IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI),
Baden-Baden, Germany, September 2016.

Joint 3D Laser and Visual Fiducial Marker based SLAM for a Micro Aerial Vehicle

Sebastian Houben, David Droeschel, and Sven Behnke

Abstract— **Laser scanners have been proven to provide reliable and highly precise environment perception for micro aerial vehicles (MAV). This oftentimes makes them the first choice for tasks like obstacle avoidance, close inspection of structures, self-localization, and mapping. However, artificial environments may pose problems if the scene is self-similar or symmetric and, hence, localization becomes ambiguous if only relying on distance measurements (e.g., when flying along a parallel aisle).**

In this paper, we propose to tackle these instances by introducing visual fiducial markers into the scene, detecting them with copter-mounted cameras and fusing these detections with laser-based self-localization in a graph optimization. Our approach abstracts the underlying multiple stages of laser-based SLAM to a slim interface that is only connected to the map building process and augments the self-localization in uncertain situations.

We demonstrate the applicability of our approach during experiments in an indoor scenario with sparsely distributed fiducial markers. The test encompasses accurate map building with both the laser scanner and video cameras and subsequent relocalization relying on the detection of fiducial markers only.

I. Introduction

Highly accurate simultaneous localization and mapping (SLAM) is one of the most important capabilities for micro air vehicles (MAV) as nearly all high-level systems like mission and trajectory planning, dynamic replanning, and obstacle avoidance rely on its robust and accurate functioning. Particularly, in GPS-denied scenarios, e.g., during indoor operation, the aforementioned tasks become even more demanding due to lack of free space.

SLAM algorithms that are based on the readings of a rotating laser scanner have been shown to yield highly precise self-localization and dense and accurate detection of permanent structures and dynamic obstacles [1], [2] for many MAV-related applications. The said setup allows for range measurements in virtually all directions, but is, as a caveat, unable to distinguish between detected materials. In consequence, if a single scan does not unambiguously match the map, e.g. in symmetric or repetitive environments, both the self-localization and mapping may suffer from motion drift or worse, wrong map matching and, subsequently, loss of localization. The combination with video cameras via visual SLAM methods might alleviate this problem, however, in particular in indoor environments structures may be repetitive in visual appearance as well.

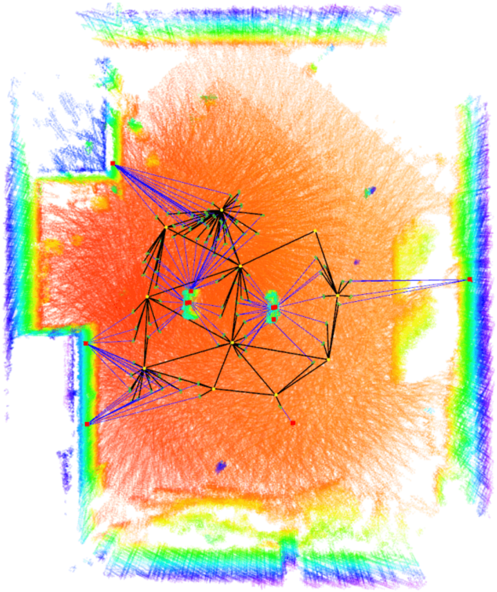


Fig. 1: The SLAM pipeline utilizes the laser scanner readings to continuously extend an accurate high-resolution map of the MAV's surroundings. Cameras are able to detect the visual markers that were sparsely distributed in the outer walls and the inner structure in order to stabilize the system in otherwise ambiguous situations. Yellow nodes depict laser-based keyframes. They are connected by registration constraints to each other and to green tag observation nodes, which are connected by blue observation constraints to the visual fiducials (red squares).

In this paper, we propose to disambiguate these situations by use of visual markers (cf. Fig. 2) that are tracked with one or more cameras. We build upon the well-known AprilTags [3] that provide an artificial aligned planar pattern whose orientation and distance can (due to known size) be estimated from a single image. They carry an encoded integer along with an error correction code to make multiple tags distinct. It is, vice versa, possible to estimate the position of the camera sensor from a single detected tag. This estimation is, however, prohibitively unstable and would require the presence of multiple markers in a close distance to the robot at any time. In combination, a laser scanner based SLAM system can provide a localization prior that is then refined by the detection of a single tag.

Since the detection is lightweight compared to the computational demand of laser-based SLAM, we consider it a

This work has been supported by a grant of the Bundesministerium für Wirtschaft und Energie in the Autonomics for Industry 4.0 project InventAIRy.

