information

Review

Semantic Mapping for Mobile Robots in Indoor Scenes: A Survey

Xiaoning Han [1,2,3], Shuailong Li [1,2,3], Xiaohui Wang [1,2,3] and Weijia Zhou [1,*]

[1] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; hanxiaoning@sia.cn (X.H.); lishuailong@sia.cn (S.L.); wangxiaohui1@sia.cn (X.W.)
[2] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[*] Correspondence: zwj@sia.cn

Abstract: Sensing and mapping its surroundings is an essential requirement for a mobile robot. Geometric maps endow robots with the capacity of basic tasks, e.g., navigation. To co-exist with human beings in indoor scenes, the need to attach semantic information to a geometric map, which is called a semantic map, has been realized in the last two decades. A semantic map can help robots to behave in human rules, plan and perform advanced tasks, and communicate with humans on the conceptual level. This survey reviews methods about semantic mapping in indoor scenes. To begin with, we answered the question, what is a semantic map for mobile robots, by its definitions. After that, we reviewed works about each of the three modules of semantic mapping, i.e., spatial mapping, acquisition of semantic information, and map representation, respectively. Finally, though great progress has been made, there is a long way to implement semantic maps in advanced tasks for robots, thus challenges and potential future directions are discussed before a conclusion at last.

Keywords: semantic mapping; mobile robots; indoor scenes; acquisition of semantic information; map representation

Citation: Han, X.; Li, S.; Wang, X.; Zhou, W. Semantic Mapping for Mobile Robots in Indoor Scenes: A Survey. *Information* **2021**, *12*, 92. https://doi.org/10.3390/info12020092

Academic Editor: Arkaitz Zubiaga

Received: 4 January 2021
Accepted: 18 February 2021
Published: 22 February 2021

1. Introduction

Sensing and modeling its surroundings is an essential requirement for a mobile robot. When moving through an indoor environment, a robot needs to plan a safe path to the destination, without collisions with obstacles. To build a map of its surroundings, the robot needs to integrate perceived data based its localization. At the same time, the robot has to compare its observation with the map to localize itself. This coupled problem is known as Simultaneous Localization And Mapping (SLAM). With the advances of SLAM, great progress has been made in spatial mapping. Geometric maps, generated by the spatial mapping, contain spatial information about the environment, either metric or topological, which allows a mobile robot to localize itself, plan a path, and avoid collisions with obstacles. Nowadays, spatial maps are popularly implemented in mobile robots.

With the improvement of life quality of human beings and the development of technology of robotics, especially for mobile robots, there is a trend that more robots will be introduced into domestic life [1], taking sweeping robots as an example. To co-exist with human-beings, robots face a series of challenges. First, a robot should behave in human rules, an unreasonable behavior may rise the antipathy of hosts, like standing a long time in front of a door or following the host too closely. Second, a robot may have to interact with the environment as some complex tasks are required, like a robot may be asked to go to another room with a door closed, the robot has to open the door to arrive the destination. Third, it is reasonable for a mobile robot to understand oral commands from its hosts, like "fetch an apple from the fridge". It is intractable for a mobile robot to handle those orders, with only spatial information of its surroundings.

The reason, preventing a mobile robot from applying in those situations, is that there is a gap between human beings and robots. A robot, equipped with a computer or microcomputer, tends to store and represent its environment in a mathematical way. While for human-beings, conceptual knowledge is preferred in the description of the environment and communication with others. For example, when asked to fetch a bottle of fresh milk, a person would walk to a refrigerator in a kitchen, and then open the door of the refrigerator to find whether there is a bottle of milk, while it could be problematic for a robot, as the robot has none of the prior knowledge about those concepts, such as kitchen, refrigerator, and milk bottle. A robot can perform such tasks with help of a human, by interpreting that information into a mathematical presentation, namely the geometric coordinates of those entities. Semantic mapping is a potential way to help robots to coexist with human beings, as it bridges the semantic gap by attaching semantic information to geometric maps.

Nowadays, semantic mapping in outdoor scenes has been developed quickly, due to its successful implementations in self-driving cars. In applications of self-driving cars, constructed geometric maps are segmented into several predefined categories, such as roads, trees, and pedestrians. While semantic mapping in indoor scenes faces different challenges. Indoor scenes are the main environment for human everyday life, thus there are more chances for a robot to interact with and serve humans. Besides, compared with several specific categories to be concerned in outdoor scenes, indoor scenes can be labeled by diverse place categories and placed with various kinds of objects. Additionally, human activities can make a change in their surroundings. For example, a person can introduce novel objects, discard useless objects, or change the positions of some instances. Thus, semantic mapping in indoor scenes has attracted the attention of many scholars in the research field of robotics. This work focuses on semantic mapping in indoor scenes and makes an effort to provide a comprehensive overview.

There are some other works, whose topics are related to semantic mapping, as listed in Table 1. Reference [2] is the first work to define a semantic map for robotics formally. The authors evaluated three topical semantic representations by this definition. The survey [3] is a comprehensive work, which reviewed the topic of semantic mapping of robotics, both in indoor and outdoor scenes. In this survey, references about semantic mapping were categorized by their primary characteristics. Furthermore, applications of semantic mapping are also introduced. While reference [4] focused on semantic information extraction from visual data. In [5], the history and trends of SLAM were introduced, and semantic mapping is reviewed as one of map representation. Reference [6] also surveyed the works of semantic maps, and focuses on their application in the navigation task. This survey is different from those works in three aspects. First, as mentioned above, we focus on semantic mapping in indoor scenes, as it faces different challenges with outdoor scenes. Second, both typical and recent semantic mapping methods are included in this survey, as listed in Table 2. Third, we try to review the semantic mapping system from a new perspective. Semantic mapping is divided into three modules, namely spatial mapping, acquisition of semantic information, and map representation, according to its definitions.

