CEA Comité Español de Automática
XLV JORNADAS AUTOMÁTICA MÁLAGA 2024
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Enfoque general y sistemático para percepción activa semántica en robótica

Morilla-Cabello, D.[1,*], Montijano, E.[1]

[a] *Universidad de Zaragoza, Instituto de Investigación en Ingeniería de Aragón (I3A), C. de Mariano Esquillor Gómez, s/n, 50018 Zaragoza, España.*

To cite this article: Morilla-Cabello, D., Montijano, E. 2024. General and Systematic Approach to Semantic Active Perception in Robotics. Jornadas de Automática, 45. https://doi.org/10.17979/ja-cea.2024.45.10938

Resumen

En este artículo, abordamos el problema de la percepción activa de información semántica, centrado en determinar las acciones que un robot móvil debe realizar para obtener información semántica de calidad del entorno. Con el auge de los algoritmos de percepción semántica, surgen nuevas oportunidades para los sistemas de planificación robóticos. Sin embargo, para aprovechar estas oportunidades, es crucial identificar los elementos esenciales que cualquier sistema de control orientado a la percepción debe tener. Para ello, proponemos una arquitectura general aplicable a cualquier sistema de percepción activa y analizamos las diferencias fundamentales que surgen al considerar información semántica en su diseño. Además, describimos una implementación preliminar de la arquitectura propuesta. Nuestro objetivo principal es proporcionar a los investigadores una formulación general y un sistema unificado y modular que facilite el avance en el campo de la percepción activa semántica.

Palabras clave: Tecnologías robóticas, Robótica móvil, Percepción y sensorizado.

General and Systematic Approach to Semantic Active Perception in Robotics

Abstract

In this article, we address the problem of active perception of semantic information, which focuses on determining the actions a mobile robot must take to obtain high-quality semantic information from the environment. The rise of semantic perception algorithms presents new opportunities for robotic planning systems. However, to capitalize on these opportunities, it is crucial to identify the essential elements that any perception-oriented control system must have. To this end, we propose a general architecture applicable to any active perception system and analyze the fundamental differences that arise when considering semantic information in its design. Additionally, we describe a preliminary implementation of the proposed architecture. Our main objective is to provide researchers with a general formulation and a unified, modular system that facilitates advancements in the field of semantic active perception.

Keywords: Robotics technology, Mobile robots, Perception and sensing.

1. Introducción

The recent advancements in robotics and the widespread availability of public code repositories have led to the establishment of *de facto* systems in various fields. These out-of-the-box solutions facilitate the research process by enabling rapid prototyping of new ideas and supporting greater complexity in proposed solutions. Additionally, by standardizing the conceptual components of the system, researchers can more easily identify module-specific problems and bottlenecks in current solutions.

Inspired by this methodology, in this paper, we analyze the modules required to implement a general semantic active perception framework. Active perception seeks to close the loop between perception and action. By reasoning on observations during the mission execution, a robot can adjust its planning to the new information, ultimately enhancing the mission per-

