KTH Computer Science and Communication

School of Computer Science and Communication
CVAP - Computational Vision and Active Perception

Understanding the Real World: Combining Objects, Appearance, Geometry and Topology for Semantic Mapping

Andrzej Pronobis and Patric Jensfelt

TRITA-CSC-CV 2011:1 CVAP319

Andrzej Pronobis and Patric Jensfelt
Understanding the Real World: Combining Objects, Appearance, Geometry and Topology for Semantic Mapping

Report number: TRITA-CSC-CV 2011:1 CVAP319
Publication date: May, 2011
E-mail of author(s): [pronobis,patric]@kth.se

Reports can be ordered from:

School of Computer Science and Communication (CSC)
Royal Institute of Technology (KTH)
SE-100 44 Stockholm
SWEDEN

telefax: +46 8 790 09 30
http://www.csc.kth.se/

Understanding the Real World: Combining Objects, Appearance, Geometry and Topology for Semantic Mapping

Andrzej Pronobis and Patric Jensfelt
Centre for Autonomous Systems
Computational Vision and Active Perception Lab
School of Computer Science and Communication
KTH, Stockholm, Sweden
pronobis@kth.se

May 10, 2011

Contents

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 because 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.

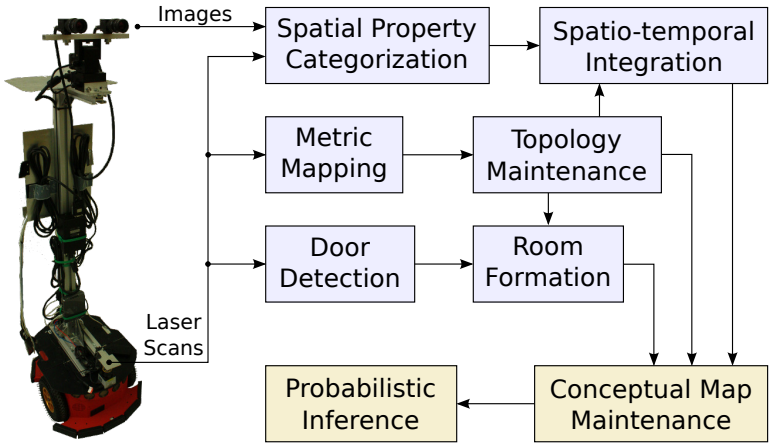


Figure 1: Dora the Explorer as well as the elements and the data flow inside the semantic mapping system.

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 [1] where we presented a system for multi-clue integration of laser and vision data for place categorization. A fundamental difference from that work is that we now have decoupled 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 [2] 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 [3] 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. That is, if two people are asked to sort the images into different categories they are likely to come up with different partitions. [3] 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 [4] another method is presented for categorization of outdoors scenes based on representing the distribution of codewords in each scene category. In [5] a new image descriptor, PACT, is presented and shown to give superior results on the datasets used in [2, 4].

In robotics, one of the early systems for place recognition is [6] where color histograms is used to model the appearance of places in a topological map and place recognition performed as a part of the localization process. Later [7] 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 [8] 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 [9] a vision-only place recognition system is presented. Super Vector Machines (SVMs) are used as classifiers. The characteristics are similar to those of [7]; cannot identify and learn new categorizes on-line, only works with data from a single source and classification was done frame by frame. In [10, 11] 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 [13] the works in [7, 9] 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 [14] the work from [5] 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 [15] 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 fil-

ter 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 [16] 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 [17] 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 [8] makes use of object observations to perform the place categorization. In [18] 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 [20] a Bayesian framework for connecting objects to place categories is presented. In [21] the work in [8] 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 [22]. 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 of 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 discretisation 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 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 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.

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 [1]. Those models are trained on features extracted directly from the robot's sensory input. Following [1], 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 [23]. 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 [1]). 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 [1].

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 multi-modal 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