Table 1. Previous surveys related to semantic maoping for robotics.

Reference	Topic	Year
Paulus and Lang [2]	Definition of Semantic Mapping	2014
Kostavelis and Gasteratos [3]	Semantic Mapping	2015
Liu et al. [4]	Semantic Information Extraction	2016
Cadena et al. [5]	History and Trends of SLAM	2016
Crespo et al. [6]	Semantic Navigation	2020

Table 2. References about semantic mapping, each work is depicted in six aspects.

Reference	Sensors	SLAM methods	Acquisition Method	Content	Map Representation	Applications
[7]	sonar ring, laser, color camera	-	simplified instances and reference	object and room categories	two hierarchies	-
[8]	3D laser range	6D SLAM	reference and model matching	plain label and instance category	-	-
[9]	2D laser and a camera	GMapping	text detection and OCR	room information	-	-
[10]	Hokuyo laser range and Wearable motion sensors	-	reference	furniture type	-	-
[11]	laser scans, cameras, odometer	EKF SLAM	instance recognition and inference and property classification	instance category, room category and geometric property	4-layer architecture	reasoning about unexplored area
[12]	RGBD camera	-	2D instance segmentation	instances category	-	-
[13]	Depth camera	SLAM++	instance matching	instance category	-	augmented reality and relocalization
[14]	RGBD camera	-	human-robot interaction	*	world knowledge and domain knowledge	-
[15]	RGBD camera	-	dense scene segmentation	object category and background	-	-
[16]	RGB camera	LSD SLAM	CNN based 2D segmentation	object category and background	-	-
[17]	RGBD camera	GMapping	place classification	scene category	-	behave in human rules
[18]	RGBD camera	KinectFusion	CNN based 2D segmentation	object category and background	-	-
[19]	RGBD camera	graph-based SLAM [20]	CNN and SVM	object category	-	-
[21]	RGBD camera	ORB SLAM	SSD	object category	-	-
[22]	RGBD camera	DVO SLAM [23]	CNN-based semantic segmentation	object category and background	-	-
[24]	RGBD camera	Kinect Fusion	FCN sementic segmentation	object category and background	-	-
[25]	RGBD camera	ORB SLAM	Faster RCNN	object category and poses	-	-
[26]	RGBD camera	-	Mask R-CNN	object category	-	-
[27]	RGBD camera	voxblox [28]	PSPNet and Mask-RCNN	object category and background	-	-
[29]	Sonar and stereo camera	-	R-FCN	object category	-	semantic navigation
[30]	RGBD camera	ORB SLAM	CRF-RNN semantic segmentation	object category	-	-

The rest of this paper is organized as follows. In Section 2, we try to answer the question, What is a semantic map, by two definitions of a semantic map for robotics in earlier literature. According to its definitions, the typical mapping process is introduced and divided into three modules. Each of the three modules is reviewed in the next three sections respectively. Next, we discussed challenging open issues and potential future directions of semantic mapping. At last, the main contributions of this work are briefly concluded.

2. Definitions of Semantic Map

As many researchers believe that a map, designed and constructed for complex tasks for mobile robots, should be attached with semantic information, the concept semantic map has been proposed. Before we go further into semantic mapping, there is an essential question, i.e., What is a semantic map for robotics. To answer this question, two formal definitions of a semantic map are found in previous works.

In [8], a semantic map for robotics is defined as follows:

> *"A semantic map for a mobile robot is a map that contains, in addition to spatial information about the environment, assignments of mapped features to entities of known classes. Further knowledge about these entities, independent of the map contents, is available for reasoning in some knowledge base with an associated reasoning engine."*

Another formal definition of semantic robotic map is stated in reference [2]:

> *"A semantic map for E limited to D is a tuple* $\mathcal{M}_{sem} = \langle \mathcal{M}, \mathcal{L}, \mathcal{A} \rangle$*... $\mathcal{A}$ is a structure, which represents knowledge about the relationships between entities, classes, and attributes, also known as common-sense knowledge about D. Generally, $\mathcal{A}$ can be defined in an arbitrary way and has to allow for inference."*

where E is a mathematical description of the local environment, D is task domain, $\mathcal{M}$ is a set of maps for E, and $\mathcal{L}$ is a set of links.

Despite the different terms and expressions in the two definitions, there are three common characteristics in semantic mapping. First, a geometric map, containing geometric information of an environment, is the main body of semantic mapping, as it is the base for attaching semantic information, and serves for basic requirements of mobile robots, e.g., localization, obstacles avoidance, and path plan. Second, semantic information, as an extensive human-understandable description of environments, links physical entities with conceptual elements in a common-knowledge base, and bridges the semantic gap between humans and robots. Last, constructed geometric maps and acquired semantic information should be organized in an appropriate structure, which endows a robot the capacity of reasoning.

A typical binary grid map, built by a 2D laser scanner, can be viewed as a simple semantic map, as shown in Figure 1, according to those three characteristics. In this type of map, an environment is divided into regular grid cells, each cell is attached with a value, either 0 or 1, which indicates whether the cell is occupied by obstacles. Other than applying in the navigation task for robots, this information can be understood by human beings.

A typical semantic mapping pipeline is depicted in Figure 2. To build a semantic map, a robot, equipped with different types of sensors, like 2D or 3D laser scanners, RGB cameras, or RGBD cameras, constructs perception data into a geometric map, which employs SLAM technology as a front end. This module is called spatial mapping in this work. Next, the robot acquires semantic information from sensor data at the same time, or from the constructed geometric map in an off-line manner. Thus links between the prior common-sense knowledge base and elements in geometric maps are built through the acquisition of semantic information. At last, constructed geometric maps and acquired semantic information is organized into an appropriate representation, which enables a robot to reason further information and plan advanced tasks. This is one of the two reasons that semantic mapping is not a simple process of attaching semantic information to a geometry map. The other reason is that spatial mapping and acquisition of semantic mapping can benefit from each other. During spatial mapping, relative poses between sequential observation are obtained and can be employed to fuse semantic information, which is detailed in Section 4.

