# Multi-Modal Loop Closing in Unstructured Planetary Environments with Visually Enriched Submaps

Riccardo Giubilato, Mallikarjuna Vayugundla, Wolfgang Stürzl, Martin J. Schuster, Armin Wedler, Rudolph Triebel

***Abstract*— Future planetary missions will rely on rovers that can autonomously explore and navigate in unstructured environments. An essential element is the ability to recognize places that were already visited or mapped. In this work we leverage the ability of stereo cameras to provide both visual and depth information, guiding the search and validation of loop closures from a multi-modal perspective. We propose to augment submaps that are created by aggregating stereo point clouds, with visual keyframes. Point clouds matches are found by comparing CSHOT descriptors and validated by clustering while visual matches are established by comparing keyframes using Bag-of-Words (BoW) and ORB descriptors. The relative transformations resulting from both keyframe and point cloud matches are then fused to provide pose constraints between submaps in our graph-based SLAM framework. Using the LRU rover, we performed several tests in both an indoor laboratory environment as well as a challenging planetary analog environment on Mount Etna, Italy. These environments consist of areas where either keyframes or point clouds alone fail to provide adequate matches, thus demonstrating the benefit of the proposed multi-modal approach.**

## I. Introduction

Simultaneous Localization and Mapping (SLAM) enables autonomous robots to explore unknown and GPS-denied environments. The ability to create drift-free maps in challenging outdoor environments is crucial for the accurate localization of scientifically relevant targets or for search-and-rescue related tasks. In the context of space exploration, stereo cameras are widely used as a mechanically simple sensor, delivering both visual and depth information about the observed environment. A major advantage of depth information is that up to a certain degree the 3D structure of an environment can be estimated independent of light conditions and viewpoints. On the other hand, areas without suitable 3D features might still provide meaningful visual cues. In addition, the range for obtaining reliable depth information, in particular from stereo vision is limited, as accuracy and resolution decreases quickly with distance. Therefore, in order to obtain sufficient information for loop closure detection and validation in challenging environments it is advantageous to exploit both modalities.

In this paper we present the following major contributions:

- We propose a loop closure framework which leverages both the 3D structure and the visual appearance provided by stereo cameras in the context of a submap-based SLAM system, allowing to maximize pose accuracy and to increase the chance of establishing inter-pose constraints.
- We test the proposed framework both in an indoor laboratory environment comprising replicas of natural features and in an outdoor environment on Mount Etna, Italy, designated as an analog environment for lunar scenarios.

All authors are with German Aerospace Center (DLR), Institute of Robotics and Mechatronics, Münchener Str. 20, 82234, Wessling, Germany {firstname.lastname@dlr.de}

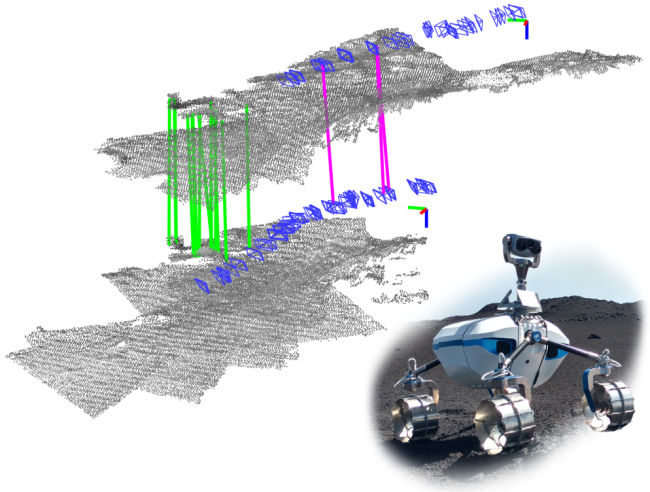


Fig. 1. Example of matching submaps from the "*Etna*" sequences using 3D descriptor and keyframe correspondences. Green lines denote matching 3D keypoints validated after Hough3D clustering, magenta lines connect matching keyframes. The two reference systems represent the origins of both submaps (Submaps are displaced in the vertical direction for visualization purposes). Bottom: LRU rover on Mount Etna, a designated Moon-analog test site in Sicily, Italy.

The paper is organized as follows: In section II we give an overview of the related work. After describing our submap-based SLAM pipeline (section III) we then, in section IV, present our approach for establishing multi-modal loop closures. In section V we evaluate the proposed system in several experiments and, finally, in section VI we draw some conclusions.

## II. Related Work

Many works exist in the literature about loop closure in SLAM by means of visual place recognition or by performing data association with other types of sensors [3]. In the context of mobile robotics and according to the motivation of this work, we distinguish the following categories based on whether correlations are searched amongst images or point clouds.

