Visual Place Recognition: A Survey

Stephanie Lowry, Niko Sünderhauf, Paul Newman, *Fellow, IEEE*, John J. Leonard, *Fellow, IEEE*, David Cox, Peter Corke, *Fellow, IEEE*, and Michael J. Milford, *Member, IEEE*

***Abstract*—Visual place recognition is a challenging problem due to the vast range of ways in which the appearance of real-world places can vary. In recent years, improvements in visual sensing capabilities, an ever-increasing focus on long-term mobile robot autonomy, and the ability to draw on state-of-the-art research in other disciplines—particularly recognition in computer vision and animal navigation in neuroscience—have all contributed to significant advances in visual place recognition systems. This paper presents a survey of the visual place recognition research landscape. We start by introducing the concepts behind place recognition—the role of place recognition in the animal kingdom, how a "place" is defined in a robotics context, and the major components of a place recognition system. Long-term robot operations have revealed that changing appearance can be a significant factor in visual place recognition failure; therefore, we discuss how place recognition solutions can implicitly or explicitly account for appearance change within the environment. Finally, we close with a discussion on the future of visual place recognition, in particular with respect to the rapid advances being made in the related fields of deep learning, semantic scene understanding, and video description.**

***Index Terms*—Visual place recognition, place recognition.**

I. Introduction

VISUAL place recognition is a well-defined but extremely challenging problem to solve in the general sense; given an image of a place, can a human, animal, or robot decide whether or not this image is of a place it has already seen? Whether referring to humans, animals, computers, or robots, there are some fundamental things a place recognition system must have and must do. First, a place recognition system must have an internal representation—a map—of the environment to compare to the incoming visual data. Second, the place recognition system must report a belief about whether or not the current visual information is from a place already included in the map, and if so, which one. Performing visual place recognition can be difficult due to a range of challenges; the appearance of a place can change drastically (see Fig. 1), multiple places in an environment may look very similar, a problem known as perceptual aliasing, and places may not always be revisited from the same viewpoint and position as before.

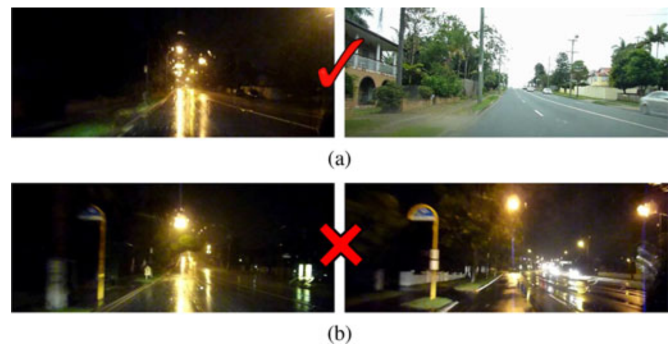

Fig. 1. Visual place recognition systems must be able to (a) successfully match very perceptually different images while (b) also rejecting incorrect matches between aliased image pairs of different places.

In robotics, this research topic is highly relevant given the ever increasing focus on long-term mobile robot autonomy and rapid improvements in visual sensing capabilities and cost. Vision is the primary sensor for many localization and place recognition algorithms [1]–[19]. Place recognition is also a growing research field, as evidenced by citation analyses and a number of dedicated place recognition workshops at recent and upcoming robotics and computer vision conferences including the IEEE International Conference on Robotics and Automation (2014, 2015) and the IEEE Conference on Computer Vision and Pattern Recognition (2015). The problem of persistent place recognition has also formed a regular component of many more general workshops including the long-running ICRA Workshop on Long-Term Autonomy (2011–2014).

Our aim in writing this survey article is to provide a comprehensive review of the current state of place recognition research that is relevant both to robotics and other fields of research including computer vision and neuroscience. The timing for such a survey is particularly fortuitous given major events across these related fields: for example, the almost universal usage of deep learning techniques in state of the art recognition systems in computer vision, and the 2014 Nobel Prize in Physiology or Medicine award to Edvard Moser, May-Britt Moser, and John O'Keefe, who discovered the key representations of place in

Manuscript received March 17, 2015; revised September 19, 2015; accepted October 23, 2015. This paper was recommended for publication by Associate Editor R. Eustice and Editor D. Fox upon evaluation of the reviewers' comments. This work was supported by an ARC Future Fellowship FT140101229, a Microsoft Research Faculty Fellowship, and the Australian Centre for Robotic Vision.

S. Lowry was with the Australian Centre for Robotic Vision, School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia. She is now with the Centre for Applied Autonomous Sensor Systems, Örebro University, 70219 Örebro, Sweden (e-mail: stephanie.lowry@oru.se).

N. Sünderhauf, P. Corke, and M. J. Milford are with the Australian Centre for Robotic Vision, School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia (e-mail: niko.suenderhauf@qut.edu.au; peter.corke@qut.edu.au; michael.milford@qut.edu.au).

P. Newman is with the Mobile Robotics Group, Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, U.K. (e-mail: pnewman@robots.ox.ac.uk).

J. J. Leonard is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jleonard@csail.mit.edu).

D. Cox is with the Department of Molecular and Cellular Biology, the School of Engineering and Applied Science, and the Center for Brain Science, Harvard University, Cambridge, MA 02138 USA (e-mail: davidcox@fas.harvard.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2015.2496823

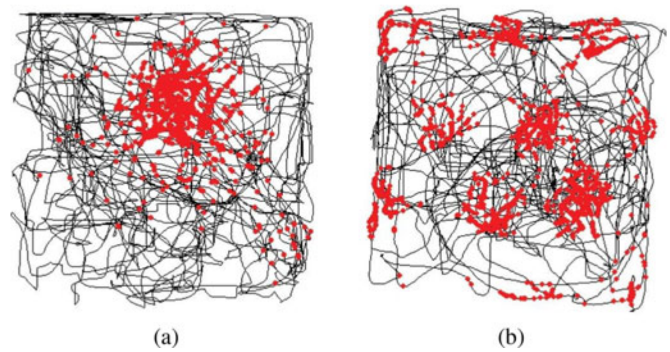

Fig. 2. Neuroscience experiments have shown that the brains of animals such as rats contain place cell and grid cell neurons. Each place cell fires strongly at one location in an environment, while each grid cell fires at multiple, regularly spaced locations. This figure shows the firing locations of (a) a place cell and (b) a grid cell placed over the path of an animal in a square environment (Annual Review of Neuroscience by Annual Reviews. Republished with permission of Annual Reviews, from [34]; permission conveyed through Copyright Clearance Center, Inc.).

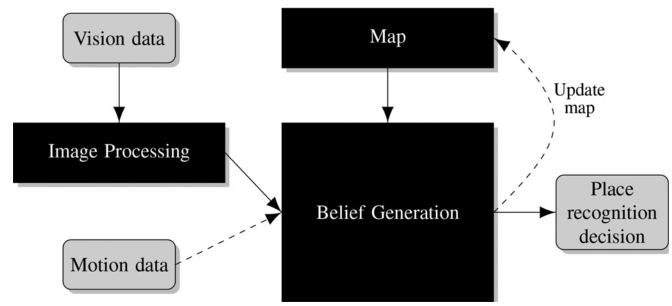

Fig. 3. Schematic of a visual place recognition system. Incoming visual data is processed by the *image processing* module. The robot's knowledge of the world is stored in the *map*. The *belief generation* module decides whether the current visual data matches a previously stored place. Motion information is also often included, and the map may be continually updated during operation.

the mammalian brain. This paper provides an overview of the place recognition problem and its relationship with many major robotics research fields including simultaneous localization and mapping (SLAM), localization, mapping, and recognition. Because of the increasing focus in the research community on long-term robot autonomy in challenging environments, we also provide a particular focus on the problem of lifelong visual place recognition for robots.

II. Concept of Place in Robotics and the Natural Kingdom

The problem of navigation and place recognition has a venerable tradition in psychology and neuroscience. In 1948, the research of Tolman [20] on rats navigating mazes motivated him to propose the cognitive map—a mental representation of the world with information about relationships between places that animals gradually learn. The concept of the cognitive map, while not without its critics [21], [22], has been influential not only in psychology and neuroscience, but also areas such as urban planning, where Lynch [23] proposed that the elements of a cognitive map be paths, edges, nodes, districts, and landmarks, and in robotics, where mapping approaches have been inspired by the cognitive map [24], [25], and by its successor, the spatial semantic hierarchy [26].

With the development of techniques to record neural activity in the brain of animals [27] came the identification of place cells in the rat hippocampus by O'Keefe and Dostrovsky [28]. Place cells fire when the rat is in a particular place in the environment [see Fig. 2(a)], and the population of place cells cover the entire environment [29], [30]. Furthermore, if a rat moves from one environment to another, the same place cells can be used to represent multiple different environments. O'Keefe and Conway [31] proposed that these place cells form a part of Tolman's cognitive map. The understanding about the relationships between neural activity and places in the world was extended by the discovery of head direction cells in the dorsal presubiculum [32] and of grid cells [33] in the medial entorhinal cortex. Head direction cells fire when an animal turns its head in a particular direction relative to its body, while grid cells fire in multiple places in the environment, in such a format that their firing fields form a regular grid [see Fig. 2(b)].

Place recognition, as observed via the firing of place cells, is triggered by both sensory cues and self-motion [29]. Studies with rats show that place cell firing is initially based on self-motion, but if the environment is changed—by altering the distance between start and end goals, for example—the place cell will update to the correct location according to the external visual landmarks [35], [36]. The correction may occur smoothly or abruptly, depending on the size of the mismatch.

Many of the same concepts arise in robotics. Most robots have access to external observation data as well as self-motion information. Topological and metric relationships between places are used in combination with sensory cues to determine the most likely place, similar to the neuronal firing of the place cells. Fig. 3 presents a schematic of a visual place recognition system. Visual place recognition systems contain three key components—an *image processing* module to interpret the incoming visual data, a *map* that maintains a representation of the robot's knowledge of the world, and a *belief generation* module, which uses the incoming sensor data in combination with the map to make a decision about whether the robot is in a familiar or novel place. A place recognition system may also use motion or transition information to inform the belief generation process. Furthermore, most place recognition systems are designed to operate online and, thus, must update the map accordingly.

This paper discusses what qualifies as a place in the context of robotic navigation. It then looks at the three key modules that make up the place recognition system: the image processing module, the mapping framework, and the belief generation module. The paper then turns to the problem of changing environments. It revisits each of the modules—the image processing module, the mapping module, and the belief generation module—and investigates how each has to be adapted to incorporate the notion of appearance change into the place recognition system's model of the world.

III. What is a Place?

The concept of places in robotics is motivated by the challenges of robotic navigation and mapping. A real robot has fallible sensors and actuators, and it is challenging to build a

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

metrically accurate map of the world, and to maintain self-localization within such a representation. The combination of both these goals, known as SLAM [37]–[41], is even more difficult to consistently achieve.

Instead of maintaining an accurate metric map, an alternative approach is to use a "relational map, which is rubbery and stretchy, rather than to try to place observations in a 2-D coordinate system" (see [40]). Such a topological map is conceptually similar to the biological notion of a cognitive map. Nodes represent the possible places in the world and edges represent the possible paths between these places. Robot navigation can then be abstractly defined as following edges between nodes. Places represent key intersections or decision points between routes [42], [43] as well as desirable end goals.

This topological approach to navigation is not without difficulties in practice. The robot has to associate abstract routes and places with physical places and paths, and the complex relationship between the robot sensors, the robot controls, and the robot's topological and metric interpretations of the world need to be defined [26]. Another issue is how a robot can generate topological maps. If the robot has access to a metric gridmap of the environment, it can extract topological information, emphasizing relevant navigation information like open spaces and passageways [44]. Alternatively, a topological map can be created by a robot from visual and transition information.

The definition of a place depends on the navigation context, and may either be considered as a precise position—"a place describes part of the environment as a zero-dimensional point" (see [26]), or as a larger area—"a place may also be defined as the abstraction of a region" where a region "represents a two-dimensional subset of the environment" (see [26]). For example, a room in a building might in some cases qualify as a single place, while in other cases, it might contain many different places. A region could also be defined as a 3-D area, depending on the requirements of the environment or robot. Unlike a robot pose, a place does not have an orientation, and an ongoing challenge in place recognition is pose invariance—ensuring recognition regardless of the orientation of the robot within the place.

The location of each place—whether a 1-D point or a larger region—can be selected based on spatial or temporal density. In this approach, a new place is added according to a particular time step, or when the robot has travelled a certain distance. Alternatively, a place can be defined in terms of its appearance. Kuipers and Byun [25] defined a place as somewhere distinctive relative to other nearby locations, according to some associated sensory information known as a place signature or place description. While the distinctiveness criterion is not always required, a topological place is defined as having a certain appearance configuration [45], [46] and the physical bounds of a place occur where the appearance changes significantly, called a "gateway" [47].