All authors are with the Autonomous Intelligent Systems Group, Computer Science Institute VI, University of Bonn, 53113 Bonn, Germany {houben, droeschel, behnke}@ais.uni-bonn.de

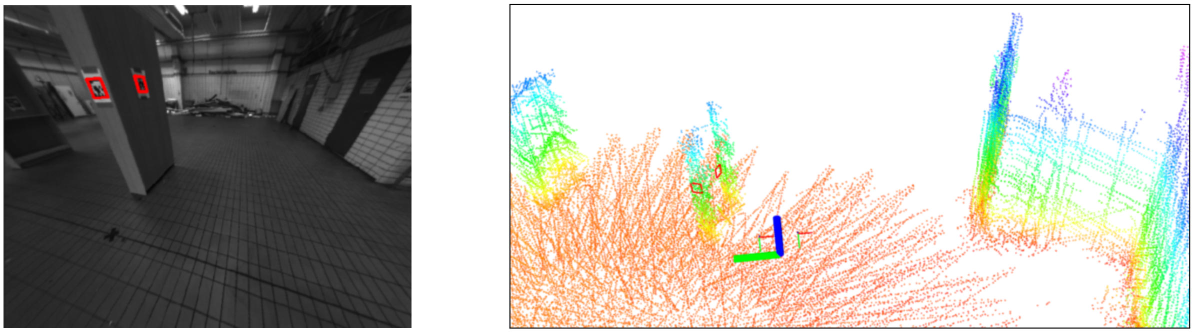


Fig. 2: A camera image taken during the described experiments (left) and the map acquired by the laser scanner up to that point (right). April tags are shown with a red border.

highly beneficial supplement at low cost.

As our two main contributions, we propose how to seamlessly integrate both the laser scanner and the camera sensors in a graph-optimization framework without the need for a detailed sensor model (Sec. III) and demonstrate the effectiveness of our setup by experiments in an indoor scenario that allow to assess the solely video-based localization and relocalization within a known map that was previously recorded using both sensors (Sec. IV).

II. Related Work

For an overview over visual fiducial markers we refer the reader to [4], [5] for the very popular ARTag system and its successors, as well as [6] covering the less common coding systems. In contrast to ARTags, AprilTags [3] are distributed as open-source and show higher robustness against occlusion, lighting conditions, and lens distortion which is why they have also been used for camera calibration [7].

Fiducial markers have been successfully deployed to estimate the pose of static and moving objects in robot experiment setups (cf. e.g., [8]) and camera-tracking in VR settings (cf. [6]). Their use for purely visual SLAM has been demonstrated in [9] and [10] in an EKF-based filter framework, but by design both methods rely on the presence of multiple tags within the field of view at any time.

As a long-term prospect, the automatic disambiguation of self-localization by human-readable text labels is desirable, as it avoids the preparation of the environment for robot operation. First attempts have been presented [11], but are only able to distinguish between a small number of words in real-time. General approaches that do not rely on predefined dictionaries and font systems [12] are promising, but far from applicable in real-time on embedded hardware.

III. Method

Our aim is to combine two types of sensors that have complementary characteristics. The laser-scanner setup yields highly accurate pose estimations with a good estimate of its covariance at a low rate of 2 Hz. Video cameras, on the other hand, are able to provide detections with a much higher frequency ($\approx 30\,\text{Hz}$), but these are sparsely distributed and do, hence, result in a much coarser localization.

We present a graph-based approach with three types of nodes—laser-based keyframes, camera poses, and tag poses—and sparse constraints between them, based on registration of keyframes, laser-based pose tracking, and tag observations. In order to integrate the measurements of the corner points of each detected AprilTag, the formulation aims at minimizing the reprojection error in image coordinates. This holds the assumption that the reprojection error is isotropic and invariant to the position of the tag within the image. Since the relative pose between a camera and a tag bears six degrees of freedom and four coordinates pairs are determined, it would be possible to derive a covariance estimation by bootstrapping the tag detections, but we neglect this possibility for its computational demand and assume a static isotropic uncertainty.

