

Human-Inspired Long-Term Indoor Localization in Human-Oriented Environment

Nicky Zimmerman[†]

Matteo Sodano[‡]

Abstract—Lifelong localization is crucial for enabling the autonomy of service robots. In this paper, we present an overview of our past research on long-term localization and mapping, exploiting geometric priors such as floor plans and integrating textual and semantic information. Our approach was validated on challenging sequences spanning over many months, and we released open source implementations.

I. INTRODUCTION

Localization has been long-studied in robotics [2], [17], [20]. Supported by high-resolution sensors, GNSS-denied robot localization research initially focused on precise, detailed geometric maps [1], [5], [10], [11]. However, humans do not require highly accurate geometric information to navigate [9], [19], but instead, extract semantic information from objects around them. Inspired by how humans navigate, we can exploit insights from human navigation to improve long-term localization, which enables robots to navigate in the same environment over extended periods, spanning several months or even years. In this work, we summarize our past contributions to robust long-term localization and mapping, exploiting long-lasting geometric, textual and semantic cues [21], [22], [23], [24].

Localization research has advanced beyond purely geometric research, and recently many works tackle the task of semantic localization and mapping [14], [16], [18]. However, they focus on getting a detailed reconstruction of the current environment, rather than long-term localization. Human-occupied environments tend to be dynamic, having both fast-moving agents and structural and quasi-static changes such as opening/closing doors and furniture changes. When localizing with a known map, similar to how humans revisit places, a critical component is which elements are contained in the map. Regardless of the localization strategy, a map containing unstable features may lead to localization failures. To support long-term localization, map elements must be selected rather than including all current observations. These elements can either be inferred with common sense or learned statistically from observing an environment from a period of time [21]. Textual cues are another element that can effectively support long-term localization. Textual information such as room numbers can serve as a unique identifier of places and is often included in floor plans. However, they are often neglected in mapping approaches. Works on text-guided localization are also scarce [3], [13]. Thus, we proposed a framework for integrating textual information in maps and exploiting it during localization [24].

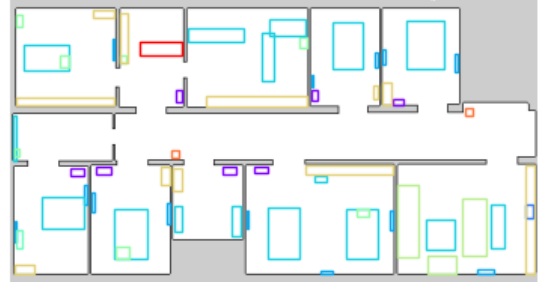


Fig. 1: A 2D abstract semantic map enriching a floor plan with semantic information, used for long term localization. Different box colors indicate different object classes.

A crucial aspect of long-term localization is the level of accuracy required. In fact, there is a trade-off between accuracy and robustness, and each task requires a different blend of the two. For example, for planning and navigating along the path of hundreds of meters, robustness (i.e., avoiding jumps in the trajectory) is more important, while high accuracy is only required in specific end-points (i.e. pickup/delivery points). When the requirement of exceptional accuracy is lifted, we can consider less precise maps, such as hand-drawn maps [6], visitor map and floor plans annotated by end users with objects [8], [21]. This relaxation of requirements can even remove the need for mapping, and allow end-users to independently edit floor-plan based maps.

In our previous works, we advocated for fusing geometric, semantic, and textual information for long-term localization in human-oriented environments. In this paper, we provide a summary of our contributions and insights.

II. APPROACH

Our goal is to globally localize in an indoor environment, exploiting semantic and textual cues, and prior geometric information, when available. We utilize semantically-enriched floor plans that capture long-lasting features, and localize by integrating semantic and textual information in a Monte Carlo localization (MCL) framework [4].

