MDPI (/)

Journals (/about/journals) Topics (/topics) Information (/authors) Author Services (/authors/english) Initiatives (/about/initiatives) About (/about)

(https://www.cookiebot.com/en/what-is-behind-powered-by-cookiebot/)

Sign In / Sign Up (/user/login)

This website uses cookies

We use cookies to personalise content and ads, to provide social media features and to analyse our traffic. We also share information about your use of our site with our social media, advertising and analytics partners who may combine it with other information that you've provided to them or that they've collected from your use of their services.

Submit (https://susy.mdpi.com/user/manuscripts/upload?journal=sensors)

Search for Articles:

Title / Keyword

Author / Affiliation / Email

Sensors

All Article Types

Search

Advanced Search

Necessary

Preferences

Statistics

Marketing

Journals (/about/journals) / Sensors (/journal/sensors) / Volume 16 (/1424-8220/16) / Issue 4 (/1424-8220/16/4) / 10.3390/s16040493 /

sensors (/journal/sensors)

Show details

Allow all

Submit to this Journal (https://susy.mdpi.com/user/manuscripts/upload?form%5Bjournal_id%5D%3D3)

Allow selection

Review for this Journal (https://susy.mdpi.com/volunteer/journals/review)

Deny

Powered by **Cookiebot by Usercentrics** (https://www.cookiebot.com/en/what-is-behind-powered-by-cookiebot/)

Propose a Special Issue (/journalproposal/sendproposalspecialissue/sensors)

► Article Menu

MDPI (/)

Article Menu

(https://www.cookiebot.com/en/what-is-behind-powered-by-cookiebot/)

Academic Editors

João Valente

Antonio Barrientos

Subscribe SciFeed (/1424-8220/16/4/493/scifeed_display)

Recommended Articles

Related Info Links

More by Authors Links

Table of Contents

Open Access Article

Building an Enhanced Vocabulary of the Robot Environment with a Ceiling Pointing Camera

by
Alejandro Rituerto [1,*] (mailto:aleritu@gmail.com) (https://orcid.org/0000-0001-6738-3382),
Henrik Andreasson [2], **Ana C. Murillo** [1],
Achim Lilienthal [2] (https://orcid.org/0000-0003-0217-9326) and **José Jesús Guerrero** [1]

[1] Instituto de Investigación en Ingeniería de Aragón, Deptartmento de Informática e Ingeniería de Sistemas, University of Zaragoza, Zaragoza 50018, Spain

[2] Centre for Applied Autonomous Sensor Systems, Deptartment of Technology, Örebro University, Örebro SE-70182, Sweden

[*] Author to whom correspondence should be addressed.

Sensors **2016**, *16*(4), 493; **https://doi.org/10.3390/s16040493 (https://doi.org/10.3390/s16040493)**

Submission received: 13 November 2015 / Revised: 28 March 2016 / Accepted: 31 March 2016 / Published: 7 April 2016

(This article belongs to the Special Issue **Robotic Sensory Systems for Environment Protection and Conservation (/journal/sensors/special_issues/RSSEPC)**)

Download Browse Figures

(https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g001.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g002.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g003.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g004.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-

MDPI (/)

00493/article_deploy/html/images/sensors-16-00493-g005.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g006.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g007.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g008.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g009.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g010.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g011.png?1461897655) (https://pub.mdpi-res.com/sensors/sensors-16-00493/article_deploy/html/images/sensors-16-00493-g012.png?1461897655)

Versions Notes (/1424-8220/16/4/493/notes)

Abstract

Mobile robots are of great help for automatic monitoring tasks in different environments. One of the first tasks that needs to be addressed when creating these kinds of robotic systems is modeling the robot environment. This work proposes a pipeline to build an enhanced visual model of a robot environment indoors. Vision based recognition approaches frequently use quantized feature spaces, commonly known as Bag of Words (BoW) or vocabulary representations. A drawback using standard BoW approaches is that semantic information is not considered as a criteria to create the visual words. To solve this challenging task, this paper studies how to leverage the standard vocabulary construction process to obtain a more meaningful visual vocabulary of the robot work environment using image sequences. We take advantage of spatio-temporal constraints and prior knowledge about the position of the camera. The key contribution of our work is the definition of a new pipeline to create a model of the environment. This pipeline incorporates (1) tracking information to the process of vocabulary construction and (2) geometric cues to the appearance descriptors. Motivated by long term robotic applications, such as the aforementioned monitoring tasks, we focus on a configuration where the robot camera points to the ceiling, which captures more stable regions of the environment. The experimental validation shows how our vocabulary models the environment in more detail than standard vocabulary approaches, without loss of recognition performance. We show different robotic tasks that could benefit of the use of our visual vocabulary approach, such as place recognition or object discovery. For this validation, we use our publicly available data-set.

Keywords: visual vocabulary (/search?q=visual+vocabulary); computer vision (/search?q=computer+vision); bag of words (/search?q=bag+of+words); robotics (/search?q=robotics); place recognition (/search?q=place+recognition); environment description (/search?q=environment+description)

1. Introduction

Bag of Words (BoW) approaches are a common way to represent images based on a quantized feature space. They are broadly used in visual recognition problems, such as object or place recognition and image retrieval [**1**,**2**,**3**,**4**,**5**,**6**]. These techniques create a catalog of image features or words and describe each image as a vector of occurrence counts of these words.

A typical drawback using standard BoW approaches is that semantic information is usually neglected when grouping the visual features into the clusters or visual words. In a general setting, including conceptual information is challenging since no assumption can be made about the type and meaning of the visual features that may appear. However, many applications could benefit from including semantic content in the vocabulary to achieve their goals. Our objective is to build an improved vocabulary that provides a more meaningful model for applications using BoW techniques. We focus on mobile robotic applications which imply long-term operations indoors, in environments such as a warehouse [**7**] or a museum [**8**].

Our approach is built after two intuitive hypotheses. First, robotic platforms provide sequential information (a sequence of images) including tracking information, while building the vocabulary helps to cluster the different appearances of a scene element. Second, cameras in robotic platforms have a fixed location to the robot, respectively. We can find spatial restrictions and make reasonable assumptions about the location of elements of the environment (e.g., the lamps are on the ceiling, the posters on the wall, ...). We use simple geometric information together with the image appearance to cluster the environment elements.

As done previously in multiple robotics applications [**9**,**10**], we benefit from having a camera pointing to the ceiling. Such camera configuration improves the operation over long periods of time, something crucial to address indoor monitoring applications with autonomous systems. Upper parts of indoor scenes are typically less dynamic than the rest of the scene and provide a more robust visual model. In this setting, we expect the environment to include a small number of repeatable elements (different kinds of lamps or windows) and a few elements which are rather unique (exit signs, posters or labels), most of them with fixed locations within the scene. We assume that these elements present a restricted set of appearances from a few points of view that are reachable by our robot.

Our approach consists of a novel hierarchical process (summarized in **Figure 1**). It creates a visual vocabulary with richer information about the environment than standard BoW methods. Instead of computing the visual words directly from the acquired data, we propose obtaining an initial grouping into classes containing similar sets of tracked key-points. Then, standard vocabularies are computed in each of those classes.

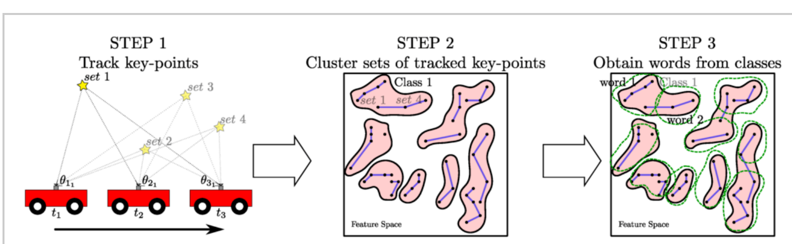


Figure 1. Diagram of our novel vocabulary construction method. During the acquisition, extracted image key-points are tracked. All the image appearances and altitude angle values of these key-points are stored and grouped together under the same set number. These sets of tracked key-points are later clustered into classes, with the elements in each class representing the same element of the environment. Finally, words are extracted running a clustering for each of the obtained classes.

The main differences and contributions of our approach with regard to prior work are the following:

- We leverage the visual vocabulary for a more meaningful representation of a given working environment;
- We propose a novel way to include spatio-temporal information in the vocabulary construction process: thanks to feature tracking, our approach automatically groups the different appearances of scene elements;
- Additionally, we propose including each key-point altitude value in the key-point descriptors to encode the different viewpoints for the scene elements.

Our experimental section demonstrates how the presented approach builds a visual model that provides higher representativity of the environment. At the same time, the created vocabulary maintains the same performance for place recognition as a standard vocabulary.

2. Related Work

How to acquire a representative visual model of the environment is a problem that has been studied for a long time. In particular, acquiring models of the environment for long term operation is a subject of great interest, since it provides intelligent systems with higher autonomy [**11**,**12**]. Additionally, enhancing those models with semantic information is a key element for human–computer interaction [**13**,**14**,**15**,**16**,**17**].

Seeking for robust models across time, indoor vision based robotic tasks have taken advantage of using visual ceiling features [**9**,**10**]. Elements on ceilings are usually more stable over time than those in floors or walls, where dynamic scene elements often appear. Recent results revisited this idea for indoor and industrial oriented robotic applications [**18**,**19**], where long term operation is required.

Robustness during long term operations is an issue of great importance when designing autonomous systems for indoor environment monitoring and surveillance [**20**,**21**,**22**]. In such applications, the robot needs to reliably localize itself in a known environment and have some semantic information that allows the system to measure parameters of interest or discover certain events, such as the presence or absence of common objects or changes in the environment infrastructure (potential changes may be due to broken components, leaks, *etc.*).

One of the most popular approaches to build visual models is the Bag of Words (BoW) representation. It is based on a quantization of the image feature descriptor space into several groups (visual words) that compose the visual vocabulary. BoW based approaches are very popular in various recognition tasks due to their good performance despite the simplified representation. Authors of [**23**] perform a survey of current visual topological mapping methods concluding that BoW approaches are better than global descriptor or local feature based approaches for large scale operation. We find good results on recognition at large scale achieved partly thanks to BoW image representation, such as appearance based localization and mapping in large scenarios [**5**], object retrieval approaches [**2**], or event detection methods [**24**].

In a seminal work from Sivic *et al.* [**1**], authors propose to use tracked features to build a visual vocabulary because they are more likely to be stable features. Inspired by these ideas, we use key-point tracking not only to find stable features but to discover the different appearances of the environment elements from a different viewpoint. In [**25**], authors extract the different appearances of the elements from a structure-from-motion point cloud of the environment. Given the different viewpoints, different image descriptors, of a 3D point, the number of appearances is reduced using mean-shift clustering. Using these set of appearances, authors localize new images of the environment in the created 3D map. Work in [**26**] makes use of feature tracks that in this case are clusters of image patches of similar appearance. Authors infer relationships between visual words given these feature tracks and learn a new metric to compare image features. This metric estimates the probability of observing a word of the vocabulary given an observed word in the query image. Probability will be high if multiple feature tracks include features assigned to both words and low otherwise. Contrary to our proposal that tracks the key-points of the scene to detect their appearance, these works group features by matching their descriptors.