A. Laser-based Localization

We built upon the setup described in our previous work [13] and refer the reader to [14], [15] for related reading. Briefly, two laser-scanners are mounted upon a rotary disk that is rigidly attached to the copter frame and allows to take range measurements in all directions over the course of an entire rotation which takes $0.5\,\text{s}$. The accumulated readings are transformed into a point cloud built around the scanner disk. As the copter moves during scan acquisition, this accumulation requires odometry estimates that can be retrieved from IMU readings and visual odometry. We will refer to such an accumulated point cloud as a 3D scan. Registering a 3D scan with the current scene representation that was built up to that point yields a highly precise copter pose estimate. Aligning all scans to a global map is done in a two-stage hierarchical fashion in order to restrict the computational cost. First, 3D scans are registered to an egocentric 3D map representation that contains the most recent observations. After a given distance, the egocentric map is registered with neighboring egocentric maps to form an allocentric map of laser-based keyframes. The data structures (scan, egocentric map, allocentric map) represent different levels of detail and densities of the accumulated readings. Thus, the computational complexity in order to match each hierarchy level also increases.

We abstract this laser-based localization and mapping pipeline into two crucial items that are used in the sensor fusion approach described below. First, the poses that result from the local-to-global registration are utilized as keyframes in the later graph-optimization. Second, the low-rate ego-localization received by registering a single scan are interpolated in order to obtain estimates for the camera position and orientation at the time of a tag detection.

B. Detection of Fiducial Markers

Olsen et al. [3] describe an efficient algorithm for detection of AprilTags even under severe lens distortion. In our setup, we use a stereo array of three pairs of wide-angle cameras but rectify them to satisfy the pinhole model (cf. [16]). Please also note that the algorithm is able to derive the orientation of the quadratic tag and can, hence, identify all four corners independently from the relative camera rotation. Let s be the size of the quadratic tag and

$$\begin{aligned} P &= \{p_{i,a} : i = 1, .., 4\} \\ &= \left\{(-\tfrac{s}{2}, -\tfrac{s}{2}, 0), (\tfrac{s}{2}, -\tfrac{s}{2}, 0), (\tfrac{s}{2}, \tfrac{s}{2}, 0), (-\tfrac{s}{2}, \tfrac{s}{2}, 0)\right\} \end{aligned}$$

its corner points in an aligned coordinate frame with origin at the tag's center. In the following, the subindices w, a, c, and l denote the coordinate frames *world*, *apriltag*, *camera*, and *laser*, respectively. If the pose of the tag represented as the 4×4 matrix $T_{A,wa} \in SE(3)$ is known, we can compute

$$\begin{pmatrix} \mathbf{p}_{i,w} \\ 1 \end{pmatrix} = T_{A,wa} \begin{pmatrix} \mathbf{p}_{i,a} \\ 1 \end{pmatrix} \text{ for all } \mathbf{p}_{i,a} \in P$$

to obtain the four corner points relative to an allocentric frame w. We define

$$\begin{aligned} \pi(x, y, z) &= (\tfrac{x}{z}, \tfrac{y}{z}) \\ \pi_{\mathcal{C}}(\mathbf{p}_w) &= \pi\left(K_C I_{3\times4} T_{C,cw} \begin{pmatrix} \mathbf{p}_w \\ 1 \end{pmatrix}\right) \end{aligned}$$

as the projection from a point in said world frame to its corresponding image coordinates. $T_{C,cw}$ denotes the pose of the camera C, K_C its calibration matrix. If an AprilTag is detected, the four corner points $\mathbf{x}_{i,c} \in \mathbb{R}^2$ for $i = 1, ..., 4$ are given in image coordinates. Via

$$\begin{aligned} &\pi_{\mathcal{C}}\left(T_{A,wa} \begin{pmatrix} \mathbf{p}_{i,a} \\ 1 \end{pmatrix}\right) \\ &= \pi\left(K_C I_{3\times4} \underbrace{T_{C,cw} T_{A,wa}}_{=:T_{A,ca}} \begin{pmatrix} \mathbf{p}_{i,a} \\ 1 \end{pmatrix}\right) = \mathbf{x}_{i,c} \end{aligned}$$

one sets up eight equations with eight unknowns in the matrix $T_{A,ca}$ to retrieve the relative pose of the detected AprilTag w.r.t. the camera frame.

C. Joint Graph-based SLAM

We propose to fuse the two sensor modalities by a graph-based formulation that represents the sparsity of the cost function that is continuously minimized for copter pose and map estimation. It is composed of three error terms

$$\underset{\mathcal{L},\mathcal{C},\mathcal{A}}{arg\ min} \left\{e_{\mathcal{P}}(\mathcal{L}, \mathcal{C}) + e_{\mathcal{B}}(\mathcal{L}) + e_{\mathcal{D}}(\mathcal{C}, \mathcal{A})\right\}, \qquad (1)$$

where $\mathcal{L}$ and $\mathcal{C}$ are poses obtained for the laser-based localization and the cameras at different points of time, and $\mathcal{A}$ represents the tags whose position is assumed static. $e_{\mathcal{P}}$ and $e_{\mathcal{D}}$ denote the discrepancy between the current configuration $(\mathcal{L}, \mathcal{C}, \mathcal{A})$ and the sensor readings in the respective point of time. All terms will be defined more precisely below. Implementation was done with help of the *g2o* framework [17].

