2021 IEEE International Conference on
Autonomous Robot Systems and Competitions (ICARSC)
April 28-29, Santa Maria da Feira, Portugal

2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) | 978-1-6654-3198-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICARSC52212.2021.9429799

Evaluation of Bags of Binary Words for Place Recognition in Challenging Scenarios

Ana Rita Gaspar
FEUP
INESC TEC
Porto, Portugal
up201402645@fe.up.pt

Alexandra Nunes
FEUP
INESC TEC
Porto, Portugal
up201402644@fe.up.pt

Aníbal Matos
FEUP
INESC TEC
Porto, Portugal
anibal@fe.up.pt

***Abstract*—To perform autonomous tasks, robots in real-world environments must be able to navigate in dynamic and unknown spaces. To do so, they must recognize previously seen places to compensate for accumulated positional deviations. This task requires effective identification of recovered landmarks to produce a consistent map, and the use of binary descriptors is increasing, especially because of their compact representation. The visual Bag-of-Words (BoW) algorithm is one of the most commonly used techniques to perform appearance-based loop closure detection quickly and robustly. Therefore, this paper presents a behavioral evaluation of a conventional BoW scheme based on Oriented FAST and Rotated BRIEF (ORB) features for image similarity detection in challenging scenarios. For each scenario, full-indexing vocabularies are created to model the operating environment and evaluate the performance for recognizing previously seen places similar to online approaches. Experiments were conducted on multiple public datasets containing scene changes, perceptual aliasing conditions, or dynamic elements. The Bag of Binary Words technique shows a good balance to deal with such severe conditions at a low computational cost.**

***Index Terms*—SLAM, appearance-based localization, bag of binary words, place recognition, loop closure**

I. Introduction

Autonomous robots operating in real-world environments need to be able to navigate in dynamic and unknown spaces, and therefore they need to map their operational environment in order to localize themselves (SLAM). For successful navigation, recognizing a place that has already been visited by the vehicle is an essential aspect [1]. Making this decision independently of the surroundings - unsupervised learning - is yet a challenge. This task requires effective identification of landmarks already seen to produce a consistent map. Given the need to use viewpoint invariant feature extractors and descriptors, and given the large extraction time required for SIFT/SURF features, binary algorithms are increasingly used for place recognition [2]. These features have a very compact representation, which in turn requires less memory and low computational time. There are several methods for place recognition, but visual Bag-of-Words (BoW) models stand out for performing loop closure detection quickly and robustly [3]. They excel in simplicity and speed in searching for similar images and provide results with less computational overhead.

FCT - Fundação para a Ciência e a Tecnologia

Moreover, these models can be arranged in hierarchical vocabulary trees, which allows efficient search. The vocabulary can be created using offline or online approaches, but it is usually learned using an unsupervised clustering algorithm such as K-Means or Random Sampling. Nevertheless, the SLAM technique is widely used in ground robotics, but some outdoor scenarios are characterized by repetitive textures, such as grass, trees or soil, which favors the phenomenon of perceptual aliasing. However, BoW approaches are sensitive to this phenomenon. In addition, outdoor scenarios can also be very dynamic, making place detection difficult, as the appearance of a place can change over time due to changes in lighting or weather. This is of particular concern for offline vocabulary-based approaches, as the vocabulary may not be representative of the scenario. Considering the advantages of using binary features, it is important to evaluate their effectiveness in extracting image features and performing scene matching in outdoor scenarios, which in themselves can present severe conditions. Therefore, this paper presents a comparative analysis of some binary descriptors, discussing the performance of the methods based on image transformations and the computation time. Moreover, given the strengths of the BoW techniques, a fundamental question is raised: *Are these models not really suitable for uncontrolled environments?* Therefore, as main contribution this paper intends to provide an evaluation of the behaviour of BoW approaches in challenging scenarios involving scene changes, perceptual aliasing conditions or dynamic elements.

This work is arranged as follows: section II presents basic works in the field. Section III gives an overview of a visual place recognition system. It also describes the used BoW-based approaches and performance metrics of place recognition. Then, section IV describes the community datasets used for evaluating the recognition of previously seen places and reports the different experiments performed. Finally, the main conclusions of this work are discussed in section V.

