SeqSLAM with Bag of Visual Words for Appearance Based Loop Closure Detection

Konstantinos A. Tsintotas(✉), Loukas Bampis, Stelios Rallis, and Antonios Gasteratos

Department of Production and Management Engineering, Democritus University of Thrace, 12 Vas. Sophias, 671 32 Xanthi, Greece
ktsintot@pme.duth.gr

Abstract. The detection of pre-visited areas in robots' traversed path, widely known as loop closure detection, is vital for drift and position correction in robotic applications, such as simultaneous localization and mapping. In this paper, we present a sequence based approach for pose estimation, by advancing the well known SeqSLAM algorithm with the usage of Bag of Words (BoW) model. A visual vocabulary is produced in an offline procedure resulting in the system 's ability to describe the incoming image stream by visual words, at the online process. Image similarity is achieved through BoW histogram comparisons instead of sum of absolute differences metric. Comparative results on several publicly-available datasets show the benefits of the proposed method offering high recall scores at 100% precision against the original one.

Keywords: SLAM · Loop Closure Detection · Bags of Words SeqSLAM · Mobile robotics

1 Introduction

As a robot navigates through an unknown area a map is incrementally constructed based on the incoming sensor data. In Simultaneous Localization and Mapping (SLAM) [1,2], place recognition is vital for robust and efficient pose estimation due to drifts that occur by mis-accurate sensor measures or field abnormalities. Recognition of pre-visited places, also known as Loop Closure Detection (LCD) [3,4], can be achieved visually resulting in the system's performance enhancement.

Since the computational power increased in autonomous systems, cameras became a kind of omni-sensor units thanks to the rich information they provide [5]. Towards this transition, the robotics community have moved to approach the development of appearance based mechanisms for LCD; systems that rely their functionality only at cameras' measurements. Such approaches are able to achieve accurate pose estimation when odometry information is noisy.

N. A. Aspragathos et al. (Eds.): RAAD 2018, MMS 67, pp. 580–587, 2019.
https://doi.org/10.1007/978-3-030-00232-9_61

Many global visual recognition systems tackle the LCD task through a greedy image-to-image search for the appropriate match between the query image and the database (pre-visited images). Sum of Absolute Differences (SAD), Bag of Words (BoW) [6] histograms and images' feature vote aggregation are some widespread techniques for images comparison. Counter to global methods, local appearance-based recognition systems address the comparison task between groups of images through the usage of the aforementioned methods. The SeqSLAM approach constitutes one of the most notable algorithms in sequence-based visual recognition systems [7]. The mechanism seeks for the proper candidate match within sequences of images in a sliding window search scheme.

In this paper, we present an improved version of SeqSLAM, based on BoW model, capable of providing accurate LCDs. In an offline procedure, a generic set of training image descriptors, extracted with the ORB [8] detector, are provided as input to a k-median hierachical clustering for the visual vocabulary (VV) production. Afterwards, when the incoming data stream enters the pipeline a quantization procedure takes place on the images' extracted local features. Employing a vocabulary tree [9] the representation of camera measurements by Visual Words (VWs) is achieved. Lastly, the pipeline follows the original version for the proper image selection though a sequence-based sliding window scheme.

Much of the related work, that try to solve the visual place recognition task, makes use of BoW techniques either in a global [10–13] or in a local [14,15] manner. Many researchers extended SeqSLAM: in [16], a faster version of the initial approach is attained with an approximate nearest neighbor technique in the histogram of gradient descriptor space; adding odometry information on the received images, a more efficient real-time SeqSLAM implementation is presented in [17]; in [18], a Bayes filter provides the system with a subset of loop closing candidate sequences. Evaluation on the selected areas proves a faster SeqSLAM system.

The main contribution offered by this paper focuses on the construction of an adaptive SeqSLAM version with robust and efficient results. The mechanism rely its performance on the BoW model for image description. This technique provides the system with orientation invariance properties resulting better performance in situations where the camera's viewpoint change but is disadvantaged in significantly varying environmental conditions (e.g., day-night), since the image's extracted local features are not illumination invariant. Comparative results on several different indoor and outdoor datasets prove the significant system's improvement offering high recall rates for 100% precision.

The rest of the paper is organized as follows: Sect. 2 contains the description of the proposed algorithm in detail. In Sect. 3, the experimental procedure and comparative results are presented while in Sect. 4 conclusions and plans for future work are discussed.

2 Methodology

In this section the proposed approach is presented. As mentioned, the main algorithm is based on the description of the incoming images by VWs. In an offline

process the VV construction takes place, while images' similarity is obtained through histogram comparisons. Following the initial version of SeqSLAM, each query time when an image arrives to the pipeline, the searching functionality seeks for the best sequence match with the database, over a sliding window scheme. An outline of the proposed algorithm is shown in Fig. 1.