In order to traverse pose hypothesis in a gradient-descent formulation, we denote $tq(T) = (t_x, t_y, t_z, q_x, q_y, q_z)$ the 6-element vector corresponding to the rigid transform T. It contains the translational part (t_x, t_y, t_z) and the vector part of the unit quaternion describing the change in orientation. Thus, one retrieves the full quaternion as

$$(q_w = 1 - \sqrt{q_x^2 + q_y^2 + q_z^2}, q_x, q_y, q_z).$$

Since the gradient descent is performed in this six-dimensional vector space, an expression for tq^{-1} is needed in order to obtain the respective transformation matrix. It is given in Equation (4). A numerically stable algorithm to compute $tq(T)$ is stated in [18].

The graph is constructed from three types of vertices (cf. Fig. 3):

- $\mathcal{L}$ denotes the set of laser-based keyframes that are added to the graph whenever a registration of the local and the global map has been performed. The vertices $L = (T_{L,lw}) \in \mathcal{L}$ carry the copter pose $T_{L,lw}$ at the time of the registration.
- $\mathcal{C}$ denotes the set of camera keyframes that are added to the graph whenever a previously unknown tag is detected or a known tag as been observed from a significantly distant position. The vertices $C = (T_{C,cw}, K_C)$ carry the camera pose $T_{C,cw}$ as well as the calibration matrix K_C of the respective camera. Please note that K_C is required to define optimization constraints (see below), but remains itself fixed for the course of the optimization. It can be obtained beforehand by means of camera calibration.
- $\mathcal{A}$ denotes the set of AprilTags which are added to the graph whenever a new tag is observed by a camera. The vertices $A = (T_{A,aw}, P)$ carry their pose and the corner points (cf. Sec. III-B) that are also assumed fixed.

Two types of edges define constraints among the vertices:

- The set $\mathcal{E}_D$ is composed of edges $E = ((C_E, A_E), \mathbf{x}_{1,E,c}, ..., \mathbf{x}_{4,E,c})$ that impose a constraint due to the detection of a tag $A_E \in \mathcal{A}$ in camera keyframe $C_E \in \mathcal{C}$. A detection is given by the four corner points of the tag $\mathbf{x}_{1,E,c}, ..., \mathbf{x}_{4,E,c}$ in image coordinates (cf. Sec. III-B)
- The set $\mathcal{E}_P$ is composed of edges $E = ((L_E, C_E), T_{E,lc}, I_E)$ imposing a relative pose constraint between a laser-based keyframe $L_E \in \mathcal{L}$ and a camera keyframe $C_E \in \mathcal{C}$. Each edge carries the relative pose as well as the information matrix $I_E \in \mathbb{R}^{6\times6}$ as the inverse of the pose covariance w.r.t. the vector representation $tq(T_{E,lc})$.

- The set $\mathcal{E}_B$ with edges $E = ((L_E, L'_E), T_{E,ll}, I_E)$ is defined analogously to $\mathcal{E}_P$ and imposes constraints between laser-based keyframes $L, L' \in \mathcal{L}$ whose local maps overlap.

With this notation set, we now specify Eq. (1):

$$\underset{\substack{T_L, T_C, T_A \\ \text{with} \\ (T_L) \in \mathcal{L} \\ (T_C, K_C) \in \mathcal{C} \\ (T_A, P_A) \in \mathcal{A}}}{arg\ min} \{e_{\mathcal{P}}(\mathcal{L}, \mathcal{C}, \mathcal{E}_{\mathcal{P}}) + e_{\mathcal{B}}(\mathcal{L}, \mathcal{E}_{\mathcal{B}}) + e_{\mathcal{D}}(\mathcal{C}, \mathcal{A}, \mathcal{E}_{\mathcal{D}})\} \tag{2}$$

