robotics

MDPI

Review

Extracting Semantic Information from Visual Data: A Survey

Qiang Liu *, Ruihao Li, Huosheng Hu and Dongbing Gu

School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK; rlig@essex.ac.uk (R.L.); hhu@essex.ac.uk (H.H.); dgu@essex.ac.uk (D.G.)

* Correspondence: qliui@essex.ac.uk; Tel.: +44-1206-874-092

Academic Editor: Hong Zhang

Received: 17 December 2015; Accepted: 23 February 2016; Published: 2 March 2016

Abstract: The traditional environment maps built by mobile robots include both metric ones and topological ones. These maps are navigation-oriented and not adequate for service robots to interact with or serve human users who normally rely on the conceptual knowledge or semantic contents of the environment. Therefore, the construction of semantic maps becomes necessary for building an effective human-robot interface for service robots. This paper reviews recent research and development in the field of visual-based semantic mapping. The main focus is placed on how to extract semantic information from visual data in terms of feature extraction, object/place recognition and semantic representation methods.

Keywords: semantic map; visual data; feature extraction; object recognition; place recognition; semantic representation

1. Introduction

Traditionally, robotic mapping is broadly divided into two categories: metric and topological mapping. Metric maps describe the geometric features of the environment, whereas topological maps involve the connectivity of different places and are used for robots to navigate from one place to another [1]. Figure 1 shows both a metric map and a topological map that were constructed. An early representative of the metric mapping approach is based on occupancy grids that model the occupied and free space. In contrast, the topological mapping approach uses nodes to represent distinct places or landmarks and curved lines to describe the paths between nodes. Recently, a new hybrid mapping that combines the metric and topological paradigm was developed to compensate the weakness of individual approaches. This mapping approach applies a metric map for accurate navigation in a local space and a global topological map for moving from one place to another.

All of these traditional mapping approaches are navigation oriented and enable mobile robots to navigate around and plan a path to reach a goal [2]. The maps built by traditional mapping approaches are relatively low-level maps since they are unable to interpret scenes or encode semantic information. To serve people, service robots should be able to communicate with humans through semantic information, such as human speech commands, "Can I have a cup of coffee?" or "Please open the window", so that they are able to interact with humans in a human-compatible way [3].

In a semantic map, nodes representing places and landmarks are named by linguistic words. Examples of these include names and categories of different objects, rooms and locations. More specifically, a semantic map can be regarded as an extension of a hybrid map, which contains a geometric description, topological connectivity and semantic interpretation [4]. It provides a friendly way for robots to communicate with humans. This survey reviews numerous literature works in visual-based semantic mapping and attempts to provide an overview of the state-of-the-art

methodologies in this field. It is mainly focused on how to extract semantic information from visual data, including feature extraction, object/place recognition and semantic representation methods. It differs from other existing comprehensive surveys on traditional robot mapping approaches [1] or general semantic mapping methodologies [5].

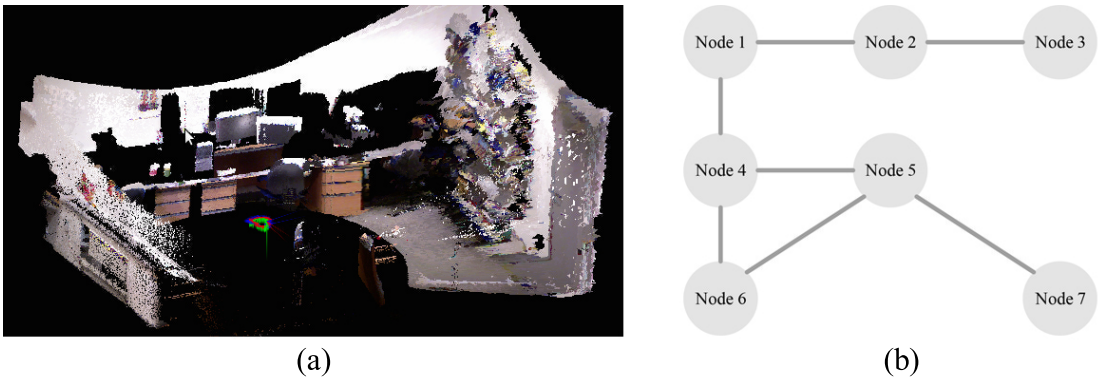

Figure 1. Maps obtained through traditional mapping approaches. (**a**) Metric map; (**b**) Topological map.

The rest of this paper is organised as follows. Section 2 overviews various visual-based approaches in the semantic mapping process, including different visual sensors, such as conventional cameras, RGB-D cameras and stereo cameras. In Section 3, visual feature extraction methodologies are outlined and classified in terms of global and local features. Section 4 describes three basic recognition approaches in semantic mapping, namely global, local and hybrid approaches. Subsequently, how to generate semantic representations of the environment is outlined in Section 5, and some typical real-world applications are presented in Section 6. Finally, a brief conclusion and discussion are given in Section 7.

2. Overview of Visual-Based Approaches to Semantic Mapping

More recently, using semantic data to represent environments has become a popular research domain and drawn enormous attention from different fields [6,7]. The application of visual data in semantic mapping systems seems to be a sensible decision, as humans perceive the world through their eyes. Visual data allow the representation of both low-level features, such as lines, corners and shapes, and high-level features, such as colours, relations and texts. In this way, a wider variety of objects can be recognized, which can highly enrich semantic maps.

2.1. General Process for Semantic Mapping

In general, a visual-based semantic mapping system consists of three parts. At first, the specific features are pre-selected based on sensor type, and feature descriptors are computed and obtained. Subsequently, features or descriptors are classified in terms of *a priori* knowledge so that objects and places can be recognized. Finally, properties are endowed with semantic meanings on the basis of topological and metric maps. Figure 2 presents the general process for semantic mapping. Note that a metric map is considered as a complementary attribute of a semantic map. In addition, some systems rely on direct image segmentation to obtain semantic information rather than using feature descriptors to represent objects or scenes.

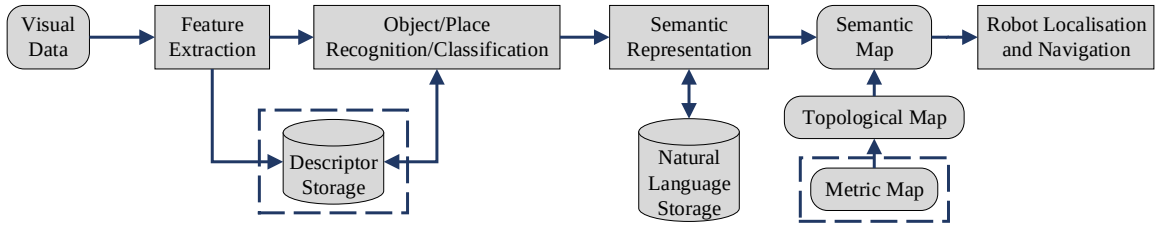

Figure 2. Overview of the general process for semantic mapping.

Many types of visual sensors have been developed to provide a variety of interpretations of the world. In addition, the subsequent processing methods are highly dependent on the data type used. To some extent, the visual sensor applied plays a key role in a semantic mapping system. The visual sensors used for robot mapping can be generally classified into two categories according to the data generated: a 2D image and a 3D point cloud.

2.2. Use of Conventional Cameras

At the early stage of development, visually recognizing objects or places was normally done by using conventional cameras that record two-dimensional images, as shown in Figure 3. In [8], models of indoor Manhattan scenes [9] were acquired from individual images generated from a 2D camera and then assigned with semantic labels. Anguelov *et al.* captured colour and motion properties with an omni-directional camera for door modelling and detection [10]. Tian *et al.* also addressed this problem, since doors are generally considered as distinguishable landmarks between rooms in indoor environments [11]. Wu *et al.* [12] and Neves dos Santos *et al.* [13] employed 2D visual data to tackle the problem of place recognition in semantic mapping, and object recognition with monocular cameras was presented by Civera *et al.* [14] and Riazuelo *et al.* [15].

Figure 3. Cameras that capture 2D images. (**a**) Monocular camera; (**b**) omni-directional camera.

2.3. Use of RGB-D Cameras

In recent years, extracting semantic information from 3D point clouds has become a new trend due to the availability of low-cost and light-weight 3D point cloud capturing devices, such as stereo cameras and RGB-D sensors, which allow the application to small robot platforms or even wearable devices easily. Compared to 2D images, 3D point clouds overcome the limitation in the data-stream itself by providing additional depth data. Moreover, humans recognize and perceive a 3D world in terms of our eyes. Therefore, object recognition through capturing 2D projections of the 3D world is inevitably inaccurate and might be even misleadingly suggested, especially when it comes to a large variety of goods in our daily life [16]. Figure 4 presents good evidence. Judging from the match between the two watches in the left image, this match seems quite successful. However, if we zoom out from the second watch on the right, it is in fact another image stitched to the surface of a mug. The semantic interpretation of a mug in this case is thus completely wrong. This is a

clear example indicating that some semantic representations cannot be entirely precise without 3D spatial information.

(a) (b)

Figure 4. Object recognition using features extracted only from 2D images. (**a**) Object matching that seems correct; (**b**) After zooming out from the image on the right, we can tell that the semantic information obtained is completely wrong.

Therefore, cameras that are capable of providing not only high-resolution 2D images, but also positioning and other characteristics of objects are widely used in semantic mapping systems. Various technologies can be used to build 3D scanning devices, such as stereoscopy, time-of-flight, structured light, conoscopic holography and modulated light. However, each technology comes with its own limitations, advantages and costs. Some of the RGB-D cameras deployed in the published works are shown in Figure 5.

(a) (b)

Figure 5. RGB-D cameras. (**a**) Kinect for Xbox 360; (**b**) Kinect for Xbox One.