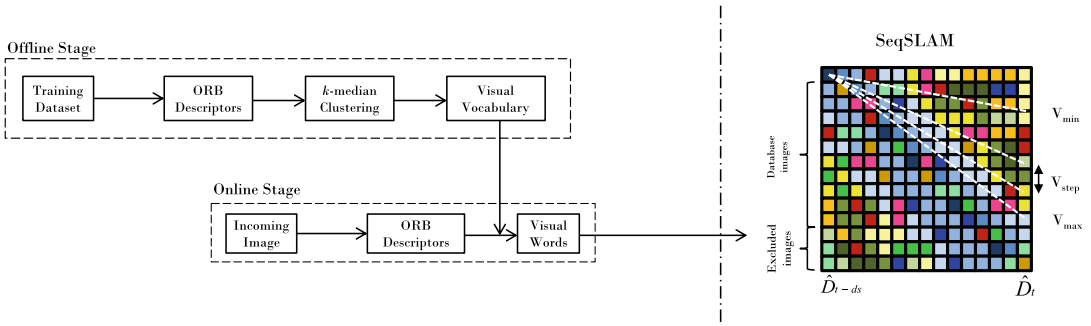

Fig. 1. An overview of the proposed appearance-based algorithm for loop closure detection. During an offline procedure, ORB [8] descriptors are extracted from a training dataset. The accumulated descriptors are used as input to a k-median clustering for the visual vocabulary production. When incoming frames enter the pipeline, local feature descriptors are extracted and converted to visual words. Similarity between images is performed through histogram comparisons. Finally, the approach follows the original SeqSLAM [7] implementation for loop closure identifications.

2.1 Vocabulary Production

Towards VV construction a training dataset is required for the descriptor space quantization. Accumulated ORB local feature descriptors are provided as input to a k-median hierarchical clustering with k-means++ seeding [19]. The Hamming is used as distance metric since the descriptors are in binary format. The vocabulary tree parametrization rely on the conclusions in [12], where the number of levels L and the number of branches B are defined as 6 and 10 respectively. Based on these parameters a total set of $\mathrm{W} = B^L$ discrete VWs is created.

2.2 SeqSLAM

As originally proposed, SeqSLAM converts the incoming data into grayscale equivalents before downsampled to a specific number of pixels. Afterwards, images are normalized in a local area of pixels and, through a SAD technique (Eq. 1), comparisons are performed between them.

$$D_{ij} = \frac{1}{d_x d_y} \sum_{x=0}^{d_x} \sum_{y=0}^{d_y} |\rho_{x,y}^{i} - \rho_{x,y}^{j}| \tag{1}$$

The d_x and d_y represent the downsampled image dimensions, while ρ^i and ρ^j denote the pixel's intensity values for i and j instances, respectively. A difference vector D is produced as it is shown in Fig. 1 (right), containing the distance metrics of image t to the database excluding only the instances close in query time. A contrast enhancement is performed on every element of D_t analogous to 1D version of patch normalization:

$$\hat{D}_t = \frac{D_t - \overline{D_\varepsilon}}{\sigma_\varepsilon} \tag{2}$$

where $\overline{D_\varepsilon}$ denotes the local mean and σ_ε the local standard deviation around database element j. The searching procedure is applied in the produced enhanced matrix $\hat{D}$.

At query time, using a sliding window scheme, the database is searched for the appropriate image match through the usage of a fix-sized sequence length. Different distances are computed for multiple velocities assumptions corresponding to different vehicle's velocities V. For each velocity case a score s is produced:

$$s = \sum_{h=i-w}^{i} \hat{D}_{h,k} \tag{3}$$

where i denotes the query index and w is the window (sequence) size. In Fig. 1 the possible trajectories are represented by the white stripes. The variable k computes the velocity assumptions in $\hat{D}$ and it is calculated as:

$$k = j + V(w - i + h) \tag{4}$$

where V is assigned with multiple values within the range of V_{min}, V_{max}, each time advancing by V_{step}. Afterwards, the minimum score s is selected for the specific database image j. Finally, a vector S containing the scores of query image t against the database ones is generated. The minimum score is selected and normalized to the second lowest one. The final decision for a loop closing match depends on the satisfaction of threshold γ.

2.3 SeqSlam with Bag of Words

The main objective of our approach is the image representation by VWs. To this end as images enter the pipeline the ORB local features are extracted. The produced descriptors are mapped into VWs through the vocabulary tree with the usage of Hamming distance. For a given instance t, a VW histogram is created $V_t = [v_1, v_2, ...v_{VV}]$ based on the widely used "term frequency-inverse document frequency" (tf-idf) technique [20].

