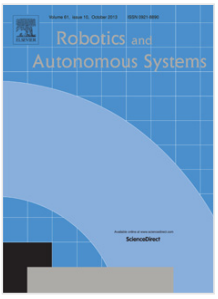
Accepted Manuscript



Semantic localization in the PCL library

Jesus Martínez-Gómez, Vicente Morell, Miguel Cazorla, Ismael García-Varea

PII: S0921-8890(15)00194-3
DOI: http://dx.doi.org/10.1016/j.robot.2015.09.006
Reference: ROBOT 2534

To appear in: *Robotics and Autonomous Systems*

Please cite this article as: J. Martínez-Gómez, V. Morell, M. Cazorla, I. García-Varea, Semantic localization in the PCL library, *Robotics and Autonomous Systems* (2015), http://dx.doi.org/10.1016/j.robot.2015.09.006

HIGHLIGHTS

- **Presentation of a BoW implementation in the Point Cloud Library**
- **Proposal of a general framework for semantic localization systems**
- **The framework allows for integrations of future 3D features and keypoints**
- **The Harris3D detector outperforms uniform sampling with fewer detected keypoints**
- **BoW descriptors obtain better results than the ESF global feature**

Semantic Localization in the PCL library

Jesus Martínez-Gómez[a,b], Vicente Morell[a], Miguel Cazorla[a,*], Ismael García-Varea[b]

[a] *Dpto. of Computer Science and Artificial Intelligence, University of Alicante., P.O. Box 99. 03080, Alicante, Spain.*
[b] *Computer System Department, University of Castilla-La Mancha, Spain.*

Abstract

The semantic localization problem in robotics consists in determining the place where a robot is located by means of semantic categories. The problem is usually addressed as a supervised classification process, where input data correspond to robot perceptions while classes to semantic categories, like kitchen or corridor.

In this paper we propose a framework, implemented in the PCL library, which provides a set of valuable tools to easily develop and evaluate semantic localization systems. The implementation includes the generation of 3D global descriptors following a Bag-of-Words approach. This allows the generation of fixed-dimensionality descriptors from any type of keypoint detector and feature extractor combinations. The framework has been designed, structured and implemented to be easily extended with different keypoint detectors, feature extractors as well as classification models.

The proposed framework has also been used to evaluate the performance of a set of already implemented descriptors, when used as input for a specific semantic localization system. The obtained results are discussed paying special attention to the internal parameters of the BoW descriptor generation process. Moreover, we also review the combination of some keypoint detectors with different 3D descriptor generation techniques.

Keywords: Semantic Localization, PCL, 3D features, classification

*Corresponding author
Email address: miguel.cazorla@ua.es (Miguel Cazorla)

1. Introduction

The semantic localization problem can be defined as the problem of determining the place where a robot is located by means of semantic categories. The problem is usually addressed as a supervised classification process, where input data correspond to robot perceptions, and classes to semantic room/place categories like kitchen, bathroom, or corridor. Commonly, this classification process is tackled by using models that require fixed-dimensionality inputs, such as SVMs [17] or Bayesian Network classifiers [34]. To transform robot perception into fixed-dimensionality descriptors, we can choose for using global features or build them from a set of local features following the well-known Bag-of-Words (BoW) approach [33].

During the last decade, the semantic location problem has attracted the attention of the scientific community, becoming one of the well-known problems in robotics. In fact, several image processing techniques, evaluation datasets, open challenges, and different approaches has been proposed so far, as it is shown in a survey paper [8] recently published. Actually, the semantic information about the place where the robot is located can be very helpful for more specific robotic tasks like autonomous navigation, high-level planning, simultaneous location and mapping (SLAM), or human-robot interaction.

The Point Cloud Library (PCL [21]) has become, in less than four years from its first release, the most widely used open source project for 2D/3D image and point cloud processing. The PCL proposes several algorithms for most of the well-known problems in computer vision: feature extraction, surface reconstruction, image registration, model fitting, and segmentation. Moreover, it implements standard machine learning techniques for clustering and supervised classification. However, PCL does not currently provide a standard procedure for generating 3D global descriptors from local ones. This could be carried out by following a BoW approach, which would allow PCL users to take advantage of all the useful 3D local features included in the library for a wider range of problems. Concretely, any type of 3D local feature could be properly used as input for the semantic localization problem.

In this article, we propose a PCL implementation of the BoW approach relying on machine learning techniques already implemented in the library. Several 3D global descriptors generated with such approach are evaluated when serving as input for the semantic localization problem. Therefore, the purpose of this work is two fold: one one hand, to propose a general framework to easily develop and evaluate semantic localization systems using 3D

point cloud information as input data; and, on the other hand, to implement it in the PCL, taking advantage of the availability of 3D image processing techniques; both with the aim at providing a set of tools to be useful for the PCL community.

