Hierarchical Spatial Model for 2D Range Data Based Room Categorization

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

Abstract— **The next generation service robots are expected to co-exist with humans in their homes. Such a mobile robot requires an efficient representation of space, which should be compact and expressive, for effective operation in real-world environments. In this paper we present a novel approach for 2D ground-plan-like laser-range-data-based room categorization that builds on a compositional hierarchical representation of space, and show how an additional abstraction layer, whose parts are formed by merging partial views of the environment followed by graph extraction, can achieve improved categorization performance. A new algorithm is presented that finds a dictionary of exemplar elements from a multi-category set, based on the affinity measure defined among pairs of elements. This algorithm is used for part selection in new layer construction. Room categorization experiments have been performed on a challenging publicly available dataset, which has been extended in this work. State-of-the-art results were obtained by achieving the most balanced performance over all categories.**

I. INTRODUCTION

Cognitive capabilities are becoming a highly desired property in the next generation service robots that would co-exist with humans in their homes [1], [2]. A cognitive ability like space perception and building efficient models of space are most important capabilities of home service robots, which remain challenging problems. The spatial model should be compact, while enable efficient representation of diverse environments.

Several approaches to modeling space have been proposed, e.g., [2], [3], [4], [5], [6]. In most of them, robots perceive their surroundings using range sensors that acquire 2D ground-plan-like data, which is also the case in our approach. But in contrast to most approaches we argue the benefits of knowledge transmission in space perception. A robot would benefit from learning basic elements of spatial representation by observing numerous spaces. These elements would form a prior that it can adapt with small amount of learning to a new environment, which is not unlike a human capability. For example, an adult person has a general representation of a bathroom, which was formed through observing numerous bathrooms. Upon encountering a new bathroom a person easily orients in the never-before observed instance of this spatial category. This is in stark contrast to the current state-of-the-art in robotics. When the robot observes a new room its representation is usually formed independently of previously modeled spaces. Therefore, the size of the generated model grows at least linearly with respect to the number of memorized places.

Spatial representations can be divided in metric and topological. Metric representations [3] strive to accurately describe the geometry of the environment, while topological models [4] use graphs to model space. Hybrid approaches [5] combine both of the above paradigms and combining these approaches on multiple levels of abstraction results in hierarchical representations [2], [6], [7]. These models have an appealing design property of viewing real-world abstractions at various levels of abstraction. In particular, compositional hierarchies from the field of shape recognition [8] possess many appealing properties that a robot might exploit. The lowest layer in such hierarchy is composed of basic elements (e.g., short line segments), which are combined to produce parts of increasing complexity at higher layers. These deep hierarchical compositional networks are well suited to model entities that can be allocated to several different categories, since they offer within-category and across-category sharing of elements [8] and result in compact models. A prior work that introduced these models to spatial recognition was the sHoP model from Ursic et al. [7]. They interpret laser-range scans as images and apply the model similar to [8] for parsing into parts of mid-level complexity. The spatial layout of the mid-level parts was then encoded by a holistic model, i.e., a spatial histogram of parts. A drawback of [7] is that it requires converting laser scans to images first and the final model (histogram) is not encoded within the philosophy of hierarchical compositional models, thus deviating from the predicted advantages of these models [8]. Another drawback of the obtained holistic model is that, like all holistic models, it is principally sensitive to noise and missing data.

A. Our Approach and Contributions

In this paper we propose a new, fourth, layer in the sHoP [7] model and the new learning routines, which is our main contribution. The three existing layers in sHoP are compositional dictionaries with parts shared across many categories and can thus be thought of as compact generative models of observed measurements. In contrast, the new fourth layer contains semantically-richer and category-specific parts required for improved discrimination. Unlike [7], [9], who build holistic descriptors by spatial histograms of lower-layer part responses, the room categorization is carried out in our approach solely by detection of the fourth-layer parts. The local specificity is introduced in the fourth-

[1] The authors are with the Faculty of Computer and Information Science, University of Ljubljana {peter.ursic, danijel.skocaj, matej.kristan}@fri.uni-lj.si