*Autor para correspondencia: davidmc@unizar.es

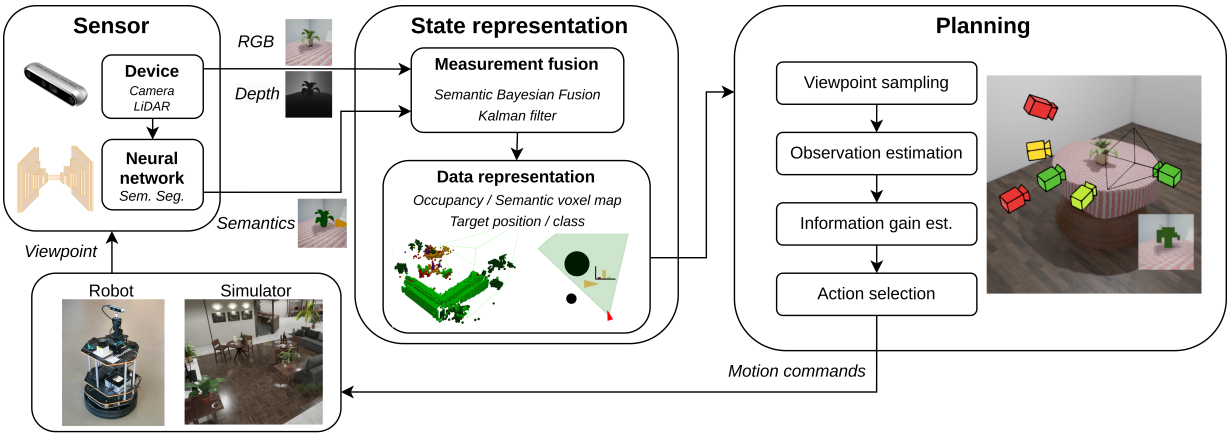

Figure 1: System diagram for semantic active perception tasks. A robot with a sensor captures measurements and fuses them into a state representation or belief to integrate information about the environment. This belief is subsequently used for planning the next actions to maximize mission performance.

formance and robustness. The particularities of dealing with semantic information represent a new research domain where the majority of existing approaches are insufficient. Among others, fundamental challenges are:

- **Integration of semantic knowledge:** The knowledge representation derived from neural networks differs from traditional geometric representations, requiring new methods for describing, fusing, and interpreting this information inside the planning.
- **Predicting the information value:** Active perception methods rely on forward estimates of potential measurements to decide where to move. Most neural networks are agnostic to the relative positioning between the sensor and the scene, which makes it very difficult to estimate the output of future measurements.
- **Testing and evaluation:** Since learning-based semantic perception methods are trained on real data, the evaluation and benchmarking of any semantic active perception algorithm must ensure the availability of photo-realistic measurements not required in previous robotic problems.

These challenges require a reconsideration and restructuring of existing systems. In this article, we present the conceptual framework necessary for semantic active perception, derived from its general problem formulation. Subsequently, we investigate the modules crucial for a planning pipeline in this context. Finally, we provide initial software implementations to address the gaps present in prior literature[1].

2. Related work

Semantic information is acquired by interpreting high-level information from raw sensor measurements such as images captured by cameras or LiDAR point clouds. To this end, several neural networks stand out for their performance and ease-of-use. Semantic information can be extracted from images using semantic segmentation, which infers per-pixel class labels as in Chen et al. (2018), object detection, detecting the position and class of individual objects as in Redmon et al. (2016), or panoptic segmentation, separating individual instances of each class. Similar approaches have been proposed for LiDAR data such as in Milioto et al. (2019).

Regarding map representations, OctoMap by Hornung et al. (2013) and Voxblox by Oleynikova et al. (2017) are two widely used volumetric systems. Extensions for semantic mapping include Asgharivaskasi and Atanasov (2023) for reasoning over semantic voxels and Grinvald et al. (2019) for modelling individual objects. More complex systems, such as Kimera proposed by Rosinol et al. (2021), integrate SLAM systems with semantic surfels representations and dynamic scene graphs.

While several libraries exist for robot navigation, such as the ROS navigation stack initially introduced in Marder-Eppstein et al. (2010), which enables a robot to plan a route between two points, planning for active perception remains less established. Recent advancements in geometric scene reconstruction have provided solutions in exploration Zhou et al. (2021) and surface reconstruction Schmid et al. (2020), incorporating active perception concepts like viewpoint sampling, observation estimation, and information gain estimation. However, these works do not incorporate semantic information. Solutions for semantic active perception have been proposed by Asgharivaskasi and Atanasov (2023) and Morilla-Cabello et al. (2023b). Nevertheless, a general and holistic solution that establishes the basis for semantic active perception research is still missing.

3. Active Perception Problem Description

Consider a robot moving in an environment and equipped with a sensor to gather data. We denote $\mathbf{r}_t$ as the robot position at time t. The robot's objective is to estimate the value of a feature of interest, denoted by $\mathbf{x}$. This feature varies across

[1] https://github.com/dvdmc/active_perception

active perception tasks such as the occupancy or class of cells on a map, the location of the robot, or the position or class of some targets.

The robot can move in the environment according to some dynamic model

$$\mathbf{r}_{t+1} = f(\mathbf{x}, \mathbf{u}_t). \tag{1}$$

It also has a perception algorithm available to obtain measurements of $\mathbf{x}_t$, denoted as $\mathbf{z}_t$, at certain positions according to

$$\mathbf{z}_t = g(\mathbf{x}, \mathbf{r}_t). \tag{2}$$

To infer $\mathbf{x}_t$ from $\mathbf{z}_t$, measurements are fused to form a probabilistic state estimate or state belief, $\hat{\mathbf{x}}_t \sim p(\mathbf{x}_t|\mathbf{z}_{0:t})$, which is updated recursively as

$$\hat{\mathbf{x}}_t = h(\hat{\mathbf{x}}_{t-1}, \mathbf{r}_t, \mathbf{z}_t), \tag{3}$$

where h is a fusion algorithm, e.g., Bayes or Kalman filter.