Then, the three major contributions of this work are:

- The generation of 3D global descriptors from PCL local features following a Bag-of-Words approach, which will allow the generation of fixed-dimensionality descriptors from any kind of keypoint detector and feature extractor combination.
- The definition of a common framework to develop and evaluate semantic localization systems within PCL. This framework has been designed and implemented to be easily extended with different and new keypoint detectors, feature extractors and classification models.
- The experimentation carried out with a challenging benchmark, which provides sequences of labeled RGB-D images acquired with a mobile robot within indoor office environments. In this experimentation, we evaluate the internal parameters that take part in the BoW approach (e.g. the dictionary size), but we also discuss the role of the keypoint detectors and feature extractors.

The rest of the paper is organized as follow: in Section 2, a more detailed description of the semantic localization problem is presented, as well as a review of some recent proposal to deal with that problem. Section 3 presents the design and development of the proposed framework. In Section 4, the specific contributions of this work to the PCL are described. In Section 5 the experimental results carried out to demonstrate the functionality and usability of this work are presented. Finally, in Section 6 the main conclusions and future works are outlined.

2. Semantic Localization

2.1. Problem definition

As stated before, the semantic localization problem can be formulated as a classical statistical pattern recognition problem as follows. Let I be a perception from a robot (in our case an RGB-D image), $d(I)$ a function that generates a specific descriptor given I, and M a classification model

that provides the class posterior probability $P_M(c|d(I))$, where c is a class label from a set of predefined class categories $\mathcal{C}$. Then, this problem can be stated, without loss of generality, as the problem of finding the optimal label $\hat{c}$ according to:

$$\hat{c} = \arg\max_{c \in \mathcal{C}} P_M(c|d(I))$$

In general, and following that approach, we can identify two main steps to be performed when designing and building a semantic localization system:

1. To carry out a descriptor generation process given the input perception.
2. To design a classifier capable of discriminating among the different types of scenes. This classifier will be trained using the descriptors generated in the previous step.

A more detailed description of these two steps is shown in Section 3.

2.2. Related work

Here we review the most outstanding approaches related to the semantic localization problem. While this problem has been extensively studied [8], we firstly detail the internal stages of this process to then identify those techniques with higher relevance.

The semantic localization problem consists of the process of acquiring an image, generate a suitable representation (that is, an image descriptor) and classify the imaged scene [32]. This classification can be performed according to a) high-level features of the environment, like detected objects [19, 30, 6], b) global image representations [16], or c) local features [28]. A method for scene classification based on global image features was presented in [27], where the temporal continuity between consecutive images was exploited using a Hidden Markov Model. In [15], a scene classifier with range data as input information and AdaBoost as the classification model is proposed. In 2006, Pronobis et al. [18] developed a visual scene classifier using composed receptive field histograms [11] and SVMs.

The use of the Bag of Words (BoW) technique [5] can also be considered a remarkable milestone for visual semantic scene classification. The BoW process starts by creating a visual dictionary of representative features, also known as visual words. Next, each extracted feature is assigned to the closest word in the dictionary. Then, a histogram representing the number of occurrences of each visual word is computed. This histogram is finally used as

the image descriptor. An extensive evaluation of BoW features representations for scene classification was presented in [33], demonstrating that visual words representations are likely to produce superior performance. In [9], an extension of the BoW technique using a spatial pyramid was proposed. Also, this work is one of the most relevant works related to scene classification allowing to merge local and global information into a single image descriptor. The spatial pyramid approach has been successfully applied to several semantic localization problems, and it can be considered a standard solution for generating descriptors.

All mentioned works used visual cameras as input devices. However, visual cameras are highly affected by changing lighting conditions. The lighting variations can occur due to different external weather conditions, but also because of the presence or lack of artificial lights. This reason makes the use of RGB-D cameras very useful in current semantic localization approaches, even to deal with real-time constraints as proposed in [10].

3. Framework Design

In this section, we describe the BoW framework proposed to manage the semantic localization problem, which has been previously defined as a classical supervised classification problem. Therefore, we assume the following initial setup. We are provided with, at least, two sequences of RGB-D images acquired with a mobile robot. The RGB-D images represent scenes from an office indoor environment, such as Universities or Government buildings. Each RGB-D image from the first sequence (training) is labeled with the semantic category of the room where it was acquired, using labels as "kitchen" or "corridor". The problem consists in determining the label for the RGB-D images from the second sequence (test).
The framework proposed includes the following steps:

1. Extract features from training and test RGB-D data. The goal of this step is to find an appropriate image representation, suitable for serving as input in subsequent steps. It involves a set of sub-tasks.
   (a) Select a keypoint detection method, which reduces the amount of points to work with and speeds up the process.
   (b) Select a feature extraction procedure. The combination of keypoints and features should present some specific characteristics: efficiency, repeatability, distinctiveness and accuracy [28, 14].