[2]Aleš Leonardis is with the School of Computer Science, University of Birmingham a.Leonardis@cs.bham.ac.uk

layer parts by building them as fully-connected local graphs, called the local-map visibility graphs (LMVG), that encode relations among pairs of third-layer parts, which extend much farther than the fixed size receptive fields of the lower layers of original sHoP. In fact, the receptive field is defined by the range of robot's sensor (laser scanner in our case). In summary, the robot performs a short translation in the environment, parses each consecutive laser scan by sHoP, and applies EKF-SLAM [10] to construct a local map. The data association is modified in the SLAM to work directly with the third-layer parts, thus building a smoothed third-level part local map, which is in turn used to construct an instance of the fourth-level part. The fourth layer dictionary of parts is constructed by applying a novel part selection approach to a large set of LMVGs obtained from observing many instances of rooms. In particular, a new combinatorial optimization algorithm MCABE is proposed to construct a dictionary of most discriminative exemplars of LMVGs. The proposed approach is tested on an extended publicly-available dataset of Domestic Rooms [7] and achieves state-of-the-art results by surpassing the holistic models built by spatial histograms of lower layers of sHoP [7], [9].

The remainder of the paper is organized as follows: Related work is discussed in Section II, Section III provides an overview of the sHoP [7] model and introduces the proposed category-specific higher layer. In Section IV the MCABE algorithm is introduced, Section V discusses our approach to room categorization, Section VI reports the experimental results, and in Section VII conclusions are drawn.

II. RELATED WORK

Several approaches combine multiple modalities to perform room categorization. In [6] great performance is achieved by combining information about the existence of objects, the appearance, geometry, and topology of space. Mozos et al. [11] combine laser range data with vision, several approaches apply vision only, e.g., [4], while a 3D Time-of-Flight infrared sensor was applied in [12].

Scarce research has been performed in the field of 2D laser-range-data-based room categorization, despite its potential for complementing predictions obtained from other modalities. Mozos et al. [13] developed a method that performs categorization based on a single scan. They used AdaBoost and distinguished between four categories. Friedman et al. [14] employed the Voronoi random fields, providing the distinction between four categories. 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 [14] requires building a holistic map of the space, meaning that the quality will significantly depend on the accuracy of the SLAM. In contrast, our approach requires only partial reconstructions and allows changes in room while the robot moves through it (e.g., moving chairs, persons walking, doors moving,...). While [14] 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. Furthermore, our representation does not require loop-closing in SLAM. In [7], a compositional model was also used to extract partial descriptions of space. However, their spatial model is a holistic spatial histogram of compositions and requires learning a support vector machine on top of a fixed three-layer architecture. This reduces the potential of online adaptations of the model and deviates from the hierarchy-of-parts paradigms [8]. In contrast, our approach does not require converting laser scans into images and performs room categorization using top-layer parts only, without requiring formation of large descriptors with support vector machines.

As part of our approach we propose a new method for discriminative parts selection, exemplars, for room categorization. A large body of literature exists on exemplar learning. Sparse coding approaches [15] model data vectors as sparse linear combinations of basic elements. Linear discriminant analysis [16] finds a linear combination of features, which characterizes two or more classes of elements. Another powerful approach for exemplar selection is the Affinity propagation [17] that applies message-passing for selecting exemplars that separate the dataset into clusters based on a prescribed affinity matrix. Summarization approaches [18] are designed to extract information from data that is both minimal and most important in the considered context. In contrast to these approaches, our proposed approach allows online formation of dictionary of exemplars for cross-category discrimination based solely on the defined affinity across the elements in the training set.

III. SPATIAL HIERARCHY OF PARTS

The sHoP model from [7] is briefly described in Section III-A. In Section III-B we detail the proposed new layer and the construction of the category-specific parts.

A. Category-Independent Lower Layers

A laser-range finder mounted on a mobile robot provides ground-plan-like observations from which the sHoP learning algorithm learns a Layered hierarchy of parts (Fig. 1). Parts are formed by promoting local compositions of basic shapes that frequently occur in the measurements. The lowest layer, layer 1, contains a fixed dictionary of eighteen line fragments at different orientations (Fig. 1a). The dictionary of the second layer is formed by promoting frequently occuring combinations of the layer 1 parts, and similarly, layer 3 contains the compositions of layer 2 parts. Parts on layers 2 and 3 are rotationally invariant. 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. The learned parts in the three layers are common to all room categories, i.e., are category-independent, and ensure good scalability with respect to the number of modeled categories.

