Unsupervised semantic clustering and localization for mobile robotics tasks

Article *in* Robotics and Autonomous Systems · May 2020

DOI: 10.1016/j.robot.2020.103567

CITATIONS
21

READS
278

4 authors:

Vasiliki Balaska
Democritus University of Thrace
10 PUBLICATIONS **71** CITATIONS
SEE PROFILE

Loukas Bampis
Democritus University of Thrace
52 PUBLICATIONS **570** CITATIONS
SEE PROFILE

Moses Boudourides
University of Patras
69 PUBLICATIONS **346** CITATIONS
SEE PROFILE

Antonios Gasteratos
Democritus University of Thrace
283 PUBLICATIONS **4,285** CITATIONS
SEE PROFILE

Unsupervised Semantic Clustering and Localization for Mobile Robotics Tasks

Vasiliki Balaska[a], Loukas Bampis[a], Moses Boudourides[b], and Antonios Gasteratos[a]

[a]Department of Production and Management Engineering, Democritus University of Thrace, Vas. Sophias 12, GR-671 32, Xanthi, Greece
[b]Department of Mathematics, University of Patras, Rio GR-265 00 Patras, Greece

Abstract

Due to its vast applicability, the semantic interpretation of regions or entities increasingly attracts the attention of scholars within the robotics community. The paper at hand introduces a novel unsupervised technique to semantically identify the position of an autonomous agent in unknown environments. When the robot explores a certain path for the first time, community detection is achieved through graph-based segmentation. This allows the agent to semantically define its surroundings in future traverses even if the environment's lighting conditions are changed. The proposed semantic clustering technique exploits the Louvain community detection algorithm, which constitutes a novel and efficient method for identifying groups of measurements with consistent similarity. The produced communities are combined with metric information, as provided by the robot's odometry through a hierarchical agglomerative clustering method. The suggested algorithm is evaluated in indoors and outdoors datasets creating topological maps capable of assisting semantic localization. We demonstrate that the system categorizes the places correctly when the robot revisits an environment despite the possible lighting variation.

Keywords: Topological mapping, Illumination invariance, Community detection, Robot localization

1. Introduction

Contemporary research in robotics endows modern autonomous systems with the ability to semantically recognize and segment regions/entities. This affords them with increased flexibility and the ability to interpret and interact with the environment on a higher level. Apart from places, robots are capable of recognizing objects and classify them in semantic areas. Overall, semantic interpretation is considered an active and fundamental research field, which relies mostly on computer vision methods for place recognition (Kostavelis and Gasteratos (2015)).

Localization (operated either on metric or topological maps) stands for a critical ability incorporated into nowadays robots, yet semantics, which can constitute an essential foundation for this capacity, is still underdeveloped. Moreover, the successful communication between humans and robots can be established only through the ability of the second to sense and classify their own surroundings by precisely recalling spatial memories (Kostavelis et al. (2016)). Metric maps are mainly used for small-scale spaces (Karaoguz and Bozma (2014)) and they are organized in a geometric manner, while the relative conceptual remains hidden. The reveal of this hidden information can be achieved by reorganizing it on the basis of a topological map (Lowry et al. (2015)). This is directly related with topology, the branch of mathematics studying characteristics invariant in continuous deformations (Erkent and Bozma (2012); Pronobis and Jensfelt (2012)). A scene's description based on topological maps retains information regarding the semantic region it belongs. These graphs are a fundamental element of a semantic map, enabling abstraction to metric maps (Karaoguz and Bozma (2014)).

In order to construct a topological map, the similarity between the acquired camera measurements needs to be computed. Zhang et al. (2010) identified the mechanisms for quantifying such similarities, as to the degree of abstracted information from the respective images, into the following categories. Image-based approaches rely on pixel differences between consecutive images to determine changes in a scene. Techniques based on local-features detect and try to associate various key-points, so as to measure the similarity between different frames (Fraundorfer et al. (2007); Korrapati et al. (2011)). Furthermore, in histogram-based approaches, images are compared by means of features' statistics (Karaoguz and Bozma (2014)). Lastly, a typical approach for simplifying the information regarding a place is to address the problem with a Bag-of-Words (BoW) representation. The method of BoW describes the input measurements as a quantized set of local features, thus reducing the searching space to gain efficiency (Sivic and Zisserman (2003)). However, a scene contains randomly scattered features, thus a meaningful way to describe it is via histograms of visual words (visual word vectors) (Fazl-Ersi and Tsotsos (2012)).

This paper addresses the problem of semantically mapping an environment. With the term semantic, we refer to the signs and the things to which they refer Kostavelis and Gasteratos (2015). Thus, within the scope of this work, the term semantic mapping is used to indicate the identification and the recoding of visual signs and symbols that

contain meaningful information. Our goal is to produce an unsupervised system to segment the robot's trajectory into different semantic regions. Then, each time the robot passes through the same area, it gets self-localized within one of the computed semantic divisions by exclusively using visual information. This essentially means that a specific label for each trajectory segment is not necessarily required to achieve semantic localization. On the contrary, our method is able to identify distinct regions that preserve semantic consistency, however, a meaningful name can be assigned to each one of them with minimum effort after the map has being completed. We make use of the BoW image representation as a means of uniformly describing the captured images and shaping semantic clusters based on their similarities. The semantic interpretation of the robot's path is achieved by applying the Louvain Community Detection Algorithm (LCDA) (Blondel et al. (2008)). Semantically important properties of a particular environment are identified with great tolerance over various lighting conditions. Extending our previous work (Balaska et al. (2019)), we improved the development of the places' semantic representation by introducing an additional hierarchical agglomerative clustering method to incorporate the information of a metric map. Using only LCDA, many snapshots of the same semantic region are redundantly grouped into different clusters due to dissimilarities occurring by observing the same scene from arbitrary orientations. This is a common problem for any appearance-based method that utilizes monocular and unidirectional cameras since the view of the environment changes significantly when different content is observed (Fraundorfer et al. (2007); Lynen et al. (2017); Bampis et al. (2018)). By means of geometrical information, such over-segmented areas can be consolidated, thus improving the quality of the map. A its final stage the map contains both semantic and metric information (topological map).

The rest of the paper is structured as follows. In Section II, we discuss representative related work in the field of semantic mapping construction. Section III contains a detailed explanation of our approach. In Section IV, we present the results of our experiments, while in the last section, we draw conclusions and present suggestions for future work.

2. Related Literature

Generally speaking, one may argue that a semantic map is a kind of topological map with purely semantic data about the places that a robot encounters together with objects they contain; those maps have been used for robot detection, mapping, navigation and classification (Rusu et al. (2008); Krishnan and Krishna (2010); Ranganathan and Dellaert (2007); Nieto-Granda et al. (2010)). Such semantic informational systems can be obtained using either supervised or unsupervised methods. In what follows, we list some of the most representative techniques for both approaches.

