Room Categorization Based on a Hierarchical Representation of Space

Regular Paper

Peter Uršič¹*, Domen Tabernik¹, Marko Boben¹, Danijel Skočaj¹, Aleš Leonardis¹ and Matej Kristan¹

¹ Vicos Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia
* Corresponding author E-mail: peter.ursic@fri.uni-lj.si

Received 31 May 2012; Accepted 17 Dec 2012

© 2013 Uršič et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract For successful operation in real-world environments, a mobile robot requires an effective spatial model. The model should be compact, should possess large expressive power and should scale well with respect to the number of modelled categories. In this paper we propose a new compositional hierarchical representation of space that is based on learning statistically significant observations, in terms of the frequency of occurrence of various shapes in the environment. We have focused on a two-dimensional space, since many robots perceive their surroundings in two dimensions with the use of a laser range finder or sonar. We also propose a new low-level image descriptor, by which we demonstrate the performance of our representation in the context of a room categorization problem. Using only the lower layers of the hierarchy, we obtain state-of-the-art categorization results in two different experimental scenarios. We also present a large, freely available, dataset, which is intended for room categorization experiments based on data obtained with a laser range finder.

Keywords Representation of Space, Compositional Hierarchies, Room Categorization, Range Data

1. Introduction

The development of cognitive systems is becoming an important area of robotics [1-3]. In the future, mobile robots will be present in our homes and workplaces, where they will be used to perform various tasks and where great flexibility and efficiency will be expected from them. One of the most basic cognitive capabilities of a mobile robot is its spatial competence, since many of its other capabilities depend on it. For example, if someone would like the robot to bring him a cup of tea, very accurate object recognition is of no great use if the robot has difficulties finding the kitchen. A mobile robot is expected to efficiently execute tasks including localization and mapping, exploration, navigation, room categorization etc. A robot’s performance at such tasks depends on many factors, but large influence can be ascribed to the chosen representation of space. Choosing a representation is thus of central importance in sensor data interpretation. The type of representation we use determines what information is made explicit in the model, the purposes for which a model can be used and the efficiency with which those purposes can be accomplished.
Designing an efficient spatial model is a challenging problem. The model should be as compact as possible, while simultaneously it needs to be able to efficiently represent a huge variety of environments. A number of different approaches to modelling space already exist [2-13]. In most of them, robots perceive their surroundings using range sensors, which is also the case in our approach. Our long-term goal is to address the scalability issue of existing spatial representations. A major drawback of current state-of-the-art robotic systems is that they usually do not make use of the vast quantity of information that has been obtained in previous observations, when they are exploring new, never before seen, environments. Consider an example based on human perception of space. Every one of us has already seen a large number of, let us say, bathrooms and therefore, we have formed some general idea about what a typical bathroom looks like. If we one day find ourselves in a friend's, never before seen, bathroom we do not act surprised. Quite the opposite, we find it easy to orientate ourselves in it. Current state-of-the-art cognitive systems do not operate in the same way. When the robot finds itself in a new environment it usually scans the entire place, before it can achieve a good performance [12]. The size of the generated model grows linearly with respect to the number of memorized spaces, because the system assumes that the new environment has nothing in common with previously observed ones. In our work we want to provide some prior knowledge about the general characteristics of space, learned from previous observations, which would enable a service robot to explore a new environment, even without recording the whole place.

In recent years, hierarchical compositional models have been shown to possess many appealing properties, which have the potential to meet our goals. They are used in the computer vision community for object category detection. A central point of these models is that their lowest layer is composed of elementary parts, which are combined to produce more complex parts in the next layer. This procedure may be recursively repeated over several layers, which gradually increases the complexity of the vocabulary of parts. An appealing aspect of compositional hierarchies is that, on the one hand, they offer sharing of object parts within each object category, while on the other hand, they can also reuse the parts at multiple levels of granularity among different categories [14]. In fact, in recent work Fidler et al. [14] have shown that a hierarchical compositional model allows incremental training and significant sharing of parts among many categories. Sharing reduces storage requirements and at the same time makes inference efficient, since hypotheses of the shared parts are verified simultaneously for multiple categories.

In this paper we adapt the hierarchical model from [14] to develop a description suitable for representation of space. The representation is two-dimensional and is based on data obtained with a laser range finder. The algorithm that is used to learn the hierarchical model is an extended version of the Learning the Hierarchy of Parts (IHOP) algorithm [14]. Our extended version is called the Spatial Hierarchy of Parts (sHoP) algorithm. To the best of our knowledge, this is the first attempt at using a hierarchical compositional model such as [14] for the representation of space on the lowest semantic level. To demonstrate the suitability of our representation, we perform a series of experiments in the context of room categorization problems based on laser scans only. Our new low-level descriptor, called the Histogram of Compositions (HoC), inspired by the work of [15], is used for this purpose. The elements of the hierarchy are used as the building blocks of the HoC descriptor, which is then used as an input for categorization with a Support Vector Machine (SVM). We use the descriptor not only to perform the categorization, but also to verify the effectiveness of our spatial model. Moreover, we consider two different experimental scenarios for room categorization in this paper. In the first scenario, which is called exploratory room categorization (exploratory RC), the categorization is performed based on a set of laser scans obtained in each room. In the second scenario, which is called single-shot room categorization (single-shot RC), the categorization is performed based only on a single scan. The experiments performed on demanding datasets show that our method delivers state-of-the-art results. Using our mobile robot, we have also obtained a large dataset, called the Domestic Rooms (DR) Dataset, which we are making publicly available. The motivation behind its creation was the need for a comprehensive set of data, corresponding to domestic environments that could be used to benchmark room categorization algorithms based on data obtained with a laser range finder. A large number of rooms that are present in our dataset also represent a great value to any algorithm that is designed to learn from observations, since the obtained knowledge possesses greater statistical significance if the data used for learning covers a wider area of population.

