4

Visual Place Recognition for Simultaneous Localization and Mapping

Konstantinos A. Tsintotas*, Loukas Bampis and Antonios Gasteratos

Democritus University of Thrace, School of Engineering, Vas. Sofias 12, Xanthi, Greece

Abstract

As a mobile robot, e.g., an aerial, underwater, or ground-moving vehicle, navigates through an unknown environment, it has to construct a map of its surroundings and simultaneously estimate its pose within this map. This technique is widely known in the robotics community as simultaneous localization and mapping (SLAM). During SLAM, a fundamental feature is loops' detection, i.e., areas earlier visited by the robot, allowing consistent map generation. Due to this reason, a place recognizer is adopted, which aims to associate the current robot's environment observation with one belonging in the map. In SLAM, visual place recognition formulates a solution, permitting loops' detection using only the scene's appearance. The main components of such a framework's structure are the image processing module, the map, and the belief generator. In this chapter, the reader is initially familiarized with each part while several visual place recognition frameworks paradigms follow. The evaluation steps for measuring the system's performance, including the most popular metrics and datasets, are also presented. Finally, their experimental results are discussed.

Keywords: Mobile robot, aerial, underwater, moving vehicle, SLAM, sensory data, visual recognition, image processing

4.1 Introduction

As a wide range of applications, such as search and rescue [15, 21, 25, 30, 84], space [20, 49], inspection [22, 23, 95] and underwater exploration [41, 66,

**Corresponding author*: ktsintot@pme.duth.gr

Romil Rawat, Purvee Bhardwaj, Upinder Kaur, Shrikant Telang, Mukesh Chouhan and K. Sakthidasan Sankaran (eds.) Autonomous Vehicles Volume 2: Smart Vehicles, (47–80) © 2023 Scrivener Publishing LLC

103], demand autonomous robots, accurate navigation is more than necessary for an intelligent system to accomplish its assigned tasks. Simultaneous localization and mapping (SLAM) [69], i.e., a robot's capability to incrementally construct a map of its working environment and subsequently estimate its position in it, has become the core of autonomous navigation over the last three decades when global positioning information is missing [24]. However, drift is inevitably accumulated over time, given the sensor signals' noise and the absence of position measurements. Hence, SLAM needs to identify when the robot revisits a previously traversed location and recall it. Thus, the system's drift error and uncertainty regarding the estimated position and orientation (pose) can be bounded and rectified, allowing consistent map generation.

> This process is widely known as loop closure detection and is achieved via a place recognition pipeline responsible for associating the incoming sensory data (query) with the map (database).

Several techniques were used for mapping the operating environment in the early years, such as range and bearing sensors, *viz.*, lasers, radars, and sonars. However, due to the available computational power, which has increased over the late years, and the findings of how animals navigate using vision [71], mapping was pushed from other sensors to vision-based ones [65]. Nowadays, such cameras are successfully utilized for mapping trajectories of up to 1000 km [40]. Beyond the sensor's low cost and its applicability to various mobile platforms, especially the ones with restricted computational abilities, e.g., unmanned aerial vehicles (UAVs) [61, 100], the main reason for its utilization is related to the rich textural information presented in images [46, 47], which provide a significant advantage over the other sensors permitting to capture the environment's appearance with high distinctiveness effectively [38]. Not surprisingly, modern robotic

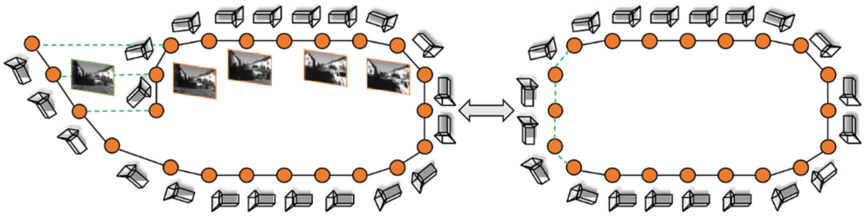


Figure 4.1 A representative example of a pose graph visual place recognition system in a simultaneous localization and mapping framework. As the global positioning information is not available, the robot generates an uncertain estimation about its position (left). When loops are detected in the traversed route, the internal map is rectified, allowing consistent map generation (right).

navigation systems are based on visual place recognition algorithms to detect loop closures [10, 11, 14, 28, 90, 93, 96] (see Figure 4.1).

As these methods seek a known location in the traversed route through the incoming visual information, at this point, it is reasonable to distinguish visual place recognition from image classification and image retrieval. More specifically, the former concerns the problem of categorizing a query image into a known class, while the latter tries to detect the most identical instances of the same class in a database. On the contrary, visual place recognition searches for similar images to the recent scene, which may belong in the same scene category but come from different places on the map. Hence, image processing holds a vital role in the system's performance since visual representations for classification tasks may not perform accordingly for place recognition and vice versa. Furthermore, as image processing efficiency heavily affects the system's confidence, mapping and matching techniques also play an important role in the final decision. In general, a visual place recognition system contains three key components [57]:

- An image processing module for interpreting the camera data.
- A map that represents the robot's knowledge about the world.
- A belief generator that decides if the robot navigates in a familiar way or not. Its decision is based on combining the incoming sensory data and the map.

In this chapter, we approach the construction of such a system. We also present several pipelines, including paradigms that use different image processing, mapping, and matching approaches. Recent advances are illustrated through the exemplar systems. After comprehending the chapter, the reader will know how to formulate a custom visual place recognition framework.

4.2 The Structure for a Visual Place Recognition System

With the aim to achieve visual place recognition, a workflow of processes is executed as summarized in the schematic of Figure 4.2. The pipeline comprises three parts: i) the image processing of the incoming visual data, ii) the trajectory mapping, and iii) the decision making through a belief generator. Lastly, as the place recognition for SLAM has to work online, it needs to update the map during navigation. Each part is described in detail in the following sections.

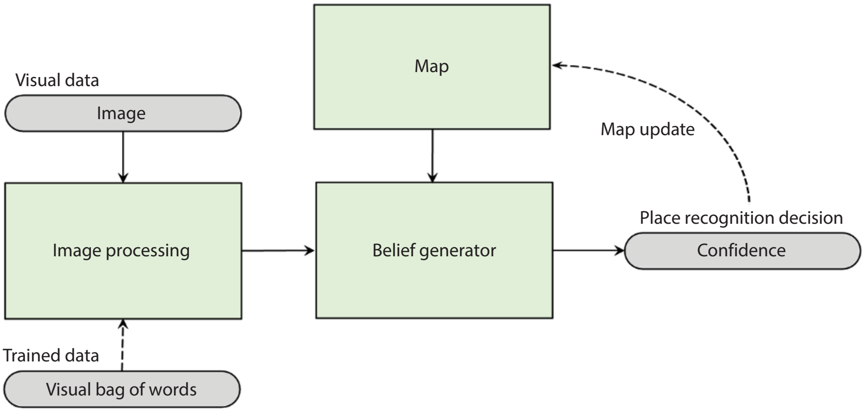


Figure 4.2 Schematic of a visual place recognition system. As the incoming sensory measurement enters the pipeline, the image processing module extracts the corresponding visual representation, i.e., the global or local description vectors. The robot's mapping, either single image- or sequence of images-based, is stored in the database, while the final part of the pipeline, *viz.*, the belief generation module, outputs a confidence regarding whether or not the robot revisits an already mapped location.

4.2.1 Image Processing

A visual place recognition framework has to detect earlier visited areas by employing only the visual data captured through the sensor; the perceived images have to be interpreted robustly, aiming for an informatively built map.

> Rather than working directly with image pixels, most methods use feature vectors extracted from the image processing module to describe the traversed route.

This way, each database image is represented through global (based on the entire image) or local (based on a region-of-interest) features.

4.2.1.1 Global Descriptor Extraction

Studies have demonstrated that humans rapidly categorize a scene using just the coarse global information or "gist" of a scene [17]. Methods based on global feature extractors describe the appearance of the image holistically via a single vector [5, 44, 50, 54, 72, 86]. Their main advantages are the compact representation and computational efficiency, allowing lower storage consumption and faster indexing while querying the database. However, their main disadvantage is the inability to handle occlusions since the geometrical information is not provided.

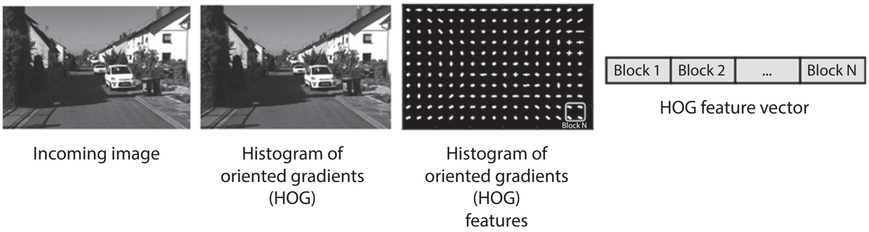

Block N
Block 1
Block 2
...
Block N
HOG feature vector
Incoming image
Histogram of oriented gradients (HOG)
Histogram of oriented gradients (HOG) features

Figure 4.3 A histogram of oriented gradients (HOG) feature vector extracted from an incoming camera measurement. Image's local shape is encoded generating this way a 1 × N vector, where N is the HOG feature-length that represents the corresponding visual data.

