


Viewpoint-tolerant Place Recognition combining 2D and 3D information for UAV navigation

Conference Paper**Author(s):**

Maffra, Fabiola; Chen, Zetao; [Chli, Margarita](#) 

Publication date:

2018

Permanent link:

<https://doi.org/10.3929/ethz-b-000249607>

Rights / license:

[In Copyright - Non-Commercial Use Permitted](#)

Originally published in:

<https://doi.org/10.1109/ICRA.2018.8460786>

Funding acknowledgement:

157585 - Collaborative vision-based perception for teams of (aerial) robots (SNF)

644128 - Collaborative Aerial Robotic Workers (SBFI)

Viewpoint-tolerant Place Recognition combining 2D and 3D information for UAV navigation

Fabiola Maffra, Zetao Chen and Margarita Chli
Vision for Robotics Lab, ETH Zurich, Switzerland

Abstract—The booming interest in Unmanned Aerial Vehicles (UAVs) is fed by their potentially great impact, however progress is hindered by their limited perception capabilities. While vision-based odometry was shown to run successfully onboard UAVs, loop-closure detection to correct for drift or to recover from tracking failures, has so far, proven particularly challenging for UAVs. At the heart of this is the problem of viewpoint-tolerant place recognition; in stark difference to ground robots, UAVs can revisit a scene from very different viewpoints. As a result, existing approaches struggle greatly as the task at hand violates underlying assumptions in assessing scene similarity. In this paper, we propose a place recognition framework, which exploits both efficient binary features and noisy estimates of the local 3D geometry, which are anyway computed for visual-inertial odometry onboard the UAV. Attaching both an appearance and a geometry signature to each ‘location’, the proposed approach demonstrates unprecedented recall for perfect precision as well as high quality loop-closing transformations on both flying and hand-held datasets exhibiting large viewpoint and appearance changes as well as perceptual aliasing.

Video—https://youtu.be/8Vkr_nSbR34

Datasets—<http://www.v4rl.ethz.ch/research/datasets-code.html>

I. INTRODUCTION

With small Unmanned Aerial Vehicles (UAVs) sparking great interest for a plethora of potential applications ranging from digitization of archaeological sites to search-and-rescue, there has been an increasing body of research dedicated in automating their navigation. As Spatial understanding forms the basis of autonomous robot navigation, a variety of techniques for robotic egomotion estimation and map building that perform SLAM (Simultaneous Localization and Mapping) have been proposed in the literature. In addition, addressing place recognition by determining whether a robot returns to a previously visited place is a key competence to enable the creation of accurate maps, relocalization and even collaboration between different robots performing SLAM, essentially opening up the way towards long-term operation of robotic platforms in real world scenarios. However, the agility and portability of small aircraft comes at the cost of small payload and as a result, limited computational capabilities. Current solutions involve restricting the onboard memory of past experiences by limiting the size of the SLAM map (e.g. as in [1]). As small estimation errors are usually

This research was supported by the Swiss National Science Foundation (SNSF, Agreement no. PP00P2_157585) and EC’s Horizon 2020 Programme under grant agreement n. 644128 (AEROWORKS).

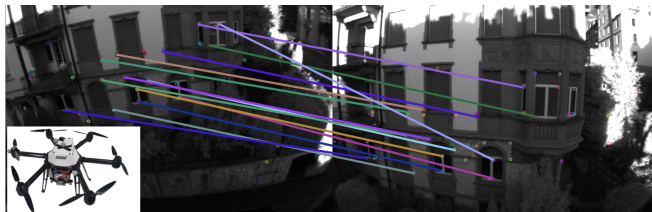


Fig. 1: A loop in the UAV dataset correctly detected by the proposed approach, despite the large viewpoint change and the mismatches caused by repetitive scene structure. This challenging dataset was captured with the UAV in the inset and it is being made publicly available.

accumulated over time, restricting the estimation process to a limited window accentuates the problem of drift even further, highlighting the need for suitable place recognition techniques. Moreover, the agility and dynamicity of UAV manoeuvres pose particular challenges in place recognition, as the same place needs to be estimated from very different viewpoints.

Inspired by the challenges of place recognition from aerial imagery, in this paper, we present a scalable framework to identify loop-closures in a robot’s trajectory using low cost, binary features suitable for UAV navigation. As UAV navigation is one of the hardest scenarios for place recognition, the portability of the proposed method to other platforms (e.g. a ground robot) with simpler motion and computational constraints should be straightforward. Moreover, while the vast majority of works in this domain restrict their operation to a decision whether there has been a loop closure or not, here, we go a step further to accurately estimate the transformation between the matching robot poses, which can be directly used in a subsequent optimization step. Designed to be interfaced with a keyframe- and vision-based odometry system, the proposed pipeline is shown to outperform the state of the art on both indoor, aerial sequences evaluated on ground-truth data from a highly accurate tracking system (i.e. Vicon), as well as outdoor hand-held and aerial urban sequences against GPS position information. To encourage further research and benchmarking in viewpoint-tolerant place recognition our challenging datasets are being made publicly available. Figure 1 illustrates an example of a successful loop detected by the proposed approach designed to cope with large viewpoint changes and perceptual aliasing. The main contributions of this work are:

- a new, carefully designed place recognition pipeline especially developed for robot navigation, which avoids

false positive loop closures at all costs, exhibiting robustness to viewpoint changes, and

- new datasets with visual-inertial information and manually-annotated ground-truth capturing viewpoint, illumination and situational changes, suitable to test place recognition approaches.

