OG-SGG: Ontology-Guided Scene Graph Generation for Semantic Scene Representation in Robotic Applications Extended Abstract

Fernando Amodeo[1] and Fernando Caballero[2] and Natalia Díaz-Rodríguez[3] and Luis Merino[1]

I. Motivation

Telepresence robots allow people to remotely interact with others. One of the main applications of these robots is in assistance tasks. For example, they allow disabled people to attend events remotely, or caregivers to interact remotely with people under their care. In particular, this work considers the application of telepresence robots for elderly care [5] (see Fig. 1).

However, controlling these systems is a complex task. The human controller needs to focus on both low-level tasks (such as controlling the robot) and high-level tasks (such as maintaining a conversation) at the same time; and this can lead to a cognitive overload, therefore reducing the attention that is given to the high-level tasks [7]. For this reason, being able to interact with these robots using only higher-level commands is preferable (i.e. *Approach a given object*, *Follow a given person*, etc.) – in this case the robot would then be in charge of low-level control. This is in fact a necessity if one considers visually-impaired people as users of the robotic system.

In all these cases, the robot needs to extract and provide semantic information about the scenario so that the scene can be described in human terms to the users, and they can in turn indicate the robot where to go next for interactions. This information is represented in the form of a scene graph based on an ontology, which then allows the robot to perform automated reasoning and fulfill other downstream tasks. While the motivation behind the work is semantic level control of telepresence robots, the same ideas can be used for many other downstream tasks involving human-robot interaction.

This work surveys existing research on the automatic generation of those scene graphs, such as [8], [6], [10], [9], and investigates their application to telepresence robots.

*This work is partially supported by Programa Operativo FEDER Andalucia 2014-2020 and Consejeria de Economía y Conocimiento (PY20_00817 DeepBot) and the project PLEC2021-007868, funded by MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR. N. Díaz-Rodríguez is supported by the Spanish Government Juan de la Cierva Incorporación contract (IJC2019-039152-I) and Google Research Scholar Programme.

[1]Service Robotics Laboratory, Universidad Pablo de Olavide, Seville, Spain. famozur@upo.es, lmercab@upo.es

[2]Service Robotics Laboratory, Universidad de Sevilla, Spain. fcaballero@us.es

[3]Computer Science and Artificial Intelligence Dept., Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Spain. nataliadiaz@ugr.es

Fig. 1: A telepresence robot for elderly care. The robot in this picture is used to remotely assist with the activities of a day care centre.

These methods extract a semantic graph for a given image, composed of the main objects present in the scene and the relations between them (e.g. Fig. 3). The main problem identified in these solutions is their inability to take existing "expert" knowledge about the domain world into account. Moreover, the existing available datasets for training the models (such as Visual Genome [3]) are quite noisy, biased and too general, as a result of how they were collected.

Therefore, a systematic way to specialize these datasets according to domain needs and improving the semantics of the output of the model must be devised. The main goal of this work is thus finding a way to reuse and repurpose existing scene graph generation models and datasets for specific robotic applications, and applying additional techniques that take into account existing domain knowledge of the application, so that we can improve the performance of a machine learning model within the reduced scope of a given problem and ontology. This is precisely what the proposed OG-SGG methodology sets out to do.

II. Key Methods

The proposed pipeline (see Fig. 2) consists of three main components: a scene graph generation network, a training dataset filtering and augmentation process, and a network output post-processing process. These last two processes, the core of this work's contribution, make use of pre-existing expert knowledge defined in the domain ontology, while the network itself can be adapted from the existing state of the art with minimal changes according to needs (such as efficiency).