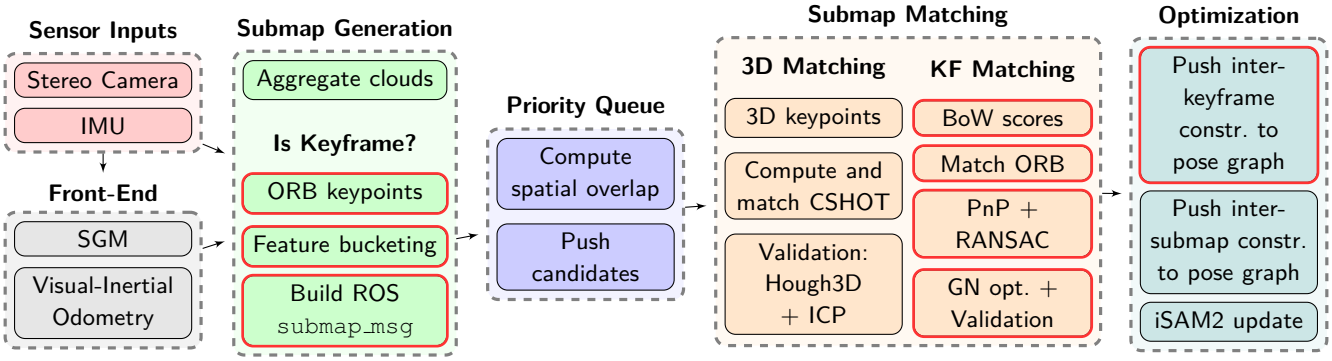



Fig. 2. Summary of our SLAM pipeline including parts from the previous 3D-only framework [1], [2] and the new contributions in this paper (highlighted in red): Stereo point clouds are aggregated into submaps from visual-inertial pose estimations. Submaps embed visual keyframes comprising a set of ORB descriptors as well as the camera pose relative to the submap origin (sec. III). Once new submaps are published in the ROS network, the spatial overlap of submaps, accounting for the pose covariance, establish a priority queue of tentative matches. Matches between submaps are validated from 3D and keyframe matches, maximizing the probability of validating loop closures. 3D matches are found by clustering CSHOT descriptor matches (sec. IV-A) while keyframe matches are established from a set of candidates using PnP+RANSAC and subsequently validated (sec. IV-B). All the constraints resulting from validated structure or keyframe matches are added to the graph, which is then optimized by the iSAM2 algorithm.

### A. *Visual Place Recognition*

Visual place recognition is traditionally performed by matching local feature descriptors, such as SURF [4], SIFT [5], BRISK [6] or ORB [7], usually leveraging clustering techniques to reduce the computational effort while preserving match precision [8], [9]. Visual SLAM systems often make use of the Bag of Words (BoW) model for loop closure detection such as ORB-SLAM2 [10], LDSO [11] or S-PTAM [12].

A large number of works describe adaptations and improvements of the BoW model for place recognition. The authors of [13] and [14] propose an incremental Bag of Binary Words formulation to remove the need to train a vocabulary, which is built and queried during the same session. In [15], the authors demonstrate the benefits of exploiting the co-occurrence of pairs of visual words to increase precision. In [16], the problem of visual place recognition is efficiently solved in a probabilistic manner removing the dependency on hard thresholds and showing robustness to perceptual aliasing. Although providing improvements over earlier works, the usual target application is for autonomous driving in urban environments, where the visual appearance is rarely ambiguous and co-occurrence of visual features is often guaranteed while revisiting the same locations.

Other approaches are targeted at improving robustness of vision-based approaches, which suffer from differences in viewpoint, non-monotonic changes in illumination and general appearance changes such as seasonal or meteorologic. The authors of [17] propose a localization system for autonomous driving systems leveraging multiple maps that are created in different environmental conditions and merged together. This, however, requires bootstrapping initial locations using GPS. The problem of place recognition despite strong changes in viewpoint is addressed in [18] by densification of the map using local meshes to gather more feature correspondences in the context of a BoW scheme. Differently from our system, this approach relies only on visual features, which in challenging outdoor scenarios might be ambiguous.

### B. *Point Cloud based Localization*

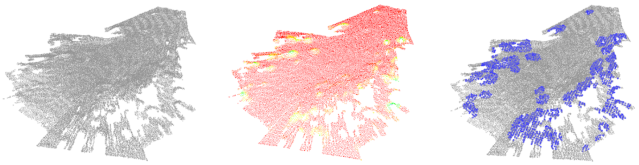
Other localization approaches rely on the usage of 3D Light Detection and Ranging (LiDAR) sensors, which return metrically accurate point clouds regardless of the visual appearance of the environment. Such is the case of many localization systems for autonomous driving, as common cityscapes suit well the problem of aligning 3D scans.

Detection of loop closures in this case relies on recalling similar point clouds, using global descriptors [19] or local descriptors such as SHOT [20] or FPFH [21]. Other approaches address the problem in way similar to the visual case by analyzing similarity of depth images [22].

In recent years, convolutional neural networks have been used to learn more efficient local descriptors [23], [24] that show higher performance than handcrafted approaches but require powerful hardware, which is usually unavailable in resource constrained vehicles. A different approach is followed by the authors of SEGMAP [25], [26] where a compact representation of segmented regions from point clouds are obtained through an autoencoder and matched in a different network. Despite the high performances in recalling similar places, this approach relies on the range and accuracy of 3D LiDARs and might be unsuitable for unstructured planetary environments, where it is unclear how to extract segments and 3D structure is often absent.