2.1. Supervised methods

In the work of Erkent and Bozma (2012), topological maps are composed of bubble surfaces, which, transformed into bubble descriptors and classified by a multi-class Support Vector Machines (SVM) classifier, are used to learn specific places. A Voronoi Random Field (VRF)-like algorithm based on Conditional Random Field and an SVM classifier have been used to achieve semantic labeling of Generalized Voronoi Graphs nodes (Shi et al. (2012); Demir and Isil Bozma (2015)). A low-level image descriptor for addressing the room classification problem is proposed by Uršič et al. (2012) and Yeh and Darrell (2008) using an SVM classifier. A semantic map construction method which combines place and object recognition –both utilizing supervised learning– is presented by Kostavelis and Gasteratos (2017). High-dimensional histograms can be used to represent an observed scene and achieve the classification of visual places through SVM (Pronobis and Jensfelt (2012)). All these methods are supervised and require the respective systems to be trained on labeled data to allow semantic characterization of a new scene. Even though such approaches have the potential to generalize knowledge in areas that the robot has not encountered before, our method of identification and clustering does not require any supervision, while also achieving high re-localization rates each time a place is revisited.

2.2. Unsupervised Method

Place categorization from visual cues operates on video and image streams and can detect "unknown" place labels (Ranganathan (2012)). Moreover, a Gaussian mixture model have been used to form a probability distribution of features extracted by image-preprocessing, while a place detector groups features into places through odometry data and a hidden Markov model (Chella et al. (2007)). Visual features of a robot's traversed environment have been represented through the BoW model and, by means of a Neural Gas, the spatial information for each scene is clustered into semantically consistent groups (Kostavelis et al. (2016)). Finally, the method proposed by Erkent et al. (2017) was based on the Single-Linkage (SLINK) agglomerative algorithm (Sibson (1973)). This unsupervised and incremental approach allows the robot to learn about organizing the observed environment and localizing in it. Finally, the authors Guillaume et al. (2011) used CENsus TRansform hISTogram (CENTRIST) and GIST descriptors, applying an unsupervised approach with a Self-Organizing Map for classification.

All the techniques cited in this subsection have applied unsupervised techniques for semantically clustering the environment and thus, they do not require labeling. In our work, since were are interested in building the entire topological map of an area without pre-training, we follow the same approach. The proposed technique differs from the aforementioned ones in that we adopt a modern graph-based clustering technique (LCDA), which considerably improves the results, and we combine the produced communities with the metric measurements in a coarse-to-fine manner.

3. Approach

In this section, our approach for semantic clustering and localization is detailed. We begin by describing the se-

lected mechanism for representing the input images in a rotation/scale-invariant manner, as well as the algorithm for clustering the feature vectors according to their similarity. The formulated semantic groups are then refined using temporal and metric information with the aim to produce the topological map. Finally, we propose two different approaches for recognizing the exact position of a query camera frame within the computed map.

3.1. Semantic Clustering

3.1.1. Image representation

As the robot explores a new environment, we retain the most prominent SURF (Bay et al. (2008)) features on each frame to produce a visual vocabulary. Instead of using a plain BoW, we construct the visual vocabulary with a tree structure (Nister and Stewenius (2006)), reducing the computational complexity during the on-line phase of similarity computation. To achieve this, we follow the typical approach of hierarchically performing k-means clustering over all feature vectors (Bampis et al. (2017)). The resulting vocabulary tree consists of L levels and K branches per level leading to a total $W = K^L$ leaf nodes (visual words). More specifically, in our approach, we set $L = 5$ and $K = 10$, thus the vocabulary size W is computed at $100K$ visual words. Based on the vocabulary tree, we extract the corresponding histograms to describe all the obtained images. Each word w_i is assigned with a weight, corresponding to its inverse document frequency term:

$$idf(i) = \log \frac{N}{N_i}, \tag{1}$$

where N is the number of training images, and N_i is the count of word w_i in them. At each time instance t, the f most prominent SURF descriptors of the input query image (I_t) are computed and quantized into the most similar visual words. Thus, I_t is converted into a BoW vector $u_t \in R^W$:

$$u_{t_i} = tf(i, I_t) * idf(i), \tag{2}$$

where $tf(i, I_t) = n_i^{I_t}/n^{I_t}$, $n_i^{I_t}$ corresponds to the total count of occurrences of word w_i in image I_t, and n^{I_t} is the number of words in I_t. Finally, we calculate the unit vector as our final BoW descriptor:

$$\bar{u}_t = \frac{u_t}{\|u_t\|_2}. \tag{3}$$

To measure the similarity between two BoW vectors $\bar{u}_1$ and $\bar{u}_2$, we calculate a score S based on the l^2 norm (Bampis et al. (2018); Newman et al. (2006)):

$$S = 1 - 0.5 * \|\bar{u}_1 - \bar{u}_2\|_2. \tag{4}$$

Subsequently, we shape a similarity matrix M quantifying the degree of resemblance between all recorded frames and indicating whether the related images belong to the same community.

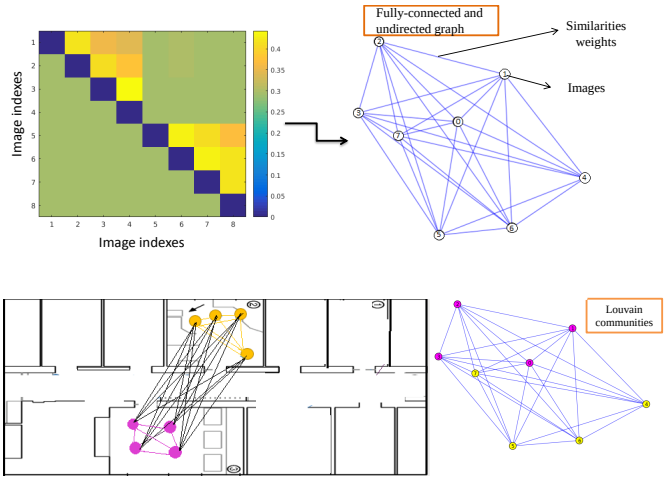


Figure 1: The process of semantic clustering through the Louvain method and a typical segmentation output between two different areas.

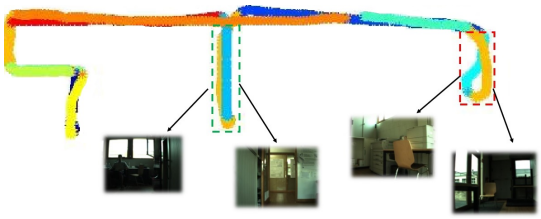
3.1.2. Clustering with similarities

M contains the information regarding the complete similarity structure of the executed route. At this point, we need to emphasize that our goal is to semantically group coherent regions from any part of the trajectory and reject time information. Thus, the whole matrix is converted into a graph, the nodes of which correspond to images and the edges to the similarity weights. Retaining only similarity measurements between images with low time proximity reduces the computational complexity, but it should fail to merge discontinuous trajectory regions (e.g., two distinct but similar offices). Therefore, applying the LCDA on the whole deriving graph essentially means that the edges will extend from the first to the last node connecting each possible pair of images. Figure 1 illustrates the above stated graph formulation procedure, as well as a typical output that LCDA algorithm would produce.