The contributions of our paper are three-fold. First, we present a new hierarchical representation of space, along with the sHoP algorithm that is used to learn the hierarchy. Second, we propose a new low-level descriptor, HoC, which can be used for room categorization. As our third contribution, we present a freely available DR Dataset, which provides a large scale-testing environment for room categorization approaches based on data obtained with a laser range finder.

The remainder of the paper is organized as follows: Related work is discussed in Section 2, Section 3 provides an overview of the IHOP algorithm and our extended version, in Section 4 the HoC descriptor is introduced and our two scenarios for room categorization are discussed.
We present the DR Dataset in Section 5 and report results of several room categorization experiments in Section 6. We conclude the paper in Section 7 and discuss future research possibilities.

2. Related Work

Several spatial models have been proposed. Metric representations [4, 5] use sensory information to accurately describe the geometry of space to some extent, topological representations [6, 7, 13] use graphs to model space, while hybrid approaches [8, 9] combine both of the above paradigms. Combining one or even both of the approaches, metric and topological, on multiple levels of abstraction results in hierarchical representations [2, 10-12]. Despite several existing approaches to modelling space, to the best of our knowledge, our work is the first attempt at using a hierarchical compositional model for the representation of space on the lowest semantic level, at which range sensors are usually used to observe the environment.

Various systems performing topological localization have been developed for room categorization. In [12] very accurate room categorization is achieved using multimodal information. Approaches using less information available for categorization have also been considered. Laser range data combined with vision was used for categorization in [16] and many approaches that use vision only for the accomplishment of this task have also been presented [13, 17]. Those most related to our work are the approaches which perform room categorization based only on data obtained with range sensors. In [18] a 3D Time-of-Flight infrared sensor was used for acquiring 3D information, which allowed the distinction between three types of rooms (office, meeting room and hall). Only range data was used in [19]. Their robot was equipped with a 360-degree field of view range sensor and they were able to distinguish between four categories (rooms, corridors, doorways and hallways). The categorization was performed with AdaBoost and it was based only on a single scan. Laser range data was also used for categorization in [20], where Voronoi random fields (VRFs) were employed to label different environments in the place, providing the distinction between four categories (rooms, hallways, junctions and doorways). Their approach uses a state-of-the-art SLAM technique to generate a metric occupancy grid map of an environment, while the Voronoi graph is then extracted from this map. For each point on the Voronoi graph, VRFs then estimate the type of place it belongs to. In our work we consider two different experimental scenarios for room categorization. The exploratory RC scenario is more challenging than the problems presented in [19, 20]. Our range sensor’s field of view is only 240 degrees. Although several scans are obtained in every room, we do not require any information about how these scans are correlated to one another. We have taken in consideration a very challenging set of room types (living room, corridor, bathroom and bedroom), which are available in our DR Dataset. We demonstrate the effectiveness of our spatial model using this scenario. We compare our approach with state-of-the-art on challenging datasets, on all of which good categorization results are obtained. The first comparison is made on our DR Dataset in the context of an exploratory RC scenario. Following the single-shot RC scenario experiments are performed on two publically available datasets presented in [19] and on the well established COsy Localization Database (COLD) [22].

Some datasets containing odometry and range sensor readings, like our dataset, are already publicly available (like [21, 22], for example). However, most of these are primarily targeted for research on Simultaneous Localization And Mapping (SLAM). Moreover, some datasets have been obtained in the outdoor environment, while those that have been obtained indoors usually correspond to office-like environments. As far as we know, there is no freely available dataset that possesses data obtained from a large number of different rooms in the domestic environment and which could be effectively used for range-scan-based room categorization.

3. Learning a Hierarchical representation of space

We start this section by briefly describing the original HoP algorithm designed for object categorization. In the second subsection we extend the algorithm to make it suitable for the representation of space.

3.1 The HoP algorithm

Learning the Hierarchy of Parts (HoP) is an efficient, biologically inspired algorithm that was originally designed for object categorization [14]. The algorithm learns a compositional hierarchy of so-called parts. On the lowest level of the hierarchy parts are represented as Gabor filters, corresponding to small fractions of oriented edges in the image. Parts in the upper layers are composed from the lower-layer parts; therefore, they increase in size and complexity with each following layer. Parts in the top-most layer are representations of entire objects.

When learning object categories, the procedure follows two important steps. First, the algorithm is given a large set of everyday images containing various objects. Based on these images, a few lower layers of the hierarchy are learned in an unsupervised fashion. All the learned parts are the ones that occur most frequently in the images, thus giving the hierarchy compactness, while maintaining high representativeness. Lower layers of the hierarchy are category-independent; therefore, they are common to all
categories. This sharing of lower-layer parts is a basic advantage of the described representation. Second, for each specific category that needs to be learned, the algorithm is given a set of images containing only objects of that particular category. Upper layers of the hierarchy are then learned based on these images, again providing only statistically-significant parts in terms of frequencies of appearance. Upper layers of the hierarchy are category-specific and are therefore characteristic for each category and are also learned with minimal supervision.

In the context of learning a spatial representation, the main drawback of IHoP is that its library parts are not rotationally invariant, which is a crucial property for obtaining a compact and expressive hierarchy. For this reason, we extended IHoP to satisfy the rotational invariance condition. To augment the set of all detectable orientations, the extended algorithm, the sHoP, uses 18 Gabor filters on the lowest layer in contrast to IHoP, which uses only 6. The learning algorithm and the inference process have also been extended to make them suitable for dealing with rotational invariance.

3.2 The sHoP algorithm

We use a laser-range finder mounted on a mobile robot to observe the environment, which provides us with range data that is ground plan-like. Laser scans are afterwards transformed into images, which are the input for the sHoP learning algorithm (see Figure 1). We are able to construct lower, category independent layers of the hierarchy with an unsupervised learning algorithm, which learns the most frequent parts observed in the images by the examination of a large amount of spatial data.