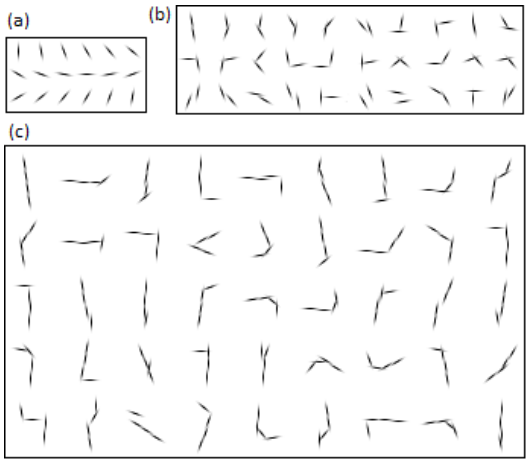

Fig. 1. Lower layers of the sHoP model: layer 1 parts (a), layer 2 parts (b) and layer 3 parts (c). Due to spatial constraints only a subset of all parts is shown for layers 2 and 3.

B. Category-Specific Higher Layer

Learning the fourth layer of the hierarchy by following the co-occurrence-based approach described in the previous section 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. The second issue is that the training data at this scale offers a poor combinatorial evidence of co-occurrence leading to overly category-instance specific parts that do not generalize well to other instances within the category. Furthermore, a more flexible structure is required to account for the diversity of real-world instances. We address the first issue by merging consecutive partial scans into a wider view, *a local map*, of the environment directly on the output of layer 3 parts. The second issue is addressed by introducing a new encoding of compositions by so-called local-map visibility graphs (LMVG) and a new part selection algorithm called multi-category affinity-based exemplar search (MCABE).

In summary, a mobile robot performs a short tour in the environment from which it obtains a set of partial views from the laser-range finder (Fig. 2). From each laser-scan, layer 1 parts are extracted, and the inference from [7] is applied to obtain layer 2 and 3 parts. The partial views from consecutive scans are then merged using SLAM into a local map, upon which the LMVG is created. In the following we describe the pipeline for layers 1 to 3 inference, compositions-based SLAM for local map creation and construction of LMVGs.

Layer 1 extraction: Layer 1 part positions are extracted from each laser-scan acquired in the short tour. In contrast to [7] we do not convert the scans to images, but perform inference directly on the raw measurements. Layer 1 part positions are calculated directly from laser-scan data, based on the differences between measurements of the neighbouring beams in the laser-scan. The obtained Layer 1 scan representation is then smoothed to reduce the measurement noise. The smoothing algorithm focuses on small local patches, in which approximately-serially positioned and similarly oriented parts are averaged to form a straight line.

Inference of layers 2 and 3: Using the learned library, layer 2, and then layer 3 part positions and orientations are inferred from each scan (Fig. 2 d-1, d-2).

Local map creation: We use feature-based EKF-SLAM [10] to merge the partial views into a local map of the environment using only the inferred layer 3 parts (Fig. 2e). We propose a new part-type matching approach for the data association in the SLAM. The observed parts are associated with parts that are already in the map according to the score

$$\Gamma(d, s) = (\delta - d)(1 + s), \tag{1}$$

where d is the distance between the predicted and observed location, δ is the threshold, and s is the discrete part-type matching function. In cases where $d > \delta$ the distance between predicted and observed location is too large and the association is not made. The value of s is equal to 1 if parts are of the same part type, and 0 otherwise. Data association is therefore performed by finding prediction-observation pairs of parts corresponding to maximal values of $\Gamma(d, s)$. Layer 3 parts forming the output of SLAM are reconstructed down to the first layer, to which smoothing is applied to improve alignment in the map.

Visibility-graph: The inferred lower-layer parts in the local map present the nodes of the visibility graph, which is the model for our layer 4 parts construction. Visibility-graph (Fig. 2f) is a fully-connected graph in which each 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 the angle $\alpha \in [0, \pi]$, 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 part. Finally, a binary variable v defines the *visibility* of the two nodes. It is determined if part corresponding to one node is visible from the part corresponding to the other node by examining sub-part compositions of both parts. We say that the two parts are visible if there exists a line segment running from a Layer-1 sub-part of the first part to a Layer-1 sub-part of the other part, lies fully within the room, and does not intersect with any other detected part. The visibility property encodes which regions correspond to room interior, since areas that are covered by the *visible* connections define the interior surface. Fig. 2g shows an example of the *visible* connections for a selected node. Each node in the graph is assigned a real value η, that measures the curvature of the associated part. In particular, η is a mapping from the space of all parts from the first three layers to interval $[0, 2]$, where value 0 corresponds to a straight line, value 1 to the right angle, value 2 to a degenerate case of acute angle of 0 degrees, while all the intermediate states are mapped in-between those values. Curvature measure is calculated from part geometry by linear transformations of angles between short line segments that constitute a part. A few examples of calculated measures for Layer 3 parts are shown in Fig. 3.