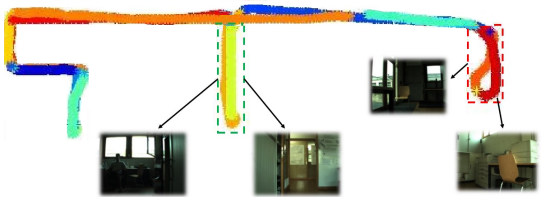
According to Louvain algorithm, given an input graph $G = (V, E)$, where V and E are the sets of nodes and edges, respectively, community detection is performed by dividing G into n sets $C = V_1, V_2, ..., V_i, ...V_n$, with each V_i being a community (semantic region or semantic cluster). LCDA is iterative and repeats until there is no additional improvement in the modularity. A modularity function (Newman et al. (2006); Traag et al. (2019)) is used to measure the strength of dividing a network into communities. Firstly, each node initializes within its own community and, for every node in a graph, the modularity value ΔQ is computed for all neighboring communities:

$$\Delta Q = \left(\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right) - \left(\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right), \tag{5}$$

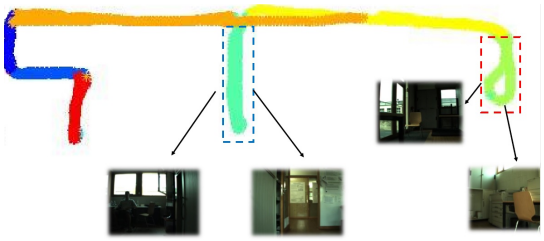
where $\sum_{in}$ is the sum of the links' weights within the community to which the node i is assigned, $\sum_{tot}$ is the sum of the links' weights associated with the nodes in the community, $k_{i,in}$ is the sum of the weights of the links from node i to the rest of the community, k_i is the sum of the link's weights incident to node i, and m is the sum of the link's weights in the network. Still,



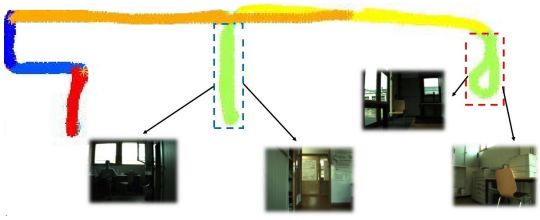
(a) Semantic segmentation outcome obtained from the LCDA algorithm. Note that the highlighted places (offices) are inaccurately divided into multiple communities due to high visual dissimilarities observed during opposite-oriented traverses of the same area.



(b) Semantic communities that do not preserve time consistency are temporarily segmented into further refined groups. Note the subdivision of the orange regions in the highlighted office areas from Fig.3a.



(c) Hierarchical agglomerative clustering based on odometry data. Clusters located in the same area are effectively unified. Semantic communities from the highlighted areas which are spatially coherent are re-grouped.



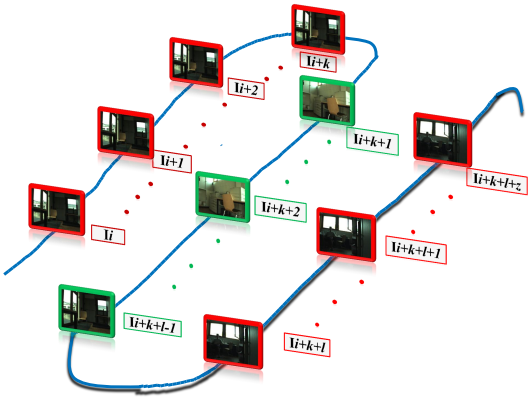
(d) Final semantic map combining the results from the semantic clustering with the Louvain method (Fig.2a) and the hierarchical agglomerative through metric data (Fig.2c). The two highlighted groups are merged into a single semantic community.

Figure 2: The proposed technique combines similarity and metric information to produce the final semantic map. Each formulated community is represented with a different color.

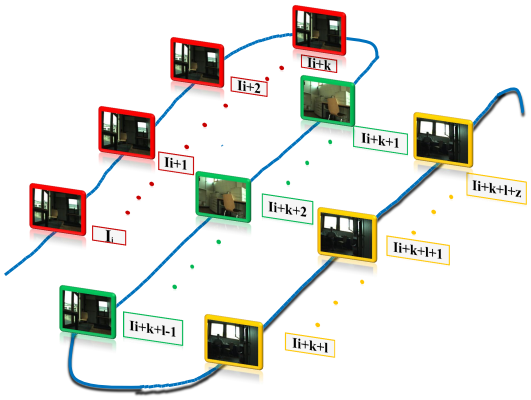
with the aim to increase the computational efficiency, a simplified expression has been adopted (Blondel et al. (2008)):

$$\Delta Q = e_{ic} - \frac{k_i \sum_{tot}}{2m}, \tag{6}$$

where e_{ic} is the sum of the links' weights between node i and the community C. Note that each node belongs to the community that yields the maximum value in modularity. This phase



(a) Semantic clustering output from the Louvain method.



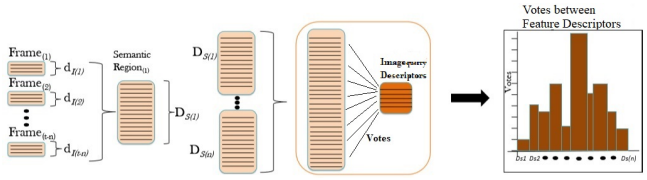
(b) The result of further dividing semantic clusters with temporal inconsistencies.

Figure 3: A representative example illustrating the notion of partitioning a semantic cluster, the image-members of which do not preserve temporal consistency.

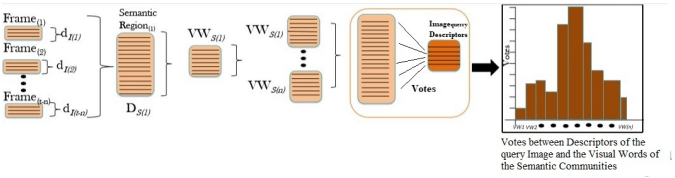
is repeated until a local maximum modularity is achieved, and then, all the communities are reduced to vertices creating a new graph. Internal community edges are converted into a single self-looping edge, and the resulting weight is the sum of all internal weights in the community. Multiple edges between two communities are suppressed into a single edge, while the corresponding weight results from accumulating the edges between them (Ozaki et al. (2016)). By means of LCDA, the images are clustered into multiple community configurations, each of which possesses a particular modularity. We choose the maximum modularity value to produce the best possible semantic segmentation. Figure 2a shows the semantic topological map created by directly applying the above described procedure on the *COLD-Freiburg part A seq.1* (Pronobis and Caputo (2009)).

3.1.3. Segmenting clusters with temporal inconsistency

For this step, since the semantic clusters may contain members from completely different places, we choose to divide the communities based on their index regularity. This step is necessary in order to correct clustering inconsistencies from the LCDA algorithm in Section 3.1.4, which are originated from low similarity scores between frames of the same semantic region (e.g., observing the same area from opposing orientations).