II. RELATED WORK

Place recognition, also referred to as loop closure detection, is usually addressed using appearance-based cues. Typically, two main tasks must be accomplished to address place recognition: (a) query a database of images to find possible similar locations and then (b) determine which, if any, of these images represents the same place as the query. Identifying whether a robot is revisiting a place by directly matching a query image to all images into a database containing its previously visited locations is very inefficient. For this reason, either a Bag of Words approach (BoW) approach [2] with an inverted file index or a descriptor voting scheme [3] are usually applied in the first task followed by a geometric consistency check in the latter one. The widely known BoW approach relies on discretizing the space of feature descriptors generated using a set of training images to build a dictionary of visual words and then representing new images as a set of visual words it contains. Several well-performing algorithms using a BoW representation were proposed in the literature, with FABMAP [4] considered to be one of the most successful pipeline for place recognition.

A less popular approach is to consider global image representations, instead of traditional feature-based representations. In PTAM [5] for example, a smaller and blurred version of the original keyframe image was used as a descriptor of a place, which implies that for relocalization (i.e. loop-closure detection) an exhaustive search across the entire database of images is necessary to identify a potential correlation match. SeqSLAM [6] has demonstrated very impressive recall rates on scenes with dramatic changes in lighting (day/night), however, the method still lacks invariance (e.g. in viewpoint) and relies on using long sequences of images to tackle perceptual aliasing of the query location. Moreover, the scalability of such methods is more limiting than with feature-based BOW approaches, where indexing and searching for matches can be done more efficiently.

More recently, Convolutional Neural Networks (CNNs) have been successfully applied to solve the place recognition problem under extreme changes in appearance (e.g. time of the day, weather, seasons as well as human activity and occlusions). While [7] and [8] train a CNN to learn a compact image representation suitable to place recognition, another common strategy cast the place recognition problem as a classification task [9], [10]. While impressive results have been obtained by using deep learning techniques, this approach still very computational expensive. While efforts to reduce the computational complexity exist [11], place recognition using deep learning remains unsuitable for real-time estimation onboard a small UAV with small payload and limited computational capabilities.

Place recognition onboard a small UAV is a particularly challenging problem; the dynamicity and agility of a small UAV means that it is very likely to approach the same scene from a wide range of viewpoints, which is by definition fatal for global image-representation techniques, while feature based BoW approaches also struggle greatly. This is inherently a very different problem from the traditional place recognition on a car in the streets of a city as addressed in [4] and [6]. The need for unique and repeatably recognizable features is all the more important in order to allow viewpoint-invariant recognition. As a result, current methods choose to work with the highest quality of feature detectors and descriptors, such as SIFT [12] and SURF [13]. These features, however, are typically far too expensive to employ onboard a small UAV, which renders most of the existing place recognition techniques unusable.

Interestingly, features with binary descriptors, such as ORB [14], BRISK [15] and FREAK [16] promise similar matching performance to SIFT or SURF at a dramatically low computation, however, it becomes far more difficult to cluster them into visual words in a BoW approach. The work in [17] was the first in the literature to use binary features for place recognition, however, the precision-recall characteristics of this method still very sensitive to noise.

Another interesting line of research that has recently appeared makes use of learning techniques to overcome the large viewpoint differences from ground to aerial images. While these wide baselines are not usually addressed in place recognition systems, novel algorithms to air-ground matching have been proposed in complementary areas [18], [19], [20]. Despite the impressive aforementioned algorithms, we still lack a robust solution that overcomes the large viewpoint differences between images captured from a UAV, while keeping onboard computation affordable for a long-term place recognition system.

Finally, while most of the place recognition systems ignores the underlying structure and geometry between features when comparing features sets, a handful of works have investigated how to incorporate some geometric information in their location models, such as in [21], where locations are represented by both visual landmarks and a distribution of the distances between them in 3D coming from range-finders or stereo cameras. Instead of relying on additional sensors to obtain 3D landmark positions, in [22] landmarks are tracked between successive images using a single camera, recording the binary covisibility between landmarks in a graph-based map of the world. In the general case, the graph matching problem in undirected graphs is an NP-hard problem. As a result, there are still open questions on how such techniques can be efficiently and sufficiently approximated to provide the robustness necessary for place recognition for a UAV.

