S-AvE: Semantic Active Vision Exploration and Mapping of Indoor Environments for Mobile Robots

José V. Jaramillo[1], Roberto Capobianco[2][0000−0002−2219−215X], Francesco Riccio[2][0000−0002−9112−8143], and Daniele Nardi[2][0000−0001−6606−200X]

Sapienza University of Rome. Deptament of Computer, Control and Management Engineering "Antonio Ruberti"
[1]josevicentejaramillo@gmail.com
[2]{lastname}@diag.uniroma1.it

Abstract. Semantic mapping is fundamental to enable cognition and high-level planning in robotics. It is a difficult task due to generalization to different scenarios and sensory data types. Hence, most techniques do not obtain a rich and accurate semantic map of the environment and of the objects therein. To tackle this issue we present a novel approach that exploits active vision and drives environment exploration aiming at improving the quality of the semantic map.

Keywords: Semantic mapping, Map exploration, Mobile robots.

1 Introduction

Simultaneous exploration and map building are fundamental skills for mobile robots. However, building a comprehensive map of the environment – spanning from raw sensory observation to high-level semantic concepts [12] – is an extremely difficult task [13]. In literature, autonomous map building processes rely upon map exploration techniques that provide robots with an effective strategy to visit unknown portions of the environment. In this context, proposed approaches [14] maximize exploration, while minimizing the time spent in building the map. However, standard solutions to this problem are limited to a geometric reconstruction of the environment and, most importantly, they formalize the exploration strategy by only considering geometric and topological landmarks of the environment [14]. Conversely, semantic mapping [6, 9] enhances geometric, metric and topological knowledge about the environment by means of semantic concepts, thus enabling improved robot cognition. In this context, classic exploration techniques are still generally used, resulting in inaccurate and incomplete semantic maps. To tackle this issue, and to improve robot capabilities in exhaustively exploring the environment at the semantic level, we introduce S-AvE (Semantic-Based Active Vision Exploration), a new map exploration technique. We refer to map exploration as the exploration of an environment with the goal of building an internal representation of it including semantic knowledge. Standard map exploration techniques, such as frontier-based exploration [4, 7], do not focus on object reconstruction and the strategy that the

robot executes aims at maximizing the amount of visited portions of the environment disregarding objects therein. Such approaches, moreover, do not exploit semantic labels to influence the map exploration strategy and often result in incomplete and poorly detailed maps, which limit robots' autonomy and abilities. S-AVE, instead, explicitly uses detected objects to drive exploration by means of active vision [2, 3, 1], and it additionally combines semantic information to state-of-the-art map exploration techniques [8] to improve the 3D representation of the environment. We validate S-AVE in five different indoor environments in the Gazebo simulator against a frontier-based algorithm and human-driven robots.

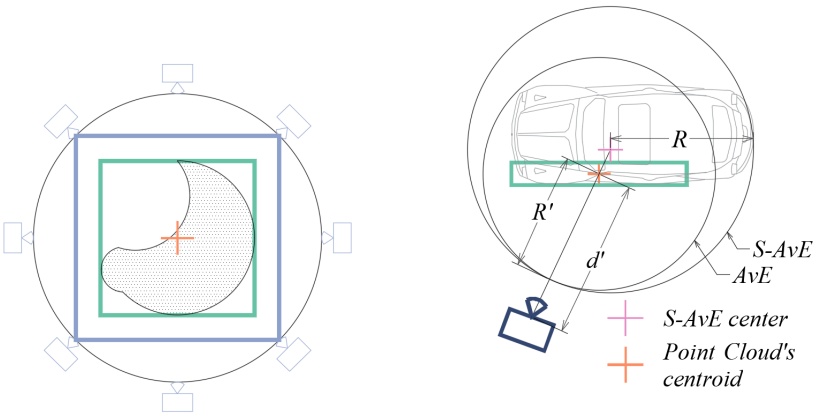
Summarizing, S-AVE is a novel technique that introduces a new paradigm in map exploration. S-AVE explicitly reasons about objects in the environment and their semantic label to improve its 3D reconstruction. The contributions to the-state-of-art are twofold: (1) it represents a novel technique for object-centered map exploration; and (2) it generates better semantic maps to support and improve robot reasoning and task execution. Moreover, it is important to highlight that this work contributes to the community by providing a ready-to-use simulation environment for map exploration and a first baseline for future benchmarking. We provide five simulated environments as free-to-download packages which can be easily used to integrate new exploration strategies to perform benchmarking[1] of semantic mapping approaches. We additionally release the source code for both S-AVE[2], as well as the data collected during our user study[3].

