Object Classification based on a Geometric Grammar with a Range Camera

Jiwon Shin, Stefan Gächter, Ahad Harati, Cédric Pradalier, and Roland Siegwart
Autonomous Systems Lab (ASL)
Swiss Federal Institute of Technology, Zurich (ETHZ)
{shin, gaechter, harati, pradalier, siegwart}@mavt.ethz.ch

***Abstract*— This paper proposes an object classification framework based on a geometric grammar aimed for mobile robotic applications. The paper first discusses the geometric grammar as a compact representation form for object categories with primitive parts as its constituent elements. The paper then discusses the object classification implemented as parsing of primitive parts. In particular, two approaches are discussed that constrain the search space in order to render the parsing of the primitive parts practical. The two approaches are experimentally verified, first, for a generic object category of chair applied to real range images acquired with a range camera mounted on a mobile robot and, second, for multiple generic object categories applied to synthetic range images. The experimental results show the practicability of the framework.**

I. Introduction

Two recent trends in technology and research are first, the advent of a novel type of range camera to capture 3D scenes [1], and second, the regained interest to solve the classification problem with object geometry, in particular, object structure [2], [3]. Object structure has been recognized as a strong characteristic for classification [4]. However, due to the lack of a reliable, compact, and affordable range image sensor, most of the algorithms and systems relied on object structure information extracted from 2D images. The novel type of range camera bears the potential to be compact and affordable as they are based on well-established technologies. The two trends have yet little converged despite of the necessity of object classification in human-robot interaction.

Structural variability within an object category may be well explained using a geometric grammar, especially a probabilistic one which can incorporate uncertain and incomplete measurements. In such an approach, object classification is reduced to detection of object parts and verification of the geometric relations among them. Since robust detection of complex object parts is as difficult as object classification itself, the representation of the overall structure of parts is reduced to cuboids, which are independent of appearance and easier to detect in point clouds.

With a grammar, objects can be represented in a hierarchical, recursive manner. The hierarchy can start at feature level and end at scene level, enabling not only object but also place classification, as demanded in mobile robotics [5]. Important is to include geometry in such a hierarchy. Classification based on object geometry has been introduced in the beginning of the computer vision era [6] and as soon as 3D range sensors became available. Such early approaches are discussed in [7], [8]. Furthermore, a grammar as a compact representation method[1] for object classification has been recognized for a long time [10]. Still, image parsing approaches have been applied mainly to 2D intensity images. Recent approaches concentrate on the comprehensive probabilistic formulation of the image parsing problem and the learning involved [10], [9]. This work explores the potentials of a grammar-based classification approach for 3D range images with an emphasis on practical solutions for mobile robotics.

The paper continues with a brief review of related work in Section II. Section III introduces a geometric grammar based representation for object classification. In Section IV, two different parsing approaches are discussed. Section V presents experimental results for a single object category on real range images and for multiple object categories on synthetic range images. The paper concludes with Section VI.

II. Related Work

The application of grammars as a representation method for object classification in 3D has been introduced by [9]. The work proposes a probabilistic geometric grammar in which geometric information about parts and their relations is represented using multivariate normal distributions. In this work, the probabilistic geometric grammar is defined by an AND-OR graph as in [10], which uses this representation to parse 2D intensity images according to a stochastic grammar. Most previous work has applied grammar-based approaches to intensity images. However, a grammar-like approach applied to 3D range images captured with a laser scanner, is presented in [11]. The mentioned work derives the parse tree from functional constraints. The present work uses 3D range images as well but differs from the cited ones in that the parsing is seen as a geometric constraint search, as discussed in [12], in context of object recognition. Therefore, the goal and contribution of this paper is to combine different approaches into a practical solution for object classification aimed for mobile robotic applications with sensory information provided by a range camera.

This work was partially supported by the EC under the FP6-IST-045350 robots@home project.

[1] The compactness of the representation is especially evident in learning, where compared to other approaches, smaller training sets are needed [9].

