



Monocular camera and laser based semantic mapping system with temporal-spatial data association for indoor mobile robots

Xu Song¹ · Zuo Zhijiang¹ · Xuan Liang¹ · Zhou Huaidong²

Received: 24 April 2022 / Revised: 25 July 2022 / Accepted: 5 February 2023 /
Published online: 7 March 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

In the future, the goal of service robots is to operate in human-centric indoor environments, requiring close cooperation with humans. In order to enable the robot to perform various interactive tasks, it is necessary for robots to perceive and understand environments from a human perspective. Semantic map is an augmented representation of the environment, containing both geometric information and high-level qualitative features. It can help the robot to comprehensively understand the environment and bridge the gap in human-robot interaction. In this paper, we propose a unified semantic mapping system for indoor mobile robots. This system utilizes the techniques of scene classification and object detection to construct semantic representations of indoor environments by fusing the data of a camera and a laser. In order to improve the accuracy of semantic mapping, the temporal-spatial correlation of semantics is leveraged to realize data association of semantic maps. Also, the proposed semantic mapping system is scalable and portable, which can be applied to different indoor scenarios. The proposed system was evaluated with collected datasets captured in indoor environments. Extensive experimental results indicate that the proposed semantic mapping system exhibits great performance in the robustness and accuracy of semantic mapping.

Keywords Human-robot interaction · Semantic mapping · Temporal-spatial correlation

1 Introduction

The autonomous navigation technology for indoor mobile robots has a wide range of applications in the service industry and medical field [18, 36]. Especially in major public health emergencies, such as the recent outbreak of COVID-19 around the world, the deployment of such robots can help humans perform dangerous and high-intensity tasks.

✉ Xu Song
xusong1991@jhun.edu.cn

¹ School of Smart Manufacturing, Jiangnan University, Wuhan, 430056, China

² School of Mechanical Engineering & Automation, Beihang University, Beijing, 100191, China

It is a common practice to build a globally consistent metric map of the indoor environment based on geometric information for robot navigation [27]. With respect to this type of map, the robot navigation is performed by providing the specific geometric coordinates in the global map reference frame. Yet, with the increasingly complex operating scenarios of indoor robots, traditional navigation methods based on metric maps have been unable to meet the needs of human beings [7]. In the future, the goal of robots is to operate in a human-centered home or industrial environment [12, 13]. Robots should process the ability to perceive and understand surrounding environments like humans, such as distinguishing rooms and corridors in the environment. In this way, the robot can complete various interactive tasks in the real world. This kind of skill is not provided by the geometric metric map [14, 37].

An attempt to achieve such a goal is provided by a semantic map, an augmented representation of indoor environments. The semantic map entails both geometrical information and high-level qualitative features, i.e., the robot can involve semantic attributes about the objects and the scenes encountered, in association with the geometrical perception of the environment. In this way, the semantic map can improve the ability to handle various navigation tasks for mobile robots [20, 24].

During the last decade, semantic mapping has received extensive attention for its rich informative representation of the environment [23, 35]. It is worth noting that they are still often challenged by typical adversities found in real environments. The work on semantic mapping has been fragmented, and it has rarely been possible to form a unified system that can be comprehensively evaluated in various environments. Most of them only adopt a single cue of environments to estimate scene categories of indoor environments, ignoring other semantics in local scenes and the temporal-spatial association between semantic representations [6, 28]. For this reason, in order to perform various manipulation tasks effectively for indoor robots, it is necessary to embed a higher level of environmental perception to recognize and locate semantics of interest with spatial and temporal relevance. Obviously, this is an essential ability to perform intelligent decision-making, especially for indoor service robots, which need to deal with different entities of interest. This meaningful internal perceptual representation enables the robot to perceive a variety of unsafe situations. For example, when a robot goes to the conference room to pick up a document, it needs to walk along the corridor and avoid pedestrians.

To this end, in this paper, we propose a unified semantic mapping system with temporal and spatial coherence for indoor mobile robots. The proposed system utilized the results of scene classification and object mapping obtained by state-of-the-art computer vision algorithms as observations to construct a semantic map based on the data of a laser and a camera. Our major contributions are summarized as follows:

- To the best of our knowledge, this paper is the first work concerned about both the scene semantics and object semantics for semantic mapping with temporal-spatial data association. The proposed system constructs the multi-dimensional semantics of indoor environments through scene semantic mapping and object-augmented semantic mapping.
- This paper proposes an object-augmented semantic mapping method. To ensure the consistency of semantic representations and improve the accuracy of semantic mapping, we design a temporal-spatial data association method.
- In scene semantic mapping, to effectively avoid misclassification of scenes, we present a scene classification model based on inception v3 and LSTM. The proposed scene

classification model takes the consecutive frames as input to better learn the correlation of adjacent scenes and enhance the discriminative ability of image features.

- To verify the effectiveness of the proposed method, systematic experiments are conducted on a real mobile robot.

2 Related work

2.1 Semantic mapping in robotics

In recent years, since indoor service robots are comprehensively integrated into the daily life of human beings, semantic mapping has attracted significant attention due to its rich informative representation for environments [1, 41]. In order to perform intelligent decision-making and human-robot interaction in human-centric indoor environment, researchers try to integrate the semantic concept of environments into the navigation process for indoor mobile robots. It can be achieved by semantic map [10, 15]. The semantic mapping methods could be divided into two categories from the perspective of the form of construction: augment-based mapping methods with semantic labels, and reasoning-based semantic map.

Typically, the core of augment-based mapping methods is to obtain the semantic labels of environments based on specific technique, such as scene recognition, object detection and semantic segmentation. For instance, Andreas et al. [19] defined the semantic mapping problem for robotics utilizing a purely appearance-based approaches based on semantic labeling. Their system enables mobile robots to recognize simple planar structures and specific objects such as doors and walls by building a 3D point cloud model of indoor environments. Based on the foundation of this work, extensive researches on semantic mapping with labels for indoor robots have emerged. McCormac et al. [17] proposed a dense 3D semantic mapping system, consisting of three modules: ElasticFusion system, VGG 16-layer network and Bayesian estimation. This work first adopts the ElasticFusion system to track the camera pose and construct geometric map. Then, they perform dense pixel-wise semantic segmentation based on CNNs. Finally, Bayesian estimation and conditional random fields are used to optimize pixel predictions from different viewpoints, constructing a dense 3D semantic map. Although it provides a real-time and loop-closure solution for indoor semantic interactive system, this work focuses on small-scale structured scenarios. Rozumnyi et al. [26] proposed a deep fusion method based on machine learning, which integrated semantic 3D reconstruction, scene construction and multi-sensor data into a learning-based framework. This method automatically extracts sensor parameters and scene attributes parameters from training data and representing them in the form of confidence values to achieve semantic mapping, which only requires a small amount of training data to obtain better generalization ability.

The reasoning-based semantic mapping method represents the conceptual relationship of the environment in a structured form, and usually builds a model of ontology semantics based on semantic reasoning rules. It provides the robot with the ability of complex task planning and intrinsic knowledge reasoning [21]. Ruiz-Sarmiento et al. [9] proposed a multiversal semantic map representation containing contextual relationships of environments to address the symbolic grounded uncertainty through the conditional random field technique. Crespo et al. [3] presented a semantic conceptual model construction method, which defined the hierarchies of concepts by designing an ontology with specific rules. The semantic conceptual model of environments provides the robot with the ability to

perform human-robot communication and complex navigation tasks. Generally, ontology-based semantic mapping methods describe the spatial hierarchy of concept definitions, which are the semantic basis for mutual communication between different subjects (such as humans, machines, software, etc.) in environments. However, the disadvantage of the ontology-based semantic mapping method is that an inference system needs to be built in advance to manage the robot ontology, and the robot needs different ontology construction methods in different environments.

2.2 Features for semantic classification

The semantic mapping methods based on scene recognition usually construct a semantic-metric map by adding the labels of scene to the metric map through supervised learning. It combines geometric features and high-level qualitative features of environments [31]. In [43], Zhou et al. constructed Places 365 image database for scene recognition which contains a total of 8 million training images and 365 scene categories. Based on this database, the performance of different state-of-the-art CNN models is evaluated. Pronobis et al. [22] presented a real-time multi-modal semantic space labeling system that integrated multiple visual features and laser data to perform place classification by means of Support Vector Machine (SVM). Brucker et al. [2] proposed a semantic mapping method which automatically assigns scene semantic labels from RGB-D data based on Conditional Random Field (CRF) approach. Compared to existing alternatives, this method can effectively improve the accuracy of scene classification, even in the case of incorrect geometric priors. Kostavelis et al. [11] proposed a semantic mapping framework, involving both geometrical and semasiological elements that represent the relationships between scenes and object in environments based on the technology of scene recognition and object detection. Niko et al. [30] proposed a transferable semantic mapping system that does not require environment-specific training. In this system, a monocular camera is employed to achieve scene classification based on the place205 model, which is fused with the grid map created by laser. However, in practical applications, the scene classification model of this method has great limitations and is prone to scene confusion, especially when the robot is in the transition area of adjacent scenes.