![Figure 1](image.png)

**Figure 1.** Laser scans are transformed into images. Figure shows three example images acquired at different positions in one of the living rooms of the DR Dataset. The contents of these images are hard to recognize even for a human, which shows the difficulties of our approach. We might get some impression about the shape of the room from the first image. However, most of the lines that seem to form the walls of the room are actually sofas, while the walls are hidden behind them. To get some sense out of the other two images is even harder. In the last image the robot was directed to a passage between the couch and the wall.

3.2.1 The Library of Parts

The sHoP learning algorithm takes as input a set of images, where each image is a representation of a single range scan (Figure 1). The output is a library of parts representing the learned hierarchy, which is composed of several layers (Figure 2). The learned rotationally invariant parts represent spatial shape primitives with a compositional structure.

On the lowest layer there are 18 Gabor filters (Figure 2-a). These model 18 different detectable orientations, where two consecutive ones differ by 10 degrees. These filters could be considered as a single rotationally invariant first-layer part, but for better computational efficiency all of them are stored in the library.

In the higher layers, each part is a composition of two parts from the previous layer. Each layer contains only those parts, which were observed most frequently in the input images and each part is stored only in a single orientation, the one in which it was observed most frequently.

![Figure 2](image.png)

**Figure 2.** (a) Orientation-specific Layer 1 parts. (b) Layer 2 parts. (c) The most frequent 200 parts on Layer 3. The image is shrunken to save space. (d) Only a few Layer 4 parts on a shrunken image. Other parts are not shown to save space.
3.2.2 Part Structure and Rotational Invariance

The library serves as a knowledge base, which is used by the robot when it observes the environment. Parts are stored in the library in a single reference orientation and are structured in the following way.

A Layer $k$ part $P$ is defined by the expression

$$P = \{\text{layer}, \text{type}, p_{1\text{pos}}, p_{1}, p_{2}\},$$

(1)

where $\text{layer}$ denotes the layer to which the part corresponds (it equals $k$ in this case), $\text{type}$ is a consecutive number of the part in the considered layer of the hierarchy and it is used to identify parts, $P_1$ and $P_2$ are parts from the previous $(k-1)$-st layer (we call them subparts), whose composition is $P$, while $p_{1\text{pos}}$ is the position of $P_1$ relative to the geometric mean of $P_1$ and $P_2$ positions (see Figure 3). The structure of $P_1$ and $P_2$ subparts is analogous to the structure of $P$ and can be described as:

$$P_1 = \{\text{layer}_1, \text{type}_1, p_{11\text{pos}}, p_{11}, p_{12}\},$$

(2)

$$P_2 = \{\text{layer}_2, \text{type}_2, p_{21\text{pos}}, p_{21}, p_{22}\},$$

(3)

where $\text{layer}_1 = \text{layer}_2 = k-1$, $\text{type}_1$ and $\text{type}_2$ are the consecutive numbers of $P_1$ and $P_2$ on layer $k-1$, $P_{11}$, $P_{12}$, $P_{21}$ and $P_{22}$ represent subparts from Layer $k-2$, while $p_{11\text{pos}}$ and $p_{21\text{pos}}$ are the positions of $P_{11}$ and $P_{21}$ relative to the geometric means of the subpart positions. To add order to part definitions, parameter $p_{1\text{pos}}$ always corresponds to the subpart with the smallest $\text{type}$ value. In other words, the condition $\text{type}_1 \leq \text{type}_2$ must always be satisfied. On the lowest layer of the library, parts are simply represented with one of the integer values from the set $\{1, 2, \ldots, 18\}$, where 18 corresponds to the number of Gabor filters.

**Figure 3.** Schematic view of the structure of an example part $P$. A letter $G$ is used to denote the geometric mean of subpart $P_1$ and $P_2$ positions. Geometric mean of parts $P_{11}$ and $P_{12}$ is denoted with $G_1$, while $G_2$ corresponds to geometric mean of $P_{21}$ and $P_{22}$. Other quantities are defined in the text.

Each laser scan that is potentially observed by the robot during the observation of the environment is represented with a list of parts at certain locations relative to the robot. The creation of this list falls under the topic of inference process, which will be further explained in the next subsection. However, at this point we describe the structure of inferred parts, which differs from the structure of parts in the library and whose purpose is to model the environment at hand.

Inferred part $I$ is defined by the expression

$$I = \{\text{layer}, \text{type}, \phi\},$$

(4)

where $\text{layer}$ denotes the layer to which the inferred part corresponds, $\text{type}$ is a reference to an appropriate part from the library and $\phi$ denotes an inferred part’s orientation. Since each inferred part is represented using only three numbers, the model has the potential to scale very well, as higher and higher layers of the hierarchy are introduced.

During the learning and inference process, which will both be described in the next subsection, parts are compared to each other. If the part at hand is an inferred part, it is first reconstructed into the form of the parts in the library, before it is used to perform any computations. Both parts from the library and inferred parts are not rigidly defined, because we allow for some variance in their structure. This means that the positions of the subparts, from which they are composed, can vary slightly, while they still represent the same part. The same holds for rotational invariance. When the decision has to be made if two parts represent the same part in two different orientations, one part is rotated into another. The result of the rotation does not need to match perfectly with other part’s orientation; rather there is some predefined threshold that defines how much they can differ.

Rotational invariance is achieved in the following way. Let us assume that we would like to compare two parts, $P$ and $Q$. We would like to determine if they represent the same part in two different orientations. Both parts need to be from the same layer (Layer $k$). The structure of $P$ is defined above, while $Q$ can be described using the analogous notation (note that inferred parts are also transformed into this structure before performing the computations):

$$Q = \{\text{layer}', \text{type}', p_{1\text{pos}}, p_3, p_4\},$$

(5)

where $\text{layer}' = k$, $\text{type}'$ is unknown (otherwise the comparison would be unnecessary) and

$$P_3 = \{\text{layer}_3, \text{type}_3, p_{3\text{pos}}, p_{31}, p_{32}\},$$

(6)

$$P_4 = \{\text{layer}_4, \text{type}_4, p_{4\text{pos}}, p_{41}, p_{42}\}.$$  

(7)

The following sequence of simple computations is performed to verify the equality of parts:
• Index type1 has to match type1 and type2 has to match type2.
• We calculate the potential angle of rotation \( \varphi \) of a part \( P \) relative to \( Q \) from \( p_{1\theta} \) and \( p_{2\theta} \).
• Rotation of \( p_{1\theta} \) for angle \( \varphi \) must result in \( p_{2\theta} \) and rotation of \( p_{2\theta} \) for angle \( \varphi \) must result in \( p_{1\theta} \).

