization and mapping for mobile robots. *IEEE Transactions on Robotics*, 21 (3):364–375, June 2005.

[109] Christian Siagian and Laurent Itti. Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[110] Robert Sim and James J. Little. Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, pages 2082–2089, Beijing, China, October 2006. ISBN 1-4244-0258-1.

[111] Thorsten Spexard, Shuyin Li, Britta Wrede, Jannik Fritsch, Gerhard Sagerer, Olaf Booij, Zoran Zivkovic, Bas Terwijn, and Ben Kröse. BIRON, where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, October 2006. ISBN 1-4244-0258-1.

[112] Cyrill Stachniss, Oscar Martinez Mozos, and Wolfram Burgard. Speeding-up multi-robot exploration by considering semantic place information. In *Proceedings of the 2006 IEEE International Conference on Robotics and Automation (ICRA'06)*, Orlando, Florida, USA, 2006.

[113] Bipin Suresh, Carl Case, Adam Coates, and Andrew Y. Ng. Autonomous sign reading for semantic mapping. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, May 2011.

[114] Nadeem Ahmed Syed, Huan Liu, and Kah Kay Sung. Incremental learning with support vector machines. In *Proceedings of the 16th International Joint Conference on Artificial intelligence (IJCAI'99)*, Stockholm, Sweden, 1999.

[115] Hashem Tamimi and Andreas Zell. Vision based localization of mobile robots using kernel approaches. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'04)*, Sendai, Japan, 2004.

[116] Adriana Tapus and Roland Siegwart. Incremental Robot Mapping with Fingerprints of Places. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Edmonton, Alberta, Canada, 2005.

[117] Sebastian Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99(1):21–71, 1998.

[118] Nicola Tomatis, Illah Nourbakhsh, and Roland Siegwart. Hybrid simultaneous localization and map building: a natural integration of topological and metric. *Robotics and Autonomous Systems (RAS)*, 44(1):3–14, July 2003.

[119] Elin Anna Topp and Henrik I. Christensen. Topological modelling for human augmented mapping. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, 2006.

[120] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision (IJCV)*, 53(2), 2003.

[121] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, 2003.

[122] Iwan Ulrich and Illah Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation (ICRA'00)*, San Francisco, CA, USA, 2000.

[123] Christoffer Valgren and Achim J. Lilienthal. Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 1856–1861, May 2008. ISBN 978-1-4244-1646-2.

[124] Christoffer Valgren and Achim J. Lilienthal. SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments. *Robotics and Autonomous Systems (RAS)*, 58(2):149–156, February 2010.

[125] Shrihari Vasudevan, Stefan Gächter, Viet Nguyen, and Roland Siegwart. Cognitive maps for mobile robots - an object based approach. *Robotics and Autonomous Systems (RAS)*, 55(5):359–371, May 2007.

[126] Shrihari Vasudevan and Roland Siegwart. Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robotics and Autonomous Systems (RAS)*, 56:522–537, 2008.

[127] Karel Čapek. *R.U.R (Rossumovi univerzální roboti)*. 1920.

[128] Pooja Viswanathan, David Meger, Tristram Southey, James J. Little, and Alan K. Mackworth. Automated spatial-semantic modeling with applications to place labeling and informed search. In *Proceedings of the 2009 Canadian Conference on Computer and Robot Vision (CRV'09)*, Kelowna, British Columbia, Canada, May 2009. ISBN 978-1-4244-4211-9.

[129] Pooja Viswanathan, Tristram Southey, James J. Little, and Alan K. Mackworth. Automated place classification using object detection. In *Proceedings of the 2010 Canadian Conference on Computer and Robot Vision (CRV'10)*, Ottawa, Ontario, Canada, June 2010. ISBN 978-1-4244-6963-5.

[130] Julia Vogel and Bernt Schiele. A semantic typicality measure for natural scene categorization. *Annual Pattern Recognition Symposium (DAGM'04)*, 2004.

[131] Jürgen Wolf, Wolfram Burgard, and Hans Burkhardt. Robust vision-based localization by combining an image retrieval system with Monte Carlo localization. *IEEE Transactions on Robotics*, 21(2):208–216, 2005.

[132] Jianxin Wu, Henrik I. Christensen, and James M. Rehg. Visual place categorization: problem, dataset, and algorithm. In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'09)*, 2009.

[133] Jianxin Wu and James M. Rehg. Where am I: place onstance and category recognition using spatial PACT. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, June 2008. ISBN 978-1-4244-2242-5.

[134] Kimitoshi Yamazaki, Ryohei Ueda, Shunichi Nozawa, Yuto Mori, Toshiaki Maki, Naotaka Hatao, Kei Okada, and Masayuki Inaba. A Demonstrative research for daily assistive robots on tasks of cleaning and tidying up rooms. In *Proceedings of the 14th Robotics Symposia*, pages 522–527, 2009.

[135] Hendrik Zender, Oscar Martinez Mozos, Patric Jensfelt, Geert-Jan M. Kruijff, and Wolfram Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems (RAS)*, 56(6):493–502, June 2008.

Part II

Included Papers

Paper A

A Realistic Benchmark for Visual Indoor Place Recognition

Andrzej Pronobis, Barbara Caputo,
Patric Jensfelt and Henrik I. Christensen

Published in
Robotics and Autonomous Systems

A Realistic Benchmark for Visual Indoor Place Recognition

Andrzej Pronobis, Barbara Caputo,
Patric Jensfelt and Henrik I. Christensen

Abstract

An important competence for a mobile robot system is the ability to localize and perform context interpretation. This is required to perform basic navigation and to facilitate local specific services. Recent advances in vision have made this modality a viable alternative to the traditional range sensors and visual place recognition algorithms emerged as a useful and widely applied tool for obtaining information about robot's position. Several place recognition methods have been proposed using vision alone or combined with sonar and/or laser. This research calls for standard benchmark datasets for development, evaluation and comparison of solutions. To this end, 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 Introduction

A fundamental competence for an autonomous agent is to know its position in the world. Providing mobile robots with abilities to build an internal representation of space and obtain robust information about their location therein can be considered as one of the most urgent problems. The topic is vastly researched. This resulted, over the years, in a broad range of approaches spanning from purely metric [27, 18, 62], to topological [58, 57, 17], and hybrid [53, 12]. As robots break down the fences and start to interact with people [63] and operate in large-scale environments [17, 57], topological models are gaining popularity for augmenting or replacing purely

metric space representations. In particular, the research on topological mapping has pushed methods for place recognition. Scalability, loop closing, and the kidnapped robot problem have been at the forefront of the issues to be addressed.

Traditionally, sonar and/or laser have been the sensory modalities of choice for place recognition and topological localization [42, 38]. The assumption that the world can be represented in terms of two dimensional geometrical information allowed for many practical implementations. Yet, the inability to capture many aspects of complex realistic environments leads to the problem of perceptual aliasing [29], and greatly limits the usefulness of purely geometrical methods. Recent advances in vision have made this modality emerge as a natural and viable solution. Vision provides richer sensory input allowing for better discrimination. It opens new possibilities for building cognitive systems, actively relying on semantic context. Not unimportant is the cost effectiveness, portability and popularity of visual sensors. As a result, this research line is attracting more and more attention, and several methods have been proposed using vision alone [55, 48, 50, 17], or combined with more traditional range sensors [28, 52, 49].

In spite of large progress, vision-based localization still represents a major challenge. First of all, visual information tends to be noisy and difficult to interpret. The visual appearance of places varies in time because of illumination changes (day and night, artificial light on and off) and because of human activities (furniture moved around, objects being taken out of drawers, and so on). Thus, the solutions must be highly robust, provide good generalization abilities and in general be adaptive. Additionally, the application puts strong constraints on the computational complexity, and the increased resolution and dimensionality of the visual data still constitutes a problem.

The fact that so many different parameters influence the accuracy of a vision-based localization system is another challenge itself, especially burdensome at the design stage. As the results depend greatly on the choice of training and test input data, which are unstable over time, it is hard to measure the influence of the different parameters on the overall performance of the system. For the same reason, it becomes nearly impossible to compare fairly solutions which are usually evaluated in different environments, under different conditions, and with different assumptions. This is a major obstacle slowing down progress in the field. There is a need for standardized benchmarks and databases which would allow for fair comparisons, simplify the experimental process and boost development of new solutions.

Databases are heavily exploited in the computer vision community, especially for object recognition and categorization [25, 4, 3]. As the community acknowledges the need for benchmarking, a lot of attention is directed towards designing new datasets, reflecting the increasing capabilities of visual algorithms [45]. Also in robotics, research on Simultaneous Localization and Mapping (SLAM) makes use of several publicly available datasets [26, 40]. Still, no database emerged as a standard benchmark for visual place recognition applied to robot localization.

This paper aims at filling this gap, and presents a benchmark consisting of two different image databases gathered in the same indoor environment. The databases

are augmented with an experimental procedure as well as extensive baseline evaluation. The datasets were carefully designed and later annotated. Three different imaging devices were used for acquisition (two mobile robot platforms and a standard camera), resulting in data of different characteristics and quality. In order to create a realistic and challenging test bed, the acquisition process was performed in an uncontrolled typical office environment, under various illumination and weather conditions (sunny, cloudy, night), and over a significant span of time. All of this makes the databases very well suited for evaluating robustness of visual place recognition algorithms, applied to the problem of robot topological localization, in presence of different types of variations often occurring in real-world indoor settings.

An important component when providing the community with a new collection of data is to provide a baseline evaluation that illustrates the nature of the dataset (see Section 5.1 for explanation). We thoroughly assessed the databases with a purely appearance-based place recognition method. The method uses two types of image descriptors, local and global, in order to extract rich visual information. Both descriptors have shown remarkable performances, coupled with computational efficiency on challenging object recognition scenarios [31, 30]. The classification step is performed using Support Vector Machines [16] and specialized kernels are used for each descriptor. Results show that the method is able to recognize places with high precision and robustness under varying illumination conditions, even when training on images from one camera device and testing on another.

The rest of the paper is organized as follows: after a review of related literature (Section 2), we discuss the problem and challenges we addressed with the benchmark (Section 3). Then, Section 4 gives a detailed description of the data acquisition process and scenario and presents the acquisition results. Finally, the algorithm used for the baseline evaluation as well as the experimental procedure are described in Section 5, and the experimental results are given in Section 6. The paper concludes with a summary (Section 7).

2 Related work

Place recognition and topological localization are vastly researched topics in the robotic community, where vision and laser range sensors are usually the privileged modalities. Although laser-based solutions have proven to be successful for certain tasks [38], their limitations inspired many researchers to turn towards vision which nowadays becomes tractable in real-time applications. The available methods employ either perspective [55, 51, 20] or omnidirectional cameras [23, 9, 58, 35, 6, 39, 59]. The main differences between the approaches relate to the way the scene is perceived, and thus the method used to extract characteristic features from the scene. Landmark-based techniques make use of either artificial or natural landmarks in order to extract information about a place. Mata *et al.* [34] proposed a system able to interpret information signs through its ability to

read text and recognize icons. Visually distinctive regions were also used as landmarks [50]. Other solutions employed mainly local image features such as SIFT [31, 6, 48], SURF [8, 39, 59], also using the bag-of-words approach [20, 22, 17], or representation based on information extracted from local patches using Kernel PCA [51]. Global features are also commonly used for place recognition. Torralba *et al.* [56, 55, 54] suggested to use a representation called the "gist" of the scene, which is a vector of principal components of outputs of a bank of spatially organized filters applied to the image. Other approaches use color histograms [58, 9], gradient orientation histograms [11], eigenspace representation of images [23], or Fourier coefficients of low frequency image components [35]. Several authors observed that robustness and efficiency of the recognition system can be improved by combining information provided by both types of visual cues (global and local) [48, 50, 61]. Although vision-based localization methods are now commonly applied, it remains difficult to compare the different approaches, as the evaluations presented by the authors usually follow different procedures and are performed on different data sets.

There are a number of heavily used standard databases in robotics and computer vision. In robotics, these databases are used mainly for testing algorithms for simultaneous localization and mapping (SLAM) [26, 40] and mostly contain odometry and range sensor data. In case of the computer vision community, the effort concentrated on creating standard benchmarks for such problems as object [25, 4, 45], action [33], scene [3], or texture recognition and categorization [2]. The MIT-CSAIL Database of Objects and Scenes [3] is a notable exception as it provides several image sequences acquired in both indoor and outdoor environments and was used to evaluate performance of a visual place recognition system.

This paper makes an important contribution by providing annotated data from visual and laser range sensors together with an experimental procedure that can be followed in order to evaluate place recognition and localization systems. In contrast to the previously available benchmarking solutions, the databases contain several sets of images and image sequences acquired in the same environment under various conditions and over a significant span of time. This makes them perfect for evaluating robustness of the algorithms under dynamic variations that often occur in realistic settings. The introduction of standard benchmark databases has made an impact on the research on such problems as object categorization or simultaneous localization and mapping (SLAM), allowing different methods to be more fairly compared in the same scenario. The authors hope that the benchmark proposed in this paper will similarly influence the research on visual place recognition in the context of mobile robot localization.

3 Design strategy

This section defines and characterizes the problem that we address with the benchmark (Section 3.1) and analyzes the difficulties and open challenges in visual place recognition that have to be considered in a realistic scenario (Section 3.2).

3.1 Problem Statement

Let us begin with a brief definition of a place and the place recognition problem that we will use throughout this paper. A place can be regarded as a usually nameable segment of a real-world environment distinguished due to different functionality, appearance or artificial boundary. In view of this definition, the place recognition or identification problem can be characterized as follows. Given a set of training sensory data, captured in each of the considered places, build models of the places reflecting their inherent properties. Next, when presented with new test data, unavailable during training, acquired in one of the same places, identify the place where the acquisition was performed (e.g. Barbara's office) based on the knowledge encoded in the models. This is different from the problem of place categorization where the task is to classify test data captured in a novel place as belonging to one of the place categories (e.g. an office). As the partition of space into different places can be based on several criteria, here we consider a supervised scenario where the algorithm has to distinguish between five areas of different functionality, selected by a teacher.

This benchmark is designed to test the performance of a visual place recognition system on images acquired within an indoor office environment. As the primary scenario, we consider the case where a place recognition system is used to provide a mobile robot with information about its location. For this reason, part of the data presented in this paper was acquired using cameras mounted on mobile robot platforms. While designing the benchmark, we concentrated on testing the ability of a visual recognition system to identify a place based on one image only. This makes the problem harder, but also makes it possible to perform global localization where no prior knowledge about the position is available (e.g. in case of the kidnapped robot problem). Spatial or temporal filtering can be used together with the presented methods to enhance performance.

We concentrate on indoor environments, since in the considered scenario, they play a crucial role, being typical spaces for the interaction between humans and service robots or robotic assistants [63]. At the same time, office environments, just like home environments, constitute an important class of indoor spaces for robotic companions. In this benchmark, our aim is to provide datasets and experimental procedures that will allow for evaluating robustness of place recognition systems based on different types of visual cues to typical variations that occur in an indoor environment for the considered scenario. These include illumination changes, variations introduced by human activity and viewpoint changes. As a consequence, instead of providing datasets spanning over a very large portion of space, we provide image sequences acquired over a time span of several months, under various illumination conditions and using different devices. The proposed evaluation framework should allow for concluding that an algorithm robust to the variations captured in the benchmark data will be robust to similar types of variations within other indoor office environments.

The benchmark is designed for evaluating vision-based methods. We choose

vision as sensory modality for several reasons. First, the visual sensor is very rich and, although also very noisy, provides great descriptive capabilities. This is crucial in indoor environments where other sensors, such as a laser range finder, suffer from the problem of perceptual aliasing (different places look the same [29]). Furthermore, the visual appearance of places encodes information about their semantics, which plays a major role in enabling systems to interact with the environment. Finally, in the era of cheap portable devices equipped with digital cameras, it is also one of the most affordable and commonly available solutions.

3.2 Challenges

Recognizing indoor places based on their visual appearance is a particularly challenging task. First of all, in case of indoor environments, there is no obvious spatial layout that once observed could be used to distinguish between different places. Moreover, viewpoint variations cause the visual sensor to capture different aspects of the same place, which often can only be learned if enough training data are provided. At the same time, real-world environments are usually dynamic and their appearance changes over time. The visual recognition system must be robust to variations introduced by changing illumination as well as human activity. 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. Moreover, if the environment is being used, the fact that people appear in the images, objects are being moved or furniture relocated may greatly influence the performance of the system. All these issues were taken into consideration while designing this benchmark in order to create a realistic test bed.

4 Data Acquisition

Based on the analysis of the problem presented in the previous section, we carefully designed and acquired two databases comprising images captured in the same indoor environment, but using different devices: the INDECS (INDoor Environment under Changing conditionS) database [47] and the IDOL (Image Database for rObot Localization) database [32]. This section describes the resulting data acquisition procedure. In case of INDECS, we acquired images of the environment 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. All three devices are shown in Fig. 1. 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. Both databases are publicly available and can be downloaded from `http://www.pronobis.pro`.

(a) Minnie (b) Camera (c) Dumbo

Figure 1: Devices employed in the acquisition: the two mobile robot platforms "Minnie" (a) and "Dumbo" (c) as well as the standard camera on a tripod (b).

The rest of the section is organized as follows: Section 4.1 presents the acquisition scenario, as to say the environment where both databases were acquired. Then, Section 4.2 provides a description of the INDECS database, and Section 4.3 gives detailed information on the robot platforms and IDOL. Finally, we perform an analysis of the obtained data in Section 4.4.

4.1 Acquisition Scenario

The acquisition was conducted within a five room subsection of a larger office environment of the Computer Vision and Active Perception Laboratory at the Royal Institute of Technology in Stockholm, Sweden. Each of the five rooms represents a different type of functional area: a one-person office, a two-persons office, a kitchen, a corridor, and a printer area (in fact a continuation of the corridor). The function that a room fulfills determines the furniture, objects, and activity that is likely to be found there. Places like the corridor, the printer area and the kitchen can be regarded as public which implies that various people may be present. On the other hand, offices were imaged usually empty or with their owners at work. In the corridor and the printer area, furniture is mostly fixed and objects are less moveable. As a result, these areas were less susceptible to variations caused by human activity in comparison to the kitchen or the offices, where furniture (e.g.

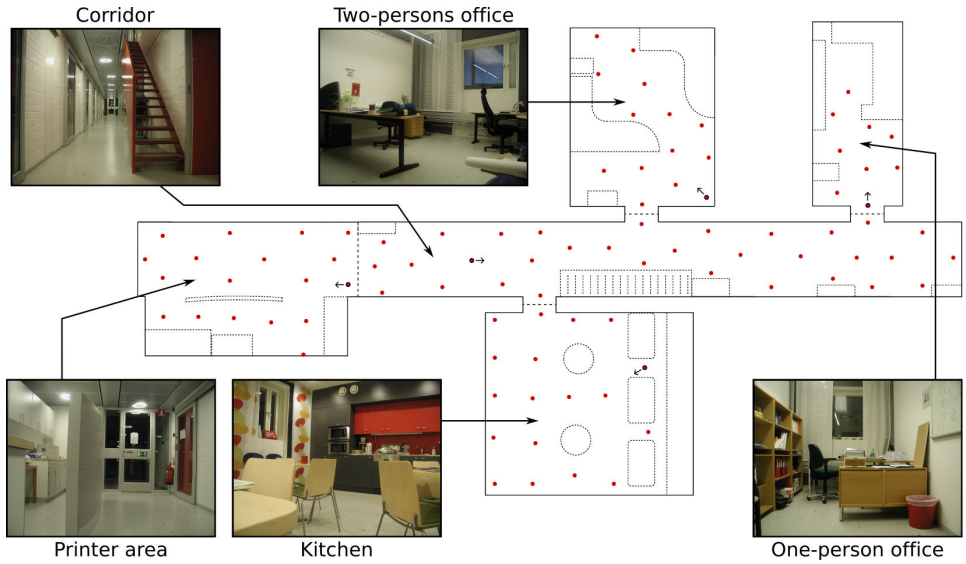


Figure 2: A general map of the part of the office environment that was imaged during acquisition of the INDECS and IDOL databases. Boundaries between the five rooms were marked with dashed lines. Dotted lines were used to draw an approximate outline of furniture. Moreover, the location of points at which the tripod was placed while recording the INDECS database were marked. The pictures are taken from the database and show the interiors of the five rooms. The small arrows were used to indicate the viewpoints at which the presented pictures were taken.

chairs) is relocated more often and objects (e.g. cups, laptops etc.) are frequently moved.

The rooms are physically separated by sliding glass doors. The printer area is an exception and was treated as a separate place only due to its different functionality (the border between the corridor and the printer area was arbitrarily defined). The laboratory contains additional rooms which were not taken into consideration while creating the database. However, because of the glass door, other parts of the environment can still be visible in the images. Examples of pictures showing the interior of each room as well as a general map of the environment are presented in Fig. 2.

As already mentioned, the visual data were acquired with three different devices. In each case, the appearance of the rooms was captured under three different illumination and weather conditions: in cloudy weather (natural and artificial light), in sunny weather (direct natural light dominates), and at night (only artificial light). Since all the rooms have windows, the influence of natural illumination was signif-

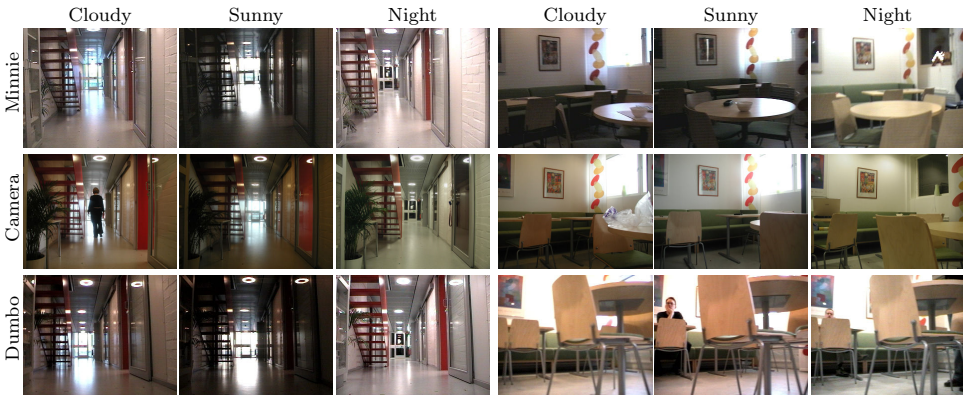
icant. The image acquisition was spread over a period of time of three months, for the INDECS database, and over two weeks for the IDOL database. Additionally, the INDECS database was acquired ten months before the experiments with the robots. In this way, we captured the visual variability that occurs in realistic environments due to varying illumination and natural activities in the rooms. Fig. 3 presents a comparison of images taken under different illumination conditions and using various devices.

4.2 The INDECS database

The INDECS database consists of pictures of the environment described above, gathered from different viewpoints using a standard camera mounted on a tripod. We marked several points in each room (approximately one meter apart) where we positioned the camera for each acquisition. The rough positions of all points are shown on the map in Fig. 2. The number of points changed with the dimension of the room, from a minimum of 9 for the one-person office to a maximum of 32 for the corridor. At each location we acquired 12 images, one every 30°, even when the tripod was located very close to a wall or furniture. Examples of images taken at the same location and from several angles are presented in Fig. 4. Images were acquired using an Olympus C-3030ZOOM digital camera and the height of the tripod was constant and equal to 76 cm. All images in the INDECS database were acquired with a resolution of 1024x768 pixels, the auto-exposure mode enabled, flash disabled, the zoom set to wide-angle mode, and the auto-focus enabled. In this paper, the INDECS images were subsampled to 512x386 before being used in the experiments. The images were labeled according to the position of the point where the acquisition happened. As a consequence, images taken, for example, from the corridor but looking into a room are labeled as the corridor. The images were acquired across a time span of three months and under varying illumination conditions (sunny, cloudy and night). For each illumination setting, we captured one full set of images. In total, there are 3264 images (324 for the one-person office, 492 for the two-persons office, 648 each for the kitchen and the printer area, and 1152 for the corridor) in the INDECS database.

4.3 The IDOL database

The IDOL database was acquired using cameras on two mobile robot platforms. Both robots, the PeopleBot Minnie and the PowerBot Dumbo, were equipped with a pan-tilt-zoom Canon VC-C4 camera, a SICK laser range finder, and wheel encoders. However, as it can be seen from Fig. 1, the cameras were mounted at different height. On Minnie, the camera was 98cm above the floor, whereas on Dumbo it was 36cm. Furthermore, the camera on Dumbo was tilted up approximately 13°, to reduce the amount of floor captured in the images. The selected positions of the cameras result in different characteristics of the environment being captured in the images. Due to the low placement of the camera on Dumbo, the captured images



(a) Corridor (b) Kitchen

(c) Two-persons office

Figure 3: Example pictures taken from the INDECS and IDOL databases acquired with the camera and the two robot platforms under various illumination conditions. The pictures show the influence of illumination (especially (a) and (c)) and illustrate the differences between images acquired in a cluttered environment using different devices (b). Additional variability caused by natural activities in the rooms is also apparent (presence of people, relocated objects and furniture).

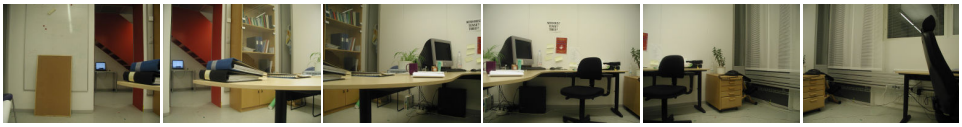


Figure 4: Pictures from the INDECS database taken from several angles at the same location in the two-persons office.

are less susceptible to variations introduced by human activity in the environment and direct sunlight coming through the windows. At the same time, the camera on Minnie was able to capture the appearance of objects located on the desks and provide more information about the semantics of a place. All images were acquired with a resolution of 320x240 pixels, with the zoom fixed to wide-angle (roughly 45° horizontal and 35° vertical field of view), the auto-exposure and the auto-focus modes enabled.

We followed the same procedure during image acquisition with both robot platforms. Each robot was manually driven (average speed around 0.3-0.35m/s) through each of the five rooms while continuously acquiring images at the rate of five frames per second. The path was roughly planned so that the robots could capture the visual appearance of all the rooms. For the different illumination conditions (sunny, cloudy, night), the acquisition procedure was performed twice, resulting in two image sequences acquired one after another giving a total of six sequences for each robot platform across a span of over two weeks. Each of the image sequences in the database is accompanied by laser scans and odometry data. Due to the manual control, the path of the robot was slightly different for every sequence. Examples of paths are presented in Fig. 7, 8, and 9. Each image sequence consists of 1000-1300 frames. To automate the process of labeling the images for the supervision, the robot pose was estimated during the acquisition process using a laser based localization method [21]. Again, each image was labeled as belonging to one of the five rooms based on the position from where it was taken.

4.4 Acquisition Results

Examples illustrating the properties of images that can be found in both databases are given in Fig. 3. First of all, we can observe a significant influence of illumination. The appearance of the rooms is affected by highlights, shadows and reflections, especially in case of strong direct sunlight. Moreover, the fact that the auto-exposure mode was on, resulted in a lower contrast in the informative parts of images, when the camera was directed towards a bright window in sunny weather. At the same time, the conditions observed during cloudy weather were much more stable and could be seen as intermediate between those during sunny weather and at night. A second important type of variability was introduced by human presence and activities. In some cases, people partially occluded the view. Furthermore, the fact

that the environment was observed for some time, allowed to capture different configurations of furniture or objects placed on the desks or kitchen tables. The fact that objects could be observed in the images makes it possible to use the database in more complex scenarios where place recognition and object recognition complement each other e.g. by contextual priming [55, 54] (especially in case of the high resolution images in the INDECS database). Finally, we can compare the images acquired using the three different devices. We see that each device captures different aspects of the same environment, mainly due to the variations in viewpoint caused by the different heights of the cameras. The influence of viewpoint is substantial, especially for cluttered scenes, when the camera was close to the furniture.

For both databases, the environment was observed from multiple viewpoints. For INDECS, the viewpoints are stable over different weather conditions, but the appearance of the rooms is almost fully captured as the images were taken in 12 directions. In case of IDOL, we observe changes in viewpoint due to manual control of the robot, but since the robot was driven in a particular direction, parts of the environment might not be observed. As previously mentioned, labelling was based on the position of the camera rather than contents of the images, and acquisition was performed even close to walls or furniture. As a result, both databases contain difficult cases, where the contents of the image is either non-informative or is weakly associated with the label.

To summarize, despite the fact that the acquisition was performed in a relatively small environment (consisting of 5 different rooms), there are several types of variability captured which pose a challenge to a recognition system. These range from different acquisition conditions to large viewpoint variations across the devices. Moreover, the acquisition procedure was carefully designed, and each single dataset offers different, but usually well specified, type of variability. As a result, the influence of different factors on the accuracy of the system can be isolated and precisely measured. The relatively small environment does not allow for concluding that a system evaluated on the data will offer similar absolute performance in a different environment. However, since the data capture the influence of a large amount of variations on the appearance of a standard office environment, we can expect that an algorithm robust to those variations will be robust to similar types of variations within other indoor office environments.

5 Baseline Evaluation

This section presents the visual place recognition system with which we assessed the INDECS and IDOL databases. We applied a fully supervised, appearance-based method. It assumes that each room is represented, during training, by a collection of images capturing its visual appearance under different viewpoints, at a given time and illumination. During testing, the algorithm is shown images of the same rooms, acquired under roughly similar viewpoints but possibly under different illumination conditions and after some time (where the time range goes from some minutes to

several months). The goal is to recognize correctly each single image seen by the system. The method is based on a large-margin discriminative classifier, namely the Support Vector Machines (SVMs) [16] and two different image representations. We use global and local image features, and we combine them with SVMs through specialized kernels. As a result, the recognition process always consists of two steps: feature extraction and classification.

In the rest of this section, we first motivate the decision to provide a baseline evaluation with the presented datasets (Section 5.1). Then, we describe the employed image representations (Section 5.2) and the classifier (Section 5.3). Finally, we explain the procedure followed in our experiments (Section 5.4).

5.1 Motivation

An important component when providing the community with a new collection of data is to give a quantitative measure of how hard the database is. Benchmark databases have become a very popular tool in several research communities during the last years [25, 33], because they provide at the same time an instrument to develop new state of the art algorithms, and a way to call attention on a research topic. When a database is used for developing a new algorithm, it is extremely useful to be able to compare the obtained results with those obtained by some other established technique: this permits to understand what are the advantages of the new method over existing approaches. At the same time, presenting a new corpus together with a baseline evaluation helps the community to quickly identify the open challenges of the problem and therefore concentrate there their research efforts. While often the baseline evaluation consists of a newly developed method, very often it is a well known, off the shelf solution: again, the goal of a baseline evaluation is not that of presenting new theory, but to provide a quantitative evaluation of how challenging the new dataset is, coupled with a well defined experimental protocol.

The computer vision community has been traditionally very open to the introduction of publicly available databases [25, 33] and associated benchmark challenges [4]. These two tools, combined together, have heavily contributed to set the research agenda of the last years. The robotics community has recently started to acknowledge the value and power of such collections, as it is witnessed by several successful benchmark evaluations [5, 1].

5.2 Feature Extraction

The feature extraction step aims at providing a representation of the input data that minimizes the within-class variability while at the same time maximizing the between-class variability. Additionally, this representation is usually more compact than raw input data and therefore allows to reduce the computational load imposed by the classification process. Features can be derived from the whole image (global features) or can be computed locally, based on its salient parts (local features).

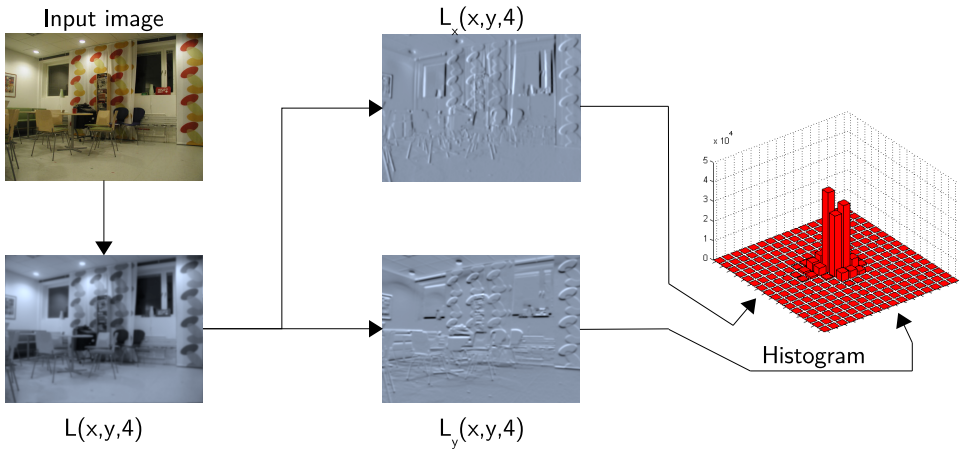


Figure 5: The process of generating multi-dimensional receptive field histograms shown on the example of the first-order derivatives computed at the same scale $t = 4$ from the illumination channel.

As environments can be described differently, depending on the considered scale, scale-space theory appears as a suitable framework for deriving effective representations here. Following this intuition, we chose to use two scale-space theory based features, one global (Composed Receptive Field Histograms, CRFH [30]) and one local (Scale Invariant Feature Transform, SIFT [31]). The rest of the section describes briefly the two approaches.

Global Features: Compose Receptive Field Histograms

CRFH is a multi-dimensional statistical representation of the occurrence of responses of several image descriptors applied to the image. This idea is illustrated in Fig. 5. Each dimension corresponds to one descriptor and the cells of the histogram count the pixels sharing similar responses of all descriptors. This approach allows to capture various properties of the image as well as relations that occur between them.

Multi-dimensional histograms can be extremely memory consuming and computationally expensive if the number of dimensions grows. For example, a 9-dimensional histogram with 16 quantization levels per dimension contains approximately $7 \cdot 10^{10}$ cells. In [30], Linde and Lindeberg suggest to exploit the fact that most of the cells are usually empty, and to store only those that are non-zero. The histogram can be stored in a sparse form as an array $[(c_1, v_1), (c_2, v_2), \ldots, (c_n, v_n)]$, where c_i denotes the index of the cell containing the non-zero value v_i. This representation allows not only to reduce the amount of memory required, but also

to perform operations such as histogram accumulation and comparison efficiently. For our experiments, we built multi-dimensional histograms using combinations of several image descriptors, applied to the scale-space representation at various scales, namely: first- and second-order Gaussian derivatives, gradient magnitude, Laplacian and Hessian determinant applied to both intensity and color channels.

Local Features: Scale Invariant Feature Transform

The idea behind *local features* is to represent the appearance of an image only around a set of characteristic points known as the *interest points*. The similarity between two images is then measured by solving the correspondence problem. Local features are known to be robust to occlusions, as the absence of some interest points does not affect the features extracted from other local patches.

The process of local feature extraction consists of two stages: *interest point detection* and *description*. The interest point detector identifies a set of characteristic points in the image that could be re-detected even in spite of various transformations (e.g. rotation and scaling) and variations in illumination conditions. The role of the descriptor is to extract robust features from the local patches located at the detected points.

In this paper, we used the scale, rotation, and translation invariant Harris-Laplace detector [36] and the commonly used SIFT descriptor [31]. Comparisons of local descriptors and interest point detectors, presented in [37], show that both algorithms are highly reliable. Moreover, the SIFT descriptor has shown to perform well for object classification ([19]) and mobile robot localization ([6, 20]).

5.3 Classification: Support Vector Machines

The choice of the classifier is the second key ingredient for an effective visual place recognition system. In this paper, we chose Support Vector Machines (SVMs) based on their state-of-the-art performances in several visual recognition domains [41, 13, 7]. The rest of this section reviews briefly the theory behind the algorithm, and describes our choices for the kernel function. We refer the readers to [16] for a thorough introduction to the subject.

Linear SVM

Consider the problem of separating a set of training data $(\boldsymbol{x}_1, y_1), ...(\boldsymbol{x}_m, y_m)$ into two classes, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label. Assuming that the two classes can be separated by a hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, then the optimal hyperplane will be the one with maximum distance to the closest points in the training set. The optimal values for $\boldsymbol{w}$ and b can be found by solving a constrained minimization problem via Lagrange multipliers, resulting in

a classification function

$$f(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \cdot \boldsymbol{x} + b\right), \tag{1}$$

where α_i and b can be found efficiently using the Sequential Minimal Optimization (SMO, [43]) algorithm. The $\boldsymbol{x}_i$ with $\alpha_i \neq 0$ are called *support vectors*.

Non-linear SVM and Kernel Functions

To obtain a nonlinear classifier, one maps the data from the input space $\Re^N$ to a higher dimensional feature space $\mathcal{H}$ by $\boldsymbol{x} \to \Phi(\boldsymbol{x}) \in \mathcal{H}$, such that the mapped data points of the two classes are linearly separable in the feature space. Assuming there exists a kernel function K such that $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$, a nonlinear SVM can be constructed by replacing the inner product $\boldsymbol{x}_i \cdot \boldsymbol{x}$ by the kernel function $K(\boldsymbol{x}_i, \boldsymbol{x})$ in Eqn. (1). This corresponds to constructing an optimal separating hyperplane in the feature space.

The choice of the kernel function is a key ingredient for the good performance of SVMs; based on results reported in the literature, we chose in this paper the χ^2 *kernel* [15] for global features and the *match kernel* [60] for local features.

The χ^2 kernel belongs to the family of exponential kernels, and is given by

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\gamma \chi^2(\boldsymbol{x}, \boldsymbol{y})\right\}, \qquad \chi^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \frac{||x_i - y_i||^2}{||x_i + y_i||}. \tag{2}$$

The match kernel is given by [60]

$$K(\boldsymbol{L}_h, \boldsymbol{L}_k) = \frac{1}{n_h} \sum_{j_h=1}^{n_h} \max_{j_k=1,\ldots,n_k} \left\{K_l(\boldsymbol{L}_h^{j_h}, \boldsymbol{L}_k^{j_k})\right\}, \tag{3}$$

where $\boldsymbol{L}_h, \boldsymbol{L}_k$ are local feature sets and $\boldsymbol{L}_h^{j_h}, \boldsymbol{L}_k^{j_k}$ are two single local features. The sum is always calculated over the smaller set of local features and only some fixed amount of best matches is considered in order to exclude outliers. The local feature similarity kernel K_l can be any Mercer kernel. We used the RBF kernel based on the Euclidean distance for the SIFT features:

$$K_l(\boldsymbol{L}_h^{j_h}, \boldsymbol{L}_k^{j_k}) = \exp\left\{-\gamma ||\boldsymbol{L}_h^{j_h} - \boldsymbol{L}_k^{j_k}||^2\right\}. \tag{4}$$

The match kernel was introduced in [60], and despite the claim in the paper, it is not a Mercer kernel [10]. Still, it can be shown that it statistically approximates a Mercer kernel in a way that makes it a suitable kernel for visual applications [10]. On the basis of this finding, and of its reported effectiveness for object categorization [41], we will use it here.

Multi-class SVM

The extension of SVM to multi class problems can be done mainly in two ways:

- *One-vs-all strategy.* If M is the number of classes, M SVMs are trained, each separating a single class from all remaining classes. The decision is then based on the distance of the classified sample to each hyperplane and the final output is the class corresponding to the hyperplane for which the distance is largest.
- *One-vs-one strategy.* In this case, $M(M-1)/2$ two-class machines are trained for each pair of classes. The final decision can then be taken in different ways, based on the $M(M-1)/2$ outputs. A popular choice is to consider as output of each classifier the class label and count votes for each class; the test image is then assigned to the class that received more votes. Another alternative is to use signed distance from the hyperplane and sum distances for each class. Other solutions based on the idea to arrange the pairwise classifiers in trees, where each tree node represents an SVM, have also been proposed [44, 16].

In this paper, for efficiency reasons, we will use the pairwise approach and the voting-based method, which we found to constantly outperform the second variant in preliminary experiments (the complexity of the SVM training algorithm is approximately $O(n^2)$ and smaller training subsets of the binary classifiers make the training procedure faster).

5.4 Experimental Setup

We conducted four series of experiments in order to assess thoroughly the INDECS and IDOL databases. For each series of experiments, we evaluated the performance of both local and global image representations. We divided the databases into several subsets with respect to the illumination conditions that prevailed during acquisition and the device employed. For the INDECS database, we considered three image sets, one for each illumination setting (cloudy, night, sunny). Since the IDOL database consists of 12 image sequences, we used each full sequence as a separate set. The system was always trained in a supervised fashion on one, two or three data sets and tested on a fourth different set. In order to test the limits of the underlying visual recognition algorithm, we considered each image in the test set separately, and as a final measure of performance, we used the percentage of properly recognized images. As the number of acquired images varied across rooms, the performance obtained for each place was considered separately during the experiments. The final classification rate was then computed as the average between all the rooms results. This procedure ensures that performance on each place contributes equally to the overall result, thus avoiding the biases towards areas with many acquired images, such as the corridor.

We started with a set of reference experiments, assessing the data acquired under stable illumination. To achieve this, for training and testing we used data

sets acquired with the same device and under similar conditions. Next, we increased the difficulty of the problem and tested the robustness of the system to changing illumination conditions as well as to other variations that may occur in real-world environments. Training and recognition were in this case performed on data sets consisting of images captured under different illumination settings and usually on different days. The third set of experiments aimed to reveal whether a model trained on an image set acquired with one device can be useful for solving localization problem with a different device (and usually after some time). Finally, we checked whether the robustness of the recognition algorithm can be increased by providing additional training data capturing a wider spectrum of visual variability. For that, we trained the system on two or three image sets gathered under different illumination conditions. Additionally, before carrying out the benchmarks described above, we conducted a set of preliminary experiments in order to select proper kernel functions and feature extractor parameters. All the results obtained on these experiments are reported in Section 6.

For all experiments, we used our extended implementation of Support Vector Machines based on the *libsvm* software [14]. We set the value of the error penalty C to be equal to 100 and we determined the kernel parameters via cross-validation.

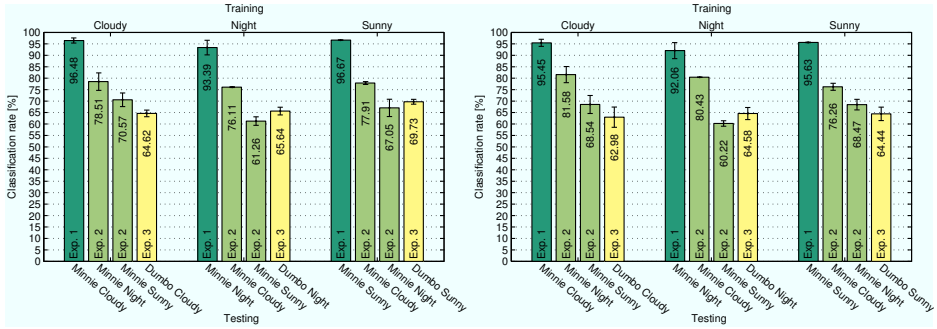
6 Experimental Results

This section reports the results of the baseline evaluation of the INDECS and IDOL databases, according to the procedure described in Section 5.4. We present the results in consecutive subsections, and we give a brief summary in Section 6.5.

As described in Section 5.4, before performing the actual benchmark, we ran a set of preliminary experiments on the INDECS database, mainly using the global features (CRFH). We evaluated the performance of the multi-dimensional histograms built from a wide variety of combinations of global image descriptors listed in Section 5.2 for several scale levels and numbers of histogram bins per dimension. A comprehensive report on the obtained results can be found in [46]. The experiments revealed that the most valuable global features can be extracted using non-isotropic, derivative-based descriptors, and that chromatic cues are more susceptible to illumination variations. As a result, here we used composed receptive field histograms of six dimensions with 28 bins per dimension, computed from second order normalized Gaussian derivative filters, applied to the illumination channel at two scales. The scale levels were different for the experiments with IDOL ($\sigma = 1$ and 4) and with INDECS ($\sigma = 2$ and 8). This was motivated by the fact that the cameras mounted on the robots obtained images of lower quality, and their movement introduced additional distortions.

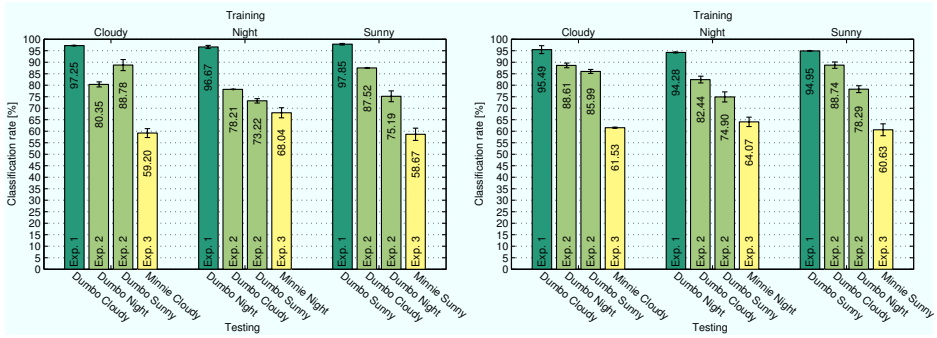
6.1 Stable Illumination Conditions

In order to evaluate our method under stable illumination conditions, we trained and tested the system on pairs of image sequences taken from the IDOL database



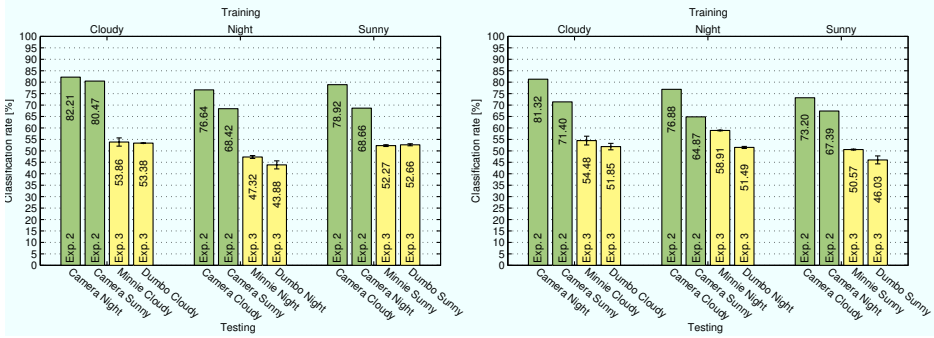
(a) Training on global features (*CRFH*) extracted from images acquired with *Minnie*.

(b) Training on local features (*SIFT*) extracted from images acquired with *Minnie*.



(c) Training on global features (*CRFH*) extracted from images acquired with *Dumbo*.

(d) Training on local features (*SIFT*) extracted from images acquired with *Dumbo*.



(e) Training on global features (*CRFH*) extracted from images acquired with the *standard camera*.

(f) Training on local features (*SIFT*) extracted from images acquired with the *standard camera*.

Figure 6: Average results of the first three experiments on the IDOL and INDECS databases with both image representations. In each figure, the results are grouped according to the type of illumination conditions under which the training images were acquired. The bottom axes indicate the platform and illumination conditions used for testing. The uncertainties are given as one standard deviation.

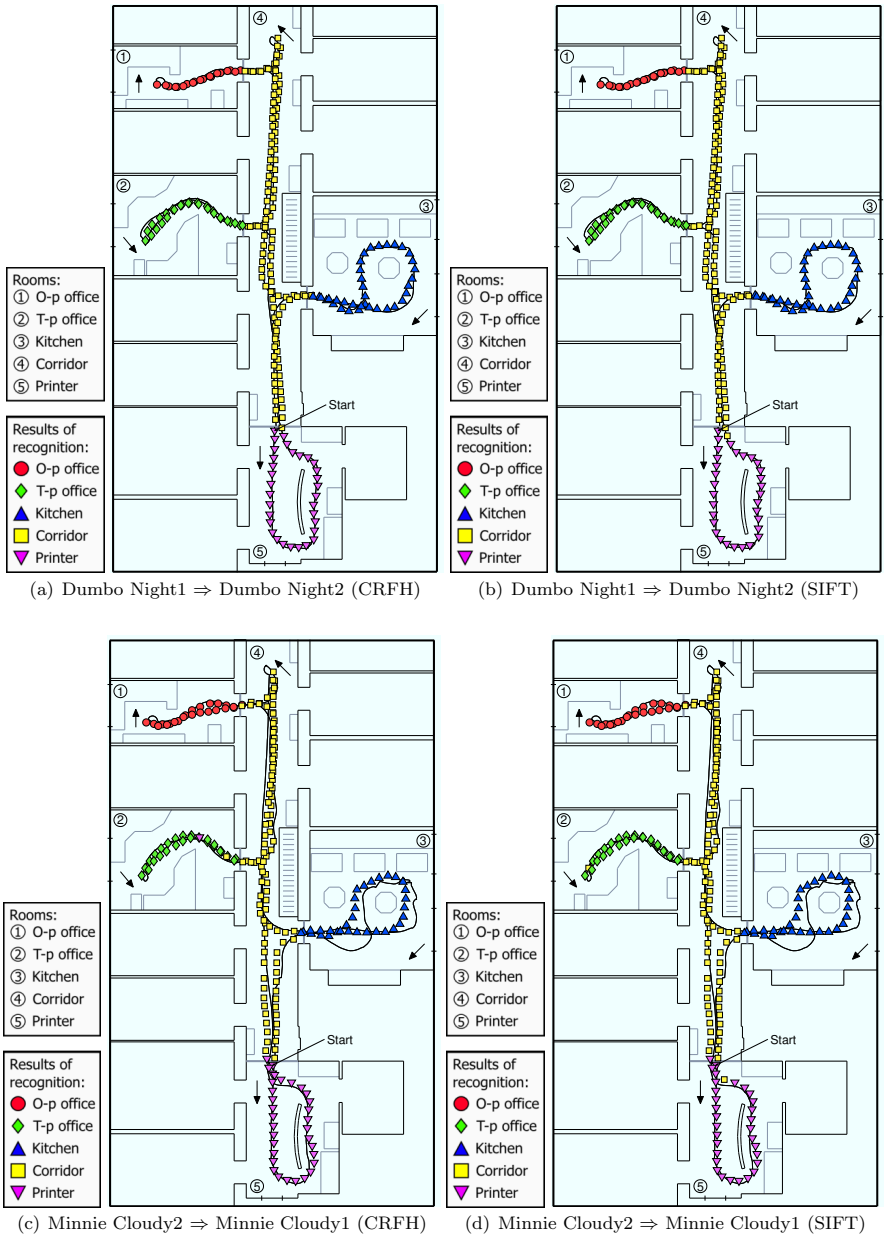
acquired one after the other using the same robot. As mentioned previously, we analyzed performance of both global (CRFH) and local (SIFT) image descriptors. We did not use the INDECS database for these experiments since only one set of data for each illumination setting was available. Although the illumination conditions for both training and test images were in this case very similar, the algorithm had to tackle other kinds of variability such as viewpoint changes (caused mainly by the manual control of the robot) and presence/absence of people. The results of the performed experiments are presented in Fig. 6a,c for CRFH and in Fig. 6b,d for SIFT. For each platform and type of illumination conditions used for training, the first bar presents an average classification rate over the two possible permutations of the image sequences in the training and test sets[1]. On average, the system classified properly 95.5% of the images acquired with Minnie and 97.3% of images acquired with Dumbo when global features were used. When local features were applied, the average recognition rates were slightly lower and equal to 94.4% and 94.9% respectively.

