Journal of Intelligent & Robotic Systems (2023) 109: (2023) 109:86
https://doi.org/10.1007/s10846-023-02021-y

REGULAR PAPER

Localization Exploiting Semantic and Metric Information in Non-static Indoor Environments

Clara Gomez[1] · Alejandra C. Hernandez[1] · Ramón Barber[1] · Cyrill Stachniss[2]

Received: 17 January 2023 / Accepted: 17 November 2023 / Published online: 1 December 2023

Abstract
Mobile robot localization is an important task in navigation and can be challenging, especially in non-static environments as the scene naturally involves movable objects and appearance changes. In this paper, we address the problem of estimating the robot's pose in non-static environments containing movable objects. We understand as non-static environments, dynamic environments in which objects might be moved or changed their appearance. We propose a probabilistic localization approach that combines metric and semantic information and takes into account both, static and movable objects. We perform a pixel-wise association of depth and semantic data from an RGB-D sensor with a semantically-augmented truncated signed distance field (TSDF) in order to estimate the robot's pose. The combination of metric and semantic information increases the robustness w.r.t. movable objects and object appearance changes. The experiments conducted in a real indoor environment and a publicly-available dataset suggest that our approach successfully estimates robot pose in non-static environments and they show an improvement compared to robot localization based only on metric or semantic information and compared to a feature-based method.

Keywords Visual localization · Semantic localization · Dynamic environments · Monte Carlo localization

1 Introduction

Localization is an essential capability for a mobile robot operating in real-world environments. Despite being widely researched, robot localization in the real world still has many challenges, especially for non-static worlds. Dynamic environments are affected by moving and movable objects, appearance changes and external changes such as lighting. All these conditions increase the difficulty in localization.

Among the changes that affect dynamic environments, we focus on objects that may be moved or change their appearance. We refer to these environments as non-static which differ from dynamic environments as the latter also considers moving objects such as cars or people. Changes in non-static environments are challenging for localization as the robot's internal representation of the environment no longer matches the current state of the world, causing the robot to make wrong estimations or even get lost. This problem strongly differs from the challenges that bring moving objects in the environment. The main issue for mapping and localization that involve moving objects is that they occlude static and trustful areas of the environment. Thus, it is relevant to study moving and movable objects separately as the solution for each of the problems can be merged in a later step.

In this paper, we present an approach to robot localization in non-static indoor environments, but also static environments can benefit from our method. Through an RGB-D sensor, the robot performs probabilistic localization using metric information from the depth image and semantic information extracted from the RGB image. This combination is advantageous for localization as semantic information provides a good estimate for coarse localization and metric

✉ Clara Gomez
clgomezb@ing.uc3m.es

Alejandra C. Hernandez
alejhern@ing.uc3m.es

Ramón Barber
rbarber@ing.uc3m.es

Cyrill Stachniss
cyrill.stachniss@igg.uni-bonn.de

[1] Robotics Lab, University Carlos III of Madrid, Getafe, Spain

[2] University of Bonn, Bonn, Germany

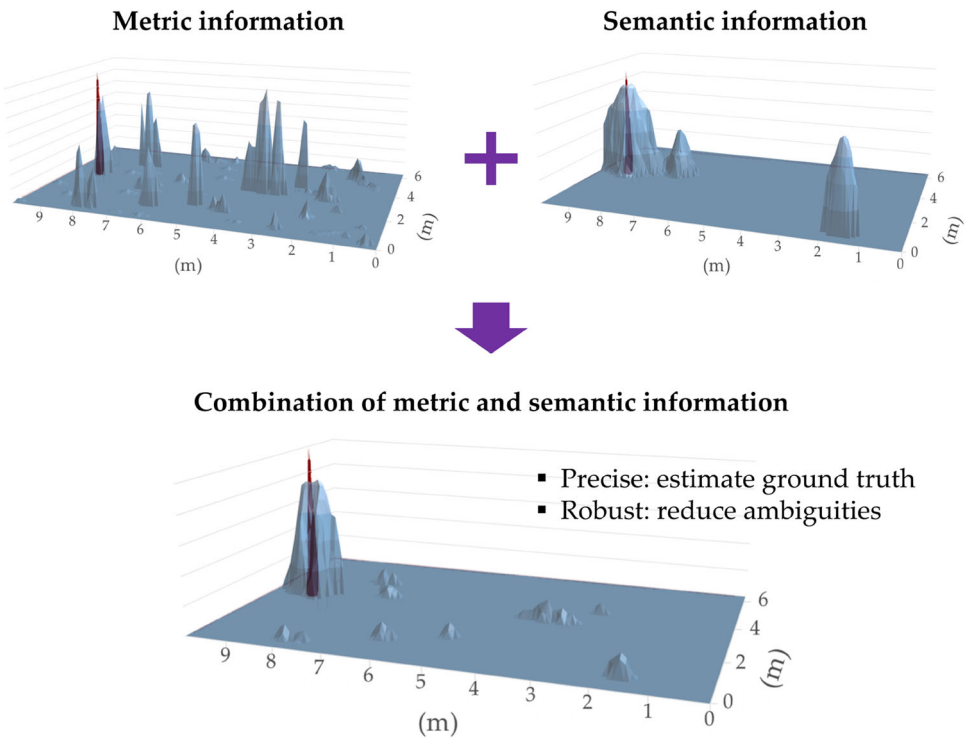

Fig. 1 Pose estimation for a given observation using only metric or only semantic information and with a combination of both. X and Y axes represent the environment size and Z axis the pose probability for the robot

information generates accurate fine-grained estimations in static environments (the degradation of the estimation based on metric information in non-static environments is compensated with the help of the semantic-based estimation). Figure 1 depicts an example of a metric observation model, a semantic model, and a joint one. The joint one provides a more peaked and less ambiguous pose estimate computed from an RGB-D image. Furthermore, most other localization approaches neglect dynamic and movable objects, using only the static parts for pose estimation [1–3]. However, the proposed method improves pose estimation using semantic and metric information of movable and static elements.

This paper proposes a novel RGB-D localization system for mobile robots operating in non-static indoor environments. The system exploits metric and semantic information for Monte Carlo localization in environments in which objects are movable or change in their appearance. We use a model representation based on truncated signed distance fields (TSDFs) augmented with semantic information. For localization, we use a particle filter to estimate the robot pose. The observation model of the particle filter is based on pixel-wise metric and semantic information based on objects. Our contributions can be summarized as follows:

- A novel localization system for mobile robots that combines metric and semantic information for pose estimation.
- The application of the localization system for the understudied scenario of non-static indoor environments.

The experimental evaluation shows that our approach succeeds in estimating robot pose in a real environment and a publicly-available dataset. Comparisons prove a better performance in non-static environments w.r.t. only-metric and only-semantic estimation and feature-based estimation. In sum, our localization approach based on metric and semantic information (i) provides better performance than metric-based localization even in static environments, and (ii) is robust to movable objects and performs localization over multiple sessions without remapping the environment.

2 Related Work

Several works use semantics for visual localization in static and dynamic environments. Semantic information is useful in static environments as it helps in pose disambiguation [4–6] and in loop closure detection [7–9]. In the work by Bavle et al. [4], semantics are included in localization by comparing the average distance of a given object in the image with the mapped elements.