In parallel with the growing popularity of vocabulary based recognition approaches, we find research results analyzing their drawbacks [**27**,**28**], such as the loss of discriminative information or under-representation of descriptor space regions with low population. Therefore, we also find approaches trying to overcome some of these issues, as well as augmenting the model semantic information. For example, [**29**] presents a novel framework for object category recognition that unifies visual codebook generation with classifier training. Traditional visual vocabulary learning and weighting is performed independently, and authors of [**30**] present Joint-ViVo, a method where words and their weights are learned jointly. In [**31**], authors improve the BoW creation by selecting informative words depending on a saliency value computed in the images. The approach described in [**32**] proposes a supervised learning algorithm which generalizes the *k*-means algorithm by allowing soft assignments and exploits supervised information to improve the discriminative power of the vocabulary. Work in [**33**] studies the problem of large scale image retrieval by developing a new class of bag-of-features to encode geometric information of objects within an image. Authors of [**34**] present a framework to integrate semantic information from Flickr labels for supervised vocabulary construction. The work in [**14**] propose a framework to solve the missing correspondences between low-level multimedia data and high-level semantics. This framework can be used to link the elements in a visual vocabulary with the elements of a semantic vocabulary.

One of our intuitions is that geometric and spatial restrictions on the scene are important cues to improve our visual model. Therefore, we include spatial information in the form of the key-points altitude as part of our key-points descriptor. Previous work has already proved the advantages of using geometric cues in visual models. For example, the work in [**35**] presents the Geometric min-Hashing. This method uses semi-local geometric information to create and increase the discriminative power of the hash keys, improving their visual model. In [**36**], spatial pyramids are used to characterize the spatial arrangement of words in the image. The methods in [**37**,**38**] include the relative spatial configuration of the words to improve the vocabulary. The first work introduces a bag of spatio-visual words representation (BoSVW) obtained by clustering of visual words correlogram. The second work uses spatial arrangement of visual words (WSA) for image retrieval and classification outperforming spatial pyramids in the retrieval scenario. In [**39**], angle and scale of image descriptors are used to improve image matching. Authors weigh the matching score of a query and a database image by verifying the consistency of angle and scale of all the matched words between the images.

Finally, there is another important group of related work regarding unsupervised learning for object or feature discovery. Unsupervised learning has been used to discover the distinctive architecture elements of certain areas [**40**], representative views of an object [**41**], object models [**42**] or the appearance of objects and their image segmentation [**43**]. Closer to our approach, the work in [**44**] finds a set of discriminative and representative image patches by using an iterative process of clustering and training. In [**45**], authors construct a compact and discriminative semantic visual vocabulary using diffusion maps and quantized mid-level features.

3. Enhanced Vocabulary Construction

This section details our approach to build an enhanced vocabulary of the environment traversed by a mobile camera. Steps are summarized in **Figure 1**: (1) feature detection and tracking that groups the features into sets of tracked key-points; (2) clustering sets of similar appearance into classes; (3) obtaining the visual words from each class.

3.1. Key-Point Detection and Tracking

We detect key-points in the scene and track them using the Lucas–Kanade tracker. For each frame, we compute an image descriptor around each key-point location. For both key-point detection and image descriptor computation, we use SURF (Speed-Up Robust Features) [**46**]. Note that the appearance of key-points is likely to change from frame to frame while they are tracked. We can exploit the knowledge of these appearance variations being generated by the same entity thanks to the tracking. Due to the camera configuration, pointing towards the ceiling, the altitude angle encodes the point of view change that produces the appearance change of the scene point. The tracked image descriptors are stored, together with the altitude θ of the corresponding location, as shown in Equation (1).

As a result of this step, our approach has identified m sets of tracked key-points, where each set, $set\ i$, contains n_i image descriptors and altitude angles corresponding to the same tracked scene point from different points of view. $\mathbf{X}_{j_i}$ denotes the jth descriptor and altitude angle pair included in $set\ i$:

$$set\ i = \{\mathbf{X}_{j_i} = [desc_{j_i},\ \theta_{j_i}]\} \quad (1)$$

with j in $[1 \ldots n_i]$ and i in $[1 \ldots m]$.

Different sets could contain features that belong to the same scene element, e.g., scene points tracked at different time while revisiting the same location or points corresponding to repeated objects in the environment. The following step of our approach groups sets that are likely to have been produced by the same object or element.

3.2. Clustering Sets of Tracked Key-Points

To cluster the sets of tracked key-points, we have evaluated two clustering methods: Hierarchical Clustering and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Both methods can work based on a similarity measure between sets.

3.2.1. Similarity Measure

The similarity measure described next estimates how likely two key-point sets are to correspond to the same scene element. The altitude value encodes the relative position of the scene point with respect to the camera. We assume that the appearance of a scene point from the same viewpoint is the same, but different scene points can look similar from different positions. We use the altitude difference to penalize these cases.

The distance between two features $\mathbf{X}_i = [desc_i,\ \theta_i]$ and $\mathbf{X}_j = [desc_j,\ \theta_j]$, is computed using:

$$d(\mathbf{X}_i, \mathbf{X}_j) = ||desc_i, desc_j|| f(|| \theta_i, \theta_j ||) \quad (2)$$

where $\|desc_i, desc_j\|$ is the Euclidean distance between the appearance descriptors and $f(x)$ is the penalization due to the altitude difference between both features obtained as follows:

$$f(x) = (1 + a\ \exp\ (b\ \exp\ (c\ x))) \tag{3}$$

where *a*, *b* and *c* are the parameters that define the shape of the penalization.

The penalization function is shown in **Figure 2**. It grows rapidly and continuously after a small difference of altitudes $(0.1$ rad), and the maximum penalization is used for values above 0.8 rad. Parameters *a*, *b* and *c* have been selected to accomplish these requirements. Other functions with the same properties are also valid. The selected function penalizes the matching of key-points observed from different altitude angles but does not modifies the matching other rotations of the key-points. In such cases, the rotation invariance of the chosen SURF descriptor allows for recognizing the key-point.

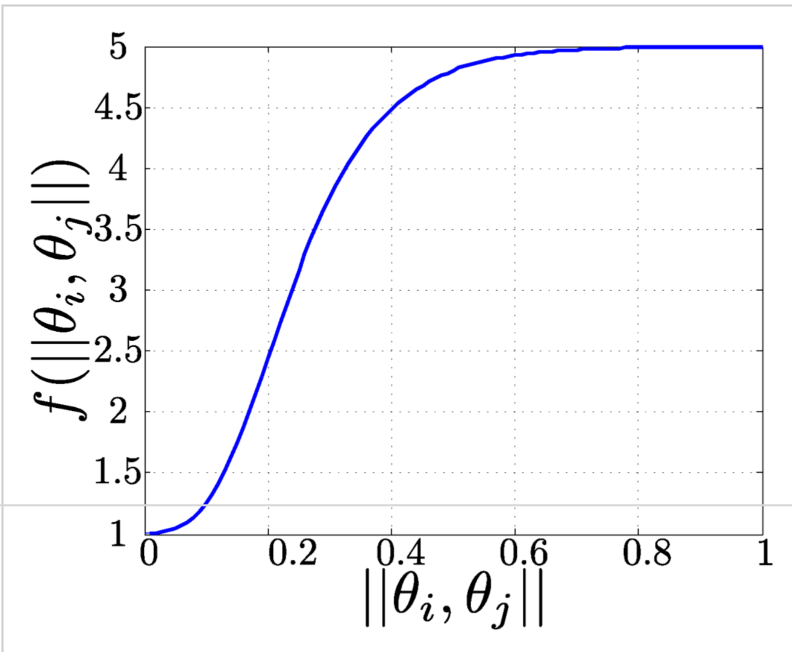


Figure 2. Penalization Equation (**3**): $a = 4$, $b = -7.5$ and $c = -10$.

To compare two sets of tracked key-points, we have to compare all the features of one set with all the features from the other set. Each feature of $set\ \ i$ is compared with all the features of $set\ \ j$, the minimum distance value is selected, and the mean of all these minimum values is considered as the distance between sets:

$$D\,(set\ \ i, set\ \ j) = \underset{\mathbf{X}_k \in set\ \ i}{mean}\ (\ \underset{\mathbf{X}_l \in set\ \ j}{\min}\ (d\,(\mathbf{X}_k, \mathbf{X}_l))) \tag{4}$$

where $D(set\ i, set\ j)$ is the distance between two different sets, $set\ i$ and $set\ j$, and $\mathbf{X}_k$ and $\mathbf{X}_l$ are features included in these sets, respectively.

3.2.2. Clustering Approaches

We consider two common clustering approaches that build on a similarity measure between the elements to be clustered.

- *Hierarchical Clustering.*

  We have implemented a clustering method based on the agglomerative Hierarchical Clustering [47], where each element starts as a cluster and each cluster is paired with its most similar cluster as we move up in the hierarchy. Our hypothesis is that elements in the same cluster are probably observations originated from the same object or scene point. We modify the standard Hierarchical Clustering by defining threshold th_S to avoid merging too dissimilar clusters (clusters are not paired if their distance is above this threshold). As a result of this modification of the standard Hierarchical Clustering, key-points sets far from any other set are not paired and compose a singleton cluster. In this method, new clusters are created on every iteration, so new distances have to be computed. We adopt the Unweighted Pair Group Method with Arithmetic Mean [48] approach, where the distance between two clusters is the mean of the distances between the sets of tracked key-points included in each cluster.

  Hierarchical Clustering is conceptually simple, that means it is easy to implement and modify. Additionally, it outputs a hierarchy of clusters, a structure more informative than flat clustering technique results. The drawback is its complexity, $O(n^3)$ in the general case, what makes it too slow for big data-sets.

- *DBSCAN (Density-Based Spatial Clustering of Applications with Noise).*

  DBSCAN [49] is a density-based clustering algorithm that uses a estimated density distribution of corresponding nodes to find clusters in the data. This algorithm is based in the notion of density reachability: Two elements, *q* and *p*, are directly density reachable if their distance is not bigger than *ε*. *q* is called density-reachable from *p* if there is a sequence of elements, $p_1 \ldots p_n$ with $p_1 = p$ and $p_n = q$, where each p_{i+1} is directly density-reachable from p_i. With these definitions, a cluster is a subset of elements mutually density-connected. To handle the noise, this method defines the parameter $minPts$, the minimum number of elements required to create a cluster. Subsets of density-connected elements with less than $minPts$ elements are considered as noise.