Detailed results for two experiments conducted on data captured with each of the platforms are shown in Fig. 7. The figure presents maps of the environment with plotted paths of the robot during acquisition of the training and test sequences used during each of the experiments. Moreover, the symbols used to draw the test path indicate the results of recognition performed using image acquired at each location. Each experiment started at the point marked with the label "Start" and the arrows show the direction of driving. The position of the furniture (plotted with gray line) is approximate and sometimes slightly varied between the experiments. It can be observed that the errors are usually not a result of viewpoint variations (compare the training and test paths in the kitchen, especially in Fig. 7c,d) and mostly occur near the borders of the rooms. This can be explained by the relatively narrow field of view of the cameras as well as the fact that the images were not labeled according to their content but to the position of the robot at the time of acquisition. Since these experiments were conducted with the sequences captured under similar conditions, we treat them as a reference for other results.

6.2 Varying Illumination Conditions

We also conducted a series of experiments aiming to test the robustness of our method to changing illumination conditions as well as to other variations caused by normal activities in the rooms. The experiments were conducted on both INDECS and IDOL databases. As with the previous experiments, the same device was used for both training and testing. This time, however, the selected training and testing data sets consisted of images acquired under different illumination conditions and usually on different days. Fig. 6a-d show average results of the experiments with the image sequences from the IDOL database acquired with both robots for each permutation of the illumination conditions used for training and testing and both

[1] Training on the first sequence, testing on the second sequence, and vice versa.



(a) Dumbo Night1 ⇒ Dumbo Night2 (CRFH)

(b) Dumbo Night1 ⇒ Dumbo Night2 (SIFT)

(c) Minnie Cloudy2 ⇒ Minnie Cloudy1 (CRFH)

(d) Minnie Cloudy2 ⇒ Minnie Cloudy1 (SIFT)

Figure 7: Maps of the environment with plotted paths of the robot during acquisition of the training (black line) and test (points) sequences taken from the IDOL database and used during the experiments with *stable illumination conditions*. The shape of each point on the test path indicates the result of recognition.

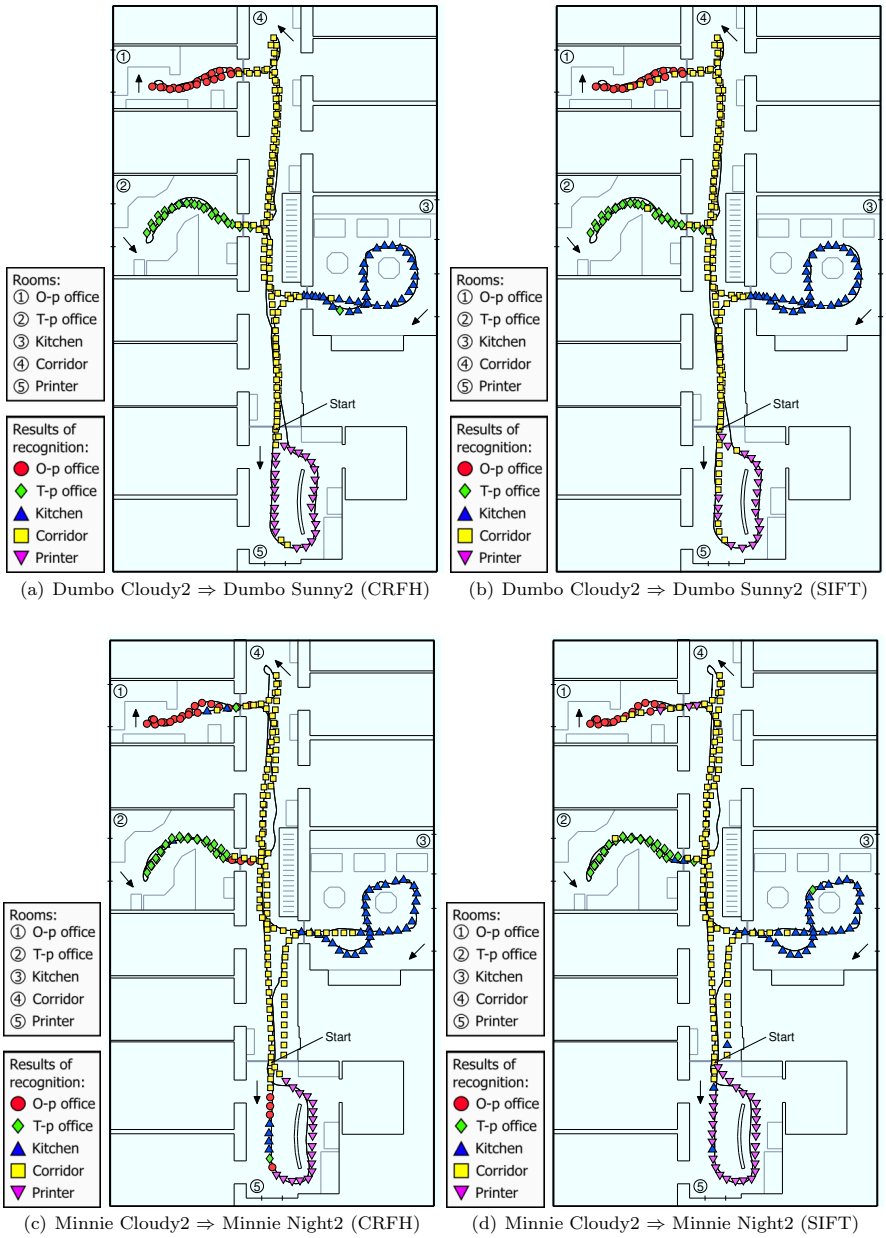
image representations (the two middle bars for each figure and type of training conditions). The presented classification rates obtained on the IDOL database were always averaged over two experiments with different image sequences. Fig. 6e,f gives corresponding results obtained on the INDECS database.

We see that, in general, the system performs best when trained on the images acquired in cloudy weather. The explanation for this is straightforward: the illumination conditions on a cloudy day can be seen as intermediate between those at night (only artificial light) and on a sunny day (direct natural light dominates). In such case, the average classification rate computed over two testing illumination conditions (sunny and night) for both CRFH and SIFT was equal to 84.6% and 87.3% for Dumbo, 74.5% and 75.1% for Minnie, and 81.3% and 76.4% for the INDECS database. In general, local features performed slightly better than the global features (in average 71.9% vs. 72.6% for Minnie and 80.5% vs. 83.2% for Dumbo), although it was usually not the case for the INDECS database (in average 75.9% vs. 72.5%). Fig. 8 presents detailed results for two example runs and both feature types. The errors occurred mainly for the same reasons as in the previous experiments and additionally in places heavily affected by the natural light e.g. when the camera was directed towards a bright window or, in particular, large glass door in the printer area. In such cases, the automatic exposure system with which all the cameras were equipped caused the pictures to darken. Minnie was more susceptible to this phenomenon due to the higher position of its camera.

6.3 Recognition Across Platforms

The third set of experiments was designed to test the portability of the acquired model across different platforms. For that purpose we trained and tested the system on image sets acquired under similar illumination conditions using different devices. We started with the experiments on image sequences from the IDOL database. We trained the system on the images acquired using either Minnie or Dumbo and tested with the images captured with the other robot. We conducted the experiments for all illumination conditions and both image representations. The main difference between the platforms from the point of view of our experiments lies in the height at which the cameras are mounted (98cm for Minnie and 36cm for Dumbo). The results presented in Fig. 6a-d indicate that our method was still able to classify correctly up to about 70% of images for CRFH and 65% of images for SIFT. There was no clear advantage of using one particular feature type. The system performed better when trained on the images captured with Minnie. This can be explained by the fact that the lower mounted camera on Dumbo provided less diagnostic information. It can also be observed from Fig. 9 that, in general, the additional errors occurred when the robot was positioned close to the walls or furniture. In such cases the height of the camera influenced the content of the images the most.

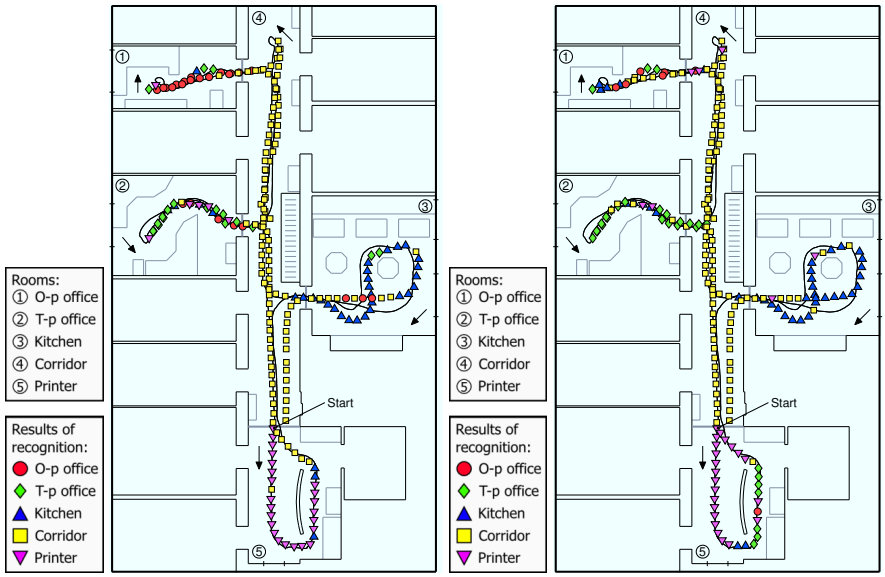
We followed a similar procedure using the INDECS database as a source of training data and different image sequences taken from the IDOL database for testing. It is important to note that the acquisition procedure differed in case of



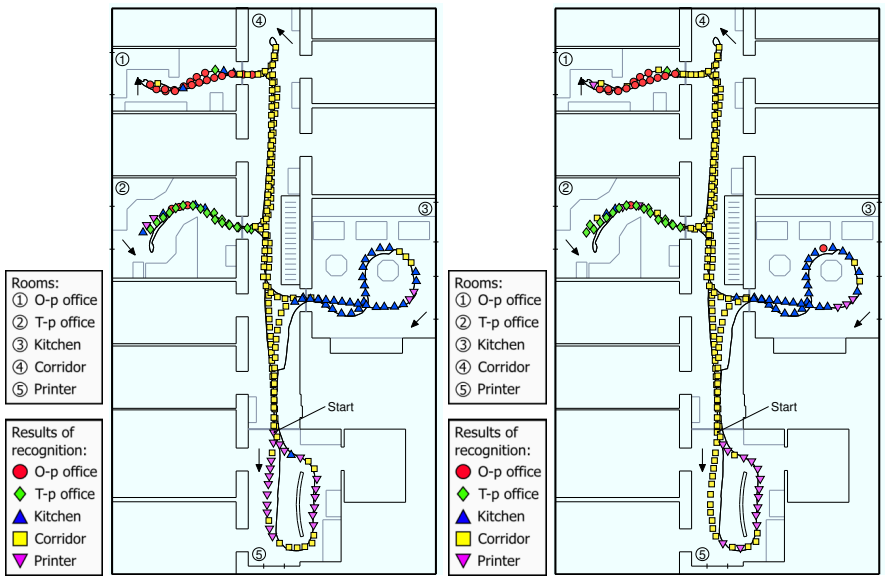
(a) Dumbo Cloudy2 ⇒ Dumbo Sunny2 (CRFH) (b) Dumbo Cloudy2 ⇒ Dumbo Sunny2 (SIFT)

(c) Minnie Cloudy2 ⇒ Minnie Night2 (CRFH) (d) Minnie Cloudy2 ⇒ Minnie Night2 (SIFT)

Figure 8: Maps of the environment with plotted paths of the robot during acquisition of the training (black line) and test (points) sequences taken from the IDOL database and used during the experiments with *varying illumination conditions*. The shape of each point on the test path indicates the result of recognition.



(a) Dumbo Night1 ⇒ Minnie Night2 (CRFH) (b) Dumbo Night1 ⇒ Minnie Night2 (SIFT)



(c) Minnie Sunny2 ⇒ Dumbo Sunny1 (CRFH) (d) Minnie Sunny2 ⇒ Dumbo Sunny1 (SIFT)

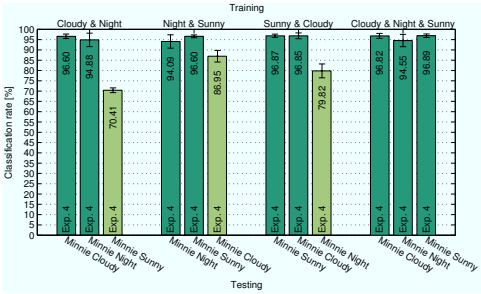
Figure 9: Maps of the environment with plotted paths of the robot during acquisition of the training (black line) and test (points) sequences taken from the IDOL database and used during the experiments with *recognition across platforms*. The shape of each point on the test path indicates the result of recognition.

both databases, and the INDECS database was gathered ten months before the acquisition of IDOL. The points at which the pictures were taken were positioned approximately 1m from each other and, in case of the kitchen, covered different area of the room due to reorganization of the furniture. Consequently, the problem required that the algorithm was invariant not only to various acquisition techniques but also offered great robustness to large changes in viewpoint and the appearance of the rooms introduced by long-time human activity. The experimental results are presented in Fig. 6e,f. We see that the algorithm obtains a recognition performance of about 50%. While this result is surely disappointing if compared to the 70% reported above, obtained for the two robot platforms, it is still quite remarkable considering the very high degree of variability between training and test data, and that results are significantly above chance (which in this case would be 20% as the datasets contain images acquired in 5 rooms).

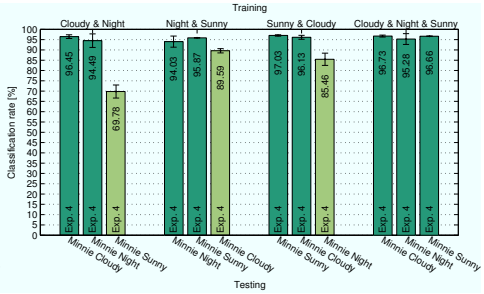
6.4 Training-based Robustness

The final series of experiments aimed at revealing whether the robustness of the recognition algorithm can be boosted by providing additional training data capturing a wider spectrum of visual variability that might occur in a real-world environment. In particular, we concentrated on invariance to changing illumination conditions as this is the kind of variability that a continuously running visual recognition system has to deal with every day. To achieve that, we trained the system on two or three image sequences from the IDOL database gathered under different illumination conditions, and we evaluated the recognition performance on another, fourth, image set. The obtained results for both platforms, all combinations of image sequences used for training as well as both CRFH and SIFT are presented in Fig. 10a-d. The darker bars indicate the results of experiments corresponding to those discussed in Section 6.1, when training was done on an image sequence acquired under conditions similar to those used for testing. The results shown using the brighter bars can be compared with those of the experiments under varying illumination conditions analyzed in Section 6.2.

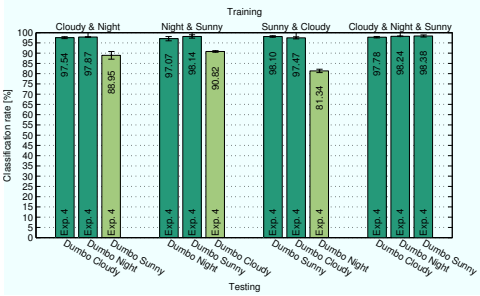
It is apparent that including images acquired under different conditions into the training set improves recognition accuracy. Although the algorithm has to incorporate much more information about each of the places into the model, the recognition accuracy for test sets acquired under similar conditions as those used for training is even greater than this obtained when each training sequence was used separately (as for the experiments discussed in Section 6.1). For example, the average recognition rate over all test sets and illumination settings for models trained on three sequences acquired using Dumbo was equal to 98.1% for CRFH and 97.1% for SIFT. At the same time, for the experiments with stable illumination conditions reported in Section 6.1 (see Fig. 6), we got only 97.3% and 94.9%. The same trend can be observed for sequences captured using Minnie. Concluding, the ability of the algorithm to handle large within-class variability is clearly not a limiting factor. It is important to note, that the recognition rate for conditions



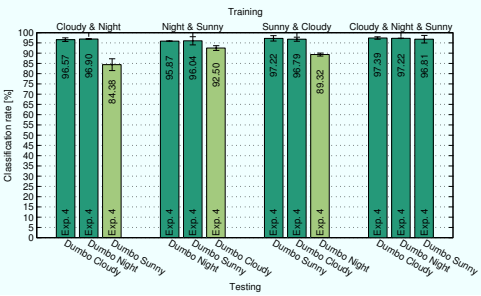
(a) Training on global features (*CRFH*) extracted from images acquired with *Minnie*.



(b) Training on local features (*SIFT*) extracted from images acquired with *Minnie*.



(c) Training on global features (*CRFH*) extracted from images acquired with *Dumbo*.



(d) Training on local features (*SIFT*) extracted from images acquired with *Dumbo*.

Figure 10: Performance of the system trained on two or three image sequences acquired under different illumination conditions for both mobile platforms and image representations. The classification rates were averaged over all possible combinations of training and test sequences. The uncertainties are given as one standard deviation.

which were not used during training is also greatly improved when more training data are provided. For example, if the system was trained using the images captured during sunny weather and at night using Minnie, the average classification rate for testing image sequence acquired with cloudy weather was equal to 86.95% for CRFH and 89.59% for SIFT. Consequently, the classification rate improved by 9.9% in case of CRFH and 11.2% in case of SIFT for testing conditions not known during training, at the same time slightly improving the rates for testing conditions used also for training.

It has to be pointed out that due to the larger number of training images capturing different types of variability, the number of support vectors stored in the final model grows as well. In such case, the user pays the price of the recognition time and the memory requirements, which in case of SVMs grow linearly with the number of support vectors.

6.5 Discussion

The results of the extensive experimental evaluation presented in this section indicate that our method is able to perform place recognition using standard visual sensors with high precision. It offers good robustness to changes in the illumination conditions as well as to additional variations introduced by the natural variability that occurs in real-world environments. At the same time, there is a difference in performance of the system between the experiments under stable and varying conditions, indicating that there is room for improvement in this matter.

As the system is to be used on a robot platform, it must not only be accurate but also efficient. For this reason we tried to provide the highest possible robustness using relatively small amount of training data acquired during only one run. We managed to achieve a recognition time of less than 200ms per frame on a Pentium IV 2.6 GHz using the global image representation. The results reported in Section 6.4 indicate that it is possible to significantly improve the robustness by incorporating images acquired during two or three runs under different illumination conditions into one training set. However, the higher performance does not come without a price. Since the number of support vectors in such case even doubles, the recognition time increased by about 50ms.

In all the experiments, we evaluated both global (CRFH) and local (SIFT) image descriptors. In general, we did not find any clear advantage of using one feature type over the other, and each representation has its strengths and weaknesses. The global features, however, clearly outperform SIFT in terms of efficiency, since the matching process required in order to compare two sets of local patches is computationally expensive. The efficiency of the solution based on local features could be improved by applying a more efficient matching algorithm (e.g. by using a pyramid match SVM kernel [24]) or faster interest point detector and more compact descriptor (e.g. SURF [8, 39]). Since global and local representations capture different aspects of a scene, the robustness of the final solution can be further improved by integrating both cues as proposed in [48, 49].

7 Summary

This paper discussed the need for standard benchmarking solutions for vision-based topological localization, with particular emphasis on visual place recognition. We defined and analyzed carefully the problem, and we specified the open challenges that need to be addressed by a realistic benchmark. We presented two new databases, acquired on the basis of this analysis. The first, the INDECS database, contains pictures captured with a standard camera mounted on a tripod. The second, the IDOL database, contains image sequences acquired using cameras mounted on two mobile robot platforms. The two databases were recorded within the same indoor office environment. They capture a wide spectrum of natural variations introduced by both changing illumination and human activity. Each database can be seen as a different approach to the problem; thus, they can be used

to analyze different properties of a place recognition system.

We assessed both databases with a large set of baseline experiments, using a fully supervised visual place recognition system. The method employs a large-margin discriminative classifier and two different image representations: a local representation, based on SIFT features, and a global representation, consisting of multidimensional histograms of receptive fields. We conducted the experiments according to an experimental procedure designed to contain problems of varying complexity and exploit most of the variability captured in the datasets. The experimental procedure can be seen as a part of the benchmark proposed in this paper. We started from experiments performed under stable illumination settings. We then performed experiments testing the robustness of the algorithms to changing illumination and human activity. Finally, we conducted experiments with large viewpoint variations and different acquisition methods.

The reported results show that the method is able to recognize places with high precision when training and testing is performed within a relatively stable environment, or when enough training data is provided. At the same time, there is space for improvement in the robustness to illumination and large viewpoint variations. The database still poses a challenge to the system which should provide stable performance in presence of variability usually observed in real-world environments.

Finally, the dependency between the overall performance of the system and the particular set of data becomes visible as the complexity of the problem grows. Moreover, different methods (in this case different image descriptors) perform differently for different types of variations. This emphasizes the need for an extensive experimental evaluation, on a common benchmark dataset, for comparison of different approaches. When realistic datasets are available, more extensive evaluation can be conducted as the data can be reused, fully exploited, and less effort is required for acquisition and annotation. The authors believe that benchmarking solutions, such as the one presented in this paper, will make an impact on the research on visual place recognition and topological localization as was the case for other localization and visual recognition problems.

Acknowledgment

This work was sponsored by the SSF through its Centre for Autonomous Systems (CAS), the EU integrated projects CoSy FP6-004250-IP, CogX ICT-215181 and DIRAC IST-027787 and the Swedish Research Council contract 2005-3600-Complex. The support is gratefully acknowledged.

References

[1] ImageCLEF Robot Vision Challenge. URL `http://www.robotvision.info`.

[2] The KTH-TIPS Image Database. URL `http://www.nada.kth.se/cvap/databases/kth-tips/`.

[3] The MIT-CSAIL Database of Objects and Scenes. URL `http://web.mit.edu/torralba/www/database.html`.

[4] The PASCAL Visual Object Classes Challenge. URL `http://www.pascal-network.org/challenges/VOC/`.

[5] The Semantic Robot Vision Challenge. URL `http://www.cs.cmu.edu/~srvc/`.

[6] Henrik Andreasson, André Treptow, and Tom Duckett. Localization for mobile robots using panoramic vision, local features and particle filter. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, 2005.

[7] Annalisa Barla, Francesca Odone, and Alessandro Verri. Histogram intersection kernel for image classification. In *Proceedings of the International Conference on Image Processing (ICIP'03)*, pages 513–516, Barcelona, Spain, 2003.

[8] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110 (3):346–359, June 2008.

[9] Paul Blaer and Peter Allen. Topological mobile robot localization using fast vision techniques. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation (ICRA'02)*, Washington, DC, USA, 2002.

[10] Sabri Boughorbel, Jean-Philippe Tarel, and François Fleuret. Non-Mercer kernels for SVM object recognition. In *Proceedings of the 15th British Machine Vision Conference (BMVC'04)*, London, England, 2004.

[11] David M. Bradley, Rashmi Patel, Nicolas Vandapel, and Scott M. Thayer. Real-time image-based topological localization in large outdoor environments. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, Edmonton, Alberta, Canada, 2005.

[12] Emma Brunskill, Thomas Kollar, and Nicholas Roy. Topological mapping using spectral clustering and classification. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 3491–3496, San Diego, CA, USA, October 2007.

[13] Barbara Caputo, Eric Hayman, and P. Mallikarjuna. Class-specific material categorisation. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, pages 1597–1604. Citeseer, 2005.

[14] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: A library for Support Vector Machines*, 2001. URL `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[15] Oliver Chapelle, Patrick Haffner, and Vladimir Vapnik. Support Vector Machines for Histogram-Based Image Classification. *IEEE Transactions on Neural Networks*, 10 (5), 1999.

[16] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. ISBN 0521780195.

[17] Mark Cummins and Paul M. Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research (IJRR)*, 27(6):647–665, 2008.

[18] Gamini Dissanayake, Paul M. Newman, Steven Clark, Hugh F. Durrant-Whyte, and M. Csorba. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241, June 2001.

[19] Gyuri Dorkó and Cordelia Schmid. Object class recognition using discriminative local features. 2005.

[20] David Filliat. A visual bag of words method for interactive qualitative localization and mapping. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[21] John Folkesson, Patric Jensfelt, and Henrik I. Christensen. Vision SLAM in the measurement subspace. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, 2005.

[22] Friedrich Fraundorfer, Christopher Engels, and David Nistér. Topological mapping, localization and navigation using image collections. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[23] José Gaspar, Niall Winters, and José Santos-Victor. Vision-based navigation and environmental representations with an omni-directional camera. *IEEE Transactions on Robotics and Automation*, 16(6):890–898, December 2000.

[24] Kristen Grauman and Trevor Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, pages 1458–1465, 2005. ISBN 0-7695-2334-X.

[25] Greg Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report 7694, Caltech, 2007.

[26] Andrew Howard and Nicholas Roy. The Robotics Data Set Repository (Radish), 2003. URL `http://radish.sourceforge.net/`.

[27] Matjaž Jogan and Aleš Leonardis. Robust localization using an omnidirectional appearance-based subspace model of environment. *Robotics and Autonomous Systems (RAS)*, 45(1):51–72, 2003.

[28] David M. Kortenkamp and Terry Weymouth. Topological mapping for mobile robots using a combination of sonar and vision sensing. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, Seattle, Washington, USA, 1994.

[29] Benjamin Kuipers and Patrick Beeson. Bootstrap learning for place recognition. *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI'02)*, pages 174–180, 2002.

[30] Oskar Linde and Tony Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proceedings of the 2004 15th International Conference on Pattern Recognition (ICPR'04)*, 2004.

[31] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[32] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. The KTH-IDOL2 database. Technical report, Kungliga Tekniska Högskolan, CVAP/CAS, September 2006.

[33] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, Miami Beach, FL, USA, 2009.

[34] Mario Mata, Jose M. Armingol, Arturo de la Escalera, and Miguel Angel Salichs. Using learned visual landmarks for intelligent topological navigation of mobile robots. In *Proceedings of the 2003 IEEE International Conference on Robotics and Automation (ICRA'03)*, 2003.

[35] Emanuele Menegatti, Mauro Zoccarato, Enrico Pagello, and Hiroshi Ishiguro. Image-based Monte-Carlo localisation with omnidirectional images. *Robotics and Autonomous Systems*, 48(1), 2004.

[36] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV'01)*, Vancouver, Canada, 2001.

[37] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, WI, USA, 2003.

[38] Oscar Martinez Mozos, Rudolph Triebel, Patric Jensfelt, Axel Rottmann, and Wolfram Burgard. Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems (RAS)*, 55(5):391–402, 2007.

[39] Ana Cris Murillo, José Jesús Guerrero, and Carlos Sagüés. SURF features for efficient robot localization with omnidirectional images. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[40] Eduardo Nebot. The Sydney Victoria Park Dataset. URL `http://www-personal.acfr.usyd.edu.au/nebot/dataset.htm`.

[41] Maria-Elena Nilsback and Barbara Caputo. Cue integration through discriminative accumulation. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, DC, USA, 2004.

[42] Illah Nourbakhsh, Rob Powers, and Stan Birchfield. Dervish: An office navigation robot. *AI Magazine*, 16(2):53–60, 1995.

[43] John C. Platt. *Fast training Support Vector Machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, November 1999.

[44] John C. Platt, Nello Cristianini, and John Shawe-Taylor. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing Systems (NIPS)*, volume 12, pages 547–553, 2000.

[45] Jean Ponce, Tamara L. Berg, Mark Everingham, David A. Forsyth, Martial Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan C. Russell, Antonio Torralba, Christopher K. I. Williams, Jianguo Zhang, and Andrew Zisserman. Dataset issues in object recognition. In *Towards Category-Level Object Recognition*, Lecture Notes in Computer Science, pages 29–48. Springer-Verlag, 2006.

[46] Andrzej Pronobis. *Indoor Place Recognition Using Support Vector Machines*. Master's thesis, Kungliga Tekniska Högskolan, Stockholm, Sweden, December 2005.

[47] Andrzej Pronobis and Barbara Caputo. The KTH-INDECS database. Technical report, Kungliga Tekniska Högskolan, CVAP/CAS, September 2005.

[48] Andrzej Pronobis and Barbara Caputo. Confidence-based cue integration for visual place recognition. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 2394–2401, San Diego, CA, USA, October 2007.

[49] Andrzej Pronobis, Oscar Martinez Mozos, and Barbara Caputo. SVM-based discriminative accumulation scheme for place recognition. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 522–529, Pasadena, CA, USA, May 2008.

[50] Christian Siagian and Laurent Itti. Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[51] Hashem Tamimi and Andreas Zell. Vision based localization of mobile robots using kernel approaches. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'04)*, Sendai, Japan, 2004.

[52] Adriana Tapus and Roland Siegwart. Incremental Robot Mapping with Fingerprints of Places. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Edmonton, Alberta, Canada, 2005.

[53] Sebastian Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99(1):21–71, 1998.

[54] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision (IJCV)*, 53(2), 2003.

[55] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, 2003.

[56] Antonio Torralba and Pawan Sinha. Recognizing Indoor Scenes. Technical Report 2001-015, MIT, July 2001.

[57] Muhammad Muneeb Ullah, Andrzej Pronobis, Barbara Caputo, Jie Luo, Patric Jensfelt, and Henrik I. Christensen. Towards robust place recognition for robot localization. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 530–537, Pasadena, CA, USA, May 2008. ISBN 978-1-4244-1646-2.

[58] Iwan Ulrich and Illah Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation (ICRA'00)*, San Francisco, CA, USA, 2000.

[59] Christoffer Valgren and Achim J. Lilienthal. Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 1856–1861, May 2008. ISBN 978-1-4244-1646-2.

[60] Christian Wallraven, Barbara Caputo, and Arnulf Graf. Recognition with local features: The kernel recipe. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, Beijing, China, 2003.

[61] Christian Weiss, Hashem Tamimi, Andreas Masselli, and Andreas Zell. A hybrid approach for vision-based outdoor robot localization using global and local image features. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[62] Jürgen Wolf, Wolfram Burgard, and Hans Burkhardt. Robust vision-based localization by combining an image retrieval system with Monte Carlo localization. *IEEE Transactions on Robotics*, 21(2):208–216, 2005.

[63] Hendrik Zender, Oscar Martinez Mozos, Patric Jensfelt, Geert-Jan M. Kruijff, and Wolfram Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems (RAS)*, 56(6):493–502, June 2008.

Paper B

The More you Learn, the Less you Store: Memory-controlled Incremental SVM for Visual Place Recognition

Andrzej Pronobis, Luo Jie and Barbara Caputo

Published in
Image and Vision Computing

The More you Learn, the Less you Store: Memory-controlled Incremental SVM for Visual Place Recognition

Andrzej Pronobis, Luo Jie and Barbara Caputo

Abstract

The capability to learn from experience is a key property for autonomous cognitive systems working in realistic settings. To this end, this paper presents an SVM-based algorithm, capable of learning model representations incrementally while keeping under control memory requirements. We combine an incremental extension of SVMs [43] with a method reducing the number of support vectors needed to build the decision function without any loss in performance [14] 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 to work 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. Experiments in both scenarios clearly show the power of our approach.

1 Introduction

Many recent advances in fields such as computer vision and robotics have been driven by the ultimate goal of creating artificial cognitive systems able to perform human-like tasks. Several attempts have been made to create integrated cognitive architectures and implement them, for instance, on mobile robots [3, 22, 1, 2]. The ability to learn and interpret complex sensory information based on the previous experience, inherently connected with cognition, has been recognized as crucial and vastly researched [41, 39, 32]. In most cases, the recognition systems used are trained offline, i.e. they are based on batch learning algorithms. However, in the real, dynamic world, learning cannot be a single act. It is simply not possible to create a static model which could explain all the variability observed over time.

Continuous information acquisition and exchange, coupled with an ongoing learning process, is necessary to provide a cognitive system with a valid world representation.

In artificial autonomous agents constrained by limited resources (such as mobile robots), continuous learning must be performed in an incremental fashion. It is obviously not feasible to rebuild the internal model from scratch every time new information arrives, neither it is possible to store all the previously acquired data for that purpose. The model must be updated and the updating process must have certain properties. First, the knowledge representation must remain compact and free from redundancy to fit into the limited memory and maintain a fixed computational complexity. We call this property *controlled memory growth*. Second, in the continuous learning scenario, a model cannot grow forever even though new information is constantly arriving. Thus, the updating process should be able to gradually filter out unnecessary information. We call this property *forgetting capability*.

Discriminative methods have become widely popular for visual recognition, achieving impressive results on several applications [47, 18, 13]. Within discriminative classifiers, SVM techniques provide powerful tools for learning models with good generalization capabilities; in some domains like object and material categorization, SVM-based algorithms are state of the art [7, 16]. This makes it worth it to investigate whether it is possible to perform continuous learning with this type of methods. Several incremental extensions of SVMs have been proposed in the machine learning community [12, 8, 43, 34]. Between these methods, the approximate techniques [12, 43] seem better suited for visual recognition because, at each incremental step, they discard non-informative training vectors, thus reducing the memory requirements. Other methods, such as [8, 34], instead require to store in memory all the training data, eventually leading to a memory explosion; this makes them unfit for real-time autonomous systems.

This paper presents an SVM-based incremental method which performs like the batch algorithm while reducing the memory requirements. We combine an approximate technique for incremental SVM [43] with an exact method that reduces the number of support vectors needed to build the decision function without any loss in performance [14]. This results in an algorithm performing as the original incremental method with a reduction in the memory requirements. We then present an extension of the method for the exact simplification of the support vector solution [14]. We introduce a parameter that links the performance of an SVM to the amount of vectors that is possible to discard. This allows a user-set trade-off between performance and memory reduction.

We evaluate the suitability of our method for autonomous cognitive systems in two challenging scenarios: adaptation in presence of dynamic changes and transfer of knowledge between autonomous agents. In both cases, we concentrate on the problem of visual place recognition applied to mobile robot topological localization. The problem is important from the point of view of engineering cognitive systems, as it allows to tie semantics with space representations and provides solutions for typical problems with purely metric localization. However, it is also a challenging

recognition problem as it requires processing of large amounts of high-dimensional visual information which is noisy and dynamic in nature. In this context, the memory and computational efficiency become one of the most important properties of the learning algorithm determining the design choice.

In our considerations, we first focus on the scenario in which the incremental learning is used to provide adaptability to different types of variations observed in real-world environments. In our previous work [38, 36], we presented a purely appearance-based model able to cope with illumination and pose changes, and we showed experimentally that it could achieve satisfactory performances when considering short time intervals between the acquisition of the training and testing data. Nevertheless, a room's appearance is doomed to change dramatically over time because it is used: chairs are pushed around, objects are taken in/out of drawers, furniture and paintings are added, or changed, or re-arranged; and so forth. As it is not possible to predict a priori how a room is going to change, the only possible strategy is to update the representation over time, learning incrementally from the new data recorded during use.

As a 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, which is a tabula rasa. The ability to transfer knowledge between different domains enables humans to learn efficiently from small number of examples. This observation inspired robotics and machine learning researchers to search for algorithms able to exploit prior knowledge so to improve performance of artificial learners and speed up the learning process. To tackle this problem, it is necessary an efficient way of exploiting the knowledge transferred from a different platform as well as updating the internal representation when new training data are available. The knowledge transfer scheme should be adaptive and privilege newest data so to prevent from accumulating outdated information. Finally, the solution obtained starting from a transferred model should gradually converge to the one learned from scratch, not only in terms of performance on a task but also of required resources (e.g. memory).

To achieve these goals, we used our memory-controlled incremental SVM and we evaluated its performance in terms of accuracy, memory growth, complexity and forgetting capability. We compare the results obtained by our method with those achieved by the batch algorithm and by two other incremental extensions of SVMs, one approximate (the fixed-partition incremental SVM, [43]) and one exact (online independent SVM, [34]). We evaluated the algorithms on a visual place recognition database acquired using two mobile robot platforms [38], which we extended with new data acquired 6 months later using the same hardware. Then, we confirmed the results on another database acquired in a different environment and using different hardware [37]. To test the adaptability of the recognition system, we performed topological localization experiments under realistic long-term variations. To test the knowledge transfer capabilities, we performed experiments in case of which visual knowledge captured in the SVM model was gradually exchanged between the two mobile robot platforms. The experiments clearly show the power of our

approach in both scenarios, while illustrating the need for incremental solutions in artificial cognitive systems.

The rest of the paper is organized as follows: after a review of related work (Section 2), Section 3 gives our working definition of visual place recognition for robot localization. Section 4 reviews SVMs, it introduces the memory-controlled incremental SVM algorithm, which will constitute a building block of the adaptive place recognition system and a base for our knowledge transfer technique, and it briefly describes two other incremental extensions of SVMs against which we will benchmark our approach. Section 5 describes our experimental setup; Section 7 concentrates on the adaptation problem and presents experimental evaluation of the algorithms in this context. Finally, Section 8 gives details of our approach to the transfer of knowledge and shows its effectiveness with a set of experiments. The paper concludes with a summary and possible directions for future work.

2 Related Work

In the last years, the need for solutions to such problems as robustness to long-term dynamic variations or transfer of knowledge is more and more acknowledged. In [39], the authors tried to deal with long-term visual variations in indoor environments by combining information acquired using two sensors of different characteristics. In [49], the problem of invariance to seasonal changes in appearance of an outdoor environment is addressed. Clearly, adaptability is a desirable property of a recognition system. At the same time, Thrun and Mitchell [46, 30] studied the issue of exchanging knowledge related to different tasks in the context of artificial neural networks and argued for the importance of knowledge-transfer schemes for lifelong robot learning. Several attempts to solve the problem have also been made from the perspective of Reinforcement Learning, including the case of transferring learned skills between different RL agents [28, 20].

The work conducted in the fields of cognitive robotics and vision stimulated the research in the machine learning community directed towards developing extensions for algorithms that were commonly used due to their superior performance but were missing the ability to be trained incrementally. As a result, methods such as Incremental PCA have been invented and successfully applied e.g. for mobile robot localization [4, 42]. As it was already mentioned, several incremental extensions have been introduced also for Support Vector Machines [12, 8, 43]. Between these methods, the approximate techniques[12, 43] seem better suited for visual recognition because, at each incremental step, they discard non-informative training vectors, thus reducing the memory requirements. Other methods, such as [8, 34], or simple KNN-based solutions, instead require to store in memory all the training data, eventually leading to a memory explosion. This limits their usefulness for complex real-world problems involving continuous learning of visual patterns.

Despite the fact that the approximate incremental SVM extensions allow to reduce the amount of data stored during the learning process, there is no guarantee

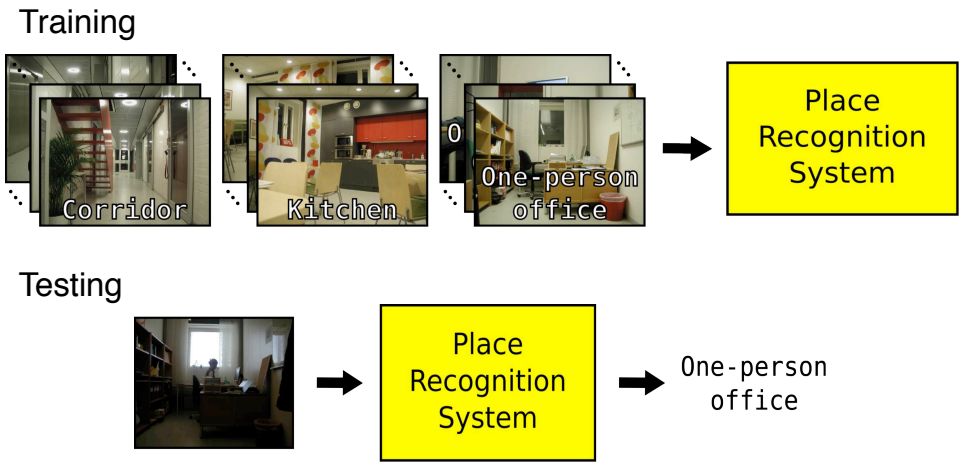


Figure 1: A schematic representation of our visual place recognition system for robot localization.

that the continuously updated model will not grow forever. Additionally, the results of experiments that can be found in the literature do not give a clear answer if it is possible to apply such methods for complex problems such as visual place recognition or transfer of visual knowledge.

3 Visual Place Recognition for Robot Localization

In this section, we give our working definition of visual place recognition, explaining how it can be applied to mobile robot topological localization. We define a place as a nameable segment of a real-world environment, uniquely identifiable because of its specific functionality and/or appearance. Examples of places, according to this definition, are a kitchen, an office, a corridor, and so forth. We adopt the appearance-based paradigm, and we assume that a realistic scene can be represented by a visual descriptor without any loss of discriminative information. We consider a fully supervised, incremental learning scenario: we assume that, at each incremental step, every room is represented by a collection of images which capture its visual appearance under different viewpoints, at a fixed time and illumination setting. During testing, the algorithm is presented with images of the same rooms, acquired under similar viewpoints but possibly under different illumination conditions and after some time, with a time range going from some minutes to several months. The goal is to recognize correctly each single image seen by the system. Fig. 1 illustrates the approach.

A typical application for an indoor place recognition system is topological robot localization. The localization problem is vastly researched. This resulted, over the

years, in a broad range of approaches spanning from purely metric [19, 11, 52, 15], to topological [48, 31, 39], and hybrid [45, 6]. Traditionally, sonar and/or laser have been the sensory modalities of choice [33, 31]. Yet, the inability to capture many aspects of complex realistic environments leads to the problem of perceptual aliasing [23], and greatly limits the usefulness of such methods for semantic mapping. Recent advances in vision have made this modality emerge as a natural and viable solution for localization problems. Vision provides richer sensory input allowing for better discrimination. It opens new possibilities for building cognitive systems, actively relying on semantic context. Not unimportant is the cost effectiveness, portability and popularity of visual sensors. As a result, despite the complexity of the problem, this research line is attracting more and more attention, and several methods have been proposed using vision alone [40, 48, 47, 36, 41], or combined with more traditional range sensors [21, 44, 39].

Our visual place recognition system uses SVM-based discriminative place models trained on global and local image features. These features are described in details in Section 5. The classification algorithm is introduced in Section 4. In our experiments, we always used only a single image as input for the recognition system. This makes the recognition problem harder, but also it makes it possible to perform global localization where no prior knowledge about the position is available (e.g. in case of the kidnapped robot problem). Spatial or temporal filtering can be used together with the presented method to enhance performance.

4 Memory-controlled Incremental SVM

This section describes our algorithmic approach to incremental learning of visual place models. We propose a fully supervised, SVM-based method with controlled memory growth that tends to privilege newest information over older data. This leads to a system able to adapt over time to the natural changes of a real-world setting, while maintaining a limited memory size and computational complexity.

The rest of this section describes the basic principles of Support Vector Machines (Section 4.1), a popular incremental extension of the basic algorithm (Section 4.2), our memory-controlled version of incremental SVM (Section 4.3) and an exact method based on a similar intuition (Section 4.4), with which we will compare our approach.

4.1 SVM: the batch algorithm

Consider the problem of separating the set of training data $(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_m, y_m)$ into two classes, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label (for multi-class extensions, we refer the reader to [10, 50]). If we assume that the two classes can be linearly separated when mapped to some higher dimensional Hilbert space $\mathcal{H}$ by $\boldsymbol{x} \rightarrow \Phi(\boldsymbol{x}) \in \mathcal{H}$ (see [10, 50] for solutions to non-separable cases), the optimal hyperplane is the one which has maximum distance to the closest points

in the training set, resulting in a classification function:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right), \tag{1}$$

where $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$ is the kernel function. Most of the α_i's take the value of zero; $\boldsymbol{x}_i$ with nonzero α_i are the Support Vectors (SV). Different kernel functions correspond to different similarity measures. Choosing a suitable kernel can therefore have a strong impact on the performance of the classifier. Based on results reported in the literature [38], here we used the two following kernels:

- The χ^2 kernel [5] for histogram-like global descriptors:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\{-\gamma\chi^2(\boldsymbol{x}, \boldsymbol{y})\}, \qquad \chi^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N} \frac{(x_i - y_i)^2}{x_i + y_i};$$

- The matching kernel [51] for local features:

$$K(\boldsymbol{L}_h, \boldsymbol{L}_k) = \frac{1}{n_h} \sum_{j_h=1}^{n_h} \max_{j_k=1,\ldots,n_k} \left\{ K_l(\boldsymbol{L}_h^{j_h}, \boldsymbol{L}_k^{j_k}) \right\},$$

  where $\boldsymbol{L}_h, \boldsymbol{L}_k$ are local feature sets and $\boldsymbol{L}_h^{j_h}, \boldsymbol{L}_k^{j_k}$ are two single local features. The sum is always calculated over the smaller set of local features and only some fixed amount of best matches is considered in order to exclude outliers. The local feature similarity kernel K_l can be any Mercer kernel. We used the RBF kernel based on the Euclidean distance for the SIFT [26] features:

$$K_l(\boldsymbol{L}_h^{j_h}, \boldsymbol{L}_k^{j_k}) = \exp\left\{-\gamma||\boldsymbol{L}_h^{j_h} - \boldsymbol{L}_k^{j_k}||^2\right\}.$$

4.2 SVM: an Incremental Extension

Among the incremental SVM extensions proposed so far [43, 12, 8], approximate methods seem to be the most suitable for visual recognition, because they discard a significant amount of the training data at each incremental step. Exact methods instead need to retain all training samples in order to preserve the convexity of the solution at each incremental step. As a consequence, they require huge amounts of memory when employed in realistic, continuous learning scenario as the one we consider here. Approximate methods avoid this problem by sacrificing the guaranteed optimality of the solution. Still, several studies showed that they generally achieve performances very similar to those obtained by an SVM trained on the complete data set (see [12] and references therein), because at each incremental step the algorithm remembers the essential class boundary information regarding the data seen so far (in form of support vectors). This information contributes properly to generate the classifier at the next iteration.

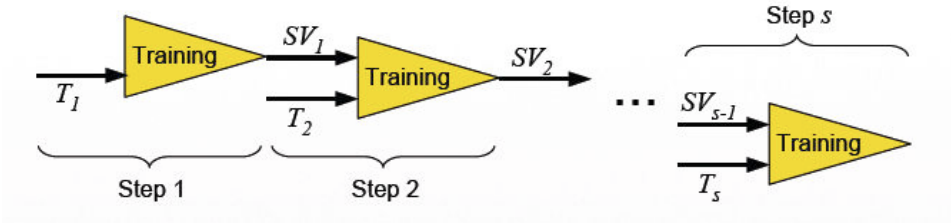


Figure 2: The fixed-partition incremental SVM algorithm.

Once a new batch of data is loaded into memory, there are different possibilities for performing the update of the current model, which might discard a part of the new data according to some fixed criteria [12, 43]. For all the techniques, at each step only the learned model from the data previously seen (preserved in form of SV) is kept in memory. In this paper we will consider the fixed-partition method [43]. Here the training data set is partitioned in batches of some size k:

$$\boldsymbol{T} = \{\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots \boldsymbol{T}_n\},$$

with $\boldsymbol{T}_i = \{(\boldsymbol{x}_j^i, y_j^i)\}_{j=1}^k$. At the first step, the model is trained on the first batch of data $\boldsymbol{T}_1$, obtaining a classification function

$$f_1(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{m_1} \alpha_i^1 y_i^1 K(\boldsymbol{x}_i^1, \boldsymbol{x}) + b^1\right). \tag{2}$$

At the second step, a new batch of data is loaded into memory and added to the current set of support vectors; then, the *new* training set becomes

$$\boldsymbol{T}_2^{inc} = \{\boldsymbol{T}_2 \cup \boldsymbol{SV}_1\}, \qquad \boldsymbol{SV}_1 = \{(\boldsymbol{x}_i^1, y_i^1)\}_{i=1}^{m_1},$$

where $\boldsymbol{SV}_1$ are the support vectors learned at the first step. The new classification function will be:

$$f_2(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{m_2} \alpha_i^2 y_i^2 K(\boldsymbol{x}_i^2, \boldsymbol{x}) + b^2\right).$$

Thus, as new batches of data points are loaded into memory, the existing support vector model is updated, so to generate the classifier at that incremental step. The method is illustrated in Fig. 2. Note that this incremental method can be seen as an approximation of the chunking technique used for training SVM [10, 50]. Indeed, the chunking algorithm is an exact decomposition which iterates through the training set to select the support vectors. The fixed-partition incremental method instead scan through the training data just once, and once discarded, does not consider them anymore. The fixed-partition incremental algorithm has been tested on several benchmark databases commonly used in the machine learning community

[12], obtaining good performances comparable to the batch algorithm and other approximate methods. An open issue is that in principle there is no limitation to the memory growth. Indeed, several experimental evaluations show that, while approximate methods generally achieve classification performances equivalent to those of batch SVM, the number of SV tends to grow proportionally to the number of incremental steps (see [12] and references therein).

4.3 Memory-controlled Incremental SVM

The core idea of the memory-controlled incremental SVM is that the set of support vectors $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^m$ in Eq. (1) is not guaranteed to be linearly independent. Based on this observation, it is possible to reduce the number of support vectors of a trained classifier, eliminating those which can be expressed as a linear combination of the others in the feature space, as proposed in [14] for reducing the complexity of the SVM solution. By updating the weights accordingly, it is ensured that the decision function is exactly the same as the original one. More specifically, let us suppose that the first r support vectors are linearly independent, and the remaining $m - r$ depend linearly on those in the feature space: $\forall j = r+1, \dots m$, $\boldsymbol{x}_j \in span\{\boldsymbol{x}_i\}_{i=1}^r$. Then it holds

$$K(\boldsymbol{x}, \boldsymbol{x}_j) = \sum_{i=1}^{r} c_{ij} K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{3}$$

and the classification function (1) can be rewritten as

$$f(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{r} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \sum_{j=r+1}^{m} \alpha_j y_j \sum_{i=1}^{r} c_{ij} K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right). \tag{4}$$

If we define the coefficients γ_{ij} such that $\alpha_j y_j c_{ij} = \alpha_i y_i \gamma_{ij}$ and $\gamma_i = \sum_{j=r+1}^{m} \gamma_{ij}$, then Eq. (4) can be written as

$$f(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{r} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \sum_{i=1}^{r} \alpha_i y_i \sum_{j=r+1}^{m} \gamma_{ij} K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right)$$

$$= \text{sgn}\left(\sum_{i=1}^{r} \alpha_i (1+\gamma_i) y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right) = \text{sgn}\left(\sum_{i=1}^{r} \widehat{\alpha}_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right), \tag{5}$$

where

$$\widehat{\alpha}_i = \alpha_i (1+\gamma_i) = \alpha_i \left(1 + \sum_{j=r+1}^{m} \frac{\alpha_j y_j c_{ij}}{\alpha_i y_i}\right).$$

The α_i coefficients can be pre-multiplied by the class labels $\alpha'_i = \alpha_i y_i$ which results in a simple equation that can be used to obtain the weights of the reduced classifier:

$$\widehat{\alpha}'_i = \begin{cases} \alpha'_i + \sum_{j=r+1}^{m} \alpha'_j c_{ij} & \text{for } i = 1, 2, \ldots, r \\ 0 & \text{for } i = r+1, r+2, \ldots, m. \end{cases} \tag{6}$$

Thus, the resulting classification function (Eq. (5)) requires now $m - r$ less kernel evaluations than the original one.