2 Semantic Active-Vision Exploration

Most of the approaches to active vision focus on evaluating a set of *candidate* poses and ranking them by expected information gain [11]. To reduce the search space of the theoretically infinite number of poses, [10] introduced the use of tessellation for discretizing the search space. In this case, a candidate pose is generated in each vertex of the tessellated sphere, directed towards the center. This method has been widely adopted in literature [3, 5]. Since these studies have been designed to model similarly sized objects located in the center of a platform, the tessellated sphere radius and center can be constants, as the object is placed on a limited size platform. In order to use the same algorithms in an open environment, we generate the candidate poses with a tessellated sphere (or circle, depending on agent's reachability) for each encountered object. Hence, the first phase of S-AVE consists in extracting, for each object in its current view, the objects sizes and centroids. We rely on a RGBD camera and an ideal object detector algorithm to acquire an object-segmented point cloud. Then, the portion of the point cloud corresponding to the closest object is used to compute the size of the bounding box and the centroid of the object.

To compute the candidate pose, S-AVE generates a sphere, in the robot cartesian, space that delimits the objects of interest. The radius R' of such a sphere as $R' = \sqrt{X'^2 + Y'^2}$ where $X' = X_{max} + offset + clearance$ and $Y' = Y_{max} + offset + clearance$. Fig. 1(a) provides a visual feedback on Equation **??**. X_{max} and Y_{max} are the coordinates

[1] https://github.com/JoseJaramillo/lucrezio_simulation_environments
[2] https://github.com/JoseJaramillo/lucrezio_semantic_explorer
[3] https://github.com/JoseJaramillo/S-AvE_ExperimentalEval



(a) Top view of an object with its Next-Best-View candidates poses

(b) Resulting object radius

Fig. 1. S-AvE adjusts the estimated radius of objects in accordance to a object ontology.

of the max 2D points of the bounding box (green). $offset$ is used to allow generating unknown volumes outside of the bounding box limits, and *clearance* is a parameter used so that the generated next-best-view candidates are separated from the object (blue). However, due to a mismatch between the actual centroid and the size of the object suuch an estimation can easily reveal inaccurate. To overcame this limit, we exploit semantic information in the process of generating candidate poses, and specifically for adjusting the radius and the center of the sphere.

The main idea consists first extracting the object missing parameters (i.e. bounding box size and centroid of the obejct) from the current view, and then searching the knowledge base of the object categories for the object's *typical size*. Then S-AvE adjusts the object parameters in accordance with a basic decision tree: if the observed object dimensions are smaller than expected, parameters are tuned accordingly; otherwise detected size and estimated centroid are preserved. The typical size of objects can be determined using several sources. In our case, we choose to simply compute the mean size among those provided on multiple online furniture stores.

It is important to notice that, in case the height of the object is much less than typical height, the robot can compute the next-view in a position where the field of view can fully cover the expected height. To this end, we compute the minimum distance to cover the object with the camera's field of view as $Dmin = max(Dmin_{top},\ Dmin_{bottom})$ where $Dmin_{top} = (H - H)/tan(\beta + \theta$ and $Dmin_{bottom} = H'/tan(\beta + \theta)$. top and $bottom$ refer to the minimum distance to cover the top and bottom of the object with the camera's field of view. H is the typical height of the object, H' is the camera height. θ is the camera pitch, and $\beta = \alpha/2$. α is the camera vertical viewing angle. Hence, we can recover the radius R as $R = d + \frac{max(X,Y)}{2}$, where $d = min(Dmin,\ Dmax)$ and X, Y are the planar object sizes found on the object's typical size database. For big objects or

tilted cameras, the computed minimum distance can be extremely large or even infinite. Hence, a maximum distance *Dmax* is set according to the environment.