(c) For each keypoint detected, extract the descriptor associated to the selected feature when possible. We can find some keypoints not meeting the feature requirements, such as a number of surrounding points within a neighborhood. This fact can reduce the final number of features extracted from the RGB-D image.

2. Transform the features extracted into global descriptors with fixed-dimensionality using a BoW approach.
   (a) Merge all the features extracted from the complete training sequence into a single set of features.
   (b) Perform a k-means clustering over this set to select a subset of k representative features. This subset of features is known as the dictionary, and its size k should have been previously defined.
   (c) For each training and test RGB-D image, assign all their (previously extracted) features with the closest word in the dictionary. Then, compute a histogram over these assignations whose dimensionality corresponds to the dictionary size. This histogram is then used as image descriptor.
3. Train a classification model using the training sequence. Based on the training descriptors generated in the previous step (and the room labels), we train a SVM classifier [29]. Thanks to the use of fixed-dimensionality inputs, most of the classifiers capable of managing continuous data could be used.
4. Classify the whole test sequence. The last step classifies each test descriptor with the SVM model computed in the training stage.

Fig. 1 shows the descriptor generation process from a set of extracted features. It can be observed how the final descriptor presents the same dimensionality for all the input images, even when a different number of features were extracted from them.

4. Point Cloud Library Contributions

In this section, we describe the two main contributions to the PCL. The source code of the provided tool is available online under the Creative Commons Attribution license (CC-BY 3.0) at

https://bitbucket.org/vmorell/semanticlocalization

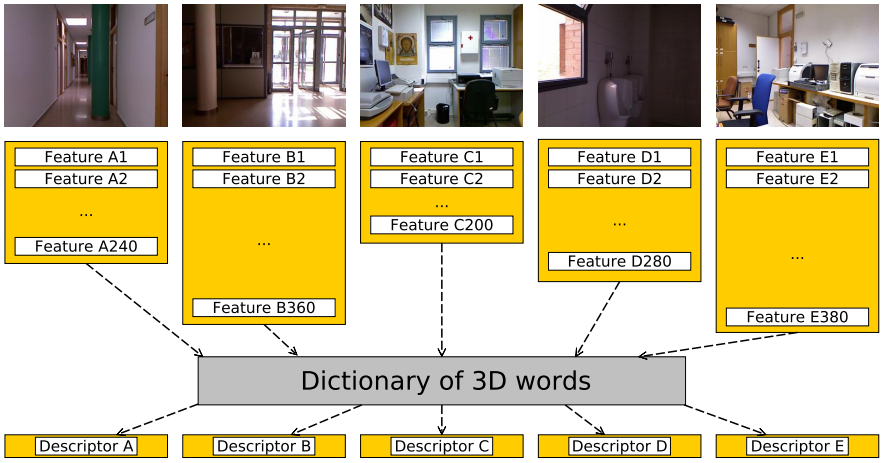


Figure 1: Descriptor generation process from the features extracted and a dictionary of 3D words previously computed.

4.1. 3D global descriptors from local features

Although there are several global descriptors for 3D data, as previously commented in Section 2, the BoW approach could be used for describing the whole point cloud using local features. This global feature, a histogram, could be used for other tasks purposes. Local features could come from any combination of 3D keypoint detectors and features. In the presented framework, it is quite easy to modify the code to include different keypoint detectors and feature methods. We provide in the code some experimentation with several 3D local keypoint detectors and feature descriptors available in the PCL. We briefly describe them.

One of the simplest detector is Uniform Sampling (US). US builds a 3D voxel grid with the input data and takes the centroid (average point inside a voxel) of the voxel grid as keypoint. The resulting point cloud is then reduced and downsampled in a uniform way. Another keypoint detector is Harris3D [23]. The implementation available in PCL takes the normals to the input pointcloud as the input for this detector. For each point, it selects points in a given neighborhood and calculates a covariance matrix of the normals at those points. Then, a value is calculated for each point based on the determinant and trace of the covariance matrix (as proposed in [7] for 2D). After a local maximum suppression method is applied, the surviving

points are the keypoints for the input point cloud.