The linearly independent subset of the support vectors as well as the coefficients c_{ij} can be found by applying methods from linear algebra to the support vector matrix given by

$$\boldsymbol{K} = \begin{bmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_1, \boldsymbol{x}_m) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_m, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_m, \boldsymbol{x}_m) \end{bmatrix}, \tag{7}$$

We employ the QR factorization with column pivoting [17] for this purpose. The QR factorization with column pivoting algorithm is a widely used method for selecting the independent columns of a matrix. The algorithm allows to reveal the numerical rank of the matrix with respect to a parameter τ, which acts as a threshold in defining the condition of linear dependence. Additionally, it performs a permutation of the columns of the matrix so that they are ordered according to the degree of their relative linear independence. Consequently, if for a given value of τ the rank of the matrix is r, then the linearly independent columns will occupy the first r positions.

The QR factorization with column pivoting of the matrix $\boldsymbol{K} \in \Re^{m \times m}$ is given by

$$\boldsymbol{K}\boldsymbol{\Pi} = \boldsymbol{Q}\boldsymbol{R}, \tag{8}$$

where $\boldsymbol{\Pi} \in \Re^{m \times m}$ is a permutation matrix, $\boldsymbol{Q} \in \Re^{m \times m}$ is orthogonal, and $\boldsymbol{R} \in \Re^{m \times m}$ is upper triangular. If we assume that the rank of the matrix K with respect to the parameter τ equals r, then the matrices can be decomposed as follows:

$$\begin{bmatrix} \boldsymbol{K}_1 & \boldsymbol{K}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{0} & \boldsymbol{R}_{22} \end{bmatrix}, \tag{9}$$

where the columns of $\boldsymbol{K}_1 \in \Re^{m \times r}$ create a linearly independent set, the columns of $\boldsymbol{K}_2 \in \Re^{m \times m-r}$ may be expressed as a linear combination of the columns of $\boldsymbol{K}_1$, $\boldsymbol{Q}_1 \in \Re^{m \times r}$, $\boldsymbol{Q}_2 \in \Re^{m \times m-r}$, $\boldsymbol{R}_{11} \in \Re^{r \times r}$, $\boldsymbol{R}_{12} \in \Re^{r \times m-r}$, $\boldsymbol{R}_{22} \in \Re^{m-r \times m-r}$.

The products of the QR factorization can be used to obtain the coefficients c_{ij} as follows

$$\boldsymbol{C} = \begin{bmatrix} c_{1,r+1} & \cdots & c_{1,m} \\ \vdots & \ddots & \vdots \\ c_{r,r+1} & \cdots & c_{r,m} \end{bmatrix} = \boldsymbol{R}_{11}^{-1} \boldsymbol{Q}_1^T \boldsymbol{K}_2. \tag{10}$$

The coefficients together with the permutation matrix $\boldsymbol{\Pi} \in \Re^{m \times m}$ and the number of the linearly independent support vectors r are sufficient to obtain the reduced

solution. Using matrix notation, Eq. (6) can be expressed as follows

$$\begin{cases} \widehat{\boldsymbol{\alpha}}'_1 = \boldsymbol{\alpha}'_1 + \boldsymbol{R}_{11}^{-1}\boldsymbol{Q}_1^T\boldsymbol{K}_2\boldsymbol{\alpha}'_2 \\ \widehat{\boldsymbol{\alpha}}'_2 = \mathbf{0} \end{cases} \tag{11}$$

The rank r of the matrix $\boldsymbol{K}$ can be estimated by thresholding $\|\boldsymbol{R}_{22}\|_2$ with the value of the parameter τ. This means that, in practice, the choice of the τ value determines the number of linearly independent support vectors retained by the algorithm. For instance, by choosing a value of τ of 0.1 one will select a number of linearly independent support vectors smaller than by choosing a τ value of 0.01. This has two concrete effects on the algorithm:

1. As the value of τ increases, the number of support vectors decreases. This means that, by tuning τ, it is possible to reduce the memory requirements and to increase speed during classification;

2. At the same time, as τ increases, Eq. (5) will become more and more an approximation of the exact solution, because we are considering as linearly dependent vectors that are not. Therefore, we are not able to preserve fully their informative content. Still, we don't lose all the information carried by the discarded support vector $\mathbf{x}_j$, as its weight α_j is used to compute the updated value of the weights $\widehat{\alpha_i}$ for the remaining support vectors. This should result in a graceful decrease of classification performance compared to the optimal solution.

We propose to combine this model simplification with the fixed-partition incremental algorithm, adding the reduction process at each incremental step. We call the new algorithm memory-controlled incremental SVM. It can be illustrated as follows:

1. **Train**. The algorithm receives the first batch of data $\boldsymbol{T}_1$. It trains an SVM and obtains a set of support vectors $\boldsymbol{SV}_1$.

2. **Find linearly dependent SVs**. The algorithm finds permutation of $\boldsymbol{SV}_1$ that orders the SVs according to the degree of their linear independence.

3. **Find** τ. The algorithm searches for the value of τ, $\tau^\star$, that satisfies certain requirements regarding the number of support vectors or estimated performance of the classifier.

4. **Reduce**. The algorithm computes the reduced solution determined by the chosen $\tau^\star$. After this step, the reduced model contains a subset of the original SVs, $\widehat{\boldsymbol{SV}_1} = red(\boldsymbol{SV}_1)$, and can be used to classify test data.

5. **Retrain**. As the new batch of data $\boldsymbol{T}_2$ arrives, step (1) is repeated using as training vectors $\widehat{\boldsymbol{T}}_2^{inc} = \{\boldsymbol{T}_2 \cup \widehat{\boldsymbol{SV}_1}\}$.

For applications that require speed and/or have limited memory requirements, at step (3) of the algorithm, one can tune τ so to obtain at each incremental step a predefined maximum number of stored SV. For applications where accuracy is more relevant, one can estimate at each incremental step the τ corresponding to a pre-defined maximum decrease in performance. This can be done on the batch of data $\boldsymbol{T}_i$ at each step, dividing $\boldsymbol{T}_i$ in two subsets and training on one and testing on the other or by applying the leave-one-out strategy. We denote with the symbol Θ the percentage of the original classification rate that is guaranteed to be preserved after the reduction in this case.

In order to apply the method to multi-class problems, we used the one-vs-one multi-class extension. In a set of preliminary experiments comparing the one-vs-one and one-vs-all algorithms, we did not observe significant differences in the behavior of both methods (for further details, we refer the reader to [35]). The one-vs-one algorithm, given M classes, trains $M(M-1)/2$ two-class SVMs, one for each pair of classes. In case of the place recognition experiments, this method obtained smaller training times due to large number of training samples and relatively small number of classes.

4.4 Online Independent Incremental SVM

The idea to exploit the linear independence in the feature space has also been implemented in an online extension of SVMs, called Online Independent Support Vector Machine (OISVM, [34]). OISVM selects incrementally basis vectors that are used to build the solution of the SVM training problem, based upon linear independence in the feature space. Vectors that are linearly dependent on already stored ones are rejected. An incremental minimization algorithm is employed to find the new minimum of the cost function. This approach reduces considerably the complexity of the solution and therefore the testing time. As OISVM is an exact method, it requires to store all data acquired by the system during its whole life span for the update of the cost function. In many cases (e.g. in case of place recognition), the data samples are multi-dimensional and require a substantial amount of storage. Additionally, the learning algorithm needs to build a gram matrix the size of which is quadratic in the number of training samples. This leads inevitably to a memory explosion when the number of incremental steps grows, as we will show experimentally. Through its heuristics, the memory-controlled algorithm allows to decrease the number of training data samples at each incremental step and thus reduce the memory consumption.

5 Experimental Setup

This section describes our experimental setup. We first describe the IDOL2 and COLD-Freiburg databases, on which we will run all the experiments reported in this paper (Sections 5.1 and 5.2), then we briefly describe the feature representa-

Figure 3: Robot platforms employed in the experiments with the IDOL2 database and images illustrating the appearance of the five rooms from the robots' the point of view.

tions used in the experiments (Section 5.3). Finally, we discuss the performance evaluation measure and parameter selection method (Section 5.4).

5.1 The IDOL2 Database

The IDOL2 (Image Database for rObot Localization 2, [27]) database contains 24 image sequences acquired by a perspective camera, mounted on two mobile robot platforms. Both mobile robot platforms, the PeopleBot Minnie and the PowerBot Dumbo, are equipped with cameras. On Minnie the camera is located 98cm above the floor, whereas on Dumbo its height is 36cm. Fig. 3 shows both robots and some sample images from the database acquired by the robots from very close viewpoints, illustrating the difference in visual content. These images were acquired under the same illumination conditions and within short time spans.

The robots were manually driven through an indoor laboratory environment and the images were acquired at a rate of 5fps. Each image sequence consists of 800-1100 frames automatically labeled with one of five different classes (Printer Area [PA], CoRridor [CR], KiTchen [KT], Two-persons Office [TO], and One-person Office [OO]). The labeling is based on the camera's position given by the laser-based localization system proposed in [15]. The acquisition procedure was repeated several times to capture the changes in illumination and varying weather conditions (sunny, cloudy, and night). Also, special care was taken to capture people's activities, change of location for objects and for furniture; for part of the environment (two-persons office) we were able to record a significant change in decoration which occurred over a time span of 6 months. Fig. 4 shows some sample images from the database, illustrating these variations. It is important to note that each single sequence captures the appearance of the considered experimental environment under stable illumination settings and during the short span of time that is required to drive the robot manually around the environment.

The 24 image sequences are divided as follows: for each robot platform and

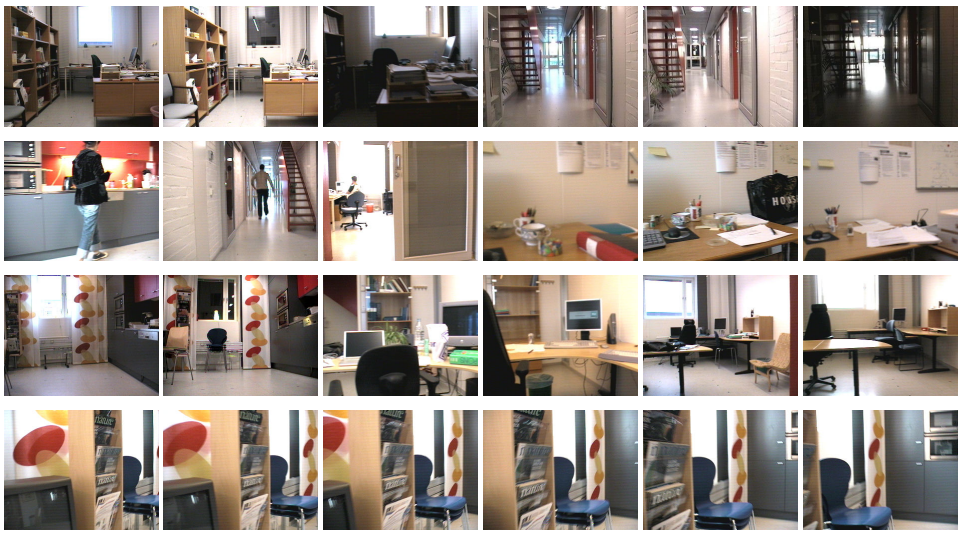


Figure 4: Sample images illustrating the variations captured in the IDOL2 database. Images in the top row show the variability introduced by changes in illumination for two rooms. The second and third rows show people appearing in the environment (first three images, second row) as well as the influence of people's activity including some larger variations which happened over a time span of 6 months. Finally, the bottom row illustrates the changes in viewpoint observed for a series of images acquired one after another in 1.2 second.

for each type of illumination conditions (cloudy, sunny, night), there are four sequences recorded. Of these four sequences, the first two were acquired six months before the last two. This means that, for every robot we always have subsets of sequences acquired under similar conditions and close in time, as well as subsets acquired under different conditions and distant in time. This makes the database useful for several types of experiments. It is important to note that, even for the sequences acquired within a short time span, variations still exist from everyday activities and viewpoint differences during acquisition. For further details, we refer the reader to [27].

5.2 The COLD-Freiburg Database

The COLD-Freiburg database is a collection of image sequences acquired at the Autonomous Intelligent System Laboratory at the University of Freiburg and constitutes a part of the COsy Localization Database (COLD, [37]). The acquisition procedure of the COLD-Freiburg database was similar to that of the IDOL2 database. Image sequences were acquired using a mobile robot platform, under

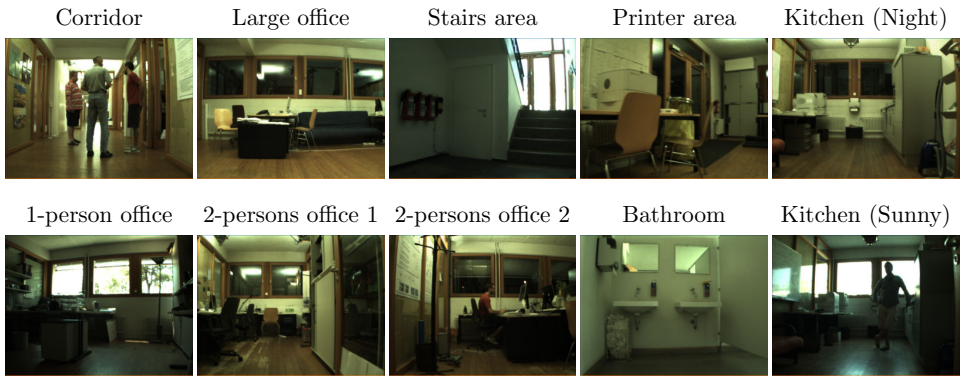


Figure 5: Sample images from the COLD-Freiburg database illustrating the rooms in which acquisition was performed and different types of captured variability introduced by human activity and changes in illumination.

several illumination conditions (sunny, cloudy, night) and across several days. As in case of IDOL2, special care was taken to capture people's activities and change of location of objects and furniture. However, the acquisition was performed using both perspective and omnidirectional cameras, in several parts of a different environment and using different hardware. For further details, we refer the reader to [37].

For our experiments, we employed only the perspective images and we selected 6 different extended sequences from the database. The extended sequences were acquired in a larger section of the environment consisting of 9 rooms of different functionality: a corridor, a printer area, a kitchen, a large office, 2 two-persons offices, a one-person office, a bathroom and a stairs area. The sequences contained on average 2547 frames. The 6 sequences were selected to mimic the organization of the IDOL2 database. For each illumination setting, we chose 2 sequences acquired under similar conditions and close in time.

5.3 Image Descriptors

Two visual descriptors, global and local, were employed during our experiments. We used Composed Receptive Field Histograms (CRFH, [25]) as global features. CRFHs are a multi-dimensional statistical representation of the occurrence of responses of several image descriptors applied to the image. This idea is illustrated in Fig. 6. Each dimension corresponds to one descriptor and the cells of the histogram count the pixels sharing similar responses of all descriptors. This approach allows to capture various properties of the image as well as relations that occur between them. Multi-dimensional histograms can be extremely memory consuming and computationally expensive if the number of dimensions grows. In [25], Linde

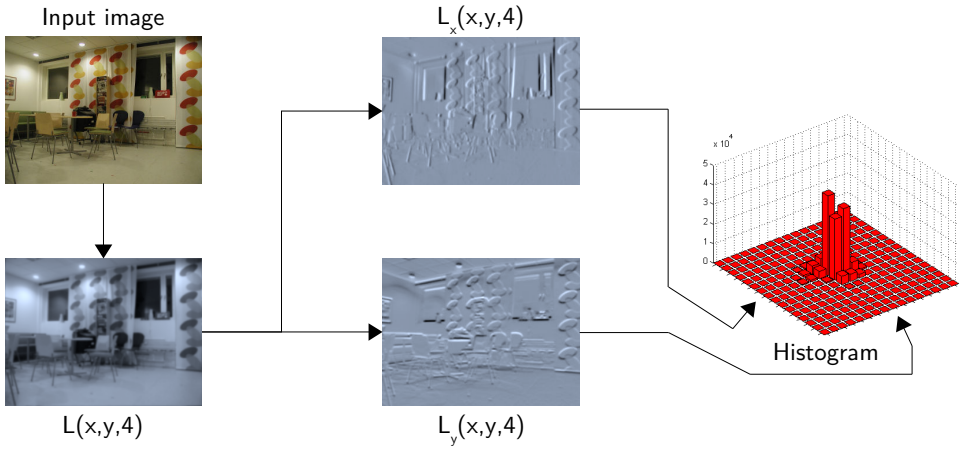


Figure 6: The process of generating multi-dimensional receptive field histograms using the first-order derivatives computed at the scale $t = 4$ and the number of bins per dimension set to 16.

Figure 7: Examples of images marked with interest points detected using the Harris-Laplace detector. The radius of the circles illustrate the scale at which the points were detected.

and Lindeberg suggest to exploit the fact that most of the cells are usually empty, and to store only those that are non-zero. This representation allows not only to reduce the amount of memory required, but also to perform operations such as histogram accumulation and comparison efficiently.

The idea behind local features is to represent the appearance of an image only around a set of characteristic points known as the interest points. The similarity between two images is then measured by solving the correspondence problem. Local features are known to be robust to occlusions and viewpoint changes, as the absence of some interest points does not affect the features extracted from other

local patches. The process of local feature extraction consists of two stages: *interest point detection* and *description*. The interest point detector identifies a set of characteristic points in the image that could be re-detected even in spite of various transformations (e.g. rotation and scaling) and variations in illumination conditions. The role of the descriptor is to extract robust features from the local patches located at the detected points. In this paper, we used the scale, rotation, and translation invariant Harris-Laplace detector [29] and the SIFT descriptor [26]. Fig. 7 shows two examples of interest point detected on images of indoor environments.

5.4 Parameter Selection and Performance Evaluation

For all experiments, the kernel parameter and the SVM cost parameter C were determined via cross validation, separately for each database. Then, the obtained values were used as constants for all the incremental learning experiments. For all experiments, we used the implementation of SVM provided by the *libsvm* library [9].

Since the employed datasets are unbalanced (e.g. in case of the IDOL2 database there are on average 443 samples for CR, 114 for 1pO, 129 for 2pO, 133 for KT and 135 for PR), as a measure of performance for the reported results and parameter selection, we used the average of classification rates obtained separately for each actual class. For each single experiment, the percentage of properly classified samples was first calculated separately for each room and then averaged with equal weights independently of the number of samples acquired in the room. This allowed to eliminate the influence that large classes could have on the performance score.

In our experiments, we observed a few percent improvement of the final results when a performance measure that is not invariant to unbalanced classes was used. This was caused by very good performance of the system for the corridor class. The was visually distinct from the other classes and was represented by the largest number of samples. As a result, in our experiments, the measure was used mainly to compensate for the influence of the corridor class.

6 Experiments on Support Vector Reduction

To begin with, we run some experiments to evaluate the behavior of the support vector reduction algorithm described in Section 4.3. We used two sequences from the IDOL2 database [27], one as train set and the other as test set. We chose CRFH as an image descriptor, and trained SVMs with four different types of kernels: linear kernel, RBF kernel, χ^2 kernel and histogram intersection (Hist.-Inte.) kernel. First, the SVM classifier was trained using the SMO algorithm. Then, starting from the obtained discriminative function, the reduction algorithm was tested, for different values of the reduction threshold τ. After each experiment (for each value of τ), the original model was reduced and the number of kept support vectors and the performance of the reduced model were tested on the same test set. If the classification rate dropped below 80% of the initial classification rate, i.e. $\Theta < 80\%$, the process was stopped. Fig. 8 reports the percentage of the reduced number of

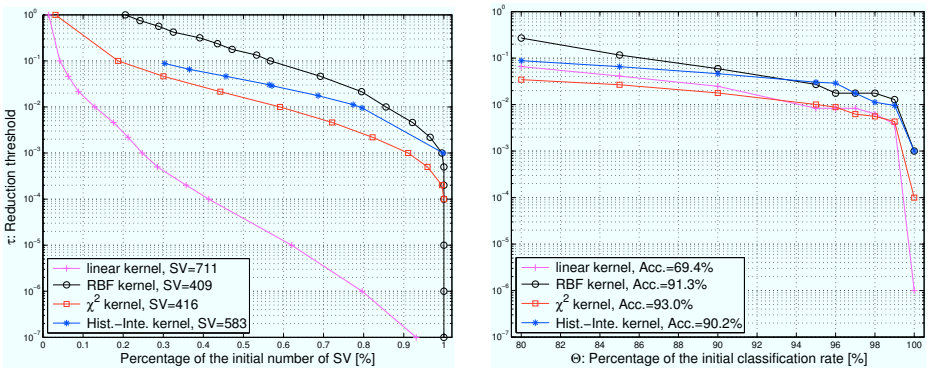


Figure 8: Percentage of the reduced number of Support Vectors (SV) compared to the initial model (left), and the percentage of the original classification rate that is preserved after the reduction (right), both as a function of different value of τ for various kernel types. The initial number of Support vectors (SV) and initial classification rate (Acc.) were reported for each kernel.

Support Vectors (SV) compared to the initial model (left), and the percentage of the initial classification rate that is preserved after the reduction (right), as a function of different value of τ. We see that, apart for the linear kernel, the algorithm behaves as expected, obtaining a gentle decrease in performance as the number of stored support vectors is being reduced. It is worth noting that the linear kernel is known for being not a good metric for histogram-like features, as instead all the other three kernels are. This might explain its different behavior.

7 Experiments on Adaptation

As a first application of our method, we present experiments on visual place recognition in highly dynamic indoor environments. We consider a realistic scenario, where places change their visual appearance because of varying illumination conditions or human activity. Specifically, we focus on the ability of the recognition algorithm to adapt to these changes over long periods of time. As it is not possible to predict in advance the type of changes that will occur, adaptation must be performed incrementally.

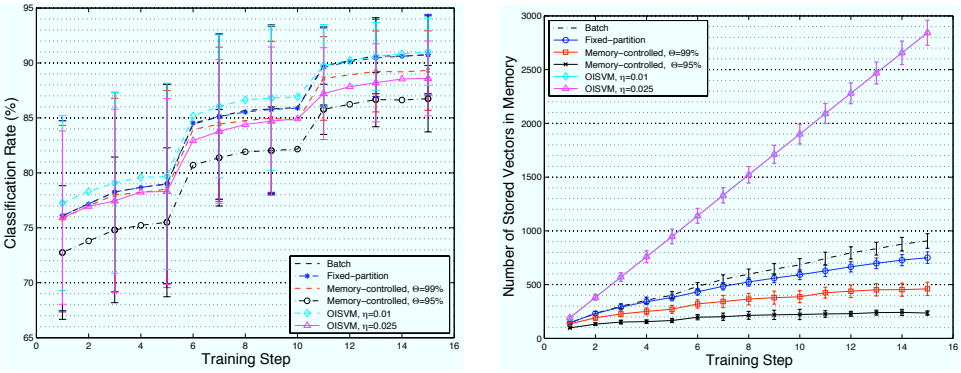
We conducted two series of experiments to evaluate the effectiveness of the memory-controlled incremental SVM for this task. In the first, we considered a case in which the variability observed by the recognition system was *constrained* to changes introduced by long-term human activity under stable illumination conditions. Such experimental procedure allowed us to thoroughly examine the properties of each of the incremental methods in a more controlled setting. The corresponding experiments are reported in Section 7.1. In the second, we considered a real-world,

unconstrained scenario where the algorithms had to incrementally gain robustness to variations introduced by changing illumination and short-term human activity, and then, to use their adaptation abilities to handle long-time environment changes. The corresponding experiments are reported in Section 7.2. In both experiments, we compared our approach with the fixed-partition incremental SVM, OISVM and the batch method. This last algorithm is used here purely as a reference, as it is not incremental. We used CRFH global image features. We tested a wide variety of combinations of image descriptors, with several scale levels [35]. On the basis of an evaluation of performance and computational cost, we built the histograms from normalized Gaussian derivative filters applied to the images at two different scales, and we used χ^2 as a kernel for SVM. We also performed experiments using SIFT local features combined with the matching kernel for SVM. Both types of features previously proved effective for the place recognition task [39, 38].

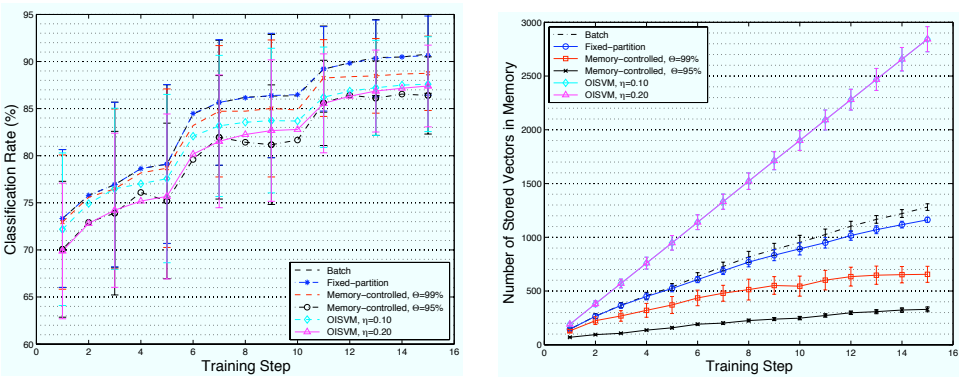
7.1 Experiments with Constrained Variability

In the first series of experiments, we evaluated the properties of the memory-controlled incremental SVM in a simplified scenario. We therefore trained the system on three sequences acquired under similar illumination conditions, with the same robot platform. The fourth sequence was used for testing. Training on each sequence was performed in 5 steps, using one subsequence at a time, resulting in 15 steps in total. We considered 36 different permutations of training and test sequences. Here we report average results obtained on both global and local features by the three incremental algorithms (fixed-partition, OISVM, and memory-controlled) as well as the batch method. We tested the memory-controlled algorithm using two different values of the parameter Θ, i.e. $\Theta = 99\%, 95\%$. This corresponds to the maximum accepted reduction of the recognition rate of 1% and 5% respectively, as explained in Section 4.3. Similarly for OISVM, we used three different values of the parameter η that determines how sparse the final solution is going to be (as in [34]).

Fig. 9, left, shows the recognition rates obtained at each incremental step by all methods and for both feature types. Fig. 9, right, reports the number of training samples that had to be stored in the memory at each step of the incremental procedure. First, we see that OISVM achieves very good performance similar to the batch method. However, both methods suffer from the same problem: they require all the training samples to be kept in the memory during the whole learning process. This makes them unsuitable for realistic scenarios, particularly in cases when the algorithm should be used on a robotic platform with intrinsically limited resources. The fixed-partition algorithm achieves identical performance as the batch method, while greatly reducing the number of training samples that need to be stored in the memory at each incremental step. However, despite that all the algorithms show plateaus in the classification rate whenever the model is trained on similar data (coming from consecutive subsequences), the number of support vectors grows roughly linearly with the number of training steps.



(a) Classification rate and number of training samples stored for global features.



(b) Classification rate and number of training samples stored for local features.

Figure 9: Average results obtained for the experiments with constrained variability for three incremental methods and the batch algorithm.

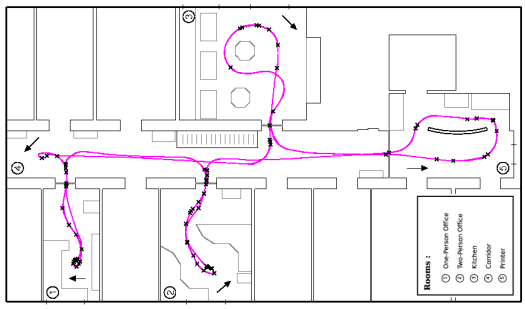
We see that for the memory-controlled incremental SVM, both the classification rate and the number of stored support vectors show plateaus every five incremental steps (as opposed to the classification rate only in case of the other methods). The method controls the memory growth much more successfully than the original fixed-partition incremental technique. For instance, when we accept only one percent reduction in classification (i.e. $\Theta = 99\%$), the number of support vectors stored after the 15 steps is 39.6% (CRFH) and 43.7% (SIFT) lower than for the fixed-partition incremental method. For $\Theta = 95\%$, the gain in memory compression is much greater than the overall decrease in performance. This feature, i.e. the possibility to trade memory for a controlled reduction in performance, can be potentially very useful for systems operating in realistic, open-ended learning scenarios and with limited memory resources. This approach would be even more

appealing for systems which can compensate the loss in performance by doing information fusion over time or from multiple sensors. It is worth underlying that the growth in the number of support vectors decreases over time (Fig. 9, bottom). For example, for CRFH and $\Theta = 99\%$, the model trained on the second sequence (step 6 to 10) grows by 115 vectors on average, but trained on the third sequence (step 11 to 15) grows only by 74 vectors. This may indicate that the number of SVs eventually tends to reach a plateau.

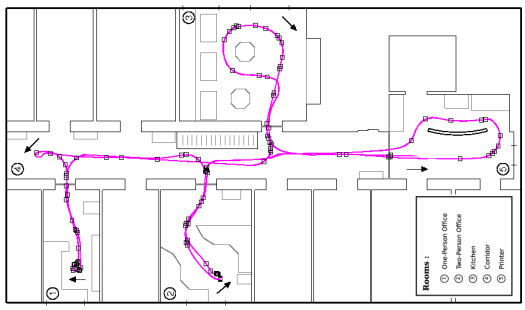
In order to gain a better understanding of the methods' behavior, we performed an additional analysis of the results. Fig. 11b shows, for the two approximate incremental techniques, the average amounts of vectors (originating from each of the three training sequences) that remained in the model after the final incremental step (note that, in our case, this analysis would be pointless for OISVM, as it requires storing all the training data). The figure illustrates how the methods weigh instances, learned at different time, when constructing the internal representation. We see that both fixed-partition and memory-controlled algorithms privilege new data, as the SVs from the last training sequence are more represented in the model. This phenomenon is stronger for the memory-controlled algorithm.

To get a feeling for how the forgetting capability works in case of the memory-controlled method, we plotted the positions where the SVs were acquired, for $\Theta = 99\%$ and the CRFH features. Fig. 10 reports results obtained for a model built after the final incremental step. The positions were marked on three maps presented in Fig. 10a,b,c so that each of the maps shows the SVs originating from only one training sequence. These SVs could be considered as landmarks selected by the visual system for the recognition task. As already shown in Fig. 11b, most of the vectors in the model come from the last training sequence. Moreover, the number of SVs from the previous training steps decreases monotonically, thus the algorithm gradually forgets the old knowledge. It is interesting to observe how the vectors from each sequence are distributed along the path of the robot. On each map, the places crowded with SVs are mainly transition areas between the rooms, regions of high variability, as well as places at which the robot rotated (thus providing a lot of different visual cues without changing position). To illustrate the point, Fig. 11a shows sample images acquired in the corridor, for which the SVs decay quickly, and one of the offices, for which they are being preserved much longer. The results indicate that the forgetting is not performed randomly. On the contrary, the algorithm tends to preserve those training vectors that are most crucial for discriminative classification, and first forgets the most redundant ones.

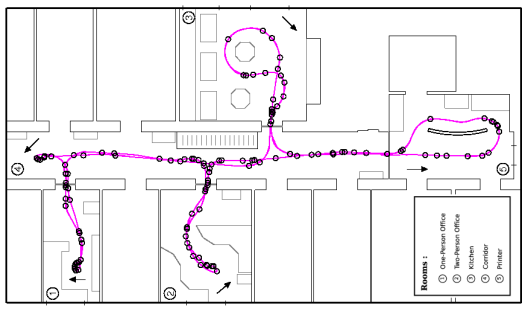
On the basis of these experimental findings, we can conclude that the memory-controlled incremental SVM is the best method for vision-based robot localization of those considered here. Therefore, in the rest of the paper we will use only this algorithm, with $\Theta = 99\%$.



(a) 78 Support Vectors from the 1st sequence.

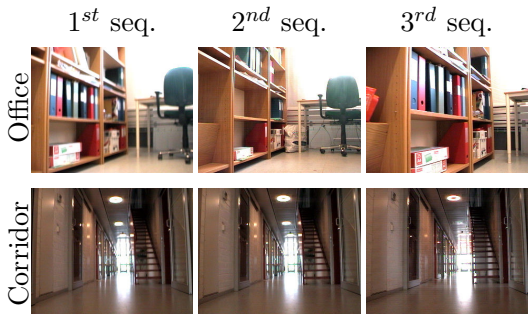


(b) 111 Support Vectors from the 2nd sequence.

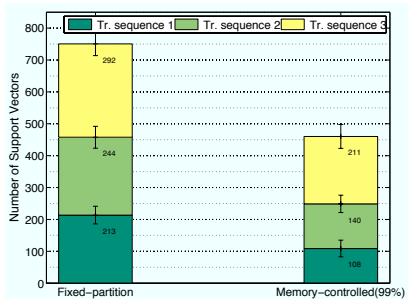


(c) 149 Support Vectors from the 3rd sequence.

Figure 10: Maps of the environment with plotted positions of the support vectors stored in the model obtained after the final incremental step for one of the experiments conducted using the memory-controlled technique with $\Theta = 99\%$. The support vectors were divided into three maps (a-c) according to the training sequence they originate from. Additionally, each map shows the path of the robot during acquisition of the sequence (arrows indicate the direction of driving). We observe that the Support Vectors from the old training sequences were gradually eliminated by the algorithm and this effect was stronger in regions with lower variability.

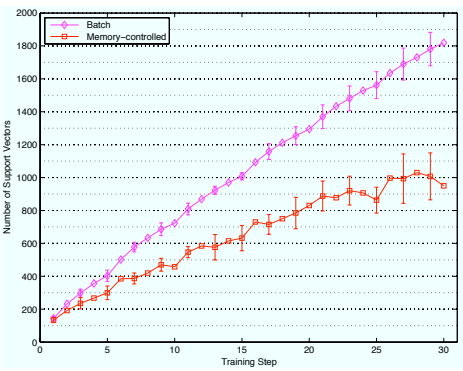


(a) Sample images from the three training sequences.

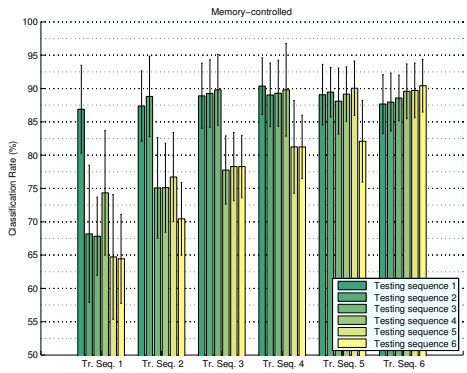


(b) Statistics of Support Vectors stored in the final approximate incremental models.

Figure 11: Sample images captured in regions of different variability (left). Comparison of the average amounts of training vectors coming from the three sequences that were stored in the final incremental model for the two approximate incremental techniques (right).

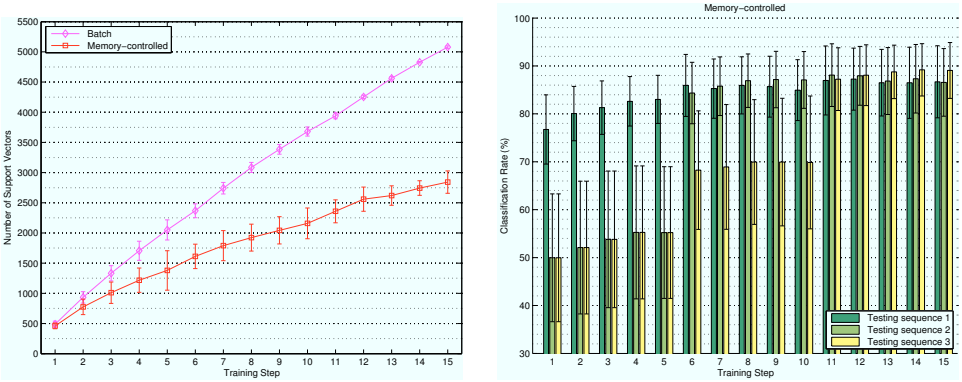


(a) Support vectors



(b) Performance of memory-reduced

Figure 12: Average results of the IDOL2 experiments in the real-world scenario. (a) compares the amounts of SVs stored in the models at each incremental step for the batch and the memory-controlled method. (b) reports the classification rate measured every fifth step (every time the system completes learning a whole sequence) with all the available test sets. The training and test sets marked with the same indices were acquired under similar conditions.



(a) Support vectors (b) Performance of memory-reduced

Figure 13: Average results of the COLD-Freiburg experiments in the real-world scenario. (a) compares the amounts of SVs stored in the models at each incremental step for the batch and the memory-controlled method. (b) reports the classification rate measured every step with all the available test sets. The consecutive training and testing sequences were acquired under similar conditions.

7.2 Experiments with Unconstrained Variability

The next step was to test our incremental method in a real-world scenario. To this purpose, we considered the case where the algorithm needed to incrementally gain robustness to variations introduced by changing illumination and human activities, while at the same time using its adaptation ability to handle long-time changes in the environment. We performed the experiments first on the IDOL2 database. Then, to confirm the behavior on a different set of data, we used the COLD-Freiburg database. We first trained the system on three IDOL2 sequences acquired at roughly similar time but under different illumination conditions. Then, we repeated the same training procedure on sequences acquired 6 months later. In order to increase the number of incremental steps and differentiate the amount of new information introduced by each set of data, each sequence was again divided into five subsequences. In total, for each experiment we performed 30 incremental steps. Since the IDOL2 database consists of pairs of sequences acquired under roughly similar conditions, each training sequence has a corresponding one which could be used for testing. Feature-wise, here we used only the global features (CRFH). Indeed, the experiments presented in the previous section showed that local features achieve an accuracy similar to that of CRFH, but at a much higher computational cost and memory requirement. Also, preliminary experiments show that this behavior is confirmed in this scenario, hence the choice to use here only the global descriptor.

We used a very similar system and experimental procedure for the experiments with the COLD-Freiburg dataset. As in case of IDOL2, we divided each sequence into 5 subsequences and used pairs of sequences acquired under roughly similar conditions for training and testing. In case of both databases, the experiment was repeated 12 times for different orderings of training sequences. Fig. 12 and 13 report the average results together with standard deviations. By observing the classification rates for a classifier trained on the first sequence only, we see that the system achieves best performance on a test set acquired under similar conditions. The classification rate is significantly lower for other test sets. In case of IDOL2, this is especially visible for images acquired 6 months later, even under similar illumination conditions. At the same time, the performance greatly improves when incremental learning is performed on new batches of data. The classification rate decreases for the old test sets; at the same time, the size of the model tends to stabilize.

7.3 Discussion

The presented results provide a clear evidence of the capability of the discriminative methods to perform incremental learning for vision-based place recognition, and their adaptability to variations in the environment. Table 1 summarizes the performance obtained by each method in terms of accuracy, speed, controlled memory growth and forgetting capability. For each algorithm (i.e. for each row), we put a cross corresponding to the property (i.e. the column) that the algorithm has shown to possess in our experiments. The fixed-partition method performs as well as batch SVM, but it is unable to control the memory growth and requires much more memory space. We also found that OISVM could get very good accuracy while achieving a low computational complexity during testing. However, none of the two methods has shown to possess an effective forgetting capability: for the fixed-partition method, the old SVs decay slowly, but the decay is neither predictable nor controllable; for OISVM, every training vector must be stored into memory. As opposed to this, the memory-controlled algorithm is able to achieve performances statistically equivalent to those of batch SVM, while at the same time providing a principled and effective way to control the memory growth. Experiments showed that this has induced a forgetting capability which privileges newly acquired data to the expenses of old one and the model growth slows down whenever new data are similar to those already processed. Furthermore, since a lot of training images can be discarded during the incremental process, the training time soon becomes significantly lower than for the batch method. For instance, in case of the second experiments, training the classifier at the last step took 25.5s for the batch algorithm and only 5.6s for the memory-controlled method on a 2.6GHZ Pentium IV machine, and recognition time was twice as fast for the memory-controlled algorithm than for the batch one.

	Accuracy	Forgetting	Memory	Speed
Fixed-partition	x	x		
OISVM	x			x
Memory-controlled	**x**	**x**	**x**	**x**

Table 1: Comparing incremental learning techniques for place recognition and robot localization applications.

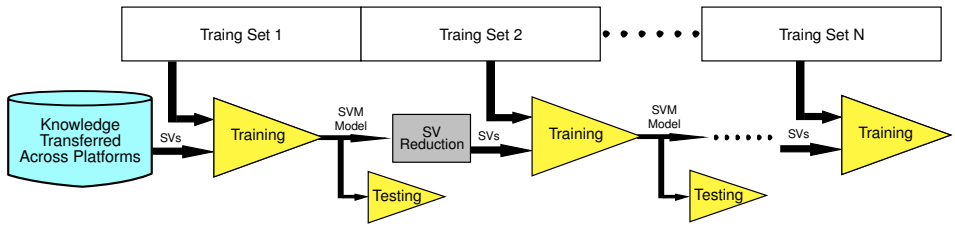


Figure 14: A diagram illustrating the data flow in the knowledge-transfer system.

8 Experiments on Knowledge Transfer

As a second application of our method, we considered the problem of transfer of knowledge between robotic platforms with different characteristics, performing vision-based recognition in the same environment. We used the IDOL2 database and the robots Minnie and Dumbo for these experiments. The main difference between the two platforms lies in the height of the cameras (see Fig. 15). They both use the memory-controlled incremental SVM as a basis for their recognition system, thus they share the same knowledge representation. The aim is to efficiently exploit the knowledge acquired e.g. by one robot so to boost the recognition performance of another robot. We propose to use our method to update the internal representation when new training data are available. Fig. 14 illustrates how our approach can be used for transfer of knowledge. We would like the knowledge transfer scheme to be adaptive, and also to privilege newest data so to avoid accumulation of outdated information. Finally, the solution obtained starting from a transferred model should gradually converge to the one learned from scratch, not only in terms of performance but also of required resources (e.g. memory).

The challenges in the transfer of knowledge will come from:

- *(a) Differences in the parameters of the two platforms*
  The cameras are mounted at two different heights, thus the informative content of the images acquired by the two platforms is different. Because of this, the knowledge acquired by one platform might not be helpful for the other one or, in the worst case, it might constitute an obstacle. Preliminary experiments showed that SIFT is more suitable for the transfer of knowledge in our scenario than CRFH. For that reason, CRFH will not be used.

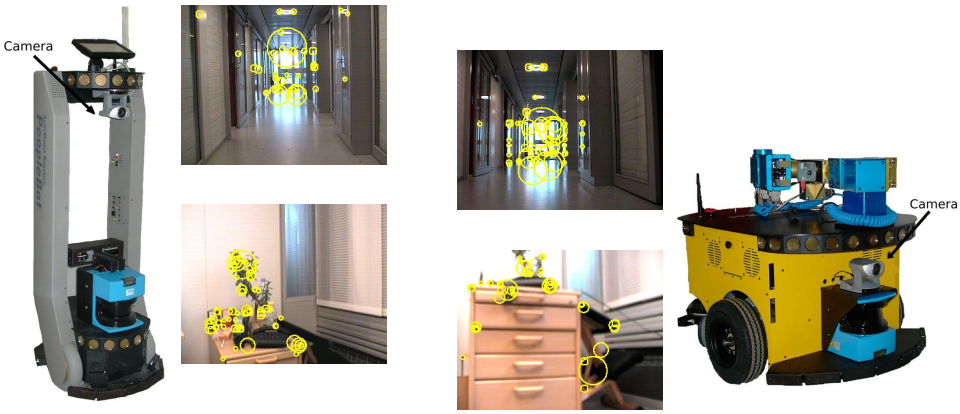


Figure 15: Knowledge transfer across robot platforms which only partially share visual information.

- *(b) Room by room/frames by frames knowledge update*
  It is desirable to update the model transferred across platforms as soon as new data are available. We will investigate the behavior of the algorithm when the update is performed room-by-room, or frames-by-frames. Both scenarios are at risk of unbalanced data with respect to the class being updated.

- *(c) Growing memory requirements*
  Building on top of an already trained classifier might lead to a solution that will be much more demanding in terms of memory usage and computational power than the one learned from scratch. Although our memory-controlled approach is capable of reducing the number of SVs, its reduction process does not take the sources of the information into consideration. In order to favor information coming from the platform currently in use, we imposed to the algorithm to discard only those SVs that were linearly dependent *and* came from the previous platform by adding meta-information on the training examples. This scheme speeds up the turnover of stored SVs, while preferring newest data and at the same time preserving relevant information.

In the IDOL2 database, for each robot and for every illumination condition, we always have two sequences acquired under similar conditions. Here, we always used such pairs of sequences, one as a training set and the other one as a test set. In all the experiments, we benchmarked against a system not using any prior knowledge.

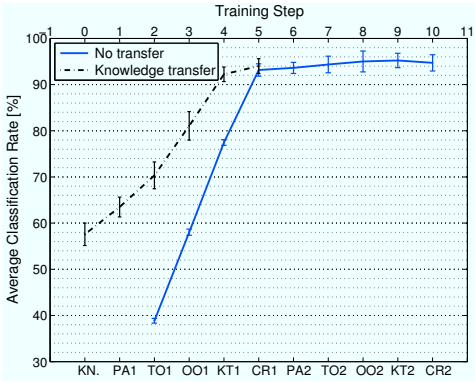
8.1 Experiments with room by room updates

In the first series of experiments, the system was updated incrementally in a room by room (i.e. class by class) scenario. The system was trained incrementally on one

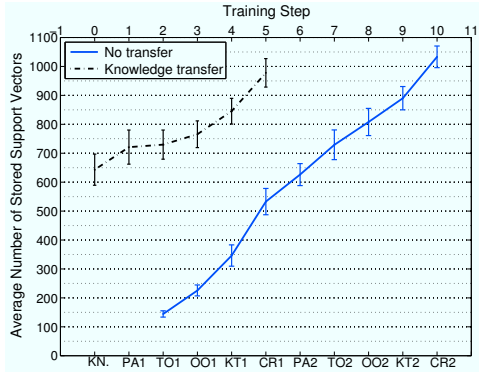
sequence; the corresponding sequence, acquired under roughly similar conditions, was used for testing. The prior-knowledge model was built using standard batch SVM from one image sequence, acquired under the same illumination conditions and at close time as the training one, but using a different platform. As there are five classes in total, training was performed in 5 steps (the algorithm learned incrementally one room at the time). In the no-transfer case, the system needed to build the model from scratch, and thus needed to acquire data from at least two classes. In this case, training on each sequence required only 4 steps since in the first step the algorithm learned to distinguish between the first two classes.

Building on top of knowledge acquired from another platform implies a growth in the memory requirements. To evaluate this behavior in relationship to its effects on performance and compare fairly to the system trained without a prior model, we incrementally updated the model without transferred knowledge on another sequence acquired under conditions similar to that of the first training sequence. This experiment makes it possible to evaluate performance and memory growth when both systems are trained on two sequences. The main difference is that in one case both sequences were acquired and processed by the same platform; in the other case, one sequence was acquired and processed by a different platform. We considered different permutations in the rooms order for the updating; for each permutation, we considered 6 different orderings of the sequences used as training, testing, and prior-knowledge sets. Due to space reasons, we report only average results for one permutation, together with standard deviations in Fig. 16.

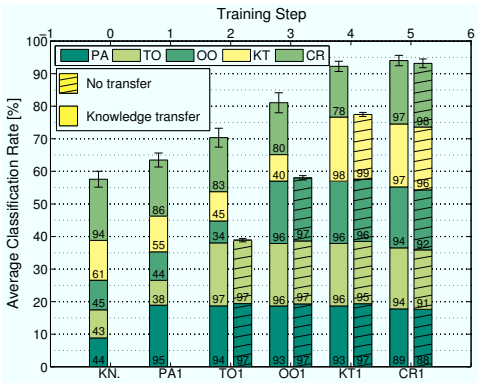
We can see that, for both approaches, the system gradually adapts to its own perception of the environment. It is clear that the knowledge-transfer system has a great advantage in terms of performance over the no-transfer system at the first steps. For instance, we see that, after the second update (TO1, Fig 16a), the knowledge-transfer system achieves a classification rate of 65.3%, while the no--transfer knowledge obtains only 37%. The advantage in classification rate for the knowledge-transfer system remains considerable for the steps OO1 and KT1. However, it is interesting to note that even when both systems have been updated on a full sequence (CR1, Fig 16a), the knowledge-transfer system still maintains an advantage in performance. Considering the differences between the two platforms, and that the transferred knowledge model was built on a single sequence, this is a remarkable result. It can also be observed from Fig. 16d that the memory-controlled algorithm facilitated the decay of knowledge from the other platform (in the first incremental step, we did not perform the reduction), while the knowledge acquired by its own sensor gradually becomes the main source for the model. As the no-transfer system continued to learn one additional sequence incrementally, its memory growth eventually exceeded the knowledge-transfer case (see Fig 16b). Although the model was built on two sequences acquired by the same platform, the knowledge-transfer system still obtains a comparable performance. We conclude that the transfer of knowledge, in a room by room updating scenario, acts as an effective boosting of performance, without any long-term growth of the memory requirements.



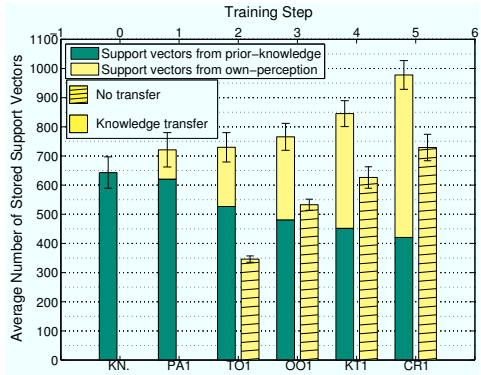
(a) Classification rates at each training step.



(b) Number of support vectors at each training step.

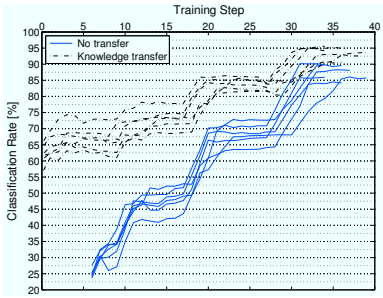


(c) Comparison of the performance at each training step.

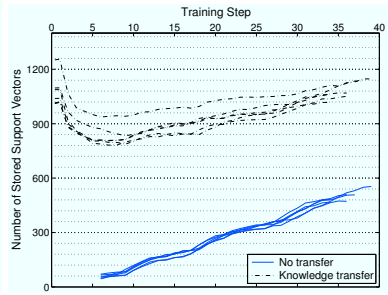


(d) Comparison of the number of support vectors at each step.