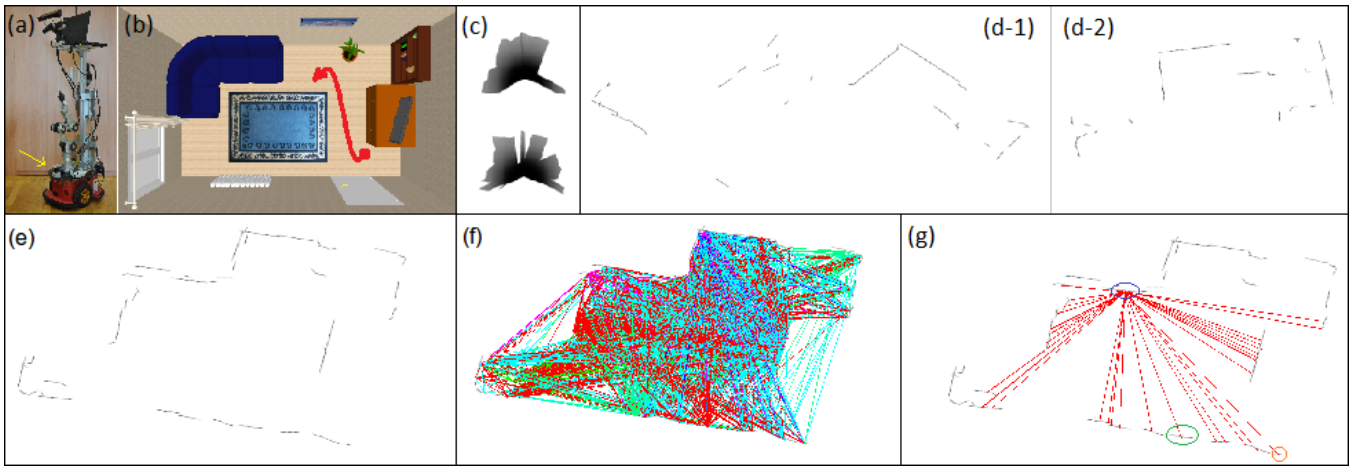

Fig. 2. LMVG creation pipeline. (a) Our robot observes the environment using a laser-range finder, which is marked with an arrow. (b) Schematic view of an example room with robots short-path marked with red. The scheme is shown only for visualisation, data used in our work was obtained in real-world environments. (c) Schematic view of two example scans obtained along the short-path. (d-1) and (d-2) Two example partial views of the room represented by Layer 3 parts, acquired at different positions along the short-path. These representations, as well as all the following images, were derived from real-world data. (e) All partial views merged into local map. (f) Visibility-graph with all the *visible* connections. (g) All *visible* connections of an example node that is a Layer 3 part encircled by a blue ellipse. Among others, for example, a visible connection exists to a Layer 3 part marked with a green ellipse, and to a Layer 1 part marked with an orange ellipse.

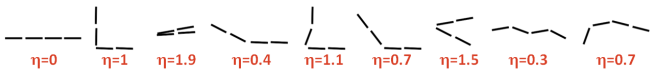

Fig. 3. Examples of η values for a few Layer 3 parts.

Comparing LMVGs: An affinity measure among LMVGs is needed for construction and inference with layer 4 parts. LMVGs are compared to each other by graph matching. We start with definition of a pair of graph-similarity functions. The first function, designed to measure the similarity between nodes of two graphs, is defined as

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

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 function is designed to measure the similarity between connections of two graphs and is defined as

$$f_2(\gamma_1, \gamma_2) = f_2((D_1, \alpha_1, v_1), (D_2, \alpha_2, v_2)) = \tag{3}$$
$$(1 - |v_1 - v_2|)(e^{-\frac{1}{2}\frac{(D_1-D_2)^2}{(\max\{D_1,D_2\}\sigma_D)^2}} + e^{-\frac{1}{2}\frac{(\alpha_1-\alpha_2)^2}{\sigma_\alpha^2}}),$$

