2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

A Bottom-up Framework for Construction of Structured Semantic 3D Scene Graph

Bangguo Yu[1], Chongyu Chen[2], Fengyu Zhou[1], Fang Wan[1], Wenmi Zhuang[1] and Yang Zhao[3]

Abstract— **For high-level human-robot interaction tasks, 3D scene understanding is important and non-trivial for autonomous robots. However, parsing and utilizing effective environment information of the 3D scene is not trivial due to the complexity of the 3D environment and the limited ability for reasoning about our visual world. Although there have been great efforts on semantic detection and scene analysis, the existing solutions for parsing and representation of the 3D scene still fail to preserve accurate semantic information and equip sufficient applicability. This study proposes a bottom-up construction framework for structured 3D scene graph generation, which efficiently describes the objects, relations and attributes of the 3D indoor environment with structured representation. In the proposed method, we adopt visual perception to capture the semantic information and inference from scene priors to calculate the optimal parse graph. Afterwards, an improved probabilistic grammar model is used to represent the scene priors. Experiment results demonstrate that the proposed framework significantly outperforms existing methods in terms of accuracy, and a demonstration is provided to verify the applicability in applying to high-level human-robot interaction tasks. The supplementary video can be accessed at the following link: `https://youtu.be/vEWNxnSwmKI`.**

I. INTRODUCTION

3D indoor scene understanding is one of the key factors for intelligent robots to execute high-level human-robot interaction tasks in indoor environment. Real indoor environment commonly contains several rooms (e.g., office, conference) and more than one objects with the same class (e.g., red cup, white cup, the cup on the desk or in the cabinet). To reason the indoor environment, robots have to find an effective way to represent the 3D scene information, such as the objects, relations and attributes. This has attracted much research attention and many achievements have been made. However, the existing systems of 3D scene representation and parsing still have many problems such as long processing time, poor accuracy and limited applicability for robot. These limitations motivate us to consider the 3D indoor scene representation and parsing method.

In the existing literature, there are some works for addressing scene representation and parsing. In principle, the point cloud map [1], probabilistic occupancy map [2] and semantic map [3] [4] can be constructed by SLAM to describe the environment. But these methods contain lots of redundancy information and can not reflect the relations of the objects. Using computer vision, the objects, relations and attributes of image can be acquired to generate semantic scene graphs. The first dataset proposed the concept of scene graph is Visual Genome [5], which provides annotated scene graphs for 100k images. Following Visual Genome dataset, most related works about scene graph have been developed, such as scene graph generation [6] [7], image retrieval [8] and scene synthesis [9]. However, those methods are only limited to 2D scene graph, which is not suitable for moblie robots in 3D scene. With the development of object recognition [10] [11], pose estimation [12] [13] and scene graph generation [6] [7], Kim et al. [14] proposed 3D scene graph structure as an environment model and expanded 2D scene graph into 3D space. In 3D scene graphs, the objects were described by nodes, the relations between the pairs of objects were described by edges. However, this method was limited by single scene and didn't make full use of the objects' 3D information during relation extraction. Armeni et al. [15] extended the Visual Genome dataset to 3D space and ground semantic information. This work confirmed the key of 3D scene graph for robots navigation, but was unable to realize real-time work for robots and needed expensive computing resource. A probabilistic grammar model of spatial And-Or graph (S-AOG) [16] was used to parse the 3D scene [17] [18] and synthesize 3D room layouts [19]. Although these grammar-based methods learn the priors to infer the best result, they didn't focus on the applicability of real-time work for intelligent robots high-level semantic tasks.

Motivated by the above observations, a bottom-up framework of 3D scene graph construction with structured representation is proposed in this paper. In this proposed framework, an improved S-AOG structure is designed to learn from scene dataset and generate the structured scene priors. Next, the visual perception is used to capture the semantic information from 3D scene. Then, the inference from scene priors are adopted to calculate the optimal parse graph as the structured 3D scene graph, which can be used to represent the objects, relations and attributes of the indoor environment. Furthermore, 3D scene graph is utilized to complete the tasks of high-level human-robot interaction navigation.

Our contributions are summarized as follows:

- We propose a bottom-up framework of the structured 3D scene graph construction which efficiently represents the semantic information of 3D indoor scene.
- An improved probabilistic grammar model is designed

The major part of this work is done when Bangguo Yu is with Darkmatter AI Research.

