Effectiveness of the Bag-of-Words approach on the object search problem in 3D domain

Vladimir Privalov
Brno University of Technology,
Faculty of Information Technology
Department of Computer Graphics
and Multimedia
Brno, Czech republic
iprivalov@fit.vutbr.cz

Vítězslav Beran
Brno University of Technology,
Faculty of Information Technology
Department of Computer Graphics
and Multimedia
Brno, Czech republic
beranv@fit.vutbr.cz

Pavel Smrž
Brno University of Technology,
Faculty of Information Technology
Department of Computer Graphics
and Multimedia
Brno, Czech republic
smrz@fit.vutbr.cz

ABSTRACT

In this work, we investigate the application of the Bag-of-Words approach for object search task in 3D domain. Image retrieval task solutions, operating on datasets of thousands and millions images, have proved the effectiveness of Bag-of-Words approach. The availability of low cost RGB-D cameras is a rise of large datasets of 3D data similar to image corpuses (e.g. RoboEarth). The results of such an investigation could be useful for many robot scenarios like place recognition from a large dataset of samples of places acquired during the long-term observation of an environment. The first goal of our research presented in this paper is focused on the sensitivity of the Bag-of-Words approach to various parameters (e.g. spacial sampling, surface description etc.) with respect to precision, stability and robustness. The experiments are carry out on two widely-used datasets in object instance identification task in 3D domain.

CCS CONCEPTS

• **Computing methodologies → Computer vision**; **Perception**;

KEYWORDS

Bag-of-Words, object search, large-scale datasets

ACM Reference Format:
Vladimir Privalov, Vítězslav Beran, and Pavel Smrž. 2017. Effectiveness of the Bag-of-Words approach on the object search problem in 3D domain. In *SCCG '17: SCCG '17: Spring Conference on Computer Graphics 2017, May 15–17, 2017, Mikulov, Czech Republic.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3154353.3154365

1 INTRODUCTION

Object localization and instance search play an important role in service robotics. Autonomous service robot operating in household or production environment should know which objects present in the environment and how to interact with them in order to perform many autonomous manipulation tasks, e.g. grasping.

The task of object search is considered as the task of finding and identifying objects from a dataset of objects in real world using an image from the camera [Alhamzi et al. 2014]. Robot should identify objects and estimate their position and orientation in the real world. The object search problem is still challenging. Firstly, the appearance of objects variates greatly due to changing illumination conditions. Secondly, occlusions and clutter appear commonly in regular household scenes. Ordinary indoor scene frequently contains many different objects. Thus the robot should accurately identify and localize objects in order to perform tasks like object manipulation.

In recent years, due to rapid developments in vision sensor technologies, advanced low cost consumer RGB-D cameras became widely available. These cameras provide high-resolution dense color and depth images making the perception of the environment complete with more details of the geometric appearance of surfaces. Using the combination of color and depth data has shown its higher efficiency and robustness under clutter and occlusions as well as changing illumination conditions in object and place recognition compared to 2D images [Aldoma et al. 2012b].

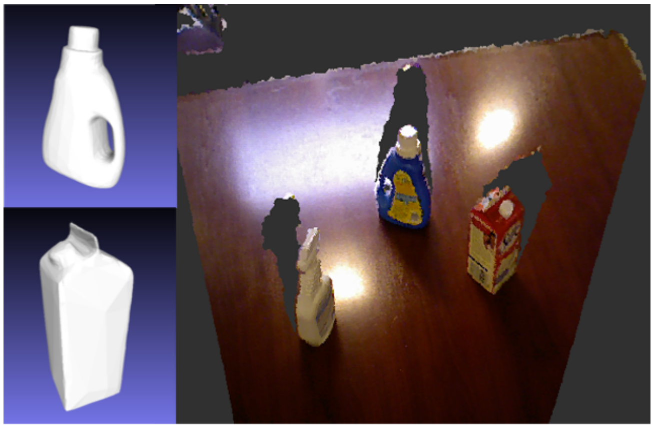

Figure 1: Test scene from Kinect dataset and two models found

The recent advent of low cost RGB-D cameras facilitated the acquisition of datasets of scenes and objects in the robotics research community providing both color and depth data. There are already datasets of real-life objects released in the robotics research

SCCG '17, May 15–17, 2017, Mikulov, Czech Republic

ACM ISBN 978-1-4503-5107-2/17/05…$15.00
https://doi.org/10.1145/3154353.3154365

community, like RoboEarth [Waibel et al. 2011]. We are expecting the ongoing advent of very large databases of 3D objects in order of thousands. New tasks appear which could benefit from large datasets of objects. Thus one of our considerations is the scalability of the Bag-of-Words approach to very large datasets of objects.