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 (3) 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:

$$M_{a,b} = \tag{4}$$
$$M_{(i_1,i_2),(j_1,j_2)} = \begin{cases} f_1(\eta_1, \eta_2) & \text{for } i_1 = j_1, i_2 = j_2 \ , \\ f_2(\gamma_1, \gamma_2) & \text{for } i_1 \neq j_1, i_2 \neq j_2 \ , \\ \rho(a, b) & \text{otherwise,} \end{cases}$$

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 evaluates to $f_1(\eta_1, \bar{\eta})$, where $\bar{\eta}$ denotes the combined curvature corresponding to i_2 and j_2, while it evaluates to 0 otherwise. Any cluster 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 $\Theta(C) = \Theta(x) = \frac{1}{2}\mathbf{x}^T\mathbf{M}\mathbf{x}$. Optimal matching $\mathbf{x}^*$ is thus the binary vector that maximizes the score

$$\mathbf{x}^* = \text{argmax}(\mathbf{x}^T\mathbf{M}\mathbf{x}). \tag{5}$$

Equation (5) is a the quadratic assignment problem (QAP), which is NP-hard. To find the approximate solution we use spectral matching [19] augmented with an integer-projected fixed-point method [20]. The normalized value of the solution is defined as an affinity measure $\psi(G_1, G_2) = 2\frac{\Theta(x^*)}{|G_1||G_2|} \in [0, 1]$, 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. Graph-similarity functions embedded into QAP enable a loose definition of LMVG structure. Semi-static objects of the environment, like chairs, do not disturb the recognition, since graph comparisons are not sensitive to small disturbances of nodes and connections.

IV. LEARNING DISCRIMINATIVE DICTIONARY

In this section we present the Multi-Category-Affinity-Based-Exemplars-search (MCABE) algorithm, which is used to find exemplar LMVGs that form the layer 4 in our hierarchy.

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 poorly describe the examples 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 k-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 $|E|$ as small as possible, which induces sparsity of the solution. The cost function is defined as

$$F(S, E, \psi) = \tag{7}$$

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

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, we have designed an iterative algorithm for finding an approximate solution.

All elements in the set are initialized as exemplars of their categories. This set is then iteratively reduced, while maximizing discrimination. In each iteration exemplars of each category are recalculated, one category at a time. For a single considered category exemplars from previous iteration are discarded, while new exemplars are being calculated with respect to current exemplars of all other categories. Let $1 < \tilde{k} < k$ be the index of the category currently being optimized. Then, optimization is performed with respect to exemplars of categories $1, 2, \ldots, \tilde{k}-1$ determined in current iteration, and with respect to exemplars of categories $\tilde{k}+1, \ldots, k-1, k$ determined in previous iteration. For category $\tilde{k}$ new exemplars are determined by arranging its elements into clusters, which are composed of an exemplar and all the elements represented by it. Clusters contain elements that are similar to each other, whereas its exemplar is the one that is least similar to exemplars of other categories. Since exemplars of other categories also represent well their corresponding non-exemplars, good discriminativity is induced against all elements of other categories. Clusters are being formed by sequential element validation. Considered 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 score that approximates the summation term of (7) is calculated. The summation terms increase the value of the cost function in presence of large within-category similarities and decrease it in presence of large across-category similarities. 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 this iteration. 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. These state-dependent action rules are detailed in Algorithm 1. By selecting exemplars following this approach the value of the cost function is being maximized. The process is repeated over all categories, which completes one iteration. The result is a set of candidate exemplars that serve as the basis for the next iteration. Note that sparsity constrains are directly incorporated in the above-described set of state-dependent action rules.

V. APPLICATION TO ROOM CATEGORIZATION

The following scenario is considered for room categorization. In training phase, sHoP hierarchy up to the new layer 4 is learned as described in Section III and MCABE from Section IV is applied for dictionary selection in the layer 4.

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. The constructed LMVGs are compared to the layer 4 LMVGs in the learned hierarchy. Based on the affinity between the measured and library parts, the room is categorized by a voting scheme as follows.