II. Related Works

The BoW algorithm is one of the most widely used method for appearance-based loop closure detection. In this context, the first vocabulary approach based on binary features (BRIEF) was proposed in [4]. The vocabulary is built offline by relying

on a hierarchical BoW model and the BRIEF descriptor showed tolerate only small scale or rotation changes. This fact limits the system to in-line trajectories and loop events with a similar viewpoint. Therefore, this work was later extended by using ORB features. It achieved higher recall (strength of the algorithm to recognize known places) than BRIEF in outdoor scenarios with larger viewpoint differences [5]. To deal with perceptual aliasing, an algorithm to validate the matched scenes and delete outliers is crucial. On the other hand, there are also some approaches that avoid vector quantization and use direct feature matching to compute similarity. This alternative shows better recall performance, but it implies higher computational cost [6]. To overcome this limitation of conventional BoW but maintain their simplicity, the use of spatial co-occurrence of words directly in the vocabulary itself was later proposed [7]. It provides better discrimination in loop closure detection and hence showed higher recall performance under such conditions. With the goal of successfully detecting loop closure situations in uncontrolled environments, a visual vocabulary procedure that is created when visual information becomes available during the vehicle navigation was proposed [8]. This appearance-based incremental vocabulary outperforms offline vocabularies, and it strongly approaches the full-indexing vocabulary - also created offline, but with images of the performed trajectory - in outdoor scenarios with some dynamic elements. However, the method is still affected by perceptual aliasing. Later, in [9], due to the importance of building the codebook online, an incremental vocabulary approach based on a hierarchical scheme (BoW) and relying on the FLANN library was proposed. This approach combines global and local binary descriptors and shows high performance, but is still vulnerable to dynamic elements. Recently, an incremental BoW strategy has been proposed where the vocabulary is created using ORB features and is based on hashing techniques [10]. It provides a word deletion policy that reduces the size of the visual vocabulary with little impact on performance and consequently requires less computation time. This approach provides good recall in indoor and outdoor environments with static and slightly dynamic objects.

III. Visual Place Recognition

Proper loop closure balances accumulated position drifts, and thus this mechanism is crucial for successful autonomous navigation. Thus, the question is: *"Given an image of a place, how can a robot decide whether this image is a place it has already seen?"*. A place recognition system consists of three main modules (green blocks), see Fig. 1. The processing module interprets the incoming data, the map contains a representation of the robot's environment, and the belief generation module compares the incoming sensor data with the map and decides whether the vehicle is in a known or new location. Finally, the map is updated accordingly.

Resorting to BoW models, a visual vocabulary is used to represent an image as a single numerical vector, i.e., each image feature is associated with a word. Typically, the vocabulary is learned using an unsupervised clustering algorithm, such as K-Means, applied to local descriptors from an image dataset. A confidence measure is performed to check whether the current visual input matches a particular location in the map representation of the robot's world. Then, a list of loop candidates is obtained. Finally, the obtained loops must be evaluated to verify whether the match is correct or incorrect, and consequently to determine the robustness of the system to detect similar places in a different viewpoint and timestamp.

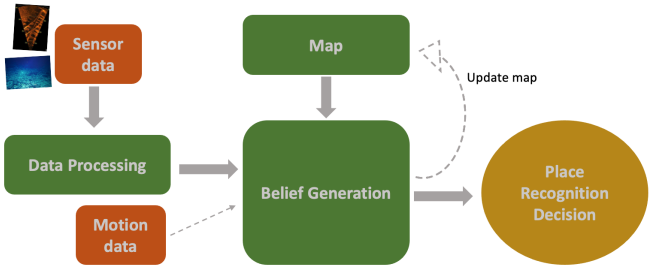

Fig. 1. Overview of a visual place recognition system.

This section describes the approaches taken in this research for place recognition system in challenging outdoor scenarios. In subsection III-A, the approach for indexing and converting images into a BoW representation is presented. In subsection III-B, the system used to detect loops in an image sequence is described in detail - corresponding to the belief generation module in a place recognition system. Finally, the subsection III-C describes the performance metrics used to evaluate loop closure detections.

A. DBoW2

DBoW2 is a C++ open-source library[1] that implements a hierarchical tree for approximating nearest neighbors in image feature space and creating a visual vocabulary [4]. Given a rich set of features extracted from some images, the vocabulary tree is based on agglomerative hierarchical K-Means++ clustering [11]. Two user-defined parameters are required: the depth factor, L, which defines the maximum level of the tree, and the branching factor, K, which defines the maximum number of nodes (groups) for each level. The clustering steps are performed for each level based on the clusters obtained in the previous level until the maximum depth L is reached. Finally, a tree of W leaves - words of the vocabulary - is obtained, where each leaf node is a group containing a single data point. Each word has an associated weight according to its relevance in the training corpus that later is used to measure the similarity between two images (bag-of-words vectors), as explained below in the III-B subsection. In addition, an inverted and direct files can be kept along the bag-of-words, allowing fast queries and feature comparisons, see Fig. 2.