Typically, when computer vision techniques are applied to the field of mobile robotics, each frame of image is processed individually [32]. To address this problem, based on the fact that what the robot sees is a temporally coherent sequence of frames, we embed a semantic classifier in a recurrent neural network to construct a real-time semantic mapping system for indoor mobile robots in this paper. The introduction of recurrent neural networks based on image sequences enables robots to correct false misclassifications combined with prior knowledge.

2.3 Object-augmented mapping

Apart from scene-based methods, the semantic mapping methods based on object-augmented mapping aim at obtaining the object-level information of environments. It mainly consists of three steps: camera tracking, object detection and semantic mapping [5, 40]. The key component of object-augmented semantic mapping is data association, correlating the objects observed by robots with the real objects existing in the semantic map. Generally, data association between different semantic objects is nonparametric, since the visual features of objects are related to the pose of observing sensors. For this reason, Iqbal et al. [8] proposed a nonparametric statistical method for data association between successive frames based on the distribution of depth data. This method is conducive to discover the

interconnection between the occurrences of semantic instances during the process of object mapping for robots. Since the visual features of different objects at different viewpoints may appear similar, resulting in misclassification. Therefore, Tchuiev et al. [34] proposed a recursive Bayesian method for semantic mapping in ambiguous scenarios that incrementally estimated the hybrid belief of object poses over time. To detected and categorized nuclear waster objects, Li et al. [33] proposed a weakly-supervised learning approach DCNN-GPC to obtain the 3D semantic map of objects. In particular, DCNN-GPC is end-to-end, scalable and Bayes-based. The experimental results verify the effectiveness in solving a novel RGB-D object detection and recognition application with limited human annotations. Renato et al. [16] proposed a learning-based object-augmented semantic mapping system, combining the environment structure and object semantics using visual and depth cues. This approach models the detected semantic classes with a constant Kalman filter module to track and update the most probable object position. Gregorio et al. [4] proposed a real-time mapping framework for robot navigation. This work can obtain the objects semantics as well as the 6 DOF poses of object instances in environments by means of object detection and visual features matching.

3 Methodology

The core of semantic map is the extraction and association of semantic representations. This paper comprehensively considers the multi-dimensional semantic representations of scenes and objects in indoor environments, and ensures the spatial and temporal consistency of semantic information through data association of semantic representations. The overview of proposed semantic mapping system is shown in Fig. 1. The semantic mapping process consists of three stages: joint calibration of sensors, semantic mapping, and semantic data association, respectively. First, the monocular camera and the laser are jointly calibrated to perform the data fusion. Then, scene classification and object detection algorithms are exploited to extract the semantic representation of environments based on the data of monocular camera and laser. On this basis, semantic mapping is performed according to the coordinate transformation. Finally, data association is realized to construct semantic maps with semantic associations for indoor environments by means of spatio-temporal correlation of semantic representations.

3.1 Object-augmented semantic mapping

In order to perform manipulating task in a human-like navigation method, the robot also needs to understand the semantic objects in local scene. To this end, it is necessary to embed a higher level of environmental perception to detect and locate specific objects of interest. It provides the robot with the capability to interact with entities in environments.

During the process of object-augmented semantic mapping, the robot may visualize the same objects from different angles at different times, which may be labeled as different instances. If the confidence is employed to solve the semantic association problem, it may cause the semantic instance to be repeatedly mapped on the map. For this reason, this paper presents an object-augmented semantic mapping method, combining semantic instances and metric maps of indoor environments for semantic association. The proposed object-augmented semantic mapping method consists of three stages: object detection, object localization and object tracking, as shown in Fig. 2. First, we use the YOLO v2 network to detect the pre-trained objects in real time, outputting the category of objects and the pixel

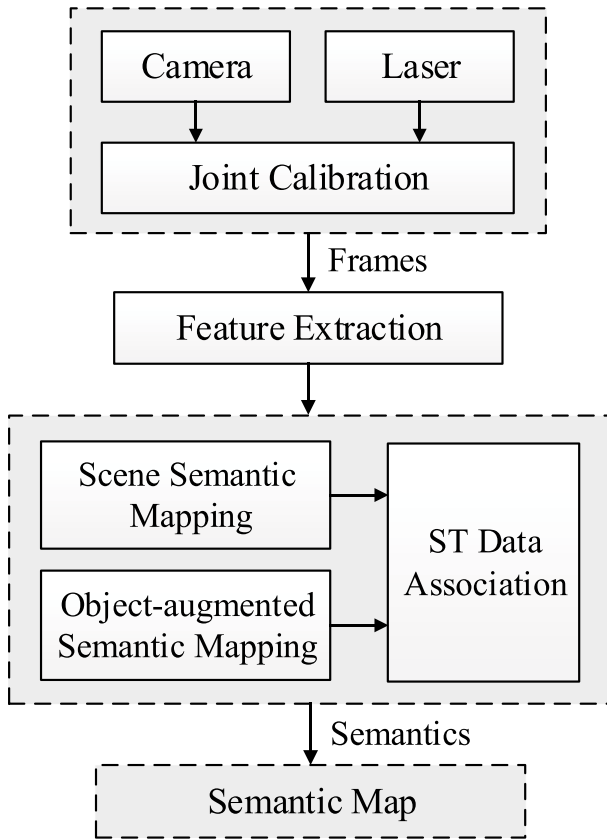


Fig. 1 Overview of the proposed semantic mapping system

coordinates of the boundary box. Then, the object is mapped from the pixel coordinate frame to the laser coordinate frame according to the relative relationship between the camera coordinate frame and the laser coordinate frame. Finally, we perform semantic association to track the mapped objects for distinguishing different instances in indoor environments.

3.1.1 Object detection

This section describes the component of object detection in object-augmented semantic mapping. We extract a preliminary semantic representation of environments by detecting objects of interest in the camera's perspective. For object detection model, we choose YOLO v2 network [25] considering the performance of recall and efficiency. The training and detection process of the YOLO v2 network is included in the same network, and it is not necessary to calculate the candidate bounding boxes of objects separately. The final output of the YOLO v2 model is the coordinates of the center point of the bounding box (b_x, b_y) , the width of the bounding box b_w , the height of the bounding box b_h , as well as the class label of the detected object l_i and the corresponding probability p_i , $p_i \in [\sigma_p, 1]$, where i represents the i -th bounding box, σ_p represents preset threshold. The detection result is output only when the label probability of object detection is greater than that σ_p .

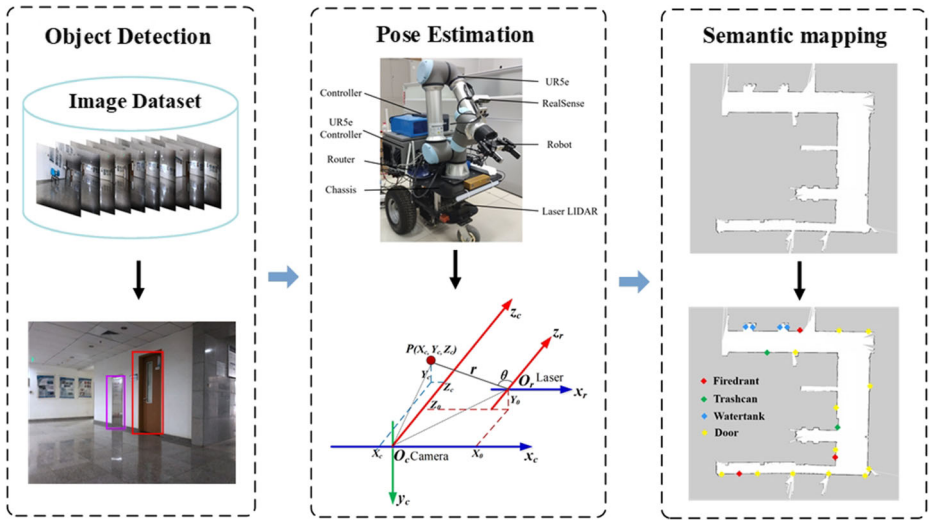


Fig. 2 Architecture of object-augmented semantic mapping

3.1.2 Object localization and mapping

(1) Joint calibration of the camera and the laser

In this paper, we use the center point of the bounding box to denote the pixel coordinates of objects in images. After obtaining the detection semantic information of objects, it needs to be mapped into the map based on the relative relationship between the pixel coordinate frame and the laser coordinate frame. Since the visual information of the camera and the laser data are not consistent, data alignment between the camera and the laser is essential by means of joint calibration, as shown in Fig. 3.