In the context of dynamic environments, semantics in indoor environments have been mainly used for dealing with moving objects [1–3, 10–13]. Most approaches eliminate dynamic elements and rely on the static parts for estimation, but other approaches present inspiring solutions that can be applicable to movable objects too. This is the case of approaches such as CubeSLAM [12] that improve camera pose estimation by the inclusion of dynamic elements,

specifically object poses, their dimensions and their semantics.

Works dealing with non-static indoor environments, such as objects that are moved or change in their appearance, have been mainly approached from a feature-based perspective [14–17]. In the work by Patel et al. [17], semantically enhanced features are used for pose estimation. Images are semantically segmented at an object level and features are extracted for each detected object. Features are only matched if they belong to the same semantic type. Other works such as the work by Dayoub et al. [15] and Derner et al. [16] assign a weight to each feature according to its stability. In the work presented by Stachniss and Burgard [18], a Rao-Blackwellized particle filter estimates the robot pose from laser information in a non-static environment. In that work, the robot collects non-static information and pose estimation considers several configurations of the non-static areas.

Localization in outdoor dynamic environments has been approached from multiple perspectives: appearance-based methods that exploit image sequences [19, 20] or navigation sequences [21] and methods that exploit semantics [22–24]. Our work is more similar to the latter, especially localization using semantically-labeled observations of static and movable elements [22, 23, 25–27]. In the work by Toft et al. [27], localization combines feature matching and semantic information. Similarly to Patel et al. [17], images are semantically labeled and each feature is assigned a semantic type. Toft et al. [27] generate camera pose hypotheses and assign a score to them according to the semantically-labeled matches. Stenborg et al. [23] propose a semantic localization in which each pixel and 3D point are compared according to their semantic type. Chen et al. [22] use semantic information to filter dynamic elements and to improve ICP matching.

We present a probabilistic localization algorithm for indoor environments that includes semantic information in the estimation similar to [4] and [23]. Unlike most approaches for non-static indoor environments [15–17], we exploit semantic segmentation for every pixel and 3D point. However, we share with Dayoub et al. [15] and Derner et al. [16] a weighting method in the estimation process. In their case, each feature is weighted according to its stability. Our case includes a weighting method between metric and semantic information that can be linked to the persistence of objects.

3 Method

Given an environment model, we estimate the robot pose from depth and semantic information obtained from an RGB-D sensor. To this end, the map is built as a semantically-augmented truncated signed distance field (TSDF). Pose estimation exploits directly the TSDF and the semantics added to the model.

3.1 Environment Model Representation

A TSDF augmented with semantic information is used as model representation through *tsdf-fusion*, the work by Zeng et al. [28]. The idea behind TSDFs is to represent the environment as a 3D voxel grid in which each voxel contains an SDF value. The SDF provides the signed distance to the nearest surface for each voxel, positive distance means that the voxel is before the object surface and negative otherwise (inside the object or occluded by it). Voxels in a surface return a distance of 0. In TSDFs, distances are truncated to a maximum value. Voxels also contain a weight that represents how reliable the SDF value is.

As an extension of *tsdf-fusion*, we include a semantic value for each voxel. The semantic value is assigned to surface voxels and indicates the class of object that the voxel belongs to. During mapping, we assume the environment to be static. However, some conflicts may appear between voxels due to perspective changes. The weight of the TSDF addresses this problem w.r.t. metric conflicts. We have implemented a similar method for the semantic value as we count the times that a certain semantic value is assigned to a voxel. Then, we resolve conflicts by assigning to each voxel the most likely semantic value.

3.2 Monte Carlo Localization in Non-static Environments

We use a particle filter [29] to estimate the pose of a ground robot as they naturally deal with multimodal probability distributions. Particle filters calculate a belief over the robot pose using a set of weighted particles and each particle represents a candidate pose. First, the particle filter predicts the robot pose based on the odometry obtained from its wheel encoders. Then, a weight is assigned to each particle that represents how well the surroundings of the particle match with the robot observation. Finally, a new set of particles is created from the resampling of the old ones, where the chance of survival of a particle is proportional to its weight.

We use a standard odometry motion model to predict the new pose of the particles in a 2D plane. The odometry motion model uses odometry measurements from the robot's wheel encoders to calculate the pose increment between previous 3-DOF pose x_{t-1} and current 3-DOF pose x_t for each particle k. Given that increment, we calculate the new distribution of the particles, $p(x_t^{[k]} \mid u_t, x_{t-1}^{[k]})$ according to the model by Thrun et al. [30] where u_t refers to the motion command.

We propose an observation model that considers depth and semantic information to compute the particles weights. Each observation obtained from the RGB-D sensor consists of a depth image and an RGB image. We are transforming RGB images into semantic images by labeling each pixel with its semantic class or unknown class. Given a pixel of

the semantic image $q = \begin{bmatrix} a & b \end{bmatrix}^\top$, we can identify its corresponding depth $D(q)$ using the depth image. Backprojecting the pixel q, we can obtain its 3D point x:

$$\mathrm{x} = \begin{bmatrix} \frac{a-c_x}{f_x} D(q) \\ \frac{b-c_y}{f_y} D(q) \\ D(q) \end{bmatrix}, \tag{1}$$

where a and b are the pixel 2D coordinates in the image and c_x, c_y, f_x and f_y are the intrinsic parameters of the camera, assuming a pinhole camera model. This transformation results in a semantically annotated 3D representation for each single image.

Through the observation model, we update the particle weight w each time a new observation z_t is received taking into account the pose of each particle $x_t^{[k]}$ and the map of the environment m. We calculate the weight of the particle as the product of the probability obtained for each individual pixel i of the observation z_t, where N represents the total number of pixels in the observation:

$$w^{[k]} = \eta \; p(z_t \mid x_t^{[k]}, m) = \eta \prod_{i=1}^{N} p(z_t^i \mid x_t^{[k]}, m). \tag{2}$$

The weights of the particles are then normalized so they all add up to 1 using the normalization factor η.

In a similar spirit as beam models for range finders [31] where the observation model for a single beam is a mixture of four densities (see [30], page 157, eq. (6.12)), we calculate the probability for every pixel as the mixture of two probabilities: p_{sdf} accounting for metric information and p_{sem} for semantic information. The two different probabilities are mixed by a weighted average defined by the weighting factors z_{sdf} and z_{sem} with $z_{sdf} + z_{sem} = 1$. We calculate the observation model at a pixel level as:

$$p(z_t^i \mid x_t^{[k]}, m) = \begin{pmatrix} z_{sdf} \\ z_{sem} \end{pmatrix}^\top \begin{pmatrix} p_{sdf}(z_t^i \mid x_t^{[k]}, m) \\ p_{sem}(z_t^i \mid x_t^{[k]}, m) \end{pmatrix}, \tag{3}$$

with i referring to each pixel of the sensor data.