This qualitative concept of topological places as regions that are visually homogeneous needs to be quantified—that is, how can a place recognition system actually segment the world into distinct places? Ranganathan [48] noted that there are similarities with the problem of change-point detection in video segmentation [49], [50], and used change-point detection algorithms such as Bayesian surprise [50] and segmented regression [51] to define places within a topological map [48], [52]. A new place is created when the appearance of the environment, determined from the sensor measurements, becomes sufficiently different from the current model of the environment. Similarly, Korrapati *et al.* [53] used image sequencing partitioning techniques to group visually similar images together as topological graph nodes, while Chapoulie *et al.* [54] combined Kalman filtering with the Neyman–Pearson Lemma. Murphy and Sibley [55] combined dynamic vocabulary building [56] and incremental topic modeling [57] to continually learn new topological places in an environment, and Volkov *et al.* [58] used coresets [59] to segment the environment. Topic modeling, coresets, and Bayesian surprise techniques can also be used for other aspects of robotic navigation, such as summarizing a robot's past experience [60]–[62], or determining exploration strategies [63].

Appearance-based and density-based place selection methods are practical to implement as they depend on measurable quantities such as distance, time, or sensor values [64]. An ongoing challenge is the enhancement of appearance information with semantic labels such as "door" or "intersection" so places can be selected online based on their value as decision points. The addition of semantic data to maps can improve planning and navigation tasks [65] and requires place recognition to be linked with other recognition and classification tasks, especially scene classification and object recognition. These relationships are symbiotic: place recognition can improve object detection by providing contextual priming for object detection as well as contextual priors for object localization [66], and conversely, object recognition can also aid place recognition [67]–[70], particularly in indoor environments where the function of a place such as "kitchen" or "office" can be inferred from the objects within it, and used to infer the location from a labeled semantic map [71].

IV. Describing Places: The Image Processing Module

Visual place description techniques fall into two broad categories: those that selectively extract parts of the image that are in some way interesting or notable; and those that describe the whole scene, without a selection phase. Examples of the first category are local feature descriptors such as scale-invariant feature transforms (SIFT) [72] and speeded-up robust features (SURF) [73]. Local feature descriptors first require a detection phase which determines the parts of the image to retain as local features [see Fig. 4(a)]. In contrast, global or whole-image descriptors such as Gist [74], [75] do not have a detection phase but process the whole image regardless of its content [see Fig. 4(b)].

A. *Local Feature Descriptors*

The development of the local feature method SIFT [72] led to its widespread use in place recognition [76]–[83]. As other local feature detection and description methods were developed, they too were applied to the visual localization and place recognition problem. For example, Ho and Newman [84] used Harris affine regions [85], Murillo *et al.* [86] and Cummins and Newman [87]

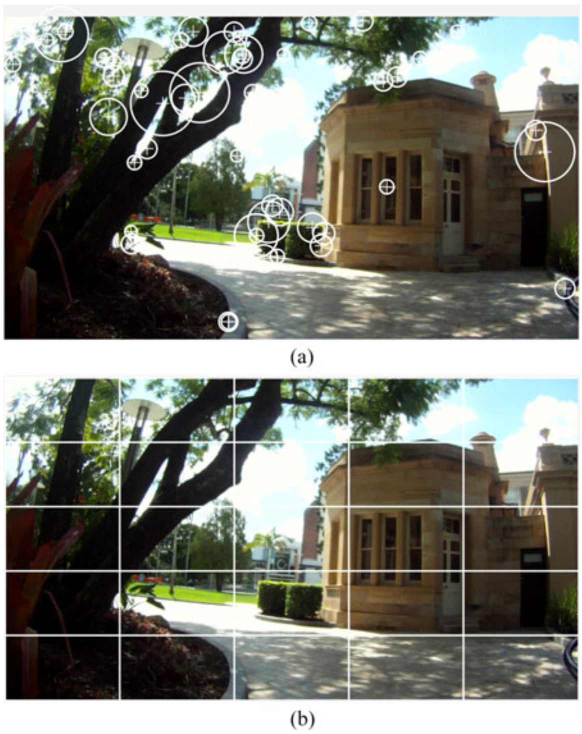

Fig. 4. Visual place description techniques fall into two broad categories. (a) Interesting or salient parts of the image are selected for extraction, description and storage. For example, SURF [73] extracts interest points in an image for description. The circles are interest points selected by SURF within this image. The number of possible features may vary depending on the number of interest points detected in the image. (b) Image is described in a predefined way such as the grid shown here without first detecting interest points. Whole-image descriptors such as Gist [74], [75] then process each block regardless of its content.

used SURF [73], while FrameSLAM [2] used CenSurE [88]. Since local feature extraction consists of two steps—detection followed by description—it is not uncommon to combine different techniques for each. For example, Mei *et al.* [89] used the detection technique FAST [90] to find keypoints in the image, which were then described by SIFT descriptors. Similarly, Churchill and Newman [15] used FAST extraction combined with BRIEF [91] descriptors.

Each image may contain hundreds of local features, and directly matching image features can be inefficient. The bag-of-words model [92], [93] increases efficiency by quantizing local features into a vocabulary that can be compared using text retrieval techniques [94]. The bag-of-words model partitions a feature space, such as SIFT or SURF descriptors, into a finite number of visual words. A typical vocabulary contains 5000–10 000 words, but a vocabulary as large as 100 000 words has been used for place recognition by FAB-MAP 2.0 [87]. For each image, every feature is assigned to a particular word, ignoring any geometric or spatial structure, thereby allowing images to be reduced to binary strings or histograms of length n, where n is the number of words in the vocabulary.

Images described using the bag-of-words model can be efficiently compared using binary string comparison such as a Hamming distance or histogram comparison techniques. Vocabulary trees [95] can make the process for large-scale place recognition even more efficient. Originally proposed for object recognition, vocabulary trees use a hierarchical model to define words, an approach that enables faster lookup of visual words and the use of a larger and thus more discriminating vocabulary. Localization systems that use the bag-of-words approach include [82], [84], [87], [96], [97] and many others.

Because the bag-of-words model ignores the geometric structure of the place it is describing, the resulting place description is pose invariant; that is, the place can be recognized regardless of the position of the robot within the place. However, the addition of geometric information to a place has been shown to improve the robustness of place matching, particularly in changing conditions [14], [87], [98]–[100]. These systems may assume a laser sensor is available for 3-D information [98], use stereo vision [14], epipolar constraints [100], [101], or simply define the scene geometry according to the position of the elements within the image [102], [103]. The tradeoff between pose invariance—recognizing places regardless of the robot orientation—and condition invariance—recognizing places when the visual appearance changes—has not yet been resolved and is a current challenge in place recognition research.

The bag-of-words model is typically predefined based on features extracted from a training image sequence. This approach can be limiting as the resulting model is environment dependent and needs to be retrained if a robot is moved into a new area. Nicosevici and Garcia [56] propose an online method to continuously update the vocabulary based on observations, while still being able to match prior observations with future observations. As a result, a bag-of-words model can be used without requiring a pretraining phase and can adapt to the environment, outperforming pretrained models despite requiring less *a priori* knowledge [56].

B. Global Descriptors

Global place descriptors used in early localization systems included color histograms [5] and descriptors based on principal component analysis [104]. Lamon *et al.* [105] used a variety of image features—such as edges [106], corners [107], and color patches—combined into a fingerprint of a location. By ordering these features in a sequence between 0° and 360°, place recognition could be reduced to string-matching. These systems used omnidirectional cameras which allowed rotation-invariant matching at each place.

Global descriptors can be generated from local feature descriptors by predefining the keypoints in the image—for example, using a grid-based pattern—and then using the chosen feature description method on the preselected keypoints. Badino *et al.* [108] used whole-image descriptors based on SURF features known as WI-SURF to perform localization and BRIEF-Gist [109] used BRIEF features [91] in a similar whole-image fashion.

A popular whole-image descriptor is Gist [74], [75], which has been used for place recognition on a number of occasions [110]–[113]. Gist uses Gabor filters at different orientations and different frequencies to extract information from the image. The results are averaged to generate a compact vector that represents the "gist" of a scene.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 5. Object proposal methods such as the Edge Boxes method [123] shown here were developed for object detection but can also be used to identify potential landmarks for place recognition. The boxes in the images above show landmarks that have been correctly matched between two viewpoints of a scene (from [122]).

C. Describing Places Using Local and Global Techniques

Local and global descriptors each have different advantages and disadvantages. Local feature descriptors are not restricted to defining a place only in terms of a previous robot pose, but can be recombined to create new places that have not previously been explicitly observed by the robot. For example, Mei *et al.* [114] defined places via covisibility: the system finds cliques in the landmark covisibility map, and these cliques define places even when the landmarks have not simultaneously been seen in a single frame. Covisibility can outperform standard image-based techniques [78]. Lynen *et al.* [115] generated a 2-D space of descriptor votes where regions of high vote density represent loop closure candidates.

Local features can also be combined with metric information to enable metric corrections to localization [2], [7], [76]. Global descriptors do not have the same flexibility, and furthermore, whole-image descriptors are more susceptible to change in the robot's pose than local descriptor methods, as whole-image descriptor comparison methods tend to assume that the camera viewpoint remains similar. This problem can be somewhat ameliorated by the use of circular shifts as in [116] or by combining a bag-of-words approach with a Gist descriptor on segments of the image [17], [110].

While global descriptors are more pose dependent than local feature descriptors, local feature descriptors perform poorly when lighting conditions change [117] and are comprehensively outperformed by global descriptors at performing place recognition in changing conditions [118], [119]. Using global descriptors on image segments rather than whole images may provide a compromise between the two approaches, as sufficiently large image segments exhibit some of the condition invariance of whole images, and sufficiently small image segments exhibit the pose invariance of local features. McManus *et al.* [120] used the global descriptor HOG [121] on image patches to learn condition invariant scene signatures, while Sünderhauf *et al.* [122] used the Edge Boxes object proposal method [123] combined with a mid-level convolutional neural network (CNN) feature [124] to identify and extract landmarks as illustrated in Fig. 5.

D. Including Three-Dimensional Information in Place Descriptions

The image processing techniques described above are appearance based—they "model the data directly in the visual domain (instead of making a geometric model)" (see [104]). However, in metric localization systems, the appearance-based models must be extended with metric information. Monocular image data is not a natural source of geometric landmarks—"the essential geometry of the world does not "pop out" of images the same way as it does from laser data" (see [125]). While many systems use data from additional sensors such as lasers [98] or RGB-D cameras [126]–[128], geometric data can also be extracted from conventional cameras to allow metric calculation of the robot pose.

Metric range information can be inferred using stereo cameras [2], [129]–[131]. Monocular cameras can also infer metric information using Structure-from-Motion algorithms [132]. Methods include MonoSLAM [7], PTAM [133], DTAM [134], LSD-SLAM [135], and ORB-SLAM [136]. Metric information can be sparse: that is, range measurements are associated with local features such as image patches as in MonoSLAM [7], SIFT features as in [76], CenSurE features as in FrameSLAM [2], or ORB features [137] as in ORB-SLAM [136]. In contrast, DTAM stores dense metric information about every pixel, and LSD-SLAM maintains semidense depth data on the parts of the image containing structure and information. Dense metric data allow a robot to perform obstacle avoidance and metric planning as well as mapping and localization; therefore, fully autonomous vision-only navigation can be performed [16].

The introduction of novel sensors, such as RGB-D cameras, that provide dense depth information as well as image data has spurred the development of dense mapping techniques [70], [126]–[128], [138], [139]. These sensors can also exploit 3-D object information to improve place recognition. SLAM++ [70] stores a database of 3-D object models, uses this database to perform object recognition during navigation, and uses these objects as high-level place features. Objects have a number of advantages over low-level place features: they provide rich semantic information and can reduce memory requirements via semantic compression, that is, storing object labels rather than full object models in the map [70].

V. Remembering Places: The Mapping Module

For a place recognition or navigation task, the system needs to refer to a map—a stored representation of the robot's knowledge of the world—to which the current observation is compared. The map framework differs depending on what data are available and what type of place recognition is being performed. Table I displays a taxonomy of mapping approaches, which depends on the level of physical abstraction in the map and whether or not metric information is included in the place description. The most concrete mapping framework listed is the topological-metric or topometric map. Although it is possible to have a globally metric map, such maps are only feasible in small geographical areas, and there are mechanisms for fusing topometric maps into globally metric maps [140]. Thus, for the purposes of place recognition, any globally metric map can be considered as a one-node topometric map.