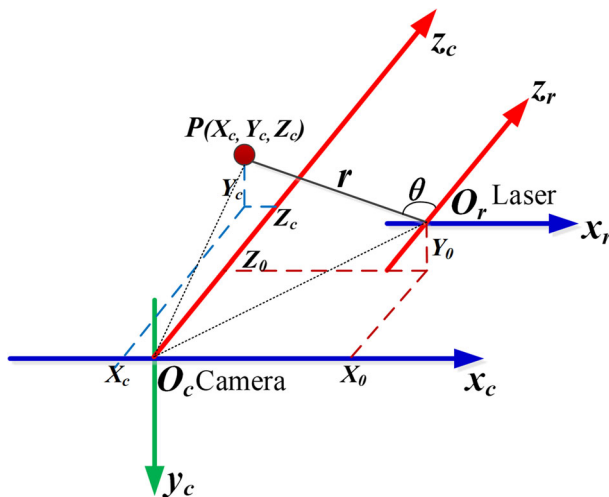


Fig. 3 Coordinate system of the camera and the laser

The coordinates of the detection object P in laser coordinate frame, camera coordinate frame and pixel coordinate frame are (r, θ) , (X_c, Y_c, Z_c) , and (u, v) , respectively. Then there are:

$$\begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = \begin{pmatrix} X_0 + r \sin \theta \\ Y_0 \\ Z_0 + r \cos \theta \end{pmatrix} \tag{1}$$

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f_x * \frac{X_0 + r \sin \theta}{Z_0 + r \cos \theta} + c_x \\ f_y * \frac{Y_0}{Z_0 + r \cos \theta} + c_y \\ 1 \end{pmatrix} \tag{2}$$

where (X_0, Y_0, Z_0) is the coordinates of the laser in the camera coordinate system, f_x, f_y, c_x, c_y are internal parameters of camera. The above formula is a typical nonlinear equation. The joint calibration process of the laser and the camera is to solve the above formula. To realize data fusion, we adopt the method in work [42] to perform joint calibration to obtain the coordinate transformation relationship between the camera and the laser. First, the parametric equation of the plane calibration board in the camera coordinate system is estimated through a checkerboard. Since the laser point falls on the plane calibration board, the laser point is mapped to the camera coordinate system according to the coordinate relationship. And then the distance error between the point and the plane is constructed. Finally, the equation is solved using the method of nonlinear least squares to obtain the optimal solution.

(2) Object-augmented semantic mapping

After observed objects are detected and mapped onto the corresponding position on the semantic map, the next step is to continuously keep track of it. That is, the robot may detect the same semantic instance from different viewpoints at different moments. At this time, it is necessary to judge whether the object is detected or a new instance. Ideally, we would like to associate each previously detected instance with the correct stored instance, and treat undetected target objects as new instances. Therefore, it is necessary to perform spatio-temporal data association of semantic information for robots. In this case, using confidence of the object class to solve the data association problem may cause the semantic instance to be repeatedly mapped on the map. For this reason, this paper presents an object-augmented semantic association method, ensuring the temporal and spatial consistency of semantic representations.

For temporal association, a same object may be detected in consecutive image frames. For this reason, we assume that the pixel coordinate P_{img} of the center of the bounding box in adjacent frames are close in pixel distance. Specifically, if the candidates in frame i and frame $i+1$ are the same object, the pixel displacement of the bounding box center in consecutive detections should theoretically be less than a threshold Th_bbc which is a pixel ratio to the total pixels of the image, e.g., $Th_bbc = 0.005$ means that the pixel displacement of the bounding box center of consecutive detected candidates is less than 0.5% to the total pixels of the image. In addition, the area similarity threshold Th_as of detected bounding box between adjacent frames is also a criterion for judging the candidate objects, e.g., $Th_as = 0.8$ means that the area similarity of the detected bounding box between consecutive frames is greater than 80%. To sum up, we consider the candidate detections in consecutive frames to be the same object if the following three criteria are met simultaneously: (1) the pixel displacement of the bounding box center in consecutive frames should theoretically be less than a threshold Th_bbc ; (2) the area similarity of detected bounding box between consecutive frames should be greater than a threshold Th_as ; (3)



Fig. 4 (a)Frame 284: three objects detected; (b)Frame 286: four objects detected, D1 is a new object

candidate instances should have been detected in at least one of the last 3 frames, but not counting the current frame. It is worth noting that consecutive detections in temporally close frames that do not belong to the same category can still be associated with the same instance, as long as the above three criteria are met simultaneously and no other instances belong to the same category. This is because consecutive detections of the same object are sometimes labeled as different categories due to mis-detections, as the robot moves towards to the object.

Figure 4 presents an example of temporal association. There three detected objects in frame 284: statue, door and fire hydrant ($A1$, $B1$ and $C1$), as shown in Fig. 4(a). Then, a new object door ($D1$) is detected in frame 286. Obviously, $A2$, $B2$ and $C2$ are associated with $A1$, $B1$ and $C1$, because all three criteria are met. However, for the object $D1$ in frame 286, no previous candidates can be associated with it. For example, compare $D1$ with $A1$ or $B1$, the criterion of Th_bbc is not met. Compare $D1$ with $C1$, $D1$ can not be associated with $C1$ although the criteria of Th_bbcc and Th_as are met. Because the $C1$ is already associated with $C2$ in current frame, so the association between $C1$ and $D1$ is forcibly terminated, as depicted in criterion (3). Therefore, the object $D1$ is constructed as a new candidate in semantic map.

For spatial association, an object will correspond to multiple instances with different poses on the map as the robot moves. To combine different object observations and fuse multiple instances with different poses on the map, the most probable pose of the object needs to be updated in real time. For this reason, we adopt Kalman Filter to construct the model of object observations and estimate the latest state of each object. In this way, we can update the most probable pose of each object based on the previous information and then generate the final spatial associated objects on the map. Since we mainly focus on static objects (such as chair and water-tank) in indoor environment during semantic mapping, the state transition equation and measurement model can be expressed as:

$$\begin{cases} x_i(k) = x_i(k - 1) + \tilde{\omega}(k) \\ y(k) = x_i(k) + \tilde{z}(k) \end{cases} \quad (3)$$

where $\tilde{\omega} \sim N(0_{3 \times 1}, W)$, $\tilde{z} \sim N(0_{3 \times 1}, Z)$, W is the process covariance matrix and Z is the measurement covariance matrix.

An example of object mapping process is shown in Fig. 5. Figure 5(a) is the current perspective image captured by the robot; Fig. 5(b) is the object detection result based on the YOLO v2 network. As can be seen, there are two categories of objects (door and fire

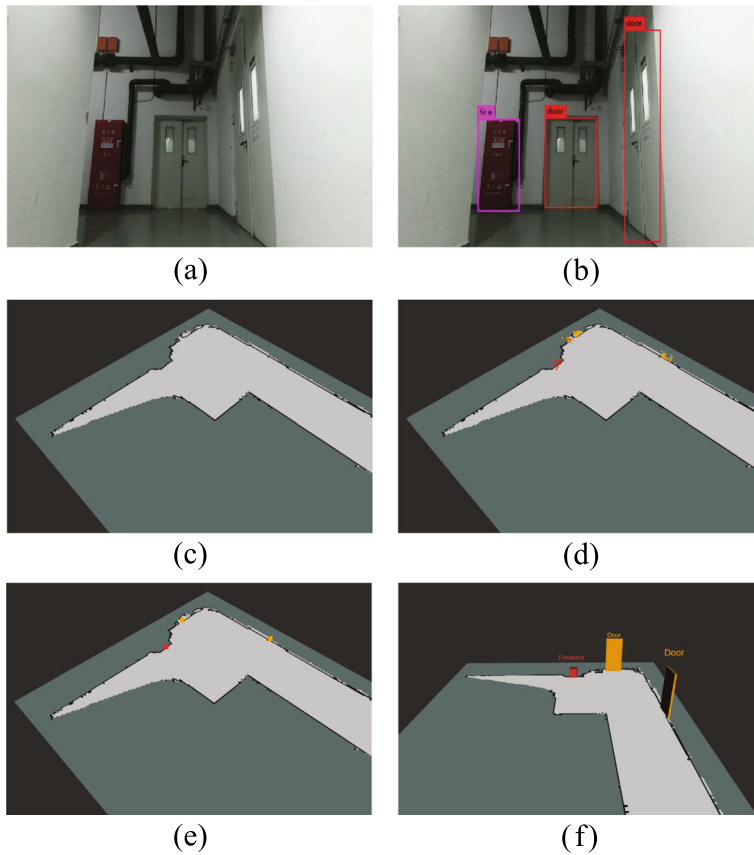


Fig. 5 The process of object localization, filtering and mapping

hydrant) in this scene; Fig. 5(c) denotes the grid map of indoor environment constructed by laser; Fig. 5(d) represents the object-augmented mapping data. In this figure, the purple dots and the red dots denote the mapping data of the fire hydrant and the door over time during the robot navigation process; Fig. 5(e) is the pose of object mapping after filtering; Fig. 5(f) is the augmented display of the corresponding pose of object mapping on the semantic map.