Finally, for effective completion of autonomous robotic tasks, the robot should be able to search objects in real time. Existing methods dealing with 3D object search mostly rely on local high-dimensional 3D feature descriptors calculated on a subset of point cloud called keypoints. The use of local features becomes a bottleneck in the total object search pipeline degrading its final computational efficiency. Hence, we are interested in fast performing methods for object search scalable to datasets of thousands objects.

In this paper, we investigate how effectively the Bag-of-Words approach could be applied in 3D domain to the object search task. The Bag-of-Words approach has shown its effectiveness on datasets of thousands and even million images in image retrieval systems. We present experiments evaluating the sensitivity of this approach to different parameters like surface description and keypoint sampling density with respect to performance and robustness. Moreover, we analyze the influence of the vocabulary size and the spatial verification on the resulting search performance as well as the generalization capabilities of the approach.

We believe that the results of our investigation could have benefits for many robot scenarios. For instance, a mobile robot captures image samples of different places by the onboard camera during an observation tour around some office environment. The robot could ultimately collect hundreds or even thousands of image samples in the environment. Given the database of samples of visited places, the robot is able to effectively recognize the place it is located at that moment looking for the most relevant candidate in the database of samples. Ordinary office environments commonly occupy large multi-floor buildings, thus the algorithms should be scalable to very large databases of places.

The paper has following structure. In section 2, we overview the state-of-the-art approaches addressing the object search problem. In section 3, we describe the Bag-of-Words approach and the object search pipeline following this approach in more detail. Then we present the experimental methodology and datasets used for the evaluation of the Bag-of-Words approach in section 4. In section 5, we present the results of the experiments and discuss our observations from these results. Finally, in section 6, we outline the main conclusions and future works.

2 RELATED WORKS

We are focusing on a specific task of object search in large database of objects in 3D domain and particularly on the application of the Bag-of-Words approach to this task. There have been merely several studies dealing with this problem in 3D domain.

The Bag-of-Words approach was originally applied to image retrieval task, where it shown high effectiveness on databases of thousands and millions images. Sivic et al. in [Sivic and Zisserman 2006] adapted the Bag-of-Words technique for shots retrieval from popular movies given an user-specified image as a query. In their work, SIFT descriptors calculated for local affine invariant regions are quantized into visual words by mean of k-means clustering. The authors used "term frequency - inversed document frequency" (TF-IDF) weighting of visual words scoring the relevance of an image to the query. They compared their Bag-of-Words based approach with a baseline method implementing standard frame to frame matching and their approach outperformed the baseline method in precision result. Nister et al. in [Nister and Stewenius 2006] proposed an object recognition approach with an indexing scheme based on Bag-of-Words approach. In this work, the authors used hierarchical k-means for the quantization of descriptors into visual words and showed this method as more optimal than the standard k-means in terms of computational complexity. Nister et al. showed that vocabulary tree greatly speeds up the search procedure in comparison with traditional k-means and allows for large vocabularies. Philbin et al. in [Philbin et al. 2007] proposed approximate k-means quantization method for the task of image retrieval and compared its performance with the hierarchical k-means method. The experimental results on different vocabulary sizes (10K, 20K, 50K and 1M) shown the peak in performance at 1M visual words, although for larger vocabulary sizes the performance curve appears quite flat. Also they shown that the search quality could be significantly improved with including an efficient spatial verification stage to re-rank the candidates obtained from the Bag-of-Words model.

Considering the 3D data, the Bag-of-Words technique is commonly used for 3D model retrieval [Lavoué 2011; Li et al. 2008; Toldo et al. 2009]. Lavoue et al. in [Lavoué 2011] used simplest Bag-of-Words based pipeline building the visual vocabulary from own local Fourier descriptors and describing each 3D shape by the histogram of word occurrences. They estimated the influence of the vocabulary size on the retrieval results and observed that the increasing of vocabulary size improves the performance (vocabulary sizes observed: 100, 200 and 300), although the difference between last two vocabulary sizes was very small. The experiments were performed on a dataset of 400 objects. Authors of [Li et al. 2008] proposed to incorporate spatial information into the Bag-of-Words model (Spatial enhanced Bag-of-Words) representing each 3D model as a collection of Bag-of-Words histograms of regions along with their relative positions. Authors observed that the spatial information improved the performance comparing to standard Bag-of-Words approach, although the improvement is not large. Martinez-Gomez et al. in [Martínez-Gómez et al. 2016a] presented a Bag-of-Words based object categorization approach using RGB-D images and an experimental evaluation of different learning techniques like k-Nearest-Neighbor, Random Forest and SVMs on the RGB-D Object dataset of 300 objects. According to the results obtained, large vocabulary can lead to higher accuracy, although the increasing vocabulary size degrades the performance for k-NN method.