[1]The authors are with the School of Control Science and Engineering, Shandong University, Jinan, 250061, P.R. China. Correspondence should be addressed to Fengyu Zhou: `zhoufengyu@sdu.edu.cn`

[2]Chongyu Chen is with DarkMatter AI Research, Guangzhou, 511458, P.R. China `chenchongyu@dm-ai.cn`

[3]Yang Zhao is School of Electrical Engineering and Automation, Qilu University of Technology (Shandong academy of sciences), Jinan, 250353, P.R. China `zdh1136@gmail.com`

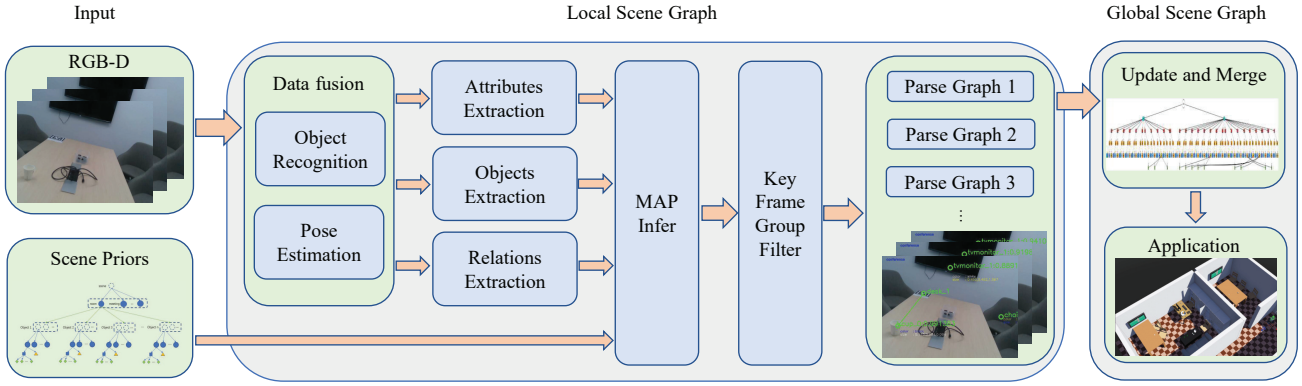


Fig. 1: Overall architecture of the scene graph construction and application framework. The framework with scene priors receives RGB-D image and generates structured 3D scene graph for application.

to denote the scene priors, which can be used to infer the optimal parse graph with structured representation.

- A demonstration is provided to show the applicability of structured 3D scene graph for high-level human-robot interaction navigation.

II. OVERALL SYSTEM ARCHITECTURE

The overall architecture of structured 3D scene graph construction is illustrated in Fig. 1. The input includes the sequence of RGB-D images and scene priors. In the first step, to extract the relations and attributes, the image should be preprocessed by object recognition and pose estimation. Specifically, we use yolov-v3 [10] and apriltag [20] to recognize object information. Limited by the recognition algorithm for big objects like desk or room, we adopt the apriltag to detect them. After object recognition, the object list of category, 2D bounding box and detection probability can be obtained. Using the depth information, the 3D bounding box is estimated by the height and width of 2D bounding box roughly. For the first frame, the relative position of camera can be estimated by SLAM. Specifically, we use ORB-SLAM2 [13]. Next, combining with the outputs of object recognition and pose estimation, the object extraction module removes the duplicate detection of the objects which have the same category and the similar 3D position. The relation extraction module is designed to obtain the relations between objects using 3D position and 3D bounding box. The attribute extraction module is designed to focus on the 3D position, size and color. Then using scene priors, a parse graph pg can be calculated for each image by maximizing a posteriori estimation (MAP), which is the best explanation for extracted objects, attributes and relations. The output of MAP is added to keyframe group filter (KGF) to remove repeated and fake detection. After KGF optimization, parse graph is added to local scene graphs and displayed in real-time. Global scene graph can be obtained by updating and merging the local scene graphs. The structured 3D scene graph is also used to perform the high-level human-robot interaction tasks, like finding a speacial object in multi-room indoor environment according to the human command.

III. STRUCTURED 3D SCENE GRAPH CONSTRUCTION

In this section, we describe the representation of structured 3D scene graph and the main modules of generation framework. These modules can improve the efficiency and reduce the inaccuracy of scene graph.

A. Representation

We follow the [19] method and use S-AOG to construct the scene priors. Spatial And-Or graph is a probabilistic grammar model which represents the hierarchical decompositions from scene (top-level) to objects (bottom-level) by a set of terminal and non-terminal nodes. The terminal ndoes represent objects, non-terminal nodes encode the grammar rules. Contextual relations encode the spatial relations through horizontal links. For meeting the requirement of parsing the environment, an improved structure of S-AOG is proposed as shown in Fig. 2. Formally, the improved S-AOG of a scene is denoted by: $G_s =< S,V,R,E,P >$, where S is the root node of scene, V is the vertex set, R is the production rules, E is the contextual relations, and P is the probability model defined on S-AOG. Different from [19], the hierarchy of scene component is removed and the representation of object category hierarchy is improved. Each member of object category hierarchy is a finite category set of Or-node, in which each Or-node can be alternatived to the instantiated object or empty. Therefore, the set of Or-node can represent different number of objects.