As mentioned before, we directly exploit the TSDF to calculate the metric probability p_{sdf}, since it directly represents the distance of the point to the nearest surface. Given the 3D point x computed for each pixel i (whose SDF value is 0 as pixels in the image are always surfaces), we can calculate its corresponding voxel in the model representation. The corresponding voxel satisfies that the 3D pose of the pixel is contained within the 3D volume of the voxel. The sdf value of the corresponding voxel is retrieved, denoted as sdf_i. The distribution p_{sdf} has a maximum at $sdf_i = 0$, which means that the point is already a surface. The variable σ_{sdf} represents the standard deviation for metric probability. We compute the likelihood for the metric information as:

$$p_{sdf}(z_t^i \mid x_t^{[k]}, m) = \exp\left(-\frac{sdf_i^2}{2\sigma_{sdf}^2}\right). \tag{4}$$

For the likelihood given semantic information, we introduce the semantic distance S_i. Given a pixel with a certain semantic class, S_i refers to the Euclidean distance between the position of the 3D projection of the pixel x and the position of the closest voxel in the map that belongs to the same semantic class. As before, p_{sem} takes the maximum value for $S_i = 0$, which means that the corresponding voxel for that point belongs to the desired semantic class. The standard deviation for semantic probability is denoted by σ_{sem}. We calculate the likelihood of the semantic cue as:

$$p_{sem}(z_t^i \mid x_t^{[k]}, m) = \exp\left(-\frac{S_i^2}{2\sigma_{sem}^2}\right). \tag{5}$$

We then resample the particles and otherwise run a standard Monte Carlo localization approach (MCL) [29].

4 Experimental Evaluation

The focus of this work is to develop a localization system based on metric and semantic information for robots operating in non-static indoor environments. Thus, we present experiments to show the capabilities of our method and to support our claims, which are: (i) our approach provides better performance than metric-based localization even in static environments, and (ii) it is robust to movable elements without the need to update the map in real environments and a publicly-available dataset [32]. In addition, a comparison to a feature-based approach, FAB-MAP [14], is provided.

4.1 Experimental Setup

Real-world experiments are performed using a Turtlebot 2 robot with an Asus Xtion RGB-D sensor. Figure 2 shows the semantically-augmented TSDF model. For the real-world experiments, manually-labeled (ground truth) semantic information and automatic semantic labeling from Mask R-CNN [33] are provided. Two examples showing the differences between manually-labeled and Mask R-CNN are included in Fig. 3. Each row in the image corresponds to a camera pose where column (a) shows automatic semantic labeling and (b) manual annotation. In the case of the publicly-available dataset just manual semantic labeling is

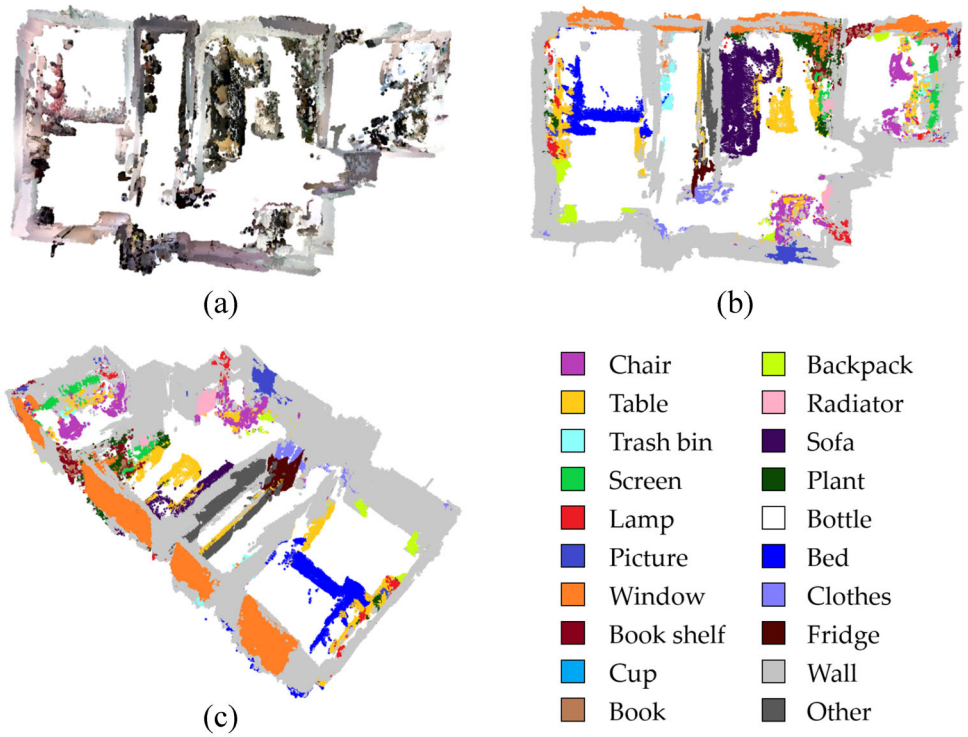

Fig. 2 TSDF of the real world used for the experiments: (a) shows an illustrative view of the TSDF with RGB information, (b) and (c) two views of the TSDF with semantic information. Color-coded semantics corresponds to the included table

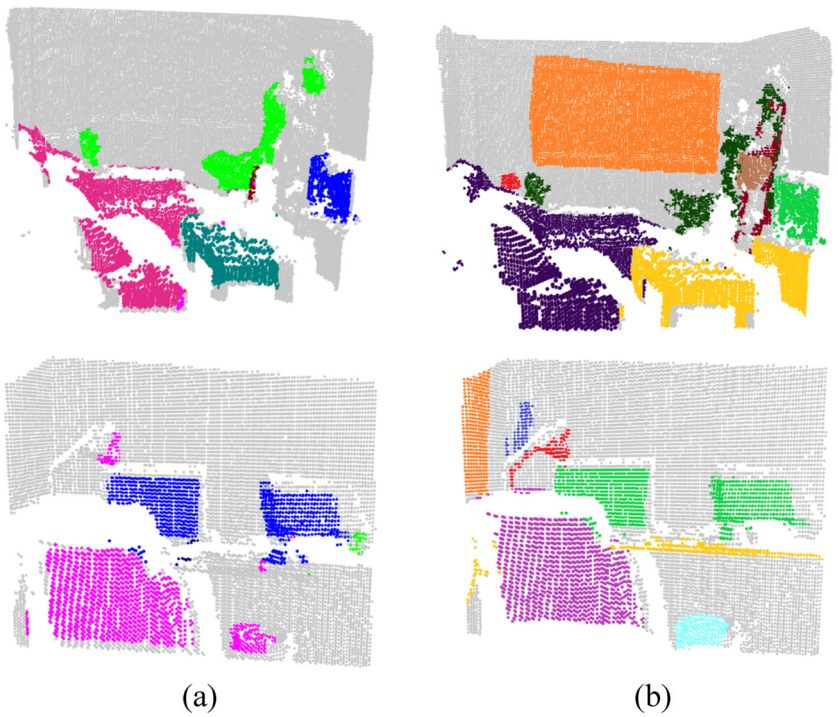

Fig. 3 Comparison of semantic segmentation methods for two images frames (each row): (a) shows automatic semantic labeling through Mask R-CNN and (b) and manual annotation. Notice that color coding is different for automatic semantic labeling and manual annotation, the latter follows the color code shown in Fig. 2. In addition, automatic labeling detects fewer object classes (i.e. the window) and fails to properly detect some object edges. The correction of the semantic labeling is out of the scope of this paper and we can expect a decrease in localization performance caused by the inaccuracies

used. Ground truth position is provided by the authors for the publicly-available dataset experiments and for the real-world environment it is extracted from an accurate laser-based sensor setup.

Regarding the observation model, metric and semantic information are weighted equally $z_{sdf} = z_{sem} = 0.5$. Another approach is setting the weight based on the change in appearance or the object class according to $z_{sdf} = 1 - \alpha$ and $z_{sem} = \alpha$ where α represents the change in appearance of the object class, see [34]. For MCL considering only metric information, we assign z_{sdf} to 1 and z_{sem} to 0 and, for semantic-only estimation, z_{sdf} to 0 and z_{sem} to 1. For all the experiments, σ_{sem} was set to 1.5m and σ_{sdf} to 2m; these parameters were experimentally defined in an initial test and they remained fixed for the multiple tests in the two environments considered in this evaluation (real-world environment and Witham Wharf dataset).