Bag-of-Words have also been used for the task of visual localization [Martínez-Gómez et al. 2016b; Wang et al. 2005]. Martinez-Gomez et al. in [Martínez-Gómez et al. 2016b] used the Bag-of-Words approach for the task of semantic localization of robot with the use of data from RGB-D sensor. The problem of semantic localization is interpreted as the classification process, where the input perception data should be assigned to a semantic category, e.g. corridor or kitchen. Authors presented comparison between the k-Nearest-Neighbor and SVM classification methods and evaluated the performance of different 3D feature descriptors and keypoint detectors in the proposed localization pipeline. According to reported results, SVM outperformed k-Nearest-Neighbor in most cases.

As it is seen from the state-of-the-art, no studies focusing on a thorough investigation of the Bag-of-Words approach has been observed in 3D domain. In this research, we present a comprehensive investigation of the sensitivity of the Bag-of-Words approach to various parameters in 3D domain. This investigation should be useful for researchers addressing the object search or similar problems in computer graphics and computer vision areas (e.g. place recognition) providing possible directions in the research on this approach and, we believe, will motivate further research in this direction.

3 BAG-OF-WORDS BASED SEARCH APPROACH

This section presents the idea behind the Bag-of-Words approach. The Bag-of-Words approach is a technique used for search documents in textual information retrieval. That technique assumed that each document can be uniquely characterized by the set of words occurred in it. Following that assumption, the technique computes frequency statistics for discriminative words on all documents in offline stage. When one specifies a query term, a search engine uses fast indexing mechanisms to quickly find the documents relevant to the query term taking the word frequency into account [Pangercic et al. 2011].

The Bag-of-Words technique can be effectively adapted to the area of computer vision. We can calculate some local feature descriptors on reference 3D models of objects and these descriptors will characterize the reference object models just as words characterize documents in text retrieval. Since feature descriptors have high dimensionality, there is need to quantize them into discrete terms referred as visual terms or words. The clustering method is the simplest method used for quantizing feature descriptors into visual words. These visual words are accumulated into a visual vocabulary. Next, each 3D model can be uniquely described by a compact sparse vector of visual word frequencies referred as "term vectors" or "Bag of words" (the elements of these vectors are referred as terms). While the term vector representation is independent of the dimensionality of the 3D feature descriptor used, we can apply the methods from text document retrieval for search of relevant 3D models.

Such a compact representation of reference models allows for fast comparison against reference models instead of making direct comparison between local features of these models. Simple similarity metrics such as cosine distance could be applied in the matching procedure.

A common object search procedure consists of two major phases: building database and searching. In the first phase, we use the local features calculated over the set of reference object models and scenes to construct a visual vocabulary and Bag-of-Words objects database. In the testing phase, given a scene point cloud from depth sensor, the search for the best relevant objects by means of compact Bag-of-Words descriptors is carried out.

For constructing the visual vocabulary, this method uses a set of object models and scenes. Scenes are included to the reference data in order to account all possible local appearance characteristics that can occur in the testing scenes and encode them by means of visual terms. That could make the object description more discriminative.

Next step after the constructing visual vocabulary is the building the Bag-of-Words object database. The Bag-of-Words object database describes each reference object model i as a term vector d_i, where each element determines the importance of concrete term (word) from visual vocabulary in object model i:

$$d_i = (w_{i,0}, w_{i,1}, \ldots w_{i,k}) \tag{1}$$

Elements $w_{i,j}$ correspond to all the terms in the vocabulary. k denotes the number of terms in the vocabulary. Each element $w_{i,j}$ associates to the term j a weight that indicates the importance of the term j in the object model i. The term weight $w_{i,j}$ is calculated using the weighting technique TF-IDF [Sivic and Zisserman 2006].

In the search phase, a simple comparison between Bag-of-Words models of scene point cloud and all the object models is performed by means of the cosine distance metric:

$$s_i = \frac{q \cdot d_i}{\|q\| \|d_i\|} \tag{2}$$

Then, all reference object models are ranked based on the similarity scores and n top-ranked ones are considered as candidates for object search.

The procedure above could be commonly considered as the prefiltering stage, as it merely returns a subset of candidates for object search. The Bag-of-Words representation initially ignores the geometric relations between visual words. Two different reference object models could be represented by the same combination of visual terms, but these terms could have different spatial arrangement on these objects. As a result, objects not relevant to the test scene could be returned as wrong candidates. Thus a spatial verification could be optionally included to the Bag-of-Words approach as a refinement resulting in the correct object candidate geometrically consistent with the query scene.

Given a set of n top-ranked object candidates, the spatial verification tests the geometric consistency of these candidates with the test scene using the reference objects database. Alignment algorithms like SAC-IA algorithm from [Rusu et al. 2009] was used for this purpose in our approach.

4 EXPERIMENTAL FRAMEWORK

Our aim is to investigate how the performance of object search is affected by different keypoint detectors, feature descriptors and main internal parameters of the approach, when the Bag-of-Words approach is used for building the reference database and object retrieval. We perform a set of experiments on two widely used RGB-D datasets. The overview of datasets used in experiments, experimental procedure and results obtained are highlighted in more detail in the next subsections.

