Towards Robust Loop Closure Detection in Weakly Textured Environments using Points and Lines

Joan P. Company-Corcoles, Emilio Garcia-Fidalgo and Alberto Ortiz
Department of Mathematics and Computer Science, *University of the Balearic Islands* and IDISBA
Palma de Mallorca, Spain
{joanpep.company, emilio.garcia, alberto.ortiz}@uib.es

***Abstract*—SLAM approaches rely on loop closure strategies to correct the inconsistencies of the generated map. These inconsistencies are mainly caused by the effect of sensor noise in odometry sources. For the case of visual SLAM, loop detection typically rely on the repetitive detection and matching of texture-based keypoints. Weakly textured environments, however, can lead to scenes lacking these kind of points and, hence, poor-performing loop detectors. An alternative for these environments is the use of geometrical cues such as line segments, which are frequently present within human-made, structured environments. Under this context, in this work, we introduce a novel appearance-based loop closure detection method that integrates lines and points to enhance performance in these scenarios. For this purpose, we build an incremental Bag-of-Binary-Words scheme for each visual cue to retrieve previously seen images from the two complementary perspectives. Furthermore, we rely on a late fusion strategy to combine the image candidates resulting for both visual vocabularies. An effective mechanism to group similar images close in time is applied next to reduce the effort of the image candidate search. Finally, we propose a novel scheme to validate geometrically the loop candidates, integrating lines into the procedure. The proposed approach compares favourably with other state-of-the-art methods for several datasets.**

I. Introduction

Simultaneous Localization and Mapping (SLAM) addresses the problem of building a map of the environment while, at the same time, localizing the robot within the generated map. These approaches typically depend on loop closure strategies, which, by identifying previously seen places, correct the accumulated position error and re-localize the robot when the tracking system fails. When images are involved in this association procedure, this process is referred to as *appearance-based* loop closure detection [1]–[4].

Many visual SLAM approaches rely on points as visual features [5]. Despite their impressive results in highly-textured scenarios, their performance degrades in weakly-textured environments, where it is typically difficult to find large sets of point features. Under this context, some visual SLAM systems have recently combined points and lines in the loop closure stage [6], [7]. However, these works rely on off-line Bag-of-Words (BoW) models [8]–[10]. This kind of approach requires a pre-training step whenever the environment changes with regard to the available, pre-trained visual vocabulary. To overcome this shortcoming, our proposal adopts an incremental dictionary-based approach [1]–[3], [11], [12] that avoids the pre-training. Furthermore, to solve the unavoidable spatial verification process for loop hypothesis validation, our solution relies only on 2D image data, contrary to other studies that require 3D information supplied by either a stereo camera or a previous mapping process [6], [7].

Summing up, this work proposes a novel appearance-based loop closure detection system that achieves a high number of loop detections by combining points and lines. As commented above, we take advantage of an on-line BoW model, based on binary descriptors [3], [9], [10], which reduces the computational effort and avoids the classical training stage of off-line schemes. Two visual dictionaries, one for each type of visual feature, are maintained. To combine the information obtained from each vocabulary, we employ a late fusion strategy based on a ranked voting system. To conclude, we introduce a novel and faster alternative than the traditional RANSAC method for the spatial verification stage, which is in charge to discard false positives obtained from the visual vocabularies as loop candidates. The proposed loop closing approach is validated using multiple datasets, recorded under different environmental conditions, and it is compared against several state-of-the-art methods.

II. Loop Closure Detection

In this section, we introduce our loop closure detection approach. For a start, we detect keypoints and lines for each sampled image, and next compute binary descriptors for each. These descriptors are then used to obtain a list of the most similar images from each visual vocabulary. The two resulting candidate lists are fused using a ranked voting system which integrates visual similarities from both visual perspectives. To avoid consecutive images to compete between them as loop closure candidates, we group them using the concept of dynamic islands [3], and a representative image of the best island is selected as loop candidate. Finally, this image is assessed geometrically against the query image by using points and lines: if the number of inliers resulting from the spatial verification process is higher than a threshold, the loop is accepted; otherwise it is rejected.