The Normal Aligned Radial Feature (NARF) [24] keypoint detector and feature descriptor uses the range image, instead of the point cloud, to calculate the descriptor. The keypoint detector finds borders in the range image and calculates a score, indicating how the surface changes on each point. After this score is calculated, a smoothing process and non-maximum suppression are applied. With regard to the feature extraction process, NARF extracts a descriptor from each keypoint and its neighborhood. A star pattern is used, and for each beam of the pattern, it calculates the intensity changes along the cells lying under the beam. Then, for each beam, a value in the range $[-0.5, 0.5]$ is obtained. To make it invariant against rotation, the predominant orientation is calculated. Another feature used in the framework is the Signature of Histograms of OrienTations (SHOT) [25]. The descriptor is calculated by concatenating a set of local histograms over the 3D volume defined by a 3D grid centered at a keypoint. For each local histogram and for each point, the angular difference between the normal in the point and the normal in the keypoint is accumulated in the histogram. A variant is Color-SHOT [26], which adds a color histogram to the original SHOT descriptor.

Another two features used in our experiments are based on the Point Feature Histogram (PFH) [22]. PFH selects from a keypoint, a set of points in a given neighborhood. For each two points in that neighborhood, PFH calculates four values which together express geometric relationship between those points. The four values are concatenated and a histogram is calculated using the values of all the possible combination of points. The first variation of PFH is the Fast Point Feature Histogram [20], which improves the efficiency of the original feature by not processing some points in the neighborhood. The second one is PFH-RGB, which includes color to the geometrical information.

Regarding the BoW implementation, it includes the following stages: features extraction (from selected keypoints), dictionary generation, features/words association (also known as features quantization), and histogram representation computing words frequencies. The first stage relies on a keypoint/feature combination and generates a set of features. A dictionary is generated from this set by using the functionalities provided by the machine learning module included in the PCL. Concretely, we take advantage of the k-means implementation using the set of features as input. The resulting k centroids are then established as dictionary words. The association between features and words, as well as the histogram computation, are also carried

out from the functionalities of the machine learning PCL module.

4.2. Framework for semantic localization

Our main contribution in this paper is the development of a framework that could be used for experimentation in semantic localization. Our main goal building this framework is the suitability for future development, i.e., it must be easy to integrate different keypoint detectors and feature descriptors, as well as to use others classification methods.

For that reason, we have defined a diagram class (see Fig. 2) where several abstract classes and methods are presented. The `SemanticLocalization` class implements some methods: `readConfiguration`, which reads a configuration file containing the point clouds to be used as input to the method; `test` and `validate` are used for testing and validating the method (these methods call the `train` and `classify` abstract methods), and finally `showResults` that shows the results of the classification. So `train` and `classifyFrame` are abstract and must be implemented in inherited classes. This class also has several attributes: `frames` are the input point clouds for the classification, while `detector` and `features` are the keypoint detector method and feature descriptor to be used in the classification, respectively.

We also provide two different classification methods, both making use of the BoW descriptors as input data. The first one is the Support Vector Machine (SVM) [1], which learns to classify elements from two different classes finding a hyperplane which provides less classification error. By other hand, we have used the k-Nearest-Neighbors ($k-$NN) [4] method that directly uses the training data as model. Given a new element to classify, the k nearest neighbors from the training data are selected. The new element is assigned to the class with more elements in the neighborhood. Other supervised classification methods could be incorporated easily.

The `SemanticLocalizationBoW` class inherits from `SemanticLocalization` and uses a BoW approach. To do that, an attribute class `dictionary` contains the dictionary to be used in the classification process. In this class, two methods are implemented: `computeDictionary` that must be called before training, and `wordsAssignation` where the words from the data are calculated. From this class, two other classes are defined, depending on the classification method used: `SemanticLocalizationBoWSVM` that needs to define a `SVMModel`, and `SemanticLocalizationBoWKNN` that does not need to define any additional attribute.

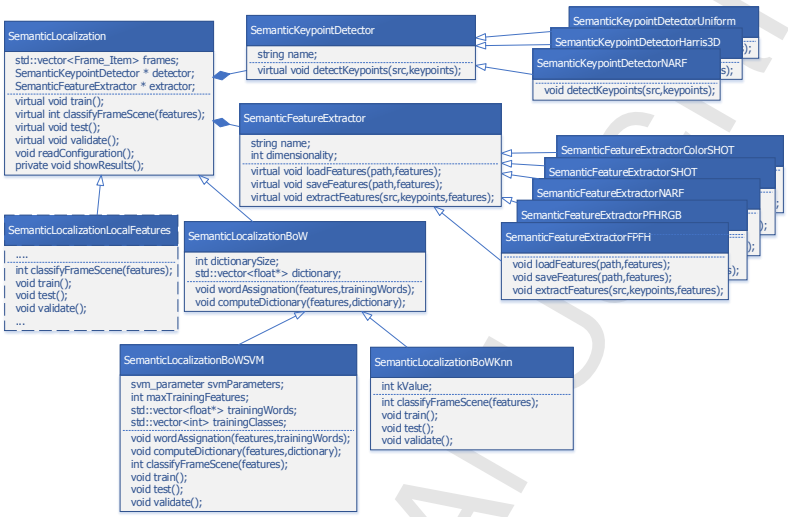


Figure 2: Class diagram of the implemented framework.