(a) The voting process for the semantic localization with features descriptors.



(b) The voting process for the semantic localization with visual words.

Figure 4: The representation of voting processes to achieve semantic localization.

Note that this further division will be reversed later on since our final goal is to produce communities with members from different parts of the trajectory. Therefore, should a semantic group V_x contain images $[I_i, I_{i+k}]$ and $[I_{i+k+l}, I_{i+k+l+z}]$, while images $[I_{i+k+1}, I_{i+k+l-1}]$ belong to a different community V_y, then V_x is divided into two subsets $V_{x_1} = [I_i, I_{i+k}]$ and V_{x_2}=$[I_{i+k+l}, I_{i+k+l+z}]$, resulting into groups with temporally neighboring members. Figure 3 depicts this process (which produces a larger number of communities that will be re-combined in a later step), on a representative simulation environment to better visualize its effect. The segmentation of the communities in Fig. 2a using temporal information results into the representation depicted in Fig. 2b.

3.1.4. Re-clustering with metric data

With the aim to merge nearby groups that do not necessarily observe the same view (yet they belong to the same semantic place), we utilize the respective robot's odometry from each location. This information can be obtained through Simultaneous Localization and Mapping (SLAM) techniques, such as ORB-SLAM2 (Mur-Artal and Tardós (2017)), or direct-indirect measurements through global positioning and inertial sensors. Thus, for each cluster, we compute its average 3D position and re-cluster those representations using a hierarchical agglomerative method (Pandey and Khanna (2014)). The combination of clusters from Sections 3.1.2 and 3.1.3 with metric data, essentially identifies and clusters common visual information within the total of each room. Considering this characteristic, our technique reassemblies a room detector which additionally retains an unsupervised nature, reducing its bias over specific training environments.

Firstly, we calculate the average 3D coordinate for each created cluster and then the minimum Euclidean distance between all possible centroid pairs. The last are merged into clusters until an inconsistency coefficient check interrupts the procedure. This coefficient is estimated as:

$$c_n = \frac{(d_n - m_n)}{std_n}, \tag{7}$$

where m_n and std_n are the mean and standard deviation, respectively, between distances d_n in cluster n (Bampis et al. (2017)). Similar cluster pairs are connected with high c_n values, while connections between dissimilar clusters are characterized by low ones. Note that input centroids, as well as the first clusters joining them, are initialized with $c_n = 0$. Pairs of groups that present the maximum c_n are merged on each level of the hierarchical clustering architecture (GhasemiGol et al. (2010); Xu et al. (2016)). Finally, using an appropriate inconsistency coefficient threshold r, the procedure is finalized for every group with $c_n > r$. Thus, the semantic regions are merged into clusters that correspond to the same physical area. The result of agglomerating upon the outcome in Fig. 2b is depicted in Fig. 2c.

3.1.5. Combining semantic and metric information

Following the above process, the final map is created which combines both the extracted semantic and metric information. The groups deriving from the agglomerative clustering in Section 3.1.4 are back-merged if they were connected by the output of LCDA (Section 3.1.2). Hence, the effect of the index-inconsistency division, which was implemented in Section 3.1.3, is reversed, allowing the final semantic places to span over multiple areas of the trajectory (Fig. 2d).

3.2. Semantic localization

When the robot revisits a mapped area, the map should support it with sufficient information allowing the categorization of any currently captured view into one the computed semantic communities. In order to assess this capability, we propose two approaches: (i) a rigorous one, based on feature matching, and (ii) an efficient one, based on the formulation of a visual vocabulary from each semantic region.

Regarding the feature matching approach, semantic localization is achieved by a vote aggregation scheme. Given a semantically clustered trajectory with its image database, the f most prominent SURF features are extracted from each frame. Therefore, every semantic area is characterized by a list of descriptors (D_s), obtained from the concatenation of its member sets for each query image. Then, we extract the d_I SURF features for each query image and compute the Euclidean distance between these descriptors and the ones of the above lists. The identified nearest neighboring pairs assign votes to the corresponding semantic regions, while the highest scoring one represents the area into which the robot is located (Fig. 4a).

The second approach to semantically localize the robot into the computed map follows the same voting structure, however in a reduced database searching space. More specifically, each of the D_s lists mentioned above is converted into a distinct BoW representation of size k. The number of visual words is calculated using the following function:

$$k = \alpha * N, \tag{8}$$

where N is the number of features in each D_s, and α is a scaling factor ($\alpha < 1$). Then, the process of converting each feature

Table 1: Number of images which are assigned in each of the semantic clusters, according to the ground-truth. *Cloudy* sequences are used for semantically mapping the trajectory, while *sunny* and *night* for semantic localization.

Ground-Truth Regions /COLD Sequences		**printer area**	**offices**	**corridor**	**stairs**	**store**	**bath**
First visit: Semantic mapping.	Fr1-1 Cloudy	179	345	641	50	190	131
Rest of the visits: Semantic localization.	Fr1-1 Sunny	120	450	400	60	335	230
	Fr1-1 Night	125	760	600	160	230	125

Ground-Truth Regions /KTH Sequences		**corridor**	**office1**	**office2**	**kitchen**	**printer area**
First visit: Semantic mapping.	KTH Cloudy	234	142	96	241	114
Rest of the visits: Semantic localization.	KTH Sunny	379	196	173	147	100
	KTH Night	461	196	175	202	147

from the query image to a visual word from this new vocabulary yields one vote to the corresponding semantic region. Note that the visual words produced at this step are different from the primitive ones in Section 3.1.1 since in this case, we seek for a finer vocabulary with the best performing size per semantic region (Tsintotas et al. (2018, 2019)). The rest of the voting steps are equivalent with the ones mentioned during the feature matching technique (Fig. 4b).

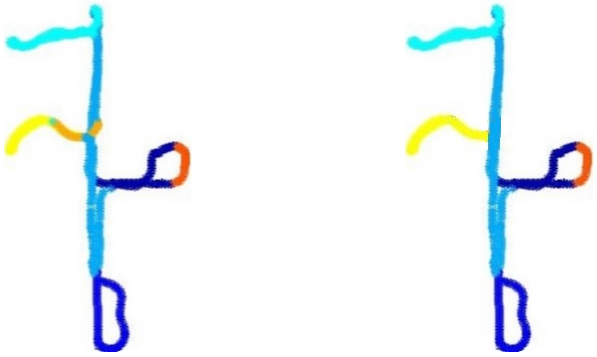
4. Experimental Results

In this section, the proposed method is evaluated both in terms of semantic clustering and semantic localization. We applied our method on the following datasets: *COLD-Freiburg* (Pronobis and Caputo (2009)), *COLD-Saarbrücken* (Pronobis and Caputo (2009)), *KTH-IDOL2* (Luo et al. (2006)), and *New College* (Smith et al. (2009)), which are comprised of explicit semantic regions. Two of the above, viz. *COLD-Freiburg part A sequence1 cloudy1* and *KTH-IDOL2 cloudy1*, were used as validation sets to assess the effect of the system's parameters on the performance. A different set of 5 testing cases was also utilized to assess the scalability of the selected parameters on additional datasets. The above experimental procedure is used in order to discriminate the effect of parameters tuning from the performance of our final system, as measured on a completely dissimilar for set with different environmental conditions (Gálvez-López and Tardos (2012)). With the aim to measure the achieved performance, we extracted the ground-truth information each dataset through their respective blueprints. Finally, during the semantic localization, we calculated the percentage of correctly detected semantic locations (true positive assignments) when the robot revisited an area of the mapped trajectory.