In principle, a scene configuration is represented by a parse graph pg, where the terminal nodes are the objects in the scene with their attributes and relations. By selecting a child node from the Or-nodes, parse graph can be the instantiation of the S-AOG. A pg can be decomposed as $pg = (pt, E_{pt})$, where pt is the hierarchical structure of pg, E_{pt} is a part of contextual relations E on parse tree. For one scene, we extract the object features Γ_O, spatial relations features Γ_S and attribute features Γ_A from a group of input RGB-D image $T_t = \{I_t^k\}_{k=1,\ldots,N}$, where t is the KFG group index, N represents the number of image in the KFG group. Based on $\Gamma =< \Gamma_S, \Gamma_O, \Gamma_A >$ and G_s, the

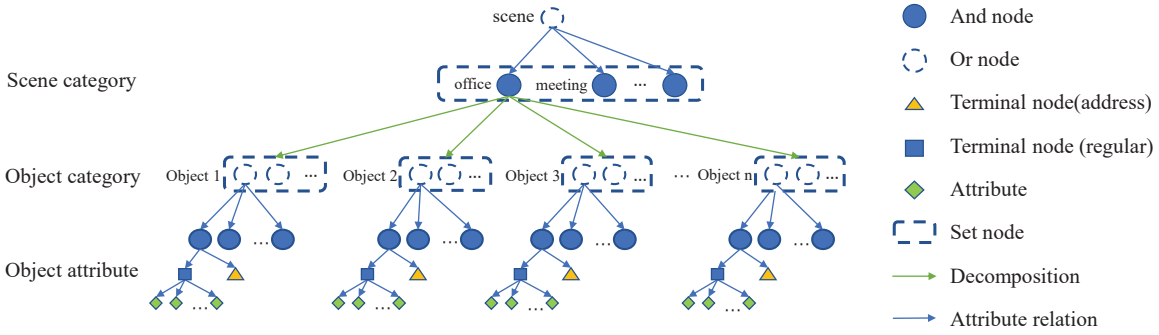


Fig. 2: The improved structure of S-AOG priors. And-node represents a decomposition of a large entity into smaller components (e.g., from office into desk). Or-node describes an alternative branch of possible objects (e.g., a scene can be a living room or office, a bed can be a big bed or small bed). Address node is a pointer to other ternimal nodes, representing the horizontal link of contextual relations, and regular node is a spatial entity in the scene with attributes.

local scene graphs can be inferred and defined as $PG_t = \{pg_t^k\}_{k=1,\dots,N}$ after the KFG optimization. The global scene graph PG_{global} can be obtained by updating and merging the set $\{\dots PG_{t-1}, PG_t, PG_{t+1} \dots\}$.

B. MAP Probabilistic Formulation

In this section, we introduce the probabilistic model defined on this framework. Given the image features Γ and scene priors G_s, the posterior probability of a parse graph sequence PG is defined as:

$$\begin{aligned} p(PG|\Gamma, G_s) &\propto p(\Gamma|PG)\,p(PG|G_s) \\ &= p(\Gamma_O, \Gamma_S, \Gamma_A|PG)\,p(PG|G_s) \\ &= \underbrace{p(\Gamma_O|PG)}_{\text{object}}\underbrace{p(\Gamma_S|PG)}_{\text{spatial}}\underbrace{p(\Gamma_A|PG)}_{\text{attribute}}\underbrace{p(PG|G_s)}_{\text{grammar prior}} \end{aligned} \tag{1}$$

The first three terms are likelihood terms for objects, spatial relations and attributes, respectively. The last term is a prior probability of the parse graph given the grammar G_s of scene.

1) Likelihood of Parse Graph: In this part we use [21] method to represent the likelihood. We assume that both the prior probability for the image $P(\Gamma)$ and $P(PG)$ for object, relation and attribute are uniformly distributed. So the likelihood of object $P(\Gamma_O)$ given a parse graph PG is defined as:

$$\begin{aligned} p(\Gamma_O|PG) = p(\Gamma_O|PG_O) &= \frac{p(PG_O|\Gamma_O)\,P(\Gamma_O)}{P(PG_O)} \\ &\propto p(PG_O|\Gamma_O) = \prod_{k=1}^{N} p\left(pg_O^k|\Gamma_O^k\right) \end{aligned} \tag{2}$$

where $p\left(pg_O^k|\Gamma_O^k\right)$ is the detection probability of object in image I^k. Similarly, the likelihood of spatial relation $P(\Gamma_S)$ can be expressed as:

$$p(\Gamma_S|PG) \propto p(PG_s|\Gamma_S) = \prod_{k=1}^{N} p\left(pg_S^k|\Gamma_S^k\right) \tag{3}$$

where $p\left(pg_S^k|\Gamma_S^k\right)$ is the detection probability of relations. The likelihood of attribute $P(\Gamma_A)$ can be denoted as:

$$p(\Gamma_A|PG) \propto p(PG_a|\Gamma_A) = \prod_{k=1}^{N} p\left(pg_A^k|\Gamma_A^k\right) \tag{4}$$

where $p\left(pg_A^k|\Gamma_A^k\right)$ is the detection probability of attributes.