Since images are fully described by VWs, the proposed algorithm avoids the usage of SeqSLAM's pre-processing stage where the incoming data are downsampled and normalized before compared to the database. We utilize BoW histogram comparisons to measure images' distances based on the $L2$-norm:

$$L2(v_q, v_d) = \frac{1}{2} \left\| \frac{v_q}{|v_q|} - \frac{v_d}{|v_d|} \right\|_2 \tag{5}$$

where v_q, v_d represent the query and database images' descriptor vectors, respectively. The obtained values are utilized for the creation of the distance matrix, similarly to D of the initial approach. The lower a value between two compared images is, the more similar are the instances.

At this point, the contrast enhancement process is performed as declared by Eq. 2. Trajectories are calculated for each database image j in a fix-sized sequence length. The score vectors S are produced and loop closure candidates are identified in a same manner with the initial version when a factor γ is met.

3 Experimental Evaluation

The following section provides a brief description of the training and tested datasets. Also, the ground truth measurements and comparative results are discussed. Experiments were performed on an Intel i7-6700HQ 2.6 GHz processor with 8 GB RAM using the Matlab based OpenSeqSLAM implementation[1].

3.1 Procedure

See Table 1.

Table 1. Properties of used datasets

Dataset	Description	Image size	Camera pose
Bivosa 2008-09-01 [21]	Indoor & Outdoor, Static	320×240	Frontal
Biccoca 2009-02-25b [21]	Indoor, static	640×480	Frontal
Lip6 Indoor [13]	Indoor, static	240×192	Frontal
Lip6 Outdoor [13]	Outdoor, highly dynamic	240×192	Frontal
Malaga 2009 Parking 6L [22]	Outdoor, slightly dynamic	1024×768	Frontal

Datasets. Publicly-available datasets are selected in a way a variety of environments, camera viewpoints and velocities to be achieved. In the case of Lip6 [13] datasets, the incoming instances are obtained by a hand-held camera representing indoor and outdoor environments. Bivosa (BV) 2008-09-01 [21] and Biccoca (BC) 2009-02-25b [21] datasets are recorded from a robotic platform, while in Malaga (MG) 2009 Parking 6L [22] the incoming image stream retrieved by a camera mounted on a car. Lip6 datasets are preferred as they provide considerable loop closure candidates and many orientation fluctuations. The BC and MG are selected due to many strong perceptual aliasing examples presented in their trajectories. Lastly, BV is preferred as the training dataset since it provides plenty indoor and outdoor scenes making it capable of computing the system's VV.

[1] Implementation of OpenSeqSLAM can be found at http://openslam.org/openseqslam.html.

Ground Truth. A binary matrix whose elements correspond to absence ($GT_{ij} = 0$) or existence ($GT_{ij} = 1$) of a loop closure event at different image timestamps is defined as Ground Truth (GT). At Lip6 datasets the GT is offered by the authors. In cases of BC and MG the used GT was constructed manually in our previous work [23] and utilized without further process.

Precision-Recall. In order to compare the achieved performance between the algorithms the precision and recall metrics are used. Precision is defined as the ratio of the system's detected true positive events over the total mechanism's identifications. Recall is the ratio between the detected true positive loop closure events and the actual loop closures declared by the GT.

3.2 Training Procedure

The VV has to be generic enough so as to accurately describe both indoor and outdoor environments. Towards this aim BV is selected as the training dataset since it includes a variety of different places in the traversed path. By using 10k instances a set of 9M local descriptors were extracted. Afterwards, they were utilized as input to a k-median hierarchical clustering resulting into a vocabulary tree with 10^6 discrete VWs.

3.3 Comparative Results

In Table 2 the algorithm's obtained results for each tested dataset are shown using the precision-recall metrics. The parameters remain the same to the OpenSeqSLAM implementation. The best performing decision factor γ was selected for each approach and dataset. The proposed method outperforms the original one. In BC and MG the system's performance exceeds the initial reaching recall rates of approximate 75% and 80% for 100% precision, respectively. For Lip6 the recall rate is more than twice the performance of the original SeqSLAM in both datasets.