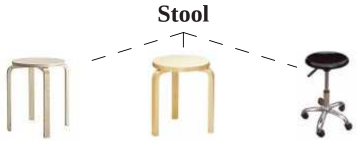


Fig. 1. An example of an OR category.

III. Object Representation

We adopt a parts-based object representation because many artificial as well as natural objects are composed of parts. This representation allows the parts to be modeled independently of the viewpoint, while incorporating missing or occluded parts. Further, decomposing an object into its parts enables a larger number of structurally similar objects to be grouped together. In this work, objects are decomposed hierarchically and are represented by a grammar.

A grammar defines the possible sequences of symbols that constitute valid statements in a given language. The verification of validity of a sequence is called parsing. Formally, a grammar $G = (T, N, S, R)$ consists of a set of terminals T and non-terminals N, a start variable $S \in N$, and a set of production rules $R : S \rightarrow T$, $S \rightarrow N$, $N \rightarrow N$, and $N \rightarrow T$. In this work, a formal grammar G describes which possible constellations of primitive parts constitute a valid object. Parsing is then classification, the verification if a given part constellation matches with a certain object category. The terminals T of the grammar are primitive parts, which can be any suitable descriptor. Here, primitive parts are physical parts abstracted as cuboids. For instance, a chair leg is represented by a stick-like cuboid. The non-terminals N are groups of primitive parts such as *chair back*, which is a collection of back and support pieces. The start variable S is a generic object category such as *chair*.

For parts detection, we employ an incremental detection algorithm presented in [13]. The parts detector extracts only the primitive parts not the relationship between parts. For example, both a leg underneath and a back support above a chair seat are simply two stick-like parts. Therefore, a context-free grammar[2], which treats all parts without contextual information, is appropriate. The production rules of the grammar can be represented by an AND-OR graph, which encodes all possible parse trees of the grammar and therefore, all structural information of an object category.

In an AND-OR graph, an OR-node codes the structural variability within an object category. Figure 1 depicts an example of such a category. A stool with four legs, three legs, or one axle are all valid variations of *stool*. Thus, an OR-node captures an object category with more than one acceptable constellation of parts. An AND-node encodes generalization of different object categories using partial similarity in the structure. The AND relationship combines different constellations of parts to obtain a more complex constellation. In Figure 2, the category of *stool* with a back constructs *chair*. Similarly, *armchair* is derived from *chair*.

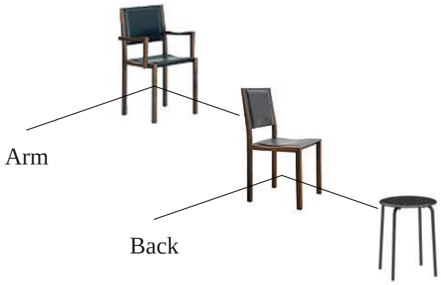


Fig. 2. An example of AND categories.

Due to uncertainty and incompleteness of the sensor data, the detected primitive parts have a limited confidence measure. Hence, the parameters for primitive parts – in this work, a cuboid defined by the center point and the span lengths – are modeled as random variables. The relations between the primitive parts – in this work, the Euclidean distance between the center points – are also modeled as random variables. Since each object is represented by geometric parts and relations in a probabilistic manner, the grammar is called a *probabilistic geometric grammar* [14].

A probabilistic geometric grammar is related to a stochastic attribute grammar, a known object representation method in computer vision.[3] An attribute grammar augments the terminals and non-terminals with attributes to define constraints for the production rules. Here, the attributes are center points and span lengths of the primitive parts, where distances between center points are used in the production rules. In a stochastic grammar, each production rule is augmented with *a-priori* defined or learned probability, which represents an *a-priori* knowledge on the object categories. However, since the object categories under investigation depend on the application, for example, office versus hospital, we assume no *a-priori* knowledge. Therefore, the probabilities of production rules in the AND-OR graph are not modeled explicitly.

