Ensuring the Consistency of Heterogeneous World Representations Using Structural Features

Matteo Luperto[1], Jose-Luis Matez-Bandera[2], Tomasz Piotr Kucner[3],
Michele Antonazzi[1], Gabriele Somaschini[1], Javier Monroy[2], Javier Gonzalez-Jimenez[2], Nicola Basilico[1]

***Abstract*—This paper states the common lack of consistency existing between the different environment representations (e.g. a 2D occupancy grid map, a 3D semantic map, or a 3D floor plan, among others) held by a mobile robot. This inconsistency means that the same features in the environment could be represented differently in each representation; this lack of correspondence can indicate that different representations can capture different complementary aspects, or that some/all of them may be incorrect. The latter can lead the robot to make contradictory decisions when using more than one representation at the same time (e.g. 2D metric map for navigation and a 3D semantic map for object interaction). To mitigate this problem, we propose a framework for building consistent heterogeneous maps. The core idea is to build reliable correspondences between the features in each map and use it to improve the other ones. It is motivated by the fact that each representation will have its own strengths not present in the other ones. We illustrate this problem with a practical example using two different representations obtained from state-of-the-art methods.**

I. Introduction

Indoor environments, such as domestic, public, or industrial buildings, are particularly challenging for the deployment of autonomous mobile robots due to the presence of dynamic and static obstacles (e.g., objects, humans, or animals) and the generally narrow areas the robot must operate in, like passages, doorways, or rooms with dense furniture. A key requisite for an autonomous mobile robot to operate in such challenging scenarios is to have a stable and robust representation of its working environment. Only in this way, the robot can reliably localise itself and other mobile agents, and, eventually, perform autonomous navigation.

The representation of the environment, usually a metric *map*, is obtained by integrating different percepts acquired with the robot's sensory system. In literature, different techniques are used to acquire such environmental representation with good results [1]–[4]; by relying on these environmental representations, robots have shown of being able to navigate such environments fairly well. Still, it is often the case that these environment representations are not entirely accurate. An inaccurate world representation, combined with the complexity of indoor environments, results in a degradation of navigation and other tasks-related capabilities [1]. Ultimately, this causes the fact that autonomous robot working indoors are likely to experience navigation failures, thus requiring an external, and often costly, recovery.

[1] ML, MA, NB, GS are with the University of Milan, Italy.
[2] JB, JM, JJ are with the University of Malaga, Spain.
[3] TK is with the Aalto University, Finland.

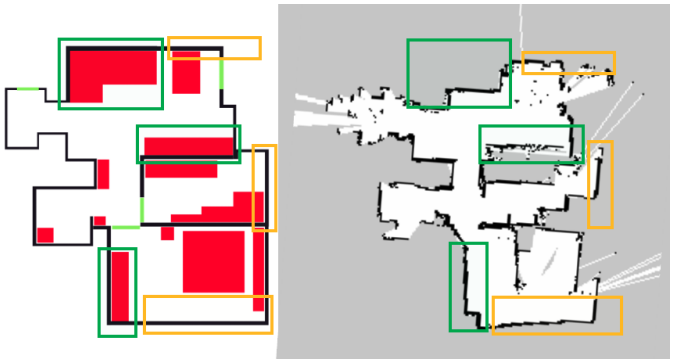


Fig. 1: The floor plan of an apartment (left), compared with its 2D map as acquired by the robot. The footprint of the furniture is indicated in red. The green boxes indicate portions of the 2D map where the walls are not observed by the 2D laser due to the occlusion from furniture but can be observed by an RGBD camera, as we show in Section IV.

Two of the most popular approaches to obtain a representation of the environment are to use 2D laser-range scanner data to build a 2D grid map of the environment, or to use RGB or RGBD images to reconstruct a 3D representation of the environment. The popularity of the use of laser-range scanners to build 2D grid maps is due to the fact that laser-range scanner data are robust to different types of environmental conditions, that well-established SLAM methods are able to build stable 2D grid maps from them, and because those maps and sensor data can be effectively exploited to perform several tasks as localisation, path planning, or path execution. The advantages of the latter approach, i.e. using vision-based data, lie in the fact that high-level information can be extracted (e.g. object class). By leveraging 3D information, it is also possible to detect openings in walls, which enriches the representation.