4.1 Datasets

For experimental evaluation we used Kinect dataset [Aldoma et al. 2012b] and Willow dataset proposed for the ICRA 2011 Solutions in Perception challenge [Aldoma et al. 2013].

Each of datasets includes a set of reference models of objects (CAD models in Kinect dataset and 3D point clouds in Willow dataset) and a set of test scenes represented as point clouds in PCD format. The object models in both datasets represent household objects commonly available in most retail stores. Each test scene in Kinect dataset includes 3 to 6 reference models, while all the scenes in Willow dataset include 6 objects. Both datasets provide ground truth of the object presence in scenes in text files. Two examples of scenes from Willow dataset are presented on Figure 2 and one from Kinect dataset on Figure 3.

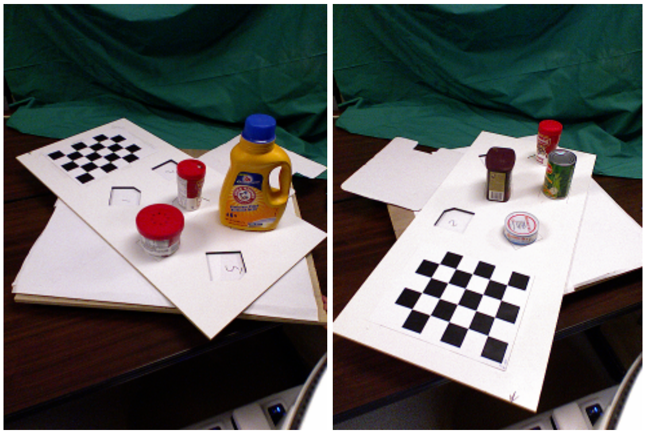

Figure 2: Scene examples from Willow dataset

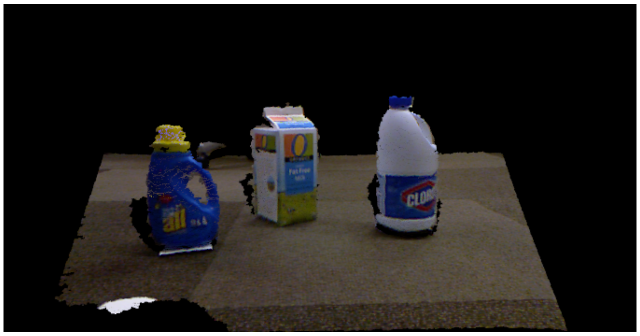

Figure 3: Scene example from Kinect dataset

We built reference and test datasets out of both Kinect and Willow datasets. Considering Kinect dataset, we have chosen 20 object models and 20 scenes for reference dataset and 20 scenes for testing dataset. Considering Willow dataset, we have chosen 30 object models and 20 scenes for reference dataset and 120 scenes for testing dataset. Several scene instances are included to both the reference and test datasets simultaneously for both Kinect and Willow datasets. Willow dataset is more challenging than Kinect dataset as it presents higher degree of occlusions and multiple nearby objects.

4.2 Experiments setup and methodology

We perform several experiments for estimating different characteristics of the approach. In first experiment, we estimate how the density of keypoint sampling influences the performance of object search by performing test on two different keypoint detectors and several feature descriptors:

- keypoint detectors: Harris3D and Uniform Sampling
- feature descriptors: FPFH, SHOT, PFHRGB and SHOTColor

Uniform Sampling detector of keypoints creates a 3D voxel grid over the input point cloud data and for each voxel takes its centroid, i.e. the avarage point inside a voxel, as keypoint. This method results in a uniform dense point cloud of keypoints. Harris3D [Sipiran and Bustos 2011] is an adaptation of the corner and edge based Harris detector used on color images. Harris3D method uses normals of the input point cloud as the input and calculates the covariance matrix in each point applying a local minimim suppression method to select keypoints. This results in a more sparse point cloud of keypoints than that when using the Uniform Sampling. Harris3D detector was used in [Filipe and Alexandre 2014]. Both these keypoint detectors were used by Martinez-Gomez in experiments on the semantic localization problem [Martínez-Gómez et al. 2016b].

FPFH (Fast Point Feature Histogram) [Rusu et al. 2009] and PFHRGB (Point Feature Histogram RGB) [Hänsch et al. 2014] description methods are based on PFH descriptor [Rusu et al. 2008]. For each keypoint, PFH method calculates four values representing geometric relationship between every two points in the neighborhood of the point. All the possible combinations of neighboring points contribute to one histogram. FPFH descriptor is designed to decrease the computational complexity of PFH descriptor. PFHRGB descriptor extends the geometrical information in PFH histogram with color. FPFH encodes 33 values, PFHRGB encodes 250 values. FPFH descriptor was used in [Huang and You 2013] and [Aldoma et al. 2012a]. SHOT [Tombari et al. 2010] method computes local histograms including geometry information of neighboring point locations in a spherical support structure. The final descriptor is built through concatenating all such local histograms resulting in 352 values. SHOTColor [Tombari et al. 2011] includes color data to SHOT descriptor resulting in 1344 values. All the considered description methods were used by Martinez-Gomez in experiments on the semantic localization problem [Martínez-Gómez et al. 2016b] and in [Alexandre 2012] for object recognition problem.