$$\begin{aligned}
e_{\mathcal{P}}(\mathcal{L}, \mathcal{C}, \mathcal{E}_{\mathcal{P}}) &= \sum_{L \in \mathcal{L}} \sum_{E \in N_{\mathcal{P}}(L)} d(tq(T_E^{-1} T_L^{-1} T_{C_E}), I_E) \\
e_{\mathcal{B}}(\mathcal{L}, \mathcal{E}_{\mathcal{B}}) &= \sum_{L \in \mathcal{L}} \sum_{E \in N_{\mathcal{B}}(L)} d(tq(T_E^{-1} T_L^{-1} T_{L'_E}), I_E)
\end{aligned}$$

$$e_{\mathcal{D}}(\mathcal{C}, \mathcal{A}, \mathcal{E}_{\mathcal{D}}) =$$

$$\sum_{C \in \mathcal{C}} \sum_{E \in N_{\mathcal{D}}(C)} \sum_{i=1}^{4} d\left(\pi_C\left(T_{A_E,wt}\begin{pmatrix}\mathbf{p}_{i,a} \\ 1\end{pmatrix}\right)\right) - \mathbf{x}_{i,E,c}, I_E\Bigg)$$

where d denotes the Mahalanobis metric

$$d : \mathbb{R}^n \times \mathbb{R}^{n \times n} \to \mathbb{R} : d(x, I) \mapsto x^T I x.$$

and $N_{\mathcal{B}}(V), N_{\mathcal{P}}(V), N_{\mathcal{D}}(V)$ refer to the subset of edges from $\mathcal{E}_{\mathcal{B}}, \mathcal{E}_{\mathcal{P}}, \mathcal{E}_{\mathcal{D}}$, respectively, that are incident to V.

D. Camera-only Localization

As camera images arrive at a much higher rate than laser-based localization can be computed it can be necessary to estimate the pose of a camera solely with the help of its marker detections. To this end, one computes

$$\underset{T_C \text{ with } (T_C, K_C) \in \mathcal{C}}{arg\ min} \quad \{e_{\mathcal{D}}(\{T_C\}, \mathcal{A}, \mathcal{E}_{\mathcal{D}})\}. \tag{3}$$

E. Implementation Details

As the minimum of the cost function from (2) is invariant under rigid transformations, one pose (usually the pose of the copter when initialized) is fixed and excluded from the optimization. The graph construction process is embedded straightforwardly into a relocalization task, i.e., retrieving the current pose within a known map. At the start of the system, the mapping and AprilTag detection is executed as usual. If an already known tag is perceived, an according edge between the new and the previously known graph is added. Unfixing the anchor pose will now align both graphs during the optimization phase.

IV. Experiments

We demonstrate the effectiveness of our approach with the help of a hexacopter (cf. Fig. 4) equipped with six wide angle cameras and the spinning laser scanner disk described in section III-A. The cameras possess a field of view of more than 180° and show, hence, severe image distortion. For the experiments either two cameras of a stereo pair were calibrated extrinsically and only a region around the optical center was undistorted and rectified leaving the effective field of view at 120°. An example is shown in Fig. 2. This allows to use the pinhole model during the optimization stage (cf. Sec. III). The calibration among cameras and laser scanner is derived from the CAD model of the copter.

In the regarded scenario, the copter is manually flown along a course of approximately 25 m inside a factory building at moderate velocity of up to $1\,\mathrm{m\,s^{-1}}$. In sparsely distributed locations, AprilTags are pasted at the height of the flying copter and on the ground. Figure 2 shows a part of the experimental setup.

The used AprilTags were printed from the family *36h11* [3] at a side length of 0.163 m. This parametrizes binary patterns with a pairwise Hamming distance of at least 11 bit which enormously robustifies the detection stage. In fact, we did never experience false positive detections or false decoding of a correctly detected tag.

In the first part of the experiment, the proposed SLAM algorithm was deployed to construct a map with both the laser scanner and the six video cameras pinpointing the positions of the detected AprilTags within this map. A result is shown in Fig. 1. The course contained a loop closing which allows the laser-based mapping to be globally consistent. In the second part, single video frames were extracted from the sequence and a localization via AprilTags was performed. This setup is supposed to mimic a relocalization task where the copter has to find its position with respect to a known map. To assess the quality of the camera-derived pose, a comparison with the laser-based self-localization was performed. This reference positioning is known to be highly accurate. The precision is reported in the range of a few centimeters [2].