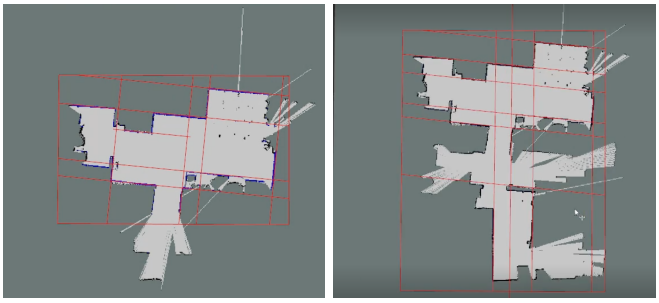
A defining property of indoor environments is that they are highly structured, as they are composed of walls, rooms, doors, and furniture, usually arranged in regular patterns as perpendicular or parallel wrt each other. This structure can be found both in *global structural features*, that are common to the entire environment, e.g. that walls in different rooms are perpendicular or parallel among them, as well as in *local structural features*, e.g. that a specific wall of a room is perceived as a straight line or it has an opening (e.g. a door) at a certain location. Maps often fail to represent this property, due to the fact that robot sensors are noisy, that

there is occlusion, and that SLAM methods are not able to compensate entirely for such inaccuracies. As an example, despite the fact that most walls in indoor environments follow a straight line, their representation in a map is often composed of a series of smaller line segments that have a similar, yet different, orientation (see highlighted areas with yellow boxes in Fig. 1). To enforce this property, several methods build upon the strong assumption of considering Manhattan worlds only, where all the environmental features are parallel or perpendicular among them. However, methods that follow this assumption are strongly limited when applied to non–Manhattan worlds.

When we consider the detection of structural features, 2D grid maps from laser–range scanner data are usually coherent with the structure of the indoor environment they represent; i.e. they represent most of the global structure features of the environment [5]–[8], as it is possible to identify in 2D grid maps the direction of walls, doorways, corners, and to segment the map in rooms [9], [10]. However, the inherent limitations of 2D laser-range scans can be appreciated in the resulting maps, showing areas where the structural features have not been mapped due to occlusions from furniture or movable objects, like those highlighted in green in Fig. 1, or the impossibility to extract semantic information regarding the nature of objects present, or the room category, to cite some. At the same time, 3D vision–based data are crucial to detect local structural features of the environment even in presence of occluding elements, and can be used to identify semantic knowledge [11]. Conversely, the task of combining different 3D local perceptions into a globally consistent map is far more challenging [12]–[14], resulting in the fact that 3D maps are not assured to be globally consistent. This implies that separate rooms in large-scale environments – that are actually aligned in the environment – are often not aligned in the 3D representation due to drift or poor localisation. Still, the understanding of the 3D structure of the environment is particularly useful for multiple tasks [15], [16]

Recent works [17], [18] have proposed to use hierarchical 2D-3D representations of the environment to overcome the limitations of 2D and 3D standalone approaches. However, 2D and 3D representations of the environment are usually obtained independently, thus resulting in two different and not consistent world representations.

The main topic of this position paper is to discuss a methodology to acquire robust representations of the environment that accurately captures the structure of indoor environments, with the aim of improving world–representation and navigation capabilities of mobile robots. To do this, we want to exploit the fact that both 2D and 3D representations are able, at different extents and with different levels of accuracy, to detect both global and local structural features of the environment. Heterogeneous map representations are not independent because the same features of the environment should manifest in each of them: enforcing the consistency of these features in different representations is the process of correctly representing such dependence. Our proposal is to ensure the consistency of local and global features detected with 2D and 3D approaches to compensate for each other errors; as a result, we can obtain two different representations (2D and 3D) that are consistent among them, and furthermore, globally consistent.



(a) (b)

Fig. 2: An example of the structural features obtained from ROSE on a 2D grid map. In red it can be seen the directions of the walls, while in blue are highlighted the components of the map that are due to global structural features. In the full 2D map (b), there are alignment issues due to mapping, so the directions of walls are not consistent with the map .