IV. Object Classification

Object classification is implemented as parsing of detected primitive parts, where each object category is described by a probabilistic geometric grammar. The difficulty of parsing an object is that unlike human languages, object parts have no natural order. For instance, there is no particular order in which the four legs of a stool should be processed. Important is their structural relations. Therefore, to reduce complexity, it is necessary to constrain the search for valid parts constellations. The following discusses two constrained search methods – a generalized Hough transform (GHT) and a joint compatibility test (JCT). A general discussion on these methods can be found in [12]. Both methods build on a set of classified primitive parts provided by a parts detector, which localizes and classifies potential parts in a sequence of 3D range images [13].

[2] Actually, context is necessary to perform the object classification, but the contextual information is not encoded in the production rules, as it is the case for a context-sensitive grammar. Instead, they are additional constraints.

[3] A short review of the various types of grammars can be found in [15].

JCBB The joint compatibility branch and bound test for object parsing given a set of classified primitive parts.

```
Data: Detected parts $\mathcal{D}$ and modeled object parts $\mathcal{O}$
Input: Current compatibility hypothesis $\mathcal{H}$, which is a set of pairings of detected and modeled parts, and the current index $i$ to the detected parts under investigation
Output: The best compatibility hypothesis $\mathcal{B}$ found
Procedure:
$I \leftarrow |\mathcal{D}|$
$J \leftarrow |\mathcal{O}|$
if $i > I$ then
    if $|\mathcal{H}| \geq |\mathcal{B}|$ and
    $overall\_score(\mathcal{H}) > overall\_score(\mathcal{B})$ then
        $\mathcal{B} \leftarrow \mathcal{H}$
    end
else
    $D_i \in \mathcal{D}$
    for $j = 1$ to $J$ do
        $O_j \in \mathcal{O}$
        if $relation\_match(\mathcal{H}, D_i, O_j)$ then
            $pair \leftarrow \{i, j\}$
            JCBB $(\mathcal{H} \cup pair, i + 1)$
        end
    end
    if $|\mathcal{H}| + I - i \geq |\mathcal{B}|$ then
        JCBB $(\mathcal{H}, i + 1)$
    end
end
```

A. Generalized Hough Transform

In the first approach, the parsing is initialized by a voting scheme and then executed from the initial node of the parse tree in a top-down order. The details of the approach can be found in [16]. Here, a brief summary is given. The voting scheme is a probabilistic GHT adapted from the implicit shape model [3], which consists of a set of object specific parts and the corresponding votes for the relative locations of the reference point with respect to the parts. The set of votes can be regarded as a spatial discrete probability distribution reflecting the learned knowledge of the relative locations. In this work, the probability distribution is approximated by a Normal distribution. Each classified primitive part casts its votes for the reference point, and once the voting is completed, the local maxima indicate potential object locations, equivalent to non-terminal nodes of a parse tree. The parsing is completed by verifying the part constellation according to the object category definition starting from the initial guess. Hence, the voting scheme constrains the search to a limited number of branches in the parse tree.

The GHT is limited in the following ways. First, the approach does not limit the number of parts that can belong to an object. Any primitive part within an acceptable distance of the reference point votes as a part of the object. For example, if there are five leg parts within the correct distance of the object reference point, then all five parts are associated with the chair. Second, it limits sharing of different part types of similar structure. For example, a leg and a back support look similar, but because the relation of a leg to the reference point of the chair is different from that of a back support, they must have two different representations. Third, the approach does not contain any hierarchy, thus, for each new object category, the relations have to be relearned from scratch. The JCT can handle the three explained cases.