In second experiment, we compare the performance of the feature descriptors above when using Uniform Sampling on Kinect dataset. We are interested in the sensitivity of the precision of the search system wrt. internal parameters of the Bag-of-Words approach. As the internal parameter we considered the vocabulary size. For experimental evaluation of the approach we used following values for vocabulary size: 200, 400, 800 and 1000.

Moreover, we performed two experiments to evaluate the performance of the spatial verification refinement and the generalization capability of the approach.

For evaluating the performance of object search we use the standard mean average precision (mAP) metric, which is commonly used for estimating the quality of search in retrieval systems. mAP is the mean of the average precision scores for all queries.

5 RESULTS AND DISCUSSION

5.1 Evaluation of keypoints sampling density

In this experiment we compare results obtained for two keypoint detectors providing different density of keypoint sampling. Uniform Sampling allows for the most dense and evenly distributed sampling of keypoints. Harris3D provides more sparse points sampling.

The amount of keypoints detected by different keypoint detection methods on test data on individual datasets are presented in Table 1.

Table 1: Number of keypoints detected by different keypoint detection methods on two datasets

dataset	Kinect	Willow
Harris3D	600 - 2K	1K - 2K
Uniform Sampling	5K - 35K	9K - 15K

Table 2: mAP results for dense and sparse keypoint sampling on two datasets

	sparse		dense	
	Kinect	Willow	Kinect	Willow
SHOTColor	0.265	0.218	0.280	0.183
FPFH	0.265	0.183	0.240	0.216
PFHRGB	0.263	0.202	0.200	0.229
SHOT	0.298	0.214	0.292	0.252

The results obtained for Uniform Sampling (dense sampling) and Harris3D (sparse sampling) keypoint detectors on both Kinect and Willow datasets for the vocabulary size 1000 are presented in Table 2.

It is worth to note that Kinect dataset provides no color data in object models and scenes, merely depth data. On the other hand, Willow dataset provides color in both object models and scenes. As it is seen from the table, FPFH, PFHRGB and SHOT features improve the mAP results for dense keypoint sampling on Willow dataset. The superior performance of Uniform Sampling when both color and depth information are available proves the observation that the increasing the number of keypoints improves the object recognition results [Bayramoglu and Alatan 2016]. Therefore, SHOTColor clearly outperforms the other descriptors for Harris3D keypoint detector on Willow dataset. SHOTColor and PFHRGB are the only two features integrating color information.

The dense sampling of keypoints degrades the results on Kinect dataset in all cases. It could be explained by inconsistency in the point cloud resolution of object models and scenes in Kinect dataset (varies from 0.0005 to 0.006 for models and pretty consistent, 0.002, for scenes). The point cloud resolution of the object models and scenes in Willow dataset, on the other hand, is fairly consistent (0.001 vs 0.002 accordingly).

It is also evident that the mAP results are better in most of cases on Kinect dataset. The performance of FPFH drops drastically on Willow dataset and FPFH gets the worst results among the feature descriptors on this dataset. It could be explained by the challenging nature of Willow dataset presenting high degree of occlusion and clutter.

5.2 Keypoint detectors and feature descriptors

The results obtained from experiments on Kinect dataset for Uniform Sampling keypoint detector are presented in Table 3. From the table it is evident that the mAP results tend to increase with increasing the vocabulary size for most of feature descriptors, although this trend is not consistent over all vocabulary sizes. mAP results raise rapidly for SHOTColor. SHOTColor and SHOT perform slightly better than another descriptors. SHOTColor differs from SHOT only in use of color. On the contrary, PFHRGB and FPFH achieve lower precision. The similarity in performance of PFHRGB and FPFH lies in similar nature of these descriptors. FPFH could be a good compromise between search performance and computational complexity.

Table 3: mAP results for Uniform Sampling keypoint detector on Kinect dataset

vocab. size	200	400	1000
SHOTColor	0.240	0.246	0.280
FPFH	0.257	0.258	0.240
PFHRGB	0.210	0.251	0.200
SHOT	0.276	0.264	0.292

The experimental results demonstrate that the search performance is very sensitive to different feature descriptors.