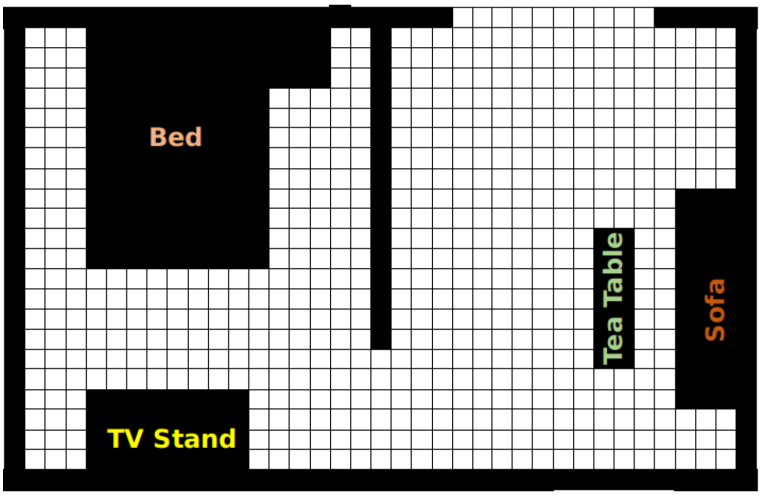

Figure 1. The 2D binary grid map, with each cell set with a 0 (empty) or 1 (occupied).

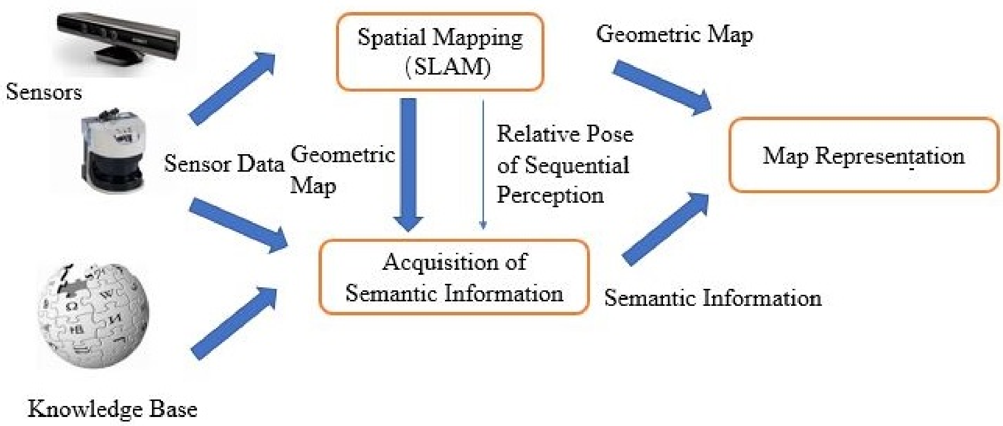

Figure 2. The typical semantic mapping pipeline contains three modules, namely spatial mapping, acquisition of semantic information, and map representation.

3. Spatial Mapping

Spatial mapping is the process to generate a geometric description of the environment from raw sensor data. To build a geometric map of an environment automatically, a robot needs the capacity of perception, which relies on equipped sensors. In the earlier era, 2D and 3D laser scanners are the most popular sensors for robots. As high-precision distance information can be obtained from reflection, in recent years, due to the popularity of commercial RGBD cameras, RGBD cameras have been widely equipped in robots and popularly utilized in spatial mapping. There is a significant advantage that RGBD cameras can obtain not only geometry information but also visual information simultaneously.

During spatial mapping, sequential perception data are organized into a geometric map incrementally. When the robot moves around in an unfamiliar scene, it localizes itself based on its perception of the environment. On the other hand, the accumulation of perception data is based on its localization. Thus, the two problems are coupled together as Simultaneous Localization and Mapping (SLAM), which has attracted the attention of scholars for more than 40 years, both filter-based [31,32] and graph-based [33] methods have been developed.

Thanks to the quick development of SLAM, which produces a geometric map, almost all works of semantic mapping take off-of-shelf SLAM methods as a front end. A particle filter-based SLAM, GMapping [34] is widely employed to build a 2D grid map of environment [7,9,17]. Based on 6D SLAM [35], 3D perception data, obtained from a 3D laser in a stop-scan-go fashion, is registered into a point cloud, via Iterative Closest Points (ICP) algorithm. Besides, a loop closure scheme is introduced to eliminate accumulated errors [8]. An off-line grid mapping method [36] based on Sonar is employed in [29] to build a metric map. In recent years, visual SLAM has attracted the interest of scholars. Reference [15] employs the fovis library [37] as an visual odometer, the observed points from each RGBD frame are projected into 3D constructed map, with Kalman filter. LSD SLAM [38] is employed in [16] to build a semi-dense map with an RGB camera, which is running on a CPU at 25 Hz. ElasitcFusion [39] is employed in [18] to construct an surfel-based map. RGBD version of ORB-SLAM2 [40] is employed in [21].

It should be addressed that, there is a difference between SLAM and spatial mapping. A map in SLAM, as a compact representation of the environment, is mainly used for localization, can be represented in a sparse or dense way. While for spatial mapping, a geometric map serves for more tasks, e.g., navigation. Navigation is a fundamental and vital requirement for an automatic mobile robot. A dense description is preferred as all obstacles are placed on the map, which enables the robot to plan a collision-free path.