  DBSCAN is a widely used clustering technique. We use the DBSCAN implementation included in ELKI open source data mining software [50]. The complexity of this method is lower than for Hierarchical Clustering, $O(n^2)$ for the basic form of the algorithm, so it is faster and more appropriate for big data-sets.

The results of this clustering step can be semantically understood as follows:

- The obtained clusters represent common scene points and include their possible appearances according to the different viewpoints under which the scene elements were observed.
- Non paired or noisy sets are unique scene points. These sets are dissimilar to the rest of sets but may be highly representative of the locations where they appear. These unique sets are clustered together after running the clustering step. The created cluster does not represent any common scene points, but includes unique key-points representative of a location. Unique sets are clusters with just one set of tracked key-points when using Hierarchical Clustering. When running DBSCAN with $minPts = 2$, these unique sets are considered as noise.

3.3. Vocabulary Construction

Our last step to obtain the visual model consists of building a vocabulary for each class resulting from previous step. The whole vocabulary is composed by a fixed number of words *K*. We assign a number of words k_i to each class *i*. k_i is proportional to the number of features included in class *i*:

$$k_i = \left\lceil K \frac{\#\, features \ \in class\ i}{\#\ total\ features} \right\rceil \tag{5}$$

K-means clustering algorithm is run with the elements within each class. Differently from previous steps, *k*-means is run using only the appearance descriptors of the features in that class. Each word gets a representative appearance descriptor from *k*-means. We add an altitude value computed as the average altitude of all features assigned to that word.

In the resulting vocabulary, larger classes are represented by more words than small ones. The class including non paired sets, which is usually big, will receive a large amount of words, which guarantee that we account for these marginal and unique scene elements.

3.4. Assigning Words to a New Feature Using the Created BoW

To classify a new feature, $\mathbf{X}_{new}$, into the discovered classes, it is compared with the words included in the vocabulary using the distance described in Equation (**2**). We assign the corresponding word *i* according to the nearest neighbor, but only if that distance is below a matching threshold, th_M.

$$i = \arg\min_{i\in[1,k]} \left(d\left(\mathbf{X}_{new}, \mathbf{X}_i\right) | d\left(\mathbf{X}_{new}, \mathbf{X}_i\right) < th_M \right) \tag{6}$$

By using this threshold, we model the fact that a new feature could not belong to any of the modeled classes.

4. Analysis of the Performance of the Hierarchical Vocabulary

We have evaluated all steps of our approach and the performance and properties of the obtained visual vocabulary.

4.1. Experimental Settings

4.1.1. Data-Sets

We validate our method with a data-set acquired from a robotic platform at the AASS laboratories and offices in Örebro University, Sweden. It includes two image sequences of two different trajectories around the same environment: 158.5 m (1879 frames) and 164 m (2142 frames). They were acquired at different days, in the same environment and follow different trajectories. The acquisition was performed at 30 frames per second, and the images have a resolution of 768×768 pixels. As explained previously, the camera has been set pointing to the ceiling, therefore the objects that appear are mostly light sources, windows and signs. **Figure 3** shows some sample images of this data-set, which is available online [**51**].

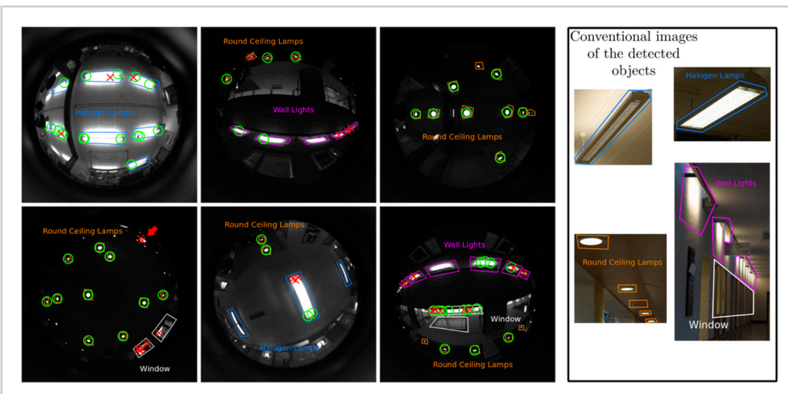


Figure 3. Examples of objects discovered in the data-set. Green circles mark correct assignments and red crosses show incorrect ones. The colored shapes show the labeled areas: round Ceiling Lamps (**orange**); Halogen lamps (**blue**); wall Lights (**magenta**); windows (**white**). Matching threshold is set to 0.15. For clarity, some conventional images of the objects in the environment are shown on the right of the figure. (Best seen in color).

A second data-set is used for qualitative evaluation of the method. This sequence has been acquired in a trajectory of about 10 m on a different indoor environment, traversing a corridor. The purpose of this sequence is to further analyze the correspondence between classes and real objects in a different scenario than the main data-set used.

4.1.2. Performance Measurements

The proposed method is evaluated attending to three different criteria:

- *Accuracy:* it evaluates the accuracy of the vocabulary to classify new features into the discovered classes. Total, A_{Total}, and average class accuracy, $A_{Average}$, of the classification are respectively computed as:

$$A_{Total} = 100 \frac{\#\ correct\ class}{\#\ test\ features} \tag{7}$$

$$A_{Average} = mean\left(100 \frac{\#\ correct\ class_i}{\#\ test\ features_i}\right) \tag{8}$$

where the *i* index represents the classes.

- *Normalized inverse pixel deviation:* it evaluates the similarity of the key-points patches included in each class. This quality measurement is based on the standard deviation of the image patches of features that have been clustered together. Given class *i*, we define the class pixel deviation, S_i, as the mean of the standard deviation of the gray level of every pixel of the features patches included in class *i*:

$$S_i = \underset{\forall (x,y)}{mean} \left(\underset{\forall j \in i}{std} \left(I_j (x,y) \right) \right) \tag{9}$$

where $j \in i$ represents all the patches of the features included in the class *i*, $I_j(x,y)$ is the gray level of pixel (x,y) from the patch of feature *j*, (x,y) values are limited to the size of the patches (32×32 pixels in our case) and $std()$ is the standard deviation.

We define the normalized inverse pixel deviation for each class, S'_i:

$$S'_i = 1 - \frac{S_i}{S_{max}} \tag{10}$$

where S_{max} is the maximum pixel deviation. More meaningful classes will have higher S'_i values.

- *Intra-class distance:* it evaluates the similarity between all the sets of key-points included in each class. Distance between all the sets of key-points is computed using Equation (**4**). The intra-class distance is the mean of these distances. Lower values of this distance mean more compact clusters, where the sets grouped are more similar.

Normalized inverse pixel deviation and intra-class distance both evaluate how similar are the elements grouped under the same class label. However, the first one computes distances between key-point patches, just the key-points appearance, while the latter computes distances between sets of key-points using tracking and viewpoint information together with the SURF descriptors.

4.1.3. Comparison with *k*-Means Vocabulary

Next, experiments compare the properties of our proposed vocabulary with those of the standard *k*-means vocabulary. We have chosen this technique as baseline for the visual vocabularies. Standard *k*-means has shown good performance in plenty of computer vision applications, but one of its drawbacks is that not semantic information is considered when building the words. In contrast, the hierarchical vocabulary presented in this work groups in the same class key-points that are likely to come from the same scene element. **Figure 4** shows the basic differences between both approaches.

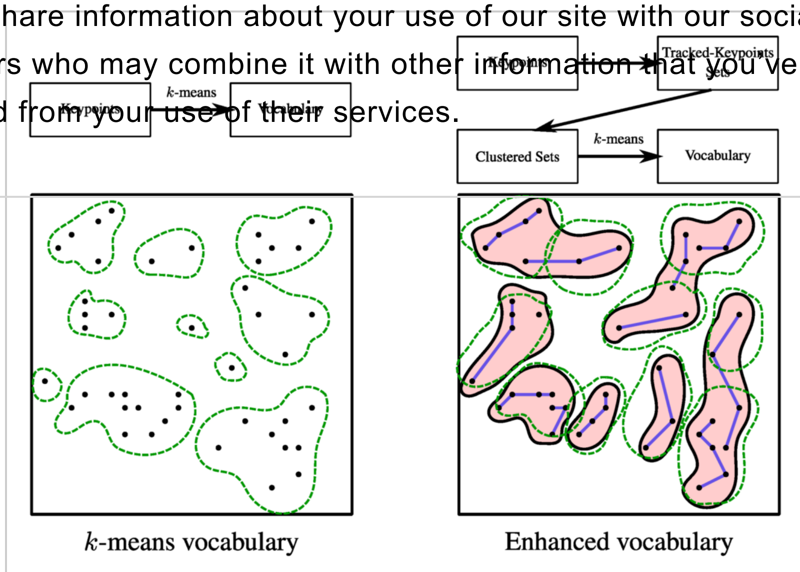


Figure 4. Differences between standard *k*-means vocabulary and our hierarchical vocabulary building process. *k*-means clusters directly in the feature descriptor space. Our enhanced vocabulary first groups by tracking, then clusters tracked key-point sets into classes, and finally clusters with a k-means within each of the found classes to create the final visual words.

4.2. Analysis of the Clustering of Tracked Key-Points Sets

Figure 5 shows the evolution of the main parameters of the clustering results for different configurations of the two evaluated clustering algorithms. We can observe how the intra-class distance and the number of non paired sets have similar behavior for values of th_S and ε between 0.05 and 0.35. However, DBSCAN creates less clusters than Hierarchical Clustering. In DBSCAN, two elements can be clustered together if they are density reachable, even if their distance is high. In contrast, in our modified Hierarchical Clustering, to pair two elements the distance between these elements must be lower than th_S. The requirement to cluster elements is more relaxed for DBSCAN, so more sets are clustered together and less clusters are created than with Hierarchical Clustering. For ε higher than 0.35, very few clusters are created with

DBSCAN. This means that dissimilar sets are clustered together affecting the quality of the results as can be seen in the evolution of S' and the intra-class distance. **Figure 5**d shows a peak on the number of non-paired sets when $th_S > 0.5$. This is caused because there are no sets which similarity distance with other set is below that threshold. As a result of this, all sets are grouped together in the non-paired cluster.

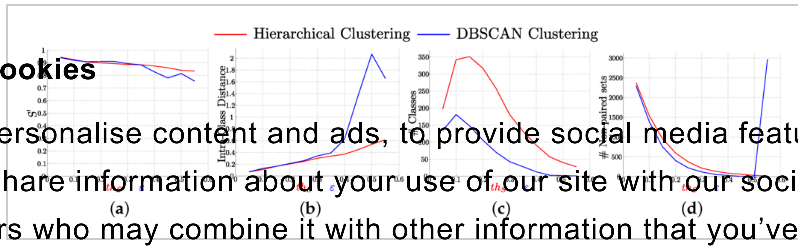


Figure 5. Comparison of the clustering results for both studied methods: Hierarchical Clustering and DBSCAN Clustering. Evaluation of the normalized inverse pixel deviation (**a**); intra-class distance (**b**); number of created classes (**c**) and number of non paired classes (**d**) for both Hierarchical Clustering (**red**) and DBSCAN (**blue**) clustering. th_S and ε are equivalent parameters for both clustering methods respectively. (Best seen in color).

