05-21

16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Invariant Spatial Information for Loop-Closure Detection

Yamamoto Ryohei
University of FUKUI
3-9-1, bunkyo, fukui,
fukui, Japan
yamamotofh26@gmail.com

Tanaka Kanji
University of FUKUI
3-9-1, bunkyo, fukui,
fukui, Japan
tnkknj@u-fukui.ac.jp

Takeda Koji
University of FUKUI
3-9-1, bunkyo, fukui,
fukui, Japan
takedakoji00@gmail.com

Abstract

Recently, Bag-of-Words (BoW) has become a de-facto standard solution for loop-closure detection (LCD) in robotic visual SLAM (rvSLAM). Whereas BoW is efficient in using appearance information as invariant feature for comparing query and mapped scenes, it is not straightforward to use spatial information as invariant feature in BoW. In this paper, we propose a new LCD approach, termed invariant spatial information (ISI), which transforms spatial information in every query/mapped image into an invariant coordinate system by exploiting the available 3D map in rvSLAM. This enables direct comparison of feature location between different images, and allows us to use both appearance and spatial information as invariant feature. Experiments show that the proposed ISI-based LCD outperforms existing LCD methods.

1 Introduction

Loop closure detection (LCD) is one of most fundamental problems in robotic-visual SLAM (rvSLAM). rvSLAM is a technology aiming to build a 3D point cloud map along the robot trajectory, in real-time, with a monocular on-board live camera [1]. LCD is the problem of detecting loop-closure events (i.e., revisiting known parts of the environment), which is crucial for resetting inherently accumulative errors over time, and obtaining a consistent map. There are two key requirements for an LCD solution [2]: (1) High processing efficiency as LCD is a subtask of the time-critical SLAM task, and (2) 100% precision as false loop-closures can lead to catastrophic map damage in a SLAM system.

Bag-of-Words (BoW) is a de-facto standard solution for LCD [3]. Its basic idea is to represent every mapped (or reference) image by an unordered collection of local features, termed visual words, which are then efficiently indexed and retrieved by inverted index. Use of BoW in object and scene retrieval was originally proposed in [4], and further extended to the LCD applications in [5]. Since then, various methods have been developed to improve efficiency and accuracy of the BoW methods. Examples include high-speed feature extraction (e.g., SURF [6], ORB [7], BRIEF [8]), generative model (e.g., FAB-MAP [2]), forming images to places (e.g., island [9]), and incremental vocabulary (e.g., incremental dictionary [10], incremental BoW [3]). Our experimental system is inspired by the state-of-the-art LCD methods, based on incremental dictionary [3] and BRIEF features [8].

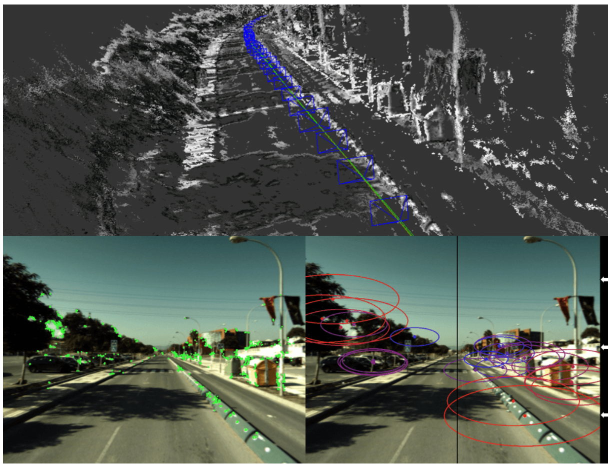


Figure 1. Overview of our approach: Spatial information in every mapped image (bottom left) is transformed into an invariant coordinate system (bottom right) by exploiting the available 3D map in rvSLAM (top).

One of main limitations of BoW is that it ignores the spatial relationships between visual words. In the field of computer vision, it has been reported that spatial information such as scene layout is useful cue for scene recognition and understanding [11]. Such spatial information would be effective also for LCD to differentiate between objects with similar appearance but with different locations, such as poles on the left and right sides of a road. However, implementing such spatial information within a BoW method is not straightforward. This is because a monocular camera does not provide depth information, but it only provides 2D projection of objects onto the image plane. Such 2D projections can be easily affected by viewpoint change and occlusions. Hence, it is often not an invariant feature. Typical BoW methods avoid this problem by simply ignoring any spatial information. However, ignoring such potentially useful spatial information naturally may lead to sub-optimal performance.