[1]https://github.com/dorian3d/DBoW2

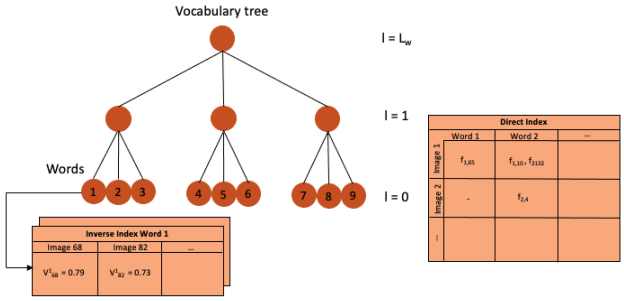

Fig. 2. Example of a vocabulary tree with direct and inverse indexes.

B. DLoopDetector

DLoopDetector is an open-source C++ library[2] for loop detection based on a BoW structure. It implements temporal and geometrical constraints between any pair of images of the loop candidates, to obtain more consistent results in the similar place detection [4].

For each current image, their features are converted into a bag-of-words vector, v_t and the database is searched for v_t, resulting in a list of matched candidates based on the weights of the words and their normalized scores (L1 distance). Only matches whose score exceeds a minimum threshold α are considered. The images that are close in time are grouped into islands and treated as only one match. Thus, only the island with the highest score is selected as the matched group. If an island passes the temporal constraint - more than an user-defined number of k consistent islands - only the match (database image) with the higher normalized score is considered as a loop closure candidate. Next, this candidate must be accepted by the geometric check to be considered a loop, by finding a fundamental matrix between the matched images using RANSAC. To compute these matches, this library takes advantage of the BoW vocabulary and resorts to the direct index, where the comparison is performed only between those features associated with the same nodes at an user-specified level number (level l) in the vocabulary tree. Fig. 3 represents the process overview.

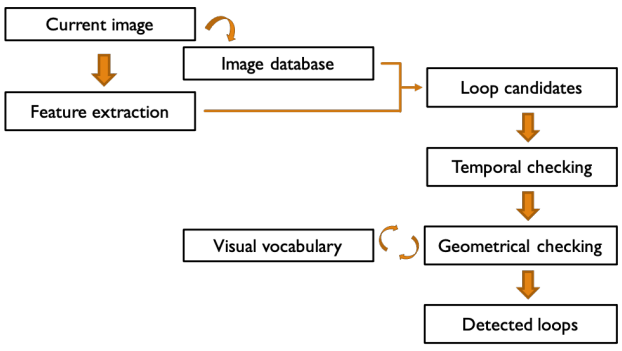

Fig. 3. Diagram of the main steps of DLoopDetector for image similarity detection.

[2]https://github.com/dorian3d/DLoopDetector

C. Loop Closure Evaluation

The performance of the place recognition algorithm is typically evaluated using precision-recall metrics/curve. The evaluation against ground-truth is done by counting true positives (TP), false positives (FP) and false negatives (FN). TP defines the cases where the algorithm successfully recognizes the queried image as a known place. In contrast, FP are the cases in which the algorithm incorrectly recognizes the queried image as a known place. The cases where the place recognition algorithm does not falsely recognize the queried image as a known place are considered as FN (undetected loops). Then, precision and recall metrics are computed:

- *Precision* is the proportion of relevant instances (TP) among all retrieved instances (TP+FP).
- *Recall* is the proportion of relevant instances (TP) among all instances that were actually retrieved (TP+FN).

Thus, precision describes the robustness of the algorithm to recognize a place without error, while recall determines the strength of the algorithm to recognize known places without loss. In the navigation context, the maximum recall that can be achieved at a precision of 100% is of particular interest: a false positive will cause the algorithm to produce inconsistent maps and consequently an unreliable pose estimate, which can lead to irreparable failures in many robotic applications. A high recall also allows for better motion estimation, as it avoids incorrect trajectory adjustments. To find an ideal combination of precision/recall, the F1-score is also computed. It is a harmonic mean of both metrics:

$$\text{F1-score } (\%) = 2 \times \frac{Precision \times Recall}{Precision + Recall},$$

where a higher value indicates the most suitable mix.