Using this scheme, the final user can focus on implementing its method, or using different keypoint detectors and feature descriptors, thus providing an easy way to make experiments in semantic localization.

5. Experimental results

5.1. Dataset description: ViDRILO

All the experimentation included in this article has been carried out using ViDRILO: the Visual and Depth Robot Indoor Localization with Objects information dataset [13]. This dataset, whose overall characteristics are shown in Table 1, provides five different sequences of RGB-D images captured by a mobile robot within an office indoor environment.

Each RGB-D image is annotated with the semantic category of the room it was acquired, from a set of ten room categories. Unreleased sequences from ViDRILO have been successfully used in the RobotVision at ImageCLEF competition [12] in 2013 and 2014 [3, 2]. Fig. 3 shows exemplar images for each one of the ten room categories using the following codes: CR (Corridor), HA (Hall), PO (Professor Office), SO (Student Office), TR (Technical

Table 1: Overall ViDRILO sequences distribution.

Sequence	Number of Frames	Floors imaged	Dark Rooms	Time Span
Sequence 1	2389	1st,2nd	0/18	0 months
Sequence 2	4579	1st,2nd	0/18	0 months
Sequence 3	2248	2nd	4/13	3 months
Sequence 4	4826	1st,2nd	6/18	6 months
Sequence 5	8412	1st,2nd	0/20	12 months

Room), TO (Toilet), SE (Secretary Office), VC (Video Conference Room), WH (Warehouse), and EA (Elevator Area).

Figure 3: Exemplar visual images for all room categories in ViDRILO.

To focus on the internal parameters of the BoW approach, the experimentation stage is limited to the use of Sequence 1 and Sequence 2 from the dataset. The room distribution for these sequences is shown in Fig. 4. We opted for this selection to evaluate the use of the BoW approach in a general semantic localization benchmark. Namely, Sequences 3-5 involve environmental modifications with respect to the initial acquisitions due to the time span (3, 6 and 12 months respectively). These modifications are caused by the human activity: some objects can change their location (e.g. trashes and phones), while others are added or removed from the environment (e.g. tabletop papers and ballpoint pens). The use of Sequences 3-5 increases the challenging of the problem, and it would require the use of non-standard solutions to cope with domain adaptation. The robot used for the acquisition

of Sequences 1-2 followed a similar path but in the opposite direction, which affects the viewpoint of the imaged scenes. Here, we can observe that we are facing a challenging problem due to the dataset is highly unbalanced: most of the RGB-D images belong to the "Corridor" category.

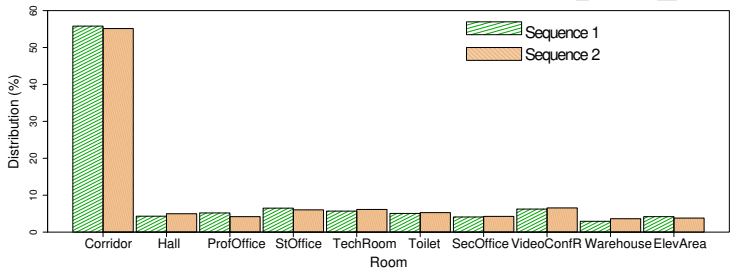


Figure 4: Room distribution for Sequences 1 and 2 in the ViDRILO dataset.

5.2. Study of keypoints detection

Three different keypoints detection methods are evaluated in this work: NARF, Harris3D, and Uniform Sampling, all of them implemented in the PCL. These methods select a subset of 3D points from an input cloud using different methods, but they differ in the average amount of selected points. In the following, we describe the internal parameters used for the experimentation. We only fixed those parameters that should be explicitly established. The rest of parameters were set to their default values. Regarding the NARF detector, we used a support size of 20 cm. This parameter represents the diameter of the sphere used to find neighboring points, and therefore to estimate if a point belongs to a border or not. With respect to the Harris3D detector, we used a threshold of 0.01 as we found it as a reasonable value to remove weak keypoints. Finally, the Uniform Sampling detector internally uses a voxel grid unsupervised downsampling method. We opted for a radius of 0.03 meters, which makes keypoints being selected using tiny 3D boxes whose volume is 3 cubic centimeters.

Fig. 5 graphically presents the keypoint detection with these three techniques. We selected NARF, Harris3D and Uniform Sampling to study the effect of detecting a small, medium and large number of keypoints respectively.