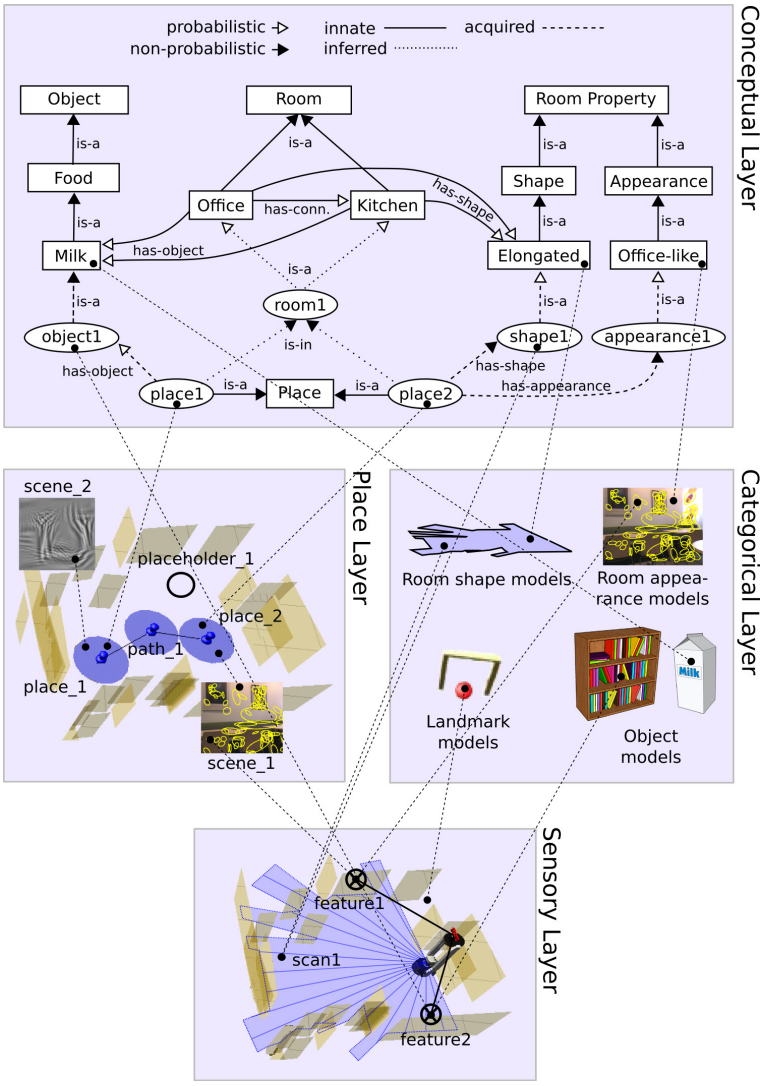


Figure 2: The layered structure of the spatial representation. The position of each layer within the representation corresponds to the level of abstraction of the spatial knowledge. The conceptual layer illustrates part of the ontology representing both instance and predefined world knowledge.

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 associated 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 relationships). As in [21], 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 [24]) 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 [25]. Chain graphs are a natural generalization of directed (Bayesian Networks) and undirected

[1] `http://openmind.hri-us.com/`

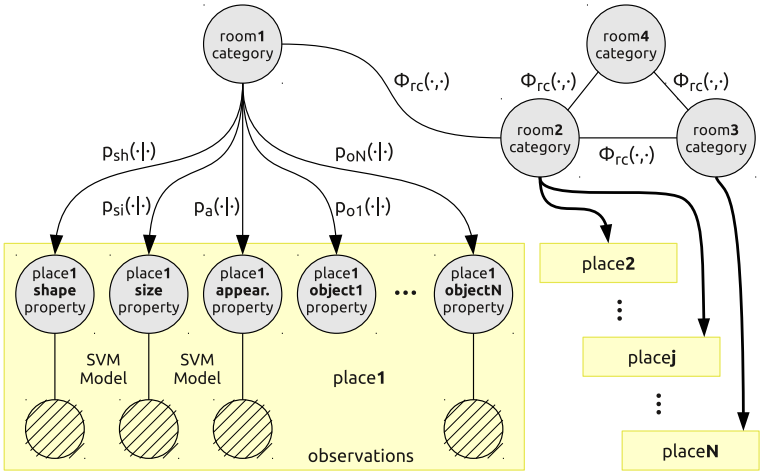


Figure 3: Structure of the chain graph model compiled from the conceptual map. The vertices represent random variables. The edges represent the directed and undirected probabilistic relationships between the random variables. The textured vertices indicate observations that correspond to sensed evidence.

(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 [25]:

$$f(x) = \prod_{\tau \in T} f(x_\tau | x_{pa(\tau)}),$$

where $pa(\tau)$ denotes the set of parents of vertices 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 [26], to comply with time constraints imposed by the robotic applications.

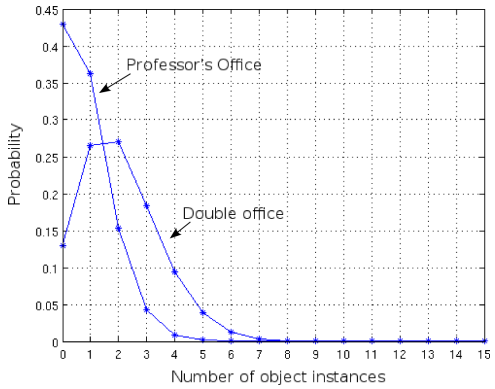


Figure 4: The Poisson distributions modelling the existence of a certain number of objects in a room on the example of computers present in a double office and a professor's office.

The 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. 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 [27] 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 [28].

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 [1] 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

[2] `http://www.cas.kth.se/COLD`

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 platform 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].

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.