2) Grammar Prior of Parse Graph: The Grammar prior of parse graph can be described by three subsets: prior of object, relation and attribute:

$$\begin{aligned} p(PG|G_s) &= p(PG_O, PG_S, PG_A|G_s) \\ &= p(PG_O|G_s)\,p(PG_S|G_s)\,p(PG_A|G_s) \\ &= \prod_{k=1}^{N} p\left(pg_O^k|G_s\right)\prod_{k=1}^{N} p\left(pg_S^k|G_s\right)\prod_{k=1}^{N} p\left(pg_A^k|G_s\right) \end{aligned} \tag{5}$$

where $p(pg_O|G_s)$, $p(pg_S|G_s)$ and $p(pg_A|G_s)$ are the prior of object, spatial relation and attribute.

3) Inference: For a group of images, the PG is found for each scene that best explains the extracted features Γ by maximizing the posterior probability:

$$\begin{aligned} PG &= \underset{PG}{\operatorname{argmax}}\, p(PG|\Gamma, G_s) \\ &= \underset{PG}{\operatorname{argmax}}\, \{p(\Gamma_O|PG_O)\,p(\Gamma_S|PG_S)\,p(\Gamma_A|PG_A) \\ &\qquad p(PG_O, PG_S, PG_A|G_s)\} \end{aligned} \tag{6}$$

C. Learning

The learning of the S-AOG consists of two main parts: 1) learn the S-AOG grammar structure of each scene, and 2) learn the parameters of the S-AOG, including the branching probabilities of the Or-nodes.

1) Grammar Structure: Due to the different prior distributions of different scenes, the distributions of seven scenes is learned from the indoor scene dataset, including bedroom, living room, kitchen, office, bathroom, dining room and conference room. Each scene has its own distribution to the objects. Each object also learns its attributes and relations with others.

2) Parameter Learning: The branching probability of Or-node is simply given by the frequency of each alternative choice:

$$(O_A \to O_{\beta_i}) = \frac{\#\left(O_A \to O_{\beta_i}\right)}{\sum_{j=1}^{n(O_A)} \#\left(O_A \to O_{\beta_i}\right)} \tag{7}$$

where O_A is the Or-node, O_{β_i} is one sub-node of O_A, $\#\left(O_A \to O_{\beta_i}\right)$ is the number of times $O_A \to O_{\beta_i}$ occurred in all dataset [16]. We compute the objects probabilities, attributes probabilities and relations probabilities with others in each scene.

D. Spatial Relation Extraction

3D space is more stable and invariant. According to the results of analyzing the Visual Genome dataset [7], we find that a number of semantic relations are concerned with person. Without person, the spatial relation makes up a large proportion in dataset. Therefore, we consider no person in indoor scene and raise three abstract spatial relations from dataset including support, parent and adjacency relations. Three relation transforms are defined such that the specific relations in dataset can be mapped in abstract relations which is Table I.

TABLE I: Transforms from Specific Relation to Abstract Relation.

abstract	specific
Support	on, over, under, cover, handing on, on bottom of, on top of ...
Parent	of, with, inside of, has, has a, in, sitting on, part of ...
Adjacency	beside, next to, around, at, in front of, behind ...

1) Support relation: Support is the most common relation in 3D indoor scene. Since the noise of camera makes it difficult to get the accurate position and size of the objects exactly. Hence, the possibility of support relation is denoted according to human knowledge:

1) The supported object is located on the supporting object (e.g., a cup on a desk), and the size of supported object is commonly larger than supporting object's size.
2) The lower surface of supported object is close to top surface of supporting object.

Then we judge the support relation between two objects $o1, o2$ by sampling the points of $o1$ on xy plane and get the statistic proportion $P_{xy}(o1,o2)$ of how many points are in $o2$'s range xy plane. For the second condition, $P_h(o1,o2)$ is defined to measure the high error of $o1$'s low surface and $o2$'s top surface

$$P_h(o1,o2) = \begin{cases} \cos\left(\frac{\pi h_e}{2H_T}\right), & h_e < H_T \\ 0, & h_e \geq H_T \end{cases} \tag{8}$$

where H_T is the threshold of the high error from $o1$'s low surface to $o2$'s top surface, h_e is the detection of the high error. Finally, the probability $P_S(o1,o2)$ of the support relation between $o1$ and $o2$ can be represented as

$$P_S(o1,o2) = P_{xy}(o1,o2) \cdot P_h(o1,o2) \tag{9}$$

2) Parent Relation: Parent relation represents that one object contains another object in 3D space. The child object can be a part of parent object (e.g., a bottle and a lid) or just exists in parent object (e.g., a milk in a refrigerator). A volume sample frequency $P_P(o1,o2)$ is adopted to measure the probability of parent relation between $o1$ and $o2$.