5.3 Spatial verification of candidates

In this experiment, we investigate how the integrating spatial verification stage affects the performance of object retrieval. For this purpose, we compared the results of Bag-of-Words object search with two different configurations: with and without spatial verification. We used Uniform Sampling keypoint detector and two feature descriptors: FPFH and PFHRGB. For the purpose of the experiment, we scored the best 10 object candidates returned from the Bag-of-Words search procedure by SAC-IA fitness and take 5 top-ranked objects as the result.

The results obtained for FPFH on Kinect dataset are presented in Table 4. As shown in the table, spatial verification consistently improves the quality of object search for vocabularies with more than 200 visual words increasing mAP almost twice for vocabulary size 800.

Table 4: mAP results for spatial verification for FPFH on Kinect dataset

vocab. size	200	400	800
verification	0.286	0.305	0.406
no verification	0.308	0.273	0.248

The results obtained for PFHRGB descriptor on Willow dataset could be seen in Table 5. As Table shows, the spatial verification improves the mAP results consistently for vocabulary larger than 200 words. The results are expectable as Willow dataset is enriched with color information and PFHRGB descriptor relies on this modality of data.

The results allow to conclude that spatial verification is a reliable refinement that gives improved robustness in challenging scenes (in particular, in presence of occlusion and clutter for Willow dataset).

Table 5: mAP results for spatial verification for PFHRGB on Willow dataset

vocab. size	200	400	800
verification	0.289	0.321	0.293
no verification	0.294	0.276	0.269

5.4 Generalization capability of the approach

Our goal in this experiment is to investigate how effectively could the algorithm perform the object search when the visual vocabulary is shared between different datasets. For the purpose of this experiment, we have built the visual vocabulary on one dataset and estimated it on the other.

The results obtained in generalization experiment for PFHRGB feature on Kinect dataset are shown in Table 6. We also can see the results obtained in previous experiment on Kinect dataset using the vocabulary built on the same dataset. From table it is seen that the results from generalization experiment are better, although merely for vocabulary sizes 200 and 1000. As Willow dataset provides a larger amount of object models than Kinect dataset, the visual vocabulary is richer accounting more local features.

We also present the results obtained for PFHRGB feature on Willow dataset in Table 7. As we can see, the results are better for two vocabulary sizes: 200 and 400, and degrade for vocabulary size 1000. Thus, although Kinect dataset has a lack of color information, the Bag-of-Words approach allows to build a descriptive visual vocabulary out of purely geometric local characteristics, which could effectively improve the results of object search on different dataset.

We can conclude that the Bag-of-Words approach has a good generalization capabilities, which could enhance the search results.

Table 6: The average precision results obtained in generalization experiment on Kinect dataset. The results obtained on vocabulary built on Willow dataset (first row) are shown in comparison with those obtained on vocabulary built on the same (Kinect) dataset (second row) for PFHRGB descriptor

vocab. size	200	400	1000
Willow	0.250	0.220	0.244
Kinect	0.210	0.251	0.200

Table 7: The average precision results obtained in generalization experiment on Willow dataset. The results obtained on vocabulary built on Kinect dataset (first row) are shown in comparison with those obtained on vocabulary built on the same (Willow) dataset (second row) for PFHRGB descriptor

vocab. size	200	400	1000
Kinect	0.254	0.210	0.199
Willow	0.251	0.116	0.229

We can see that the mAP results obtained in all experiments are quite poor. The reasons behind could be: 1) We used full 3D point clouds of reference models for building the Bag-of-Words object database and tested the approach on 2.5D point clouds, presenting merely the surfaces seen from the viewpoint of the depth sensor; 2) Both datasets include a large number of similar looking objects (This problem more relates to Willow dataset, as the dataset includes a number of objects having similar shape and appearance like ones presented in Figure 4); 3) We applied quite simple methods in our object search pipeline. Probably there is need in more advanced techniques and methods for indexing and weighting Bag-of-Words vectors like soft-weighting; 4) We used relatively small vocabularies for our object search system. Larger vocabularies would improve the performance of object search.

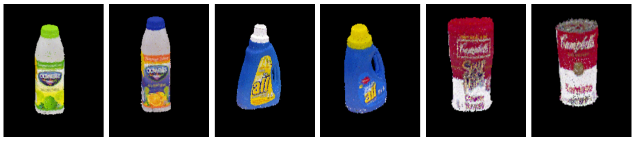

Figure 4: Examples of challenging objects in Willow dataset

Considering similar shape of objects, that challenge is more crucial for Kinect dataset, as the dataset provides no color. For Willow dataset, the feature descriptors relying on color information could cope with this problem.

These results are preliminary as this is the first step in our investigation on the applicability of the Bag-of-Words approach to the object search task. We think that our system for object search based on the Bag-of-Words approach is in its early stage and further improvements are needed for its practical applicability in real-life scenarios. We will continue to work on that problem further on and attempt to achieve better results.

5.5 Discussion

The experimental results lead us to several findings about the practical application of the Bag-of-Words approach to the object search problem.