If the results of the above computations lie within the predefined bounds, we can conclude that the parts are of the same type.

3.2.3 Learning Algorithm and Inference Process

In this subsection learning of the library of parts and inference process are described in more detail. The input for the learning algorithm is a set of images obtained from range data and the output is a hierarchy of rotationally invariant parts. While the inference process is also used as an intermediate step in the learning procedure, its main purpose is to detect and extract part information from input images, once we already have a learnt library.

Learning starts with a library containing a single layer \( L_1 \) with 18 Gabor filters. In the first step edge detection is used to find small fragments of oriented edges in the input images. Positions and response intensities of edge fragments (\( L_1 \) parts) are then used as an input for learning the second layer of the hierarchy. The first layer is the only layer that is fixed and that is not learned from observations.

In all of the following steps the sequential interchange of layer learning and inference process is performed. The hierarchy is being learned layer-by-layer, by repeating the following two steps:

• Layer learning: Positions, confidence values and orientation information of Layer \( k \) parts corresponding to every input image are used to learn Layer \( k+1 \) of the hierarchy. We address the notion of confidence values in the context of inference below.
• Inference process: Using the inferred parts from the previous \( k \)-th layer, and Layer \( k+1 \) of the library, positions, orientations and types of Layer \( k+1 \) parts in all of the input images are inferred (see Figure 4).

When inferring the positions and orientations of parts in the images each inferred part is assigned two confidence values. These values carry information about how well a certain instantiation of a part in the image represents the corresponding part from the library. The first value is based on the response intensities of edge detection and the second is a measure of how precisely an instantiation of a part can be rotated into a corresponding part in the library.

Figure 4. Positions of Layer 3 parts inferred from an example image acquired in one of the living rooms (middle image in Figure 1). Different colours correspond to different part types.

Now let us take a closer look at the learning of Layer \( k+1 \) from Layer \( k \). This is performed in four steps: Finding Neighbourhoods, Finding Local Maxima, Forming a Sequence of Rotationally Invariant Parts and Adding Parts to the Library, which are all described below.

Finding Neighbourhoods All of the inferred images, constructed from Layer \( k \) parts, are inspected for pairs of nearby parts, while the following actions are being performed:

• For every part \( C \) that is observed in the images and for every orientation \( O_c \) of part \( C \) that is observed, assume \( \{C,O_c\} \) is the central part of the pair.
• For every part \( N \) and every orientation \( O_N \) of part \( N \) that is observed in the proximity of \( \{C,O_c\} \), assume \( \{N,O_N\} \) is the neighbour part in the pair.
• For every such pair a neighbourhood \( Nb(\{C,O_c\};\{N,O_N\}) \) is maintained.

These neighbourhoods are used to store the information about relative positions and position frequencies of parts. That is, each neighbourhood tells us in which positions \( \{N,O_N\} \) has been observed relative to \( \{C,O_c\} \) in the input images and also how many times these positions were observed.

Finding Local Maxima For every observed combination of part types and orientations, the most frequent relative positions of parts are obtained by searching for the maximum values of relative position occurrences in the neighbourhoods. Usually, one or two maxima are found for each neighbourhood.

Forming a Sequence of Rotationally Invariant Parts All of the inferred images containing Layer \( k \) parts are inspected. The images are searched for pairs of parts in specific orientations, while considering only their relative positions defined by neighbourhood maxima. The number of occurrences of all of the observed pairs satisfying this condition is counted, while the results representing the same pairs in different orientations are summed together. This procedure thus constructs a list of
pairs with their corresponding frequencies, which is sorted according to the frequency of occurrence in descending order at the end. In this way a list of rotationally invariant Layer \( k+1 \) part candidates is obtained.

**Adding Parts to the Library** The sequence of part candidates constructed in the previous step contains the most frequent parts at the beginning and the least frequent ones at the end. Therefore, depending on how large a representation we want, a portion of parts at the beginning of the list is declared as Layer \( k+1 \) of the library, while the rest is discarded.

4. Application to Room Categorization

In this section we present our new low-level image descriptor (HoC) and we describe how we use it in two different scenarios of room categorization.

4.1 The Histogram of Compositions

Our Histogram of Compositions (HoC) descriptor is created from a single laser range measurement, while the previously learnt library is required for the creation process. The descriptor is formed through the following steps:

- A range measurement obtained by the mobile robot is transformed into an image, on which nearby points are artificially connected. Connecting the points is necessary because pure range data provides a discrete set, on which edge detection is not efficient. Some example images, obtained at different locations in one of the living rooms of the DR Dataset, are shown in Figure 1.
- An inference process is used to infer positions and orientations of parts from the image, for some chosen layer of the hierarchy (see Figure 4).
- The positions of parts in the inferred image are rotated into a reference position, with the use of principal component analysis (PCA).
- The image is divided into 24 regions as shown in Figure 5. The robot is positioned in the centre.

![Figure 5](image.png)

**Figure 5.** The 24 regions of the image. The dashed arrows correspond to the principal axes.

- A sequence of 24 histograms is created, where each histogram corresponds to one region of the image. The number of bins in each histogram is equal to the number of parts in the considered layer of the hierarchy. Each bin corresponds to one part type from that layer, while the value corresponding to the height of the bin equals the sum of confidences of parts in that region. All of the histograms are then concatenated into a single feature vector, forming our HoC descriptor.