3) Adjacency Relation: Considering the stardard of adjacency relation is subjective in 3D space, and the judgement of adjacency depending on the current view of camera, we only judge the relation in spatial state by finding the closest object of what we want, like "the mouse next to the keyboard".

After MAP module, the relation whose probability lower than a threshold value are deleted from the graphs.

E. KeyFrame Group Filter

Due to the process of detection unavoidably produce the erroneous results, and the object recognition do not detect every object perfectly, the keyframe group is imported to remove repeated and fake detection. Two types of groups are defined:

1) Active KeyFrame Group: Active keyframe group accepts the image frame results when the object detection module outputs. When one frame is captured, the objects, relations and attributes can be calculated as outputs and taken in active keyframe group. When the number of frame reaches N, copy the group to optimal keyframe group for optimizing and removing before half part of active keyframe group. When the next frame result comes in, index it as the $N/2+1$.

2) Optimal KeyFrame Group: Optimal keyframe group inherit active keyframe group when active group is full. After receiving the date, we proceed as the following:

- Filter the fake detection which appears discrete and infrequent.
- After removing repeated and fake detection, the mean values of the position and size for each object are calculated.
- Improve the relations and attributes using optimized results for each object.
- Take the optimized results pg in the local scene graphs.

Afterwards, the optimized local scene graphs can be obtained because of reducing the inaccurate effectively.

F. Local and Global Scene Graph Construction

Each frame I^i only captures an incomplete part of scene, and the result can be represented by pg^i which can be regarded as view-centric parse graph. When a sequence of pg is jointed, the scene-centric scene graph can be got. The illustration is shown in Fig. 3. When the number of frame for each group reach N, the local scene graphs can be represented by $PG = \left\{pg^k\right\}_{k=1,\ldots,N}$. With the development of local

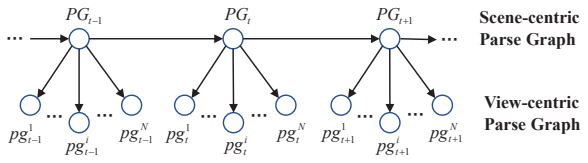


Fig. 3: Parse graph of scene-centric and view-centric

scene graphs, the global scene graph PG_{global} can be obtained by updating and merging the set $\{\ldots PG_{t-1}, PG_t, PG_{t+1}\ldots\}$.

To calculate PG_{global}, the similarity score is proposed to merge the same object node from the set of PG. The color

distribution is unstable in real scene duo to illumination and view. We define that the similarity score $Score(o1,o2)$ of two objects $o1$, $o2$ depends on the category $S_c(o1,o2)$ and 3D position $S_p(o1,o2)$. The category similarity $S_c(o1,o2)$ is denoted as

$$S_c(o1,o2) = \begin{cases} max\{p_o(o1), p_o(o2)\}, & C_{o1} = C_{o2} \\ 0, & C_{o1} \neq C_{o2} \end{cases} \quad (10)$$

where $p_o(o1)$ and $p_o(o2)$ are the detection probabilities of objects, C_{o1} and C_{o2} are the categories of objects. For 3D position similarity, the $S_p(o1,o2)$ is represented as

$$S_p(o1,o2) = max\left\{1 - \frac{abs(Pose(o1) - Pose(o2))}{\alpha \cdot max\{Size_M(o1), Size_M(o2)\}}, 0\right\} \quad (11)$$

where $Pose(o1)$ and $Pose(o2)$ are the 3D position of $o1$ and $o2$, $Size_M(o1)$ and $Size_M(o2)$ are the largest length of objects' size, α is a parameter which depends on the size of object to improve the detection accuracy. The smaller the object's size, the larger the α. Therefore, the similarity score $Score(o1,o2)$ can be described as

$$Score(o1,o2) = \beta S_c(o1,o2) + (1-\beta) S_p(o1,o2) \quad (12)$$

When the similarity score $Score(o1,o2)$ higher than a threshold, the $o1$ and $o2$ can be detected as the same object.

IV. EXPERIMENT

In this section, we focus on the accuracy of structured 3D scene graph construction, and the applicability of scene graph based robots application. First, the proposed method is evaluated in real indoor environment by reporting the performance. Next, the scene graph is used in a high-level human-robot interaction navigation.

A. Structure 3D Scene Graph Construction

1) Experiment Details: We choose two real indoor scenes which consist of an office room and a conference room to assess the algorithm. The *realsense* camera is used to obtain the RGB-D image. The resolutions of the image frames are 640 * 480 (color) and 640 * 480 (depth) with 30-Hz frame rate, and all the color and depth images are aligned. The camera parameters are provided as well. Finally, the number of image sequence we get is 7103.