4.3. Influence of th_S, ε and K Parameters in the Performance of the Resulting Vocabulary

For this analysis, we randomly split the sequences and use 70% as training, and 30% as test data. The clustering of sets of tracked key-points into classes is performed on the whole sequence to define a ground truth class assignment for both train and test data. The vocabulary is built using only the training data and the next results correspond to the classification of the test data into vocabulary classes. **Figure 6** shows the Total and Average accuracy of the classification and the Normalized inverse pixel deviation.

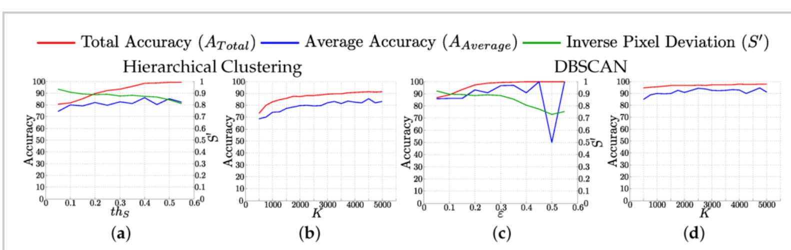


Figure 6. Influence of parameters, th_S, ε and *K* in the performance of the vocabulary. We show the evolution of total (**red**) and average (**blue**) accuracy and normalized inverse pixel deviation (**green**) for th_S (**a**) and ε (**c**) with $K = 3000$ and for *K*; (**b**) and (**d**), for th_S and ε equal to 0.1.

For Hierarchical Clustering, A_{Total} grows with th_S, while $A_{Average}$ remains constant. This means that larger classes get even larger by clustering dissimilar sets of tracked key-points.

A_{Total} increases due to the good performance of these large classes, but $A_{Average}$ remains constant due to the poor performance of the small classes. As expected, the quality of the classes, represented by S', decreases when th_S grows.

For DBSCAN clustering, we can observe how both, A_{Total} and $A_{Average}$ grow with ε while S' decreases. The high values observed for A_{Total}, are effect of the low number of clusters created. The chance of classifying correctly a key-point when most of the features are part of the same cluster is very high. A_{Total} and $A_{Average}$ remain almost constant when *k* grows. Those values are in all cases higher than for Hierarchical Clustering because the first class created by DBSCAN clustering is too big, as we show later in **Section 4.5**. Most of the test key-points are classified under that class, so the accuracy is high. However, this class is compounded by dissimilar sets of tracked key-points. High values of *K* produce large vocabularies that better fit the feature space. However, one of the goals of using vocabularies is to reduce the number of elements included in the model, so small *K* values are preferred.

4.4. Influence of the Robot Motion in the Detection of Key-Point Classes

This experiment analyzes the suitability and correctness of our vocabulary to model the environment. We classify key-points from a test sequence not used to build the vocabulary into the classes discovered by our approach. **Figure 7** shows examples of key-points classified along different test sub-sequences. Each column of the figure shows key-points detected in those frames together with the word and corresponding class assigned to that key-point. Note that as the key-point appearance and position varies, it is assigned to different words but is still classified under the same class. This validation demonstrates for a given scene element, *i.e.*, one of our classes, how the different words encode the element appearances and viewpoints.

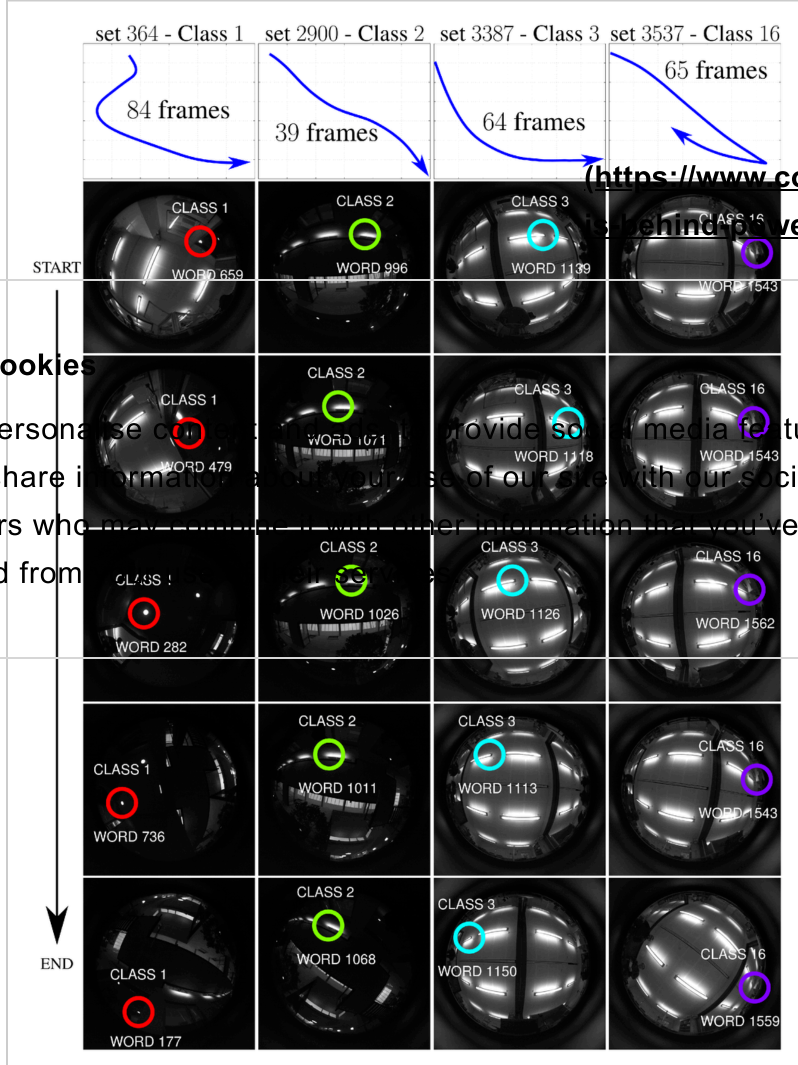


Figure 7. Key-point classification into words and classes as the robot moves. The plots in the first row represent the trajectories followed by the robot while acquiring the frames. The trajectories include rotation, translation and combinations of both. Each column shows how the key-points of a set of key-points are individually classified as being of the same class. All the key-points of a set correspond to the same element of the scene that has been tracked. (Best seen in color).

4.5. Analysis of the Object Information Included in the Vocabulary

One interesting property of our method is that the classes created by clustering sets of tracked key-points are related to scene elements of the environment. We want to analyze if these discovered classes actually correspond to objects or parts of the environment (note that every step of our method is unsupervised, so we do not have any concept name associated with any of the classes).

For this experiment, we use both sequences, one for obtaining the vocabulary and the second one for testing. In both sequences, we have labeled manually 10% of the images with bounding boxes around the four most repeated objects in the environment: Halogen Lamp, Round Ceiling Lamp, Wall Light and Window (see **Figure 3**). Since the camera is pointing to the ceiling, the objects that appear are mainly different kinds of lamps and windows.

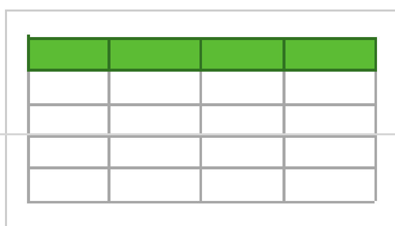
4.5.1. Relationship between Words and Classes with the Environment Objects

First, we analyze how our proposal creates a visual vocabulary where the classes (and words subsequently) are related with the environment objects. A quantitative analysis of the relation between scene objects and the classes created by our vocabulary can be seen in **Table 1**. It shows the normalized entropy values for the classification of the objects into classes and words. Normalized entropy measures the unpredictability of a random variable. In our case, the random variable is the classification of a key-point generated by a scene element into words or classes. If the key-points of an object are always classified in the same word, the entropy will be zero. However, if object key-points are assigned into a lot of different words, the entropy value will be high. To compare the entropy of different vocabularies, we use the Normalized Entropy:

$$Normalized\ Entropy = \frac{\left(-\sum_{i=1}^{n} f_i \log_2 (f_i)\right)}{\log_2 (n)} \quad (11)$$

where, for each object, f_i is the proportion of occurrences of that object in the class word *i*, and *n* is the number of classes or words.

Table 1. Normalized entropy of the object classification into classes or words.



Analyzing the classification at word level, the three vocabularies have a similar normalized entropy value (about 0.48). However, looking at the values for the classification into classes, the values are much lower (0.275 and 0.394 for our approach with DBSCAN and with Hierarchical Clustering, respectively). Using a standard *k*-means vocabulary (c), there is no relation between words and scene objects and words generated from the same objects are not related in any way.

4.5.2. Classes Representing Objects

The following experiments show how the relation between the classes created by our approach and the objects of the environment can be used to detect the different object occurrences.

First, we define the concept of representative classes. A class is representative of an object if most of the class key-points are also key-points of that object:

$$100 \frac{\#\ class_i \text{ key-points in } object_j}{\#\ class_i \text{ key-points}} \geq th_R \tag{12}$$

where th_R is the representativity threshold which models how unique a class needs to be to be representative of an object.

Figure 8 shows the number of representative classes for each object and the number of representative classes associated to more than one object for different values of th_R. Even for high values of th_R, we find representative classes for all the objects. The number of representative classes associated to more than one object is small for low values of th_R, and null for th_R higher than 35%

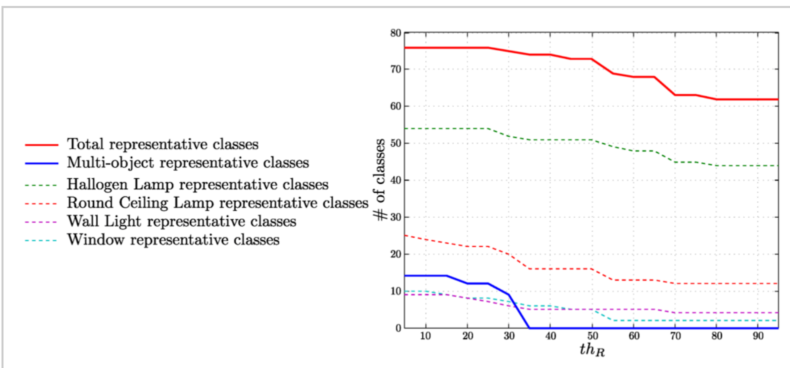


Figure 8. Representative classes for each object and representative classes associated to more than one object (multi-object classes) for different values of th_R. In the analyzed vocabulary, 342 classes were created.