There are three main categories of geometric maps, namely metric, topological, or hybrid. The direct outputs of SLAM are metric representations of an environment. a 2D metric map is built in [10], which provides a spatial presentation to attach semantic information. In a 2D planar environment, grid occupancy map is widely applied, as it has been proven its success in navigation, with the assistance of layered costmap [41]. While in 3D scenarios, point cloud [38] or surfel-based representation [12] are widely applied. While in other works [11], the topological structure is extracted from a metric map to store geometric information of the environment in a hybrid way. In which, a metric map is discretized into places, those places are connected by a path, which generates a topological map of the environment.

4. Acquisition of Semantic Information

During or after spatial mapping, semantic information, as a human-understandable description of the environment, is attached to geometric maps, which builds a bridge between a spatial model of the environment and a common-sense knowledge base. In other work, this process is called anchoring [7] or semantic information extraction [4]. By anchoring, physical entities observed by sensors are attached to conceptual elements in a knowledge base. While semantic information extraction focuses on object recognition or detection from visual images. To unify the process of describing the environment in a human-being understandable way, the term acquisition of semantic information is utilized in this work, due to two differences. First, only objects and room categories are considered in those works, while more information, such as temperature and humidity, can also enrich the map with more semantic descriptions. Second, semantic information can be obtained via diverse sensors and various methods. Thus, the acquisition of semantic information is preferred in this work as it describes this process in a more general way.

As shown in Table 2, categories of objects and rooms are the most common semantic information. The reason is that instances and rooms are the main elements in the real world, which has distinguishable boundaries. For example, in [7], sensory data are abstracted into two main physical entities, i.e., rooms, and objects. On the other hand, category information is also a vital concept in human common-sense knowledge. Thus the bridge between a geometric map and a common-sense knowledge base is built by categorizing physical instances and rooms. Besides the categories, other semantic information also plays an important role in robotic practical applications. For example, conceptual information is often utilized in the exploration of unobserved objects, and geometric properties, such as pose and size of objects, are necessary for pick and play tasks.

There are plenty of ways to acquire that semantic information. We classify those methods into three groups, namely human input, sensor-based, and inference, according to different sources used in this process.

4.1. Human Input

The most intuitive way to create a semantic map is to acquire semantic information by taking human input into the process. A semantic map can be generated by adding conceptual information into a constructed geometric map by hand-coding. More automatically, before exploring the environment, artificial landmarks, which encode semantic information, are placed in the environment, or attached to big objects. A robot obtains semantic information by moving around and detecting those landmarks. Reference [9] can be viewed as one of the examples. A fully automatic method based on off-of-shelf text detection and optical character recognition (OCR) is employed in [9]. A robot moves along in the hallway, scans walls around it to detect text from room signs, and then recognizes characters with an OCR system. Textual information, such as owner, capacity, and room number, is extracted and attached to the generated geometric map.

Furthermore, human-robot interaction is employed in [14] as a source of semantic information. In this work, instances are pointed with a commercial laser pointer, the robot detects the dot and segments the instance based on its size as prior knowledge. Once the instance is detected, both commands and object description are obtained through dialogue with humans, which is enabled with an automatic speech recognition module.

There is an obvious advantage of taking human input as a semantic information source. Information loss can be avoided as the information is transferred directly from a human to a robot. Otherwise, the content of semantic information is not restricted, as a user can add arbitrary knowledge to the map. However, the process is heavily dependent on humans. To overcome this limitation, many automatic acquisition methods are proposed based on sensor data or inference.

4.2. Sensor-Based Methods

In the early years, the object category is determined by instance recognition, with an object database prepared beforehand. When mapping, robots observe the environment and recognize predefined objects through visual or geometric features. To that end, RGB cameras, 3D laser scanners, as well as RGB-D cameras are widely employed for the acquisition of semantic information. Objects are recognized with BLORT toolkit [42] with prepared object models. Reference [11] uses SIFT (Scale-Invariant Feature Transform) features to recognize object instances, only six categories are considered, namely, a book, a cereal box, a computer, a robot, a stapler, and a roll of toilet paper. SLAM++ [13] aims at building an object-oriented map, with repeated instances represented in this map. An object database, consisting of 3D models of instance, is prepared beforehand using KinectFusion [43], and then instances are recognized with Generalized Hough Transform based on Point Pair Features. In [8], after scene interpreting, planes are removed from the point cloud, objects are detected with prior knowledge in an object database, using a trained Support Vector Machine (SVM) or a Gentle Adaptive Boost Algorithm (Adaboost) [44] with features in rendered depth and reflective images. While preparing such an object database can be cumbersome, as there can be plenty of instances in an environment, and each category can have diverse instances. Some works are focusing on extract general category information by training on datasets with huge samples.

Besides instance recognition, there are some works focusing on scene interpreting by semantic segmentation. Reference [12] extracts semantic information by segmenting images pixel-wise, with a Random Forest. Both depth and color region features are taken into consideration. Once objects are segmented in a single frame, class-wise surfels are projected into a 3D reconstructed map based on camera trajectory estimated in spatial mapping. Similar to [12], a 2D segmentation algorithm, based on Randomized Decision Forests, is proposed to segment each RGBD frame in [15]. An RGBD frame is segmented via a trained Randomized Decision Forest, accelerated by well-designed features and a keyframe-based scheme. A dense conditional random field is employed to improve spatial consistency both in 2D and 3D refinement with visual and geometrical similarity.