4.2 Performance in a Real Static Environment

To support the claim that the combination of metric and semantic information improves robot localization in static environments, we run the localization algorithm in a real environment for different paths without including any change.

4.2.1 Illustrative Example for Single Observation

First of all, we show how pose estimation for a single observation varies when considering only metric or only semantic information, and the combination of both. Given the observation shown in Fig. 4(a), the metric-based estimation is shown in Fig. 4(b), the semantic-based estimation in Fig. 4(c) and the estimation combining both sources of information in Fig. 4(d). For this experiment, we evaluated the observation model on a dense grid of 10 cm. For each grid position, we evaluate 16 orientations, the one that gives the maximum weight gets represented in Fig. 4 in which greater z-values represent higher probability for that x-y position. Analyzing the results, we can see that the metric-based model is multimodal as many places in the environment could match the proposed observation. The semantic-based model identifies two modes (very close to each other) that correspond to the two book shelves with books that are in the environment. Finally, the combined estimation benefits from both models and the belief is peaked around the ground truth (shown in red). This simple and isolated example illustrates how position probabilities react for metric and semantic information and the benefits of the combination of both. Additionally, it offers a ground understanding for the following experiments.

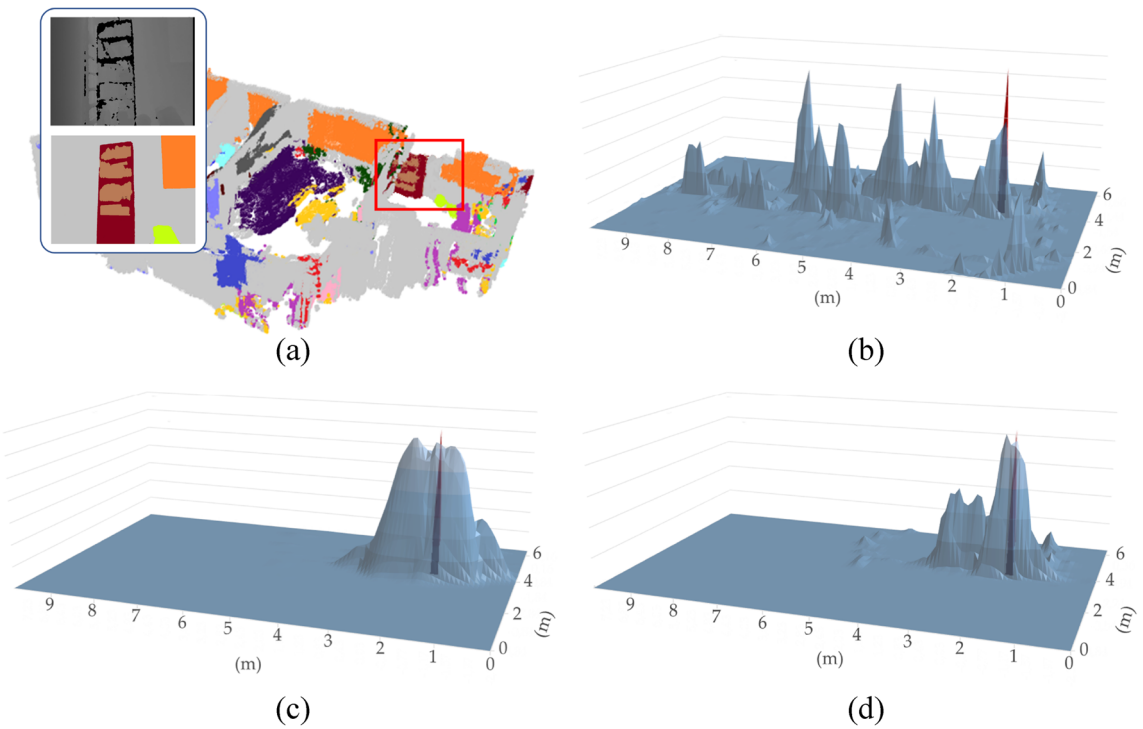

Fig. 4 Estimation given a single observation in the static environment: (a) shows the depth and semantic images for the observation and the TSDF with a red square highlighting the observed area; (b), (c) and (d) show the metric-based, semantic-based and combined estimation, respectively. The ground truth is represented in red. X and Y axes represent the environment size and Z axis the pose probability for the robot

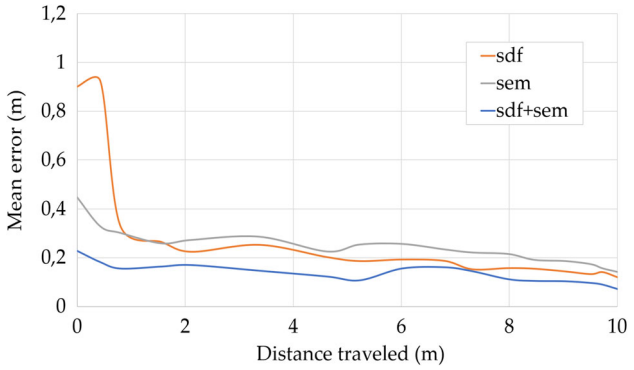

Fig. 5 Position error (m) w.r.t. distance traveled in the static environment for the three studied cases: metric (sdf), semantic (sem) and combined (sdf+sem)

4.2.2 Pose Estimation while Navigating in the Real Static Environment

Results for a path in the initial environment show quantitatively how the position error differs between the three cases: metric (sdf), semantic (sem) and combined (sdf+sem). We have initialized the particle filter with 3000 particles and run the experiment 5 times to obtain average results. Figure 5 shows how the mean error evolves along the path. Once the filter has converged, our approach (sdf+sem) obtains an average error of 13.1 cm with a variance of 1.2 cm, using manually-annotated semantics, and 14.9 cm with a variance of 1.54 cm using Mask R-CNN. In addition, the solution with combined metric and semantic information is the fastest to converge. The average pose results after convergence are summarized in Table 1. Mean and variance in position error, ε_p and σ_p^2, and heading error, ε_h and σ_h^2, are included for the three cases. Results including semantics are split according to manual annotation (sem_gt) or automatic labeling (sem_rcnn). Metric-based estimation fails to predict the pose in one execution and thus it obtained high error values. If the erroneous execution is overlooked (also included in Table 1) the mean position error would be 24.2 cm, still twice as large as for our approach. In addition, we can observe that the usage of manually-annotated or automatically-labeled semantic data does not strongly impact the results, as using Mask R-CNN decreases accuracy by only 1.82 cm compared to "perfect" manual annotations. This experiment demonstrates that the combination of metric and semantic information can improve the accuracy of pose estimation in real-world environments, as the two experiments using this combination outperform the other approaches.

4.2.3 Pose Estimation based on Number of Particles

In addition to evaluating the accuracy and the convergence with a given number of particles, we wanted to show the efficiency and robustness of the approaches to the number of particles. A lower number of particles leads to a faster pose estimation. Figure 6 shows the evolution of position error w.r.t. the number of particles. Our approach requires less than 1000 particles to obtain an average error of 20 cm, whereas metric-based and semantic-based estimation require approximately 4000 and 3000 particles, respectively. These results are obtained from the average of 5 executions for each number of particles configuration. This experiment demonstrates that our method could run faster than the other approaches and still obtain an acceptable accuracy. Additionally, it suggests that our approach is more robust against the random initialization of the particles.