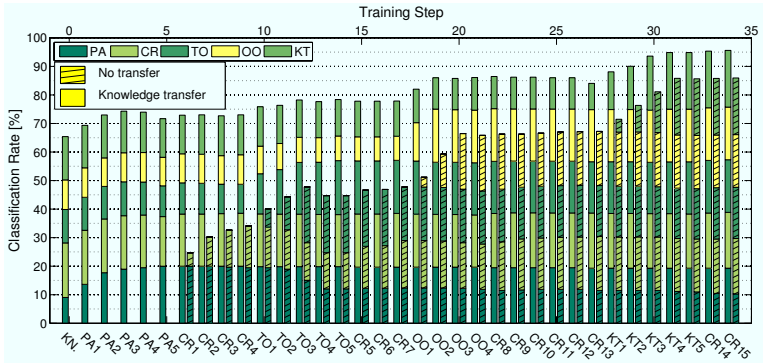
Figure 16: Average results obtained for the system incrementally trained with and without transfer of knowledge in the room by room fashion. Fig. 16a,b compare the final recognition rates and the total number of support vectors for both cases. Fig. 16c,d present a detailed analysis: classification rates obtained for each of the rooms and the amount of support vectors in the final model that originate from the transferred knowledge. In all the plots, the first step "KN." corresponds to the results obtained for the transferred knowledge before any update was performed.



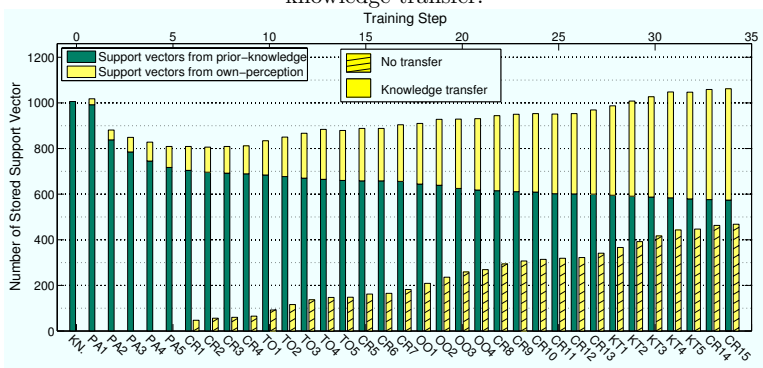
(a) Classification rates at each training step.



(b) Number of support vectors at each training step.



(c) Detailed comparison of the performance of the system with and without knowledge-transfer.



(d) Number of stored support vectors of incremental experiment with and without knowledge-transfer at each step.

Figure 17: Average results obtained for the system incrementally trained with and without transfer of knowledge in the frames by frames fashion. The labels below each bar indicate the batch of data used for the incremental update. Again, the first step labeled as "KN." corresponds to the results obtained for the transferred knowledge before any update was performed.

8.2 Experiments with frame by frame updates

The second series of experiments explored the behavior of the system in a frames by frames updating scenario. Here, for each incremental update, we used a certain number of consecutive frames taken from the training image sequence. Again, the system was trained incrementally on one sequence, and a corresponding sequence was used as a test set. We examined the performance of the system for the case when updating was performed using 30 frames per step[1]. Thus, for each experiment, it took more than 30 incremental steps in total to complete a sequence. The prior-knowledge model was built using two complete sequences acquired by the other platform, under the same illumination conditions and very close in time. This provided a better start-up performance than in case of the previous experiments. Again, we benchmarked against the system not using any prior knowledge. In this case, in order to fulfill the requirement of training using at least 2 classes, the first training set consisted of all the images captured in the first room plus the first 30 frames captured in the second room. As a consequence, the full training process required five to six less steps than in case of equivalent experiments using the knowledge-transfer scheme. The experiment was repeated 6 times for different orderings of training sequences. Since the number of training steps varied (due to a different number of images in each sequence), we report all the results separately. Fig. 17a,b report the amount of stored SVs and classification rates at each step, for all the experiments. This shows the general behavior for both approaches. Fig. 17c,d present results for one of the 6 experiments, so to allow a detailed analysis.

By observing the classification rates obtained at each step in both cases, we see that the advantage of the knowledge-transfer scheme is even more visible here than for the room by room updating scenario. This might be due to the fact that some of the training sets used for the no-transfer case are highly unbalanced. We can observe from Fig. 17c that the performance of the system for previously learned rooms can drop considerably when a new batch of frames is loaded; this is not the case for the knowledge-transfer system. The twelfth step, when the system was updated with frames from the two-persons office (TO3, Fig. 17c), is a typical example. Note that this is a general phenomenon present, although less pronounced, also in the room by room updating scenario. Our interpretation is that the model of the prior--knowledge contains information about the overall distribution of the data. This helps to find a balanced solution when dealing with non-separable instances using soft-margin SVM [10]. As a last remark the knowledge from the transferred model is gradually removed over time (see Fig. 17d).

[1]Experiments conducted for 10 and 50 frames per training step gave analogous results, and for space reasons are not reported here.

9 Summary and Conclusions

In this paper we presented a novel extension of SVM to incremental learning that achieves the same recognition performance of the standard, batch method while limiting the memory growth over time. This is achieved by discarding, at each incremental step, all the support vectors that are not linearly independent. The information they carry is not lost, as it is retained into the algorithm's decision function in the form of weighting coefficients of the remaining support vectors. We call this method memory-controlled incremental SVM. We applied it to the problem of place recognition for robot topological localization, focusing on two distinct scenarios: adaptation in presence of dynamic changes and transfer of knowledge between two robot platforms engaged in the same task. Experiments show clearly the effectiveness of our approach in terms of accuracy, speed, reduced memory and capability to forget redundant, outdated information.

We plan to extend this work in several ways. First, we want to use the memory-controlled algorithm in multi-modal learning scenarios, for instance using laser-based features combined with visual ones, as done in [39], in an incremental setting. Here we should be able to exploit fully the properties of the method, and aggressively trade memory for accuracy on single modalities, while retaining an high overall performance. Second, we would like to investigate further the knowledge transfer scenario, and incorporate in our framework ways to select the data to be transferred, as proposed in [24]. Future work will concentrate in these directions.

Acknowledgments

This work was sponsored by the EU FP7 project CogX (A. Pronobis) and IST-027787 DIRAC (B. Caputo, L. Jie), and the Swedish Research Council contract 2005-3600-Complex (A. Pronobis). The support is gratefully acknowledged.

References

[1] EU FP6 Integrated Project COGNIRON: The Cognitive Robot Companion. URL `http://www.cogniron.org`.

[2] EU FP6 Integrated Project RobotCub. URL `http://www.robotcub.org/`.

[3] EU FP6 IST Cognitive Systems Integrated Project CoSy: Cognitive Systems for Cognitive Assistants. URL `http://www.cognitivesystems.org/`.

[4] Matej Artač, Matjaž Jogan, and Aleš Leonardis. Mobile robot localization using an incremental eigenspace model. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation (ICRA'02)*, pages 1025–1030, May 2002. ISBN 0-7803-7272-7.

[5] Serge Belongie, Charless Fowlkes, Fan Chung, and Jitendra Malik. Spectral partitioning with indefinite kernels using the Nyström extension. In *Proceedings of the 7th European Conference on Computer Vision (ECCV'02)*, pages 531—-542, 2002.

[6] Emma Brunskill, Thomas Kollar, and Nicholas Roy. Topological mapping using spectral clustering and classification. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 3491–3496, San Diego, CA, USA, October 2007.

[7] Barbara Caputo, Eric Hayman, and P. Mallikarjuna. Class-specific material categorisation. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, pages 1597–1604. Citeseer, 2005.

[8] Gert Cauwenberghs and Tomaso Poggio. Incremental and decremental support vector machine learning. *Advances in Neural Information Processing Systems (NIPS)*, 13, 2001.

[9] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: A library for Support Vector Machines*, 2001. URL `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[10] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. ISBN 0521780195.

[11] Gamini Dissanayake, Paul M. Newman, Steven Clark, Hugh F. Durrant-Whyte, and M. Csorba. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241, June 2001.

[12] Carlotta Domeniconi and Dimitrios Gunopulos. Incremental support vector machine construction. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM'01)*, pages 589–592, 2001. ISBN 0-7695-1119-8.

[13] Gyuri Dorkó and Cordelia Schmid. Object class recognition using discriminative local features. 2005.

[14] Tom Downs, Kevin E. Gates, and Annette Masters. Exact simplification of support vector solutions. *Journal of Machine Learning Research (JMLR)*, 2(2):293–297, May 2002.

[15] John Folkesson, Patric Jensfelt, and Henrik I. Christensen. Vision SLAM in the measurement subspace. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, 2005.

[16] Mario Fritz, Bastian Leibe, Barbara Caputo, and Bernt Schiele. Integrating representative and discriminant models for object category detection. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, 2005.

[17] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[18] Kristen Grauman and Trevor Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, pages 1458–1465, 2005. ISBN 0-7695-2334-X.

[19] Matjaž Jogan and Aleš Leonardis. Robust localization using an omnidirectional appearance-based subspace model of environment. *Robotics and Autonomous Systems (RAS)*, 45(1):51–72, 2003.

[20] George Konidaris and Andrew G. Barto. Autonomous shaping: Knowledge transfer in reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, 2006.

[21] David M. Kortenkamp and Terry Weymouth. Topological mapping for mobile robots using a combination of sonar and vision sensing. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, Seattle, Washington, USA, 1994.

[22] Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems, Special Issue on Human and Robot Interactive Communication*, 4(1):125–138, 2007.

[23] Benjamin Kuipers and Patrick Beeson. Bootstrap learning for place recognition. *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI'02)*, pages 174–180, 2002.

[24] Alessandro Lazaric, Marcello Restelli, and Andrea Bonarini. Transfer of samples in batch reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, pages 544–551, Helsinki, Finland, 2008.

[25] Oskar Linde and Tony Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proceedings of the 2004 15th International Conference on Pattern Recognition (ICPR'04)*, 2004.

[26] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[27] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. The KTH-IDOL2 database. Technical report, Kungliga Tekniska Högskolan, CVAP/CAS, September 2006.

[28] Richard J. Malak and Pradeep K. Khosla. A framework for the adaptive transfer of robot skill knowledge using reinforcement learning agents. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation (ICRA'01)*, Seoul, Korea, May 2001.

[29] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV'01)*, Vancouver, Canada, 2001.

[30] Tom M. Mitchell. The discipline of machine learning. Technical Report CMU-ML-06-108, Carnegie Mellon University, 2006.

[31] Oscar Martinez Mozos, Rudolph Triebel, Patric Jensfelt, Axel Rottmann, and Wolfram Burgard. Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems (RAS)*, 55(5):391–402, 2007.

[32] Ana Cris Murillo, Jana Košecká, J J Guerrero, and C Sagüés. Visual door detection integrating appearance and shape cues. *Robotics and Autonomous Systems (RAS)*, 56(6), 2008.

[33] Illah Nourbakhsh, Rob Powers, and Stan Birchfield. Dervish: An office navigation robot. *AI Magazine*, 16(2):53–60, 1995.

[34] Francesco Orabona, Claudio Castellini, Barbara Caputo, Jie Luo, and Giulio Sandini. Indoor place recognition using online independent Support Vector Machines.

In *Proceedings of the British Machine Vision Conference (BMVC'07)*, Warwick, UK, 2007.

[35] Andrzej Pronobis. *Indoor Place Recognition Using Support Vector Machines*. Master's thesis, Kungliga Tekniska Högskolan, Stockholm, Sweden, December 2005.

[36] Andrzej Pronobis and Barbara Caputo. Confidence-based cue integration for visual place recognition. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 2394–2401, San Diego, CA, USA, October 2007.

[37] Andrzej Pronobis and Barbara Caputo. COLD: The CoSy localization database. *The International Journal of Robotics Research (IJRR)*, 28(5):588–594, 2009.

[38] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I. Christensen. A discriminative approach to robust visual place recognition. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, October 2006.

[39] Andrzej Pronobis, Oscar Martinez Mozos, and Barbara Caputo. SVM-based discriminative accumulation scheme for place recognition. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 522–529, Pasadena, CA, USA, May 2008.

[40] Stephen Se, David G. Lowe, and James J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation (ICRA'01)*, Seoul, Korea, 2001.

[41] Christian Siagian and Laurent Itti. Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[42] Danijel Skočaj and Aleš Leonardis. Weighted and robust incremental method for subspace learning. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, pages 1494–1501, 2003. ISBN 0-7695-1950-4.

[43] Nadeem Ahmed Syed, Huan Liu, and Kah Kay Sung. Incremental learning with support vector machines. In *Proceedings of the 16th International Joint Conference on Artificial intelligence (IJCAI'99)*, Stockholm, Sweden, 1999.

[44] Adriana Tapus and Roland Siegwart. Incremental Robot Mapping with Fingerprints of Places. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Edmonton, Alberta, Canada, 2005.

[45] Sebastian Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99(1):21–71, 1998.

[46] Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems (RAS)*, 15(1-2):25–46, 1995.

[47] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, 2003.

[48] Iwan Ulrich and Illah Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation (ICRA'00)*, San Francisco, CA, USA, 2000.

[49] Christoffer Valgren and Achim J. Lilienthal. SIFT, SURF and seasons: Long-term outdoor localization using local features. In *Proceedings of the European Conference on Mobile Robots (ECMR'07)*, 2007.

[50] Vladimir Vapnik. *Statistical Learning Theory*. Wiley and Son, New York, 1998.

[51] Christian Wallraven, Barbara Caputo, and Arnulf Graf. Recognition with local features: The kernel recipe. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, Beijing, China, 2003.

[52] Jürgen Wolf, Wolfram Burgard, and Hans Burkhardt. Robust vision-based localization by combining an image retrieval system with Monte Carlo localization. *IEEE Transactions on Robotics*, 21(2):208–216, 2005.

Paper C

Confidence-based Cue Integration for Visual Place Recognition

Andrzej Pronobis and Barbara Caputo

Published in
IEEE/RSJ International Conference on Intelligent Robots and Systems

Confidence-based Cue Integration for Visual Place Recognition

Andrzej Pronobis and Barbara Caputo

Abstract

A distinctive feature of intelligent systems is their capability to analyze their level of expertise for a given task; in other words, they know what they know. As a way towards this ambitious goal, 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 and 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.

1 Introduction

A key competence for an autonomous agent is the ability to localize itself in the world. Vision-based localization represents a challenge for the research community, because the visual information tends to be noisy and difficult to analyze. Still, this research line is attracting more and more attention, and several methods have been proposed using vision alone [16, 18, 15], or combined with more traditional range sensors [6, 17]. The increasing activity in this research area comes firstly from the portability and cost-effectiveness of visual sensors; secondly, from the specific type of information that only these sensors can bring. This is the case for instance in place categorization or understanding, where the semantic information plays a crucial role. Furthermore, visual place recognition can be applied as a method for loop closing, scalability issues, or recovery from the kidnapped robot problem.

A vast majority of algorithms proposed so far were designed to provide as output a hard decision: the system is trained to recognize a fixed and pre-defined set of environments (e.g. kitchen, corridor, office etc.) and then, when presented with a test image, it classifies it as one of the possible places, but little or nothing is generally said regarding the *confidence* of this decision or other possible hypotheses. Measuring confidence, or knowing what is known, is a fundamental concept for autonomous robots. Indeed, in many real-world applications it is more desirable to abstain from action because of a self-recognized lack of confidence, rather than take a hard decision which might result in a costly error. Thus, introducing a confidence measure in a recognition algorithm allows to provide reliability despite constrained performance of the algorithm or lack of updated information, and makes it possible to evaluate when it is necessary to seek for extra information (e.g. from multiple cues or modalities) in order to achieve a confident decision.

It is possible to define a confidence measure for any pattern recognition algorithm: for probabilistic methods, it will be related to the posterior probability of the image at hand; for discriminative classifiers, it will be related to the distance from the separating hyperplane. In this paper, we will focus on large margin classifiers, specifically on Support Vector Machines (SVMs), even if most of the concepts and ideas we will propose can be easily extended to any margin-based discriminative method. We build on our previous work on place recognition, where we presented an SVM-based method able to recognize indoor environments under severe illumination changes and across a time span of several weeks [15]. Our first contribution is the introduction of a method for ranking the hypotheses generated by the classifier and measuring their confidence. The method is based on the distance from the hyperplane and the average distance of each training class. We present experiments showing that our confidence measure gives a better performance compared with the classic hard decision SVM and, more important, a decision that is more informative of the level of knowledge of the robot.

Once a system is able to output not only its guess, but also the level of confidence of the guess, action should be taken. Indeed, we can expect that when an image is classified with a low level of confidence, it is because the algorithm doesn't have enough information. A possible way to increase the knowledge, and thus the confidence, is to use additional information such as both global and local visual descriptors, or laser-based geometrical data, and combine them through an integration scheme. An effective method for visual cue integration using SVMs has been proposed in [11], called Discriminative Accumulation Scheme (DAS). A second contribution of this paper is to apply that algorithm to the domain of vision for robotics. We also propose its generalized version (Generalized DAS), that can be built on top of our confidence estimation method. Experiments confirm the effectiveness of the approach and show that G-DAS consistently outperforms the original DAS.

While using multiple visual cues improves both classification accuracy and relative confidence, it is computationally expensive (more features to compute and classify), which is undesirable for an autonomous agent. Ideally, a system should

use additional information only when necessary, i.e. only when the level of confidence of a single cue is not such to obtain a reliable decision. The final contribution of this paper is to combine the G-DAS framework with the confidence estimation approach, so that multiple cues are used only when they are necessary to disambiguate low-confidence cases. Our experiments on local and global visual cues show that the proposed approach reduces the computational load of about 55% in average, achieving the same performance obtained by using G-DAS on all the images.

The rest of the paper is organized as follows: after an overview of previous work on confidence measures and cue integration (Section 2), Section 3 gives a brief description of the methodology used further. Section 4 describes our confidence estimation method and evaluates its effectiveness for visual place recognition. Section 5 reviews DAS, presents our generalized version of the algorithm and assesses its performance; Section 6 shows how by combining the two techniques we achieve a better overall performace while reducing the computational load. The paper concludes with a summary and possible avenues for future research.

2 Related Work

We are not aware of confidence estimation and/or cue integration methods within the robotics literature for visual place recognition. However, computing confidence estimates for discriminative classifiers is an open problem in machine learning. Although classifiers like K-NN, ANN, or SVM output numeric scores for class membership, some experiments show that, when used directly, they are not well correlated with classification confidence [4]. Several authors attacked this problem by developing more sophisticated measures such as probability estimates obtained by trained sigmoid function [12] with extensions for multi-class problems [20], or relative distance from the separating hyperplane, normalized with the average class distance from the plane [5]. More comments on their performance can be found in Section 4.

Visual cue integration via accumulation was first proposed in a probabilistic framework by Poggio *et al.*[14], and then further explored by Aloimonos and Shulman [1]. The idea was then extended to SVMs by Nilsback and Caputo [11] (DAS). The resulting method showed remarkable performances on object recognition applications and together with its generalized version (G-DAS) is used here as a cue integration scheme for disambiguating classes with low confidence estimate.

3 A Few Landmarks

This section serves as a base for the results and theory presented further. We describe the common scenario and methodology used during all experimental evaluations (Sections 3.1 and 3.2), we briefly review SVMs (Section 3.4) and the visual descriptors used throughout the paper (Section 3.3).

3.1 Experimental Scenario

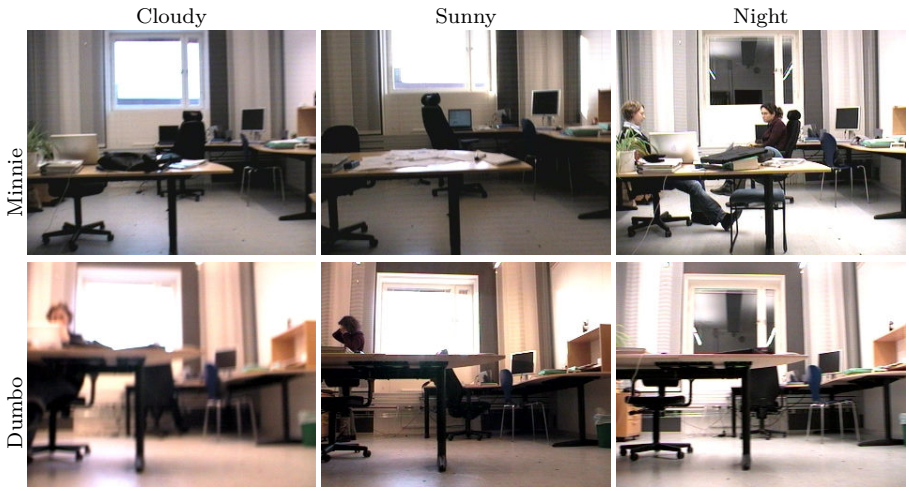
The algorithms presented in this paper have been tested in the domain of mobile robot topological localization. As benchmarking data we used the IDOL (Image Database for rObot Localization [9]) database which was introduced in [15] in order to test robustness of our discriminative approach to visual place recognition in real-world scenario and under varying illumination conditions. The database comprises sequences of images of places acquired using cameras of resolution 320x240 pixels mounted at different heights (98cm and 36cm) on two mobile robot platforms, the PeopleBot Minnie and the PowerBot Dumbo. The acquisition was performed in a five room subsection of a larger office environment, selected in such way that each of the five rooms represented a different functional area: a one-person office, a two-persons office, a kitchen, a corridor, and a printer area (part of the corridor). Example pictures showing interiors of the rooms are presented in Fig. 1.

The appearance of the rooms was captured under three different illumination conditions: in cloudy weather, in sunny weather, and at night. The robots were manually driven through each of the five rooms while continuously acquiring images at the rate of 5 frames per second. Each image was then labelled as belonging to one of the rooms according to the position of the robot estimated using a laser-based localization method. The acquisition was performed twice for each robot and illumination condition, resulting in 12 image sequences in total over a span of time of more than two weeks. Thus, the sequences captured variability introduced not only by illumination but also natural activities in the environment (presence/absence of people, furniture relocated etc.). Example images illustrating the captured variability for both robot platforms are shown in Fig. 1.

3.2 Experimental Procedure

As a basis for the experiments, we used the place recognition system presented in [15], which is built around Support Vector Machines [3], and a rich global descriptor [7]. While designing the system, we followed the assumption that the global configuration of a scene is informative enough for recognition and obtained good performance despite variations captured in the IDOL database. In this work, in order to increase robustness, we additionally used the SIFT descriptor [8] that has already been proved successful in the domain of vision-based localization [16].

Following [15], we took a fully supervised approach and assumed that during training each room is represented by a collection of images capturing its visual appearance under various viewpoints, at fixed time and illumination setting. During testing, the algorithm is presented with images of the same rooms, acquired under roughly similar viewpoints but possibly under different illumination conditions, and after some time. The goal is to recognize each single image seen by the system. As in [15], we considered three sets of experiments for three types of problems of different complexity. In case of each single experiment, training was always performed on one image sequence subsampled to 1 fps (every fifth image), and



(a) Two-persons office

(b) Corridor

(c) Remaining rooms (Minnie at night)

Figure 1: Example pictures taken from the IDOL database showing the interiors of the rooms, variations occurring across platforms, as well as introduced by illumination changes and natural activity in the environment.

testing was done using a full sequence. The first set consisted of 12 experiments performed on different combinations of training and test data, acquired using the same robot platform and under similar illumination conditions. For the second set of experiments, we used 24 pairs of sequences captured under different illumination conditions. Finally, the third set was performed on 24 pairs of training and test sequences acquired under similar illumination settings but using a different robot. As a measure of performance we used the percentage of properly classified images calculated separately for each of the rooms and then averaged with equal weights independently of the number of images acquired in each room.

3.3 Image Representations

In this work, we employed two types of visual cues, global and local, extracted from the same image frame. As global representation we used the Composed Receptive Field Histograms (CRFH) [7], a multi-dimensional statistical representation of responses of several image filters. Computational costs were reduced by using a sparse and ordered histogram representation, as proposed in [7]. Following [15], we used histograms of 6 dimensions, with 28 bins per dimension, computed from second order normalized Gaussian derivative filters applied to the illumination channel at two scales.

We used the SIFT descriptor [8] in order to obtain the local image representation. SIFT represents the local image patches around interest points characterized by coordinates in the scalespace in the form of histograms of gradient directions. In order to find the coordinates of the interest points, we used the Harris-Laplacian detector [10], a scale invariant extension of the Harris corner detector.

3.4 Support Vector Machines

Consider the problem of separating the set of training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ into two classes, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label. If we assume that the two classes can be separated by a hyperplane in some Hilbert space $\mathcal{H}$, then the optimal separating hyperplane is the one which has maximum distance to the closest points in the training set resulting in a discriminant function

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{1}$$

The classification result is then given by the sign of $f(\boldsymbol{x})$. The values of α_i and b are found by solving a constrained minimization problem, which can be done efficiently using the SMO algorithm [3]. Most of the α_i's take the value of zero; those $\boldsymbol{x}_i$ with nonzero α_i are the "support vectors". In case where the two classes are non-separable, the optimization is formulated in such way that the classification error is minimized and the final solution remains identical.

The mapping between the input space and the usually high dimensional feature space $\mathcal{H}$ is done using the kernel function $K(\boldsymbol{x}_i, \boldsymbol{x})$. Several kernel functions have

been proposed for visual applications; in this paper we will use the χ^2 kernel [2] for the global CRFH descriptors, and the match kernel proposed in [19] for the local SIFT descriptors. Both have been used in our previous work on SVM-based place recognition, obtaining good performances.

The extension of SVM to multi class problems can be done mainly in two ways:

1. *One-against-All (OaA) strategy.* If M is the number of classes, M SVMs are trained, each separating a single class from all remaining classes. The decision is then based on the distance of the classified sample to each hyperplane. Typically algebraic distance ($f(\boldsymbol{x})$) is used and the final output is the class corresponding to the hyperplane for which the distance is largest.

2. *One-against-One (OaO) strategy.* In this case, $M(M-1)/2$ two-class machines are trained for each pair of classes. The final decision can then be taken in different ways, based on the $M(M-1)/2$ outputs. A popular choice is to consider as output of each classifier the class label and count votes for each class; the test image is then assigned to the class that received more votes. Another alternative is to use signed distance from the hyperplane and sum distances for each class. Other solutions based on the idea to arrange the pairwise classifiers in trees, where each tree node represents an SVM, have also been proposed [13, 3]. In this paper, we will use the voting-based method, which we found to constantly outperform the second alternative in our preliminary experiments.

4 Confidence Estimation

This section presents our approach to the problem of ranking hypotheses generated by the classifier and measuring their confidence. We first describe two methods based on the standard OaO and OaA multi-class extensions and our modified version of the OaA principle. Then, we show benchmark experiments evaluating the performance of the methods on visual data. The algorithms presented here will be one of the building block of the confidence-based cue integration scheme we will introduce in Section 6.

4.1 The algorithms

As already mentioned, discriminative classifiers do not provide any out-of-the-box solution for estimating confidence of the decision; however, it is possible to derive confidence information and hypotheses ranking from the distances between the samples and the hyperplanes. In case of SVMs, this can be done very efficiently thanks to the use of kernel functions and does not require additional processing in the training phase (as opposed to probability estimation methods like [12]). As it will be shown by experiments, despite its simplicity, such approach can yield good results when applied to complex problems such as visual place recognition.

We stress that, since it is based on the generated hyperplanes, performance will depend on how well the model reflects the statistics of the test data. In other words, the approach will work best for cases where the difficulty comes from the inability to perfectly separate the training samples and still provide good generalization capabilities.

It is straightforward to extend the standard OaO and OaA multi-class methods so that additional information about the decision becomes available. Let us present the methods using a more general notation and introduce a variable $V_h(\boldsymbol{x})$, which will be a distance-based score assigned by the hyperplane h to the sample $\boldsymbol{x}$. In case of the two standard algorithms, the score will just be equal to the distance of the test sample to the hyperplane: $V_h(\boldsymbol{x}) = D_h(\boldsymbol{x})$. Typically, the value of the discriminant function is used as a distance measure ($D_h(\boldsymbol{x}) = f_h(\boldsymbol{x})$). In order to find the best hypothesis j^*, we follow the rules described in Section 3.4:

- *for the OaO strategy:*

$$j^* = \underset{j=1\ldots M}{\mathrm{a}rgmax} \left|\{i : i \in \{1\ldots M\},\ i \neq j,\ V_{i,j}(\boldsymbol{x}) > 0\}\right|,$$

  where the indices i, j are used to denote the hyperplane separating class i from class j.

- *for the OaA strategy:*

$$j^* = \underset{j=1\ldots M}{\mathrm{a}rgmax} \left\{V_j(\boldsymbol{x})\right\},$$

  where V_j is the score assigned by the hyperplane separating class j from the other classes.

If now we think of the confidence as a measure of unambiguity of the decision, we can define it as:

- *for OaO,* the minimal score (distance) to the hyperplanes separating the first hypothesis and the other classes:

$$C(\boldsymbol{x}) = \min j = 1\ldots M,\ j \neq j^* \left\{V_{j^*,j}(\boldsymbol{x})\right\}$$

- *for OaA,* the difference between the maximal and the next largest score:

$$C(\boldsymbol{x}) = V_{j^*}(\boldsymbol{x}) - \max j = 1\ldots M,\ j \neq j^* \left\{V_j(\boldsymbol{x})\right\}$$

The value $C(\boldsymbol{x})$ can be thresholded for obtaining a binary confidence information. Confidence is then assumed if $C(\boldsymbol{x}) > \tau$ for threshold τ. The values $V_{j^*,j}(\boldsymbol{x})$ (for OaO) and $V_j(\boldsymbol{x})$ (for OaA) can also be used to rank the hypotheses and find between which of them the classifier is uncertain.

The output of the algorithms described above depends only on the distances of the test sample to the hyperplanes, that for SVMs is determined by the vectors lying

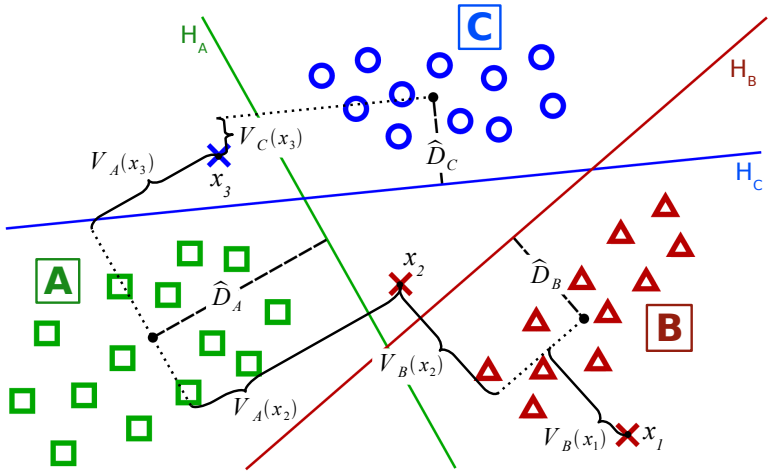


Figure 2: Artificial classification problem illustrating the way the scores $V_j(\boldsymbol{x})$ are calculated for classified samples in case of the modified OaA approach. We can observe that although the points $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are located approximately in the same distance from two hyperplanes, they are classified as belonging to class B and C respectively with high confidence.

close to the class boundaries (the support vectors). To make it more dependent on the distribution of all available training data, we suggest to use the OaA principle and redefine the score $V_j(\boldsymbol{x})$ to be equal to the distance from the average distance of the training samples to the hyperplane (see Fig. 2 for an illustration):

$$V_j(\boldsymbol{x}) = \left| \widehat{D}_j - D_j(\boldsymbol{x}) \right| .$$

Thus, we do not measure how far the test sample is from the hyperplane, but how close it is to the training data belonging to one of the classes. The best hypothesis can be determined by the following rule:

$$j^* = \underset{j=1\ldots M}{argmin} \left\{ V_j(\boldsymbol{x}) \right\} . \tag{2}$$

Using the same definition of confidence as above, we get:

$$C(\boldsymbol{x}) = \min j = 1 \ldots M,\ j \neq j^* \left\{ V_j(\boldsymbol{x}) \right\} - V_{j^*}(\boldsymbol{x}). \tag{3}$$

As in case of the previous algorithms, we can order the hypotheses using the values of $V_j(\boldsymbol{x})$ and obtain hard confidence information by thresholding. An explanation on a real example from one of our experiments is shown in Fig. 3.

4.2 Experimental Evaluation

We preformed a benchmark evaluation of the three confidence estimation methods (the two methods based on the standard OaO and OaA multi-class extensions and

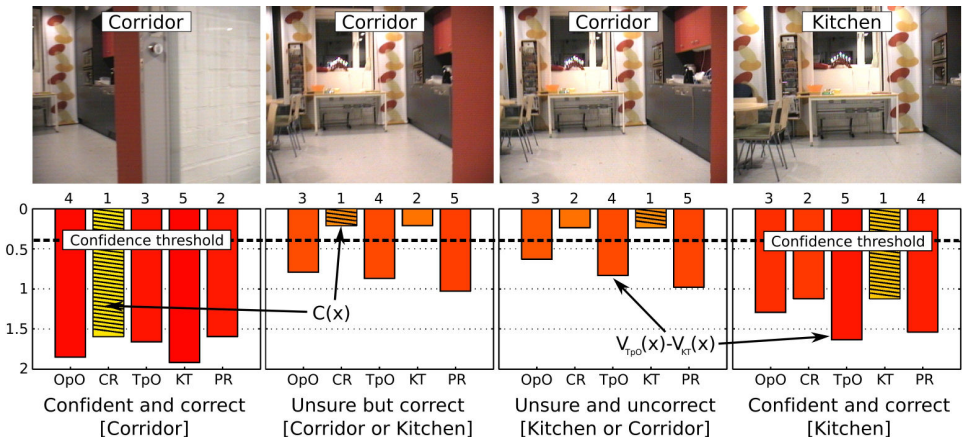
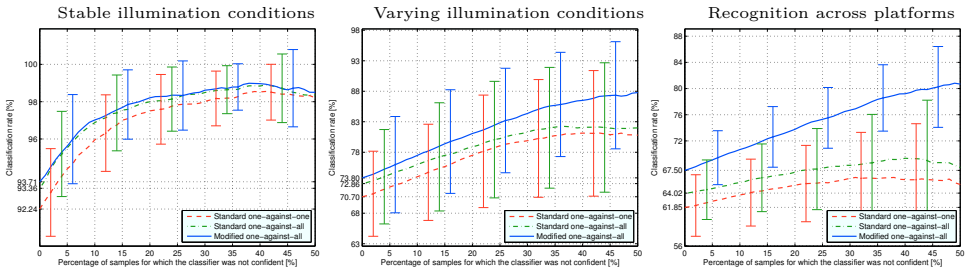


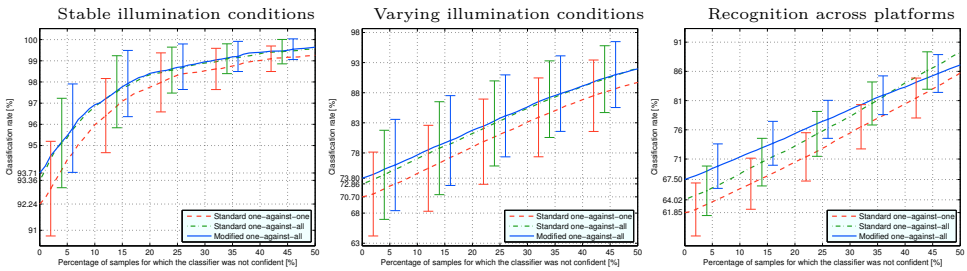
Figure 3: Real confidence estimates obtained using the modified OaA algorithm for four images acquired by the robot Minnie turning from the corridor towards the kitchen. According to the laser-based localization system used as ground truth, the first three images were acquired in the corridor, while the fourth image was already captured in the kitchen. The bar charts show the ranking of hypotheses (top axis), the estimated confidence of the decision (shaded bar), and the difference between the score for the best hypothesis and the others (remaining bars). For the confidence threshold set as shown in the figure, we obtain two soft decisions (suggesting the correct hypothesis) for the cases of lowest confidence.

the method based on the new modified version of OaA) on the IDOL database. As described in Section 3.2, we performed three sets of experiments: training and testing under stable illumination conditions, varying illumination conditions, and recognition across different robotic platforms. For all experiments we measured the performance of the algorithms for a range of values of the confidence threshold. We used two measures of performance in order to analyse different properties of the methods. First, for each value of the confidence threshold, we calculated the classification rate (percentage of properly classified test images) only for those test samples for which the decision was regarded as confident. As a second measure we used the classification rate calculated for all samples and including additional hypotheses between which the algorithm was unsure when the confidence was below the threshold. For example, if for a given threshold, the decision was "kitchen or corridor" and the test image was acquired in one of these rooms, the decision was counted as correct.

The average results obtained for the global features (CRFH) are presented in Fig. 4. The experiments were repeated also for local features (SIFT); however, the results showed the same trends and thus are omitted for space reasons (classification rates for local features and hard-decision SVMs can be found in Section 5). To obtain these results, we used the value of the discriminant function as a distance



(a) Classification rates for confident samples only



(b) Classification rates including all hypotheses below confidence threshold

Figure 4: Results of evaluation of the three confidence estimation algorithms on three types of problems.

measure ($D_h(\boldsymbol{x}) = f_h(\boldsymbol{x})$). We performed identical experiments for two other distance measures: the distance of a sample to its normal projection onto the hyperplane and relative distance normalized by the average class distance to the plane [5]; however, the results clearly showed the advantage of the solutions based on the value of $f(\boldsymbol{x})$.

The plots presented in Fig. 4 show the dependency between the classification rates and the percentage of images of the test sequence for which the classifier was not confident, given some value of threshold. The classification rates for hard-decision SVM are marked on the vertical axis (initial values, all decisions treated as confident). It can be observed that the classification rate calculated for the confident decisions only (Fig. 4a) is increasing for all methods as the percentage of unsure decisions grows. This means that the algorithms tend to eliminate the misclassified samples. It is clear that the modified OaA approach performs best with respect to this measure. Moreover, we can see that the method consistently delivers best classification rates when hard decisions are considered. The advantage in terms of classification rate varyes from +0.4% to +3.5% with respect to standard OaA and +1.5% to +5.7% with respect to standard OaO and grows with the complexity of the problem.

Additional conclusions can be drawn from the analysis of the second performance measure (Fig. 4b). First, we see that if we tolerate soft decisions in e.g.

30% of cases, the resulting classification rate increases from +5.2% (Fig. 4b, left) to +12% (Fig. 4b, right) and can even reach 99% in case of experiments performed for similar illumination conditions. Second, it is still visible that both OaA-based methods consistently outperform the OaO-based algorithm, and the modified version of the OaA strategy in general achieves the best performance. This time, however, the advantage of the modified OaA with respect to the algorithm based on the standard approach is smaller and decreases as the number of unsure decisions grows. This makes us conclude that the modified OaA method is better when it comes to finding and estimating confidence of the best hypothesis. However, the standard OaA-based algorithm is similarly or even more (Fig. 4b, right) efficient for ranking hypotheses. This property of the modified algorithm may become important if additional information could be used to improve classification results for the decisions in cases when the classifier is not confident enough.

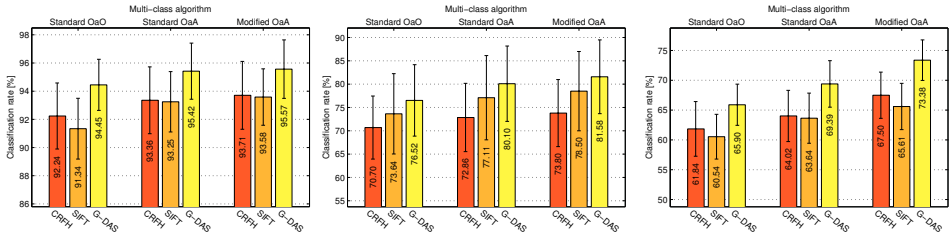
5 Cue Integration

Last section showed the importance of defining an effective confidence measure for SVM, and its impact on classification accuracy. Still, once the algorithm is able to measure an unsatisfactory level of confidence, it should react accordingly. The most desirable action should of course lead to higher confidence and accurate classification; one of the possible way to achieve this result is to use effectively multiple cues. In this section, we introduce a generalization of the integration scheme proposed in [11] to a wider class of multi-class extensions, and we present experimental evidence of its efficiency. How to combine these cue integration schemes with confidence-based classification approach will be the subject of Section 6.

5.1 Generalized Discriminative Accumulation Scheme

Suppose we are given M visual classes and, for each class, a set of n_j training images $\{\boldsymbol{I}_i^j\}_{i=1}^{n_j}$, $j = 1, \ldots, M$. Suppose also that, from each image, we extract a set of P different cues $\{T_p(\boldsymbol{I}_i^j)\}_{p=1}^{P}$ (the cues could also be different modalities). The goal is to perform recognition using all the cues. The original DAS algorithm consists of two steps:

1. *Single-cue SVMs.* From the original training set $\{\{\boldsymbol{I}_i^j\}_{i=1}^{n_j}\}_{j=1}^{M}$, containing images belonging to all M classes, define P new training sets $\{\{T_p(\boldsymbol{I}_i^j)\}_{i=1}^{n_j}\}_{j=1}^{M}$, $p = 1, \ldots, P$, each relative to a single cue. For each new training set train a multi-class SVM. In general, kernel functions may differ from cue to cue. Model parameters can be estimated during the training step via cross validation. In case of the original DAS algorithm, the standard OaA multi-class extension was used. Then, given a test image $\boldsymbol{I}$, for each single-cue SVM the algebraic distance to each hyperplane $f_j^p(T_p(\boldsymbol{I}))$, $j = 1, \ldots, M$ was computed according to Eq. 1.



(a) Stable illumination (b) Varying illumination (c) Recognition across platforms

Figure 5: Average results for G-DAS based on different multi-class extensions, for three different series of experiments.

2. *Discriminative Accumulation.* After all the distances were collected $\{f_j^p\}_{p=1}^P$, for all the M hyperplanes and the P cues, the image $\boldsymbol{I}$ was classified using their linear combination:

$$j^* = \underset{j=1}{\overset{M}{\arg\max}} \left\{ \sum_{p=1}^{P} a_p f_j^p(T_p(\boldsymbol{I})) \right\}, \; a_p \in \Re^+.$$

The coefficients $\{a_p\}_{p=1}^P$ can also be evaluated via cross validation during training.

The original algorithm performed accumulation at the level of algebraic distances from the hyperplanes f_j^p, obtained from a standard OaA multi-class SVM. As shown in Section 4, there are other methods available, and it is possible to introduce more effective multi-class algorithms based on the OaA principle. We thus propose to extend the DAS framework to be applicable also for the other methods; we call this extension the Generalized Discriminative Accumulation Scheme (G-DAS). The discriminative accumulation is here performed at the level of the scores V_h (see Section 4):

$$V_h^{\Sigma P}(\boldsymbol{I}) = \sum_{p=1}^{P} a_p V_h^p(T_p(\boldsymbol{I})), \; a_p \in \Re^+. \tag{4}$$

As a result, any multi-class extension can be used within the G-DAS framework (both OaA and OaO based).

5.2 Experimental Evaluation

We evaluated the effectiveness of G-DAS for the visual place recognition problem by running the three series of experiments described in Section 3.2. SIFT and CRFH were used as features, χ^2 and match kernel as similarity measures for the nonlinear SVMs, and kernel parameters as well as weighting coefficients for the accumulation schemes were determined via cross validation.

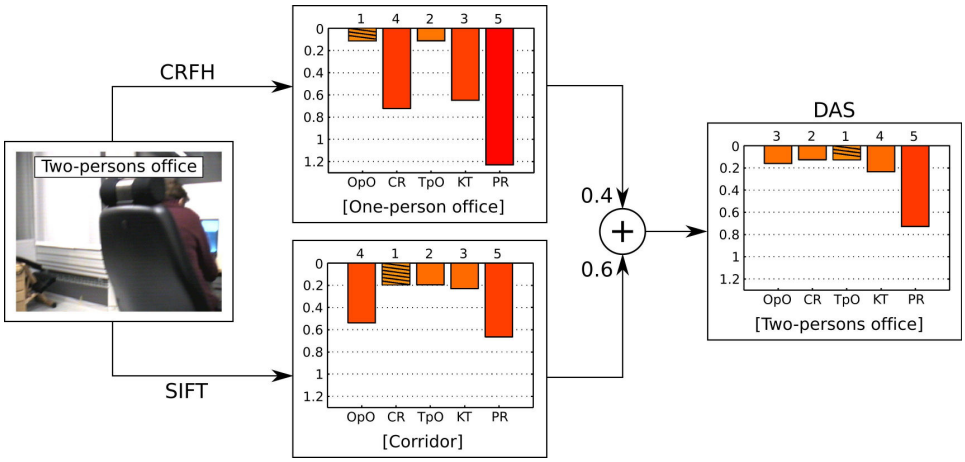
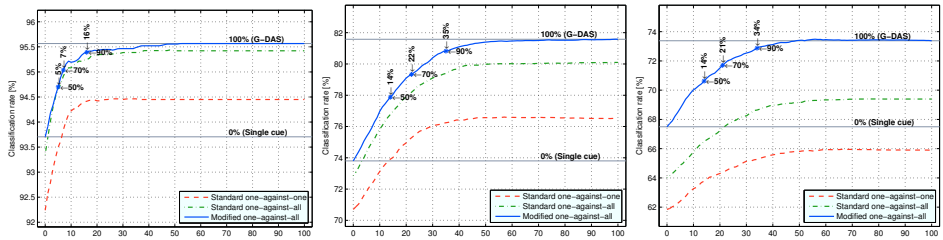


Figure 6: An example of test image misclassified by using a single cue, but classified correctly by using G-DAS with modified OaA multi-class extension (see Fig. 3 for an explanation of the bar charts).

Fig. 5 shows the recognition results obtained using a single cue SVM, with global or local descriptors, and those obtained using the G-DAS algorithm. For all those three different approaches, we used three different multi-class extensions: standard OaO, standard OaA and our new modified OaA. Note that, when using standard OaA, G-DAS corresponds to the original DAS algorithm.

A first comment is that for all three different multi-class extensions, for all the three series of experiments, the accumulation scheme clearly achieves consistently better results than the single cues approaches. The gain in performance goes from a minimum of +1.9% in accuracy, obtained for the stable illumination condition experiments (Fig. 5a) to a maximum of +7.8%, obtained for the varying illumination (Fig. 5b), with respect to the CRFH only, using the modified OaA approach. The increase in performance grows with the difficulty of the task and is on average a +2% for stable illumination, +5% for varying illumination and +6% for recognition across platforms. A second comment is that G-DAS with our modified OaA consistently performs better than the original DAS, for all the three scenarios; this confirms the effectiveness of this confidence measure for visual recognition. An important property of the DAS algorithm, which is also preserved by G-DAS, is the ability to classify correctly images even when each of the single cues used gives misleading information. Fig. 6 shows an example of this behavior: the test image is misclassified as 'one-person office' by using CRFH, and as 'corridor' by using SIFT; by combining these two cues in G-DAS, the image is correctly classified as 'two-persons office'. We can then conclude that G-DAS is an effective method for cue integration for visual place recognition in realistic settings.



(a) Stable illumination (b) Varying illumination (c) Recogn. across platforms

Figure 7: Dependencies between the average classification rates obtained for the confidence-based cue integration strategy with CRFH used as a primary cue and the percentage of test samples for which both cues were used. The horizontal lines indicate the performance of CRFH only and G-DAS.

6 Confidence-based Cue Integration

As motivated in Section 5, a desirable behaviour of a system aware of its own ignorance would be to search for additional sources of information in order to achieve higher confidence. We showed that G-DAS can be effectively used for visual cue integration; however, it requires that both cues are available and used for classification even in cases when one cue is sufficient to obtain correct result, and the additional computational effort could be avoided. In this section, we present and experimentally evaluate a strategy allowing to greatly decrease this computational load and still maintain the high level of accuracy provided by multiple cues. We propose to employ G-DAS for cue accumulation and extract the additional information only in cases when the confidence of a decision based on the cues available is not satisfactory.

6.1 The method

It is reasonable to assume that the confidence estimation methods presented in Section 4 can be used as efficient filters, filtering out the images for which G-DAS would be either not required or not effective. First, the experiments reported in Section 4 proved that the confidence estimation methods are able to eliminate the incorrect decisions. Second, both methods and the G-DAS framework operate on the distances calculated in the high dimensional feature space, and G-DAS is expected to be most effective in cases of low confidence (see the example in Fig. 6).

Suppose again that we are able to extract P different cues $\{T_p(\boldsymbol{I})\}_{p=1}^{P}$ from the input image $\boldsymbol{I}$. Let us assume that the cues are ordered. The order of the cues can be motivated by the computational cost associated with feature extraction and classification. To obtain the final decision we use the following algorithm:

1. Set $k = 1$.

2. Extract features for the k^{th} cue ($T_k(\boldsymbol{I})$).

3. Perform classification for the k^{th} cue and obtain the scores $V_h^k(T_k(\boldsymbol{I}))$ for all hyperplanes.

4. Perform cue integration for the cues $1 \dots k$ according to Eq. 4 and obtain the accumulative scores $V_h^{\Sigma k}(\boldsymbol{I})$.

5. Find the best hypothesis j_k^* and confidence estimates $C_k(I)$ based on the scores $V_h^{\Sigma k}(\boldsymbol{I})$.

6. If the confidence is below the threshold ($C_k(I) < \tau$) and $k < P$, increment k and go to step 2. Otherwise, use the obtained hypothesis as final decision ($j^* = j_k^*$).

The threshold τ is a parameter of the algorithm and allows to trade the accuracy for computational cost.

6.2 Experimental Evaluation

We performed an experimental evaluation of the confidence-based cue integration strategy for the global (CRFH) and local (SIFT) visual cues. We tested solutions based on both CRFH and SIFT used as a primary cue. The experiments showed the advantage of the CRFH-based solution in terms of the number of images for which both cues had to be used to obtain accuracy identical with the one offered by G-DAS. Moreover, the local features are much more computationally expensive mainly due to matching process performed during classification.

In this paper, we report results for CRFH used as a primary cue. The plots presented in Fig. 7 clearly show that in order to obtain accuracy comparable with the one delivered by G-DAS used for all test images, it is necessary to use the second cue only in approximately 40% of cases. This is for the modified OaA multi-class extension, which once more outperformed the other confidence estimation methods. As already mentioned, in our case, feature extraction and classification was more costly for the local cue. As a result, the strategy presented here allowed to reduce the amount of computations by about 55% in average compared to G-DAS. Since the dependency between the number of images for which the second cue is used and the classification rate is highly non-linear, it can be advantageous to trade the accuracy for computational cost; e.g. to achieve gain of 70% of the one provided by G-DAS, the second cue should be used in 7% (stable illumination conditions, Fig. 7a) to 22% (varying illumination conditions, Fig. 7b) of cases only. Concluding, the power of multiple cues can be achieved for much lower computational cost, if information about the classifier's confidence is exploited.