Next, we want to quantify how new occurrences of objects are recognized using the representative classes. The representative classes of each annotated object are selected in the training data with th_R equal 50%. In the test data, key-points detected as being of a representative class of an object are labeled as being generated by that object. **Figure 9**a shows precision and recall when classifying features into object labels for different values of the similarity threshold, th_M.

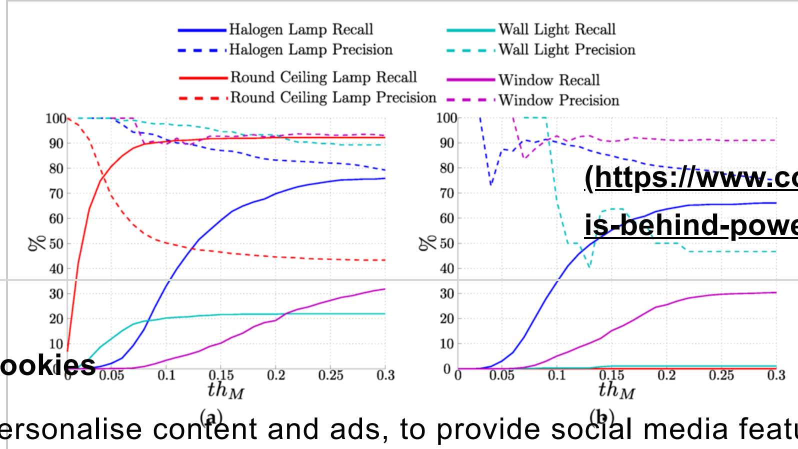


Figure 9. Precision (**dashed line**) and Recall (**solid line**) for the detection of real objects as the matching threshold th_M varies. We can see better performance when including altitude in the descriptor, together with SURF appearance (**a**) that using only SURF (**b**).

The highest recall, above 90%, was achieved for the detection of Round Ceiling Lamps, which appear in almost all the rooms traversed during the sequence. However, the precision is low for this class, below 50% for higher values of th_M. For the rest of the objects, precision is above 90%. For Halogen Lamps, recall reaches almost 80%, while for Wall Lights and Windows recall values are lower, about 20% and 30%, respectively. Those objects have similar appearance and the areas of the image where they appear are the same so they are really hard to distinguish.

Inclusion of the Altitude Values

Figure 9b shows the same plot that **Figure 9**a, but, in this case, the altitude has not been used in the process. While the results are similar for Halogen Lamps and Windows, the accuracy for Wall Lights decreases from around 20% to 2% when we do not use the altitude. Even more dramatic is the change in the detection of Round Ceiling Lamps: without the altitude value, no Round Lamps are detected. In our case, the image descriptors of features around the Round Ceiling Lamps are similar to the descriptors created by other entities of the environment, so the appearance descriptor is not enough to distinguish this object.

4.5.3. Qualitative Object-Class Correspondence

Figure 10 shows examples of the correspondence between the classes created by our vocabulary and the elements of the environment. For this experiment, we have run our vocabulary creation method in the second sequence presented. Again, most of the objects detected are lamps, halogen lamps (on and off) and emergency lights. Ceiling tubes, posters and labels are also detected. Occurrences in different frames are shown for most of the classes.

Note that classes 33 and 17, which correspond to wall signs, are detected in both sides of the corridor.

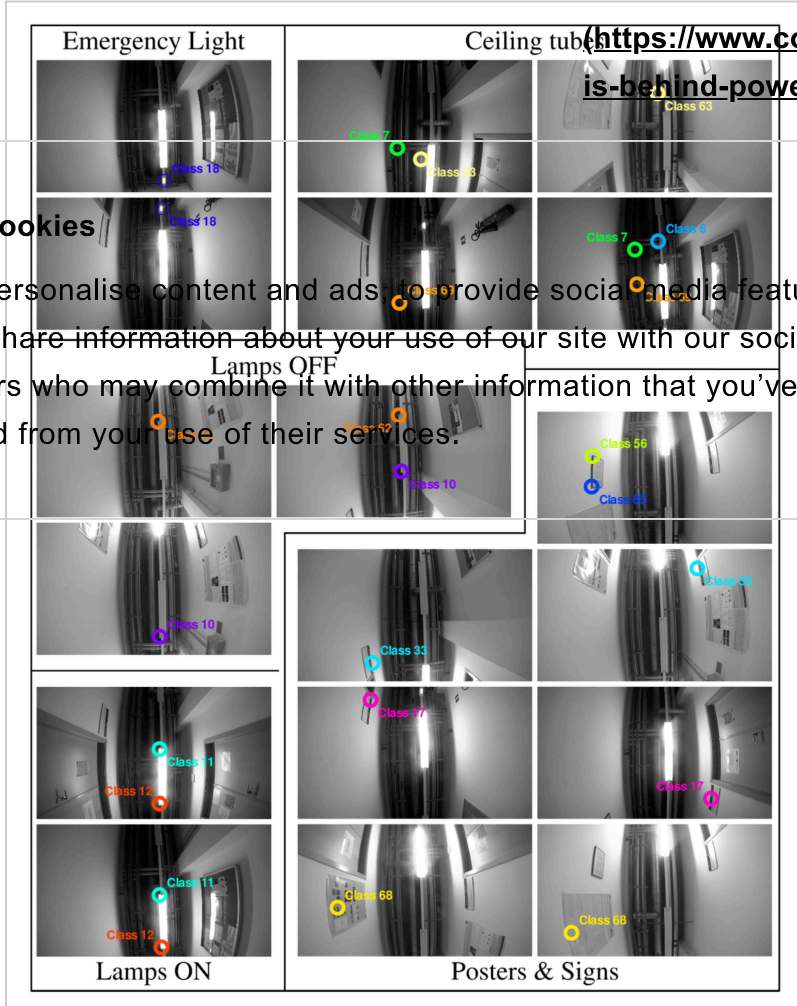


Figure 10. Examples of classes created by our method that correspond to real scene elements. Most of the classes are detected on various frames, and on different sides of the images. (Best seen in color).

5. Applications Using the Proposed Vocabulary

This section describes two possible applications of the vocabulary presented in this work. Place recognition and object detection are studied here.

5.1. Place Recognition

One of the applications where BoW representations have been widely used is place recognition. The next experiment shows the performance using a vocabulary created with our proposal and compares it with the performance using standard *k*-means vocabulary.

The trajectories of the two sequences included in our data-set are shown in **Figure 11**. The ground truth of the sequences are aligned so the positions in both trajectories can be compared.

We obtained this aligned ground truth using the g2o optimization tool [52].

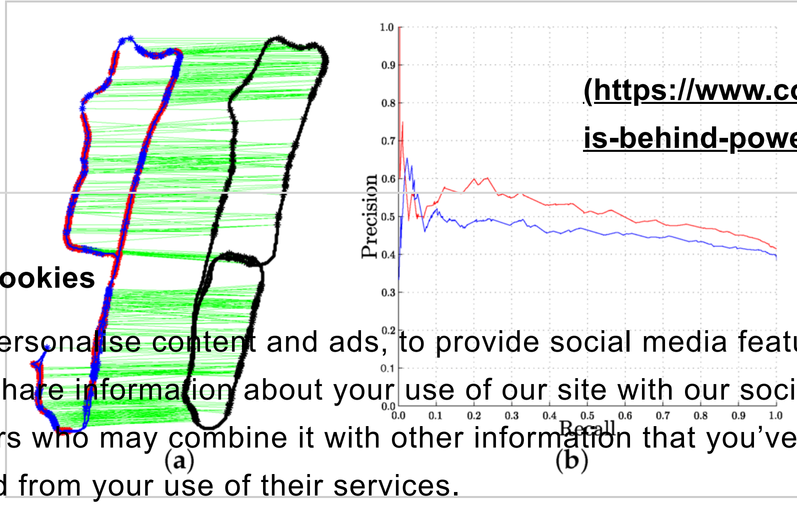


Figure 11. Our vocabulary *vs.* standard *k*-means vocabulary for Place Recognition. (**a**) odometry of the two sequences used: Test (**blue**) and Train (**black**). Correct localization results are shown as green lines, and errors are shown with red points in the test trajectory; (**b**) precision-recall curves using our enhanced vocabulary (**red**) and *k*-means vocabulary (**blue**). These curves have been obtained varying $th_{HistDist}$. (Best seen in color).

According to BoW representations, each image is represented by a histogram of the number of occurrences of each word in the image. Each histogram is normalized with the total number of words in the corresponding image, and the image similarity is obtained according to these histograms distance. The localization of a test image is considered to be the same as the localization of the most similar training image found. If the most similar training image found is within a similarity distance larger than a threshold ($th_{HistDist} = 0.2$ in our case), we consider that the location is unknown or uncertain, therefore no answer is given. The localization will be considered correct if the match and the test odometry positions are within 1.5 m. **Figure 11** shows the results of the experiment, including the visualization of the test and train trajectories and the precision-recall curve. We can observe how our approach gives better results than those obtained using a standard *k*-means vocabulary.

Finally, we show some examples of correct place localization when the robot orientation is very different between train and test in **Figure 12**. In these cases, our model was robust enough to classify those images as being the same location.

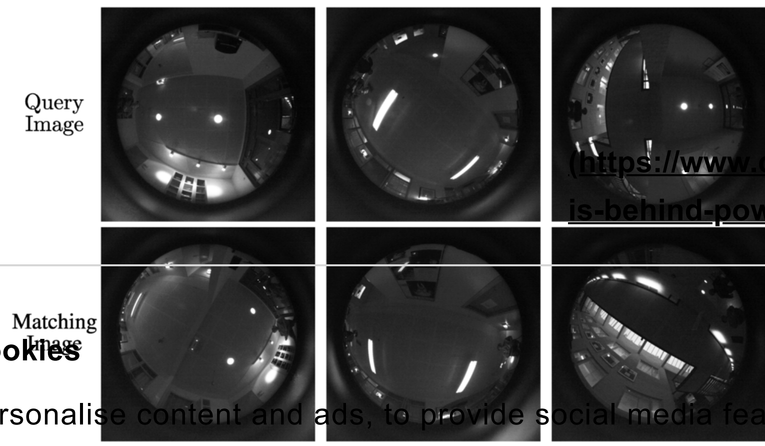


Figure 12. Examples of correct place recognition where the robot location is rotated significantly between test and training images. The left image of each row shows the query test image, and the right image show the correctly matched image from the training data.

5.2. Object Detection

Section 4.5 shows how the vocabulary represents some of the objects found in the environment. **Figure 3** shows the detection of objects in different frames. For this experiment, we have used the representative classes of each object. Key-points detected as being of a representative class of an object are labeled as being generated by that object. The position of the markers, green circles for correct detections and red crosses for incorrect ones, correspond to the position where the key-points where detected in the image. In those examples, we can see how most of the Round Ceiling Lamps are correctly detected. In the first image of the bottom row, the red arrow shows points classified as Round Ceiling Lamp that correspond to a different kind of lamp that newly appears in this room. We can also see how the areas labelled as windows are very dissimilar due to the objects seen through these windows and their different shapes.