In this paper, we propose a new LCD framework which augments visual words with an invariant spa-

© 2019 MVA Organization

tial information (ISI) (Fig.1). The proposed approach modifies both the map-building and LCD processes such that they extract and exploit ISI. In the map-building process, it analyzes the spatial structure of the 3D map being built and extracts useful spatial information *before* the 3D map is compressed into the compact BoW representation. Furthermore, the spatial information is transformed into an invariant coordinate system such that keypoint locations of visual words serve as ISI. In the LCD process, BoW based image retrieval is performed using ISI as an additional cue. We implemented the proposed framework on top of the state-of-the-art LSD-SLAM framework [1] and incremental dictionary framework [3]. Experiments show that the proposed ISI-based LCD outperforms existing LCD methods.

2 Related Work

Recently, BoW has been a hot research focus in the LCD community. In [2], the authors assumed the observation likelihood to be independent of all past observations and approximated it by the Chow Liu tree to capture co-occurrence information for discriminative matching. In [9], the authors developed a robust LCD framework based on database query, match grouping ("islands"), and temporal/geometrical consistency. In [12], the authors presented a versatile and accurate monocular SLAM system, and they further extended it to a keyframe-based Visual-Inertial SLAM that is able to metrically close loops in real-time and reuse the map that is being built online. In [3], the authors presented a novel appearance-based LCD method which makes use of an incremental BoW dictionary based on binary descriptors and the concept of dynamic islands for match grouping, which achieved a high accuracy and outperformed other state-of-the-art methods. However, these existing LCD frameworks do not make use of spatial information, as explained in Section 1. Hence, these approaches are orthogonal to ours and could be used to further boost performance of our LCD method proposed in the current paper.

A very few methods have addressed the issue of using spatial information in LCD. In [13], the authors assumed the availability of image sequence as query input and additionally incorporates the environmental structure into the scene descriptor, by treating bunches of visual words with similar optical flow measurements as single similarity vote. In [14], the authors presented a reliable LCD method for keyframe-based SLAM in urban scenes, which estimates the most salient plane in the live view, converts 3D scenes into orthophoto representations, by which 3D LCD can be re-formulated as an image retrieval problem. In our study, we do not assume the availability of query image sequence nor a-priori knowledge on salient 3D planes, but instead we make use of the 3D map being built online by rvSLAM as prior.

Our ISI-based approach, which uses 3D information in 2D matching, is related to but different from the well-studied computer vision applications of 2d-to-3d matching [15]. Whereas these previous applications typically assumed that an optimal 3D reference model is a-priori built in offline. The current LCD applications do not have separate offline/online processes, and the 3D model must be incrementally built online in real-time. Therefore, typical LCD solutions use just the BoW representations of every query/mapped image, which is much more compact than the raw map data. However, matching such BoW representations poses a significant challenge due to spatial sparseness and information lost in the vector quantization (i.e., image-to-BoW conversion), which is our focus in the current paper.

3 Approach

3.1 The LCD Framework

We suppose a SLAM process runs in parallel to the LCD process. The SLAM process reconstructs a 3D map with trajectory estimation, in real-time, from the sequence of live monocular images. In this paper, we adopt the state-of-the-art direct image-alignment based SLAM, LSD-SLAM in [1], as the SLAM algorithm. However, directly memorizing the entire map in main memory requires a linear cost to the number of images, which is prohibitive in large-size environments. Therefore, we do not memorize the entire 3D map in the main memory. Instead, we extract two kinds of compact information, visual odometry and ISI, from the map and memorize them.

The proposed LCD framework consists of two stages: mapping and localization stages.

The mapping stage aims to build and update an efficient and compact image database, with their viewpoint estimate, in real-time, from live images. First, a collection of BRIEF descriptors [8] is extracted from the current live image. Then, each binary BRIEF descriptor is viewed as a visual word and indexed by the inverted file. Then, ISI (Section 3.2) at each BRIEF keypoint is extracted and assigned to the visual word. Then, each visual word is indexed by the inverted index. Finally, each visual word is checked whether it is already a member of the incremental dictionary [3] and if not, it is inserted as a new visual word to the dictionary.