7 Summary and Conclusion

This paper presented an effective approach to the problems of confidence estimation and cue integration for large-margin discriminative classifiers. We showed by extensive experiments, on problems of different complexity from the domain of visual place recognition, that exploiting available confidence information encoded in the classifier's outputs can greatly increase reliability of a system. When combined with a cue integration scheme, this results in a significantly increased performance for a relatively low computational cost. We used SVMs and combined local and global cues extracted from the same visual stimuli; all the presented methods could easily be extended to other large margin classifiers and to multiple modalities.

The potential of this approach can be used in many ways. First, we plan to incorporate confidence information to an incremental learning framework and use it to trigger the learning procedure. Second, we want to create an active system able to autonomously search for cues in order to obtain confident result. Finally, we will test our method in a multi-modal system and for integration of a larger number of cues.

References

[1] John Aloimonos and David Shulman. *Integration of visual modules: An extension of the Marr paradigm.* Academic Press, 1989.

[2] Oliver Chapelle, Patrick Haffner, and Vladimir Vapnik. Support Vector Machines for Histogram-Based Image Classification. *IEEE Transactions on Neural Networks*, 10 (5), 1999.

[3] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, 2000. ISBN 0521780195.

[4] Sarah Jane Delany, Pádraig Cunningham, Dónal Doyle, and Anton Zamolotskikh. Generating estimates of classification confidence for a case-based spam filter. In *Proceedings of the 6th International Conference on Case-Based Reasoning (ICCBR'05)*, Chicago, IL, USA, 2005.

[5] Jae-Jin Kim, Bon-Woo Hwang, and Seong-Whan Lee. Retrieval of the top N matches with Support Vector Machines. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'00)*, Barcelona, Spain.

[6] David M. Kortenkamp and Terry Weymouth. Topological mapping for mobile robots using a combination of sonar and vision sensing. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, Seattle, Washington, USA, 1994.

[7] Oskar Linde and Tony Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proceedings of the 2004 15th International Conference on Pattern Recognition (ICPR'04)*, 2004.

[8] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[9] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. The KTH-IDOL2 database. Technical report, Kungliga Tekniska Högskolan, CVAP/CAS, September 2006.

[10] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV'01)*, Vancouver, Canada, 2001.

[11] Maria-Elena Nilsback and Barbara Caputo. Cue integration through discriminative accumulation. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, DC, USA, 2004.

[12] John C. Platt. Probabilistic outputs for Support Vector Machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 1999.

[13] John C. Platt, Nello Cristianini, and John Shawe-Taylor. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing Systems (NIPS)*, volume 12, pages 547–553, 2000.

[14] Tomaso Poggio, Vincent Torre, and Christof Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[15] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I. Christensen. A discriminative approach to robust visual place recognition. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, October 2006.

[16] Stephen Se, David G. Lowe, and James J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation (ICRA'01)*, Seoul, Korea, 2001.

[17] Adriana Tapus and Roland Siegwart. Incremental Robot Mapping with Fingerprints of Places. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Edmonton, Alberta, Canada, 2005.

[18] Iwan Ulrich and Illah Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation (ICRA'00)*, San Francisco, CA, USA, 2000.

[19] Christian Wallraven, Barbara Caputo, and Arnulf Graf. Recognition with local features: The kernel recipe. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, Beijing, China, 2003.

[20] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research (JMLR)*, 5:975–1005, 2004.

Paper D

Multi-modal Semantic Place Classification

Andrzej Pronobis, Oscar Martínez Mozos,
Barbara Caputo and Patric Jensfelt

Published in
The International Journal of Robotics Research

Multi-modal Semantic Place Classification

Andrzej Pronobis, Oscar Martínez Mozos,
Barbara Caputo and Patric Jensfelt

Abstract

The ability to represent knowledge about space and its position therein is crucial for a mobile robot. To this end, topological and semantic descriptions are gaining popularity for augmenting purely metric space representations. 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 and 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. The results clearly demonstrate the effectiveness of our cue integration scheme and its value for robust place classification under varying conditions.

1 Introduction

The most fundamental competence for an autonomous mobile agent is to know its position in the world. This can be represented in terms of raw metric coordinates, topological location, or even semantic description. Recently, there has been a growing interest in augmenting (or even replacing) purely metric space representations with topological and semantic place information. Several attempts have been made to build autonomous cognitive agents able to perform human-like tasks [2, 1]. Enhancing the space representation to be more meaningful from the point of view of spatial reasoning and human-robot interaction have been at the forefront of the issues being addressed [64, 25, 53]. Indeed, in the concrete case of indoor environments, the ability to understand the existing topological relations and associate semantic terms like "corridor" or "office" with places, gives a much more intuitive idea of the position of the robot than global metric coordinates. Additionally, the

semantic information about places can extend the capabilities of a robot in other tasks such as localization [45], exploration [49], or navigation [21].

Nowadays, robots are usually equipped with several sensors providing both geometrical and visual information about the environment. Previous work on place classification relied on sonars and/or laser range data as robust sensory modalities [33]. However, the advantages of geometric solutions, such as invariance to visual variations and low dimensionality of the processed information, quickly became their weaknesses. The inability to capture many aspects of complex environments leads to the problem of perceptual aliasing [26] and can limit the usefulness of such methods for topological and semantic mapping. Recent advances in vision have made this modality emerge as a natural and viable alternative. Vision provides richer sensory input allowing for better discrimination. Moreover, a large share of the semantic description of a place is encoded in its visual appearance. However, visual information tends to be noisy and difficult to interpret as the appearance of places varies over time due to changing illumination and human activity. At the same time, the visual variability within place classes is huge, making the semantic place classification a challenging problem. Clearly, each modality has its own characteristics. Interestingly, the weaknesses of one often correspond to the strengths of the other.

In this paper, we propose an approach to semantic place classification which combines the stability of geometrical solutions with the versatility of vision. First, we present a recognition system implemented on a mobile robot platform integrating multiple cues and modalities. The system is able to perform robust place classification under different types of variations that occur in indoor environments over a span of time of several months. This comprises variations in illumination conditions and in configuration of furniture and small objects. The system relies on different types of visual information provided by global and local descriptors and on geometric cues derived from laser range scans. For the vision channel we apply the Scale-invariant Feature Transform (SIFT, [28]) and Composed Receptive Fields Histograms (CRFH, [27]). For the laser channel we use the features proposed in [33, 34].

We combine the cues using a new high-level accumulation scheme, which builds on our previous work [42, 36]. We train for each cue a large margin classifier which outputs a set of scores encoding confidence of the decision. Integration is then achieved by feeding the scores to a Support Vector Machine (SVM, [14]). Such an approach allows to optimally combine cues, even obtained using different types of models, with a complex, possibly non-linear function. We call this algorithm the SVM-based Discriminative Accumulation Scheme (SVM-DAS).

Finally, we show how to build a self-contained semantic space labeling system, which relies on multi-modal place classification as one of its components. The system is implemented as a part of an integrated cognitive robotic architecture [2, 23] and runs on-line on a mobile robot platform. While the robot explores the environment, the system acquires evidences about the semantic category of the current area produced by the place classification component and accumulates them both

over time and space. As soon as the system is confident about its decision, the area is assigned a semantic label. We integrate the system with a Simultaneous Localization and Mapping (SLAM) algorithm and show how a metric and topological space representation can be augmented with a semantic description.

We evaluated the robustness of the presented methods in several sets of extensive experiments. We conducted experiments on two different robot platforms, in two different environments and for two different scenarios. First, we run a series of off-line experiments of increasing difficulty on the IDOL2 database [29] to precisely measure the performance of the place classification algorithm in presence of different types of variations. These ranged from short-term visual variations caused by changing illumination to long-term changes which occurred in the office environment over several months. Second, we run a live experiment where a robot performs SLAM and semantic labeling in a new environment using prebuilt models of place categories. Results show that integrating different visual cues, as well as different modalities, allows to greatly increase the robustness of the recognition system, achieving high accuracy under severe dynamic variations. Moreover, the place classification system, when used in the framework of semantic space labeling, can yield a fully correct semantic representation even for a new, unknown environment.

The rest of the paper is organized as follows: after a review of the related literature (Section 2), Section 3 presents the main principle behind our multi-modal place classification algorithm and describes the methods used to extract each cue. Then, Section 4 gives details about the new cue integration scheme and Section 5 describes the architecture of the semantic labeling system. Finally, Section 6 presents detailed experimental evaluation of the place classification system and Section 7 reports results of the live experiment with semantic labeling of space. The paper concludes with a summary and possible avenues for future research.

2 Related Work

Place classification is a vastly researched topic in the robotic community. Purely geometric solutions based on laser range data have proven to be successful for certain tasks and several approaches were proposed using laser scanners as the only sensors. [24] used a pre-programmed routine to detect doorways from range data. Additionally, [4] used line features to detect corridors and doorways. In their work, [10] 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. Finally, [33] applied boosting to create a classifier based on a set of geometrical features extracted from range data to classify different places in indoor environments. A similar idea was used by [53] to describe regions from laser readings.

The limitations of geometric solutions inspired many researchers to turn towards vision which nowadays becomes tractable in real-time applications. The proposed methods employed either perspective [55, 50, 18] or omnidirectional cameras [22,

8, 58, 32, 5, 35, 59]. The main differences between the approaches relate to the way the scene is perceived, and thus the method used to extract characteristic features from the scene. Landmark-based techniques make use of either artificial or natural landmarks in order to extract information about a place. [47] relied on visually distinctive image regions as landmarks. Many other solutions employed local image features, with SIFT being the most frequently applied [28, 46, 5, 42]. [65] used the SIFT descriptor to build a topological representation by clustering a graph representing relations between images. Other approaches used the bag-of-words technique [18, 19], the SURF features [7, 35, 59], or representation based on information extracted from local patches using Kernel PCA [50]. Global features are also commonly used for place recognition. Torralba *et al.* [56, 55, 54] suggested to use a representation called the "gist" of a scene, which is a vector of principal components of outputs of a bank of spatially organized filters applied to the image. Other approaches use color histograms [58, 8], gradient orientation histograms [9], eigenspace representation of images [22], or Fourier coefficients of low frequency image components [32].

In all the previous approaches only one modality is used for the recognition of places. Recently, several authors observed that robustness and efficiency of the recognition system can be improved by combining information provided by different visual cues. [47, 61] used a global representation of the images together with local visual landmarks to localize a robot in outdoor environments. [42] used two cues composed of global and local image features to recognize places in indoor environments. The cues were combined using discriminative accumulation. Here, we extend this approach by integrating information provided by a laser range sensor using a more sophisticated algorithm.

Other approaches also employed a combination of different sensors, mainly laser and vision. [51] used and omnidirectional camera and two lasers covering 360 degrees field of view to extract fingerprints of places for topological mapping. This approach was not used for extracting semantic information about the environment. [41] and [16] relied on range data and vision for recognition of objects in outdoor environments (e.g. grass, walls or cars). Finally, [45] used a combination of both modalities to categorize places in indoor environments. Each observation was composed of a set of geometrical features and a set of objects found in the scene. The geometrical features were calculated from laser scans and the objects were detected using Haar-like features from images. The extracted information was integrated at the feature level. In contrast, the method presented in this work learns how to combine and weight outputs of several classifiers, keeping features and therefore the information from different modalities separated.

Various cue integration methods have been proposed in the robotics and machine learning community [51, 36, 42, 39, 57, 31]. These approaches can be described according to various criteria. For instance, [13] suggest to classify them into two main groups, *weak coupling* and *strong coupling*. Assuming that each cue is used as input of a different classifier, weak coupling is when the output of two or more independent classifiers are combined. Strong coupling is instead when the output of

one classifier is affected by the output of another classifier, so that their outputs are not anymore independent. Another possible classification is into *low level* and *high level* integration methods, where the emphasis is on the level at which integration happens. We call *low level integration methods* those algorithms where cues are combined together at the feature level, and then used as input to a single classifier. This approach has been used successfully for object recognition using multiple visual cues [31], and for topological mapping using multiple sensor modalities [51]. In spite of remarkable performances for specific tasks, there are several drawbacks of the low level methods. First, if one of the cues gives misleading information, it is quite probable that the new feature vector will be adversely affected influencing the whole performance. Second, we can expect the dimension of such a feature vector to increase as the number of cues grows, and each of the cues needs to be used even if one would allow for correct classification. This implies longer learning and recognition times, greater memory requirements and possibly curse of dimensionality effects. Another strategy is to keep the cues separated and to integrate the outputs of individual classifiers, each trained on a different cue [39, 36, 42]. We call such algorithms *high level integration methods*, of which voting is the most popular [17]. These techniques are more robust with respect to noisy cues or sensory channels, allow to use different classifiers adapted to the characteristics of each single cue and decide on the number of cues that should be extracted and used for each particular classification task [42]. In this paper, we focus on a weak coupling, high level integration method called *accumulation*. The underlying idea is that information from different cues can be summed together, thus accumulated. The idea was first proposed in probabilistic framework by [39] and further explored by [3]. The method was then extended to discriminative methods in [36, 42].

3 Multi-modal Place Classification

The ability to integrate multiple cues, possibly extracted from different sensors, is an important skill for a mobile robot. Different sensors usually capture different aspects of the environment. Therefore using multiple cues leads to obtaining a more descriptive representation. The visual sensor is an irreplaceable source of distinctive information about a place. However, this information tends to be noisy and difficult to analyze due to the susceptibility to variations introduced by changing illumination and everyday activities in the environment. At the same time, most recent robotic platforms are equipped with a laser range scanner which provides much more stable and robust geometric cues. These cues, however, are unable to uniquely represent the properties of different places (perceptual aliasing, [26]). Clearly performance could increase if different cues were combined effectively. Note that even alternative interpretations of the information obtained by the same sensor can be valuable, as we will show experimentally in Section 6.

This section describes our approach to multi-modal place classification. Our method is fully supervised and assumes that during training, each place (room) is

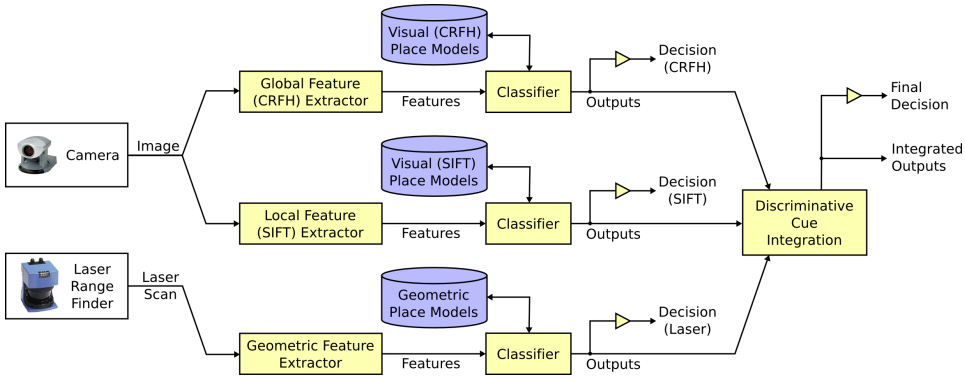


Figure 1: Architecture of the multi-modal place classification system.

represented by a collection of labeled data which captures its intrinsic visual and geometric properties under various viewpoints, at a fixed time and illumination setting. During testing, the algorithm is presented with data samples acquired in the same places, under roughly similar viewpoints but possibly under different conditions (e.g. illumination), and after some time (where the time range goes from some minutes to several months). The goal is to recognize correctly each single data sample provided to the system.

The architecture of the system is illustrated in Figure 1. We see that there is a separate path for each cue. We use two different visual cues corresponding to two types of image features (local and global) as well as simple geometrical features extracted from laser range scans. Each path consists of two main building blocks: a feature extractor and a classifier. Thus, separate decisions can be obtained for each of the cues in case only one cue is available. Alternatively, our method could decide when to acquire additional information (e.g. only in difficult cases, [42]). In cases when several cues are available, the outputs encoding the confidence of the single-cue classifiers are combined using an efficient discriminative accumulation scheme.

The rest of this section gives details about the algorithms used to extract and classify each of the cues for the vision-based paths (Section 3.1) and laser-based path (Section 3.2). A comprehensive description of the algorithms used for cue integration is given in Section 4.

3.1 Vision-based Place Classification

As a basis for the vision-based channel, we used the place recognition system presented in [43, 42], which is built around a Support Vector Machine (SVM) classifier [14] and two types of visual features, global and local, extracted from the same image frame. We used Composed Receptive Field Histograms (CRFH) [27] as global

features, and SIFT [28] as local features. Both have already been proved successful in the domain of vision-based place recognition [43, 42] and localization and mapping [46, 5].

CRFHs are a sparse multi-dimensional statistical representation of responses of several image filters applied to the input image. Following [43], we used histograms of 6 dimensions, with 28 bins per dimension, computed from second order normalized Gaussian derivative filters applied to the illumination channel at two scales. The SIFT descriptor instead represents local image patches around interest points characterized by coordinates in the scalespace in the form of histograms of gradient directions. To find the coordinates of the interest points, we used a scale and affine invariant region detector based on the difference-of-Gaussians (DoG) operator [44].

We used SVMs for creating models from both visual cues. A review of the theory behind SVMs can be found in Section 4.1. In case of SVMs, special care must be taken in choosing an appropriate kernel function. Here we used the χ^2 kernel [12] for CRFH, and the match kernel proposed by [60] for SIFT. Both have been used in our previous work on SVM-based place recognition, obtaining good performances.

3.2 Laser-based Place Classification

In addition to the visual channel, we used a laser range sensor. A single 2D laser scan covered a field of view of 180° in front of the robot. A laser observation $z = \{b_0, ..., b_{M-1}\}$ contains a set of beams b_i, in which each beam b_i consists of a tuple (α_i, d_i), where α_i is the angle of the beam relative to the robot and d_i is the length of the beam.

For each laser observation, we calculated a set of simple geometric features represented by single real values. The features were introduced for place classification by [33] where laser observations covering a 360° field of view were used. The complete set of features consists of two subsets. The first subset contains geometrical features calculated directly from the laser beams. The second subset comprises geometrical features extracted from a polygon approximation of the laser observation. This polygon is created by connecting the end points of the beams. The selection of features is based on the results presented in [33, 34].

As classifiers for the laser-based channel, we tried both AdaBoost [20], following the work in [34], and SVMs. In the rest of the paper, we will refer to the two laser-based models as L-AB and L-SVM, respectively. For the geometric features, we used a Radial Basis Function (RBF) kernel [14] with SVMs, chosen through a set of reference experiments. In case of AdaBoost, we constructed a multi-class classifier by arranging several binary classifiers into a decision list in which each element corresponded to one specific class. Both classifiers were benchmarked on the laser-based place classification task. Results presented in Section 6.2 show an advantage of the more complex SVM classifier.

4 Discriminative Cue Integration

This section describes our approach to cue integration from one or multiple modalities. We propose an SVM-based Discriminative Accumulation Scheme (SVM-DAS), a technique performing non-linear cue integration by discriminative accumulation. For each cue, we train a separate large margin classifier which outputs a set of scores (outputs), encoding the confidence of the decision. We achieve integration by feeding the scores to an SVM. Compared to previous accumulation methods [39, 11, 36, 42], SVM-DAS gives several advantages: *(a)* discriminative accumulation schemes achieve consistently better performances than probabilistic ones [39, 11], as shown in [36]; *(b)* compared to previous discriminative accumulation schemes [36, 42], our approach accumulates cues with a more complex, possibly non-linear function, by using the SVM framework and kernels [14]. Such an approach makes it possible to integrate outputs of different classifiers such as SVM and AdaBoost. At the same time, it learns the weights for each cue very efficiently, therefore making it possible to accumulate large numbers of cues without computational problems.

In the rest of the section we first sketch the theory behind SVMs (Section 4.1), a crucial component in our approach. We then describe the Generalized Discriminative Accumulation Scheme (G-DAS, [42], Section 4.2) on which to a large extent we build. Finally, we introduce the new algorithm and discuss its advantages in Section 4.3.

4.1 Support Vector Machines

Consider the problem of separating the set of labeled training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n), \ldots, (\boldsymbol{x}_N, y_N)$ into two classes, where $\boldsymbol{x}_n \in \Re^L$ is a feature vector and $y_n \in \{-1, +1\}$ its class label. Assuming that the two classes can be separated by a hyperplane in some Hilbert space $\mathcal{H}$, then the optimal separating hyperplane is the one which has maximum distance to the closest points in the training set resulting in a discriminant function

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n y_n \mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}) + \beta.$$

The classification result is then given by the sign of $f(\boldsymbol{x})$. The values of α_n and β are found by solving a constrained minimization problem, which can be done efficiently using the SMO algorithm [38]. Most of the α_n's take the value of zero; those $\boldsymbol{x}_n$ with nonzero α_n are the "support vectors". In case where the two classes are non-separable, the optimization is formulated in such way that the classification error is minimized and the final solution remains identical. The mapping between the input space and the usually high dimensional feature space $\mathcal{H}$ is done using kernels $\mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x})$.

The extension of SVM to multi class problems can be done in several ways. Here we will mention three approaches used throughout the paper:

1. *Standard one-against-all (OaA) strategy.* If M is the number of classes, M SVMs are trained, each separating a single class from all other classes. The decision is then based on the distance of the classified sample to each hyperplane, and the sample is assigned to the class corresponding to the hyperplane for which the distance is largest.

2. *Modified one-against-all strategy.* In [42], a modified version of the OaA principle was proposed. The authors suggested to use distances to precomputed average distances of training samples to the hyperplanes (separately for each of the classes), instead of the distances to the hyperplanes directly. In this case, the sample is assigned to the class corresponding to the hyperplane for which the distance is smallest. Experiments presented in this paper and in [42] show that in many applications this approach outperforms the standard OaA technique.

3. *One-against-one (OaO) strategy.* In this case, $M(M-1)/2$ two-class SVMs are trained for each pair of classes. The final decision can then be taken in different ways, based on the $M(M-1)/2$ outputs. A popular choice is to consider as output of each classifier the class label and count votes for each class; the test image is then assigned to the class that received more votes.

Support Vector Machines do not provide any out-of-the-box solution for estimating confidence of the decision; however, it is possible to derive confidence information and hypotheses ranking from the distances between the samples and the hyperplanes. In the work presented in this paper, we applied the distance-based methods proposed in [42], which define confidence as a measure of unambiguity of the final decision related to the differences between the distances calculated for each of the binary classifiers.

4.2 Generalized Discriminative Accumulation Scheme

The Generalized Discriminative Accumulation Scheme (G-DAS) was proposed first in [42], as a more effective generalization of the algorithm presented in [36]. It accumulates multiple cues, possibly from different modalities, by turning classifiers into experts. The basic idea is to consider real-valued outputs of a multi-class discriminative classifier (e.g. SVM) as an indication of a soft decision for each class. Then, all the outputs obtained from the various cues are summed together, therefore linearly accumulated. Specifically, suppose we are given M classes and, for each class, a set of N_m training samples $\{\{\boldsymbol{s}_{m,n}\}_{n=1}^{N_m}\}_{m=1}^{M}$. Suppose also that, from each sample, we extract a set of T different cues $\{\mathcal{T}_t(\boldsymbol{s}_{m,n})\}_{t=1}^{T}$. The goal is to perform recognition using all of them. The G-DAS algorithm consists of two steps:

1. *Single-cue Models.* From the original training set $\{\{\boldsymbol{s}_{m,n}\}_{n=1}^{N_m}\}_{m=1}^{M}$, containing samples belonging to all M classes, define T new training sets $\{\{\mathcal{T}_t(\boldsymbol{s}_{m,n})\}_{n=1}^{N_m}\}_{m=1}^{M}$, $t = 1, \ldots, T$, each relative to a single cue. For each

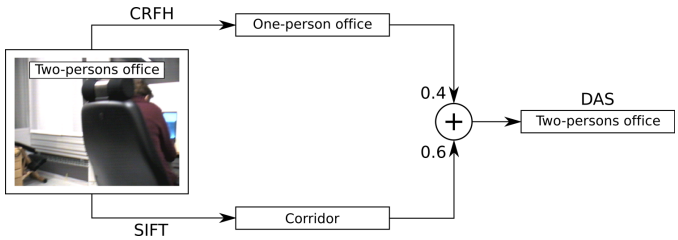


Figure 2: A real example of test image misclassified by each of the single cues, but classified correctly using G-DAS.

new training set train a multi-class classifier. Then, given a test sample $\boldsymbol{s}$, for each of the T single-cue classifiers estimate a set of outputs $\{\mathcal{V}_{t,v}(\mathcal{T}_t(\boldsymbol{s}))\}_{v=1}^{V}$ reflecting the relation of the sample to the model. In case of the SVMs with standard OaO and OaA multi-class extensions, the outputs would be values of the discriminant functions learned by the SVM algorithm during training, i.e. $\mathcal{V}_{t,v}(\mathcal{T}_t(\boldsymbol{s})) = f_{t,v}(\mathcal{T}_t(\boldsymbol{s}))$, $v = 1, \ldots, V$, and $V = M(M-1)/2$ for OaO or $V = M$ for OaA.

2. *Discriminative Accumulation.* After all the outputs are computed for all the cues, combine them with different weights by a linear function:

$$\mathcal{V}_v(\boldsymbol{s}) = \sum_{t=1}^{T} \sigma_t \mathcal{V}_{t,v}(\mathcal{T}_t(\boldsymbol{s})),\ \sigma_t \in \Re^+,\ v = 1, \ldots, V.$$

The final decision can be estimated with any method commonly used for multi-class, single-cue SVM.

An important property of accumulation is the ability to perform correct classification even when each of the single cues gives misleading information. This behavior is illustrated on a real example in Figure 2. Despite these advantages, G-DAS presents some potential limitations: first, it uses only one weight for all outputs of each cue. This simplifies the parameter estimation process (usually, an extensive search is performed to find the coefficients $\{\sigma_t\}_{t=1}^{T}$), but also constraints the ability of the algorithm to adapt to the properties of each single cue. Second, accumulation is obtained via a linear function, which might not be sufficient in case of complex problems. The next section shows how our new accumulation scheme, SVM-DAS, addresses these issues.

4.3 SVM-based Discriminative Accumulation Scheme

The Support Vector Machines-based Discriminative Accumulation Scheme (SVM-DAS) accumulates the outputs generated by single-cue classifiers by using a more complex, possibly non-linear function. The outputs are used as an input to an SVM,

and the parameters of the integration function are learned during the optimization process, for instance using the SMO algorithm [38]. These characteristics address the potential drawbacks of G-DAS discussed in the previous section.

More specifically, the new SVM-DAS accumulation function is given by:

$$\mathcal{V}_u(\boldsymbol{s}) = \sum_{n=1}^{N} \alpha_{u,n} y_n \mathcal{K}(\boldsymbol{v}_n, \boldsymbol{v}) + \beta_u, \;\; u = 1, \ldots, U,$$

where $\boldsymbol{v}$ is a vector containing all the outputs for all T cues:

$$\boldsymbol{v} = \left[\{\mathcal{V}_{1,v}(\mathcal{T}_1(\boldsymbol{s}))\}_{v=1}^{V_1}, \ldots, \{\mathcal{V}_{T,v}(\mathcal{T}_T(\boldsymbol{s}))\}_{v=1}^{V_T} \right].$$

The parameters $\alpha_{u,n}$, y_n, β_u, and the support vectors $\boldsymbol{v}_n$ are inferred from the training data either directly or efficiently during the optimization process. The number of the final outputs U and the way of obtaining the final decision depends on the multi-class extension used with SVM-DAS. We use the one-against-one extension throughout the paper for which $U = M(M-1)/2$.

The nonlinearity is given by the choice of the kernel function $\mathcal{K}$, thus in the case of the linear kernel the method is still linear. In this sense, SVM-DAS is more general than G-DAS, while it preserves all its important properties (e.g. the ability to give correct results for two misleading cues, see Figure 2). Also, for SVM-DAS each of the integrated outputs depends on all the outputs from single-cue classifiers, and the coefficients are learned optimally. Note that the outputs $\mathcal{V}_{t,v}(\mathcal{T}_t(\boldsymbol{s}))$ can be derived from a combination of different large margin classifiers, and not only from SVM.

SVM-DAS can be seen as a variation of ensemble learning methods that employ multiple models to improve the recognition performance. The key reason why ensemble algorithms obtain better results is because the individual classifiers make errors on different data points. Typically, different training data is used for each classifier [40]. In our experiments, we use data representing different types of information e.g. obtained using different sensors.

5 Place Classification for Semantic Space Labeling

One of the applications of a place classification system is semantic labeling of space. This section provides a brief overview of the problem and describes how we employed our multi-modal place classification method to build a semantic labeling system. We evaluated the system in a live experiment described in Section 7.

5.1 Semantic Labeling of Space

The problem of semantic labeling can be described as assigning meaningful semantic descriptions (e.g. “corridor” or “kitchen”) to areas in the environment. Typically, semantic labeling is used as a way of augmenting the internal space representation

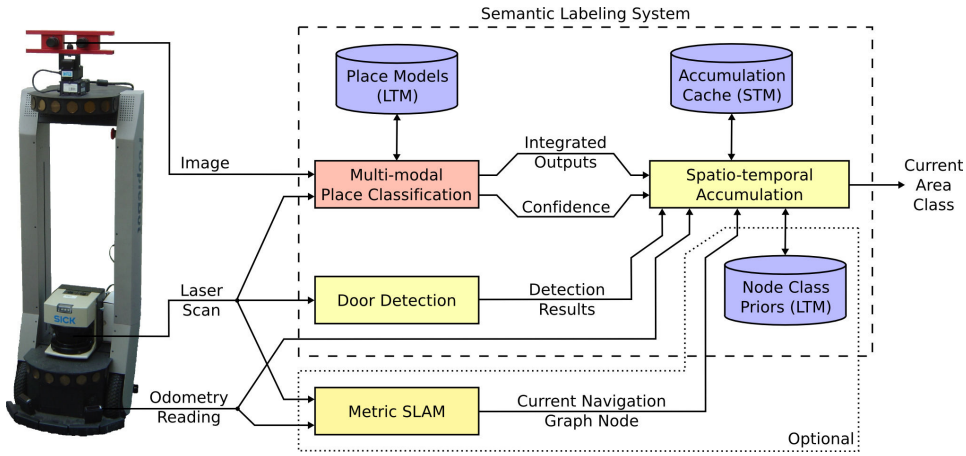


Figure 3: Architecture of the semantic space labeling system based on place classification (LTM: Long-Term Memory, STM: Short-Term Memory).

with additional information. This can be used by the agent to reason about space and to enhance communication with a human user. In case of most typical environments, it is sufficient to distinguish between semantic categories which are usually associated with rooms [63], such as "office", "meeting room" or "corridor". It is labeling at this level that we will discuss in this paper.

As will be shown through experiments in Sections 6 and 7, the place classification system described in this paper can yield a place class with high accuracy given a single sample of multi-modal data (e.g. one image and a laser scan). However, when used for semantic labeling, the algorithm is requested to provide a label for the whole area under exploration. At the same time, the system must be resilient and able to deal with such problems as temporary lack of informative cues, continuous stream of similar information or long-term occlusions. Given that the system is operating on a mobile robot, crude information about its movement is available from wheel encoders. This information can be used to ensure robustness to the typical variations that occur in the environment but also to the problems mentioned above. Finally, the system should be able to measure its own confidence and restrain from making a decision until some confidence level is reached. All these assumptions and requirements have been taken into consideration while designing the system described in the following section.

5.2 Architecture of the System

The architecture of our system is presented in Figure 3. The system relies on three sensor modalities typically found on a mobile robot platform: a monocular camera, a single 2D laser scanner, and wheel encoders. The images from the camera,

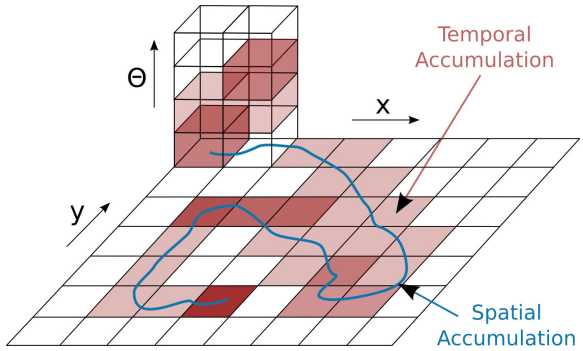


Figure 4: Illustration of the spatio-temporal accumulation process. As the robot explores the environment, the beliefs collected on the way accumulate over time within the bin corresponding to the current pose (x, y, θ) and over space in different bins.

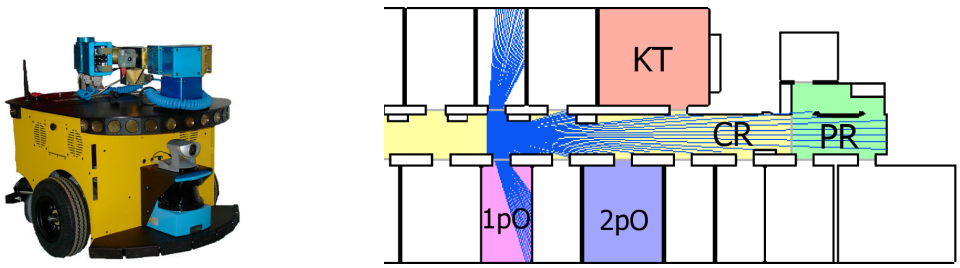
together with the laser scans are used as an input for the multi-modal place classification component. For each pair of data samples, place classification provides its beliefs about the semantic category to which the samples belong. These beliefs are encoded in the integrated outputs as discussed in Section 4. Moreover, the confidence of the decision is also measured and provided by the classification component.

A labeling system should provide a robust and stable output over the whole area. Since the sensors employed are not omni-directional, it is necessary to accumulate and fuse information over time. Moreover, the data that the robot gathers are not evenly spread over different viewpoints. As a possible solution, we propose to use a confidence-based spatio-temporal accumulation method. The principle behind the method is illustrated in Figure 4. As the robot explores the environment, it moves with a varying speed. The robot has information about its own movement (odometry) provided by the wheel encoders. As errors accumulate over time, this information can only be used to estimate relative movement rather than absolute position. This is sufficient for our application. The spatio-temporal accumulation process creates a sparse histogram along the robot pose trajectory given by the odometry and described by the metric position (x, y) and heading (θ) as shown in Figure 4. The size of the histogram bins are adjusted so that each bin roughly corresponds to a single viewpoint. Then, as the robot moves, the beliefs about the current semantic category accumulate within the bins as in case of G-DAS (with equal weights). This is what we call the temporal accumulation. It prevents a single viewpoint from becoming dominant due to long-term observation. Since each viewpoint observed by the robot will correspond to a different bin, performing accumulation across the bins (this time spatially) allows to generate the final outputs to which each viewpoint contributes equally. In order to exclude most of the

misclassifications before they get accumulated, we filter the decisions based on the confidence value provided by the place classification component. Moreover, as the odometry information is unreliable in the long term, the contents of bins visited a certain amount of viewpoints ago are invalidated. Note that semantic labeling is an application of the method presented in this paper and not the main focus of the paper. The accumulation scheme we present here builds on the ideas of discriminative accumulation and confidence estimation to further illustrate their usefulness. If the emphasis was on labeling, more advanced methods based on Hidden Markov Models [45], probabilistic relaxation [48] or Conditional Random Fields [16] should be taken into consideration. The advantages of our method are seamless integration with other components of the system and simplicity (the method does not require training or making assumptions on the transition probabilities between locations or areas).

The accumulation process ensures robustness and stability of the generated label for a single area. However, another mechanism is required to provide the system with information about area boundaries. This is required for the accumulation process not to fuse the beliefs across different areas. Here, we propose two solutions to that problem. As described in the previous sections, we can assume that each room of the environment should be assigned one semantic label. In case of indoor environments, rooms are usually separated by a door or other narrow openings. Thus, as one solution, we propose to use a simple laser-based door detector which generates hypotheses about doors based on the width of the opening which the robot passes. Such a simple algorithm will surely generate a lot of false positives. However, this does not cause problems in the presented architecture as false positives only lead to oversegmentation. This is a problem mainly for other components relying on precise segmentation rather than for the labeling process itself. In fact, the labeling system could be used to identify false doors and improve the segmentation by looking for directly connected areas classified as being of the same category.

As a second solution, we propose to use another localization and mapping system in order to generate the space representation which will then be augmented with semantic labels. Here we take the multi-layered approach proposed in [64]. The method presented by [64] builds a global metric map as the first layer and a navigation graph as the second. As the robot navigates through the environment, a marker or navigation node is dropped whenever the robot has traveled a certain distance from the closest existing node. Nodes are connected following the order in which they were generated. If information about the current node is provided to the spatio-temporal accumulation process, labels can be generated for each of the nodes separately. Moreover, as it is possible to detect whether the robot revisited an existing node, the accumulated information can be saved and used as a prior the next time the node is visited. For the live experiment described in this paper, we used the detected doors to bound the areas and navigation graph nodes to keep the priors. We then propagated the current area label to all the nodes in the area.



(a) (b)

Figure 5: 5(a) The mobile robot platform used in the experiments; 5(b) Map of the environment used during data acquisition and an example laser scan simulated in the corridor. The rooms used during the experiments are annotated.

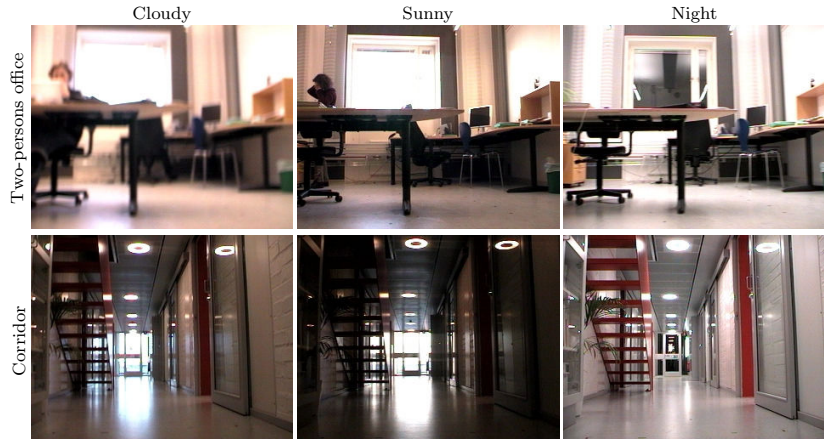
6 Experiments with Place Classification

We conducted several series of experiments to evaluate the performance of our place classification system. We tested its robustness to different types of variations, such as those introduced by changing illumination or human activity over a long period of time. The evaluation was performed on data acquired using a mobile robot platform over a time span of 6 months, taken from the IDOL2 database (Image Database for rObot Localization 2 [30]). Details about the database and experimental setup are given in Section 6.1. The experiments were performed for single-cue models and models based on different combinations of cues and modalities. We present the results in Section 6.2 and Section 6.3 respectively. Additionally, we analyze performance and properties of different cue integration schemes in Section 6.4.

6.1 Experimental Setup

The IDOL2 database was introduced in [30]. It comprises 24 image sequences accompanied by laser scans and odometry data acquired using two mobile robot platforms (PeopleBot and PowerBot). The images were captured with a Canon VC-C4 perspective camera using the resolution of 320x240 pixels. In this paper, we will use only the 12 data sequences acquired with the PowerBot, shown in Figure 5a.

The acquisition was performed in a five room subsection of a larger office environment, selected in such way that each of the five rooms represented a different functional area: a one-person office (1pO), a two-persons office (2pO), a kitchen (KT), a corridor (CR), and a printer area (PR). The map of the environment and an example laser scan are shown in Figure 5b. Example pictures showing interiors of the rooms are presented in Figure 6. The appearance of the rooms was captured under three different illumination conditions: in cloudy weather, in sunny weather, and at night. The robot was manually driven through each of the five rooms while



(a) Variations introduced by illumination

(b) Variations observed over time

(c) Remaining rooms (at night)

Figure 6: Examples of pictures taken from the IDOL2 database showing the interiors of the rooms, variations observed over time and caused by activity in the environment as well as introduced by changing illumination.

continuously acquiring images and laser range scans at a rate of 5fps. Each data sample was then labelled as belonging to one of the rooms according to the position of the robot during acquisition. Extension 1 contains a video illustrating the acquisition process of a typical data sequence in the database. The acquisition was conducted in two phases. Two sequences were acquired for each type of illumination conditions over the time span of more than two weeks, and another two sequences for each setting were recorded 6 months later (12 sequences in total). Thus, the sequences captured variability introduced not only by illumination but also natural activities in the environment (presence/absence of people, furniture/objects relocated etc.). Example images illustrating the captured variability are shown in Figure 6.

We conducted four sets of experiments, first for each cue separately and then for cues combined. In order to simplify the experiments with multiple cues, we matched images with closest laser scans on the basis of the acquisition timestamp. In case of each single experiment, both training and testing were performed on one data sequence. The first set consisted of 12 experiments, performed on different combinations of training and test data acquired closely in time and under similar illumination conditions. In this case, the variability comes from human activity and viewpoint differences. For the second set of experiments, we used 24 pairs of sequences captured still at relatively close times, but under different illumination conditions. In this way, we increased the complexity of the problem [43, 42]. In the third set of experiments, we tested the robustness of the system to long-term variations in the environment. Therefore, we conducted 12 experiments, where we tested on data acquired 6 months later, or earlier, than the training data, again under similar illumination conditions. Finally, we combined both types of variations and performed experiments on 24 pairs of training and test sets, obtained 6 months from each other and under different illumination settings. Note that in the last two sets of experiments described, the task becomes more and more challenging as the difference between training and test set increases. By doing this, we aim at testing the gain in robustness expected from cue integration in very difficult, but still realistic, scenarios.

For all experiments, model parameters were determined via cross validation. Since the datasets in the IDOL2 database are unbalanced (on average 443 samples for CR, 114 for 1pO, 129 for 2pO, 133 for KT and 135 for PR), as a measure of performance for the reported results and parameter selection, we used the average of classification rates obtained separately for each actual class (average per-class recall). For each single experiment, the percentage of properly classified samples was first calculated separately for each room and then averaged with equal weights independently of the number of samples acquired in the room. This allowed to eliminate the influence that large classes could have on the performance score. Statistical significance of the presented results was verified using the Wilcoxon signed-ranks test (when performance of two methods was compared) or Friedman and post-hoc Nemenyi test (when multiple methods were compared) at a confidence level of $\alpha = 0.05$ as suggested in [15]. The results of the post-hoc tests were

visualized using critical difference diagrams. The diagrams show average ranks of the compared methods and the groups of methods that are not significantly different are connected (the difference is smaller than the critical difference presented above the main axis of the diagram). The reader is referred to [15] for details on the applied tests and the critical difference diagrams presented below.

6.2 Experiments with Separate Cues

We first evaluated the performance of all four types of single-cue models: the two SVM models based on visual features (CRFH, SIFT), the AdaBoost and the SVM models trained on the laser range cues (referred to as L-AB and L-SVM). For SVM, we tried the three multi-class extensions described in Section 4.1. The results of all four sets of experiments for these models are presented in Figure 7-10 (the first four bar groups). Moreover, the results of statistical significance tests comparing the models based on the combined results of all four experiments are illustrated in Figure 11. We first note that, as expected, CRFH and SIFT suffer from changes in illumination (-15.3% and -11.0% respectively), while the geometrical features do not (-1.9% for L-AB and -0.6% for L-SVM). Long-term variations pose a challenge for both modalities (-7.0% – -10.2% for vision and -3.7% – -7.9% for laser). We also see differences in performance between the two visual cues: CRFH suffers more from changes in illumination, while SIFT is less robust to variations induced by human activities. It is also interesting to note that under stable conditions, the vision-based methods outperform the systems based on laser range cues (95.1% for CRFH and 92.5% for L-SVM; the difference is statistically significant). This illustrates the potential of visual cues, but also stresses the need for more robust solutions.

These experiments are also a comparison between two recognition algorithms using laser-range features, namely the boosting-based implementation (L-AB) presented in [33] and the current SVM-based implementation (L-SVM). Figures 7-10 and 11 show the results. We can see that the difference in performance is statistically significant in favor of the SVM-based method for all three multi-class extensions (from 6.1% for Exp. 1 to 10.3% for Exp. 4 in average). The classification results for the L-AB are worse than the results of the original paper by [33]. There are two main reasons for that. First, the number of classes is increased to five, while in [33] was of a maximum of four. Second, in these experiments, we used a restricted field of view of 180 degrees, whilst in [33] the field of view was of 360 degrees. This decreases the classification rate, as it has been shown in previous work [34].

As already mentioned, all the experiments with SVMs were repeated for three different multi-class extensions: standard OaO and OaA as well as modified OaA algorithm. The obtained results are in agreement with [42] - in case of single cue and G-DAS experiments, the modified version gives the best performance with a statistically significant difference independently of the modality on which the classifier was trained.

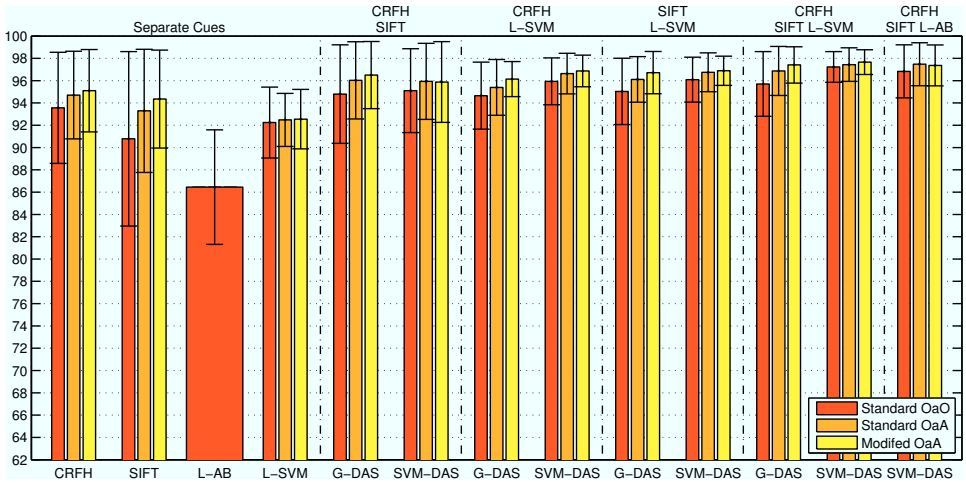


Figure 7: Classification results for Exp. 1: stable illumination conditions, close in time.

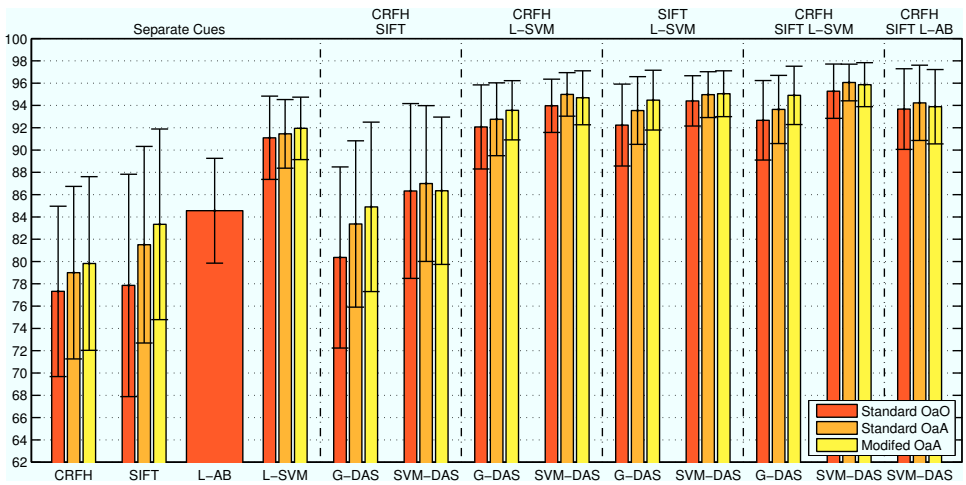


Figure 8: Classification results for Exp. 2: varying illumination conditions, close in time.

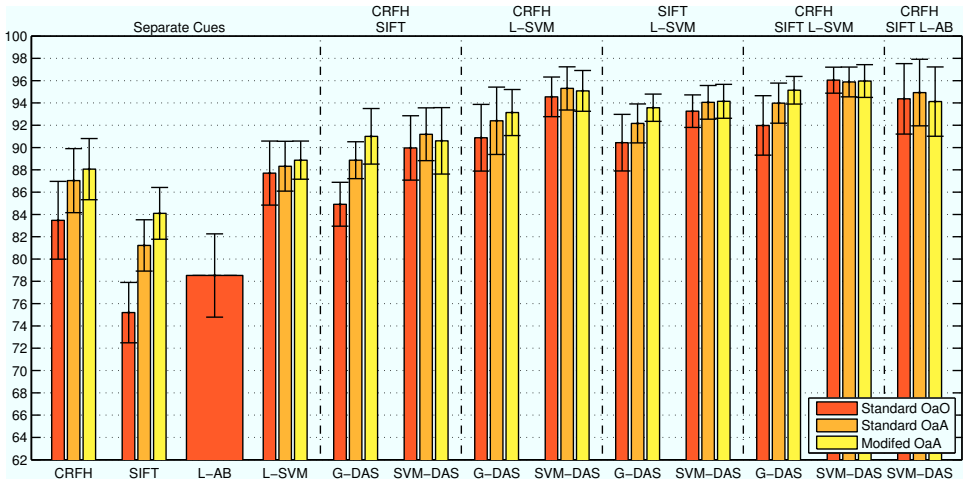


Figure 9: Classification results for Exp. 3: stable illumination conditions, distant in time.

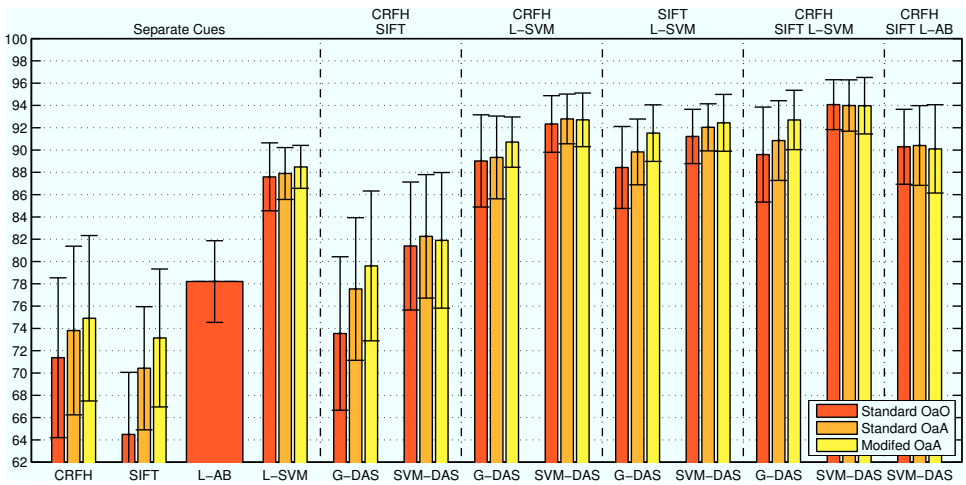


Figure 10: Classification results for Exp. 4: varying illumination conditions, distant in time.