This work is partially supported by EU-H2020 projects BUGWRIGHT2 (GA 871260) and ROBINS (GA 779776), and by projects PGC2018-095709-B-C21 (MCIU/AEI/FEDER, UE), and PROCOE/4/2017 (Govern Balear, 50% P.O. FEDER 2014-2020 Illes Balears). This publication reflects only the authors views and the European Union is not liable for any use that may be made of the information contained therein.

A. Image Description

An image I_t sampled at time t is described as $\phi(I_t) = \{P_t, L_t\}$, being P_t a set of local keypoint descriptors and L_t a set of line descriptors extracted from the image. Point detection and description is performed using ORB [13], while line segments are detected using LSD [14] and described using a binary form of LBD [15]. The set of the m point descriptors found at image I_t is defined as $P_t = \{d_0^t, d_1^t, \dots, d_{m-1}^t\}$, whereas the set of the n line descriptors at I_t are defined as $L_t = \{l_0^t, l_1^t, \dots, l_{n-1}^t\}$. As will be shown later, the combination of these two descriptors enhances the retrieval results in a wider range of scenarios than only using points. This is due to the fact that some environments may be described more distinctively using lines than points (i.e. weakly-textured, structured scenes), or vice versa.

B. Retrieval of Loop Closure Candidates

Loop closure candidates are obtained using *OBIndex2* [3], which combines an incremental Bag-of-Binary-Words (BoBW) scheme jointly with an inverted file to rapidly obtain similar images. *OBIndex2* allows managing efficiently an increasing number of visual words using a hierarchical tree structure. In our proposal, we maintain two instances of *OBIndex2*: one for points and one for lines. When an image I_t is available, its features are used to retrieve the list of the most similar images from the two visual dictionaries: on the one hand, the list of m most similar images using points $C_p^t = \{I_{p_0}^t, \dots, I_{p_{m-1}}^t\}$, and, on the other hand, the list of the n most similar images using lines is $C_l^t = \{I_{l_0}^t, \dots, I_{l_{n-1}}^t\}$. These lists are sorted according to their associated scores $s_p^t(I_t, I_j^t)$ and $s_l^t(I_t, I_j^t)$, which are based on a term frequency-inverse document frequency (tf-idf) scoring scheme. Next, scores are min-max normalized to the range $[0, 1]$ [3], what allows controlling the differences in range caused by the distribution of the visual words on each vocabulary. Finally, we limit the number of candidates per list filtering those images whose normalized score $\tilde{s}_k^t$ is lower than a threshold.

C. Fusion of Lists of Candidates

The next step is to merge the two candidate lists C_p^t and C_l^t to obtain a joint perspective of the retrieved loop closure candidates. To this end, in this work, we rely on a *late* fusion approach [16] by means of a ranked voted system using the *Borda count* [17], a simple data fusion method based on democratic election strategies. In our proposal, a voter is defined for each visual dictionary. Each voter emits an ordered list of candidates C_k^t of different size. The number of candidates c that votes for each set is the minimum length of the two candidate lists. Next, the top-c images on each list C_k^t are ranked with a score b_k defined as:

$$b_k(I_j^t) = (c - j)\, \widetilde{s}_k^t\left(I_t, I_j^t\right) , \tag{1}$$

where j denotes the order of the image I_j in the list C_k^t and $\widetilde{s}_k^t\left(I_t, I_j^t\right)$ is the normalized score of the image in that list. For each image that appears in both lists, a combined Borda score β is computed as the geometric mean of the individual scores using equation 2:

$$\beta(I_j^t) = \sqrt{b_p(I_j^t)\, b_l(I_j^t)}\,. \tag{2}$$

The geometric mean allows us to reduce the influence of false positive image candidates that can appear in one of the lists. The resulting list C_{pl}^t combines thus the information of the two visual vocabularies.

Next, to deal with the fact that some environments mostly exhibit one type of the features, images that only appear in one of the lists are also placed into C_{pl}^t, although penalized by a constant factor. Finally, C_{pl}^t is sorted according to the scores $\beta(I_j^t)$ of all the retrieved image candidates.