Table 1 Mean error and variance in static environment

Method	ε_p (m)	σ_p^2 (cm)	ε_h (rad)	σ_h^2 (rad)
sdf (with error)	0.4371	59.31	0.0954	0.01819
sdf (without error)	0.2423	3.58	0.0697	0.01284
sem_gt	0.1894	2.50	0.0681	0.0085
sem_rcnn	0.1741	1.99	0.0739	0.0092
sdf+sem_gt	**0.1309**	**1.23**	**0.0541**	**0.0047**
sdf+sem_rcnn	0.1491	1.54	0.0615	0.0053

Bold indicates the best results among the compared ones (lowest mean error and variance)

4.3 Performance in a Real Non-static Environment

This experiment supports the claim that our approach deals with changes through the combination of metric and semantic information. To this end, we run pose estimation for paths of approximately 10 m in different mapping sessions that involve objects that changed in their appearance or objects that were moved, removed or added, as shown in Fig. 7.

First of all, we evaluate the convergence capability with movable objects for the studied cases: metric (sdf), semantic (sem) and our approach (sdf+sem). For this experiment, the particle filter is initialized with 3000 particles and we

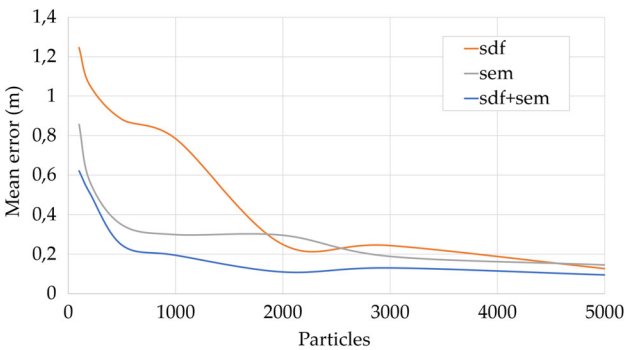

Fig. 6 Position error (m) w.r.t. number of particles for the studied cases: metric (sdf), semantic (sem) and combined (sdf+sem)

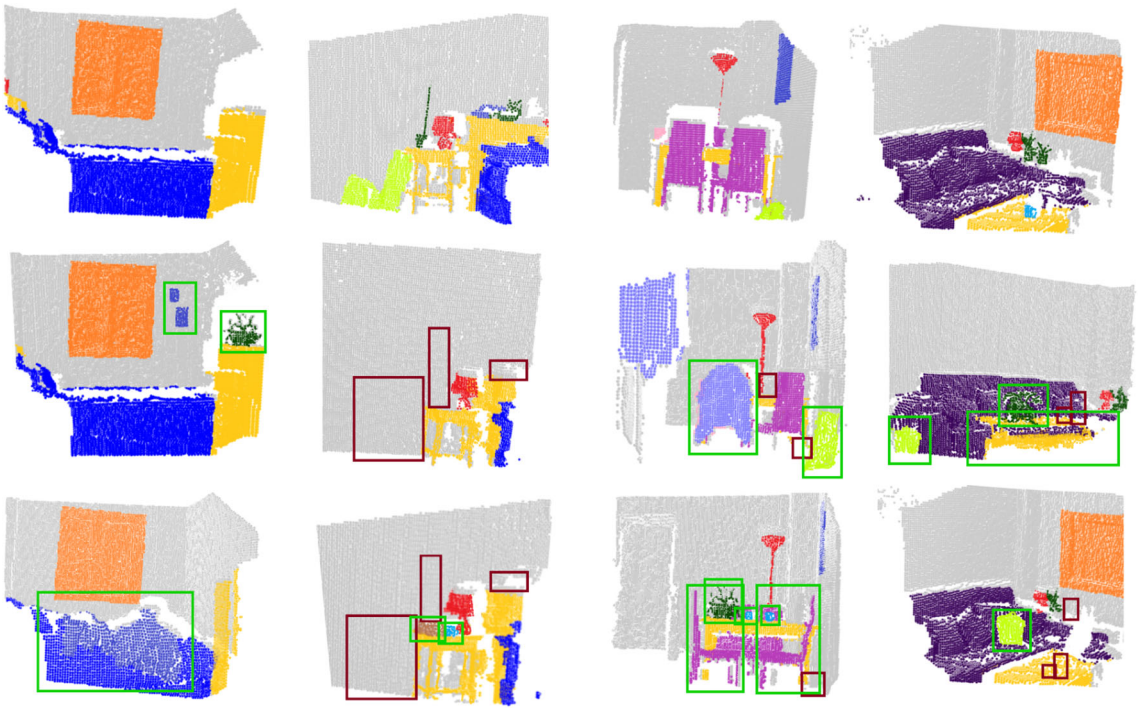

Fig. 7 Examples of changes: the first row corresponds to the static environment and the second and third rows include changes such as removed objects (red) and new or appearance chance (green)

have executed the 4 different paths 3 times for the three cases. Figure 8 illustrates how the position error evolves as the robot moves. We can distinguish an initialization phase until the distance traveled is approximately 1.75 m. During pose tracking, the three methods show occasional increases in mean position error due to movable objects and changes. Our approach (blue) obtains lower mean error during path execution and it is more robust as it always keeps a mean error close to 20 cm. Figure 9 shows the estimation for one of the paths. Our approach (blue) is the first to converge to the ground truth and it successfully tracks robot pose along the path. Table 2 shows average pose errors for the three cases after convergence. Position and heading errors refer to the mean error and variance, ε_p and σ_p^2, and ε_h and σ_h^2. Results including semantics are split according to manual annotation (sem_gt) or automatic labeling (sem_rcnn). Comparing to the static scenario, the mean position error for metric-only increases 2.16 times, for semantic-only using manual semantics 1.92 times and using automatic semantics 1.94 times and for our method 1.46 times (with manual semantics) and 1.49 times (with automatic semantics). In addition, mean position error for our approach in the non-static environment is lower than metric-only error in the static one. As in the static environment case, the accuracy is not highly affected by the use of automatically-labeled or manually-annotated. These results suggest that appearance changes have a low impact on our method exploiting metric and semantic information jointly.

4.4 Time Performance in a Real Environment

The pose estimation method has not been optimized for real-time execution, however the aim of this evaluation is to show that computing metric and semantic information does not create additional overhead compared to metric-only estimation. This experiment is executed in an Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz with 16GB RAM and

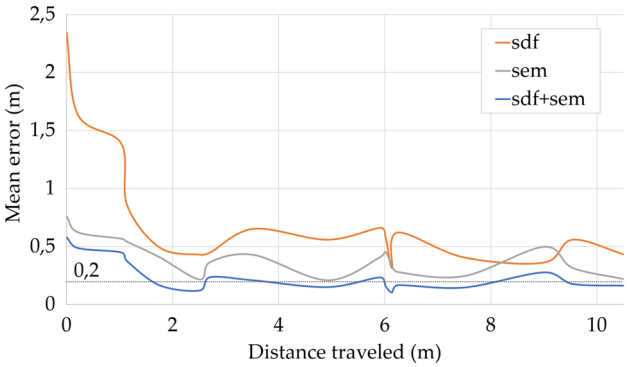

Fig. 8 Position error (m) w.r.t. distance traveled in non-static environment for the three studied cases: metric (sdf), semantic (sem) and combined (sdf+sem)