RGB-D cameras generate 3D images through structured light or time-of-flight technology, both of which can provide depth information directly. In terms of a structured light camera, the camera projects a known pattern onto objects and perceives the deformation of the pattern by an infrared camera to calculate the depth and surface information of the objects, as shown in Figure 6. For a time-of-flight camera, the camera obtains depth information by measuring the time of flight of a light signal between the camera and objects. Compared to RGB-D cameras based on time-of-flight technology (e.g., Kinect for Xbox One), the structured light sensors (e.g., Kinect for Xbox 360) are sensitive to illumination. This limits their applicability in direct sunlight [17].

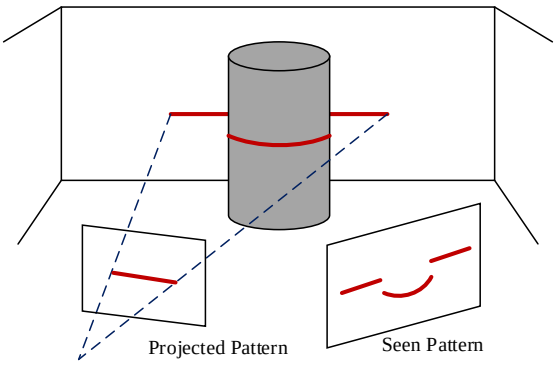

Figure 6. Principle of RGB-D cameras based on structured light technology.

2.4. Use of Stereo Cameras

A stereo camera has two or more lenses with a separate image sensor or film frame for each lens, as shown in Figure 7. It simulates human binocular vision and therefore gives the ability to capture 3D images [18], even though the process only involves capturing 2D images with each lens. Figure 8 shows the simplest schematic representation and the basic principle of many 3D stereo camera systems. A point in the real world is projected onto two film frames differently by two cameras due to their disparate positions. The point in the left camera image is shifted by a given distance in the right camera image. If the relative position of each point is known in each camera (sometimes a hard task), the depth value can then be obtained by applying the following equation to each point pair as observed by both cameras:

$$D = f(b/d) \tag{1}$$

where D is the depth value, f is the focal length of the camera, b is the baseline between the cameras and d is the disparity of the same point in different film frames. In this case for Point 2, $d = d_1 + d_2$.

Figure 7. Stereo cameras. (**a**) Minoru 3D webcam; (**b**) Ensenso N10 stereo 3D camera.

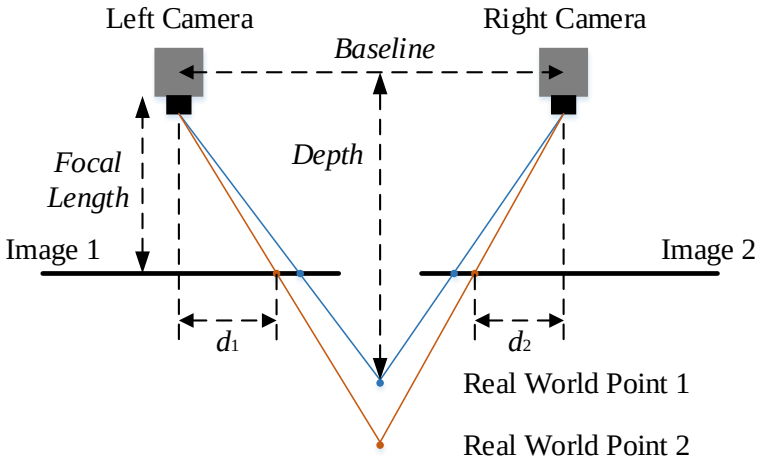

Figure 8. Principle of most 3D stereo cameras.

3. Visual Features Applied in Semantic Mapping Systems

In the last decade, some researchers have reported systems in which semantic interpretation of certain scenes was obtained [19,20]. However, the acquisition was done through conversations between humans and robots or even hand-coded into the systems, rather than using the robots' own sensors [21]. Visual features describe the elementary characteristics, such as shapes, colours, textures, motions and relations, between pixels in raw visual data and can be broadly divided into two categories: global and local features [22].

Global features represent an image as a whole without directly describing the spatial layout of the properties in the image. More specifically, the statistics of all of the pixels in a movable

fixed-size bounding box is extracted to generate feature vectors, which can determine the likelihood for image matches. Such features are suitable for large-scale environment recognition, e.g., roads, lawns in outdoor environments and rooms in buildings. However, global features are sensitive to cluttered background and occlusion due to their essential attributes. Therefore, their performance drops relatively in the case of object recognition in indoor environments where direct specification of the content in an image is required or when an object is not enclosed by the bounding box. Local features, on the other hand, rely on individual pixels or discrete regions. Typically, salient features of highly textured objects are extracted by feature detectors and represented by compendious feature descriptors. The representation of the content in an image is thus more robust to scale, viewpoint or illumination changes.

3.1. Global Features

Despite the limitation of global features, they are still useful in cases where a rough place or object classification is required. Global features consist of the statistics extracted from a whole image or a bounding box, such as contour, shape, texture, colour or a combination of them [23]. They generally represent an image with a single high-dimensional feature vector and, thus, can be easily applied with any standard classification methods [24]. Moreover, thanks to the compact representation and low computational cost, they have been employed by some semantic mapping systems in real time. Table 1 gives the differences and similarities of the global features used for object recognition in semantic mapping systems.

Table 1. Global features for object recognition in semantic mapping systems. DOT, dominant orientation template.

Feature	Classification	Performance
Haar-like [25]	Texture	Robust to illumination changes
Colour histograms [26]	Colour	Robust to viewing angles
HOG [27]	Template	Robust to illumination and shadowing changes, sensitive to object orientations
DOT [28]	Template	Computationally efficient, robust to small shifting and deformations
High dimensional composed receptive field histograms [29]	Combination	Robust to illumination and minor scene changes
GIST [30]	Combination	Robust to occlusion and viewpoint changes; noise in isolated regions is ignored

Inspired by the application of the Haar-like feature in human face detection [25], a small number of objects were first recognized as landmarks in [31]. The recognized objects were then applied as supplementaries to the geometrical features in order to distinguish rooms that had similar geometrical structure and could only be further identified by the objects found there. The Haar wavelets present the average intensity differences between regions and likewise can be used to compare the differences between the sum of pixel values in bounding boxes, which allows relatively high robustness to illumination changes.

Ulrich and Nourbakhshthus implemented colour histograms for place recognition by comparing query images with limited images of an entire dataset [32]. Applying colour histograms for image matching was first brought up by Swain and Ballard [26]. The number of colour histograms is based on the number of the colour bands used, e.g., red, green and blue. Each histogram is built by simply counting the number of pixels with a specific intensity in different colour bands. Such a feature is robust to viewing angle changes in the case when properties in the environment remain fixed. Furthermore, it provides a highly compact representation of an image and, thus, requires less memory space. However, one obvious drawback is that it fails to describe spatial relations, which limits its applicability. Filliat *et al.* also adopted this feature to discriminate identical chairs of different colours [33].

The spatial information, such as feature location, was not included by the holistic methods presented above due to the lack of a segmentation step. Some features have managed to divide an image into small discrete regions and then compute the global statistics within individual regions in

order to obtain some rough spatial information. Grimmett *et al.* used the histogram of oriented gradient descriptor (HOG) [27] to represent training data in order to detect a parking space [34]. HOG describes objects by concatenating histograms of the gradient directions computed from the pixels within individual regions divided from an image, called cells. Each histogram is then contrast-normalized across a group of cells, called a block, to decrease the susceptibility to illumination or shadowing changes, except for object orientations. Such a feature has a high performance for pedestrian detection if they maintain a roughly upright position.

In the case of texture-less object detection, Hinterstoisser *et al.* proposed a feature named dominant orientation templates (DOT) based on local region management [28], which was applied in [35] to match the objects in a predefined database. Compared to the HOG feature, DOT relies on dominant orientations rather than histograms. It only considers the orientations of gradients with the biggest magnitude values, thus providing invariance to small shifting and deformations. DOT is also computational time saving for object matching thanks to the ignorance of gradient values and the implementation of orientations rather than directions.

Global features have also been combined in some systems to provide richer representations of the environment. A high dimensional composed receptive field histogram was applied in [29], which consists of normalized Gaussian derivatives, differential invariants and chromatic cues. Siagian *et al.* attempted to incorporate context using the GISTdescriptor [30] for scene classification [36,37]. Orientation, colour and intensity channels are employed by GIST to filter input images with Gabor filters at multiple spatial scales to extract the gistof images. Hinterstoisser *et al.* presented another 3D feature as a complement to the DOT feature, named LINE-MOD, by computing the object surface normal with a depth sensor [38]. These methods tend to be relatively more robust than using a single global feature, since the random noise produced by individual features can be averaged out.

3.2. Local Features

Local features that are widely used in semantic mapping systems for object and place recognition can be further divided into three categories: edge-, corner- and blob-based approaches [39]. Figure 9 shows the definition of local visual features in computer vision. An edge is a set of pixels with strong gradient magnitudes or located where the image intensities change sharply. This normally refers to the boundaries between distinguishable regions. A corner is a pixel at which two edges intersect or has edges with two or more directions in the neighbourhood. The term corner is additionally used in some cases, which differ from our common sense, e.g., a small white spot (corner) on a black background, since apart from relying on explicit edge detection, a corner can also be computed from the curvature in the image gradient. A blob is a group of connected pixels with similar characteristics. It refers to an interest point, as well, because many interest point detection methods are essentially based on corner detection at multiple scales.