3.2 Scene semantic mapping

Typically, humans determine spatial locations based on scene concepts. If asked about motion trajectories in environments, people tend to express scene concepts such as “I’m in the hall, I’m going to the kitchen for a glass of water”, rather than coordinates or nodes on a map. Semantic mapping based on scene classification connects the scene semantics and geometric features of the environment, which is beneficial to intelligent decision-making and human-robot interaction when mobile robots perform various high-level tasks. The proposed scene semantic mapping system consists of two parts: indoor scene classification and semantic mapping. First, with visual images as input, a CNN model based on inception v3 and LSTM is exploited to perform indoor scene classification. Then, semantic mapping is performed on the corresponding grid cells based on the confidence of scene classification.

3.2.1 Scene classification model

In the field of semantic scene classification, it is usually assumed that the scene labels of consecutive frame images are independent of each other. However, there are multiple associations between scene category labels at different locations in indoor environments. For example, there is a hierarchical relationship between adjacent scenes, and some image frames may belong to multiple scene categories at the same time. Especially when the robot is in the transition area between adjacent scenes, if a single-frame image is employed to identify the scene semantics, it is easy to cause scene confusion, resulting in misclassification.

To address the above problem, in this paper, based on the fact that what the robot sees is a temporally coherent sequence of image data, we present a scene classification model based on LSTM, which embeds the semantic classifier into a recurrent neural network. The scene classification model takes the image sequence of continuous frames as input to better learn the correlation of different scenes and enhance the discriminative ability of image features. It can effectively avoid misclassification caused by insufficient feature information of a single frame image.

The scene classification model in this paper is shown in Fig. 6, which consists of two parts: feature extraction from image sequence, analysis of the feature sequence and prediction of scene categories. For the input image sequence $\{x_1, x_2, \dots, x_N\}$, we first use the inception v3 network to extract the convolutional features of each frame. The feature map from the last convolutional layer in inception v3 network is regard as the output spatial feature $\{y_1, y_2, \dots, y_N\}$. Assuming that the feature extraction process of the inception v3 is defined as follow:

$$y_i = f(x_i), i = 1, 2, \dots, N \tag{4}$$

where y_i is the feature vector of the frame i in the image sequence, its size is 2048 dimensions, $y_i \in R^{2048}$.

Then, the convolution features extracted by the inception v3 model are input into LSTM to obtain the image features with temporal correlation. LSTM is a recurrent neural network

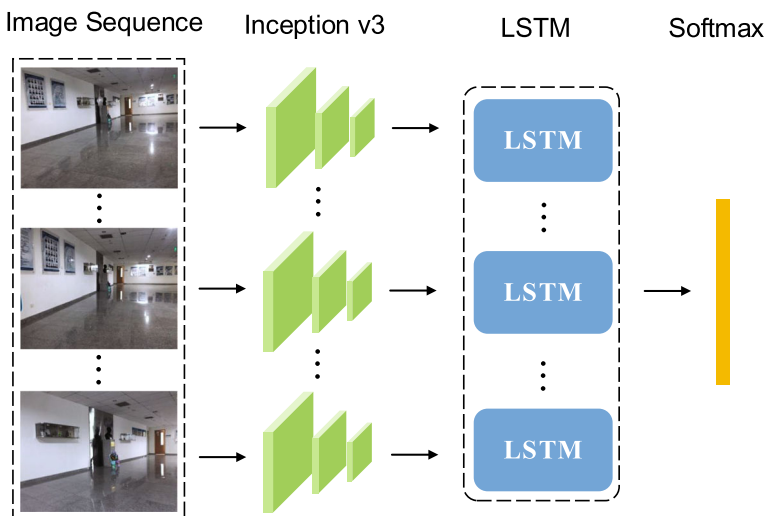


Fig. 6 The scene classification model based on Inception v3 and LSTM

(RNN) with memory function for analyzing temporal sequence data, which has been successfully applied to various field, such as automatic speech recognition and human action recognition [29, 38, 39]. When dealing with the problem of scene classification, we should analyze the differences and consistency in the intrinsic characteristics of indoor scenes. Compared with traditional RNN, LSTM model can avoid the effects of gradient explosion and gradient disappearance. Since LSTM model contains processing units of “forget gate”, “input gate” and “output gate”, it can selectively store historical data according to data information and selectively transmit current input and output information. In this way, the integration of the inception v3 model and the LSTM model provides the robot with the ability to better learn the contextual information of image sequences, enabling the robot to utilize prior knowledge to correct misclassifications.

After that, the softmax layer mapping the feature vector extracted from LSTM to $(0, 1)$, and outputs the probability of scene category:

$$\zeta = - \sum_{i=1}^m \log(p_i) \quad (5)$$

$$p_i = \frac{\exp(W_{y_i}^T z_i + b_{y_i})}{\sum_{j=1}^n \exp(W_j^T z_i + b_j)} \quad (6)$$

where p_i represents the class probability output from softmax classifier, ζ is the loss value of softmax function, b is the bias vector.

3.2.2 Incremental semantic mapping based on probability distribution

The set of scene category labels is defined as $\hat{c} = (c_0, c_1, c_2 \dots, c_n)$. For the image at time t , the discrete probability distribution of the known categories estimated by the scene classification model is defined as $p(c_i|I_t)$, then the corresponding likelihood function is

$$L(I_t|\hat{c}_t) = (p(c_0|I_t), p(c_1|I_t), \dots, p(c_n|I_t)) \quad (7)$$

where $\hat{c} = (c_0, c_1, c_2 \dots, c_n)$ represents the vector set of known scene category labels, $p(c_i|I_t)$ represents the probability that the image I_t at time t belongs to scene category c_i .

The semantic mapping process based on scene classification can be cast as a probability estimation problem based on Bayesian estimation due to the temporal continuity of image frames. Assuming that the first-order Markov property is satisfied, the following Bayesian estimation formula can be obtained:

$$p(\hat{c}_t|I_t) = L(I_t|\hat{c}_t) \cdot p(\hat{c}_{t-1}|I_{t-1}) \quad (8)$$

For the image data at time t , we employ the proposed scene classification model to output the probability distribution of indoor scene where the robot is located. At the same time, the Gmapping algorithm is used to construct a grid map of the robot’s current position. Then, the semantic information of indoor scene is integrated with the metric information of grid map to realize the scene semantic mapping.

The constructed scene semantic grid cell is shown in Fig. 7. We take the scene category with the highest probability as the semantic label of the current grid cell. The color of the grid cell represents the corresponding scene category. Compare with traditional grid maps, the grid cell on the semantic map not only represents the probability of occupancy, but also the probability that the grid cell belongs to a certain semantic scene category.

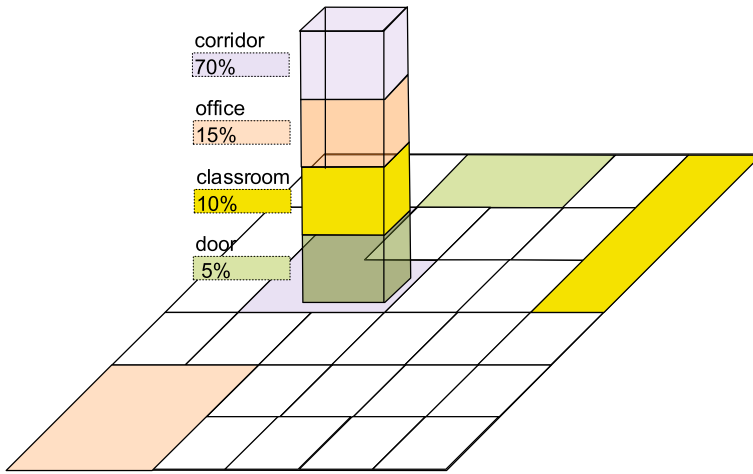


Fig. 7 The constructed scene semantic grid cells

In addition, the perspective of the camera and the laser is not consistent during the experiment, as shown in Fig. 8. The horizontal viewing angle of the monocular camera used in this paper is about 70° , corresponding to the blue area in Fig. 5, while the scanning angle of the laser used in this paper is about 270° , corresponding to the gray area in Fig. 8. In the process of semantic mapping, only the laser data within the effective viewing angle of the camera is taken to build a semantic map. Therefore, we take all the areas located within $0.5\text{m}\sim 4\text{m}$ in front of the mobile robot and coincide with the horizontal viewing angle of the camera as the target area for scene mapping at the current moment, corresponding to the olive area in Fig. 8. After obtaining the real-time scene classification results of the image

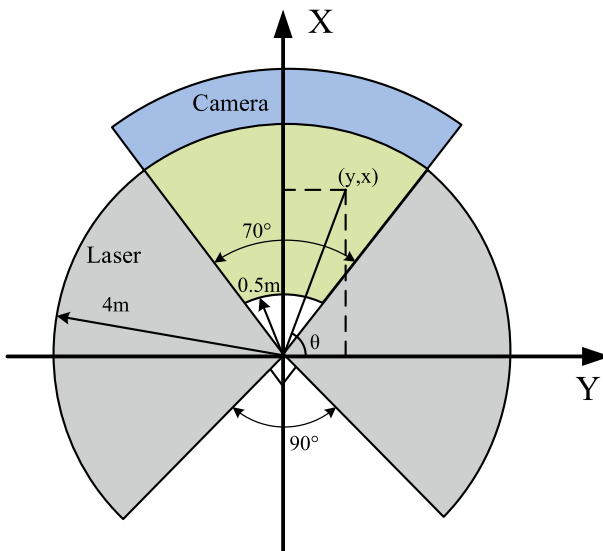


Fig. 8 The perspective of the camera and the laser

frames, for each grid cell covered by the current laser view within the camera view, we update the semantic scene label using Bayesian estimation.