4.1. Datasets

As mentioned above, we applied our method on 5 public available datasets, which allows us to present comparative and reproducible results. In what follows, the most representative characteristics of each individual set are detailed:

COLD dataset: The COLD dataset was created through three different mobile robotic platforms, viz., the ActivMedia PeopleBot, the ActivMedia Pioneer-3, and the iRobot ATRV-Mini. The PeopleBot and Pioneer-3 were used for recording three individual groups (cloudy, sunny, night) of laboratory areas from Saarbrücken and Freiburg, respectively, and they were equipped with a monocular camera sensor, an omni-directional one, and a SICK laser scanner. The iRobot platform also carried a monocular and an omni-directional sensor, and captured an office area at Ljubljana. In each case, the same settings of camera was used for image acquisition with a resolution of 640×480 pixels. From the available sensory data, we made use of the collected monocular measurements from the three different illumination condition, in order to perform our experiments.

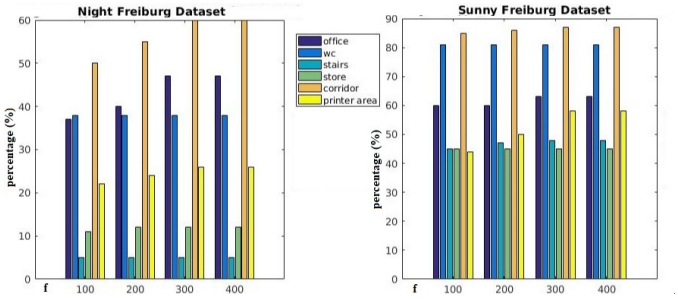


(a) $f = 100$ (b) $f = \{200, 300, 400\}$

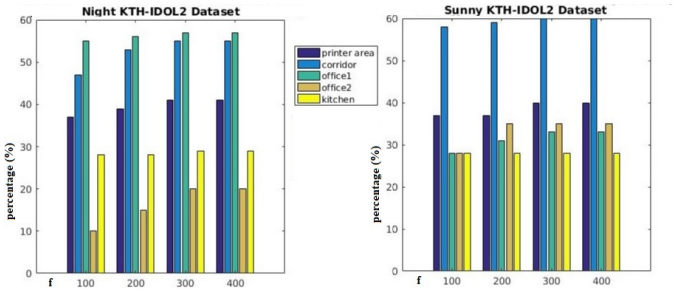
Figure 5: The effect of different values of f in *KTH-IDOL2 cloudy* dataset. Each semantic community is represented with a different color.

KTH-IDOL2 Dataset: All the measurements included this dataset were acquired in the Computational Vision and Active Perception Laboratory, at the Royal Institute of Technology, Sweden. The image sequences, laser scans, and odometry data were acquired using two mobile robotic platforms, the PeopleBot Minnie and the PowerBot Dumbo. The created image dataset consists of scenes from 5 different indoors regions (one-person office, two-persons office, corridor, kitchen, and printer area) under various illumination conditions (in cloudy weather, in sunny weather, and during the night). All of the images were acquired at a resolution of 320×240 pixels.

New College Dataset: The dataset was collected within the New College grounds in Oxford during early November 2008

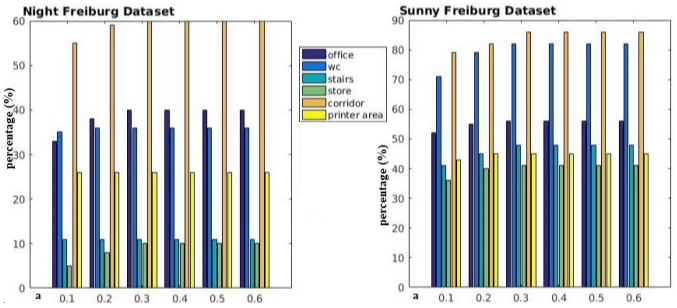


(a) COLD sunny and night

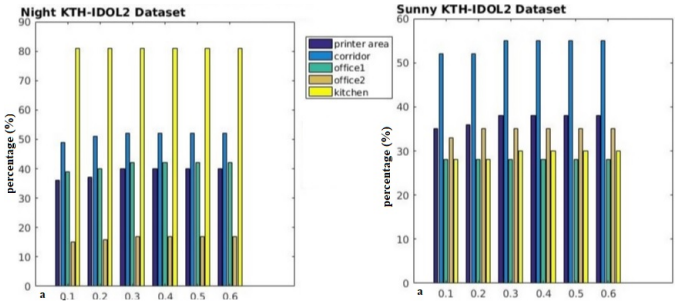


(b) KTH-IDOL2 sunny and night

Figure 6: The effect of a different f values over the feature's aggregation process.



(a) COLD sunny and night



(b) KTH-IDOL2 sunny and night

Figure 7: The effect of different α values over the visual word's aggregation process.

in the afternoon. The robot captured stereo and panoramic images, together with laser scans, from three regions (Campus, Quad, and Parkland). The Campus environment includes medieval buildings. The Quad is an enclosure centered around an oval lawn. A short tunnel leads from the Quad to a space surrounded by old buildings, finally reaching the Parkland area. In our work, we only utilized one of the stereo streams resulting in a sequence of 512×384 pixels per frame.

4.2. Parameters Evaluation

In order to assess the introduced parameters' effect over the proposed system, a number of experiments were performed on the two evaluation datasets, analyzed in Table 1. In such a way, the final performance (presented in Section 4.2) is not directly influenced by the selected parameterization, proving the expandability of our approach.

4.2.1. Number of features

Our first experiment refers to the evaluation of the maximum number of features per image used during semantic clustering and localization. The presented results correspond to the most representative ones which preserved the lowest computational complexity: $f = \{100, 200, 300, 400\}$. Four different semantic maps were created for each of the *COLD-Freiburg sequence1 cloudy1* and *KTH-IDOL2 cloudy1* cases (one for every assessed value of f). In both datasets, we noticed that by choosing $r = 0.78$ yielded a suitable and consistent semantic segmentation in the hierarchical tree for all the eight different maps. Nevertheless, meaningful results were also obtained for r values within the range of $[0.71, 1.2]$. For r values lower than 0.71 none of the evaluated datasets produced valid semantic segmentation results. Considering the formulated ground-truth, it was found that the semantic maps resulting for each distinct number of features are similar for the case of *COLD-Freiburg*, as presented in Fig. 2d. However, in *KTH-IDOL2 cloudy*, the semantic map created with a maximum of 100 features differs from the ground-truth and the rest of the evaluated values. Fig. 5a shows the map corresponding to 100 features, while in Fig. 5b we show the same map as produced by each of the 200, 300, and 400 cases. The evaluation of the maximum allowed number of used features was also carried out, related to the process of voting between a query image and the ones in the database. More specifically, the *COLD-Freiburg* and *KTH-IDOL2* datasets were tested for semantically localizing the camera measurements, when the robot revisited the same area at different time periods and particularly under different lighting conditions. The evaluation presented in Fig. 6 shows the percentage of query features which were matched to the appropriate D_s list. The results imply that a robust semantic localization is achieved, for the sunny and night sequences of the corresponding datasets, when considering a maximum number of 300 or 400 features in both cases. From the above, the value of $f = 300$ is the most appropriate since it restrains the system's computational complexity.