As a case study, we investigate the integration between a 2D and a 3D structure identification, ROSE^2 [19] and Sigma-FP [20]. We briefly present the two methods, by highlighting their strength and limitations, which are typical for 2D and 3D approaches. We then discuss how to obtain an integration between them by focusing on some relevant examples as obtained with real-world noisy data. We show how an agreement between 2D and 3D features can be used to reduce the uncertainty of both approaches. At the same time, of particular interest is a mismatch between 2D laser-based and 3D vision-based features. In this case, the two methods can consider the uncertainty of the two perceptions and integrate them accordingly.

II. ROSE AND ROSE^2

ROSE^2 [19] is a method for identifying global structural features in 2d grid maps. The method is composed of two main parts. The first one, called ROSE [21], identifies all the occupied cells that belong to the main structure (e.g., walls); at the same time, it filters out map components that are due to clutter, noise, and non-structural components. As a result of this, ROSE obtains a set of main *directions* of walls (i.e., the direction of the lines on which the walls lie). Commonly, in indoor environments, most walls follow one or two directions being either perpendicular or parallel wrt each other.

The second part focuses on grouping locally perceived portions of walls along the identified main directions of walls, following the assumption that a wall could be shared by multiple rooms. This allows identifying global structural features. The aim of this is to obtain a geometrical floor plan-like representation of the environment that is resilient to clutter and partial observations, and which is used to segment, i.e., to identify rooms in, the original occupancy grid map. An

example of the output of ROSE^2 can be seen, in red, in Fig. 2a. In this map, obtained in the same environment of Fig. 1 ROSE identifies the two main (perpendicular) directions common to all walls correctly, while red lines are the direction of single walls.

The main advantages of ROSE^2 are in the fact that it is able to retrieve global structural features from 2D maps even in the presence of severe clutter and inaccuracies. At the same time, ROSE^2 results are inaccurate in contexts where the map fails to represent the global features, as in those of Fig. 2b. In such a map, that represents the same environment of Fig. 1, the walls of rooms in the top part of the map are not aligned with those of the bottom part of the map, as they should be.

In settings like this, ROSE^2 is able to detect that the set of walls in the map follow the same direction, but it is not able to detect the correct directions robustly nor to compensate for this type of inconsistency.

III. SIGMA-FP

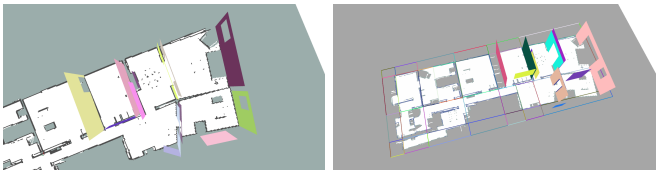
Sigma-FP [20] is a plane-based method that incrementally builds the 3D floor plan of an environment from a sequence of RGB-D images (see Fig. 3) while dealing with the unavoidable uncertainties of robot localization and plane extraction. Leveraging RGB-D data, Sigma-FP is able to delimit the walls even in case of high occlusion. Yet, it should be noted that for a decent reconstruction, the robot should navigate not far from walls as both, the field–of–view and the reliable range of the depth sensors, are usually limited compared to other employed sensors (e.g. 2D laser).

The output of Sigma-FP comprises the ensemble of detected walls, where each is defined by a set of *local structural features*: i) the number of openings (i.e. doors and windows) and their location on the wall, ii) the wall boundaries in Cartesian space, and iii) a multivariate Gaussian distribution over wall-plane representation in the Plane Parameter Space (PPS) –a minimum space for plane parameterisation–. While this set of local features provides a comprehensive depiction of each wall, walls are treated as individual entities, entailing a lack of *global structural features* and hence, failing to ensure global consistency. The latter could be compensated by assuming a Manhattan World, but the method would be impractical for non-Manhattan scenarios. Instead, Sigma-FP operates on the relaxed Atlanta World constraint, where it is challenging to identify the actual global structure of the environment.

