Learning part-based spatial models for laser-vision-based room categorization

The International Journal of Robotics Research
XX(X):1–18

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Peter Uršič[1] and Aleš Leonardis[1,2] and Danijel Skočaj[1] and Matej Kristan[1]

Abstract

Room categorization, i.e., recognizing the functionality of a never before seen room, is a crucial capability for a household mobile robot. We present a new approach for room categorization that is based on 2D laser range data. The method is based on a novel spatial model consisting of mid-level parts that are built on top of a low-level part-based representation. The approach is then fused with a vision-based method for room categorization, which is also based on a spatial model consisting of mid-level visual-parts. In addition, we propose a new discriminative dictionary learning technique that is applied for part-dictionary selection in both laser-based and vision-based modalities. Finally, we present a comparative analysis between laser-based, vision-based, and laser-vision-fusion-based approaches in a uniform part-based framework that is evaluated on a large dataset with several categories of rooms from the domestic environments.

Keywords

Room categorization, part-based models, discriminative dictionary learning, laser-vision fusion

1 Introduction

High-level perception of spatial categories is a crucial capability for household mobile robots. One of such tasks is room categorization, i.e., inferring the type of the newly encountered, never before seen, room, based on the knowledge learned from past observations. Transferring the robot into a new environment should not affect its operation, and in such case, room-category inference provides conceptual information that is a basis for performing many other tasks with greater efficiency. It can be applied to improve navigation, localization, and exploration. For example, a robot might not be able to fetch a cup of coffee if it has difficulties finding a kitchen.

The complexity of room categorization task largely depends on the type of sensors being used. The laser range-data-based approaches (Mozos et al., 2005; Friedman et al., 2007; Uršič et al., 2012) deal with a point-wise two-dimensional data that represents a scarce information source. A more abundant data of greater dimensionality is available to the methods that apply a 3D infrared sensor (Swadzba et al., 2010), while a relatively highly descriptive information is employed in the vision-based approaches (Parizi et al., 2012; Dixit et al., 2015; Mandeljc et al., 2016). To further improve the recognition, researchers have considered joining information from multiple modalities. Laser range data has been joined with vision (Mozos et al., 2007; Shi et al., 2013), others used vision in conjunction with behavioral data obtained by recording the movements of the robot to avoid walls and obstacles while executing a given behavioral pattern (Yi et al., 2011), several approaches apply RGB-D sensors that provide 3D depth information along with images (Ruiz-Sarmiento et al., 2015), whereas a method combining information about object observations, the appearance, geometry, and topology of space, along with human input has been presented in (Pronobis et al., 2012).

This paper focuses on application of two of the most popular sensor modalities used in household mobile robots, i.e., the 2D laser-range scans and vision. Room categorization based on such data is a challenging problem (see Figure 1). An apparent division between holistic and part-based approaches can be drawn in both, laser-based and vision-based, fields. Holistic methods interpret input data as a whole, whereas part-based methods decompose it into smaller fragments, i.e., parts, and then perform the reasoning with respect to the per-part interpretations. We focus on part-based models, since several recent studies point out the appealing aspects of these approaches, which seem superior to the holistic methods: They result in compact and expressive representations (Fidler et al., 2009; Uršič et al., 2013), are robust to distortions and occlusion (Mandeljc et al., 2016), and parts can be recombined in a model to express a combinatorial number of variants of a place or a scene (Parizi et al., 2015). Despite various existing part-based approaches, several open challenges remain: (i) Considering previously proposed range-data-based models, an increase in categorization performance is expected with further abstraction of parts by following a hierarchical approach (Uršič et al., 2013). However, it is unclear how learning should proceed, due to incompatibility of

[1] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[2] School of Computer Science, University of Birmingham, UK

Corresponding author:
Peter Uršič, Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia
Email: peter.ursic@fri.uni-lj.si

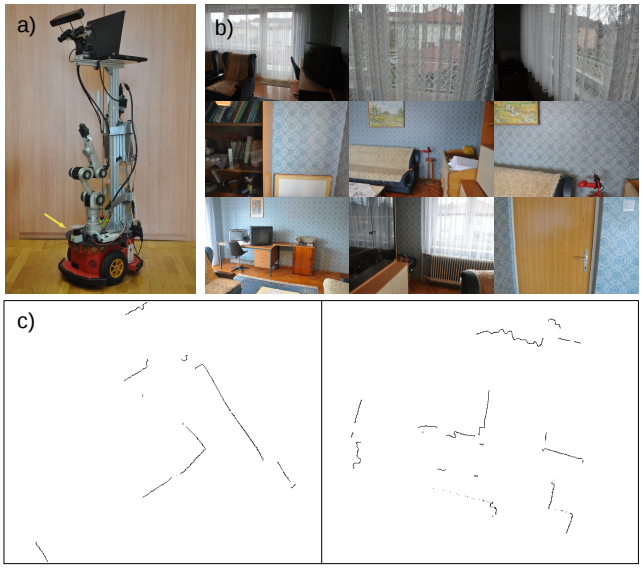

Figure 1. Room categorization based on laser and vision. (a) A mobile robot observes the environment with a laser-range finder (marked with an arrow) and a camera. (b) A few images acquired in an example living room. Due to sub-optimal viewpoints, these observations impose a challenging room categorization problem. (c) Two laser scans acquired in the same room. It is hard to reason about room category from such data even for a human.

increased part receptive field sizes with view limitations of the range scanner. (ii) Parts at various levels of granularity are employed in different modalities, laser-based and vision-based. A question remains how these models could be combined in a unified fusion scheme for improved room categorization performance. (iii) Part-based representations require building a compact and discriminative dictionary suitable for categorization. It is unclear what methods should be applied in different modalities to obtain compatible models that could be effectively joined in laser-vision fusion.

Considering the open challenges regarding part-based representations for room categorization, this work provides the following contributions:

- We present a novel approach for room categorization in the laser-based domain. A new spatial model built on top of a low-level part-based representation (Uršič et al., 2013) is proposed. The model consists of mid-level parts that are built through local map creation using Simultaneous Localization and Mapping (SLAM) and subsequent fully-connected visibility-graph extraction. The overlaid model avoids the limitations of the low-level part-based representation and provides state-of-the-art room categorization performance.
- We propose a new approach for room categorization based on laser-vision fusion in a uniform part-based framework, and we present a comparative analysis between laser-based, vision-based, and laser-vision-fusion-based part-based methods that is evaluated on several categories of rooms from the domestic environments.
- We present a new discriminative dictionary learning technique that finds a set of category-specific exemplars based on pair-wise affinities. The method is general and is in this work applied for part-dictionary selection in both laser-based and vision-based modalities, which enables a natural formulation of the fusion scheme.

This paper is a significant extension of our preliminary work that was published in (Uršič et al., 2016). The proposed discriminative dictionary learning paradigm has been upgraded, the range-data-based model for room categorization has been joined with a vision-based approach in a part-based fusion framework, and additional experiments have been carried out.

The remainder of the paper is organized as follows: Related work is discussed in Section 2, Section 3 introduces the new mid-level parts for the laser-based spatial model, Section 4 overviews the parts that we use in the vision-based model, whereas Section 5 presents our learning algorithm used for part-dictionary selection in both modalities. The approaches for laser-based, vision-based, and laser-vision-fusion-based room categorization are introduced in Section 6, experimental evaluation is reported in Section 7, and in Section 8 conclusions are drawn.

2 Related work

Several approaches for 2D range-data-based room categorization have been presented, with vast majority of them being holistic. Perhaps the most popular is the work of (Mozos et al., 2005), who performed the categorization using AdaBoost (Freund et al., 1997), which was applied to boost simple features (like the area covered by the scan or the average distance of consecutive beams) to a strong classifier. Their method was based on a single scan and they distinguished between four categories (rooms, corridors, doorways and hallways). A set of simple features extracted from a laser scan was also applied for categorization in (Sousa et al., 2007), who focused only on two basic categories (rooms and corridors). A subset of features proposed by (Mozos et al., 2005) was used in (Premebida et al., 2015). They applied dynamic bayesian mixture models for categorization and performed their experiments in an office environment with four categories. Three categories of rooms of the office environment were considered in (Shi et al., 2012), where Voronoi graphs were employed and conditional random fields (CRF) (Sutton et al., 2012) with support vector machine (SVM) (Cortes et al., 1995) ware applied to semantically label the graph nodes. The ideas of part based models were followed in (Shi et al., 2010), who proposed an approach that is able to classify different areas of a single laser scan into different semantic labels. They chose a set of dominant features introduced by (Mozos et al., 2005) and distinguished between two simple categories (rooms and corridors). Somewhat related to our work is the approach of (Friedman et al., 2007), where Voronoi random fields were employed, providing the distinction between rooms, hallways, junctions and doorways. Similarly to our approach, their method also uses SLAM to obtain maps of the environment and then extracts graphs from those maps, whereas the representations used are significantly different. The approach of (Friedman et al., 2007) requires building a holistic map of the space represented as a metric occupancy grid, meaning that the quality will significantly depend on

the accuracy of the SLAM and that it will be sensitive to any changes in the room. In contrast, our approach requires only partial reconstructions that are less prone to such problems. Moreover, this also eliminates the need for any loop-closing in SLAM. While (Friedman et al., 2007) form a spatial representation by a Voronoi graph, our graph-based representations are build as local constellations of parts and enjoy advantages of part-based models over holistic ones. The most related to the work presented in this paper is our previously proposed room categorization approach, which is based on the Spatial Hierarchy of Parts (sHoP) model (Uršič et al., 2012; Uršič et al., 2013). We applied the model similar to (Fidler et al., 2009) for parsing laser scans into parts of low-level complexity and were able to distinguish between four challenging categories of rooms from the domestic environments (living rooms, corridors, bathrooms, and bedrooms). However, a drawback of (Uršič et al., 2013) is that the final categorization model encodes spatial layout of the low-level parts in a holistic manner, i.e., a spatial histogram of parts, and therefore deviates from the predicted advantages of part-based compositional models (Fidler et al., 2009). Another drawback of the obtained model is that, like all holistic models, it is principally sensitive to noise and missing data.

Extensive research has been performed in the field of vision-based room categorization. Most approaches apply holistic image descriptors. The representations are often formed in a bag-of-words fashion (Csurka et al., 2004). For example, bag-of-words model using Scale-Invariant Feature Transform (SIFT) from (Lowe, 1999) was used in (Ayers et al., 2007), (Wu et al., 2009) apply local texture descriptors as visual words, (Margolin et al., 2014) employed oriented texture curves, while (Gemert et al., 2006) applied proto-concepts as region-based visual words. Several researchers (Li et al., 2010; Yao et al., 2012; Mesnil et al., 2015) apply a large number of object detectors, and reason about room category based on the occurrences of the detected objects. In (Juneja et al., 2013), authors learn distinctive region detectors, which are in turn used to construct a holistic bag of parts descriptor, while (Parizi et al., 2015) jointly learn a set of prototype regions and image classifiers by random sampling of prototype regions and form a holistic image representation by concatenation of the region responses. Several researchers apply spatial pyramid histograms (Sadeghi et al., 2012; Singh et al., 2012), or use them in combination with object (Sadovnik et al., 2012) or region (Doersch et al., 2013) detectors. A great boost in performance has been recently achieved by the application of the convolutional neural networks (CNN) (Zhou et al., 2014). To further improve the performance, CNN features have been used in combination with additional features (Zuo et al., 2014; Gong et al., 2014), or have been extracted from various image patches (Jie et al., 2015; Dixit et al., 2015). To exploit the rich structure that is present in images, part-based models have been considered. In (Parizi et al., 2012) reconfigurable models have been proposed, which split the images into pre-defined grid and classify each cell into a predefined class. Only a few parts per image with highly constrained position were considered by (Quattoni et al., 2009), (Pandey et al., 2011) adapted the deformable parts model (DPM) in combination with a greedy sliding window search, while (Mandeljc et al., 2016) employ region proposals in combination with CNN features.

A few approaches that combine laser range data with vision have also been presented. In (Mozos et al., 2007), objects extracted from vision observations are represented as features that are appended to a laser-based feature set, which is then categorized using the AdaBoost (Freund et al., 1997) classifier. Their experiments have been performed in an office environment and were constrained to a single building. Range data was also fused with vision in (Shi et al., 2013), who combined a set of features extracted from laser scans with the Object Bank (Li et al., 2010) image-based features. Their approach performs simultaneous classification of places and objects based on CRF (Sutton et al., 2012), while their experimentation was also constrained to an office environment of a single building. Laser-vision fusion was also presented in (Pronobis et al., 2010). They apply features proposed by (Mozos et al., 2005) for the laser channel and SIFTs (Lowe, 1999) for the vision channel, and perform the fusion with a discriminative accumulation scheme that is based on SVM (Cortes et al., 1995). However, their approach performs classification of previously visited places, not room categorization.