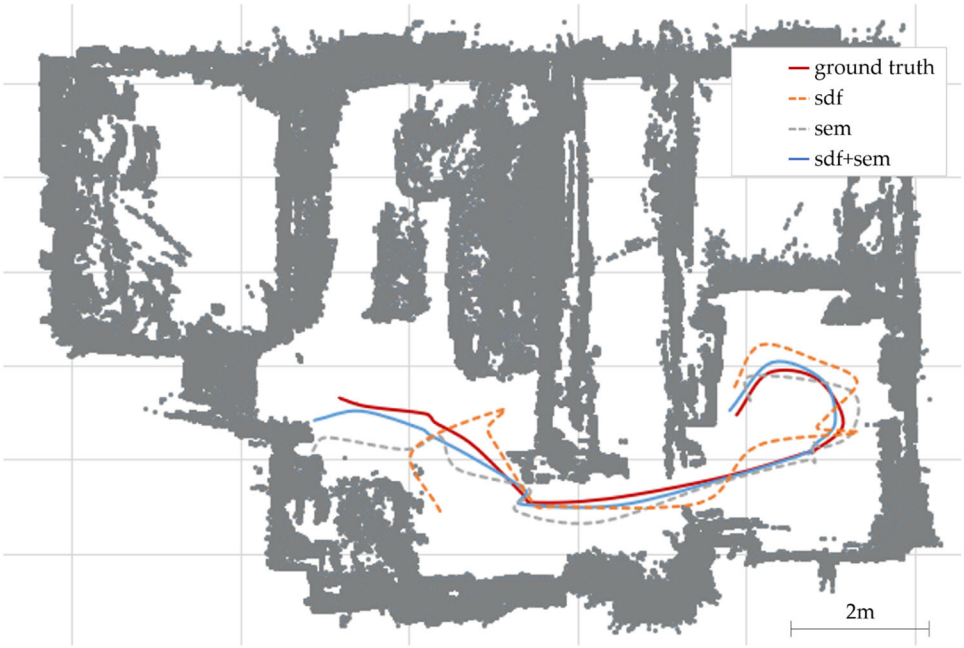

Fig. 9 Pose estimation for a path after changes. Ground truth and the estimation for the three studied cases is included

NVIDIA GeForce GTX 1050 Ti GPU using the pose estimation with 3000 particles. Metric-only estimation takes on average 32.13 ms to calculate each particle weight, semantic-only takes 9.34 ms and metric and semantic estimation takes 29.46 ms. The difference between metric- and semantic-only estimations is caused by the reduced number of operations to calculate the semantic weight. And the difference between metric-only and metric and semantic estimation is caused by the corresponding voxel search duration. The latter is on average more accurate and this speeds up the search for the corresponding voxel.

4.5 Performance Evaluation on Witham Wharf Dataset

This experiment supports the statement that the combination of metric and semantic information also leads to an improvement of pose estimation in other environments. Witham Wharf [32] is selected as comparing dataset as it is a real-world environment that contains multiple visits to the same office environment in which mainly objects have been moved around, which is the target of the proposed method. A partial 3D map with manually-annotated semantics is shown in Fig. 10. This is a challenging environment for semantic mapping as there are fewer, sparser and more repetitive objects. We used one daylight session of *training_Nov* to build the TSDF and evaluate our approach in the static scenario and 5 random daylight sessions from *testing_Dec* for the non-static scenario. Figure 11 shows the average position error as the robot moves in both scenarios. We have used dashed lines to represent position errors in the static scenario and solid lines for the non-static cases. We observe on average higher errors for both scenarios than in the real-world experiment and longer times to convergence to acceptable error values. This can be caused by the greater dimensions of this environment and the sparsity and repeatability of the semantic elements. Even with these issues, we see that the configuration using metric and semantic information is overall the one with lower position error. To see these results in detail, Table 3 shows average pose errors after convergence. Position and heading errors are listed according to the mean error and variance. In the table we can see, the greater average errors in position and heading for the three cases compared to the real-world experiment. In the case of our approach, we can see specially a degradation in heading estimation, being lower the error for the metric case. However, we still can see the improvement due to the joint information in Witham Wharf dataset.

Table 2 Mean error and variance in environment with changes

Method	ε_p (m)	σ_p^2 (cm)	ε_h (rad)	σ_h^2 (rad)
sdf	0.5269	16.93	0.1031	0.0146
sem_gt	0.365	7.31	0.2135	0.0481
sem_rcnn	0.3370	9.88	0.2014	0.0394
sdf+sem_gt	**0.1912**	**2.47**	**0.0861**	**0.0117**
sdf+sem_rcnn	0.2222	6.74	0.1271	0.0183

Bold indicates the best results among the compared ones (lowest mean error and variance)

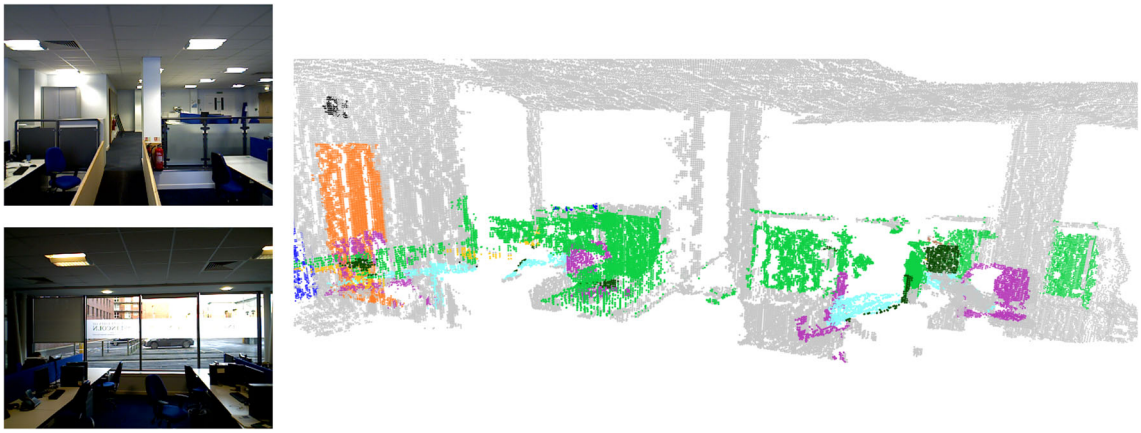

Fig. 10 RGB images from Witham Wharf dataset and partial semantically-annotated 3D map

4.6 Comparison to a Feature-based Method

As mentioned, localization for non-static environments has been mainly approached with feature-based methods. Thus, it is interesting to compare our approach with a feature-based method, FAB-MAP [14], which is a place recognition method that determines if a new observation comes from a known place or from a new one. The authors show that FAB-MAP is able to deal with appearance changes, which is applicable to our experimental scenario. We have executed FAB-MAP in the same paths in the real environment and the same mapping sessions on Witham Wharf dataset for the static and non-static scenarios. We perform place recognition with FAB-MAP for 17 query images in the experiment in the real environment and 8 query images for the dataset experiment. Figure 12 shows the results for the real environment and Fig. 13 for Witham Wharf dataset. Each box represents the probability assigned by FAB-MAP to the ground truth image, being green a high probability and red a low one. At first sight, we can see how FAB-MAP overall assigns a high probability (green) to the ground truth image in the static scenarios, which indicates its good performance. However, it clearly degrades when changes are introduced as ground truth images do not get assigned the higher probabilities (represented in yellow and orange). Quantitatively, in the static scenarios FAB-MAP obtains an average probability for the correct image of 0.74 in the real environment and 0.99 in the dataset. However, the estimation in the non-static scenario obtains worse results, especially in the real environment. The average probability of the ground truth image in the real environment is 0.17 and in Witham Wharf dataset 0.42. Although we cannot perform a direct comparison due to the different goals of each approach, we observe higher errors in FAB-MAP when dealing with changes. FAB-MAP fails to predict the correct place for most observations and our method successfully maintains a low pose error.