### C. *Heterogeneous Approaches*

As mentioned in the review paper [27], the combination of visual appearance and geometry can lead to better performances for place recognition. In [28], cameras are localized on a 3D map built from LiDAR scans by aligning local reconstructions from a purely visual pipeline to the map. Similarly, in [29], stereo frames are localized with respect to a 3D map by aligning depth images. Both approaches require to bootstrap an initial pose estimate for localization.



(a) Original Cloud (b) Curvatures (c) Keypoints

Fig. 3. Sampling keypoints on submaps from high curvature areas: After computing normals on the original submap point cloud (a), curvature is estimated from PCA (b) and points with high curvatures are used as seeds to sample 3D keypoints (c).

As a solution to this problem, the authors of [30] train a 3D and a 2D network which create a shared embedding space producing similar descriptors both for images and point clouds when they refer to the same place. In these works, localization depends on a strong correlation between visual information and structure. Contrarily, in our SLAM system, image and structure similarity are recalled independently from each other, increasing robustness for arbitrary viewpoints and trajectories.

## III. Submap-Based SLAM

In this section we briefly summarize the submap-based SLAM system for exploration vehicles equipped with stereo cameras introduced in our previous publications [1], [2], [31]. A schematic overview is given in Fig. 2, highlighting the components most relevant to this paper. Visual-inertial odometry provides locally accurate pose estimates which are used to merge local stereo point clouds into submaps. Creation of new submaps is triggered by enforcing constraints on the length of the travelled path or by the growth of pose uncertainty. Thus, submaps can be used as rigid point clouds for the purpose of loop closure and global map building. The origin of each submap is constrained by relative pose constraints and constitute nodes in a graph optimized by the iSAM2 algorithm [32] from the GTSAM library.

In this work, we extend our previous point cloud based formulation of submaps to embed visual information in the form of keyframes. While building submap $S$, the extracted keyframes are emplaced in the submap according to the pose from visual-inertial odometry relative to the submap origin. The relative position of keyframes is kept constant within each submap under the assumption that pose estimates are locally accurate. New keyframes are saved when either the translation or rotation exceeds a threshold with respect to the last keyframes. ORB features are extracted on each new frame and bucketed across the image to achieve a uniform distribution. Depth is associated to the detected features from stereo matching computed by SGM (Semi-Global Matching [33]) running on a dedicated FPGA. Each keyframe contains the transformation from the submap origin $\mathbf{T}_k^s$ as well as its covariance $\Sigma_k^s$.

## IV. Submap Matching

As a new submap is generated, loop closures are searched for by examining the overlap with older submaps, inflated according to their spatial uncertainty. As the overlap exceeds a set score [1], the corresponding pair of submaps is evaluated for possible 3D or keyframe matches.

### A. 3D Keypoint Matching

Upon first evaluation for matching, 3D keypoints are extracted either on segmented obstacle regions [31] or on high curvatures regions [34], as illustrated in Fig. 3. The later approach is better suited for unstructured outdoor environments where the distinction between traversable regions and obstacles is less evident. CSHOT descriptors [20] are matched using kd-trees on candidate submap pairs. Hough3D clustering [35] is then applied on the obtained descriptor correspondences in order to gather multiple hypotheses of transformations between the point clouds. In case, multiple hypotheses are compatible with a relative pose prior, for geometric consistency, the one with the highest number of voters is selected. The match is then refined using ICP, and a pose constraint is added to the graph.

### B. Keyframe Matching

Let submaps $S_0$ and $S_1$ be candidate matches containing two sets of visual keyframes $\mathrm{K}_0 \in S_0$ and $\mathrm{K}_1 \in S_1$.

**BoW scoring**: The first step is to compute a BoW representation for all keyframes, storing it in a cache to avoid recomputing in case the same submap is recalled for further evaluations. We use the DBoW2 library [9] for generating BoW vectors and scoring them. For every pair of BoW vectors $\mathbf{v}_i$ and $\mathbf{v}_j$, we compute the L1 score $s(\mathbf{v}_i, \mathbf{v}_j)$ proposed in [9]. For every keyframe in $\mathrm{K}_1$ we search for the keyframe in $\mathrm{K}_0$ that maximizes the score $s$ and such that the scores in a window of 3 keyframes centered on the candidate from $\mathrm{K}_0$ exceed the $60\%$ of the maximum value. In fact, if adjacent keyframes in $\mathrm{K}_0$ have comparable scores with the query in $\mathrm{K}_1$, they probably share visual words, as if they were captured consecutively. Furthermore, we impose a global constraint on BoW scores: we keep track of the highest and lowest score observed so far ($s_{\max}$ and $s_{\min}$) and accept a candidate matching keyframe pair only if

$$s(\mathbf{v}_i, \mathbf{v}_j) > 0.6 \cdot (s_{\max} - s_{\min}) + s_{\min}. \quad (1)$$