[3] `http://www.robotvision.info`

[4] `http://www.cas.kth.se/cold-stockholm`

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.

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.

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 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

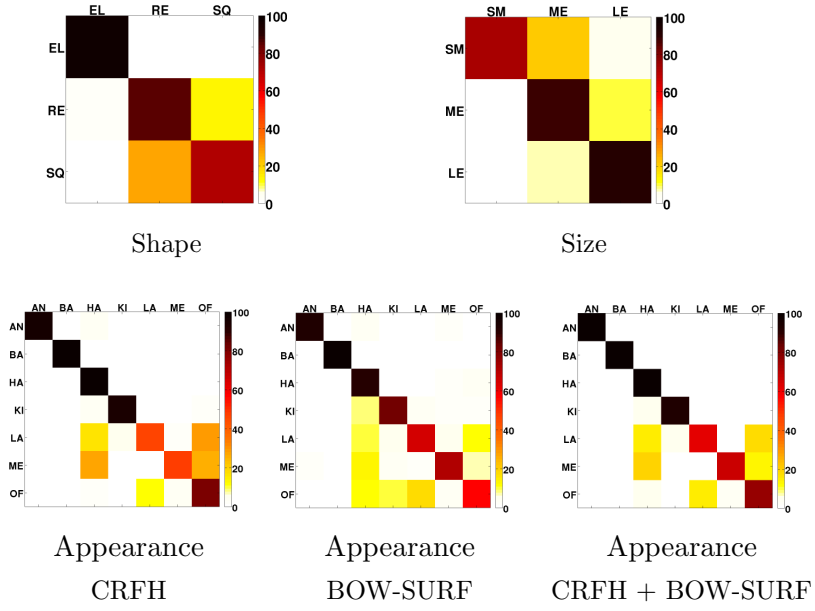


Figure 6: Confusion matrices for the offline experiments with sensory categorical models of each of the properties.

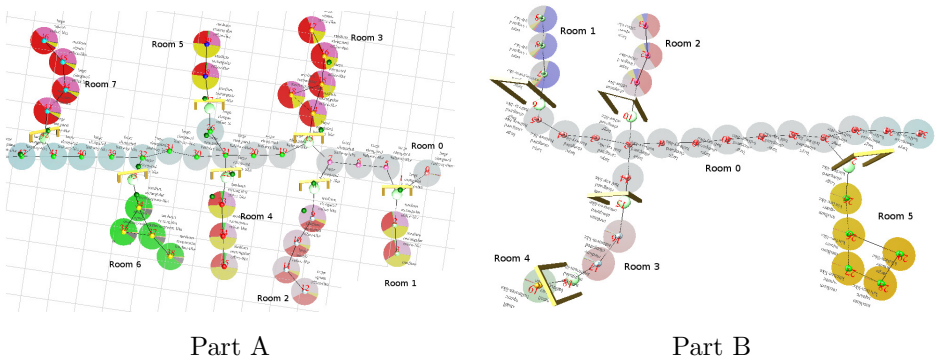


Figure 7: Topological maps of the environment anchored to a metric map indicating the outcomes of room segmentation and categorization (best viewed in color). The pie charts indicate the location of places in the environment and the probability distribution over the inferred room categories (each color corresponds to a room category). For the detailed information about the inferred categories, see Fig. 8.

source of that event is 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.

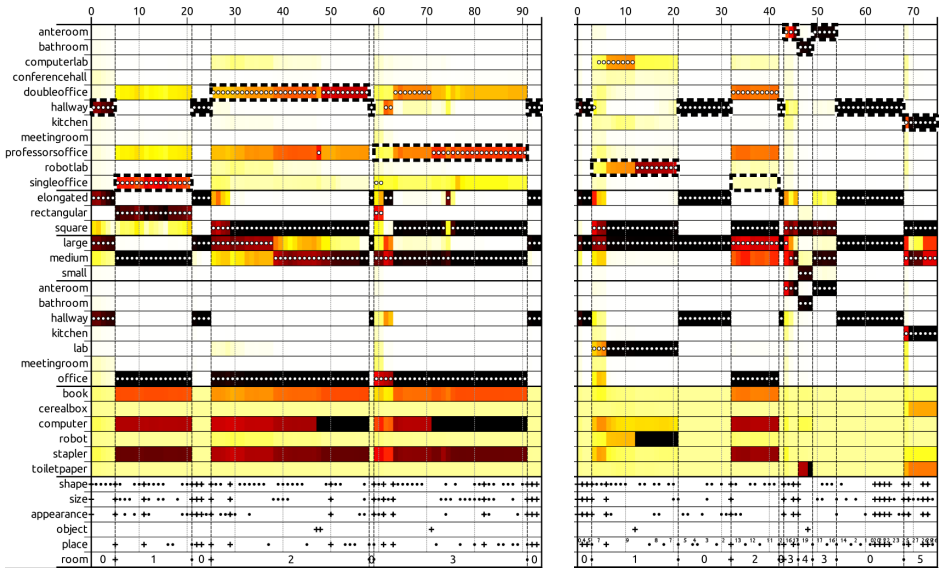
Looking at part B, we see that the system identified most of the room categories correctly with the exception of a single office which due to a misclassification of size was incorrectly recognized as a double office. The experiment proved that the system can deliver an almost perfect performance by integrating 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 [15]. We further plan to investigate more applications where the semantic information provided by our system can be utilized.



