

The ETH-MAV Team in the MBZ International Robotics Challenge

Rik Bähnemann* Michael Pantic* Marija Popović* Dominik Schindler*

Marco Tranzatto* Mina Kamel Marius Grimm Jakob Widauer Roland Siegwart

Juan Nieto

Autonomous Systems Lab (ASL)
ETH Zurich - Swiss Federal Institute of Technology
Zurich, Switzerland
Corresponding author: brik@ethz.ch

Abstract

This article describes the hardware and software systems of the Micro Aerial Vehicle (MAV) platforms used by the ETH Zurich team in the 2017 Mohamed Bin Zayed International Robotics Challenge (MBZIRC). The aim was to develop robust outdoor platforms with the autonomous capabilities required for the competition, by applying and integrating knowledge from various fields, including computer vision, sensor fusion, optimal control, and probabilistic robotics. This paper presents the major components and structures of the system architectures, and reports on experimental findings for the MAV-based challenges in the competition. Main highlights include securing second place both in the individual search, pick, and place task of Challenge 3 and the Grand Challenge, with autonomous landing executed in less than one minute and a visual servoing success rate of over 90% for object pickups.

Supplementary Material

For a supplementary video see: <https://youtu.be/DXYFAkjHeho>.
For open-source components visit: <https://github.com/ethz-asl>.

1 Introduction

The Mohamed Bin Zayed International Robotics Challenge (MBZIRC) is a biennial competition aiming to demonstrate the state-of-the-art in applied robotics and inspire its future. The inaugurating event took place on March 16-18, 2017 at the Yas Marina Circuit in Abu Dhabi, UAE, with total prize and sponsorship money of USD 5M and 25 participating teams. The competition consisted of three individual challenges and a triathlon-type Grand Challenge combining all three in a single event. Challenge 1 required an Micro Aerial Vehicle (MAV) to locate, track, and land on a moving vehicle. Challenge 2 required a Unmanned

*The authors contributed equally to this work. Their names are listed in alphabetical order.

Ground Vehicle (UGV) to locate and reach a panel, and operate a valve stem on it. Challenge 3 required a collaborative team of MAVs to locate, track, pick, and deliver a set of static and moving objects on a field. The Grand Challenge required the MAVs and UGV to complete all three challenges simultaneously.

This paper details the hardware and software systems of the MAVs used by our team for the relevant competition tasks (Challenges 1 and 3, and the Grand Challenge). The developments, driven by a diverse team of researchers from the Autonomous Systems Lab (ASL), sought to further multi-agent autonomous aerial systems for outdoor applications. Ultimately, we achieved second place in both Challenge 3 and the Grand Challenge. A full description of the approach of our team for Challenge 2 is described in (Carius et al., 2018).

The main challenge we encountered was building robust systems comprising the different individual functionalities required for the aforementioned tasks. This led to the development of two complex system pipelines, advancing the applicability of outdoor MAVs on both the level of stand-alone modules as well as complete system integration. Methods were implemented and interfaced in the areas of precise state estimation, accurate position control, agent allocation, object detection and tracking, and object gripping, with respect to the competition requirements. The key elements of our algorithms leverage concepts from various areas, including computer vision, sensor fusion, optimal control, and probabilistic robotics. This paper is a systems article describing the software and hardware architectures we designed for the competition. Its main contributions are a detailed report on the development of our infrastructure and a discussion of our experimental findings in context of the MBZIRC. We hope that our experiences provide valuable insights into outdoor robotics applications and benefit future competing teams.

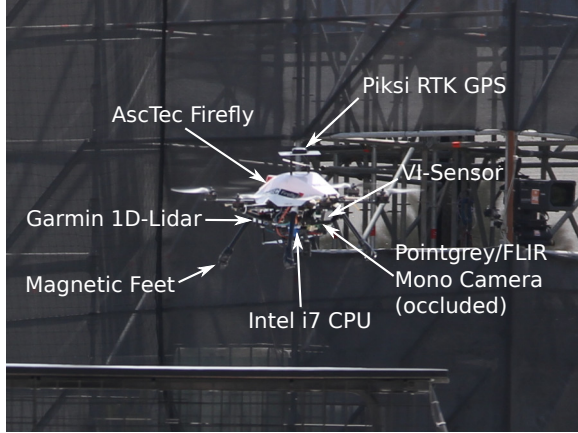
This paper is organized as follows: our MAV platforms are introduced in Section 2, while state estimation and control methods are presented in Sections 3 and 4, respectively. These elements are common to both challenges. Sections 5 and 6 detail our approaches to each challenge. Section 7 sketches the development progress. We present results obtained from our working systems in Section 8 before concluding in Section 9.

2 Platforms

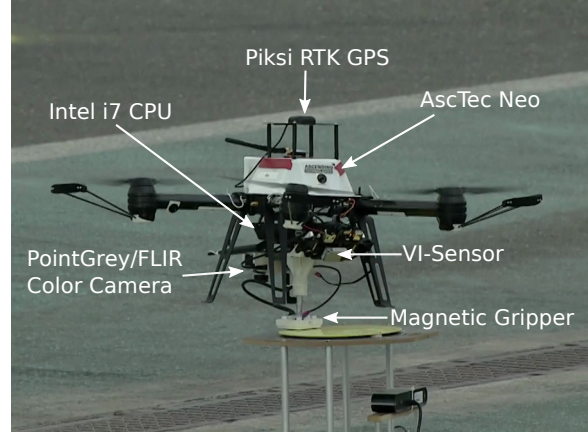
Our hardware decisions were led by the need to create synergies between Challenge 1 and Challenge 3, as well as among previous projects in our group. Given our previous experience with and the availability of Ascending Technologies (AscTec) multirotor platforms, we decided to use those for both challenges (Figure 1). In Challenge 1, we used an AscTec Firefly hexacopter with an AscTec AutoPilot (Figure 1a). In Challenge 3, we used three AscTec Neo hexacopters with AscTec Trinity flight controllers (Figure 1b). Both autopilots provide access to on-board filtered Inertial Measurement Unit (IMU) and magnetometer data, as well as attitude control. An integrated safety switch allows for taking over remote attitude control in cases where autonomous algorithms fail. The grippers have integrated Hall sensors for contact detection, as described in detail in Section 6.4.

All platforms are equipped with a *perception unit* which uses similar hardware for state estimation and operating the software stack. This unit is built around an on-board computer with Intel i7 processor running Ubuntu and ROS as middleware to exchange messages between different modules. The core sensors for localization are the on-board IMU, a slightly downward-facing VI-Sensor, and a Piksi RTK GPS receiver. The VI-Sensor was developed by the ASL and Skybotix AG (Nikolic et al., 2014) to obtain fully time-synchronized and factory calibrated IMU and stereo-camera datastreams, and a forward-facing monocular camera for Visual-Inertial Odometry (VIO). Piksi hardware version V2 is used as the RTK receiver (Piksi Datasheet V2, 2017). RTK GPS is able to achieve much higher positioning precision by mitigating the sources of error typically affecting stand-alone GPS, and can reach an accuracy of a few centimeters.

Challenge-specific sensors, on the other hand, are designed to fulfill individual task specifications and thus differ for both platform types. In Challenge 1, the landing platform is detected with a downward-facing



(a) The AscTec Firefly hexacopter shortly before landing on the moving target (Challenge 1). For the landing task it has a downward-facing monocular camera and Lidar, and custom landing gear.



(b) One of three identical AscTec Neo hexacopters grasping a moving object (Challenge 3). For aerial gripping each drone has a high resolution color camera and an EPM gripper with Hall sensors.

Figure 1: The MAV platforms used by the ETH team. All MAVs have a VI-Sensor and RTK GPS for state estimation. Computation is done on-board using an Intel i7 CPU running Ubuntu and ROS.

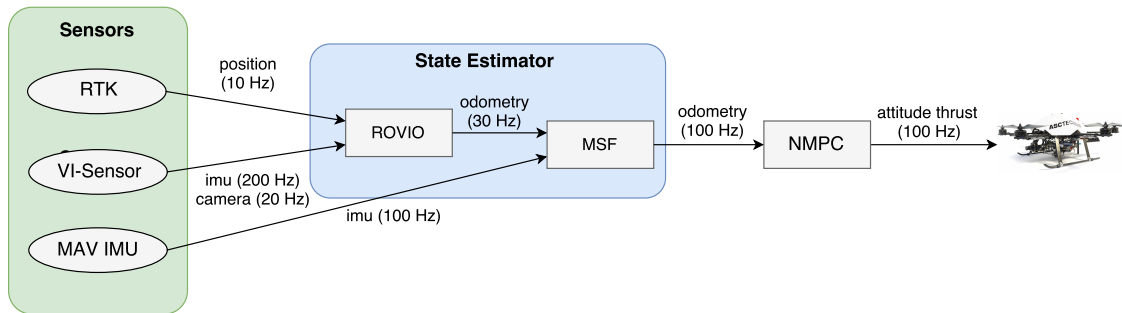


Figure 2: The state estimator comprises Robust Visual Inertial Odometry (ROVIO), which fuses visual-inertial data with GPS positions, and Multi Sensor Fusion (MSF) which robustly fuses the pose estimate from ROVIO with the on-board IMU for control.

PointGrey/FLIR Chameleon USB 2.0 monocular camera with 1.3MP and a fisheye lens (Chameleon 2 Datasheet, 2017). The distance to the platform is detected with a Garmin LIDAR-Lite V3 Lidar sensor (Lidar Datasheet, 2017). Moreover, additional commercial landing gear is integrated in the existing platform. For Challenge 3, we equip each MAV with a downward-facing PointGrey/FLIR 3.2 MP Chameleon USB 3.0 color camera to detect objects (Chameleon 3 Datasheet, 2017). For object gripping, the MAVs are equipped with modified NicaDrone OpenGrab EPM grippers (NicaDrone OpenGrab EPM Datasheet v3, 2017).