**Pose Estimation and Validation**: As a candidate keyframe pair is evaluated, the first step is to match all ORB descriptors, rejecting descriptor pairs whose Hamming distance is higher than 50. As mentioned previously (Sec. III), part of the image has depth information from disparity evaluated on a FPGA. We leverage this information to align $k_1 \in \mathrm{K}_1$ to $k_0 \in \mathrm{K}_0$ in a 3D-to-2D fashion, selecting from the matched feature pairs only those where depth is available in both frames. Specifically, we use the P3P algorithm [36] embedded in a RANSAC scheme to compute a tentative 6D transformation $\mathbf{T}_{k_0}^{k_1}$ using the full 3D landmarks from $k_0$ and the ORB features from $k_1$. If the number of inliers after the RANSAC test is lower than the $60\%$ of the input 3D-2D correspondences we discard the match. This value is selected from empirical considerations, in order to reject false matches but accounting for a moderate amount of

outliers. For the matches that passed the RANSAC test, we use the depth information in $k_1$ to check if the aligned sparse 3D clouds are coherent. Let $\mathbf{X}_j^{k_0} = \{x, y, z\}_j^{k_0}$ and $\mathbf{X}_j^{k_1} = \{x, y, z\}_j^{k_1}$ be two matched 3D landmarks belonging respectively to keyframes $k_0$ and $k_1$. We therefore compute the number of ORB matched pairs such that

$$|z_j^{k_1} - (R_{31}x_j^{k_0} + R_{32}y_j^{k_0} + R_{33}z_j^{k_0} + t_z)| < 0.1\,\mathrm{m} \quad (2)$$

where $0.1\,\mathrm{m}$ is the estimated maximum depth uncertainty of the stereo camera and $j = 1, 2, \ldots, n_{\text{inl}}$. $R_{31}, R_{32}, R_{33}$ and $t_z$ are coefficients from the bottom row of the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ returned after RANSAC. Thus, (2) is the difference of the z camera coordinates of an aligned pair of landmarks. If the fraction of keypoint pairs that satisfy (2) is lower than the $75\%$, the match is rejected as it may indicate that the RANSAC search selected a wrong transformation which satisfied a minimum consensus from a set of few wrong keypoint matches. This step is followed by a non-linear optimization step, refining the transformation computed after P3P-RANSAC and, furthermore, providing a covariance for the estimated pose. This optimization step aims at solving the following problem for all feature pairs that satisfied the previous validation test (2):

$$\underset{\mathbf{T}_{k_0}^{k_1}}{\arg\min} \sum_{j=1}^{n} \rho(r_j) \qquad r_j = |\mathbf{x}_j - \pi(\mathbf{X}_j^{k_0}, \mathbf{T}_{k_0}^{k_1})| \quad (3)$$

where $\rho()$ refers to the Cauchy robust loss function [37], $\mathbf{x}_j$ denotes the location of the keypoint $j$ in the image of $k_1$, $\mathbf{X}_j^{k_0}$ denotes the 3D coordinates of landmark $j$ in the reference frame of $k_0$ and $\pi$ is the projection function to the image of $k_1$. After solving (3), we obtain the covariance of the pose by extracting the marginal covariance related to the pose parameters. Equation (3) is solved with the Gauss-Newton method using the GTSAM library.

Finally, before accepting a keyframe match, we compute the transformation between the submap origins induced by the keyframe match using (4). Let $s_0$ and $s_1$ be the origins of two submaps containing the matching keyframes $k_0$ and $k_1$ respectively. Given the transformation between keyframes $\mathbf{T}_{k_1}^{k_0}$ and visual-inertial constraints, the transformation between submap origins is defined by

$$\begin{aligned} \mathbf{T}_{s_0}^{s_1} &= \mathbf{T}_{k_1}^{s_1}\, \mathbf{T}_{k_0}^{k_1}\, \mathbf{T}_{s_0}^{k_0} \\ &= \mathbf{T}_{k_1}^{s_1}\, \mathbf{T}_{k_0}^{k_1}\, (\mathbf{T}_{k_0}^{s_0})^{-1} \end{aligned} \quad (4)$$

where $\mathbf{T}_{k_0}^{s_0}$ and $\mathbf{T}_{k_1}^{s_1}$ are poses from visual-inertial odometry with respect to each submap origin and $\mathbf{T}_{k_0}^{k_1}$ results from the keyframe match. As the IMU makes the roll and pitch angles observable and submaps are gravity-aligned as provided, we verify that the roll and pitch components of $\mathbf{T}_{k_0}^{k_1}$ are close to zero or negligible. If a keyframe match satisfies all these checks, it is used to add a constraint to the graph as described in the following.

**Loop closure constraints**: Once a match is validated between keyframes $k_0$ and $k_1$, the respective poses are added as nodes to the graph (Sec. III) and are constrained with respect to the origins of the parent submaps with $[\mathbf{T}|\Sigma]_{k_0}^{s_0}$ and $[\mathbf{T}|\Sigma]_{k_1}^{s_1}$ (see Fig 4). The pose and covariance $[\mathbf{T}|\Sigma]_{k_0}^{k_1}$ instead constrain the global poses of the matching keyframes. In addition, eventual submap matches based on 3D correspondences (Sec. IV-A) establish a constraint between the origins of submaps $S_0$ and $S_1$.