For time t , the observation of the camera is I_t . Then, the probability of scene category of the current grid cell is:

$$p_k(\hat{c}_i|I_{1:t}) = \left[1 + \frac{1 - p_k(\hat{c}_i|I_t)}{p(\hat{c}_i|I_t)} \frac{1 - p_k(\hat{c}_i|I_{1:t-1})}{p(\hat{c}_i|I_{1:t-1})} \frac{p(\hat{c}_i)}{1 - p(\hat{c}_i)} \right]^{-1} \quad (9)$$

where $p_k(\hat{c}_i|I_t)$ and $p_k(\hat{c}_i|I_{t-1})$ represent the probability that the grid cell belongs to the scene category \hat{c}_i for the observation at time t and $t-1$, respectively, $p(\hat{c}_i)$ represents a prior probability.

4 Experiments

To evaluate the performance of our semantic mapping system, the experiment has been performed in various indoor scenes of Beihang University with a mobile robot, as shown in Fig. 9. The maximum linear velocity and maximum angular velocity of the mobile robot are set to 0.5 m/s and 0.5 rad/s, respectively. The mobile robot is equipped with a camera and a 2D laser. The camera utilized to perceive visual features of indoor environments is 1024×768 pixels in resolution. The laser developed by HOKUYO covers 30 m and 270°.

4.1 Experimental setup

During the experiments of semantic mapping, we steered the mobile robot through the indoor environment online and collected the environmental information from different sensors. Then, the semantic map of environments was constructed by means of the proposed semantic mapping system offline based on the previously acquired data.

In the experiment of object-augmented semantic mapping, Th_bbc and Th_as are set to 0.008 and 0.8, respectively. Actually, the parameters in different scenarios are different. For example, a large-scale indoor environment requires a low value of Th_bbc . Since in large-scale indoor environment, the objects are generally farther from the robot, the ratio of pixel displacement of the bounding box center in consecutive detections to the total pixels of the image is relatively low.

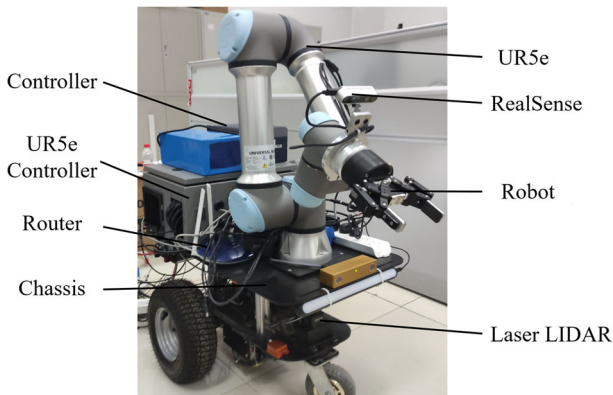


Fig. 9 The mobile robot platform for experiments

4.2 Image dataset for experiments

Since existing available indoor image datasets did not contain annotation images with required object and scene category, we choose to collect images and label the datasets in various indoor scenes.

For dataset of object detection, since our semantic mapping system mainly in indoor scenes, the following criteria need to be followed when constructing semantic objects: (1) the selected objects should be common in indoor environments; (2) the selected objects should have typical visual features to facilitate identification and detection; (3) the selected objects should facilitate the robot to perform semantic interaction tasks. Therefore, in this paper, we select the following four types of objects for object-augmented semantic mapping: door, fire hydrant, watertank and elevator, which generally do not change position over time and are suitable for semantic mapping, as shown in Fig. 10. It is worth noting that a single image may contain multiple objects. We collected a total of 3675 images of the above four categories of objects from indoor environments. Among them, the training dataset, validation dataset and test dataset contain 2535, 386 and 754 images, respectively. With respect to the training details of object detection model, we used a learning rate of 0.001 for the first 1000 iterations and 0.0001 for the next iterations. The weight decay and momentum are set to 0.0005 and 0.9 respectively.

For dataset of scene classification, in order to verify the performance of the proposed scene classification model based on inception v3 and LSTM, we select the following seven scenes for scene semantic mapping experiment according to the characteristics of indoor environments: conference room, elevator, corridor, warehouse, laboratory, staircase and drinking room, as shown in Fig. 11. The image dataset is captured from image sequences in these seven categories of scenes. In order to improve the accuracy of scene classification, it is necessary to follow the following criteria when collecting image sequences of indoor scenes: (1) collect images from environment with different perspectives, different scales and different lighting to maintain the diversity of images in the dataset; (2) the scene labels



Fig. 10 The dataset for indoor object detection



Fig. 11 The dataset for indoor scene classification

of image sequences in the dataset should be sorted by time step to meet the requirements of the LSTM network. We collected a total of 4900 images from the above seven categories of scenes in indoor environments, and then resized the size of the images in the dataset to 299×299 . Among them, the training dataset, validation dataset and test dataset contain 3506, 480 and 914 images, respectively. With respect to the training details of scene classification model, we use an initial learning rate of 0.00001 and a batch size of 128. The Adam is adopted for optimization to adjust the learning rate adaptively.

4.3 Evaluation of the proposed scene classification model

Figure 12 depicts the normalized confusion matrix of the scene classification model proposed in this paper, where the vertical axis represents the actual scene label and the horizontal axis represents the predicted scene label. The confusion matrix denotes the probability that the scene on the vertical axis is classified as the scene on the horizontal axis. It can be seen that the average accuracy of the proposed scene classification model is about 85%. Additionally, “warehouse” had the lowest average classification accuracy at 72%. This is mainly due to the miscellaneous items stored in warehouse, which cause the visual features of the “warehouse” scene to be easily confused with other scenes.

Figure 13 shows CAM (Class Activation Map) and classification probability of indoor scenes. The CAM represents the class activation map of the scene classification model and is a visualization of CNN features. In the CAM, the darker the pixel area (the red area in

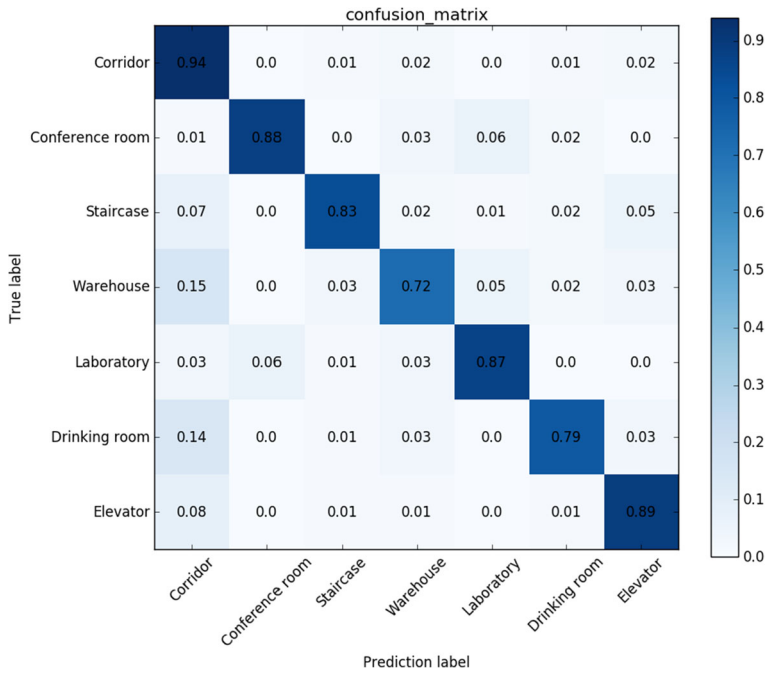


Fig. 12 The normalized confusion matrix of the proposed scene classification model