3 State Estimation

Precise and robust state estimation is a key element for executing the fast and dynamic maneuvers required by both challenges. This section presents our state estimation pipeline, which consists of a cascade of Extended Kalman Filters (EKFs), denoted by the blue box in Figure 2. The following subsections detail our architecture and the major design decisions behind it. The last subsection highlights some insights we attained when integrating RTK GPS.

Symbol	Name	Origin	Description
A	<i>arena</i>	Center of arena	Fixed orientation and origin with respect to <i>enu</i> frame
B	<i>mav-imu</i>	MAV-IMU	Aligned with MAV-IMU axes
C	<i>camera</i>	Focal point camera	z -axis pointing out from the lens
E	<i>enu</i>	Center of the arena	Local tangent plane, aligned with ENU directions
G	<i>gripper</i>	Center gripper surface	z -axis pointing out from gripper surface
I	<i>vi-imu</i>	VI-Sensor-IMU	Aligned with VI-Sensor IMU axes
L	<i>lidar</i>	Lidar lens	z -axis pointing out from the lens
O	<i>odom</i>	MAV starting position	Aligned with <i>enu</i> frame after ROVIO reset
T	<i>target</i>	Landing target center	Platform center estimated by the Challenge 1 tracker

Table 1: Description of the frames used.

3.1 Conventions and Notations

In this paper, we utilize the following conventions and notations: ${}^A\mathbf{p}_{B,C}$ refers to a vector from point B to point C expressed in coordinate frame A . The homogeneous transformation matrix ${}^A\mathbf{T}_{B,C} \in \mathbb{R}^{4 \times 4}$ expressed in A converts a homogeneous vector ${}^C\mathbf{p}_{C,D}$ to the vector ${}^A\mathbf{p}_{B,D}$, i.e., ${}^A\mathbf{p}_{B,D} = {}^A\mathbf{T}_{B,C} {}^C\mathbf{p}_{C,D}$. Coordinate frame names are typed lower case italic, e.g., *enu* corresponds to the coordinate frame representing the local tangent plane aligned with East-North-Up (ENU) directions. Table 1 presents a complete description of all frames mentioned in the remainder of this work.

3.2 ROVIO

The first block of the pipeline is a monocular VIO estimator called ROVIO (Bloesch et al., 2015). It achieves accurate tracking performance by leveraging the pixel intensity errors of image patches, which are directly used for tracking multilevel patch features in an underlying EKF. Its software implementation is open-source and publicly available (Rovio Github, 2017). In its standard configuration, ROVIO takes as input from the VI-Sensor one monocular camera stream synchronized with IMU and outputs odometry data. An odometry message is composed of information about position, orientation, angular, and linear velocities of a certain frame with respect to another. This filter outputs odometry messages regarding the state of the VI-Sensor IMU frame, *vi-imu*, with respect to the *odom* frame, which is gravity-aligned with the origin placed at the MAV initialization point.

3.3 RTK GPS as External Pose in ROVIO

In our configuration, ROVIO fuses VIO information, which is locally accurate but can diverge in the long-term (Scaramuzza and Fraundorfer, 2011), with accurate and precise RTK measurements, which arrive at a low frequency but do not suffer from drift (Kaplan, 2005). Our system involves two RTK receivers: a base station and at least one MAV. The base station is typically a stationary receiver configured to broadcast RTK corrections, often through a radio link. The MAV receiver is configured through the radio pair to receive these corrections from the base station, and applies them to solve for a centimeter-level accurate vector between the units.

RTK measurements, expressed as (*latitude, longitude, altitude*), are first converted to a local Cartesian coordinate system and then transmitted to ROVIO as external position measurements. ENU is set as the local reference frame *enu* with an arbitrary origin close to the MAV initialization position. This fusion procedure

prevents ROVIO odometry from drifting, as it is continuously corrected by external RTK measurements. Moreover, the output odometry can be seen as a “global” state, since it conveys information about the *vi-imu* frame with respect to the *enu* frame. This “global” property is due to the independent origin and orientation of the *enu* frame with respect to the initial position and orientation of the MAV. In order to successfully integrate VIO and RTK, the *odom* and *enu* frames must be aligned so that the previous two quantities are expressed in the same coordinate system. To do so, we exploit orientation data from the on-board IMU.

Given the transformations from the *vi-imu* frame I to the *mav-imu* frame B ${}_B\mathbf{T}_{B,I}$ obtained using the Kalibr framework (Kalibr Github, 2017) and the transformation from *mav-imu* frame B to the *enu* frame E ${}_E\mathbf{T}_{E,B}$ obtained from RTK measurements, the aforementioned alignment operation is performed by imposing:

$${}_O\mathbf{T}_{O,I} = {}_E\mathbf{T}_{E,B} {}_B\mathbf{T}_{B,I}. \quad (1)$$

The resulting local *enu* positions, expressed by (*east, north, up*), can be directly used by ROVIO. Note that, while the orientation of these two frames is the same, they can have different origins.

3.4 MSF

The second block of the pipeline consists of the MSF framework (Lynen et al., 2013): an EKF able to process delayed measurements, both relative and absolute, from a theoretically unlimited number of different sensors and sensor types with on-line self-calibration. Its software implementation is open-source and publicly available (MSF Github, 2017).

MSF fuses the odometry states, output by ROVIO at ~ 30 Hz, with the on-board IMU, as shown in Figure 2. This module improves the estimate of the current state from ROVIO by incorporating the inertial information available from the flight control unit. The final MSF output state is a high-frequency odometry message at ~ 100 Hz expressing the transformation, and angular and linear velocities of the MAV IMU frame, *mav-imu*, with respect to the *enu* frame, providing a “global” state estimate. This odometry information is then conveyed to the MAV position controller.

3.5 A Drift-Free and Global State Estimation Pipeline

There were two main motivations behind the cascade configuration (Figure 2): (i) obtaining a global estimate of the state, and (ii) exploiting the duality between the local accuracy of VIO and the driftless property of RTK.

A global state estimate describes the current status of the MAV expressed in a fixed frame. We use the *enu* frame as a common reference, as the MAVs fly in the same, rather small, area simultaneously. This permits sharing odometry information between multiple agents, since they are expressed with respect to a common reference. This exchange is a core requirement in Challenge 3 and the Grand Challenge, so that each agent can avoid and maintain a minimum safety distance given the position and traveling directions of the others, as detailed in Section 4 and Section 6.2. Moreover, this method enables the MAV in Challenge 1 to make additional prior assumptions for platform tracking, as explained in Section 5.2.

The integration of ROVIO and MSF with RTK GPS ensures the robustness of our state estimation pipeline against known issues of the two individual systems. A pure VIO state estimator cascading ROVIO and MSF, without any external position estimation, would suffer from increasing drift in the local position, but still be accurate within short time frames while providing a high-rate output. On the other hand, RTK GPS measurements are sporadic (~ 10 Hz), and the position fix can be lost, but the estimated geodetic coordinates have precise 1 to 5 cm horizontal position accuracy, 8 to 15 cm vertical position accuracy, and do not drift over time (Piksi Accuracy, 2017). This combined solution achieves driftless tracking of the current state, mainly due to RTK GPS, while providing a robust solution against RTK fix loss, since VIO alone

computes the current status until an external RTK position is available again. As a result, during the entire MBZIRC, our MAVs operated without any localization issues.

3.6 RTK GPS Integration

RTK GPS was integrated on the MAVs five months before the MBZIRC, and evolved during our field trials. The precision of this system, however, is accompanied by several weaknesses. Firstly, the GPS antenna must be placed as far away as possible from any device and/or cable using USB 3.0 technology. USB 3.0 devices and cables may interfere with wireless devices operating in the 2.4 GHz ISM band (USB 3.0 Interference, 2017). Many tests showed a significant drop in the signal-to-noise ratio of GPS L1 1.575 GHz carrier. A Software Defined Radio (SDR) receiver was used to identify noise sources. Simple solutions to this problem were either to remove any USB 3.0 device, to increase the distance of the antenna from any component using USB 3.0, or add additional shielding between antenna and USB 3.0 devices and cables. Secondly, the throughput of corrections sent from RTK base station to the MAVs must be kept as high as possible. We addressed this by establishing redundant communication from the base station and the MAV: corrections are sent both over a 5 GHz WiFi network as User Datagram Protocol (UDP) subnet broadcast as well as over a 2.4 GHz radio link. In this way, if either link experienced connectivity issues, the other could still deliver the desired correction messages. Finally, the time required to gain a RTK fix is highly dependent on the number and signal strength of common satellites between the base station and the MAVs. Our experience showed that an average number of over eight common satellites, with a signal strength higher than 40 dB – Hz, yields an RTK fix in ~ 10 min¹.

ASL released the ROS driver used during MBZIRC, which is available on-line (Piksi Github, 2018). This on-line repository contains ROS drivers for Pixsi V2.3 hardware version and for Pixsi Multi. Moreover it includes a collection of utilities, such as a ROS package that allows fusing RTK measurements into ROVIO. The main advantages of our ROS driver are: WGS84 coordinates are converted into ENU and output directly from the driver; RTK corrections can be sent both over a radio link and Wifi. This creates a redundant link for streaming corrections, which improves the robustness of the system.

4 Reactive and Adaptive Trajectory Tracking Control

In order for the MAVs to perform complex tasks, such as landing on fast-moving platforms, precisely picking up small objects, and transporting loads, robust, high-bandwidth trajectory tracking control is crucial. Moreover, in the challenges, up to four MAVs, a UGV, and several static obstacles are required to share a common workspace, which demands an additional safety layer for dynamic collision avoidance.

Our proposed trajectory tracking solution is based on a standard cascaded control scheme where a slower outer trajectory tracking control loop generates attitude and thrust references for a faster inner attitude control loop (Achtelik et al., 2011). The AscTec autopilot provides an adaptive and reliable attitude controller. For trajectory control, we use a Nonlinear Model Predictive Controller (NMPC) developed at our lab (MAV Control Github, 2017; Kamel et al., 2017b; Kamel et al., 2017a).