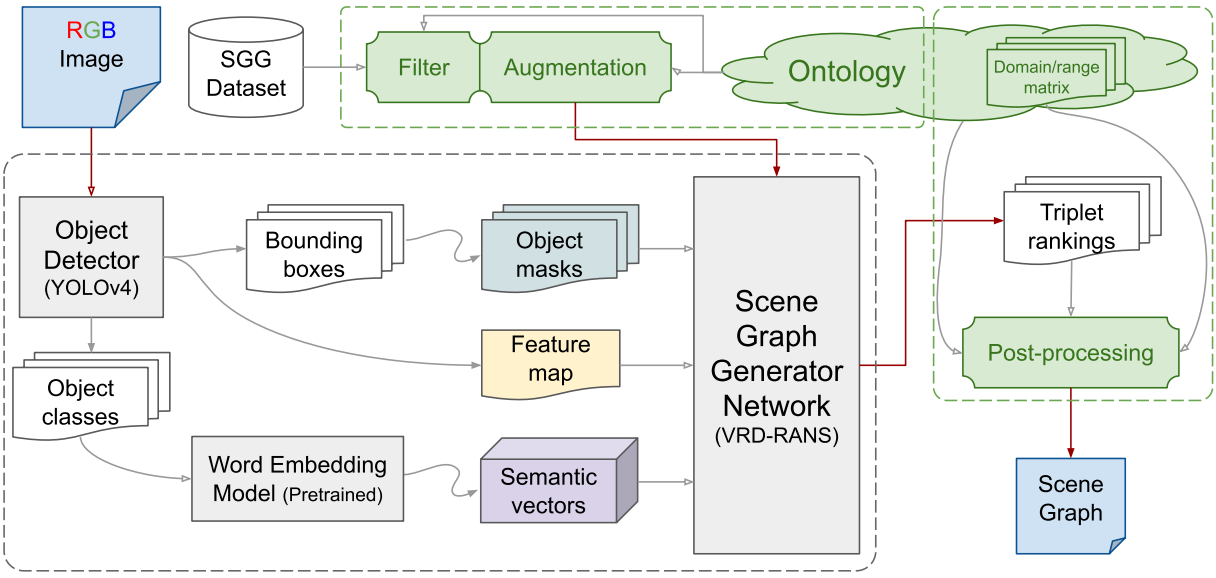


Fig. 2: Full OG-SGG pipeline. The diagram shows three main components in full detail, along with both internal and external data flow. The proposed ontology-aware additions are highlighted in green, which include the filter/augmentation component, and the post-processing component. A pre-existing object detection and scene graph generation component is also included in the pipeline.

Dataset		**R@K** (k = 1)			**R@K** (k = 8)			**mR@K** (k = 1)			**mR@K** (k = 8)		
Training	Test	20	50	100	20	50	100	20	50	100	20	50	100
VG-SGG	TERESA	27.0	34.7	41.9	23.8	34.8	51.1	19.1	30.6	36.4	29.3	42.7	57.0
VG-SGG (filtered)	TERESA	**44.7**	**47.9**	**53.2**	**46.5**	**53.4**	**66.3**	**44.0**	**51.2**	**53.6**	**44.7**	**53.9**	**60.5**
VG-SGG	AI2THOR	**19.1**	24.4	32.1	**21.9**	27.2	40.5	7.0	11.8	16.2	9.4	15.2	24.6
VG-SGG (filtered)	AI2THOR	17.3	**28.7**	**38.7**	18.9	**33.1**	**52.5**	**11.6**	**21.1**	**27.9**	**13.8**	**28.0**	**48.3**
VG-indoor	TERESA	26.1	33.7	40.2	26.0	41.4	56.1	10.5	20.6	25.2	19.8	35.5	53.6
VG-indoor (filtered)	TERESA	**43.6**	**45.3**	**51.8**	**44.2**	**51.5**	**62.8**	**45.0**	**52.9**	**59.6**	**47.1**	**57.7**	**69.3**
VG-indoor	AI2THOR	18.7	21.4	26.0	22.1	26.5	37.4	9.3	17.7	23.0	12.2	22.3	31.5
VG-indoor (filtered)	AI2THOR	**21.4**	**25.7**	**31.4**	**24.7**	**32.9**	**45.9**	**14.5**	**24.1**	**30.8**	**16.9**	**32.5**	**48.0**

TABLE I: Quantitative results.

The scene graph generation network, called VRD-RANS, was adapted from an existing work [8]; and reimplemented from scratch due to a lack of publicly available code. This network was chosen because of its use of semantic vectors (improving generalization), a single global feature map instead of the more common RoI pooling approach (improving runtime efficiency for embedded GPUs used in robotics), and a novel training strategy that has built-in data augmentation in the form of negative sampling.