6. Conclusions

This work presents a new method to create an enhanced visual vocabulary from an image sequence that learns semantic relationships between visual words of the working environment. The key elements of the method are the use of tracked scene points and the inclusion of information of the key-points altitude when building the visual words. Our approach is focused on long term indoor robotics monitoring applications. With this purpose, we consider systems where the camera points to the ceiling, which facilitates the acquisition of more stable and repetitive scene elements. The experimental validation, with indoor sequences acquired from a mobile robotic platform, shows the performance and the enhanced semantic properties for different method parameters. Comparisons with the standard *k*-means vocabulary present our method as a richer alternative for the usual BoW approach to build a visual vocabulary. Our method provides more representative semantic information of the environment, including relationships between visual words and environment objects or parts. At the same time, our method has shown better results in a place recognition application than the usual BoW approach, and promising results for object discovery and semantic representation for long term place recognition in a robot operating environment.

Acknowledgments

This work was supported by Spanish Government and European Union under project DPI2015-65962-R.

Author Contributions

Alejandro Rituerto implemented the algorithm, run the experiments, contributed on the data acquisition and wrote the present document. Henrik Andreasson developed the acquisition code and also contributed on the data acquisition. All the authors participated in analysis of the results and the proofreading process.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Nice, France, 13–16 October 2003; pp. 1470–1477.
2. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.

3. Jégou, H.; Douze, M.; Schmid, C. Improving bag-of-features for large scale image search. *Int. J. Comput. Vis.* **2010**, *87*, 316–336. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Improving+bag-of-features+for+large+scale+image+search&author=J%C3%A9gou,+H.&author=Douze,+M.&author=Schmid,+C.&publication_year=2010&journal=Int.+J.+Comput.+Vis.&volume=87&pages=316%E2%80%93336&doi=10.1007/s11263-009-0285-2)**] [**CrossRef (https://dx.doi.org/10.1007/s11263-009-0285-2)**] [**Green Version (https://core.ac.uk/download/pdf/48239857.pdf)**]

4. Tuytelaars, T.; Lampert, C.H.; Blaschko, M.B.; Buntine, W. Unsupervised object discovery: A comparison. *Int. J. Comput. Vis.* **2010**, *88*, 284–302. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Unsupervised+object+discovery:+A+comparison&author=Tuytelaars,+T.&author=Lampert,+C.H.&author=Blaschko,+M.B.&author=Buntine,+W.&publication_year=2010&journal=Int.+J.+Comput.+Vis.&volume=88&pages=284%E2%80%93302&doi=10.1007/s11263-009-0271-8)**] [**CrossRef (https://dx.doi.org/10.1007/s11263-009-0271-8)**]

5. Cummins, M.; Newman, P. Appearance-only SLAM at large scale with FAB-MAP 2.0. *Int. J. Robot. Res.* **2011**, *30*, 1100–1123. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Appearance-only+SLAM+at+large+scale+with+FAB-MAP+2.0&author=Cummins,+M.&author=Newman,+P.&publication_year=2011&journal=Int.+J.+Robot.+Res.&volume=30&pages=1100%E2%80%931123&doi=10.1177/0278364910385483)**] [**CrossRef (https://dx.doi.org/10.1177/0278364910385483)**]

6. Kostavelis, I.; Gasteratos, A. Learning spatially semantic representations for cognitive robot navigation. *Robot. Auton. Syst.* **2013**, *61*, 1460–1475. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Learning+spatially+semantic+representations+for+cognitive+robot+navigation&author=Kostavelis,+I.&author=Gasteratos,+A.&publication_year=2013&journal=Robot.+Auton.+Syst.&volume=61&pages=1460%E2%80%931475&doi=10.1016/j.robot.2013.07.008)**] [**CrossRef (https://dx.doi.org/10.1016/j.robot.2013.07.008)**]

7. Everett, H.R.; Gage, D.W.; Gilbreath, G.A.; Laird, R.T.; Smurlo, R.P. Real-world issues in warehouse navigation. *Photonics Ind. Appl. Int. Soc. Opt. Photonics* **1995**, *2352*, 249–259. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Real-world+issues+in+warehouse+navigation&author=Everett,+H.R.&author=Gage,+D.W.&author=Gilbreath,+G.A.&author=Laird,+R.T.&author=Smurlo,+R.P.&publication_year=1995&journal=Photonics+Ind.+Appl.+Int.+Soc.+Opt.+Photonics&volume=2352&pages=249%E2%80%93259)**]

8. Burgard, W.; Cremers, A.B.; Fox, D.; Hähnel, D.; Lakemeyer, G.; Schulz, D.; Steiner, W.; Thrun, S. Experiences with an interactive museum tour-guide robot. *Artif. Intell.* **1999**, *114*, 3–55. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Experiences+with+an+interactive+museum+tour-guide+robot&author=Burgard,+W.&author=Cremers,+A.B.&author=Fox,+D.&author=H%C3%A4hnel,+D.&author=Lakemeyer,+G.&author=Schulz,+D.&author=Steiner,+W.&author=Thrun,+S.&publication_year=1999&journal=Artif.+Intell.&volume=114&pages=3%E2%80%9355&doi=10.1016/S0004-3702(99)00070-3)**] [**CrossRef (https://dx.doi.org/10.1016/S0004-3702(99)00070-3)**]

9. Fukuda, T.; Yokoyama, Y.; Arai, F.; Shimojima, K.; Ito, S.; Abe, Y.; Tanaka, K.; Tanaka, Y. Navigation system based on ceiling landmark recognition for autonomous mobile robot-position/orientation control by landmark recognition with plus and minus primitives. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Minneapolis, MN, USA, 22–28 April 1996; pp. 1720–1725.

10. Wulf, O.; Lecking, D.; Wagner, B. Robust self-localization in industrial environments based on 3D ceiling structures. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, China, 9–15 October 2006; pp. 1530–1534.

11. Konolige, K.; Bowman, J. Towards lifelong visual maps. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), St. Louis, MO, USA, 9–15 October 2009; pp. 1156–1163.

12. Valgren, C.; Lilienthal, A.J. SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments. *Robot. Auton. Syst.* **2010**, *58*, 149–156. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=SIFT,+SURF+&+seasons:+Appearance-based+long-term+localization+in+outdoor+environments&author=Valgren,+C.&author=Lilienthal,+A.J.&publication_year=2010&journal=Robot.+Auton.+Syst.&volume=58&pages=149%E2%80%93156)**]

13. Bredeche, N.; Chevaleyre, Y.; Zucker, J.D.; Drogoul, A.; Sabah, G. A meta-learning approach to ground symbols from visual percepts. *Robot. Auton. Syst.* **2003**, *43*, 149–162. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=A+meta-learning+approach+to+ground+symbols+from+visual+percepts&author=Bredeche,+N.&author=Chevaleyre,+Y.&author=Zucker,+J.D.&author=Drogoul,+A.&author=Sabah,+G.&publication_year=2003&journal=Robot.+Auton.+Syst.&volume=43&pages=149%E2%80%93162&doi=10.1016/S0921-8890(02)00356-1)**] [**CrossRef (https://dx.doi.org/10.1016/S0921-8890(02)00356-1)**]

14. Duygulu, P.; Bastan, M. Multimedia translation for linking visual data to semantics in videos. *Mach. Vis. Appl.* **2011**, *22*, 99–115. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Multimedia+translation+for+linking+visual+data+to+semantics+in+videos&author=Duygulu,+P.&author=Ba%C5%9Ftan,+M.&publication_year=2011&journal=Mach.+Vis.+Appl.&volume=22&pages=99%E2%80%93115&doi=10.1007/s00138-009-0217-8)**] [**CrossRef (https://dx.doi.org/10.1007/s00138-009-0217-8)**] [**Green Version (https://core.ac.uk/download/pdf/52922645.pdf)**]

15. Rituerto, A.; Murillo, A.; Guerrero, J. Semantic labeling for indoor topological mapping using a wearable catadioptric system. *Robot. Auton. Syst.* **2014**, *62*, 685–695. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Semantic+labeling+for+indoor+topological+mapping+using+a+wearable+catadioptric+system&author=Rituerto,+A.&author=Murillo,+A.&author=Guerrero,+J.&publication_year=2014&journal=Robot.+Auton.+Syst.&volume=62&pages=685%E2%80%93695&doi=10.1016/j.robot.2012.10.002)**] [**CrossRef (https://dx.doi.org/10.1016/j.robot.2012.10.002)**]

16. Astua, C.; Barber, R.; Crespo, J.; Jardon, A. Object Detection Techniques Applied on Mobile Robot Semantic Navigation. *Sensors* **2014**, *14*, 6734–6757. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Object+Detection+Techniques+Applied+on+Mobile+Robot+Semantic+Navigation&author=Astua,+C.&author=Barber,+R.&author=Crespo,+J.&author=Jardon,+A.&publication_year=2014&journal=Sensors&volume=14&pages=6734%E2%80%936757&doi=10.3390/s140406734&pmid=24732101)**] [**CrossRef (https://dx.doi.org/10.3390/s140406734)**] [**PubMed (http://www.ncbi.nlm.nih.gov/pubmed/24732101)**]

17. Arandjelović, R.; Zisserman, A. Visual Vocabulary with a Semantic Twist. In *Asian Conference on Computer Vision (ACCV)*; Springer: Berlin, Germany, 2015; pp. 178–195. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Visual+Vocabulary+with+a+Semantic+Twist&author=Arandjelovi%C4%87,+R.&author=Zisserman,+A.&publication_year=2015&pages=178%E2%80%93195)**]

18. Xu, D.; Han, L.; Tan, M.; Li, Y.F. Ceiling-based visual positioning for an indoor mobile robot with monocular vision. *IEEE Trans. Ind. Electron.* **2009**, *56*, 1617–1628. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Ceiling-based+visual+positioning+for+an+indoor+mobile+robot+with+monocular+vision&author=Xu,+D.&author=Han,+L.&author=Tan,+M.&author=Li,+Y.F.&publication_year=2009&journal=IEEE+Trans.+Ind.+Electron.&volume=56&pages=1617%E2%80%931628)**]

19. Hwang, S.Y.; Song, J.B. Monocular vision-based SLAM in indoor environment using corner, lamp, and door features from upward-looking camera. *IEEE Trans. Ind. Electron.* **2011**, *58*, 4804–4812. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Monocular+vision-based+SLAM+in+indoor+environment+using+corner,+lamp,+and+door+features+from+upward-looking+camera&author=Hwang,+S.Y.&author=Song,+J.B.&publication_year=2011&journal=IEEE+Trans.+Ind.+Electron.&volume=58&pages=4804%E2%80%934812&doi=10.1109/TIE.2011.2109333)**] [**CrossRef (https://dx.doi.org/10.1109/TIE.2011.2109333)**]

20. Vieira, M.; Faria, D.R.; Nunes, U. Real-Time Application for Monitoring Human Daily Activity and Risk Situations in Robot-Assisted Living. In *Robot 2015: Second Iberian Robotics Conference*; Springer International Publishing: Lisbon, Portugal, 2016; pp. 449–461. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Real-Time+Application+for+Monitoring+Human+Daily+Activity+and+Risk+Situations+in+Robot-Assisted+Living&author=Vieira,+M.&author=Faria,+D.R.&author=Nunes,+U.&publication_year=2016&pages=449%E2%80%93461)**]

21. Kani, S.; Miura, J. Mobile monitoring of physical states of indoor environments for personal support. In Proceedings of the IEEE/SICE International Symposium on System Integration (SII), Nagoya, Japan, 11–13 December 2015; pp. 393–398.