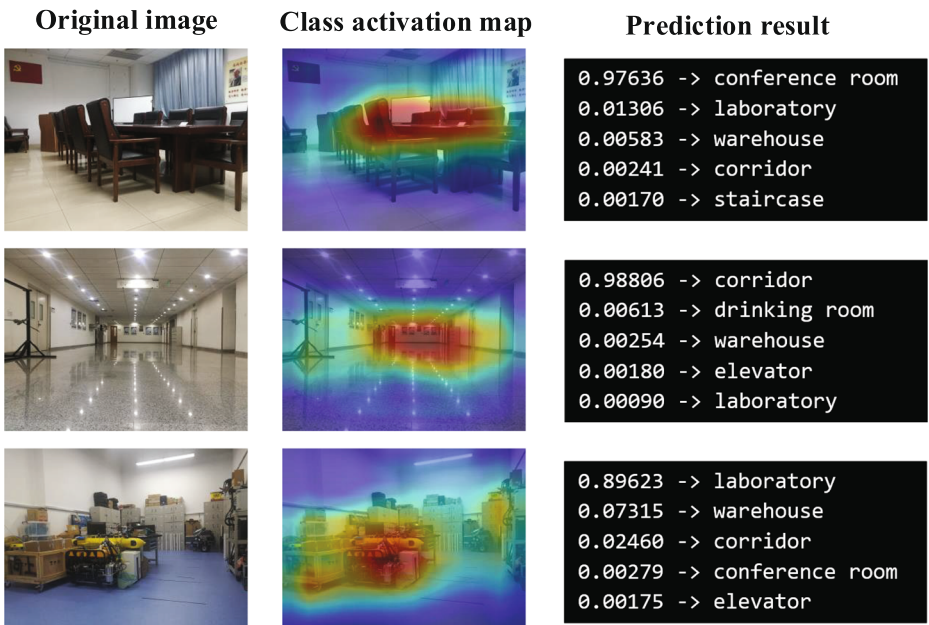


Fig. 13 The dataset for indoor scene classification

Table 1 Precision and recall of different scene classification models

Scene classification model	Precision(%)	Recall(%)
AlexNet	67.9	70.4
VGGNet	69.8	71.2
Inception v3	72.3	74.1
Proposed model	84.9	83.5

Fig. 13), the greater the contribution to the scene classification result. It can be seen that the classification probabilities of the three scenes of “conference room”, “corridor” and “lab” have reached 97.6%, 98.8% and 89.6%, respectively.

Additionally, in order to comprehensively evaluate the performance of the proposed scene classification model, we conduct comparative experiments of different methods on the established dataset, in which Precision and Recall are used as evaluation indicators, as shown in Table 1. For a specific category, the Precision represents the ratio of the number of samples correctly predicted as this category to the total number of samples predicted as this category, and the Recall represents the ratio of the number of samples correctly predicted as this category to the total number of samples of the category. Here, the Precision represents the average of each category’s precision, and the Recall represents the average recall for each category.

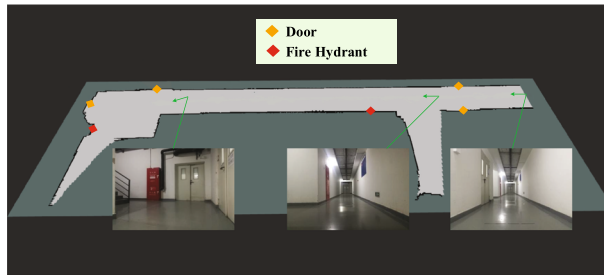
As depict in Table 1, it can be seen that the performance of the inception v3 model is better than AlexNet and VGGNet. In addition, compare with AlexNet, VGGNet, and inception v3, the proposed model performs obviously better in both precision and recall, reaching 84.9% and 83.5%, respectively. This is because the proposed scene classification model based on inception v3 and LSTM utilizes the temporal correlation between image sequences to improve accuracy and robustness of scene classification.

4.4 Results of semantic mapping

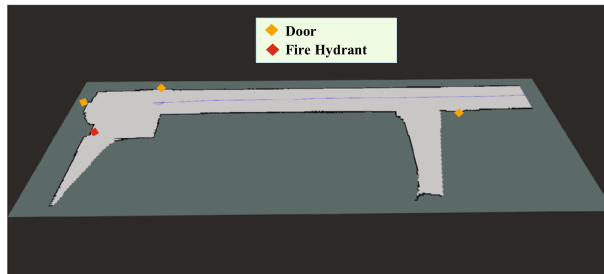
4.4.1 Object-augmented semantic mapping

In this section, comparative experiments of object-augmented semantic mapping are conducted in two different scenes. Figures 14 and 15 present the 2D ground-truth map, object-augmented semantic map of the proposed method and object-augmented semantic map of Martins et al. [16] in two different scenes, respectively. Martins et al. [16] constructs the augmented semantic map with object semantics based on RGB-D camera. In [16], the technique of 3D semantic segmentation is used to model object semantics.

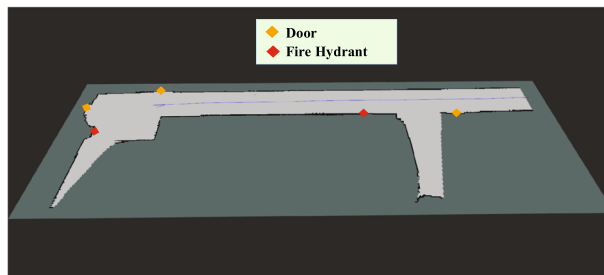
In Figs. 14 and 15, the 2D ground-truth map depicts the ground-truth location of semantic object on the semantic map. On the object-augmented semantic map, the yellow semantic instance denotes “door”, and the red instance denotes “fire hydrant”, in which all semantic instances are represented in global coordinates. The solid blue lines in each map are the odometer trajectories recorded by the mobile robot. Additionally, some perspective images of different locations in the indoor environment are shown on the semantic map. As can be seen, in the semantic map of Martins et al. [16], some objects are even projected outside the map, especially when the robot rotates rapidly. In contrast, the proposed system can correctly realize the object-augmented semantic mapping.



(a)



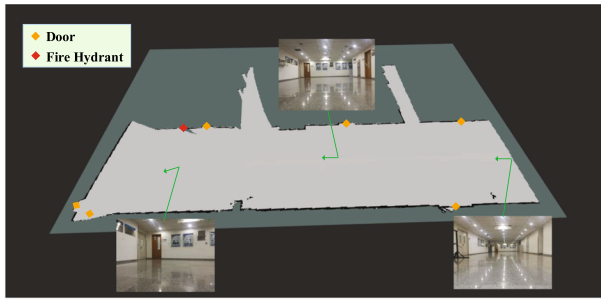
(b)



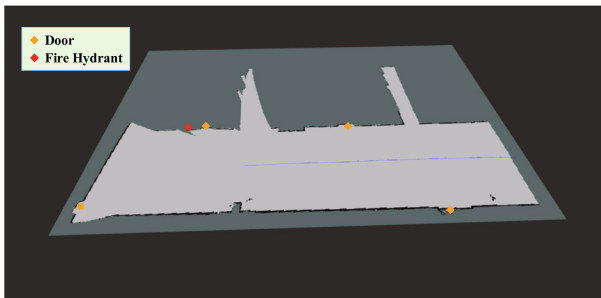
(c)

Fig. 14 (a) 2D ground-truth map (covered mapped area of $46\text{m} \times 22\text{m}$) with the mapped object location of scene 1; (b) Object-augmented semantic map of Martins et al. [16]; (c) Object-augmented semantic map of the proposed method

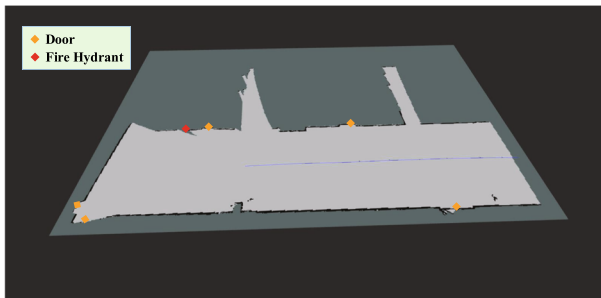
To quantitatively evaluate the performance of the object-augmented semantic mapping method, the quantitative results of comparative experiments among Martins et al. [16] and the proposed method are shown in Table 2. We choose the metrics of FP, FN and average error to evaluate the performance of different methods. FP is false positive that represents the number of objects that are wrongly instantiated, and FN is false negative that represents the number of objects that are not mapping in the semantic map. The average error denotes the mean of the mapping errors of all objects in the semantic map. The semantic objects are continuously detected and mapped during the movement of the mobile robot. An object will be mapped to multiple unfiltered instances on the semantic map over time. In this case, errors between the position of the instance on the semantic map and its ground-truth position may occur when filtering and updating multiple unfiltered instances, which can be seen



(a)



(b)



(c)

Fig. 15 (a) 2D ground-truth map (covered mapped area of $33\text{m} \times 16\text{m}$) with the mapped object location of scene 2; (b) Object-augmented semantic map of Martins et al. [16]; (c) Object-augmented semantic map of the proposed method

from the comparison of the 2D ground-truth map and object-augmented semantic map. As can be seen, the performance of the proposed object-augmented mapping method is significantly better than Martins et al. [16] in terms of average error. The average errors of Martins et al. [16] are 0.37m and 0.49m in the two scenes. In contrast, the average errors of the proposed method are 0.16m and 0.22m. In my opinion, there are two reason for this result: (1) the semantic mapping method in [16] performs 3D semantic segmentation to model object semantics based on 3D point clouds, causing ROS internal inter-process communication delay. In this case, with the movement of the robot, the pose of the objects cannot

Table 2 Object detection and mapping results of different scenes

		Detection	FP	FN	Average error(m)
Scene1	Martins et al. [16]	4	0	2	0.37
	Proposed method	5	0	1	0.16
Scene2	Martins et al. [16]	6	0	1	0.49
	Proposed method	6	0	1	0.22

be updated in time, which will magnify the error of object mapping; (2) the semantic mapping method in [16] can not perform data association for object semantics accurately and timely. In contrast, the proposed method contains the module of temporal-spatial data association, ensuring the consistency of semantic representations and improving the accuracy of semantic mapping.