The goal of active perception is to move the robot to maximize the information gathered from the measurements to obtain the best possible estimate of the true value of $\mathbf{x}$ as efficiently as possible. Formally, this requires planning the set of actions $\mathcal{U} = (\mathbf{u}_0, \ldots, \mathbf{u}_N), N > 0$ that optimizes a desired information quality criterion, I, over the state estimation process

$$\max_{\mathbf{u}_0,\ldots,\mathbf{u}_N} \sum_{k=0}^{N} I_k(\hat{\mathbf{x}}_k), \text{ s.t. (1), (2), (3)} \tag{4}$$

Commonly used information criteria are the differential entropy of the estimation $I(\hat{\mathbf{x}}_t) = \Delta H^k_{k-1} = H(\hat{\mathbf{x}}_t) - H(\hat{\mathbf{x}}_{t-1})$ and choices for $H(\cdot)$ depend on the specific distributions. The dependence of I on the estimated state does not allow solving (4) online, needing to adapt the planning as new knowledge about $\mathbf{x}$ is acquired. The mission completion might be determined by some threshold on H when a goal is considered to be achieved, or when a certain budget is exhausted.

Based on the previous problem description, solving (4) requires perceiving the environment, keeping a state belief, planning according to the state belief, and moving the robot to acquire new beneficial observations. In the following sections we analyze existing solutions for the modules of the proposed system, explain the differences that arise when applying them to semantic active perception, and offer alternatives to the limitations.

4. Perception

In traditional active perception, measurements in (2), come from a laser sensor, RGB-D camera, or SLAM system that can be modelled geometrically. In the case of semantic observations, $g(\cdot)$ not only includes a sensor to obtain raw data such as images or 3D point clouds but also a neural network module, NN_θ, to extract high-level semantic information (see Figure 2). Denote the raw data captured at position $\mathbf{r}_t$ as the vector $\mathbf{Y}$. The semantic neural network will output a set of confidences per class k

$$\mathbf{p} = NN_\theta(\mathbf{Y}). \tag{5}$$

The semantic measurement is modelled after a categorical-distributed likelihood, which parameters $\mathbf{p} = (p_1, ..., p_k)$ are the confidences returned by the neural network

$$\mathbf{z}_t \sim p(l = l_i|\mathbf{p}). \tag{6}$$

(a) RGB image (b) Depth image (c) Semantic image

Figure 2: Example of typical visual semantic output. 2a shows the image acquired by the camera. 2b shows the per-pixel depth projected on an image as a grayscale map (lighter means further), 2c shows the per-pixel label output by a semantic segmentation network over the RGB image (where there are no labels, the output is the trivial *background* class).

Previous approaches using semantic measurements follow a similar formulation. However, they differ in measurement representation. Approaches such as Grinvald et al. (2019) only account for the label with the maximum confidence, neglecting the probabilistic information. Others, such as Rosinol et al. (2021) consider the confidences from the neural network. In Morilla-Cabello et al. (2023a), we also incorporate the epistemic uncertainty obtained from the neural network to model a more advanced sensor noise.

5. State Representation

The state representation defined in Section 3 depends on the problem at hand. Figure 3 shows different examples explained below. In most robotics applications, the state contains geometric variables such as the robot's pose and an occupancy map. These solutions are widely covered by existing literature for classic active perception and exploration, like Schmid et al. (2020) or other works described in Section 2. However, for semantic active perception, the representation must incorporate semantic information per voxel or object. Formally, per voxel or object $m \in \mathcal{M}$ in the scene, we assume a label $l_m \in \mathcal{K} = \{1, \ldots, K\}$, where $\mathcal{K}$ is the set of predefined possible classes. We define a categorical distribution over the set of possible semantic classes

$$p(l_m = l_i|\mathbf{p}) = \prod_{j=1}^{k} p_j^{[j=i]}, \tag{7}$$

where $\mathbf{p} = (p_1, \ldots, p_k)$ such that p_j represents the probability of the voxel belonging to class j, the exponent, $[j = i]$, is the indicator function, and $\sum_{j=1}^{k} p_j = 1$.