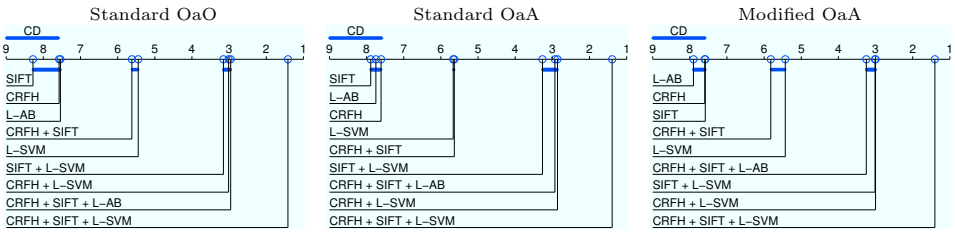
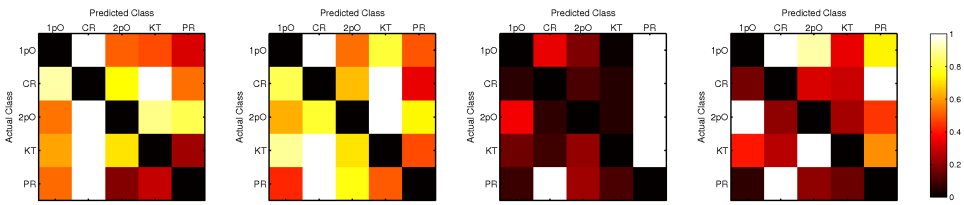


Figure 11: Critical difference diagrams comparing four single-cue models and solutions based on multiple cues integrated using SVM-DAS with the Nemenyi test for a confidence level of $\alpha = 0.05$. The comparison is based on the combined results of Exp. 1-4 and presented separately for each multi-class extension. The average ranks of the methods are plotted on the axis and the groups of methods that are not significantly different are connected.



(a) CRFH (b) SIFT (c) L-AB (d) L-SVM

Figure 12: Distribution of errors made by the four models for each actual class (bright colors indicate errors). The diagonal elements were removed.

Figure 12 shows the distribution of errors for each actual class (room) made by the four models. It is apparent that each of the cues makes errors according to a different pattern. At the same time, similarities occur between the same modalities. We see that visual models are biased towards the corridor, while the geometrical models tend to misclassify places as the printer area. A possible explanation is that the vision-based models were trained on images acquired with perspective camera with constrained viewing angle. As a result, similar visual stimuli coming from the corridor are present in the images captured by the robot leaving each of the rooms. The same area close to a doorway, from the geometrical point of view, is similar to the narrow passage in the printer area. This analysis is a strong motivation to integrate these various cues with a stack of classifiers, as theory indicates that this is the ideal condition for exploiting the different informative content [40].

6.3 Experiments with Cue Integration

For the final experiments, we selected four different cue accumulation methods: G-DAS and SVM-DAS with three kernel types (linear, RBF, and histogram in-

		Predicted Class				
		1pO	CR	2pO	KT	PR
Actual Class	1pO	**11.20 (93.71)**	0.36 (3.06)	0.16 (1.33)	0.11 (0.96)	0.11 (0.94)
	CR	0.25 (0.53)	**45.36 (97.73)**	0.19 (0.42)	0.33 (0.70)	0.29 (0.62)
	2pO	0.17 (1.22)	0.11 (0.8)	**13.26 (96.92)**	0.06 (0.46)	0.08 (0.60)
	KT	0.17 (1.18)	0.35 (2.45)	0.08 (0.57)	**13.42 (95.12)**	0.09 (0.67)
	PR	0.09 (0.65)	0.77 (5.59)	0.03 (0.19)	0.05 (0.33)	**12.90 (93.24)**

Table 1: Confusion matrix for the multi-cue system based on CRFH, SIFT and L-SVM integrated using SVM-DAS. Normalized average values in percentage over all experiments are reported. The values in brackets were normalized separately for each actual class (row). The presented results are only for the standard OaO multi-class extension since the results for the remaining extensions were comparable.

tersection (HI) kernel [6]). The parameters of the algorithms (weights in case of G-DAS and SVM model in case of SVM-DAS) were always adjusted on the basis of outputs generated during all experiments with single-cue models trained on one particular data sequence. Then, during testing, the previously obtained integration scheme was applied to all experiments with models trained on a different sequence, acquired under similar illumination and closely in time. This way, the generalization abilities of each of the methods were tested in a realistic scenario. In all experiments, we found that SVM-DAS with an RBF kernel outperforms the other methods and the difference in performance with respect to G-DAS was statistically significant for all combinations of cues and multi-class extensions (Wilcoxon test). For space reasons, we report results of each of the experiments only using SVM-DAS based on the RBF kernel and G-DAS for comparison (Figure 7-10, last 9 bar groups). A detailed comparison of all variants of SVM-DAS for the most complex problem (Exp. 4) is given in Figure 13. Results of statistical significance tests comparing the multi-cue solutions with single-cue models based on the combined results of all experiments are illustrated in Figure 11.

We tested the methods with several combinations of different cues and modalities. First, we combined the two visual cues. We see that the generalization of a purely visual recognition system can be significantly improved by integrating different types of cues, in this case local and global. This can be observed especially for Exp. 4, where the algorithms had to tackle the largest variability. Despite that, according to the error distributions in Figure 12, we should expect largest gain when different modalities are combined. As we can see from Figure 7-10 this is the case indeed. By combining one visual cue and one laser range cue (e.g. CRFH + L-SVM), we exploit the descriptive power of vision in case of stable illumination conditions and the invariance of geometrical features to the visual noise. Moreover, if the computational cost is not an issue, the performance can be further improved by using both visual cues instead of just one. As can be seen from Figure 11, by integrating single-cue models or adding another cue to a multi-cue system, we always get an improvement statistically significant.

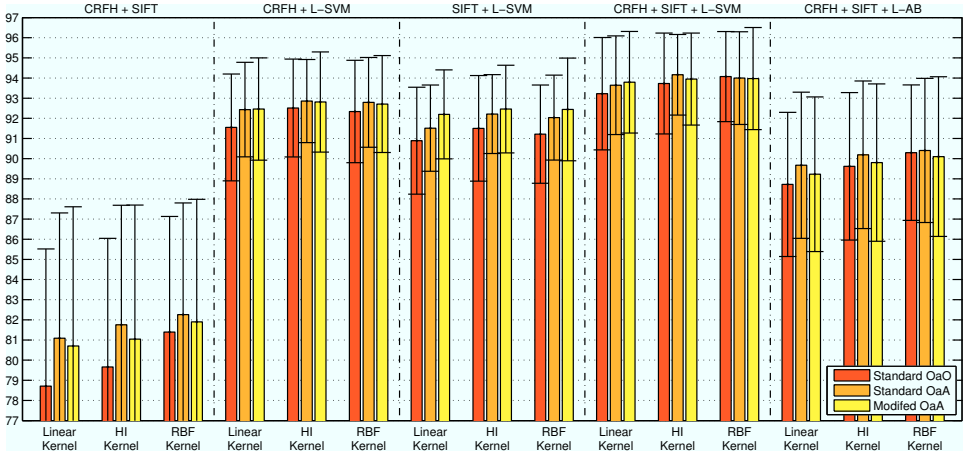


Figure 13: Comparison of performance of SVM-DAS based on different kernel functions for the most complex problem (Exp. 4).

We performed a more detailed analysis of the best results. Table 1 contains the confusion matrix for the multi-cue system based on CRFH, SIFT and L-SVM integrated using SVM-DAS with an RBF kernel. We see that even if the the corridor class contained on average four times more samples than each of the room classes and was visually and geometrically distinctive, the results are balanced and the recognition rates for each actual class are similar. In general, during our experiments, more balanced solutions were preferred due to the performance metric used (average of the diagonal values in brackets in Table 1).

As it was mentioned in Section 4.3, SVM-DAS can be applied for problems where outputs of different classifiers need to be integrated. To test this in practice, we combined the SVM models trained on visual cues with AdaBoost model based on geometrical features (L-AB). As usual, for SVM we used several multi-class extensions that in most cases produced outputs having different interpretation than those generated by the multi-class algorithm used for AdaBoost. In those cases G-DAS could not be applied. We present the results in Figure 7-10 (last bar group) and Figure 11. The method obtained large and statistically significant improvements compared to each of the individual cues. For instance for Exp. 4, the recognition rate increased by 12.2% in average. This proves the versatility of our approach.

6.4 Analysis of Cue Integration Schemes

Results presented so far clearly show that SVM-DAS performs significantly better than G-DAS and, by using more sophisticated kernel types for SVM-DAS, it is possible to perform non-linear cue accumulation. Moreover, the experiments (see Figure 13) show that we can expect better results with the RBF kernel (especially

Cues (**Primary cue**)	Cue integration method	
	G-DAS	SVM-DAS RBF Kernel
CRFH + SIFT	25.971±18.503	29.453±22.139
CRFH + **L-SVM**	21.230±20.199	32.736±20.256
SIFT + **L-SVM**	28.820±20.982	33.344±22.425
SIFT + CRFH + **L-SVM**	31.858±20.474	40.833±21.916

Table 2: Average percentages (with standard deviations) of test samples for which all cues had to be used in order to obtain the maximal recognition rate.

for the OaO multi-class extension), although there is no drastic improvement. We therefore suggest to choose the kernel according to constraints on the computational cost of the solution. Since there are fast implementations of linear SVMs, it might be beneficial to use a linear kernel in cases when the integration scheme must be trained on a very large number of samples. In applications where only the number of training parameters is an issue, the non-parametric histogram intersection kernel can be used instead of RBF.

We now further discuss differences between high-level (e.g. SVM-DAS) and low-level (feature level) cue integration. There are several advantages in integrating multiple cues with a high level strategy. First, different learning algorithms can be used for each single cue. In our experiments, this allowed to combine SVM-based models employing different kernel functions (e.g. the χ^2 kernel for CRFH and the match kernel for SIFT) or even different classifiers (AdaBoost and SVM). Moreover, parameters can be tuned separately for each of the cues. Second, both the training and recognition tasks can be divided into smaller subproblems that can be easily parallelized. Finally, it is possible to decide on the number of cues that should be extracted and used for each particular classification task. This is an important feature, since, in most cases, decisions based on a subset of cues are correct while extraction and classification of additional features introduces additional cost. For example, a solution based on global visual features, laser range cues and SVM-DAS runs in real-time at a rate of approximately 5fps, which would not be possible if an additional visual cue like SIFT was used. The computational cost can be significantly reduced by taking the approach presented in [42]. By combining confidence estimation methods with cue integration, we can use additional sources of information only when necessary - when the decision based on one cue only is not confident enough. This scheme is referred to as Confidence-based Cue Integration. Table 2 presents the results of applying the scheme to the experiments presented in this section. We see that, in general, we can base our decision on the fastest model (marked with bold font in Table 2), such as the efficient and low-dimensional model based on simple laser-range features, and we can retain the maximal performance by using additional cues only in approximately 30% of cases. This greatly reduces the computational time required on average e.g. approximately three times for

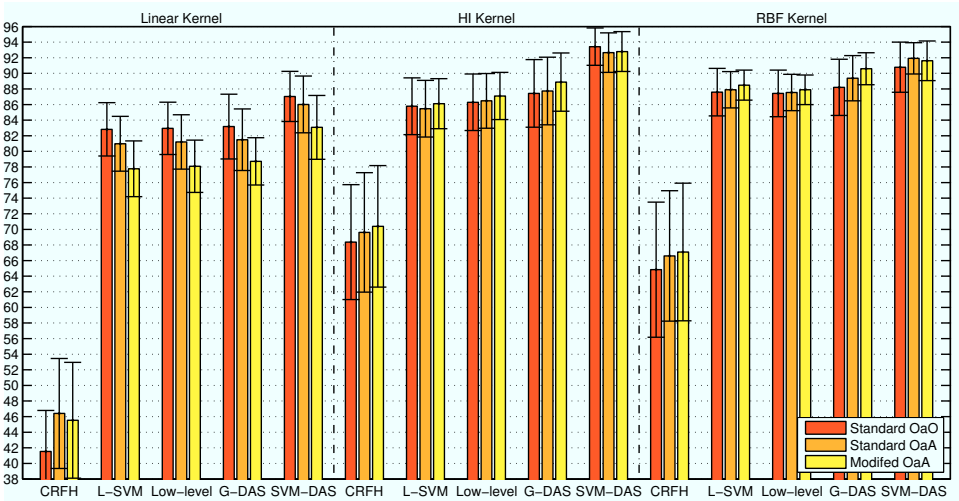


Figure 14: Comparison of performance of two single-cue models and solutions based on the cues integrated on both low and high level for the most complex problem (Exp. 4).

CRFH, L-SVM and SVM-DAS. Additional cues will be used more often when the variability is large, and rarely for less difficult cases. This is not possible in case of low-level integration where all the cues must be extracted and classified in order to obtain a decision.

Another important factor is performance. During our experiments, we compared the performance of G-DAS and SVM-DAS (with an RBF kernel) with models build on cues combined on the feature level. We performed three different sets of comparisons. First, we built single-cue models and models based on features combined on the low level using SVM and the non-parametric linear kernel, using the same values of the SVM training parameters for all models. Then, we integrated the outputs of the single-cue models using G-DAS and SVM-DAS. In case when G-DAS was used, the solution remained linear. In the second comparison, for building the models we used the non-linear, non-parametric histogram intersection kernel. Finally, we used an RBF kernel and performed parameter selection for each of the models. All comparisons were based on CRFH and laser-range cues, since the dedicated kernel function required by SIFT could not be used with any of the other features for low-level integration.

The results for the most complex problem (Exp. 4) are given in Figure 14 and statistical significance tests comparing the solutions are illustrated in Figure 15. It can be observed that, in every case, the high-level integration significantly outperformed solutions based on features combined on the low level. In only one case there was no significant difference between G-DAS and low-level integration; how-

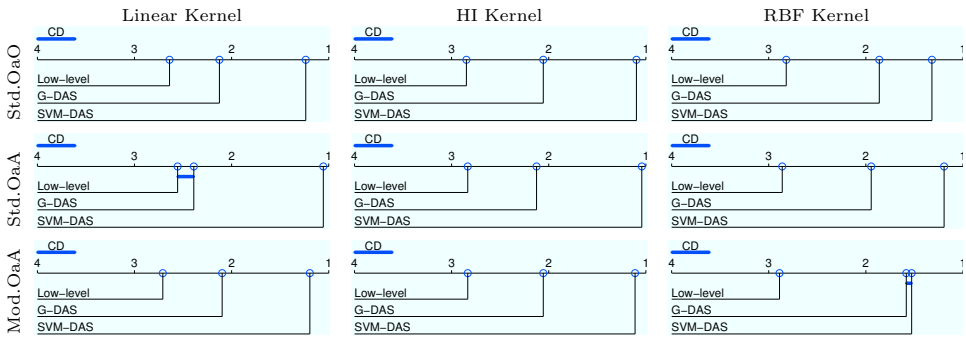


Figure 15: Critical difference diagrams comparing two single-cue models and solutions based on the cues integrated on both low and high level with the Nemenyi test for a confidence level of $\alpha = 0.05$. The comparison is based on the combined results of Exp. 1-4 and presented separately for three kernel functions and multi-class extensions used with SVM. The average ranks of the methods are plotted on the axis and the groups of methods that are not significantly different are connected.

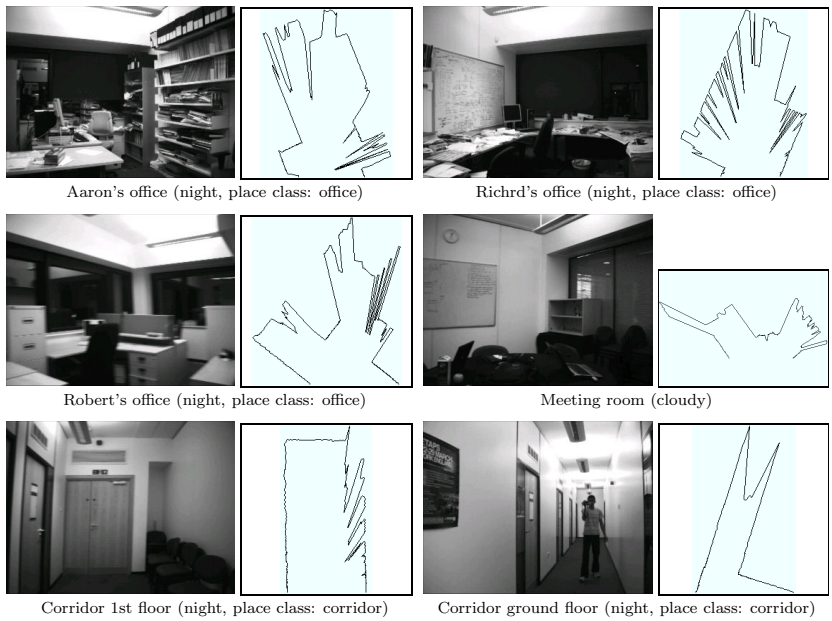
ever, SVM-DAS still performed better than the other solutions. This is in agreement with the results reported in [52, 36] and can be explained by greater robustness of the high-level methods to noisy cues or sensory channels and the ability of different classifiers to adapt to the characteristics of each single cue.

7 Experiments with Semantic Space Labeling

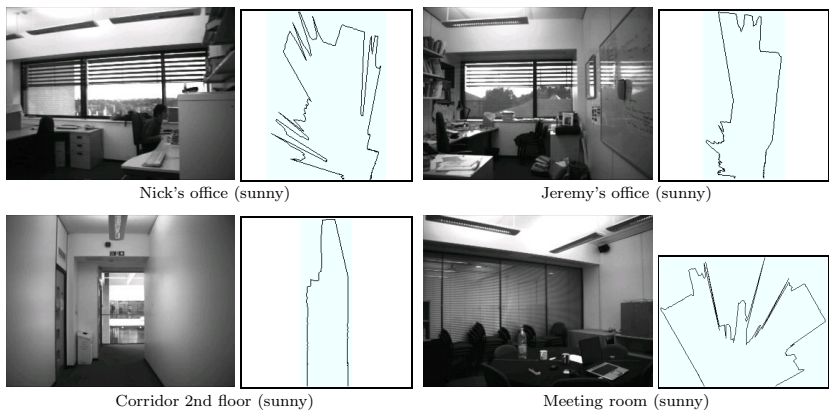
We performed an independent live experiment to test our multi-modal semantic space labeling system running in real-time on a mobile robot platform. The experiment was performed during working hours in a typical office environment. Both the environment and the robot platform were different than in case of the off-line evaluation described in Section 6. The whole experiment was videotaped and a video presenting the setup, experimental procedure and visualization of the results can be found in Extension 2.

7.1 Experimental Setup

The experiment was performed between the 7^{th} and 10^{th} of September 2008 in the building of the School of Computer Science at the University of Birmingham, Birmingham, United Kingdom. The interior of the building consists of several office environments located on three floors. For our experiments, we selected three semantic categories of rooms that could be found in the building: a corridor, an office and a meeting room. To build the model of an office, we acquired data in three different offices: Aaron's office (1^{st} floor), Robert's office (1^{st} floor) and



(a) Samples from the data sequences used to train the models of place classes.



(b) Samples acquired during the test run.

Figure 16: Examples of images and laser scans (synchronized) taken from the data sequences used for training the models of place classes (a) and acquired during the test run (b) in each of the rooms considered during the experiment. The figure illustrates the within-category variations for corridors and offices as well as other types of variability observed for each place class (e.g. different illumination conditions, activity in the environment).

Richard's office (ground floor). To create a representation of the corridor class, we recorded data in 2 corridors, one on the ground floor and one on the 1^{st} floor. The acquisition was performed at night. Finally, to train the model of a meeting room, we used an instance on the 2^{nd} floor. All training data expect the one from the meeting room was acquired in another part of the building than the one used testing. The data for this class were recorded during the day. A video illustrating the whole data acquisition process is available as Extension 3. The interiors of the rooms are presented in Figure 16a, as seen by vision and laser. The robot was manually driven around each room and data samples were recorded at the rate of 5 fps. All the collected training data are available as Extension 6. In case of the meeting room, corridor on the $1st$ floor as well as Aaron's and Richard's offices, the acquisition was repeated twice.

For the real-time experiment, we built the system as described in Section 5. Following the findings of the off-line experiments, we used SVM-DAS with the RBF kernel to integrate the classifier outputs for vision and laser range data. For efficiency reasons, we used only global features (CRFH) for the vision channel. We used the one-against-all multi-class SVM extension for the place models. Other parameters were set as described in Section 6.

We trained the place models separately for each modality on a dataset created from one data sequence recorded in each of the rooms. One of the advantages of SVM-DAS is the ability to infer the integration function from the training data, after training the models. We used the additional data sequences acquired in some of the rooms and trained SVM-DAS on the outputs of the uni-modal models tested on these data.

The PeopleBot robot platform shown in Figure 3 was used for data acquisition and the final experiment. The robot was equipped with a SICK laser range finder and Videre STH-MDCS2 stereo head (only one of the cameras was used). The images were acquired at the resolution of 320x240 pixels. The whole system was implemented in the CAST (The CoSy Architecture Schema Toolkit, [23]) framework and run on a standard 2.5GHz dual-core laptop. The processing for both modalities was executed in parallel using both of the CPU cores.

7.2 Experimental Procedure and Results

Three days after the training data were collected, we performed a live experiment in the lab on the 2^{nd} floor in the same building. The experiment was conducted during the day with sunny weather. The part of the environment that was explored by the robot consisted of 2 offices (Nick's office and Jeremy's office), a corridor and a meeting room. The interiors of the rooms and the influence of illumination can be seen in the images in Figure 16b.

The SLAM system of the robot constructs a metric map and navigation graph. In this experiment, the task is to semantically label the navigation graph nodes and areas as the map is being built. The only knowledge given to the robot before the experiment consisted of the models of the three place classes: "office", "corridor"

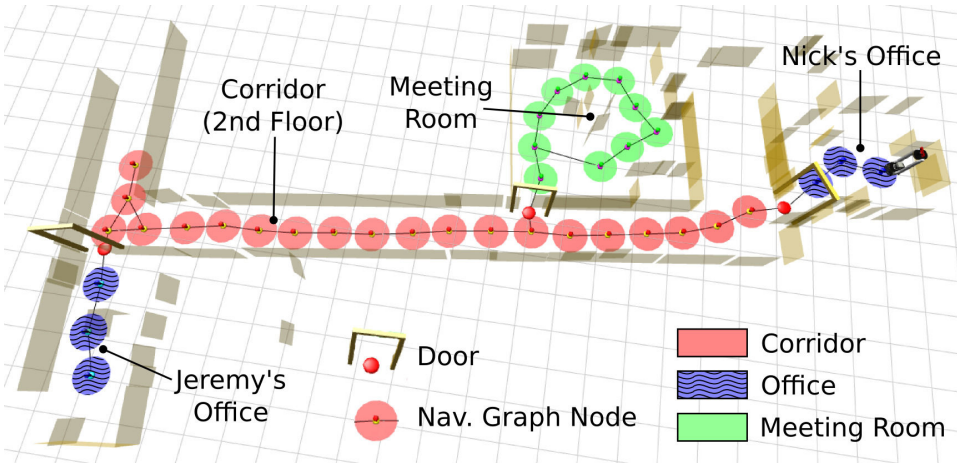


Figure 17: Final map obtained after the test run. The navigation graph is overlaid on the metric map and the color of the circles around the graph nodes indicate the place class assigned to each area bounded by detected doors. The system correctly labeled all the areas in the environment.

and "meeting room". As stated in Section 5, every time the robot created or revisited a node, the accumulated beliefs about the semantic category of the area were used to label the node and saved as a future prior. The label was also propagated to the whole area. We used detected doors to assign nodes to areas.

The whole experiment was videotaped and a video presenting the experimental setup, the test run and visualization of the obtained results can be found in in Extension 2. The robot started in Nick's office, and was manually driven through the corridor to Jeremy's office. Then, it was taken to the meeting room where the autonomous exploration mode was turned on. The robot used a frontier-based algorithm based on [62]. Laser data was limited to 2m distance in the exploration to make sure that the robot not just perceived how the environment looked but also covered it to build the navigation graph. After the meeting room was explored, the robot was manually driven back to Nick's office where the experiment finished. A video presenting visualization of the full test run is available in Extension 4. The labeling process was running on-line and the place classification was performed approximately at the rate of 5 times per second. The final semantic map build during the run is shown in Figure 17. We can see that the system correctly labeled all the areas in the environment.

The sensory data acquired during the test run are available as Extension 6. Moreover, a video presenting the sequence of images and laser scans is presented in Extension 5. The fact that the data were stored allowed for additional performance analysis of the multi-modal place classification system, similar to the one presented

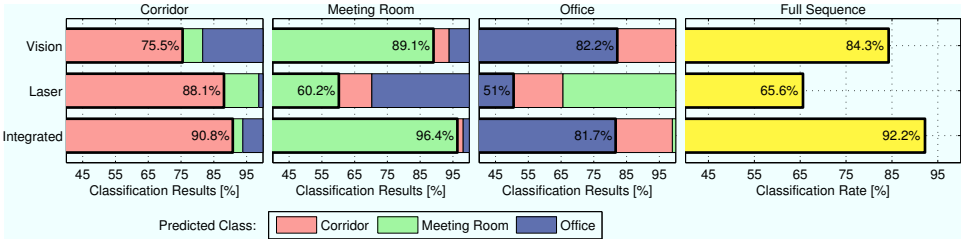


Figure 18: Place classification results obtained on the dataset recorded during the test run. The first three bar charts show the results separately for each place class: "corridor", "meeting room" and "office". The charts show the percentage of the samples that were properly classified (most left bars marked with thick lines), but also how the misclassifications were distributed. The chart on the right presents the percentage of properly classified samples during the whole run. The two top rows give results for single modalities, while the bottom row shows results for the multi-modal system.

in Section 6. The results are displayed in Figure 18. When we look at the overall classification rate for all the data samples in the test sequence, we see that vision significantly outperformed laser in this experiment (66% v.s. 84%). Still, the performance of the system was boosted by additional 8% compared to vision alone when the two modalities were integrated. The gain is even more apparent if we look at the detailed results for each of the classes (the first three charts in Figure 18). We see that the modalities achieved different performance, but also different error patterns, for each class. Clearly, the system based on laser range data is a very good corridor detector. On the other hand, vision was able to distinguish between the offices and the meeting room almost perfectly. Finally, the integrated system always achieved the performance of the more reliable modality and for two out of three classes outperformed the uni-modal systems. As can be seen in the video in Extension 2 and 4, this provided stable performance for each of the classes and a robust base for the semantic labeling system.

8 Conclusions

This paper addressed the problem of place classification and showed how it can be applied to semantic knowledge extraction in robotic systems. This is an important and challenging task, where multiple sensor modalities are necessary in order to achieve generality and robustness, and enable systems to work in realistic settings. To this end, we presented a new cue integration method able to combine multiple cues derived by a single modality, as well as cues obtained by multiple sensors. The method was thoroughly tested in off-line experiments on realistic data collected under varying conditions and as part of a real-time system running on a robotic

platform. The results obtained using multiple visual cues alone, and combined with laser range features, clearly show the value of our approach. Finally, we showed that the system can successfully be applied for the space labeling problem where it can be used to augment the internal space representation with semantic place information. All the data used in the paper are available as extensions to the paper and from the IDOL2 database [29].

In the future, we plan to extend this method and attack the scalability issue, with particular attention to indoor office environments. These are usually characterized by a large number of rooms with very similar characteristics; we expect that in such a domain our approach will be particularly effective. Another important aspect of place classification is the intrinsic dynamics in the sensory information: as rooms are used daily, furniture is moved around, objects are taken in and out of drawers and people appears. All of this affects the sensor inputs of places in time. We plan to combine our approach with incremental extensions of the SVM algorithm [30, 37] and to extend these methods from fully supervised to semi-supervised learning, so to obtain a system able to learn continuously from multiple sensors.

Acknowledgment

Special thanks go to Sagar Behere for his great help with running the integrated system on the robotic platform, data acquisition and videotaping. This work was sponsored by the EU integrated projects FP6-004250-IP CoSy (A. Pronobis, O. Martínez Mozos, P. Jensfelt), ICT-215181-CogX (A. Pronobis and P. Jensfelt) and IST-027787 DIRAC (B. Caputo), and the Swedish Research Council contract 2005-3600-Complex (A. Pronobis). The support is gratefully acknowledged.

1 Index to Multimedia Extensions

The multimedia extensions to this article are at: http://www.ijrr.org.

Extension	Type	Description
1	Video	The acquisition procedure of a typical data sequence in the IDOL2 database.
2	Video	The setup, procedure and visualization of the experiment with semantic space labeling based on multi-modal place classification.
3	Video	The process of acquiring data for training the models of places for the experiment with semantic space labeling.
4	Video	Visualization of the complete test run and results obtained during the experiment with semantic space labeling.
5	Video	The complete sequence of images and laser scans acquired during the test run of the experiment with semantic space labeling.
6	Data	The dataset (sequences of images and laser scans) collected during the experiment with semantic labeling of space.

Table 3: Table of multimedia extensions.

References

[1] EU FP6 Integrated Project COGNIRON: The Cognitive Robot Companion. URL `http://www.cogniron.org`.

[2] EU FP6 IST Cognitive Systems Integrated Project CoSy: Cognitive Systems for Cognitive Assistants. URL `http://www.cognitivesystems.org/`.

[3] John Aloimonos and David Shulman. *Integration of visual modules: An extension of the Marr paradigm*. Academic Press, 1989.

[4] Philipp Althaus and Henrik I. Christensen. Behaviour coordination in structured environments. *Advanced Robotics*, 17(7):657–674, 2003.

[5] Henrik Andreasson, André Treptow, and Tom Duckett. Localization for mobile robots using panoramic vision, local features and particle filter. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, 2005.

[6] Annalisa Barla, Francesca Odone, and Alessandro Verri. Histogram intersection kernel for image classification. In *Proceedings of the International Conference on Image Processing (ICIP'03)*, pages 513–516, Barcelona, Spain, 2003.

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110 (3):346–359, June 2008.

[8] Paul Blaer and Peter Allen. Topological mobile robot localization using fast vision techniques. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation (ICRA'02)*, Washington, DC, USA, 2002.

[9] David M. Bradley, Rashmi Patel, Nicolas Vandapel, and Scott M. Thayer. Real-time image-based topological localization in large outdoor environments. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, Edmonton, Alberta, Canada, 2005.

[10] Pär Buschka and Alessandro Saffiotti. A virtual sensor for room detection. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'02)*, Lausanne, Switzerland, 2002.

[11] Barbara Caputo and Gyuri Dorkó. How to combine color and shape information for 3D object recognition: Kernels do the trick. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2002.

[12] Oliver Chapelle, Patrick Haffner, and Vladimir Vapnik. Support Vector Machines for Histogram-Based Image Classification. *IEEE Transactions on Neural Networks*, 10 (5), 1999.

[13] James J. Clark and Alan L. Yuille. *Data Fusion for Sensory Information Processing Systems*. Kluwer Academic Publisher, 1990.

[14] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. ISBN 0521780195.

[15] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research (JMLR)*, 7:1–30, 2006.

[16] Bertrand Douillard, Dieter Fox, and Fabio Ramos. A spatio-temporal probabilistic model for multi-sensor object recognition. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[17] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2nd edition, 2000.

[18] David Filliat. A visual bag of words method for interactive qualitative localization and mapping. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[19] Friedrich Fraundorfer, Christopher Engels, and David Nistér. Topological mapping, localization and navigation using image collections. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[20] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Journal of Computer and System Sciences*, volume 55, pages 119–139, Barcelona, Spain, 1997.

[21] Cipriano Galindo, Alessandro Saffiotti, Silvia Coradeschi, Pär Buschka, Juan Antonio Fernández-Madrigal, and Javier González. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, 2005. ISBN 0-7803-8912-3.

[22] José Gaspar, Niall Winters, and José Santos-Victor. Vision-based navigation and environmental representations with an omni-directional camera. *IEEE Transactions on Robotics and Automation*, 16(6):890–898, December 2000.

[23] Nick A. Hawes and Jeremy L. Wyatt. Engineering intelligent information-processing systems with CAST. *Advanced Engineering Informatics*, 24(1):27–39, January 2010.

[24] Sven Koenig and Reid G. Simmons. Xavier: A robot navigation architecture based on partially observable Markov decision process models. In David M. Kortenkamp, R Peter Bonasso, and Robin R. Murphy, editors, *Artificial Intelligence Based Mobile Robotics: Case Studies of Successful Robot Systems*, pages 91–122. {MIT} Press, 1998.

[25] Benjamin Kuipers. An intellectual history of the Spatial Semantic Hierarchy. In Margaret Jefferies and Albert Yeap, editors, *Robot and Cognitive Approaches to Spatial Mapping*. Springer Verlag, 2008.

[26] Benjamin Kuipers and Patrick Beeson. Bootstrap learning for place recognition. *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI'02)*, pages 174–180, 2002.

[27] Oskar Linde and Tony Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proceedings of the 2004 15th International Conference on Pattern Recognition (ICPR'04)*, 2004.

[28] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[29] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. The KTH-IDOL2 database. Technical report, Kungliga Tekniska Högskolan, CVAP/CAS, September 2006.

[30] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. Incremental learning for place recognition in dynamic environments. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 721–728, San Diego, CA, USA, October 2007.

[31] Jiri Matas, R. Marik, and Josef Kittler. On representation and matching of multi-coloured objects. In *Proceedings of the 5th IEEE International Conference on Computer Vision (ICCV'95)*, Boston, MA, USA, 1995.

[32] Emanuele Menegatti, Mauro Zoccarato, Enrico Pagello, and Hiroshi Ishiguro. Image-based Monte-Carlo localisation with omnidirectional images. *Robotics and Autonomous Systems*, 48(1), 2004.

[33] Oscar Martinez Mozos, Cyrill Stachniss, and Wolfram Burgard. Supervised learning of places from range data using AdaBoost. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, 2005.

[34] Oscar Martinez Mozos, Rudolph Triebel, Patric Jensfelt, Axel Rottmann, and Wolfram Burgard. Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems (RAS)*, 55(5):391–402, 2007.

[35] Ana Cris Murillo, José Jesús Guerrero, and Carlos Sagüés. SURF features for efficient robot localization with omnidirectional images. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[36] Maria-Elena Nilsback and Barbara Caputo. Cue integration through discriminative accumulation. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, DC, USA, 2004.

[37] Francesco Orabona, Claudio Castellini, Barbara Caputo, Jie Luo, and Giulio Sandini. Indoor place recognition using online independent Support Vector Machines. In *Proceedings of the British Machine Vision Conference (BMVC'07)*, Warwick, UK, 2007.

[38] John C. Platt. *Fast training Support Vector Machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, November 1999.

[39] Tomaso Poggio, Vincent Torre, and Christof Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[40] Rudi Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6:21–45, 2006.

[41] Ingmar Posner, Derik Schroeter, and Paul M. Newman. Describing composite urban workspaces. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[42] Andrzej Pronobis and Barbara Caputo. Confidence-based cue integration for visual place recognition. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 2394–2401, San Diego, CA, USA, October 2007.

[43] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I. Christensen. A discriminative approach to robust visual place recognition. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, October 2006.

[44] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision (IJCV)*, 66(3), 2006.

[45] Axel Rottmann, Oscar Martinez Mozos, Cyrill Stachniss, and Wolfram Burgard. Semantic place classification of indoor environments with mobile robots using boosting. In *Proceedings of the 20nd National Conference on Artificial Intelligence (AAAI'05)*, pages 1306–1311, Pittsburgh, Pennsylvania, USA, July 2005. ISBN 1-57735-236-x.

[46] Stephen Se, David G. Lowe, and James J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation (ICRA'01)*, Seoul, Korea, 2001.

[47] Christian Siagian and Laurent Itti. Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[48] Cyrill Stachniss, G Grisetti, Oscar Martinez Mozos, and Wolfram Burgard. Efficiently learning metric and topological maps with autonomous service robots. *Information Technology*, 49:232–237, 2007.

[49] Cyrill Stachniss, Oscar Martinez Mozos, and Wolfram Burgard. Speeding-up multi-robot exploration by considering semantic place information. In *Proceedings of the 2006 IEEE International Conference on Robotics and Automation (ICRA'06)*, Orlando, Florida, USA, 2006.

[50] Hashem Tamimi and Andreas Zell. Vision based localization of mobile robots using kernel approaches. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'04)*, Sendai, Japan, 2004.

[51] Adriana Tapus and Roland Siegwart. Incremental Robot Mapping with Fingerprints of Places. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Edmonton, Alberta, Canada, 2005.

[52] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Discriminative cue integration for medical image annotation. *Pattern Recognition Letters, Special Issue on IMageCLEF Med Benchmark Evaluation*, 29(15):1996–2002, 2008.

[53] Elin Anna Topp and Henrik I. Christensen. Topological modelling for human augmented mapping. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, 2006.

[54] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision (IJCV)*, 53(2), 2003.

[55] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, 2003.

[56] Antonio Torralba and Pawan Sinha. Recognizing Indoor Scenes. Technical Report 2001-015, MIT, July 2001.

[57] Jochen Triesch and Christian Eckes. Object recognition with multiple feature types. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN'98)*, Skövde, Sweden, 1998.

[58] Iwan Ulrich and Illah Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation (ICRA'00)*, San Francisco, CA, USA, 2000.

[59] Christoffer Valgren and Achim J. Lilienthal. Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 1856–1861, May 2008. ISBN 978-1-4244-1646-2.

[60] Christian Wallraven, Barbara Caputo, and Arnulf Graf. Recognition with local features: The kernel recipe. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, Beijing, China, 2003.

[61] Christian Weiss, Hashem Tamimi, Andreas Masselli, and Andreas Zell. A hybrid approach for vision-based outdoor robot localization using global and local image features. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, 2007.

[62] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings of the International Symposium on Computational Intelligence in Robotics and Automation*, Monterey, California, USA, 1997.

[63] Hendrik Zender, Patric Jensfelt, Oscar Martinez Mozos, Geert-Jan M. Kruijff, and Wolfram Burgard. An integrated robotic system for spatial understanding and situated interaction in indoor environments. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI'07)*, Vancouver, British Columbia, Canada, 2007.

[64] Hendrik Zender, Oscar Martinez Mozos, Patric Jensfelt, Geert-Jan M. Kruijff, and Wolfram Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems (RAS)*, 56(6):493–502, June 2008.

[65] Zoran Zivkovic, Bram Bakker, and Ben Kröse. Hierarchical map building using visual landmarks and geometric constraints. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, Edmonton, Alberta, Canada, 2005.

Paper E

Representing Spatial Knowledge in Mobile Cognitive Systems

Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir,
Adrian N. Bishop and Patric Jensfelt

Published in
11th International Conference on Intelligent Autonomous Systems

Representing Spatial Knowledge in Mobile Cognitive Systems

Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir,
Adrian N. Bishop and Patric Jensfelt

Abstract

A cornerstone for cognitive mobile agents is to represent the vast body of knowledge about space in which they operate. In order to be robust and efficient, such representation must address requirements imposed on the integrated system as a whole, but also resulting from properties of its components. In this paper, we carefully analyze the problem and design a structure of 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.

1 Introduction

Many recent advances in the fields of robotics and artificial intelligence have been driven by the ultimate goal of creating artificial cognitive systems able to perform human-like tasks. Several attempts have been made to create integrated cognitive architectures and implement them on mobile robots [4, 3, 13, 1, 2]. There is an increasing interest in, and demand for, robots that are capable of dealing with complex and dynamic environments outside the traditional industrial workplaces. 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.

Spatial knowledge constitutes a fundamental component of the knowledge base of a cognitive agent providing a basis for navigation, reasoning, planning and episodic memories. Moreover, it is a common ground for communication between a robot and a human. In order for the process of acquisition, interpreting, storing and recalling of the spatial knowledge to be robust and efficient under limited

resources and in realistic settings, the knowledge must be properly structured and represented. Such knowledge representation must address requirements imposed on the integrated system as a whole, but also resulting from properties of its components. Due to this central role, the design of a spatial knowledge representation should be one of the first steps in building a cognitive system.

In this work, we develop a structure of a spatial knowledge representation for a cognitive mobile system that we call COARSE (Cognitive lAyered Representation of Spatial knowledgE). We carefully analyze the role of a spatial representation and formulate design assumptions and requirements imposed by the functionality and components of an integrated system. Our representation is layered and represents knowledge at different levels of abstraction, from low-level sensory input to high level conceptual symbols. It is designed for representing complex, cross-modal, spatial knowledge that is inherently uncertain and dynamic and includes discrete symbols that facilitate communication with the user and components of the system. Moreover, we propose models and algorithms that could be used as instantiations of each layer of the representation.

This paper is motivated by the desire to create a framework that is powerful, robust and efficient, but most importantly suited for mobile agents performing typical human-like tasks. The literature contains many algorithms for spatial mapping and instantiations of mobile robotic systems. However, the existing representations are either designed for a very specific domain [7, 12], they concentrate on a fraction of the spatial knowledge [20, 23] or are designed to solve a single algorithmic task very efficiently rather than for use within a larger system [8, 10, 18]. The idea of this paper, is to take a step back, focus on structuring the whole body of spatial knowledge and see how an analysis of requirements can lead the way towards a powerful spatial representation for a cognitive mobile robot.

2 Related Work

There exists a broad literature on mobile robot localization, navigation and mapping and many algorithms relying on spatial knowledge have been proposed. These include solutions to such problems as Simultaneous Localization and Mapping (SLAM) [8, 15, 10, 18] or place classification [16, 20]. Every such algorithm maintains a representation of spatial knowledge. 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 [14] concentrate on lower levels of spatial knowledge abstraction and do not support higher-level conceptualization or representation of categorical information.

At the same time, we witness a growing interest in building artificial mobile cognitive systems [4, 3, 1, 2]. These are complex, usually modular, systems that require a unified and integrated approach to spatial knowledge representation. The central role of spatial knowledge in those systems has been recognized and several

authors proposed subsystems processing spatial knowledge integrated with other components such as dialogue systems [24, 22]. However, neither of those provides a clear structure of the represented knowledge, perform a thorough analysis of the needs of different components of a mobile cognitive system or encapsulates all major aspects of spatial knowledge.

The most comprehensive relevant representation has been proposed in [24]. However, it has several major drawbacks that makes it unsuitable for systems that deal with dynamic and uncertain knowledge within large-scale, complex environments. First of all, the knowledge is never fully abstracted and is always grounded in an accurate global metric map. This makes the system less robust and scalable. Moreover, the categorical knowledge is not explicitly represented. The high-level conceptualization relies on rigid ontologies and ignores uncertainties associated with represented symbols. Finally, it is modality-specific and does not allow for knowledge fusion from multiple sources. In the rest of the paper, we propose an approach to spatial knowledge representation that addresses those problems.

3 Analysis of the Problem

Before designing a representation of spatial knowledge, it is important to review the aspects a representation should focus on. In this section, we analyze those aspects and propose our definition of a generic spatial knowledge representation. Then, we formulate the problem within the context of cognitive systems.

3.1 What is a Spatial Knowledge Representation?

Following the analysis by Davis [9], we formulate several points that characterize a general representation of spatial knowledge. A spatial representation can be seen as:

a) A substitution (surrogate) for the world that allows the agent to perform reasoning about the parts of the environment which are beyond its sensory horizon. Such a surrogate is naturally imperfect, and is incomplete (some aspects are not represented), inaccurate (captured with uncertainty), and will become invalid (e.g. due to dynamics of the world that cannot be observed and is too complex to be captured by the representation). Moreover, since the representation cannot be perfect, all the inferences based on that representation, such as the outcomes of the localization process, are uncertain. The only perfect representation of the world or the environment in which the agent operates is the environment itself.

b) A set of ontological commitments that determine the terms in which the agent thinks about space. The representation defines the aspects of the world that should be represented. Moreover, it defines the level of detail at which they should be represented as well as their persistence. The ontology should be understood in more general terms, from spatial concepts and their relations to categorical models or types of features extracted from the sensory input.

c) A set of definitions that determine the reasoning that can be (and that should be) performed within the framework and the possible inferences and their outcomes. The reasoning will typically correspond to determining the current location with respect to the internal map (topologically, semantically etc.), providing necessary knowledge for the navigation process, determining the properties of a location in space etc. Moreover, the representation defines how the location of the agent is represented and in what terms it is possible to refer to points in space (e.g. in terms of metric coordinates, semantic category of a place etc.).

d) A way of structuring the spatial information so that it is computationally feasible to perform all the necessary processing and inferences in a specified time (e.g. in real time) despite limited resources.

e) A medium of communication between the agent and human. If the agent is supposed to exchange information with humans, the representation must be designed in a way that allows the agent to interpret human expressions and generate expressions that are comprehensible to humans.

f) Similarly, a medium of communication between components of an integrated system.

3.2 Spatial Representation for Mobile Cognitive Systems

In this work, we narrow the focus to mobile cognitive systems. Based on the analysis of existing approaches [3, 1, 23] as well as ongoing research on artificial cognitive systems [4], we have identified several areas of functionality, usually realized through separate subsystems, that must be supported by the representation. These include localization, navigation, and autonomous exploration, but also understanding and exploiting semantics associated with space, human-like conceptualization and categorization of space, reasoning about spatial units and their relations, human-robot communication, action planning, object finding and visual servoing, and finally recording and recalling episodic memories.

Having in mind the aforementioned functionalities, aspects covered by a representation of spatial knowledge as well as limitations resulting from practical implementations, we have identified several desired properties and designed a representation reflecting those properties.

Complex, cross-modal, spatial knowledge in realistic environments is inherently uncertain and dynamic. Therefore, it is futile to represent the environment as accurately as possible. A very accurate representation must be complex, require a substantial effort to synchronize with the world and still cannot guarantee that sound inferences will lead to correct conclusions [9]. Our primary assumption is that the representation should instead be minimal and inherently coarse and the spatial knowledge should be represented only as accurately as it is required to support the functionality of the system. Furthermore, redundancy should be avoided and whenever possible and affordable, new knowledge should should be inferred from the existing information. It is important to note that uncertainties associated with represented symbols should be explicitly modeled.

Information should be abstracted as much as possible to make it robust to dynamic changes. Moreover, representations that are more abstract should be used for longer-term storage. At the same time, knowledge extracted from immediate observations can be much more accurate (e.g. for the purpose of visual servoing). In other words, the agent should use the world as an accurate representation whenever possible. It is important to mention that rich and detailed representations should not constitute a permanent base for more abstract ones (as is the case in [24]). Similarly, space should be represented on different spatial scales from single scenes to whole environments.

Space should be discretized into a finite number of spatial units. Discretization of continuous space is one of the most important abstracting steps as it allows to make the representation robust, compact and tractable. Discretization drastically reduces the number of states that have to be considered e.g. during the planning process [11] and serves as a basis for higher level conceptualization [24].

A representation should allow not only for representing instantiations of spatial segments visited by the robot. It is equally important to provide means for representing unexplored space. Furthermore, categorical knowledge should be represented that is not specific to any particular location and instead corresponds to general knowledge about the world. Typical examples would be categorical models of appearance of places [20] or objects [19].

Finally, we focus on the fundamental role of the representation in human-robot interaction. Spatial knowledge representation should model correspondence between the represented symbols and human concepts of space. Spatial properties (e.g. shape, size), semantic categories of rooms (e.g. kitchen, office) or spatial segments (e.g. rooms, floors, buildings) recognized by humans are examples of such concepts. This correspondence could be used to generate and resolve spatial referring expressions [24] or path descriptions.

4 Structure of the Representation

In this section, we propose a representation of spatial knowledge that adheres to the desired properties formulated above. Figure 1 gives a general overview of the structure of the representation. It is sub-divided into four layers which can be regarded as sub-representations focusing on different aspects of the world, abstraction levels of the spatial knowledge and different spatial scales. Moreover, each layer defines its own spatial entities and the way the agent's position in the world is represented. The properties of each layer are summarized in Table 1.

At the lowest abstraction level, we have the sensory layer which maintains an accurate representation of the robot's immediate environment extracted directly from the robot's sensory input. Higher, we have the place and categorical layers. The place layer provides fundamental discretisation of the continuous space into a set of distinct places. The categorical layer focuses on low-level, long-term categorical models of the robot's sensory information. Finally, at the top, we have

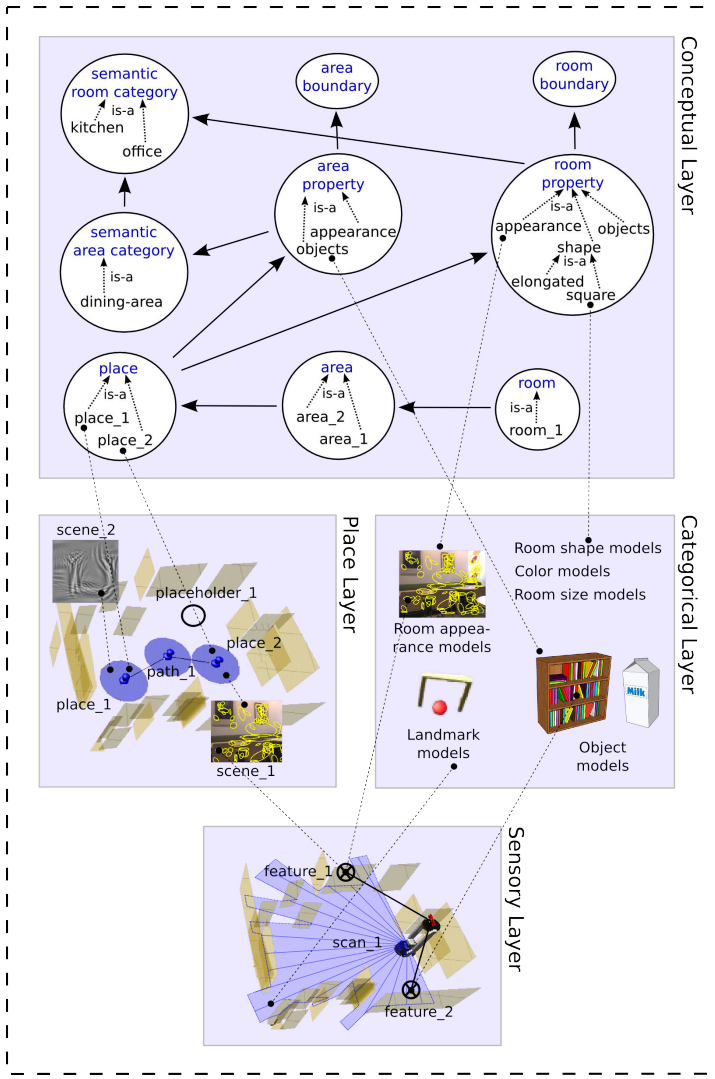


Figure 1: The layered structure of the spatial representation. The position of each layer within the representation corresponds to the level of abstraction of the spatial knowledge.