22. Mantha, B.R.; Feng, C.; Menassa, C.C.; Kamat, V.R. Real-time building energy and comfort parameter data collection using mobile indoor robots. In Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC), Oulu, Finland, 15–18 June 2015; Volume 32, p. 1.

23. Garcia-Fidalgo, E.; Ortiz, A. Vision-based topological mapping and localization methods: A survey. *Robot. Auton. Syst.* **2015**, *64*, 1–20. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Vision-based+topological+mapping+and+localization+methods:+A+survey&author=Garcia-Fidalgo,+E.&author=Ortiz,+A.&publication_year=2015&journal=Robot.+Auton.+Syst.&volume=64&pages=1%E2%80%9320&doi=10.1016/j.robot.2014.11.009)**] [**CrossRef (https://dx.doi.org/10.1016/j.robot.2014.11.009)**]

24. Myers, G.K.; Nallapati, R.; van Hout, J.; Pancoast, S.; Nevatia, R.; Sun, C.; Habibian, A.; Koelma, D.C.; van de Sande, K.E.; Smeulders, A.W.; *et al.* Evaluating multimedia features and fusion for example-based event detection. *Mach. Vis. Appl.* **2014**, *25*, 17–32. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Evaluating+multimedia+features+and+fusion+for+example-based+event+detection&author=Myers,+G.K.&author=Nallapati,+R.&author=van+Hout,+J.&author=Pancoast,+S.&author=Nevatia,+R.&author=Sun,+C.&author=Habibian,+A.&author=Koelma,+D.C.&author=van+de+Sande,+K.E.&author=Smeulders,+A.W.&publication_year=2014&journal=Mach.+Vis.+Appl.&volume=25&pages=17%E2%80%9332&doi=10.1007/s00138-013-0527-8)**] [**CrossRef (https://dx.doi.org/10.1007/s00138-013-0527-8)**]

25. Irschara, A.; Zach, C.; Frahm, J.M.; Bischof, H. From structure-from-motion point clouds to fast location recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 2599–2606.

26. Mikulik, A.; Perdoch, M.; Chum, O.; Matas, J. Learning vocabularies over a fine quantization. *Int. J. Comput. Vis.* **2013**, *103*, 163–175. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Learning+vocabularies+over+a+fine+quantization&author=Mikulik,+A.&author=Perdoch,+M.&author=Chum,+O.&author=Matas,+J.&publication_year=2013&journal=Int.+J.+Comput.+Vis.&volume=103&pages=163%E2%80%93175&doi=10.1007/s11263-012-0600-1)**] [**CrossRef (https://dx.doi.org/10.1007/s11263-012-0600-1)**]

27. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Lost in quantization: Improving particular object retrieval in large scale image databases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
28. Boiman, O.; Shechtman, E.; Irani, M. In defense of nearest-neighbor based image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
29. Yang, L.; Jin, R.; Sukthankar, R.; Jurie, F. Unifying discriminative visual codebook generation with classifier training for object category recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
30. Wang, J.J.Y.; Bensmail, H.; Gao, X. Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification. *Pattern Recognit.* **2013**, *46*, 3249–3255. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Joint+learning+and+weighting+of+visual+vocabulary+for+bag-of-feature+based+tissue+classification&author=Wang,+J.J.Y.&author=Bensmail,+H.&author=Gao,+X.&publication_year=2013&journal=Pattern+Recognit.&volume=46&pages=3249%E2%80%933255&doi=10.1016/j.patcog.2013.05.001)**] [**CrossRef (https://dx.doi.org/10.1016/j.patcog.2013.05.001)**]
31. Bai, S.; Matsumoto, T.; Takeuchi, Y.; Kudo, H.; Ohnishi, N. Informative patches sampling for image classification by utilizing bottom-up and top-down information. *Mach. Vis. Appl.* **2013**, *24*, 959–970. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Informative+patches+sampling+for+image+classification+by+utilizing+bottom-up+and+top-down+information&author=Bai,+S.&author=Matsumoto,+T.&author=Takeuchi,+Y.&author=Kudo,+H.&author=Ohnishi,+N.&publication_year=2013&journal=Mach.+Vis.+Appl.&volume=24&pages=959%E2%80%93970&doi=10.1007/s00138-012-0473-x)**] [**CrossRef (https://dx.doi.org/10.1007/s00138-012-0473-x)**]

32. Fernando, B.; Fromont, E.; Muselet, D.; Sebban, M. Supervised learning of Gaussian mixture models for visual vocabulary generation. *Pattern Recognit.* **2012**, *45*, 897–907. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Supervised+learning+of+Gaussian+mixture+models+for+visual+vocabulary+generation&author=Fernando,+B.&author=Fromont,+E.&author=Muselet,+D.&author=Sebban,+M.&publication_year=2012&journal=Pattern+Recognit.&volume=45&pages=897%E2%80%93907&doi=10.1016/j.patcog.2011.07.021)**] [**CrossRef (https://dx.doi.org/10.1016/j.patcog.2011.07.021)**] [**Green Version (https://core.ac.uk/download/pdf/52652004.pdf)**]

33. Cao, Y.; Wang, C.; Li, Z.; Zhang, L.; Zhang, L. Spatial-bag-of-features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3352–3359.

34. Ji, R.; Yao, H.; Sun, X.; Zhong, B.; Gao, W. Towards semantic embedding in visual vocabulary. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 918–925.

35. Chum, O.; Perdoch, M.; Matas, J. Geometric min-hashing: Finding a (thick) needle in a haystack. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 17–24.

36. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1465–1472.

37. Bolovinou, A.; Pratikakis, I.; Perantonis, S. Bag of spatio-visual words for context inference in scene classification. *Pattern Recognit.* **2013**, *46*, 1039–1053. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Bag+of+spatio-visual+words+for+context+inference+in+scene+classification&author=Bolovinou,+A.&author=Pratikakis,+I.&author=Perantonis,+S.&publication_year=2013&journal=Pattern+Recognit.&volume=46&pages=1039%E2%80%931053&doi=10.1016/j.patcog.2012.07.024)**] [**CrossRef (https://dx.doi.org/10.1016/j.patcog.2012.07.024)**]

38. Penatti, O.; Silva, F.B.; Valle, E.; Gouet-Brunet, V.; Torres, R.D.S. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognit.* **2014**, *47*, 705–720. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Visual+word+spatial+arrangement+for+image+retrieval+and+classification&author=Penatti,+O.&author=Silva,+F.B.&author=Valle,+E.&author=Gouet-Brunet,+V.&author=Torres,+R.D.S.&publication_year=2014&journal=Pattern+Recognit.&volume=47&pages=705%E2%80%93720&doi=10.1016/j.patcog.2013.08.012)**] [**CrossRef (https://dx.doi.org/10.1016/j.patcog.2013.08.012)**]

39. Jegou, H.; Douze, M.; Schmid, C. Hamming embedding and weak geometric consistency for large scale image search. In Proceedings of the European Conference on Computer Vision (ECCV), Marseille, France, 12–18 October 2008; pp. 304–317.

40. Doersch, C.; Singh, S.; Gupta, A.; Sivic, J.; Efros, A.A. What makes Paris look like Paris? *ACM Trans. Graph.* **2012**, *31*, 101. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=What+makes+Paris+look+like+Paris?&author=Doersch,+C.&author=Singh,+S.&author=Gupta,+A.&author=Sivic,+J.&author=Efros,+A.A.&publication_year=2012&journal=ACM+Trans.+Graph.&volume=31&pages=101&doi=10.1145/2185520.2185597)**] [**CrossRef (https://dx.doi.org/10.1145/2185520.2185597)**]

41. Berg, T.L.; Berg, A.C. Finding iconic images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Miami, FL, USA, 20–25 June 2009; pp. 1–8.

42. Fergus, R.; Perona, P.; Zisserman, A. Object class recognition by unsupervised scale-invariant learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Madison, WI, USA, 18–20 June 2003; Volume 2, pp. II-264–II-271.

43. Russell, B.C.; Freeman, W.T.; Efros, A.A.; Sivic, J.; Zisserman, A. Using multiple segmentations to discover objects and their extent in image collections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; pp. 1605–1614.

44. Singh, S.; Gupta, A.; Efros, A.A. Unsupervised discovery of mid-level discriminative patches. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012; pp. 73–86.

45. Liu, J.; Yang, Y.; Shah, M. Learning semantic visual vocabularies using diffusion distance. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 461–468.

46. Bay, H.; Ess, A.; Tuytelaars, T.; van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Speeded-Up+Robust+Features+(SURF)&author=Bay,+H.&author=Ess,+A.&author=Tuytelaars,+T.&author=van+Gool,+L.&publication_year=2008&journal=Comput.+Vis.+Image+Underst.&volume=110&pages=346%E2%80%93359&doi=10.1016/j.cviu.2007.09.014)**] [**CrossRef (https://dx.doi.org/10.1016/j.cviu.2007.09.014)**]

47. Ward, J.; Joe, H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=Hierarchical+grouping+to+optimize+an+objective+function&author=Ward,+J.&author=Joe,+H.&publication_year=1963&journal=J.+Am.+Stat.+Assoc.&volume=58&pages=236%E2%80%93244&doi=10.1080/01621459.1963.10500845)**] [**CrossRef (https://dx.doi.org/10.1080/01621459.1963.10500845)**]

48. Sokal, R.; Michener, C. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **1958**, *6*, 1409–1438. [**Google Scholar (https://scholar.google.com/scholar_lookup?title=A+statistical+method+for+evaluating+systematic+relationships&author=Sokal,+R.&author=Michener,+C.&publication_year=1958&journal=Univ.+Kansas+Sci.+Bull.&volume=6&pages=1409%E2%80%931438)**]

49. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, USA, 2–4 August 1996; pp. 226–231.