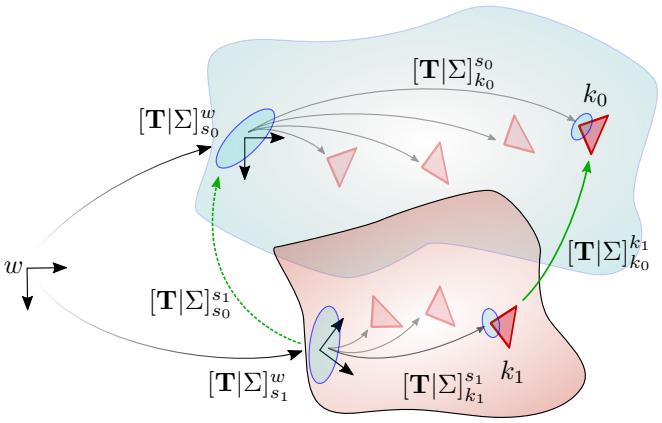



Fig. 4. Overview of the transformations involved in the process of submap matching. Green arrows denote matching constraints: the solid line denotes inter-keyframe constraints (visual matches, Sec. IV-B) while the dotted line denotes inter-submap constraints (3D matches, Sec. IV-A). Black arrows denote instead constraints from visual-inertial odometry: from world to submap origins and from submap origins to keyframes.

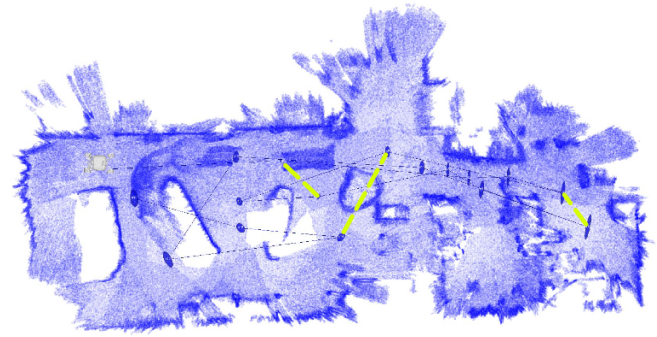
## V. Experiments

In this section we describe tests of our SLAM system in two different environments demonstrating the benefit of using multiple modalities for matching submaps built from stereo camera measurements. The test sites were a laboratory environment at the DLR Institute of Robotics and Mechatronics, Germany, and a designated planetary analog environment on Mount Etna, Italy [38]–[40].

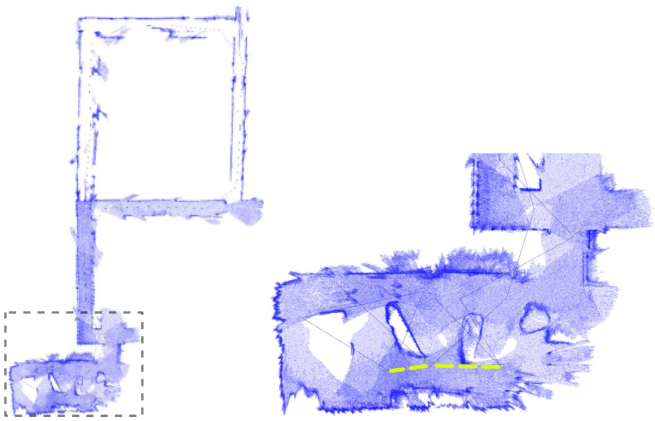
### A. Indoor Sequences

The first test environment is an indoor laboratory featuring a number of large stone replicas which provide the look of a rocky outdoor environments. The laboratory features a ceiling mounted Vicon tracking system which provides absolute poses (rotation and translation) for an accurate ground truth with a coverage of about $70\,\mathrm{m}^2$. In these sequences, the obstacles, or non-traversable regions, provide unique 3D structures where the extracted CSHOT descriptors should be unambiguous. As the obstacles are easily distinguishable from the remaining environment (walls and floor), we extract 3D keypoints on segmented obstacles from depth images [1], [31].

To evaluate the quality of pose estimation, we compute position and angular errors given the difference between Vicon poses and submap origins, which are the one optimized by iSAM2. Let $\mathbf{T}_{s_i}^{w}$ and $\hat{\mathbf{T}}_{s_i}^{w}$ be the estimated and true transformations from a global reference system to the origin of the $i$-th submap. Thus, the transformation $\mathbf{T}_{\text{err}} = \mathbf{T}_{s_i}^{w}\ (\hat{\mathbf{T}}_{s_i}^{w})^{-1}$ describes the difference between the true and estimated pose of submap $i$. Being $\mathbf{T}_{\text{err}} = [\mathbf{R}_{\text{err}} \mid \mathbf{t}_{\text{err}}]$, we estimate the



(a) *rmc_lab_long* top view



(b) *rmc_corr* full top view (c) *rmc_corr* detail

Fig. 5. Top views of the *rmc_lab_long* (a) and the *rmc_corr* (b) sequences; (c) shows of detail of *rmc_corr* as indicated by dashed gray rectangle in (b). Dashed yellow lines highlight the inter-submaps constraints after establishing loop closures. The accuracy of pose estimates after SLAM updates can be visually inferred from the coherence of the point clouds in the proximity of the walls and rocks in the lab.

position error using the L2 norm of the translation part $\|\mathbf{t}_{\text{err}}\|^2$ and the orientation accuracy by computing the total angle from the rotation term $\mathbf{R}_{\text{err}}$ as

$$\phi_{\text{err}} = \arccos\left(\frac{\text{Tr}(\mathbf{R}_{\text{err}}) - 1}{2}\right) \tag{5}$$

where $\text{Tr}(\mathbf{R}_{\text{err}})$ is the trace of the rotation matrix.