The localization stage aims to localize and verify location estimates, in real-time, using live images as query. First, a collection of BRIEF visual words are extracted from the current live image, in the same manner as in the mapping stage. Then, a nearest neighbor search over the database is performed using each visual word as query. For the sake of reliability, the Hamming distance of the nearest neighbor BRIEF descriptor is checked by the ratio test with a pre-set ratio threshold of 0.8 as suggested in [16]. Then, correct matches are established by island based match grouping. Then, TF-IDF score is computed from the matches. Finally, the top-ranked match is scored in terms of inlier count by RANSAC post-verification.

3.2 Invariant Spatial Information (ISI)

The key concept of the ISI approach is to transform local feature keypoints into an invariant coordinate system. This enables a direct comparison of the spatial layout between different query and mapped images. Computational cost for such a transformation is negligibly low, as the transformation of each mapped image can be done as a part of the map building process, and thus we only have to transform just one image per query.

The success of the proposed ISI approach depends on the reliability of the invariant coordinate transformation. The transformation should be invariant against small environment variations, originated from dynamic objects, clutters, and the vehicle's trajectories of query and mapped images.

Our ISI approach consists of two distinct steps: (1) We determine the origin of the invariant coordinate system, termed center-of-scene (CoS) (Section 3.2.1). (2) We then localize keypoints with respect to the invariant coordinate system (Section 3.2.2).

3.2.1 Invariant Coordinate (IC)

The invariant coordinate system is simply represented by a single dominant invariant 2D landmark point on the u-v image plane, termed center-of-scene (CoS). Then, the displacement Δu of the landmark from the original image center along the horizontal axis u is computed. Then, the transformation is defined as a mapping which shifts each keypoint in the image by Δu along the horizontal axis.

In this way, the coordinate and transformation are simply represented by 2D and 1D parameters. Such a simple representation is relevant to many mobile robotics applications including autonomous driving, in which the vertical displacement Δv is often not so significant compared with the horizontal displacement Δu. Such a low-dimensional localization is more robust than high-dimensional ones from the perspective of recognition performance.

In this study, we try to detect and use a vanishing point (VP) in the scene as CoS (black vertical line in the bottom right panel of Fig.1). To extract a VP, the input image is cropped by eliminating upper/bottom 20% image regions and then processed by the VP detection algorithm in [17].

Note that such a coordinate transformation comes with risk. As shown above, the coordinate transformation relies on the success of CoS estimation, which is a high-level pattern recognition problem and whose solution is far from perfect. When the CoS estimation fails, the transformation will do more harm than good. Therefore, we need to take into account not only effect but also risk of using ISI. In this paper, we track the detected VP over frames and if Euclidean distance in 2D VP location between the current and previous frame's VPs is larger than a pre-set threshold T_{vp}, we simply do not use the ISI strategy for such a scene. T_{vp} is set 20% of the image width.

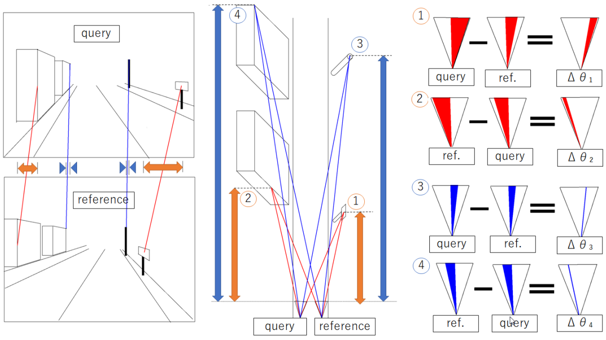


Figure 2. Invariant Spatial Information.

3.2.2 Spatial Matching (SM)

Note that the availability of 3D information is very different between query and mapped viewpoints. For the mapped viewpoint, its surrounding 3D regions are often already mapped. Hence, their rich 3D information can be provided by the map. For the query viewpoint, its surrounding 3D regions are often not mapped yet or currently under reconstruction. Based on the consideration, our LCD approach assumes 3D information is available only for mapped images.

Given 3D information for the mapped image, each keypoint in the mapped image can be localized in a local region R in the query image, which serves as a matching region (MR) for the keypoint. For simplicity, we model this MR as a circular region centered at the keypoint location (colored circles in the bottom right panel in Fig.1). Fig.2 illustrates MRs for four objects with different depth for query and reference images. As can be seen the radius r of the circular region is a function of the depth d at the keypoint, and it is inverse proportional to the depth at the feature keypoint that is predicted from the point cloud: $r = w/20d^{-1}$, where w is the image width. Finally, the top-ranked match is scored in terms of inlier count by RANSAC post-verification.