III. METHODOLOGY

The proposed framework is designed to be employed within the loop of robot navigation, so we assume that a vision-based SLAM/odometry system using a keyframes paradigm runs on a separate thread. A hierarchical Bag of

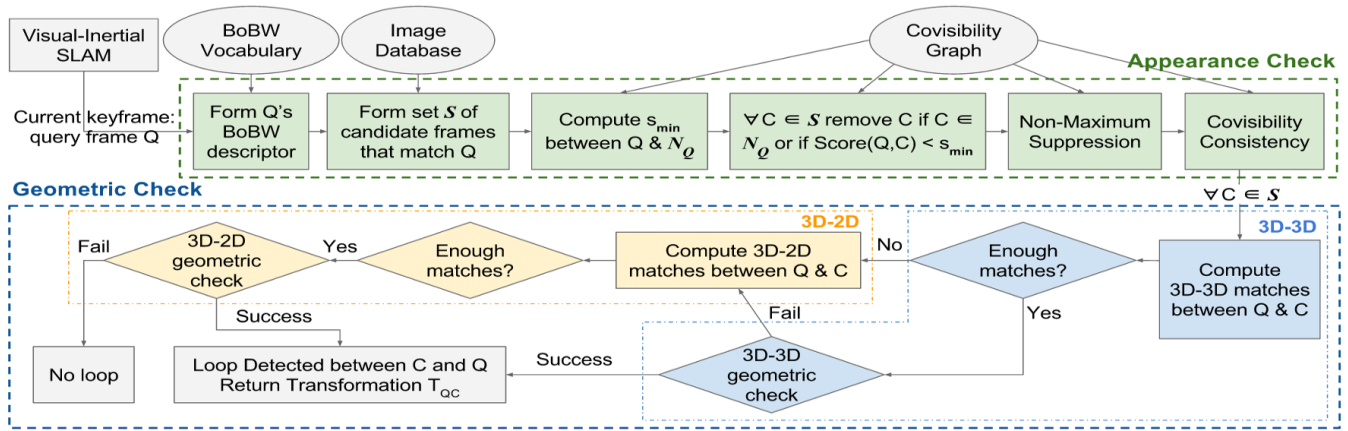


Fig. 2: The proposed Place Recognition Pipeline, first queries the current keyframe Q for an appearance match consulting the BoBW vocabulary, the database of visited keyframes (i.e. known ‘locations’) and the keyframes’ covisibility graph maintained by the SLAM system. If Q appears similar enough to a database image, the candidate matching keyframes are checked for similarity in their geometry of features both in 3D (local map space) and in 2D (image space).

Binary Words (BoBW) visual vocabulary is formed in binary descriptors’ space with an inverted file index to efficiently query at runtime, the database of keyframes captured during the robot’s trajectory for loop-closures. The workflow comprises of two consecutive checks as illustrated in Figure 2; an Appearance Check making use of the keyframe-covisibility information captured by SLAM refines and removes erroneous loop-closure candidates suggested by the BoBW descriptors, before a Geometric Check tests for matches in the configuration of features (in 3D and in 2D) in the candidate keyframe matches that survive the Appearance Check. A successful Geometric Check denotes loop closure detection; in this case the system does not only provide the matched keyframes, but also the best rigid transformation found between them.

A. Visual-Inertial Keyframe-based SLAM

With the ultimate goal of place recognition for a UAV, we assume that a nominal monocular-inertial SLAM system is running in the background, as this is a widely accepted sensor setup for small aircraft with limited payload [1], permitting absolute scale estimation. The proposed system, however, is agnostic to the keyframe- and vision-based SLAM system to be used (i.e. no inertial sensing is necessary). In this work and throughout our experiments, we employ the open-sourced OKVIS visual-inertial SLAM/odometry framework of [23], [24], while we have developed a Covisibility Graph data-structure similarly to [25], where any two keyframes (nodes) share an edge if they share enough 3D landmarks. This approach is more adaptive than choosing a fixed number of consecutive images to represent a location. As SLAM keyframes can provide both the detected features (in this case BRISK [15]) in image space and the local 3D map, a new entry is created in the Image Database for every new SLAM keyframe. Each such entry comprises of an appearance signature of the corresponding keyframe, namely its BoBW descriptor and a geometry signature that is the local, sparse 3D map of keypoints that this keyframe has

been associated with.

B. Building the BoBW Visual Vocabulary

Opting for a hierarchical visual vocabulary [17], the proposed method describes an image as a collection of words combined with an inverted file index allowing efficient retrieval in a large database of images. While features, such as SURF [13] and SIFT [12] are well-established and known to provide stable detection and high quality description of corresponding image areas, their prohibitively high computational cost has been driving research in robotic visual perception towards the computationally cheaper binary-descriptor alternatives, such as BRISK, ORB [14] and BRIEF [26]. Clustering binary instead of floating point descriptors, however, to form visual words is still subject to research, as a bit flip could potentially change a descriptor’s mapping to the words space, resulting to low word repeatability and thus, violating one of the basic assumptions of the bag-of-words approach.

Following the approach suggested by Galvez and Tardos [17], we create a visual vocabulary adapted to the binary features used by OKVIS, namely BRISK [15], effectively reusing any features extracted in the loop of SLAM. The aim here is to exploit any scale and rotation invariance offered by BRISK. It should be noted that the descriptor size used within OKVIS consists of 48 bytes (instead of 64 as in the original implementation) and the feature orientation is aligned with the gravity since the inertial sensor provides this. To compute the BoBW vocabulary we discretize the 48-byte BRISK descriptors’ space using about 3500 training images in total. These depict indoor and outdoor environments and are different to the ones used at runtime. The resulting vocabulary tree has 10 branches and 6 depth levels resulting to a vocabulary of a million words.