A. Abstract Semantic map

We propose the concept of abstract semantic map, a map representation that captures the most long-lasting structures of an environment. This map includes a floor plan (or an occupancy grid map), which is overlaid by layers of semantic and textual information. Objects with semantic significance are represented at high-level by their semantic class and 2D rectangles overlying the occupancy grid. Textual information is encoded through the text likelihood maps (Sec. II-

[†]IDSIA, Universita della Svizzera italiana, Switzerland

[‡]University of Bonn, Germany

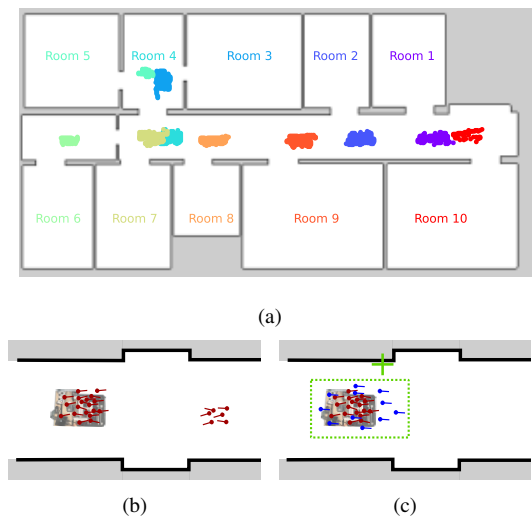


Fig. 2: (a) The text likelihood maps (b) Particle distribution prior to text spotting, with multi-modal distribution (c) Particle injection based on the text likelihood map after a textual cue was detected.

B). The floor plan is segmented into rooms, where each room contains a list of semantic objects, a room category representing a higher level of semantic understanding and a name corresponding to a text sign. While we offer a mapping approach that can construct a 3D metric-semantic map from RGB images [23], it is also possible to edit the abstract semantic map manually, using our GUI application, MAPHisto¹.

To determine which semantic classes hold significance for the purpose of long-term localization, we estimated how transient instances of these classes were. As part of our work [21], we collected data for several months in a dynamic environment of a university lab, including office rooms, corridors and a kitchen. Employing an object detection model on images with known poses, we were able to detect when objects have been moved, and assign a stability score to a semantic class based on how often instances of that classes moved around.

B. Text Likelihood Maps

Textual information is integrated into our semantic maps through likelihood functions, describing where the robot might be when detecting a specific textual cue. In our work [24], we adopt a text spotting pipeline based on the work of Liao et al. [7] and Shi et al. [15], who proposed a detection and recognition models for text, respectively. We decide to focus on room numbers as a textual localization cue, which can be either assigned manually to the abstract semantic map, or extracted from a floor plan. We record posed images and apply the text spotting, matching each detection to our list of rooms. For each room number, we construct a 2D histogram of locations where the text was successfully detected. These sampled locations provide a sparse representation of the likelihood of detecting a specific textual tag, which we refer to as text likelihood maps (Fig. 2).

¹<https://github.com/FullMetalNicky/Maphisto>

We create a continuous, dense representation to approximate the likelihood with a uniform distribution within an axis-aligned bounding box enclosing all sampled locations where the detection rate is above threshold τ for this textual cue. When collecting posed images is not possible, one can manually annotate these bounding boxes.

C. MCL Integration

Our approach extends Monte Carlo localization [4]. We provide different sensor models for integrating textual and semantic cues.

When a textual cue is detected, we inject particles into the corresponding area of the map (Fig. 2), based on the text likelihood maps. This technique also allows delocalized robots to get new hypotheses based on the detection, enriching their particle filters with the proposed poses, and is sometimes referred to in literature as reciprocal sampling [12].

For 2D object detection predictions, we propose a sensor model based on cosine distance between the detected heading and the semantically-annotated map [21]. We also provided a sensor model for 3D bounding box prediction [23], which considers the IoU between a 3D bounding box prediction and the metric-semantic, accounting for the typical detection noise for that specific semantic class. We use the detected semantic objects to infer the room category and utilize this hierarchical semantic information to initialize the particles in the filter in the corresponding rooms.

III. EXPERIMENTAL EVALUATION

In the course of our research, we performed various experiments on datasets recorded for several months, capturing different scenarios such as dynamic obstacles, opening and closing of doors and rearranging furniture.

The outcome of the semantic stability analysis [21], suggested that it is not necessary to carry this analysis for each type of environment, as the classes that were singled out through rigorous data collection and statistical analysis perfectly matched what we expected to see by using common sense. Additionally, throughout our research, we discovered that the classes chosen for one instance of a specific environment (i.e. "office") transferred successfully to other instances. We used the same classes for deployments in ETH Zurich and ABB facility in Sweden, as well as our later work on semantic localization on resource-constrained platforms [22], where we ported our human-inspired localization to an ultra-low-power compute platform with 1.5 MB of memory.