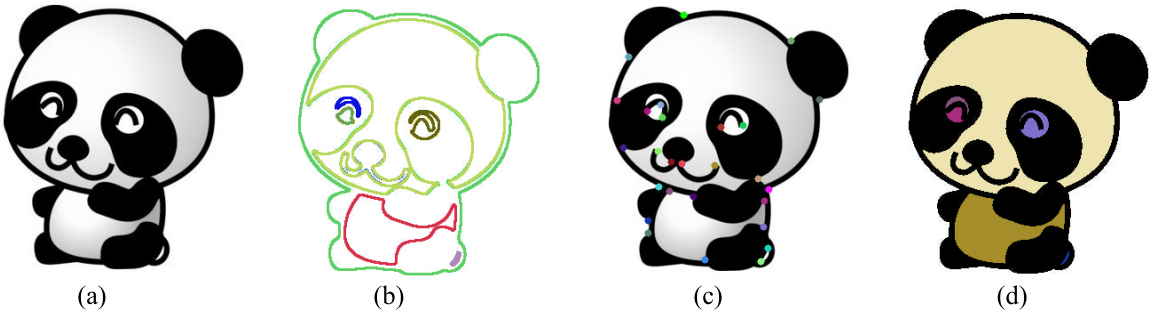

Figure 9. Definition of local visual features in computer vision. (**a**) Original image; (**b**) Edge; (**c**) Corner; (**d**) Blob.

In this section, the local features are presented accordingly. Table 2 gives the differences and similarities of the local features used for object recognition in semantic mapping systems.

Table 2. Local features for object recognition in semantic mapping systems. KLT, Kanade–Lucas–Tomasi; FAST, features from accelerated segment test; AGAST, adaptive and generic accelerated segment test; BRIEF, binary robust independent elementary features; CenSurE, centre surround extrema; ORB, oriented FAST and rotated BRIEF; BRISK, binary robust invariant scalable keypoints; FREAK, fast retina keypoint; MSER, maximally-stable extremal region.

Category	Classification	Feature	Performance
Edge based	Differentiation based	Sobel	Computationally efficient, high error rate
		Canny [40]	High accuracy, computationally expensive
Corner based	Gradient based	Harris [41]	Sensitive to noise, computationally expensive
		KLT [42]	Computationally efficient, sensitive to noise
Corner based	Template based	FAST [43]	Computationally efficient, low level of generality
		AGAST [44]	High level of generality, computationally efficient
		BRIEF [45]	Computationally efficient, sensitive to viewpoint rotations
Blob based (keypoint)	PDE based	SIFT [46]	Robust to scale and transformation changes, computationally expensive
		SURF [47]	Robust to scale and transformation changes, computationally efficient
		CenSurE [48]	High accuracy, computationally efficient
Blob based (keypoint)	Template based	ORB [49]	Computationally efficient, robust to viewpoint rotations
		BRISK [50]	Robust to scale changes, computationally efficient
		FREAK [51]	Computationally efficient, robust to scale changes
Blob based (region)	Intensity based	MSER [52]	Robust to affine transformations, computationally efficient

3.2.1. Edge-Based Approaches

The primary characteristic of edges in an image is a sharp change, which is commonly used and captured by classical differentiation-based edge detectors. Currently, such edge detectors are only used to generate fundamental cues to construct more robust features or provide complementary information for semantic mapping systems [53]. Ranganathan and Dellaert [54] converted each training image into a set of regions of interest with the Canny edge detector [40]. Clustered edges were obtained to facilitate modelling texture-less objects, like desks. The Canny edge detector sets three general criteria for edge detection: low error rate, precise localization on the centre of edges, and a given edge in an image should only be marked once. Owing to these criteria, it is one of the most strictly-defined methods that provides robust and reliable edge detection. Wu *et al.* attempted to filter each input image with a Sobel operator beforehand, since they were interested in the spatial structure property of an image rather than detailed textural information [12].

In recent years, edge detection in computer vision has been extended to a broader concept, which is quite similar to object segmentation, named boundary detection. Boundary detection considers an object as a whole. It suppresses the internal edges extracted from the textures within objects and only presents the edges between objects and background. Multiple low-level features are combined to detect boundaries based on machine learning algorithms. However, simply using 2D images tends to be computationally more expensive or less reliable than applying an additional depth channel for them, since it is relatively straightforward to obtain object boundaries from a 3D image. Thus, boundary detection using only 2D images is rarely implemented in semantic mapping.

3.2.2. Corner-Based Approaches

Primitive corner-based approaches rely on gradient assessment, which is a theoretical concept abstracted from our common sense understanding for the term corner. In [54,55], a Harris corner detector [41] was used to facilitate training the database. It computes the differential of autocorrelation according to directions directly. A similar detector, named Kanade–Lucas–Tomasi (KLT) [42], was employed in [56] for efficient and continuous tracking. Compared to the Harris detector, KLT has an

additional greedy corner selection criterion; thus, it is computationally more efficient. However, these corner detectors are not reliable in all circumstances during semantic mapping, since the gradient assessment method is highly sensitive to noise.

In order to decrease the complexity of gradient assessment and increase computational efficiency, some methods based on a template have been implemented. Such features extract corners by comparing the intensity of a pixel with other pixels in the local neighbourhood, *i.e.*, a predefined template. Henry *et al.* [57] and Gálvez-López *et al.* [58] attempted to apply features from accelerated segment test (FAST) [43] for indoor mapping and loop closure, respectively. Based on machine learning algorithms, FAST yields a large speed increase, thus being widely employed by real-time systems. FAST uses a circular template of 16 pixels to evaluate whether a candidate pixel is actually a corner. The candidate pixel is classified as a corner in cases when a certain number of contiguous pixels in the circle are all brighter than the intensity of the candidate pixel plus a threshold value or all darker than the intensity of the candidate pixel minus a threshold value. During the high-speed test for rejecting non-corner points, a decision tree is applied to address the correct rules of the chosen detector. However, FAST suffers from a low level of generality, since it has to be trained for specific scenes before applied. FAST-ER [59] and adaptive and generic accelerated segment test (AGAST) [44] increase the performance of FAST in terms of repeatability and generality, by widening the thickness of the Bresenham's circle and training a set of decision trees rather than relying on one tree, respectively.

Due to the high demand for real-time applications, Gálvez-López and Tardós adopted binary robust independent elementary features (BRIEF) [45] to find the best frame-to-frame matches for real-time localization over long periods [58]. BRIEF is a binary string constructed by classifying image patches according to pairwise intensity comparisons, which leads to small memory usage and is highly computational efficient during recognition.

3.2.3. Blob-Based Approaches

Blob-based approaches rely on identifying the unique regions in an image by comparing local properties (e.g., intensity and colour) to their neighbouring regions. In a blob, specific properties of all of the points remain constant or approximately constant, *i.e.*, to some extent, the points are similar to each other. Blob-based approaches can be further divided into two categories: keypoint- and interest region-based approaches. Keypoint-based approaches are focused on finding local extrema in scale spaces, whereas interest region-based approaches aim at segmenting regions. A scale space is a representation of gradually-smoothed images obtained by the rules that can describe the basic properties of interest. The scale space presents an image with an additional third dimension. Note that a corner can also be regarded as a keypoint at a specific scale.

Classical interest point-based methods are based on partial differential equations (PDE), among which the Laplacian of Gaussian (LoG) is one of the most widely-used methods. An input image is first convolved by a Gaussian function at a certain scale to represent the image in a Gaussian scale space. The Laplace operator is then applied to obtain strong responses for bright and dark blobs. Compared to LoG, the difference of Gaussians (DoG) computes the Laplacian of Gaussian operator by the difference between two continuous images smoothed by a Gaussian function. DoG can also be viewed as an approximation of the Laplacian operator, thus being computationally more efficient. A hybrid blob detector Hessian–Laplacian combining the Laplacian with the determinant of the Hessian (DoH) blob detectors has also been proposed, where spatial selection is done by the determinant of the Hessian matrix and scale selection is performed with the scale-normalised Laplacian.

Based on DoG and the Hessian matrix, Lowe proposed scale invariant feature transform (SIFT) [46], which has been widely applied by robot SLAM and object recognition systems [3,60–62]. The original image is convolved with DoG to identify potential interest points that are invariant to scale and orientation changes. The points selected from the training image, which usually lie on high-contrast regions of images, such as edges and corners, are detectable even under changes in image scale, noise and illumination. Another property of these points is that the relative positions between

them in the original image remain stable from one image to another. Subsequently, low contrast and unstable points are rejected based on their locations and scales. Orientations are then assigned to the points based on gradient directions, thus providing invariance to transformations. Finally, SIFT computes a descriptor vector (histogram of oriented gradient) as a representation for each keypoint. Compared to other feature descriptors, SIFT is highly robust to scale and transformation changes, but is computationally expensive. A refinement of SIFT was proposed by Mikolajczyk and Schmid, named gradient location and orientation histogram (GLOH) [63], which proves to be more distinctive than SIFT, yet requires even more computational cost.

Riazuelo *et al.* initially extracted speeded up robust features (SURF) from each image and stored them for later object recognition in the RoboEarth database, which is a knowledge-based system providing web and cloud services [15]. SURF has claimed to be several times faster than SIFT, and the accuracy still remains relatively acceptable. SURF employs integral images and uses square-shaped filters to approximate the determinant of the Hessian matrix during Gaussian smoothing, thus being more computational efficient. Morisset *et al.* used centre surround extrema (CenSurE) [48] to obtain visual odometer in real time [64]. CenSurE is another approximation of LoG. Compared to SIFT and SURF, CenSurE features are evaluated for all of the pixels across all scales in raw images. This leads to higher accuracy. Moreover, even seeking extrema at all scales, it still maintains a relatively low computational cost by adopting a set of simple centre surround filters. Implementations of these refinements in semantic mapping systems can also be found in [33,65,66].