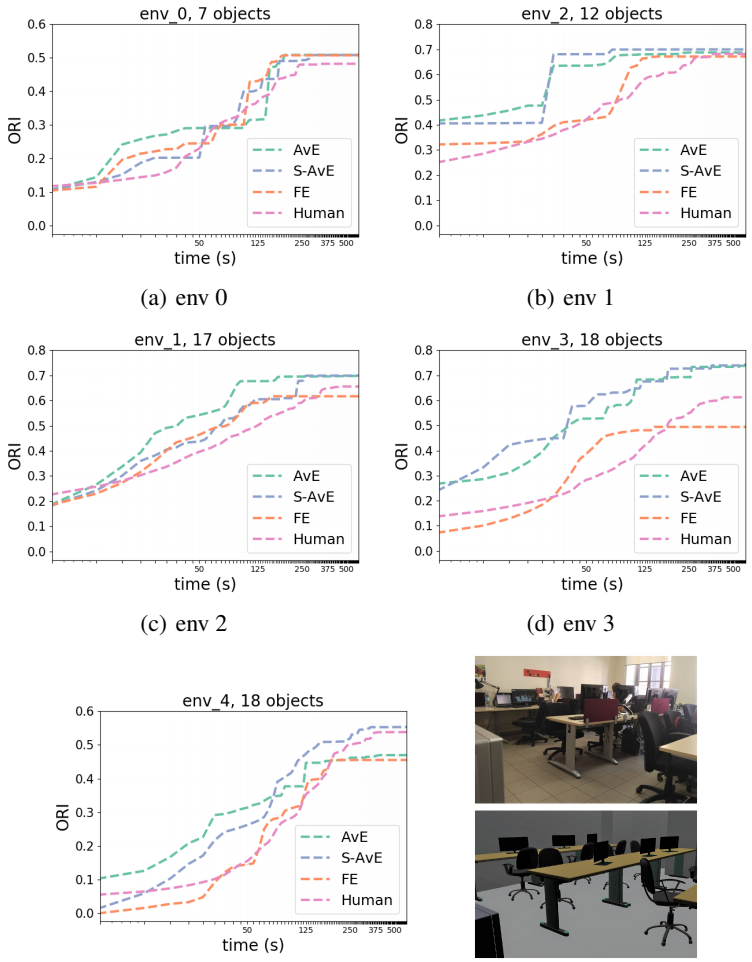
Then, to approximate the extracted centroid of a point cloud to the true centroid of the object. We compute the adjusted center in the direction formed by the camera center and the object point cloud's centroid, as shown in Fig.fig:distances. To compute the position of the center δ as $\delta = \sigma + l[cos(\alpha), sin(\alpha)]^T$, where α is the angle formed from the camera center σ to the point cloud's centroid ε, computed as $\alpha = atan2(\varepsilon.Y - \sigma.Y,\ \varepsilon.X - \sigma.X)$. Here, l is the length from camera center σ and to the object center δ, which is computed as $l = R - R' + d'$, where d' is the distance from the camera center to the point cloud centroid, and R' is the radius previously computed.

Finally, in order to get the goal pose, we evaluate the candidate poses by an active vision method. In this work, we use the *unobserved voxel volumetric information* introduced in [3], which returns the set of candidate poses ranked by information gain. Hence, by evaluating the poses in rank order, the next-best-view is the first reachable pose by the robot.

3 Experimental Evaluation

S-AvE is evaluated in five different scenarios, and each of them is a replica of real world environments – Fig.fig:sim-real-env compares a real environment (top) and its simulated world (bottom). Moreover, in order to validate the performance improvement of our approach, we include in the experimental evaluation two baseline methods: a frontier-based exploration (FE) technique [4]; and results of a user-study (Human) in which we let humans drive the robot to explore the environment. In particular, FE aims at showing the improvement of our solution with respect standard in map exploration techniques – in which no semantic knowledge is exploited; while Human is used to define an interval on the spectrum of possible strategies. In fact, as observed during the user-study, humans rely on spatial semantic knowledge and perform differently to standard FE methods and disregard metric information. We include in the evaluation also a stripped down version of S-AvE which only considers the detected object size to leverage the exploration algorithm. We refer to such an algorithm as AvE.

To assess S-AvE performance we rely on the object reconstruction index (ORI) which is computed against the groundtruth of the environments which contains the true number of objects, and objects shapes and dimensions as voxels. We compute the ORI index as in: $ORI = \frac{1}{N}\sum_{n=0}^{N} pv(n)/V_n$ where n denotes the n-*th* object, N is the total numbe r of objects, V_n is the total number of voxels representing n and $pv(\cdot)$ is a function that returns the number of perceived voxels of n. Each environment has a different structure, topology and a varying number of objects. Respectively for each environment the robot is challenged with 7, 12, 17, 18 and again 18 objects. All objects are common indoor objects such as, chairs, desks, cabinets, tables, fridges and other alike furniture. To evaluate our solution, in each of these environments we deploy a simulated differential-drive wheeled robot equipped with a depth camera sensor and laser range-sensor. The robot is configured to execute each of our competing algorithms (one at time) and to store metrics parameters. The plots in Fig. 2 report the results in reconstructing all the objects for each of the environment. As it can be notice, S-AvE has a