IV. Experimental Results

In this section, the ability of conventional BoW techniques to deal with the common challenges that exist in outdoor scenarios is evaluated. First, the datasets used for the different experiments are described in subsection IV-A. Next, subsection IV-B provides an analysis of the behavior of binary features for working in such scenarios. Then, subsection IV-C demonstrates the performance of BoW techniques for recognizing previously seen places, in different operating conditions, namely in the presence of scene changes or dynamic objects and under perceptual aliasing conditions. All experiments were performed using an AMD Ryzen 7 2700X @ 3.7Ghz with 16GB RAM computer.

A. Datasets

Behavioral evaluation of BoW techniques in challenging scenarios was performed with several outdoor community datasets corresponding to different operational conditions. The KITTI odometry benchmark[3] consists of 22 sequences in urban and highway environments [12]. Sequence 05 that describes an urban scenario and has various loop closures

[3]http://www.cvlibs.net/datasets/kitti/eval_odometry.php

was used. The New College dataset[4] was also used [13]. The images are captured by a robot as it moves through the environment, which includes buildings, open spaces, and vegetated areas. The interest for this dataset lies in the high perceptual aliasing conditions. Finally, the traditional BoW was validated against a St. Lucia dataset part[5]. It includes construction sites, bright scenes, dark scenes, and multi-lane roads [14]. Fig. 4 shows a sample scene from each dataset. The details used of each dataset are summarized in Table I.

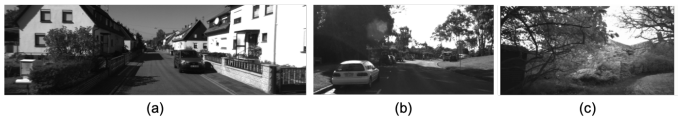

Fig. 4. Example scenario shown in KITTI 05 (a), St. Lucia (b) and New College (c) dataset.

TABLE I
DATASETS USED TO VALIDATE BOW TECHNIQUES

Dataset	# Images	Size (px)	Rate (Hz)	Length (km)
New College	281	1280 x 960	20	0.42
KITTTI 05	2761	1226 x 370	10	2.2
St. Lucia	2901	1024 x 518	10	2.7

B. Descriptors effectiveness

When using image processing for navigation tasks, it is necessary to extract reliable image features both for matching keypoints and detect loop closures. To achieve this, there are several detectors/descriptors, but the binary category has been increasingly applied in this context. Based on the invariance assumed by each descriptor, ORB and BRISK are among the best scale-invariant, rotation-invariant and affine-invariant feature detectors [15] [2]. Therefore, an evaluation of their operational capabilities under some challenging conditions is presented in this section. For this experiment both ORB and BRISK vocabularies were created and the ability to detect similarity between images is checked based on DLoopDetector process, as detailed above in section IV-C.

Fig. 5 shows the scenarios selected from the KITTI 05 dataset to evaluate the behavior of the feature descriptors to detect similar places in some challenging situations, namely scale invariance, changes in viewpoint (including rotation and orientation), and variations in illumination. Each case represents the same place but acquired in different conditions and timestamp. Both descriptors showed be able to recognize places as known, demonstrating their robustness to these scene changes. BRISK identifies more matching points between images in most cases. However, after removing outliers - finding a fundamental matrix with RANSAC - it obtains fewer final matching points. This means that the ORB matches are effective and it is, therefore, less sensitive to changes in the scene, especially in terms of orientation and illumination variations. Moreover, both descriptors are computationally efficient, but BRISK is slightly more expensive: about three times more for KITTI 05. Therefore, the next experiments will be based on the ORB descriptor, adapting DLoopDetector accordingly.

[4]https://www.robots.ox.ac.uk/∼mobile/IJRR_2008_Dataset/data.html
[5]https://wiki.qut.edu.au/display/raq/UQStLucia

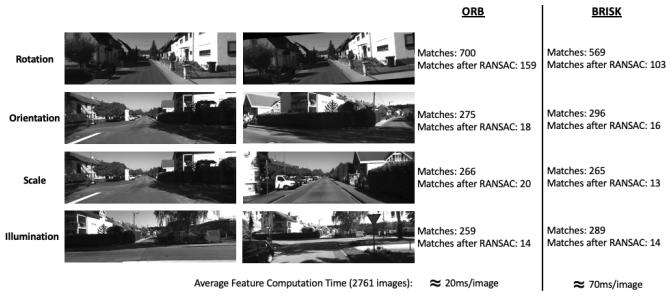

Fig. 5. Scenes selected from the KITTI 05 dataset to evaluate the rotation, scale, viewpoint and illumination invariances of the ORB and BRISK descriptors.