The trajectory tracking controller provides three functionalities: (i) it tracks the reference trajectories generated by the challenge submodules, (ii) it compensates for changes in mass and wind with an EKF disturbance observer, and (iii) it uses the agents’ global odometry information for reactive collision avoidance. For the latter, the controller includes obstacles into its trajectory tracking optimization that guarantees safety distances between the agents (Kamel et al., 2017a). The signal flow is depicted in Figure 3.

The obstacle avoidance feature and our global state estimation together serve as a safety layer in Challenge 3 and the Grand Challenge, where we program the MAVs to avoid all known obstacles and agents. This

¹With the new Pixsi Multi GNSS Module, RTK fixes are obtained in 3 min on average.

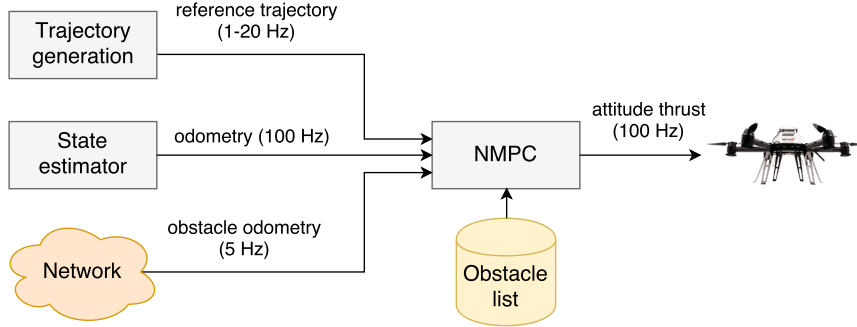


Figure 3: The trajectory control flow used in both challenges. The controller allows trajectory tracking as well as reactive collision avoidance. If an obstacle is listed in the loaded obstacle list and its state is broadcasted over the network, the NMPC guarantees a minimum distance to the object.

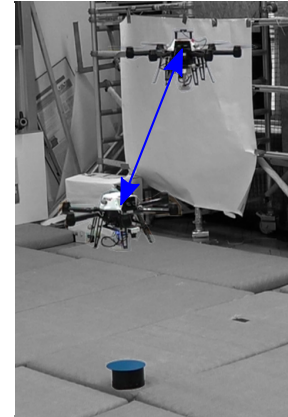
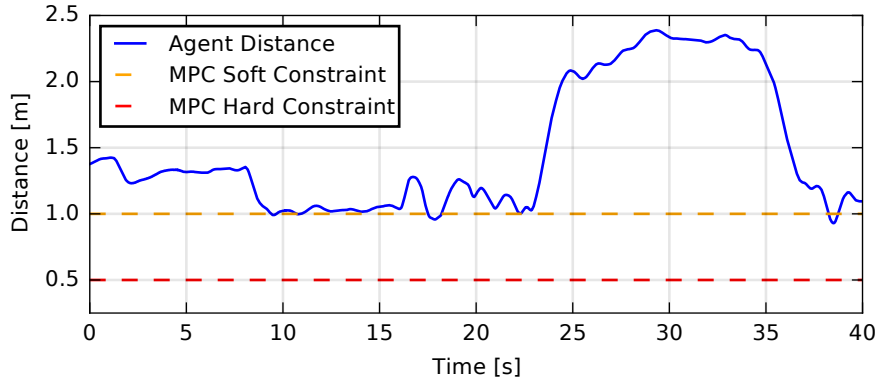


Figure 4: Two MAVs attempting to pick up the same object in a motion capture environment. Sharing global position over the wireless network allows the NMPC to prevent collisions without explicitly communicating the MAV intentions.

paradigm ensures maintaining a minimum distance to an obstacle, provided its odometry and name are transmitted to the controller. For this purpose, each MAV broadcasts its global odometry over the wireless network, and a ground station relays the drop box position. We throttle the messages to 5 Hz to reduce network traffic. Essentially, this is the only inter-robot communication on the network.

Figure 4 shows an example scenario where two MAVs attempt to pick up the same object. The MAV further away from the desired object perceives the other vehicle as an obstacle by receiving its odometry over wireless. Even though the MAVs do not broadcast their intentions, the controller can maintain a minimum distance between them.

5 Challenge 1: Landing on a Moving Platform

Challenge 1 requires an MAV to land on a moving platform. The landing platform, also referred to in the following as the “target”, moves inside the arena, on an eight-shaped path with a constant velocity, and is characterized by a specific mark composed of a square containing a circle and a cross (Figure 6a). Challenge 1 can be decomposed into seven major tasks (Figure 5): (i) MAV pose estimation (“State estimator” block), (ii) MAV control (“NMPC” block), (iii) landing platform detection (“Detector” block), (iv) landing platform

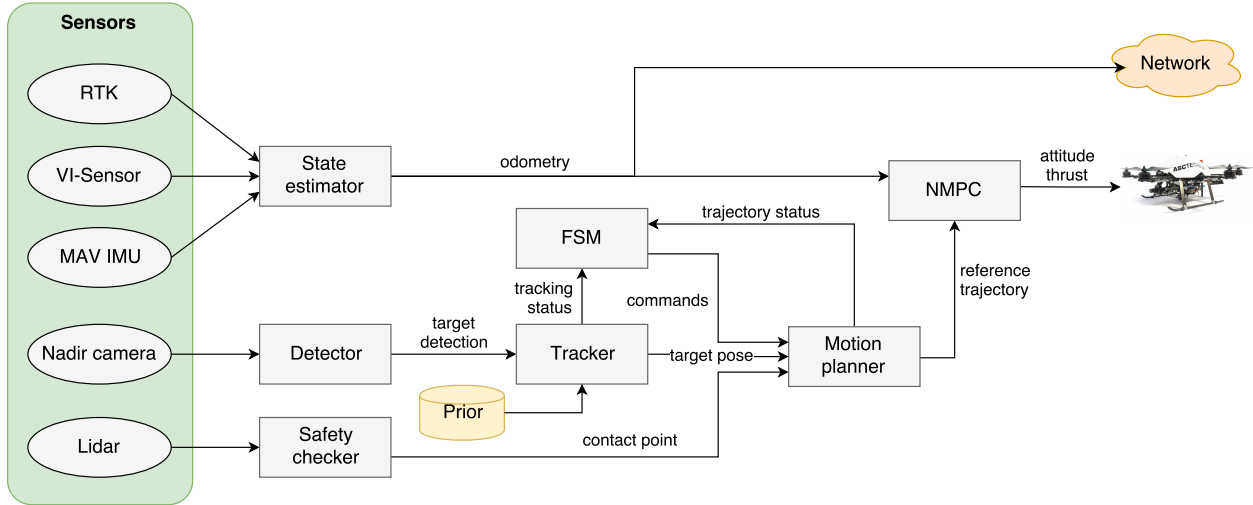


Figure 5: Block overview of the modules used for Challenge 1.

tracking (“Tracker” block), (v) planning the landing maneuver from the current MAV position to above the landing platform (“Motion planner” block), (vi) executing a final safety check before switching off the propellers (“Safety checker” block), and (vii) steering and coordinating all the previous modules together with a Finite State Machine (FSM) (“FSM” block).

As (i) and (ii) were addressed in Sections 3 and 4, only the remaining tasks are discussed in the following.

Based on the challenge specifications, we made the following assumptions to design our approach:

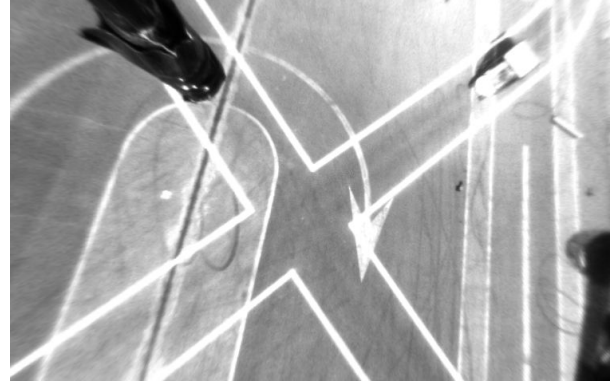
- the center of the arena and the path are known/measurable using RTK GPS,
- the landing platform is horizontal and of known size,
- the height of landing platform is known up to a few centimeters,
- the width of platform markings is known, and
- the approximate speed of the platform is known.

5.1 Platform Detection

A Point Grey Chameleon USB 2.0 camera (752 px × 480 px) with a fisheye lens in nadir configuration is used for platform detection. Given the known scale, height above ground, and planar orientation of the platform and its markings, the complete relative pose of the platform is obtainable using monocular vision. As the landing platform moves at up to 15 km/h, a high frame rate for the detector is desirable. Even at 30 fps and 100% recall and visibility, the platform moves up to 14 cm between detections. As a result, our design uses two independent detectors, invariant to scale, rotation, and perspective, in parallel. A quadrilateral detector identifies the outline of the platform when far away or when its markings are barely visible (as in Figure 6b), and a cross detector relies on the center markings of the platform and performs well in close-range situations (as in Figure 6a).



(a) During final approach: Good visibility of center markings and partial visibility of outline markings.



(b) During wait/search from high altitude: No center markings visible, outline barely visible.

Figure 6: Rectified images from the nadir fish-eye camera for detection. Due to lighting, distance and view point variations, the appearance of the landing target may change significantly.

5.1.1 Quadrilateral Detector

As the lens distortion is rectified, a square landing platform appears as a general quadrilateral under perspective projection. Thus, a simple quadrilateral detector can pinpoint the landing platform regardless of scale, rotation, and perspective. We used the quadrilateral detector proposed in the AprilTags algorithm (Olson, 2011). It combines line segments that end respectively start sufficiently close to each other into sequences of four. To increase throughput of the quadrilateral detector, the line segment detection of the official AprilTags implementation is replaced by the EDLines algorithm as proposed in (Akinlar and Topal, 2011). As the quadrilateral detector does not consider any of the markings within the platform, it is very robust against difficult lighting conditions while detecting many false positives. Thus, a rigorous outlier rejection procedure is necessary.