B. Joint Compatibility Test

The second approach is based on a constrained search in an interpretation tree. The search is actively guided during the parsing, not just at the initialization as the GHT. Object recognition as a constrained search problem has been widely explored, see for example [17]. Here, the search is constrained using the joint compatibility branch and bound algorithm [18]. The algorithm was initially designed to handle data association in simultaneous localization and mapping. In this work, assigning the detected primitive parts to the modeled object parts can be seen as a data association problem. An interpretation tree represents all possible correspondences between the detected primitive parts and the modeled object parts. The JCT traverses the interpretation tree in search for the hypothesis that includes the largest number of jointly compatible pairings of primitive parts and the model, see algorithm JCBB. The implementation used here relies on two evaluations: part matching and relation matching. The part matching determines if a detected part can be an instance of an object part, which is already realized in the parts detection phase. The relation matching examines the arrangement of the detected parts against the relations modeled by the grammar, where the quality of the matching is expressed by a score.

The parsing begins by selecting a seed part D_0 from the set of detected primitive parts $\mathcal{D}$, and pairing it with an appropriate[4] object part O_0 from the set of object parts $\mathcal{O}$ of the grammar, to initiate a compatibility hypothesis $\mathcal{H}_0$. Then, another part $D_1 \in \mathcal{D} \setminus \{D_0\}$ is selected and is paired with $O_1 \in \mathcal{O} \setminus \{O_0\}$. If no appropriate O_1 exists, D_1 is discarded. If it does, the relation between D_0 and D_1 is tested against all possible relations between O_0 and O_1. If matched, the pair is added to $\mathcal{H}_0$. If not, D_1 is discarded and O_1 is freed. A third part $D_2 \in \mathcal{D} \setminus \{D_0, D_1\}$ is selected and is accepted if its relations to all the parts already in $\mathcal{H}_0$ are preserved. The process continues until all parts in $\mathcal{D}$ have been tested against $\mathcal{H}_0$. The parsing is repeated with different seed parts, and the hypothesis that yields the highest match score is selected. Thus, the depth of the search in a branch of the tree is bounded by the best available hypothesis, and the branching is controlled by the best available relation matching. The branch and bound aspect of the algorithm is in its ability to choose which hypotheses to grow or to discard without evaluating all possible correspondences.

[4] If D_0 is a stick-like cuboid, then it can only be paired with a stick-like O_0 such as a chair leg or back support.

The relation matching score is based on a χ^2 test of the Mahalanobis distance to account for the uncertainty of parts. Parts of high variance are discarded. We use the distance between the center points as the relations. Other relations such as angle can also be used. Objects are assumed to be in their natural upright pose. Thus, the distance components in *z* direction are treated independently of *x* and *y* to ensure the preservation of vertical relations between parts while enabling invariance in *x* and *y*. The relation matching score r is derived from the Mahalanobis distance d as

$$r = \exp(-\lambda\, d), \tag{1}$$

where $\lambda = 0.5$ is a design parameter included to generate a probabilistic measure. The part matching score p is equal to the detection probability provided by the parts detector. The part and relation matching are expressed as likelihoods to handle the varying number of parts and relations during the parsing. Assuming statistical independence among parts and relations, the total matching score t for a hypothesis is

$$t = \sum_{i \in parts} \log \frac{p_i}{P_0} + \sum_{j \in relations} \log \frac{r_i}{R_0}, \tag{2}$$

where $P_0 = 0.5$ and $R_0 = 0.5$ are the part and relation scores for the null hypothesis. Finding the best hypothesis is then equal to maximizing this sum of log-likelihood ratios.