Inspired by FAST and BRIEF corner detector based on a template, Rublee *et al.* presented oriented FAST and rotated BRIEF (ORB) by estimating the patch orientation [49], thus being invariant to viewpoint rotations. The scale pyramid is also applied to increase its robustness to scale changes. Such a method was employed in [67] to generate the photometric feature for RGB-D mapping. Grimmett *et al.* used binary robust invariant scalable keypoints (BRISK) [50] to build maps for automated parking [68]. BRISK applies the AGAST detector in both the image plane and scale space to classify keypoints, so that it is invariant to scale changes. A keypoint detector motivated by human retina and derived from BRISK was presented by Alahi *et al.* [51], named fast retina keypoint (FREAK), which was applied in [69] for facial point detection and emotion recognition. Compared to BRISK, FREAK has a higher density of points near the centre of the sampling grid.

Meger *et al.* [60] and Sun *et al.* [70] applied the maximally-stable extremal region (MSER) [52] in their systems to provide object location information for an attentive system and to extract lane marking features, respectively. MSER is one of the most widely-used methods for interest region detection. It is robust to affine transformations and is highly computationally efficient. However, it is sensitive to image blur changes. Moreover, MSER is a region detector in essence, thus being only suitable to distinguish objects with little variation in colour from high contrast scenes.

3.2.4. Discussion

One of the most important factors to evaluate the feature detectors and descriptors implemented in semantic mapping systems is their accuracy (reliability). To assess it, a repeatability criterion presented by Schmid *et al.* measures whether or not the same feature is detected in two or more different images of the same scene under varying viewing conditions [71]. The repeatability is the ratio between the accurate pairing number and the minimum number of keypoints detected in the given images. The repeatability of some local features is shown in Figure 10 [22]. Three image transformations are considered: zoom factor, viewpoint and rotation.

With respect to some SLAM or object recognition systems running in real time, the computational complexity of the applied feature detectors also plays a key role. Figure 11 [22] shows the efficiency evaluation of some widely-used keypoint detectors in semantic mapping systems. The first graph assesses the average processing time based on the number of keypoints detected. FAST and AGAST are computationally more efficient than other detectors. Another noticeable difference is that the processing time of CenSurE and ORB remains constant, whereas with the increase of the keypoint number, the

processing time of other detectors grows linearly. The second graph presents the influence of image scale changes on the detectors. The processing time for all of the detectors rises as a quadratic function with the increase of image scale. Again, SIFT and SURF are several times more time consuming.

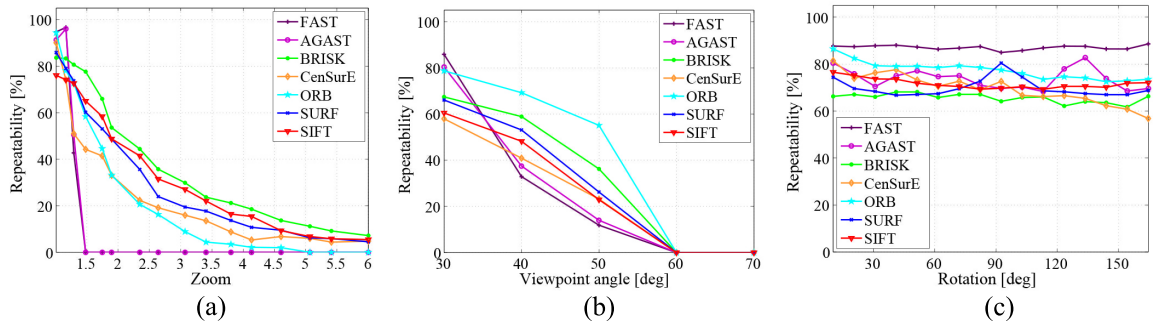

Figure 10. Repeatability based on image: (**a**) Zoom; (**b**) Viewpoint angle; and (**c**) Rotation transformations [22]. Reproduced with permission from Dr. Canclini.

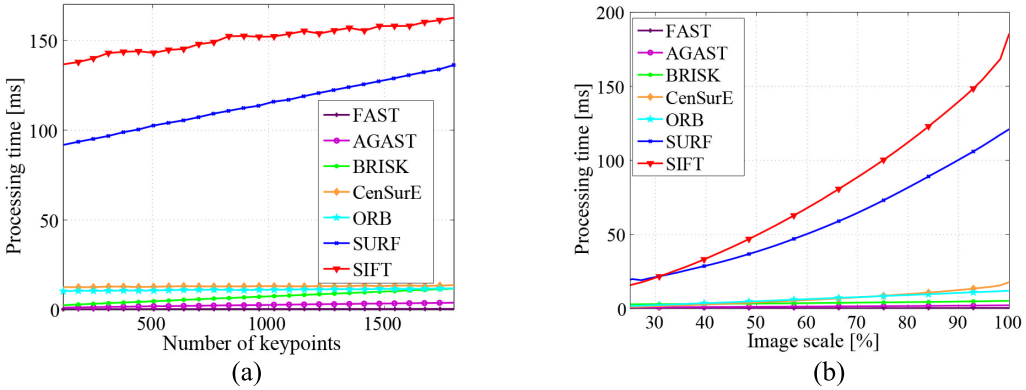

Figure 11. Processing time evaluation based on: (**a**) The number of keypoints; and (**b**) Image scale changes [22]. Reproduced with permission from Dr. Canclini.

4. Recognition Approaches

This section presents some basic object/place recognition approaches in semantic mapping systems, namely global, local and hybrid approaches [72]. Object recognition methods based on global features are classified into global approaches. Such approaches also consist of some place recognition methods that employ image segmentation algorithms directly, rather than referring to the properties in the environment. Local approaches include pixel-wise operations on the basis of local feature extraction and the straightforward sampling of pixel contributions. Some systems combine global approaches with local approaches to achieve a more robust performance, which is discussed in hybrid approaches. In addition, information that is retrieved to distinguish individual instances within an object class (e.g., shampoo or conditioner, someone's office) is also discussed.

4.1. Global Approaches

Based on the global statistic feature retrieved from texture, the hidden Markov model (HMM) was applied in [73] for place recognition and new place categorisation. For HMM, the states that represent different locations are not directly visible, whereas the output acquired from the states is visible. Compared to using a uniform transition matrix, HMM provides a significant increase in recognition performance. Furthermore, the computational cost is quite low and can be neglected during mapping. However, it is only applicable for a small database. Mozos *et al.* implemented a cascade of classifiers [74], which depended on boosting to detect eight different objects in order to

recognize six places [31]. Boosting is a supervised learning-based method combing several simple weak classifiers to achieve a relatively high performance. For each of the weak classifiers used, the requirement is that its accuracy should be better than random guessing. The accuracy of the weak classifiers leads to their distributions once they are added. The cascade of classifiers is essentially a degenerated decision tree that rejects non-object regions at each stage and retains interest regions for further classification. Although the training time is long, the prediction can be run in real time. Gentle AdaBoost was also applied by Murphy *et al.* for object and scene recognition [30].

In the case of colour histogram features, Ulrich and Nourbakhsh used a simple unanimous voting scheme to classify places [32]. The input images were voted by each colour band with the smallest minimum matching distance. A certain place was classified when the colour bands unanimously voted for the same place and the total confidence was above a threshold. Such a method is quite straightforward and computationally efficient. However, one important prerequisite is that the visible properties in scenes should remain relatively fixed, and its performance drops when it comes to a large database (over 100 images).

A support vector machine(SVM) [75] was applied with the HOG feature by Dalal and Triggs [27]. SVM is a set of supervised models with associated learning algorithms widely used for data analysis and pattern recognition. The training process tries to build a model by the given examples to assign new examples into two categories, making it a non-probabilistic binary linear classifier. This step is essentially a process to find a model with high performance, *i.e.*, a clear gap that is as wide as possible. In this paper, positive examples (images that contain pedestrians) and negative examples (person-free images) were provided for SVM training. The implementation of linear SVM rather than using the kernel decreased the computational cost of the system. Pronobis *et al.* also applied SVM to recognize places [29], yet based on the kernel [76], which proved to achieve better performance for histogram-like features. Results in this paper showed that the places were recognized with high precision and robustness, even when training on images from one camera device and testing on another. Inspired by Taylor and Drummond [77], the streaming SIMDextension (SSE) was applied to efficiently compute error functions [28].

4.2. Local Approaches

Apart from the approaches mentioned above, one of the most promising works has been done by Lowe [46]. Once the SIFT features are detected and described, recognizing objects becomes a problem of finding groups of similar descriptions that have all undergone the same transformation. More specifically, an interest point in the test image is compared to an interest point in the reference image by the differences between their description vectors, which is based on Euclidean distance. For rapid computation against large databases, the features are put in a KD-tree, which is a data structure based on nearest neighbour searching for large databases. The Hough transform is used to cluster the features that belong to the same object. Clusters of at least three features that agree on the object and its pose are identified as candidate matches. A least-square approximation is then made to obtain the best estimated affine projection parameters, which are further applied to decide whether to keep or reject the matches. This method has been widely implemented in robot semantic mapping systems [3,60,61] thanks to its high robustness. However, due to the complexity of the SIFT feature, the recognition process still suffer from high computational cost.

In the case of the binary features inspired by modern computer architecture, such as BRIEF, BRISK, ORB and FREAK, the Hamming distance is used for matching. The Hamming distance between two feature vectors is the number of positions at which the corresponding symbols are different. Such a matching method is highly computationally efficient. However, the accuracy is lower than the method presented by Lowe.