A. Pure Image Retrieval

The most abstract form of mapping framework for place recognition only stores appearance information about each place

TABLE I
MAPPING FRAMEWORKS FOR VISUAL PLACE RECOGNITION

Level of map abstraction	Place description type	Comments
Pure image retrieval	Appearance-based	No position information
Topological	Appearance-based	Includes transition information
Topological-metric	Appearance-based	Includes metric information between but not within places
Topological-metric	Sparse metric information (landmark maps)	SLAM system – includes metric information between and within places
Topological-metric	Dense metric information (occupancy grid maps)	SLAM system – includes metric information between and within places

in the environment, with no associated position information. Pure image retrieval assumes that matching is based solely on appearance similarity and applies image retrieval techniques from computer vision that are not specific to place-based information [3]. Although valuable information is lost by not including relative position information, there are computationally efficient indexing techniques that can be exploited.

A key concern with place recognition is system scalability—as the robot visits more and more places, storage requirements will increase and search efficiency will decrease. As a result, maps need to be designed to ensure large-scale efficiency. If a bag-of-words model is used to quantize the descriptor space, image retrieval can be accelerated using inverted indices; the image ID numbers are stored against the words that appear in the image, rather than the words being stored against the image IDs. Inverted indices allow quicker elimination of unlikely images, rather than requiring a linear search of all images in the database.

Schindler *et al.* [3] used a hierarchical vocabulary tree [95] to achieve efficient visual place recognition of a city-sized dataset (a 20-km traversal with around 100 million features). This paper showed that place recognition performance improves if only the most informative features from each image are used, where information gain is measured using a conditional entropy calculation. Improved place recognition with a reduced feature set was also observed by Li and Košecká [141].

FAB-MAP 2.0 [87], [142] also used an inverted index with a bag-of-words model to demonstrate visual place recognition along a 1000-km path. While Schindler *et al.* [3] used a voting scheme to match locations, FAB-MAPs probabilistic model that included negative observations—words that did *not* appear in the image—as well as positive observations required simplification before the inverted index approach could be applied.

Place recognition can also be made more efficient by using hierarchical searching at the place level as well as at the vocabulary level. Mohan *et al.* [143] selected the most likely environment using cooccurent feature matrices. Preselecting a subset of the global environment reduced the search space, thereby increasing the efficiency of the place recognition process.

B. Topological Maps

Pure topological maps contain information about relative positions of places but do not store metric information regarding how these places are related [5], [6], [118], [119]. Topological information can be used to both increase the number of correct place matches and filter out incorrect matches [14], [84]. A probabilistic system like FAB-MAP can be run as a pure image retrieval process by assuming a uniform location prior at all steps, but performance improves when transition information is included through Bayesian filtering or similar techniques.

While image retrieval techniques can use an inverted index to improve efficiency, topological maps can use a location prior to speed up matching, that is, the place recognition system only has to search places known to be close to the robot's current position. A sampling-based method such as a particle filter can be used to sample possible places [12], [13], [111], [144]. The particles are resampled according to which places are the most likely and can stay close by the current robot location if it is well localized, or spread out across the whole environment if the robot is lost. Computation time is thus proportional to the number of particles, not the size of the environment [145].

Alternatively, since the number of loop closures in an environment is naturally sparse, Latif *et al.* [19] used topological information to formulate place recognition as a sparse convex L_1-minimization problem and applied efficient homotopy methods [146] to provide loop closure hypotheses.

The addition of topological information into the recognition process allows place recognition using low-resolution data and thus lower memory requirements. Using the sparse convex L_1-minimization formulation, successful place recognition was achieved using images as small as 48 pixels [19]. Even in challenging scenarios where images were blurred or observed under different environment conditions such as different times of day, the use of topological information allowed visual place recognition using as few as 32 4-bit pixels per image [147].

C. Topological-Metric Maps

As image retrieval can be enhanced by adding topological information, topological maps can be enhanced by including metric information—distance, direction, or both—on the map edges. For example, both FAB-MAP [6] and SeqSLAM [118] are originally purely topological systems, but the addition of odometry information has been demonstrated to improve each system's place recognition performance by CAT-SLAM [13] and SMART [148] respectively.

These topological-metric maps can be appearance-based, in which case metric information is only included as relative poses between each place node [149]–[152]. However, metric information about the position of landmarks or objects in a place can also be stored within each node [1], [2], [26], [140], [153]–[156].

The metric information within the topological place node can be stored as a sparse landmark map [2], [7], [76], or as a dense occupancy grid map [134] if depth information is extracted from the image data. Although the notion of dense spatial modeling using a truncated signed distance function representation can be traced back to the work of Moravec and Elfes [39] in the mid-1980s, it has only become feasible in the past few years with the advent of GPU technology [134].

VI. Recognizing Places: The Belief Generation Module

Ultimately the purpose of a place recognition system is to determine whether a place has been seen before. Thus, the central goal of any place recognition system is reconciling visual input with the stored map data to generate a belief distribution. This distribution provides a measure of likelihood or confidence that the current visual input matches a particular location in the robot's map representation of the world. There is a general understanding that if two place descriptions appear similar there is a greater likelihood of them being captured at the same physical location, but the degree to which this is true depends on the particular environment. For example, repetitive environments may exhibit perceptual aliasing where different places are indistinguishable. Conversely, changing conditions may cause the same place to appear drastically different at different times.

A. Place Recognition and Simultaneous Localization and Mapping

Place recognition plays an important role in pose graph SLAM algorithms by providing loop closure candidates [157]. Pose graphs, also known as view-based representations [158], [159], are widely utilized in modern SLAM systems because of their computational efficiency for fixed size maps, although they can suffer from an increase in computational requirements for long duration missions. Loop closure is vital for consistent mapping as it allows the system to correct drift in local odometry measurements [160], [161]. Loop closure can be decoupled from the online local update step, and many systems independently perform both SLAM-like local metric correction and topological-like loop closure [1], [2], [80], [161]; a system can perform local metric correction using laser scan data [80], [161] or visual odometry [1], [2] while a separate global process looks for matches in order to close large loops.

If the place descriptions are appearance based and do not contain any metric information, but the map contains metric distances between places, the system can still use the loop closures to perform metric correction at the place level [149]–[152]. However, if the place descriptions contain metric information associated with the image features, as is the case for FrameSLAM [2], then a more precise correction can be performed. Maps that are purely topological do not provide any metric pose correction. In these cases, localization at a topological level occurs; that is, the system simply identifies the most likely location.

The place recognition maps that contain metric information both within and between the place descriptions can be used to perform a full metric SLAM solution. There are a wide range of SLAM techniques available, as summarized in [162]–[164]. Thrun and Leonard [164] identify three key SLAM paradigms: extended Kalman filters (EKF) [37], [38], [165]–[167] and Rao-Blackwellized particle filters [168], as well as the pose graph approach discussed above [160], [161], [169]–[171]. Vision-based systems utilize all these methods: MonoSLAM [7] uses an EKF, Rao-Blackwellized particle filters are used in [12], [172], and [173], and pose graph optimization techniques are used in [2] and [174].

B. Topological Place Recognition

If multiple streams of data are available, a voting scheme [3], [5], [79], [96], [175] can be used. Ulrich and Nourbakhsh [5] used multiple color bands, each of which voted for what it considered the most likely location. Depending on the votes, the system could be confident, uncertain, or confused. If the confident bands were unanimous and the total confidence was above a certain threshold, the system was confident about its location; if none of the bands were sufficiently confident or the total confidence value was too low, the system was uncertain; and if the confident bands disagreed on the location, the system was confused.

If a system uses the bag-of-words model, inspired as it is by text-based document analysis, it may use the related Term frequency-inverse document frequency (TF-IDF) score [56], [114], [176]. Each visual word in an image has a TF-IDF score, which is made up of two parts: the term frequency, which measures how often the word appears in the image, and the inverse document frequency, which measures whether the word is common across all images. The TF-IDF score is then the product of these two values.

A probabilistic calculation can also be used to compute place matching likelihood, using a calculation based on Bayes theorem. Early examples of appearance-based probabilistic localization used Gaussians to represent probability [177], or a mixture of Gaussians combined with expectation maximization [178], or a Gaussian kernel [179] with Parzen smoothing [104]. Other choices for the observation likelihood include the use of TF-IDF for the observation likelihood, if a bag-of-words model is being used [83], [180]. Siagian and Itti [111], [181] used Monte Carlo localization with two observation update steps each with an independent observation likelihood: one based on the segment likelihood and one based on the object likelihood. Garcia-Fidalgo and Rodriguez [182] used the observation likelihood that relates the number of feature matches between two images to the overall number of features in the image, scaled by a normalizing constant.

The observational likelihood can also be computed via a data-driven approach. FAB-MAP [6], [87] is a probabilistic appearance-based localization system that uses a data-driven approach to calculating an observational likelihood. FAB-MAP uses a bag-of-words model with SIFT or SURF features for image description and calculates the distinctiveness of each word during a training phase. As a bag-of-words model may have many words—FAB-MAP has been used with a 100 000-word vocabulary [87]—the full joint probability distribution of the

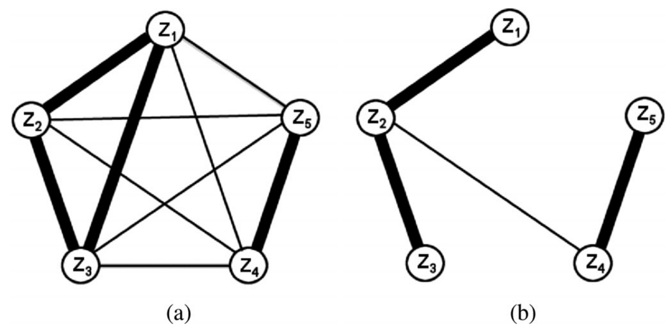

Fig. 6. FAB-MAP learns a probabilistic model of the relationship between word appearance and place recognition. (a) Full joint distribution takes into account the relationships between words (the thick lines between words represent those with the largest mutual information). (b) Chow–Liu tree approximates the full joint distribution as a junction tree where each word depends only on one other word (from [184]).

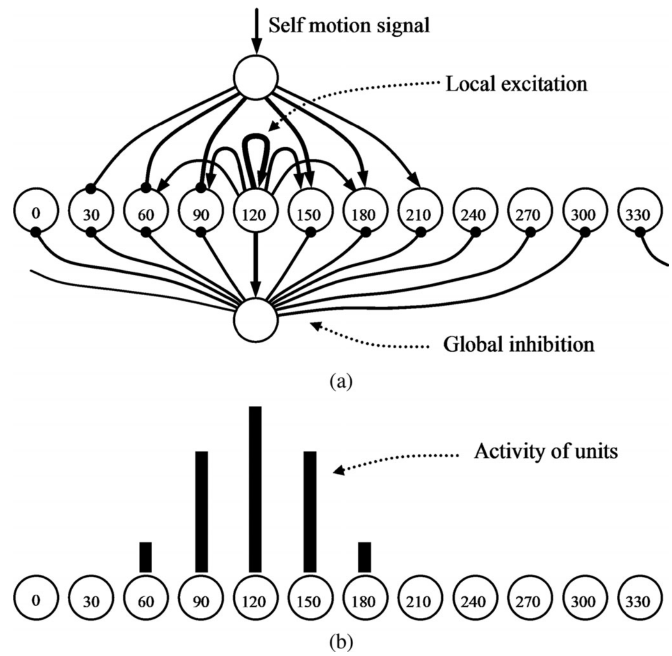

Fig. 7. CANs are a type of neural network that can be used to model the behavior of place cells, head direction cells, and grid cells. (a) Example of a CAN used to model head direction cells. Each cell excites itself and units near itself (see local excitation arrows) and inhibits other cells. (b) Stable activity packet centered at 120° generated by the combination of local excitation and global inhibition with input from a motion input performs place recognition by exciting nearby pose cells and inhibiting those that are far away through a combination of odometry and visual input (from [116]).

observed words [see Fig. 6(a)] can be approximated by a naïve Bayes assumption or a Chow–Liu tree [183] [see Fig. 6(b)].

FAB-MAP handled the perceptual aliasing problem by considering not only whether two locations were similar in the sense that they have many visual words in common, but also whether the words in common were sufficiently rare that the locations could be considered distinctive. As a result, if two locations looked similar but the words that appear were frequently observed, FAB-MAP generated a low matching probability by using the denominator as a normalizing constant that was calculated over the set of all previously seen locations and the set of all locations that have not yet been visited.