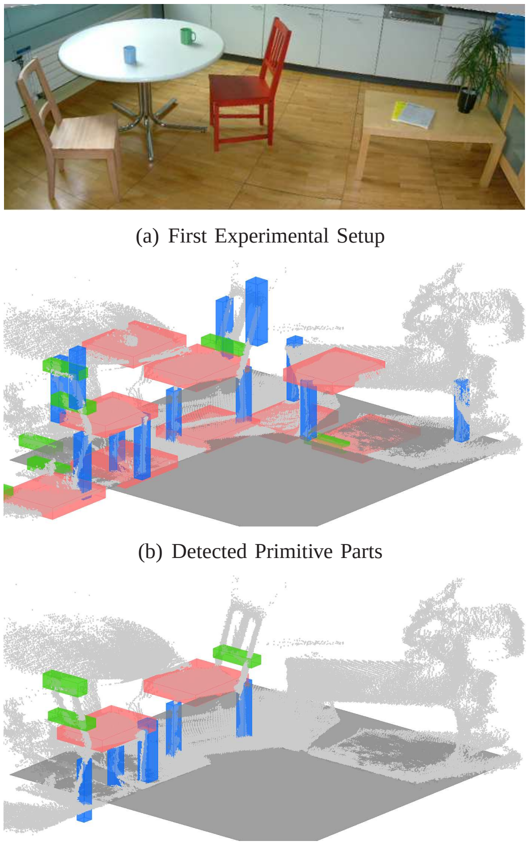
Multiple occurrence of objects are detected by running the parser repeatedly and removing the primitive parts of the best hypothesis from $\mathcal{D}$ at each iteration. For intra-class variations, the highest overall matching score is selected among all possible variations. Partial inter-class similarity is handled by growing the overall score until the detected primitive parts match the object model. For example, a stool is searched first, then from the remaining primitive parts, a back to form a chair. The relation matching between the stool and the back results in the overall score for the chair.

V. Results

The two parsing approaches are verified experimentally to demonstrate their practicability. The parsing is first tested on real data from an indoor environment for a generic object category of *chair*. Then, the parsing is tested for a more complex but simulated scenario to examine its capability of classifying multiple object categories.

A. Single Generic Object Class

The setup consists of two chairs in front of a dining table and a coffee table on the right of them as depicted in Figure 3(a). The scene is captured using a SR-2 range camera [1] mounted on a robot. Primitive parts are detected incrementally using the method described in [13] and classified as objects at every frame. In total, five data sets of 450 to 720 frames are acquired, each set with different robot trajectory, type of chairs, and distances between objects. Figure 3(b) is an example (No. 2 in Tab. I) of primitive parts detected in a sequence of 550 range images. During the detection process, some physical parts are undetected, some are detected incorrectly, and some, which do not belong to the object under investigation, are detected. Thus, the parsers have to cope with missing, incorrect, and clutter parts.



(a) First Experimental Setup

(b) Detected Primitive Parts

(c) Detected Objects

Fig. 3. (a) Setup of the first experiment (Setup of No. 2 in Tab. I). The robot travels first toward the wooden chair on the left and then turns slowly toward the coffee table on the right, while visiting the red chair in the center of the scene. Snapshot of the detected primitive parts (b) before and (c) after the object classification. The depicted parts are a selection of all the parts detected and classified in the whole sequence. Stick-like parts are depicted in blue, bar-like parts in green, and plate-like parts in red. Three point clouds out of the whole sequence are added for clarity.

TABLE I

	Generalized Hough Transform		Joint Compatibility Test	
No.	Detection Rate	Precision	Detection Rate	Precision
1	97.0 %	100 %	99.5 %	100 %
2	96.6 %	100 %	98.6 %	84.8 %
3	84.5 %	63.8 %	51.3 %	53.8 %
4	92.9 %	83.0 %	61.5 %	36.8 %
5	90.0 %	100 %	99.0 %	84.8 %

Figure 4 shows the grammar for *chair*. The JCT approach makes use of the entire parse tree, whereas the GHT one does not consider the two stick-like parts of *back* because it requires different part definitions for the different functional parts, as explained in Section IV-A. The GHT approach requires at least three parts that belong to *chair* to initialize the parsing; the JCT approach requires at least three parts for *stool* and at least one part for *back*.

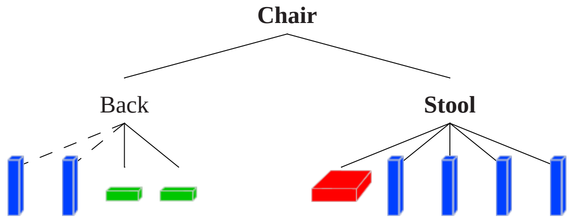


Fig. 4. The single category grammar used for testing. OR relations are indicated by *dashed* lines. AND relations are indicated by *solid* lines. Object categories are indicated by **bold** fonts.