4.3. Hybrid Approaches

Some systems adopted global features, local features and depth information to generate a more robust recognition performance. Depth information additionally provides spatial dimensions of objects and represents objects in more detail, thus leading to a higher recognition accuracy compared to using solely 2D features. Histograms of oriented energy and colour were directly applied for object detection in [78]. Stückler *et al.* employed region features in both colour and depth space and applied object-class segmentation algorithms for semantic mapping [79], based on random decision forests (RFs), which is an ensemble learning method for classification and has been demonstrated to achieve comparable performance to SVM [80]. In this work, a subset of images from the training set was randomly selected as a sample to train the decision trees. Small objects were better sampled for training; thus, the actual individual distributions of class labels were reassigned according to this. One advantage of RFs is the high computational efficiency during outputting, yet the training time is still relatively long.

Filliat [81] and Martínez-Gómez *et al.* [82] employed a bag of binary words (BoW) model [83] to incrementally learn to recognize different rooms from any robot position. The model is inspired by the technique in document classification. BoW consists of two phases, representation (indexing) and recognition (retrieval). Image features that are robust to intensity, rotation, scale and affine are detected and described by independent feature descriptors with vectors, such as SURF and FAST (SIFT, colour histograms and normalised grey level histogram in this paper). Subsequently, the vectors are clustered by vector quantisation algorithms, e.g., K-means clustering [84]. The predefined codewords (words in documents) are then assigned to the clusters to generate a codebook (a word dictionary); thus, the images are represented by a histogram of codewords. In the case of the recognition stage, generative or discriminative models, such as the naive Bayes classifier, SVM and AdaBoost, are applied as the classifiers. Such a method is quite flexible in terms of both applicable features and recognition approaches. However, the spatial relationships among the clusters are ignored when BoW is used alone, which has been compensated by Lazebnik *et al.* [85].

Text or signs can provide location information directly. Most text recognition systems implemented optical character recognition (OCR) for classification [86–88], which is an off-the-shelf technology to convert images of typed, handwritten or printed text into machine-encoded text. Sami *et al.* [87] adopted a back-projection of the colour histogram to locate interest regions and applied the Canny edge detector to remove background. A pan/tilt/zoom camera was used in [88] to provide better focusing performance on potential text regions in the wild. However, text retrieval still suffers from low performance in cluttered environments, which limits its practicability.

5. Semantic Representation

Semantic representation is the interpretation process from objects or places to a human-compatible or even human-like language. Some systems presented above apply classification or segmentation methods for the purpose of recognizing specific objects or scenes; thus, the semantic information is directly obtained. In this section, we mainly focus on the semantic information inferred by robots.

Early systems [3,21] adopted the idea that a semantic map consisted of two separate, but tightly interconnected parts: a spatial part and a terminological part [89]. This is a typical structure of hybrid knowledge representation (KR) systems [90], as shown in Figure 12 (redrawn according to [21]). The spatial part contains raw images from the sensors, geometric information of the environment and connectivity between the rooms, whereas the terminological part consists of general semantic knowledge about the environment, giving meanings to the features of the corresponding properties in the environment in terms of general concepts and relations. These two hierarchies are interrelated by the concept of anchoring [91]. In [21], NeoClassicAI language was employed to establish the conceptual hierarchy and provided the robot with inference capability. However, the conceptual knowledge was hand-coded into the system, and uncertainties about the properties in the environment were not included in the representation.

Vasudevan and Siegwart attempted to classify places based on objects, as well [92], but the system was fully probabilistic. In their system, objects were grouped into predefined clusters and conceptualised during the training process. A simple naive Bayesian classifier (NBC) was then employed to infer and identify the place categories on the basis of the clusters.

Meger *et al.* [60] developed an attentive system projecting the location and semantic information of the recognized objects back into the grid map. Since the object recognition subsystem were trained by collecting object model data through submitting text-based queries to Internet image search engines, the semantic information was thus easily incorporated into the object models. Once an object was observed, the semantic information was directly acquired. A similar work [93] built the spatial-semantic object models based on the LabelMe database [94]. Bayes' theorem was applied to define place categories by the recognized objects. More specifically, a cluster model grouped objects on the basis of their locations; thus, the place categories were just another representation of the clusters.

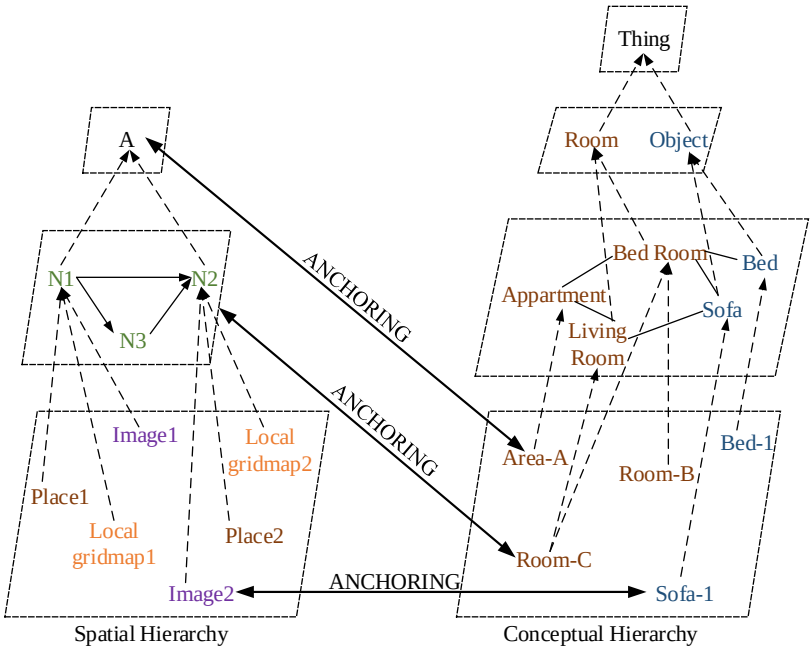

Figure 12. The spatial and conceptual hierarchy interrelated by anchoring [21]. Redrawn with permission from Prof. Galindo.

Zender *et al.* [61] and Capobianco *et al.* [95] encoded conceptual knowledge into a Web Ontology Language-description logic (OWL-DL) ontology. In [61], a description-logic reasoner employed some situated dialogues between a robot and a user to provide new knowledge for the robot to further infer. In addition, a laser sensor was implemented for place classification. More specifically, a navigation node (a marker) was placed in the metric map after the robot moved one metre away from the last node. The nodes were then connected and classified for room type identification. Pronobis *et al.* extended such a method and applied the doors detected in indoor environments to bound areas [96]. The final map is shown in Figure 13. In [95], a standard methodology for representing and evaluating semantic maps was proposed. The formalisation consisted of a reference frame, spatial information and a set of logic predicates. With this system structure, the performance of semantic representations can then be compared against those of other systems.

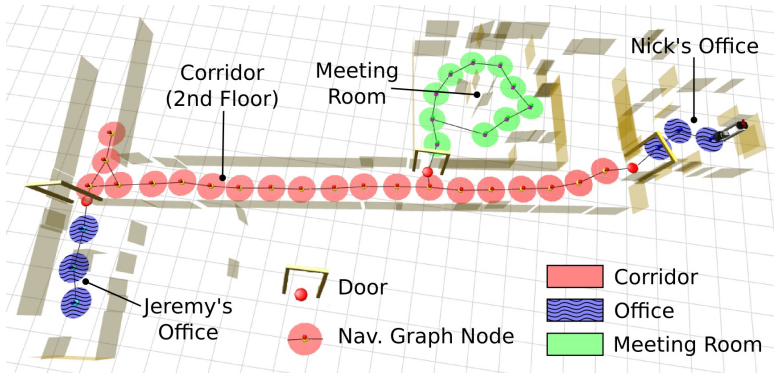

Figure 13. The final semantic map obtained in [96]. Reproduced with permission from Dr. Pronobis.

6. Typical Applications

Service robots are gradually working their way into our daily lives to become household servants, healthcare systems and even cognitive companions. The primary responsibility of service robots is to obey the orders given by humans and perform tasks with high efficiency and accuracy. A semantic map provides a friendly human-robot interface and enables these service robots to be used by the general public without the need for training.

6.1. Indoor Applications

Some applications can be found in indoor environments. Galindo *et al.* presented a typical autonomous navigation method based on a pre-built semantic map [21]. In their experiment, the robot was given a command "go to the bathroom". Following this command, the robot inference system found a node in the topological map, and the spatial information in the metric map that connected to the node was retrieved by anchoring. Thus, the command was translated to "go to the node" and then executed between the topological and metric hierarchies. Figure 14 describes this navigation method.

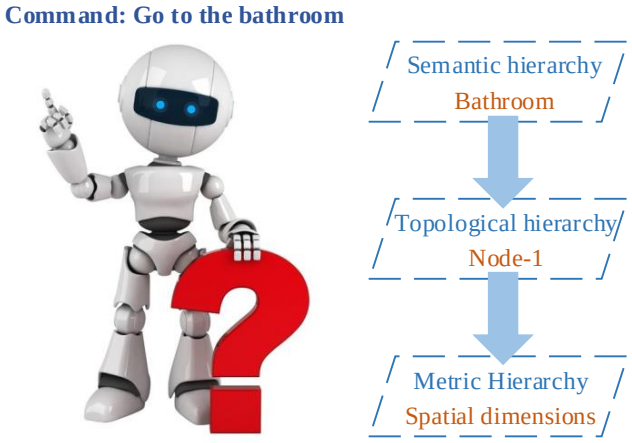

Figure 14. Robot navigation based on a semantic map.

The authors in [97] enhanced the inference capability of a robot with a semantic map. The common knowledge known by almost everyone was applied to detect deviations from normal conditions. A goal was then autonomously generated by the encoded information about how things should be, e.g., if a bottle of milk was observed on a table, the robot would set a goal by itself and bring the milk into a fridge. Crespo *et al.* also presented a reasoning module to infer new knowledge for mobile robots [98]. A relational database used for storing objects and perceiving information was implemented to provide inference capability based on objects' links.