C. Loop Closure Detection

Next, the performance of DLoopDetector was evaluated based on the BoW scheme for detecting previously seen places. To obtain performance measures, a ground-truth for loop closure was generated for each dataset. Each one provides either a pose file which is used to generate the ground-truth using 4m, 15m and 4.5m as thresholds for placing two images at the same location for the KITTI 05, St. Lucia and New College respectively. Moreover, only the loops where the distance to ground-truth is greater than the frequency F are considered as final loops to prevent multiple loops from being detected in a second. Finally, the situations where the camera motion is stopped were discarded. Evaluation against ground-truth is based on precision and recall metrics. An ideal precision/recall mix is found by the F1-score. Thus, for each sequence, TP, FP and FN are counted as described in section III-C. For all experiments, full-indexing vocabularies are used instead of using vocabularies with a training set, as suggested by the DLoopDetector library and common in BoW techniques. Thus, the operating environment is modeled and the performance for detecting previously seen places is evaluated similarly to online approaches.

The precision-recall curves for KITTI 05 and St. Lucia varying the threshold similarity parameter, α, between 0.1 and 0.45, and extracting 1000 keypoints per image are shown in Fig. 6. For clarity, the best results at 100% of precision (Pr) are shown in Table II. In both datasets, a high recall rate (Re) is achieved with no false positives. For KITTI 05, a low value of similarity means that false loop situations are detected (FP), and the higher the value of similarity, the stronger the algorithm is able to detect loop situations without losing any. For St. Lucia, the best performance is achieved at the expense of simplifying the conditions required to recognize a place as known (low similarity threshold, α = 0.15). Nevertheless, overall, more recognitions are lost

compared to KITTI 05, reaching a maximum recall of 45.90% for α = 0.45. When the default similarity is required for both scenarios (α = 0.3), the algorithm is not robust enough to recognize places without errors in St. Lucia, which means that FP are detected and does not achieve 100% of precision. Thus, the keypoints extracted for each image represent well a simple environment. Conversely, 1000 features are not at all sufficient to model sophisticated scenes for which the algorithm is able to recognize places.

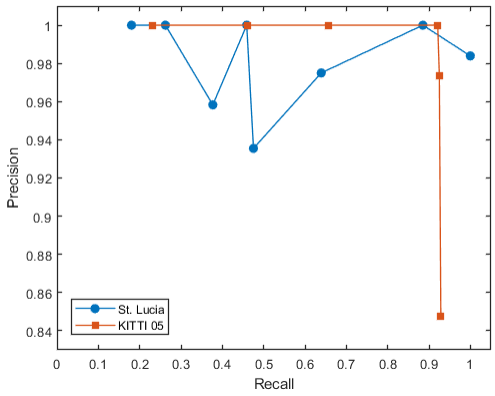

Fig. 6. Precision-recall curves in the KITTI 05 and St. Lucia datasets for different values of the similarity threshold , α, extracting 1000 keypoints per image.

TABLE II
RESULTS AT 100% OF PRECISION IN THE KITTI 05 AND ST. LUCIA DATASETS, FOR SEVERAL VALUES OF SIMILARITY THRESHOLD α.

Dataset	Pr	Re
KITTTI 05	100	92.11
St. Lucia	100	88.52

Table III shows the performance obtained in the KITTI 05 and St. Lucia datasets by varying the number of keypoints used to build the visual vocabularies (Voc). Specifically, vocabularies using 1000 and 1500 features are created. For detecting image similarities, 2000 keypoints per image are extracted and the default similarity threshold (α = 0.3) are considered for both datasets.

TABLE III
RESULTS IN THE KITTI 05 AND ST. LUCIA DATASETS, VARYING THE KEY POINTS NUMBER EXTRACTED IN VOCABULARY.