4.2.2. Number of visual words

Instead of performing localization directly on the detected feature space, the required computations can be reduced by producing a visual vocabulary specifically designed for each of the computed semantic areas. To that end, we also evaluate the scaling factor α (equation 8). By assessing a wide variety of

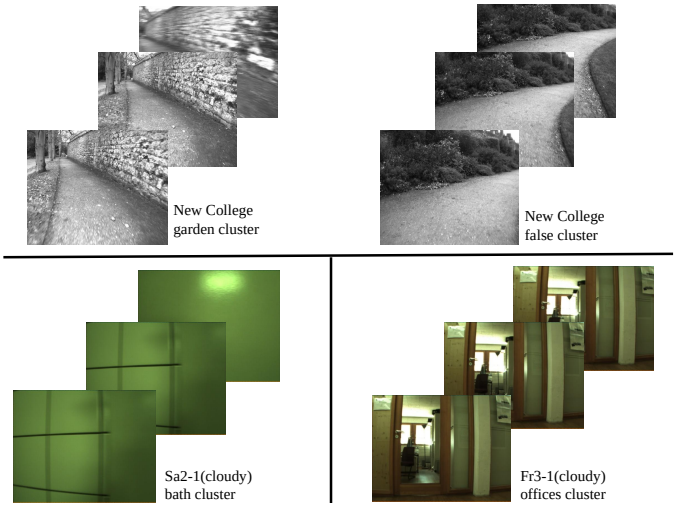


Figure 8: Representative examples of "miss-clustered" images from each of the evaluated datasets.

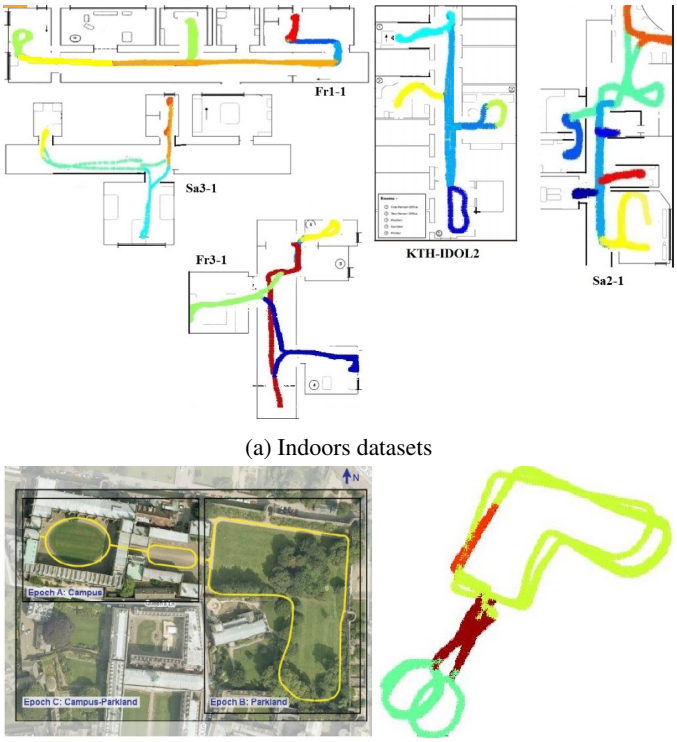
possible values, the results shown in Fig. 7 indicate that high localization performance is attained by choosing $\alpha = 0.3$ as the scaling factor for the vocabulary, while still resulting in a constrained number of visual words.

4.3. Overall Performance

The following experiments include all the available datasets and are designed to assess the final performance of our system. Among the rest, even though *New College* refers to a challenging outdoor and dynamic environment, it is not composed by different sequences in which the robot repeated the full trajectory at a different period of time. Nevertheless, to the best our knowledge, this is the only outdoor dataset in the literature with well-recognizable semantic locations, while also containing plenty of revisited regions. For this reasons, we have chosen to test with *New College*, although other similar methods do not make use of this dataset; no comparative results can be extracted for outdoors. For this case, the available images were divided into two subsets, each of which was treated as two different routes of the area.

4.3.1. Semantic clustering

Using the aforementioned set of parameters, we measured the total count of images correctly clustered into the respective semantic location. Table 2 shows the percentages of frames being clustered in a fault semantic region according to ground-truth. The rest of each dataset's instances were grouped in an appropriate segment. In the *New College* dataset, a semantic region (165 images) was falsely created, which is located in the semantic area of the garden. This inconsistency is due to the fact that the robot observes a totally different view with respect to the rest of that area. Such cases of falsely formulated clusters could potentially be solved by using of omni-directional cameras. In addition, 50% of *COLD-Saarbrücken seg.2 cloudy1* images was "miss-clustered", mainly due to the camera's instability, which resulted into blurred frames with inaccurate local point descriptors , or plane views with no discriminative visual



(a) Indoors datasets

(b) Outdoor dataset

Figure 9: Final semantic maps with their ground-truth blueprints.

information. However, such failures could potentially be resolved using a more sophisticated image-description technique, e.g., deep learning feature vectors. Also, in the case of *Freiburg seg.3 cloudy1*, during a small part of its route through the corridors, the view of the robot is directed towards the office area. This characteristic resulted in the categorization of 30% of its content in a different semantic cluster. Figure 8 presents some examples of wrongly clustered images and Fig. 9 shows the final created semantic maps for all datasets imposed over the ground-truth blueprints.

4.3.2. Localization and comparative results

In addition, our final system was tested with regards to the achieved semantic localization performance. The selected measurements refer to the percentage of correctly localized image instances to the respective ground-truth semantic areas. Table 3 shows the obtained results on every evaluated dataset. Our approach corresponds to an unsupervised method and does not require any labeled training data to extract semantic regions, while still achieving a high percentage of correct localization instances. More specifically, the comparative methods use the same datasets and similar experimental procedure as ours, therefore is justified correctly and robust, the direct acquisition of percentages from the papers, since that we did not find an public available code to execute tests. Hence, we provide our source code[1] so it can be used to evaluate our method in different environments. Due to their large presence in the literature,

[1] https://github.com/VasiaBalaska/Semantic-Clustering-and-Localization

Table 2: Percentage of "miss-clustered" ground-truth images in every semantic region.