4.2 Two Scenarios for Room Categorization

We consider two scenarios for room categorization in this work. The first scenario, exploratory RC, is based on a set of range measurements, while the second scenario, single-shot RC, is based on a single observation. In each scenario categorization is based on HoC, while the feature vectors used as input for SVM differ as described below.

The first scenario is as follows. Consider a mobile robot equipped with a door detection system. The robot enters a new room by passing the door, which triggers the room categorization procedure. The robot starts exploring the room and then after a short tour it decides upon its type. We undertake this problem by first transforming a set of laser scans obtained in a single room to a set of HoC descriptors. Then a single feature vector is created as a composition of an average and standard deviation of all the descriptors in the set, forming a compact representation of the room. Therefore, the input for the SVM consists of a feature vector for each room.

The idea of the second scenario is to perform room categorization based only on a single range measurement. In this case, a single range measurement represented as a HoC descriptor is directly used as a feature vector for the input to SVM. Therefore, when following this approach, there are as many feature vectors as there are laser scans obtained in a room.

5. The DR Dataset

The DR Dataset contains robot observations of several rooms from a domestic environment. We have gathered our data with a Pioneer P3-DX robot, which was manually driven through different rooms using a joystick, while observing the space with a Hokuyo URG laser range-finder (see Figure 6). The range sensor has a field of view of 240 degrees, a range of 5.6m and was positioned approximately 30cm above the floor. The robot was guided for a short tour through each room (at speeds between 0.1 and 0.5 m/s), while odometry readings and laser range data were recorded with a frequency of 10 frames per second. We have posed some restrictions while gathering the data. There were no people in any of the rooms and all of the doors were closed, meaning that no other room was visible from any room, other than the one the robot was currently observing. Exceptions are spaces, where two rooms are not clearly separated. For example, it is common that a
living room and a kitchen share the same space in an apartment. In those cases, the robot was guided through the room following a trajectory that avoided the views of the other room type as much as possible. These restrictions provided us with a controlled environment, which could also be used as a reference in any future experiments about the robustness of the algorithms, when other variables will also be present in the data.

**Figure 6.** Our robot. We used only the base and the laser range finder, which is indicated with an arrow, to gather our data.

24 different homes have been scanned in this way, consisting of houses and apartments from 4 different cities, which contribute to the high variability of our data. There are 90 rooms in our dataset, consisting of 21 living rooms, 6 corridors, 35 bathrooms and 28 bedrooms. Typical corridors, which are long, straight and narrow spaces, are not very common in homes and this is the reason why we were only able to observe a small number of them. Different numbers of laser scans were obtained in each room. Their number is dependent on room size and varies from 100 to 600 scans. Some examples of obtained scans are shown in Figure 1. For convenience and better visualisation we applied a SLAM algorithm [23] to our data, but it is not used for our categorization method. Ground plan images that were obtained in this way for a few example rooms are shown in Figure 7.

The dataset is organized in the following way. There are four directories, where each corresponds to one room type. In each directory there are as many files as there are examples of rooms that have been scanned for the corresponding room type. Therefore, each file contains all the data that has been gathered in a single room. Rows of the file represent consecutive observations, while each contains the following information: timestamp, corresponding room type, odometry reading, sensor characteristics and range measurements. The dataset is freely available at http://go.vicos.si/drdataset.

**Figure 7.** Ground plan views of some of the example rooms. The images were generated with Rao-Blackwelized Particle Filter SLAM algorithm [23], provided by The Mobile Robot Programming Toolkit. Two examples per room type are shown in a single row. Living rooms are displayed in the first row, corridors in the second, bathrooms in the third and bedrooms in the fourth. Path of the robot is displayed in light blue, while the detected obstacles are shown in dark blue.

6. Experimental Results

Following two different scenarios for room categorization we have conducted two sets of experiments. The aim of the first set, referring to exploratory RC, was to show that (i) even the lower layers of our hierarchy are rich with information required for room categorization and (ii) that the HoC descriptor is an effective tool for room categorization. We demonstrate this using the DR Dataset. To compare our work with current state-of-the-art techniques, an equivalent experiment was designed using the state-of-the-art algorithm proposed by [19] on the same dataset. The second set of experiments refers to single-shot RC. To evaluate our approach in the context of other algorithms even more thoroughly, two publicly available datasets, presented in [19], were used to conduct the experiments. Moreover, experiments were also performed on the well-established COLD [22] dataset.

A separate dataset that was also acquired in our work, consisting of a large number of rooms and corresponding
to several categories, was used to learn the category independent layers of our hierarchy, which were then used in all of the following experiments as our spatial representation. In all of the experiments training and testing data were completely separate, therefore the tests were always performed on previously unseen places. We used a LIBSVM library [24] for the categorization with SVM.

6.1 Exploratory Room Categorization

In this subsection we first describe the extensive experiments that were performed using our approach, to evaluate the effectiveness of our spatial model. Then, we describe the experiment that compares our results with the state-of-the-art results.

6.1.1 Proposed Spatial Model Evaluation

Using the DR Dataset we performed a set of categorization experiments in which a single feature vector per room was used as an input to SVM (as described in Subsection 4.2). Different layers of the hierarchy were tested to evaluate the suitability of our representation of space. We have also performed an experiment in which our representation was not used. In this case the number of scan points was counted in each region corresponding to the HoC descriptor. In preliminary research, we considered several different divisions of the image into regions for the formation of the HoC. The optimal regions are shown in Figure 5 in Section 4. The radii that correspond to the circles have the values of 0.8, 1.5 and 2.5 m if scaled up to real world size, while the angles of the straight lines have values of 20, 90, 160, 200, 270 and 340 degrees.

In the categorization procedure we performed 1000 trials of each experiment, in which we distinguished between 4 types of rooms. In each trial, training (80%) and testing (20%) data was randomly chosen from the set of all the available feature vectors (note that each room was represented using a single feature vector). A linear kernel was used for categorization in every experiment. We used 4-fold cross validation on the training set to find the best parameter C for the SVM. The learned model’s performance was then tested on the testing data.