In summary, our evaluation suggests that our method increases localization accuracy in non-static indoor environments. It also upholds that the combination of metric and semantic information is a key aspect for real-world operation in static and non-static environments.

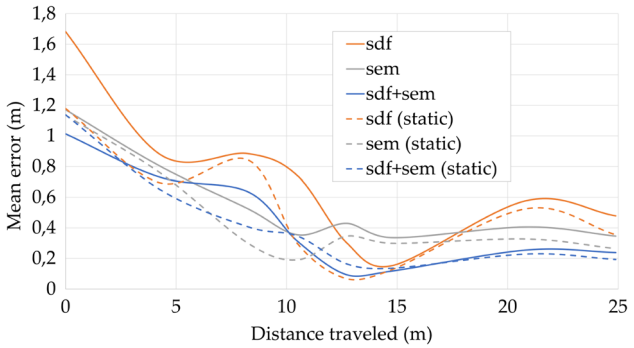

Fig. 11 Position error (m) w.r.t. distance on Witham Wharf dataset for the static (dashed line) and non-static scenarios in the studied cases: metric (sdf), semantic (sem) and combined (sdf+sem)

Table 3 Mean error and variance on Witham Wharf dataset

Method	ε_p (m)	σ_p^2 (cm)	ε_h (rad)	σ_h^2 (rad)
sdf (static)	0.4157	24.35	**0.1766**	**0.0667**
sem (static)	0.355	15.57	0.5077	0.345
sdf+sem (static)	**0.3041**	**12.08**	0.3911	0.1462
sdf	0.5741	39.78	**0.182**	**0.0556**
sem	0.4573	23.28	0.5118	0.3272
sdf+sem	**0.3402**	**16.84**	0.2639	0.087

Bold indicates the best results among the compared ones (lowest mean error and variance) both for static and non-static scenarios

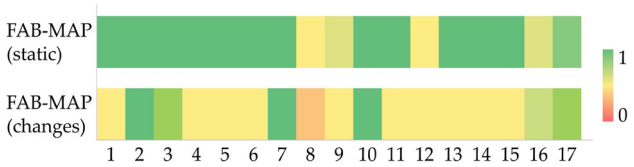

Fig. 12 FAB-MAP estimation for the real environment. The probability of the correct image is coded from green (high) to red (low)

5 Conclusion and Future Work

In this paper, we presented a novel approach to visual localization in non-static indoor environments. Our approach combines metric and semantic information to perform robot pose estimation. Our method uses information from movable objects (in addition to the widely-used static objects) because movable objects also provide valuable cues for localization. This allows us to successfully estimate robot pose even when the representation of the environment does not match the current state of the world. We evaluated our approach in a real indoor environment and Witham Wharf dataset and provided comparisons to only-metric and only-semantic methods and to FAB-MAP feature-based approach. With the experimental evaluation, we supported the claims made in this paper which were a better localization performance in static and non-static environments and higher robustness to movable objects without the need of remapping. The results suggest that the combination of metric and semantic information makes the approach more robust as fewer particles are needed for similar pose errors. In addition, the overall pose error is reduced both in static and non-static environments.

Despite these encouraging results, the proposed system has several limitations. Firstly, our method is meant for environments that have dense and varied objects, otherwise the semantic estimation will not provide a significant improvement. Secondly, adding semantics helps in the coarse localization but still requires that most objects remain in the same coarse locations between mapping sessions. The last limitation is the high processing time, which is far from real time. To overcome the mentioned limitations, future lines of work would be: first, to include semantics for structural elements in environments with low density of objects. Additionally, adapting the weights of the measurement model of the particle filter (as explained in the experimental setup) could help to identify which objects should not be matched for the new mapping session (because they are probably not anymore in that position). And lastly, instead of time-consuming dense voxel-based calculation of particle weights a more sparse representation could be used.

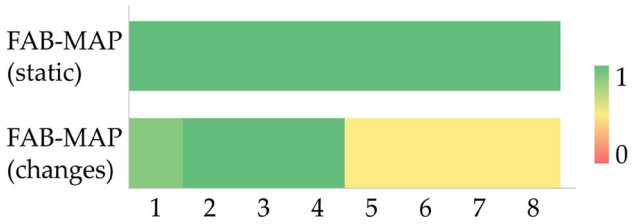

Fig. 13 FAB-MAP estimation on Witham Wharf dataset. The probability of the correct image is coded from green (high) to red (low)

Author Contributions All authors contributed to the study conception and design. Implementation, data collection and analysis were performed by Clara Gomez. The first draft of the manuscript was written by Clara Gomez and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work has partially been funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017008 (Harmony), also by HEROITEA: Heterogeneous Intelligent Multi-Robot Team for Assistance of Elderly People (RTI2018-095599-B-C21), funded by Spanish Ministerio de Economia y Competitividad, and the RoboCity2030 – DIH-CM project (S2018/NMT-4331, RoboCity2030 – Madrid Robotics Digital Innovation Hub).

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Han, S., Xi, Z.: Dynamic Scene Semantics SLAM Based on Semantic Segmentation. IEEE Access **8**, 43563–43570 (2020)
2. Palazzolo, E., Behley, J., Lottes, P., Giguère, P., Stachniss, C.: ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. arXiv preprint (2019)

