 DOI: 10.3217/978-3-85125-603-1-23

Vision-based Autonomous Feeding Robot

Matthias Schörghuber[1,4], Marco Wallner[1], Roland Jung[2], Martin Humenberger[3], Margrit Gelautz[4]

***Abstract*— This paper tackles the problem of vision-based indoor navigation for robotic platforms. Contrary to methods using adaptions of the infrastructure (e.g. magnets, rails), vision-based methods try to use natural landmarks for localization. However, this imposes the challenge of robustly establishing correspondences between query images and the natural environment which can further be used for pose estimation. We propose a monocular and stereo VSLAM algorithm which is able to, first, generate a map of the target environment and, second, use this map to robustly localize a robot. Our hybrid VSLAM approach is able to utilize map points from the previously generated map to (i) increase robustness of its local mapping against challenging situations such as rapid movements, dominant rotations, motion blur or inappropriate exposure time, and to (ii) continuously assess the quality of the local map. We evaluated our approach in a real-world environment as well as using public benchmark datasets. The results show that our hybrid approach improves the performance in comparison to VSLAM without an offline map.**

I. INTRODUCTION

In order to perform autonomous navigation, robot platforms need to be able to robustly and reliably localize themselves and track their position within their operation area. Good examples are commercially available robot platforms fulfilling logistic tasks, e.g. within hospitals or warehouses. For localization, these systems rely on magnetic markers, rails or other infrastructure-based guidance systems. To continuously track the pose between such markers, many robot platforms perform dead reckoning approaches such as wheel odometry. The target robot platform in this paper, namely Wasserbauer's "Butler Gold" feeding robot (shown in Fig. 1), currently uses a very similar approach for autonomous navigation. Even if this and related approaches perform well in target environments, they require costly adaptions of the infrastructure and thus only work within a very well-defined area or even only along well-defined paths.

Since, on the one hand, the application fields of robots are not limited to industrial environments where necessary adaptions can be made, and on the other hand, the costs and efforts of these installations should be reduced, alternative approaches are investigated. The robotics community suggests to use cameras (mounted on the robot) for navigation since they provide rich information about the environment and are easy to install in comparison to other technologies (a survey can be found in [10]). Following this idea, in this paper, we present a visual navigation system, especially designed for autonomous robot indoor navigation. The goal is to robustly localize and track the robot's position within a certain area using passive cameras only. We excluded active light emitting technologies to not interfere with the environment and to enable a possible extension for outdoor usage. In the nature of the application, the vision system has to operate in a challenging environment as dynamic objects (moving cows) are present and the structure (feed, tools, constructions) changes from mission to mission. Furthermore, the system has to be robust against a variety of environmental conditions such as dirt, dust, moisture, lighting or occlusions.

Fig. 1: Target robot platform for feeding with front-facing stereo camera system. Right image: © Wasserbauer

II. RELATED WORK

Similar to many other navigation tasks, for visual localization and pose estimation, certain landmarks are needed. We roughly differentiate between artificial and natural landmarks. Artificial landmarks, such as QR-Codes, are well defined and need to be placed manually. Natural landmarks, such as significant corners or well-textured areas in the image, need to be identified automatically using proper feature extraction methods. In this work, we focus on visual navigation using natural landmarks, since our target is a general approach where the environment does not need to be adapted. The problem of visual localization using natural landmarks is addressed in several ways. Structure-from-Motion (SfM, e.g. [8]) uses multiple images to estimate their positions and to reconstruct the captured environment. Image-based localization (e.g. [11]) uses the resulting maps to estimate the pose of query images. While applications

The research leading to these results has received funding from the Austrian Ministry for Transport, Innovation and Technology (BMVIT) within the ICT of the Future Programme of the Austrian Research Promotion Agency (FFG) under grant agreement no. 849909 (FarmDrive) and Industrienahe Dissertation Programme under grant agreement no. 848518 (AVIS).