Nowadays, with successful applications in computer vision, convolutional neural networks (CNNs) have attracted the attention of plenty of scholars in semantic mapping [45]. Reference [16] employed DeepLab v2 [46] to segment an RGBD frame, which is a dilated convolution network, with spatial pyramid pooling to handle both small and large objects. Similar to [12], keyframe-based scheme is also utilized in [16]. In [17], CNN based place categorization method is combined with spatial mapping. The Places205 network [47] is transferred to the real-world environment without any fine-tuning. To overcome the close-set limitation, the network is expanded by adding a one-vs-all classifier with renormalization. CT-MAP was proposed by [25], in which not only object categories but object poses are detected. In this work, Faster RCNN [48] is deployed as an object detector. To boost object detection performance, category-level relations, which are obtained from public datasets, are organized in the conditional random field (CRF) and temporal consistency is also considered with the assumption that an object would stay in a short period. Reference [21] combines ORB-SLAM2 with single-stage object detector Single-Shot Detector (SSD) [4] to build an object-oriented meaningful map. The object detector is running in every single RGB image, and a series of bounding boxes are returned. To segment objects from depth images, the depth image is over segmented into super voxels, similar ones are get together by construct an adjacency graph and cutting edges based on convex/concave information. In sequential observation, object candidates are associated by measuring Euler distance between each other. with the k-d tree to speed up. The single detection labels are accumulated and determined by max value. R-FCN [36] is employed in [29] to extract object category information. First, a point cloud is built with ORB SLAM [40] or LSD SLAM [38]. Then, a user specifies a particular part of the point cloud, the robot moves around to observe from different viewpoints to detect it. Objects are detected and projected to a 2D occupy map. Then minimal bounding rectangles (MBRs) are used to represent objects in the 2D occupy map.

Acquisition of semantic information is not a simple categorization problem. When mapping, a robot moves around in the environment, a target object may be observed in a series of frames from different viewpoints. To keep temporal consistency, sequential information should be fused. For example, the Bayesian framework is employed to fuse the label results of each frame in an incremental manner [15,17,22]. Besides, there are some potential rules in indoor scenes based on human living habits, such as an oven is more likely placed in a kitchen. Those contextual relations are vital cues to keep global consistency. References [15,18,25] employ CRF to boost performance of categorization.

4.3. Inference

Instead of obtained from raw sensor data, some works try to extract semantic information based on obtained semantic information, with help of implicit or explicit contextual relations. In reference [7], the room categories are defined in a common-sense knowledge base with NeoClass language, which enables a robot to reason the category of a room, based on objects detected in this room via PTLplan [49]. Once the point cloud of the environment is obtained, Random Sample Consensus (RANSAC) [50] is utilized to extract plains in scenes in [8]. Those plains are labeled as walls, doors, floors, or ceilings, by a constraint network, which contains geometric relations of each type of planes. This problem is solved with Prolog. Instead of using low-level sensor data and monomodal information, a multimodal property-based reasoning architecture is proposed in [11], in which properties, such as objects, and appearance, are obtained from sensor data directly, and serves as a mediate level between lower sensory information and high-level conceptual. Both room categories and properties are jointly estimated with a probabilistic chain graph model, which is a generalization of Bayesian Networks and Markov Random Filed. The geometric properties are classified by an Support Vector Machine (SVM), which is trained with a dataset of several room instances. The object-room relation can be obtained in the Open Mind Indoor Common Sense database.

Despite based on conceptual information in the environment, authors in reference [10] argued that there is a correlation between furniture type and human activities and proposed a novel method to reason furniture type by human interaction. To that end, human activities are recognized from multi-wearable motion sensors via neural networks and hidden Markov models. Once activities are recognized, furniture type can be determined by a predefined activity-furniture table.

As it is based on other obtained semantic information, the inference is employed as a post-process following other acquisition methods. For example, when an oven is discovered in the scene, the room can be inferred as a kitchen. The capacity of inference relies on reasoning rules, thus it matters how to prepare contextual information, In [11], Open Mind Indoor Common Sense is utilized to extract related conceptual information, While some works extract such information from public computer vision dataset. In [17], place categories are used to boost object recognition, with object-scene as prior knowledge, extracted from NYU depth dataset.

5. Map Representation

Once a geometric map is built and semantic information is acquired, both spatial information and conceptual information should be organized under certain rules. While the main purpose of most works is to show the acquisition results of semantic information, thus there is a way to attach semantic labels to physical elements in geometric maps and those elements with different colors in visual representation. For example, authors in [21] employ ORB-SLAM2 [40] to build a point cloud model of the environment, and several kinds of objects are detected and shown in different colors. As a result of semantic mapping, Reference [8] also labels scenes and objects with different colors. A semantic map is constructed in [29] by inserting minimum bounding rectangles (MBRs) of an object into the corresponding area. With semantic information attached to geometric maps, it is straightforward to interpret semantic labels into geometric position, which enables a robot the capacity of semantic navigation.

Besides simple attachment, there is some typical and sophisticated structure to represent a semantic map, which endows the capacity of reasoning. Hierarchy structure is a popular way to represent environment information, with specific information placed in lower levels, and the abstract in the upper. Two hierarchies are employed in reference [7]. One is for spatial information, and the other for conceptual. The former contains three levels, local grid maps and object images are placed in the lowest level, the middle one is a topological graph of places, and the top is a node of the whole space. The conceptual hierarchy is coded beforehand with NeoClassic language. The top is a *Thing* node, including two main branches, *Room* and *Object*, categories of rooms and objects are lying in the third level and the lowest for specific instances. The links between spatial hierarchies and conceptual hierarchies are built by anchoring.

In [11], the spatial knowledge is represented in a 4-layer structure, namely sensory layer, place layer, categorical layer, and conceptual layer. With sensor data on the bottom layer and abstract concept knowledge on the top layer. The sensory layer includes a metric map of the environment. On the place level, a topological graph is maintained with places as nodes, and paths as edges. Objects and landmarks are placed on the categorical layer, and spatial properties and visual models are also included in this layer. The conceptual layer contains the relations linking lower-level sensory information to common-sense knowledge, as well as a static ontology of common-sense knowledge. As a probabilistic chain graph [51] is employed in [11], a probability reasoning engine Loopy Belief Propagation [52] is utilized to perform inference over semantic information.