In our experiments, the scene priors are learned from Visual Genome dataset. Our system runs in real time and mixes other modules in ROS. C++ and python are used to integrate our work with other techniques. All experiments are carried out in Inter Core i7-7700HQ (four cores @2.80GHz), 8GB RAM and GTX1060 with Max-Q. For the keyframe group, N is set as 50. The threshold of object and relation after MAP module is set as 0.05, and the threshold of support relation H_t is set as the supported object's high. The similarity score threshold is set as 0.5, and β is set as 0.2. The global scene graph is saved in the JSON format and displayed in real-time.

2) Algorithms: For assessing the accuracy of the scene graph, a few baseline methods are shown as follow:

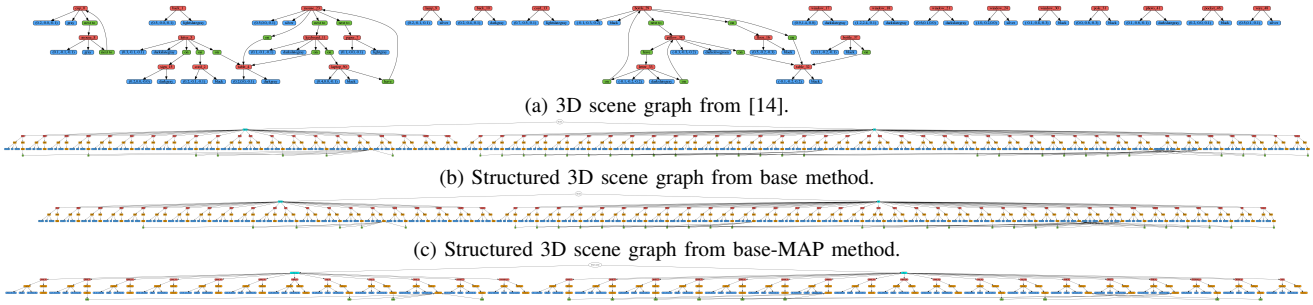
- The 3D scene graph construction framework [14] is used as baseline.
- The attributes extraction, objects extraction and relation extraction modules are adopted to construct the 3D scene graph as the base method.
- The base-MAP method add scene priors in MAP probabilistic formulation to infer the optimal scene graph based on base method.
- The last method is the full method, the proposed structured 3D scene graph construction framework, which adds keyframe group filter to the base-MAP method.

3) Evaluation Framework: Since the attribute in dataset is incomplete and random, the prior distribution of attribute is poor, the attribute is extracted from scene directly. In our experiment, we only assess the performance of object and relation. A human judgement metric is used to evaluate the accuracy of each method. We follow the [14] method to recruited five experiment participants and gathered five responses for each graph. For each resulting graphs, the number of false entities (objects and relations) and missing entities is counted by the participants to calculated the precision (PR), recall (RE) and F1-measure (F1-M) for each method. We measured the runtime of each method to compare the efficiency as well.

TABLE II: Results of Comparative Study.

Method	Object			Relation			Average Runtime (sec/frame)
	PR	RE	F1-M	PR	RE	F1-M	
3D scene [14]	0.684	0.464	0.553	0.545	0.375	0.444	3.267
base	0.450	0.941	0.606	0.483	0.875	0.622	0.101
baseMAP	0.635	0.943	0.759	0.560	0.875	0.683	0.103
full	**0.968**	**0.882**	**0.923**	**0.929**	**0.867**	**0.897**	0.117

4) Result: The quantitative results of the comparision study are reported in TABLE II. The scene graphs results of 3D scene [14] and our methods are shown in Fig. 4. It can be seen from the base method that the precision is too low and the recall is high by erroneous and repetitive detection of the objects and relations. Similarly, the resulting scene graphs from base-MAP method contain repetitive detection of the objects and relations, so the recall is the same with base method. However, the precision of object and relation improves a little. Comparison between base and base-MAP verifies the performance of MAP in boosting the precision. But affected of the repetitive data, MAP can only remove the low probabilistic objects and relations. The full method adds the KGF to the base-MAP method and makes a successive improvement of the accuracy performance. Although a little objects and relations are missing in the resulting scene graphs (the recall is reduced), KGF rejects most of the repetitive objects and relations, and improves the precision of the framework. The number of erroneous detection in scene graphs from full is much less than base-MAP. From TABLE II, we can alse see that our method outperforms the 3D scene



(a) 3D scene graph from [14].

(b) Structured 3D scene graph from base method.