5.1.2 Cross Detector

Because the quadrilateral detector is designed for far-range detection and does not work in partial visibility conditions, we use a secondary detector, called the “cross detector”, based on platform markings for close-range maneuvers. The cross detector uses the same detected line segments as output by the EDLines algorithm (Akinlar and Topal, 2011) as input and processes them as follows:

1. For each found line segment the corresponding line in polar form is calculated (Figure 7).
2. Clusters of two line segments whose corresponding lines are sufficiently similar (difference of angle and offset small) are formed and their averaged corresponding line is stored.
3. Clusters whose averaged corresponding lines are sufficiently parallel are combined, thus yielding a set of four line segments in a 2 by 2 parallel configuration (Figure 7).
4. The orientation of line segments is determined such that two line segments on the same corresponding line point toward each other, thus allowing the center point of the detected structure to be computed.

Similarly to the quadrilateral detector, the cross detector is scale- and rotation-invariant. However, the assumption of parallel corresponding lines (Step 3) does not hold under perspective transformation. This

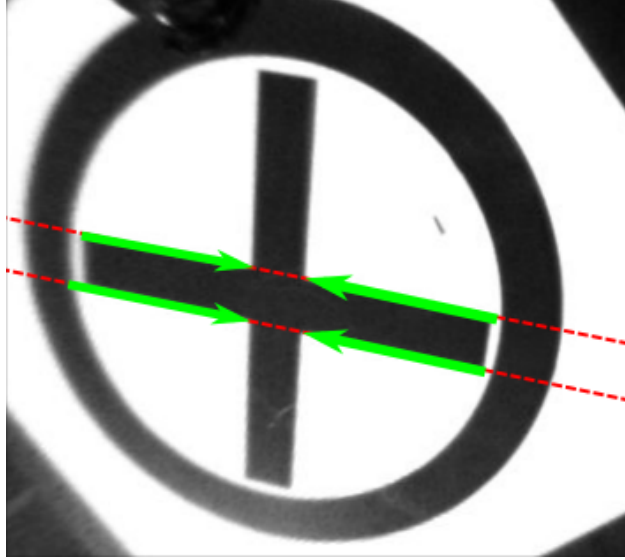


Figure 7: Selected line segments (green) and their corresponding line (red). The orientation of the line segments is indicated by the arrow tip.

can be compensated by using larger angular tolerances when comparing lines and by the fact that the detector is only needed during close-range maneuvers when the camera is relatively close to above the center of the platform.

The cross detector also generates many false-positive matches which are removed in the subsequent outlier rejection stage.

5.1.3 Outlier Rejection

As the landing platform is assumed to be aligned horizontally at a known height, the observed 2D image coordinates can be directly reprojected into 3D arena coordinates.

For the output of the quadrilateral detector, this enables calculating true relative scale and ratio based on the four corner points. Similarly, the observed line width of the cross detector output can be determined. Together with the prior probability of observing a platform in a certain location, these metrics are used to compute an overall probability of having observed the true landing platform. Figure 8 visualizes a simple example for both detectors.

For each detected quadrilateral, we compute the ratio of sides r and relative scale s from the diagonal lengths, d_1 and d_2 :

$$r = \begin{cases} \frac{d_1}{d_2} & , \text{ if } d_2 > 0, \\ 0 & , \text{ if } d_2 = 0, \end{cases} \quad c = \frac{d_1 + d_2}{2 \cdot t}, \quad (2)$$

where $t = 1.5 \cdot \sqrt{2}$ m is the nominal size of the platform diagonal according to the challenge specifications.

Based on ratio r , scale c , and the calculated x - y -position, the ratio likelihood $l_r = L(r, 1, \sigma_r)$, scale likelihood $l_c = L(c, 1, \sigma_c)$, and position likelihood $l_p = \mathcal{L}_{\text{track}}(x, y)$ are calculated. $L(y, \mu, \sigma)$ is modeled as a Gaussian likelihood function and $\mathcal{L}_{\text{track}}$ is a joint likelihood of the along and across-track distribution according to the path described in Section 5.2.1. Note that $\mathcal{L}_{\text{track}}$ is a numerical approximation, as the two distributions are not truly independent. Depending on the tuning of σ , the response of the likelihood functions is sharp, and thus a threshold on the approximate joint likelihood $l_r \cdot l_c \cdot l_p$ yields good outlier rejection.

Similarly, the measurements c_1, c_2, w_1, w_2 are taken for each response of the cross detector and their individual likelihood is calculated and combined with a position likelihood according to the path. Here the nominal value (mean) for w_1 and w_2 is 15 cm, and for c_1 and c_2 it is $15 \cdot \sqrt{2}$ cm.

Adjusting the covariances of each likelihood function and the threshold value allows the outlier rejection procedure to be configured to a desired degree of strictness.

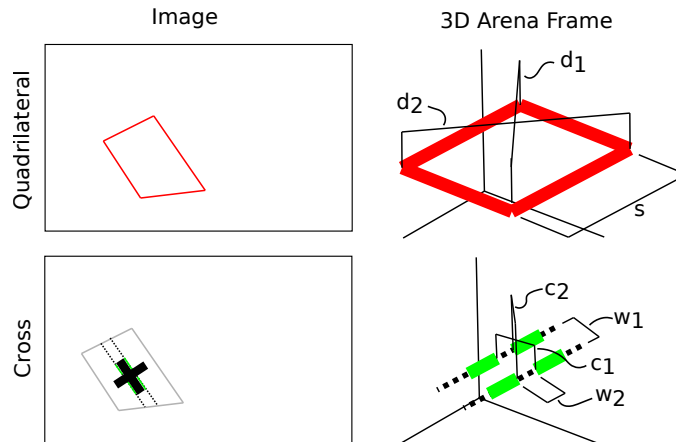


Figure 8: Calculation of measurements for outlier rejection: Image coordinates are reprojected onto a fixed z plane (given by the z -height of the platform) and several simple measurements ($d_1, d_2, s, c_1, c_2, w_1, w_2$) are taken. Note that reprojections of false positive detections that are not planar in the real world lead to highly distorted objects on the fixed z plane, and thus have no chance passing the outlier rejection.

5.2 Platform Tracking

After successfully detecting the landing target, the MAV executes fast and high-tilt maneuvers in order to accelerate towards it. During these maneuvers, the target may exit the Field of View (FoV) for a significant period of time, leading to sparse detections. Thus, a key requirement for the tracker is an ability to deduce the most probable target location in the absence of measurements.

In order to maximize the use of a priori knowledge about the possible target on-track location, along-track movement and speed, we chose a non-linear particle filter. Particle filters are widely used in robotics for non-linear filtering (Thrun, 2002), as they approximate the true a posteriori of an arbitrary complex process and measurement model. In contrast to the different variants of Kalman filters, particle filters can track multiple hypotheses (multi-modal distributions) and are not constrained by linearization or Gaussianity assumptions, but generally are more costly to compute.

The following elements are needed for our particle filter implementation based on the Bayesian Filtering Library (Gadeyne, 2001):

- *State.* Each particle represents a sampling (x_A, y_A, θ) , where x_A and y_A are the location in arena frame A , and θ corresponds to the movement direction along the track. As z_A and velocity are assumed, they are not part of the state space.
- *A priori distribution.* See Section 5.2.2.
- *Process model.* Generic model (Section 5.2.3) of a vehicle moving with steering and velocity noise along a path (Section 5.2.1).
- *Measurement model.* A simple Gaussian measurement model is employed. Particles are weighted according a 2D Gaussian distribution centered about a detected location.

- *Re-sampling*. Sampling importance re-sampling is used.

The resulting a posteriori distribution can then be used for planning and decision-making, e.g., aborting a landing approach if the distribution has not converged sufficiently or tracking multiple hypotheses until the next detection.

5.2.1 Platform Path Formulation

Piecewise cubic splines are used to define the path along which the target is allowed to move within certain bounds. The main advantage of such a formulation is its generality, as any path can be accurately and precisely parametrized. However, measuring distance along or finding a closest point on a cubic spline is non-trivial, given that no closed form solution exists. Instead, we leverage iterative algorithms to pre-compute and cache the resulting data, thus providing predictable runtime and high refresh rates. Since our on-board computer is equipped with sufficient memory, we store the following mappings:

- Parametric form with range $[0, \dots, 1]$ to length along the track, and vice-versa.
- x_A, y_A position in the *arena* frame A to parametric form using nearest neighbor search.

A major benefit of this approach is that it requires no iterative algorithms during runtime. Figure 9 provides an illustrative example. To facilitate efficient lookup and nearest-neighbor search, the samplings are stored in indexed KD-trees (Blanco, 2014).