The map representation in [14] is divided into two parts. World knowledge contains acquired knowledge of a certain environment. The other, named domain knowledge, contains general knowledge in a task domain. In world knowledge, an occupancy grid map is maintained, and cell maps are utilized to represent local areas, a topological map maintains the graph model of the whole environment. Besides, instances with their categories and properties are saved in this knowledge.

It can be concluded that a hierarchy structure is preferred in map representation. In a hierarchy structure, the information in different levels of abstraction is clearly distinguished by placing specific information on the lower layer, otherwise on higher. Ontology is an appropriate way to represent conceptual knowledge, as other information, such as properties and relations, are expressed in an organized way. An representation example of indoor scene, which in shown in Figure 1, is depicted in Figure 3.

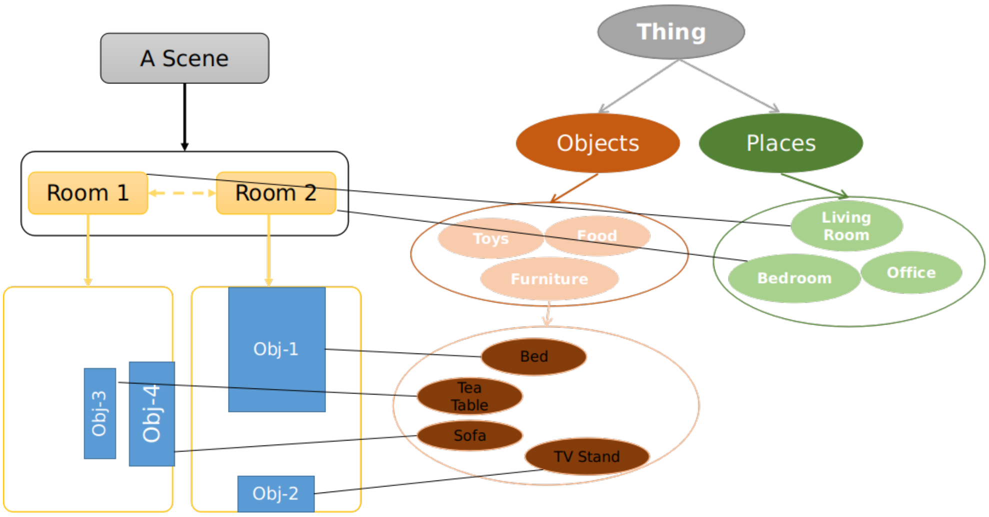

Figure 3. Presentation of a semantic map of the scene shown in Figure 1, a hierarchy structure of the geometric map (**left**), and onlogy of common-sense knowledge (**right**).

6. Open Issues and Potential Directions

Though many impressive semantic mapping methods have been proposed, there are still some limitations in implements of practical applications. Besides, there are some potential directions for more advanced semantic mapping.

6.1. Heterogeneous Sensor Fusion

Currently, in the process of semantic mapping, lasers and cameras are widely employed in most works. There is still a potential to fuse different information from heterogeneous sensors. One impressive example is the success of Visual-Inertial Odometry (VIO) [53,54], which fuses camera data with inertial measurement units to localize with higher precision. Furthermore, with specific sensors, some further information can be obtained directly. For example, equipped with a temperature sensor, a robot can get temperature information. Tactile sensors are helpful to recognize objects when they are not visually distinguished [55]. With multi-modal information, a more comprehensive description of the environment can be generated.

6.2. Dynamic Scenes and Open World

As the main place of human activities, indoor scenes can change over time. Those changes will pose challenges for semantic mapping. First of all, most geometric maps are designed to model a scene in still mode. While for dynamic scenes, it is impossible to map the environment at each time. Once a scene has been changed, the reconstructed map will lose its effect. Moreover, a human may introduce novel objects into the environment, which is known as open-world challenge [56]. There is a need for a robot to learn in an incremental mode. To deal with this issue, [17] proposes an expandable classification system by combining ConvNet with one-vs-all classifiers. Reference [14] also addressed this problem, and proposed a method to add semantic information into a in an incremental way. For map representation in dynamic scenes, few works have addressed this problem, there is a need for an efficient way to represent a dynamic scene.

6.3. Cloud Robotics

In recent years, cloud robotics has attracted the attention of many researchers. There are a few works that explore the potential to implement cloud robotics in semantic mapping. For example, [56] introduces RoboEarth, which defines several recipes for robots, and has shown its effect to build a semantic map with a cloud robot. While different scenes may share similar knowledge, as there are some common rules, which are obeyed by different people. Thus it can be researched to endow a robot with semantic mapping capacity with the assistance of a cloud robot.

6.4. Task-Oriented Map Representation

Even though some typical map representations have been reviewed, it is still not clear that how can we implement those representations in task planning. In the field of computer science, a computer program consists of algorithms and data structures. Compared with computer programming, one possible way to implement task planning is to find a appreciate way to represent a robot task into a description of robot capacities and a representation of a semantic map. Thus the problem can be decoupled into two parts. one is a mathematical or logical description of robot capacity, and a map representation, which should consist of necessary elements for task planning.

7. Conclusions

In this paper, semantic mapping in indoor scenes for mobile robots has been reviewed. According to definitions in different works, semantic mapping consists of three modules, namely spatial mapping, acquisition of semantic information, and map representation.

Dense SLAM technology is usually employed as a front-end to produce geometric maps, with spatial information of the environment. Thus, spatial mapping benefits from the impressive advances of SLAM technology, such as GMapping [34], ORB SLAM2 [40], and ElasticFusion [39].

Semantic information in maps for mobile robots provides a human-understandable description of the environment. Acquisition of semantic information builds a bridge between physical elements in the environment and conceptual elements in a knowledge base. At first, with prior knowledge of instances, object recognition is utilized to categorize instances. While with the development of computer vision, vision-based segmentation on each frame is implemented and frame-based segmentation results are fused incrementally based on the Bayes framework. In recent years, due to advances in deep learning, there is a trend to employ off-the-shelf networks in instance categorization or scene interpreting.