Originally, the set of unvisited locations was modeled by randomly sampling from the Chow–Liu tree, and the probability that the robot was at a location that had not yet been observed was a user-defined parameter. However, Paul and Newman [60], [62], [185] presented an iterative learning mechanism to generate a representative set of the true distribution of the appearance of the world. Latent Dirichlet Allocation [186] was used to cluster images into major topics that summarize how the world, as seen so far by the robot, appears. These topics were used to generate a sampling set that is proportional to what is common in the world; for example, foliage occurs frequently in many environments so should not be considered distinctive. The system learned incrementally; after each deployment, a better sampling set was created as the system discovered more about the world. Furthermore, an online–offline learning process was proposed, whereby during the robot's "down-time," it could download and integrate further relevant image data from the internet.

Olson [187] observed that "correct hypotheses generally agree with each other, whereas incorrect hypotheses tend to disagree with each other" and used this property to eliminate false positive matches by calculating a pairwise consistency matrix between possible hypotheses to find the most consistent set of hypotheses from the dominant eigenvectors. The same paper also observed that the amount of information required to generate a belief match should scale with the robot's positional uncertainty. The system ensured this by requiring that local hypothesis matches covered a large physical space in comparison with the robot's positional uncertainty to ensure that the robot was not incorrectly located within its uncertainty ellipse [187].

This approach contrasts with FAB-MAP's requirement of a few highly distinctive matches. Instead, many matches are required, but these matches do not need to be particularly distinctive, as the geometrical relationship between the matches ensures the uniqueness of the hypothesis.

Biologically inspired methods for place recognition mimic the known place cells structure in the rat hippocampus [116], [188]. In RatSLAM [116], a type of neural network known as a continuous attractor network (CAN) was used to model place cells (see Fig. 7). This CAN used a combination of local excitation and global inhibition combined with input from ego-motion and visual sensors to perform localization. In a similar manner Giovannangeli *et al.* [188] used a place cell model to perform vision-based navigation in indoor and outdoor environments without a metric map.

C. Evaluation of Place Recognition Systems

Topological place recognition systems are typically evaluated using precision and recall metrics and their relationship via a precision-recall curve. A system selects matches based on a particular confidence measure. The correct matches are known as *true positives*, the incorrect matches are *false positives*, and matches that the system erroneously discards are *false negative* matches. Precision is defined as the proportion of selected matches that are true positive matches, and recall is the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

proportion of true positives to the total number of actual matches, that is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

A perfect system would be one that achieves precision of 100% and recall of 100%. Precision and recall are often related to each other via a precision–recall curve, which plots recall against precision for a range of confidence values.

Until recently, place recognition prioritized avoidance of false positive matches [6], as introducing false matches into a map could cause catastrophic failure. As a result, recall at 100% precision was the key metric for place recognition success. However, several methods for using topological information to correct false positive matches have been proposed [189]–[191], and attention has turned from eliminating all false positives to finding many potential place matches and then correcting any mismatches in a topological post-processing step. Increasing the number of potential matches is particularly important when performing place recognition in changing environments, when strict matching methods are liable to fail.

Furthermore, as place recognition systems transition from "demonstration" (typically with prerecorded data sets) to "deployment" (operating in real time on autonomous vehicles), the performance evaluation methodology may change further to include a consideration of the spatial distribution of place matches within the environment. For example, McManus *et al.* [192] used the probability of travelling a given distance without a successful match as a measure of place recognition success. This metric expresses how evenly distributed the place matches are across the environment and is an important measure for the overall integrity of a navigation system that uses place recognition as a module.

VII. Visual Place Recognition in Changing Environments

Early place recognition systems often implicitly used the simplifying assumption that the visual appearance of each place would not change over the course of the experiment. However, as robotic systems operate in ever-larger uncontrolled environments and for longer time periods, it has rapidly become apparent that this assumption is no longer valid. Consequently, in recent years, there has been a growing focus on creating persistent robotic navigation systems, including persistent place recognition techniques. The ability to localize in and generate maps of dynamic environments has been identified as being of key importance [193]. This section revisits each of the previous concepts: place description techniques, mapping frameworks, and the belief generation process, and discusses how each has to be adapted to manage a changing environment.

A. Describing Places in Changing Environments

It is clear that the appearance of a place can vary greatly over time due to a large number of causes including changes in lighting and weather (see Fig. 1). There are two approaches for performing place recognition when faced with appearance change—the first tries to find a condition-invariant description of the place, the way local feature descriptors are designed to be scale-, rotation-, and illumination-invariant. The second method tries to learn how appearance change occurs.

1) Invariant Methods: The difficulty of matching places in changing environments using conventional local features is a significant one for persistent robot navigation. Furgale and Barfoot [117] observed that the non-repeatability of SURF features due to changing appearance, particularly lighting change, was a major cause of failure during visual-teach-and-repeat experiments. Existing image description methods have been tested to determine their robustness to illumination and other change. In [194], Valgren and Lilienthal tested SIFT features and a number of SURF variants across change in lighting, cloud cover, and seasonal conditions. The SURF variants all outperformed SIFT, but none of the tested features were found to be robust across all conditions. However, in later work [100], the authors combined U-SURF [73], the most successful SURF variant, with a consistency check using the epipolar constraint, and achieved between 80% and 100% correct matching within small (40 image) datasets.

Ross *et al.* [195], [196] studied the effect of lighting change on features using time-lapse footage across full days to determine the illumination sensitivity of each descriptor. The feature keypoints were predefined within each image, and only the variance of the feature descriptor was tested, in contrast with the work of Valgren and Lilienthal [100], [194], which tested the combined effect of feature detector and descriptor. The U-SIFT [72] descriptor was shown to display the greatest lighting invariance of the tested descriptors.

Instead of using point features such as SIFT or SURF, other descriptors can be chosen. Whole-image descriptors have been used in systems such as SeqSLAM and others [118], [119], [197] that demonstrate robustness against environmental change. However, as for other description methods, too drastic a change in appearance will cause system failure [111] and whole-image descriptors also suffer from the additional problem of sensitivity to viewpoint change [198].

Edge features can be used in appropriate environments [172], [199], as they are invariant to lighting, orientation and scale [199]. Nuske *et al.* [199] used line-based localization to localize against an existing map with a fish-eye camera and tested it in an outdoor industrial area under various lighting conditions across times of day from 7:00 to 17:00. Borges *et al.* [200] extended this system to generate its own edge map using 3-D laser data for localization. However, data association using edge features can be challenging [172].

Techniques such as shadow removal [201] and the use of an illumination invariant color space [192] can lessen the effect of appearance variability caused by illumination change. Alternatively, a hardware-based solution to place recognition in variable lighting conditions can be used. McManus *et al.* [202] used scanning laser-rangefinders to create "camera-like" images that were not affected by the illumination of the scene. This solution had the advantage of being applicable in complete darkness. A

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

long-wave infrared thermal imaging camera is another sensor that can be deployed in a manner similar to a standard camera but which responds differently to lighting variance. Maddern and Vidas [203] showed thermal imaging cameras can provide improved place recognition at night-time when visible light cameras fail.

CNNs have recently been used as robust feature extractors for place recognition in changing environments. Exploring the utility of CNNs for place recognition has been motivated by their ability to learn generic features that are transferrable to a variety of related but different visual tasks [204], [205]. The authors of [206], [207] utilized CNN features as holistic image descriptors and analyzed the robustness of different layers against visual appearance and viewpoint changes. They concluded that mid-level features exhibit a robustness against appearance changes, while higher level features are more robust against changes in viewpoint and carry more semantic information that can be used to partition the search space [207].

One aspect of visual data that has not been investigated in depth for place recognition in changing environments is that of color. While conventional images descriptors such as SURF and BRIEF operate on grayscale images, most available cameras capture color images, which have the potential to provide new and interesting information about place recognition in changing environments. Color information presents an interesting paradox for place recognition in changing environments: it is known to perform poorly as a feature when the illumination of a scene changes [195], but conversely, relative color information contains information about lighting that can improve place recognition dramatically by identifying and removing shadows [201]. Illumination invariant images use relative color information and are more reliable for place recognition during the day, but are outperformed by color images at night, when the underlying assumptions about black-body illumination are violated [208].

2) Learning Methods: The alternative to invariant approaches is to learn a relationship between how places appear at different times. These methods assume that places change appearance in a similar way across an environment, and therefore, change learned during training can be generalized to previously unseen locations. This assumption has been tested by observing static webcams from different locations [209], [210] and demonstrating that the most significant transformations across time are similar across different places. Furthermore, a training set of locations can be used to compute a principal component basis that encodes new locations with only a small loss of accuracy.

Ranganathan *et al.* [211] learned a fine vocabulary [212]; a fine vocabulary is similar to a bag-of-words model in that it segments a descriptor space, such as SIFT descriptors, but it does so very finely—into over 16 million words in [212]. The system then learned a probability distribution over these words. The motivation for the fine vocabulary is the observation that descriptors transform in a highly non-linear way due to illumination change, changing viewpoint, and other effects, and learning a distribution of alternative words allows these changes to be learned and quantified. In [211], the distribution was learned from multiple training runs over the same environment, and features were matched across different illumination conditions to generate the probability distribution. Improved performance was reported compared to using a conventional vocabulary tree [95], with an additional 10–15% of the dataset being correctly matched. The distance metric was also compared and the symmetric Kullback–Leibler divergence was shown to outperform either the standard descriptor distance metric or a probability distance metric.

Using webcam footage, Carlevaris-Bianco and Eustice [213] tracked image patches over different lighting conditions to generate a large set (3 million features) of positive and negative examples. From these data, a neural network learning technique [214] mapped the patches into a new space in which positive matches were close together, according to the Euclidean distance, and negative matches were further away. The mapped descriptors were shown to be substantially more successful at place recognition than SIFT and SURF descriptors—compared with SURF descriptors, an additional 10% of the test locations were correctly matched.

Neubert *et al.* [18] learned a visual translation between two different seasons. Training images from two different seasons were segmented using SLIC superpixels [215]. The superpixels were described using a color histogram and a SURF descriptor, and a dictionary of translations of superpixels from one season to another season was learned. Similarly, Lowry *et al.* [216] learned a linear transformation from images captured in the morning to images captured in the late afternoon. However, for such appearance translation to be successful, the pairs of training images must be well aligned.

Learning-based methods frequently require a supervised training phase, which implies that the likely appearance change is known and that relevant training data is available. Lowry *et al.* [217] proposed an unsupervised learning method for place recognition in changing environments. This system identified and removed aspects from each observation that were widespread across the environment. Removing commonly occurring elements reduced the risk of widespread place recognition failure and increased the stability of the place descriptions.

B. Remembering Places in Changing Environments

If the environment is changing, the map also needs to change to continue faithfully representing the environment. The system must determine what to remember and what to forget. It may also be beneficial for the system to maintain multiple representations of a place, as places can vary between different configurations. This section presents mapping frameworks for place recognition that have the capacity to handle changing environments in one of these two ways—either by deciding what to remember and what to forget, and/or by remembering multiple different representations. These systems are not all specific to vision-based systems, and many have been designed to handle laser data, but demonstrate concepts that are relevant to any sensor modality or map framework.

1) Remembering and Forgetting Data: In a dynamic environment, each place representation must be updated as new observations are obtained by the robot. A balance has to be found between using recent observations to overwrite obsolete

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

information, and not allowing fleeting events to overwrite the status quo. However, it is difficult to determine which events are transient and which are worth remembering. Drawing inspiration from concepts in neuroscience, Biber and Duckett [218] referred to this as the "stability-plasticity dilemma." Biological brains can inspire solutions for coping with this dilemma; concepts such as sensory memory, short-term memory, and long-term memory found in human memory models have been co-opted to create decision models for remembering and forgetting.

One biologically inspired mapping system passes sensor information through an analog of sensory memory to short-term memory and long-term memory storage areas [219], [220]. In the first stage, a selective attention mechanism decides which information will be upgraded from sensory memory to short-term memory, based on information from the long-term memory. The second stage involves using a rehearsal mechanism to determine which information will be transferred from short-term to long-term memory. Using attention and rehearsal mechanisms ensures that more persistent, stable, and frequently occurring features are remembered, while transient features are forgotten. Elements must be seen and recognized sufficiently often before they are considered for promotion to a higher level of memory. Furthermore, obsolete features are slowly filtered out of the long-term memory. There is a complementary problem of which elements to remember, which typically uses similar criteria [219], [221] to the forgetting process.

Andrade-Cetto and Sanfeliu [222] required that features be trustworthy and reliable as well as up-to-date in order to be retained, while Bailey [221] considered a usefulness criteria based on visibility—a feature that can be blocked by other elements of the environment is liable to suffer from occlusion errors and be less useful in the future. Johns and Yang [102] and Hafez *et al.* [223] used a bag-of-words model and applied a quality measure to determine useful features to retain, considering both feature distinctiveness and feature reliability when generating a model of a location. Johns and Yang [224] also proposed a generative bag-of-words model that considered the variance as well as the mean value of each data point when matching scenes.