The investigation of two keypoint sampling strategies has shown that the superiority of a particular sampling method highly depends on the character of underlying data used for the evaluation of the approach. For instance, the dense keypoint sampling can benefit for the performance of object search in case of a dataset providing object models and scenes with consistent point cloud resolution. Otherwise, if the object models and scenes have different point cloud resolution, sparse keypoint sampling methods like Harris3D can be a proper choice. Uniform Sampling merely samples points uniformly regardless the distinctiveness of regions, while Harris3D detects more "meaningfull" points, particularly ones appearing in regions of drastic change in surface normals. Thus we think that sparse sampling allows to tackle the inconsistency in the spatial resolution of point clouds. In addition, it is worth noting that dense keypoint sampling is more expensive than sparse one in terms of computational complexity, as it increases the number of points to deal with.

As the experimental results demonstrate, the integrating spatial verification stage is useful in any cases, particularly for datasets enriched with color information, as the SAC-IA algorithm relies on all the available data modalities, i.e. color and depth, depending on the type of feature descriptor used.

Therefore, in different cases, some descriptors are more efficient than the other. SHOT descriptor has shown better results than another descriptors in most cases. The performance of FPFH descriptor degrades under the presence of occlusions and nearby objects as was also reported by the authors of [Aldoma et al. 2012a]. It could be seen that the descriptors having similar nature like SHOT and SHOTColor do not differ radically in performance, thus either of those could be used depending on available data modalities in the dataset.

6 CONCLUSION

In this paper, we investigate the application of the Bag-of-Words approach for object search task in 3D domain. We have performed the experimental evaluation of the approach on two widely-used datasets. During the experiments we evaluated the effect of the density of keypoint sampling on the performance of object search, several feature descriptors designed for RGB-D data and several vocabulary sizes.

The obtained results have demonstrated that the performance of object search increases when the keypoints are densely sampled leading to high-quality object representations and is highly sensitive to different feature descriptors. The results also confirmed the benefits of including color information in the description of object models for search. Experiments on spatial verification refinement applied to object candidates have shown that this additional refinement stage could improve the performance of object search even in presence of occlusions and clutter in scene. Moreover, the experimental results on generalization capability of the approach obtained when the visual vocabulary was built on one dataset and the search performed on another, have shown the good reliability of the approach.

As future work, besides working with other larger datasets, we plan to improve the quality of the approach by including weighting scheme (like soft-weighting). The robustness of the object search approach to appearance and geometric similarity of objects could probably be improved by using partial 2.5D views of object models captured from different viewpoints rather than full 3D point clouds in the vocabulary construction stage. That could allow to retrieve a correct partial view of object seen from the depth sensor in the scene, thus accounting the viewpoint of the camera. Further we plan to investigate the possibility of adaptation of the vocabulary or applied weights for tasks where the system is operating in various conditions (data domains) 24/7.

REFERENCES

Aitor Aldoma, Zoltan C. Marton, Federico Tombari, Walter Wohlkinger, Christian Potthast, Bernhard Zeisl, Radu B. Rusu, Suat Gedikli, and Markus Vincze. 2012a. Tutorial: Point Cloud Library: Three-Dimensional Object Recognition and 6 DOF Pose Estimation. *IEEE Robotics Automation Magazine* 19, 3 (Sept 2012), 80–91. https://doi.org/10.1109/MRA.2012.2206675

Aitor Aldoma, Federico Tombari, Johann Prankl, Andreas Richtsfeld, Luigi Di Stefano, and Markus Vincze. 2013. Multimodal cue integration through Hypotheses Verification for RGB-D object recognition and 6DOF pose estimation. In *2013 IEEE International Conference on Robotics and Automation*. 2104–2111. https://doi.org/10.1109/ICRA.2013.6630859

Aitor Aldoma, Federico Tombari, Luigi Di Stefano, and Markus Vincze. 2012b. *A Global Hypotheses Verification Method for 3D Object Recognition*. Springer Berlin Heidelberg, Berlin, Heidelberg, 511–524. https://doi.org/10.1007/978-3-642-33712-3_37

Lu'is A. Alexandre. 2012. 3D Descriptors for Object and Category Recognition: a Comparative Evaluation. In *IEEE International Conf. on Intelligent Robotic Systems - IROS*, Vol. Workshop on Color-Depth Camera Fusion in Robotics. 1–6.

Khaled Alhamzi, Mohammed Elmogy, and Sherif Barakat. 2014. 3D Object Recognition Based on Image Features: A Survey. *International Journal of Computer and Information Technology* 3, 03 (2014), 651–660.

Neslihan Bayramoglu and A. Aydin Alatan. 2016. Comparison of 3D Local and Global Descriptors for Similarity Retrieval of Range Data. *Neurocomput.* 184, C (April 2016), 13–27. https://doi.org/10.1016/j.neucom.2015.08.105

Silvio Filipe and Luís A. Alexandre. 2014. A comparative evaluation of 3D keypoint detectors in a RGB-D Object Dataset. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*.