Figure 5 covers the evaluation of the pose estimated from single camera frames. In total, 13 AprilTags were detected in 749 images. Due to the sparse distribution of tags 697 frames contained a single detection and 50 contained two of them. Tags were detected at a distance from 0.61 m to 4.99 m. Figures 5(a) and (b) show the distributions of the relocalization error. The average accuracy is at 0.50 m±0.85 m (median: 0.19 m) / 10°±15° (median: 4°) for frames with a single detection and 0.78 m±1.22 m (median: 0.22 m) / 14°±21° (median: 4°) for frames with two tag detections. Indeed, both distribution show a clear mode at a reasonable accuracy, but suffer from strong outliers. The distance estimation to the detected tag, which is depicted in Fig. 5(c) is performed very accurately at 3.15 % / 3.82 % on average. For the scope of our experiments no clear correlation between self-localization accuracy and detection characteristics like distance to the tag or view angle could be established (cf. Fig. 5(d), (e), (f)). Also, all tags show a very similar error distribution leaving none of them particularly unsuited or hard to use for relocalization.

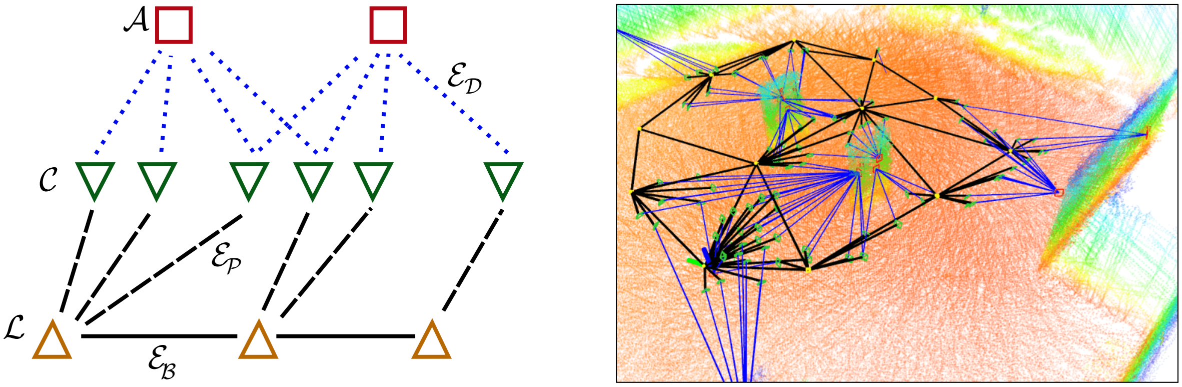

$\mathcal{A}$
$\mathcal{E}_{\mathcal{D}}$
$\mathcal{C}$
$\mathcal{E}_{\mathcal{P}}$
$\mathcal{L}$
$\mathcal{E}_{\mathcal{B}}$

Fig. 3: Graph-representation of the optimization problem at hand: The vertex sets $\mathcal{A}, \mathcal{C}$, and $\mathcal{L}$ correspond to the estimation variables, the edge sets $\mathcal{E}_{\mathcal{P}}$ and $\mathcal{E}_{\mathcal{B}}$ to pose constraints, and $\mathcal{E}_{\mathcal{D}}$ to a reprojection constraint. The same color coding is used in both depictions.

$$tq^{-1}(t,q) = \begin{pmatrix} q_w^2 + q_x^2 - q_y^2 - q_z^2 & 2(q_x q_y - q_w q_z) & 2(q_z q_x + q_w q_y) & t_x \\ 2(q_x q_y + q_w q_z) & q_w^2 - q_x^2 + q_y^2 - q_z^2 & 2(q_y q_z - q_w q_x) & t_y \\ 2(q_z q_x - q_w q_y) & 2(q_y q_z + q_w q_x) & q_w^2 - q_x^2 - q_y^2 + q_z^2 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4}$$

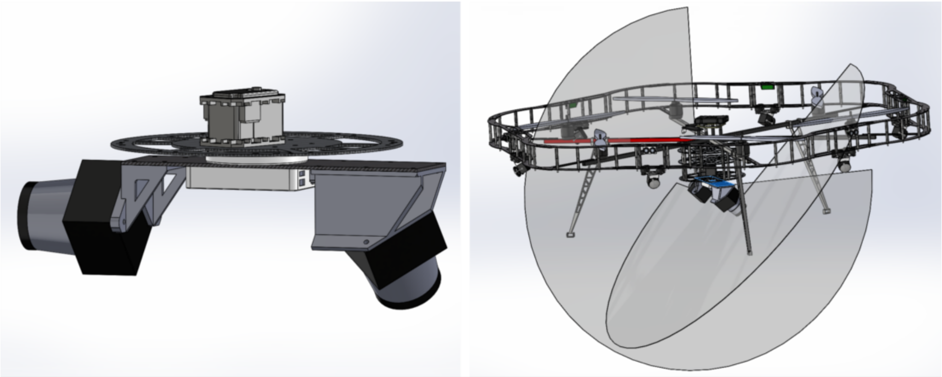


Fig. 4: Scheme of the deployed copter (right) and the laser scanner (left). The spinning disk with mutually skewed two laser scanners and their respective scan range is depicted below the central core. The outer mounting carries three stereo pairs of wide-angle cameras.