Aiming to facilitate the reader's understanding, a global descriptor adopted by a wide range of techniques is presented in this chapter. This method is based on image histograms. Different forms, e.g., color histograms [48, 97], histogram-of-oriented-gradients (HOG) [79], or composed receptive field histograms [58], are used for visual place recognition; however, HOG, which was initially designed for object detection tasks [78], is the most frequently selected. Its structure is based on calculating every pixel's gradient and subsequently creating a histogram according to the results. In Figure 4.3, an illustrative example is given.

4.2.1.2 Local Descriptors Extraction

On the other hand, local features have shown significant advantages compared to the global nature of the previous category. These features are extracted by detecting and describing point-of-interest in an image [1, 13, 16, 53, 56, 75, 76], and they have shown robustness against various image deformations that a freely moving camera may induce, such as scale, rotation, and partial occlusions. However, their extraction process constitutes the bottleneck for any visual place recognition system. This is due to the multitude of possible detections, which can reach the range of thousands, especially in highly textured environments [96]. Therefore, the robotics community has adopted more sophisticated solutions that quantize the corresponding descriptors' space to address this redundancy, yielding the widely known visual bag of words model [80]. This technique originates from text retrieval [6] tasks and compresses the otherwise meaningful information while permitting faster indexing while searching the database. This data representation is referred to as visual vocabulary consisting of a specific quantity of visual words. According to the vocabulary generation, place recognition frameworks are distinguished into i) pre-trained and

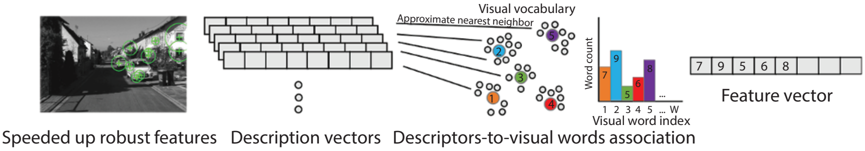


Figure 4.4 An example of an image representation based on the offline formulation of a pre-trained visual vocabulary model. Speeded up robust features [16] are extracted from regions-of-interest in the incoming image, and subsequently, their descriptors are associated with the most similar entries in the vocabulary. The output 1 × W vector, where W is the vocabulary's size, denotes a descriptive histogram of the visual words' count.

ii) online or incremental approaches. Methods that utilize learning data [7, 10–12, 28, 34, 63, 67, 73, 83], i.e., quantizing a sample of local descriptors through a clustering technique [60], belong to the first category, whereas algorithms that generate their visual vocabulary during navigation belong to the second one [4, 31, 43, 45, 52, 70, 89, 91, 92, 101].

Regarding the first category, when the incoming camera measurement enters the image processing module, a visual word is assigned to each extracted descriptor, i.e., local features are classified according to the available vocabulary. This outputs a descriptive histogram vector which indicates the words appearing in each frame. An overview of this pipeline is outlined in Figure 4.4. Nevertheless, while pre-trained methods are able to achieve high performances under computational constraints, their success is highly dependent on their visual vocabulary and, in turn, on the quality of the data used during training. Following this realization, to avoid a performance failure, incremental approaches that "learn" the working environment online are developed. In most cases, these pipelines cluster consecutive local descriptors via different features' matching techniques during navigation.

4.2.2 Map

In every visual place recognition system, map representation constitutes a vital functionality that refers to the model followed for "remembering" the traversed route. Through how the robot maps its route, appearance-based systems are differentiated into single image- and sequence of images-based. Most of the time, the scenario with which the robot would deal is the one that determines the model of the map, e.g., indoors environments with prolonged trajectory segments, such as corridors, provide better results when sequence mapping is adopted.

4.2.2.1 *Single Image-Based*

Approaches that belong to the first category use only the latest image to seek candidate loop closure detections. However, even if an instant view is utilized during the query, single image-based techniques are divided into i) dense and ii) hierarchical mapping.

(i) Dense: During dense map representation, each incoming visual sensory data is associated with a distinct location in the trajectory [28, 96, 102]. When the most recent (query) image is captured, the database is exhaustively searched to identify the most similar entry. As shown in Figure 4.5 (left), the query is associated with image 3 after an exhaustive comparison to the rest of the views is performed.

(ii) Hierarchical: Direct feature matching can be time-consuming since an exhaustive search can significantly burden the computational complexity. Even in the case of global descriptors, this problem becomes intractable for long trajectories. As an efficient solution, a hierarchical structure to the trajectory mapping can be implemented [39]. This technique is discovered in the mammalian brain, both in the hippocampus's grid cells [82] and the visual cortex's pathway [51]. By clustering the camera's stream into tractable groups, i.e., images exhibiting time or content proximity, a hierarchical structure of places is formed [35, 89]. When querying the database images, scalability is achieved by inspecting only the most promising location, as depicted in Figure 4.5 (right).

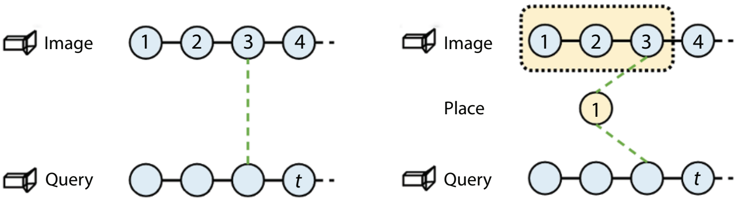


Figure 4.5 In single image-based trajectory mapping, two categories are presented: dense and hierarchical. The first category uses distinct data to represent the traversed path (left), while the second groups images with similar visual content into a common representation (right). During database querying, the computational complexity is higher in approaches belonging to the first category, while methods in the second one tackle this issue by only inspecting the most promising candidates.

4.2.2.2 Sequence of Images-Based

Methods of the second category create sub-maps, i.e., groups of single images, along the navigation course [8, 10, 11, 34, 83, 88, 90, 93, 94, 98]. These submaps, also referred to as sequences or places, are described by common data. For instance, when a pre-trained visual bag of words is utilized, description is achieved by a common visual words histogram [12] in which words from the total of image members are voting. However, during the query process, the latest sequence is utilized to search the database, in contrast to single image-based techniques, where querying is implemented using only the most recent one. As illustrated in Figure 4.6, images belonging to place *Q* are associated with the ones of place 1, insomuch as the identification between places is achieved.

4.2.3 Belief Generator

Given a query view, a visual place recognizer has to determine if the location the incoming view represents can be found in the database and subsequently match the image with robustness against viewpoint and conditional variations under runtime and memory constraints. Aiming to decide whether or not the robot navigates in a previously seen area, a similarity score among the query and the database is computed. The belief generator performs data comparisons depending on how the incoming images are processed to gain the necessary confidence. The most common techniques are discussed in this section. Moreover, to manage the challenges related to the loops' detection in cases of perceptual aliasing, both temporal (loop closures will only be considered if others exist nearby) and

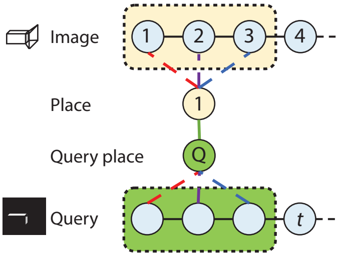


Figure 4.6 Sequence of images-based mapping divides the trajectory into sub-maps. In contrast to the hierarchical approach, images are grouped into places; however, the latest generated sequence is used for querying the database. When a proper match is identified, an image-to-image association is then performed in a coarse-to-fine manner.

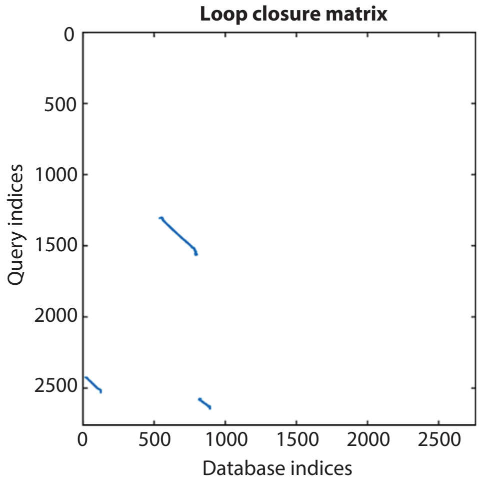


Figure 4.7 Loop closure detection matrix provided by a visual place recognition system for a dataset containing 2,761 images. The loop closure events are illustrated by the off-diagonal elements (blue lines). The system's performance is measured based on this information.

geometric (a valid transformation has to be calculated between the chosen pair) constraints are considered. Then, the belief generator's output is organized in a square matrix, like the one presented in Figure 4.7, whose off-diagonal non-zero elements denote the corresponding loops identified by the system.

4.2.3.1 *Pixel-Wise Similarity*

The most naïve solution based on pixel-wise comparisons is the sum of absolute differences (SAD):

$$D_{ij} = \frac{1\Sigma_{x=0}^{i} R_x \Sigma_{y=0}^{j} R_y}{R_x R_y} \left| \rho_{x,y} - \rho_{x,y} \right| \tag{4.1}$$

where R_x and R_y denote the dimensions of the images, while ρ represents each pixel's intensity value. This technique is selected mainly when raw sensory data is used in the system, avoiding the widely used image processing methods. It is a computationally costly process that is unable to cope with rotation and scale variations. However, using a downsampling scheme and sequence of images-based mapping [64], this method can detect loops with high efficiency [55, 79, 85, 90, 93, 94, 99].