C. Appearance Check

The first step to place recognition is to check the current query keyframe Q against the Image Database for any entry

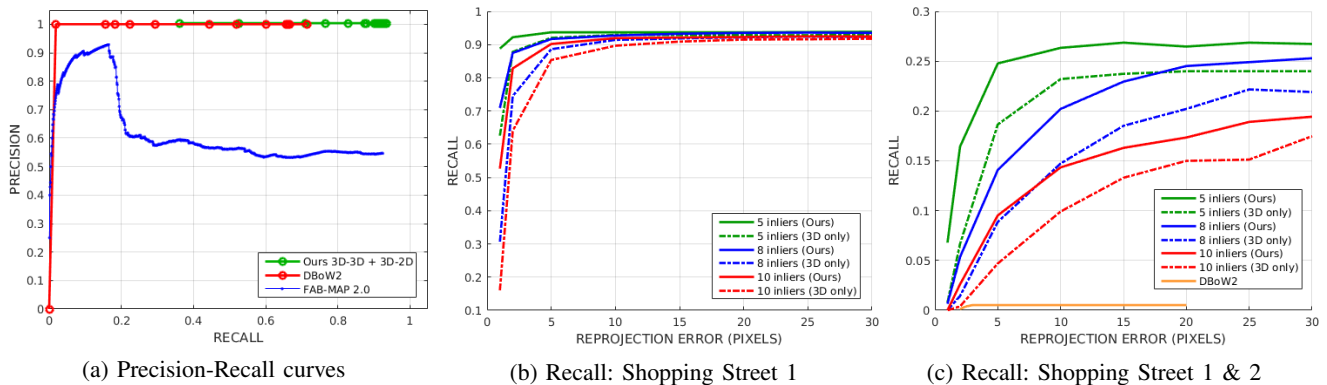


Fig. 3: Precision & Recall analysis. Testing in Shopping Street 1, in (a) the proposed method outperforms FAB-MAP 2.0 and DBoW2. Maintaining perfect precision, in (b) and (c), recall is monitored for variable reprojection error thresholds for the proposed method in full (i.e. using all geometric checks) and using 3D checks only. Accepted inliers are varied from 10 (very restrictive) to 5 (most relaxed). Even in the more challenging dataset used in (c), the proposed method outperforms the 3D only approach and DBoW2 by a large amount.

with similar appearance. To this end, the BoBW descriptor of Q is scored based on its L1-distance to Database entries, using a ‘term frequency–inverse document frequency’ (tf-idf) weighting scheme [27] to suppress commonly occurring words to form the set \mathcal{S} of matching keyframe candidates. Following the approach of [25], the set \mathcal{N}_Q of immediate neighbours of Q in the Covisibility Graph (i.e. depicting common scene structure) is formed, recording the minimum similarity score s_{min} between Q and any member of \mathcal{N}_Q computed as the L1-distance between their BoBW descriptors. Any candidate matching keyframes in \mathcal{S} that score lower similarity to Q than s_{min} or already belong to \mathcal{N}_Q are removed from \mathcal{S} . All remaining members of \mathcal{S} undergo non-maximum suppression within their immediate neighbourhood in the Covisibility Graph; all members of such a covisibility group are scored for their similarity to Q and the corresponding entry in \mathcal{S} is replaced with the highest scoring keyframe in each group. If the sum of the highest N scores in one such group does not reach at least 75% of the best score across all groups, the corresponding entry in \mathcal{S} is removed entirely. Finally, every surviving candidate in \mathcal{S} is checked for covisibility consistency with at least 3 candidate matches surviving the last Appearance Checks (i.e. corresponding to the two previous query keyframes). Two keyframes are defined to be covisibility-consistent if their covisibility groups share at least one keyframe. This last step aims to eliminate candidates in \mathcal{S} that do not share similar appearance with the previous query keyframes.

D. Geometric Check

The BoW approach discards all spatial information between visual words by definition, accepting as a match two different images having the same words regardless of their constellation. While in ground robot navigation scenarios this might be enough [4], in UAV navigation, where very different viewpoints are expected, geometric verification of an appearance match is imperative. Moreover, while traditionally, place recognition techniques stop short of

estimating a relative transformation between the matching frames (e.g. this would be enough in image retrieval), in robot navigation, this information constitutes very useful input to a subsequent optimization step to enforce the loop closure that is detected and avoid local minima. Realising this, [25] implement a geometrical validation step employing the Horn method [28], which given two sets of 3D map points with known correspondences, estimates a 3D rigid transformation between them if enough inliers are found. However, for dynamic camera motion with large viewpoint changes, SLAM systems struggle to find enough correct 3D map points needed for a successful Horn test resulting to much fewer loops detected than actually experienced.

The first priority in place recognition is to avoid false positive loop detections, however, false negatives become of particular interest in viewpoint-challenging cases as they occur far more commonly than in any other scenario, effectively limiting our ability to correct for accumulated drift. In this spirit, here we propose to first use the 3D-3D Horn’s geometric verification and if this proves unsuccessful, check for a 2D-3D geometric consistency using the method of [29]. This provides a closed-form solution to the Perspective-Three-Point (P3P) problem for the full transformation between two camera poses in the world reference frame using at least three 2D-3D point correspondences.