Datasets	**Fr1-1(cloudy)**						**Fr3-1(cloudy)**				**Sa3-1(cloudy)**			
semantic communities	corridor	printer area	offices	store	stairs	bath	corridor	offices	stairs	bath	corridor	offices	bath	printer area
% of miss-clustered images	0%	27%	5%	0%	0%	22%	15%	**30%**	20%	8%	2%	3%	3%	3%

Datasets	**Sa2-1(cloudy)**						**KTH-IDOL2(1-1)(cloudy)**				
semantic communities	Lab	printer area	offices	bath	corridor	kitchen	printer area	corridor	office1	kitchen	office2
% of miss-clustered images	8%	4%	0%	**50%**	0%	25%	3%	1%	7%	20%	7%

Datasets	**New College**			
semantic communities	cycle	stoa	garden	false cluster
% of miss-clustered images	0%	0%	5%	**100%**

Table 3: Percentage of correctly localized instances on the semantic maps formulated through *cloudy* sequences. The below compared works applied **a supervised method**. Table entries marked with "–" correspond to measurements which are not available in the literature.

Methods /Datasets	**Proposed Method (F.-voting)**	**Proposed Method (BoW-voting)**	**Ullah et all., 2008**	**Pronobis et al., 2012**	**Wang et al., 2011**	**Rubio et al., 2014**	**Hasasneh et al., 2012**	**Fazl-Ersi et al., 2012**	**Mancini et al. 2017**
Fr1-1 Sunny	96%	95%	80.33%	-	81%	-	82%	-	-
Fr1-1 Night	87%	86%	78.57%	-	63.10%	-	88.76%	-	-
KTH Sunny	85%	85%	-	70.51%	-	64.41%	-	92.01%	93.62%
KTH Night	81%	81%	-	78.51%	-	60.93%	-	92.01%	93.62%
Fr3-1 Sunny	87%	86%	68.65%	-	68.9%	-	79.20%	-	-
Sa2-1 Night	90%	89%	69.11%	-	69.45%	-	71.2%	-	-
Sa3-1 Sunny	97%	95%	85.07%	-	91.30%	-	92.72%	-	-
Sa3-1 Night	90%	89%	85.22%	-	79.70%	-	81.94%	-	-

Table 3 contains additional comparative results with representative supervised methods for semantic localization. In the *New College* dataset, our proposed feature-voting method achieves 78%, while the accuracy of BoW-voting is found at 75%. As mentioned above, comparative semantic mapping methods are not available in the literature for this dataset, therefore we did not record it in the table. As can be seen, our approach achieved higher or comparable performance with the majority of semantic localization techniques. Specifically, our method is highly accurate as compared to the ones proposed by Pronobis and Jensfelt (2012), Ullah et al. (2008), Wang and Lin (2011), Rubio et al. (2014), and Hasasneh et al. (2012), which demonstrates its efficacy in relation to classification schemes (SVM and Bayes classifier) used in the above works. However, the methods in (Fazl-Ersi and Tsotsos (2012)) and (Mancini et al. (2017)) achieved higher results. We assign this behavior to the fact that both works utilized more accurate, although expensive to compute, feature vectors for their classifiers which are derived from Convolutional Neural Networks (CNNs) and Oriented Uniform Patterns, respectively. Another reason is that, contrary to both the aforementioned techniques, our method does not require labeled training samples to formulate the semantic regions. Even though, all the above supervised algorithms are typically capable to generalize in similar data, training off-line their model and applying it on arbitrary scenarios requires careful and time-consuming labeling of learning examples.

Table 4 contains additional comparative results with two unsupervised methods. Our higher performance as compared to the work of Erkent et al. (2017) is justified due to the usage of local SURF features from images instead of the bubble descriptors. Even though Guillaume et al. (2011) used sophisticated image descriptors, their method achieves lower accuracy than the proposed system since their parameters for the Self-Organizing Map do not generalize in each dataset.

5. Conclusions

In this paper a complete architecture for semantic segmentation and localization has been described. The available visual and odomentry data were combined in a clustering pipeline to produce the topological map of a previously unexplored environment. The results that emerged were constructive, concluding that by properly formulating the semantic map of an area, the correct categorization is achieved even under different lighting conditions.

Table 4: Percentage of correctly localized instances on the semantic maps formulated through *cloudy* sequences. The below compared works applied **an unsupervised method**. Table entries marked with "–" correspond to measurements which are not available in the literature.

Methods /Datasets	**Proposed Method (F.-voting)**	**Proposed Method (BoW-voting)**	**Erkent et al., 2017**	**Guillaume et al., 2011 (GIST descriptors)**	**Guillaume et al., 2011 (CENTRIST descriptors)**
Fr1-1 Sunny	96%	95%	66.20%	90.61%	88.58%
Fr1-1 Night	87%	86%	45.40%	85.90%	77.68%
KTH Sunny	85%	85%	-	-	-
KTH Night	81%	81%	-	-	-
Fr3-1 Sunny	87%	86%	-	-	-
Sa2-1 Night	90%	89%	44%	68.82%	88.24%
Sa3-1 Sunny	97%	95%	56.20%	95.30%	92.39%
Sa3-1 Night	90%	89%	44%	81.42%	89.68%

Nowadays, the scientific interest has been fixed on mapping techniques under extreme seasonal changes. To that end, our method could be applied straightforwardly on such applications as well, only by replacing the utilized local features/BoW with CNN-derived feature vectors (Sünderhauf et al. (2015a,b); Kenshimov et al. (2017)). In addition, our plans for future work include the combination of the semantic information from multiple agents (robots) to produce a unified map.

ACKNOWLEDGMENT

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Program «Human Resources Development, Education and Lifelong Learning» in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKϒ).

References

Balaska, V., Bambis, L., Gasteratos, A., 2019. Graph-based semantic segmentation, in: Advances in Service and Industrial Robotics, pp. 1–8.

Bampis, L., Amanatiadis, A., Gasteratos, A., 2017. High order visual words for structure-aware and viewpoint-invariant loop closure detection, in: Proceedings of the IEEE/RSJ International Conference onIntelligent Robots and Systems, pp. 4268–4275.

Bampis, L., Amanatiadis, A., Gasteratos, A., 2018. Fast loop-closure detection using visual-word-vectors from image sequences. The International Journal of Robotics Research 37, 62–82.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-Up Robust Features (SURF). Computer vision and image understanding 110, 346–359.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008, P10008.

Chella, A., Macaluso, I., Riano, L., 2007. Automatic place detection and localization in autonomous robotics, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 741–746.

Demir, M., Isil Bozma, H., 2015. Video summarization via segments summary graphs, in: Proceedings of IEEE International Conference on Computer Vision Workshops, pp. 19–25.

Erkent, Ö., Bozma, I., 2012. Place representation in topological maps based on bubble space, in: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 3497–3502.

Erkent, Ö., Karaoguz, H., Bozma, H.I., 2017. Hierarchically self-organizing visual place memory. Advanced Robotics 31, 865–879.

Fazl-Ersi, E., Tsotsos, J.K., 2012. Histogram of oriented uniform patterns for robust place recognition and categorization. The International Journal of Robotics Research 31, 468–483.