4.2.3.2 Euclidean or Cosine Distance

Regarding images represented as points or vectors in the feature space, the most common distances to compare are the Euclidean and the cosine distance, respectively:

$$d_{Euclidean}\left(\bar{p},\bar{q}\right)=\sum\nolimits_{(i=1)}^{n}\frac{vt}{\left(\bar{q}_i-\bar{p}_i\right)_2} \tag{4.2}$$

$$d_{cosine}\left(\bar{p},\bar{q}\right)=1-\frac{\bar{p}\cdot\bar{q}}{k\bar{p}kk\bar{q}l} \tag{4.3}$$

In the above, d denotes the visual distance among the query I_q and the database instances I_p, while q^- and p^- are their description vector representations. Evidently, the smaller the distance, the higher the similarity of the candidate pair [2, 3, 9, 10–12, 28, 29, 34, 42].

4.2.3.3 Vote Density

Besides, when local extractors are adopted for image or place representation, loop closures are highlighted via voting schemes [26, 36, 59, 89, 91, 96]. The query's local detectors distribute votes in the database to their nearest neighbors, i.e., the descriptors presenting the minimum distance. After this polling, a voting score is received that is used to evaluate the similarity (see Figure 4.8). A naïve approach is to count the number of votes and apply heuristic normalization [26]; yet, in most cases, thresholding the votes' density is not intuitive and varies depending on the environment. Probabilistic voting schemes, such as the binomial density function [36], also consider the total of database's accumulated votes to calculate a score that points to pre-visited locations [89, 91, 96].

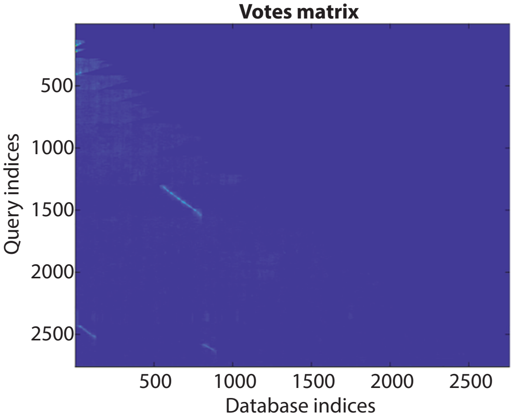


Figure 4.8 Votes matrix as provided by a visual place recognition system for a dataset including 2,761 images. As shown by the off-diagonal lines, the votes' density is higher in regions that correspond to loop closure events.

4.2.3.4 Temporal Consistency

Unlike classification tasks or image retrieval, during robot navigation, visual data are captured sequentially. As the detection of several fault-free loop closures is the prime goal of contemporary robots, missing a few identifications is minor since temporal continuity affords many chances to retrieve them in the frames following. This characteristic is exploited in every sequence of images-based mapping technique, while for single image-based methods, temporal consistency checks are adopted. In such a scheme, multiple matching image pairs need to be identified before accepting a loop closure event [2, 45, 77, 89, 96]. Another line of approaches for incorporating temporal information utilizes more sophisticated techniques based on probabilistic models, such as the Bayes filter [4, 29, 35, 52, 68, 92].

4.2.3.5 Geometrical Verification

An optional step before a visual place recognition pipeline accepts a loop closing match is the selected pair's geometrical verification when data association is performed [3, 4, 19, 34, 35, 62, 74, 89, 91, 96]. This process is based on the local features' spatial information, and it is achieved through the computation of a fundamental/essential matrix or using epipolar constraints. As a result, approaches that use a single vector for representing the incoming images (either through global descriptors or visual word histograms), ignoring the scene's geometry, have to extract local features to perform this check additionally. Typically, geometrical verification is performed using a variant of RANSAC (random sample consensus [32]) and accepts a loop closure event only if a minimum number of local point inliers is identified. Otherwise, the respective candidate is rejected.

4.3 Evaluation

In this section, the protocol of evaluation is presented in detail, based on which the estimated parameters of a custom visual place recognition system are assessed. The main components needed for measuring a system's performance include the utilized datasets, their respective ground truth, and the metrics typically used to assess the performance. The following subsections describe each of these parts.

4.3.1 Ground Truth

Ground truth is typically formed in the shape of a binary matrix of equal dimensions with the similarity one, highlighting the real loop closure events occurring in the dataset. In most cases, this information is provided along with the dataset and indicates pairs of images that capture the same area.

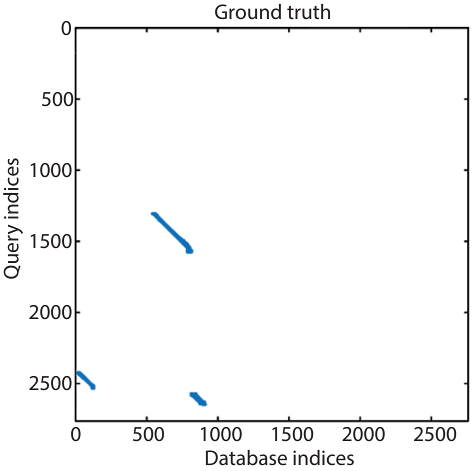


Figure 4.9 Ground truth matrix for a dataset containing 2,761 images. As depicted, the off-diagonal elements (blue lines) indicate the actual loop closure events existing in the trajectory. Detections that fall on the ground truth are the true-positives, while identifications that fall outside this area are false-positives. Using this information, the system's performance is measured.

The matrix's columns and rows depict images at distinct time-stamps, while its boolean data are set to 1 to denote the existence of a loop event ($y_{i,j} = 1$) and 0 otherwise ($y_{i,j} = 0$). An illustrative example is given in Figure 4.9.

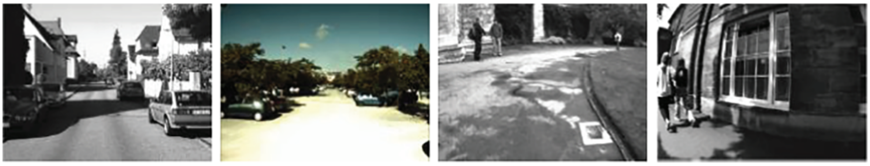
> Ground truth data are combined with the similarity matrix extracted through the recognition process to measure the system's performance.

4.3.2 Datasets

Publicly available datasets are used in most experiments aiming to provide a common performance baseline among different approaches while also covering a wide range of visual sensory properties, e.g., image resolution, frequency, and robot's velocity. The evaluation of a method is administered via several tests on different datasets, aiming to prove its performance capability. Although various cases exist in the literature, the most acknowledged and widely used are presented for this chapter. The selected dataset present outdoor, dynamic, and static environments containing urban views at their most. A summary of each one is provided in Table 4.1, while Figure 4.10 contains some representative samples. Two out of the five datasets come from the KITTI visual suite [37], mainly consisting of trees, cars, and houses. The incoming visual stream is registered through the mounted stereo camera system of a forward-moving car providing accurate odometry information along with high-resolution data (image's size and camera frequency). Sequences 00 and 05 are selected since they contain the most notable loop examples. Malaga 2009 Parking6L has been recorded by the vision system of an electric platform, while New College and City Centre via a mobile robot. The perceived data represent a university campus

Table 4.1 Description of the most commonly used benchmark datasets for evaluating visual place recognition techniques within simultaneous localization and mapping.

Dataset label	Camera position	Image resolution	Frequency
[37] KITTI vision suite (00, 05)	Frontal	1241 × 376	10 Hz
[18] Malaga 2009 Parking 6L	Frontal	1024 × 768	7.5 Hz
[81] New College	Frontal	512 × 384	20 Hz
[28] City Centre	Lateral	1024 × 768	7 Hz



KITTI (05) Malaga 2009 parking 6L New College City Centre

Figure 4.10 Example images of the presented datasets. From left to right: KITTI vision suite (05) [37], Malaga 2009 parking 6L [18], New College [81], and City Centre [28].

parking lot containing cars and trees, while New College and City Centre depict mostly buildings and pedestrians. They are incorporated since they refer to significantly different operational conditions (e.g., camera orientation, resolution, frequency, traveled distance) and plenty of loop closure events. However, for New College, the incoming camera measurements are resampled to 1 Hz, from the initial 20 Hz, owing to the low velocity of the robot and the high camera frequency. This way, the characteristics of modern robotic platforms are simulated more accurately.

4.3.3 Evaluation Metrics

The evaluation metrics presented in this chapter are the most frequent in the literature for visual place recognition. The precision and recall metrics against the ground truth information are utilized to assess an algorithm's performance on selected datasets. Precision is defined as the ratio between accurately identified loop closure events (true-positives) and the total of the system's detections:

$$\text{Precision} = \frac{\text{True-positive}}{\text{positives} + \text{False-positives}} \cdot \text{True} \tag{4.4}$$

More specifically, a true-positive is indicated by the ground truth information and represents matches between actually loop closing samples. On the contrary, a false-positive match characterizes an image pair association not included in the ground truth. Besides, recall is defined as the number of true-positives over the sum of loop closure events included in the ground truth:

$$\text{Recall} = \frac{\text{True} - \text{positives}}{\text{positives} + \text{False-negatives}} \cdot \text{True} \tag{4.5}$$

False-negative detections represent the locations that ought to have been recognized, but the method failed to.

The R_{P100} metric

> For developing a genuinely autonomous robot capable of generating consistent maps, visual place recognition mechanisms should operate at 100% precision since a single erroneous detection can cause a total failure for SLAM.