Property	Sensory Layer	Place Layer	Categorical Layer	Conceptual Layer
Aspects represented	Accurate geometry and appearance	Local spatial relations, coarse appearance, geometry	Perceptual categorical knowledge	High-level spatial concepts / Links concepts $\leftrightarrow$ entities
Agent's position	Pose within the local map	Place ID	Relationship to the categorical models	Expressed in terms of high level spatial concepts
Spatial scope	Small-scale, local	Large-scale	Global	Global
Knowledge persistence	Short-term	Long-term	Very long-term	Life-long / Very long-term

Table 1: Comparison of properties of the four layers of the spatial representation.

the conceptual layer, which associates human concepts with the categorical models in the categorical layer and groups places into human-compatible spatial segments such as rooms. The following sections provide details about each of the layers.

4.1 Sensory Layer

In the sensory layer, a detailed robocentric model of the robot's immediate environment is represented based on direct sensory input as well as data fusion over space around the robot and short time intervals. The sensory layer stores low-level features and landmarks extracted from the sensory input together with their exact position with respect to the robot. Measures of uncertainty are also included in this representation. Landmarks that move beyond a certain distance are forgotten and replaced by new information. Thus, this representation is akin to a sliding window, with robocentric and up-to-date direct perceptual information. It is also essentially bottom-up only, though directives and criteria, such as guiding the attentional process, may be imposed from upper layers.

The representation in the sensory layer helps to maintain stable and accurate information about the relative movements of the robot. Moreover, it allows for maintaining and tracking the position of various features while they are nearby. This can be useful for providing "virtual sensing" such as 360° laser scans based on short-term temporal sensory integration as well as generation of features based on spatial constellations of landmarks located outside the field of view of the sensor. Additionally, it could be used for temporal filtering of sensory input or providing robustness to occlusions. Finally, the sensory layer can provide the low level robotic movement systems with data for deriving basic control laws such as for obstacle avoidance or visual servoing.

4.2 Place Layer

The place layer is responsible for the fundamental, bottom-up discretization of continuous space. In the place layer, the world is represented as a collection of basic spatial entities called places as well as their spatial relations. Each place is defined in terms of features that are represented in the sensory layer, but also spatial relations to other places. The aim of this representation is not to represent the world as accurately as possible, but at the level of accuracy sufficient for performing required actions and robust localization despite uncertainty and dynamic variations. Similarly, the relations do not have to be globally consistent as long as they are preserved locally with sufficient accuracy. The representation of places in the place layer persists over long term.

Besides places, the place layer also defines paths between them. The semantic significance of a path between two places is the possibility of moving directly between one and the other. This does not necessarily imply that the robot has traveled this path previously. A link might be created for unexplored place e.g. based on top-down cues resulting from the dialogue with the user (e.g. when the user indicates part of the environment that should be of interest to the robot, but not immediately). In addition, the place layer explicitly represents unexplored space. Tentative places are represented which the robot would probably uncover if it moved in a certain direction.

The place layer operates on distinct places as well as their connectivity and spatial relations to neighboring places. No global representation of the whole environment is maintained. Still, since the local connectivity is available, global representation (e.g. a global metric map) can be derived when needed. This representation will not be accurate, but will preserve the connectivity and relaxed spatial relations between all the places.

4.3 Categorical Layer

The categorical layer contains long-term, low-level representations of categorical models of the robot's sensory information. The knowledge represented in this layer is not specific to any particular location in the environment. Instead, it represents a general long-term knowledge about the world at the sensory level. In this layer models of landmarks, objects or appearance-based room category or other properties of spatial segments such as shape, size or color are defined in terms of low-level features. The position of this layer in the spatial representation reflects the assumption that the ability to categorize and group sensory observations is the most fundamental one and can be performed in a feed-forward manner without any need for higher-level feedback from cognitive processes.

The categorical models stored in this layer give rise to properties that are utilized by conceptual layer. In many cases, the values of those properties will correspond to human spatial concepts, not to internal concepts of the robot (e.g. office-like appearance or elongated shape). The properties might require complicated models

that can only be inferred from training data samples. In case of models that correspond to human concepts, they can be learned in a supervised fashion, using a top-down supervision signal.

4.4 Conceptual Layer

The conceptual layer provides an ontology that represents taxonomy of the spatial concepts and properties of spatial entities that are linked to the low-level categorical models stored in the categorical layer. This associates semantic interpretations with the low-level models and can be used to specify which properties are meaningful e.g. from the point of view of human-robot interaction. Moreover, the conceptual layer represents relations between the concepts and instances of those concepts linked to the spatial entities represented in the place layer. This makes the layer central for verbalization of spatial knowledge and interpreting and disambiguating verbal expressions referring to spatial entities.

The second important role of the conceptual layer is to provide definitions of the spatial concepts related to the semantic segmentation of space based on the properties of segments observed in the environment. A building, floor, room or area are examples of such concepts. The conceptual layer contains information that floors are usually separated by staircases or elevators and that rooms usually share the same general appearance and are separated by doorways. Those definitions can be either given or learned based on asserted knowledge about the structure of a training environment introduced to the system.

Finally, the conceptual layer provides definitions of semantic categories of segments of space (e.g. rooms) in terms of values of properties of those segments. The properties can reflect the general appearance of a segment as observed from a place, its geometrical features or objects that are likely to be found in that place.

5 Instantiations

This section indicates specific models and algorithms maintaining those models that we propose to use for representing knowledge stored in each layer.

We propose to realize the sensory layer using a robocentric, metric SLAM [6, 5]. Robocentric mapping reflects the properties of the sensory layer and allows for a straightforward treatment of forgetting knowledge that falls outside a certain horizon around the robot. The robocentric map can be seen as a sliding window centered on the robot and containing a detailed view of the world, which allows the robot to maintain a drift free estimate of the pose as long as it stays in a local region of space. The SLAM algorithm explicitly represents the uncertainty associated with the pose of the robot and the location of all landmarks in the local surrounding using a multivariate Gaussian distribution [6, 5].

We propose to instantiate the place layer based on the mapping framework proposed in [21]. Central to the approach is the place map represented as a collection of places. A place is defined by a subset of values of arbitrary, possibly complex,

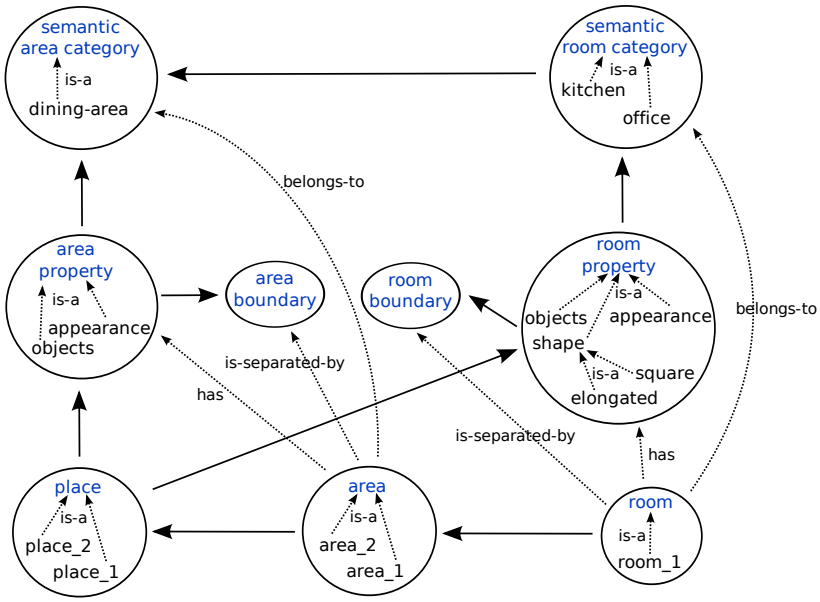


Figure 2: Overview of a possible instantiation of the conceptual layer. The solid arrows represent dependencies, while the dashed arrows illustrate the ontology that represents the taxonomy of spatial concepts and properties of spatial entities.

distinctive features and spatial relations reflecting the structure of the environment. The features provide information about the world and can be perceived by an agent when at that place. In this sense, the places build on the perception of the agent and are based on its perceptual capabilities.

The categorical layer can be seen as an ensemble of categorical models of the robot's sensory information. The literature provides a broad range of models that could be used for this purpose. First, in order to represent visual and geometrical properties of areas in the environment, we suggest to use the multi-modal place classification algorithm presented in [20]. Other methods can be employed for representing landmarks (e.g. doors [17]) and object categories [19].

For the conceptual layer, we propose a possible instantiation presented in Figure 2. The conceptual layer provides an ontology that represents the taxonomy of the spatial concepts and properties as well as dependencies between the concepts, properties and instances of spatial entities. We use a fixed, handcrafted ontology for representing the taxonomy and a probabilistic model for representing the dependencies. In such an approach, the ontology is largely encoded in the structure of the probabilistic model. We represent the location of the robot within segments of space (e.g. a room or an area such as a dining area), the observed properties of areas and rooms as well as semantic categories of areas and rooms in terms of

random variables. In the illustration in Figure 2, we can consider the circles as random variables and the solid arrows as dependencies within a graphical model. At the same time, the *is-a* relations link the random variables with their values. Further, the model represents the spatial hierarchy of segments of space. There is a dependency between the location of the robot at different levels of this hierarchy (e.g. a room and an area within the room). Moreover, the dependency between the instance of a place and the properties of areas and rooms observed from this place is represented. Those properties in turn influence the semantic categories of areas or rooms to which the place belongs. Finally, the proposed model represents the dependency between the area and room properties observed as the robot explores the environment and the probability that the robot crossed a boundary of a spatial segment. This link effectively defines the concepts of a room and an area and can be used to provide semantic segmentation of space.

6 Conclusions and Future Works

In this paper, we presented an analysis of the requirements for a spatial knowledge representation for cognitive systems and proposed a layered representation that conforms to those requirements. The representation provides a unified and coherent view on the structure of spatial knowledge and a basis for designing artificial cognitive systems. We further proposed specific models and algorithms as possible instantiations. Future work will focus on integrating those algorithms, which so far were only evaluated in separation, into a complete spatial subsystem providing spatial understanding capabilities for a mobile robot.

References

[1] EU FP6 Integrated Project COGNIRON: The Cognitive Robot Companion. URL `http://www.cogniron.org`.

[2] EU FP6 Integrated Project RobotCub. URL `http://www.robotcub.org/`.

[3] EU FP6 IST Cognitive Systems Integrated Project CoSy: Cognitive Systems for Cognitive Assistants. URL `http://www.cognitivesystems.org/`.

[4] EU FP7 ICT Cognitive Systems Large-Scale Integrating Project CogX: Cognitive Systems that Self-Understand and Self-Extend. URL `http://cogx.eu/`.

[5] Adrian N. Bishop and Patric Jensfelt. Stochastically convergent localization of objects and actively controllable sensor-object pose. In *Proceedings of the 10th European Control Conference (ECC'09)*, Budapest, Hungary.

[6] José A. Castellanos, Ruben Martinez-Cantin, Juan D. Tardós, and José Neira. Robocentric map joining: Improving the consistency of EKF-SLAM. *Robotics and Autonomous Systems (RAS)*, 55(1):21–29, 2007.

[7] Antonio Chella and Irene Macaluso. The perception loop in CiceRobot, a museum guide robot. *Neurocomputing*, 72(4-6), 2009.

[8] Mark Cummins and Paul M. Newman. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Proceedings of Robotics: Science and Systems (RSS'09)*, 2009.

[9] Randall Davis, Howard Shrobe, and Peter Szolovits. What is a Knowledge Representation? *AI Magazine*, 14(1):17–33, 1993.

[10] Udo Frese and Lutz Schröder. Closing a million-landmarks loop. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, 2006.

[11] Nick A. Hawes, Hendrik Zender, Kristoffer Sjöö, Michael Brenner, Geert-Jan M. Kruijff, and Patric Jensfelt. Planning and acting with an integrated sense of space. In *Proceedings of the 1st International Workshop on Hybrid Control of Autonomous Systems - Integrating Learning, Deliberation and Reactive Control (HYCAS)*, Pasadena, CA, USA, July 2009.

[12] Patric Jensfelt, Erik Förell, and Per Ljunggren. Automating the marking process for exhibitions and fairs - the making of Harry Plotter. *IEEE Robotics and Automation Magazine*, 14(3):35–42, 2007.

[13] Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems, Special Issue on Human and Robot Interactive Communication*, 4(1):125–138, 2007.

[14] Benjamin Kuipers. The Spatial Semantic Hierarchy. *Artificial Intelligence*, 119(1-2): 191–233, 2000.

[15] Michael J. Milford and Gordon F. Wyeth. Mapping a suburb with a single camera using a biologically inspired SLAM system. *IEEE Transactions on Robotics*, 24(5): 1038–1053, October 2008.

[16] Oscar Martinez Mozos, Rudolph Triebel, Patric Jensfelt, Axel Rottmann, and Wolfram Burgard. Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems (RAS)*, 55(5):391–402, 2007.

[17] Ana Cris Murillo, Jana Košecká, J J Guerrero, and C Sagüés. Visual door detection integrating appearance and shape cues. *Robotics and Autonomous Systems (RAS)*, 56(6), 2008.

[18] Lina María Paz, Patric Jensfelt, Juan D. Tardós, and José Neira. EKF SLAM updates in O(n) with Divide and Conquer SLAM. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, Roma, Italy, 2007.

[19] Axel Pinz. Object categorization. *Foundations and Trends in Computer Graphics and Vision*, 1(4):255–353, 2006.

[20] Andrzej Pronobis, Oscar Martinez Mozos, Barbara Caputo, and Patric Jensfelt. Multi-modal semantic place classification. *The International Journal of Robotics Research (IJRR), Special Issue on Robotic Vision*, 29(2-3):298–320, February 2010.

[21] Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Adrian N. Bishop, and Patric Jensfelt. A framework for robust cognitive spatial mapping. In *Proceedings of the 14th International Conference on Advanced Robotics (ICAR'09)*, Munich, Germany, June 2009.

[22] Thorsten Spexard, Shuyin Li, Britta Wrede, Jannik Fritsch, Gerhard Sagerer, Olaf Booij, Zoran Zivkovic, Bas Terwijn, and Ben Kröse. BIRON, where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, October 2006. ISBN 1-4244-0258-1.

[23] Sebastian Thrun, Arno Bücken, Wolfram Burgard, Dieter Fox, Thorsten Fröhlinghaus, Daniel Henning, Thomas Hofmann, Michael Krell, and Timo Schmidt. Map learning and high-speed navigation in RHINO. In David M. Kortenkamp, R. P. Bonasso, and R. Murphy, editors, *AI-based Mobile Robots: Case Studies of Successful Robot Systems*. MIT Press, 1998.

[24] Hendrik Zender, Oscar Martinez Mozos, Patric Jensfelt, Geert-Jan M. Kruijff, and Wolfram Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems (RAS)*, 56(6):493–502, June 2008.

Paper F

Understanding the Real World: Combining Objects, Appearance, Geometry and Topology for Semantic Mapping

Andrzej Pronobis, and Patric Jensfelt

Published in
Technical Report

Understanding the Real World: Combining Objects, Appearance, Geometry and Topology for Semantic Mapping

Andrzej Pronobis, and Patric Jensfelt

Abstract

A cornerstone for mobile robots operating in man-made environments and interacting with humans is representing and understanding the human semantic concepts of space. In this report, 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 experiments offline and online on a mobile robot showing the efficiency and usefulness of our system.

1 Introduction

In this report we focus on the understanding of space to, for example, facilitate interaction between humans and robots and increase the efficiency of the robot performing tasks in man-made environments. We consider applications where the robot is operating in an indoor office or domestic environment, i.e. environments which have been made for and are, up until now, almost exclusively inhabited by humans. In such an environment human concepts such as rooms and objects and properties such as the size and shape of rooms are important, not only for the sake of the interaction with humans but also for knowledge representation and abstraction of spatial knowledge. We will describe the system in the context of a mobile robot (see Fig. 1) but most of the system would remain unchanged if used as part of, e.g., a wearable device.

The main contribution of this work is a way of combining information about the existence of objects, the appearance, geometry and topology of space for semantic mapping in a principled manner. It builds on our previous work [16] where we presented a system for multi-cue integration of laser and vision data for place categorization. A fundamental difference from that work is that we now have de-

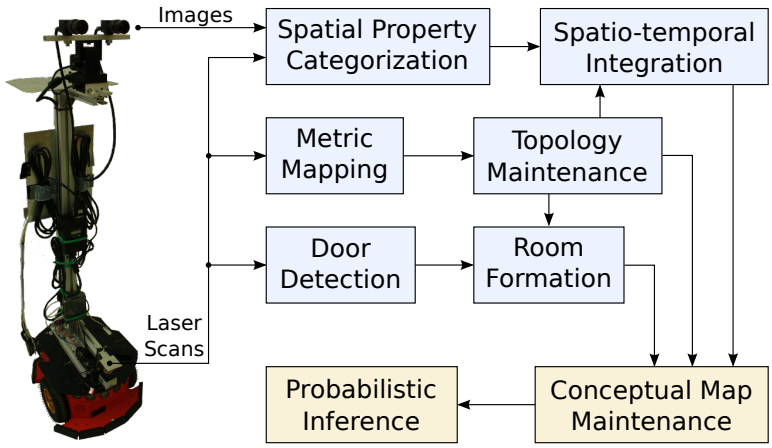


Figure 1: Dora the Explorer as well as the elements and the data flow inside the semantic mapping system.

coupled the lowest levels of information from the categorization by introducing the so called properties. This allows us to incorporate additional sources of knowledge and describe the space at much finer level of granularity.

Comparing to our previous approach, the vision and laser pipe lines now feed into modules estimating the values of appearance and geometric properties of space instead of directly categorizing rooms. In this way room categories are not directly defined based on the low level sensor data but in terms of spatial properties. This has several advantages. It paves the way for better scalability. It makes training of new categories easier and is therefore an important step towards a system that is able to support life-long learning. The properties can correspond to human concepts of space. The use of such human understandable properties provides better support for verbalization of knowledge, e.g., the corridor is large (size property) and elongated (shape property) as well as the dual, i.e., interpreting what a human says and ultimately learning models for new categorizes based on human input. Finally, additional spatial properties such as based on objects or even actions observed in the environment can be easily incorporated.

We believe that objects play an important role in understanding space. Introducing properties describing the existence of certain objects provides a seamless way to integrate objects in the above mentioned system. A by-product of the system presented in this report is that it allows for predicting the existence of objects and the values of spatial properties. That is, given that, for example, appearance and shape indicate that a certain room is a kitchen the object properties associated with such room category might suggest that it is likely to find a cereal box in that room.

Our new system also allows for human input being treated in the same principled

way as the information from a camera or a laser scanner. That is, if a human tells us that there is a certain object nearby or that we are in the room next to the kitchen this type of information can be incorporated. Furthermore, by incorporating information about the topology of space we can infer properties of space even without having made any observations there. For example, starting in an office the system would be able to say that it is very likely that the neighbouring room is a corridor because that is the typical topology.

Another advantage with the property based system is that it allows us to train the level above the properties in the system directly using ground truth information. That is, instead of training based on the outcome from the low level processing we can train with data from common sense databases or crawling the internet for information about typical topologies, objects-room relations, etc.

The presented approach is evaluated offline on a comprehensive database, COLD-Stockholm, capturing appearance and geometry of almost 50 rooms belonging to different semantic categories as well as online in the same environment on a mobile robot.

1.1 Outline

Section 2 relates the work in this report to what has been presented in the literature. Section 3 goes into a bit more detail about spatial understanding and more specifically about the employed spatial model. In Section 4 we describe the properties we use in the system and in Section 5 we describe our conceptual map. Section 6 gives a system overview and describes how its components are connected. Section 7 describes the experimental setup and Section 8 presents experimental results for room categorization both offline and online. Finally, Section 9 draws conclusions and suggests avenues for future research.

2 Related Work

The system we present here provides a much broader functionality than a plain place categorization system, but as place categorization is one of its typical uses, we give an overview of the work in that research field. Place categorization has been addressed both by the computer vision and the robotics community. In computer vision the problem is often referred to as scene categorization. Although also related, object categorization methods are not covered here. However, as already mentioned, we believe that objects are key to understanding space and we will include them in our representation but will make use of standard methods for recognizing/categorizing them.

In computer vision one of the first works to address the problem of place categorization is [20] based on the so called "gist" of a scene. One of the key insights in the paper is that the context is very important for recognition and categorization of both places and objects and that these processes are intimately connected. Place recognition is formulated in the context of localization and information about the

connectivity of space is utilized in a Hidden Markov Model (HMM). Place categorization is also addressed using an HMM. In [24] the problem of grouping images into semantic categories is addressed. It is pointed out that many natural scenes are ambiguous and that the performance of the system is often quite subjective. That is, if two people are asked to sort the images into different categories they are likely to come up with different partitions. [24] 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 when used in the categorization, i.e. the saliency of that cue. The system is evaluated in natural outdoor scenes. In [2] another method is presented for categorization of outdoors scenes based on representing the distribution of codewords in each scene category. In [26] a new image descriptor, PACT, is presented and shown to give superior results on the datasets used in [20, 2].

In robotics, one of the early systems for place recognition is that of [21] where color histograms are used to model the appearance of places in a topological map and place recognition performed as a part of the localization process. Later [10] uses laser data to extract a large number of features used to train classifiers using AdaBoost. This system shows impressive results based on laser data alone. The system is not able to identify and learn new categories: adding a new category required off-line re-training, no measure of certainty and it segmented space only implicitly by providing an estimate of the category for every point in space. In [11] this work is extended to also incorporate visual information in the form of object detections. Furthermore, this work also adds an 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. In [13, 14] a vision-only place recognition system is presented. Super Vector Machines (SVMs) are used as classifiers. The characteristics are similar to those of [10]; cannot identify and learn new categorizes on-line, only works with data from a single source and classification was done frame by frame. In [8, 15] a version of the system supporting incremental learning is presented. The other limitations remains the same. In [12] a measure of confidence is introduce as a means to better fuse different cues and also provide the consumer of the information with some information about the certainty in the end result. In [16] the works in [10, 13, 14] are combined using an SVM on top of the laser and vision based classifiers. This allows the system to learn what cues to rely on in what room category. For example, in a corridor the laser based classifier is more reliable than vision whereas in rooms the laser does not distinguish between different room types. Segmentation of space is done based on detecting doors that are assumed to delimit the rooms. Evidence is accumulated within a room to provide a more robust and stable classification. It is also shown that the method support categorization and not only recognition. In [25] the work from [26] is extended with a new image descriptor, CENTRIS, and a focus on visual place categorization in indoor environment for robotics. A database, VPC, for benchmarking of vision based place categorization systems is also presented. A Bayesian filtering scheme is added on top of the frame based categorization to

increase robustness and give smoother category estimates. In [18] the problem of place categorization is addressed in a drastically different and novel way. The problem is cast in a fully probabilistic framework which operates on sequences 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 to allow for real-time performance. All information deemed to belong to the same segment is used to estimate the category for that segment using a bag-of-words technique. In [28] a system for clustering panoramic images into convex regions of space indoors is presented. These regions correspond roughly with the human concept of rooms and are defined by the similarity between the images. In [22] panoramic images from indoor and outdoor scenes are clustered into topological regions using incremental spectral clustering. 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.

As mentioned above [11] makes use of object observations to perform the place categorization. In [4] objects also play a key role in the creation of semantic maps and the *anchoring* problem, i.e., that of associating sensor level information with the same entity at the symbolic level, is studied. In [19] a 3D model centered around objects is presented as a way to model places and to support place recognition. In [23] a Bayesian framework for connecting objects to place categories is presented. In [27] the work in [11] is combined with detections of objects to deduce the specific category of a room in a first-order logic way.

3 Semantic Spatial Understanding

In order to build a semantic mapping system, it is necessary to make certain assumptions about how the vast body of the spatial knowledge should be represented. The functionality of our system is centred around the representation of complex, cross-modal, spatial knowledge that is inherently uncertain and dynamic. The representation employed here follows the principles presented in [17]. In addition to supporting standard applications such as localisation and path planning, it integrates instance knowledge with conceptual world knowledge using a probabilistic framework. Below, we first describe the fundamental concepts that we use to describe space and then present an overview of a spatial knowledge representation on top of which our system is built.

3.1 The Ontology of Space

Our primary assumption is that spatial knowledge should be abstracted. This keeps the complexity under control, makes the knowledge more robust to dynamic changes, and allows to infer additional knowledge about the environment. One of the most important steps in abstraction of spatial knowledge is discretization of continuous space. In our view, the environment is decomposed into discrete areas

called places. Places connect to other places using paths which are generated as the robot travels the distance between them. Thus, places and paths constitute the fundamental topological graph of the environment.

An important concept employed by humans in order to group locations is that of a room. Rooms tend to share similar functionality and semantics which make them a good candidate for integrating semantic knowledge over space. In the case of indoor environments, rooms are usually separated by doors or other narrow openings. Thus, we propose to use a door detector and perform reasoning about the segmentation of space into rooms based on the doorway hypotheses.

Many other concepts than simply those related to the topology are being used by humans to describe space. In this work, we focus on the combination of objects, which we believe are strongly related to the semantic category of a place where they are typically located, with other spatial properties. As properties, we identify shape of a room (e.g. elongated), size of a room (e.g. large, compared to other typical rooms) as well as the general appearance of a room (e.g. office-like appearance).

3.2 Spatial Knowledge Representation

The spatial knowledge representation on top of which we build our system is presented in Fig. 2. It consists of four layers corresponding to different levels of abstraction, from low-level sensory input to high-level conceptual symbols. Each layer defines its own spatial entities and the way the agent's position in the world is represented.

The knowledge is abstracted and represented only as accurately as necessary, and uncertainty is present at all levels. This keeps the complexity of the representation under control, makes the knowledge more robust to dynamic changes, and allows to infer additional knowledge about the environment.

The lowest level of our representation is the sensory layer. This maintains an accurate representation of the robot's immediate environment. Above this are the place and categorical layers. The place layer discretises 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 (in our case pre-trained) of the robot's sensory information which are not specific to any particular location or environment. These could be the sensory models of object categories, but also values of spatial properties such as an elongated shape or office-like appearance. On top of this, the conceptual layer creates a unified representation relating sensed instance knowledge to general conceptual knowledge. It includes a 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. In the following sections, we focus on the concrete implementations of the principles outlined here and algorithms maintaining the representations in each of the layers.

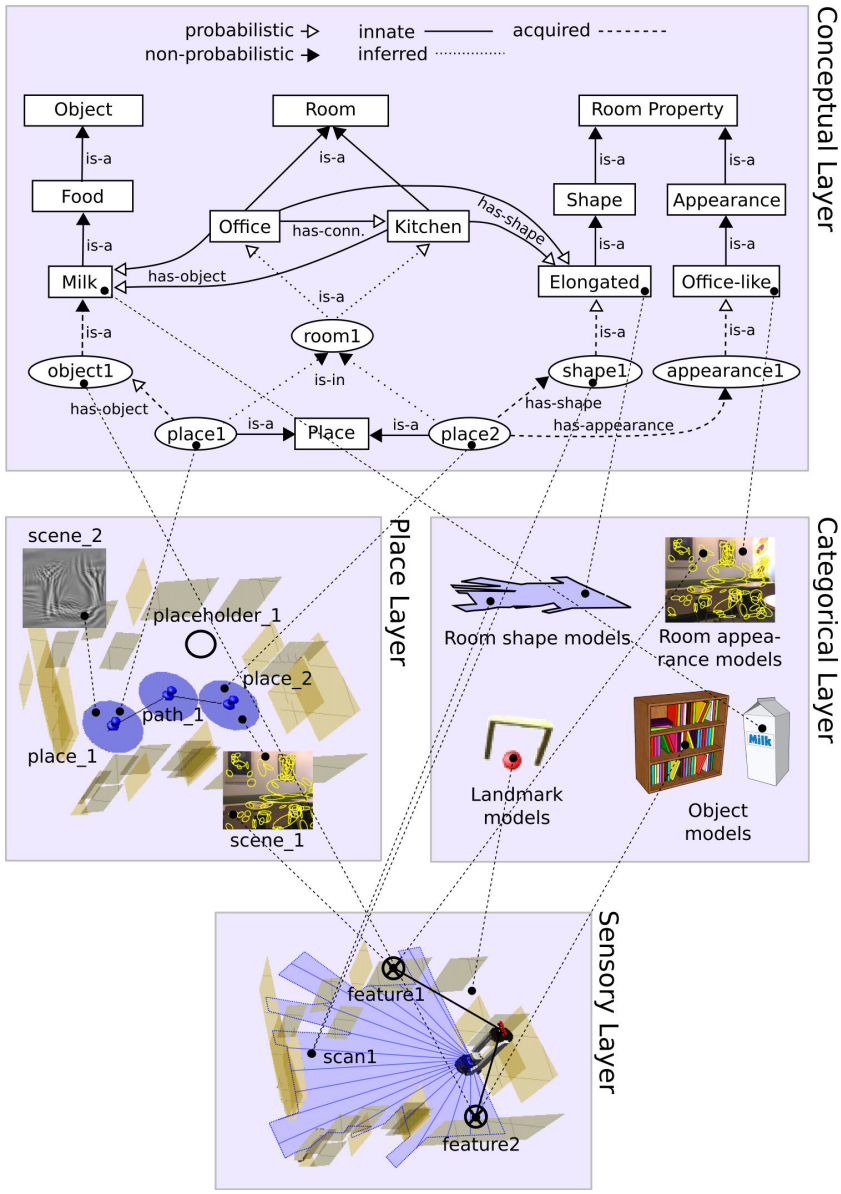


Figure 2: The layered structure of the spatial representation. The position of each layer within the representation corresponds to the level of abstraction of the spatial knowledge. The conceptual layer illustrates part of the ontology representing both instance and predefined world knowledge.

4 Categorical Models of Sensory Information

The system employs categorical models of sensory information which abstract the information into a set of spatial concepts. These models correspond to the categorical layer of the spatial representation.

Independent models of shape, size and appearance properties are built. To provide sufficient robustness and tractability in the presence of noisy, high-dimensional information, we use non-linear kernel-based discriminative classifier models, namely Support Vector Machines, as proposed in [16]. Those models are trained on features extracted directly from the robot's sensory input. Following [16], we use a set of simple geometrical features extracted from laser range data in order to train the shape and size models. The appearance models are build from two types of visual cues, global, Composed Receptive Field Histograms (CRFH) and local based on the SURF features discretized into visual words [1]. We compute CRFH from second order normalised Gaussian derivative filters applied to the illumination channel at two scales. The two visual features are further integrated using the Generalized Discriminative Accumulation Scheme (G-DAS [16]). In case of SVMs, special care must be taken in choosing an appropriate kernel function. Here we used the RBF kernel for the geometrical shape model and χ^2-kernel for the visual appearance model.

The models are trained from sequences of images and laser range data recorded in multiple instances of rooms belonging to different categories and under various different illumination settings (during the day and at night). By including several different room instances into training, the acquired model can generalise sufficiently to provide categorisation rather than instance recognition. In order to measure the uncertainty associated with the generated hypotheses, confidence measures are derived from the distances between the classified samples and discriminative model hyperplanes [16].

5 The Conceptual Map

The key component of our semantic mapping approach is the probabilistic conceptual map which can be seen as a realization of the conceptual layer of the spatial representation. In order to fully exploit the uncertainties provided by the multimodal lower-level models, the robot needs to be capable of uncertain reasoning on the conceptual level. Below, we first present the uncertain ontology of the conceptual map relating sensed instance knowledge and general conceptual knowledge. Then, we provide an implementation of the map in terms of a probabilistic graphical model.

5.1 Uncertain Ontology

The ontology of spatial concepts and instances of those concepts implemented in the conceptual map is presented in Fig. 2. In order to represent the uncertainty associ-

ated with some of the relationships, we extended the standard ontology notation by annotating relations as either probabilistic or non-probabilistic. The resulting ontology defines a taxonomy of concepts through hyponym relationships (is-a) as well as relations between concepts (has-a). As in [27], the ontology distinguishes three primary sources of knowledge: *predefined* (taxonomy and conceptual common-sense knowledge, e.g. the likelihood that cornflakes occur in kitchens), *acquired* (knowledge acquired using the robot's sensors), and finally *inferred* (knowledge generated internally, e.g. that the room is likely to be a kitchen, because you are likely to have observed cornflakes in it). We could further differentiate between acquired knowledge and *asserted* knowledge which can be obtained by interaction with a human.

The ontology ties the concepts to instance symbols derived from the lower level representations. The instance knowledge includes the presence of objects and sensed spatial properties such as shape, size, appearance and topology. The conceptual knowledge comprises common-sense knowledge about the occurrence of objects in rooms of different semantic categories, and the relations between these categories and the aforementioned spatial properties.

In our system, the "has-a" relations for rooms, objects, shapes, sizes and appearances were acquired by analysing common-sense knowledge available through the world wide web (for details see [5]) as well as annotations available together with the database described in this report. The relation linking rooms and objects was first bootstrapped using a part of the *Open Mind Indoor Common Sense* database[1]. Obtained object-location pairs were then used to generate '*obj* in the *loc*' queries to an online image search engine. The number of returned hits was used to obtain the probabilities of existence of an object of a certain category in a certain type of room. All relations that were not directly present in the obtained results, were assumed to hold with a certain constant probability.

5.2 Probabilistic Inference

The conceptual map constructed according to the ontology presented above was implemented using a chain graph probabilistic model [7]. Chain graphs are a natural generalization of directed (Bayesian Networks) and undirected (Markov Random Fields) graphical models. As such, they allow for modelling both "directed" causal as well as "undirected" symmetric or associative relationships, including circular dependencies.

The joint density f of a distribution that satisfies the Markov property associated with a chain graph can be written as [7]:

$$f(x) = \prod_{\tau \in T} f(x_\tau | x_{pa(\tau)}),$$

[1] `http://openmind.hri-us.com/`

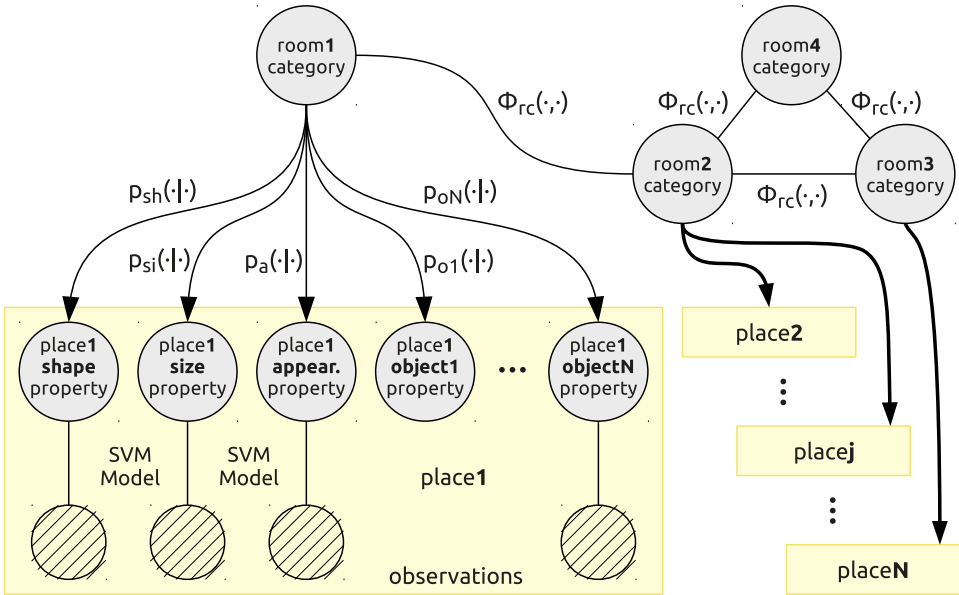


Figure 3: Structure of the chain graph model compiled from the conceptual map. The vertices represent random variables. The edges represent the directed and undirected probabilistic relationships between the random variables. The textured vertices indicate observations that correspond to sensed evidence.

where $pa(\tau)$ denotes the set of parents of vertice τ. This corresponds to an outer factorization which can be viewed as a directed acyclic graph with vertices representing the multivariate random variables X_τ, for τ in T (one for each chain component). Each factor $f(x_\tau|x_{pa(\tau)})$ factorizes further into:

$$f(x_\tau|x_{pa(\tau)}) = \frac{1}{Z(x_{pa(\tau)})} \prod_{\alpha \in A(\tau)} \phi_\alpha(x_\alpha),$$

where $A(\tau)$ represents sets of vertices in the normalized undirected graph $\mathcal{G}_{\tau \cup pa(\tau)}$, such that in every set, there exist edges between every pair of vertices in the set. The factor Z normalizes $f(x_\tau|x_{pa(\tau)})$ into a proper distribution.

In order to perform inference on the chain graph, we first convert it into a factor graph representation and apply an approximate inference engine, namely Loopy Belief Propagation [9], to comply with time constraints imposed by the robotic applications.

The structure of the chain graph model is presented in Fig. 3. The structure of the model depends on the topology of the environment. Each discrete place is represented by a set of random variables connected to variables representing semantic category of a room. Moreover, the room category variables are connected by undirected links to one another according to the topology of the environment.

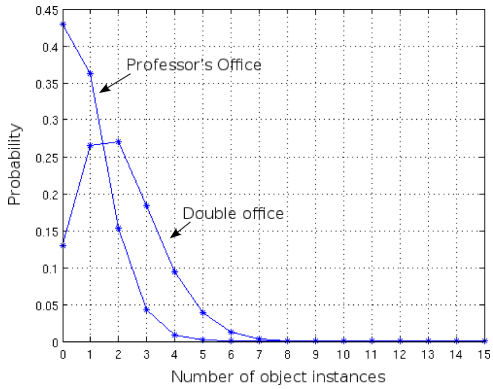


Figure 4: The Poisson distributions modelling the existence of a certain number of objects in a room on the example of computers present in a double office and a professor's office.

The potential functions $\phi_{rc}(\cdot,\cdot)$ represent the type knowledge about the connectivity of rooms of certain semantic categories.

The remaining variables represent shape, size and appearance properties of space and presence of a certain number of instances of objects as observed from each place. These can be connected to observations of features extracted directly from the sensory input. As explained in Section 4, these links are quantified by the categorical models in the categorical layer. Finally, the functions $p_{sh}(\cdot|\cdot)$, $p_{si}(\cdot|\cdot)$, $p_a(\cdot|\cdot)$, $p_{o_i}(\cdot|\cdot)$ utilise the common sense knowledge about object, spatial property and room category co-occurrence to allow for reasoning about other properties and room categories.

The conditional probability distributions $p_{o_i}(\cdot|\cdot)$ are represented by Poisson distributions. The parameter λ of the distribution allows to set the expected number of object occurrences. This is exemplified in Fig. 4 presenting two distributions corresponding to the relation between the number of computers in different types of offices used later in the experiments. In the specific case of the double office, we set the expected number of computers to two. However, in all remaining cases, including the professor's office, the parameter λ was calculated to match the probability of there being no objects of a certain category as provided by the common sense knowledge databases. The result is that the room is more likely to be a double office rather than a professor's office if there are multiple computers in it.

6 System Overview

Having described the representations and the primary elements of the system, we now explain the data flow through the system and mention all the remaining components. A coarse visualization of the data flow is presented in Fig. 1.

The layered structure of the spatial knowledge representation naturally permits the existence of data driven processes that abstract knowledge existing in the lower-level layers to contribute to knowledge in higher-level layers. This is the general principle reflected by the data flow described below. In order to make those processes tractable, the updates are performed only if a substantial change (according to a modality-specific heuristic) has occurred.

First, mapping and topology maintenance processes create the place map. A SLAM algorithm [3] builds a metric map of the environment which can be seen as the sensory layer of the representation. The metric map is further discretized into places distributed spatially in the metric map. The places together with paths obtained by traversing from one place to another constitute the place map of the place layer. Then, based on the information about the connectivity of places and the output of a template-based laser door detector, a process forms rooms by clustering places that are transitively interconnected without passing a doorway. Since the door detection algorithm can produce false positives and false negatives, room formation must be a *non-monotonic* process to allow for knowledge revision. Room formation and maintenance is handled by a general purpose rule engine, which is able to make non-monotonic inferences in its symbolic knowledge. The approach is an adaptation of the one by [6].

The categorical models are provided with sensory information from the laser range finder and a camera. This information is classified and confidence estimates are provided indicating the similarity of the sensory input to each of the categorical models. The estimated confidence information is then accumulated over each of the viewpoints observed by the robot while being in a certain place [16] and further normalised to form potentials. The categorisation results are fed back into the chain graph triggering an inference in the probabilistic model. Accordingly, *room categorisation* is performed as a result of the reasoning process in the conceptual map.

7 Experimental Scenario

All the categorical models used in the experiments were trained on the COLD-Stockholm database. COLD-Stockholm is a new database acquired as an extension of the COLD database[2]. Several parts of the database were previously used during the RobotVision@ImageCLEF[3] contests and proved to be challenging in the context of room categorization.

7.1 The COLD-Stockholm Database

The database consists of multiple sequences of image, laser range and odometry data. The sequences were acquired using the MobileRobots PowerBot robot plat-

[2] `http://www.cas.kth.se/COLD`
[3] `http://www.robotvision.info`

Figure 5: Examples of images from the COLD-Stockholm database acquired in 9 different rooms. A video illustrating the acquisition process is available on the website of the database.

form equipped with a stereo camera system in addition to a laser scanner. The acquisition was performed on four different floors (4th to 7th) of an office environment, consisting of 47 areas (usually corresponding to separate rooms) belonging to 15 different semantic and functional categories and under several different illumination settings (cloudy weather, sunny weather and at night). The floors are structurally similar but the indivudual rooms are quite different. The robot was manually driven through the different floors of the environment while continuously acquiring images at a rate of 5fps. Each data sample was then labelled as belonging to one of the areas according to the position of the robot during acquisition. Examples of images from the COLD-Stockholm database are shown in Fig. 5. More detailed information about the database can be found online[4].

[4] `http://www.cas.kth.se/cold-stockholm`

Property	Cues	Classification rate
Shape	Geometric features	84.9%
Size	Geometric features	84.5%
Appearance	CRFH	80.5%
Appearance	BOW-SURF	79.9%
Appearance	CRFH + BOW-SURF	84.9%

Table 1: Classification rates obtained for each of the properties and cues.

7.2 Experimental Setup

In order to guarantee that the system will never be tested in the same environment in which it was trained, we have divided the COLD-Stockholm database into two subsets. For training and validation, we used the data acquired on floors 4, 5 and 7. The data acquired on floor 6 were used for testing during our offline experiments and the online experiment was performed on the same floor.

For the purpose of the experiments presented in this report, we have extended the annotation of the COLD-Stockholm database to include 3 room shapes, 3 room sizes as well as 7 general appearances. The room size and shape, were decided based on the length ratio and maximum length of edges of a rectangle fitted to the room outline. These properties together with 6 object types defined 11 room categories used in our experiments. The values of the properties as well as the room categories are listed in Fig. 8.

8 Experiments

We performed two types of experiments. First, offline to evaluate the performance of our property classifiers. Then, we used the models obtained during the offline experiments and performed real-time semantic mapping on a mobile robot.

8.1 Offline Experiments

The offline experiments evaluated the performance of each of the property categorizers separately. First, the rooms having the same values of properties were grouped to form the training and validation datasets. Then, parameters of the models were obtained by cross-validation. Finally, all training and validation data were collected together and used for training the final models which were evaluated on test data acquired in previously unseen rooms.

The classification rates obtained during those experiments for each of the properties and cues are presented in Tab. 1. The models were trained and tested on 3 different shapes, 3 different sizes and 7 different appearances. The rates were obtained separately for each of the classes and then averaged in order to exclude the influence of unbalanced testing set. We can see that all classifiers provided a

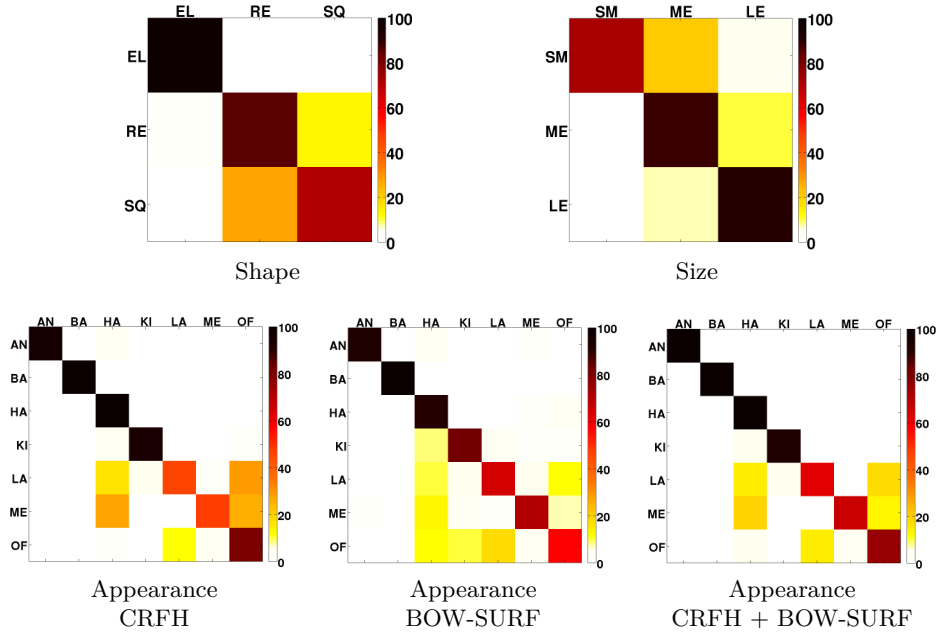


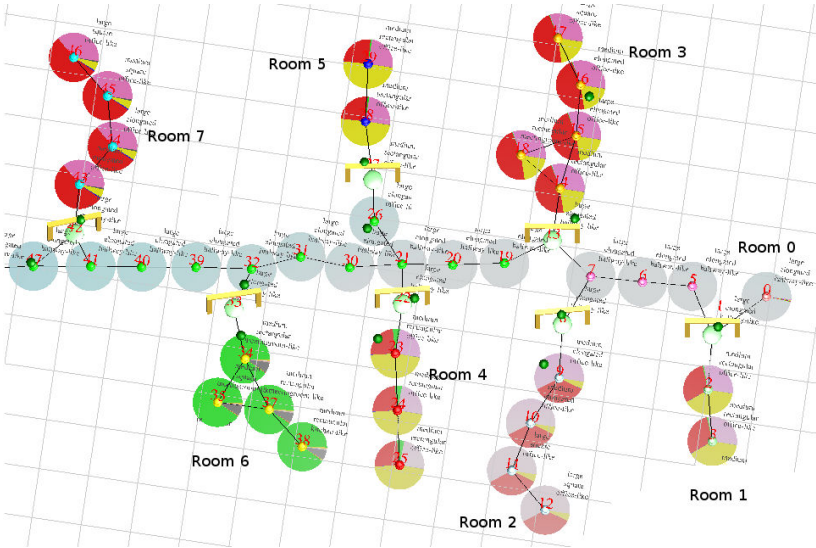
Figure 6: Confusion matrices for the offline experiments with sensory categorical models of each of the properties.

recognition rate above 80%. Additionally, we see that integrating two visual cues (CRFH and BOW-SURF) increased the classification rate of the appearance property by almost 5%. Moreover, from the confusion matrices in Fig. 6 we see that the confusion occurs always between property values being semantically close and in case of appearance is largely reduced by cue integration.

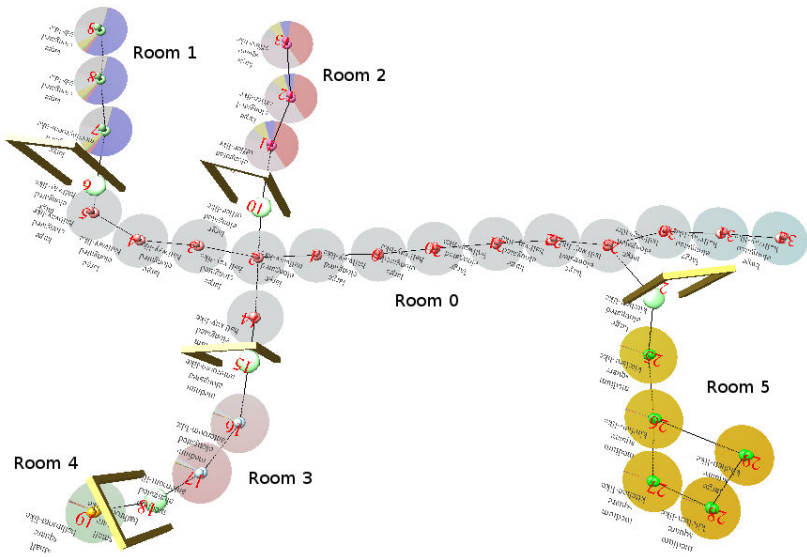
8.2 Online Experiments

The models obtained during the offline experiments were used in the semantic mapping system during the online experiments. The experiments were performed on the 6th floor of the building where the COLD-Stockholm database was acquired, i.e. in the part which was not used for training. The robot was manually driven through 15 different rooms while performing real-time semantic mapping without relying on any previous observations of the environment. The obtained maps of parts of the environment (A and B) are presented in Fig. 7.

The robot recorded beliefs about the shapes, sizes, appearances, objects found and the room categories for every significant change event in the conceptual map. The results for the two parts of the environment are presented in Fig. 8. Each column in the plot corresponds to a single event and the source of that event is



Part A



Part B

Figure 7: Topological maps of the environment anchored to a metric map indicating the outcomes of room segmentation and categorization (best viewed in color). The pie charts indicate the location of places in the environment and the probability distribution over the inferred room categories (each color corresponds to a room category). For the detailed information about the inferred categories, see Fig. 8.

indicated using dots (changes) and crosses (additions) at the bottom. At certain points in time, the robot was provided with asserted human knowledge about the presence of objects in the environment.

By analysing the events and beliefs for part A, we see that the system correctly identified the first two rooms as a hallway and a single office using purely shape, size and general appearance (there are no object related events for those rooms). The next room was properly classified as a double office, and that belief was further enhanced by the presence of two computers. The next room was initially identified as a double office until the robot was given information that there is a single computer in this room. This was an indication that the room is a single person office that due to its dimensions is likely to belong to a professor.

Looking at part B, we see that the system identified most of the room categories correctly with the exception of a single office which due to a misclassification of size was incorrectly recognized as a double office. The experiment proved that the system can deliver an almost perfect performance by integrating multiple sources of semantic information.

A video illustrating showing the system in action is available online at `http://www.pronobis.pro/research/semantic-mapping`.