Any keypoint whose distance exceeds 90% of the maximum range of the distance function are regarded as unreliable and not used for matching. We term the above spatially-constrained matching strategy as spatial matching (SM).

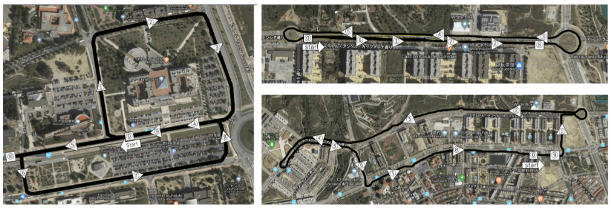


Figure 3. Experimental environments of dataset A (left), B (right bottom), and C (right top).

3.3 Scoring

As mentioned, we have introduced two differing ISI-based methods: invariant coordinate (IC) and spatial matching (SM). These two methods usually provide different scoring results. The question is how to fuse these different scoring results to obtain a final decision. To address this issue, we loosely follow the idea of weighted-sum score in spatial pyramid matching (SPM) [11], and we re-define the score function as a weighted sum of scores with and without IC/SM:

$$S = \sum_{f \in F_0} S_{tfidf}(f) + \sum_{f \in F_1} \frac{1}{A(f)} S_{tfidf}(f). \qquad (1)$$

F_0 is the entire feature set of the query image, and F_1 is the subset whose keypoints are located with the MR R. The first and the second terms of the Eq.1 are similar in concept with as the level 0 and 1 similarity scores in SPM. $S_{tfidf}(f)$ is the TF-IDF score of the feature f. $A(f)$ is area [pixels] of the feature f normalized by the image area [pixels]. Accordingly, the inlier count for RANSAC is computed by adding inlier counts for F_0 and F_1. Unlike F_0, F_1 tends to be a sparse feature set, and if it is very sparse, it is not reliable. To address this issue, we evaluate the number of level-1 features that passed SM $N_F = |F_1|$ and the number of relevant reference images that passed SM N_I, and compute average number of level-1 features per relevant reference image N_F/N_I. Then, if the number of feature matches is smaller than $10 \times N_F/N_I$, then we ignore the TF-IDF score and the inlier count score for F_1.

4 Experiments

In this section, our ISI approach is compared with previous LCD methods in terms of precision and recall. We compare our approach with three representative LCD methods: FAB-MAP [2], DLoopDetector [9], and iBoW-LCD [3]. The first is the probabilistic LCD approach introduced in Section 1. The second is the island-based LCD approach. The third is the recently-developed LCD approach based on dynamic island and incremental dictionary.

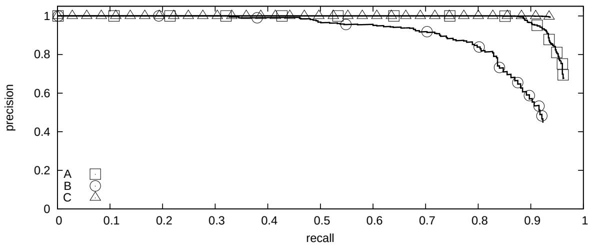
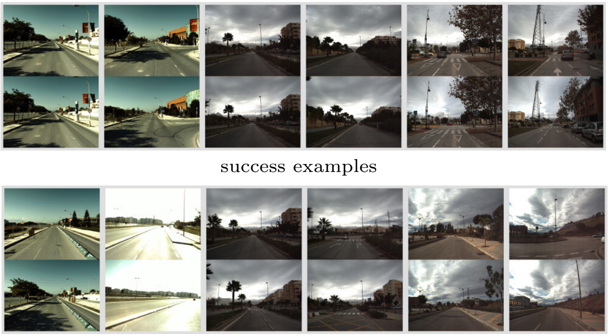


Figure 4. Precision recall curves.



success examples

failure examples

Figure 5. Examples of detected loop-closures. (Top: query images. Bottom: mapped images.)