Therefore, for measuring the performance of a visual place recognition system, the most common indicator is the highest achieved recall at 100% precision (R_{P100}), which denotes the highest possible recall score with no false-positive detections.

4.4 Paradigms

In this section, paradigms of visual place recognition are shown, and several solutions are introduced. Then, dealing with all the aspects of a visual place recognition system, e.g., different image processing techniques, mapping, and belief generators, four exemplar systems are presented. The first two examples address the problem by utilizing a pre-trained visual bag of words model and sequence of images-based mapping. Incremental vocabularies for map building are included in the subsequent systems. A single image-based and hierarchical map is presented in the third example, while a dense approach follows in the last one.

4.4.1 Sequence of Images-Based Visual Word Histograms

This paradigm describes a representative pipeline [10] for sequence of images-based visual place recognition, which combines appearance information from multiple frames to describe the entire content of a physical scene. During an autonomous mission, the input camera measurements are clustered into groups based on a metric distance threshold of 5 meters on the traversed route. Each member of a produced image-sequence is processed to extract local feature descriptors, which are then converted into their corresponding visual words from a pre-trained vocabulary of W entries. These words are accumulated into a single visual word histogram of size W capable of describing the total of the sequence's respective scene. Furthermore, the same words are also used to create single

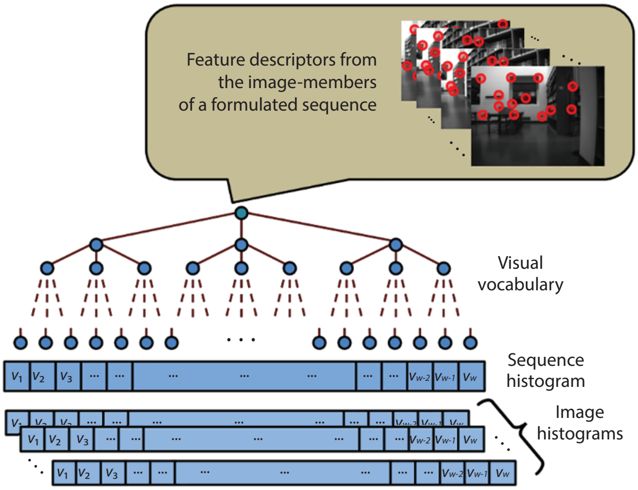


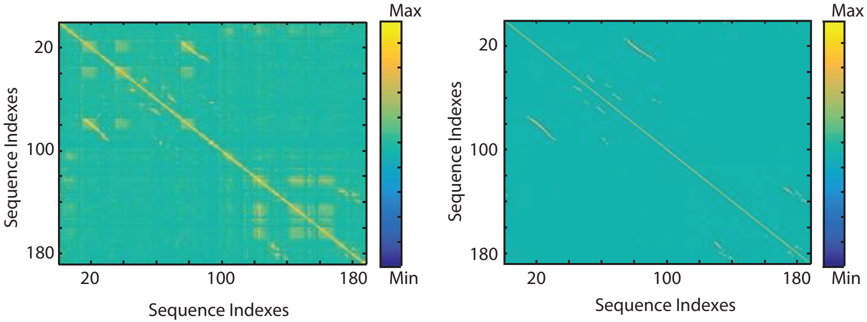
Figure 4.11 Formulating descriptive histograms based on the visual words of a pre-trained vocabulary for image-sequences and single instances.

image histograms and characterize each individual member. Figure 4.11 depicts procedure mentioned above for producing descriptive vectors for image-sequences and single instances.

In order to match the individual sequences, a metric based on *L*2-norm is used. More specifically, similarity is measured based on the *L*2-score between a query (S_q) and a database (S_d) sequence:

$$L2VS^q, V\left(S^d\right) = 1 - 0.5 \frac{VS_q}{q2 - kV\left(S^d\right)k_2} \frac{V\left(S^d\right)}{2 \cdot VS} \tag{4.6}$$

In the above, k...k$_2$ denotes the *L*2-norm, which in turn implies that the *L*2-scores are in the [0, 1] range, with higher values being associated with sequence pairs that correspond to visually similar scenes. As the trajectory grows, the computed values can be arranged to form a similarity matrix *M* incrementally, similar to the one presented in Figure 4.12a. This matrix is symmetric with each element (i, j) containing a corresponding normalized score [34]:



(**a**) Unfiltered similarity matrix M.

(**b**) Filtered similarity matrix M^0.

Figure 4.12 The proposed convolutional filter's impact on a sequence of images-based similarity matrix. Filtered elements which correspond to system's loop closure events are highlighted and can be straightforwardly separated from the ones which are not loop closures.

$$\frac{L2_0 V\left(S_i\right) VS_j = L2V\left(S_i\right), VS_j}{L2V\left(S_i\right), V\left(S_{i-1}\right)} \tag{4.7}$$

M is post-processed by a convolutional filter to further enhance entries with high similarity values that jointly escalate among its two directions. The respective kernel is defined as follows:

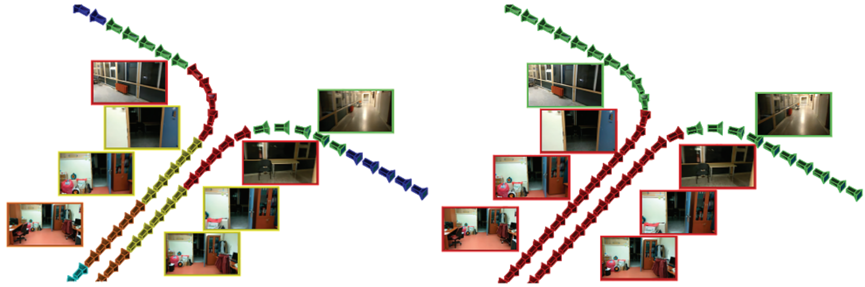
$$K = \begin{bmatrix} 0.5 & -0.5 & 0 \\ -0.5 & 1 & -0.5 \\ 0 & -0.5 & 0.5 \end{bmatrix} \tag{4.8}$$

and it can be applied incrementally, while new camera measurements are obtained and *M* is filled with new $L2^0$-scoring values. The resulting filtered matrix M^0 is shown in Figure 4.12b. Finally, M^0 values that overpass a predefined threshold $k = 0.32$ are considered to represent sequence pairs that contain loop closing image candidates. Finally, image-to-image associations are achieved by identifying the highest *L*2-scores among the respective visual word histograms of the individual image-members.

4.4.2 Dynamic Sequence Segmentation

A significant drawback of the approach mentioned above is the lack of a dynamic procedure to segment the traversed route effectively. More specifically, the utilization of fixed distance intervals does not guarantee that the whole visual information of any two sequences will overlap, although they conform to the exact location. This fact is especially highlighted when considering two sequences that include turning points, like the ones depicted in Figure 4.13a. If that is the case, each pair of actually matching sequences includes several highly dissimilar visual entries, inevitably decreasing the corresponding similarity values. With the aim to address this inconsistency, the method in [12] implements traversed route's dynamic segmentation based on the visual consistency of the observed environment. This way, sequence members are grouped using the information provided by their visual content, thus providing well-defined sequence boundaries (see Figure 4.13b).

The technique followed utilizes the received visual words' variance. During the formulation of each individual sequence, a binary vector *vb* is retained to monitor the already observed visual words. This vector has the same length with the visual vocabulary, with each entry denoting the absence (0) or existence (1) of the corresponding visual word in the group of images. As new observations are processed, the extracted local



(**a**) Image-sequences formulated by making use of a fixed traversed length for segmenting the trajectory. Sequence pairs contain some significantly unrelated images

(**b**) Sequence segmentation based on the *visual word variance* among successive images. Sequences from the same area are well-aligned resulting in visual word histograms with similar structure.

Figure 4.13 Highlighting the importance of similarity consistence sequence segmentation. Color-coding represents different sequences in which camera poses and image instances belong to.

descriptors are mapped into their respective visual words. By checking their index with v_b, words are assigned with a label N, if they are seen for the first time, or O otherwise. Therefore, a *visual word variance* metric is computed as $\sigma_b = N/(N + O)$, marking the current sequence's completion and the initialization of a new one each time $\sigma_b > 0.75$. Note that N and O represent the visual words' number marked as N and O, respectively. When a new sequence is signaled, vector v_b is set to zero, and the same process follows to the next ones. In cases of $\sigma_b \leq 0.75$, v_b is updated with the visual words marked as N, and the following camera frame is noted as member of the ongoing sequence.