D. Computation of Dynamic Islands

An additional temporal consistency verification procedure is next performed to avoid consecutive images to compete among them as loop candidates. To this end, we rely on the concept of *dynamic islands* [3]. A dynamic island Υ_n^m is a group of images whose timestamps range from m to n. A set of islands is built for each image I_t. To build this set, images $I_i \in C_{pl}^t$ are evaluated sequentially. If its timestamp lies in the $[m, n]$ interval, the image is associated to its corresponding island Υ_n^m. If its timestamp does not overlap with any of the existing islands, a new island is created. After processing all images in C_{pl}^t, a global score g is computed for each island as:

$$g(\Upsilon_n^m) = \frac{\sum_{i=m}^{n} \beta(I_i^t)}{n - m + 1}\,. \tag{3}$$

The resulting set of islands Γ_t is sorted in descending order according to g. This global score represents the average of the Borda scores, integrating hence points and lines information from all images associated to an island. Next, a representative island $\Upsilon^*(t)$ is selected among the set of resulting islands to determine which area of the environment is the most likely to close a loop with I_t. For this purpose, iBoW-LCD is based on the concept of *priority islands*. Priority islands are defined as the ones of Γ_t that overlap in time with the island selected at time $t-1$, $\Upsilon^*(t-1)$. In iBoW-LCD, the island finally selected corresponds to the priority island with the highest score g, if any. This selection is only based on the appearance of the images. Nonetheless, in some weakly textured environments, this policy can fail, due to perceptual aliasing, leading to incorrect island associations. To overcome this problem, in this proposal, an island is retained for the next time step only if the final selected loop candidate satisfies the spatial verification procedure, as explained in the next section. When the best island $\Upsilon^*(t)$ is identified, the image I_c with the highest Borda score β of $\Upsilon^*(t)$ is selected and used in the next stage to validate the loop.

E. Spatial Verification

Loop closure detection methods based on BoW schemes are only based on appearance and ignore the spatial arrangement

of the image features, which can result into false detections. To address this problem, a geometric verification procedure is performed to validate the selected candidate I_c. To implement the spatial verification step, RANSAC is typically used through a specific transformation model between mages [1], [3]. Although quite robust, RANSAC is still affected by a large amount of outliers. To minimize this, the Nearest Neighbour Distance Ratio (NNDR) [18] test can be applied before RANSAC to pre-filter certain incorrect matches. However, this test only considers the image appearance and, hence, when using line features, a large amount of correct line matches can be discarded due to the similarity between descriptors. This fact arises particularly in low-textured environments, where a low number of points is detected and lines become the prominent visual feature. New structural matching constraints have been recently introduced, such as Local Geometric Support (LOGOS) [19] or Grid-based Motion Statistics (GMS) [20], to deal with this issue. These methods determine the set of inliers between images without requiring neither RANSAC nor the ratio test. They are based on the existent relationships between local feature neighbourhoods, and, thus, they achieve a higher amount of matches per frame, resulting into a reduction of false positives loop closure detections.

For this work, we introduce an alternative use of GMS to be able to deal with lines. In short, we employ a point representation for each of the two end-points of a line segment, so that a line is regarded as a correct match if GMS accepts one of the two end-points. If the global number of matches produced by GMS is higher than a threshold, then the loop candidate is accepted; otherwise it is rejected. As will be shown in the experiments, this alternative version of GMS offers a good balance between performance and computational times.

III. Experimental Results

This section reports on a set of experiments to validate the proposed approach. We also compare the performance of our approach with other methods of the state of the art. As usual, the evaluation is performed in terms of precision-recall (P-R). To evaluate the combination of points and lines proposed in this approach, we have selected several publicly available datasets of different nature: from weakly-textured scenes, which usually contain more lines than points, to highly-textured scenes with the opposite characteristics, as well as intermediate cases. The datasets considered for the evaluation are: CityCentre [21] (CC), KITTI 00 [22] (K00), KITTI 06 [22] (K06) and Lip6Outdoor [1] (L6O). For each dataset, we use the ground truth from the original authors except for the KITTI sequences, where the ground truth provided by [23] is employed. All experiments were performed on an Intel Core i7-9750H (2.60 GHz) processor with 16 GB RAM.