Furthermore, while the fusion of occupancy information has been well-studied in the literature as in Elfes (1989), the fusion of semantic measurements is still an open problem. In McCormac et al. (2017), a Bayesian approach is proposed to fuse semantic measurements:

$$p(l_i|\mathbf{X}_{1:t}) \propto p(l_i|\mathbf{X}_{1:t-1}) \prod_j p(z_j = l_i|\mathbf{X}_t). \tag{8}$$

(a) Virtual environment (b) Voxblox (c) Semantic voxel map (d) Target based map

Figure 3: Different state representations for the problem of active perception. For the environment shown in 3a, the Voxblox map in 3b is used to store occupancy information for geometric active perception, the semantic voxel map in 3c is used for active semantic mapping. In other cases, a simpler target-based representation with information about the target position and class, and other environment information like occluder positions might be enough as shown in 3d.

In Morilla-Cabello et al. (2023a), we study the influence of correctly characterizing uncertainty from neural networks in the semantic fusion process for semantic mapping, improving over (8) and other classic fusion methods. Generally, mapping systems embed the update or fusion method within the data structure implementation. Based on the insights from previous work, we argue that semantic active perception systems must separate the data representation from the information fusion to facilitate research on alternative fusion methods.

6. Planning

In order to solve (4) for each task, the solution might use different state representations and planning strategies. However, there are common modules required in every case. We separate the active perception planning phases into viewpoint sampling, observation estimation, information gain estimation, and action selection (Figure 4). Formally, these steps decompose (4) into sub-problems to tackle them independently.

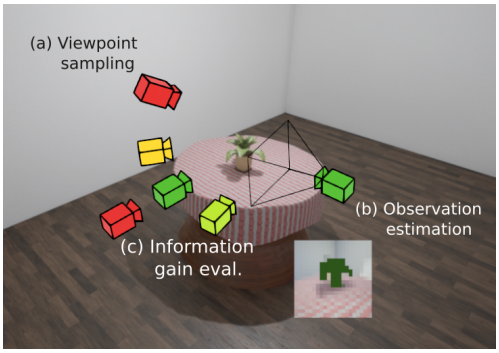

Figure 4: Steps for semantic active perception planning in a classification task. (a) Viewpoints are sampled using random or informed sampling. (b) Approximated observations are estimated for potential viewpoints. In this case, the estimated confidence for the classification. (c) The predictions are used to compute the estimated information gain represented as red-green for low-high estimated values respectively.

6.1. Viewpoint Sampling

This step involves selecting viewpoints for evaluation. Given that the viewpoints space, represented by the domain of $\mathbf{r}$, can be arbitrarily large, only a subset of potential viewpoints can feasibly be evaluated. Common practices include random selection or the use of heuristics such as targeting regions of interest to perform informed sampling as in Kompis et al. (2021) and Dai et al. (2020). This problem is common for metric and semantic active perception. Similarly to previous work, in Papatheodorou et al. (2023), viewpoints are sampled around semantic objects to reconstruct them. Viewpoints can also undergo further optimization in subsequent steps.

6.2. Observation Estimation

For any potential viewpoint, the main challenge in semantic active perception is to approximate Equation (2), from the information about the environment in $\hat{\mathbf{x}}_t$,

$$\hat{\mathbf{z}}_t = g'(\hat{\mathbf{x}}_t, \mathbf{r}_t), \tag{9}$$

where $g'(\cdot)$ is a surrogate for the observation model.

Similar to metric active perception, the first step is to assess the information that the sensor will capture. Performing raycasting on the viewpoints' camera frustum and checking the occupancy of the observed space is a common (but expensive) solution. In the case of semantic active perception, the observed volume can similarly be estimated. However, there is a key difference from metric approaches: observing a surface might be sufficient to determine the occupancy, but semantic active perception requires estimating the semantic output provided by the neural network from different viewpoints.

Since $\mathbf{Y}$ is also unavailable, a surrogate model must reason about the current state belief. This model must incorporate priors about the neural network behaviour to make informed decisions. A common approximation used in Popović et al. (2020) involves employing a distance-based confidence model. Other approaches, such as those by Liu et al. (2023) and Feldman and Indelman (2020), build a prior model of neural network outputs based on the relative object-camera pose.