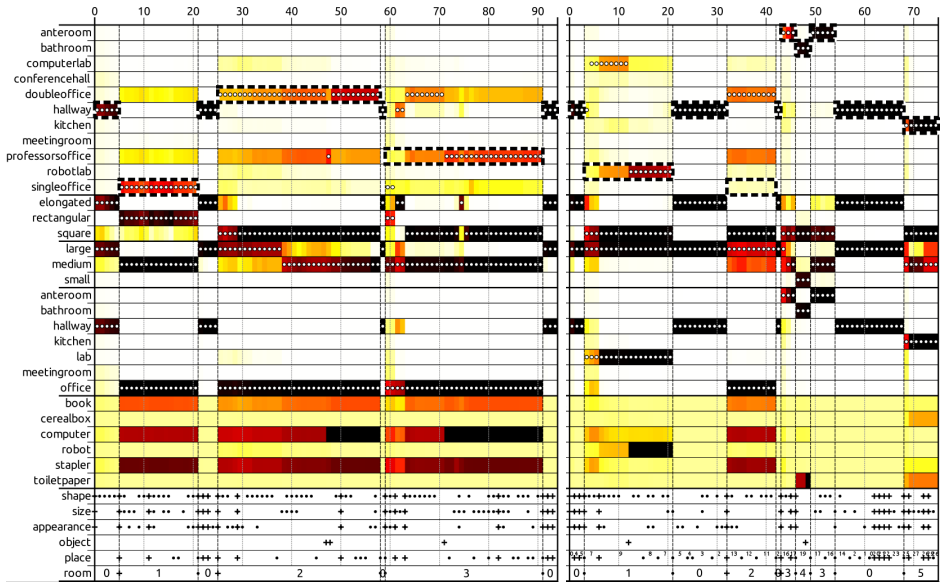
9 Conclusions and Future Works

In this report we have presented a probabilistic framework combining heterogenous, uncertain, information such as object observations, the shape, size and appearance of rooms for semantic mapping. A graphical model, more specifically a chain-graph, is used to represented the semantic information and perform the inference over it. We introduced the concept of properties between the low level sensory data and the high level concepts such as room categorizes. The properties allowed us to decouple the learning processes at the different levels and pave the way for better scalability. By making the properties understandable to humans, possibilities open in terms of spatial knowledge verbalization and interpretation of human input.

There are several ways in which the work presented in this report can be extended. We intend to look closer at how to define new categories based on a human description. For example, a person might describe a student canteen as a large room, kitchen like with very many tables. We will also look at ways to make the segmentation of space part of the estimation process as is made in PLISS [18]. We further plan to investigate more applications where the semantic information provided by our system can be utilized.

Acknowledgement

This work was supported by the SSF through its Centre for Autonomous Systems (CAS) and the EU FP7 project CogX. The help by Alper Aydemir, Moritz Göbelbecker and Kristoffer Sjöö is also gratefully acknowledged.



Part A Part B

Figure 8: Visualization of the events registered by the system during exploration and its beliefs about the categories of the rooms as well as the values of the properties. The room category ground truth is marked with thick dashed lines while the MAP value is indicated with white dots. A video showcasing the system is available at: `http://www.pronobis.pro/research/semantic-mapping`.

References

[1] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110 (3):346–359, June 2008.

[2] Li Fei-Fei and Pietro Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005.

[3] John Folkesson, Patric Jensfelt, and Henrik I. Christensen. The M-space feature representation for SLAM. *IEEE Transactions on Robotics*, 23(5):1024–1035, 2007.

[4] Cipriano Galindo, Alessandro Saffiotti, Silvia Coradeschi, Pär Buschka, Juan Antonio Fernández-Madrigal, and Javier González. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, 2005. ISBN 0-7803-8912-3.

[5] Marc Hanheide, Charles Gretton, Richard W. Dearden, Nick A. Hawes, Jeremy L. Wyatt, Andrzej Pronobis, Alper Aydemir, Moritz Göbelbecker, and Hendrik Zender. Exploiting probabilistic knowledge under uncertain sensing for efficient robot

behaviour. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, Barcelona, Spain, 2011.

[6] Nick A. Hawes, Marc Hanheide, Jack Hargreaves, Ben Page, and Hendrik Zender. Home alone: Autonomous extension and correction of spatial representations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, 2011.

[7] Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal Of The Royal Statistical Society Series B*, 64(3):321–348, 2002.

[8] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. Incremental learning for place recognition in dynamic environments. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 721–728, San Diego, CA, USA, October 2007.

[9] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research (JMLR)*, 11: 2169–2173, 2010.

[10] Oscar Martinez Mozos, Cyrill Stachniss, and Wolfram Burgard. Supervised learning of places from range data using AdaBoost. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, 2005.

[11] Oscar Martinez Mozos, Rudolph Triebel, Patric Jensfelt, Axel Rottmann, and Wolfram Burgard. Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems (RAS)*, 55(5):391–402, 2007.

[12] Andrzej Pronobis and Barbara Caputo. Confidence-based cue integration for visual place recognition. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 2394–2401, San Diego, CA, USA, October 2007.

[13] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I. Christensen. A discriminative approach to robust visual place recognition. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, October 2006.

[14] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I. Christensen. A realistic benchmark for visual indoor place recognition. *Robotics and Autonomous Systems (RAS)*, 58(1):81–96, January 2010. URL `http://linkinghub.elsevier.com/retrieve/pii/S0921889009001146`.

[15] Andrzej Pronobis, Jie Luo, and Barbara Caputo. The more you learn, the less you store: memory-controlled incremental SVM for visual place recognition. *Image and Vision Computing (IMAVIS)*, 28(7):1080–1097, July 2010. URL `http://linkinghub.elsevier.com/retrieve/pii/S0262885610000314`.

[16] Andrzej Pronobis, Oscar Martinez Mozos, Barbara Caputo, and Patric Jensfelt. Multi-modal semantic place classification. *The International Journal of Robotics Research (IJRR), Special Issue on Robotic Vision*, 29(2-3):298–320, February 2010.

[17] Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Adrian N. Bishop, and Patric Jensfelt. Representing spatial knowledge in mobile cognitive systems. In *11th International Conference on Intelligent Autonomous Systems (IAS-11)*, Ottawa, Canada,

August 2010. URL `http://www.pronobis.pro/publications/pronobis2010sprep.pdf`.

[18] Ananth Ranganathan. PLISS: Detecting and labeling places using online change-point detection. In *Proceedings of Robotics: Science and Systems (RSS'10)*, Zaragoza, Spain, June 2010.

[19] Ananth Ranganathan and Frank Dellaert. Semantic modeling of places using objects. In *Proceedings of Robotics: Science and Systems (RSS'07)*, 2007.

[20] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*, 2003.

[21] Iwan Ulrich and Illah Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation (ICRA'00)*, San Francisco, CA, USA, 2000.

[22] Christoffer Valgren and Achim J. Lilienthal. Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 1856–1861, May 2008. ISBN 978-1-4244-1646-2.

[23] Shrihari Vasudevan and Roland Siegwart. Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robotics and Autonomous Systems (RAS)*, 56:522–537, 2008.

[24] Julia Vogel and Bernt Schiele. A semantic typicality measure for natural scene categorization. *Annual Pattern Recognition Symposium (DAGM'04)*, 2004.

[25] Jianxin Wu, Henrik I. Christensen, and James M. Rehg. Visual place categorization: problem, dataset, and algorithm. In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'09)*, 2009.

[26] Jianxin Wu and James M. Rehg. Where am I: place onstance and category recognition using spatial PACT. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, June 2008. ISBN 978-1-4244-2242-5.

[27] Hendrik Zender, Oscar Martinez Mozos, Patric Jensfelt, Geert-Jan M. Kruijff, and Wolfram Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems (RAS)*, 56(6):493–502, June 2008.

[28] Zoran Zivkovic, Olaf Booij, and Ben Kröse. From images to rooms. *Robotics and Autonomous Systems (RAS), Special Issue From Sensors to Human Spatial Concepts*, 55(5):411–418, 2007.

Paper G

Exploiting Probabilistic Knowledge under Uncertain Sensing for Efficient Robot Behaviour

Marc Hanheide, Charles Gretton, Richard Dearden, Nick Hawes, Jeremy Wyatt, Andrzej Pronobis, Alper Aydemir, Moritz Göbelbecker and Hendrik Zender

Published in
International Joint Conference on Artificial Intelligence

Exploiting Probabilistic Knowledge under Uncertain Sensing for Efficient Robot Behaviour

Marc Hanheide, Charles Gretton, Richard Dearden,
Nick Hawes, Jeremy Wyatt, Andrzej Pronobis,
Alper Aydemir, Moritz Göbelbecker and Hendrik Zender

Abstract

Robots must perform tasks efficiently and reliably while acting under uncertainty. One way to achieve efficiency *is to give the robot common-sense knowledge about the structure of the world.* Reliable *robot behaviour can be achieved by modelling the uncertainty in the world probabilistically. We present a robot system that combines these two approaches and demonstrate the improvements in efficiency and reliability that result. 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 continual planning system which is able to plan in 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 behaviour.*

1 Introduction

One dream of the AI community is to build a robot capable of performing tasks on demand in dynamic real-world environments like homes and offices. Such a robot must perform task and observation planning under uncertainty in pursuit of its current goals. It must do this while exploiting knowledge about the nature of the environments in which it is expected to operate. Towards realising the stated dream, this paper presents a robot system that uses a new planning approach to reason with new representations of space. Our approach integrates probabilistic models of common-sense conceptual knowledge, with models of the visual appearance of objects and of room categories, to represent an object search task. In order to allow the robot to effectively exploit this knowledge, we have developed a novel system

Figure 1: Our object search robot in a home environment composed of rooms of different categories. An example use case of the system is to find cornflakes located in the kitchen. The mobile robot is equipped with a laser scanner and a stereo camera rig.

for continual planning that automatically switches between using decision-theoretic and classical procedures to synthesise efficient action strategies.

We have implemented our approach on the mobile robot depicted in Fig. 1, and evaluated that system by having it perform *object search* tasks in real-world home and office environments. The objects it is able to search for are all *instances* of *categories*, e.g. a specific box of cornflakes in the kitchen, as depicted in Fig. 1, is an instance of the category of cornflakes boxes, which is itself a sub-category of cereal boxes. The robot uses structured representations of knowledge at this *conceptual* level – e.g. cereal boxes are often located in kitchens or dining rooms, and sofas are often located in living rooms. Such *relational structure* expresses generalisations across multiple environments, and can be naturally represented probabilistically in order to support intelligent decision making across multiple environments. We have compiled a common-sense knowledge base in an offline manner. Our two key novel contributions are:

1. A probabilistic conceptual map that combines general purpose and contingent spatial knowledge in a single structure, together with processes for creating, maintaining, and reasoning with it. This relational structure models the uncertain contingent knowledge the robot has about instances (e.g. what category of room it thinks room 1 is) in conjunction with its – also uncertain – common-sense conceptual knowledge (e.g. which types of objects are located in a particular category of room).

2. A switching continual planner that synthesises action strategies for the very large partially observable decision processes posed by the tasks we consider. Our approach is to switch between decision-theoretic and classical modes of planning at different levels of abstraction. The classical system quickly solves a determinisation

of the problem at hand, interpreting probabilistic information in terms of a cost model. The decision-theoretic system quickly solves abstract decision problems derived using the classical plan and the probabilistic belief-state. Overall, this approach allows the system to exploit our rich representation of spatial knowledge, and generate intelligent behaviour under uncertainty in a timely manner.

2 Related Work

Probabilistic representations are employed for many localised functions in robots operating in the real world. For example, [23] use such representations in most of their system's individual components, but their robot behaviour is generated using a reactive controller rather than a domain-independent planner as here. [11] treat sensing deterministically and beliefs qualitatively during planning. We are not aware of any robot system that features both a unifying probabilistic representation, and a domain-independent planner which is able to reason quickly over that unified decision-theoretic model to generate behaviour.

Object search with mobile robots has been studied for almost 20 years, yet no previous system plans with probabilistic conceptual knowledge about both room and object categories. Instead, most dedicated systems treat the problem as a geometric one. For example, recently [20] propose how a robot can optimally locate an object in a mostly unknown 3D space. Closest to our approach is the work by [1] who used probabilistic spatial relations and static properties of rooms to pose the object search problem as a fully-observable Markov decision process (MDP). This work employed background knowledge to inform an MDP planner of good locations (e.g. room1) to search for a particular object. Earlier work by [8] proposed to make this relationship bi-directional: objects give evidence for room categories, and room categories provide information about where objects can be found. In [2] this approach was extended to treat some of the conceptual knowledge as uncertain, although sensing here is restricted to object occurrence and the planner does not use a stochastic model of sensing. [24] went beyond this to perform room categorisation using Bayesian reasoning about the presence of objects, but did not (as none of these did) include observation models for planning.

Compared to these existing approaches, we utilise a richer spatial representation combining visual room appearance, room geometry, presence of objects and the topological structure of space, extending our previous work [16] which only combined visual appearance of rooms and their geometry for the purpose of room categorisation. Also, our system is a successor to a robot that was able to exploit only deterministic conceptual and instance knowledge [9].

3 Conceptual Map

The *conceptual map* realises the highest layer of the our qualitative spatial framework [17, 16]. This framework comprises several layers of abstraction from sensor

readings, up to a topological map represented as a graph of interconnected *places* which each form a part of a *room*. On top of the topological map the framework includes the conceptual map populated by instances, of pre-defined concepts, generated by dedicated processes. An excerpt of a conceptual map for our object search task is shown in Fig. 2. The conceptual map is *relational*, describing common-sense knowledge as relations between concepts, and describing instance knowledge as relations between either instances and concepts, or instances and other instances. Relations in the conceptual map are either predefined, acquired, or inferred, and can either be deterministic or probabilistic. A non-existing relation in the conceptual map is thought of as having probability 0. An acquired relation is one that is grounded in observation and generated as a result of a sensing process. Predefined relations are given (and quantified in the case they are probabilistic) as part of a fixed ontology of default knowledge. Any inferred relations are the result of inference processes operating solely on the conceptual map.

In our implementation of the conceptual map, the concepts, and relations between these, were selected to enable our robot to reason about conceptual knowledge for efficient object search. The representation defines a taxonomy of concepts using hyponym relationships (is-a) as well as directed relations between rooms and objects (has-a). However, we also represent undirected associative relations (such as the connectivity between rooms) in our model. The processes that populate the model with instances and acquired relations between those are shown at the bottom of Fig. 2 and are as follows.

3.1 Sensing & Acting

In our system sensing is managed by a collection of processes which abstract from odometry data, laser scans and video sequences to maintain *instances* and the *probabilistic relations* which link these instances to concepts and other *instances*. We distinguish *continuous* and *active* sensing. The former is passive, continuously revising the robot's subjective beliefs about the world. It is lightweight, and does not require a planner that might schedule information gathering actions. In contrast, active sensing is deliberately planned for.

Mapping and Topology Maintenance is a continuous process that uses a SLAM algorithm [7] to maintain metric and topological maps of the environment and localise the robot in those maps. It discretises space into metrically localised *places* approximately $1m$ apart (represented by discs in Fig. 5). It also maintains a navigation graph that supports movement from one place to another. Place existence and connectivity is treated deterministically in our current system. In order that topological places be interpreted with respect to higher level spatial concepts, mapping also features door frame detection from laser data. Using the non-monotonic reasoning approach of [9] places are grouped into rooms based on these detected door frames. The results of this continuously running process are instances of places and rooms with acquired connectivity relations.

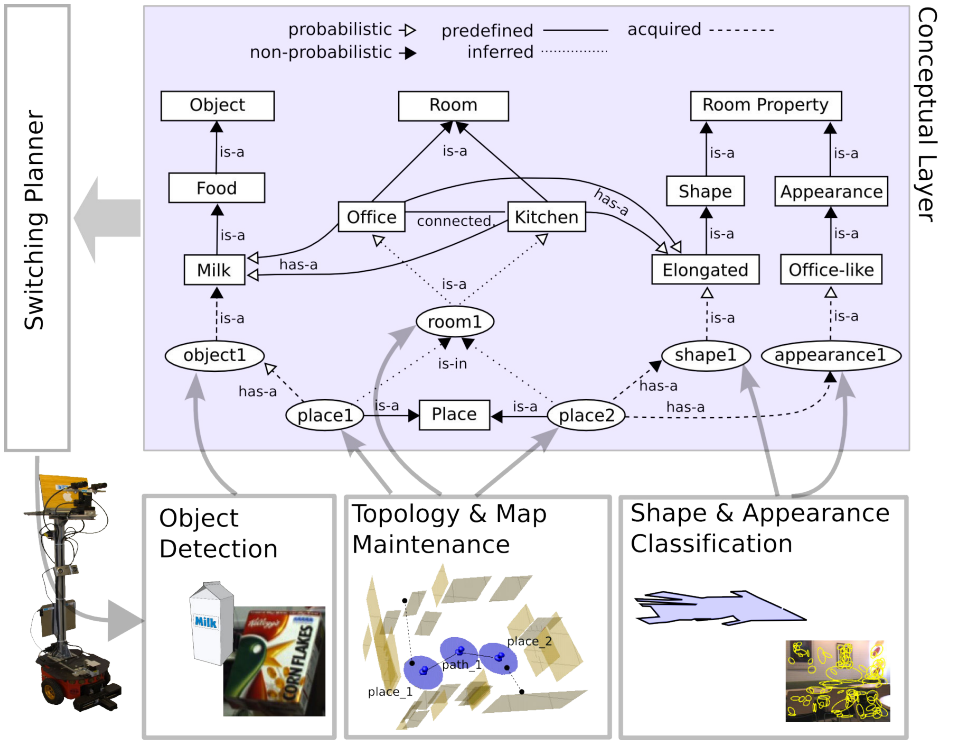


Figure 2: An abstract view of the processes and representations of the system. Sensing processes (at the bottom) discretise and categorise sensor input into instances (shown as ellipses) and acquired relations in conceptual layer. This layer also comprises knowledge about concepts (rectangles) of which only an excerpt in shown. The switching planner (cf. Sec. 4) reasons upon the state distribution given by the conceptual map.

Shape and Appearance Classification is achieved by continuous sensing of shape and appearance properties at topological places. Following [16], for a small discrete set of views at each place the robot senses low level features: (a) about the geometric shape from that view according to laser scans, and (b) about its visual appearance according to Composed Receptive Field Histograms. Those features are evaluated on the basis of Support Vector Machine (SVM) models of *specialised concepts* of "Room Property" – e.g. elongated, office-like, etc. Accumulated confidences gained from the SVM models across views are normalised to gain probabilities. These are represented in the probabilistic "is-a" relation that ties property instances to the specific concepts (cf. Fig. 2).

Object Detection is the only active sensing process in our system. It is triggered when the robot executes a visual sensing action. The underlying vision algorithms

for object detection are from the BLORT toolkit [14], and are applied to images from the robot's cameras. Object detection exhibits false positive and false negative detection rates that characterise the observation model for planning. Observation models allow the robot to reason that it might not have sensed an object despite it being perceivable, and vice-versa. The robot can then quantify the effects of active sensing processes on the conceptual map. For example, a detected object leads to the creation of a "has-object" relation for the specific instance the robot was looking for (cf. Fig. 2).

Actions in our system are all triggered by the planner. The planner typically solves two sub-problems: *navigation* and *local active visual search*. Navigation in the world is planned using the navigation graph defined by the connectivity relations. Movement between places is executed by the navigation component and includes local object avoidance. Local active visual search first requires an action to trigger the generation of discrete *viewpoints*. Following [1], the generation action is executed as a Monte-Carlo sampling of local metric maps yielding information about the probability of object presence. Viewpoints are assigned an observation probability for a set of objects. The planner then reasons using actions to move to a viewpoint and trigger goal-directed object detection for appropriate objects.

3.2 The Chain Graph Representation

The conceptual map features probabilistic relations whose probability values cannot directly be acquired through sensing processes but have to be inferred (cf. Fig. 2). In order to support Bayesian inference in the conceptual map, the relational representation is compiled into a *chain graph* [12] representation, whose structure is adapted online according to the state of underlying topological map. Chain graphs provide a natural generalisation of directed (Bayesian Networks) and undirected (Markov Random Fields) graphical models, allowing us to model both "directed" causal (such as "is-a" relations) as well as "undirected" symmetric or associative relations (such as connectivity). The use of a chain graph allows us to model circular dependencies originating from possible loops in the topological graph, as well as direct use of the probabilistic relations between the concepts. In our implementation, chain graph inference is event-driven. For example, if an appearance property, or object detection alters the probability of a relation, inference proceeds to propagate the consequences throughout the graph. In our work, the underlying inference is approximate, and uses the fast Loopy Belief Propagation [13] procedure.

An exemplary chain graph corresponding to the conceptual map shown in Fig. 2 is presented in Fig. 3. Each discrete place instance is represented by a set of random variables, one for each class of relation linked to that place. These are each connected to a random variable over the categories of rooms, representing the "is-a" relation between rooms and their categories in Fig. 2. Moreover, the room category variables are connected by undirected links to one another according to the topological map. Here, the potential functions $\phi_{rc}(\cdot,\cdot)$ describe the type knowledge

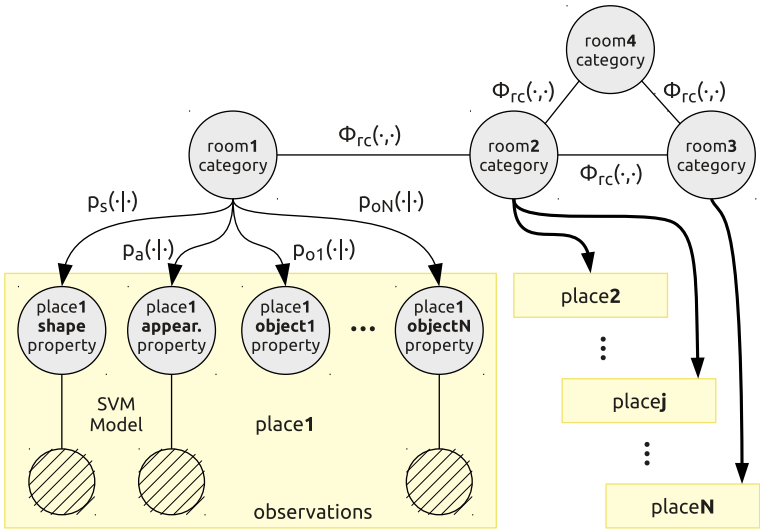


Figure 3: Structure of the chain graph model compiled from the conceptual map. The vertices represent random variables. The edges represent the directed and undirected probabilistic relationships between the random variables. The textured vertices indicate observations that correspond to sensed evidence.

about the connectivity of rooms of certain categories (e.g. that kitchens are more likely to be connected to corridors than to other kitchens).

The remaining variables represent: shape and appearance properties of space as observed from each place, and the presence of objects. These are connected to observations of features extracted directly from the sensory input. As explained in Sec. 3.1, these links are quantified by the categorical models of sensory information. Finally, the distributions $p_s(\cdot|\cdot)$, $p_a(\cdot|\cdot)$, $p_{o_i}(\cdot|\cdot)$ represent the common sense knowledge about shape, appearance, and object co-occurrence, respectively. They allow for inference about other properties and room categories e.g. that the room is likely to be a kitchen, because you are likely to have observed cornflakes in it.

3.3 Quantifying Probabilistic Relations

Our robot appeals to common-sense conceptual knowledge in order to act intelligently in indoor environments. That knowledge encapsulates, for instance, how likely it is that cereals will be found in kitchens, how hallways are usually long and offices visually cluttered, and how rooms of different types are typically connected. A question that remains is how we can quantify the probabilistic relations, such as "has-object", in Fig. 2. Our approach is to leverage common-sense knowledge available through the world wide web (WWW) to yield *object-location cooccurrence*

priors.

In our system, object and location concepts are taken from the 'locations' database provided by the *Open Mind Indoor Common Sense* (OMICS)[1] project. Compiled with the express aim of making indoor mobile robots more intelligent, that database comprises 5,800 user-given associations between common everyday objects (ca. 2,900 unique categories) and their typical locations (ca. 500 unique categories). It does not, however, quantify the likelihood of object-location pairs. Where o is a specialisation of object (e.g. "milk") and l a location (e.g. "office"), we obtain *cooccurrence frequency estimates* by counting the number of *hits* an image search engine[2] returns when resolving "o in the l" queries for each of the 1.5 million object-location pairs from OMICS. Writing $\#q(o\&l)$ for the number of hits returned by that query, and $\#q(l)$ for the number when we query the noun term l, then the *cooccurrence prior* $c(o,l)$, that o is located in l, is given by Eq. 1.

$$c(o,l) = \left(\frac{\sqrt{\#q(o\&l)}}{\sqrt{\#q(l)}} \right)^{B} \tag{1}$$

In Eq. 1, $B{=}\frac{1}{2}$ for pairs that are said to occur together in the OMICS *locations* database, and is otherwise 1. We avoid using the raw frequencies from the search engine results to mitigate the problems of: (1) occluded objects being underrepresented in image search queries – e.g., cups are stored in cupboards, and (2) image search queries are often biased to human interest, and omit the mundane and ordinary – e.g., ducks and baths are common, however faucets and baths are rarely mentioned together. We mitigate those problems by first applying the square root function to the counts. The B term has the effect of biasing the estimates to coocurrences that are deterministically represented in the OMICS database.

4 Switching Continual Planner

To generate flexible goal-oriented behaviour our system employs a domain-independent planner. This takes a starting belief-state description compiled from the probabilistic conceptual map. From a planning perspective our mobile robot domain poses important but contrary challenges. On the one hand, planning and execution monitoring must be lightweight, robust, timely, and should span the lifetime of the robot. Those processes must seamlessly accommodate exogenous events, changing objectives, and the underlying *unpredictability* of the environment. On the other hand, in order to act intelligently the robot must perform computationally expensive reasoning about *contingencies*, and possible revisions of its subjective belief according to quantitatively modelled uncertainty in acting and sensing. Addressing specifically this second challenge, [22] identify *continual planning* in the presence of detailed probabilistic models as an important direction for research.

[1] `openmind.hri-us.com`, Honda Research Institute USA

[2] `images.bing.com`

There has been much recent work scaling POMDP solution procedures to medium-sized instances. In the case of general domain-independent factored systems, the state-of-the-art scales to relatively small problems with 2^{22} states [19].[3] At their limit, these procedures take over an hour to converge. For classes of POMDP that feature exploitable structures, for example, no actions with negative effects, problems with as many as 10^{30} states can be targeted by offline procedures [4]. Moving someway towards addressing all the challenges we have outlined, recent online POMDP solution procedures have been developed which can exploit highly approximate value functions – typically computed using a point-based procedure – and heuristics in forward search [18]. These approaches are applicable in relatively small problems, and can require expensive *problem-specific* offline processing in order to yield good behaviours. A *very* recent and promising online approach for large POMDPs employs Monte-Carlo sampling to break the curse of dimensionality in situations where goal reachability is easy [21]. Although we believe it an interesting item for future work to pursue that direction, it should be noted that ease of goal reachability is not guaranteed in the problems we face.

Our work takes a concrete step towards addressing all the challenges we outlined. We have developed a *switching* domain-independent planning system that operates according to the continual planning paradigm [3]. It uses first-order declarative problem and domain representations, expressed in a novel extension of PPDDL [26] called *Decision-Theoretic (DT)PDDL*, for modelling stochastic decision problems that feature partial observability. In this paper we restrict our attention to DTPDDL models that correspond to deterministic-action goal-oriented POMDPs where all actions have non-zero cost[4] – i.e., an optimal policy can be formatted as a finite horizon contingent plan. Also, without a loss of generality we assume goals (i.e., conditions for reward) and action preconditions are conjuncts over *positive* propositions(/facts).

Our continual planning system *switches*, in the sense that the underlying planning procedure changes depending on our robot's subjective degrees of belief, and progress in plan execution. The system is continual in the usual sense that, whatever the session, plans are adapted and rebuilt online in reaction to changes to the planning model – e.g. when objectives are modified, or when our robot's path is obstructed by a door being closed. When the underlying planner is a deterministic sequential planner, i.e., a *classical* planner, we say planning is in a *sequential* session, and otherwise it is in a *DT* session. By autonomously mixing these two types of sessions our robot is able to be robust and responsive to changes in its environment, and make appropriate decisions in the face of uncertainty.

[3] Considering only room categories and distribution of objects, problems we consider in this paper have over $\sim 10^{27}$ states. The details of view points from local active visual search, and those of robot location further increase that figure. Therefore, not only because they are offline, but also because they have limited scalability these approaches are infeasible in our setting.

[4] In the case of finite-horizon planning, POMDPs with stochastic actions can be compiled into equivalent deterministic-action POMDPs, where all the original action uncertainty is expressed in the starting-state distribution [15].

During a sequential session a rewarding *trace* of a possible execution is computed, in our experiments using a modified version of *Temporal Fast Downward (TFD)* [6]. Taking the form of a classical plan, the trace specifies a sequence of actions that achieves the objectives following a deterministic approximation of the problem at hand, i.e., a *determinisation* [25]. A trace is a sequence of elements that are either: (i) actions from the DTPDDL description of the world, or (ii) atomic *assumptions*, modelled as deterministic actions, made about the truth value of facts that can only be determined at runtime – e.g., that the cereal is located in the kitchen. During sequential sessions the planner trades action costs, goal rewards, and determinacy, finding a highly valuable plan $\pi = s_0, a_0, s_1, a_1, .., s_N$ according to Equation 2.

$$V(\pi) = \prod_{i=1..N-1} \rho_i \sum_{i=1..N-1} \mathsf{R}(s_i, a_i) \tag{2}$$

Here, ρ_i is the probability that the outcome, state s_{i+1}, of the i^{th} sequenced action a_i occurs, and $\mathsf{R}(s_j, a_j)$ is the instantaneous reward received for executing action a_j in state s_j. The system always begins with a sequential session, and once TFD produces a trace, plan execution proceeds by applying actions in sequence from that trace until $\rho_i < .95$ for the next scheduled action a_i. A DT session then begins which tailors sensory processing to determine whether the assumptions made in the trace hold, or which otherwise acts to achieve the overall objectives.

Because DT planning in large problems is too slow for our purposes, DT sessions plan in an abstract decision process determined by the current trace and underlying belief-state. The abstract process posed to the DT planner is constructed by first constraining as statically false all propositions except those which are true with probability 1, or which are assumed true in the current trace. For example, if coffee cups are not necessarily in the kitchen, and the serial plan does not schedule actions whose outcomes are preconditioned on a cup being somewhere in particular (e.g., searching for or retrieving the cup from a view in the kitchen), then on first construction the states of the abstract process do not mention cups. Next, those static constraints are removed, one proposition at a time, until the number of states that can be true with non-zero probability in the initial belief of the abstract process reaches a given threshold (in our real-world experiments, 150 states). In detail, for each statically-false proposition we compute the *entropy* of the state assumptions of the current trace *conditional* on that proposition. Let X be a set of propositions and 2^X the powerset of X, then taking

$$\chi = \{ \bigwedge_{x \in X' \cap X} x \ \wedge \bigwedge_{x \in X \setminus X'} \neg x \mid X' \in 2^X \},$$

we have that χ is a set of conjunctions each of which corresponds to one truth assignment to elements in X. Where $p(\phi)$ gives the probability that a conjunction ϕ holds in the belief-state of the DTPDDL process, the entropy of X *conditional* on a proposition y, written $\mathrm{H}(X|y)$, is given by Equation 3.

(A) Partially constrained abstract belief-state

```
(:init (=(is-in Robot)kitchen)
 (.6(and(=(is-in cereal)kitchen)
        (.9(=(is-in milk) kitchen))
         .1(=(is-in milk)office))
  .4(and(=(is-in cereal)office)
        (.1(=(is-in milk)kitchen))
         .9(=(is-in milk)office)))
```

(B) Underlying DTPDDL belief

```
(:init (=(is-in Robot)office)
 (.6(and(=(is-in cereal)kitchen)
        (.9(=(is-in milk)kitchen))
         .1(=(is-in milk) office))
  .4(and(= (is-in cereal)office)
        (.1(=(is-in milk)kitchen))
         .9(=(is-in milk)office)))
 (.6(=(is-in cup)office)
  .4(=(is-in cup)kitchen)))
```

(C) Fully constrained abstract belief-state

```
(:init (=(is-in Robot)kitchen)
 (.6(=(is-in cereal)kitchen)))
```

Figure 4: Simplified examples of abstract belief-states from DT sessions.

$$\mathrm{H}(X|y) = \sum_{x \in \chi, y' \in \{y, \neg y\}} p(x \wedge y') \log_2 \frac{p(y')}{p(x \wedge y')} \tag{3}$$

A low $\mathrm{H}(X|y)$ value suggests that knowing the truth value of y is useful for determining whether or not some assumptions X hold. When removing a static constraint on propositions during the abstract process construction, y_i is considered before y_j if $\mathrm{H}(X|y_i) < \mathrm{H}(X|y_j)$. For example, if the serial plan assumes the robot is in a kitchen, then propositions about the contents of kitchens, e.g. that there is a cup in the kitchen, are added to characterise the abstract process' states. If sensing scheduled during the DT session fails to find a cup in the room, then the kitchen assumption can be judged during DT deliberations. To the abstract model we add *disconfirm* and *confirm* actions that judge each assumption in the trace. These actions yield a small reward if the corresponding judgement is true and small penalty otherwise. Once a judgement action is scheduled for execution the DT session is terminated, and a new sequential session begins.

Abstract Process Example Following the syntax and semantics of PPDDL,[5] for a simplified object search task the current belief is described by the expression in Fig. 4B. This admits 8 states with non-zero probability. For example, with probability .324 the robot is in the office, cereal and milk are in the kitchen, and a cup is in the office. Suppose a serial session plans: (1) to relocate the robot to the kitchen, (2) observe the cereal, and (3) report that cereal is located in the kitchen to a user. Characterising the abstract problem posed to the DT session at step (2), Fig. 4C gives the belief in the case where all static constraints hold. Taking

[5] Omitting the "probabilistic" string at the start of the corresponding PPDDL blocks to keep the descriptions small.

assumption X to be `(=(is-in cereal)kitchen)`, in relaxing static constraints the following entropies are calculated:

```
.47 = H(X|(=(is-in milk)office))
    = H(X|(=(is-in milk)kitchen))
.97 = H(X|(=(is-in cup)office))
    = H(X|(=(is-in cup)kitchen))
```

Therefore, the first static constraint to be relaxed is `(=(is-in milk)office)`, or equivalently `(=(is-in milk)kitchen)`, giving a refined abstract belief state depicted in Fig. 4A. Summarising, if the DT session is restricted to belief-states with fewer than 8 elements, then the starting belief-state of the DT session does not mention a "cup".

5 Experimental Evaluation

To evaluate the implemented representational and planning techniques, we first analysed our robot system performing an object search task in two different real-world environments: a larger office (O, 13 places in 3 rooms) and a smaller home (H, 7 places in 3 rooms). A sketch of the object search setting in environment H is depicted in Fig. 1. Our evaluation compares the full system described in this paper, exploiting probabilistic conceptual knowledge and evidence from shape and appearance classification, to a baseline system that cannot make use of the conceptual knowledge. We refer to these as the "full" and "lesioned" systems respectively. In the lesioned system continuous sensing of shape and appearance properties is disabled, emulating the limited reasoning available in our previous system [9]. Therefore it can neither use these properties to infer the categories of rooms nor can it exploit conceptual knowledge about object-location co-occurrence.

In all runs, a box of cornflakes (the object to search for) was placed in the environment among many other objects belonging to the nine categories the robot has been trained to detect. In the experiments, only a subset of all object-location co-occurrence frequencies consisting of 152 relations between the 19 selected object concepts (cornflakes among them) and seven given room concepts (among them kitchen, living room, corridor, and office) was employed. In a first set of runs directly comparing the full system (FC) to the lesioned case (LC), the box of cornflakes was in the kitchen, which is the canonical location for this type of object according to the common-sense conceptual knowledge (with a probability of $P(cornflakes|kitchen) = 0.336$). In a second set of runs, the object was at a non-canonical location ($P(cornflakes|living\_room) = 0.035$) to test the full configuration (results denoted as FNC).

Hypotheses The hypotheses leading to this study design are that (i) the exploitation of the conceptual knowledge in the full system will enable the robot to achieve the task quicker in canonical cases when compared to the lesioned system in the

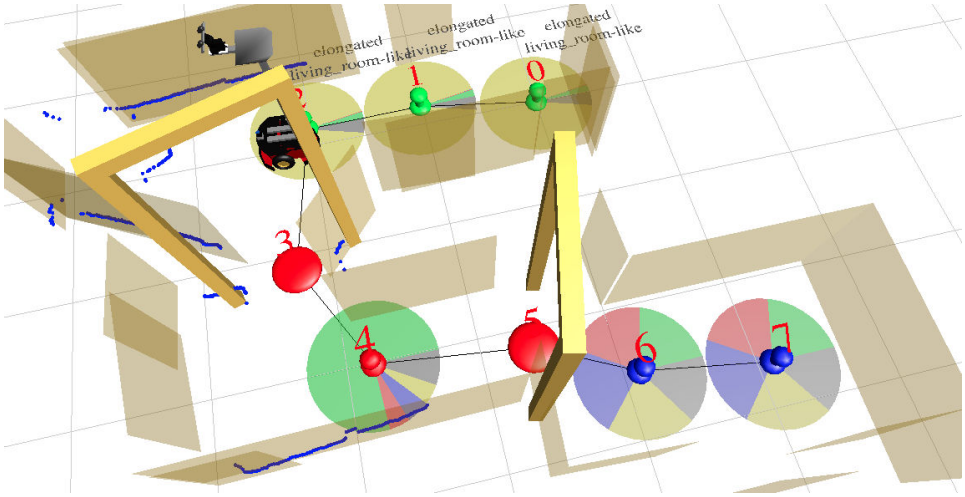
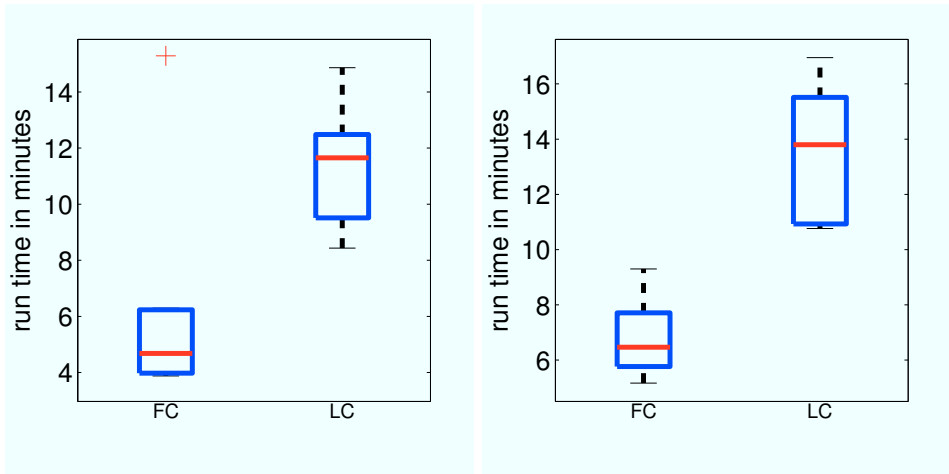


Figure 5: Environment H with numbered *places*, and pie charts indicating probabilities of room categories (yellow=living room, red=kitchen, green=corridor, blue=office, grey=others). The labels attached to place node in the living room (upper right) indicate the most likely values in the distribution of classified *shape* (top) and *visual* (bottom) properties, respectively. Detected doors, used for room partitioning, are shown as door frames. The kitchen is at the lower right.

conf.	obj. loc.	lesion	♯succ./♯tot. H	♯succ./♯tot. O	avg. time H	avg. time O
FC	kitchen	no	10/10	5/6	5.8min	6.8min
LC	kitchen	yes	9/10	5/5	11.3min	13.5min
FNC	liv. room, office	no	3/3	n/a	10.2min	n/a

Table 1: Runtimes for the three cases tested: full system (FC) and lesioned system (LC), both with object in canonical position; NFC: full system with object in non-canonical position. Total time to solve the task reported in minutes. The FNC case was only tested in environment H.

same experimental setup, (ii) although more efficient in the average case, the system will be robust in the presence of sensing errors, and (iii) that the system will still be able to achieve its goal, even relatively efficiently, in non-canonical setups. In all runs, before the robot was given the goal to find the object, it performed a short exploration of adjacent places to sense room properties in order to infer the category of the room (if this evidence was not lesioned). The acquired map after this short exploration for the H environment is shown in Fig. 5. The robot has already sensed room properties and hence evidence about the category of the room it is in is available in the conceptual map when making a first plan.



(a) env. H. (b) env. O.

Figure 6: Box and whisker diagrams of total runtime to achieve the given task in two environments comparing the 'full (FC)' to the 'lesioned (LC)' case. In the FC case for environment H one run took longer than 14 minutes indicated by the outlier.

Quantitative Results from Real-World Experiments The cornflakes box was found by the robot in 32 of the 34 runs. In the two failed runs the robot maneuvered itself into a corner of the room and required human intervention. The total execution time of the successful FC and LC runs are plotted in Fig. 6.[6] Hypothesis (i) claims that the robot is able to exploit the evidence gained from perceiving its environment by integrating this with conceptual knowledge about the commonalities of such environments. That our system does this is confirmed by a significant difference (Mann-Whitney test $p < 0.01$ for both environments) in average runtime reported in Tab. 1 for these two configurations. Looking at the typical sequence of actions for the FC configuration it becomes apparent that planning inferred lower costs for driving into the kitchen to begin searching (despite that being an extra distance to travel without looking for objects). This observation also explains the relative improvement of FC in the larger O environment, comprising larger space that has to be searched exhaustively, being comparatively higher than in H.

In all our runs, total running time was dominated by action execution with planning being a distant second. The total time spent on planning was between 16.9 seconds for the FC runs in the H environment and 64.8 seconds for the LC

[6] A video available at `http://youtu.be/0QcmSDZR-c4` illustrates an exemplary FC run in the H environment.

configuration in the larger O environment. The time was divided roughly equally between the serial and DT sessions. The time spent on planning *per planner call* was slightly higher in the FC configurations, but this was offset by a much lower number of calls in the full configuration (on average 6 compared to 13 for H, 5.6 to 20.6 for O).

Qualitative Discussion of Results The results comparing FC and LC runs could admittedly have been obtained using a system that *determinsitically* exploits structural knowledge about cornflakes being found in kitchens instead of making use of the probabilistic formulation of the knowledge. Hence, the results so far only confirm hypothesis (i). With regard to hypothesis (iii) we can confirm that the robot was able to solve non-canonical configurations 100% of the time. In these runs, the robot also first searched in the kitchen before returning to the other room and eventually finding the object there. A system entirely dogmatic about the cornflakes being in kitchens (having modeled this relation as deterministic) would not have been able to consider this alternative. Another interesting case we encountered is evident in Fig. 6(a). The single outlier in this figure is related to hypothesis (ii), indicating that that our system can cope with uncertainty in sensing. In this case the robot also first drove to the kitchen (following its initial sequential plan), entered a DT session to find the object, but eventually failed to detect the object (due to an object detection false negative). Hence, the DT session disconfirmed the original assumption that was made. Accordingly a plan was created that drove the robot back to the living room to continue its search there; again due to the remaining probability of finding objects also in non-canonical locations. However, after a comprehensive search, the likelihood of finding the object in the kitchen by looking again became higher, so the robot went back and eventually found it. In general, realistically, object detection was very reliable in our system, with observation probabilities estimated as .05 for false positives and false negatives. Accordingly, we only observed one such case of a sensing error in our runs. In order to assess the potential of the switching planner in greater detail we conducted further experiments in a simulated setting, where we were in control of sensing characteristics.

Planning with Noisy Sensing We integrated our switching planner in an enhanced version of the MAPSIM simulation environment [3] and performed a detailed experimental evaluation, comparing switching using our DT procedure with a greedy dual-mode [5] *baseline* we called "cp". For experimentation in simulation the base planner during sequential sessions is a cost-optimising version of *Fast Downward* [10]. For the dual-mode baseline, when a scheduled action triggers a switch to a DT session, the system plans to a single entropy reduction action whose execution can provide evidence regarding the truth value of an assumption from the current trace. Control is returned to a new sequential session as soon as a sensing action is executed.

We evaluate those approaches in simulation on a number of tasks, including a 6-room environment comprising 26-places and 21-objects. In simulation we considered

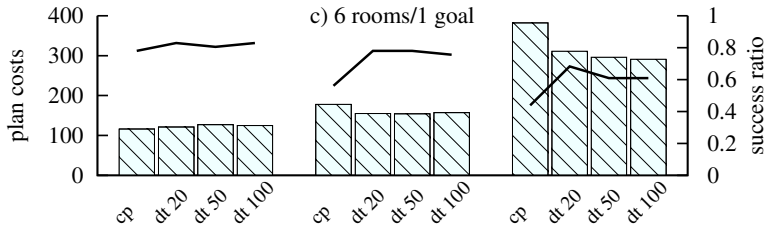


Figure 7: Average run cost (bars) and portion of successful runs (line) in simulation. For the left plot sensing is *reliable*, middle is *semi-reliable*, and right is *noisy*.

3 levels of reliability in sensing: *reliable* sensors have a .1 probability of a false negative, *semi-reliable* have a chance of 0.3 of false negative and 0.1 of false positive, and *noisy* sensors with probabilities of 0.5 and 0.2 respectively. Our evaluation examines DT sessions with initial belief-states admitting between 20 (written "dt 20") and 100 abstract states with non-zero probability. We run 50 simulations in each configuration, and plot quality data in Fig. 7 for observed behaviours in the 6-room environment where the goal is to find and report the location of a target object to a user. Here, compared to the simple greedy sensing strategy of *cp*, we have that DT sessions yield a higher rate of success, and lower expected cost of achieving the goal. Moreover, as the reliability of sensing degrades there is a clear benefit in performing expensive DT planning. Although there is insufficient space to present the results here, we observed that the overall time spent planning increases linearly as we move from *cp* to progressively refined abstractions in DT sessions.

6 Conclusion

We presented a mobile robot system that integrates two original approaches for representing and reasoning about uncertainty. The first is a conceptual map, a representation of space that combines knowledge about its qualitative structure (e.g. a topological map), with probabilistic knowledge (e.g. that cereal boxes are found in kitchens 33% of the time). The second is a continual planning and execution monitoring system that employs *switching* to plan for *very* large partially observable problems that are posed by this representation. It is important to note that the *integration* of these approaches is crucial to the success of our work. Without the novel planner, the representation would not be capable of influencing behaviour. Without the novel representation, the planner would not be able to reason over both probabilistic instance *and* conceptual knowledge at the same time. We evaluated this combination in our robot in two real-world environments, and found that it is able to yield efficient and robust behaviours in an object search task.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement No. 215181, CogX.

References

[1] Alper Aydemir, Kristoffer Sjöö, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, 2011.

[2] Abdelbaki Bouguerra, Lars Karlsson, and Alessandro Saffiotti. Handling uncertainty in semantic-knowledge based execution monitoring. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 437–443, San Diego, CA, USA, October 2007.

[3] Michael Brenner and Bernhard Nebel. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems*, 19(3):297–331, 2009.

[4] Emma Brunskill and Stuart Russell. RAPID: A reachable anytime planner for imprecisely-sensed domains. In *Uncertainty in Artificial Intelligence (UAI)*, 2010.

[5] Anthony R. Cassandra, Leslie Pack Kaelbling, and James A. Kurien. Acting under uncertainty: Discrete Bayesian models for mobile-robot navigation. In *Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'96)*, 1996.

[6] Patrick Eyerich, Robert Mattmüller, and Gabriele Röger. Using the context-enhanced additive heuristic for temporal and numeric planning. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, Thessaloniki, Greece, 2009.

[7] John Folkesson, Patric Jensfelt, and Henrik I. Christensen. The M-space feature representation for SLAM. *IEEE Transactions on Robotics*, 23(5):1024–1035, 2007.

[8] Cipriano Galindo, Alessandro Saffiotti, Silvia Coradeschi, Pär Buschka, Juan Antonio Fernández-Madrigal, and Javier González. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, 2005. ISBN 0-7803-8912-3.

[9] Nick A. Hawes, Marc Hanheide, Jack Hargreaves, Ben Page, and Hendrik Zender. Home alone: Autonomous extension and correction of spatial representations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, 2011.

[10] Malte Helmert. The fast downward planning system. *Journal of Artificial Intelligence Research*, 26:191–246, 2006.

[11] Dirk Kraft, Emre Baseski, Mila Popović, Anna Batog, Anders Kjæ r Nielsen, Norbert Krüger, Ronald P. A. Petrick, Christopher Geib, Nicolas Pugeault, Mark Steedman,

Tamim Asfour, Rüdiger Dillmann, Sinan Kalkan, Florentin Wörgötter, Bernhard Hommel, Renaud Detry, and Justus Piater. Exploration and planning in a three-level cognitive architecture. In *Proceedings of the International Conference on Cognitive Systems (CogSys'08)*, 2008.

[12] Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal Of The Royal Statistical Society Series B*, 64(3):321–348, 2002.

[13] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research (JMLR)*, 11: 2169–2173, 2010.

[14] Thomas Mörwald, Johann Prankl, Andreas Richtsfeld, Michael Zillich, and Markus Vincze. BLORT - The blocks world robotic vision toolbox. In *Proceedings of the ICRA Workshop Best Practice in 3D Perception and Modeling for Mobile Manipulation*, 2010.

[15] Andrew Y. Ng and Michael I. Jordan. PEGASUS: A policy search method for large MDPs and POMDPs. In *Uncertainty in Artificial Intelligence (UAI)*, pages 406–415. Morgan Kaufmann Publishers Inc., 2000.

[16] Andrzej Pronobis, Oscar Martinez Mozos, Barbara Caputo, and Patric Jensfelt. Multi-modal semantic place classification. *The International Journal of Robotics Research (IJRR), Special Issue on Robotic Vision*, 29(2-3):298–320, February 2010.

[17] Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Adrian N. Bishop, and Patric Jensfelt. Representing spatial knowledge in mobile cognitive systems. In *11th International Conference on Intelligent Autonomous Systems (IAS-11)*, Ottawa, Canada, August 2010. URL `http://www.pronobis.pro/publications/pronobis2010sprep.pdf`.

[18] Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-draa. Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, 32(1): 663–704, 2008.

[19] Guy Shani, Ronen I. Brafman, Shimony E. Shimony, and Pascal Poupart. Efficient ADD operations for point-based algorithms. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS'08)*, 2008.

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

[21] David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[22] Kartik Talamadupula, J. Benton, Subbarao Kambhampati, Paul Schermerhorn, and Matthias Scheutz. Planning for human-robot teaming in open worlds. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1(2), 2010.

[23] Sebastian Thrun, Michael Beetz, M Bennewitz, Wolfram Burgard, A B Cremers, Frank Dellaert, Dieter Fox, D Hähnel, C Rosenberg, Nicholas Roy, J Schulte, and D Schulz. Probabilistic algorithms and the interactive museum tour-guide robot Minerva. *The International Journal of Robotics Research (IJRR)*, 19(11):972–999, 2000.

[24] Shrihari Vasudevan and Roland Siegwart. Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robotics and Autonomous Systems (RAS)*, 56:522–537, 2008.

[25] Sungwook Yoon, Alan Fern, and Robert Givan. FF-Replan: A baseline for probabilistic planning. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS'07)*, Providence, Rhode Island, USA, 2007.

[26] Håkan L. S. Younes, Michael L. Littman, David Weissman, and John Asmuth. The first probabilistic track of the international planning competition. *Journal of Artificial Intelligence Research*, 24:851–887, 2005.