A. General Performance

Figure 1 illustrates loop closures detected using points, lines, the combination of both and the ground truth for the L6O dataset. As can be observed, the combination of both features

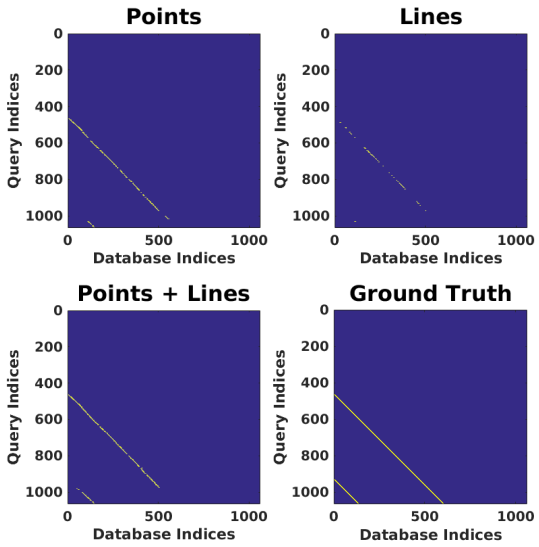

Fig. 1. Loop closure detections found in the L6O dataset using different visual features (Points, Lines, Points + Lines) and the corresponding ground truth. White dots represent a detected loop closure.

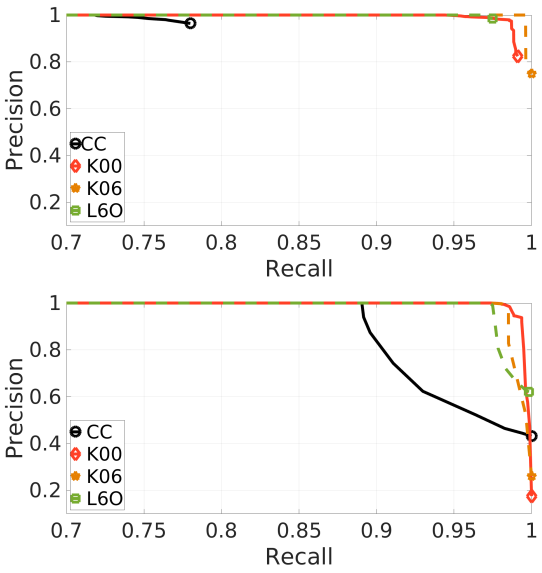

Fig. 2. P-R curves for each dataset. P is 1.0 for all R values lower than 0.7. The proposed loop closure detector is computed using two different spatial verification procedures, GMS (top) and RANSAC (bottom).

increase the number of loop closure detections. However, this combination does not imply increasing the processing time per image in comparison with the use of only points, as in [3]. The average time to process an image for the K00 dataset in iBoW-LCD is 432.38 ms, while for our proposal we need 387.82 ms. This can be attributed to the parallel execution of some parts of our algorithm.

Figure 2 shows the P-R curves obtained for each dataset using either GMS (top) and RANSAC (bottom) as method for the spatial verification procedure. Although GMS does not achieve a recall as high as RANSAC, it is more reliable for a SLAM system where false positives are critical. This fact is observed in the precision axis, where lower values at the max-

TABLE I
AVERAGE TIME (MS) REQUIRED BY GMS AND RANSAC.

	CC	K00	K06	L6O
GMS	8.76	13.18	6.11	7.79
RANSAC	15.07	15.09	14.27	6.14

TABLE II
MAXIMUM RECALL AT 100% PRECISION.

	CC	K00	K06	L6O
Bampis [4]	71.14	96.53	n.a.	58.32
Gálvez-López [9]	31.61	n.a.	n.a.	n.a.
Mur-Artal [10]	43.03	n.a.	n.a.	n.a.
Cummins [8]	38.77	49.2	55.34	n.a.
Gomez-Ojeda [6]	n.a.	75.9	56.9	n.a.
Tsintotas [12]	n.a.	97.50	n.a.	50.0
Tsintotas [11]	n.a.	93.2	n.a.	n.a.
Angeli [1]	n.a.	n.a.	n.a.	23.59
Gehrig [24]	n.a.	93.1	n.a.	n.a.
Khan [2]	38.92	n.a.	n.a.	25.58
Garcia-Fidalgo [3]	**88.25**	76.50	95.53	85.24
Proposed	71.81	**98.82**	**98.88**	**95.54**

imum recall indicate higher false positives. Another advantage of GMS against RANSAC is its reduction in computation time, as can be observed in Table I, where spatial verification times for both cases are shown.