For map representation, many works represent a constructed map for visualization, namely labeling semantic elements with different colors. While it is a clear way to show semantic information acquisition results, but not intuitive to implement it in robot tasks, while there is a consensus to represent knowledge in ontology. Some typical methods have been proposed to utilize hierarchical layers to organize the different level of knowledge, by placing abstract information in higher levels.

Despite the progress that has been achieved during the last decades, there is still a long way to implement semantic mapping in more practical applications of mobile robots. To that end, some future potential directions have been pointed out.

Author Contributions: Investigation, X.H., S.L. and X.W.; writing—original draft preparation, X.W. and S.L.; writing—review and editing, X.H.; visualization, X.H.; supervision, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ding, H.; Yang, X.; Zheng, N.; Li, M.; Lai, Y.; Wu, H. Tri-Co Robot: A Chinese robotic research initiative for enhanced robot interaction capabilities. *Natl. Sci. Rev.* **2017**, *5*, 799–801. [CrossRef]
2. Paulus, D.; Lang, D. Semantic Maps for Robotics. 2014. Available online: http://people.csail.mit.edu/gdk/iros-airob14/papers/Lang_finalSubmission_SemantiCmapsForRobots.pdf (accessed on 18 February 2021).
3. Kostavelis, I.; Gasteratos, A. Semantic mapping for mobile robotics tasks: A survey. *Robot. Auton. Syst.* **2015**, *66*, 86–103. [CrossRef]
4. Liu, Q.; Li, R.; Hu, H.; Gu, D. Extracting semantic information from visual data: A survey. *Robotics* **2016**, *5*, 8. [CrossRef]
5. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J.J. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Trans. Robot.* **2016**, *32*, 1309–1332. [CrossRef]
6. Crespo, J.; Castillo, J.C.; Mozos, O.; Barber, R. Semantic Information for Robot Navigation: A Survey. *Appl. Sci.* **2020**, *10*, 497. [CrossRef]
7. Galindo, C.; Saffiotti, A.; Coradeschi, S.; Buschka, P.; Fernandez-Madrigal, J.A.; González, J. Multi-hierarchical semantic maps for mobile robotics. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 2278–2283.
8. Nüchter, A.; Hertzberg, J. Towards semantic maps for mobile robots. *Robot. Auton. Syst.* **2008**, *56*, 915–926. [CrossRef]
9. Case, C.; Suresh, B.; Coates, A.; Ng, A.Y. Autonomous sign reading for semantic mapping. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3297–3303.
10. Li, G.; Zhu, C.; Du, J.; Cheng, Q.; Sheng, W.; Chen, H. Robot semantic mapping through wearable sensor-based human activity recognition. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 5228–5233.
11. Pronobis, A.; Jensfelt, P. Large-scale semantic mapping and reasoning with heterogeneous modalities. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 3515–3522.
12. Stückler, J.; Biresev, N.; Behnke, S. Semantic mapping using object-class segmentation of RGB-D images. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 3005–3010.
13. Salas-Moreno, R.F.; Newcombe, R.A.; Strasdat, H.; Kelly, P.H.; Davison, A.J. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.
14. Bastianelli, E.; Bloisi, D.D.; Capobianco, R.; Cossu, F.; Gemignani, G.; Iocchi, L.; Nardi, D. On-line semantic mapping. In Proceedings of the 2013 16th International Conference on Advanced Robotics (ICAR), Montevideo, Uruguay, 25–29 November 2013; pp. 1–6.
15. Hermans, A.; Floros, G.; Leibe, B. Dense 3D semantic mapping of indoor scenes from RGB-D images. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 2631–2638.
16. Li, X.; Belaroussi, R. Semi-Dense 3D Semantic Mapping from Monocular SLAM. *arXiv* **2016**, arXiv:1611.04144.
17. Sünderhauf, N.; Dayoub, F.; McMahon, S.; Talbot, B.; Schulz, R.; Corke, P.; Wyeth, G.; Upcroft, B.; Milford, M. Place categorization and semantic mapping on a mobile robot. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 5729–5736.
18. McCormac, J.; Handa, A.; Davison, A.; Leutenegger, S. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 4628–4635.
19. Himstedt, M.; Maehle, E. Online semantic mapping of logistic environments using RGB-D cameras. *Int. J. Adv. Robot. Syst.* **2017**, *14*, 1729881417720781. [CrossRef]
20. Himstedt, M.; Keil, S.; Hellbach, S.; Böhme, H.J. A Robust Graph Based Framework for Building Precise Maps from Laser Range Scans. Available online: https://www.tu-chemnitz.de/etit/proaut/ICRAWorkshopFactorGraphs/ICRA_Workshop_on_Robust_and_Multimodal_Inference_in_Factor_Graphs/Program_files/2%20-%20PreciseMaps%20Slides.pdf (accessed on 18 February 2021).
21. Sünderhauf, N.; Pham, T.T.; Latif, Y.; Milford, M.; Reid, I. Meaningful maps with object-oriented semantic mapping. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5079–5085.