Fraundorfer, F., Engels, C., Nistér, D., 2007. Topological mapping, localization and navigation using image collections, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3872–3877.

Gálvez-López, D., Tardos, J.D., 2012. Bags of binary words for fast place recognition in image sequences. IEEE Transactions on Robotics 28, 1188–1197.

GhasemiGol, M., Sadoghi Yazdi, H., Monsefi, R., 2010. A new hierarchical clustering algorithm on fuzzy data (FHCA). International Journal of computer and electrical engineering 2, —.

Guillaume, H., Dubois, M., Emmanuelle, F., Tarroux, P., 2011. Temporal bag-of-words-a generative model for visual place recognition using temporal integration, in: Proceedings of the VISAPP-International Conference on Computer Vision Theory and Applications.

Hasasneh, A., Frenoux, E., Tarroux, P., 2012. Semantic Place Recognition based on Deep Belief Networks and Tiny Images, in: Proceedings of International Conference on Informatics in Control, Automation and Robotics, pp. 236–241.

Karaoguz, H., Bozma, H.I., 2014. Reliable topological place detection in bubble space, in: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 697–702.

Kenshimov, C., Bampis, L., Amirgaliyev, B., Arslanov, M., Gasteratos, A., 2017. Deep learning features exception for cross-season visual place recognition. Pattern Recognition Letters 100, 124–130.

Korrapati, H., Mezouar, Y., Martinet, P., 2011. Efficient Topological Mapping with Image Sequence Partitioning, in: Proceedings of the IEEE European Conference on Mobile Robots ECMR, pp. 227–232.

Kostavelis, I., Charalampous, K., Gasteratos, A., Tsotsos, J.K., 2016. Robot navigation via spatial and temporal coherent semantic maps. Engineering Applications of Artificial Intelligence 48, 173–187.

Kostavelis, I., Gasteratos, A., 2015. Semantic mapping for mobile robotics tasks: A survey. Robotics and Autonomous Systems 66, 86–103.

Kostavelis, I., Gasteratos, A., 2017. Semantic maps from multiple visual cues. Expert Systems with Applications 68, 45–57.

Krishnan, A.K., Krishna, K.M., 2010. A visual exploration algorithm using semantic cues that constructs image based hybrid maps, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1316–1321.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., Milford, M.J., 2015. Visual place recognition: A survey. IEEE Transactions on Robotics 32, 1–19.

Luo, J., Pronobis, A., Caputo, B., Jensfelt, P., 2006. The kth-idol2 database. KTH, CAS/CVAP, Tech. Rep 304.

Lynen, S., Bosse, M., Siegwart, R., 2017. Trajectory-based place-recognition for efficient large scale localization. International Journal of Computer Vision 124, 49–64.

Mancini, M., Bulò, S.R., Ricci, E., Caputo, B., 2017. Learning deep nbnn representations for robust place categorization. IEEE Robotics and Automation Letters 2, 1794–1801.

Mur-Artal, R., Tardós, J.D., 2017. ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics 33, 1255–1262.

Newman, M., Barabási, A., Watts, D., 2006. The structure and dynamics of networks Princeton University Press.

Nieto-Granda, C., Rogers, J.G., Trevor, A.J., Christensen, H.I., 2010. Semantic map partitioning in indoor environments using regional analysis, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1451–1456.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree, in: Proceedings of the IEEE computer society conference on Computer vision and pattern recognition, pp. 2161–2168.

Ozaki, N., Tezuka, H., Inaba, M., 2016. A Simple Acceleration Method for the Louvain Algorithm. International Journal of Computer and Electrical Engineering 8, 207.

Pandey, S., Khanna, P., 2014. A hierarchical clustering approach for image datasets, in: Proceedings of the IEEE International Conference on Industrial and Information Systems, pp. 1–6.

Pronobis, A., Caputo, B., 2009. COLD: The CoSy localization database. The International Journal of Robotics Research 28, 588–594.

Pronobis, A., Jensfelt, P., 2012. Large-scale semantic mapping and reasoning with heterogeneous modalities, in: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 3515–3522.

Ranganathan, A., 2012. PLISS: labeling places using online changepoint detection. Autonomous Robots 32, 351–368.

Ranganathan, A., Dellaert, F., 2007. Semantic modeling of places using objects.

Rubio, F., Flores, M.J., Gómez, J.M., Nicholson, A., 2014. Dynamic bayesian networks for semantic localization in robotics, in: Proceedings of the XV Workshop of Physical Agents: Book of Proceedings, pp. 144–155.

Rusu, R.B., Marton, Z.C., Blodow, N., Dolha, M., Beetz, M., 2008. Towards 3D point cloud based object maps for household environments. Robotics and Autonomous Systems 56, 927–941.

Shi, L., Kodagoda, S., Dissanayake, G., 2012. Application of semi-supervised learning with voronoi graph for place classification, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2991–2996.

Sibson, R., 1973. SLINK: An Optimally Efficient Algorithm For The Single-Link Cluster Method. The computer journal 16, 30–34.

Sivic, J., Zisserman, A., 2003. Video Google: A text retrieval approach to object matching in videos, in: null, p. 1470.

Smith, M., Baldwin, I., Churchill, W., Paul, R., Newman, P., 2009. The new college vision and laser data set. The International Journal of Robotics Research 28, 595–599.

Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.J., 2015a. On the Performance of ConvNet Features for Place Recognition, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4297–4304.

Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M., 2015b. Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free. Proceedings of Robotics: Science and Systems .

Traag, V.A., Waltman, L., van Eck, N.J., 2019. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports 9.

Tsintotas, K.A., Bampis, L., Amanatiadis, A., Gasteratos, A., 2018. Assigning visual words to places for loop closure detection, in: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 5979–5985.

Tsintotas, K.A., Bampis, L., Gasteratos, A., 2019. Probabilistic Appearance-Based Place Recognition Through Bag of Tracked Words. IEEE Robotics and Automation Letters 4, 1737–1744.

Ullah, M.M., Pronobis, A., Caputo, B., Luo, J., Jensfelt, P., Christensen, H.I., 2008. Towards robust place recognition for robot localization, in: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 530–537.

Uršič, P., Kristan, M., Skočaj, D., Leonardis, A., 2012. Room classification using a hierarchical representation of space, in: Proceedings of the IEEE/RSJ Intelligent Robots and Systems, pp. 1371–1378.

Wang, M.L., Lin, H.Y., 2011. An extended-HCT semantic description for visual place recognition. The International Journal of Robotics Research 30, 1403–1420.

Xu, Z., Xuan, J., Liu, J., Cui, X., 2016. MICHAC: Defect prediction via feature selection based on maximal information coefficient with hierarchical agglomerative clustering, in: Proceedings of the IEEE International Conference on Software Analysis, Evolution, and Reengineering,, pp. 370–381.

Yeh, T., Darrell, T., 2008. Dynamic visual category learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Zhang, H., Li, B., Yang, D., 2010. Keyframe Detection for Appearance-Based Visual SLAM, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2071–2076.