Dataset	Voc	Pr	Re	F1-score
KITTTI 05	1000	100	86.44	92.96
	1500	100	65.79	79.37
St. Lucia	1000	87.88	47.54	61.70
	1500	100	29.51	45.57

Overall, with the same number of features used to build the codebook, the algorithm performs better in the KITTI 05 environment. Using 1000 features in this scenario achieves high recall at 100% of precision, while in St. Lucia the algorithm is not robust enough (87.88% of precision) to detect familiar places without loss, and detects many false positives. Using 1500 features instead, the algorithm is able to achieve 100% of precision in the St. Lucia scenario but loses more loop closures. This means that a more representative vocabulary contributes to the robustness of the algorithm, but the required similarity is high to detect all loop closure situations in a difficult scenario like St. Lucia. The precision/recall mix is higher extracting 1000 features per image, but since there are false detections (FP), this is not the better option in the navigation context, since it may cause an inappropriate pose estimation. In KITTI 05, the robustness of the algorithm is not lost, but the recall decreases: since this scenario is simple, a larger number of features implies a too detailed discrimination of the environment, which in combination with a high similarity threshold causes the algorithm to lose the strength to detect familiar places.

The best results at 100% of precision, by varying the threshold similarity parameter α, and the average computation time (ms) obtained in the KITTI 05, St. Lucia and New College datasets are shown in Table IV. For clarity, 2000 keypoints per image are used for all methods.

TABLE IV
RESULTS IN THE KITTI05, ST. LUCIA AND NEWCOLLEGE DATASETS

Dataset	Pr	Re	Pr(0.25)	Re(0.25)	Time
KITTTI 05	100	90	100	82.21	39.3
St. Lucia	100	68.85	100	50.82	32.8
New College	100	90	94.74	90	47.6

As can be observed, more image features are required to distinguish the environment and to improve the performance in the St. Lucia dataset, since this environment is very dynamic and has scene changes. Nevertheless, this result is obtained with a low similarity threshold, and still about 1/3 of the loop closures are not detected. The used sequence of the New College dataset is strongly composed of walls and green/outdoor areas (trees, grass and shrubs), with the interest of evaluating the behavior of the algorithm under many key points, but diverse repetitive patterns. The proposed approach is able to deal with high perceptual aliasing conditions, as the algorithm successfully recognizes the queried image as a known place (100% of precision) and identifies almost all loop situations (90% of recall). However, this best result is achieved at the cost of a high similarity threshold used compared to the other two scenarios. In terms of computational time, the algorithm proves to be efficient. The biggest difference occurs in the feature computation phase for the New College dataset, where the required maximum keypoints in all images are extracted. Table IV also shows the results obtained by specifying an equal similarity threshold to better analyse the effect of setting this parameter to a fixed value in terms of performance. For this purpose, the datasets were processed for α = 0.25. As can be observed, the recall in the KITTI 05 and St. Lucia datasets decreases somewhat due to the increase in the similarity threshold (more sensitivity). In contrast, for

the New College dataset, the algorithm misdetects situations with loop closure. Since this scenario has similar features, the required similarity is not enough for distinctiveness. Following the results presented in Table IV, the Fig. 7 shows the loops detected by DLoopDetector based on ORB-features at 100% of precision for each dataset. Note that most of the loops present were detected, especially in the KITTI 05 and New College datasets.

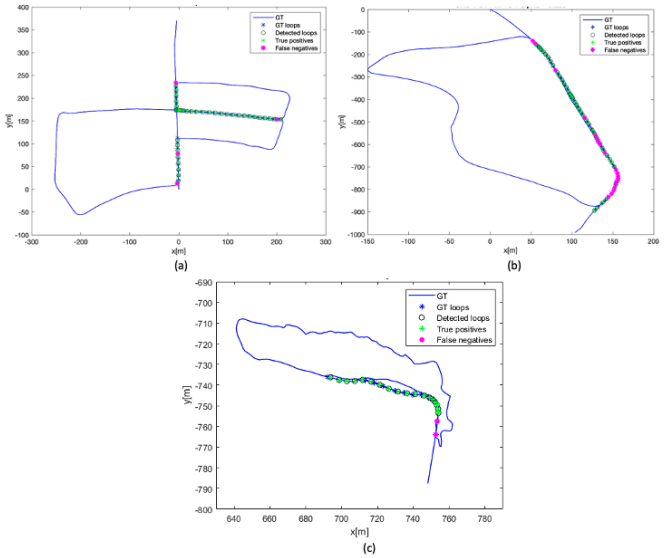

Fig. 7. Appearence-based loop closure results for KTTI05 (a), St. Lucia (b) and New College (c) datasets.