Figure 5: Keypoint detection with NARF (bottom left), Harris3D (bottom center) and Uniform Sampling (botoom right) for a sample RGB-D image (top). The number indicates the amount of keypoints detected with each method

5.3. Semantic Localization results

We test our approach for the generation of semantic localization systems on the ViDRILO dataset. Concretely, we evaluated the generalization capabilities by generating classifiers using Sequence 2 (4579 RGB-D images) for training. These systems are then used to classify the 2389 RGB-D images from Sequence 1. Both sequences were acquired in the same building during two consecutive days.

The following internal parameters are evaluated:

- 3 Keypoint detectors: NARF, Harris3D, and Uniform Sampling.
- 5 Feature extractors: NARF, SHOT, Color-SHOT, PFH-RGB, and FPFH.
- 4 Dictionary sizes: 25, 50, 100, and 200.
- 2 Classification models:

– SVM classifier (exponential chi-square kernel).
– k-Nearest-Neighbor ($k = 7$).

Fig. 6 shows the accuracies obtained with all the semantic classifiers, and we can extract some remarks from these results. Firstly, we can observe that the SVM classifier outperforms the use of k-NN in most of the cases. The two classification models evaluated in this work behave different with respect to the dictionary size. Increasing the size of the dictionary always has a positive impact on the accuracy when using SVM, but not with k-NN. Regarding the keypoint detection method, NARF is the one presenting the worst results, as it could have been expected. At this point, we should outline the bad behavior of the combination of NARF as keypoint detector and feature extraction techniques. The main differences between Harris3D and Uniform Sampling are related to the classification models. That is, the improvement obtained thanks to the use of Uniform Sampling (with respect to Harris3D) is notoriously greater when using k-NN as classification model.

An analysis of the feature extraction methods exposes PFHRGB and Color-SHOT as the most promising techniques. On the contrary, NARF, FPFH and SHOT features present the lower accuracies. It should be taken into account that PFHRGB and Color-SHOT are the only two features that integrate color information. The overall highest accuracy (69.17) was obtained with a SVM and a combination of Harris3D and PFHRGB as keypoint detector and feature extractor respectively. Therefore, we can conclude that the use of Uniform Sampling is not needed unless a k-NN classifier is used. The use of Harris3D as keypoint detection technique notoriously reduces the amount of data to work with and speeds up the 3D processing.

We also evaluated the use of one of the state-of-the-art global 3D feature: the Ensemble of Shape Functions (ESF) [31]. Using the ESF descriptor, we trained both SVM and k-NN classifiers from Sequence 2 and tested against Sequence 1. We obtained an accuracy value of 58.48% with k-NN and 64.49% with the SVM classifier. Consequently, the BoW approach allowed us to outperform the ESF global descriptor. Moreover, we obtained better results using descriptors whose dimensionality is notoriously lower than for the ESF descriptor (200 vs 640). This difference in the descriptor dimensionality would result in classification models that can be trained in a lower amount of time, and perform RGB-D images classification much faster.