2) Multiple Representations of the Environment: Not only do places change in appearance over time, but they may also change in a cyclic manner that cannot be represented by a single description. During a two-week office-based experiment [225], Milford and Wyeth noted that "the weakness is that the system deals rather inefficiently with cyclic changes such as day-night time cycles. Over a full night of operation, the pruning process gradually develops the experience map representation into one suited to localization at night time, somewhat hindering localization in the morning." These observations were corroborated by Ranganathan *et al.* [211], who stated that for an indoor office environment, consistently good localization through the 24-h cycle would require around three to four images per location. Rather than continuously remembering and forgetting information, the map should hold multiple representations of the area—either at a place level or at a whole-map level.

A place recognition system can use multiple maps of the same environment. In the work of Biber and Duckett, each map encoded a different timescale [226]. Some of these maps represented short-term memory and were updated frequently, while others were analogous to long-term memory and were not updated for hours, days, or weeks. Keeping maps that updated at different timescales ensured that old mapping data was not immediately overwritten by a temporary change in the environment. Instead static elements were reinforced over time, whilst transient events were filtered out. Place recognition was performed by selecting the local map that fitted the current sensor data best.

Systems that maintain multiple maps of the same environment may also add new map configurations only when they are necessary, rather than according to a pre-set timeframe [220]. Furthermore, Stachniss and Burgard [227] noted that not every place needs multiple representations—certain areas such as doorways may exhibit more change than the rest of the environment. Such areas may only possess a few key configurations—for example, a door may be open or closed—therefore, the world can be described sufficiently accurately using a finite number of submaps. Each region in which dynamic activity is observed was segmented from the rest of the map in a submap. Fuzzy k-means clustering was used with the Bayesian information criterion to determine the optimal number of typical configurations of this area. Using submaps to segregate dynamic areas allowed multiple environmental configurations where necessary while keeping the map manageable.

Elements of a scene that are moving when the robot observes them must be detected and may also be removed [229], [230]. However, there are often semistatic elements that are not obviously moving but which appear and disappear over time. While these elements can simply be removed as unreliable [69], [231], it is also possible that such elements may be temporarily useful for localization in specific parts of an environment [232]. Meyer-Delius *et al.* [232] used the example of a car park building, where the static elements such as the walls can be far away and are not distinctive, while the semistatic parked cars are many and relatively distinctive, and can be used for localization for a matter of hours or a day, before being forgotten. If this is the case, temporary maps are created when the robot observations do not match the expected results of the provided static map. The temporary maps are discarded when they fail to adequately match the robot observations over multiple consecutive time steps.

The systems presented above [220], [226], [227], [232] were designed for metric systems. Multiple representations can also be generated for appearance-based systems if multiple training runs are available. Johns and Yang [102] used feature co-occurrence maps generated during five training runs on a 20-km urban road-based dataset between 14:00 and 22:00. Localization can then be achieved on the same route at times interpolated between the five runs.

McManus *et al.* [120] used multiple training runs through an environment to learn scene signatures—locally distinctive elements of a place that are also stable over changes in appearance. For each location within the environment, image patches were selected that specifically demonstrate both distinctiveness

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 8. The varying appearance of a changing environment may require a system to store multiple representations of each place. This image (from [228]) shows the number of robot "experiences" stored during repeated traversals of a path over a number of months. While most places require five to ten experiences, some regions require as many as 30.

and stability. The selected patches were described using HOG descriptors [121] and used to train an SVM classifier for each location. Using scene signatures for each place allowed 100% correct place recognition in a 31 location dataset, while SURF features performed poorly, particular in rainy and foggy conditions.

If the appearance of the environment is assumed to be affected by a series of hidden periodic processes, spectral analysis such as Fourier analysis can be used to predict the most likely appearance of a location from multiple training passes at a particular time in the future. Krajnik *et al.* [233] learned and modeled these processes over an environment and demonstrated that this information can halve the number of place recognition errors when localizing three months later.

All of the systems described above share an underlying assumption—that the robot knows where it is sufficiently well to match different representations of the same location together, even if the representations are visually dissimilar. A map cannot be updated if the system does not know which location to update and, in a changing environment, it may not be possible to know exactly where the robot is. To avoid this assumption, Churchill and Newman proposed a plastic map formulation [15] that explicitly localizes within robot "experiences" rather than physical locations. A new experience is generated each time a robot visits a location that it does not recognize, and the map may implicitly have multiple representations of each location, depending on the difficulty of matching at that particular location (see Fig. 8). However, unlike the systems discussed previously, the multiple representations will not necessarily be linked together as the same physical place. The plastic map is more informative if the system can recognize and link more experiences together. However, it is a pragmatic approach that allows for graceful place recognition failure without catastrophic map collapse.

Retaining multiple representations of each location increases the place recognition search space and can decrease efficiency unless only a subset of representations is used for comparison. Because observations captured at similar times tend to demonstrate similar appearance characteristics, future potential matches can be probabilistically selected based on the system's current localization belief. Carlevaris-Bianco and Eustice [234] approximated the likelihood of two location exemplars being "co-observed" within a short time-frame with a Chow–Liu tree, while Linegar *et al.* [235] used "path memory" to select past experiences as candidate matches and improve place recognition without increasing computation time.

C. Recognizing Places in Changing Environments

Integrating appearance change into a place recognition system requires some key alterations to the belief generation process. First, as discussed above, changing environments require multiple representations of each place. If this is the case, a system may select the best map given its current sensor data [226] or it may try to predict the most likely appearance matches [18], [233]–[235].

Alternatively, the place recognition system may run multiple hypotheses in parallel. Churchill and Newman [15] assigned every saved experience its own localizer that reports whether or not the robot is successfully localized within that environment, while Morris *et al.* [220] performed filtering over possible map configurations as well as possible robot poses. Instead of selecting the single map that best matches the current sensor data, the system instead actively tracks the N best navigation hypotheses in multiple maps, while pending hypotheses are maintained and swapped out when an active hypothesis drops below the best pending hypothesis. Using multiple map hypotheses was reported to decrease the mean path error in an indoor office experiment by as much as 80%.

One factor for place recognition in changing environments is that topological information becomes more important as incoming sensor data becomes less reliable and more difficult to match to previous observations [118], [119]. It has been observed that matching image sequences rather than individual images can improve place recognition in general, and particularly in changing environments [14], [84], [118], [147].

Place recognition in changing environments benefits from exploiting the assumption that the system is not just passing through a particular place, but traversing the same or a very similar path through the environment. SeqSLAM [118] used image sequences to perform place recognition through particularly visually challenging environments. The original version assumed a similar velocity profile between traversals, while modified versions searched nonlinear paths as well as linear paths through the image similarity matrix [102] or used odometry input to linearize the signal [148]. Liu and Zhang [236] used a particle filter to improve the computation efficiency over the exhaustive search process and achieved a 10 times speed-up factor with equivalent performance at 100% precision.

Naseer *et al.* [119] exploited sequence information by formulating image matching as a minimum cost flow. Flow networks are directed graphs with a source node and a sink node, which for path-based place recognition represent the start of the traversal and the end of the traversal, respectively. By equating image comparison values to flow cost, the formulation found the optimal sequence through the environment. Differing velocity profiles were handled by allowing nodes to be either matching or hidden. Similarly, Hansen and Browning [237] used Hidden Markov Models to determine the most likely path through an environment using the Viterbi algorithm.

VIII. Conclusion

Visual place recognition has made great advances in the last 15 years, but we are still a long way from a universal place recognition system for robots that is robust and widely applicable across a range of robotic platforms and varying environments. Here, we highlight several promising avenues of ongoing and future research that are moving us closer toward this outcome.

Visual place recognition is benefitting from research in other fields, particularly the great strides being achieved in computer vision in the fields of deep learning, image classification, object recognition, and video description. While techniques such as CNNs depend on Big Data and Big Compute, techniques such as cloud robotics and online/offline processing paradigms could be exploited even by small, cheap mobile platforms. Developments in GPU hardware and novel camera sensors will inspire new concepts in place recognition as well as improving the efficiency and robustness of existing approaches.

Research in place recognition can also benefit from the ongoing research in object detection and scene classification. By exploiting object detections, it is possible to learn that objects such as buildings are useful for long-term place recognition, objects such as pedestrians should be ignored, and objects such as cars might be useful depending on the semantic and temporal context. An increased robustness to structural changes can be achieved by exploiting knowledge about which objects are dynamic or static and how that property depends on the temporal and semantic context—for example, cars in a parking garage can temporarily provide useful place recognition cues. Exploiting the expressiveness of CNNs by training or fine-tuning such networks specifically for the task of place recognition is a worthwhile direction for future research.

Visual place recognition systems can also exploit context. Although places change drastically in appearance, the relative location of places remains unchanged. This fact is integrated into belief generation modules by using location priors, recursive filtering, and path-based sequences of images, and the dependence on these techniques increases as the variation in the visual appearance of the environment increases. The use of other sources of contextual information also has the potential to improve place recognition capability—knowledge about the time of day, or the current weather conditions can also change how the place recognition system interprets the incoming visual data.

Semantic scene context can furthermore limit the search space for place recognition to ensure scalability towards long-term autonomy. Semantic context can support learning and predicting the changes in a scene and help increase robustness against environmental condition changes. Semantic mapping also has the potential to reduce memory requirements—imagine a house map only requiring words such as "kitchen," "bedroom," and "bathroom" to describe places—and current research in topic modeling, coresets and other semantic compression methods is already showing promise, as is the use of objects as high-level place recognition features.

Finally, what can visual place recognition offer to other research tasks? By necessity and opportunity, visual place recognition has taken up the challenge to solve condition invariant recognition to a degree that many fields have not, albeit under a more tightly constrained task specification than other tasks such as scene interpretation. The experience gained may have valuable applications, both in other robotic tasks such as object recognition and object classification in the wild, and in a diverse range of other areas including remote sensing, environmental monitoring, and other tasks that require recognition and identification in uncontrolled environments.

References

[1] G. Sibley, C. Mei, I. Reid, and P. Newman, "Vast-scale outdoor navigation using adaptive relative bundle adjustment," *Int. J. Robot. Res.*, vol. 29, no. 8, pp. 958–980, 2010.

[2] K. Konolige and M. Agrawal, "FrameSLAM: From bundle adjustment to real-time visual mapping," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1066–1077, Oct. 2008.

[3] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–7.

[4] M. Milford, G. Wyeth, and D. Prasser, "RatSLAM: A hippocampal model for simultaneous localization and mapping," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2004, vol. 1, pp. 403–408.

[5] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, vol. 2, pp. 1023–1029.

[6] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, 2008.

[7] A. Davison, I. Reid, N. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007.

[8] C. Harris and J. Pike, "3D positional integration from image sequences," *Image Vis. Comput.*, vol. 6, no. 2, pp. 87–90, 1988.

[9] J. Neira, M. Ribeiro, and J. Tardós, "Mobile robot localization and map building using monocular vision," in *Proc. 5th Symp. Intell. Robot. Syst.*, 1997, pp. 275–284.

[10] M. Bosse, R. Rikoski, J. Leonard, and S. Teller, "Vanishing points and three-dimensional lines from omni-directional video," in *Proc. IEEE Int. Conf. Image Process.*, Jun. 2002, vol. 3, pp. 513–516.

[11] R. Eustice, H. Singh, J. Leonard, M. Walter, and R. Ballard, "Visually navigating the RMS Titanic with SLAM information filters," presented at the Robot. Sci. Syst. Conf., Cambridge, MA, USA, Jun. 2005.

[12] R. Sim, P. Elinas, M. Griffin, and J. Little, "Vision-based SLAM using the Rao-Blackwellised particle filter," in *Proc. IJCAI Workshop Reason. Uncertainty Robot.*, 2005, pp. 9–16.

[13] W. Maddern, M. Milford, and G. Wyeth, "CAT-SLAM: Probabilistic localisation and mapping using a continuous appearance-based trajectory," *Int. J. Robot. Res.*, vol. 31, no. 4, pp. 429–451, 2012.

[14] C. Cadena, D. Gálvez-López, J. Tardós, and J. Neira, "Robust place recognition with stereo sequences," *IEEE Trans. Robot.*, vol. 28, no. 4, pp. 871–885, Aug. 2012.

[15] W. Churchill and P. Newman, "Experience-based navigation for long-term localisation," *Int. J. Robot. Res.*, vol. 32, no. 14, pp. 1645–1661, 2013.