Blodow *et al.* extracted semantic information by a laser scanner and a high-resolution 2D camera for a PR2 robot to perform tasks in kitchen environments [99]. Segmentation algorithms were applied to identify and distinguish certain kitchen facilities and their functional components. The robot was capable of analysing the task-relevant objects, locating them in the map and acting on them, e.g., using the handle to open the drawer and close it, as shown in Figure 15.

A semantic map can be inversely utilized to further improve robot localization capabilities. One basic method is to identify room categories by specific objects. For example, the room is classified as a kitchen once an oven is found inside. Initial localization errors can be reduced by reasoning about the expected location of the recognized objects [21]. Ko *et al.* extended such a method by continuously searching for other memorized objects as landmarks before referring to spatial relationships, since one object may not be enough to infer the accurate location [100]. In addition, the times for searching regions in topological and metric space were largely reduced by discarding the irrelevant areas. The computational cost in the initial stage during robot localization was thus minimized [89].

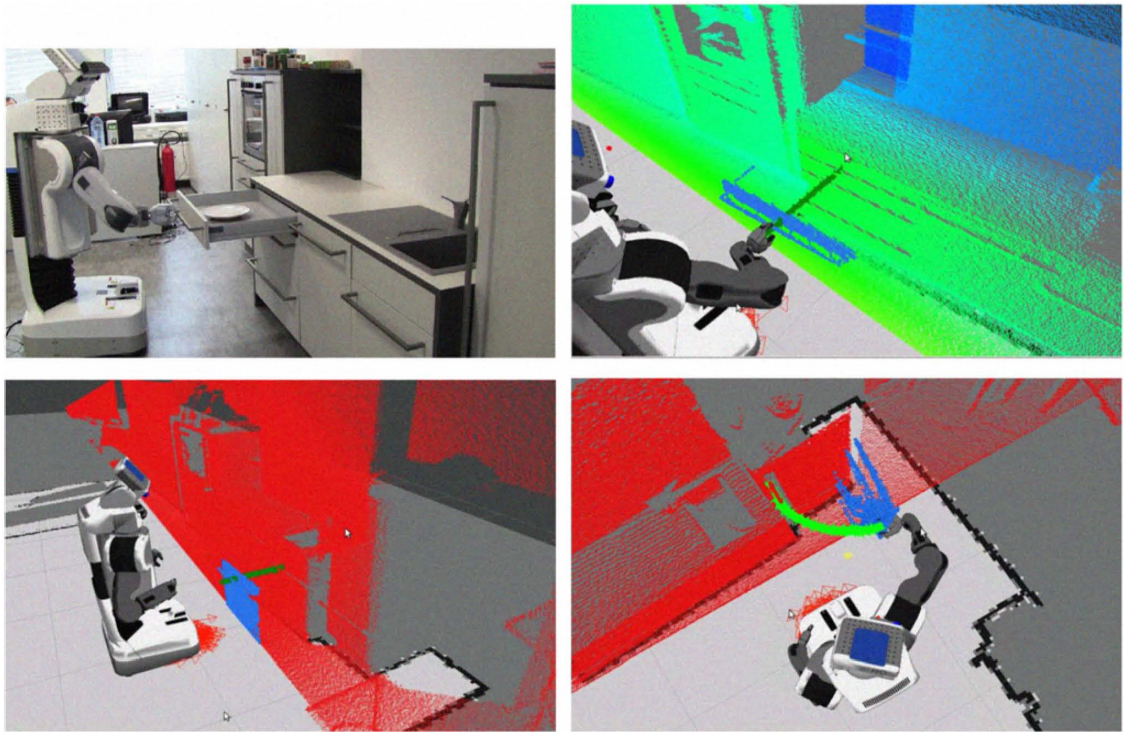

Figure 15. Interaction between the robot and the kitchen facilities [99]. Reproduced with permission from Dr. Pangercic.

6.2. Outdoor Applications

Semantic maps can also be applied to outdoor environments. Wolf and Sukhatme analysed and classified terrain to resolve issues related to non-navigable areas during path planning [2]. Boularias *et al.* additionally adopted natural language to command a mobile robot for navigation in outdoor environments, e.g., "Navigate to the building behind the pole" [101]. Bernuy and Solar presented a graph-based topological semantic mapping method for autonomous off-road driving [102].

Other applications include self-driving cars and augmented reality (AR). Reasoning about environments with additional semantic information plays a key role for self-driving cars, since the cars should be able to recognize roads, pedestrian crossings, humans, traffic lights, *etc*. However, current self-driving cars still have difficulty in identifying some kinds of objects, such as trash, which is harmless, but causes the vehicles to veer unnecessarily. When a police officer signals them to stop, they may not react accordingly. These problems can be solved by associating the metric map with semantic

information. With respect to AR, the augmentation should be conventional in semantic context with environmental elements, e.g., directing the way by virtual paths. Semantic maps are necessary for identifying objects and destinations in real environments.

7. Conclusion and Discussion

This paper has presented a survey on current semantic mapping system architectures and the state-of-the-art algorithms. The visual sensor-based approaches to semantic mapping were first introduced, and the features extracted from images were then detailed. Subsequently, the recognition and classification methods based on the extracted features were discussed, as well as the direct segmentation methods. Lastly, the semantic representation strategies and typical applications were presented.

Although semantic mapping has made significant progress, some challenging problems remain and can be summarized as follows.

- A 3D visual sensor with high accuracy and resolution is needed for abstracting semantic information, since a 2D camera is not adequate for the recognition of a wider range of objects and places. Furthermore, the blurred images produced by low accuracy and resolution 3D cameras cannot be used for a robot to recognize objects in the real world. Apart from the improvement of hardware, several software methodologies, such as super-resolution or data fusion, could be deployed for accurate semantic mapping. Super-resolution is a class of techniques that enhances the resolution of an imaging system. Data fusion is a process of integrating multiple data and knowledge representing the same scene in order to produce an accurate output.
- The current feature detectors and descriptors need to be extended in semantic mapping systems. Although local visual features are robust to geometric transformations and illumination changes, the features are quite limited to the appearance of objects, such as edges, corners and their relations. Extracting the semantic inherent characteristic of objects might be a solution, e.g., the legs of a chair, the keyboard and display of a laptop, *etc*. Another solution is to abstract the text information that can normally be found on the packaging of products. The text is quite easy to distinguish between a bottle of shampoo and conditioner by their labels.
- Classifiers should be adaptive to the dynamic changes in the real world. The current semantic mapping systems need pre-training and can only recognize the trained objects or certain scenes. However, the real-world environments are changing dynamically, and object appearances are changing all the time. Any semantic mapping algorithms need the ability of self-learning to adapt these changes and recognize new objects. Solutions might be found in deep learning algorithms.
- Semantic mapping systems should be able to detect novelty and learn novel concepts about the environment continuously and in real time. The conceptual definitions that are initially encoded by using common sense knowledge should only be used for bootstrapping, which should be updated or extended based on new experience. For instance, a robot operating in a home environment should link its actions to rooms and objects in order to bridge the gap between the metric map and semantic knowledge. Moreover, this conceptual learning performance opens new possibilities in terms of truly autonomous semantic mapping and navigation. Thus, an incremental adaptive learning model should be built by using machine learning algorithms.

Acknowledgements: Qiang Liu and Ruihao Li are financially supported by China Scholarship Council and Essex University Scholarship for their PhD studies. We also thank the support of the EU COALAS project: Cognitive Assisted Living Ambient Systems (http://www.coalas-project.eu/), which has been selected in the context of the INTERREG IVA France (Channel) - England European Cross-border cooperation programme, and co-financed by the ERDF. Our thanks also go to Robin Dowling and Ian Dukes for their technical support.

Author Contributions: All authors have made substantial contributions to this survey. Huosheng Hu supervised the research. Qiang Liu analysed and organised the methods presented in this paper and wrote the survey. Dongbing Gu and Ruihao Li provided suggestions on structuring the survey. All authors discussed and commented on the manuscript at all stages.