As part of our approach we propose a new method for discriminative parts selection, exemplars, for room categorization. The idea of associating a new instance with a previously observed entity can be traced all the way back to the British empiricists. The paradigm has spread to various fields of research, like exemplar theory in cognitive psychology (Nosofsky et al., 2011), case-based reasoning (Kolodner, 1987) and prototype-based representations (Sowa, 1984) in artificial intelligence, instance-based methods in machine learning (Kolodner, 1983), data-driven transfer in graphics (Shrivastava et al., 2011), applications in computer vision (Malisiewicz et al., 2011), etc. Perhaps one of the most notable branches of the problem, which has recently been subject to extensive research in the field of pattern recognition, entitles sparse coding methods (Wright et al., 2010; Jiang et al., 2013) that model data vectors as sparse linear combinations of basic elements. A so-called sparse dictionary of a set of basic elements, also named atoms, is learned from the data. Another powerful approach for exemplar selection is the Affinity propagation (Frey et al., 2007) that applies message-passing for selecting exemplars that separate the dataset into clusters based on a prescribed affinity matrix. Linear discriminant analysis (Fisher, 1936) finds a linear combination of features, which characterizes two or more classes of elements. Multiple instance learning methods (Amores, 2013) learn a classifier to discriminate between bags of multiple instances corresponding to various classes. Summarization approaches (Tschiatschek et al., 2014) are designed to extract information from data that is both minimal and most important in the considered context. In contrast to these methods, our approach is designed for online formation of dictionary of exemplars for cross-category discrimination based solely on the defined affinity across the elements in the training set.

3 Modeling laser-based spatial parts

This section presents a novel spatial model for room categorization that is based on 2D laser range observations. We propose a formulation of mid-level parts that are built on top of a low-level part-based representation (Uršič et al., 2013). Section 3.1 overviews the low-level parts and points out their limitations, whereas Section 3.2 introduces mid-level parts that enable access to the additional hidden potential of the low-level model.

3.1 The low-level parts

The notion of parts was introduced into range-data-based spatial modeling in our previous work (Uršič et al., 2012; Uršič et al., 2013). We proposed a so-called Spatial Hierarchy of Parts (sHoP) model, which is learned from observations and is composed of three layers of parts (Figure 2). The representation is learned in an unsupervised fashion by promoting local compositions of basic shapes that frequently occur in the range measurements. The lowest layer, layer 1, contains a fixed dictionary of eighteen line fragments at different orientations (Figure 2a). On the succeeding layers, parts are represented as compositions of previous layer parts, are rotationally invariant, and are increasing in size and complexity with each following layer. Each layer stores only the parts that are statistically significant, i.e., the ones that were observed most frequently in the input laser scans. Each part is stored only in the most frequently observed orientation. If two compositions vary in structure to some allowed extent, they are considered to represent the same part. Therefore, small flexibility of part structure is allowed, and we say that such parts correspond to the same *part type*. Since line fragments on the first layer are relatively small, their compositions are well suited to model various shapes of the environment, even round ones. Parts in these layers are common to all room categories, i.e., are category-independent, which induces good scalability with respect to the number of modeled categories.

We applied the co-occurrence principle from (Uršič et al., 2013) for learning parts in layers higher than third. It turns out that such approach leads to a poorly constructed dictionary. The first issue is that local scans form a too small receptive field to account for the size of the layer 4 parts. This problem can be overcome by merging consecutive partial scans into a wider view, a local map, of the environment using SLAM (Dissanayake et al., 2001). Obtained maps consisting of 3rd layer parts are suitable for learning an arbitrary number of additional layers. However, it turns out that the training data at this scale offers a poor combinatorial evidence of co-occurrence leading to overly instance-specific parts that do not generalize well to other instances within individual room categories (see Section 7.1.2). A more flexible model structure is required to account for the high diversity of real-world instances.

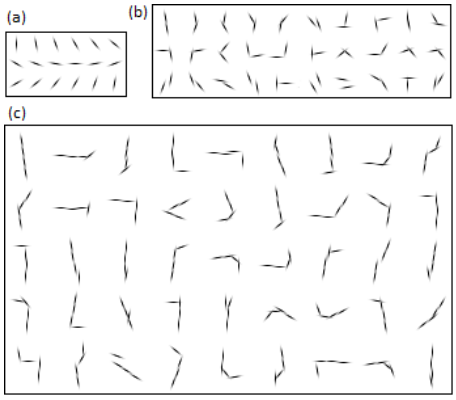

Figure 2. A hierarchy of low-level spatial parts. (a) Orientation-specific layer 1 parts. (b) A subset of layer 2 parts. (c) A subset of layer 3 parts. Note that the size of a layer 1 part corresponds only to a few pixels when scaled to the dimensions of Figure 1c.

3.2 The new mid-level parts

To overcome the limitations of low-level part-based representations (Uršič et al., 2013), we propose a new encoding of compositions by so-called local-map visibility graphs (LMVG), which represent mid-level parts that are specific for each room category. The proposed parts are insensitive to minor room differences, are rotationally invariant, but are not invariant to room size. Nevertheless, the laser-based categorization model (Section 6.1) is room-size independent, since, unlike holistic methods, our approach performs recognition based on the parts detected in the room. Therefore, even recognition of rooms with extreme sizes is not problematic, as long as individual parts of the room resemble its category well.

A single LMVG creation pipeline is presented in Figure 3: A mobile robot travels a short path in the environment (Figure 3a), from which it obtains a set of partial views (laser scans) using the laser-range finder (Figure 3b). From each laser-scan, low-level parts are extracted up to the 3rd layer using sHoP inference (Uršič et al., 2013), as shown in Figure 3c. Low-level parts from consecutive views serve as features that are mapped into a unified local map (Figure 3d) by a SLAM algorithm (Dissanayake et al., 2001). Finally, the so called visibility-graph is created upon the local map (Figure 3e).

The LMVG is a fully-connected graph built by extracting part characteristics and pair-wise part-to-part relations from the low-level local map representation. Each inferred low-level part in the map represents a graph node. Each node is assigned a real value η, that measures the curvature of the associated part. The curvature measure is calculated from part geometry by linear transformations of angles between short line segments that constitute a part. Each graph connection is represented by a triplet (D, α, v). Here, D is the distance between the two nodes forming a connection. Relative orientation of parts corresponding to those nodes is encoded by $\alpha \in [0, \pi]$, which is calculated as an angle between normals on the corresponding parts, facing towards interior of the room. Direction of room interior is obtained at laser scan acquisition step, and is calculated from relative position between the robot and observed low-level part. Finally, a binary variable v defines the mutual *visibility* of the two nodes. We say that a connection is visible if part corresponding to one node is visible from the part corresponding to the other node. Formally, the visibility property is determined by examining sub-part compositions of both parts. A connection is declared visible if at least a single line segment exists, for which the following conditions are met: (i) the segment connects some layer-1 sub-part of

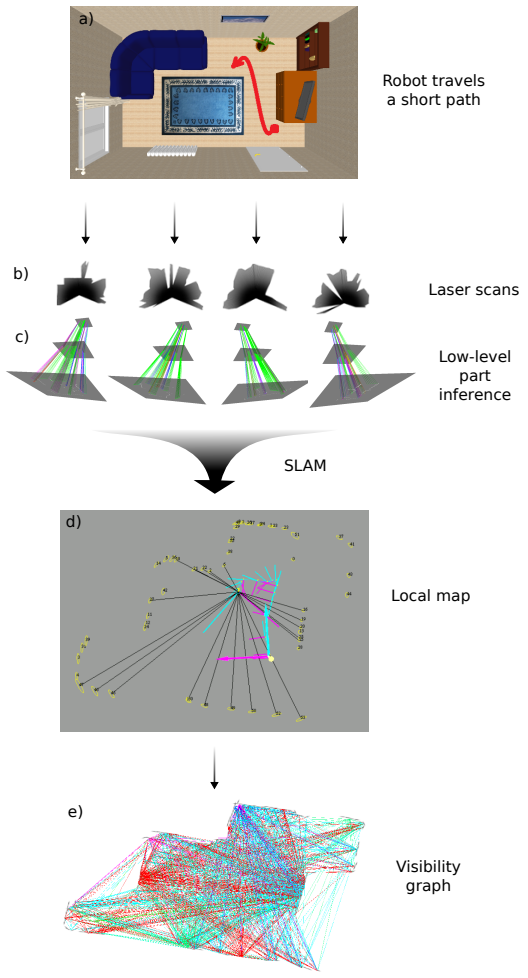

Figure 3. LMVG creation pipeline. (a) Schematic view of an example room, in which a short path traversed by the robot is marked with red. The scheme is shown only for visualisation, data used in our work was obtained in real-world environments. (b) Schematic view of a few example scans acquired along the short-path. (c) Low-level parts are extracted from each partial view (from each range-scan). (d) All partial views merged into a local map by applying SLAM. This representation, as well as the following one, was derived from real-world data. (e) Visibility-graph is extracted from the map.

the first part with some layer-1 sub-part of the other part, (ii) it lies on the interior sides of both parts, (iii) and it does not intersect with any other detected part (see Figure 4). The regions of the map that are covered by the visible connections represent the interior surface of the room that is traversable by the robot.

3.2.1 Comparing LMVGs. An affinity measure among LMVGs is needed to operate with the newly defined mid-level parts. LMVGs are compared to each other through graph matching by measuring the similarities between nodes and connections of the compared graphs. Such definition of graph-affinities imposes a loose formulation of the LMVG structure. Minor differences between the within-category instances do not disturb the recognition, since graph comparisons are not sensitive to small disturbances of nodes and connections.

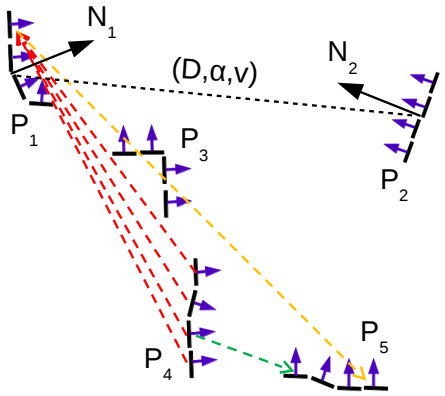

Figure 4. The visibility-graph characteristics. A visible connection ($v = 1$) exists between P_1 and P_2 low-level layer 3 parts. A connection also holds the information about the distance D between P_1 and P_2, and an angle α between N_1 and N_2. Other connections are not shown for clarity. For example, there is a visible connection also between P_4 and P_5, since the first layer 1 sub-part of P_5 is seen from the third layer 1 sub-part of P_4 (marked with a green arrow). On the other hand, there is a non-visible connection between P_1 and P_4, since sub-parts of P_1 face towards the outer (non-interior) side of P_4, (marked with red arrows), whereas a non-visible connection also exists between P_1 and P_5, because P_3 is blocking the view (marked with a yellow arrow).

An affinity measure is based on calculations using a pair of graph-similarity functions. The first function, designed to measure the similarity between nodes of two graphs, is defined as

$$f_1(\eta_1, \eta_2) = e^{-\frac{1}{2}\frac{(\eta_1-\eta_2)^2}{\sigma_\eta^2}}, \quad (1)$$

where η_1 and η_2 are the curvatures of the corresponding nodes, and σ_η is a parameter that determines the extent of the node-wise structural flexibility of LMVGs. High values of f_1 correspond to pairs of nodes with similar curvature. The second graph-similarity function is designed to measure the similarity between connections of two graphs and is defined as

$$f_2(\gamma_1, \gamma_2) = f_2((D_1, \alpha_1, v_1), (D_2, \alpha_2, v_2)) = \quad (2)$$

$$(1 - |v_1 - v_2|)(e^{-\frac{1}{2}\frac{(D_1-D_2)^2}{(\max\{D_1,D_2\}\sigma_D)^2}} + e^{-\frac{1}{2}\frac{(\alpha_1-\alpha_2)^2}{\sigma_\alpha^2}}),$$

where $\gamma_1 = (D_1, \alpha_1, v_1)$ and $\gamma_2 = (D_2, \alpha_2, v_2)$ represent the two considered connections, while σ_D and σ_α are the parameters that control the extent of the connection-wise structural flexibility of LMVGs. High values of f_2 correspond to pairs of similar connections. The first term in (2) states that similar connections should have same visibilities, while the second term allows for some discrepancies between D and α values.