4.4.2 Scene semantic mapping

According to the characteristic of the indoor experiment environment, the following seven categories of scenes are selected for scene semantic mapping environments: conference room, elevator, corridor, warehouse, laboratory, staircase and drinking room. Since all possible indoor scene semantic labels in the experiment are known, the set of scene category labels can be defined as:

$$\hat{\omega} = \{corridor, conference\ room, staircase, warehouse, laboratory, drinking\ room, elevator\} \quad (10)$$

During the experiment, the current frames of the mobile robot are captured in real time through the camera, and the image sequences of consecutive frames are input to the scene category model to output the discrete probability $p(\omega_i|X)$ of scene category from the prob layer. Then we update the grid cell covered by the current laser view within the camera view based on the scene labels with the highest probability estimated by equation. On the constructed semantic map, different colors represent different scenes.

During this process, the grid cell in the edge area of semantic map may not be mapped or the semantic representation is inaccurate, due to the limitations of the perspective of the camera and the laser, or the semantic map is not update in time, etc. In this case, it is necessary to perform spatial data association of scene semantics for robots. For the grid cell with low-confidence of scenes or unmapped on semantic map, we adopt the scene semantics of neighboring region with high confidence to label it. A typical example is the laboratory scene shown in Fig. 16. Some unoccupied grid cells in the edge area of semantic map were not mapped efficiently, because the perspective of robot did not scan every corner of the laboratory scene. The semantic maps of the laboratory scene before and after data association are shown in Fig. 17(a) and (b), respectively. As can be seen, the processed semantic map more accurately represents the scene semantics.

Additionally, we conducted experiments of scene semantic mapping in different indoor local scenes. The results are shown in Fig. 18. In the figure, the label colors represent different scenes, and the label of each grid cell is determined by the scene category with the highest confidence. As can be seen, the performance of the proposed scene mapping system performs well in terms of accuracy. It is worth noting that a local area may contain mixed categories of scenes, as shown in the upper left in Fig. 18, which contains both corridor



Fig. 16 The laboratory scene

(green) and drinking room (red). Also, the label of the drinking room extends into the corridor area. This is because when the robot moves to the corridor scene near the drinking room, the perspective of the camera also covers the area of drinking room, causing the robot to recognize the location as the drinking room. In addition, not all semantic maps are assigned the correct labels. There is misclassification of scenes in a specific scene, as shown in the

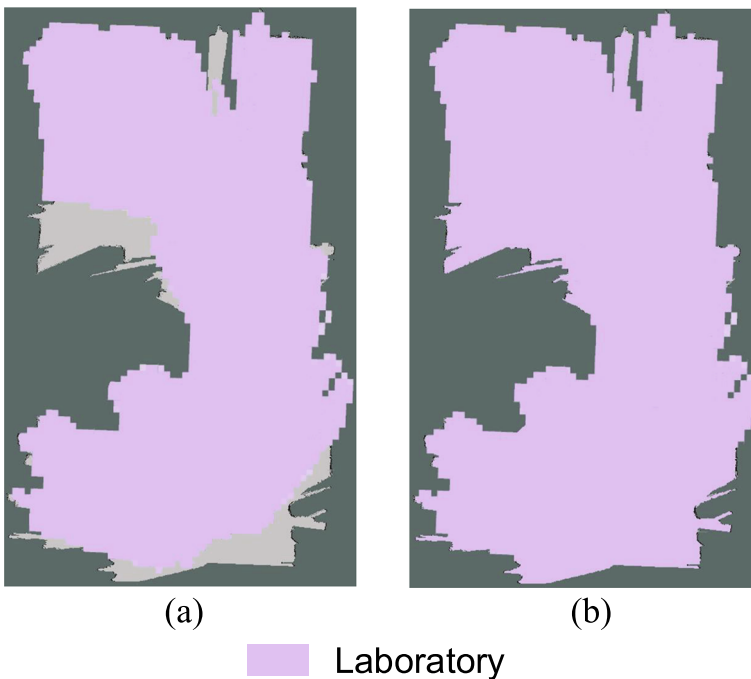


Fig. 17 The semantic map before and after spatial data association

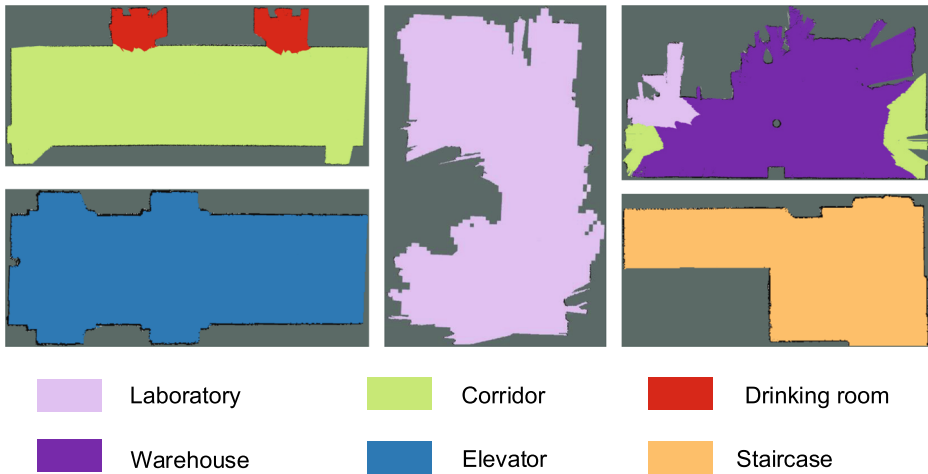


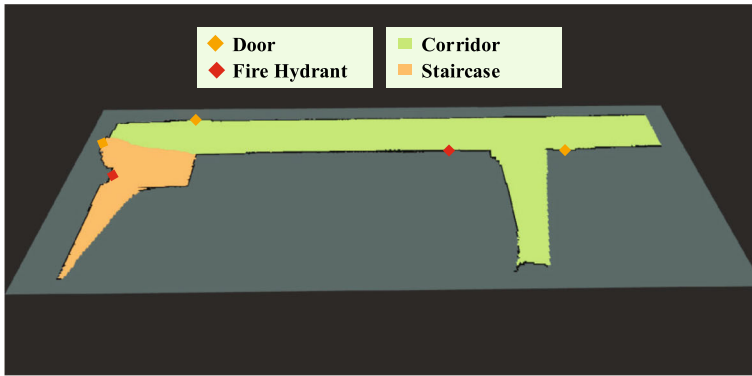
Fig. 18 Scene semantic mapping in different scenes

upper right in Fig. 18. Partial region of warehouse which is a particular challenging map is identified as corridors and laboratory. This is due to the miscellaneous items in the warehouse, resulting in the misclassification of indoor scenes. In contrast, the semantic mapping of scenes such as elevator, staircase and laboratory is more accurate.

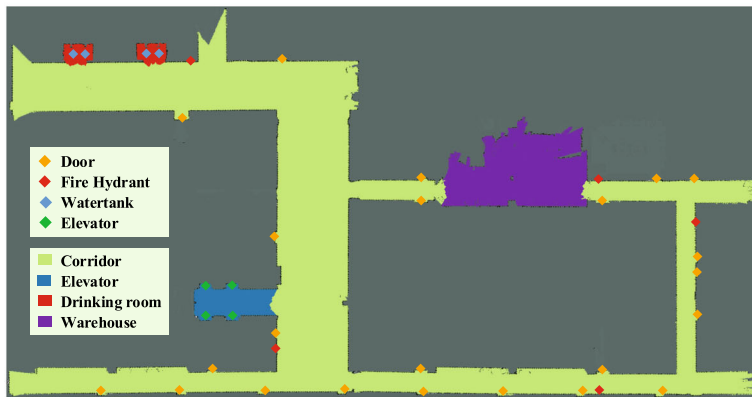
Finally, semantic mapping experiments that fuse scene semantics and object semantics are conducted in different indoor environments. In the experiment, we utilize object-augmented and scene category techniques to construct the object semantics and scene semantics with spatial and temporal correlation based on the data of camera and laser. The experiment results are shown in Fig. 19. The yellow, red, blue and green rhombic instances represent the object of door, fire hydrant, watertank and elevator, respectively. The green, blue, red and purple grid cell represent the scene of corridor, elevator, drinking room, warehouse, respectively. Our proposed semantic mapping method effectively and accurately represents semantics in indoor environments. These semantics are beneficial for robots to perform intelligent decision-making and human-robot interaction in indoor human-robot coexisting environments.