50. Achtert, E.; Kriegel, H.P.; Schubert, E.; Zimek, A. Interactive data mining with 3D-parallel-coordinate-trees. In Proceedings of the ACM Conference on Special Interest Group on Management of Data (SIGMOD), New York, NY, USA, 22–27 June 2013; pp. 1009–1012.

51. Rituerto, A.; Andreasson, H.; Murillo, A.C.; Lilienthal, A.; Guerrero, J.J. Hierarchical Vocabulary—Evaluation Data. Available online: **http://aass.oru.se/Research/Learning/datasets.html (http://aass.oru.se/Research/Learning/datasets.html)** (accessed on 4 April 2016).

52. Kuemmerle, R.; Grisetti, G.; Strasdat, H.; Konolige, K.; Burgard, W. g2o: A General Framework for Graph Optimization. Available online: **http://openslam.org/g2o.html (http://openslam.org/g2o.html)** (accessed on 4 April 2016).

Share and Cite

(mailto:?&subject=From%20MDPI%3A%20%22Building%20an%20Enhanced%20Vocabulary%20of%20the%20Robot%20Environment%20with%20a%20Ceiling%20Pointing%20Camera"&body=https://www.mdpi.com/135510%3A%0A%0ABuilding%20an%20Enhanced%20Vocabulary%20of%20the%20Robot%20Environment%20with%20a%20Ceiling%20Pointing%20Camera%0A%0AAbstract%3A%20Mobile%20robots%20are%20of%20great%20help%20for%20automatic%20monitoring%20tasks%20in%20different%20environments.%20One%20of%20the%20first%20tasks%20that%20needs%20to%20be%20addressed%20when%20creating%20these%20kinds%20of%20robotic%20systems%20is%20modeling%20the%20robot%20environment.%20This%20work%20proposes%20a%20pipeline%20to%20build%20an%20enhanced%20visual%20model%20of%20a%20robot%20environment%20indoors.%20Vision%20based%20recognition%20approaches%20frequently%20use%20quantized%20feature%20spaces%2C%20commonly%20known%20as%20Bag%20of%20Words%20%28BoW%29%20or%20vocabulary%20representations.%20A%20drawback%20using%20standard%20BoW%20approaches%20is%20that%20semantic%20information%20is%20not%20considered%20as%20a%20criteria%20to%20create%20the%20visual%20words.%20To%20solve%20this%20challenging%20task%2C%20this%20paper%20studies%20how%20to%20leverage%20the%20standard%20vocabulary%20construction%20process%20to%20obtain%20a%20more%20meaningful%20visual%20vocabulary%20of%20the%20robot%20work%20environment%20using%20image%20sequences.%20We%20take%20advantage%20of%20spatio-temporal%20constraints%20and%20prior%20knowledge%20about%20the%20position%20of%20the%20camera.%20The%20key%20contribution%20of%20our%20work%20is%20the%20definition%20of%20a%20new%20pipeline%20to%20create%20a%20model%20of%20the%20environment.%20This%20pipeline%20incorporates%20%281%29%20tracking%20information%20to%20the%20process%20of%20vocabulary%20construction%20and%20%282%29%20geometric%20cues%20to%20the%20appearance%20descriptors.%20Motiva

ted%20by%20long%20term%20robotic%20applications%2C%20such%20as%20the[...]) (https://twitter.com/intent/tweet?text=Building+an+Enhanced+Vocabulary+of+the+Robot+Environment+with+a+Ceiling+Pointing+Camera&hashtags=mdpisensors&url=https%3A%2F%2Fwww.mdpi.com%2F135510&via=Sensors_MDPI) (http://www.linkedin.com/shareArticle?mini=true&url=https%3A%2F%2Fwww.mdpi.com%2F135510&title=Building%20an%20Enhanced%20Vocabulary%20of%20the%20Robot%20Environment%20with%20a%20Ceiling%20Pointing%20Camera%26source%3Dhttps%3A%2F%2Fwww.mdpi.com%26summary%3DMobile%20robots%20are%20of%20great%20help%20for%20automatic%20monitoring%20tasks%20in%20different%20environments.%20One%20of%20the%20first%20tasks%20that%20needs%20to%20be%20addressed%20when%20creating%20these%20kinds%20of%20robotic%20systems%20is%20modeling%20the%20robot%20environment.%20This%20work%20proposes%20a%20pipeline%20%5B...%5D) (https://www.facebook.com/sharer.php?u=https://www.mdpi.com/135510) (http://www.reddit.com/submit?url=https://www.mdpi.com/135510) (http://www.mendeley.com/import/?url=https://www.mdpi.com/135510)

MDPI and ACS Style

Rituerto, A.; Andreasson, H.; Murillo, A.C.; Lilienthal, A.; Guerrero, J.J. Building an Enhanced Vocabulary of the Robot Environment with a Ceiling Pointing Camera. *Sensors* **2016**, *16*, 493. https://doi.org/10.3390/s16040493

AMA Style

Rituerto A, Andreasson H, Murillo AC, Lilienthal A, Guerrero JJ. Building an Enhanced Vocabulary of the Robot Environment with a Ceiling Pointing Camera. *Sensors*. 2016; 16(4):493. https://doi.org/10.3390/s16040493

Chicago/Turabian Style

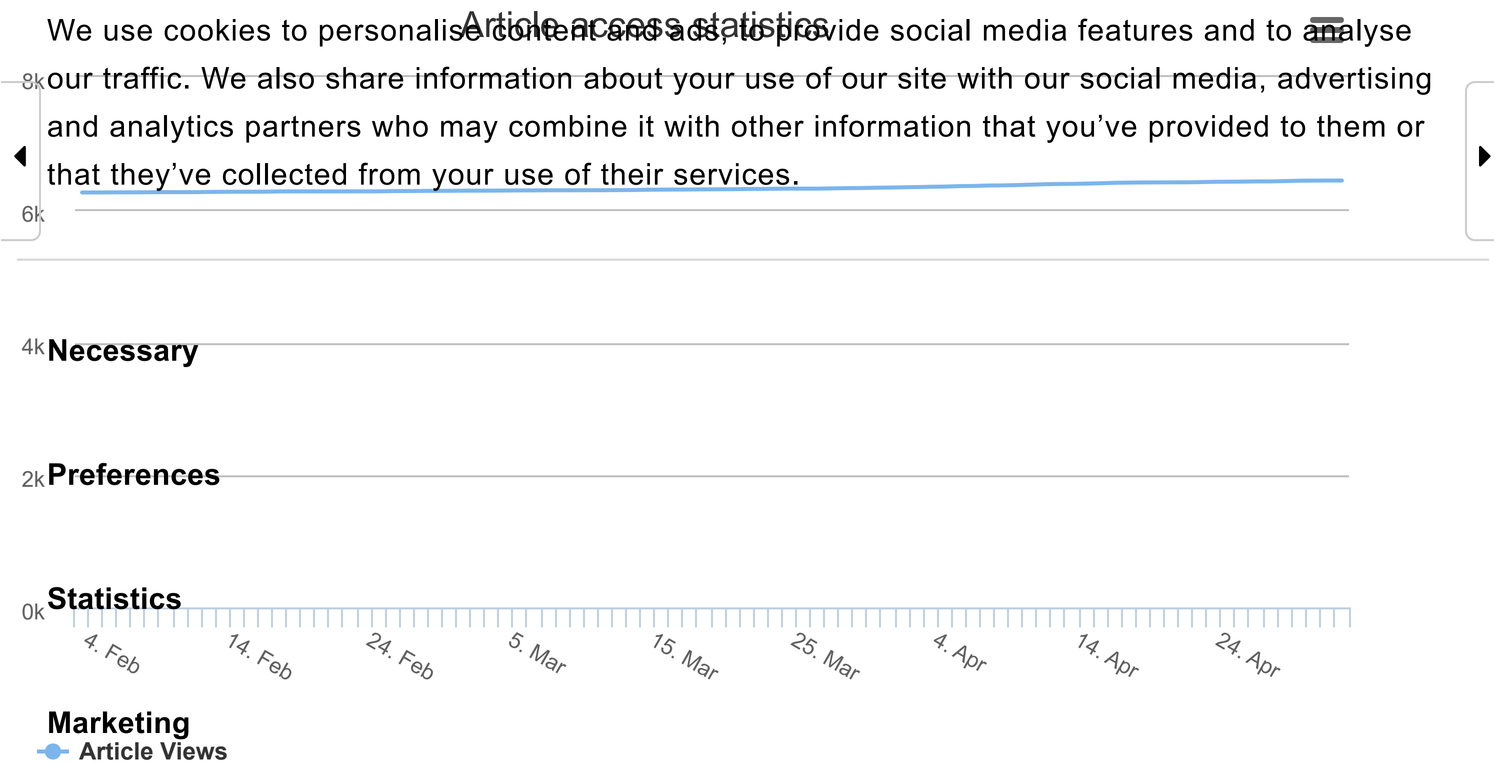
Rituerto, Alejandro, Henrik Andreasson, Ana C. Murillo, Achim Lilienthal, and José Jesús Guerrero. 2016. "Building an Enhanced Vocabulary of the Robot Environment with a Ceiling Pointing Camera" *Sensors* 16, no. 4: 493. https://doi.org/10.3390/s16040493

Note that from the first issue of 2016, this journal uses article numbers instead of page numbers. See further details **here (https://www.mdpi.com/about/announcements/784)**.

Article Metrics

Citations

Article Access Statistics



For more information on the journal statistics, click **here (/journal/sensors/stats)**.

Multiple requests from the same IP address are counted as one view. **Show details**

***Sensors* (/journal/sensors)**, EISSN 1424-8220, Published by MDPI

RSS (/rss/journal/sensors) **Content Alert (/journal/sensors/toc-alert)**

Further Information

Article Processing Charges (/apc)

Pay an Invoice (/about/payment)

Open Access Policy (/openaccess)

Contact MDPI (/about/contact)

Jobs at MDPI (https://careers.mdpi.com)

MDPI (/)

Guidelines

For Authors (/authors)

For Reviewers (/reviewers)

For Editors (/editors)

For Librarians (/librarians)

For Publishers (/publishing_services)

For Societies (/societies)

For Conference Organizers (/conference_organizers)

MDPI Initiatives

Sciforum (https://sciforum.net)

MDPI Books (https://www.mdpi.com/books)

Preprints.org (https://www.preprints.org)

Scilit (https://www.scilit.net)

SciProfiles (https://sciprofiles.com?utm_source=mpdi.com&utm_medium=bottom_menu&utm_campaign=initiative)

Encyclopedia (https://encyclopedia.pub)

JAMS (https://jams.pub)

Proceedings Series (/about/proceedings)

Follow MDPI

LinkedIn (https://www.linkedin.com/company/mdpi)

Facebook (https://www.facebook.com/MDPIOpenAccessPublishing)

Twitter (https://twitter.com/MDPIOpenAccess)

Subscribe to receive issue release notifications and newsletters from MDPI journals

Select options