[16] F. Dayoub, T. Morris, B. Upcroft, and P. Corke, "Vision-only autonomous navigation using topometric maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Nov. 2013, pp. 1923–1929.

[17] A. Murillo, G. Singh, J. Košecká, and J. Guerrero, "Localization in urban environments using a panoramic gist descriptor," *IEEE Trans. Robot.*, vol. 29, no. 1, pp. 146–160, Feb. 2013.

[18] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Robot. Auton. Syst.*, vol. 69, pp. 15–27, 2014.
[19] Y. Latif, G. Huang, J. Leonard, and J. Neira, "An online sparsity-cognizant loop-closure algorithm for visual navigation," presented at the Robot. Sci. Syst. Conf., Berkeley, CA, USA, Jul. 2014.
[20] E. Tolman, "Cognitive maps in rats and men," *Psychol. Rev.*, vol. 55, no. 4, pp. 189–208, 1948.
[21] A. Bennett, "Do animals have cognitive maps?" *J. Exp. Biol.*, vol. 199, no. 1, pp. 219–224, 1996.
[22] R. Jensen, "Behaviorism, latent learning, and cognitive maps: Needed revisions in introductory psychology textbooks," *Behav. Anal.*, vol. 29, no. 2, pp. 187–209, 2006.
[23] K. Lynch, *The Image of the City*. Cambridge, MA, USA: MIT Press, 1960.
[24] B. Kuipers, "Modeling spatial knowledge," *Cogn. Sci.*, vol. 2, no. 2, pp. 129–153, 1978.
[25] B. Kuipers and Y.-T. Byun, "A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations," *Robot. Auton. Syst.*, vol. 8, no. 12, pp. 47–63, 1991.
[26] B. Kuipers, "The spatial semantic hierarchy," *Artif. Intell.*, vol. 119, no. 12, pp. 191–233, 2000.
[27] F. Strumwasser, "Long-term recording from single neurons in brain of unrestrained mammals," *Science*, vol. 127, no. 3296, pp. 469–470, 1958.
[28] J. O'Keefe and J. Dostrovsky, "The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat," *Brain Res.*, vol. 34, no. 1, pp. 171–175, 1971.
[29] J. O'Keefe, "Place units in the hippocampus of the freely moving rat," *Exp. Neurol.*, vol. 51, no. 1, pp. 78–109, 1976.
[30] M. Wilson and B. McNaughton, "Dynamics of the hippocampal ensemble code for space," *Science*, vol. 261, no. 5124, pp. 1055–1058, 1993.
[31] J. O'Keefe and D. Conway, " Hippocampal place units in the freely moving rat: Why they fire where they fire," *Exp. Brain Res.*, vol. 31, no. 4, pp. 573–590, 1978.
[32] J. Taube, R. Muller, and J. Ranck, "Head-direction cells recorded from the postsubiculum in freely moving rats," *J. Neurosci.*, vol. 10, no. 2, pp. 420–447, 1990.
[33] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, pp. 801–806, Aug. 2005.
[34] E. Moser, E. Kropff, and M.-B. Moser, "Place cells, grid cells, and the brain's spatial representation system," *Annu. Rev. Neurosci.*, vol. 31, no. 1, pp. 69–89, 2008.
[35] K. Gothard, W. Skaggs, and B. McNaughton, "Dynamics of mismatch correction in the hippocampal ensemble code for space: interaction between path integration and environmental cues," *J. Neurosci.*, vol. 16, no. 24, pp. 8027–8040, 1996.
[36] A. Redish, E. Rosenzweig, J. Bohanick, B. McNaughton, and C. Barnes, "Dynamics of hippocampal ensemble activity realignment: Time versus space," *J. Neurosci.*, vol. 20, no. 24, pp. 9298–9309, 2000.
[37] R. Chatila and J. Laumond, "Position referencing and consistent world modeling for mobile robots," in *Proc. IEEE Int. Conf. Robot. Autom.*, Mar. 1985, vol. 2, pp. 138–145.
[38] R. Smith, M. Self, and P. Cheeseman, "A stochastic map for uncertain spatial relationships," in *Proc. 4th Int. Symp. Robot. Res.*, 1988, pp. 467–474.
[39] H. Moravec and A. Elfes, "High resolution maps from wide angle sonar," in *Proc. IEEE Int. Conf. Robot. Autom.*, Mar. 1985, vol. 2, pp. 116–121.
[40] R. Brooks, "Visual map making for a mobile robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, Mar. 1985, vol. 2, pp. 824–829.
[41] H. Durrant-Whyte, D. Rye, and E. Nebot, "Localization of autonomous guided vehicles," in *Proc. 8th Int. Symp. Robot. Res.*, 1995, pp. 613–625.
[42] H. Shatkay and L. Kaelbling, "Learning geometrically-constrained hidden Markov models for robot navigation: bridging the geometrical-topological gap," *J. Artif. Intell. Res.*, vol. 16, pp. 167–207, 2002.
[43] G. Dudek, M. Jenkin, E. Milios, and D. Wilkes, "Robotic exploration as graph construction," *IEEE Trans. Robot.*, vol. 7, no. 6, pp. 859–865, Dec. 1991.
[44] E. Fabrizi and A. Saffiotti, "Extracting topology-based maps from gridmaps," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, vol. 3, pp. 2972–2978.
[45] T. Bailey, E. Nebot, J. Rosenblatt, and H. Durrant-Whyte, "Robust distinctive place recognition for topological maps," presented at the Int. Conf. Field Service Robot., Pittsburgh PA, USA, 1999.
[46] P. Beeson, N. Jong, and B. Kuipers, "Towards autonomous topological place detection using the extended Voronoi graph," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2005, pp. 4373–4379.
[47] D. Kortenkamp, L. Baker, and T. Weymouth, "Using gateways to build a route map," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Jul. 1992, vol. 3, pp. 2209–2214.
[48] A. Ranganathan, "PLISS: Detecting and labeling places using online change-point detection," presented at the Robot. Sci. Syst. Conf., Zaragoza, Spain, Jun. 2010.
[49] G. Tsechpenakis, D. Metaxas, C. Neidle, and O. Hadjiliadis, "Robust online change-point detection in video sequences," in *Proc. Comput. Vis. Patt. Recog. Work.*, Jun. 2006, pp. 155–155.
[50] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. Comput. Vis. Patt. Recognit. (CVPR 2005), IEEE Comput. Soc. Conf.*, Jun. 2005, vol. 1, pp. 631–637.
[51] S. Esterby and A. El-Shaarawi, "Inference about the point of change in a regression model," *Appl. Statist.*, vol. 30, no. 3, pp. 277–285, 1981.
[52] A. Ranganathan and F. Dellaert, "Bayesian surprise and landmark detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 2017–2023.
[53] H. Korrapati, J. Courbon, Y. Mezouar, and P. Martinet, "Image sequence partitioning for outdoor mapping," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1650–1655.
[54] A. Chapoulie, P. Rives, and D. Filliat, "Appearance-based segmentation of indoors/outdoors sequences of spherical views," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Nov. 2013, pp. 1946–1951.
[55] L. Murphy and G. Sibley, "Incremental unsupervised topological place discovery," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 1312–1318.
[56] T. Nicosevici and R. Garcia, "Automatic visual bag-of-words for online robot navigation and mapping," *IEEE Trans. Robot.*, vol. 28, no. 4, pp. 886–898, Aug. 2012.
[57] T.-C. Chou and M. C. Chen, "Using incremental PLSI for threshold-resilient online event analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 3, pp. 289–299, Mar. 2008.
[58] M. Volkov, G. Rosman, D. Feldman, J. Fisher, and D. Rus, "Coresets for visual summarization with applications to loop closure," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2015, pp. 3638–3645.
[59] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for *k*-means, PCA and projective clustering," in *Proc. ACM-SIAM Symp. Discrete Alg.*, 2013, pp. 1434–1453.
[60] R. Paul, D. Rus, and P. Newman, "How was your day? Online visual workspace summaries using incremental clustering in topic space," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 4058–4065.
[61] Y. Girdhar and G. Dudek, "Efficient on-line data summarization using extremum summaries," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 3490–3496.
[62] R. Paul, D. Feldman, D. Rus, and P. Newman, "Visual precis generation using coresets," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 1304–1311.
[63] Y. Girdhar, P. Giguère, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," *Int. J. Robot. Res.*, vol. 33, no. 4, pp. 645–657, 2014.
[64] D. Filliat and J.-A. Meyer, "Map-based navigation in mobile robots—. A review of localisation strategies," *Cogn. Syst. Res.*, vol. 4, no. 4, pp. 243–282, 2003.
[65] M. Walter, S. Hemachandra, B. Homberg, S. Tellex, and S. Teller, "Learning semantic maps from natural language descriptions," presented at the Robot. Sci. Syst. Conf., Berlin, Germany, Jun. 2013.
[66] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 1, pp. 273–280.
[67] D. Filliat, E. Battesti, S. Bazeille, G. Duceux, A. Gepperth, L. Harrath, I. Jebari, R. Pereira, A. Tapus, C. Meyer, S.-H. Ieng, R. Benosman, E. Cizeron, J.-C. Mamanna, and B. Pothier, "RGBD object recognition and visual texture classification for indoor semantic mapping," in *Proc. IEEE Int. Conf. Tech. Prac. Robot Appl.*, Apr. 2012, pp. 127–132.
[68] A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 3515–3522.
[69] R. Biswas, B. Limketkai, S. Sanner, and S. Thrun, "Towards object mapping in non-stationary environments with mobile robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2002, vol. 1, pp. 1014–1019.
[70] R. Salas-Moreno, R. Newcombe, H. Strasdat, P. Kelly, and A. Davison, "SLAM++: Simultaneous localisation and mapping at the level of ob-