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

References

1. Thrun, S. Robotic Mapping: A Survey. In *Exploring Artificial Intelligence in the New Millennium*; Morgan Kaufmann: San Francisco, CA, USA, 2002; pp. 1–35.
2. Wolf, D.F.; Sukhatme, G.S. Semantic Mapping Using Mobile Robots. *IEEE Trans. Robot.* **2008**, *24*, 245–258.
3. Vasudevan, S.; Gächter, S.; Nguyen, V.; Siegwart, R. Cognitive Maps for Mobile Robots—An Object Based Approach. *Robot. Auton. Syst.* **2007**, *55*, 359–371.
4. Nüchter, A.; Hertzberg, J. Towards Semantic Maps for Mobile Robots. *Robot. Auton. Syst.* **2008**, *56*, 915–926.
5. Kostavelis, I.; Gasteratos, A. Semantic Mapping for Mobile Robotics Tasks: A survey. *Robot. Auton. Syst.* **2015**, *66*, 86–103.
6. Harmandas, V.; Sanderson, M.; Dunlop, M. Image Retrieval by Hypertext Links. In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA, 27–31 July 1997; pp. 296–303.
7. Zhuge, H. Retrieve Images by Understanding Semantic Links and Clustering Image Fragments. *J. Syst. Softw.* **2004**, *73*, 455–466.
8. Flint, A.; Mei, C.; Murray, D.; Reid, I. A Dynamic Programming Approach to Reconstructing Building Interiors. In Proceedings of the 11th European Conference on Computer Vision (ECCV), Crete, Greece, 5–11 September 2010; pp. 394–407.
9. Coughlan, J.M.; Yuille, A.L. Manhattan World: Compass Direction from a Single Image by Bayesian Inference. In Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV), Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 941–947.
10. Anguelov, D.; Koller, D.; Parker, E.; Thrun, S. Detecting and Modeling Doors with Mobile Robots. In Proceedings of the 2004 IEEE International Conference on Robotics and Automation (ICRA), New Orleans, LA, USA, 26 April–1 May 2004; Volume 4, pp. 3777–3784.
11. Tian, Y.; Yang, X.; Arditi, A. Computer Vision-Based Door Detection for Accessibility of Unfamiliar Environments to Blind Persons. In Proceedings of the 12th International Conference on Computers Helping People with Special Needs (ICCHP), Vienna, Austria, 14–16 July 2010; pp. 263–270.
12. Wu, J.; Christensen, H.; Rehg, J.M. Visual Place Categorization: Problem, Dataset, and Algorithm. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), St. Louis, MO, USA, 11–15 October 2009; pp. 4763–4770.
13. Neves dos Santos, F.; Costa, P.; Moreira, A.P. A Visual Place Recognition Procedure with a Markov Chain Based Filter. In Proceedings of the 2014 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), Espinho, Portugal, 14–15 May 2014; pp. 333–338.
14. Civera, J.; Gálvez-López, D.; Riazuelo, L.; Tardós, J.D.; Montiel, J. Towards Semantic SLAM Using a Monocular Camera. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011; pp. 1277–1284.
15. Riazuelo, L.; Tenorth, M.; Di Marco, D.; Salas, M.; Gálvez-López, D.; Mosenlechner, L.; Kunze, L.; Beetz, M.; Tardos, J.D.; Montano, L.; *et al.* RoboEarth Semantic Mapping: A Cloud Enabled Knowledge-Based Approach. *IEEE Trans. Autom. Sci. Eng.* **2015**, *12*, 432–443.
16. Rusu, R.B. Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. *KI-Künstliche Intell.* **2010**, *24*, 345–348.
17. Sarbolandi, H.; Lefloch, D.; Kolb, A. Kinect Range Sensing: Structured-Light versus Time-of-Flight Kinect. *Comput. Vis. Image Underst.* **2015**, *139*, 1–20.
18. Jain, R.; Kasturi, R.; Schunck, B.G. *Machine Vision*; McGraw-Hill: New York, NY, USA, 1995.
19. Theobalt, C.; Bos, J.; Chapman, T.; Espinosa-Romero, A.; Fraser, M.; Hayes, G.; Klein, E.; Oka, T.; Reeve, R. Talking to Godot: Dialogue with a Mobile Robot. In Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), EPFL, Lausanne, Switzerland, 30 September–4 October 2002; Volume 2, pp. 1338–1343.
20. Skubic, M.; Perzanowski, D.; Blisard, S.; Schultz, A.; Adams, W.; Bugajska, M.; Brock, D. Spatial Language for Human-Robot Dialogs. *IEEE Trans. Syst. Man Cybern.* **2004**, *34*, 154–167.

21. Galindo, C.; Saffiotti, A.; Coradeschi, S.; Buschka, P.; Fernández-Madrigal, J.A.; Gonzalez, J. Multi-Hierarchical Semantic Maps for Mobile Robotics. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Edmonton, AB, Canada, 2–6 August 2005; pp. 2278–2283.
22. Canclini, A.; Cesana, M.; Redondi, A.; Tagliasacchi, M.; Ascenso, J.; Cilla, R. Evaluation of Low-Complexity Visual Feature Detectors and Descriptors. In Proceedings of the 2013 18th IEEE International Conference on Digital Signal Processing (DSP), Fira, Santorini, Greece, 1–3 July 2013; pp. 1–7.
23. Siagian, C.; Itti, L. Biologically Inspired Mobile Robot Vision Localization. *IEEE Trans. Robot.* **2009**, *25*, 861–873.
24. Lisin, D.A.; Mattar, M.; Blaschko, M.B.; Learned-Miller, E.G.; Benfield, M.C. Combining Local and Global Image Features for Object Class Recognition. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 47–47.
25. Papageorgiou, C.P.; Oren, M.; Poggio, T. A General Framework for Object Detection. In Proceedings of the 6th IEEE International Conference on Computer Vision (ICCV), Bombay, India, 7 January 1998; pp. 555–562.
26. Swain, M.J.; Ballard, D.H. Color Indexing. *Int. J. Comput. Vis. (IJCV)* **1991**, *7*, 11–32.
27. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
28. Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Fua, P.; Navab, N. Dominant Orientation Templates for Real-Time Detection of Texture-Less Objects. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2257–2264.
29. Pronobis, A.; Caputo, B.; Jensfelt, P.; Christensen, H.I. A Discriminative Approach to Robust Visual Place Recognition. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, China, 9–15 October 2006; pp. 3829–3836.
30. Murphy, K.; Torralba, A.; Freeman, W. Using the Forest to See the Trees: A Graphical Model Relating Features, Objects and Scenes. *Adv. Neural Inf. Process. Syst. (NIPS)* **2003**, *16*, 1499–1506.
31. Mozos, O.M.; Triebel, R.; Jensfelt, P.; Rottmann, A.; Burgard, W. Supervised Semantic Labeling of Places Using Information Extracted from Sensor Data. *Robot. Auton. Syst.* **2007**, *55*, 391–402.
32. Ulrich, I.; Nourbakhsh, I. Appearance-Based Place Recognition for Topological Localization. In Proceedings of the 2000 IEEE International Conference on Robotics and Automation (ICRA), San Francisco, CA, USA, 24–28 April 2000; Volume 2, pp. 1023–1029.
33. Filliat, D.; Battesti, E.; Bazeille, S.; Duceux, G.; Gepperth, A.; Harrath, L.; Jebari, I.; Pereira, R.; Tapus, A.; Meyer, C.; *et al.* RGBD Object Recognition and Visual Texture Classification for Indoor Semantic Mapping. In Proceedings of the 2012 IEEE International Conference on Technologies for Practical Robot Applications (TePRA), Woburn, MA, USA, 23–24 April 2012; pp. 127–132.
34. Grimmett, H.; Buerki, M.; Paz, L.; Pinies, P.; Furgale, P.; Posner, I.; Newman, P. Integrating Metric and Semantic Maps for Vision-Only Automated Parking. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 2159–2166.
35. Matignon, L.; Jeanpierre, L.; Mouaddib, A.I. Decentralized Multi-Robot Planning to Explore and Perceive. *Acta Polytech.* **2015**, *55*, 169–176.
36. Siagian, C.; Itti, L. Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 300–312.
37. Li, K.; Meng, M.Q.H. Indoor Scene Recognition via Probabilistic Semantic Map. In Proceedings of the 2012 IEEE International Conference on Automation and Logistics (ICAL), Zhengzhou, China, 15–17 August 2012; pp. 352–357.
38. Hinterstoisser, S.; Holzer, S.; Cagniart, C.; Ilic, S.; Konolige, K.; Navab, N.; Lepetit, V. Multimodal Templates for Real-Time Detection of Texture-Less Objects in Heavily Cluttered Scenes. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 858–865.
39. Li, Y.; Wang, S.; Tian, Q.; Ding, X. A Survey of Recent Advances in Visual Feature Detection. *Neurocomputing* **2015**, *149*, 736–751.
40. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *8*, 679–698.

41. Harris, C.; Stephens, M. A Combined Corner and Edge Detector. In Proceedings of the 4th Alvey Vision Conference, Manchester, UK, 31 August–2 September 1988; Volume 15, pp. 50.
42. Tomasi, C.; Kanade, T. Detection and Tracking of Point Features. *Int. J. Comput. Vis. (IJCV)* **1991**, *9*, 137–154.
43. Rosten, E.; Drummond, T. Machine Learning for High-Speed Corner Detection. In Proceedings of the 9th European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006; pp. 430–443.
44. Mair, E.; Hager, G.D.; Burschka, D.; Suppa, M.; Hirzinger, G. Adaptive and Generic Corner Detection Based on the Accelerated Segment Test. In Proceedings of the 11th European Conference on Computer Vision (ECCV), Crete, Greece, 5–11 September 2010; pp. 183–196.
45. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. BRIEF: Binary Robust Independent Elementary Features. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 778–792.
46. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis. (IJCV)* **2004**, *60*, 91–110.
47. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded Up Robust Features. In Proceedings of the 9th European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006; pp. 404–417.
48. Agrawal, M.; Konolige, K.; Blas, M.R. CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. In Proceedings of the 10th European Conference on Computer Vision (ECCV), Marseille, France, 12–18 October 2008; pp. 102–115.
49. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An Efficient Alternative to SIFT or SURF. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
50. Leutenegger, S.; Chli, M.; Siegwart, R.Y. BRISK: Binary Robust Invariant Scalable Keypoints. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2548–2555.
51. Alahi, A.; Ortiz, R.; Vandergheynst, P. FREAK: Fast Retina Keypoint. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 510–517.
52. Matas, J.; Chum, O.; Urban, M.; Pajdla, T. Robust Wide-Baseline Stereo from Maximally Stable Extremal Regions. *Image Vis. Comput.* **2004**, *22*, 761–767.
53. Basu, M. Gaussian-Based Edge-Detection Methods—A Survey. *IEEE Trans. Syst. Man Cybern.* **2002**, *32*, 252–260.
54. Ranganathan, A.; Dellaert, F. Semantic Modeling of Places Using Objects. In Proceedings of the 2007 Robotics: Science and Systems Conference, Atlanta, GA, USA, 27–30 June 2007; Volume 3, pp. 27–30.
55. Kim, S.; Shim, M.S. Biologically Motivated Novel Localization Paradigm by High-Level Multiple Object Recognition in Panoramic Images. *Sci. World J.* **2015**, *2015*, doi:10.1155/2015/465290.
56. Tsai, G.; Kuipers, B. Dynamic Visual Understanding of the Local Environment for an Indoor Navigating Robot. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vilamoura, Algarve, Portugal, 7–12 October 2012; pp. 4695–4701.
57. Henry, P.; Krainin, M.; Herbst, E.; Ren, X.; Fox, D. RGB-D Mapping: Using Kinect-Style Depth Cameras for Dense 3D Modeling of Indoor Environments. *Int. J. Robot. Res. (IJRR)* **2012**, *31*, 647–663.
58. Gálvez-López, D.; Tardós, J.D. Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE Trans. Robot.* **2012**, *28*, 1188–1197.
59. Rosten, E.; Porter, R.; Drummond, T. Faster and Better: A Machine Learning Approach to Corner Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 105–119.
60. Meger, D.; Forssén, P.E.; Lai, K.; Helmer, S.; McCann, S.; Southey, T.; Baumann, M.; Little, J.J.; Lowe, D.G. Curious George: An Attentive Semantic Robot. *Robot. Auton. Syst.* **2008**, *56*, 503–511.
61. Zender, H.; Mozos, O.M.; Jensfelt, P.; Kruijff, G.J.; Burgard, W. Conceptual Spatial Representations for Indoor Mobile Robots. *Robot. Auton. Syst.* **2008**, *56*, 493–502.
62. Kostavelis, I.; Charalampous, K.; Gasteratos, A.; Tsotsos, J.K. Robot Navigation via Spatial and Temporal Coherent Semantic Maps. *Eng. Appl. Artif. Intell.* **2016**, *48*, 173–187.
63. Mikolajczyk, K.; Schmid, C. A Performance Evaluation of Local Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1615–1630.