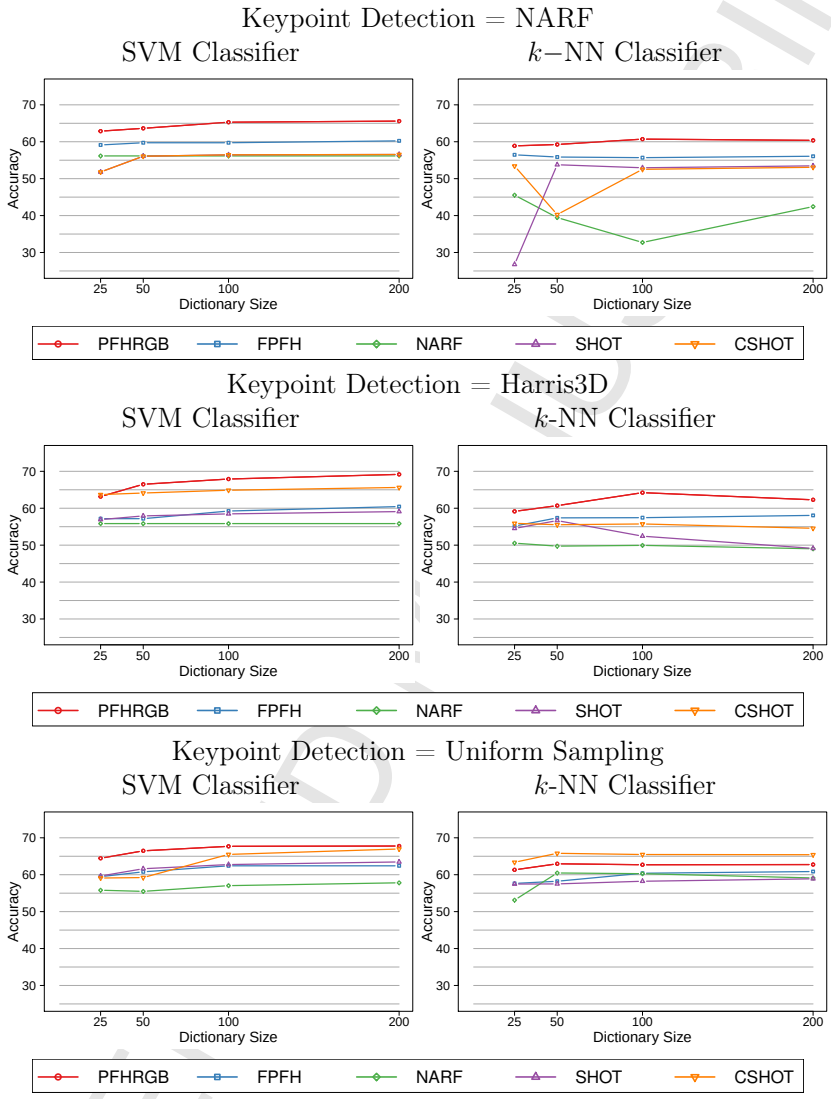


Figure 6: Semantic localization overall results. Accuracy values obtained by training a SVM (left) or k-NN classifier (right) with Sequence 2 and evaluating against Sequence 1 from the ViDRILO dataset [13].

6. Conclusions and future work

Semantic localization is a challenging problem in robotics. We have presented in this article a framework for the generation of global 3D descriptors from local ones following a BoW approach. This framework has been implemented in the Point Cloud Library and evaluated in the semantic localization problem.

Based on the experimentation stage, we can affirm that PFHRGB and Color-SHOT are the two 3D local features with the best performance. Harris3D exposed as the most appropriate keypoint detection method, due to it notoriously reduces the amount of data to work with respect to Uniform Sampling. The proposed BoW framework obtained higher accuracies than the use of the well-known global 3D feature ESF.

As future work, we have in mind the experimentation with a wider variety of 3D features and keypoint detection methods. Moreover, larger dictionary sizes will also be considered.

Acknowledgments

This work was supported by grant DPI2013-40534-R of the Ministerio de Economia y Competitividad of the Spanish Government, supported with Feder funds, and by Consejería de Educación, Cultura y Deportes of the JCCM regional government through project PPII-2014-015-P. Jesus Martínez-Gómez is also funded by the JCCM grant POST2014/8171.

References

[1] Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

[2] Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., Garcia-Varea, I., and Morell, V. (2014). ImageCLEF 2014: Overview and analysis of the results. In *CLEF proceedings*, Lecture Notes in Computer Science. Springer Berlin Heidelberg.

[3] Caputo, B., Muller, H., Thomee, B., Villegas, M., Paredes, R., Zellhofer, D., Goeau, H., Joly, A., Bonnet, P., Martinez-Gomez, J., Garcia-Varea, I., and Cazorla, M. (2013). Imageclef 2013: the vision, the data and the open challenges. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 250–268. Springer.

[4] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

[5] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

[6] Espinace, P., Kollar, T., Roy, N., and Soto, A. (2013). Indoor scene recognition by a mobile robot through adaptive object detection. *Robot. Auton. Syst.*, 61(9):932–947.

[7] H., C. and S., M. (1988). A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151.

[8] Kostavelis, I. and Gasteratos, A. (2015). Semantic mapping for mobile robotics tasks: A survey. *Robotics and Autonomous Systems*, 66(0):86–103.

[9] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA. IEEE Computer Society.

[10] Lim, H. and Sinha, S. (2012). Towards real-time semantic localization. In *ICRA Workshop on Semantic Perception and Mappin*.

[11] Linde, O. and Lindeberg, T. (2004). Object recognition using composed receptive field histograms of higher dimensionality. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, volume 2, pages 1–6. IEEE.

[12] Martínez-Gómez, J., Caputo, B., Cazorla, M., Christensen, H., Fornoni, M., García-Varea, I., and Pronobis, A. (2015). The robot vision challenge. where are we after 5 editions? *IEEE Robotics and Automation Magazine*.

[13] Martinez-Gomez, J., Cazorla, M., Garcia-Varea, I., and Morell, V. (2015). Vidrilo: The visual and depth robot indoor localization with objects information dataset. *International Journal of Robotics Research*.

[14] Martínez-Gómez, J., Fernández-Caballero, A., García-Varea, I., Rodríguez, L., and Romero-González, C. (2014). A taxonomy of vision systems for ground mobile robots. *Int J Adv Robot Syst*, 11:1–11.