LMVG G_1 is matched to LMVG G_2 by finding the cluster C of assignments $a = (i_1, i_2)$, denoting that a node i_1 of G_1 is matched to node i_2 of G_2, such that the inter-cluster score $\Theta(C) = \sum_{a,b\in C} M_{a,b}$ is maximized. Here, $\mathbf{M}$ is the adjacency matrix of potential node-to-node and connection-to-connection assignments, calculated using our graph-similarity functions:

$$M_{a,b} = \quad (3)$$

$$M_{(i_1,i_2),(j_1,j_2)} = \begin{cases} f_1(\eta_1,\eta_2) & \text{for } i_1 = j_1, i_2 = j_2\ , \\ f_2(\gamma_1,\gamma_2) & \text{for } i_1 \neq j_1, i_2 \neq j_2\ , \\ \rho(a,b) & \text{otherwise,} \end{cases}$$

where η_1 represents the curvature of node i_1 from G_1 and η_2 denotes the curvature of node i_2 from G_2, γ_1 refers to a connection between i_1 and j_1 in G_1, while γ_2 refers to a connection between i_2 and j_2 in G_2. Notation $\rho(a,b)$ refers to a potential assignment of a single node from G_1 to two nodes from G_2, or vice versa. Let us consider only the former case, since latter is analogous. Such an assignment is allowed if i_2 and j_2 are positioned close to each other, and if G_2 contains more nodes than G_1. If the conditions are met, the expression equals $f_1(\eta_1,\bar{\eta})$, where $\bar{\eta}$ denotes the combined curvature corresponding to i_2 and j_2, while it equals 0 otherwise. Any clustering of assignments C can be represented by an indicator vector $\mathbf{x}$, such that $x_a = 1$ if $a \in C$ and 0 otherwise. The total inter-cluster score can be rewritten as $\Theta(C) = \Theta(x) = \frac{1}{2}\mathbf{x}^T\mathbf{M}\mathbf{x}$, and the optimal matching $\mathbf{x}^*$ is the binary vector that maximizes the score

$$\mathbf{x}^* = \underset{x}{\operatorname{argmax}}(\mathbf{x}^T\mathbf{M}\mathbf{x}). \tag{4}$$

Equation (4) is a quadratic assignment problem (QAP), which is NP-hard. To find the approximate solution we use spectral matching (Leordeanu et al., 2005) augmented with an integer-projected fixed-point method (Leordeanu et al., 2009). The normalized value of the solution is defined as an affinity measure

$$\psi(G_1,G_2) = 2\frac{\Theta(x^*)}{|G_1||G_2|} \in [0,1], \tag{5}$$

where $|G_1|$ and $|G_2|$ denote the number of nodes in G_1 and G_2 respectively. If graphs are similar then $\psi(G_1,G_2)$ is close to 1, and it is close to 0 otherwise. Figure 5 displays similarity values between a few pairs of artificially generated example graphs.

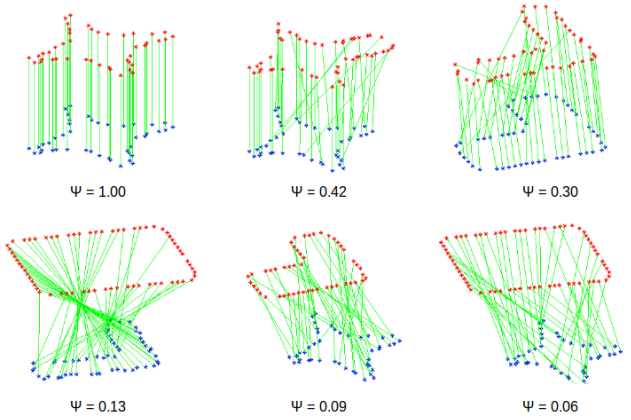

Figure 5. Similarities between artificially generated LMVGs. Maps of small rooms with simple generic shapes have been created in simulation, from which visibility graphs were extracted. Examples demonstrate the obtained optimal matching between nodes of a pair of graphs with a corresponding similarity measure. Graph connections and low-level parts associated with graph nodes are omitted for clarity.

4 Modeling vision-based spatial parts

To follow the conceptual consistency with the laser-based model, we adapt the mid-level part-based image representation for the vision modality as well. We employ a robust model that we recently presented in (Mandeljc et al., 2016). The proposed part definition enables invariance to size and rotation (Mandeljc et al., 2016). Vision-based part extraction is performed in two stages. The corresponding pipeline is shown in Figure 6.

In the first stage, object-type-agnostic region proposal algorithm is applied to images. In particular, the gestalt-principle-inspired selective search (Uijlings et al., 2013) is employed, which yields a set of salient regions that likely contain highly informative structures. The method combines the strengths of an exhaustive search and segmentation, since it aims to capture all possible object locations, while using a low-level image structure to guide the sampling process. A hierarchical grouping of segmented images is applied at various scales (Figure 6b) to propose regions that may correspond to individual objects, visually-coherent compositions of objects, or simply segments of scene with coherent characteristics like color or texture (Figure 6c).

In the second stage, each extracted region is encoded in a feature space (Figure 6d). A convolutional neural network (CNN) is well suited for such purpose, since many recent studies report exceptional performance of CNNs when applied as generic feature extractors (Razavian et al., 2014). We employ a deep network that was pre-trained by (Zhou

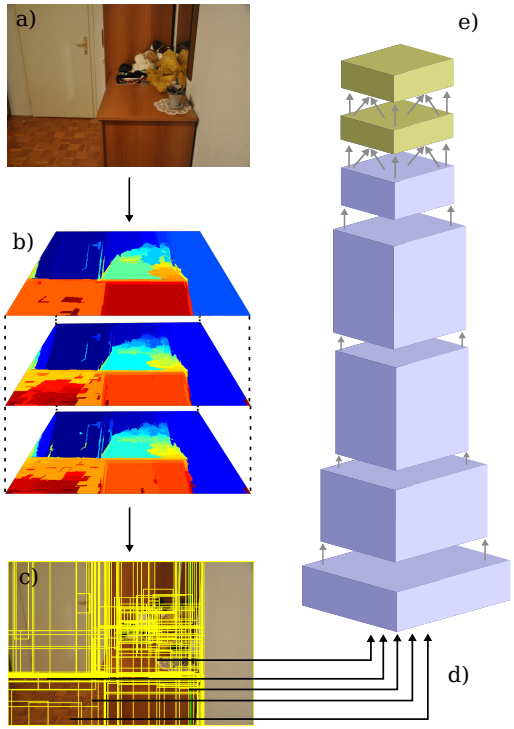

Figure 6. Vision-based part extraction pipeline. (a) Input RGB image. (b) Selective search (Uijlings et al., 2013) performs hierarchical grouping of segmented images at different scales. (c) Proposed image regions. (d) Regions are fed into a CNN (Zhou et al., 2014) that performs feature extraction. (e) The vision-based part is represented with the output of the last fully connected network layer for the corresponding input region.

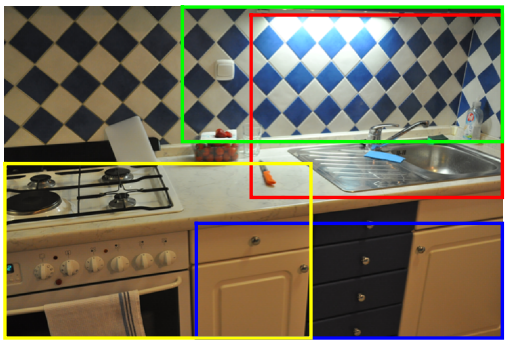

Figure 7. A few regions corresponding to mid-level visual parts extracted from an image that was acquired in a kitchen.

et al., 2014). The regions generated by the region proposal algorithm are forward-propagated through five convolutional and two fully-connected layers. From each input region a single vision-based part is created, which is represented as a 4096-dimensional feature vector. The feature vector is the response from the seventh, fully-connected (FC7), layer of the network (Figure 6e). For additional details on the network architecture and training procedure, we refer the reader to (Zhou et al., 2014). A few regions corresponding to extracted parts are shown on an example image in Figure 7.

5 Discriminative dictionary learning

From a large number of parts acquired from the training data, we construct a pair of compact dictionaries corresponding to each of the two modalities, laser-based and vision-based. Each dictionary consists of a set of discriminative parts, i.e., exemplars, that are specific for each of the modeled categories. Since laser-based parts are represented as LMVGs, standard vector-based dictionary learning approaches, like sparse coding (Wright et al., 2010), for example, are not applicable. A multi-category set of exemplar parts needs to be learned based on pair-wise affinities between LMVGs. For this purpose we present here the Multi-Category Affinity-Based Exemplars search (MCABE) algorithm, which is general and is not constrained only to the LMVG-based representation. The method is applicable to the vision modality as well. Therefore, the vision-based dictionary is learned using the same algorithm, which demonstrates the generality of the approach. Moreover, MCABE allows online dictionary creation, which makes it memory efficient. The proposed method consists of two key components, i.e., the MCABE-core algorithm, presented in Section 5.1, and the sequential mini-batch wrapper, introduced in Section 5.2.

5.1 The MCABE-core algorithm

Suppose we have a large set of elements corresponding to several categories and want to find a small subset of elements that maximally discriminate these categories. In particular, we seek a set of exemplars that well generalize within their own category and discriminate against the elements from other categories according to a defined affinity measure. The problem is formalized as follows. Assume we have a set S of n elements and k categories, such that $n = n_1 + n_2 + \cdots + n_k$ with n_i being the number of elements in the i-th category. We are also given an affinity measure ψ, which determines the similarity between pairs of elements. Without loss of generality, we can assume $\psi(x, y) \in [0, 1]$ for each pair of elements x and y, with $\psi(x, y) = 1$ corresponding to maximum similarity. The task is to find a subset of elements $E \subset S$ of size $m \ll n$, from which m_1 of them belong to category 1, m_2 of them belong to category 2, ..., and m_k of them belong to category k, where $m_1 + m_2 + ... + m_k = m$, and $m_i > 0$, for $i \in \{1, 2, ..., k\}$, subject to an optimization criteria function $F(S, E, \psi)$, which determines how well the set of exemplars E represents the original set S, i.e.,

$$E^* = \underset{E}{\operatorname{argmax}}\, F(S, E, \psi), \tag{6}$$

and with respect to keeping $|E|$ small, which induces sparsity of the solution. The cost function is defined as

$$F(S, E, \psi) = \sum_{i=1}^{k} \sum_{x \in S_i \setminus E_i} (\max_{y \in E_i} \psi(x, y) - \max_{z \in E \setminus E_i} \psi(x, z)), \tag{7}$$

where S_i corresponds to the set of all elements from category i and E_i corresponds to the set of exemplars belonging to the category i. Therefore, $E = E_1 \cup E_2 \cup ... \cup E_k$ and $S = S_1 \cup S_2 \cup ... \cup S_k$.

Equations (6) and (7) represent a combinatorial optimization problem. Since finding a solution by brute force is intractable for large n, an iterative algorithm has been designed, MCABE-core, for finding an approximate solution.

The MCABE-core algorithm proceeds as follows. All elements in the set are initialized as exemplars of their categories. This set is then gradually reduced, while maximizing discrimination. The exemplar selection process is performed by a two-level iteration procedure, the inner and the outer loop. A single step of the outer iteration refers to refinement of the exemplars for all categories, one category at a time. The inner iteration refers to refinement of the exemplars of a single category with respect to a fixed set of exemplars from all other categories. The algorithm is stopped at an outer iteration step that refers to exemplars for which a local maximum of the cost function is obtained. To refine exemplars of a single considered category i by the inner loop, previously calculated exemplars, E_i, are discarded, while the new ones are being calculated with respect to current exemplars of all other categories. Let $1 < i < k$, then, optimization is performed with respect to exemplars of categories $1, 2, \ldots, i-1$ determined in the current outer iteration step, and with respect to exemplars of categories $i+1, \ldots, k-1, k$ determined in the previous outer iteration step.

Exemplars for the i-th category are determined by arranging all elements corresponding to this category into clusters. A single cluster is composed of an exemplar and all of the elements that are represented by this exemplar. Clusters consist of elements that are similar to each other, whereas its exemplar is the one that is least similar to all the exemplars from all other categories. Since exemplars of other categories also represent well their corresponding non-exemplars, good exemplar discriminativity is induced against all elements of other categories. This property is encoded in the cost function summation terms of (7), which increase

the value of the cost function in presence of large within-category similarities and decrease it in presence of large across-category similarities.