IV. Case Study: Integrating 2D and 3D Structural Features

In this section, we show some preliminary examples of the integration of 2D and 3D structural feature detection methods. More precisely, we use the global features perceived in 2D maps - wall main directions as discovered by ROSE [21] - to align local features - wall planes detected by Sigma–FP [20]. Fig. 4a-b shows an example where neither of the 3 illustrated representations (i.e. 2D occupancy map, 2D-based features map, and 3D floor plan) are mutually consistent. The

Fig. 3: An instance of a 3D floor plan built by using Sigma-FP.



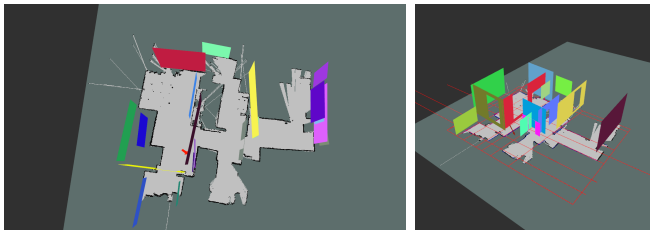
(a) (b)

Fig. 4: A practical scenario where heterogeneous representations of the same environment are not consistent with each other.In Fig. (b), 3D planes from Sigma-FP are aligned with the main wall directions extracted by ROSE.

explanation for this fact is that while the 2D occupancy map and the 3D floor plan representations include minor errors due to sensor noise and localisation drift, the 2D-based features map fails to represent properly the structure of the 2 rooms on the right side. The latter is because the environment mainly follows a Manhattan world except for these 2 rooms, hence the main directions captured by ROSE are Manhattan world and as consequence, ROSE considers the non-Manhattan rooms as mapping errors. In this sense, it can be noticed that none of the representations retain the entire structure of the environment correctly.

However, leveraging the strengths of each representation (i.e. while ROSE^2 is better at identifying global structural features, Sigma-FP finds particular value in detecting local structural features), we are able to compensate each other. Concretely, the small errors in the 3D floor plan can be rectified by aligning the walls with similar directions to the main directions given by ROSE^2. Referring to the non-Manhattan rooms misrepresented by the 2D-based features map, we can use the information from the 3D floor plan as it is more accurate locally and is obtained with low uncertainty.

The example of Fig. 5 shows a similar, yet different situation. In such an example, the 2D occupancy map does not represent accurately the environment. The latter prevents the robot from localizing properly and as a consequence, Sigma-FP represents some walls in the wrong location (as the one in yellow in the middle of Fig. 5a). In this sense, the integration with ROSE allows us to correct this error, as shown in Fig. 5b. Likewise occurs with the walls located in the room on the right



(a) (b)

Fig. 5: Another example where 2D– and 3D– features captures different aspects of the environment.

of Fig. 5a), which are aligned and merged into a single wall (in brown) in Fig. 5b. Conversely, occlusion prevents ROSE to observe some of the walls, as the green and yellow walls in the left room of Fig. 5a). In this case, Sigma-FP, which is able to observe these walls as illustrated in Fig. 5b, can add this information to the 2D-based features map.

V. Conclusion and Future Work

In this position paper, we have shown how 2D and 3D world representations, albeit useful, are often not coherent with the structural and shape of the environment. We have shown how the strengths and weaknesses of structural features from 2D and 3D data can compensate for each other, improving the overall coherence of structural features of both 2D and 3D representations. A mismatch between 2D and 3D data and 2D and 3D percepts can signal that there are inconsistencies between those maps and the actual shape of environments. In future work, we plan to research towards automatically identifying and compensating mismatches between heterogeneous representations, as well as leveraging their strengths to improve their mutual consistency.

References

[1] L. Kunze, N. Hawes, T. Duckett, M. Hanheide, and T. Krajnik, "Artificial Intelligence for Long-Term Robot Autonomy: A Survey," *IEEE RA-L*, vol. 3, no. 4, pp. 4023–4030, 2018.