For every keyframe candidate C (member of \mathcal{S}) to match Q that reaches the Geometric Check we compute the BRISK correspondences between them, limiting the correspondence search only to the keypoints that have a 3D landmark associated with them. Erroneous correspondences are removed using a second Nearest Neighbour (2nd NN) test [12], while we also apply bidirectional matching to discard ambiguous matches. If enough 3D-3D correspondences are found, we attempt to verify the 3D-3D geometry between Q and C by estimating their rigid transformation T_{QC} using Horn within a RANSAC scheme. However, if this approach fails to estimate a transformation with at least N inliers the 2D-3D geometry verification is attempted. In order to expand

the set of correspondences to consider, the 2D keypoints in Q are tested for matches with the image projections of all 3D landmarks present in C , following the strict bidirectional and 2nd NN tests. If enough 3D-2D correspondences are available we use the P3P method of [29] in a RANSAC scheme to try to estimate T_{QC} . If a transformation that satisfies a minimum threshold on the average reprojection error in pixels is found, C is accepted as a loop closure for Q . After looping through all the candidates in S for a Geometric Check, the proposed method returns the T_{QC} with the highest number of inliers (i.e. points with a reprojection error is smaller than a pre-defined threshold) and the corresponding C . For our tests we usually define this threshold to be smaller than 2 pixels, the minimum number of matches as 12 and the number of inliers to accept a loop as 8.

IV. DATASETS

While datasets containing outdoor visual and inertial information, such as KITTI [30] exist, they are typically unsuitable to evaluate place recognition methods on. In KITTI for example, most sequences exhibit mainly forward camera motion with a front-looking camera, rendering it very difficult to correctly label the images for ground truth. For this reason, the datasets used in this work were recorded especially for place recognition applications using both flying and hand-held setups in the city center of Zurich with a side-looking camera, permitting clear decisions on ground truth labelling. These manually labelled datasets are being made publicly available, given that there are no other public datasets suited to place recognition providing ground truth, visual and inertial data as well as posing viewpoint and situational challenges as described below.

While we use our recorded datasets to assess the quality of the proposed pipeline in deciding whether the camera's trajectory experiences a loop closure, in order to test the quality of the proposed transformation, we use the publicly available EuRoC Micro Aerial Vehicle (MAV) dataset [31] providing indoor visual and inertial data from a flying UAV, which has its poses recorded by a Vicon external tracking system, providing very accurate full pose information. All the datasets in this work were recorded with a Visual-Inertial (VI) sensor [32] providing grayscale global-shutter images at 20 Hz and synchronized inertial measurements. In our tests, we perform monocular-inertial estimation by using only the information provided by one of the cameras of the sensor.

A. Shopping Street 1 and 2

These two datasets were recorded in a busy shopping street in the city center of Zurich using two different configurations. Shopping street 1 uses a hand-held setup, while Shopping Street 2 was recorded months later in the same area using a 4m-long rod held vertically in order to capture the same scene from very different viewpoints. Shopping Street 1 consists of two traverses in the same street exhibiting small viewpoint changes, perceptual aliasing and appearance changes. We combine both sequences Shopping Street 1 and 2 obtaining a challenging dataset for place recognition,



Fig. 4: Example loop-closures from the Shopping Street dataset tested with the proposed approach. The loop-closures in (a) demonstrate robustness of the proposed approach to viewpoint changes and small motion blur (bottom left). In (b), the top image is an example of a loop detected across the Shopping Street 1 and 2 sequences exhibiting big changes in viewpoint and scene appearance, while the bottom image depicts a false negative, where the viewpoint and illumination changes proved too large for a loop-closure match.

with major changes in the scene appearance, challenging lighting conditions and also strong viewpoint variations. Examples are shown in Figure 4. These sequences were already successfully applied in a place recognition scenario in our previous work [33].

B. UAV dataset

This sequence was recorded along a residential street using the VI sensor mounted on the bottom of an AscTec Neo UAV (visible in the inset of Figure 1) in a front-looking configuration, while performing lateral movements with the UAV in both directions. This sequence exhibits perceptual aliasing as well as large variance in viewpoints and difficult lighting conditions as evident in Figure 7 and Figure 1.

V. RESULTS

We evaluate the proposed approach on datasets labelled with ground truth as described in Section IV and compare to the state of the art by analyzing their precision-recall characteristics. Moreover, as the proposed pipeline does not only provide a yes-or-no decision, but goes on to suggest a