Assume that category-specific layer 4 contains k sets of exemplars, 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

Algorithm 1 The MCABE algorithm

procedure MCABE($S_1, S_2, ..., S_k, \psi$)
  $E_i \leftarrow S_i \ \forall i \in \{1, 2, ..., k\}$
  $E_i^{old} \leftarrow \varnothing \ \forall i \in \{1, 2, ..., k\}$
  while $\exists i \ni: E_i^{old} \neq E_i$ **do**
    $E_i^{old} \leftarrow E_i \ \forall i \in \{1, 2, ..., k\}$
    for $j = 1, 2, ..., k$ **do**
      $\Omega \leftarrow \varnothing$
      $E_j \leftarrow \varnothing$
      $\hat{E_x} \leftarrow \varnothing \ \forall x \in S_j$
      for each $x \in S_j$ **do**
        $y \leftarrow \underset{y \in S_j \setminus \{x\}}{\operatorname{argmax}} \ \psi(x, y)$
        $f_x \leftarrow \psi(x, y) - \max_{z \in E \setminus E_j} \psi(x, z)$
        $f_y \leftarrow \psi(x, y) - \max_{z \in E \setminus E_j} \psi(y, z)$
        if $x \notin \Omega$ & $y \notin \Omega$ **then**
          If $f_x < f_y$ then $E_j = E_j \cup \{y\}$ and
          $\hat{E_y} = f_y$, otherwise $E_j = E_j \cup \{x\}$
          and $\hat{E_x} = f_x$
        else if $x \notin \Omega$ & $y \in \Omega$ **then**
          $e \leftarrow \underset{e \in E_j}{\operatorname{argmax}} \ \psi(y, e)$
          If $f_x > \hat{E_e}$ then $E_j = (E_j \setminus e) \cup \{x\}$
          and $\hat{E_x} = f_x$
        else if $x \in \Omega$ & $y \notin \Omega$ **then**
          If $x \in E_j$ and $\hat{E_x} < f_y$ then
          $E_j = (E_j \setminus x) \cup \{y\}$ and $\hat{E_y} = f_y$
          otherwise if $x \notin E_j$ and
          $e \leftarrow \underset{e \in E_j}{\operatorname{argmax}} \ \psi(x, e)$ and
          $\hat{E_e} < f_y$ then $E_j = E_j \cup \{y\}$
          and $\hat{E_y} = f_y$
        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
    end
  end
  return $E_1, E_2, ..., E_k$

exemplars in E_i. The probability that G belongs to category C_i is calculated by normalization

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

and the G casts the vote to the category $C_{\max} = \arg\max_{C_i} P(G \in C_i)$ proportionally to the probability $P(G \in C_{\max})$. These calculations are repeated for all LMVGs obtained in a room, and then a category assignment vote with the highest probability is declared as a chosen room category.

VI. EXPERIMENTAL RESULTS

Our approach was experimentally evaluated on the DR dataset [7] which we have augmented by adding two room categories. The extended dataset is publicly available at http://go.vicos.si/drdataset. The dataset was obtained using a Hokuyo URG laser range-finder mounted on a Pioneer P3-DX robot while moving in several real apartments (see [7] for further details on the robotic system used). The original dataset [7] 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.

Several experiments have been performed, each of them being cast as leave-one-out crossvalidation. 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.

The parameters of node-wise and connection-wise structural flexibilities in LMVGs (2), (3) were determined through simulation, not containing any test data, and were set to $\sigma_\eta = 0.2$, $\sigma_D = 0.2$, and $\sigma_\alpha = 0.35$. The threshold for data-association score in (1) was set to $\delta = 0.5$, which reflected optimal SLAM reconstruction. To reduce computation time of the graph matching procedures, each graph was down-sampled to contain only a third of its original nodes. The MCABE algorithm achieved convergence in 98.78% of all runs. It converged after about five iterations and selected from 26% to 29% of all LMVGs as exemplars. In the runs where convergence was not achieved the algorithm oscillated between two solutions and the one that maximized the optimization function was chosen. Graph matching is computationally the most demanding part of the proposed method. Computation times depend on matched graph sizes and on a laptop with 2.3 GHz dual-core processor range from 0.2 sec for a pair of smallest graphs, to 453.5 sec for the largest. Further parallelization could be applied for performance speedup. To categorize an LMVG about 60 matchings were performed in this work.

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 I). The mean accuracy 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. Lowering this value only gradually reduces performance, since the obtained LMVGs become less discriminative. On the other hand, increasing this value increases the probability of error accumulation in SLAM, which also leads to performance reduction. Usually, one to at most three LMVGs were obtained in each room.