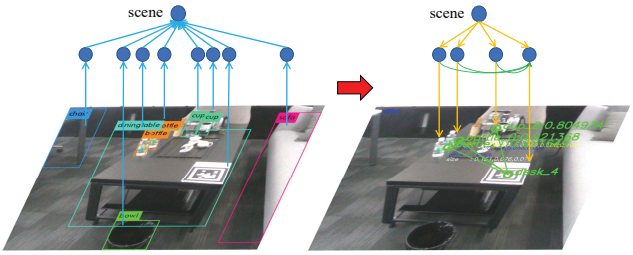
(c) Structured 3D scene graph from base-MAP method.

(d) Structured 3D scene graph from full method.

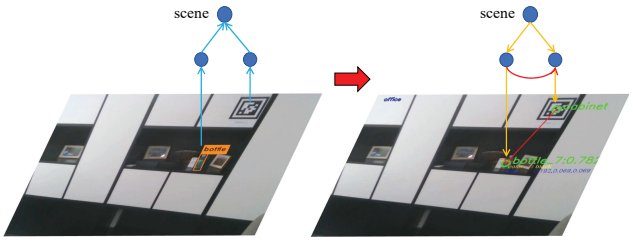
Fig. 4: The display of the experiment results. Wathet, red, green, blue and orange denote the scene, object, relation, attribute and terminal node, respectively. From base method to full method, the performance of the results become more accuracy.

[14] in terms of precision, recall, F1-measure and average runtime.

In summary, the experiment results verify the performance of the proposed structured 3D scene graph generation framework in accuracy and efficiency. The performance of visual perception and inference are shown in Fig. 5. Each pair of images are the results of object recognition module and inference module. In the step of visual perception, objects, relations and attributes are detected completely. After inference, the entity whose probability lower than the related threshold is deleted from the graphs. We also use the prior relations to enhance the detection result by updating objects probabilities, but the module can not improve the performance. We suspect the prior distributions of relations are not enough to assure the performance.



(a)

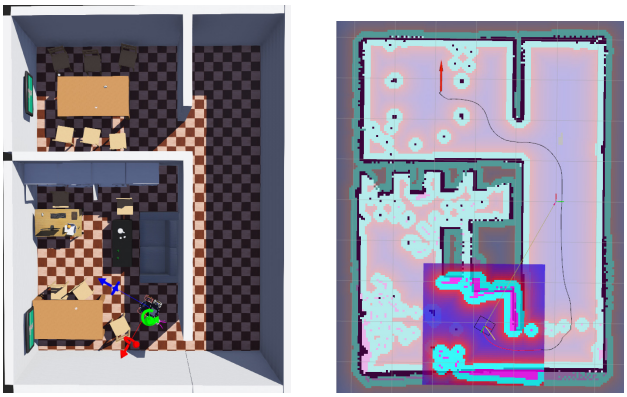


(b)

Fig. 5: The results of the visual perception and the inference from scene priors. The blue lines represent the process of visual perception, orange lines represent the inference. After inference, the optimal objects and relations can be extracted. The green line and red line describe support and parent relations. (a) shows the support relation that two bottles and one cup are on the desk. (b) shows the parent relation that bottle is in cabinet.

B. Applicability Demonstration

To verify the applicability, the structured 3D scene graph is used in human-robot interaction navigation.



(a) (b)

Fig. 6: Robots navigation in webots simulation. (a) The conference and office room in webots simulation environment. (b) The 2D map construction and navigation using SLAM and other ros packages.

1) Experiment Details: A conference and an office room simulation environment modeled from the real environment are constructed, so that the structured 3D scene graph from the real scene can be directly used in the simulation. We implement the simulation environment using Webots. The Pioneer 3-AT is used in webots simulation which has 4 motors and 16 sonar sensors with power supply, and kinect camera and laser are added to it. The demonstration is implemented in ROS as well.

2) Result: In the first step, the 3D scene graph construction framework is run in real environment and then scene graphs are obtained through camera. The 2-D map is built by Pioneer 3-AT to provide location in simulation. The simulation and map are shown in Fig. 6 and the specific scenes in conference and office are shown in Fig. 7. Then, the scene graphs can be utilized to navigate in the same scene simulation environment according to the human command which containes the object and the relations. Based on the semantic command, we use graph search algorithm to find

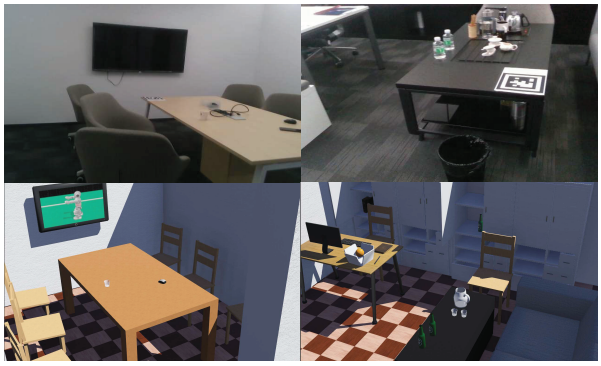


Fig. 7: The real scene and simulation scene in conference and office.

the semantic goal in the structured scene graphs quickly. According to the Fig. 6, Pioneer 3-AT is in office. After receiving the human command "the mouse on the desk and in the conference", Pioneer 3-AT can extract the object "mouse" and relations "on the desk" and "in the conference". According to these objects and relations, Pioneer 3-AT can search the goal in 3D scene graphs and confirm the position shown as the red arrow. Then ros packages are used to plan the path to conference and the navigation can be executed.