The training dataset filtering and augmentation method can be summarized as a process that takes a source dataset and applies a series of ontology-guided transformations. These include mapping predicates to object properties in the ontology, discarding (filtering) triplets without a suitable match, and extracting (augmentation) new inferred triplets according to the axioms defined in the ontology. Likewise, output post-processing prunes triplets that are deemed as inconsistent with said axioms, which are then solved by selecting the highest scoring candidate triplets that form a consistent graph.

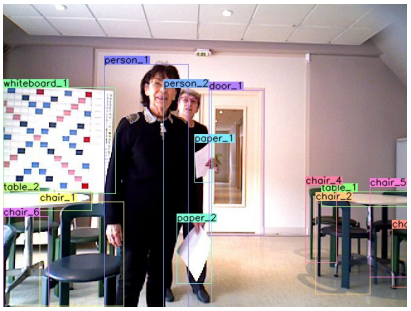
III. Key Results

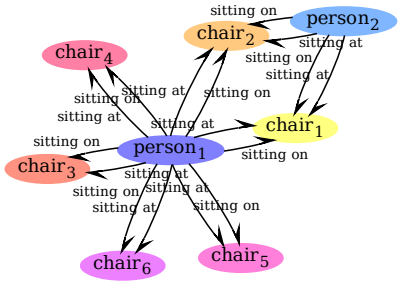
Our code is publicly available online[1].

In order to evaluate the effects of the ontology-guided scene graph generation (OG-SGG) framework, we applied it to a telepresence robotics use case. Specifically, we utilized data from the TERESA [5] European Project, which involved a telepresence robot being used within an elderly day-care centre. Additionally, we also decided to test another similar but different robotics scenario provided by the AI2THOR framework [2] – a near photo-realistic interactable framework for embodied AI agents, with the goal of facilitating the creation of visually intelligent models and pushing the research forward in that domain.

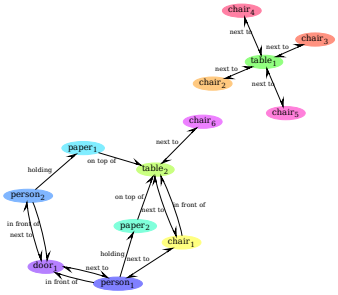
We carried out several experiments in order to prove that OG-SGG delivers the desired performance improvements in a specific application, when the source dataset used for training is unrelated to the subject matter (robotics), by virtue of being collected from general images downloaded from the internet. Our experiments were carried out by training the scene graph generation network on a series of training dataset splits extracted from the Visual Genome (VG) dataset, and

[1] https://github.com/robotics-upo/og-sgg

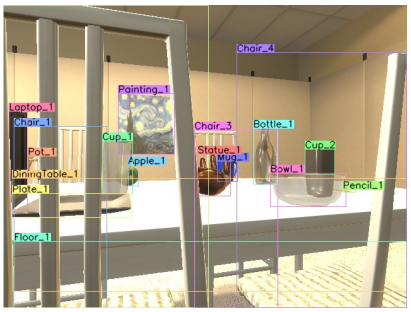


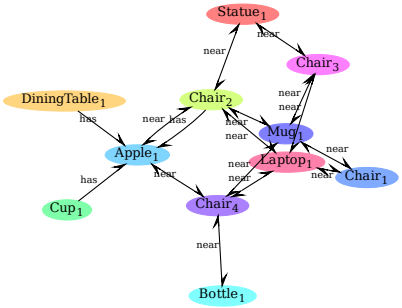


Person 1 is sitting on chairs 1, 2, 3, 4, 5 and 6.
Person 2 is sitting on chairs 1 and 2.



Person 1 and 2 are in front of a door.
Person 1 is holding a piece of paper.
Person 1 is next to chair 1.
Person 2 is holding a piece of paper.