jects," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1352–1359.
[71] J. Wu, H. Christensen, and J. Rehg, "Visual place categorization: Problem, dataset, and algorithm," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2009, pp. 4763–4770.
[72] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.
[73] H. Bay, T. Tuytelaars, and L. Van Gool, " SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.
[74] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," in *Visual Perception – Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception*. New York, NY, USA: Elsevier, 2006, pp. 23–36.
[75] A. Oliva and A. Torralba, " Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[76] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int. J. Rob. Res.*, vol. 21, no. 8, pp. 735–758, 2002.
[77] H. Andreasson and T. Duckett, "Topological localization for mobile robots using omni-directional vision and local features," presented at the 5th IFAC Symp. Intell. Auton. Veh., Lisbon, Portugal, 2004.
[78] E. Stumm, C. Mei, and S. Lacroix, "Probabilistic place recognition with covisibility maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Nov. 2013, pp. 4158–4163.
[79] J. Košecká, F. Li, and X. Yang, "Global localization and relative positioning based on scale-invariant keypoints," *Robot. Auton. Syst.*, vol. 52, no. 1, pp. 27–38, 2005.
[80] P. Newman and K. Ho, "SLAM—Loop closing with visually salient features," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2005, pp. 635–642.
[81] A. Gil, O. Reinoso, O. Mozos, C. Stachniss, and W. Burgard, "Improving data association in vision-based SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2006, pp. 2076–2081.
[82] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Incremental vision-based topological SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2008, pp. 1031–1036.
[83] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1027–1037, Oct. 2008.
[84] K. Ho and P. Newman, " Detecting loop closure with scene sequences," *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 261–286, 2007.
[85] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, vol. 1, pp. 525–531.
[86] A. Murillo, J. Guerrero, and C. Sagues, "SURF features for efficient robot localization with omnidirectional images," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3901–3907.
[87] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *Int. J. Robot. Res.*, vol. 30, no. 9, pp. 1100–1123, 2011.
[88] M. Agrawal, K. Konolige, and M. Blas, " CenSurE: Center surround extremas for realtime feature detection and matching," in *Proc. 10th Eur. Conf. Comput. Vision*, 2008, pp. 102–115.
[89] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid, "A constant-time efficient stereo SLAM system," presented at the Brit. Mach. Vis. Conf., London, U.K., 2009.
[90] E. Rosten and T. Drummond, " Machine learning for high-speed corner detection," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 430–443.
[91] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.
[92] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 2, pp. 1470–1477.
[93] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 2, pp. 524–531.
[94] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley, 1999.
[95] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2006, vol. 2, pp. 2161–2168.
[96] J. Wang, H. Zha, and R. Cipolla, "Combining interest points and edges for content-based image retrieval," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2005, vol. 3, pp. III–1256–9.
[97] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3921–3926.
[98] R. Paul and P. Newman, "FAB-MAP 3D: Topological mapping with spatial and visual appearance," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 2649–2656.
[99] P. Newman, G. Sibley, M. Smith, M. Cummins, A. Harrison, C. Mei, I. Posner, R. Shade, D. Schroeter, L. Murphy, W. Churchill, D. Cole, and I. Reid, "Navigating, recognizing and describing urban spaces with vision and lasers," *Int. J. Robot. Res.*, vol. 28, no. 11–12, pp. 1406–1433, 2009.
[100] C. Valgren and A. Lilienthal, "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments," *Robot. Auton. Syst.*, vol. 58, no. 2, pp. 149–156, 2010.
[101] C. Cadena and J. Neira, "A learning algorithm for place recognition," presented at the ICRA Workshop Long-Term Autonomy, Shanghai, China, 2011.
[102] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 3212–3218.
[103] M. Milford, W. Scheirer, E. Vig, A. Glover, O. Baumann, J. Mattingley, and D. Cox, "Condition-invariant, top-down visual place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 5571–5577.
[104] B. Kröse, N. Vlassis, R. Bunschoten, and Y. Motomura, "A probabilistic model for appearance-based robot localization," *Image Vis. Comput.*, vol. 19, no. 6, pp. 381–391, 2001.
[105] P. Lamon, I. Nourbakhsh, B. Jensen, and R. Siegwart, "Deriving and matching image fingerprint sequences for mobile robot localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2001, vol. 2, pp. 1609–1614.
[106] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[107] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vis. Conf.*, 1988, pp. 147–151.
[108] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1635–1642.
[109] N. Sünderhauf and P. Protzel, "BRIEF-Gist—Closing the loop by simple means," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2011, pp. 1234–1241.
[110] A. Murillo and J. Košecká, "Experiments in place recognition using gist panoramas," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Sep. 2009, pp. 2196–2203.
[111] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Trans. Robot.*, vol. 25, no. 4, pp. 861–873, Aug. 2009.
[112] G. Singh and J. Košecká, "Visual loop closing using gist descriptors in manhattan world," presented at the ICRA Worksop Omnidirectional Robot Vision, Anchorage, AK, 2010.
[113] Y. Liu and H. Zhang, "Visual loop closure detection with a compact image descriptor," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2012, pp. 1051–1056.
[114] C. Mei, G. Sibley, and P. Newman, "Closing loops without places," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2010, pp. 3738–3744.
[115] S. Lynen, M. Bosse, P. Furgale, and R. Siegwart, "Placeless place-recognition," in *Proc. Int. Conf. 3D Vis.*, Dec. 2014, vol. 1, pp. 303–310.
[116] M. Milford and G. Wyeth, "Mapping a suburb with a single camera using a biologically inspired SLAM system," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1038–1053, Oct. 2008.
[117] P. Furgale and T. Barfoot, "Visual teach and repeat for long-range rover autonomy," *J. Field Robot.*, vol. 27, no. 5, pp. 534–560, 2010.
[118] M. Milford and G. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1643–1649.
[119] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2564–2570.
[120] C. McManus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," presented at the Robot. Sci. Syst. Conf., Berkeley, CA, USA, Jul. 2014.
[121] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.
[122] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with ConvNet land-

marks: Viewpoint-robust, condition-robust, training-free," presented at the Robot. Sci. Syst. Conf., Rome, Italy, Jul. 2015.

[123] C. Zitnick and P. Dollár, "Edge Boxes: Locating object proposals from edges," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[124] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[125] J. Neira, A. Davison, and J. Leonard, "Guest editorial—Special issue on visual SLAM," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 929–931, Oct. 2008.

[126] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the RGB-D SLAM system," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1691–1696.

[127] M. Labbe and F. Michaud, "Online global loop closure detection for large-scale multi-session graph-based SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2014, pp. 2661–2666.

[128] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. Leonard, and J. McDonald, "Real-time large-scale dense RGB-D SLAM with volumetric fusion," *Int. J. Robot. Res.*, vol. 34, nos. 4/5, pp. 598–626, 2015.

[129] A. Davison and D. Murray, "Simultaneous localization and map-building using active vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 865–880, Jul. 2002.

[130] K. Konolige and J. Bowman, "Towards lifelong visual maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2009, pp. 1156–1163.

[131] K. Konolige, J. Bowman, J. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua, "View-based maps," *Int. J. Robot. Res.*, vol. 29, no. 8, pp. 941–957, 2010.

[132] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, " Bundle adjustment—A modern synthesis," in *Vision Algorithms: Theory and Practice*. Berlin, Germany: Springer, 2000, pp. 298–372.

[133] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. IEEE Int. Symp. Mixed Aug. Reality*, Nov. 2007, pp. 225–234.

[134] R. Newcombe, S. Lovegrove, and A. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2320–2327.

[135] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[136] R. Mur-Artal, J. Montiel, and J. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[137] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[138] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. IEEE Int. Symp. Mixed Aug. Reality*, Oct. 2011, pp. 127–136.

[139] R. Finman, L. Paull, and J. Leonard, "Toward object-based place recognition in dense RGB-D maps," presented at the ICRA Workshop Visual Place Recognition in Changing Environments, Seattle, WA, 2015.

[140] M. Bosse, P. Newman, J. Leonard, and S. Teller, "Simultaneous localization and map building in large-scale cyclic environments using the Atlas framework," *Int. J. Robot. Res.*, vol. 23, no. 12, pp. 1113–1139, 2004.

[141] F. Li and J. Košecká, "Probabilistic location recognition using reduced feature set," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2006, pp. 3405–3410.

[142] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM—FAB-MAP 2.0," presented at the Robot. Sci. Syst. Conf., Seattle, WA, USA, Jun. 2009.

[143] M. Mohan, D. Gálvez-López, C. Monteleoni, and G. Sibley, "Environment selection and hierarchical place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2015, pp. 5487–5494.

[144] H. Andreasson, A. Treptow, and T. Duckett, "Localization for mobile robots using panoramic vision, local features and particle filter," in *Proc.IEEE Int. Conf. Robot. Autom.*, Apr. 2005, pp. 3348–3353.

[145] W. Maddern, M. Milford, and G. Wyeth, "Capping computation time and storage requirements for appearance-based localization with CAT-SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 822–827.

[146] D. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[147] M. Milford, "Vision-based place recognition: How low can you go?" *Int. J. Robot. Res.*, vol. 32, no. 7, pp. 766–789, 2013.

[148] E. Pepperell, P. Corke, and M. Milford, "All-environment visual place recognition with SMART," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 1612–1618.

[149] T. Duckett, S. Marsland, and J. Shapiro, "Learning globally consistent maps by relaxation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, vol. 4, pp. 3841–3846.

[150] D. Filliat and J.-A. Meyer, "Global localization and topological map learning for robot navigation," in *Proc. Int. Conf. Sim. Adapt. Behav.*, 2002, pp. 131–140.

[151] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Visual topological slam and global localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 4300–4305.

[152] S. Bazeille and D. Filliat, "Incremental topo-metric slam using vision and robot odometry," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 4067–4073.

[153] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid, " RSLAM: A system for large-scale mapping in constant-time using stereo," *Int. J. Comput. Vis.*, vol. 94, no. 2, pp. 198–214, 2011.

[154] G. Sibley, C. Mei, I. Reid, and P. Newman, "Adaptive relative bundle adjustment," presented at the Robot. Sci. Syst. Conf., Seattle, WA, USA, Jun. 2009.

[155] K. Konolige, E. Marder-Eppstein, and B. Marthi, "Navigation in hybrid metric-topological maps," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3041–3047.

[156] P. Beeson, J. Modayil, and B. Kuipers, "Factoring the mapping problem: Mobile robot map-building in the hybrid spatial semantic hierarchy," *Int. J. Robot. Res.*, vol. 29, no. 4, pp. 428–459, 2010.

[157] H. Johannsson, M. Kaess, M. Fallon, and J. Leonard, "Temporally scalable visual SLAM using a reduced pose graph," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 54–61.

[158] I. Mahon, S. Williams, O. Pizarro, and M. Johnson-Roberson, "Efficient view-based SLAM using visual loop closures," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1002–1014, Oct. 2008.

[159] R. Eustice, H. Singh, J. Leonard, and M. Walter, "Visually mapping the RMS Titanic: Conservative covariance estimates for SLAM information filters," *Int. J. Robot. Res.*, vol. 25, no. 12, pp. 1223–1242, 2006.

[160] F. Lu and E. Milios, " Globally consistent range scan alignment for environment mapping," *Auton. Robots*, vol. 4, no. 4, pp. 333–349, 1997.

[161] J.-S. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *Proc. IEEE Int. Symp. Comput. Intell. Robot. Autom.*, 1999, pp. 318–325.

[162] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[163] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): Part II," *IEEE Robot. Autom. Mag.*, vol. 13, no. 3, pp. 108–117, Sep. 2006.

[164] S. Thrun and J. Leonard, " Simultaneous localization and mapping," in *Springer Handbook of Robotics*. Berlin, Germany: Springer, 2008, pp. 871–889.

[165] R. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *Int. J. Robot. Res.*, vol. 5, no. 4, pp. 56–68, 1986.

[166] J. Guivant and E. Nebot, "Optimization of the simultaneous localization and map-building algorithm for real-time implementation," *IEEE Trans. Robot.*, vol. 17, no. 3, pp. 242–257, Jun. 2001.

[167] M. Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem," *IEEE Trans. Robot.*, vol. 17, no. 3, pp. 229–241, Jun. 2001.

[168] M. Montemerlo and S. Thrun, *FastSLAM: A Scalable Method for the Simultaneous Localization and Mapping Problem in Robotics*. Berlin, Germany: Springer, 2007.

[169] S. Thrun and M. Montemerlo, "The graph SLAM algorithm with applications to large-scale mapping of urban structures," *Int. J. Robot. Res.*, vol. 25, no. 5-6, pp. 403–429, 2006.

[170] E. Olson, J. Leonard, and S. Teller, "Fast iterative alignment of pose graphs with poor initial estimates," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2006, pp. 2262–2269.

[171] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Trans. Robot.*, vol. 24, no. 6, pp. 1365–1378, Dec. 2008.

[172] E. Eade and T. Drummond, "Edge landmarks in monocular SLAM," in *Proc. Brit. Mach. Vis. Conf.*, 2006, pp. 7–16.

[173] E. Eade and T. Drummond, "Scalable monocular SLAM," in *Proc. Comput. Vis. Patt. Recog.*, Jun. 2006, vol. 1, pp. 469–476.

[174] E. Eade and T. Drummond, "Unified loop closing and recovery for real time monocular SLAM," presented at the Brit. Mach. Vis. Conf. Leeds, U.K., 2008.

[175] A. Jacobson, Z. Chen, and M. Milford, "Autonomous multisensor calibration and closed-loop fusion for SLAM," *J. Field Robot.*, vol. 32, no. 1, pp. 85–122, 2015.

[176] D. Gálvez-López and J. Tardós, "Real-time loop detection with bags of binary words," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2011, pp. 51–58.

[177] N. Vlassis and B. Kröse, "Robot environment modeling via principal component regression," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 1999, vol. 2, pp. 677–682.

[178] N. Vlassis and B. Kröse, "Mixture conditional density estimation with the EM algorithm," in *Proc. Int. Conf. Artif. Neural Netw.*, 1999, vol. 2, pp. 821–825.

[179] B. Ripley and N. Hjort, *Pattern Recognition and Neural Networks*, 1st ed. New York, NY, USA: Cambridge Univ. Press, 1995.

[180] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Real-time visual loop-closure detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2008, pp. 1842–1847.

[181] C. Siagian and L. Itti, "Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2007, pp. 1723–1730.

[182] E. Garcia-Fidalgo and A. Rodriguez, "Probabilistic appearance-based mapping and localization using visual features," in *Proc. Iberian Conf. Pattern Recog. Image Anal.*, 2013, pp. 277–285.

[183] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.

[184] M. Cummins and P. Newman, "Probabilistic appearance based navigation and loop closing," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 2042–2048.

[185] R. Paul and P. Newman, "Self-help: Seeking out perplexing images for ever improving topological mapping," *Int. J. Robot. Res.*, vol. 32, no. 14, pp. 1742–1766, 2013.

[186] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[187] E. Olson, "Recognizing places using spectrally clustered local matches," *Robot. Auton. Syst.*, vol. 57, no. 12, pp. 1157–1172, 2009.

[188] C. Giovannangeli, P. Gaussier, and G. Desilles, "Robust mapless outdoor vision-based navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2006, pp. 3293–3300.

[189] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *Int. J. Robot. Res.*, vol. 32, no. 7, pp. 826–840, 2013.