The results of the categorization are shown as confusion matrices in Table 1. By Layer 0 we denoted the experiment without SHoP processing, in which raw scans were used instead (scan points counting in each region). We could say that this is analogous to having only a single part in the library. As mentioned earlier, in Layer 1 there are a fixed number of orientation specific parts. Our learning algorithm learned 12 parts in Layer 2. In Layer 3 a large number of different parts were observed, but only the most frequent ones were retained. How many of them are stored in the library is considered a parameter. We performed the tests with 80, 200 and 500 parts in Layer 3. The confusion matrix is shown only for the case with 200 parts, because this one yielded the best results. We have also tested Layer 4 of the library. Parts on that layer are relatively large and therefore cover the scans quite poorly. A lot of information is lost in this way, which causes reduced categorization performance. We therefore focus only on Layers 1, 2 and 3 of the hierarchy.

| Layer | Lr | Cr | Ba | Be |
|-------|----|----|----|----|
| 0     |    |    |    |    |
|       | 76.96 | 3.50 | 0.02 | 19.52 |
| Cr    | 6.60 | 82.70 | 0.20 | 10.30 |
| Ba    | 0.14 | 3.21 | 87.50 | 9.14 |
| Be    | 10.92 | 0.04 | 15.18 | 73.86 |
| 1     |    |    |    |    |
|       | 79.68 | 0.22 | 0.04 | 20.06 |
| Cr    | 0 | 100.00 | 0 | 0 |
| Ba    | 0.19 | 1.30 | 83.50 | 15.01 |
| Be    | 15.44 | 0 | 10.32 | 74.24 |
| 2     |    |    |    |    |
|       | 75.16 | 0 | 3.46 | 21.38 |
| Cr    | 0 | 82.50 | 0 | 17.50 |
| Ba    | 0.04 | 0.51 | 92.19 | 7.26 |
| Be    | 15.68 | 0.06 | 12.14 | 72.12 |
| 3     |    |    |    |    |
|       | 82.74 | 0 | 0.08 | 17.18 |
| Cr    | 0 | 91.60 | 0 | 8.40 |
| Ba    | 0.20 | 0.29 | 92.96 | 6.56 |
| Be    | 13.32 | 0.22 | 16.24 | 70.22 |

Table 1. Confusion matrices obtained with SVM categorization using 4 categories and different layers of the hierarchy. There were 200 parts used in layer 3. Labels at the borders of the table denote: Lr-living room, Cr-corridor, Ba-bathroom, Be-bedroom. The entries are in percents. Rows correspond to ground-truth, while the columns correspond to predicted categories.

The accuracies of the categorization, computed as a percentage of correctly categorized examples, averaged over all trials of the experiment, are shown with their corresponding standard deviations in Figure 8 (a). We have also performed a series of t-tests, with which we compared each of the obtained results. We tested the null hypothesis, at α=0.01 significance level, that the two distributions of calculated accuracies have equal means. The results are shown as the above diagonal elements in Figure 8 (c). The ‘+’ signs denote a failure to reject the null hypothesis, which implies that the two observed accuracies are statistically equivalent. On the contrary, the ‘-’ signs denote the rejection of the null hypothesis at α=0.01 significance level, which means that there is a 99% probability that the two accuracies truly represent a different result. The results suggest that Layers 0, 1 and 2 show equal categorization performance, which means that parts on the lowest layers of the hierarchy are too small to possess any significant information about the local structure of space. On the other hand, Layer 3 provides statistically significantly better results. In particular, the accuracy of 80.52% in Layer 0 is increased to 83.73%. The results of the t-tests also suggest that the number of parts in Layer 3 is of no significant importance.
To evaluate the results from a different perspective we analysed another measure of categorization performance. We call this measure the mean success rate, which is computed as a mean of the diagonal entries of the confusion matrix, averaged over all trials of the experiment. Mean values with standard deviations are displayed in Figure 8 (b), while the corresponding t-test results are shown as the below diagonal entries in Figure 8 (c). This view of the results confirms that Layer 0 and 2 perform equally well, but Layer 1 stands out, showing significantly better performance. The reason for this is that on Layer 1 we actually perform the categorization based on the orientation of an edge. A specific orientation, as expected, very well characterizes corridors, but not other rooms, which can be confirmed by examining the confusion matrices. In Layer 2 this orientation information is not incorporated anymore and therefore the performance is reduced. This measure also clarifies that Layer 2 and Layer 3, with a low number of parts, show similar performance. A greater number of parts in Layer 3 significantly improves the categorization performance, from 80.26% in Layer 0 to 84.38%. The results also suggest that continuously increasing the number of parts does not increase the mean success rate. Therefore, using only the most frequent parts and also not too many of them, seems a reasonable thing to do.

Overall, bathrooms and corridors are the rooms that are categorized most accurately, followed by living rooms and then bedrooms. We believe that what makes these categorizations possible at all are general room shapes and properties. At least in homes corresponding to the DR Dataset, the characteristics of each category could be the following: Bathrooms are small rooms with square or slightly rectangular shape, bedrooms have a distinctive property of small passages between the bed and nearby walls, living rooms are usually larger and stuffed with all sorts of furniture and objects, while corridors represent long and narrow spaces.

6.1.2 Comparison With State-of-the-Art Techniques

To compare our results with the state-of-the-art we designed an equivalent experiment using the approach presented in [19]. The approach uses AdaBoost to boost simple features to a strong classifier and is based on a single scan. At the parameter determination step we determined the optimal number of hypotheses used and the optimal order of binary classifiers (see [19] for details about the algorithm). The optimal decision list turns out to be corridor, bathroom and living room. During the experiment, every laser scan obtained in a particular room has been categorized using their algorithm, while at the end majority voting has been used to determine the room type.