(a) env 0 (b) env 1

(c) env 2 (d) env 3

(e) env 4 (f) Sim. vs Real

Fig. 2. Object reconstruction index (ORI). The plots represent the results obtained in reconstructing the objects in the environments by each of the used approaches: user-study results (Human, red); frontier-based exploration (FE, orange); active-vision exploration (AvE, green); and semantic active-vision exploration (S-AvE, blue). Fig.fig:sim-real-env shows the comparison between a real world environment and its virtual replica.

better performance in all of them. It is worth mentioning that, the last scenarion resembles a common apartment composed by three areas with a less predictable structure and characterized by different object categories. In is worth noticing that AvE is not able to reach the usual performance and compares with the FE approach. Conversely, humans perform better and compare with S-AvE both in reconstructing the objects. Intuitively, the two top-scoring approaches are the ones that explicitly model object semantics and estimate object dimensions to navigate the environment. This suggests that in richer

scenarios, only relying on object appearance is not enough to exhaustively explore the environment. Moreover, by looking at the profile of the ORI and ONI indexes, S-AVE is able show a better performance with respect to all baselines.

4 Conclusion

Our paper introduces a novel approach that integrates active-vision and autonomous map exploration and that enables a robot to generate more accurate and complete semantic maps. To alleviate the computational demand of AV methods, we aim at enhancing S-AVE with generative networks that can be used to infer object dimensions after their classification.

References

1. Bajcsy, R., Aloimonos, Y., Tsotsos, J.K.: Revisiting active perception. Autonomous Robots **42**(2), 177–196 (02 2018). https://doi.org/10.1007/s10514-017-9615-3
2. Blake, A., Yuille, A. (eds.): Active Vision. MIT Press, Cambridge, MA, USA (1993)
3. Delmerico, J., Isler, S., Sabzevari, R., Scaramuzza, D.: A comparison of volumetric information gain metrics for active 3d object reconstruction. Autonomous Robots (04 2017). https://doi.org/10.1007/s10514-017-9634-0
4. Faria, M., Maza, I., Viguria, A.: Applying frontier cells based exploration and lazy theta* path planning over single grid-based world representation for autonomous inspection of large 3d structures with an uas. Journal of Intelligent & Robotic Systems **93**(1-2), 113–133 (2019)
5. de Figueiredo, R.P., Bernardino, A., Santos-Victor, J., Araújo, H.: On the advantages of foveal mechanisms for active stereo systems in visual search tasks. Autonomous Robots **42**(2), 459–476 (2018)
6. Gemignani, G., Capobianco, R., Bastianelli, E., Bloisi, D.D., Iocchi, L., Nardi, D.: Living with robots. Robot. Auton. Syst. **78**(C), 1–16 (Apr 2016). https://doi.org/10.1016/j.robot.2015.11.001
7. Hidaka, K., Kameyama, N.: Hybrid sensor-based and frontier-based exploration algorithm for autonomous transport vehicle map generation. In: 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE). pp. 994–999. IEEE (2018)
8. Nardi, F.: High-Level Environment Representations for Mobile Robots. Ph.D. thesis, Sapienza University of Rome (2019)
9. Nüchter, A., Hertzberg, J.: Towards semantic maps for mobile robots. Robotics and Autonomous Systems **56**(11), 915–926 (2008)
10. Panerai, F.M., Capurro, C., Sandini, G.: Space-variant vision for an active camera mount. In: Visual Information Processing IV. vol. 2488, pp. 284–297. International Society for Optics and Photonics (1995)
11. Pito, R.: A solution to the next best view problem for automated surface acquisition. IEEE Trans. Pattern Anal. Mach. Intell. **21**(10), 1016–1030 (Oct 1999). https://doi.org/10.1109/34.799908
12. Pronobis, A., Riccio, F., Rao, R.P.N.: Deep Spatial Affordance Hierarchy: Spatial knowledge representation for planning in large-scale environments. In: RSS 2017 Workshop on Spatial-Semantic Representations in Robotics. Boston, MA, USA (Jul 2017)
13. Taketomi, T., Uchiyama, H., Ikeda, S.: Visual slam algorithms: a survey from 2010 to 2016. IPSJ Transactions on Computer Vision and Applications **9**(1), 16 (2017)
14. Yamauchi, B.: A frontier-based approach for autonomous exploration. In: cira. vol. 97, p. 146 (1997)