V. Conclusions

We have demonstrated that fiducial markers present a viable approach to supplement laser-based self-localization and mapping algorithms. In particular, AprilTag patterns provide a simple and fast method to amend a location with unique markers that can be detected and decoded efficiently and very robustly. These properties make them ideal candidates to supplement SLAM approaches. The presented graph-optimization ansatz allows a seamless and continuous integration of their position and information in a detailed map. We have demonstrated that this map is eligible to allow for a solely camera-based relocalization although additional filtering due to the presence of detection outliers will be necessary. Indeed, we experienced that particularly if the estimation of the tag is poor and a wrong tilt is perceivable, the localization error grows rapidly.

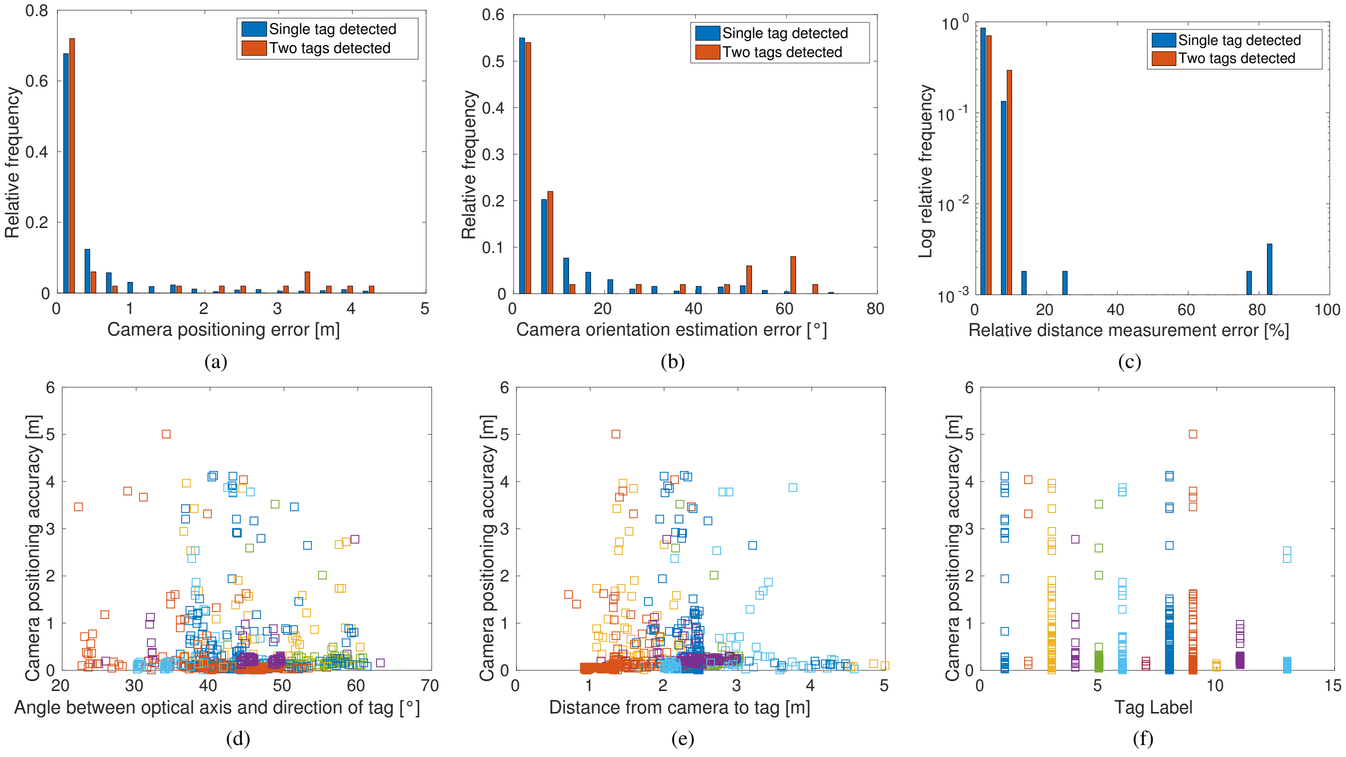
In the future, we plan to investigate methods to integrate the map by the laser scanner and fiducial markers more directly. We expect the knowledge of planar surfaces in the scene to provide helpful priors for initializing newly detected tags dramatically reducing the problem of wrong tilt.