22. Ma, L.; Stückler, J.; Kerl, C.; Cremers, D. Multi-view deep learning for consistent semantic mapping with RGB-D cameras. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 598–605. [CrossRef]
23. Kerl, C.; Sturm, J.; Cremers, D. Dense visual SLAM for RGB-D cameras. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; pp. 2100–2106. [CrossRef]
24. Xiang, Y.; Fox, D. DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks. *arXiv* **2017**, arXiv:1703.03098. .
25. Zeng, Z.; Zhou, Y.; Jenkins, O.C.; Desingh, K. Semantic Mapping with Simultaneous Object Detection and Localization. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 911–918.
26. Grinvald, M.; Furrer, F.; Novkovic, T.; Chung, J.J.; Cadena, C.; Siegwart, R.; Nieto, J. Volumetric Instance-Aware Semantic Mapping and 3D Object Discovery. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3037–3044. [CrossRef]
27. Narita, G.; Seno, T.; Ishikawa, T.; Kaji, Y. PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things. *arXiv* **2019**, arXiv:1903.01177.
28. Oleynikova, H.; Taylor, Z.; Fehr, M.; Siegwart, R.; Nieto, J. Voxblox: Incremental 3D Euclidean Signed Distance Fields for on-board MAV planning. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 1366–1373.
29. Qi, X.; Wang, W.; Yuan, M.; Wang, Y.; Li, M.; Xue, L.; Sun, Y. Building semantic grid maps for domestic robot navigation. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881419900066. [CrossRef]
30. Cheng, J.; Sun, Y.; Meng, M.Q.H. Robust Semantic Mapping in Challenging Environments. *Robotica* **2020**, *38*, 256–270. [CrossRef]
31. Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* **2006**, *13*, 99–110. [CrossRef]
32. Bailey, T.S.; Durrantwhyte, H. Simultaneous localization and mapping (SLAM): Part II. *IEEE Robot. Autom. Mag.* **2006**, *13*, 108–117. [CrossRef]
33. Grisetti, G.; Kummerle, R.; Stachniss, C.; Burgard, W. A Tutorial on Graph-Based SLAM. *IEEE Intell. Transp. Syst. Mag.* **2010**, *2*, 31–43. [CrossRef]
34. Grisetti, G.; Stachniss, C.; Burgard, W. Improved Techniques for Grid Mapping With Rao-Blackwellized Particle Filters. *IEEE Trans. Robot.* **2007**, *23*, 34–46. [CrossRef]
35. Nüchter, A.; Lingemann, K.; Hertzberg, J.; Surmann, H. 6D SLAM—3D mapping outdoor environments. *J. Field Robot.* **2007**, *24*, 699–722. [CrossRef]
36. Lee, K.; Lee, S.J.; Kölsch, M.; Chung, W.K. Enhanced maximum likelihood grid map with reprocessing incorrect sonar measurements. *Auton. Robot.* **2013**, *35*, 123–141. [CrossRef]
37. Huang, A.S.; Bachrach, A.; Henry, P.; Krainin, M.; Maturana, D.; Fox, D.; Roy, N., Visual Odometry and Mapping for Autonomous Flight Using an RGB-D Camera. In *Robotics Research : The 15th International Symposium ISRR*; Christensen, H.I., Khatib, O., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 235–252.
38. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 834–849.
39. Whelan, T.; Salas-Moreno, R.F.; Glocker, B.; Davison, A.J.; Leutenegger, S. ElasticFusion: Real-time dense SLAM and light source estimation. *Int. J. Robot. Res.* **2016**, *35*, 1697–1716. [CrossRef]
40. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
41. Lu, D.V.; Hershberger, D.; Smart, W.D. Layered Costmaps for Context-Sensitive Navigation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Chicago, IL, USA, 14–18 September 2014.
42. Mörwald, T.; Prankl, J.; Richtsfeld, A.; Zillich, M.; Vincze, M. BlLORT—The Blocks World Robotic Vision Toolbox. Available online: http://users.acin.tuwien.ac.at/mzillich/files/moerwald2010blort.pdf (accessed on 18 February 2021).
43. Newcombe, R.A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A.J.; Kohi, P.; Shotton, J.; Hodges, S.; Fitzgibbon, A. KinectFusion: Real-time dense surface mapping and tracking. In Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; pp. 127–136.
44. Freund, Y.; Schapire, R.E. Experiments with a New Boosting Algorithm. Available online: https://cseweb.ucsd.edu/~yfreund/papers/boostingexperiments.pdf (accessed on 18 February 2021).
45. Cebollada, S.; Payá, L.; Flores, M.; Peidró, A.; Reinoso, O. A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data. *Expert Syst. Appl.* **2020**, 114195. [CrossRef]
46. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
47. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 487–495.
48. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 91–99.

49. Karlsson, L. Conditional Progressive Planning under Uncertainty. Available online: https://www.researchgate.net/publication/2927504_Conditional_Progressive_Planning_under_Uncertainty (accessed on 18 February 2021).
50. Fischler, M.A.; Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
51. Lauritzen, S.L.; Richardson, T.S. Chain graph models and their causal interpretations. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2002**, *64*, 321–348. [CrossRef]
52. Mooij, J.M. libDAI: A Free and Open Source C++ Library for Discrete Approximate Inference in Graphical Models. *J. Mach. Learn. Res.* **2010**, *11*, 2169–2173.
53. Yang, Z.; Shen, S. Monocular Visual–Inertial State Estimation With Online Initialization and Camera–IMU Extrinsic Calibration. *IEEE Trans. Autom. Sci. Eng.* **2017**, *14*, 39–51. [CrossRef]
54. Kang, R.; Xiong, L.; Xu, M.; Zhao, J.; Zhang, P. VINS-Vehicle: A Tightly-Coupled Vehicle Dynamics Extension to Visual-Inertial State Estimator. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 3593–3600.
55. Liu, H.; Yu, Y.; Sun, F.; Gu, J. Visual–Tactile Fusion for Object Recognition. *IEEE Trans. Autom. Sci. Eng.* **2017**, *14*, 996–1008. [CrossRef]
56. Riazuelo, L.; Tenorth, M.; Di Marco, D.; Salas, M.; Gálvez-López, D.; Mösenlechner, L.; Kunze, L.; Beetz, M.; Tardós, J.D.; Montano, L.; et al. RoboEarth Semantic Mapping: A Cloud Enabled Knowledge-Based Approach. *IEEE Trans. Autom. Sci. Eng.* **2015**, *12*, 432–443. [CrossRef]