Clusters are being formed by sequential element validation. Before the inner iteration is applied, exemplars for the i-th category are re-initialized by assessing all the elements of the category, one after another. The element u_1 is validated by first calculating the summation term of (7) and a most similar element, u_2, from within the same category is determined for which the summation term of (7) is also obtained. Then, different actions are taken depending on the current state of u_1 and u_2. Element u_1, u_2, or both might have already been observed in the current re-initialization process. If neither was observed they form a new cluster, in which exemplar becomes the one with the highest score. On the other hand, each observed element has already been assigned to some cluster, or it might have even been assigned as the cluster's exemplar. A set of rules covering all combinations of these states define the actions of adding and removing exemplars, by preferring the ones with the higher scores. In the inner loop that follows, all elements of S_i are randomly permuted at the beginning of each step. Then, according to the given permutation, elements are rearranged into clusters with corresponding exemplars sequentially. Again, a set of state-dependent action-rules, similar to the ones of the re-initialization step, determine whether an element is added or removed from the existing set of exemplars. In this way the value of the cost function (7) is being maximized. The inner iteration is stopped when a set of exemplars E_i stabilizes, or when oscillation is detected. The approach is summarized in Algorithm 1. Note that sparsity constrains are directly incorporated in the above-described set of state-dependent action-rules.

Algorithm 1 The MCABE-core algorithm

procedure MCABE-CORE($S_1, S_2, ..., S_k, \psi$)
  $E_i \leftarrow S_i \; \forall i \in \{1, 2, ..., k\}$ ▷ Exemplars initialized to all elements
  $E_i^{old} \leftarrow \varnothing \; \forall i \in \{1, 2, ..., k\}$ ▷ Exemplars from the previous step
  ▷ The outer iteration loop:
  while $F(S, E_1, ..., E_k, \psi) > F(S, E_1^{old}, ..., E_k^{old}, \psi)$ **do**
    $E_i^{old} \leftarrow E_i \; \forall i \in \{1, 2, ..., k\}$
    for $j = 1, 2, ..., k$ **do** ▷ For each category
      $\Omega \leftarrow \varnothing$ ▷ A set of observed elements
      $E_j \leftarrow \varnothing$
      ▷ Estimated cost of exemplar x in the re-initialization step:
      $\hat{E_x} \leftarrow \varnothing \; \forall x \in S_j$
      ▷ The re-initialization step:
      for each $x \in S_j$ **do**
        $y \leftarrow \underset{y' \in S_j \setminus \{x\}}{\operatorname{argmax}} \; \psi(x, y')$
        $\tilde{f}_x \leftarrow \psi(x, y) - \max_{z \in E \setminus E_j} \psi(x, z)$
        $\tilde{f}_y \leftarrow \max_{y' \in S_j \setminus \{y\}} \psi(y, y') - \max_{z \in E \setminus E_j} \psi(y, z)$
        ▷ State-dependent action-rules:
        if $x \notin \Omega \;\&\; y \notin \Omega$ **then**
          If $\tilde{f}_x < \tilde{f}_y$ then $E_j = E_j \cup \{y\}$ and $\hat{E_y} = \tilde{f}_y$,
          otherwise $E_j = E_j \cup \{x\}$ and $\hat{E_x} = \tilde{f}_x$
        else if $x \notin \Omega \;\&\; y \in \Omega$ **then**
          $e \leftarrow \underset{e \in E_j}{\operatorname{argmax}} \; \psi(y, e)$
          If $\tilde{f}_x > \hat{E_e}$ then $E_j = (E_j \setminus e) \cup \{x\}$ and $\hat{E_x} = \tilde{f}_x$
        else if $x \in \Omega \;\&\; y \notin \Omega$ **then**
          If $x \in E_j$ and $\hat{E_x} < \tilde{f}_y$ then $E_j = (E_j \setminus x) \cup \{y\}$
          and $\hat{E_y} = \tilde{f}_y$, otherwise
          if $x \notin E_j$ and $e \leftarrow \underset{e \in E_j}{\operatorname{argmax}} \; \psi(x, e)$ and $\hat{E_e} < \tilde{f}_y$
          then $E_j = E_j \cup \{y\}$ and $\hat{E_y} = \tilde{f}_y$
        else if $x \in E_j$ and $y \in E_j$ **then**
          If $\hat{E_x} < \hat{E_y}$ then $E_j = E_j \setminus \{x\}$
          otherwise $E_j = E_j \setminus \{y\}$
        end
        $\Omega \leftarrow \Omega \cup \{x, y\}$
      end
      ▷ The inner iteration loop:
      while E_j changes & not oscillation **do**
        $\hat{S_j} \leftarrow$ Random permutation of S_j
        for each $x \in \hat{S_j}$ **do**
          ▷ State-dependent action-rules:
          if If $x \in E_j$ **then**
            ▷ Elements in a cluster corresponding to x (without x)
            $C_x \leftarrow Cluster(x)$
            if $C_x = \varnothing$ **then**
              $\tilde{f}_x \leftarrow \max_{y \in S_j \setminus \{x\}} \psi(x, y) - \ldots$
              $\ldots - \max_{z \in E \setminus E_j} \psi(x, z)$
              If $\tilde{f}_x > 0$ then $E_j = E_j \setminus \{x\}$, break
            else
              $f_c \leftarrow \psi(c, x) - \max_{z \in E \setminus E_j} \psi(c, z) \; \forall c \in C_x$
              if $\exists c \in C_x \ni: f_c \leq 0$ **then**
                $E_j = E_j \cup \{\underset{c \in C_x}{\operatorname{argmin}} f_c\}$, break
              else
                $f'_c \leftarrow \max_{y \in E_j \setminus \{x\}} \psi(c, y) - \ldots$
                $\ldots - \max_{z \in E \setminus E_j} \psi(c, z) \; \forall c \in C_x \cup \{x\}$
                If $f'_c > 0 \; \forall c \in C_x \cup \{x\}$
                then $E_j = E_j \setminus \{x\}$, break
              end
            end
          else
            $f_x \leftarrow \max_{y \in E_j} \psi(x, y) - \max_{z \in E \setminus E_j} \psi(x, z)$
            If $f_x < 0$ then $E_j = E_j \cup \{x\}$, break
          end
        end
      end
    end
  end
  return $E_1^{old}, E_2^{old}, ..., E_k^{old}$

5.2 The sequential mini-batch wrapper

In practice, dictionary learning often requires processing of huge amount of data. When exemplars are being searched for in a large dataset, a problem may become intractable when no additional approximations are put into learning. In such cases, an extremely large computational power may be needed to calculate the affinities between all possible pairs of elements. On the other hand, not all data may be fully observable in advance. Additional data chunks may arrive in batches, which should also be considered and should therefore be used to refine the model. Both scenarios call for sequential dictionary learning. To cope with the considered difficulties, the above introduced MCABE-core procedure is wrapped in a sequential mini-batch technique that processes input data in batches, merges the results, and then repeats the cycle.

The proposed method is summarized in Figure 8 in the context of a large dataset scenario. The application to sequential dictionary learning scenario is similar and will be discussed later. In the initialization, a set of elements for each category $C_1, C_2, \ldots, C_k$ is divided into n_1 subsets. Then, for each $i \in \{1, \ldots, n_1\}$, the i-th subset of each category is arranged into the i-th batch (Figure 8a). Affinities $\psi(x, y)$ between all pairs of elements x and y from each batch are in turn processed by the MCABE-core algorithm to obtain the exemplars of the first epoch (Figure 8b). The obtained exemplars from all batches are stacked together according to their corresponding category and then divided into n_2 batches by following the same principle as in initialization step (Figure 8c). Exemplars of the second epoch are then obtained using the MCABE-core algorithm for each new batch (Figure 8d). The cycle is then repeated until only

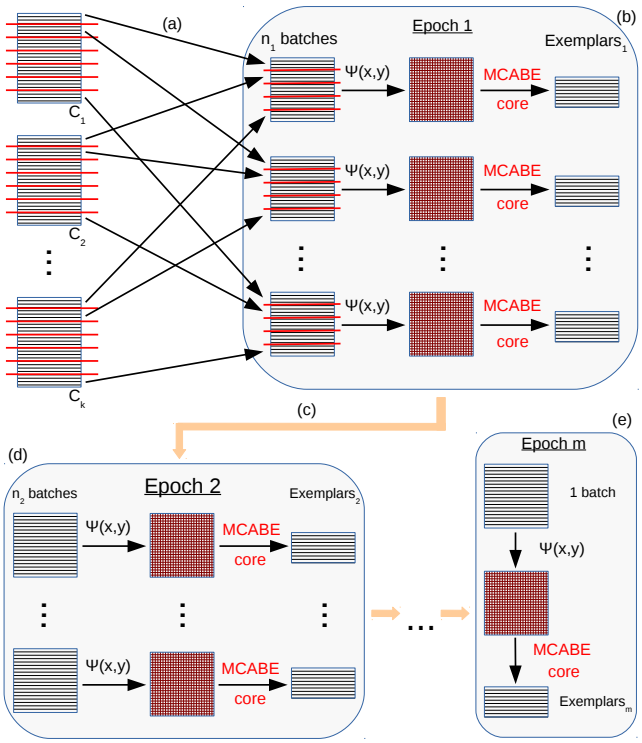

Figure 8. The MCABE algorithm: sequential mini-batch wrapper demonstrated in the context of a large dataset scenario. See text for detailed figure description.

a single batch has been processed for the entire epoch (Figure 8e).

Similar approach can be applied for sequential learning of category exemplars. The proposed method is suitable for the situation in which a total number of categories is fixed, whereas new data arrives in batches that contain a few samples for each of the categories. Exemplars for the new batch are obtained using the MCABE-core algorithm and then stacked together with previously obtained exemplars. When their number exceeds the required model compactness, a second epoch is executed, etc.

6 Laser-vision fusion for room categorization

The following scenario is considered for room categorization. In training phase, the robot extracts a large set of laser-based parts (Section 3) from the laser-range measurements and a large set of vision-based parts (Section 4) from a set of images, both acquired in various rooms corresponding to the modeled categories. MCABE is applied for dictionary selection (Section 5) in each modality. An affinity measure that is applied for comparison of laser-based parts was presented in Section 3.2.1, whereas a negative euclidean distance is used as the affinity among the vision-based parts encoded by the last fully-connected-layer CNN features. At test time, the robot enters a new room that was not in the training set. It starts exploring the room by making a few short tours and constructing a single LMVG per tour. Meanwhile, a set of images is acquired along the way, using a visual sensor (Figure 1). Observed laser-based parts (LMVGs) are categorized with respect to the laser-based dictionary (Section 6.1), parts extracted from the acquired images are categorized with respect to the vision-based dictionary (Section 6.2), whereas the combined recognition is obtained by joining the two single-modal predictions (Section 6.3).

6.1 Laser-only room categorization

Based on the affinity between the measured LMVGs and parts in the learned dictionary, the room is categorized by a voting scheme as follows. A category-specific dictionary contains k sets of exemplar parts, i.e., $\{E_1, E_2, ..., E_k\}$ with E_i being a set of exemplars, LMVGs, for i-th category. Let G be one of the LMVGs obtained in the room whose category needs to be determined. The matching score for the i-th category is calculated as the mean of the affinities, μ_i, between G and the exemplars in E_i. The probability that G belongs to category C_i is calculated by normalization

$$P(G \in C_i) = \frac{\mu_i}{\sum_{j=1}^{k} \mu_j}. \quad (8)$$

A room category is determined by following a majority voting scheme applied to all parts obtained in a room, with each part casting a vote to the category C_i proportionally to the probability $P(G \in C_i)$.

6.2 Vision-only room categorization

The vision-only room categorization follows the same principles as presented in the laser-only domain. However, since measure range of the euclidean affinities used in the vision modality equals $(-\infty, 0]$, a few simple adaptations are made to account for the negative values. In the learning stage, a summation instead of subtraction is performed in the MCABE cost function (7). To categorize a single image, image part affinities to most similar category exemplars are first fed to an exponential function to obtain the probabilities (8), which are in turn applied to the majority voting scheme. To obtain the room-level categorization, the same majority voting scheme is applied across all images acquired in a room.

6.3 Laser-vision fusion

Posterior probabilities across the considered categories are first proposed by each of the modalities separately, and then, they are joined to produce the final prediction. We apply a common approach for combining probability distributions, i.e., the linear opinion pool of experts (Clemen et al., 1999). The combined probability distribution $P(C)$ for an unknown room category C is defined as

$$P(C) = \sum_{i=1}^{n} w_i P_i(C), \quad (9)$$

where n is the number of experts, i.e., the number of considered models to be merged (two in our case), $P_i(C)$ represents the i-th distribution, and w_i are the non-negative weights, which sum to one. The chosen category in the fusion scheme then corresponds to C with maximum a posteriori (MAP) probability.

7 Experimental analysis

We report here a thorough experimental evaluation of the proposed methods. Section 7.1 presents results of

the experiments with our laser-based approach for room categorization, along with detailed performance analysis of the new mid-level parts, LMVGs, and the MCABE algorithm. Experimentation with the vision-based parts is discussed in Section 7.2. Finally, Section 7.3 provides a comparison between laser-based, vision-based, and laser-vision-fusion-based approaches.