Table I summarizes the results of classifying the five test sets. An example (No. 2 in Tab. I) of two correctly classified chairs using the GHT approach is depicted in Figure 3(c). The performance of the parsers are measured by the detection rate and precision.[5] Overall, the two approaches perform similarly well. Both have over 90 % detection rate and over 85 % precision for the first, second, and fifth data set. (Because a minimum number of primitive parts is required for the classification, the precision can reach 100 % here.) The drop in precision for the third and fourth set indicates that the object classification can fail for certain scene configurations. Such a drop can be caused by wrongly classified clutter or by neighboring primitive parts that misguide the parser. The detection rate drops as low as 50 % for the JCT approach because frequent misclassification of the coffee table legs with the legs of the nearby chair left too few chair parts for it to be classified. Such misclassification was much rarer for the GHT approach because it is more conservative. But, this advantage can reduce the flexibility to design the grammar as discussed in Section IV-A. Thus, the classification approach is extended to multiple object categories in order to analyze generalizability of the JCT approach.

B. Multiple Generic Object Classes

Because the current hardware setup with the range camera is not yet reliable enough to extract primitive parts for a larger set of various object categories with the algorithm proposed in [13], the experiment is run in a simulated environment. The simulated environment is designed in Blender and tested in Morsel, a 3D simulation software based on Panda3D. The range sensor is attached to a robot and scans the environment with a field-of-view of 180° horizontally and 90° vertically with 1° angular resolution. The grammar used for the experiment, shown in Figure 5, has three generic object categories: chair, bench, and table. (The parts are omitted for clarity.) *Chair* is a hierarchical category composed of *back* and *stool*. Twelve objects are sampled – seven chairs, two stools, two tables, and one bench – with different shape and structure as shown in Figure 6(a).

Since the focus is not the parts detection, the object parts are color-coded to ease the parts localization and classification. The parts are detected by clustering the point cloud by

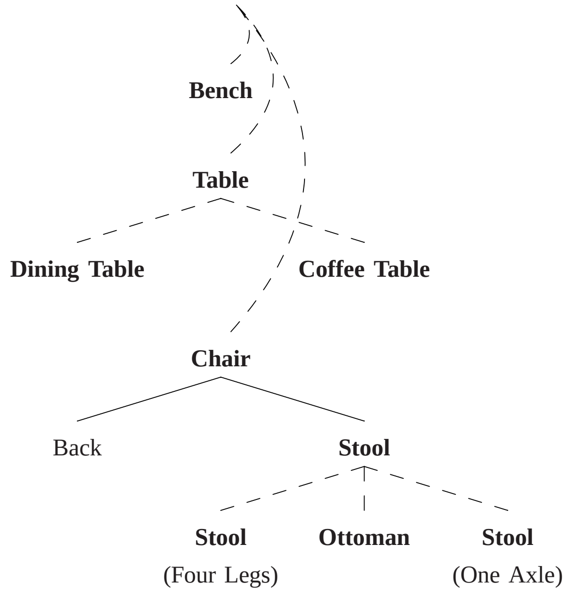
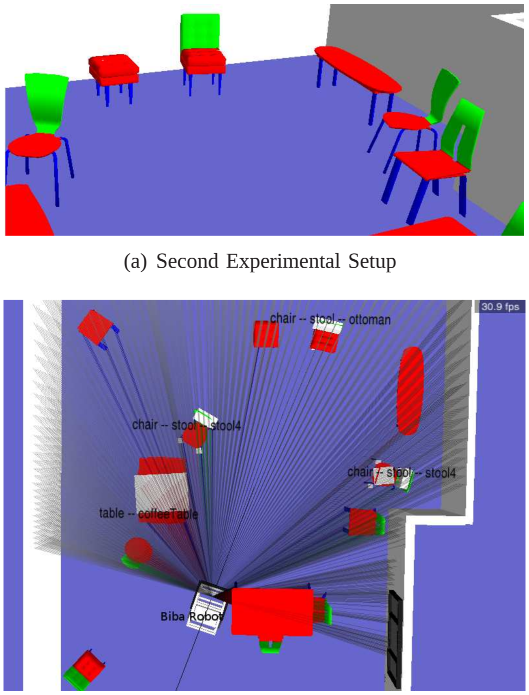


Fig. 5. The multiple categories grammar used for testing. *Dashed* lines indicate OR relations and *solid* lines, AND relations. *Bold* fonts indicate object categories.