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`.

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.

References

[1] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *IJRR*, vol. 29, no. 2-3, Feb. 2010.

[2] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV'03)*, 2003, pp. 273–280.

[3] J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," *Pattern Recognition*, pp. 195–203, 2004.

[4] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[5] J. Wu and J. M. Rehg, "Where am i: Place instance and category recognition using spatial pact," in *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, June 2008.

[6] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA'00)*, vol. 2, Apr. 2000, pp. 1023–1029.

[7] O. Martínez Mozos, C. Stachniss, and W. Burgard, "Supervised learning of places from range data using adaboost," in *ICRA'05*, 2005.

[8] O. M. Mozos, R. Triebel, P. Jensfelt, A. Rottmann, and W. Burgard, "Supervised semantic labeling of places using information extracted from laser and vision sensor data," *Robotics and Autonomous Systems Journal*, vol. 55, no. 5, pp. 391–402, May 2007.

[9] A. Pronobis, B. Caputo, P. Jensfelt, and H. Christensen, "A discriminative approach to robust visual place recognition," in *IROS'06*, 2006.

[10] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "Incremental learning for place recognition in dynamic environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS07)*, San Diego, CA, USA, October 2007.

[11] A. Pronobis, J. Luo, and B. Caputo, "The more you learn, the less you store: Memory-controlled incremental SVM for visual place recognition," *Image and Vision Computing (IMAVIS)*, Mar. 2010.

[12] A. Pronobis and B. Caputo, "Confidence-based cue integration for visual place recognition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS07)*, San Diego, CA, USA, October 2007.

[13] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *IJRR*, vol. 29, no. 2-3, Feb. 2010.

[14] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual place categorization: Problem, dataset, and algorithm," in *In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS'09)*, 2009.

[15] A. Ranganathan, "Pliss: Detecting and labeling places using online change-point detection," in *RSS*, 2010.

[16] Z. Zivkovic, O. Booij, and B. Kröse, "From images to rooms," *Robotics and Autonomous Systems, special issue From Sensors to Human Spatial Concepts*, vol. 55, no. 5, pp. 411–418, May 2007.

[17] C. Valgren and A. Lilienthal, "Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments," in *ICRA 2008*. IEEE, 2008, pp. 1856–1861.

[18] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. A. Fernandez-Madrigal, and J. Gonzalez, "Multi-hierarchical semantic maps for mobile robotics," in *IROS*, August 2005.

[19] A. Ranganathan and F. Dellaert, "Semantic modeling of places using objects," in *RSS*, 2007.

[20] S. Vasudevan and R. Siegwart, "Bayesian space conceptualization and place classification for semantic maps in mobile robotics," *Robot. Auton. Syst.*, vol. 56, pp. 522–537, June 2008.

[21] H. Zender, O. M. Mozos, P. Jensfelt, G.-J. M. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 493–502, June 2008.

[22] A. Pronobis, K. Sjöö, A. N. Aydemir, Alper an d Bishop, and P. Jensfelt, "Representing spatial knowledge in mobile cognitive systems," in *IAS-11*, Aug. 2010.

[23] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. of ECCV'06*, 2006.

[24] M. Hanheide, N. Hawes, C. Gretton, A. Aydemir, H. Zender, A. Pronobis, J. Wyatt, and M. Göbelbecker, "Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour," in *IJCAI'11*, 2011.

[25] S. L. Lauritzen and T. S. Richardson, "Chain graph models and their causal interpretations," *J. Roy. Statistical Society, Series B*, vol. 64, no. 3, pp. 321–348, 2002.

[26] J. M. Mooij, "libDAI: A free and open source C++ library for discrete approximate inference in graphical models," *J. Mach. Learn. Res.*, vol. 11, pp. 2169–2173, Aug. 2010.

[27] J. Folkesson, P. Jensfelt, and H. I. Christensen, "The m-space feature representation for SLAM," *IEEE Trans. Robotics*, vol. 23, no. 5, pp. 1024–1035, Oct. 2007.

[28] N. Hawes, M. Hanheide, J. Hargreaves, B. Page, and H. Zender, "Home alone: Autonomous extension and correction of spatial representations," in *ICRA*, 2011, to appear.