Overall, the demonstration verify the applicability of high-level human-robot interaction tasks using structured 3D scene graph in multi-room environment. In traditional methods, the position of navigation is given by humans and the robot can not understand any semantic information. The application of 3D scene graphs can fuse geometry and semantic information, which help robots to execute the semantic navigation tasks intelligently.

V. CONCLUSIONS AND FUTURE WORK

This paper presented a bottom-up framework of structured 3D scene graph generation from RGB-D image and scene priors. The proposed framework contains an improved grammar model, which is used to learned from scene dataset and describe the scene priors. The visual perception is used to capture the objects, relations and attributes from 3D scene, and the inference is adopted to obtain the optimal parse graph from scene priors. We implement our framework in a real indoor scene to demonstrate its accuracy in representing the semantic information, and the applicability of the high-level human-robot interaction navigation tasks is also verified in multi-room scene using human command. For further research, the dynamic semantic segmentation system can be used to improve the visual perception. The structure of the S-AOG is easy to expand, with the hierarchy of the S-AOG addition can help robots to execute more complex semantic tasks, such as autonomous exploration in unseen scene and the improvement of the visual perception using context grammar.

REFERENCES

[1] M. Labbe and F. Michaud, "Appearance-Based Loop Closure Detection for Online Large-Scale and Long-Term Operation," *IEEE Transactions on Robotics*, vol. 29, pp. 734–745, jun 2013.

[2] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2D LIDAR SLAM," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1271–1278, IEEE, may 2016.

[3] C. Yu, Z. Liu, X.-j. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1168–1174, IEEE, oct 2018.

[4] M. Grinvald, F. Furrer, T. Novkovic, J. J. Chung, C. Cadena, R. Siegwart, and J. Nieto, "Volumetric Instance-Aware Semantic Mapping and 3D Object Discovery," *IEEE Robotics and Automation Letters*, vol. 4, pp. 3037–3044, jul 2019.

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision*, vol. 123, pp. 32–73, may 2017.

[6] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to Compose Dynamic Tree Structures for Visual Contexts," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6621, IEEE, jun 2019.

[7] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural Motifs: Scene Graph Parsing with Global Context," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5831–5840, IEEE, jun 2018.

[8] J. Johnson, R. Krishna, M. Stark, L. J. Li, D. A. Shamma, M. S. Bernstein, and F. F. Li, "Image retrieval using scene graphs," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 3668–3678, 2015.

[9] Y. Zhou, Z. While, and E. Kalogerakis, "SceneGraphNet: Neural Message Passing for 3D Indoor Scene Augmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7383–7391, IEEE, oct 2019.

[10] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," in *arXiv e-prints*, apr 2018.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[12] T. Qin and S. Shen, "Online Temporal Calibration for Monocular Visual-Inertial Systems," *IEEE International Conference on Intelligent Robots and Systems*, pp. 3662–3669, 2018.

[13] R. Mur-Artal, J. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[14] U.-H. Kim, J.-M. Park, T.-J. Song, and J.-H. Kim, "3-D Scene Graph: A Sparse and Semantic Representation of Physical Environments for Intelligent Agents," *IEEE Transactions on Cybernetics*, pp. 1–13, 2019.

[15] I. Armeni, Z.-Y. He, A. Zamir, J. Gwak, J. Malik, M. Fischer, and S. Savarese, "3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5663–5672, IEEE, oct 2019.

[16] S. C. Zhu and D. Mumford, "A stochastic grammar of images," *Foundations and Trends in Computer Graphics and Vision*, vol. 2, no. 4, pp. 259–262, 2006.

[17] X. Liu, Y. Zhao, and S.-c. Zhu, "Single-View 3D Scene Parsing by Attributed Grammar," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 684–691, IEEE, jun 2014.

[18] Y. I. B, Y. Kobayashi, and K. Takahashi, *Computer Vision – ECCV 2018*, vol. 11211 of *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2018.

[19] S. Qi, Y. Zhu, S. Huang, C. Jiang, and S. C. Zhu, "Human-Centric Indoor Scene Synthesis Using Stochastic Grammar," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5899–5908, 2018.

[20] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2016-Novem, pp. 4193–4198, 2016.

[21] S. Qi, S. Huang, P. Wei, and S. C. Zhu, "Predicting Human Activities Using Stochastic Grammar," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 1173–1181, IEEE, oct 2017.