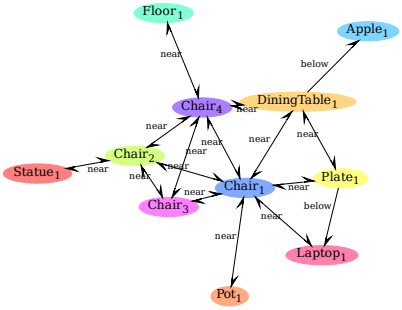


Fig. 3: Qualitative examples using TERESA and AI2THOR (top to bottom)
Left: Image with object annotations. Center: Network trained on VG-SGG. Right: Network trained on VG-SGG (filtered).

evaluating it on a completely different dataset consisting of images extracted from our desired source (TERESA or AI2THOR). The original VG data was used both unmodified and preprocessed with our proposed filtering/augmentation method, using a simple ontology created for each test dataset. In addition, a subset of VG, dubbed "VG-indoor" (only containing indoor images) was also used during the experiment, in order to observe the effects of using a smaller, coarsely filtered source dataset.

We make use of the standard metrics for scene graph generation [4], [6], namely Recall at K (**R@K**) and Mean Recall at K (**mR@K**). These metrics calculate the percentage of the ground truth relationship annotations that are correctly generated by the system among the **K** highest scoring triplets. **mR@K** in addition calculates separate values for each predicate type, which are then averaged; this results in a less biased picture of the generalization capability of the system.

Table I shows quantitative results using both training and both testing datasets. It is worth pointing out that the reported results for filtered training datasets also contain the post-processing filter. Full ablation test results can be found in our paper [1], which also confirms the major improvements brought by filtering at either end. A significant improvement in all **mR@K** variants can be seen when using filtered datasets, which indicates greater generalization capability in the network. As for **R@K**, more modest improvements can be seen mainly in variants with larger **K** and **k**. Smaller values of those parameters bring about no discernible improvement, possibly because more frequent predicates are more heavily weighted in those variants. These more frequent predicates will be overfitted in models with smaller generalization capability.

Fig. 3 shows two selected qualitative examples. The graphs were generated by running the images through the model and picking the 16 highest scoring generated triplets. The pruning effects of the post-processing rules can clearly be seen – certain structures arise in the new graphs, such as people holding cups or multiple chairs being next to a table. Limitations can also be seen, such as the lack of depth perception, or trouble with higher order reasoning (i.e. the network being unable to understand that an object can either be on a table or held by a person, but not both).

IV. Conclusions

While existing scene graph generation networks can theoretically output any combination of triplets, OG-SGG is able to leverage the ontology to reduce the set of possibilities and thus improve the quality of the generated scene graphs. Another important observation is that only a small amount of effort had to be spent in engineering an ontology for the experiment in order to obtain these results. It can be explained that OG-SGG leverages the effect that biased datasets have on neural networks, precisely by creating a new version of the dataset that is *biased* in favor of existing prior knowledge.

REFERENCES

[1] Fernando Amodeo, Fernando Caballero, Natalia Díaz-Rodriguez, and Luis Merino. OG-SGG: Ontology-Guided Scene Graph Generation. A Case Study in Transfer Learning for Telepresence Robotics. *arXiv*, 2022.

[2] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[4] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual Relationship Detection with Language Priors. In *European Conference on Computer Vision*, 2016.

[5] K. Shiarlis, J. Messias, M. van Someren, S. Whiteson, J Kim, J Vroon, G. Englebienne, K. Truong, V. Evers, N. Perez-Higueras, I. Perez-Hurtado, R. Ramon-Vigo, F. Caballero, L. Merino, J. Shen, S. Petridis, M. Pantic, L. Hedman, M. Scherlund, R. Koster, and H. Michel. TERESA: A Socially Intelligent Semi-autonomous Telepresence System. In *Workshop on Machine Learning for Social Robotics at ICRA-2015 in Seattle*, 2015.

[6] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang. Unbiased Scene Graph Generation From Biased Training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3722, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[7] Katherine M Tsui, Munjal Desai, Holly A Yanco, and Chris Uhlik. Exploring use cases for telepresence robots. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 11–18. IEEE, 2011.

[8] Lei Wang, Peizhen Lin, Jun Cheng, Feng Liu, Xiaoliang Ma, and Jianqin Yin. Visual relationship detection with recurrent attention and negative sampling. *Neurocomputing*, 434:55–66, 2021.

[9] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene Graph Generation by Iterative Message Passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[10] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene Graph Parsing with Global Context. In *Conference on Computer Vision and Pattern Recognition*, 2018.