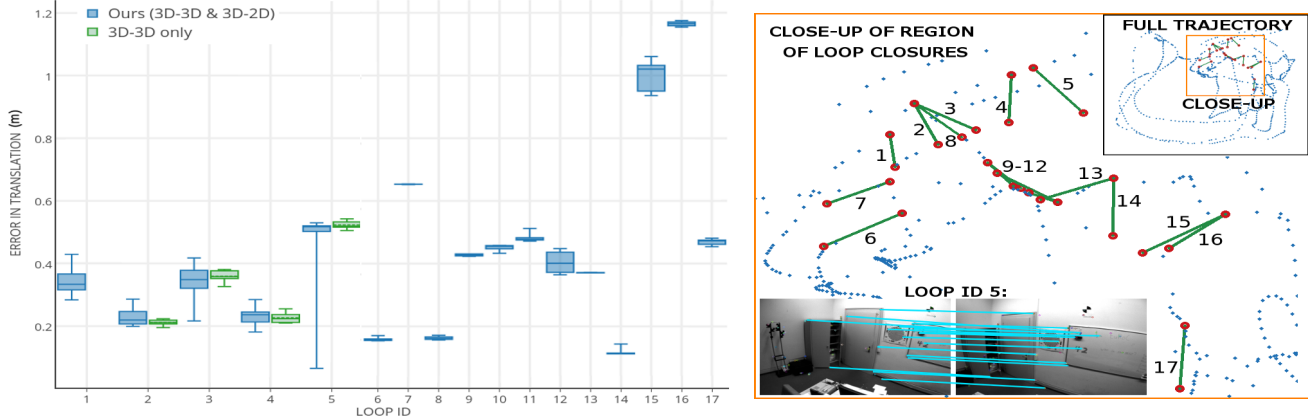


Fig. 5: Left: Error in the translation estimates of the transformation T_{QC} between two loop-closing keyframes (each pair represented by one Loop ID) averaged over 10 runs. Right: the UAV poses obtained with OKVIS (blue dots) and the loop closures with their corresponding ID annotated in green. Loop ID 5 is shown in the inset.

transformation between the matching keyframes to be used in a subsequent optimization step to enforce loop closure, we also evaluate the quality of these estimates. We present quantitative and qualitative evaluations on both hand-held and aerial scenarios.

A. Precision-Recall Characteristics

We record the precision-recall characteristics of the proposed method against FAB-MAP 2.0 [4], which is considered as the most well-established place recognition pipeline designed to combat perceptual aliasing. Moreover, as the method proposed in this paper employs binary features and draws inspiration from the DBoW2 approach of [17] we also compare to its performance. These tests are conducted on the Shopping Street 1 sequence. We test the proposed approach using a vocabulary composed of outdoor images captured in Zurich different to the ones used for testing. FAB-MAP and DBoW2 are tested using their corresponding original vocabularies. As evident in Figure 3 (a), the proposed approach achieves higher recall across all methods for perfect precision (i.e. equal to 1). The robustness of the proposed method is illustrated qualitatively in Figure 4. FAB-MAP is particularly challenged as it employs appearance-only checks in deciding for a loop-closure, while our approach and DBoW2 incorporate also geometric checks. DBoW2 exhibits high recall for perfect precision in Shopping Street 1, however, our improved geometric checks result to improved recall, which becomes particularly evident when testing with Shopping Street 1 & 2, where the viewpoint and other challenges are far greater. FAB-MAP precision-recall rates drop drastically (both to less than 0.1) in this case and DBoW2 detects four loops only. Despite that all of them are correct, they are far fewer than the total number of loop closures. The yellow curve in Figure 3(c) illustrates the recall reached by DBoW2 while varying the reprojection error.

As the proposed pipeline aims at greater robustness to viewpoint changes as well as to clean up false appearance matches, we employ both a 3D-3D geometric test similarly to ORB-SLAM [25], as well as a 3D-2D geometric test. A

comparison on precision versus recall to ORB-SLAM would not be fair, however, as it was designed to conduct loop-closure tests that are well spaced in time instead of testing at every keyframe as in the proposed method. The type and quality of features used as well as the estimation processes involved in ORB-SLAM in comparison to OKVIS have a direct impact on the quality of the performance of place recognition. So, here we isolate the effect of the 3D-3D and the 3D-2D geometric tests of the proposed pipeline to analyse the performance in both Shopping Street 1 alone and the dataset comprised of both Shopping Street 1 and 2 as shown in Figure 3 (b) and (c), respectively.

Retaining perfect precision, we monitor the recall obtained for variable reprojection error dictating the number of inliers agreeing with the transformation proposed using RANSAC. While one might expect that introducing the 3D-2D geometric checks as a second chance for a candidate loop-closure following a failed 3D-3D geometric check would have a negative impact on the precision-vs-recall trade-off, Figure 3(b) shows that higher recall can be achieved for the combined tests while retaining perfect precision. The added challenges in the Shopping Street 1 & 2 setup (greater changes in illumination, viewpoint and appearance as seen in Figure 4), indeed causes lower overall recall in Figure 3(c), but the combined 3D and 2D tests of the proposed approach still outperform the 3D only checks without compromising precision.

Traditionally, the answer to the question posed by place recognition techniques on whether we are re-visiting an already known place is binary (i.e. yes or no). Since our aim is to employ viewpoint-tolerant place recognition to indicate loop closures within SLAM, a first suggestion of the relative transformation between the loop closing frames (defined as T_{QC}) is not only very useful to a subsequent optimization step, but also an indication of the quality of the geometric checks used to decide for a loop closure in the first place. In the proposed scheme, the estimation of T_{QC} comes as a by-product of the Geometric Check step.