[15] Mozos, O. M., Stachniss, C., and Burgard, W. (2005). Supervised learning of places from range data using adaboost. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 1730–1735. IEEE.

[16] Oliva, A. and Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.*, 155:23 – 36.

[17] Orabona, F. and Castellini, C. (2007). Indoor place recognition using online independent support vector machines. In *Proceedings of the British Machine Vision Conference 2007, University of Warwick, UK, September 10-13, 2007*, pages 1–10.

[18] Pronobis, A., Caputo, B., Jensfelt, P., and Christensen, H. (2006). A discriminative approach to robust visual place recognition. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 3829–3836. IEEE.

[19] Ranganathan, A. and Dellaert, F. (2007). Semantic modeling of places using objects. In *Robotics: Science and Systems (RSS)*, Atlanta; USA.

[20] Rusu, R., Blodow, N., and Beetz, M. (2009). Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 3212–3217.

[21] Rusu, R. and Cousins, S. (2011). 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China.

[22] Rusu, R. B., Blodow, N., Marton, Z. C., and Beetz, M. (2008). Aligning point cloud views using persistent feature histograms. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3384–3391. IEEE.

[23] Sipiran, I. and Bustos, B. (2011). Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes. *The Visual Computer*, 27(11):963–976.

[24] Steder, B., Rusu, R. B., Konolige, K., and Burgard, W. (2010). Narf: 3d range image features for object recognition. In *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, volume 44.

[25] Tombari, F., Salti, S., and Di Stefano, L. (2010). Unique signatures of histograms for local surface description. In *Computer Vision–ECCV 2010*, pages 356–369. Springer.

[26] Tombari, F., Salti, S., and Di Stefano, L. (2011). A combined texture-shape descriptor for enhanced 3d feature matching. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 809–812.

[27] Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003*, pages 273–280. IEEE.

[28] Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280.

[29] Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer.

[30] Vasudevan, S. and Siegwart, R. (2008). Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robot. Auton. Syst.*, 56(6):522–537.

[31] Wohlkinger, W. and Vincze, M. (2011). Ensemble of shape functions for 3d object classification. In *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, pages 2987–2992. IEEE.

[32] Wu, J., Christensen, H., and Rehg, J. (2009). Visual place categorization: Problem, dataset, and algorithm. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 4763–4770. IEEE.

[33] Yang, J., Jiang, Y.-G., Hauptmann, A. G., and Ngo, C. (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, MIR '07, pages 197–206, New York, NY, USA. ACM.

[34] Yi, C., Suh, I. H., Lim, G. H., and Choi, B.-U. (2009). Bayesian robot localization using spatial object contexts. In *IROS*, pages 3467–3473. IEEE.

Vicente Morell
Vicente Morell received a BSc , MSc and PhD in Computer Science from the University of Alicante (Spain) in 2008, 2010 and 2014 respectively. He is currently a post-doc student and researcher in the Artificial Intelligence Department at the University of Alicante. His research interests are focused on computer vision, robotics and neural networks.

Miguel Cazorla Miguel Cazorla received a BS degree in Computer Science from the University of Alicante (Spain) in 1995 and a PhD in Computer Science from the same University in 2000. He is currently Associate Professor in the Dept Computer Science and Artificial Intelligence at the University of Alicante. His research interests are focused on computer vision and mobile robotics (mainly using vision to implement robotics tasks).He has published more than 100 papers in JCR journals and international conferences.

Jesus Martinez-Gomez received the MSc. and PhD degrees in Computer Science in 2008 and 2011, respectively, from the University of Castilla-La Mancha, Spain. He is currently member of the Intelligent Systems and Datamining research group in the Albacete Research Institute of Informatics I3A and postdoc researcher at the University of Alicante. His research includes robotics, place classification. artificial intelligence, multimodal human-robot interaction and computer vision. Since 2012, he has been the main organizer of three editions of the Robot Vision at ImageCLEF competition.

Ismael García-Varea received the M.S. in Computer Science and the PhD degree in Pattern Recognition and Artificial Intelligence from the Universitat Politécnica València (UPV), Spain, in 1996 and 2003, respectively. In 1999 he joined the Computing Systems Department of the University of Castilla–La Mancha (UCLM), where he is until now serving as an Assistant Professor. His current research interests include the areas of syntactic and statistical pattern recognition, machine learning, data mining, and robotics. Dr. García-Varea is currently an active member of the Data Mining and Intelligent Systems (SIMD) research group of the UCLM.

***Photo of each author**

Vicente Morell

Miguel Cazorla

Jesus Martínez-Gómez

Ismael García-Varea