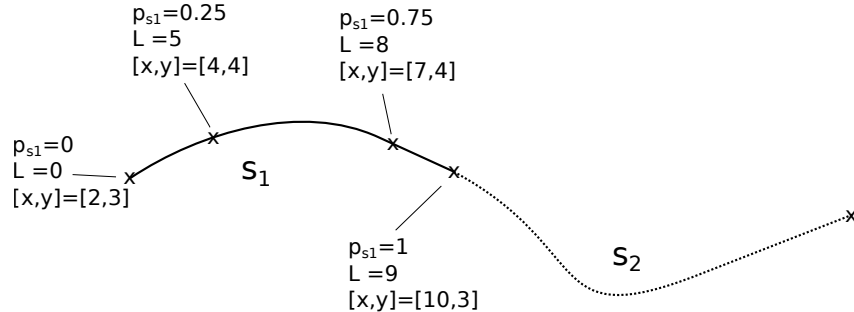


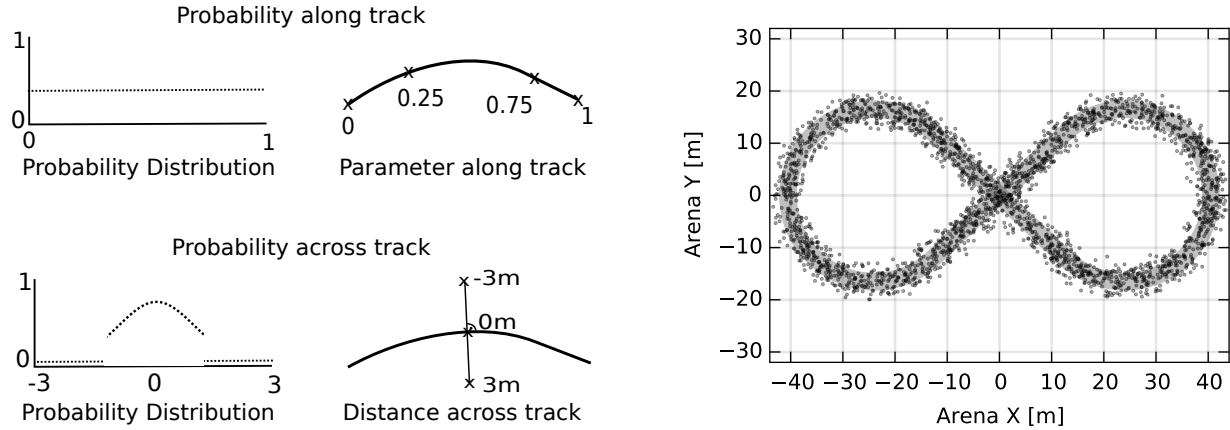
Figure 9: Simplified view of piece-wise cubic spline formulation. p_{s1} is the parameter along the first cubic spline piece, l the length along the track, and xy indicates the coordinates in *arena* frame. The l and xy values are cached in KD-trees for efficient lookup of the corresponding parameter p .

5.2.2 Prior Distribution

The location of each particle is initially sampled from two independent distributions, where (i) the location along the track is modeled as a uniform distribution along the full length of the track, i.e., the target can be anywhere, and (ii) the location across the track is modeled as a truncated Gaussian distribution, i.e, we believe that the platform tends to be more towards the center of the path, and cannot be outside the road limits. Figure 10 shows a schematic and sampled image of the prior. Note that this is a simplification, as the two distributions are not truly independent, e.g., they overlap along curves.

5.2.3 Process Model

We chose a generic process model that is independent of the vehicle type, e.g., Ackermann, unicycle or differential drive. Another motivation behind selecting this model was its execution time, as the following



(a) Independent prior distributions: Uniform along the track (along parameter p of the splines), truncated normal distribution across the track (perpendicular distance in meters).

(b) Sampled prior distribution: Each dot corresponds to one of 2500 samples on the 3 m wide platform path (gray). Note that the cut-off of the truncated Gaussian distribution is chosen to leave some slack in case of slight positioning errors of the track.

Figure 10: Schematic and sampled prior distributions.

process is applied to ~ 2500 individual particles at ~ 50 Hz. It is based on the assumption that there is an ideal displacement vector of movement along the track between two sufficiently small timesteps, as exemplified by the vector between C_k and C_{k+1} in Figure 11. In order to obtain this vector, the current position of a particle P_k is mapped onto the closest location on-track C_k and then displaced along the track according to ideal speed and timestep size, thus obtaining C_{k+1} . Note that these operations are very efficient based on the cached samples discussed in subsection 5.2.1.

The ideal displacement vector is then added to the current true position P_k and disturbed by steering noise φ_{noise} and speed noise v_{noise} , both of which are sampled from zero-mean Gaussian distributions.

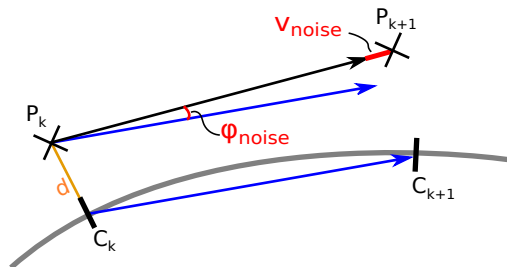
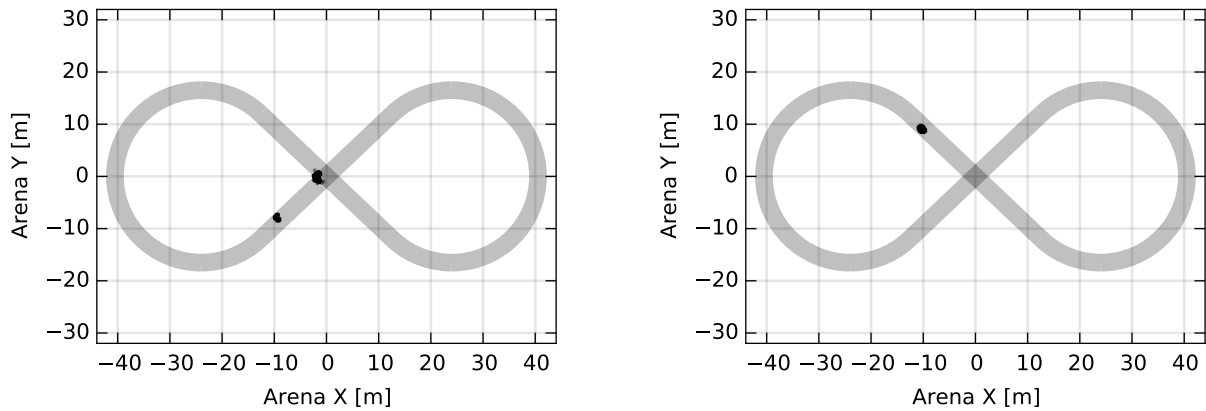


Figure 11: Illustration of the process model. P_k and P_{k+1} are the position of an individual particle at time k or $k + 1$ respectively. C_k is the closest point to P_k on the cubic spline path center. C_{k+1} corresponds to the position after moving along the path at speed s for Δt seconds.

5.2.4 Convergence Criteria

The target location distribution calculated by the particle filter is used for autonomous decision-making, such as starting or aborting a landing approach. Each timestep, the tracker determines whether the state has converged by simply checking if more than a certain threshold of the probability mass lies inside a circle with a given radius (subsequently called “convergence radius”), centered on the current weighted mean (Figure 13). This criterion can be determined in linear time and has proven to be sufficiently precise for our purposes with a probability mass threshold of 0.75 with a convergence radius of 1 m. Figure 12 shows

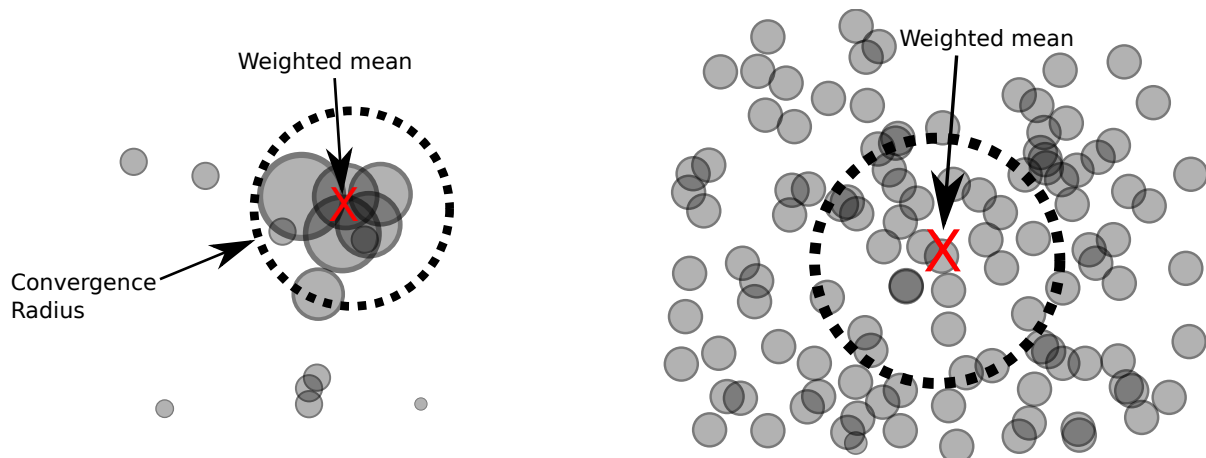


(a) State after exactly one detection and a few cycles without detection: Direction not determined yet, yielding two separate possible hypotheses that move in opposite directions.

(b) Converged tracker: Single, consistent hypothesis.

Figure 12: Visualization of particle filter states.

a typical split-state of the particle filter as it occurs after a single detection. The two particle groups are propagated independently along the track, and converge on one hypothesis as soon as a second measurement is available.



(a) Example of a converged state. A large majority of probability mass (summed weight) lies inside the convergence radius.

(b) Example of a diverged state.

Figure 13: Calculation of the convergence criteria. Each particle is visualized using a gray circle whose radius represents the current weight of the particle.

As the process model increases the particle scattering if no measurements are present, the filter automatically detects divergence after a period with no valid detections.

To calculate the future position of the target, a subsampling of the current particles is propagated forward in discrete timesteps using the process model only. This gives a spatio-temporal 4D trajectory of the predicted future state of the target over a given time horizon, which then can be used for planning.

5.3 Motion Planning

The motion planning task can be divided based on 3 different modes of operation:

1. *Searching.* The current state of the tracker is ignored and a fixed location or path that maximizes chances of observing the target is followed. Here, the MAV hovers 10m above the center of the eight-shaped path.
2. *Following.* The MAV should follow the target closely, e.g., 2m above its center, with the same velocity.
3. *Final approach.* The MAV should approach the target so that a landing is possible. The constraints are as follows: relative position and velocity in x and y directions is zero, relative velocity in z direction is chosen so that the airframe can absorb the impact shock (here: < 0.75 m/s). The first order derivatives of relative x and y speeds are also fixed to zero.

As mentioned, the tracker outputs a sampled spatio-temporal 4D trajectory of the predicted target position. Using this data, the MAV trajectory is calculated such that its position for each sample of the predicted target position and time coincides with the constraints as quickly as possible while satisfying flight envelope restrictions.