7.1 The proposed laser-based model

7.1.1 The dataset. The proposed range-data-based room categorization approach was experimentally evaluated on the DR dataset (Uršič et al., 2013), which we have augmented by adding two room categories. The dataset was obtained using a Hokuyo URG laser range-finder mounted on a Pioneer P3-DX robot (Figure 1a) while moving across rooms in several real apartments. See (Uršič et al., 2013) for further details on the robotic system used. The original dataset (Uršič et al., 2013) contains 2D range and odometry measurements acquired in 21 living rooms, 6 corridors, 35 bathrooms and 28 bedrooms. In present work two additional categories obtained by the same acquisition system have been added to the dataset. The new data has been gathered from 21 kitchens and 12 toilets, thus creating an even more challenging room categorization dataset.

7.1.2 Inadequacy of the low-level parts. In Section 3.1 we discussed that low-level part learning paradigm from (Uršič et al., 2013) is inappropriate for use with fourth-layer parts since the descriptors become overly instance specific. In this section we experimentally support our discussion by comparing the performance of (Uršič et al., 2013) applied to layer 4 parts as opposed to the original layer 3 application from (Uršič et al., 2013). We consider the original four categories from the DR dataset (living room, corridor, bathroom, and bedroom). Since layer 4 parts are too large to be extracted directly from individual laser scans, a SLAM algorithm (Dissanayake et al., 2001) was applied to create larger local maps. This allowed satisfactory extraction of the fourth layer parts.

Following (Uršič et al., 2013) several local maps were extracted from a single room, a HoC descriptor was constructed from each local map, and each room was represented by a concatenation of the average and standard deviation of all the descriptors obtained in the room. Categorization was performed by a linear SVM trained with 4-fold cross-validation to obtain the C parameter. The experiment was repeated 30 times with 20% of the rooms randomly chosen as the test rooms in each trial. Results are summarized as confusion matrices in Table 1. Note that layer 4 classification is significantly poorer than layer 3 results.

7.1.3 Parameters and model properties. The parameters of node-wise and connection-wise structural flexibilities in LMVGs (1), (2) were determined through simulation. In particular, we generated a set of artificial local maps that simulated small rooms resembling various simple shapes (a square, a rectangle, a star, an L-shaped room, etc.), from which visibility graphs were extracted. Some examples of such artificially generated LMVGs are shown in Figure 5. We then performed a greedy search to obtain the parameter values $\sigma_\eta = 0.2$, $\sigma_D = 0.2$, and $\sigma_\alpha = 0.35$, which provided maximal matching scores between pairs of visually similar

Table 1. Confusion matrices for the experiments with low-level parts learned using the co-occurrence principle from (Uršič et al., 2013). Category abbreviations: LR-living room, CO-corridor, BA-bathroom, BE-bedroom.

Layer 3	LR	CO	BA	BE
LR	**82.74**	0.00	0.08	17.18
CO	0.00	**91.60**	0.00	8.40
BA	0.20	0.29	**92.96**	6.56
BE	13.32	0.22	16.24	**70.22**
Layer 4	**LR**	**CO**	**BA**	**BE**
LR	**39.63**	0.00	25.00	35.37
CO	33.87	**0.00**	17.74	48.39
BA	6.09	0.00	**71.74**	22.17
BE	22.10	0.57	51.05	**26.29**

graphs and minimal matching scores between pairs of visually dissimilar graphs. To reduce computation time of the graph matching procedures, each graph was down-sampled to contain only a third of its original nodes.

In all of the laser-based experiments we used a bold MCABE-core implementation, i.e., a single epoch with a single batch. Application of the sequential mini-batch formulation was unnecessary because of the relatively low memory requirements of the LMVGs. The inner iteration of the MCABE-core algorithm achieved convergence usually within 30 steps. In cases the convergence was not achieved, the algorithm oscillated between two or more solutions, from which the one that maximized the optimization function was chosen. The outer iteration reached local maxima within approximately five steps, which resulted in about 60% of all LMVGs being chosen as exemplars of their categories.

Graph matching is computationally the most demanding part of the proposed method. Computation times depend on matched graph sizes (number of nodes in the graph). On a laptop with 2.3 GHz dual-core processor, the processing time ranges from 0.2 sec for a pair of smallest graphs, to 453.5 sec for the largest. Computation times are plotted with respect to the matched graph sizes in Figure 9. Majority of the graphs acquired from rooms in our dataset contain less than about 100 nodes, as shown in Figure 10. However, living room and bedroom categories contain most of the large-sized samples, therefore, it can be expected that comparing instances from these two categories is most time consuming. Further

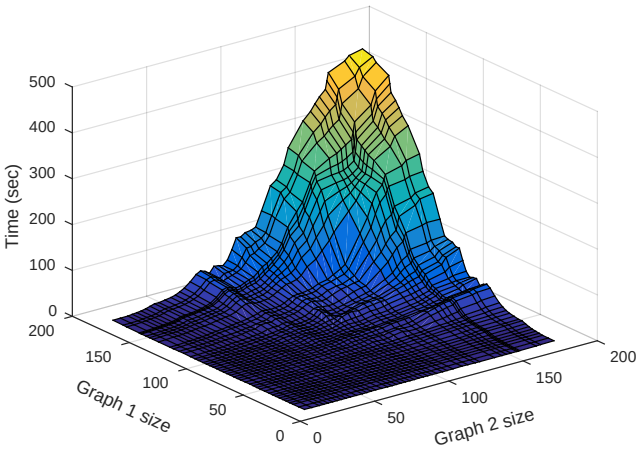

Figure 9. Computation times of the graph matching procedure shown with respect to the matched graph sizes.

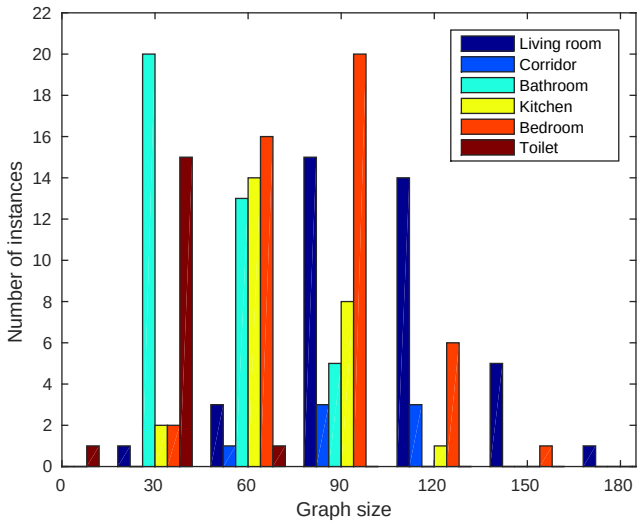

Figure 10. A histogram displaying the number of graphs per each category with respect to the graph size (number of nodes).

parallelization could be applied for performance speedup. To categorize an LMVG approximately 100 matchings were performed in this work.

7.1.4 Experimental setup. All of the following experiments were performed by applying a leave-one-out cross-validation scheme. In a single trial of each experiment a single room was chosen as a test room, while all the other rooms were used for training. There were as many trials performed as there are rooms in the dataset, so that each room has been used exactly once as a test room.

7.1.5 Short tour lengths. The length of the short-tour of local map creation step is determined by the number of consecutive laser scans used to form the map. Results of the experiments with different short-tour lengths have been evaluated with three different measures (Table 2). The mean success rate is calculated as a mean of the diagonal entries of the confusion matrix, accuracy refers to the overall percentage of correctly categorized examples, whereas standard deviation, calculated across the diagonal entries of the confusion matrix, measures how balanced the results are across all categories (lower values correspond to a better balanced performance). Optimal short-tour length is attained at 225 consecutive scans, which is the value used in the rest of our work. For lower values the obtained LMVGs are less discriminative, which causes reduced performance. On the other hand, increasing this value increases the probability of error accumulation in SLAM, which also leads to performance reduction. Usually, a single or at most three LMVGs were obtained in each room.

7.1.6 Evaluation of the MCABE. The MCABE algorithm was compared to two baseline approaches for category-specific LMVGs selection. The first is the greedy approach in which all parts obtained in training form a dictionary. This is obviously not computationally and memory efficient. In the second approach exemplars are calculated using the Affinity Propagation (AP) algorithm (Frey et al., 2007). In this case, exemplars for each category are determined by searching for most representative parts within each category, independently of the other categories, thus disregarding

Table 2. Results of the experiments with different short tour lengths in terms of a number of consecutive laser scans. Measure abbreviations: Msr - mean success rate, Acc - accuracy, Std - standard deviation.

Short tour length	Msr	Acc	Std
100	57.66	45.53	33.91
125	47.78	33.33	35.72
150	54.60	44.72	27.80
175	52.74	41.46	33.69
200	51.47	41.46	27.32
225	**61.59**	**53.66**	**23.50**
250	56.39	45.53	31.08

the across-category similarities. Results of our proposed MCABE approach are shown in comparison with the two baseline approaches in Table 3.

The proposed method, MCABE, outperforms the greedy approach in distinguishing living rooms, corridors, bathrooms, and bedrooms, while the greedy approach performs better with kitchens, and toilets. The MCABE is least accurate with bathrooms. A closer inspection shows that bathrooms are mostly confused by toilets which appears reasonable, since some of the bathrooms and toilets look quite similar from the perspective of a laser range sensor. The same miscategorization is present in the greedy approach. The fact that the non-efficient greedy method also performs quite well means that the proposed mid-level part structure, LMVG, is indeed the informative part model. Note that considering dictionary construction, the categorization accuracy and mean success rate increase when MCABE is applied, from 49.59% to 53.66%, and from 57.54% to 61.59%, respectively.

MCABE outperforms the AP (Frey et al., 2007) on living rooms, kitchens, bedrooms and toilets, performs equally well with corridors, and is outperformed with bathrooms. It can be seen that the AP significantly over-fitted the corridors category, since majority of the living room, kitchen, and bedroom examples were categorized as corridors. In the DR dataset rooms within each category are quite diverse. On the other hand, several pairs of rooms can be found that correspond to different categories and nevertheless look quite similar when viewed from the perspective of laser scans. The proposed method selects exemplars that represent their category well and that simultaneously ensure good across-category discriminativity, which enables a more balanced performance over all categories in comparison to the AP.

7.1.7 Comparison with state-of-the-art. We compared our proposed mid-level part-based approach to the approach of (Mozos et al., 2005), for which the code was provided to us by the authors, and to the recently presented low-level part-based model (Uršič et al., 2013). The approach of (Mozos et al., 2005) uses AdaBoost (Freund et al., 1997) 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 (Mozos et al., 2005) for details of the algorithm. The optimal decision list turns out to be toilet, corridor, bathroom, living room, and kitchen. In 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

Table 3. Confusion matrices for the laser-based experiments with greedy approach, Affinity Propagation - AP (Frey et al., 2007), and proposed MCABE approach. Category abbreviations: LR-living room, CO-corridor, BA-bathroom, KI-kitchen, BE-bedroom, WC-toilet.

Greedy	LR	CO	BA	KI	BE	WC
LR	**47.62**	23.81	4.76	4.76	19.05	0.00
CO	0.00	**66.67**	0.00	16.67	0.00	16.67
BA	0.00	5.71	**28.57**	14.29	2.86	48.57
KI	0.00	4.76	28.57	**52.38**	9.52	4.76
BE	3.57	14.29	17.86	14.29	**50.00**	0.00
WC	0.00	0.00	0.00	0.00	0.00	**100.00**
AP	LR	CO	BA	KI	BE	WC
LR	**23.81**	71.43	4.76	0.00	0.00	0.00
CO	0.00	**83.33**	16.67	0.00	0.00	0.00
BA	0.00	31.43	**40.00**	0.00	5.71	22.86
KI	9.52	47.62	9.52	**19.05**	9.52	4.76
BE	10.71	53.57	3.57	3.57	**28.57**	0.00
WC	0.00	0.00	16.67	0.00	0.00	**83.33**
MCABE	LR	CO	BA	KI	BE	WC
LR	**52.38**	23.81	0.00	14.29	9.52	0.00
CO	0.00	**83.33**	16.67	0.00	0.00	0.00
BA	0.00	5.71	**31.43**	8.57	17.14	37.14
KI	0.00	19.05	4.76	**42.86**	23.81	9.52
BE	7.14	14.29	3.57	7.14	**67.86**	0.00
WC	0.00	0.00	8.33	0.00	0.00	**91.67**

the room category. In (Uršič et al., 2013) Histogram of Compositions (HoC) descriptor was used to perform room categorization. Each laser-scan was represented by a HoC descriptor, which was generated using category-independent low-level 3rd layer parts of the hierarchy. A composition of an average and standard deviation of all the descriptors obtained in a room was used as an input for categorization with a support vector machine with a linear kernel.