We use the EuRoC Vicor Room 2 03 sequence of the

EuRoC MAV dataset, which provides high-precision ground-truth poses for the UAV throughout this sequence. Upon the detection of a loop closure, we evaluate the quality of T_{QC} against ground-truth testing for both the full pipeline described in Section III and when using the 3D-3D geometric checks only. For both variants of our pipeline, we accumulate the estimated translation error across 10 runs as illustrated in Figure 5. It should be noted that due to the randomised nature of RANSAC, some loops are not detected in all runs. For completeness, we also analyse the translation error in the loop-closing transformations estimated by ORB-SLAM in the same scenario, seen on the right of Figure 5. Relocalization was triggered many times due to ORB-SLAM losing track, while different keyframes are selected in each run, rendering it harder to detect the same loops across different runs than with OKVIS. Even without considering the lower recall of ORB-SLAM, Figure 6 illustrates that the translation error in T_{QC} is much larger than with the proposed approach.

As evident in Figure 5, the inclusion of the 3D-2D geometric tests can sometimes result to bigger translation error in the estimation of T_{QC} , as expected. In fact, loops 15 and 16 result to considerable error given the size of the room, where the dataset was recorded. However, out of the 17 loops detected by the full pipeline, only 4 have been detected when using the 3D checks only. It should be highlighted that many of the loop-closing transformations estimated by the full approach were still computed using Horn’s 3D-3D method, since the covisibility consistency check did not fail (as in the 3D-only case); given that 3 consecutive consistent keyframe matches are needed before accepting a loop closure, the additional loop detections provided by the 3D-2D checks lead to correct detection of more true-positives. The vast majority of the additional detections exhibit error of the same order as the more restrictive 3D only checks (i.e. less than 50cm), in stark contrast to the much larger error characteristics of ORB-SLAM in Figure 6.

In conclusion, while the addition of inertial sensing can indeed result to better quality maps in OKVIS in comparison to ORB-SLAM, even when isolating the 3D only checks used in ORB-SLAM but using OKVIS maps, the proposed approach is evidently boosting recall and achieves better quality of loop-closing transformations T_{QC} . While T_{QC} is only a suggestion subject to further optimization in a bundle adjustment or pose-graph optimization step, the closer the estimate is to reality, the better the chances of subsequent convergence of the map to the global minimum. As a result, while the proposed use of additional 3D-2D checks can result to noisier transformations, these are still better than in ORB-SLAM and the sometimes dramatic increase in recall is evidently beneficial and can really make a difference in viewpoint-challenging scenarios.

B. UAV Experiments

The proposed approach was tested using the UAV dataset, exhibiting the biggest challenge for viewpoint-tolerant place recognition as visible in Figure 7. Added challenges, such as in illumination can cause false negatives as feature detection

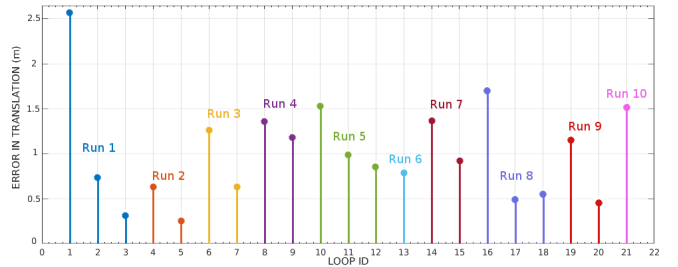


Fig. 6: The translation error in T_{QC} as estimated by ORB-SLAM for the scenario of Figure 5 (note: the loops detected here are different). Colors represent loops closed in each of the 10 runs, as different loops are detected every time.

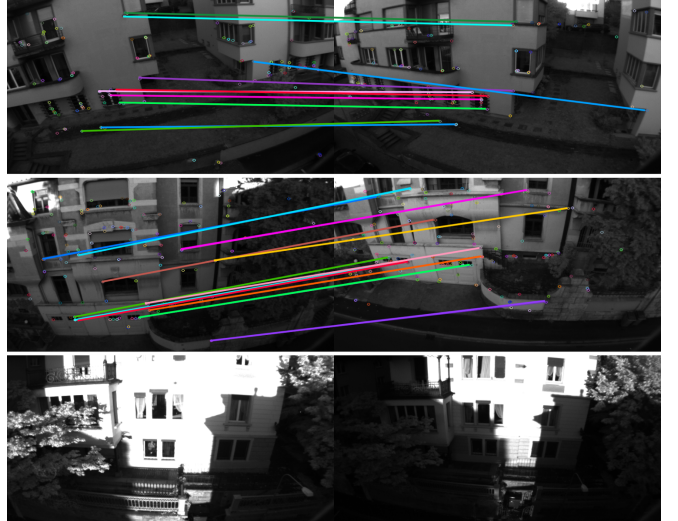


Fig. 7: Loop-closures in the UAV dataset tested with the proposed approach. Large viewpoint changes are successfully handled (top two rows), while strong lighting can wipe crucial features out resulting to false negatives (bottom).

is compromised. The loop-closures detected by our approach are visible (in green) in Figure 8. ORB-SLAM was also tested using this sequence, but no loops were detected.