We optimized the pipeline such that replanning at a frequency of 50 Hz is possible, as the trajectory is updated after each step of the particle filter. This ensures that trajectory planning is always based on the most recent available information. The generated trajectory consists of smoothly joined motion primitives up to acceleration, as proposed by (Mueller et al., 2015). In order to select a valid set of primitives, a graph is generated and the cheapest path is selected using Dijkstra’s algorithm. The set of graph vertices consists of the current MAV position and time (marked as start edge), possible interception points, and multiple possible end positions according to the predicted target trajectory and the constraints (Figure 14). Each graph edge represents a possible motion primitive between two vertices. Vertices can have defined position, time and/or velocity. The start state is fully defined as all properties are known, whereas intermediate points might only have their position specified, and possible end-points have fixed time, position, and velocity according to the prediction.

Dijkstra’s algorithm is then used to find the shortest path between the start and possible end-states. Intermediate states are updated with a full state (position, velocity and time) as calculated by the motion primitive whenever their Dijkstra distance is updated. Motion primitives between adjacent vertices are generated as follows:

- If the start and end times of two vertices are set, then the trajectory is chosen such that it minimizes jerk.
- If the end time is not set, then the path is chosen such that it selects the fastest trajectory within the flight envelope.

The cost function used to weigh edges is simply the duration of the motion primitive representing the edge times a multiplier. The multiplier is 1 for all edges that are between predicted platform locations, and l^3 for all edges connecting the start state with possible intersection points, where l is the distance between points. Effectively, this results in an automatic selection of the intersection point, with a heavy bias to intersect as soon as possible.

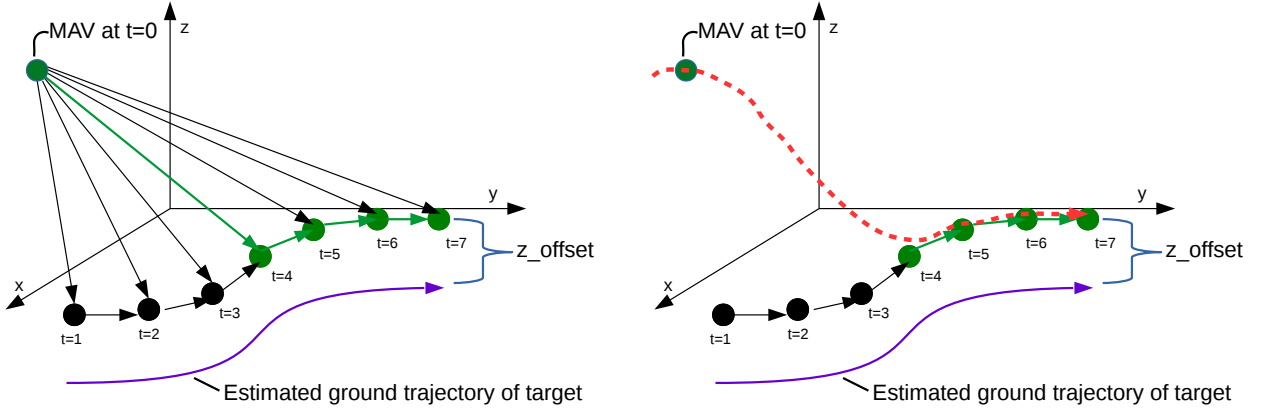


Figure 14: Directed graph based on current location and trajectory estimation of the target, with indicated shortest (cost-wise) path (green edges/vertices). Each edge represents a possible motion primitive. The red dashed line indicates the resulting smooth trajectory through the selected vertices.

5.4 Lidar Landing Safety Check

A successful Challenge 1 trial requires the MAV to come to rest in the landing zone with the platform intact and its propellers producing no thrust. Starting the final descent and then switching off the propellers on time was mission-critical for the overall result. These commands must be issued only when the MAV is considered to be above the landing platform, with a height small enough to allow its magnetic legs to attach to the metal part of the target. The range output of a Lidar sensor is employed to check the distance of the MAV from the landing platform. Even though the MAV already has all necessary information to land available at this stage, i.e., its global position and an estimate of the current platform location, an extra safety check is performed using the Lidar output: before triggering the final part of the landing procedure, it is necessary that the Lidar detects the real platform below the MAV.

This final safety check is executed using a stand-alone ROS node, which receives the MAV odometry (from the state estimation module), the estimated position of the landing platform (from the tracking module) and raw distance measurements from the Lidar. Raw measurements are provided in the form of ${}^L\mathbf{d}_{L,C_p} = (0, 0, z)^\top$, where z is the distance measurement from the sensor lens to a contact point C_p of the laser beam. They are expressed in a *lidar* frame L , whose z axis points out from the sensor lens. Raw measurements are converted in *arena* frame A distance measurements ${}^A\mathbf{d}_{A,C_p}$, by using the MAV odometry ${}^A\mathbf{T}_{A,B}$ and the qualitative displacement from *lidar* frame L to *nav-imu* frame I ${}^B\mathbf{T}_{B,L}$:

$${}^A\mathbf{d}_{A,C_p} = {}^A\mathbf{T}_{A,B} {}^B\mathbf{T}_{B,L} {}^L\mathbf{d}_{L,C_p}, \quad (3)$$

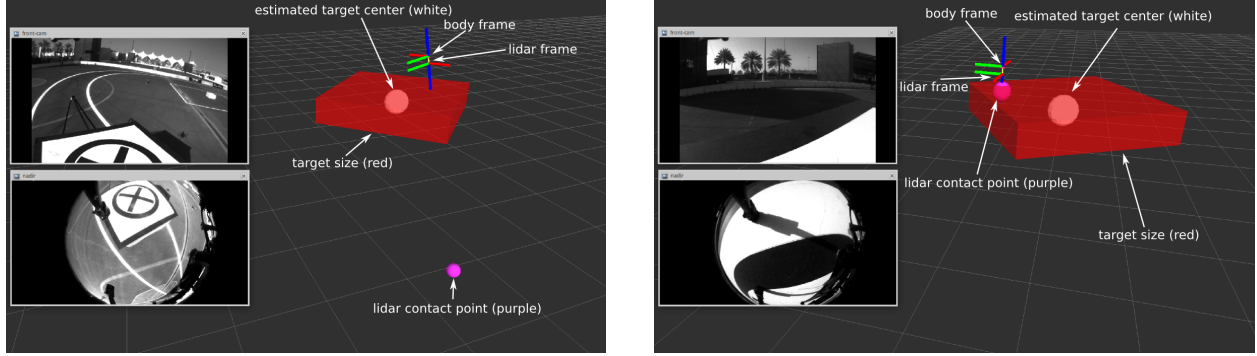
where ${}^L\mathbf{d}_{L,C_p}$ indicates the raw measurement in *lidar* frame L . Vector ${}^A\mathbf{d}_{A,C_p}$ can be seen as a “contact point” of the Lidar beam, expressed in global coordinates.

The aforementioned ROS node signals that the MAV is actually above the platform if two conditions are met: (i) the global MAV position is above the estimated global platform position, regardless of its altitude, and (ii) the Lidar contact point intercepts the estimated position of the platform. A Mahalanobis distance-based approach is employed to verify these conditions:

$$d_M(x) = \sqrt{(x - \mu)^\top \mathbf{S}^{-1} (x - \mu)}. \quad (4)$$

This is applied with different vector and matrix dimensions to check the conditions above:

- (i) $\mathbf{x} = (x_B, y_B)^\top$, $\boldsymbol{\mu} = (x_T, y_T)^\top$, and $\mathbf{S} = \text{diag}(\sigma_{x_T}^2, \sigma_{y_T}^2)$, where \mathbf{x} and $\boldsymbol{\mu}$ denote 2D global positions



(a) First landing attempt: MAV body frame was above the estimated location of the landing platform but, since the contact point of the Lidar beam was on the ground and not on the supposed platform position, the final maneuver was aborted.

(b) Third landing attempt: MAV body frame was above the estimated location of the landing platform and, since the contact point of the Lidar beam was on the supposed platform position, the final maneuver was considered completed and the motors were switched off.

Figure 15: Visualization of Lidar data recorded during the first Grand Challenge trial. Both images show the positions of the MAV (body frame) and of the Lidar, the images of the two cameras (front image from VI-Sensor and bottom image from mono camera), the estimated platform center provided by the tracker module (white ball), the available space to land if the target were on the estimated position (red parallelepiped), and the contact point of the Lidar beam.

of the MAV and the estimated platform center, respectively. The diagonal matrix \mathbf{S} contains the covariance of the estimated platform position, provided by the tracker module.

- (ii) $\mathbf{x} = {}_A\mathbf{d}_{A,C_p}$, $\boldsymbol{\mu} = (x_T, y_T, z_T)^\top$, and $\mathbf{S} = \text{diag}(\sigma_{x_T}^2, \sigma_{y_T}^2, \sigma_{z_T}^2)$, where \mathbf{x} and $\boldsymbol{\mu}$ denote the 3D Lidar beam contact point and the estimated platform center, respectively. The diagonal matrix \mathbf{S} contains the covariance of the estimated 3D platform position, which are provided by the tracker module.

The two conditions have different thresholds, $\delta_i = 1.0$, and $\delta_{ii} = 1.5$, that can be set and adjusted dynamically. Each condition is satisfied if its Mahalanobis distance is below the correspondence threshold. Even though these two conditions are tightly related, since ${}_A\mathbf{d}_{A,C_p}$ is computed using ${}_A\mathbf{T}_{A,B}$ in Equation 3, they are handled separately. This is because, by setting different thresholds, a higher importance is given to condition (ii) than to (i). This strategy allowed for a less error-prone final approach of the landing maneuver. To demonstrate this, Figure 15 shows two different landing attempts executed during the first Grand Challenge trial. In the first two attempts, the final maneuver was aborted before completion, since only the first condition was met. In the third attempt, the final approach was triggered as both conditions were met.