The experimental results are summarized as confusion matrices in Table 4. The MCABE approach outperformed the boosting-based (Mozos et al., 2005) with kitchens, bedrooms, and toilets, while (Mozos et al., 2005) performed better with living rooms, corridors, and bathrooms. The boosting-based approach performed poorly with kitchens, which were mainly categorized as living rooms, whereas bedrooms were mainly categorized as bathrooms. It is evident from Table 4 that (Mozos et al., 2005) significantly over-fitted to the first three categories, which resulted in a heavily unbalanced performance. For example, 76% of kitchen examples were incorrectly categorized as one of the first three categories. Since corridors, on which (Mozos et al., 2005) achieved 100% categorization performance, represent the category containing the smallest number of examples, whereas great performance was also achieved on living rooms and bathrooms containing lots of examples, over-fitting of (Mozos et al., 2005) cannot be conditioned on category sizes. Results demonstrate that the approach performs best on categories that are positioned at the beginning of the binary classifier decision list, while performance largely decreases at its end. The reason that toilets and bathrooms are being confused with each other, even though they are at the beginning of the list, is the same as observed with MCABE approach, i.e., some of the examples look alike from the perspective of the range sensor.

The proposed approach outperformed the HoC (Uršič et al., 2013) on corridors, kitchens, bedrooms and toilets,

Table 4. Confusion matrices for the experiments with the approaches of (Mozos et al., 2005) - Moz and (Uršič et al., 2013) - HoC. Categories: LR-Living room, CO-Corridor, BA-Bathroom, KI-Kitchen, BE-Bedroom, WC-Toilet.

Moz	LR	CO	BA	KI	BE	WC
LR	**95.24**	0.00	4.76	0.00	0.00	0.00
CO	0.00	**100.00**	0.00	0.00	0.00	0.00
BA	0.00	0.00	**80.00**	0.00	2.86	17.14
KI	42.86	14.29	19.05	**14.29**	4.76	4.76
BE	25.00	3.57	35.71	3.57	**32.14**	0.00
WC	0.00	0.00	25.00	0.00	0.00	**75.00**
HoC	LR	CO	BA	KI	BE	WC
LR	**76.19**	0.00	0.00	4.76	19.05	0.00
CO	16.67	**16.67**	16.67	50.00	0.00	0.00
BA	0.00	0.00	**77.14**	8.57	8.57	5.71
KI	14.29	4.76	14.29	**28.57**	38.10	0.00
BE	10.71	0.00	10.71	14.29	**64.29**	0.00
WC	0.00	0.00	58.33	0.00	0.00	**41.67**

while HoC performed better with living rooms and bathrooms. Similarly to other techniques, HoC confused toilets with bathrooms. The most difficulties were observed at recognizing corridors, which were confused mostly with kitchens, while kitchens were mainly categorized as bedrooms. Similarly to (Mozos et al., 2005), HoC also demonstrated an unbalanced performance. For example, 66% of the kitchen examples were incorrectly categorized as one of the three other categories. Kitchens in our dataset vary significantly in shapes and sizes. When viewed from the perspective of a 2D range sensor, some of them resemble corridor-like structure, while at the other extreme, a few resemble large spaces like some of the bedrooms or living rooms. The rigid grid into which space is divided to form the HoC descriptor can not cope with such large within-category variability and across-category similarities of rooms. On the other hand, the structure of LMVGs coupled with principled exemplar selection process is designed for such purpose, which results in improved performance.

Note that the average accuracy over the categories for MCABE is 53.66%, which is lower than for competing approaches, i.e., 60,98% for (Mozos et al., 2005) and 59,35% for (Uršič et al., 2013). A close inspection reveals that the improvement in (Mozos et al., 2005) and (Uršič et al., 2013) comes from over-fitting. The average performance is boosted at the cost of very poor categorization of a few categories. The over-fitting is quantified by standard deviation of category-wise categorization accuracies, which are 23.50%, 25.55%, and 34.95% for MCABE, (Uršič et al., 2013), and (Mozos et al., 2005), respectively. The standard deviation is the lowest for MCABE indicating the best balance in categorization and least over-fitting.

7.2 The vision-based model

In the vision-based part extraction pipeline we initially resize each image to a fixed width of 500 pixels. Selective search's "fast mode" is used to obtain the proposed regions, which are afterwards additionally filtered. The Caffe (Jia et al., 2014) implementation of a deep network called hybrid-CNN (Zhou et al., 2014) is used to perform feature extraction. To learn the part dictionary we use the sequential mini-batch formulation of MCABE, which enabled us to cope with the

large memory requirements of the vision-based data. We have employed two epochs with the first epoch consisting of two batches (see Section 7.2.4 for the strategy on mini-batch-related parameter selection).

In the following, we first report the findings of the parameter estimation analysis and provide some insights about the proposed design decisions for the part creation pipeline. A separate dataset has been utilized for this purpose. In particular, room categorization performance was optimized by observing category predictions using various model settings on a set of indoor scenes corresponding to eight household categories from (Mandeljc et al., 2016), i.e., bathroom, bedroom, children room, closet, corridor, dining room, kitchen, and living room. Using this dataset we provide insights on additional part filtering (Section 7.2.1), CNN feature selection (Section 7.2.2), considered feature perturbations (Section 7.2.3), and sequential mini-batch MCABE related parameters (Section 7.2.4). Second, we present a large dataset, i.e., the vision part of the DR dataset, and discuss the experimental scenario (Section 7.2.5), which is utilized to verify the vision-based model performance (Section 7.2.6).

7.2.1 Additional part filtering. Depending on individual image content, selective search produced various numbers of regions per image. The numbers varied from approximately 300 up to approximately 2000. To form a compact image representation, we performed additional filtering on the set of obtained regions. A straightforward approach would be to use only a limited number of best ranked region proposals, which are scored by the selective search (Uijlings et al., 2013). However, this strategy produced sub-optimal results in our case. We noticed that performance improved by applying appropriate filtering with respect to the region size. Moreover, it turns out that using various region size limitations for both, lower and upper limit, greatly affects the quality of the representation. We measure the size of the region by the length of its diagonal, and perform the size-based filtering with respect to the percentage of image diagonal. Very small regions are not informative and therefore greatly reduce the categorization performance. If the lower limit is too high the part-based model benefits are lost, and the performance is reduced again. The optimal lower limit (50% of the image diagonal) resulted in 10% overall performance boost. Small variations of this parameter (in the range of about 5%) do not influence the results significantly. On the other hand, imposing an upper limit on a region size decreases categorization accuracy, since a holistic view of the scene can be quite informative. Finally, regions with a shorter-to-wider-side ratio less than $\frac{1}{3}$ have been removed, and regions whose surface area overlaps with another region by more than 90% have been discarded. Figure 11 shows boxplots demonstrating the number of parts per image obtained after filtering with respect to various categories in our dataset for experimental evaluation (presented in Section 7.2.5).

7.2.2 CNN feature selection. Although CNNs are well suited for use as generic feature extractors (Razavian et al., 2014), the choice of the specific model can largely influence the system performance. We have tested two pre-trained CNN models that possess the same architecture. The first

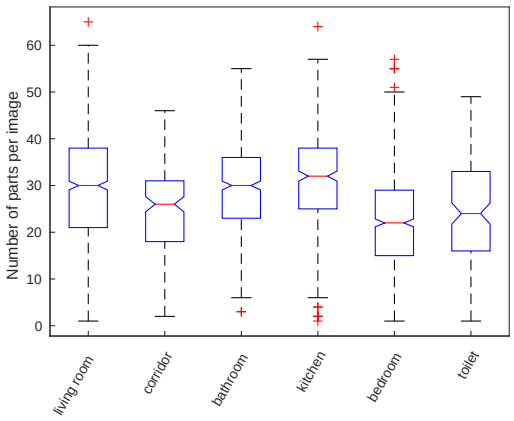

Figure 11. Boxplots indicating the statistics of the number of extracted vision-based parts per image for each of the categories in the DR dataset.

model, called ImageNet-CNN, was pre-trained by (Girshick et al., 2014) on an object-centric dataset (Deng et al., 2009). The second model, called hybrid-CNN, was pre-trained by (Zhou et al., 2014) on a combination of object-centric (Deng et al., 2009) and scene-centric (Zhou et al., 2014) images. One might expect that ImageNet-CNN would provide similar or even better categorization performance than hybrid-CNN, since selective search was designed with an aim of providing high-quality regions with respect to potential object locations (Uijlings et al., 2013), and a CNN trained merely on object-centric dataset usually performs better on images of objects (Zhou et al., 2014). But it turns out that in our case the extracted regions do not necessarily enclose only individual objects, but also larger fractions of a scene. For this reason, the categorization performance is increased for approximately 8% when hybrid-CNN is applied, since features of (Zhou et al., 2014) are more descriptive in this context.

7.2.3 Feature perturbations. In order to compute deep features from a region proposal, the chosen network architecture (Zhou et al., 2014) requires the corresponding image patch to be first resized to a fixed size of 227×227 pixels, without preserving its aspect ratio. In addition to this default approach, two alternative techniques have been considered. In the first one, regions were resized, by retaining aspect ratio, such that the smallest dimension was equal to 227 pixels. Then, three crops of the size 227×227 were extracted from the region, one from the center and one from each side along the wider dimension. CNN features were then extracted from each crop and joined using average-pooling to form a final feature vector. In the second approach, the findings of (Chatfield et al., 2014) have been followed. The features were formed using data augmentation, i.e., each image region was resized, by retaining aspect ratio, such that the smallest dimension was equal to 256 pixels. Afterwards, 227×227 crops were extracted from the four corners and the center of the image, whereupon crops were flipped about the y-axis, producing 10 perturbed samples per each region. CNN features extracted from each sample were at the end joined using sum-pooling. We have observed no gain in performance by following each of the two approaches in our application.

7.2.4 The sequential mini-batch MCABE. A relatively large set of parts extracted from training images yields a significant computational burden for MCABE algorithm. For this reason the sequential mini-batch formulation is applied in the vision-based domain. We experimentally assess the effects of the minibatch-related parameters (the number of epochs and the number of batch sizes) to categorization performance on a subset of household scene images from (Mandeljc et al., 2016). We found out that, in general, larger batch sizes usually lead to better categorization results, whereas relatively small variations in batch size (in the range of about 5%) do not influence the results significantly. It has been determined that categorization performance slightly decreases when sequential mini-batch approach is engaged, compared to a bare MCABE-core implementation. Therefore, an optimal strategy is to set batch sizes as large as memory resources allow, and adjust the number of epochs accordingly.

7.2.5 Dataset and experimental setup. All of the following experiments were performed on a subset of rooms from the DR dataset (introduced in Section 7.1.1), i.e., on a set of rooms (100 out of in total 123) for which data from visual sensor modality was acquired along the robots path. This set contains data acquired in 17 living rooms, 5 corridors, 29 bathrooms, 18 kitchens, 20 bedrooms, and 11 toilets. RGB images were acquired from a height of 1.45 m above the floor with cameras downward tilt of 5 degrees relative to the horizon. The shots were taken at acquisition points 30 cm apart along the robots trajectory. On positions at which the robot performed a turn on the spot, the rotation was split into intervals of 45 degrees, at which image acquisitions were performed. Depending on the room size, approximately 10 to 50 images were obtained in it. A few example images acquired in one of the living rooms are shown in Figure 1b. The same experimental scheme was applied as in the laser-based experiments, i.e., the room-wise leave-one-out cross-validation.

7.2.6 Experimental results. The performance of our vision-based method that uses MCABE for part dictionary learning is compared to two baseline approaches, i.e., the non-efficient greedy method in which all parts obtained in training form a dictionary, and an approach that uses affinity propagation (AP) algorithm (Frey et al., 2007) to learn exemplars for each category independently of all other categories. Results are shown as confusion matrices in Table 5.

The proposed method, MCABE, outperformed the greedy approach in distinguishing bedrooms and toilets, obtained equal performance on images of living rooms, corridors, and kitchens, and got outperformed on images of bathrooms. In both approaches, the vision-based parts demonstrated poor performance on images of corridors, which were mainly categorized as bedrooms. Image parts displaying plain walls and tall wardrobes are frequently present in both categories, which makes them hard to distinguish. On the other hand, toilets were largely miss-categorized as bathrooms. A great challenge represents the frequent occurrence of tiles in the two categories, whereas the most prominent feature of the toilet category, i.e., the toilet bowl, is not present in every image. Overall, the non-efficient greedy method and

Table 5. Confusion matrices for the vision-based experiments with greedy method, Affinity Propagation - AP (Frey et al., 2007), and proposed MCABE approach. Category abbreviations: LR-living room, CO-corridor, BA-bathroom, KI-kitchen, BE-bedroom, WC-toilet.