In addition to the above mechanism for dynamically segmenting the executed trajectory into intervals with consistent visual cues, the method in [12] additionally improves the similarity matrix's filtering approach by producing a trained convolutional kernel. Avoiding the manual selection of the filter's values, a cost-function minimization scheme is designed by utilizing the ground truth information from loop closure detection datasets. In specific, a multivariate polynomial is defined as $\theta = [\theta_0, \theta_1, \ldots, \theta_9]^T$, with which $m \cdot \theta \geq 0$ denoting the existence of a loop closure event. In the above, m= [1, m_1, ..., $m9$] represents the normalized values of a 3 × 3 sub-matrix from **M**, which are rearranged into a feature vector format. The coefficients of θ from θ_1 to θ_9 represent the convolutional kernel's values, while $k^0 = -\theta_0$ is treated as the loop closure detection threshold. The cost-function is minimized under the logistic regression classifier [27]:

$$\theta = \underset{\theta}{\operatorname{argmin}}\, J(\theta), \tag{4.9}$$

$$J\left(\theta\right) = -\frac{1}{l}\sum_{i=1}^{l} y_{tr}^{(i)} log h_{\theta}\left(m_{tr}^{(i)},\theta\right) + 1 - y_{tr}^{(i)} log 1 - h_{\theta}\left(m_{tr}^{(i)},\theta\right) \tag{4.10}$$

$$\left(\boldsymbol{m},\boldsymbol{\theta}\right) = \frac{1h_{\theta}}{1+\boldsymbol{em}\cdot\boldsymbol{\theta}} \tag{4.11}$$

and it is solved through gradient-descent. In the above, $\mathrm{y}_{\mathrm{tr}}^{(\mathrm{i})}$ and $\mathrm{m}_{\mathrm{tr}}^{(\mathrm{i})}$ represent a single training sample and its corresponding ground truth, respectively, and l is the (i) available learning set's size. Given that two classes are required, $y_{tr}= 1$ is assigned (i) to the ground truth elements that correspond to a loop closure event and $y_{tr} = 0$ otherwise. The resulting θ can be converted back to a square 3 × 3 filtering kernel with the following form:

$$K^0 = \begin{bmatrix} 2.31 & -0.57 & -1.88 \\ -0.41 & 2.19 & -0.75 \\ -1.83 & -0.34 & 2.15 \end{bmatrix} \tag{4.12}$$

with $k^0 = -\theta^0 = 3.5$.

4.4.3 Hierarchical Mapping Through an Incremental Visual Vocabulary

Most pre-trained visual word methods, like the previous ones, provide high execution frequency while searching the database and computing similarities. Although such systems have proved robust when dealing with loop closures, their performance drops when the robot navigates to a dissimilar environment to the training images since the visual vocabulary is generated a priori. Aiming to overcome this weakness, incremental mapping is used that constructs the visual vocabulary in an online manner [89]. These techniques typically induce complex computations due to their incremental nature and the exhaustive search during database querying; therefore, hierarchical methods for faster indexing are adopted.

During such a process, a feature matching coherency check determines new places, according to which segmentation to the incoming image stream is achieved when the last n images' local descriptors correlation stops existing. This way, image frames demonstrating content and time proximity are grouped, resulting in sequences of images with common visual information. Subsequently, a clustering method based on the growing neural gas [33] is executed over the gathered descriptors producing the visual words corresponding to each specific area.

As opposed to the pre-trained approaches, where visual word histogram comparisons are implemented, methods based on incremental maps, such as the one presented, utilize a voting scheme. During a query, local

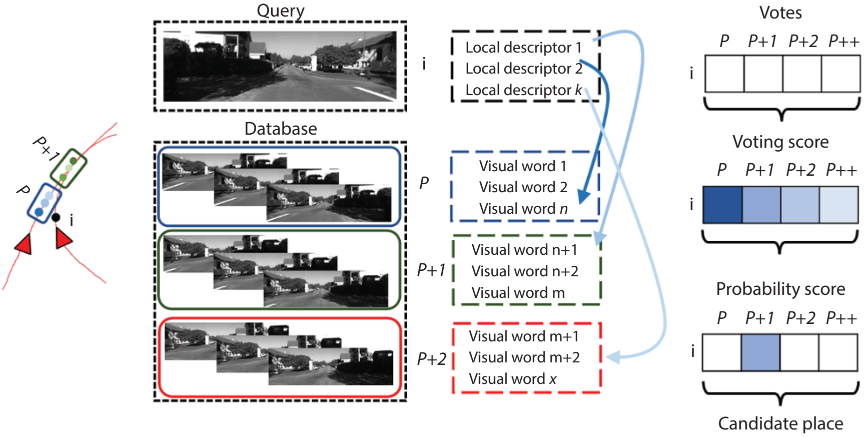


Figure 4.14 The query process of a hierarchical visual place recognition method. As the incoming image stream is processed, votes are distributed to places based on the local-descriptors-to-visual-words association. Subsequently, the candidate place is indicated through a binomial distribution function over the accumulated votes.

descriptors, extracted from the recent image, search for the most identical visual words in the generated database via the nearest neighbor scheme. Hence, each descriptor-to-visual word association corresponds to a new vote for the place. When the places' pooling is finished, a binomial density function determines the similarity scores through a probabilistic model [36]. The place which satisfies the loop closure probability threshold is selected as the candidate one and is furthermore searched for image-to-image correspondences (Figure 4.14). Finally, the chosen image has to satisfy a temporal constraint before accepted, as well as a geometrical verification through RANSAC.

4.4.4 Bag of Tracked Words for Incremental Visual Place Recognition

Unlike the previous one, the second incrementally based vocabulary mapping system uses a dense representation for environment confronted by the robot. However, since a feature matching technique would be impractical if applied at every incoming image frame, the vocabulary is generated via a point tracking technique [91].

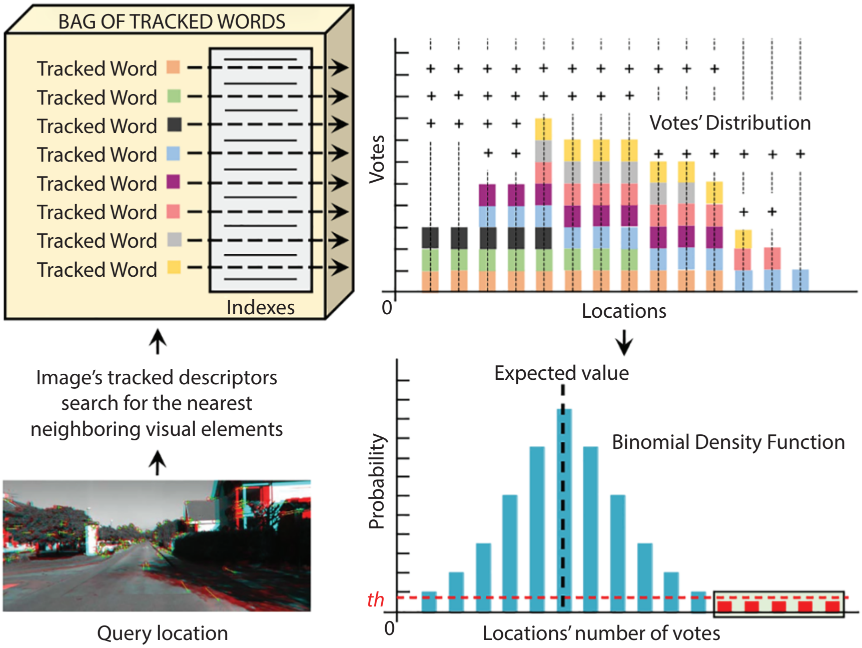


Figure 4.15 In the course of a query, local descriptors distribute votes to the locations in the database where the nearest neighboring tracked words are formed. The colored blocks depict the votes cast by different tracked words. Subsequently, the candidate locations are selected if a loop closure threshold *th* is satisfied applied on the binomial density function's probabilistic score (highlighted red area).

As shown in Figure 4.15, when new visual sensory data are processed, local points are detected and described. Subsequently, these points are tracked in the following robot view through the Kanade-Lucas-Tomasi (KLT) tracker [87]. The relevant feature is selected based on a guided-feature-detection technique that searches for the most similar descriptor located near the tracked point. Points that lose track during navigation are transformed into visual words, referred to as tracked words, while their total constitutes the bag of tracked words. These new elements are assigned to the traversed map at the location from which they are originated.

Using the previous example's belief generator, a voting scheme determines the pre-visited locations. When a query image is handled, its descriptors distribute votes into their nearest-neighboring database of tracked words. A binomial probability density function converts the accumulated votes into probabilistic scores, thus, indicating the candidate loop closures. Finally, a geometrical verification step through RANSAC ensures the robustness of the visual place recognition system.

4.5 Experimental Results

The pipelines introduced have been implemented and tested on the five datasets presented in Section 4.2. However, it is noteworthy that as the incoming sensory information in every dataset comes from a stereo camera rig, only the monocular stream is utilized. By varying the loop closure decision threshold, the precision and recall curves for the KITTI (00) dataset are shown in Figure 4.16. For the reader's convenience, the highest recall metrics at 100% precision are indicated in circles. In addition, the highest achieved scores for the other datasets are also presented in Table 4.2 intending for illustrating the full potential for each method.

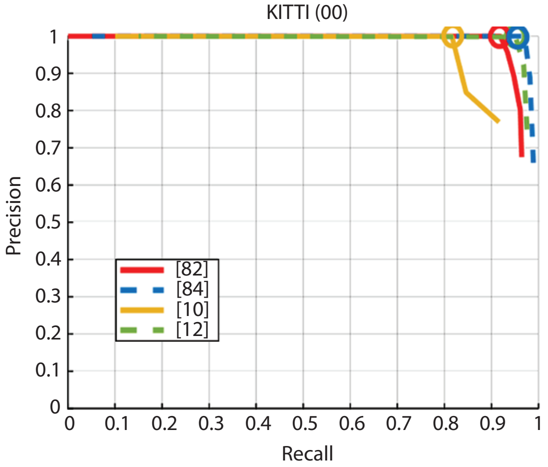


Figure 4.16 Precision and recall curves for the presented examples, which are monitored by altering the paradigms' loop closure detection thresholds. Experiments are performed on the KITTI vision suite (00). Color markers (cycles) on the top of the graphs show the highest recall for perfect precision (RP_{100}).