The three sequences, *rmc_lab_long*, *rmc_corr* and *rmc_lab* always start and end inside the laboratory, visible in Fig. 5c, such that ground truth is always available at the beginning and end of the trajectory. Table I reports the errors evaluated in all the indoor sequences. We test our SLAM system enabling the 3D and keyframe matching (KF) modules independently to observe which configuration leads to higher performances. We also report the errors obtained from the visual-inertial odometry which, as expected, are the highest. Overall, the combination of the two matching strategies leads to the best performances as the chance of validating submap matches is higher. While in our previous works [1], [34] the presence of unique 3D structures was necessary, in the proposed framework, submaps can match either from similarity in structure, visual appearance or both. In the 3D+KF case, the number of matched submaps is usually higher, as is consequently the number of constraints pushed to the optimization graph. In the laboratory datasets, furthermore, the number of matched submaps using the 3D strategy is usually higher than those using keyframes. The extent of each submap make them comprise many 3D structures which can easily be matched, while keyframe matches need the camera to return very close to previously visited places, as the stereo head is usually tilted towards the ground in order to observe the presence of obstacles. Figure 5 shows the top views of two laboratory sequences highlighting the connection between submaps from odometry and from map matches, using both 3D structure and keyframes. Examples of submap matches with matches from keyframes and 3D descriptors highlighted in different colors are presented in Figures 7c and 7d.

TABLE I

RMSE ERRORS IN THE RMC SEQUENCES WITH MULTIPLE MATCHING STRATEGIES. LOWEST ERROR IN BOLD, HIGHEST IN BRACKETS

| Sequence | RMSE | No Matches | 3D | KF | 3D+KF |
|---|---|---|---|---|---|
| *rmc_lab_long* | pos [m] | 0.47 | (0.54) | 0.27 | **0.21** |
| | z [m] | (0.20) | 0.08 | 0.05 | **0.08** |
| | angle [°] | (4.18) | 1.47 | 0.91 | **0.90** |
| | $n_{\text{matches}}$ | - | 4 | 2 | **5** |
| *rmc_corr* | pos [m] | (1.31) | 0.24 | 0.94 | **0.20** |
| | z [m] | (0.25) | 0.09 | 0.20 | **0.09** |
| | angle [°] | **0.81** | 2.18 | (4.07) | 0.86 |
| | $n_{\text{matches}}$ | - | 4 | 1 | **5** |
| *rmc_lab* | pos [m] | (0.51) | 0.33 | 0.31 | **0.28** |
| | z [m] | (0.15) | **0.09** | **0.09** | **0.09** |
| | angle [°] | (2.68) | 2.56 | **1.65** | 2.48 |
| | $n_{\text{matches}}$ | - | 3 | 1 | 3 |

TABLE II

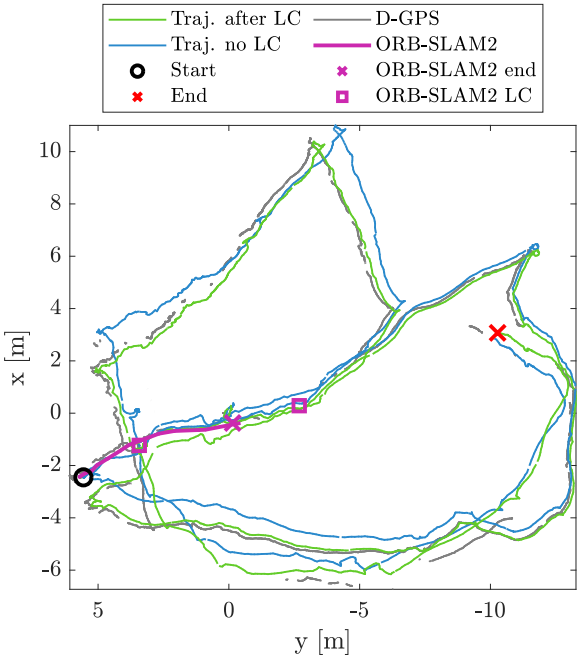
POSE ERRORS AFTER ALIGNMENT WITH THE DGPS GROUND TRUTH IN THE ETNA SEQUENCE FOR MULTIPLE MAP MATCHING STRATEGIES

| | No Matches | 3D only | KF only | 3D+KF |
|---|---|---|---|---|
| RMSE [m] | (0.38) | 0.32 | 0.30 | **0.24** |
| $n_{\text{matches}}$ | - | 2 | 4 | **5** |