6.3. Information Gain Estimation

This step requires simulating the fusion of the estimated measurement with the state belief. In traditional volumetric reconstruction, this process is approximated by volumetric gain measurements proposed by Isler et al. (2016). In semantic active perception, a different probabilistic approach must be considered. The prior and posterior state beliefs are utilized to compute the gain in information. Given that the state belief is represented as a probability distribution, the information gain is naturally defined as the differential entropy after incorporating the observation. Due to the difficulty of modelling semantics from a probabilistic standpoint, previous work has also incorporated heuristics Marques et al. (2023). In our recent work Morilla-Cabello et al. (2023b), we explored

the importance of modelling information gain estimation by examining measurement correlations with the environment.

Other works approximate active perception planning using reinforcement learning (RL) approaches Álvaro et al. (2023) and Rückin et al. (2022). In this methodology, the observation estimation and information gain estimation are approximated with a learned policy for the actions.

7. Execution

Simulators and Benchmarks. A key difference between traditional robotic systems and those used for semantic active perception is the simulator. In robot navigation, scene appearance is not critical. However, computer vision methods, especially neural networks, are trained on real images and applied to real-world scenarios. On the other hand, when evaluating computer vision systems, there is no control over sensor positioning or any concept of re-observation. Semantic active perception requires environments that closely resemble real-world scenarios and allow for actions to improve measurement quality. This motivates the use of photorealistic simulators. Recent work has introduced several, like AirSim in Shah et al. (2018) and Flightmare in Song et al. (2021). These simulators, built on game engines like Unity and Unreal, offer controllability and photorealism but lack generality, being less flexible than traditional simulators like Gazebo. Other benchmarks, such as Habitat-Sim in Puig et al. (2024), are designed for specific tasks like object-goal navigation and manipulation, limiting their modularity and generality.

Robot Deployment. When deploying semantic active perception systems on real platforms, there are significant differences compared to traditional solutions. Raw measurements need to be processed by neural networks, which typically require GPUs for computational efficiency. Recently, reduced GPU-enabled devices have been proposed, allowing the use of smaller neural networks on mobile robots. These constrained models usually present reduced performance compared to models run on workstations. We believe that semantic active perception can address this issue. By leveraging re-observation, it is possible to make better use of less powerful deep learning models instead of attempting to run oversized models on constrained platforms.

8. Software

To bridge the current gaps in existing systems, we are in the process of developing a software framework depicted in Figure 5. This system is rooted in cited work introduced throughout Sections 4,5,6, and 7.The cited solutions have the problem of being too specific to their context to allow fast prototyping of new solutions. The goal of this implementation is not to achieve better results than previous work, but to simplify and bring different modules together to improve the generalization and ease of use of tools for semantic active perception. The software can be accessed at: `https://github.com/dvdmc/active_perception`

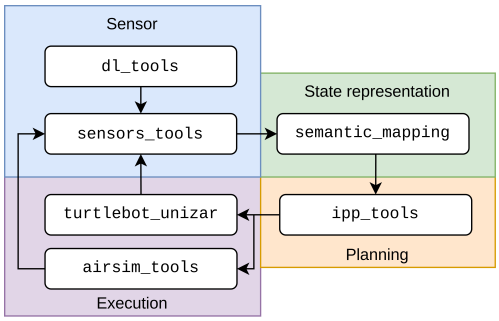

Figure 5: General architecture of the semantic active perception software framework under development.

Sensor. Our proposed system currently includes two modules regarding the sensor. `dl_tools` offers a setup for training (mainly fine-tuning), and evaluating neural networks. In `sensor_tools`, we implement a deployment framework. We include *bridges* to interface with different sources such as sensors or simulators. These tools are generally not included in other active semantic systems as the network is assumed to be given or approximated from the ground truth in simulators (Asgharivaskasi and Atanasov (2021),Rosinol et al. (2021)) Neural networks are commonly implemented using the Pytorch [2] library in Python, while traditional sensor drivers and other parts of the robotic stack are implemented in C/C++. ROS offers a good solution for bridging these programming languages. However, at the moment, there is no standard implementation of ROS messages for semantic measurements. We provide a semantic point cloud message generation that includes semantic classification probabilities.