In our localization experiments, the semantically-guided localization outperformed the purely-geometric MCL on all sequences. Despite the imprecise, hand-annotated maps, we were able to robustly localize long-term, even when objects differed from their actual size by 62.5%, or up to 1.25 m. Similarly, integrating textual cues proved to improve localization stability when localizing in environments with high geometric symmetry and lack of semantic features. Further details on the experiments, qualitative and quantitative results, ablation studies, and insights can be found in the main papers [21], [22], [23], [24].

REFERENCES

- [1] E. Bylow, J. Sturm, C. Kerl, F. Kahl, and D. Cremers. Real-Time Camera Tracking and 3D Reconstruction Using Signed Distance Functions. In *Proc. of Robotics: Science and Systems (RSS)*, volume 2, 2013.
- [2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. Leonard. Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age. *IEEE Trans. on Robotics (TRO)*, 32:1309–1332, 2016.
- [3] L. Cui, C. Rong, J. Huang, A. Rosendo, and L. Kneip. Monte-Carlo Localization in Underground Parking Lots Using Parking Slot Numbers. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2021.
- [4] F. Dellaert, D. Fox, W. Burgard, and S. Thrun. Monte carlo localization for mobile robots. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 1999.
- [5] G. Grisetti, C. Stachniss, and W. Burgard. Improved Techniques for Grid Mapping with Rao-Blackwellized Particle Filters. *IEEE Trans. on Robotics (TRO)*, 23(1):34–46, 2007.
- [6] L. Li, B. Yang, M. Liang, W. Zeng, and M. Ren. End-to-end Contextual Perception and Prediction with Interaction Transformer. *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.
- [7] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481, 2020.
- [8] J. Loo and D. Hsu. Scene Action Maps: Behavioural Maps for Navigation without Metric Information. *arXiv preprint arXiv:2405.07948*, 2024.
- [9] O. Mendez, S. Hadfield, N. Pugeault, and R. Bowden. Sedar-semantic detection and ranging: Humans can localise without lidar, can robots? In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2018.
- [10] H.P. Moravec. Sensor Fusion in Certainty Grids for Mobile Robots. In *Sensor Devices and Systems for Robotics (SDSR)*, 1989.
- [11] R.A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A.J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proc. of the Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136, 2011.
- [12] A. Prorok and A. Martinoli. A reciprocal sampling algorithm for lightweight distributed multi-robot localization. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2011.
- [13] N. Radwan, G. Tipaldi, L. Spinello, and W. Burgard. Do You See the Bakery? Leveraging Geo-Referenced Texts for Global Localization in Public Maps. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2016.
- [14] A. Rosinol, M. Abate, Y. Chang, and L. Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2020.
- [15] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.
- [16] Y. Tao, X. Liu, I. Spasojevic, S. Agarwal, and V. Kumar. 3D Active Metric-Semantic SLAM. *IEEE Robotics and Automation Letters (RA-L)*, 9(3):2989–2996, 2024.
- [17] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.
- [18] K. Wada, E. Sucar, S. James, D. Lenton, and A.J. Davison. MoreFusion: Multi-object Reasoning for 6D Pose Estimation from Volumetric Fusion. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [19] C. Yi, I.H. Suh, G.H. Lim, and B.U. Choi. Bayesian robot localization using spatial object contexts. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2009.
- [20] F. Zafari, A. Gkelias, and K.K. Leung. A Survey of Indoor Localization Systems and Technologies. *IEEE Communications Surveys Tutorials (CST)*, 21(3):2568–2599, 2019.
- [21] N. Zimmerman, T. Guadagnino, X. Chen, J. Behley, and C. Stachniss. Long-Term Localization Using Semantic Cues in Floor Plan Maps. *IEEE Robotics and Automation Letters (RA-L)*, 8(1):176–183, 2023.
- [22] N. Zimmerman, H. Müller, M. Magno, and L. Benini. Fully Onboard Low-Power Localization with Semantic Sensor Fusion on a Nano-UAV using Floor Plans. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*. IEEE, 2024.
- [23] N. Zimmerman, M. Sodano, E. Marks, J. Behley, and C. Stachniss. Constructing Metric-Semantic Maps Using Floor Plan Priors for Long-Term Indoor Localization. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.
- [24] N. Zimmerman, L. Wiesmann, T. Guadagnino, T. Läbe, J. Behley, and C. Stachniss. Robust Onboard Localization in Changing Environments Exploiting Text Spotting. *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022.