Table 2. Comparison results

Dataset	Approach	Precision (%)	Recall (%)
Biccoca 2009-02-25b [21]	SeqSLAM	100	13.90
	BoW SeqSLAM	**100**	**73.15**
Lip6 Indoor [13]	SeqSLAM	100	20.91
	BoW SeqSLAM	**100**	**54.43**
Lip6 Outdoor [13]	SeqSLAM	100	5.63
	BoW SeqSLAM	**100**	**39.95**
Malaga 2009 Parking 6L [22]	SeqSLAM	100	15.09
	BoW SeqSLAM	**100**	**80.48**

4 Conclusions

In this paper, an improved version of the well-known SeqSLAM algorithm is presented for freely moving cameras. By adopting the BoW model for image representation we address the weakness of the initial version to identify loop closure events due to its similarity metric. Possible camera orientation changes in robot's navigated path highlight the SAD inability to produce satisfactorily results. In an offline procedure, VWs are generated through a k-median clustering on the training set of ORB descriptors. As input data arrive to the system are described by VWs, while the images' comparison is performed through BoW histograms. The integration of BoW model to SeqSLAM advances the system's recall rates for 100% precision in the four tested datasets. Furthermore, in order to extend our work a dynamic sequence segmentation can be assessed offering a more generic solution (Fig. 2).

Fig. 2. Examples of loop closures detections. From left to right: Biccoca 2009-02-25b [21], Lip6 Indoor & Outdoor [13], Malaga 2009 Parking 6L [22].

References

1. Thrun, S., Leonard, J.J.: Simultaneous localization and mapping. In: Handbook of Robotics, pp. 871–889 (2008)
2. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part I. IEEE Robot. Autom. Mag. **13**(2), 99–110 (2006)
3. Ho, K.L., Newman, P.: Detecting loop closure with scene sequences. Int. J. Comput. Vis. **74**(3), 261–286 (2007)
4. Liu, Y., Zhang, H.: Visual loop closure detection with a compact image descriptor. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1051–1056 (2016)
5. Lowry, S., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., Milford, M.J.: Visual place recognition: a survey. IEEE Trans. Robot. **32**(1), 1–19 (2016)
6. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval, vol. 463. ACM Press, New York (1999)
7. Milford, M.J., Wyeth, G.F.: SeqSLAM: visual route-based navigation for sunny summer days and stormy winter nights. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 1643–1649 (2012)

8. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: Proceedings of International Conference on Computer Vision, pp. 2564–2571 (2011)
9. Nister, D., Stewenius, H.: Scalable Recognition with a vocabulary tree. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2161–2168 (2006)
10. Cummins, M., Newman, P.: Appearance-only SLAM at large scale with FAB-MAP 2.0. Int. J. Robot. Res. **30**(9), 1100–1123 (2011)
11. Bampis, L., Amanatiadis, A., Gasteratos, A.: High order visual words for structure-aware and viewpoint-invariant loop closure detection. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4268–4275 (2017)
12. Gálvez-López, D., Tardos, J.D.: Bags of binary words for fast place recognition in image sequences. IEEE Trans. Robot. **28**(5), 1188–1197 (2012)
13. Angeli, A., Filliat, D., Doncieux, S.: Fast and incremental method for loop-closure detection using bags of visual words. IEEE Trans. Robot. **24**(5), 1027–1037 (2008)
14. Bampis, L., Amanatiadis, A., Gasteratos, A.: Encoding the description of image sequences: a two-layered pipeline for loop closure detection. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4530–4536 (2016)
15. Bampis, L., Amanatiadis, A., Gasteratos, A.: Fast loop-closure detection using visual-word-vectors from image sequences. Int. J. Robot. Res. **37**, 62–82 (2017)
16. Siam, S.M., Zhang, H.: Fast-SeqSLAM: a fast appearance based place recognition algorithm. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 5702–5708 (2017)
17. Wang, Y., Hu, X., Lian, J., Zhang, L., Kong, X.: Improved SeqSLAM for real-time place recognition and navigation error correction. IEEE Conf. Intell. Hum. Mach. Syst. Cybern. **1**, 260–264 (2015)
18. Liu, Y., Zhang, H.: Towards improving the efficiency of sequence-based SLAM. In: IEEE International Conference on Mechatronics and Automation, pp. 1261–1266. IEEE, Takamatsu (2013)
19. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
20. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos, p. 1470 (2003)
21. RAWSEEDS: Robotics Advancement through Web-publicing of Sensorial and Elaborated Extensive Data Sets (Project FP6-IST-045144) (2007–2009). http://www.rawseeds.org/rs/datasets
22. Blanco, J.L., Moreno, F.A., Gonzalez, J.: A collection of outdoor robotic datasets with centimeter-accuracy ground truth. Auton. Robot. **27**(4), 327–352 (2009)
23. Tsintotas, K.A., Bampis, L., Gasteratos, A.: Assigning visual words to places for loop closure detection. In: Proceedings of IEEE International Conference on Robotics and Automation (2018)