State representation. Mainly focusing on voxel semantic mapping, we propose a simple and easy-to-modify `semantic_mapping` system. The data structure is a voxel hash-grid as in Nießner et al. (2013) that includes the confidence of the voxel belonging to each class and associated uncertainty. As discussed in Section 5, we separate the data structure from the fusion method, allowing benchmarking of different semantic fusion solutions. Previous work generally extends existing metric representations to semantics for specific cases as in Grinvald et al. (2019), resulting in non-general solutions.

Planning. In `ipp_tools` (ipp for informative path planning), we include modules necessary for active perception planning, focusing on viewpoint sampling, observation estimation, and information gain estimation. In previous work, these modules might be added as individual contributions within the planning process. We prioritize modularity in planning, recognizing it as a critical research direction in semantic active perception. Our framework incorporates observation estimation methods for estimating observed voxels, including ray-casting, rastering, and culling techniques. Additionally, we introduce information gain estimation methods, such as the one presented in Popović et al. (2020).

[2] `https://pytorch.org/`

Deployment. In `sensor_tools` and `airsim_tools`, we provide bridges to interface with various simulators like Airsim or Habitat-Sim. Other bridges might be added incrementally to minimize duplication of solutions and enhance maintainability. Additionally, in `turtlebot_unizar`, we introduce a minimal framework for deploying methods on a Turtlebot robot. We anticipate similar repositories for new robots in the future. Our aim is to offer a unified and well-documented protocol for real robot experiments.

Use cases. The modules related to the *sensor* and *state representation* have been used in Morilla-Cabello et al. (2023a) for semantic mapping. Additionally, the modules related to *planning* and *execution* have been used in real experiments for active object detection in Morilla-Cabello et al. (2023b). The overall system will be integrated into an active semantic mapping pipeline.

9. Conclusions

Despite the existence of widely used systems in robotics, the modules necessary for semantic active perception are not readily available in a generalized form. Consequently, there is a multiplicity of formulations and implementations for this problem. In this article, we present a general active perception framework based on a comprehensive problem description for semantic active perception. We identify the essential components of the active perception problem and provide a general software architecture to integrate them. We aim for this contribution to serve as a unified reference for both future theoretical formulations and software implementations in the field of active perception for high-level scene understanding.

Acknowledgements

This work has been supported by DGA project T45_23R, MCIN/AEI/ERDF/European Union NextGenerationEU/PRTR project PID2021-125514NB-I00 and grant FPU20-06563.

References

Álvaro, S.-G., Montijano, E., Böhmer, W., Alonso-Mora, J., 2023. Active classification of moving targets with learned control policies. IEEE Robotics and Automation Letters 8 (6), 3717–3724.

Asgharivaskasi, A., Atanasov, N., 2021. Active bayesian multi-class mapping from range and semantic segmentation observations. In: 2021 IEEE International Conference on Robotics and Automation. pp. 1–7.

Asgharivaskasi, A., Atanasov, N., 2023. Semantic octree mapping and shannon mutual information computation for robot exploration. IEEE Transactions on Robotics 39 (3), 1910–1928.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision.

Dai, A., Papatheodorou, S., Funk, N., Tzoumanikas, D., Leutenegger, S., 2020. Fast frontier-based information-driven autonomous exploration with an mav. In: IEEE International Conference on Robotics and Automation. pp. 9570–9576.

Elfes, A., 1989. Using occupancy grids for mobile robot perception and navigation. Computer 22 (6), 46–57.

Feldman, Y., Indelman, V., 2020. Spatially-dependent bayesian semantic perception under model and localization uncertainty. Autonomous Robots 44 (6), 1091–1119.

Grinvald, M., Furrer, F., Novkovic, T., Chung, J. J., Cadena, C., Siegwart, R., Nieto, J., 2019. Volumetric instance-aware semantic mapping and 3d object discovery. IEEE Robotics and Automation Letters 4 (3), 3037–3044.

Hornung, A., Wurm, K. M., Bennewitz, M., Stachniss, C., Burgard, W., Apr 2013. Octomap: an efficient probabilistic 3d mapping framework based on octrees. Autonomous Robots 34 (3), 189–206.

Isler, S., Sabzevari, R., Delmerico, J., Scaramuzza, D., 2016. An information gain formulation for active volumetric 3d reconstruction. In: IEEE International Conference on Robotics and Automation. pp. 3477–3484.