[1]Matthias Schörghuber and Marco Wallner are with the Austrian Institute of Technology {matthias.schoerghuber,marco.wallner}@ait.ac.at

[2]Roland Jung is with the Alpen-Adria Universität Klagenfurt roland.jung@aau.at

[3]Martin Humenberger is with Naver Labs Europe martin.humenberger@naverlabs.com

[4]Matthias Schörghuber and Margrit Gelautz are with the Vienna University of Technology margrit.gelautz@tuwien.ac.at

of the mentioned approaches are often found in large-scale and offline localization tasks where memory consumption and processing time play minor roles, in robotics these two issues are critical. Addressing these challenges, important and relevant methods for visual pose estimation such as visual odometry (VO, [6], [4], [15]), visual inertial odometry (VIO, [12], [1]), and visual simultaneous localization and mapping (VSLAM, [9], [5]) were introduced. We differentiate VO from VSLAM by the property that VO does not implement global map optimization or loop-closure, i.e., the process of recognizing that a place was visited before to reduce drift. Therefore, VO algorithms typically maintain only a local map and "forget" about the past. While VO may exhibit more drift than VSLAM, it is computationally more efficient. VIO additionally uses data from inertial measurement units (IMUs) to combine measures from inertial sensors (gyroscope, accelerometer) with visual information. A recent in-depth overview of SLAM discussing common architectures, history, the present, and future is presented in [3] and [14].

Image-based localization has its strength in absolute localization and VO/VSLAM in relative pose estimation, yet an approach which robustly combines them is still missing. In this paper, we propose an approach to overcome this problem. Recently maplab [12] was published where this problem is addressed as well. The authors perform VIO locally and simultaneously fuse the estimated relative pose with absolute localizations within the existing map. In contrast, our approach tracks and uses the existing map points directly.

We present a VSLAM algorithm which is able to combine robust geo-referenced global offline maps, previously generated using VSLAM (in this paper we use the presented algorithm) or SfM, with local online maps which are generated during run-time.

III. Visual Navigation

We propose a visual navigation system which allows to localize and track the pose of the feeding robot within the operational area. To be robust against the changing environment, we perform an initial mapping process where a map is first generated using our VSLAM algorithm and in a second step, registered onto a floor plan (to enable absolute localization). This map is then used in subsequent missions as reference and referred to as offline map. However, in the nature of the application, the existing map is outdated due to the dynamic environment as camera occlusion, structural changes or moving objects can occur. Leveraging the combination of using the initial offline map and performing online mapping, we aim to (i) operate in a global coordinate system defined by the offline map, (ii) allow flexible movements and robustness against structural changes by performing online mapping and (iii) suppress map degeneration on challenging scenes by simultaneous validation of the online mapping with the existing offline map.

The vision pipeline for pose estimation, map generation and localization is inspired by modern SLAM systems (see Section II). It uses a feature based approach with the map consisting of keyframes, 3D points and their 2D observations within the keyframes. The most similar approach to our implementation is ORB_SLAM [9]. It presents a VSLAM system with a localization only mode (i.e. without mapping), however, neither loading or storing nor usage of pre-existing maps is possible.

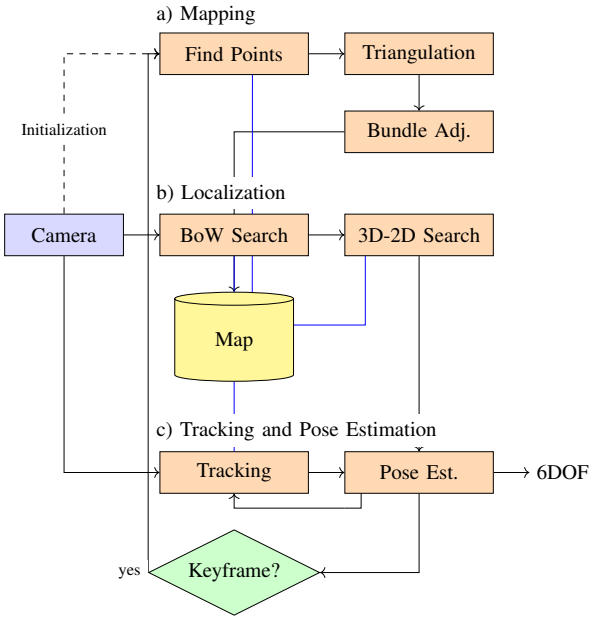


Fig. 2: Algorithmic overview of the main operational modes of the VSLAM implementation.