The overall accuracy of the categorization equals 83.70%, which is similar to our result (83.73%). The confusion matrix is shown in Table 2. The mean success rate equals 86.10%, which is almost 2% better than our approach. However, it can be seen from the confusion matrix that the two algorithms give complementary results. Our method achieves better performance with bathrooms and bedrooms, while the approach of [19] provides better results with living rooms and corridors.

|   | Lr | Cr | Ba2 | Be  |
|---|----|----|-----|-----|
| Lr| 91.33 | 0  | 4.00 | 4.67 |
| Cr| 0  | 100.00 | 0  | 0  |
| Ba| 0  | 0  | 92.38 | 7.62 |
| Be| 18.00 | 2.00 | 19.33 | 60.67 |

Table 2. Confusion matrix obtained with the approach based on the work of [19]. Labels at the borders of the table denote: Lr-living room, Cr-corridor, Ba-bathroom, Be-bedroom. The entries are in percents. Rows correspond to ground-truth, while the columns correspond to predicted categories.

6.2 Single-Shot Room Categorization

In this subsection we present single-shot-based experiments. We first describe the experiments that were
performed to compare our approach to the work of [19]. Then, we describe an additional experiment that was performed on the COLD [22] dataset.

6.2.1 Experiments on Freiburg II datasets

Single-scan-based experiments have been performed on two datasets from [19]. The first one was an office environment in Building 79 at the University of Freiburg, while the second one is Building 101 at the same University. Using the former one, [19] distinguished between three categories, which are corridor, room and doorway and managed to achieve an accuracy of 93.94%. With the latter one, they distinguished between four categories, where hallway was added to previous ones. Their obtained accuracy in this case was 89.52%.

The laser used in [19] had a maximum range of 80m, whereas the laser used in the acquisition of the DR Dataset had a range of 5.6m. For this reason, two extra radii were used in the formation of the HoC descriptor in these experiments, corresponding to circles around the robot with radii of 6 and 10 m, which resulted in 12 additional regions of the descriptor. We used Layer 3 of the hierarchy with 200 parts. The experiments were analogous to the ones in [19]. Part of the environment specified for training was used to train the SVM and parameter estimation, while the part specified for testing was used for testing. Like [19], we performed the categorization based on a single scan and used the same category definitions. Authors in [19] reported their results with overall accuracy of the categorization and the image of the environment, on which different colours indicate the predicted category for the corresponding point in the environment. Therefore, to allow proper comparison we present our results in the same manner.

In the Building 79 experiment we obtained an accuracy of 98.32%, while in the Building 101 experiment our result was 90.39%. The results are summarized in Table 3. The images corresponding to both experiments are shown in Figure 9 and Figure 10. Positions categorized as corresponding to a room are coloured dark blue, red corresponds to corridor, yellow to doorway and light blue to hall. Similarly to [19], our method also has problems detecting doorways. As noted in [19], a doorway is typically a very small region so only a few training examples are available. Furthermore, if a robot stands in a doorway the scan typically covers nearby rooms or corridors, which makes it hard to distinguish the doorway from such places. Furthermore, our method is meant for room categorization, not door detection. It can be seen in Figure 10 that some places inside some of the rooms are categorized as corridors. The reason for this is that these regions correspond to narrow areas in the room that look very similar to corridors. This can be verified by inspecting the environment map shown in [19]. Overall, we conclude that our approach delivers comparable results to the state-of-the-art [19].

|                  | [19]  | HoC |
|------------------|-------|-----|
| Building 79      | 93.94 | 98.32|
| Building 101     | 89.52 | 90.39|

Table 3. Accuracy of the categorization in the single-scan-based room categorization experiments on two publicly available datasets.

![Figure 9](image1.png) Categorization results obtained on Freiburg B79 dataset. Different colours correspond to different predicted categories: dark blue-room, red-corridor, yellow-doorway.

![Figure 10](image2.png) Categorization results obtained on Freiburg B101 dataset. Different colours correspond to different predicted categories: dark blue-room, red-corridor, yellow-doorway, light blue-hall.

6.2.2 Experiment on COLD-Freiburg dataset

The COLD dataset [22] contains data gathered from three laboratories located in three different cities: AIS lab in Freiburg, LT lab in Saarbrucken and VICOS lab in Ljubljana. There was no laser range data gathered in VICOS, therefore we are left with the two remaining ones. In both labs, AIS and LT, the environment was divided into two parts, A and B. Therefore, there are 4 environments with available laser range data. In most of these environments two different paths were followed by the robot during data acquisition: (i) The standard path, where the robot was driven across rooms that are most likely to be found in most labs; (ii) the extended path, where the robot was additionally driven across the rooms that were specific to each lab. Because the rooms in the
The recognized categorization experiment. The experiments focused on the corridor and the lab. We performed the categorization using four categories: Corridor (CR), two-person office (2PO), stairs area (SA) and toilet (TL). The categorization was based on a single shot, while the HoC descriptor and SVM parameters were the same as in the experiment presented in the previous subsection. For testing the categorization power of our approach, we trained our classifier on part A and tested on B and afterwards, analogously, we trained it on part B and tested on A. The results, averaged over both trials of the experiment, are shown in table 4.

|       | CR  | 2PO | SA  | TL  |
|-------|-----|-----|-----|-----|
| CR    | 90.84 | 3.14 | 2.22 | 3.81 |
| 2PO   | 19.23 | 46.13 | 30.90 | 3.74 |
| SA    | 7.44 | 13.82 | 70.52 | 8.22 |
| TL    | 4.18 | 3.72 | 14.56 | 77.54 |

Table 4. Confusion matrix displaying the results of the room categorization experiments performed on A and B parts of the AIS lab. The results are averaged over both trials of the experiment.

The best results were obtained for the corridors, since they are very well characterized by their distinctive shape. For improved categorization performance with other categories, we expect our range-data-based classifier could benefit from fusion with vision-based approaches. An interesting note is that the categorization performance in the work of [22] was also the best for corridors, as in our results, although their approach is not directly comparable to ours. The difference is that they performed the categorization based on data obtained with a camera and they performed the experiments on slightly different classes and environments from the COLD dataset.