B. Comparison with other Solutions

Table II shows the maximum recall achieved at 100% precision for each dataset. The proposed method is compared to other off-line and on-line approaches. The reported results come from the original works, except for [6], which has been obtained using the default parameters and the visual vocabularies provided by their authors. Not available results are indicated as *n.a.*. The proposed method provides in most cases a higher recall than the other solutions. Furthermore, our proposal outperforms the results reported by [6], which is perhaps the most similar solution to our method.

IV. CONCLUSIONS

This paper introduces an appearance-based loop closure detection method that combines points and lines to achieve a higher number of loop closure identifications, especially in weakly textured environments. This is accomplished by means of a dual BoBW scheme, one for each visual feature, to supply similar images from both perspectives in a fast way. Then, a ranked voting system is used for merging both lists of candidates. To validate the loop candidate hypothesis, we propose a geometrical check stage using a modified version of GMS as main approach, adapted to deal with both points and lines. Experimental results to validate our approach have been reported, showing that our proposal compares favourably against several state-of-the-art methods.

REFERENCES

[1] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "A fast and incremental method for loop-closure detection using bags of visual words," *IEEE Trans. on Robotics*, vol. 24, no. 5, pp. 1027–1037, 2008.

[2] S. Khan and D. Wollherr, "IBuILD: Incremental bag of binary words for appearance based loop closure detection," in *IEEE Int. Conf. on Robotics and Automation*, 2015, pp. 5441–5447.

[3] 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.

[4] L. Bampis, A. Amanatiadis, and A. Gasteratos, "Fast loop-closure detection using visual-word-vectors from image sequences," *Int. Journal of Robotics Research*, vol. 37, no. 1, pp. 62–82, 2018.

[5] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[6] R. Gomez-Ojeda, D. Zuñiga-Noël, F.-A. Moreno, D. Scaramuzza, and J. Gonzalez-Jimenez, "PL-SLAM: a stereo SLAM system through the combination of points and line segments," *arXiv preprint arXiv:1705.09479*, 2017.

[7] X. Zuo, X. Xie, Y. Liu, and G. Huang, "Robust visual SLAM with point and line features," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2017, pp. 1775–1782.

[8] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *Int. Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[9] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[10] R. Mur-Artal and J. D. Tardós, "Fast relocalisation and loop closing in keyframe-based SLAM," in *IEEE Int. Conf. on Robotics and Automation*, 2014, pp. 846–853.

[11] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "Assigning visual words to places for loop closure detection," in *IEEE Int. Conf. on Robotics and Automation*, 2018, pp. 5979–5985.

[12] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "Probabilistic appearance-based place recognition through bag of tracked words," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1737–1744, 2019.

[13] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Int. Conf. on Computer Vision*, 2011, pp. 2564–2571.

[14] R. Grompone von Gioi, J. Jakubowicz, J. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 722–732, 2010.

[15] L. Zhang and R. Koch, "An efficient and robust line segment matching approach based on LBD descriptor and pairwise geometric consistency," *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 794 – 805, 2013.

[16] N. Bhowmik, R. González V., V. Gouet-Brunet, H. Pedrini, and G. Bloch, "Efficient fusion of multidimensional descriptors for image retrieval," in *IEEE Int. Conf. on Image Processing*, 2014, pp. 5766–5770.

[17] S. Y. Jeong, K. Kim, B.-T. Chun, J. Lee, and Y. Bae, "An effective method for combining multiple features of image retrieval," in *IEEE Region 10 Conf. TENCON*, vol. 2, 1999, pp. 982–985 vol.2.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] S. Lowry and H. Andreasson, "Logos: Local geometric support for high-outlier spatial verification," in *IEEE Int. Conf. on Robotics and Automation*, 2018, pp. 7262–7269.

[20] J. Bian, W. Lin, Y. Matsushita, S. Yeung, T. Nguyen, and M. Cheng, "Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Int. Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 2828–2837.

[21] M. Cummins and P. Newman, "FAB-MAP: probabilistic localization and mapping in the space of appearance," *Int. Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Int. Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[23] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, J. J. Yebes, and S. Bronte, "Fast and effective visual place recognition using binary codes and disparity information," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2014, pp. 3089–3094.

[24] M. Gehrig, E. Stumm, T. Hinzmann, and R. Siegwart, "Visual place recognition with probabilistic voting," in *IEEE Int. Conf. on Robotics and Automation*, 2017, pp. 3192–3199.