The core blocks of our implementation are a) Mapping, b) Localization and c) Tracking and Pose Estimation as shown in Fig. 2 and described in the following.

a) Mapping: The mapping process aims to find and triangulate new reliable 3D points between the existing map and new input images. In the *Find Points* procedure, a correspondence search of image features not associated with 3D points is performed with features of neighboring keyframes. If no map is present during initialization of the algorithm, the correspondence search is performed either with the stereo image pair or with two consecutive images with sufficient translation in between in case of monocular input. The new 3D point candidates are triangulated and validated using geometric properties. Finally, *Bundle Adjustment* is performed with keyframes and 3D points affected by the new measurements. We distinguish between 3D points originating from the online and offline map within this optimization step. The 3D points of the existing global map are assumed to be fixed and reliable, thus they are higher prioritized. This prioritization prevents a drift or degeneration between the online and the offline map, as new measurements are aligned to the coordinate system defined by the offline map even in cases where online map points are dominant. To prevent map degeneration and undefined behaviour in subsequent missions, the offline map is not updated with new measurements in our current application.

b) Localization: Localization aims to find the pose of a query image within a given map. Our implementation performs a two-step process. First, the most similar keyframes are retrieved with a visual vocabulary based approach. In such a visual vocabulary the image descriptors are clustered into words of a pre-defined vocabulary (bag of words [7]). This methodology allows to efficiently compare images on this reduced set of words. Image similarity is determined by the presence of shared words of the vocabulary between images.

The second step performs a correspondence search between 3D points visible in the retrieved keyframe and 2D image points of the query image. Finally, the position and orientation of the query image can be estimated using multiple correspondences with non-linear optimization by minimizing the reprojection error.

c) Tracking and Pose Estimation: This module aims to estimate the pose of the current image by tracking an existing pose (from last tracks or successful localization) within the map. First, using the extrapolated input pose, 3D points within the camera frustum are projected onto the image plane. Second, using these estimates, guided matching is performed to establish 2D-3D correspondences between the 3D points and the 2D keypoints of the current image. The new pose is estimated by minimization of the reprojection error between the 3D point projections and corresponding 2D image feature observations.

Similar to the mapping process, we distinguish between 3D points originating from the online and offline map and prioritize the measurements within the pose estimation step accordingly. Furthermore, the tracking of 3D points from the offline map allows to determine map consistency and prevents degeneration in challenging scenes. If no 3D points of the offline map can be found, the robot either left the operational area (which can be validated using previous poses) or the camera provides no valid information for global localization (e.g. occlusions or close-up scenes). As a consequence, the algorithm tries to relocalize within the offline map. If this is not successful, a navigation error occurred and a recovery mode has to be triggered. Based on the number of visible 3D points originating from the offline map and total number of 3D points, it is decided when new keyframes shall be generated and added to the online map.

The VSLAM performance strongly depends on the robot movement. If fast movements are present, the algorithm has to deal with motion blur and large unobserved gaps between frames. This especially occurs when motions with a mainly rotational component are present. Furthermore, with monocular input, the triangulation requires a translational component, which defines the theoretical accuracy of the 3D reconstruction. The proposed approach, using an offline map as basis, increases robustness to such challenging robot movements because even if triangulation of new 3D points fails, 3D points from the offline map can be tracked. In the same way, if tracking is lost, the robot can relocalize itself within the offline map. Furthermore, the number of visible offline points can be used as quality indicator of the online map.

In order to provide the poses within the application specific world coordinate frame, the map is registered onto a geo-referenced 2D floor plan. This rigid transformation is estimated using manually selected correspondences between map points and salient features on the floor plan such as wall corners.