In Figure 15a, it can be seen in the two camera images that the landing target was actually ahead the MAV during the final approach. Even though the first condition was satisfied, such that the 2D position of the MAV was above the estimated platform position, the Lidar contact point was on the ground. This clearly indicates that the real platform was in a different location, and that probably the output of the tracker reached a weak convergence state. Relying only on the estimated target position would have led to an incorrect descent maneuver; likely resulting in flying against the UGV carrying the target. On the other hand, Figure 15b shows the system state when both conditions were met and the landing approach was considered complete.

5.5 Finite State Machine

A high-level FSM was implemented using the SMACH ROS package (Smach, 2017) to integrate the Challenge 1 submodules. The main states in the FSM are displayed in Figure 16 and briefly explained below.

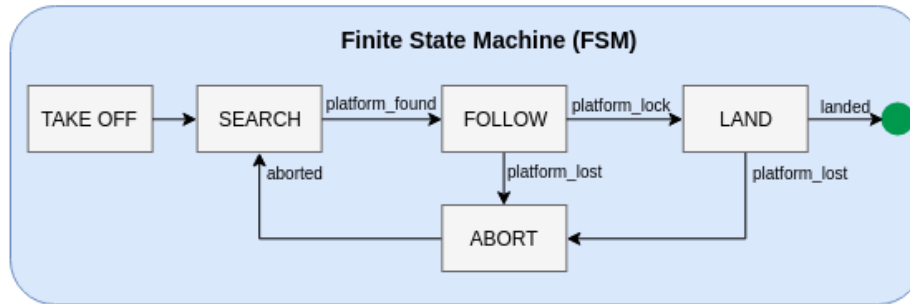


Figure 16: Task-level architecture for Challenge 1.

SEARCH: The MAV is commanded to fly above the center of the arena, and the tracker module begins searching for the target platform. When the motion direction of platform is successfully estimated and the target considered locked, the state switches to “FOLLOW”.

FOLLOW: The MAV is commanded to closely follow the target as long as either the platform is detected at a high rate, e.g., more than 7 Hz, or the tracker module decides the lock on the platform is lost. In the former case, the MAV is considered to be close enough to the platform, as the high detection rate is indicating, and the state is switched to “LAND”. In the latter case, the platform track is considered lost, so “ABORT” mode is triggered.

LAND: The MAV first continues following the platform while gradually decreasing its altitude. Once it is below a predefined decision height, the output of the Lidar is used to determine whether the estimated position of the landing platform overlaps with the actual target position. If that is the case, a fast descent is commanded and the motors are switched off as soon as a non-decreasing motion on the z -axis is detected, i.e., the MAV lands on the platform to conclude Challenge 1. If any of the previous safety checks fail, the state switches to “ABORT”.

ABORT: The MAV is first commanded to increase its altitude to a safety value in order to avoid any possibly dangerous situations. Then, the FSM transitions to “SEARCH” to begin seeking the platform again.

6 Challenge 3: Search, Pick Up, and Relocate Objects

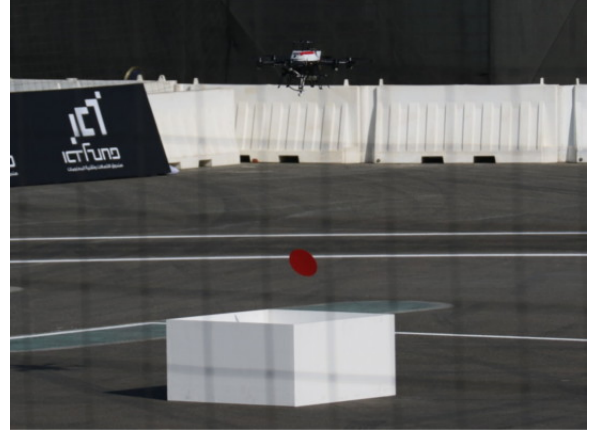
Challenge 3 requires a team of up to three MAVs equipped with grippers to search, find, pick, and relocate a set of static and moving objects in a $60\text{ m} \times 90\text{ m}$ planar arena (Figure 17). The arena has 6 moving and 10 stationary small objects as well as 3 stationary large objects. The moving objects move at random velocities under 5 km/h. Small objects have a cylindrical shape with a diameter of 200 mm and a maximum weight of 500 g. Large objects have a rectangular shape with dimensions $150\text{ mm} \times 2000\text{ mm}$, a maximum weight of 2 kg, and require collaborative transportation. Note that the presented system ignores large objects since our collaborative approach was not integrated at the time of the challenge (Tagliabue et al., 2017a; Tagliabue et al., 2017b).² The color of an object determines its type and score. Teams gain points for each successful delivery to the drop container or dropping zone.

The challenge evokes several integration and research questions. Firstly, our system must cater for the limited preparation and challenge execution time at the Abu Dhabi venue. Every team had two 20 min testing slots, two 20 min challenge slots, and two 20 min Grand Challenge slots, with flying otherwise prohibited. Furthermore, each slot only had ~ 30 min preparation time to setup the network, MAVs, FSM, and tune daylight-dependent detector parameters. These time constraints require a well-prepared system and a set of tools for deployment and debugging. Thus, a simple and clean *system architecture* is a key component.

²Also note that no team attempted to perform the collaborative transportation task during the challenge.



(a) An overview of the arena with 6 moving, 10 small static and 3 large static objects.



(b) An MAV dropping a red disc in the drop box container.

Figure 17: In Challenge 3 up to three MAVs need to collect as many objects as possible in the arena and deliver them to the drop box or dropping zone.

A second challenge is the *multi-agent system*. In order to deploy three MAVs simultaneously, we developed solutions for workspace allocation, exploration planning, and collision avoidance.

The third and most important task in this challenge is *aerial gripping*. Being able to reliably pick up 200 mm discs in windy outdoor environments is challenging for detection, tracking, visual servoing, and physical interaction. Given that only delivered objects yield points, a major motivation was to design a reliable aerial gripping system.

Driven by these challenges, we developed an autonomous multi-agent system for collecting moving and static small objects with unknown locations. In the following we describe our solutions to the three key problems above. The work behind this infrastructure is a combined effort of (Gawel et al., 2017; Bähnemann et al., 2017).

6.1 System Architecture

During development, it was recognized that a major difficulty for our system is the simultaneous deployment of more than one MAV. We also recognized network communication to be a known bottleneck in competitions and general multi-robot applications. With these considerations, we opted for a decentralized system in which each agent can fulfill its task independently while sharing minimal information.

Challenge 3 can be decomposed into four major tasks: (i) state estimation, (ii) control, (iii) waypoint navigation, and (iv) aerial gripping. A block diagram of this system for a single MAV is shown in Figure 18. The key components for autonomous flight consist of the *State Estimator* and *Reactive Position Control* presented in Section 3 and Section 4. The *Detector*, *Multitarget Tracker*, and *Servoing* blocks form the aerial gripping pipeline. The *Waypoint Navigator* handles simple navigation tasks, such as take off, exploration or object drop off. Each MAV communicates its odometry and controls a drop box semaphore over the *Network* and receives prior information about the task. The *Prior* data consist of offside and onside parameters. The object sizes, arena corners, and workspace allocation were taken from the challenge descriptions. The object color thresholding, camera white balancing, the drop box position, and the number of MAVs to engage was defined during the on-stage preparation time.

All modules are organized and scheduled in a high-level SMACH *FSM* (Smach, 2017). Its full decentralized workflow is depicted in Figure 19. After take off, each MAV alternates between exploring a predefined area

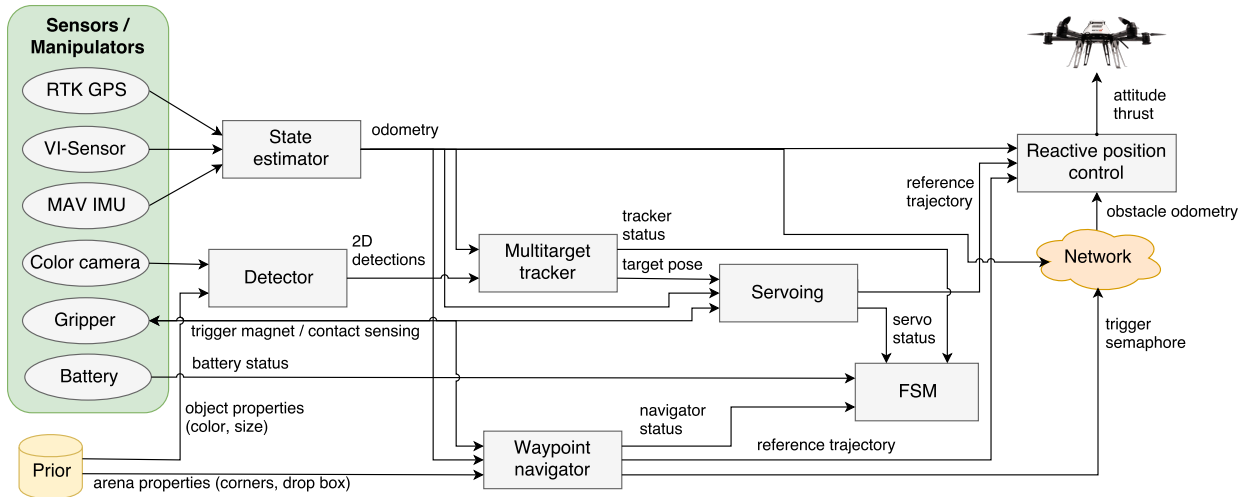


Figure 18: System diagram for Challenge 3. In addition to the mandatory state estimation and control modules, the system includes a waypoint navigator to explore the arena and fly to pre-programmed waypoints, and an elaborate object detection, tracking, and servoing pipeline.