Greedy	LR	CO	BA	KI	BE	WC
LR	**88.24**	0.00	0.00	0.00	11.76	0.00
CO	0.00	**20.00**	0.00	0.00	80.00	0.00
BA	3.45	0.00	**82.76**	0.00	6.90	6.90
KI	33.33	0.00	0.00	**61.11**	5.56	0.00
BE	10.00	0.00	0.00	0.00	**90.00**	0.00
WC	0.00	0.00	54.55	0.00	9.09	**36.36**
AP	LR	CO	BA	KI	BE	WC
LR	**88.24**	0.00	0.00	0.00	11.76	0.00
CO	0.00	**0.00**	0.00	0.00	100.00	0.00
BA	0.00	0.00	**86.21**	0.00	3.45	10.34
KI	33.33	0.00	0.00	**55.56**	11.11	0.00
BE	10.00	0.00	0.00	0.00	**90.00**	0.00
WC	0.00	0.00	63.64	0.00	9.09	**27.27**
MCABE	LR	CO	BA	KI	BE	WC
LR	**88.24**	0.00	0.00	0.00	11.76	0.00
CO	0.00	**20.00**	0.00	0.00	80.00	0.00
BA	3.45	0.00	**72.41**	0.00	6.90	17.24
KI	27.78	0.00	0.00	**61.11**	11.11	0.00
BE	5.00	0.00	0.00	0.00	**95.00**	0.00
WC	0.00	0.00	45.45	0.00	9.09	**45.45**

MCABE achieved comparable performance. The accuracy of the categorization equals 73.00% for greedy and 72.00% for MCABE approach, mean success rate is 63.08% and 63.70% in favor of MCABE, whereas a standard deviation of the diagonal elements of the confusion matrix is improved from 29.39% to 27.95% when MCABE is applied.

MCABE outperformed AP on images of corridors, kitchens, bedrooms, and toilets, obtained equal performance on living rooms, and got outperformed on bathrooms. Similarly to other techniques, but with even more drastic deviation, AP demonstrated poor performance on corridors and toilets. Application of AP resulted in a more unbalanced categorization, since exemplars of each category are in this approach being selected without considering the cross-category similarities of individual parts. The measures indicating overall categorization performance for the AP show that MCABE performed better, i.e., the accuracy equals 71.00%, mean success rate is 57.88%, whereas a standard deviation of the diagonal elements of the confusion matrix equals 37.54%.

7.3 Laser and vision fusion

The fusion related experiments were performed on a subset of rooms from the DR dataset, for which data from both, range and visual, sensor modalities were acquired along the robots path (a set of 100 rooms presented in Section 7.2.5). The same experimental procedure was applied as in previous sections, i.e., the room-wise leave-one-out cross-validation.

7.3.1 Laser-only performance. To obtain a valid baseline, laser-based experiment was reapplied to the corresponding subset of rooms. The obtained confusion matrix is shown in Table 6. The overall accuracy of categorization equals 53.00%, mean success rate is 61.28%, and a standard deviation of diagonal elements of the confusion matrix equals 24.16%. Figure 12 shows a failure case of the laser-based model for an example of a bathroom, i.e, an instance of

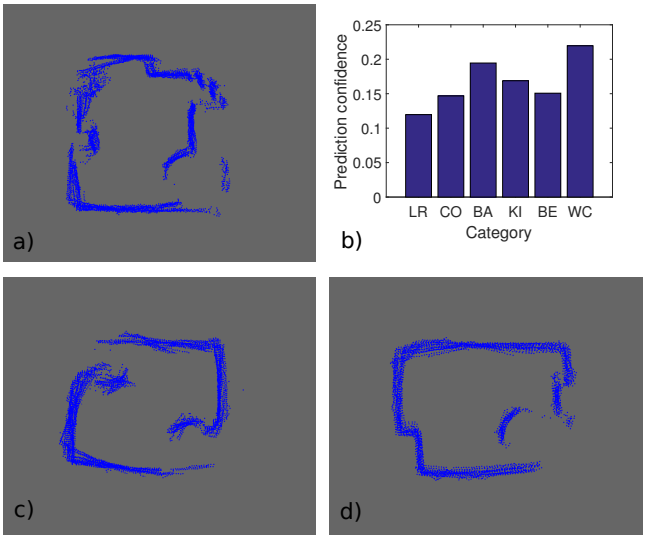

Figure 12. A failure case of the laser-based model, demonstrating challenging nature of the data. (a) Ground-plane view of an example of a bathroom that was miss-categorized as a toilet. (b) The corresponding per-category prediction confidences. (c, d) Ground-plane views of two example toilets, which show that it is difficult to distinguish between room types solely from laser-range data. These ground-plane views were created by a Rao-Blackwellized Particle Filter SLAM of (Grisetti et al., 2007), provided by The Mobile Robot Programming Toolkit, and are used only for visualization.

Table 6. Results of the laser-based experiment applied to the subset of rooms from the DR-dataset (Section 7.2.5). Category abbreviations: LR-living room, CO-corridor, BA-bathroom, KI-kitchen, BE-bedroom, WC-toilet.

	LR	CO	BA	KI	BE	WC
LR	**47.06**	17.65	0.00	23.53	11.76	0.00
CO	0.00	**80.00**	20.00	0.00	0.00	0.00
BA	0.00	6.90	**24.14**	20.69	6.90	41.38
KI	0.00	16.67	0.00	**55.56**	16.67	11.11
BE	10.00	10.00	0.00	10.00	**70.00**	0.00
WC	0.00	0.00	9.09	0.00	0.00	**90.91**

the category on which the laser-based method demonstrated poor performance. It can be seen that categorizing some of the rooms can be highly challenging when solely a 2D laser scanner is employed.

7.3.2 Vision-only performance. A confusion matrix corresponding to the results of the vision-based approach is shown in the bottom part of Table 5. With respect to the laser-based system, the method performed better with living rooms, bathrooms, kitchens, and bedrooms, whereas it got outperformed with corridors and toilets. A failure case corresponding to an instance of a corridor, i.e., a category for which the vision-based system demonstrated poor performance, is shown in Figure 13. Since corridors in domestic environments are usually quite short, the only highly characteristic views can be obtained at both ends, when facing towards the other end of the corridor. Therefore, several images acquired along the robots path are highly uninformative. On the other hand, presence of clothes and warm colors in this example caused the extracted parts to outvote the correct category predictions.

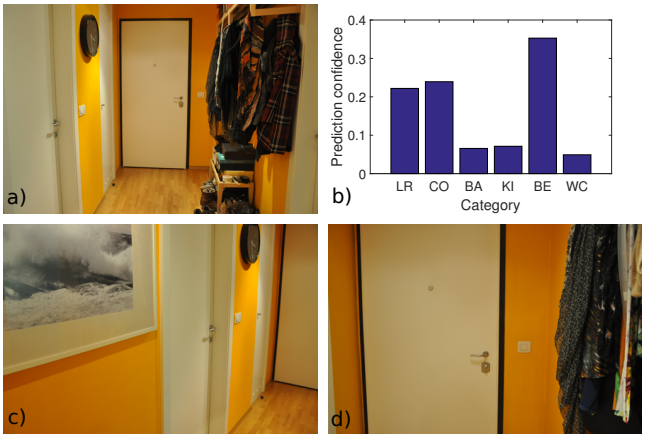

Figure 13. A failure case of the vision-based model. (a) One of visually unambiguous images extracted in a corridor. Based on all images obtained in a room, this corridor was miss-categorized as a bedroom. (b) The corresponding room-level per-category prediction confidences. (c, d) Two example images in the same corridor whose maximal confidences were ascribed to a bedroom. Presence of clothes and uninformative views resulted in incorrect category assignment.

7.3.3 The fusion. We chose the weights for the fusion model (9) in the combined laser-vision-based approach by observing the model performance using various weight combinations, and then selecting the best-performing combination. In particular, we performed an experiment in which the laser-data weight w_1 was varied from 0 to 1 by increments of 0.01. Note that the vision-based weight is defined as $w_2 = 1 - w_1$. Figure 14 shows the obtained accuracies and mean success rates for all considered weight values. The left-most points of the graph correspond to the pure vision-based categorizations, the right most points correspond to the pure laser-based model, while the intermediate points correspond to the combined predictions with intermediate weight ratios. The obtained combined predictions demonstrate that the laser-vision fusion provides better categorization performance than each of the single-modal methods. Best performance was observed at $w_1 = 0.88$ at accuracy of 70.00% and mean success rate of 71.78%. The peak performance is shifted to the right side of the graph in Figure 14 because laser-based room category predictions are in general ascribed significantly smaller confidence values than predictions of the vision-based modality (see Figure 15). A close inspection of the confusion matrices obtained at various w_1 values verified that performance of the fusion approach at w_1 values lower than approximately 0.8 does not differ significantly from the pure vision-based methods performance. Therefore, assignment of appropriate weights to each of the modalities serves as calibration of individual sensor contributions in our sensor fusion.

The confusion matrix corresponding to the optimal fusion weights is shown in Table 7. Compared to laser-based approach, the combined method provides better results with four categories, whereas the performance decreases with corridors and toilets. These two categories are the weakest for the vision-based system, which is reflected in the combined predictions. Compared to the vision-based approach, the combined method shows improvement for three of the weak-performing categories, results remain

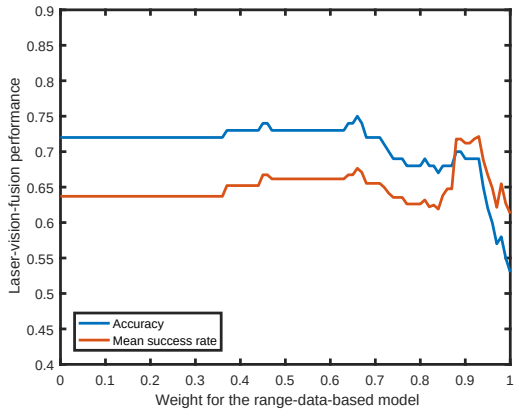

Figure 14. Laser-vision-fusion room categorization performance with respect to the weight of the range-data-based model.

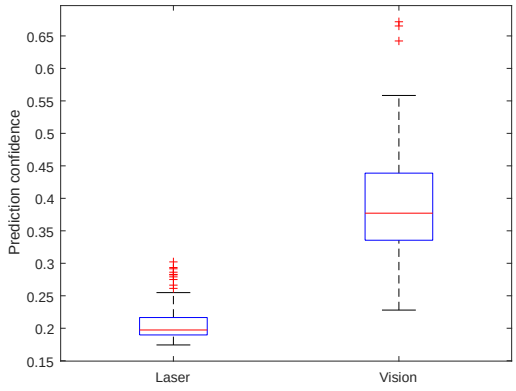

Figure 15. Boxplots of prediction confidence values across all test samples for the laser-based and vision-based modalities. Note the significantly higher mean value of the vision modality.

unchanged for the bedrooms, whereas a decrease in performance is observed for living rooms and bathrooms, i.e., the two weakest categories for the laser-based method. The combined method thus effected as a balancing factor on both systems. The standard deviation of the diagonal elements of the confusion matrix is improved from 24.16% of laser-based and 27.95% of vision-based systems, to 18.13% of the laser-vision-based method. Therefore, the combined method performed better than both systems, laser-based and vision-based, when considered individually.

8 Conclusion

We presented a new approach for laser-vision-fusion-based room categorization that is formulated in a uniform part-based framework. We addressed the open challenges regarding part-based spatial models in two of the most commonly used sensor modalities in household mobile robots, the 2D laser range scans and vision: (i) The limitations of the low-level part-based spatial model (Uršič et al., 2013) for laser-based room categorization were pointed out. It is natural to expect an increase in categorization performance by introducing higher layers of abstraction into this hierarchical representation. However, it turns out that when modeling space on a larger scale, i.e., on layer 4 of the sHoP model (Uršič et al., 2013), local scans form a receptive field that is too small to account for the size of the low-level parts and that sHoP learning

Table 7. A confusion matrix demonstrating the performance of the laser-vision-fusion-based room categorization approach in a uniform part-based framework. Category abbreviations: LR-living room, CO-corridor, BA-bathroom, KI-kitchen, BE-bedroom, WC-toilet.