In the Section VI-A we analyze our proposed spatial model and MCABE and in Section VI-B our approach is compared to the state-of-the-art.

A. Proposed Spatial Model Evaluation

The MCABE algorithm was compared to two baseline approaches for category-specific LMVGs selection. The first

TABLE I

RESULTS OF THE EXPERIMENTS WITH DIFFERENT SHORT TOUR LENGTHS. SEE TEXT FOR MEASURE EXPLANATIONS.

Short tour length	Mean accuracy	Accuracy	Standard deviation
100	48.13	42.28	32.48
125	50.16	42.28	24.55
150	53.65	43.09	28.32
175	56.27	45.53	32.58
200	58.57	52.85	23.59
225	56.59	53.66	16.81
250	50.20	41.46	28.87

is the greedy approach in which all LMVGs obtained in training are stored in the new layer. This is obviously not computationally and memory efficient. In the second approach exemplars are calculated using the Affinity Propagation (AP) algorithm [17]. In this case, exemplars for each category are determined by searching for most representative LMVGs within each category, independently of the other categories, thus disregarding the across-category similarities. Results of our proposed MCABE approach are shown in comparison with the two baseline approaches in Table II.

TABLE II

CONFUSION MATRICES FOR THE EXPERIMENTS WITH GREEDY APPROACH, AFFINITY PROPAGATION [17], AND PROPOSED MCABE APPROACH. CATEGORIES: LR-LIVING ROOM, CO-CORRIDOR, BA-BATHROOM, KI-KITCHEN, BE-BEDROOM, WC-TOILET.

Greedy	LR	CO	BA	KI	BE	WC
LR	**52.38**	33.33	4.76	9.52	0.00	0.00
CO	0.00	**83.33**	16.67	0.00	0.00	0.00
BA	0.00	5.71	**34.29**	17.14	8.57	34.29
KI	4.76	14.29	9.52	**52.38**	14.29	4.76
BE	7.14	32.14	3.57	14.29	**42.86**	0.00
WC	0.00	0.00	8.33	0.00	0.00	**91.67**
AP [17]	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	**61.90**	19.05	4.76	4.76	9.52	0.00
CO	16.67	**50.00**	16.67	0.00	16.67	0.00
BA	0.00	8.57	**37.14**	8.57	20.00	25.71
KI	0.00	19.05	4.76	**42.86**	28.57	4.76
BE	10.71	10.71	3.57	10.71	**64.29**	0.00
WC	0.00	0.00	16.67	0.00	0.00	**83.33**

The proposed method, MCABE, outperforms the greedy approach in distinguishing living rooms, bathrooms, and bedrooms, while the greedy approach performs better with corridors, kitchens, and toilets. 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 when viewed as ground-plans. The same miscategorization is also present in the greedy approach. While the non-efficient greedy method also performs quite well, the overall accuracy of categorization is increased from 50.40% to 53.66% when computationally and memory more effective MCABE approach is used.

MCABE outperforms the AP on living rooms, kitchens, and bedrooms, performs equally well with toilets, and is outperformed with corridors and 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 as ground plans. 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.

B. Comparison With State-of-the-Art

We compared our proposed approach to the approach of [13], for which the code was provided to us by the authors, and to the recently presented hierarchical model [7]. The approach of Mozos et al. [13] uses AdaBoost to boost simple features to a strong classifier and is based on a single scan. At the parameter determination step we determined the optimal number of hypotheses used and the optimal order of binary classifiers (see [13] for details). 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 the room category. In [7] HoC descriptor was used to perform room categorization. Each laser-scan was represented by a HoC descriptor, which was generated using category-independent layer 3 parts of the hierarchy. Then, 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 III. Our approach outperformed the boosting-based [13] with kitchens, bedrooms, and toilets, while [13] performed better with living rooms, corridors, and bathrooms. The boosting-based approach [13] performed poorly with kitchens, which were mainly categorized as living rooms, whereas bedrooms were mainly categorized as bathrooms. It is evident from Table III that [13] 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 [13] 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 [13] 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.

TABLE III

CONFUSION MATRICES FOR THE EXPERIMENTS WITH THE APPROACHES OF [13] AND [7]. CATEGORIES: LR-LIVING ROOM, CO-CORRIDOR, BA-BATHROOM, KI-KITCHEN, BE-BEDROOM, WC-TOILET.