## B. Outdoor Sequences

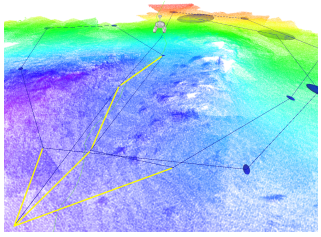
In this section we evaluate the performances of our pipeline on a sequence recorded on Mount Etna, Italy. The moon-like environment includes a distribution of small volcanic rocks resulting in an ambiguous visual appearance (see Fig. 6b for an example camera view). A DGPS setup provides global position measurements as ground truth. Thus, contrarily to the indoor sequences, we evaluate here only position errors after aligning the SLAM results to the DGPS track given correspondences between the timestamps. Poses are aligned to the ground truth using Horn's quaternion method [41] optimizing rotations and fixing the relative pose from the first timestamp correspondence.

Figure 6 shows the aligned trajectories computed with loop closure disabled and enabled with the 3D+KF matching strategy. The trajectories in blue and green are plotted by




(a) Aligned trajectories with DGPS ground truth

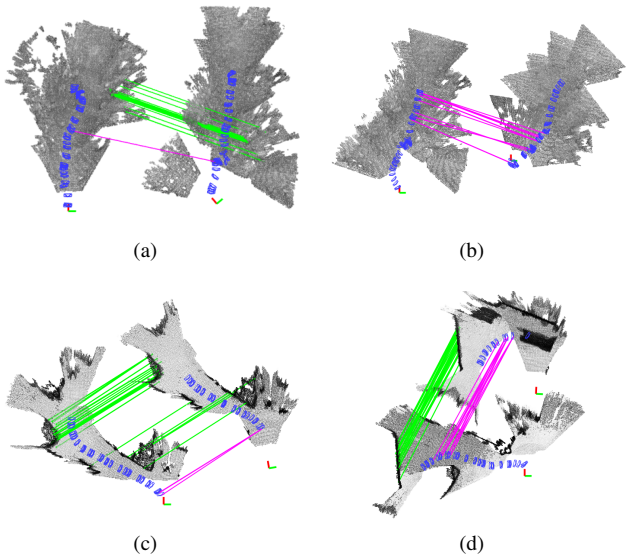
(b) Example image



(c) Inter-submap constraints

Fig. 6. Full trajectories from the Etna sequence aligned to the DGPS ground truth with Loop Closure enabled and disabled. The green path corresponds to the KF+3D map matching strategy, which is the most accurate according to Table II. The magenta path is estimated by ORB-SLAM2 before a tracking failure (magenta cross). Squares denote the positions where relocalization occurred for ORB-SLAM2, however followed by other tracking failures. (b) Example of camera image from this Etna sequence. (c) Detail of map and pose graph with inter-submap constraints after matching (yellow lines).

transforming local trajectories relative to each submap into world coordinates, using the transformations from submaps to world as estimated by the most recent iSAM2 update. The plot in Fig. 6a shows that the SLAM results after loop closures better fit the ground truth, while the trajectory from visual-inertial odometry-only accumulates drift that occur especially during rotations. In fact, the RMSE pose errors reported in Table II with loop closure enabled decrease in the order RMSE(3D-only)>RMSE(KF-only)>RMSE(KF+3D). This is because the combination of different modalities can establish a larger number of pose constraints that are also more widely distributed along the graph, see Fig. 6c for an example of loop closures. Figures 7a and 7b show two examples of submap pairs from the Etna sequence where matches are obtained predominantly either with 3D descriptors (Fig. 7a) or with keyframes (Fig. 7b), motivating the need for combining different modalities in



(a) (b) (c) (d)

Fig. 7. Heterogeneous submap matches: green lines are 3D keypoint matches from CSHOT correspondences, magenta lines are keyframe matches from ORB descriptor correspondences. In (b) the submap match is established only by keyframe correspondences, while in (a) and (c) 3D matches are predominant. (a) and (b) are submap matches from the Etna sequence, while (c) and (d) are example matches from *rmc_lab_long*.

such challenging scenarios. Note that, with respect to Table II, the improvement of a single submap match is quite significant: submaps span areas of approximately 7 meters in lenght and therefore represent large portions of the mapping session.

To compare the performance of our system with other state of the art SLAM approaches, we tested ORB-SLAM2 in RGB-D configuration using the recorded depth from SGM, computed on an FPGA onboard the rover. As can be seen in Fig. 6, the estimated trajectory stopped after about 7 meters from the start as a result of tracking failures due to the repetitive and ambiguous appearance of the challenging moon-like scenario. Note the presence of two successful relocalizations (magenta boxes) followed by an immediate tracking failure.

## VI. Conclusions

In this work, we presented a novel approach to establishing loop closures in a submap-based SLAM system by leveraging structure and appearance similarity. Candidate submap matches, selected from prior knowledge of the robot pose and uncertainty, are validated by searching correspondences both across 3D keypoint descriptors and visual keyframes. We tested the proposed system both in indoor laboratory environments with replicas of natural structures as well as in a sequence captured in a planetary analogous scenario on Mount Etna, Italy, showing the benefits in terms of pose accuracy given by matching submaps in different modalities. Future developments of this pipeline involve extending the existing framework to a multi-robot system comprising both ground and aerial vehicles [42].