[190] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2012, pp. 1879–1884.

[191] Y. Latif, C. Cadena, and J. Neira, "Robust loop closing over time for pose graph SLAM," *Int. J. Robot. Res.*, vol. 32, no. 14, pp. 1611–1626, 2013.

[192] C. McManus, W. Churchill, W. Maddern, A. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 901–906.

[193] T. Barfoot, J. Kelly, and G. Sibley, "Guest editorial—Special issue on long-term autonomy," *Int. J. Robot. Res.*, vol. 32, no. 14, pp. 1609–1610, 2013.

[194] C. Valgren and A. Lilienthal, "SIFT, SURF and seasons: Long-term outdoor localization using local features," presented at the 3rd Eur. Conf. Mob. Robot., Freiburg, Germany, 2007.

[195] P. Ross, A. English, D. Ball, B. Upcroft, G. Wyeth, and P. Corke, "A novel method for analysing lighting variance," in *Proc. Robot. Autom. (ACRA 2013), Australas. Conf.*, 2013.

[196] P. Ross, A. English, D. Ball, and P. Corke, "A method to quantify a descriptor's illumination variance," presented at the Australas. Conf. Robot. Autom., Melbourne, Australia, 2014.

[197] M. Milford, I. Turner, and P. Corke, "Long exposure localization in darkness using consumer cameras," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 3755–3761.

[198] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons," presented at the ICRA Workshop Long-Term Autonomy, Karlsruhe, Germany, 2013.

[199] S. Nuske, J. Roberts, and G. Wyeth, "Robust outdoor visual localization using a three-dimensional-edge map," *J. Field Robot.*, vol. 26, no. 9, pp. 728–756, 2009.

[200] P. Borges, R. Zlot, M. Bosse, S. Nuske, and A. Tews, "Vision-based localization using an edge map extracted from 3D laser range data," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 4902–4909.

[201] P. Corke, R. Paul, W. Churchill, and P. Newman, "Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Nov. 2013, pp. 2085–2092.

[202] C. McManus, P. Furgale, and T. Barfoot, "Towards lighting-invariant visual navigation: An appearance-based approach using scanning laser-rangefinders," *Robot. Auton. Syst.*, vol. 61, no. 8, pp. 836–852, 2013.

[203] W. Maddern and S. Vidas, "Towards robust night and day place recognition using visible and thermal imaging," presented at the RSS Work. Alt. Sens. Tech. Robot. Percep., Sydney, Australia, Jul. 2012.

[204] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. Comput. Vis. Pattern Recog. Workshop*, Jun. 2014, pp. 512–519.

[205] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1717–1724.

[206] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," presented at the Australas. Conf. Robot. Autom., Melbourne, Australia, 2014.

[207] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2015, pp. 4297–4304.

[208] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," presented at the ICRA Workshop Vis. Place Recog., Hong Kong, 2014.

[209] N. Jacobs, N. Roman, and R. Pless, "Consistent temporal variations in many outdoor scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–6.

[210] A. Abrams, E. Feder, and R. Pless, "Exploratory analysis of time-lapse imagery with fast subset PCA," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2011, pp. 336–343.

[211] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 3791–3798.

[212] A. Mikulík, M. Perdoch, O. Chum, and J. Matas, "Learning a fine vocabulary," in *Proc. Eur. Conf. Comput. Vis*, vol. 3, pp. 1–14, 2010.

[213] N. Carlevaris-Bianco and R. Eustice, "Learning visual feature descriptors for dynamic lighting conditions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2014, pp. 2769–2776.

[214] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 1994, pp. 737–744.

[215] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[216] S. Lowry, M. Milford, and G. Wyeth, "Transforming morning to afternoon using linear regression techniques," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 3950–3955.

[217] S. Lowry, G. Wyeth, and M. Milford, "Unsupervised online learning of condition-invariant images for place recognition," presented at the Australas. Conf. Robot. Autom., Melbourne, Australia, 2014.

[218] P. Biber and T. Duckett, "Dynamic maps for long-term operation of mobile service robots," presented at the Robot. Sci. Syst. Conf., Cambridge, MA, USA, Jun. 2005.

[219] F. Dayoub and T. Duckett, "An adaptive appearance-based map for long-term topological localization of mobile robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2008, pp. 3364–3369.

[220] T. Morris, F. Dayoub, P. Corke, G. Wyeth, and B. Upcroft, "Multiple map hypotheses for planning and navigating in non-stationary environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 2765–2770.

[221] T. Bailey, “Mobile robot localisation and mapping in extensive outdoor environments,” Ph.D. dissertation, Univ. Sydney, Sydney, Australia, 2002.

[222] J. Andrade-Cetto and A. Sanfeliu, “Concurrent map building and localization on indoor dynamic environments,” *Int. J. Pattern Recog. Artif. Intell.*, vol. 16, no. 03, pp. 361–374, 2002.

[223] A. Hafez, M. Singh, K. Krishna, and C. Jawahar, “Visual localization in highly crowded urban environments,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Nov. 2013, pp. 2778–2783.

[224] E. Johns and G.-Z. Yang, “Generative methods for long-term place recognition in dynamic scenes,” *Int. J. Comput. Vis.*, vol. 106, no. 3, pp. 297–314, 2014.

[225] M. Milford and G. Wyeth, “Persistent navigation and mapping using a biologically inspired SLAM system,” *Int. J. Robot. Res.*, vol. 29, no. 9, pp. 1131–1153, 2010.

[226] P. Biber and T. Duckett, “Experimental analysis of sample-based maps for long-term SLAM,” *Int. J. Robot. Res.*, vol. 28, no. 1, pp. 20–33, 2009.

[227] C. Stachniss and W. Burgard, “Mobile robot mapping and localization in non-static environments,” presented at the Nat. Conf. Artif. Intell., Pittsburgh, PA, USA, 2005.

[228] W. Churchill and P. Newman, “Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation,” in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 4525–4532.

[229] C.-C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, “Simultaneous localization, mapping and moving object tracking,” *Int. J. Robot. Res.*, vol. 26, no. 9, pp. 889–916, 2007.

[230] J. Dong, S. Wijesoma, and A. Shacklock, “Extended Rao-Blackwellised genetic algorithmic filter SLAM in dynamic environment with raw sensor measurement,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2007, pp. 1473–1478.

[231] D. Wolf and G. Sukhatme, “ Mobile robot simultaneous localization and mapping in dynamic environments,” *Auton. Robots*, vol. 19, no. 1, pp. 53–65, 2005.

[232] D. Meyer-Delius, J. Hess, G. Grisetti, and W. Burgard, “Temporary maps for robust localization in semi-static environments,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2010, pp. 5750–5755.

[233] T. Krajnik, J. Fentanes, O. Mozos, T. Duckett, J. Ekekrantz, and M. Hanheide, “Long-term topological localisation for service robots in dynamic environments using spectral maps,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2014, pp. 4537–4542.

[234] N. Carlevaris-Bianco and R. Eustice, “Learning temporal co-observability relationships for lifelong robotic mapping,” presented at the IROS Worksop Lifelong Learn. Mob. Robot. Appl., Vilamoura, Portugal, Oct. 2012.

[235] C. Linegar, W. Churchill, and P. Newman, “Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation,” in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2015, pp. 90–97.

[236] Y. Liu and H. Zhang, “Towards improving the efficiency of sequence-based SLAM,” in *Proc. IEEE Int. Conf. Mechatron. Autom.*, Aug. 2013, pp. 1261–1266.

[237] P. Hansen and B. Browning, “Visual place recognition using HMM sequence matching,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2014, pp. 4549–4555.

Stephanie Lowry received the B.Sc. (Hons.) degree in logic and computation and the M.Sc. degree in computer science, both from Victoria University of Wellington, Wellington, New Zealand, and the Ph.D. degree in engineering from Queensland University of Technology, Brisbane, Australia, in 2014.

She was previously a Postdoctoral Researcher with the Australian Research Council Centre of Excellence in Robotic Vision, Queensland University of Technology. In June 2015, she joined the Centre for Applied Autonomous Sensor Systems, Örebro University, Örebro, Sweden. Her research interests include long-term robot autonomy and the application of machine learning to robotic vision.

Dr. Lowry received the 2015 Siganto Foundation Medal for excellence in engineering Ph.D. research.

Niko Sünderhauf received the Ph.D. degree in 2012 and the Diploma (M.Sc.) degree in computer science in 2006, both from Technische Universität Chemnitz, Chemnitz, Germany.

He was a Research Fellow with Technische Universität Chemnitz between 2006 and 2014. In March 2014, he joined Queensland University of Technology, Brisbane, Australia, and is currently a Research Fellow with the Australian Research Council Centre of Excellence in Robotic Vision. His research interests include robust visual perception, place recognition in changing conditions, deep learning, SLAM, long-term autonomy, and probabilistic estimation with graphical models. Apart from mobile robotics, his research covers robust methods for sensor fusion, especially NLOS-mitigation for satellite-based localization systems.

Paul Newman (F’15) received the M.Eng. degree in engineering science from University of Oxford, Oxford, U.K., in 1995 and the Ph.D. degree in autonomous navigation from the Australian Center for Field Robotics, University of Sydney, Sydney, Australia.

He is currently a BP Professor of information engineering with the Department of Engineering Science, University of Oxford. He heads the Oxford Mobile Robotics Research Group. In 1999, he began to work in the commercial subsea navigation industry in the U.K. In late 2000, he joined the Department of Ocean Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA, where as a Postdoctoral Researcher and later a Research Scientist he worked on algorithms and software for robust autonomous navigation for both land and subsea agents. In early 2003, he returned to University of Oxford as a Departmental Lecturer in engineering science before being appointed to a University Lectureship in 2005. His research interests include autonomous navigation, especially over large spatial and temporal scales.

John J. Leonard (F’13) received the B.S.E. degree in electrical engineering and science from University of Pennsylvania, Philadelphia, PA, USA, and the D.Phil. degree in engineering science from University of Oxford, Oxford, U.K., in 1994.

He is a Professor of mechanical and ocean engineering with the Department of Mechanical Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. He is also a Member of the MIT Computer Science and Artificial Intelligence Laboratory. His research interests include the problems of navigation, mapping, and persistent autonomy for autonomous mobile robots operating in unstructured environments.

David Cox received the A.B. degree in biology and psychology from Harvard University, Cambridge, MA, USA, in 2000 and the Ph.D. degree in computational neuroscience from Massachusetts Institute of Technology, Cambridge, in 2007.

He is currently an Assistant Professor of molecular and cellular biology and of computer science with Harvard University. He was previously a Junior Fellow with the Rowland Institute, Harvard University, a multidisciplinary institute focused on high-risk high-reward scientific research at the boundaries of traditional fields. His laboratory seeks to understand the computational underpinnings of high-level visual processing through concerted efforts in both reverse- and forward-engineering. To this end, his group employs a wide range of experimental techniques (ranging from microelectrode recordings in living brains to visual psychophysics in humans) to probe natural systems, while at the same time actively developing practical computer vision systems based on what is learned about the brain.

Peter Corke (F'07) received the Bachelor's of Engineering, Master's of Engineering, and Ph.D. degrees from University of Melbourne, Melbourne, Australia.

He is currently the Director of the Australian Research Council Centre of Excellence in Robotic Vision, and a Professor of robotics and control with the Queensland University of Technology (QUT), Brisbane, Australia. His research interests include robotic vision, flying robots, and robots for agriculture. He was with University of Melbourne, first as a Research Assistant and later as a Lecturer. In 1984, he commenced with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), where he founded the Autonomous Systems Laboratory, CSIRO ICT Centre, and served as Research Director from 2004 to 2007. He was a Senior Principal Research Scientist when he left to take up a chair with QUT in 2010.

Dr. Corke was the Editor-in-Chief of *IEEE Robotics and Automation Magazine* from 2009 to 2013 and is a Founding Editor of *Journal of Field Robotics*.

Michael J. Milford (S'06–M'07) received the Ph.D. degree in electrical engineering and the Bachelor's of Mechanical and Space Engineering degree from University of Queensland, Brisbane, Australia.

He is currently an Associate Professor and Australian Research Council Future Fellow with Queensland University of Technology (QUT), Brisbane, Australia, and a Chief Investigator for the Australian Centre of Excellence for Robotic Vision. He was a Research Fellow on the Thinking Systems Project with Queensland Brain Institute on the Thinking Systems Project until 2010, when he became a Lecturer with QUT. He conducts interdisciplinary research into navigation across the fields of robotics, neuroscience, and computer vision.

Dr. Milford received an inaugural Australian Research Council Discovery Early Career Researcher Award in 2012 and became a Microsoft Research Faculty Fellow in 2013.