C. Computational Cost

Feature extraction is usually the bottleneck in place recognition systems. With this in mind, the proposed method is re-using features extracted during the estimation of SLAM, enabling loop-closure detection at frame rate (i.e. 20Hz) in all the experiments presented in this paper. As the BRISK descriptor used within OKVIS consists of 48 bytes only, this restricts its descriptability posing bigger problems in loop detection, but makes descriptor comparisons even more efficient. Moreover, more relaxed conditions in the RANSAC scheme can be created in order to improve even more the performance, but the quality of transformations can also be affected.

VI. CONCLUSIONS

This paper proposes a novel pipeline for viewpoint-tolerant place recognition that makes use of promising leads from existing works, combining them in a way that enables unprecedented robustness to a wide range of common challenges

(i.e. tolerance to viewpoint, lighting changes, occlusions, perceptual aliasing, etc). The proposed pipeline was carefully designed to support low-burden computation and to take advantage of any scale and rotation invariance offered by BRISK using combined geometric checks that exploit not only the 2D information inherent in images but also the 3D information provided by a SLAM system.

Evaluation on newly recorded challenging outdoor datasets with both hand-held and aerial footage demonstrates that the proposed pipeline achieves better, or even drastically increased at times, recall in comparison to the state of the art, while maintaining perfect precision. Since no other such dataset appears in the literature, we make our testbed publicly available. Further evaluation on the quality of the estimated loop-closure transformation on an existing, indoor aerial dataset with pose ground truth reveals better quality of estimation than state of the art. Future work will study more extreme viewpoint changes and their impact on both similarity of appearance (e.g. consistency of word assignments) as well as geometry estimated by SLAM.

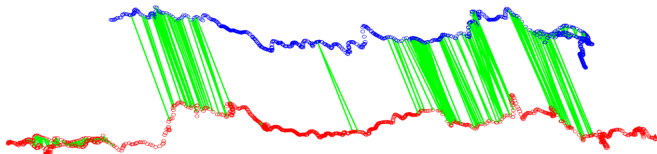


Fig. 8: Trajectory followed by the UAV in the UAV dataset. In blue/red are the UAV trajectories when travelling in opposite directions and in green are the loop-closures detected.

REFERENCES

- [1] S. Weiss, M. W. Achtelik, S. Lynen, M. C. Achtelik, L. Kneip, M. Chli, and R. Siegwart, "Monocular Vision for Long-term MAV Navigation: A Compendium," *Journal of Field Robotics (JFR)*, 2013.
- [2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2003.
- [3] M. Gehrig, E. Stumm, T. Hinzmann, and R. Siegwart, "Visual place recognition with probabilistic voting," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.
- [4] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *International Journal of Robotics Research (IJRR)*, 2011.
- [5] G. Klein and D. W. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007.
- [6] M. J. Milford, "Vision-based place recognition: how low can you go?" *International Journal of Robotics Research (IJRR)*, 2013.
- [7] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [8] F. Radenović, G. Toliás, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- [9] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla, "Learning and calibrating per-location classifiers for visual place recognition," *International Journal of Computer Vision (IJCV)*, 2016.
- [10] S. Cao and N. Snavely, "Graph-based discriminative learning for location recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [11] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2017.
- [12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, 2004.
- [13] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)*, 2008.
- [14] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT and SURF," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.
- [15] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.
- [16] A. Alahi, R. Ortiz, and P. Vanderghenst, "FREAK: Fast Retina Keypoint," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [17] D. Galvez-Lopez and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics (T-RO)*, 2012.
- [18] H. Altwaijry, E. Trulls, J. Hays, P. Fua, and S. Belongie, "Learning to match aerial images with deep attentive architectures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [19] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [20] A. L. Majdik, D. Verda, Y. Albers-Schoenberg, and D. Scaramuzza, "Air-ground matching: Appearance-based gps-denied urban localization of micro aerial vehicles," *Journal of Field Robotics (JFR)*, 2015.
- [21] R. Paul and P. Newman, "Fab-map 3d: Topological mapping with spatial and visual appearance," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2010.
- [22] E. Stumm, C. Mei, S. Lacroix, and M. Chli, "Location graphs for visual place recognition," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015.
- [23] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, and R. Siegwart, "Keyframe-based Visual-Inertial SLAM using Nonlinear Optimization," in *Proceedings of Robotics: Science and Systems (RSS)*, 2013.
- [24] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *International Journal of Robotics Research (IJRR)*, 2015.
- [25] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics (T-RO)*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [26] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.
- [27] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *International Journal of Robotics Research (IJRR)*, vol. 27, no. 6, pp. 647–665, 2008.
- [28] B. K. Horn, "Closed-form solution of absolute orientation using unit quaternions," *JOSA A*, vol. 4, no. 4, pp. 629–642, 1987.
- [29] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [30] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research (IJRR)*, 2013.
- [31] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, 2016.
- [32] J. Nikolic, J. Rehder, M. Burri, P. Gohl, S. Leutenegger, P. T. Furgale, and R. Siegwart, "A synchronized visual-inertial sensor system with fpga pre-processing for accurate real-time slam," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014.
- [33] F. Maffra, Z. Chen, and M. Chli, "Loop-closure detection in urban scenes for autonomous robot navigation," in *3D Vision (3DV)*, 2017.