We augment Malaga dataset (“#5”, “#8” in [18], and “CAMPUS-2L” in [19]) with ground-truth loop-closure annotations, as our task aims at LCD, which are respectively termed dataset A, B, and C. These datasets are gathered entirely in urban scenarios with a car equipped with several on-board sensors (Fig.3). We use left images of the front-facing on-board stereo camera, as monocular input images to our SLAM and LCD processes. Image size is 1,024×678. We use three different sections of datasets: “CAMPUS-2L” [19], “#8” and “#5” in [18], each of which consists of image sets with size 4,675, 10,026, and 4,816, and corresponds to travel distances 2,000m, 5,000m, and 2,300m. We manually analyzed the GPS information available in the dataset and annotated viewpoint pairs whose distance is less than 20m as ground-truth loop-closures.

Following the literature, we evaluate LCD performance in terms of recall@100%precision. First, loop-closure predictions output by an LCD method of interest are merged over all the query viewpoints and sorted in the descending order of RANSAC score. Then, precision values at all the 100% precision points are computed in the sorted list. Then, the highest of them is output as recall@100%precision. Fig.4 shows precision-recall curve for the proposed approach on the three datasets.

Tab.1 shows performance results. It can be seen

Table 1. Performance results.
(recall@100%precision)

method	A	B	C
FAB-MAP [2]	53.5	7.3	16.9
DLoopDetector [9]	85.8	17.8	26.1
iBoW-LCD [3]	85.1	26.8	72.5
ISI w/o SM	87.2	19.3	55.2
ISI w/o IC	84.4	37.1	84.7
ISI	86.7	40.2	82.5

that the proposed ISI method outperforms the other methods for all the datasets considered here. In addition, performance of the ISI method is better when the strategies SM and IC are used than when they are not used. As expected, iBoW-LCD performed best among the other comparing methods. However, the proposed method achieved higher performance, mainly owing to the fact it was often successful in filtering out false positive loop-closures. Overall, the strategy IC was effective in improving true positive detection by transforming features to the invariant coordinate system, while SM was effective for suppressing the false positive detection.

Fig.5 shows examples of detected loop-closures. As shown in the figure, a main source of failure was rapid appearance change of scenes which led to inconsistent loop-closure hypotheses.

5 Conclusions

We presented a method for loop closure detection based on an invariant coordinate system and invariant spatial information (ISI). We showed its efficacy on three benchmark datasets and provided results outperforming existing state-of-the-art approaches. We plan to extend our work to incorporate more semantic information for an improved ISI, as well as learning of parameters to accommodate different ISI strategies.

Acknowledgement

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 26330297, and (C) 17K00361.

References

[1] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014.

[2] Mark Cummins and Paul Newman. Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011.

[3] Emilio Garcia-Fidalgo and Alberto Ortiz. ibow-lcd: An appearance-based loop-closure detection approach using incremental bags of binary words. *IEEE Robotics and Automation Letters*, 3(4):3051–3057, Oct 2018.

[4] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.

[5] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[7] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. ORB: an efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision, ICCV*, pages 2564–2571, 2011.

[8] Michael Calonder, Vincent Lepetit, Mustafa Özuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua. BRIEF: computing a local binary descriptor very fast. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1281–1298, 2012.

[9] Dorian Gálvez-López and J. D. Tardós. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, October 2012.

[10] Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Trans. Robotics*, 24(5):1027–1037, 2008.

[11] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[12] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robotics*, 31(5):1147–1163, 2015.

[13] Loukas Bampis, Angelos Amanatiadis, and Antonios Gasteratos. High order visual words for structure-aware and viewpoint-invariant loop closure detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4268–4275, 2017.

[14] Fabiola Maffra, Lucas Teixeira, Zetao Chen, and Margarita Chli. Loop-closure detection in urban scenes for autonomous robot navigation. In *2017 International Conference on 3D Vision*, pages 356–364, 2017.

[15] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 2599–2606. IEEE, 2009.

[16] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[17] A. Dhall* and Chandak* Y. Vanishing Point Detection using Least Squares. https://github.com/ankitdhall/Vanishing-Point-Detector, 2015.

[18] José-Luis Blanco-Claraco, Francisco-Ángel Moreno-Dueñas, and Javier González-Jiménez. The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario. *The International Journal of Robotics Research*, 33(2):207–214, 2014.

[19] José-Luis Blanco, Francisco-Angel Moreno, and Javier González. A collection of outdoor robotic datasets with centimeter-accuracy ground truth. *Autonomous Robots*, 27(4):327–351, November 2009.