Moz.[13]	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[7]	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**

The proposed approach outperformed HoC on corridors, kitchens, and toilets. Equal categorization performance was obtained on bedrooms, 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 [13], 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 as ground-plans, 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. Rigid grid into which space is divided to form the HoC descriptor can not cope with such variability. On the other hand, the structure of LMVGs coupled with principled exemplar selection process is designed to represent well such large within-category variability.

Note that the average accuracy over the categories for MCABE is 53.66%, which is in fact lower than for competing approaches [13] (60,98%) and [7] (59,35%). A close inspection reveals that the improvement in [13] and [7] come from overfitting - the average performance is boosted at the cost of very poor classification of a few categories. The overfitting is quantified by standard deviation of category-wise classification accuracies, which are 16.81%, 25.55% and 34.95% for MCABE, [7], and [13], respectively. The standard deviation is the lowest for MCABE indicating the best balance in categorization and least ovefitting.

VII. CONCLUSION AND FUTURE WORK

We presented a new, fourth, layer in the sHoP [7] model consisting of category-specific parts, called LMVGs. We also presented a new combinatorial optimization algorithm MCABE, which is used for dictionary construction of most discriminative exemplars of LMVGs that form our new layer. Room categorization is carried out in the proposed approach solely by detection of the fourth-layer parts. The performance of the model is evaluated on the extended publicly available DR dataset [7], to which two additional categories have been added. The proposed approach provided state-of-the-art results by achieving the most balanced categorization performance over all categories. The presented exemplars learning algorithm is general and can easily be extended to incorporate other modalities. In the future work we plan to augment our model with data obtained by the visual sensors to further improve the categorization performance.

REFERENCES

[1] D. Skočaj, M. Janiček, M. Kristan, G.-J. M. Kruijff, A. Leonardis, P. Lison, A. Vrečko, M. Zillich, A basic cognitive system for interactive continuous learning of visual concepts, ICAIR, pp. 30-36, 2010

[2] B. Kuipers, R. Browning, B. Gribble, M. Hewett, E. Remolina, The Spatial Semantic Hierarchy, Artificial Intelligence, vol. 119, pp. 191-233, 2000

[3] A. Elfes, Occupancy grids: a stochastic spatial representation for active robot perception, UAI, pp. 136-146, 1989

[4] Z. Zivkovic, O. Booij, B. Kröse, From images to rooms, Robotics and Autonomous Systems, vol. 55, no. 5, pp. 411-418, 2007

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

[6] A. Pronobis, P. Jensfelt, Large-scale semantic mapping and reasoning with heterogeneous modalities, ICRA, pp. 3515-3522, 2012

[7] P. Uršič, D. Tabernik, M. Boben, D. Skočaj, A. Leonardis, M. Kristan, Room Categorization Based on a Hierarchical Representation of Space, IJARS, vol. 10, 2013

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

[9] P. Uršič, M. Kristan, D. Skočaj, A. Leonardis, Room classification using a hierarchical representation of space, IROS, pp. 1371-1378, 2012

[10] S. Thrun, Robotic Mapping: A Survey, Exploring Artificial Intelligence in the New Millenium, pp. 1-35, 2002

[11] O. Mozos, A. Rottmann, R. Triebel, P. Jensfelt, W. Burgard, Supervised semantic labeling of places using information extracted from sensor data, Robot. Auton. Syst., vol. 55, pp. 391-402, 2007

[12] A. Swadzba, S. Wachsmuth, Indoor Scene Classification Using Combined 3D and Gist Features, ACCV, pp. 201-215, 2010

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

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

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

[16] R. A. Fisher, The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics, vol. 7, pp. 179-188, 1936

[17] B. J. Frey, D. Dueck, Clustering by Passing Messages Between Data Points, Science, vol. 315, pp. 972-976, 2007

[18] S. Tschiatschek, R. K. Iyer, H. Wei, J. A. Bilmes, Learning Mixtures of Submodular Functions for Image Collection Summarization, NIPS, pp. 1413-1421, 2014

[19] M. Leordeanu, M. Hebert, A Spectral Technique for Correspondence Problems Using Pairwise Constraints, ICCV, pp. 1482-1489, 2005

[20] M. Leordeanu, M. Hebert, R. Sukthankar, An Integer Projected Fixed Point Method for Graph Matching and MAP Inference, NIPS, 2009