IV. Tests and Evaluation

We performed two experiments to evaluate our VSLAM system. First, in a real-world environment with recordings acquired in a cowshed and second, with a benchmark dataset from the community.

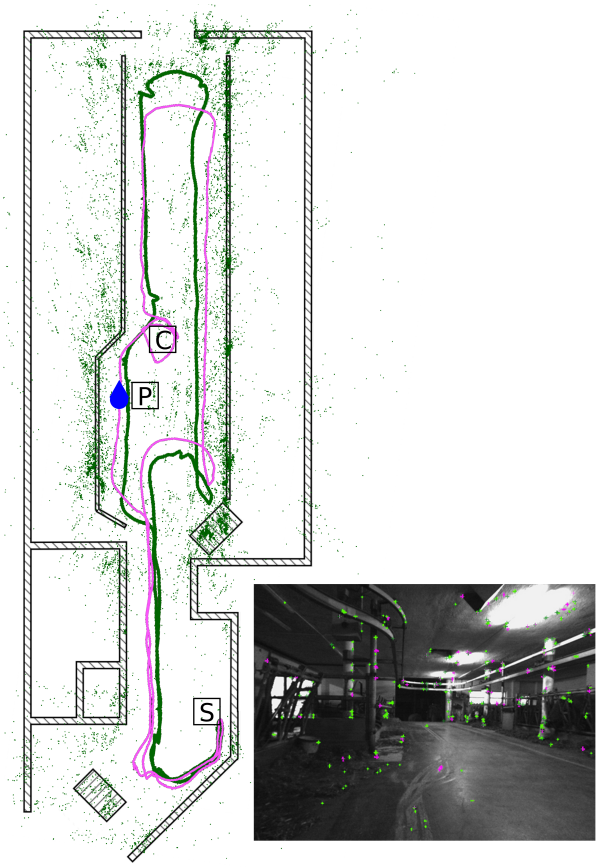
For the real-world test, we used one recording[1] to generate an offline map and a second one for combining offline maps with online mapping. The trajectories are slightly different, as can be seen in the purple and green trajectory in Fig. 3 (a). The green trajectory represents the trajectory of the initial mapping run after registration onto the floor plan. The green dots indicate the locations of the 3D points generated during this run. The purple line represents the trajectory of the second recording using the 3D map from the green mission as the offline map. Both missions have a common starting and end point indicated at the position *S* within Fig. 3 (a), hence the double line in the entry path of the cowshed.

The camera image in Fig. 3 (b) shows the 2D projections of the observed 3D points as seen from the position *P* marked by a blue circle and an arrow indicating the viewing direction in Fig. 3 (a). The green points visible within the camera image in (b) correspond to 3D points originating from the offline map (and correspond to the green dots in 3 (a)). The purple points represent newly generated 3D points in the online map.

The consistency of the purple trajectory with the floor plan confirms that the algorithm was able to perform its mission within the application specific world coordinate frame. Furthermore, with the online mapping, the system was able to estimate the pose even when the robot moved differently in comparison to the data available from the offline map; this is especially visible at the circular movement at position C in Fig. 3 (a).

Since no ground truth was available for quantitative analysis of the cowshed data, we used the *EuRoC* [2] dataset to evaluate our assumptions of increased robustness and accuracy that can be achieved by additionally including 3D points of an offline map into the core computations of the algorithm. This dataset was created with a multi-rotor unmanned aerial vehicle (UAV) equipped with a stereo camera sensor providing 20 frames per second. The dataset consists of sequences captured in a machine hall (*mh*) scenario with positional ground truth and two laboratories (*v1*, *v2*) equipped with a motion capture system providing 6 DOF ground truth. The sequences were recorded with varying difficulties. A higher difficulty implies faster translational

[1]The actual robot control was performed using its navigation system.