(a) Second Experimental Setup

(b) Detected Objects

Fig. 6. (a) Setup of the simulated environment. The robot enters the scene from the lower left corner, travels toward the center and returns in a big loop to the starting point. (b) Snapshot right after the parser classified the objects. Parts that are classified as belonging to an object are displayed as cuboids in shade of gray. Whiter color indicates a higher match score for the object class. The object categories are displayed in text.

[5] If P is the actual number, TP the number of correctly classified and FP the number of falsely classified objects, then the detection rate is TP/P and the precision is $TP/(TP+FP)$ over the whole sequences.

TABLE II

	Chair (Stool 4 Legs)	Chair (Ottoman)	Chair (Stool 1 Axle)	Stool (4 Legs)	Ottoman	Stool (1 Axle)	Coffee Table	Dining Table	Bench	No Detection
Chair (S4L)	55	6	–	30	4	–	–	–	–	36
Chair (O)	–	13	–	–	10	–	–	–	–	6
Chair (S1A)	–	–	–	–	–	–	–	–	–	6
Stool (4L)	–	–	–	19	–	–	–	–	–	15
Ottoman	–	–	–	–	9	–	–	–	–	10
Stool (1A)	–	–	–	–	–	–	–	–	–	–
Table (C)	–	–	–	–	–	–	17	–	3	1
Table (D)	–	–	–	–	–	–	–	–	–	4
Bench	–	–	–	–	–	–	1	–	8	–
Not Object	3	–	–	1	–	–	–	–	–	–

color and computing a bounding-box for each cluster. The detected cuboids are classified into the appropriate primitive part types based on its color and size. The part probability is set according to the number of points per bounding-box.

A sequence of 108 frames are acquired while the robot is traveling through the simulation environment. Each frame contains an average of 23 extracted primitive parts from an average of 5 objects. The primitive parts are parsed with the JCT approach using the grammar defined in Figure 5. Figure 6(b) depicts a snapshot[6] right after the parser classified the objects. The performance is analyzed by counting, for each frame, the number of objects the parser could potentially classify, the correctly and falsely classified objects, and the missed objects. The result for all generic object categories is: 171 true positives, 8 false positives, and 78 false negatives. Details are in Table II. The column indicates the nominal object categories, and the row is the different object categories issue of the classification. The parser is able to classify all objects except the dining table and the single axle chair. Since *chair* is derived from *stool*, actual chairs are often taken as stools when the back parts are undetected. This explains the comparably large number of misclassification in this category. Furthermore, there is only a small difference between a four-legged stool and a ottoman – the latter has shorter legs but a thicker seat than the former. Many objects are not detected because objects are often observed partially, which results in geometrically incorrect primitive parts. Compared to the best cases of Section V-A, the detection rate of *chair* is lower at 67 % while the precision is of similar degree at 97 %. This shows that with increased scene complexity, the reliable detection of primitive parts and its correct classification become more challenging. However, given a sufficient number of geometrically correct parts, the parser is able to classify the primitive parts correctly.

VI. Conclusion

This paper presented an object classification framework based on a geometric grammar aimed for mobile robotic applications. It employs a parts-based geometric grammar where a parts detector provides the primitive parts. For efficient parsing, two constrained search methods are discussed. They are experimentally verified, first, for a generic object category of chair applied to real range images and second, for multiple generic object categories applied to synthetic range images. The results show that the parsers are capable of grouping together structurally-different objects of the same object category while distinguishing structurally-similar objects of different object categories apart. In addition, the approaches can handle clutter, missing, and uncertain parts.