Table 4.2 Maximum recall at 100% precision (R_{P100}) for the presented visual place recognition paradigms.

Approach	**KITTI (00)**	**KITTI (05)**	**Malaga 2009 6L**	**New College**	**City Centre**
Sequence of images-based visual word histograms [10]	81.5	84.8	81.5	77.6	68.5
Dynamic sequence segmentation [12]	96.5	97.3	87.6	92.7	71.1
Hierarchical mapping [89]	93.1	94.2	87.9	88.0	16.3
Bag of tracked words [91]	97.5	92.6	85.0	83.0	20.0

4.6 Future Trends and Conclusion

Past and recent trends for place recognition in the SLAM context were reported here. Such a system is of paramount importance for detecting loops in the robot's traversed path, permitting consistent map generation. Early solutions were based on range sensors; however, cameras have evolved to the primary perception module in recent autonomous platforms owing to the qualitative information provided by vision and their low cost. A visual place recognizer consists of three components: image processing, map, and belief generator. At last, it could be said that having read this chapter, the reader has gained adequate knowledge to construct a visual place recognition system, including the parts needed, the comparison techniques, and the evaluation metrics which are used.

In conclusion, the reader has been introduced to four systems covering all aspects of feature-based visual place recognition systems. At first, by presenting two pre-trained visual bag of words methods, the reader walked through the global description of an image, which constitutes the most frequently used technique when low complexity is required. Moreover, it is shown how a sequence of images-based system is constructed, using both fixed and dynamic group length, and how its similarity is measured. Then, due to the importance of adopting incremental visual vocabularies for trajectory mapping, two additional pipelines are presented. Besides, their comparison techniques are provided. Finally, it is fair to say that they perform satisfactorily in speed and accuracy regarding the methods apposed.

Future works should be based on new ways of representing the incoming visual stream aiming for more robust and low complexity representations.

Additionally, efficient mapping techniques should be investigated as robots' operational conditions are getting longer, i.e., long-term navigation.

References

1. Alcantarilla, P.F., Bartoli, A., Davison, A.J.: KAZE Features. In: *Proceeding of the European Conference on Computer Vision*, pp. 214–227. Florence, Italy (2012). DOI 10.1007/978-3642-33783-3_16
2. An, S., Che, G., Zhou, F., Liu, X., Ma, X., Chen, Y.: Fast and Incremental Loop Closure Detection Using Proximity Graphs. In: *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 378–385. Macau, China (2019). DOI 10.1109/IROS40897.2019.8968043
3. An, S., Zhu, H., Wei, D., Tsintotas, K.A., Gasteratos, A.: Fast and Incremental Loop Closure Detection with Deep Features and Proximity Graphs. *Journal of Field Robotics* (2022). DOI 10.1002/rob.22060
4. Angeli, A., Filliat, D., Doncieux, S., Meyer, J.A.: Fast and Incremental Method for LoopClosure Detection using Bags of Visual Words. *IEEE Transactions on Robotics* **24**(5), 1027–1037 (2008). DOI 10.1109/TRO.2008.2004514
5. Arroyo, R., Alcantarilla, P.F., Bergasa, L.M., Yebes, J.J., Bronte, S.: Fast and Effective Visual Place Recognition using Binary Codes and Disparity Information. In: *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3089–3094. Chicago, IL, USA (2014). DOI 10.1109/IROS.2014.6942989
6. Baeza-Yates, R., Ribeiro-Neto, B., *et al.*: *Modern Information Retrieval*, vol. 463. ACM press New York (1999)
7. Balaska, V., Bampis, L., Boudourides, M., Gasteratos, A.: Unsupervised Semantic Clustering and Localization for Mobile Robotics Tasks. *Robotics and Autonomous Systems* **131**, 103567 (2020). DOI 10.1016/j.robot.2020.103567
8. Balaska, V., Bampis, L., Gasteratos, A.: Graph-based Semantic Segmentation. In: *Proceeding of the International Conference on Robotics*, pp. 572–579. Patras, Greece (2018). DOI 10.1007/978-3-030-00232-9_60
9. Balaska, V., Bampis, L., Kansizoglou, I., Gasteratos, A.: Enhancing Satellite Semantic Maps with Ground-Level Imagery. *Robotics and Autonomous Systems* **139**, 103760 (2021). DOI 10.1016/j.robot.2021.103760
10. Bampis, L., Amanatiadis, A., Gasteratos, A.: Encoding the Description of Image Sequences: A Two-Layered Pipeline for Loop Closure Detection. In: *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4530–4536. Daejeon, Korea (South) (2016). DOI 10.1109/IROS.2016.7759667
11. Bampis, L., Amanatiadis, A., Gasteratos, A.: High Order Visual Words for Structure-Aware and Viewpoint-Invariant Loop Closure Detection. In: *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots*

and Systems, pp. 4268–4275. Vancouver, BC, Canada (2017). DOI 10.1109/IROS.2017.8206289

12. Bampis, L., Amanatiadis, A., Gasteratos, A.: Fast Loop-Closure Detection using Visual Word-Vectors from Image Sequences. *International Journal of Robotics Research* **37**(1), 62–82 (2018). DOI 10.1177/0278364917740639
13. Bampis, L., Chatzichristofis, S., Iakovidou, C., Amanatiadis, A., Boutalis, Y., Gasteratos, A.: A LoCATe-based visual place recognition system for mobile robotics and GPGPUs. *Concurrency and Computation: Practice and Experience* **30**(7), e4146(2018). DOI10.1002/cpe.4146
14. Bampis, L., Gasteratos, A.: Revisiting the Bag-of-Visual-Words Model: A Hierarchical Localization Architecture for Mobile Systems. *Robotics and Autonomous Systems* **113**, 104–119 (2019). DOI 10.1016/j.robot.2019.01.004
15. Baudoin, Y., Doroftei, D., De Cubber, G., Berrabah, S.A., Pinzon, C., Warlet, F., Gancet, J., Motard, E., Ilzkovitz, M., Nalpantidis, L., *et al.*: View-Finder: Robotics Assistance to Fire-Fighting Services and Crisis Management. In: *Proceeding of the IEEE International Workshop on Safety, Security & Rescue Robotics*, pp. 1–6. Denver, CO, USA (2009)
16. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: *Proceeding of the European Conference on Computer Vision*, pp. 404–417. Graz, Austria (2006). DOI 10.1007/11744023_32
17. Biederman, I.: Aspects and Extensions of a Theory of Human Image Understanding. Computational Processes in *Human Vision: An Interdisciplinary Perspective* pp. 370–428 (1988)
18. Blanco, J.L., Moreno, F.A., Gonzalez, J.: A Collection of Outdoor Robotic Datasets with Centimeter-Accuracy Ground Truth. *Autonomous Robots* **27**(4), 327–351 (2009). DOI 10.1007/s10514-009-9138-7
19. Bosse, M., Newman, P., Leonard, J., Soika, M., Feiten, W., Teller, S.: An ATLAS Framework for Scalable Mapping. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, vol. 2, pp. 1899–1906. Taipei, Taiwan (2003). DOI 10.1109/ROBOT.2003.1241872
20. Boukas, E., Gasteratos, A., Visentin, G.: Introducing a Globally Consistent Orbital-Based Localization System. *Journal of Field Robotics* **35**(2), 275–298 (2018). DOI 10.1002/rob.21739
21. Boukas, E., Kostavelis, I., Gasteratos, A., Sirakoulis, G.C.: Robot guided crowd evacuation. *IEEE Transactions on Automation Science and Engineering* **12**(2), 739–751 (2014). DOI 10.1109/TASE.2014.2323175
22. Brogaard, R.Y., Andersen, R., Kovac, L., Zajaczkowski, M., Boukas, E.: Towards an Autonomous, Visual Inspection-aware 3D Exploration and Mapping System for Water Ballast Tanks of Marine Vessels. In: *Proceeding of the IEEE International Conference on Imaging Systems and Techniques*, pp. 1–6. New York, USA (2021)
23. Brogaard, R.Y., Ravn, O., Boukas, E.: GPU-accelerated Localization in Confined Spaces using Deep Geometric Features. In: *Proceeding of the IEEE*

International Conference on Imaging Systems and Techniques, pp. 1–6. New York, USA (2021)

24. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Transactions on Robotics* **32**(6), 1309–1332 (2016). DOI 10.1109/TRO.2016.2624754
25. Capezio, F., Mastrogiovanni, F., Sgorbissa, A., Zaccaria, R.: Robot-Assisted Surveillance in Large Environments. *Journal of Computing and Information Technology* **17**(1), 95–108 (2009). DOI 10.2498/cit.1001180
26. Cieslewski, T., Stumm, E., Gawel, A., Bosse, M., Lynen, S., Siegwart, R.: Point Cloud Descriptors for Place Recognition using Sparse Visual Information. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 4830–4836. Stockholm, Sweden (2016). DOI 10.1109/ICRA.2016.7487687
27. Crone, S.F., Finlay, S.: Instance Sampling in Credit Scoring: An Empirical Study of Sample Size and Balancing. *International Journal of Forecasting* **28**(1), 224–238 (2012). DOI 10.1016/j.ijforecast.2011.07.006
28. Cummins, M., Newman, P.: FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *International Journal of Robotics Research* **27**(6), 647–665 (2008). DOI 10.1177/0278364908090961
29. Cummins, M., Newman, P.: Appearance-Only SLAM at Large Scale with FAB-MAP 2.0. *International Journal of Robotics Research* **30**(9), 1100–1123 (2011). DOI 10.1177/0278364910385483
30. Feraru, V.A., Andersen, R.E., Boukas, E.: Towards an autonomous UAV-based system to assist search and rescue operations in man overboard incidents. In: *Proceeding of the IEEE International Symposium on Safety, Security, and Rescue Robotics*, pp. 57–64. Abu Dhabi, United Arab Emirates (2020). DOI 10.1109/SSRR50563.2020.9292632
31. Filliat, D.: A Visual Bag of Words Method for Interactive Qualitative Localization and Mapping. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 3921–3926 (2007). DOI 10.1109/ROBOT.2007.364080
32. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM* **24**(6), 381–395 (1981).
33. Fritzke, B., *et al.*: A Growing Neural Gas Network Learns Topologies. *Advances in Neural Information Processing Systems* **7**, 625–632 (1995)
34. Gálvez-López, D., Tardos, J.D.: Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE Transactions on Robotics* **28**(5), 1188–1197 (2012). DOI 10.1109/TRO.2012.2197158
35. Garcia-Fidalgo, E., Ortiz, A.: Hierarchical Place Recognition for Topological Mapping. *IEEE Transactions on Robotics* **33**(5), 1061–1074 (2017). DOI 10.1109/TRO.2017.2704598

36. Gehrig, M., Stumm, E.,Hinzmann, T., Siegwart, R.: Visual Place Recognition with Probabilistic Voting. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 3192–3199. Singapore (2017). DOI 10.1109/ICRA.2017.7989362
37. Geiger, A., Lenz, P., Urtasun, R.: Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. Providence, RI, USA (2012). DOI 10.1109/CVPR.2012.6248074
38. Hausler, S., Jacobson, A., Milford, M.: Multi-Process Fusion: Visual Place Recognition Using Multiple Image Processing Methods. *IEEE Robotics and Automation Letters* **4**(2),1924–1931 (2019). DOI 10.1109/LRA.2019.2898427
39. Hausler, S., Milford, M.: Hierarchical Multi-Process Fusion for Visual Place Recognition. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 3327– 3333. Paris, France (2020). DOI 10.1109/ICRA40945.2020.9197360
40. Ho, K.L., Newman, P.: Loop closure detection in slam by combining visual and spatial appearance. *Robotics and Autonomous Systems* **54**(9), 740–749 (2006). DOI 10.1016/j.robot.2006.04.016
41. Jiang, M., Song, S., Herrmann, J.M., Li, J.H., Li, Y., Hu, Z., Li, Z., Liu, J., Li, S., Feng, X.: Underwater Loop-Closure Detection for Mechanical Scanning Imaging Sonar by Filtering the Similarity Matrix with Probability Hypothesis Density Filter. *IEEE Access* **7**, 166614– 166628 (2019). DOI 10.1109/ACCESS.2019.2952445
42. Kansizoglou, I., Bampis, L., Gasteratos, A.: Deep Feature Space: A Geometrical Perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). DOI 10.1109/TPAMI.2021.3094625
43. Kawewong, A., Tongprasit, N., Tangruamsub, S., Hasegawa, O.: Online and Incremental Appearance-Based SLAM in Highly Dynamic Environments. *International Journal of Robotics Research* **30**(1), 33–55 (2011). DOI 10.1177/0278364910371855
44. Kenshimov, C., Bampis, L., Amirgaliyev, B., Arslanov, M., Gasteratos, A.: Deep Learning Features Exception for Cross-Season Visual Place Recognition. *Pattern Recognition Letters* **100**, 124–130 (2017). DOI 10.1016/j.patrec.2017.10.028
45. Khan, S., Wollherr, D.: IBuILD: Incremental Bag of Binary Words for Appearance Based Loop Closure Detection. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 5441–5447. Seattle, WA, USA (2015). DOI 10.1109/ICRA.2015.7139959
46. Konstantinidis, F., Kansizoglou, I., Tsintotas, K.A., Mouroutsos, S.G., Gasteratos, A.: The role of machine vision in industry 4.0: A textile manufacturing perspective. In: *Proceeding of the IEEE International Conference on Imaging Systems and Techniques*, pp. 1–6. New York, USA (2021).
47. Konstantinidis, F., Mouroutsos, S.G., Gasteratos, A.: The role of machine vision in industry 4.0: An automotive manufacturing perspective. In:

Proceeding of the IEEE International Conference on Imaging Systems and Techniques, pp. 1–6. New York, USA (2021).

48. Konstantinidis, K., Gasteratos, A., Andreadis, I.: Image Retrieval Based on Fuzzy Color Histogram Processing. *Optics Communications* **248**(4-6), 375–386 (2005).
49. Kostavelis, I., Nalpantidis, L., Boukas, E., Rodrigalvarez, M.A., Stamoulias, I., Lentaris, G., Diamantopoulos, D., Siozios, K., Soudris, D., Gasteratos, A.: SPARTAN: Developing a Vision System for Future Autonomous Space Exploration Robots. *Journal of Field Robotics* **31**(1), 107–140 (2014).
50. Kröse, B.J., Vlassis, N., Bunschoten, R., Motomura, Y.: A Probabilistic Model for Appearance-Based Robot Localization. *Image and Vision Computing* **19**(6), 381–391 (2001). DOI 10.1016/S0262-8856(00)00086-X
51. Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A.J., Wiskott, L.: Deep Hierarchies in the Primate Visual Cortex: What Can We Learn for Computer Vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8), 1847–1871 (2012). DOI 10.1109/TPAMI.2012.272
52. Labbe, M., Michaud, F.: Appearance-Based Loop Closure Detection for Online Large-Scale and Long-Term Operation. *IEEE Transactions on Robotics* **29**(3), 734–745 (2013). DOI 10.1109/TRO.2013.2242375
53. Leutenegger, S., Chli, M., Siegwart, R.Y.: BRISK: Binary Robust Invariant Scalable Keypoints. In: *Proceeding of the International Conference on Computer Vision*, pp. 2548–2555. Barcelona, Spain (2011). DOI 10.1109/ICCV.2011.6126542
54. Liu, Y., Zhang, H.: Visual Loop Closure Detection with a Compact Image Descriptor. In: *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1051–1056. Vilamoura-Algarve, Portugal (2012). DOI 10.1109/IROS.2012.6386145
55. Liu, Y., Zhang, H.:Towards Improving the Efficiency of Sequence-Based SLAM. In: *Proceeding of the IEEE International Conference on Mechatronics and Automation*, pp. 1261–1266. Takamatsu, Japan (2013). DOI 10.1109/ICMA.2013.6618095
56. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* **60**(2), 91–110 (2004). DOI 10.1023/B:VISI.0000029664.99615.94
57. Lowry, S., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., Milford, M.J.: Visual Place Recognition: A Survey. *IEEE Transactions on Robotics* **32**(1), 1–19 (2016). DOI 10.1109/TRO.2015.2496823
58. Luo, J., Pronobis, A., Caputo, B., Jensfelt, P.: Incremental Learning for Place Recognition in Dynamic Environments. In: *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 721–728 (2007). DOI 10.1109/IROS.2007.4398986

59. Lynen, S., Bosse, M., Furgale, P., Siegwart, R.: Placeless Place-Recognition. In: *Proceeding of the International Conference on 3D Vision*, vol. 1, pp. 303–310. Tokyo, Japan (2014). DOI 10.1109/3DV.2014.36
60. MacQueen, J., *et al.*: Some Methods for Classification and Analysis of Multivariate Observations. In: *Proceeding of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297 (1967).
61. Maffra, F., Chen, Z., Chli, M.: Tolerant Place Recognition Combining 2D and 3D Information for UAV Navigation. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 2542–2549. Brisbane, QLD, Australia (2018). DOI10.1109/ICRA.2018.8460786
62. Mei, C., Sibley, G., Cummins, M., Newman, P.M., Reid, I.: A Constant-Time Efficient Stereo SLAM System. In: *Proceeding of the British Machine Vision Conference*, pp. 1–11. London, UK (2009). DOI 10.5244/C.23.54
63. Mei, C., Sibley, G., Newman, P.: Closing Loops Without Places. In: *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3738–3744. Taipei, Taiwan (2010). DOI 10.1109/IROS.2010.5652266
64. Milford, M.J., Wyeth, G.F.: SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 1643–1649. Saint Paul, MN, USA (2012). DOI10.1109/ICRA.2012.6224623
65. Milford, M.J., Wyeth, G.F., Prasser, D.: RatSLAM: A Hippocampal Model for Simultaneous Localization and Mapping. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, vol. 1, pp. 403–408. New Orleans, LA, USA (2004). DOI10.1109/ROBOT.2004.1307183
66. Muhammad, N., Fuentes-Perez, J.F., Tuhtan, J.A., Toming, G., Musall, M., Kruusmaa, M.: Map-Based Localization and Loop-Closure Detection from a Moving Underwater Platform using Flow Features. *Autonomous Robots* **43**(6), 1419–1434 (2019). DOI 10.1007/s10514018-9797-3
67. Mur-Artal, R., Tardós, J.D.: Fast Relocalisation and Loop Closing in Keyframe-based SLAM. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 846– 853. Hong Kong, China (2014). DOI 10.1109/ICRA.2014.6906953
68. Murillo, A.C., Singh, G., Kosecka, J., Guerrero, J.J.: Localization in Urban Environments using a Panoramic GIST Descriptor. *IEEE Transactions on Robotics* **29**(1), 146–160 (2012). DOI 10.1109/TRO.2012.2220211
69. Nalpantidis, L., Sirakoulis, G.C., Gasteratos, A.: Non-probabilistic cellular automataenhanced stereo vision simultaneous localization and mapping. *Measurement Science and Technology* **22**(11), 114027 (2011)
70. Nicosevici, T., Garcia, R.: Automatic Visual Bag-of-Words for Online Robot Navigation and Mapping. *IEEE Transactions on Robotics* **28**(4), 886–898 (2012). DOI 10.1109/TRO.2012.2192013
71. O'Keefe, J., Conway, D.: Hippocampal Place Units in the Freely Moving Rat: Why They Fire Where They Fire. *Experimental Brain Research* **31**(4), 573–590 (1978).