## ACKNOWLEDGMENT

This work was supported by the Helmholtz Association, project alliance ROBEX (contract number HA-304) and project ARCHES (contract number ZT-0033).

## REFERENCES

[1] M. J. Schuster, *et al.*, "Distributed stereo vision-based 6D localization and mapping for multi-robot teams," *Journal of Field Robotics*, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21812

[2] M. J. Schuster, "Collaborative Localization and Mapping for Autonomous Planetary Exploration: Distributed Stereo Vision-Based 6D SLAM in GNSS-Denied Environments," Ph.D. dissertation, University of Bremen, 2019. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:gbv:46-00107650-19

[3] C. Cadena, *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 404–417.

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

[6] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 2548–2555.

[7] E. Rublee, *et al.*, "ORB: An efficient alternative to SIFT or SURF," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

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

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

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

[11] X. Gao, *et al.*, "LDSO: Direct sparse odometry with loop closure," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2198–2204.

[12] T. Pire, *et al.*, "S-PTAM: Stereo parallel tracking and mapping," *Robotics and Autonomous Systems*, vol. 93, pp. 27–42, 2017.

[13] S. Khan and D. Wollherr, "IBuILD: Incremental bag of binary words for appearance based loop closure detection," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 5441–5447.

[14] 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.

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

[16] M. Gehrig, *et al.*, "Visual place recognition with probabilistic voting," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 3192–3199.

[17] M. Bürki, *et al.*, "VIZARD: Reliable visual localization for autonomous vehicles in urban outdoor environments," *arXiv preprint arXiv:1902.04343*, 2019.

[18] F. Maffra, *et al.*, "Real-time wide-baseline place recognition using depth completion," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1525–1532, 2019. [Online]. Available: https://doi.org/10.1109/LRA.2019.2895826

[19] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 4802–4809.

[20] F. Tombari, S. Salti, and L. Di Stefano, "A combined texture-shape descriptor for enhanced 3D feature matching," in *IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 809–812.

[21] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 3212–3217.

[22] B. Steder, *et al.*, "NARF: 3D range image features for object recognition," in *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 44, 2010.

[23] Z. Gojcic, *et al.*, "The perfect match: 3D point cloud matching with smoothed densities," *arXiv preprint arXiv:1811.06879*, 2018.

[24] Z. J. Yew and G. H. Lee, "3DFeat-Net: Weakly supervised local 3d features for point cloud registration," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 630–646.

[25] R. Dubé, *et al.*, "Incremental-segment-based localization in 3-D point clouds," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1832–1839, 2018.

[26] R. Dubé, *et al.*, "SegMap: Segment-based mapping and localization using data-driven descriptors," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 339–355, 2020.

[27] N. Piasco, *et al.*, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, 2018.

[28] T. Caselitz, *et al.*, "Monocular camera localization in 3D LiDAR maps," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 1926–1931.

[29] Y. Kim, J. Jeong, and A. Kim, "Stereo camera localization in 3D LiDAR maps," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 1–9.

[30] D. Cattaneo, *et al.*, "Global visual localization in LiDAR-maps through shared 2D-3D embedding space," *arXiv preprint arXiv:1910.04871*, 2019.

[31] C. Brand, *et al.*, "Stereo-vision based obstacle mapping for indoor/outdoor SLAM," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014, pp. 1846–1853.

[32] M. Kaess, *et al.*, "iSAM2: incremental smoothing and mapping using the bayes tree," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 216–235, 2012.

[33] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[34] R. Giubilato, *et al.*, "Relocalization with submaps: Multi-session mapping for planetary rovers equipped with stereo cameras," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 580–587, 2020.

[35] F. Tombari and L. Di Stefano, "Object recognition in 3d scenes with occlusions and clutter by hough voting," in *Pacific-Rim Symposium on Image and Video Technology*, 2010, pp. 349–355.

[36] X.-S. Gao, *et al.*, "Complete solution classification for the perspective-three-point problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[37] J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4331–4339.

[38] A. Wedler, *et al.*, "ROBEX – components and methods for the planetary exploration demonstration mission," in *13th Symposium on Advanced Space Technologies in Robotics and Automation*, 2015.

[39] ——, "First results of the ROBEX analogue mission campaign: Robotic deployment of seismic networks for future lunar missions," in *68th International Astronautical Congress (IAC)*, 2017.

[40] M. Vayugundla, *et al.*, "Datasets of Long Range Navigation Experiments in a Moon Analogue Environment on Mount Etna," in *ISR 2018; 50th International Symposium on Robotics*, 2018, pp. 1–7.

[41] B. K. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, vol. 4, no. 4, pp. 629–642, 1987.

[42] M. G. Müller, *et al.*, "Robust visual-inertial state estimation with multiple odometries and efficient mapping on an MAV with ultra-wide FOV stereo vision," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 3701–3708.