There are two directions of improvement for the framework. First, to apply the framework on objects without distinctive structure, it is necessary to represent the parts and their relations in a more general form. Machine learning can be used to determine the most important parts and relations for a given object. Second, the framework can be extended to place classification, where object classification plays an important role. The hierarchical nature of a grammar can enable a seamless transition from one to the other, increasing the spatial awareness of mobile robots.

[6] The paper is accompanied by a video showing the parser in action.

References

[1] MESA Imaging AG, http://www.swissranger.ch/ (13.9.2007).

[2] P. Felzenszwalb and D. Huttenlocher, “Pictorial structures for object recognition,” *Int. J. of Comp. Vision*, vol. 61, pp. 55–79, 2005.

[3] B. Leibe, A. Leonardis, and B. Schiele, “An implicit shape model for combined object categorization and segmentation,” in *Towards Category-Level Object Recognition*. Springer, 2006, pp. 496–510.

[4] G. Medioni and A. Francois, “3d structures for generic object recognition,” *Int. Conf. Pattern Recognition*, vol. 1, pp. 30–37, 2000.

[5] S. Vasudevan, “Spatial cognition for mobile robots : A hierarchical probabilistic concept-oriented representation of space,” Ph.D. dissertation, ASL - ETH Zurich, Switzerland, Diss. ETH No. 17612, 2008.

[6] J. Mundy, *Toward Category-Level Object Recognition*, ser. Lecture Notes in Comp. Sci. Springer, 2006, vol. 4170, ch. Object Recognition in the Geometric Era: A Retrospective, pp. 3–28.

[7] R. B. Fisher, *From Surfaces to Objects - Computer Vision and Three Dimensional Scene Analysis*. John Wiley & Sons Ltd., UK, 1989.

[8] P. J. Besl, *Surfaces In Range Image Understanding*. Springer-Verlag Inc., New York, 1988.

[9] M. Aycinena, L. Kaelbling, and T. Lozano-Perez, “Learning grammatical models for object recognition,” MIT-CSAIL-TR-2008-011, Tech. Rep., 2008.

[10] S.-C. Zhu and D. Mumford, “A stochastic grammar of images,” *Found. and Trends Comp. Graph. and Vis.*, vol. 2, no. 4, pp. 259–362, 2006.

[11] M. Pechuk, O. Soldea, and E. Rivlin, “Learning function-based object classification from 3d imagery,” *Comp. Vis. and Image Understanding*, vol. 110, no. 2, pp. 173–191, 2008.

[12] W. E. L. Grimson, *Object recognition by computer: the role of geometric constraints*. MIT Press, 1990.

[13] S. Gächter, A. Harati, and R. Siegwart, “Incremental object part detection toward object classification in a sequence of noisy range images,” in *IEEE Int. Conf. Robotics and Automation*, 2008.

[14] M. A. Aycinena, “Probabilistic geometric grammars for object recognition,” Master’s thesis, MIT, 2005.

[15] J. Schmittwilken, J. Saatkamp, W. Förstner, T. H. Kolbe, and L. Plümer, “A semantic model of stairs in building collars,” *Photogrammetrie, Fernerkundung, Geoinf.*, vol. 6, pp. 415–428, 2007.

[16] S. Gächter, A. Harati, and R. Siegwart, “Structure verification toward object classification using a range camera,” in *Int. Conf. Intelligent Autonomous Systems*, 2008.

[17] O. D. Faugeras and M. Hebert, “The representation, recognition, and locating of 3-d objects,” *Int. J. Robotics Research*, vol. 5, no. 3, pp. 27–52, 1986.

[18] J. Neira and J. D. Tardos, “Data association in stochastic mapping using the jointcompatibility test,” *IEEE Trans. Robotics and Automation*, vol. 17, no. 6, pp. 890–897, 2001.