	LR	CO	BA	KI	BE	WC
LR	**82.35**	0.00	0.00	5.88	11.76	0.00
CO	0.00	**60.00**	0.00	0.00	40.00	0.00
BA	3.45	0.00	**44.83**	10.34	6.90	34.48
KI	11.11	0.00	5.56	**66.67**	16.67	0.00
BE	5.00	0.00	0.00	0.00	**95.00**	0.00
WC	0.00	0.00	18.18	0.00	0.00	**81.82**

leads to a poorly constructed dictionary at this scale. To avoid these limitations, we presented a novel spatial model consisting of mid-level parts that are built on top of the low-level part-based representation. State-of-the-art results were attained on a new challenging dataset of domestic rooms, by achieving the most balanced categorization performance over all categories. (ii) The newly presented part-based model in the laser domain allowed a natural integration with a vision-based model (Mandeljc et al., 2016), since both representations consist of mid-level parts built at the same level of granularity. A novel uniform part-based laser-vision-fusion scheme for room categorization demonstrated that the combined approach outperforms both, laser-based and vision-based, methods, when considered individually. (iii) We presented a novel part selection approach, MCABE, that finds a discriminative set of exemplars based on pair-wise part affinities. The algorithm is general and allows building compact dictionaries for part-based representations suitable for categorization. The proposed learning was applied for discriminative mid-level parts dictionary construction in both, laser-based and vision-based, modalities, resulting in good room categorization performance.

The proposed discriminative part selection framework and proposed fusion scheme are general and can be applied to other modalities. We expect the performance could be further boosted by including 3D depth sensors. Moreover, some prior knowledge about the distribution of rooms in general apartments could also be integrated in a joint multi-modal categorization framework, which would likely also improve the recognition. These will be the topics of our future work.

Acknowledgements

This work was supported by the applied project L2-6765 by the Slovenian Research Agency and a research programme P2-0214. We also acknowledge MoD/Dstl and EPSRC for providing the grant to support the UK academics (Aleš Leonardis) involvement in a Department of Defense funded MURI project.

References

Amores J (2013) Multiple instance classification: Review, taxonomy and comparative study. In: *Artificial Intelligence*, vol. 201, pp. 81-105.

Ayers B and Boutell M (2007) Home Interior Classification using SIFT Keypoint Histograms. In: *Proceedings of IEEE CVPR*.

Chatfield K, Simonyan K, Vedaldi A and Zisserman A (2014) Return of the devil in the details: Delving deep into convolutional nets. In: *BMVC*.

Clemen R T and Winkler R L (1999) Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis*, vol. 19, no. 2, pp. 187-203.

Cortes C and Vapnik V (1995) Support-Vector Networks. *Machine Learning*, vol. 20, no. 3, pp. 273-297.

Csurka G, Dance C R, Fan L, Willamowski J and Bray C (2004) Visual categorization with bags of keypoints. In: *Proceedings of ECCV*, pp. 1-22.

Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: *Proceedings of IEEE CVPR*, pp. 248255.

Dissanayake M W M G, Newman P, Clark S, Durrant-Whyte H F and Csorba M (2001) A solution to the simultaneous localization and map building (SLAM) problem. *Robotics and Automation*, vol. 17, no. 3, pp. 229-241.

Dixit M, Chen S, Gao D, Rasiwasia N and Vasconcelos N () Scene Classification with Semantic Fisher Vectors. In: *Proceedings of IEEE CVPR*.

Doersch C, Gupta A and Efros A A (2013) Mid-Level Visual Element Discovery as Discriminative Mode Seeking. In: *NIPS*, pp. 494-502.

Fidler S, Boben M and Leonardis A (2009) Evaluating multi-class learning strategies in a generative hierarchical framework for object detection. In: *NIPS*, vol. 22, pp. 531-539.

Fisher R A (1936) The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, vol. 7, no. 7, pp. 179-188.

Freund Y and Schapire R E (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139.

Frey B J and Dueck D (2007) Clustering by passing messages between data points. *Science*, vol. 315, pp. 972-976.

Friedman S, Pasula H and Fox D (2007) Voronoi random fields: Extracting the topological structure of indoor environments via place labeling. In: *Proceedings of IJCAI*, pp. 2109-2114.

Gemert J C, Geusebroek J, Veenman C J, Snoek C G M and Smeulders A W M (2006) Robust Scene Categorization by Learning Image Statistics in Context. In: *CVPR Workshop*, pp. 105-105.

Girshick R, Donahue J, Darrell T and Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of IEEE CVPR*.

Gong Y, Wang L, Guo R and Lazebnik S (2014) Multi-scale Orderless Pooling of Deep Convolutional Activation Features. In: *Proceedings of ECCV*, pp. 392-407.

Grisetti G, Stachniss C, and Burgard W (2007) Improved techniques for grid mapping with Rao-Blackwellized particle filters. *Transactions on Robotics*, vol. 23, no. 1, pp. 3446.

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S and Darrell T (2014) Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*.

Jiang Z, Lin Z and Davis L S (2013) Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. *TPAMI*, vol. 35, no. 11, pp. 2651-2664.

Jie Z and Yan S (2015) Robust Scene Classification with Cross-Level LLC Coding on CNN Features. In: *Proceedings of ACCV*, pp. 376-390.

Juneja M, Vedaldi A, Jawahar C V and Zisserman A (2013) Blocks that Shout: Distinctive Parts for Scene Classification. In: *Proceedings of IEEE CVPR*.

Kolodner J L (1983) Maintaining Organization in a Dynamic Long-Term Memory. *Cognitive Science*, vol. 7, pp. 243-280.

Kolodner J L (1987) Extending Problem Solver Capabilities Through Case-Based Inference. In: *Machine Learning Workshop*, pp. 21-30.

Leordeanu M and Hebert M (2005) A spectral technique for correspondence problems using pairwise constraints. In: *Proceedings of IEEE ICCV*, vol. 2, pp. 1482-1489.

Leordeanu M, Hebert M and Sukthankar R (2009) An Integer Projected Fixed Point Method for Graph Matching and MAP Inference. In: *NIPS*.

Li L J, Su H, Xing E P and Fei-Fei L (2010) Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: *NIPS*, vol. 23, pp. 1378-1386.

Lowe D G (1999) Object recognition from local scale-invariant features. In: *Proceedings of ICCV*, vol. 2, pp. 1150-1157.

Malisiewicz T, Gupta A and Efros A A (2011) Ensemble of Exemplar-SVMs for Object Detection and Beyond. In: *Proceedings of IEEE ICCV*.

Mandeljc R, Uršič P, Leonardis A and Kristan M (2016) Part-Based Room Categorization for Household Service Robots. In: *Proceedings of IEEE ICRA*.

Margolin R, Zelnik-Manor L and Tal A (2014) OTC: A Novel Local Descriptor for Scene Classification. In: *Proceedings of ECCV*.

Mesnil G, Rifai S, Bordes A, Glorot X, Bengio Y and Vincent P (2015) Unsupervised Learning of Semantics of Object Detections for Scene Categorization. *Pattern Recognition Applications and Methods*, vol. 318, pp. 209-224.

Mozos O M, Stachniss C and Burgard W (2005) Supervised Learning of Places from Range Data using AdaBoost. In: *Proceedings of IEEE ICRA*, pp. 1730-1735.

Mozos O M, Rottmann A, Triebel R, Jensfelt P and Burgard W (2007) Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems*, vol. 55, pp. 391-402.

Nosofsky R M, Pothos E M and Wills A J (2011) The Generalized Context Model: An Exemplar Model of Classification. In: *Formal Approaches to Categorization*, pp. 18-39.

Pandey M and Lazebnik S (2011) Scene Recognition and Weakly Supervised Object Localization with Deformable Part-based Models. In: *Proceedings of ICCV*, pp. 1307-1314.

Parizi S N, Oberlin J G and Felzenszwalb P F (2012) Reconfigurable models for scene recognition. In: *Proceedings of IEEE CVPR*, pp. 2775-2782.

Parizi S N, Vedaldi A, Zisserman A and Felzenszwalb P (2015) Automatic discovery and optimization of parts for image classification. In: *Proceedings of ICLR*.

Premebida C, Faria D R, Souza F A and Nunes U (2015) Applying probabilistic Mixture Models to semantic place classification in mobile robotics. In: *Proceedings of IEEE IROS*, pp. 4265-4270.

Pronobis A, Mozos O M, Caputo B and Jensfelt P (2010) Multi-modal Semantic Place Classification. *IJRR*, vol. 29, no. 2-3, pp. 298-320.

Pronobis A and Jensfelt P (2012) Large-scale semantic mapping and reasoning with heterogeneous modalities. In: *Proceedings of IEEE ICRA*, pp. 3515-3522.

Quattoni A and Torralba A (2009) Recognizing indoor scenes. In: *Proceedings of IEEE CVPR*, pp. 413-420.

Razavian A S, Azizpour H, Sullivan J and Carlsson S (2014) Cnn features off-the-shelf: An astounding baseline for recognition. In: *CVPR Workshops*, pp. 512519.

Ruiz-Sarmiento J R, Galindo C and Gonzalez-Jimenez J (2015) Joint categorization of objects and rooms for mobile robots. In: *Proceedings of IEEE IROS*, pp. 2523-2528.

Sadovnik A and Chen T (2012) Hierarchical object groups for scene classification. In: *Proceedings of IEEE ICIP*, pp. 1881-1884.

Sadeghi F and Tappen M F (2012) Latent Pyramidal Regions for Recognizing Scenes. In: *Proceedings of ECCV*, pp. 228-241.

Shi L, Kodagoda S and Dissanayake G (2010) Laser range data based semantic labeling of places. In: *Proceedings of IEEE IROS*, pp. 5941-5946.

Shi L, Kodagoda S and Dissanayake G (2012) Application of semi-supervised learning with Voronoi Graph for place classification. In: *Proceedings of IEEE IROS*, pp. 2991-2996.

Shi L, Kodagoda S and Piccardi M (2013) Towards simultaneous place classification and object detection based on conditional random field with multiple cues. In: *Proceedings of IEEE IROS*, pp. 2806-2811.

Shrivastava A, Malisiewicz T, Gupta A and Efros A A (2011) Data-driven Visual Similarity for Cross-domain Image Matching. *ACM Transaction of Graphics*, vol. 30, no. 6.

Singh S, Gupta A and Efros A A (2012) Unsupervised Discovery of Mid-level Discriminative Patches. In: *Proceedings of ECCV*.

Sousa P, Araujo R and Nunes U (2007) Real-Time Labeling of Places using Support Vector Machines. In: *IEEE International Symposium on Industrial Electronics*, pp. 2022-2027.

Sowa J F(1984) *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley Longman Publishing Co., Inc.

Sutton C and McCallum A (2012) An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267373.

Swadzba A and Wachsmuth S (2010) Indoor Scene Classification Using Combined 3D and Gist Features. In: *Proceedings of ACCV*, vol. 6493, pp. 201-215.

Tschiatschek S, Iyer R, Wei H and Bilmes J (2014) Learning Mixtures of Submodular Functions for Image Collection Summarization. In: *NIPS*, pp. 1413-1421.

Uijlings J R R, van de Sande K E A, Gevers T and Smeulders A W M (2013) Selective Search for Object Recognition. *International Journal of Computer Vision*.

Uršič P, Kristan M, Skočaj D and Leonardis A (2012) Room classification using a hierarchical representation of space. In: *Proceedings of IEEE IROS*, pp. 13711378.

Uršič P, Tabernik D, Boben M, Skočaj D, Leonardis A and Kristan M (2013) Room categorization based on a hierarchical representation of space. *IJARS*, vol. 10.

Uršič P, Leonardis A, Skočaj D and Kristan M (2016) Hierarchical Spatial Model for 2D Range Data Based Room Categorization. In: *Proceedings of IEEE ICRA*.

Wright J, Yi M, Mairal J, Sapiro G, Huang T S and Shuicheng Y(2010) Sparse Representation for Computer Vision and Pattern Recognition. *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031-1044.

Wu J, Christensen H I and Rehg J M (2009) Visual Place Categorization: Problem, Dataset, and Algorithm. In: *Proceedings of IEEE IROS*.

Yao J, Fidler S and Urtasun R (2012) Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In: *Proceedings of IEEE CVPR*, pp. 702-709.

Yi C, Oh Y C, Suh I H and Choi B U (2011) Indoor Place Classification Using Robot Behavior and Vision Data. *IJARS*, vol. 8, no. 5.

Zhou B, Lapedriza A, Xiao J, Torralba A and Oliva A (2014) Learning Deep Features for Scene Recognition using Places Database. In: *NIPS*, pp. 487-495.

Zuo Z, Wang G, Shuai B, Zhao L, Yang Q and Jiang X (2014) Learning Discriminative and Shareable Features for Scene Classification. In: *Proceedings of ECCV*, pp. 552-568.