3. Li, S., Lee, D.: RGB-D SLAM in dynamic environments using static point weighting. IEEE Robot. Autom. Lett. (RA-L) **2**(4), 2263–2270 (2017)
4. Bavle, H., Manthe, S., de la Puente, P., Rodriguez-Ramos, A., Sampedro, C., Campoy, P.: Stereo visual odometry and semantics based localization of aerial robots in indoor environments. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 1018–1023 (2018)
5. Zhang, W., Liu, G., Tian, G.: A coarse to fine indoor visual localization method using environmental semantic information. IEEE Access **7**, 21963–21970 (2019)
6. Cramariuc, A., Tschopp, F., Alatur, N., Benz, S., Falck, T., Brühlmeier, M., Hahn, B., Nieto, J., Siegwart, R.: SemSegMap–3D segment-based semantic localization. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 1183–1190 (2021)
7. Hughes, N., Chang, Y., Carlone, L.: Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization. Robotics: Science and Systems (RSS) (2022)
8. Yu, J., Shen, S.: SemanticLoop: loop closure with 3D semantic graph matching. IEEE Robot. Autom. Lett. (RA-L) **8**(2), 568–575 (2022)
9. Kim, J.J., Urschler, M., Riddle, P.J., Wicker, J.S.: Closing the Loop: Graph Networks to Unify Semantic Objects and Visual Features for Multi-object Scenes. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 4352–4358 (2022)
10. Xiao, L., Wang, J., Qiu, X., Rong, Z., Zou, X.: Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. J Robot. Autonomous Syst. (RAS) **117**, 1–16 (2019)
11. Xu, B., Li, W., Tzoumanikas, D., Bloesch, M., Davison, A., Leutenegger, S.: MID-Fusion: Octree-based object-level multi-instance dynamic SLAM. In: Proceedings of the IEEE International Conference on Robotics & Automation (ICRA), IEEE, pp. 5231–5237 (2019)
12. Yang, S., Scherer, S.: CubeSLAM: Monocular 3D object SLAM. IEEE Trans. Robot. **35**(4), 925–938 (2019)
13. Singh, G., Wu, M., Lam, S.-K., et al.: Hierarchical Loop Closure Detection for Long-term Visual SLAM with Semantic-Geometric Descriptors. In: IEEE International Intelligent Transportation Systems Conference (ITSC), IEEE, pp. 2909–2916 (2021)
14. Cummins, M., Newman, P.: FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. Intl. J Robot. Res. (IJRR) **27**(6), 647–665 (2008)
15. Dayoub, F., Duckett, T.: An adaptive appearance-based map for long-term topological localization of mobile robots. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 3364–3369 (2008)
16. Derner, E., Gomez, C., Hernandez, A.C., Barber, R., Babuška, R.: Towards life-long autonomy of mobile robots through feature-based change detection. In: Proceedings of the European Conference on Mobile Robotics (ECMR), IEEE, pp. 1–6 (2019)
17. Patel, N., Khorrami, F., Krishnamurthy, P., Tzes, A.: Tightly Coupled Semantic RGB-D Inertial Odometry for Accurate Long-Term Localization and Mapping. In: Proceedings of the International Conference on Advanced Robotics (ICAR), IEEE, pp. 523–528 (2019)
18. Stachniss, C., Burgard, W.: Mobile Robot Mapping and Localization in Non-Static Environments. In: Proceedings of the National Conference on Artificial Intelligence (AAAI), Pittsburgh, PA, USA, pp. 1324–1329 (2005)
19. Vysotska, O., Stachniss, C.: Lazy Data Association For Image Sequences Matching Under Substantial Appearance Changes. IEEE Robot. Autom. Lett. (RA-L) **1**(1), 213–220 (2016)
20. Vysotska, O., Stachniss, C.: Effective Visual Place Recognition Using Multi-Sequence Maps. IEEE Robot. Autom. Lett. (RA-L) **4**, 1730–1736 (2019)
21. Milford, M., Wyeth, G.F.: SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: Proceedings of the IEEE International Conference on Robotics & Automation (ICRA) (2012)
22. Chen, X., Milioto, A., Palazzolo, E., Giguère, P., Behley, J., Stachniss, C.: SuMa++: Efficient LiDAR-based Semantic SLAM. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2019)
23. Stenborg, E., Toft, C., Hammarstrand, L.: Long-term visual localization using semantically segmented images. In: Proceedings of the IEEE International Conference on Robotics & Automation (ICRA), IEEE, pp. 6484–6490 (2018)
24. Arshad, S., Kim, G.-W.: Leveraging Semantics in Appearance based Loop Closure Detection for Long-Term Visual SLAM. In: IEEE International Conference on Big Data and Smart Computing (BigComp), IEEE, pp. 371–372 (2023)
25. Chebrolu, N., Lottes, P., Läbe, T., Stachniss, C.: Robot Localization Based on Aerial Images for Precision Agriculture Tasks in Crop Fields. In: Proceedings of the IEEE International Conference on Robotics & Automation (ICRA), IEEE, pp. 1787–1793 (2019)
26. Radwan, N., Valada, A., Burgard, W.: Vlocnet++: Deep multitask learning for semantic visual localization and odometry. IEEE Robot. Autom. Lett. (RA-L) **3**(4), 4407–4414 (2018)
27. Toft, C., Stenborg, E., Hammarstrand, L., Brynte, L., Pollefeys, M., Sattler, T., Kahl, F.: Semantic match consistency for long-term visual localization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 383–399 (2018)
28. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1802–1811 (2017)
29. Dellaert, F., Fox, D., Burgard, W., Thrun, S.: Monte Carlo Localization for Mobile Robots, booktitle = IEEE International Conference on Robotics and Automation (ICRA). (1999)
30. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics (Intelligent Robotics and Autonomous Agents). The MIT Press, Cambridge, Massachusetts (2005)
31. Fox, D., Burgard, W., Thrun, S.: Markov localization for mobile robots in dynamic environments. J Artif. Intell. Res. **11**, 391–427 (1999)
32. Krajník, T., Fentanes, J.P., Mozos, O., Duckett, T., Ekekrantz, J., Hanheide, M.: Long-term topological localisation for service robots in dynamic environments using spectral maps. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 4537–4542 (2014)
33. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
34. Gomez, C., Hernandez, A.C., Derner, E., Barber, R., Babuška, R.: Object-Based Pose Graph for Dynamic Indoor Environments. IEEE Robot. Autom. Lett. (RA-L) **5**(4), 5401–5408 (2020)

Clara Gomez received her Ph. D. (Cum Laude) degree in Electric, Electronic and Automation Engineering in 2021 from University Carlos III of Madrid (UC3M) in which developed semantically-enhanced robot navigation and mapping tasks. She also obtained her M. Sc. and B. Sc. degrees at University Carlos III of Madrid. Concretely, she earned her M. Sc. (Cum Laude) degree in Robotics and Automation in 2016 and her B. Sc. degree in Electronic Engineering and Automation in 2013. Currently, she is a Senior Researcher at Ericsson Research, Stockholm, Sweden. Her research focuses on semantic mapping and localization, radio-based localization and computer vision and machine learning for resource-constrained devices.

Alejandra C. Hernandez received her Ph.D. (Cum Laude) in Electrical Engineering, Electronics and Automation (2021) from University Carlos III of Madrid, Spain (UC3M). Previously, she obtained her M.Sc. degree in Robotics and Automation from UC3M and her B.Sc. degree (Cum Laude) in Systems Engineering from Antonio José de Sucre National Experimental Polytechnic University (UNEXPO), Venezuela. Currently, she is a Senior Researcher at Ericsson Research, Stockholm, Sweden. Her research areas include semantic understanding for autonomous robots, computer vision, machine learning and distributed applications for resource-constrained devices.

Ramón Barber is Associate Professor of the System Engineering and Automation Department, at the University Carlos III of Madrid, Spain. He received the B.Sc. degree in Industrial Engineering from Polytechnic University of Madrid (1995), and the Ph. D. degree in Industrial Technologies from the University Carlos III (2000). His research topics are focused on Robotics, Mobile Robotics and Mobile manipulators including robot control, perception of the environment, environment modelling, planning, localization, and navigation tasks, considering geometrical, topological, and semantic representations. He is member of the International Federation of Automatic Control (IFAC) and Vice President of the IEEE Spanish Chapter of Robotics and Automation Society (RAS).

Cyrill Stachniss is a full professor at the University of Bonn, a Visiting Professor in Engineering at the University of Oxford, and with the Lamarr Institute for Machine Learning and AI. He is the spokesperson of the DFG Cluster of Excellence PhenoRob at the University of Bonn. Before his appointment in Bonn, he was with the University of Freiburg and ETH Zurich. His research focuses on probabilistic techniques and learning approaches for mobile robotics, perception, and navigation. The main application areas of his research are agricultural robotics, self-driving cars, and service robots. He has co-authored over 300 peer-reviewed publications and has coordinated multiple large-scale research projects on the national and European levels.