64. Morisset, B.; Rusu, R.B.; Sundaresan, A.; Hauser, K.; Agrawal, M.; Latombe, J.C.; Beetz, M. Leaving Flatland: Toward Real-Time 3D Navigation. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan, 12–17 May 2009; pp. 3786–3793.
65. Kostavelis, I.; Gasteratos, A. Learning Spatially Semantic Representations for Cognitive Robot Navigation. *Robot. Auton. Syst.* **2013**, *61*, 1460–1475.
66. Waibel, M.; Beetz, M.; Civera, J.; dâĂŹAndrea, R.; Elfring, J.; Galvez-Lopez, D.; Haussermann, K.; Janssen, R.; Montiel, J.; Perzylo, A.; *et al.* A World Wide Web for Robots. *IEEE Robot. Autom. Mag.* **2011**, *18*, 69–82.
67. Yousif, K.; Bab-Hadiashar, A.; Hoseinnezhad, R. A Real-Time RGB-D Registration and Mapping Approach by Heuristically Switching between Photometric and Geometric Information. In Proceedings of the 17th International Conference on Information Fusion (FUSION), Salamanca, Spain, 7–10 July 2014; pp. 1–8.
68. Grimmett, H.; Buerki, M.; Paz, L.; Pinies, P.; Furgale, P.; Posner, I.; Newman, P. Integrating Metric and Semantic Maps for Vision-Only Automated Parking. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 2159–2166.
69. Zhang, L.; Mistry, K.; Jiang, M.; Neoh, S.C.; Hossain, M.A. Adaptive Facial Point Detection and Emotion Recognition for a Humanoid Robot. *Comput. Vis. Image Underst.* **2015**, *140*, 93–114.
70. Sun, H.; Wang, C.; El-Sheimy, N. Automatic Traffic Lane Detection for Mobile Mapping Systems. In Proceedings of the 2011 International Workshop on Multi-Platform/Multi-Sensor Remote Sensing and Mapping (M2RSM), Xiamen, China, 10–12 January 2011; pp. 1–5.
71. Schmid, C.; Mohr, R.; Bauckhage, C. Evaluation of Interest Point Detectors. *Int. J. Comput. Vis. (IJCV)* **2000**, *37*, 151–172.
72. Salas-Moreno, R.F. Dense Semantic SLAM. PhD Thesis, Imperial College, London, UK, 2014.
73. Torralba, A.; Murphy, K.P.; Freeman, W.T.; Rubin, M. Context-Based Vision System for Place and Object Recognition. In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV), Madison, WI, USA, 16–22 June 2003; pp. 273–280.
74. Lienhart, R.; Kuranov, A.; Pisarevsky, V. Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. *Pattern Recognit.* **2003**, *2781*, 297–304.
75. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
76. Belongie, S.; Fowlkes, C.; Chung, F.; Malik, J. Spectral Partitioning with Indefinite Kernels Using the Nyström Extension. In Proceedings of the 7th European Conference on Computer Vision (ECCV), Copenhagen, Denmark, 28–31 May 2002; pp. 531–542.
77. Taylor, S.; Drummond, T. Multiple Target Localisation at over 100 FPS. In Proceedings of the 2009 British Machine Vision Conference (BMVC), London, UK, 7–10 September 2009; pp. 1–11.
78. Anati, R.; Scaramuzza, D.; Derpanis, K.G.; Daniilidis, K. Robot Localization Using Soft Object Detection. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA), St Paul, MN, USA, 14–18 May 2012; pp. 4992–4999.
79. 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 (IROS), Vilamoura, Algarve, Portugal, 7–12 October 2012; pp. 3005–3010.
80. Bosch, A.; Zisserman, A.; Munoz, X. Image Classification Using Random Forests and Ferns. In Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
81. Filliat, D. A Visual Bag of Words Method for Interactive Qualitative Localization and Mapping. In Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA), Roma, Italy, 10–14 April 2007; pp. 3921–3926.
82. Martínez-Gómez, J.; Morell, V.; Cazorla, M.; García-Varea, I. Semantic Localization in the PCL Library. *Robot. Auton. Syst.* **2016**, *75*, 641–648.
83. Sivic, J.; Zisserman, A. Video Google: A Text Retrieval Approach to Object Matching in Videos. In Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV), Nice, France, 13–16 October 2003; pp. 1470–1477.
84. Leung, T.; Malik, J. Representing and Recognizing the Visual Appearance of Materials Using Three-Dimensional Textons. *Int. J. Comput. Vis. (IJCV)* **2001**, *43*, 29–44.

85. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2169–2178.
86. 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 (ICRA), Shanghai, China, 9–13 May 2011; pp. 3297–3303.
87. Sami, M.; Ayaz, Y.; Jamil, M.; Gilani, S.O.; Naveed, M. Text Detection and Recognition for Semantic Mapping in Indoor Navigation. In Proceedings of the 2015 5th International Conference on IT Convergence and Security (ICITCS), Kuala Lumpur, Malaysia, 25–27 August 2015; pp. 1–4.
88. Wyss, M.; Corke, P.I. Active Text Perception for Mobile Robots. *QUT EPrints.* **2012**, unpublished.
89. Galindo, C.; Fernández-Madrigal, J.A.; González, J.; Saffiotti, A. Robot Task Planning Using Semantic Maps. *Robot. Auton. Syst.* **2008**, *56*, 955–966.
90. Baader, F. *The Description Logic Handbook: Theory, Implementation, and Applications*; Cambridge University Press: Cambridge, UK, 2003.
91. Coradeschi, S.; Saffiotti, A. An Introduction to the Anchoring Problem. *Robot. Auton. Syst.* **2003**, *43*, 85–96.
92. Vasudevan, S.; Siegwart, R. Bayesian Space Conceptualization and Place Classification for Semantic Maps in Mobile Robotics. *Robot. Auton. Syst.* **2008**, *56*, 522–537.
93. Viswanathan, P.; Meger, D.; Southey, T.; Little, J.J.; Mackworth, A. Automated Spatial-Semantic Modeling with Applications to Place Labeling and Informed Search. In Proceedings of the 6th Canadian Conference on Computer and Robot Vision (CRV), Kelowna, BC, Canada, 25–27 May 2009; pp. 284–291.
94. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis. (IJCV)* **2008**, *77*, 157–173.
95. Capobianco, R.; Serafin, J.; Dichtl, J.; Grisetti, G.; Iocchi, L.; Nardi, D. A Proposal for Semantic Map Representation and Evaluation. In Proceedings of the 2015 European Conference on Mobile Robots (ECMR), Lincoln, UK, 2–4 September 2015; pp. 1–6.
96. Pronobis, A.; Mozos, O.M.; Caputo, B.; Jensfelt, P. Multi-Modal Semantic Place Classification. *Int. J. Robot. Res. (IJRR)* **2009**, *29*, 298–320.
97. Galindo, C.; González, J.; Fernández-Madrigal, J.A.; Saffiotti, A. Robots that change their world: Inferring Goals from Semantic Knowledge. In Proceedings of the 5th European Conference on Mobile Robots (ECMR), Örebro, Sweden, 7–9 September 2011; pp. 1–6.
98. Crespo Herrero, J.; Barber Castano, R.I.; Martinez Mozos, O. An Inferring Semantic System Based on Relational Models for Mobile Robotics. In Proceedings of the 2015 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), Vila Real, Portugal, 8–10 April 2015; pp. 83–88.
99. Blodow, N.; Goron, L.C.; Marton, Z.C.; Pangercic, D.; Rühr, T.; Tenorth, M.; Beetz, M. Autonomous Semantic Mapping for Robots Performing Everyday Manipulation Tasks in Kitchen Environments. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011; pp. 4263–4270.
100. Ko, D.W.; Yi, C.; Suh, I.H. Semantic Mapping and Navigation: A Bayesian Approach. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan, 3–7 November 2013; pp. 2630–2636.
101. Boularias, A.; Duvallet, F.; Oh, J.; Stentz, A. Grounding Spatial Relations for Outdoor Robot Navigation. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 1976–1982.
102. Bernuy, F.; Solar, J. Semantic Mapping of Large-Scale Outdoor Scenes for Autonomous Off-Road Driving. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 11–12 December 2015; pp. 35–41.