(a) (b)

Fig. 3: (a) 3D map from the VSLAM system registered onto the cowshed floor plan; (green) 3D points and trajectory of initial map generation; (purple) trajectory of a subsequent mission. (b) 3D points of map projected onto camera image captured from the blue position *P* in (a). Green dots represent observed 3D points from the initially generated offline map, and purple dots the online generated 3D points.

and rotational movements of the UAV, operation in a low-textured environment or exposure differences due to auto shutter effects. We processed the sequences with our VSLAM algorithm, which is able to provide pose information at camera frame rate, and estimated the accuracy of each sequence by comparing it with the ground truth using the root mean square errors of absolute trajectory error (ATE) and relative pose error (RPE). The latter consists of a translational (RPE_t) and a rotational (RPE_r) component. These metrics are defined in [13].

The results are shown in TABLE I. It can be seen, as expected, that the trajectory could be estimated more precisely on less complex sequences *mh_01*, *mh_02*, *v1_01* and *v2_01* than on the more difficult *mh_04*, *mh_05*, *v1_02* and *v2_02* sequences.

In a second run, marked with an asterisk (*), the VSLAM uses as offline map the map generated from the corresponding underlined sequence (e.g. *mh_02** used the map from *mh_01* as offline map as indicated in TABLE I). In all cases, the VSLAM was able to improve the pose estimation compared to runs without an offline map, as can be seen from the bold figures in TABLE I.

Sequence	$ATE[m]$	Sequence	$ATE[m]$	$RPE_t[m]$	$RPE_r[deg]$
mh_01	0.177	v1_01	0.138	0.056	0.984
mh_02	0.126	v1_02	0.187	0.182	2.488
mh_02*	**0.121**	v1_02*	**0.124**	**0.145**	**1.97**
mh_04	0.484	v2_01	0.103	0.080	4.501
mh_05	0.389	v2_02	0.273	0.243	9.385
mh_05*	**0.290**	v2_02*	**0.184**	**0.224**	**8.77**

TABLE I: Comparison of estimated VSLAM trajectories to a ground truth with the ATE and RPE metric [13] on the *EuRoC* [2] dataset. Sequences marked with Asterisk (*) use an offline map generated from the corresponding underlined sequence. All values represent the average over multiple executions.

V. CONCLUSIONS

In this paper, we presented a novel VSLAM approach for navigation of an autonomous feeding robot. We applied our algorithm to recordings from a cowshed and showed the successful operation within this application domain. For quantitative evaluation of the approach, we used a community established dataset where ground truth data is available. The results show that our approach of combining an offline map with online mapping successfully improves the accuracy of the pose estimation.

The increased accuracy is achieved by incorporating reliable 3D points from the offline map in order to robustly triangulate new accurate 3D points. Consequently, the new 3D points are inherently aligned to the mission specific world coordinate frame. Furthermore, in the case of lost pose tracking, the robot can relocalize itself in the offline map and continue its mission.

Possible future work is the analysis of the robustness towards application specific environmental conditions such as dust along with further improvements to this aspect and methods for updating the offline map.

REFERENCES

[1] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 298–304.

[2] 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*, vol. 35, no. 10, pp. 1157–1163, 2016.

[3] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "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] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[5] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM Large-scale direct monocular SLAM," in *European Conference on Computer Vision*, 2014, pp. 834–849.

[6] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2017.
[7] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.
[8] P. Moulon, P. Monasse, and R. Marlet, "OpenMVG. an open multiple view geometry library." https://github.com/openMVG/openMVG.
[9] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
[10] W. Sakpere, M. Adeyeye-Oshin, and N. B. Mlitwa, "A state-of-the-art survey of indoor positioning and navigation systems and technologies," *South African Computer Journal*, vol. 29, no. 3, pp. 145–197, 2017.
[11] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient effective prioritized matching for large-scale image-based localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1744–1756, 2017.
[12] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "maplab: An open framework for research in visual-inertial mapping and localization," *arXiv preprint arXiv:1711.10250*, 2018.
[13] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.
[14] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM algorithms: a survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, 2017.
[15] R. Wang, M. Schwörer, and D. Cremers, "Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras," in *IEEE International Conference on Computer Vision*, 2017, pp. 3923–3931.