V. Conclusion

In this paper, a behavioral evaluation of a binary BoW technique based on a traditional learning approach was presented for uncontrolled environments involving scene changes, perceptual aliasing conditions, or dynamic elements. The KITTI 05 was preferably used to evaluate the effectiveness of binary features (ORB and BRISK) in challenging situations, namely scale invariance, changes in viewpoint, and variations in illumination. Due to its simplicity, place recognition requires a small number of keypoints and a low similarity threshold so as not to compromise recall. In contrast, St. Lucia requires a high number of features to represent the environment and achieve a recall of about 70% at 100% of precision without high similarity thresholds. Finally, for the New College, a high number of features combined with a high similarity is required to ensure the distinctiveness for place recognition, due to similar features. Thus, it was proved that the more challenging an environment is, a more representative visual vocabulary contributes to the robustness of the loop closure algorithm. From the obtained results, the BoW techniques were shown to be able to handle severe conditions, albeit at the expense of the extraction of 2000 features per image. Nevertheless, the obtained results proved that the required computational effort is reduced, and with a suitable hardware selection, this approach can be applied in a real application.

In terms of future research, tuning of BoW parameters and comparison with state-of-the-art methods will be performed. An evaluation of some indexing methods, such as K-D trees or hashing techniques, will be also considered. In addition, a performance evaluation of this conventional BoW technique based on binary features will be envisaged in an underwater environment involving some repetitive patterns with few features.

Acknowledgment

This work is financed by FCT - Fundação para a Ciência e a Tecnologia - and by FSE - Fundo Social Europeu through of the Norte 2020 – Programa Operacional Regional do Norte - through of the doctoral scholarship SFRH/BD/146460/2019. This work is also partially financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

[1] J. Melo and A. Matos, "Survey on advances on terrain based navigation for autonomous underwater vehicles," *Ocean Engineering*, vol. 139, pp. 250–264, 2017.

[2] S. A. K. Tareen and Z. Saleem, "A Comparative Analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK," in *International Conference on Computing, Mathematics and Engineering Technologies*, 2018.

[3] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, pp. 55–81, 2015.

[4] D. Gálvez-López and J. D. Tardós, "Real-Time Loop Detection with Bags of Binary Words," in *International Conference on Intelligent Robots and Systems*, pp. 51–58, 2011.

[5] R. Mur-Artal and J. D. Tardós, "Fast Relocalisation and Loop Closing in Keyframe-Based SLAM," in *IEEE International Conference on Robotics and Automation*, no. June, pp. 846–853, 2014.

[6] M. T. Law, N. Thome, and M. Cord, "Bag-of-Words Image Representation: Key Ideas and Further Insight," in *Fusion in Computer Vision - Understanding Complex Visual Content*, ch. 2, pp. 29–52, Springer International, 2014.

[7] N. Kejriwal, S. Kumar, and T. Shibata, "High performance loop closure detection using bag of word pairs," *Robotics and Autonomous Systems*, vol. 77, pp. 55–65, 2016.

[8] T. Nicosevici and R. Garcia, "Automatic Visual Bag-of-Words for Online Robot Navigation and Mapping," *IEEE Transactions on Robotics*, vol. 28, no. 4, pp. 886–898, 2012.

[9] E. Garcia-Fidalgo and A. Ortiz, "Hierarchical Place Recognition for Topological Mapping," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1061–1074, 2017.

[10] E. Garcia-Fidalgo and A. Ortiz, "IBoW-LCD: An Appearance-based Loop-Closure Detection Approach using Incremental Bags of Binary Words," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3051–3057, 2018.

[11] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in *SODA '07: Proceedings of the eighteenth annual ACMSIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.

[12] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI Vision Benchmark Suite," in *Conference on Computer Vision and Pattern Recognition*, 2012.

[13] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman, "The New College Vision and Laser Data Set," *International Journal of Robotics Research*, vol. 28, no. 5, pp. 595–599, 2009.

[14] M. Warren, D. McKinnon, H. He, and B. Upcroft, "Unaided Stereo Vision Based Pose Estimation," in *Australasian Conference on Robotics and Automation*, 2010.

[15] H. Chatoux, F. Lecellier, C. Fernandez-maloigne, H. Chatoux, F. Lecellier, C. F.-m. Comparative, and H. Chatoux, "Comparative Study of Descriptors with Dense Key points," in *23rd International Conference on Pattern Recognition*, 2016.