Kompis, Y., Bartolomei, L., Mascaro, R., Teixeira, L., Chli, M., 2021. Informed sampling exploration path planner for 3d reconstruction of large scenes. IEEE Robotics and Automation Letters 6 (4), 7893–7900.

Liu, X., Prabhu, A., Cladera, F., Miller, I. D., Zhou, L., Taylor, C. J., Kumar, V., 2023. Active metric-semantic mapping by multiple aerial robots. In: IEEE International Conference on Robotics and Automation. pp. 3282–3288.

Marder-Eppstein, E., Berger, E., Foote, T., Gerkey, B., Konolige, K., 2010. The office marathon: Robust navigation in an indoor office environment. In: IEEE International Conference on Robotics and Automation. pp. 300–307.

Marques, J. M. C., Zhai, A., Wang, S., Hauser, K., 2023. On the overconfidence problem in semantic 3d mapping. arXiv preprint arXiv:2311.10018.

McCormac, J., Handa, A., Davison, A., Leutenegger, S., 2017. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In: IEEE International Conference on Robotics and Automation. pp. 4628–4635.

Milioto, A., Vizzo, I., Behley, J., Stachniss, C., 2019. Rangenet ++: Fast and accurate lidar semantic segmentation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4213–4220.

Morilla-Cabello, D., Mur-Labadia, L., Martinez-Cantin, R., Montijano, E., 2023a. Robust fusion for bayesian semantic mapping. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 76–81.

Morilla-Cabello, D., Westheider, J., Popovic, M., Montijano, E., 2023b. Perceptual factors for environmental modeling in robotic active perception. arXiv preprint arXiv:2309.10620.

Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M., nov 2013. Real-time 3d reconstruction at scale using voxel hashing. ACM Trans. Graph. 32 (6).

Oleynikova, H., Taylor, Z., Fehr, M., Siegwart, R., Nieto, J., 2017. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1366–1373.

Papatheodorou, S., Funk, N., Tzoumanikas, D., Choi, C., Xu, B., Leutenegger, S., 2023. Finding things in the unknown: Semantic object-centric exploration with an mav. In: IEEE International Conference on Robotics and Automation. pp. 3339–3345.

Popović, M., Vidal-Calleja, T., Hitz, G., Chung, J. J., Sa, I., Siegwart, R., Nieto, J., 2020. An informative path planning framework for uav-based terrain monitoring. Autonomous Robots 44 (6), 889–911.

Puig, X., Undersander, E., Szot, A., Cote, M. D., Partsey, R., Yang, J., Desai, R., Clegg, A. W., Hlavac, M., Min, T., Gervet, T., Vondrus, V., Berges, V.-P., Turner, J., Maksymets, O., Kira, Z., Kalakrishnan, M., Malik, J., Chaplot, D. S., Jain, U., Batra, D., Rai, A., Mottaghi, R., 2024. Habitat 3.0: A co-habitat for humans, avatars and robots.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Rosinol, A., Violette, A., Abate, M., Hughes, N., Chang, Y., Shi, J., Gupta, A., Carlone, L., 2021. Kimera: From slam to spatial perception with 3d dynamic scene graphs. The International Journal of Robotics Research 40 (12-14), 1510–1546.

Rückin, J., Jin, L., Popović, M., 2022. Adaptive informative path planning using deep reinforcement learning for uav-based active sensing. In: International Conference on Robotics and Automation. pp. 4473–4479.

Schmid, L., Pantic, M., Khanna, R., Ott, L., Siegwart, R., Nieto, J., 2020. An efficient sampling-based method for online informative path planning in unknown environments. IEEE Robotics and Automation Letters 5 (2), 1500–1507.

Shah, S., Dey, D., Lovett, C., Kapoor, A., 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In: Field and Service Robotics. pp. 621–635.

Song, Y., Naji, S., Kaufmann, E., Loquercio, A., Scaramuzza, D., 2021. Flightmare: A flexible quadrotor simulator. In: Conference on Robot Learning. pp. 1147–1157.

Zhou, B., Zhang, Y., Chen, X., Shen, S., 2021. Fuel: Fast uav exploration using incremental frontier structure and hierarchical planning. IEEE Robotics and Automation Letters 6 (2), 779–786.