Ronny Hänsch, Thomas Weber, and Olaf Hellwich. 2014. Comparison of 3D interest point detectors and descriptors for point cloud fusion. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* (Aug. 2014), 57–64. https://doi.org/10.5194/isprsannals-II-3-57-2014

Jing Huang and Suya You. 2013. Detecting Objects in Scene Point Cloud: A Combinational Approach. In *2013 International Conference on 3D Vision - 3DV 2013*. 175–182. https://doi.org/10.1109/3DV.2013.31

Guillaume Lavoué. 2011. Bag of Words and Local Spectral Descriptor for 3D Partial Shape Retrieval. In *Proceedings of the 4th Eurographics Conference on 3D Object Retrieval (3DOR '11)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 41–48. https://doi.org/10.2312/3DOR/3DOR11/041-048

Xiaolan Li, Afzal Godil, and Asim Wagan. 2008. *Spatially Enhanced Bags of Words for 3D Shape Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, 349–358. https://doi.org/10.1007/978-3-540-89639-5_34

Jesus Martínez-Gómez, Miguel Cazorla, Ismael García-Varea, and Cristina Romero-González. 2016a. *Object Categorization from RGB-D Local Features and Bag of Words*. Springer International Publishing, Cham, 635–644. https://doi.org/10.1007/978-3-319-27149-1_49

Jesus Martínez-Gómez, Vicente Morell, Miguel Cazorla, and Ismael García-Varea. 2016b. Semantic localization in the PCL library. *Robotics and Autonomous Systems* 75, Part B (2016), 641 – 648. https://doi.org/10.1016/j.robot.2015.09.006

David Nister and Henrik Stewenius. 2006. Scalable Recognition with a Vocabulary Tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR '06)*. IEEE Computer Society, Washington, DC, USA, 2161–2168. https://doi.org/10.1109/CVPR.2006.264

Dejan Pangercic, Vladimir Haltakov, and Michael Beetz. 2011. Fast and Robust Object Detection in Household Environments Using Vocabulary Trees with SIFT Descriptors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on Active Semantic Perception and Object Search in the Real World*. San Francisco, CA, USA.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching.. In *CVPR*. IEEE Computer Society. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2007.html#PhilbinCISZ07

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. 2009. Fast Point Feature Histograms (FPFH) for 3D Registration. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan, May 12-17*.

Radu Bogdan Rusu, Zoltan Csaba Marton, Nico Blodow, and Michael Beetz. 2008. Persistent Point Feature Histograms for 3D Point Clouds. In *Proceedings of the 10th International Conference on Intelligent Autonomous Systems (IAS-10), Baden-Baden, Germany*.

Ivan Sipiran and Benjamin Bustos. 2011. Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes. *The Visual Computer* 27 (2011), 963–976.

Josef Sivic and Andrew Zisserman. 2006. *Video Google: Efficient Visual Search of Videos*. Springer Berlin Heidelberg, Berlin, Heidelberg, 127–144. https://doi.org/10.1007/11957959_7

Roberto Toldo, Umberto Castellani, and Andrea Fusiello. 2009. A Bag of Words Approach for 3D Object Categorization. In *Proceedings of the 4th International Conference on Computer Vision/Computer Graphics CollaborationTechniques (MIRAGE '09)*. Springer-Verlag, Berlin, Heidelberg, 116–127. https://doi.org/10.1007/978-3-642-01811-4_11

Federico Tombari, Samuele Salti, and Luigi Di Stefano. 2010. Unique Signatures of Histograms for Local Surface Description. In *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III (ECCV'10)*. Springer-Verlag, Berlin, Heidelberg, 356–369. http://dl.acm.org/citation.cfm?id=1927006.1927035

F. Tombari, S. Salti, and L. Di Stefano. 2011. A combined texture-shape descriptor for enhanced 3D feature matching. In *2011 18th IEEE International Conference on Image Processing*. 809–812. https://doi.org/10.1109/ICIP.2011.6116679

Markus Waibel, Michael Beetz, Raffaello D'Andrea, Rob Janssen, Moritz Tenorth, Javier Civera, Jos Elfring, Dorian Gálvez-López, Kai Häussermann, J.M.M. Montiel, Alexander Perzylo, Björn Schießle, Oliver Zweigle, and René van de Molengraft. 2011. RoboEarth - A World Wide Web for Robots. *Robotics & Automation Magazine* 18, 2 (2011), 69–82.

Junqiu Wang, Roberto Cipolla, and Hongbin Zha. 2005. Vision-based Global Localization Using a Visual Vocabulary. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. 4230–4235. https://doi.org/10.1109/ROBOT.2005.1570770