7. Conclusion and Future Work

We have presented a new hierarchical representation of space, which is learned using our sHoP algorithm. The hierarchy is constructed from parts that are rotationally invariant and statistically significant. To our knowledge this is the first application of a hierarchical compositional model to a spatial representation on the lowest semantic level. We have also presented a low-level image descriptor, which was used in two different approaches for room categorization and for the validation of our spatial model. We have obtained good categorization results on demanding datasets, which indicate that the proposed hierarchy of parts is suitable for representing space, and that our approach delivers state-of-the-art results on room categorization. Furthermore, we have also presented a large freely available dataset, which can be used to benchmark laser-scan-based room categorization algorithms.

There are several research venues we intend to pursue in our future work. From current research we have determined that three is the maximum number of layers that is reasonable to be learned from separate images. In the future we plan to align the images, obtained in a certain room, to a single map of that room. These maps will provide us with a more complete view of the environment, which in contrast cannot be obtained using a single laser scan. Obtained maps will serve as a basis for learning higher, category-specific layers of the hierarchy. We expect that room recognition performance will be increased with the abstraction introduced by these layers. Moreover, we also expect that room categorization will improve by using our category-specific parts from higher layers. The learned hierarchy will be used as prior knowledge for a service robot, providing the basis for recognition of new, never before seen, environments and simultaneously ensuring good scalability of the model.

8. Acknowledgments

The research leading to these results has received funding from the EU FP7 project CogX. We would also like to thank Oscar M. Mozos for sharing his algorithms with us.
9. References

[1] Škocaj D, Janiček M, Kristan M, Kruijf G J M, Leonardis A, Lison P, Vreeko A, Zillich M (2010) A basic cognitive system for interactive continuous learning of visual concepts. ICRA 2010 workshop ICAIR. pp. 30-36.

[2] Kuipers B (2000) The Spatial Semantic Hierarchy. Artificial intelligence. Vol. 119, pp. 191-233.

[3] Pronobis A, Sjöö K, Aydemir A, Bishop A N, Jensfelt P (2010) Representing spatial knowledge in mobile cognitive systems. Robotics. Available: http://www.pronobis.pro/publications/pronobis2010s prep.pdf. Accessed 2012 May 14.

[4] Elffes A (1990) Occupancy Grids: A Stochastic Spatial Representation for Active Robot Perception. UAI. pp. 136-146.

[5] Arras K O (2003) Feature-Based Robot Navigation in Known and Unknown Environments. Ph.D. thesis. EPFL Lausanne.

[6] Tapus A (2005) Topological SLAM - Simultaneous Localization And Mapping with fingerprints of places. Ph.D. thesis. EPFL Lausanne.

[7] Choset H, Nagatani K (2001) Topological simultaneous localization and mapping (SLAM): Toward exact localization without explicit localization. IEEE Trans. Robot. Autom. Vol. 17, no. 2, pp. 125-137.

[8] Thrun S (1998) Learning Metric-Topological Maps for Indoor Mobile Robot Navigation. Artificial Intelligence. Vol. 99, no. 1, pp. 21-71.

[9] Thrun S, Gutmann J S, Fox D, Burgard W, Kuipers B (1998) Integrating topological and metric maps for mobile robot navigation: A statistical approach. AAAL. pp. 989-995.

[10] Mozos O M (2008) Semantic Labeling of Places with Mobile Robots. Ph.D. thesis. Albert-Ludwigs-Universität Freiburg.

[11] Vasudevan S, Gächter S, Harati A, Siegwart R (2007) A Hierarchical Concept Oriented Representation for Spatial Cognition in Mobile Robots. In: Lungarella M, Pfeifer R, Iida F, Bongard J, editors. 50 years of artificial intelligence. Springer-Verlag.

[12] Pronobis A, Jensfelt P (2012) Large-scale Semantic Mapping and Reasoning with Heterogeneous Modalities. ICRA. pp. 3515-3522.

[13] Zivkovic Z, Booij O, Kröse B (2007) From images to rooms. Robot. Auton. Syst. Vol. 55, no. 5, pp. 411-418.

[14] Fidler S, Boben M, Leonardis A (2009) Evaluating multi-class learning strategies in a hierarchical framework for object detection. NIPS. Vol. 22, pp. 531-539.

[15] Fidler S, Boben M, Leonardis A (2008) Similarity-based cross-layered hierarchical representation for object categorization. CVPR. pp. 1-8.

[16] Mozos O M, Triebel R, Jensfelt P, Rottmann A, Burgard W (2007) Supervised semantic labeling of places using information extracted from sensor data. Robot. Auton. Syst. Vol. 55, no. 5, pp. 391-402.

[17] Wu J, Christensen H I, Rehg J M (2009) Visual Place Categorization: Problem, dataset, and algorithm. IROS. pp. 4763-4770.

[18] Swadzba A, Wachsmuth S (2008) Categorizing Perceptions of Indoor Rooms Using 3D Features. Lecture Notes in Computer Science: Structural, Syntactic, and Statistical Pattern Recognition. pp. 734-744.

[19] Mozos O M, Stachniss C, Burgard W (2005) Supervised Learning of Places from Range Data using AdaBoost. ICRA. pp. 1730-1735.

[20] Friedman S, Pasula H, Fox D (2007) Voronoi random fields: extracting the topological structure of indoor environments via place labeling. IJCAI. pp. 2109-2114.

[21] Howard A, Roy N (2003) The Robotics Data Set Repository (Radish). http://radish.sourceforge.net/.

[22] Ullah M.M, Pronobis A, Caputo B, Luo J, Jensfelt P, Christensen H I (2008) Towards robust place recognition for robot localization. ICRA. pp. 530-537.

[23] Doucet A, De Freitas N, Murphy K, Russell S (2000) Rao-Blackwellised particle filtering for dynamic Bayesian networks. UAI. pp. 176-183.

[24] Chang C C, Lin C J (2011) LIBSV: a library for support vector machines. TIST. Vol. 2, no. 3, pp. 27:1-27:27.