Acknowledgment

We thank Yongfeng Zhong for providing the interface to the AprilTag detection libraries and Nicola Krombach, Marius Beul, and Matthias Nieuwenhuisen for help and support during the experiments.

References

[1] D. Droeschel, M. Nieuwenhuisen, M. Beul, D. Holz, J. Stückler, and S. Behnke, "Multilayered mapping and navigation for autonomous



(a) (b) (c)

(d) (e) (f)

Fig. 5: Assessment of the solely video-based relocalization by means of a map from the proposed laser-supported SLAM approach.
(a) – (b): Distribution of localization and orientation error for video-based relocalization by means of a single image with the help of one or two detected tags
(c): Distribution of the error for the distance estimate from the camera to the detected tag
(d) – (f): Correlation between positioning error during relocalization and view angle / distance to detected tag, and tag id for reference. In all plots, the color encodes the estimations based on the same tag, respectively.

micro aerial vehicles," *Journal of Field Robotics*, vol. 33, no. 4, pp. 451–475, 2016.

[2] D. Droeschel, D. Holz, and S. Behnke, "Omnidirectional perception for lightweight mavs using a continously rotating 3d laser scanner," *Photogrammetrie Fernerkundung Geoinformation*, vol. 5, pp. 451–464, 2014.

[3] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *International Conference on Robotics and Automation*. IEEE, 2011, pp. 3400–3407.

[4] M. Fiala, "Artag, a fiducial marker system using digital techniques," in *Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2005, pp. 590–596.

[5] H. Kato and M. Billinghurst, "Marker tracking and hmd calibration for a video-based augmented reality conferencing system," in *International Workshop on Augmented Reality*. IEEE, 1999, pp. 85–94.

[6] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[7] A. Richardson, J.-P. Strom, and E. Olson, "Aprilcal: Assisted and repeatable camera calibration," in *International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1814–1821.

[8] M. Rubenstein, A. Cabrera, J. Werfel, G. Habibi, J. McLurkin, and R. Nagpal, "Collective transport of complex objects by simple robots: Theory and experiments," in *International Conference on Autonomous Agents and Multi-agent Systems*, 2013, pp. 47–54.

[9] T. Yamada, T. Yairi, S. H. Bener, and K. Machida, "A study on slam for indoor blimp with visual markers," in *ICROS-SICE International Joint Conference*. IEEE, 2009, pp. 647–652.

[10] H. Lim and Y. S. Lee, "Real-time single camera slam using fiducial markers," in *ICROS-SICE International Joint Conference*. IEEE, 2009, pp. 177–182.

[11] H.-C. Wang, C. Finn, L. Paull, M. Kaess, R. Rosenholtz, S. Teller, and J. Leonard, "Bridging text spotting and slam with junction features," in *International Conference on Intelligent Robots and Systems*. IEEE, 2015, pp. 3701–3708.

[12] L. Gomez-Bigorda and D. Karatzas, "Textproposals: A text-specific selective search algorithm for word spotting in the wild," *arXiv preprint arXiv:1604.02619*, 2016.

[13] M. Beul, N. Krombach, Y. Zhong, D. Droeschel, M. Nieuwenhuisen, and S. Behnke, "A high-performance mav for autonomous navigation in complex 3d environments," in *International Conference on Unmanned Aircraft Systems*. IEEE, 2015, pp. 1241–1250.

[14] J. Zhang and S. Singh, "Visual-lidar odometry and mapping: Low-drift, robust, and fast," in *International Conference on Robotics and Automation*. IEEE, 2015, pp. 2174–2181.

[15] D. Droeschel, J. Stückler, and S. Behnke, "Local multi-resolution representation for 6d motion estimation and mapping with a continuously rotating 3d laser scanner," in *International Conference on Robotics and Automation*. IEEE, 2014, pp. 5221–5226.

[16] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[17] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *International Conference on Robotics and Automation*, 2011, pp. 3607–3613.

[18] I. Y. Bar-Itzhack, "New method for extracting the quaternion from a rotation matrix," *Journal of Guidance, Control, and Dynamics*, vol. 23, no. 6, pp. 1085–1087, 2000.