[2] N. Hawes, C. Burbridge, F. Jovan, L. Kunze, B. Lacerda, L. Mudrova, J. Young, J. Wyatt, D. Hebesberger, T. Kortner *et al.*, "The Strands project: Long-term autonomy in everyday environments," *IEEE RAM*, vol. 24, no. 3, pp. 146–156, 2017.

[3] M. Luperto, M. Romeo, J. Monroy, J. Renoux, A. Vuono, F.-A. Moreno, J. Gonzalez-Jimenez, N. Basilico, and N. A. Borghese, "User feedback and remote supervision for assisted living with mobile robots: A field study in long-term autonomy," *Robotics and Autonomous Systems*, vol. 155, p. 104170, 2022.

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE T Robot*, vol. 32, no. 6, pp. 1309–1332, 2016.

[5] S.-Y. An and J. Kim, "Extracting statistical signatures of geometry and structure in 2d occupancy grid maps for global localization," *IEEE RA-L*, 2022.

[6] J. Chang, G. Lee, Y. Lu, and C. Hu, "P-SLAM: Simultaneous localization and mapping with environmental-structure prediction," *IEEE T Robot*, vol. 23, no. 2, pp. 281–293, 2007.

[7] R. Shrestha, F. Tian, W. Feng, P. Tan, and R. Vaughan, "Learned map prediction for enhanced mobile robot exploration," in *Proc. ICRA*, 2019, pp. 1197–1204.

[8] R. Bormann, F. Jordan, W. Li, J. Hampp, and M. Hägele, "Room segmentation: Survey, implementation, and analysis," in *Proc. ICRA*, 2016, pp. 1019–1026.

[9] A. Kleiner, R. Baravalle, A. Kolling, P. Pilotti, and M. Munich, "A solution to room-by-room coverage for autonomous cleaning robots," in *Proc. IROS*, 2017, pp. 5346–5352.

[10] F. Foroughi, J. Wang, A. Nemati, Z. Chen, and H. Pei, "Mapsegnet: A fully automated model based on the encoder-decoder architecture for indoor map segmentation," *IEEE Access*, vol. 9, 2021.

[11] R. Ambruş, N. Bore, J. Folkesson, and P. Jensfelt, "Meta-rooms: Building and maintaining long term spatial models in a dynamic world," in *Proc. IROS*, 2014, pp. 1854–1861.

[12] R. Ambruş, S. Claici, and A. Wendt, "Automatic room segmentation from unstructured 3-D data of indoor environments," *IEEE RA-L*, vol. 2, no. 2, pp. 749–756, 2017.

[13] S. Oesau, F. Lafarge, and P. Alliez, "Indoor scene reconstruction using feature sensitive primitive extraction and graph-cut," *ISPRS J Photogramm*, vol. 90, pp. 68–82, 2014.

[14] E. Turner, P. Cheng, and A. Zakhor, "Fast, automated, scalable generation of textured 3d models of indoor environments," *IIEEE J Sel Top Signa*, vol. 9, no. 3, pp. 409–421, 2015.

[15] H. Howard-Jenkins, J.-R. Ruiz-Sarmiento, and V. A. Prisacariu, "Lalaloc: Latent layout localisation in dynamic, unvisited environments," in *Proc. CVPR*, 2021, pp. 10 107–10 116.

[16] 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*, vol. 8, no. 1, pp. 176–183, 2023.

[17] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *Proc. ICRA*. IEEE, 2020, pp. 1689–1696.

[18] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception engine for 3d scene graph construction and optimization," in *Proc. RSS*, 2022.

[19] M. Luperto, T. P. Kucner, A. Tassi, M. Magnusson, and F. Amigoni, "Robust structure identification and room segmentation of cluttered indoor environments from occupancy grid maps," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7974–7981, 2022.

[20] J.-L. Matez-Bandera, J. Monroy, and J. Gonzalez-Jimenez, "Sigma-fp: Robot mapping of 3d floor plans with an rgb-d camera under uncertainty," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 539–12 546, 2022.

[21] T. P. Kucner, M. Luperto, S. Lowry, M. Magnusson, and A. J. Lilienthal, "Robust frequency-based structure extraction," in *Proc. ICRA*, 2021, pp. 1715–1721.