72. Oliva, A., Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer* Vision **42**(3), 145–175 (2001). DOI 10.1023/A:1011139631724
73. Papapetros, I.T., Balaska, V., Gasteratos, A.: Multi-Layer Map: Augmenting Semantic Visual Memory. In: *Proceeding of the International Conference on Unmanned Aircraft Systems*, pp. 1206–1212. Athens, Greece (2020). DOI 10.1109/ICUAS48674.2020.9213923
74. Rohou, S., Franek, P., Aubry, C., Jaulin, L.: Proving the Existence of Loops in Robot Trajectories. *International Journal of Robotics Research* **37**(12), 1500–1516 (2018). DOI10.1177/0278364918808367
75. Rosten, E., Drummond, T.: Fusing Points and Lines for High Performance Tracking. In: *Proceeding of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1508– 1515. Beijing, China (2005). DOI 10.1109/ICCV.2005.104
76. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An Efficient Alternative to SIFT or SURF. In: *Proceeding of the International Conference on Computer Vision*, pp. 2564–2571. Barcelona, Spain (2011). DOI 10.1109/ICCV.2011.6126544
77. Sánchez-Belenguer, C., Wolfart, E., Sequeira, V.: RISE: A Novel Indoor Visual Place Recogniser. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 265–271. Paris, France (2020). DOI 10.1109/ICRA40945.2020.9196871
78. Schiele, B., Crowley, J.L.: Object Recognition using Multidimensional Receptive Field Histograms. In: *Proceeding of the European Conference on Computer Vision*, pp. 610–619. Cambridge, UK (1996). DOI 10.1007/BFb0015571
79. Siam, S.M., Zhang, H.: Fast-SeqSLAM: A Fast Appearance Based Place Recognition Algorithm. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 5702–5708. Singapore (2017). DOI 10.1109/ICRA.2017.7989671
80. Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. In: *Proceeding of the IEEE International Conference on Computer Vision*, vol. 3, pp. 1470–1470 (2003). DOI 10.1109/ICCV.2003.1238663
81. Smith, M., Baldwin, I., Churchill, W., Paul, R., Newman, P.: The New College Vision and Laser Data Set. *International Journal of Robotics Research* **28**(5), 595–599 (2009). DOI 10.1177/0278364909103911
82. Stensola, H., Stensola, T., Solstad, T., Frøland, K., Moser, M.B., Moser, E.I.: The Entorhinal Grid Map Is Discretized. *Nature* **492**(7427), 72–78 (2012). DOI 10.1038/nature11649
83. Stumm, E.S., Mei, C., Lacroix, S.: Building Location Models for Visual Place Recognition. *International Journal of Robotics Research* **35**(4), 334–356 (2016). DOI 10.1177/0278364915570140

84. Sugiyama, H., Tsujioka, T., Murata, M.: Collaborative Movement of Rescue Robots for Reliable and Effective Networking in Disaster Area. In: *Proceeding of the International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 7–pp. San Jose, CA, USA (2005). DOI 10.1109/COLCOM.2005.1651217
85. Sünderhauf, N., Neubert, P., Protzel, P.: Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across all four Seasons. In: *Proceeding of the IEEE International Conference on Robotics and Automation. Workshop on Long-Term Autonomy*, p. 2013. Karlsruhe, Germany (2013).
86. Sünderhauf, N., Protzel, P.: Brief-GIST-Closing the Loop by Simple Means. In: *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1234–1241. San Francisco, CA, USA (2011). DOI 10.1109/IROS.2011.6094921
87. Tomasi, C., Kanade, T.: Detection and Tracking of Point. Tech. rep., features. Technical Report CMU-CS-91-132, Carnegie, Mellon University (1991).
88. Tsintotas, K.A., Bampis, L., An, S., Fragulis, G.F., Mouroutsos, S.G., Gasteratos, A.: Sequence-based mapping for probabilistic visual loop-closure detection. In: *Proceeding of the IEEE International Conference on Imaging Systems and Techniques*, pp. 1–6. New York, USA (2021).
89. Tsintotas, K.A., Bampis, L., Gasteratos, A.: Assigning Visual Words to Places for Loop ClosureDetection. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 5979–5985. Brisbane, QLD, Australia (2018). DOI 10.1109/ICRA.2018.8461146
90. Tsintotas, K.A., Bampis, L., Gasteratos, A.: DOSeqSLAM: Dynamic On-line Sequence based loop closure detection algorithm for SLAM. In: *Proceeding of the IEEE International Conference on Imaging Systems and Techniques*, pp. 1–6. Krakow, Poland (2018). DOI 10.1109/IST.2018.8577113
91. Tsintotas, K.A., Bampis, L., Gasteratos, A.: Probabilistic Appearance-Based Place Recognition Through Bag of Tracked Words. *IEEE Robotics and Automation Letters* **4**(2), 1737–1744 (2019). DOI 10.1109/LRA.2019.2897151
92. Tsintotas, K.A., Bampis, L., Gasteratos, A.: Modest-vocabulary loop-closure detection with incremental bag of tracked words. *Robotics and Autonomous Systems* **141**, 103782 (2021). DOI 10.1016/j.robot.2021.103782
93. Tsintotas, K.A., Bampis, L., Gasteratos, A.: Tracking-DOSeqSLAM: A Dynamic SequenceBased Visual Place Recognition Paradigm. *IET Computer Vision* **15**(4), 258–273 (2021). DOI 10.1049/cvi2.12041
94. Tsintotas, K.A., Bampis,L., Rallis,S., Gasteratos,A.: SeqSLAM with Bag of Visual Words for Appearance Based Loop Closure Detection. In: *Proceeding of the International Conference on Robotics*, pp. 580–587. Patras, Greece (2018). DOI 10.1007/978-3-030-00232-9_61
95. Tsintotas,K.A., Bampis,L., Taitzoglou,A., Kansizoglou,I., Gasteratos,A.: Safe UAV landing: A low-complexity pipeline for surface conditions recognition. In: *Proceeding of the IEEE International Conference on Imaging Systems and Techniques*, pp. 1–6. New York, USA (2021).

96. Tsintotas, K.A., Giannis, P., Bampis, L., Gasteratos, A.: Appearance-based loop closure detection with scale-restrictive visual features. In: *Proceeding of the International Conference on Computer Vision Systems*, pp. 75–87. Thessaloniki, Greece (2019). DOI 10.1007/978-3030-34995-0_7
97. Ulrich,I., Nourbakhsh, I.: Appearance-Based Place Recognition for Topological Localization. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, vol. 2, pp. 1023–1029. San Francisco, CA, USA (2000). DOI 10.1109/ROBOT.2000.844734
98. Vysotska, O., Naseer, T., Spinello, L., Burgard, W., Stachniss, C.: Efficient and Effective Matching of Image Sequences Under Substantial Appearance Changes Exploiting GPS Priors. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 2774– 2779. Seattle, WA, USA (2015). DOI 10.1109/ICRA.2015.7139576
99. Wang, Y., Hu, X., Lian, J., Zhang, L., Kong, X.: Improved SeqSLAM for Real-Time Place Recognition and Navigation Error Correction. In: *Proceeding of the International Conference on Intelligent Human-Machine Systems and Cybernetics*, vol. 1, pp. 260–264. Hangzhou, China (2015). DOI 10.1109/IHMSC.2015.23
100. Yang, S., Scherer, S.A., Yi, X., Zell, A.: Multi-Camera Visual SLAM for Autonomous Navigation of Micro Aerial Vehicles. *Robotics and Autonomous Systems* **93**, 116–134 (2017). DOI 10.1016/j.robot.2017.03.018
101. Zhang, G., Lilly, M.J., Vela, P.A.: Learning Binary Features Online from Motion Dynamics for Incremental Loop-Closure Detection and Place Recognition. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 765–772. Stockholm, Sweden (2016). DOI 10.1109/ICRA.2016.7487205
102. Zhang, H.: BoRF: Loop-Closure Detection with Scale Invariant Visual Features. In: *Proceeding of the IEEE International Conference on Robotics and Automation*, pp. 3125–3130. Shanghai, China (2011). DOI 10.1109/ICRA.2011.5980273
103. Zhang, S., He, B., Nian, R., Liang, Y., Yan, T.: SLAM and a Novel Loop Closure Detection for Autonomous Underwater Vehicles. In: *Proceeding of the OCEANS*, pp. 1–4. San Diego, CA, USA (2013). DOI 10.23919/OCEANS.2013.6741257