5 Conclusion

In this paper, we propose a unified semantic mapping system for indoor mobile robots. This semantic mapping system constructs the multi-dimensional semantics of scenes and objects in indoor environments through scene semantic mapping and object-augmented semantic mapping. In scene semantic mapping, a scene classification model based on LSTM is utilized to learn the temporal correlation of consecutive frames to enhance the discriminative ability of image features. In object-augmented semantic mapping, in order to associate different instances in semantic map, we present an object-augmented semantic association method to ensure the temporal and spatial consistency of semantic representations.



(a)



(b)

Fig. 19 The constructed semantic map of different indoor scenes

Systematic experiments were conducted on a real mobile robot and the results indicate that the proposed system exhibits great performance in the robustness and accuracy of semantic mapping.

In future work, we plan to construct the coupling model between the semantics of different categories to explore the intrinsic reasoning relationship of different semantics in indoor environments. In this way, it can provide a multi-scale and richer semantic map for robots.

Funding This work was supported by the School-level Scientific Research Project Funding Program of Jiangnan University (Grant No. 2022XKZX33) and the National Key R&D Program of China (Grant No. 2019YFB1310802).

Data Availability The datasets generated during and/or analysed during the current study are not publicly available due to it is relevant for our future research but are available from the corresponding author on reasonable request.

Declarations

Conflict of Interests The authors declare no conflict of interest.

References

1. Authors U (2012) Automatic dense visual semantic mapping from street-level imagery. In: IEEE International conference on intelligent robots & systems
2. Brucker M, Durner M et al (2018) Semantic labeling of indoor environments from 3d rgb maps. In: IEEE International Conference on Robotics and Automation (ICRA), pp 1871–1878
3. Crespo J, Barber R et al (2018) Reasoning systems for semantic navigation in mobile robots. In: IEEE International conference on intelligent robots and Systems (IROS), pp 5654–5659
4. De Gregorio D, Cavallari T, Di Stefano L (2017) Skimap++: real-time mapping and object recognition for robotics. In: Proceedings of the IEEE international conference on computer vision workshops, pp 660–668
5. de Oliveira FelipeDB, da Silva MR, Araújo AluizioFR (2021) Spatio-temporal data association for object-augmented mapping. *J Intell Robot Syst* 103(1):1–19
6. Garg S, Jacobson A, Kumar S, Milford M (2017) Improving condition-and environment-invariant place recognition with semantic place categorization. In: IEEE International conference on intelligent robots and systems (IROS), pp 6863–6870
7. Huang G (2019) Visual-inertial navigation: a concise review. In: IEEE International conference on robotics and automation (ICRA), pp 9572–9582
8. Iqbal A, Gans NR (2020) Data association and localization of classified objects in visual slam. *J Intell Rob Syst* 100(1):113–130
9. Jose-Raul, Ruiz-Sarmiento, Cipriano, Galindo, Javier, Gonzalez-Jimenez (2017) Building multiversal semantic maps for mobile robot operation - sciencedirect. *Knowl-Based Syst* 119:257–272
10. Kostavelis I, Charalampous K et al (2016) Robot navigation via spatial and temporal coherent semantic maps. *Eng Appl Artif Intell* 48:173–187
11. Kostavelis I, Gasteratos A (2017) Semantic maps from multiple visual cues. *Expert Syst Appl* 68:45–57
12. Liu P, Yu H, Cang S (2019) Adaptive neural network tracking control for underactuated systems with matched and mismatched disturbances. *Nonlinear Dyn*
13. Liu S, Liu P (2021) Benchmarking and optimization of robot motion planning with motion planning pipeline. *The International Journal of Advanced Manufacturing Technology*
14. Lowry S, Sünderhauf N et al (2015) Visual place recognition: a survey. *IEEE Trans Rob* 32(1):1–19
15. Luo RC, Chen CJ (2017) Recursive neural network based semantic navigation of an autonomous mobile robot through understanding human verbal instructions. In: IEEE International conference on intelligent robots & systems, pp 1519–1524
16. Martins R, Bersan D et al (2020) Extending maps with semantic and contextual object information for robot navigation: a learning-based framework using visual and depth cues. *J Intell Robot Syst* 99(3):555–569
17. McCormac J, Handa A et al (2017) Semanticfusion: dense 3d semantic mapping with convolutional neural networks. In: IEEE International conference on robotics and automation (ICRA), pp 4628–4635
18. Mertens A, Reiser U, Brenken B, Lüdtke M, Hägele M, Verl A, Brandl C, Schlick C (2011) Assistive robots in eldercare and daily living: automation of individual services for senior citizens. In: International conference on intelligent robotics and applications, pp 542–552
19. Nüchter A, Hertzberg J (2008) Towards semantic maps for mobile robots. *Robot Auton Syst* 56(11):915–926
20. Niu L, Cai J, Veerarahavan A, Zhang L (2018) Zero-shot learning via category-specific visual-semantic mapping and label refinement. *IEEE Trans Image Process* 28(2):965–979
21. Pronobis A, Jensfelt P (2012) Large-scale semantic mapping and reasoning with heterogeneous modalities. *IEEE International conference on robotics and automation*, 3515–3522
22. Pronobis A, Martinez Mozos O et al (2010) Multi-modal semantic place classification. *Int J Robot Res* 29(2–3):298–320
23. Qi X, Wang W, Yuan M, Wang Y, Li M, Xue L, Sun Y (2020) Building semantic grid maps for domestic robot navigation. *Int J Adv Rob Syst* 17(1):1–12
24. Rangel JC, Cazorla M, Garcia-Varea I, Romero-Gonzalez C, Martínez-Gomez J (2019) Automatic semantic maps generation from lexical annotations. *Auton Robot* 43(3):697–712
25. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: IEEE Conference on computer vision and pattern recognition, pp 7263–7271
26. Rozumnyi D, Cherabier I et al (2019) Learned semantic multi-sensor depth map fusion. *arXiv e-prints*
27. Sharma K (2018) Improved visual slam: a novel approach to mapping and localization using visual landmarks in consecutive frames. *Multimed Tools Applic* 77(7):7955–7976

28. Shi L, Kodagoda S, Dissanayake G (2010) Laser range data based semantic labeling of places. In: IEEE International conference on intelligent robots & systems
29. Singh D, Merdivan E, Hanke S, Kropf J, Geist M, Holzinger A (2017) Convolutional and recurrent neural networks for activity recognition in smart environment. In: Towards integrative machine learning and knowledge extraction, pp 194–205
30. Sinderhauf N, Dayoub F et al (2016) Place categorization and semantic mapping on a mobile robot. In: IEEE International Conference on Robotics and Automation (ICRA)
31. Song X, Liang X, Zhijiang Z, Huaidong Z (2022) A semantic mapping system based on scene classification for indoor mobile robots. In: IEEE International conference on information, communication and automation technologies (ICAT), pp 1–6
32. Sorkhi AG, Hassanpour H, Fateh M (2020) A comprehensive system for image scene classification. *Multimed Tools Applic* 79(25):18033–18058
33. Sun L, Zhao C, Yan Z, Liu P, Duckett T, Stolkin R (2018) A novel weakly-supervised approach for rgb-d-based nuclear waste object detection and categorization. *IEEE Sens J* 19(9):3487–3500
34. Tchuiev V, Feldman Y, Indelman V (2019) Data association aware semantic mapping and localization via a viewpoint-dependent classifier model. In: IEEE International conference on intelligent robots and systems (IROS), pp 7742–7749
35. Tchuiev V, Indelman V (2020) Distributed consistent multi-robot semantic localization and mapping. *IEEE Robot Autom Lett* 5(3):4649–4656
36. Thrun S (2002) Probabilistic robotics. *Commun ACM* 45(3):52–57
37. Truong X-T, Ngo TD (2017) Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model. *IEEE Trans Autom Sci Eng* 14(4):1743–1760
38. Ullah A, Ahmad J, Muhammad K et al (2017) Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE access* 6:1155–1166
39. Varior RR, Shuai B et al (2016) A siamese long short-term memory architecture for human re-identification. In: European Conference on computer vision, pp 135–153
40. Vasudevan S, Gächter S et al (2007) Cognitive maps for mobile robots—an object based approach. *Robot Auton Syst* 55(5):359–371
41. Vineet V, Miksik O et al (2015) Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In: IEEE International conference on robotics and automation (ICRA)
42. Zhang Q, Pless R (2004) Extrinsic calibration of a camera and laser range finder (improves camera calibration). In: IEEE International conference on intelligent robots and systems (IROS), vol 3, pp 2301–2306
43. Zhou B, Lapedriza A et al (2018) Places: a 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1–1

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.