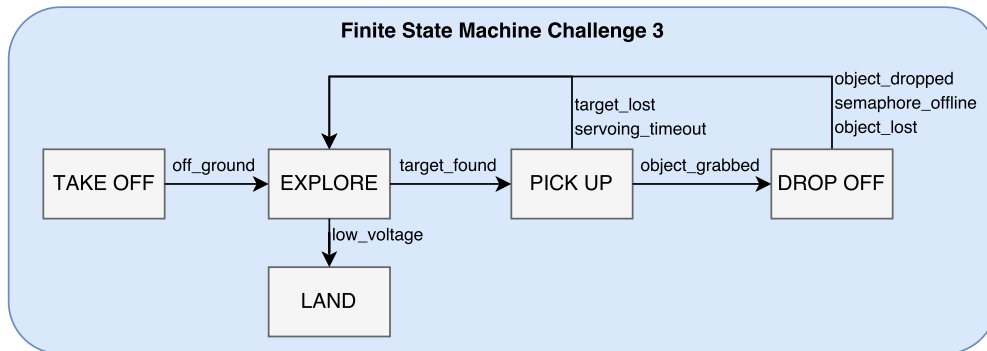


Figure 19: The FSM architecture for Challenge 3. Each MAV alternates between exploration and greedy small object pickup before landing with low battery voltage.

and greedily picking up and delivering the closest object.

TAKE OFF: a consistency checker verifies the state estimation concerning drift, RTK GPS fix, etc., and takes the MAV to a predefined exploration altitude. After taking off, the MAV state switches to “EXPLORE”.

EXPLORE: the MAV follows a predefined zig-zag exploration path at a constant altitude using the waypoint navigator. While searching, the object tracker uses the full resolution camera image to find targets from heights between 4m and 7m. If one or more valid targets are detected, the closest target is locked by the tracker and the MAV state switches to “PICK UP”. A valid target is one that is classified as small object, lies within the assigned arena bounds and outside the drop box, and has not had too many pickup attempts.

The mission is terminated if the battery has low voltage. The MAV state switches to “LAND”, which takes it to the starting position. As challenge resets with battery replacement were unlimited, this state was never entered during the challenge.

PICK UP: the object tracking and servoing algorithms run concurrently to pick up an object. The detection

processes the camera images at a quarter of their resolution to provide high rate feedback to the servoing algorithm. If the gripper’s Hall sensors do not report a successful pickup, the target was either lost by the tracker or the servoing timed out. The MAV state reverts to “EXPLORE”. Otherwise, the servoing was successful and the state switches to “DROP OFF”.

DROP OFF: the MAV uses the waypoint navigator to travel in a straight line at the exploration altitude to a waiting point predefined based on the hard-coded drop box position. Next, the drop container semaphore is queried until it can be locked. Once the drop box is available, the MAV navigates above the predefined drop box position, reduces its altitude to a dropping height, and releases the object. Then, it frees the drop box and semaphore and returns to “EXPLORE”. A drop off action can fail if the object is lost during transport or the semaphore server cannot be reached. It is worth mentioning that, in our preliminary FSM design, the MAV used a drop box detection algorithm based on the nadir camera. However, hard coding the drop box position proved to be sufficient for minimal computational load.

6.2 Multi-Agent Coverage Planning and Waypoint Navigation

The main requirements for multi-agent allocation in this system are, in order of decreasing priority: collision avoidance, robustness to agent failure, robustness to network errors, full coverage, simplicity, and time optimality.

Besides using reactive collision avoidance (Section 4), we separate the arena into one to three regions depending on the number of MAVs as illustrated in Figure 20. Each MAV explores and picks up objects only within its assigned region. Additionally, we predefine a different constant navigation altitude for each MAV to avoid interference during start and landing. While this setup renders the trajectories collision free by construction, corner cases, such as a moving object transitioning between regions or narrow exploration paths, can be addressed by the reactive control scheme.

Our system architecture was tailored for robustness to agent failures. The proposed decentralized system and arena splitting allows deployment of a different number of MAVs in each challenge trial and even between runs. Automated scripts divide the arena and calculate exploration patterns.

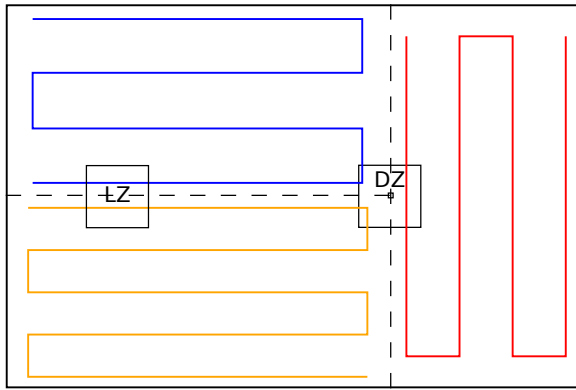
Our design minimizes the communication requirements to achieve robustness to network failures. Inter-MAV communication, i.e., odometry and drop box semaphore, is not mandatory. This way, even in failure cases, an MAV can continue operating its state machine within its arena region. All algorithms run on-board the MAVs. Only the drop box semaphore may not be locked anymore, but objects will still be dropped within the dropping zone. Furthermore, our system does not require human supervision except for initialization and safety piloting, which does not rely on wireless connectivity.

For Coverage Path Planning (CPP) we implement a geometric sweep planning algorithm to ensure object detections. The algorithm automatically calculates a lawnmower path for a given convex region as shown in Figure 20d. The maximum distance d_{\max} between two line sweeps is calculated based on the camera’s lateral FoV α , the MAV altitude h , and a user defined view overlap $\delta \in [0 \dots 1]$:

$$d_{\max} = (1 - \delta) \cdot 2 \cdot h \cdot \tan \frac{\alpha}{2}. \tag{5}$$

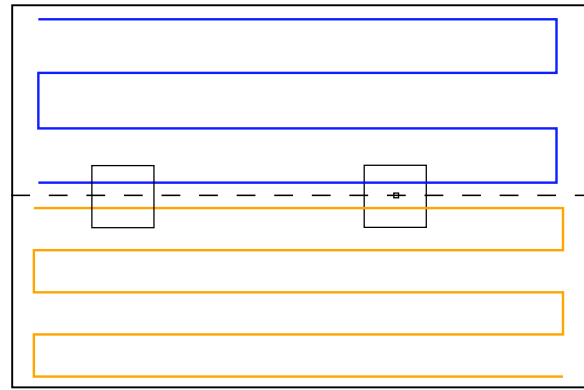
This distance is rounded down to create the minimum number of equally spaced sweeps covering the full polygon. During exploration, the waypoint navigator keeps track of the current waypoint in case exploration is continued after an interruption.

In order to process simple navigation tasks, such as take off, exploration, delivery or landing, we have a waypoint navigator module that commands the MAV to requested poses, provides feedback on arrival, and flies predefined maneuvers, such as dropping an object in the drop box. The navigator generates velocity

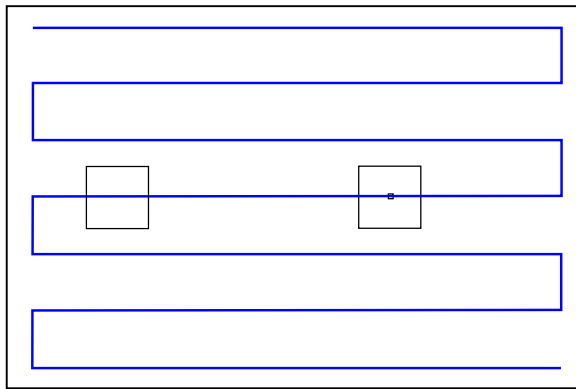


--- MAV 1 --- MAV 2 --- MAV 3

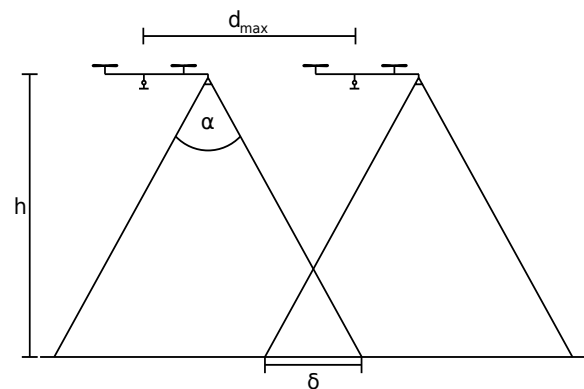
(a) Three MAVs.



(b) Two MAVs.



(c) One MAV.



(d) The sweep distance.

Figure 20: For object search, the arena is divided into convex regions based on the arena corners, dropbox in the Dropping Zone (DZ), Landing Zone (LZ) and number of MAVs. Each MAV explores its region with a zig-zag path. The maximum sweep distance d_{\max} is a function of altitude h , camera FoV α , and overlap δ .

ramp trajectories between current and goal poses. This motion primitive is a straight line connection and thus collision free by construction, as well as easy to interpret and tune.

The main design principles behind our navigation framework are simplicity, coverage completeness, and ease of restarts, which were preferable over time optimality provided by alternative methods, such as informative path planning, optimal decision-making, or shared workspaces. Even when exploring the full arena with one MAV at 5 m altitude with 2 m/s maximum velocity and 4 m/s² acceleration, the total coverage time is less than 5 min. Furthermore, our practical experience has shown that the main difficulty in this challenge is aerial gripping, rather than agent allocation. Our MAVs typically flew over all objects in the assigned arena, but struggled to detect or grip them accurately.

6.3 Object Detection, Tracking, and Servoing

Our aerial gripping pipeline is based on visual servoing, a well-established technique where information extracted from images is used to control robot motion. In general, visual servoing is independent of the underlying state estimation, allowing us to correctly position an MAV relative to a target object without external position information. In particular, we use a pose-based visual servoing algorithm. In this approach, the object pose is first estimated from the image stream. Then, the MAV is commanded to move towards the object to perform grasping.

The challenge rules specify the colors and shapes of the object to be found. However, the floor material and exact color code/gloss type of the objects were unknown until the challenge start. Thus, we decided to develop a blob-detector with hand-crafted shape classification that uses the known object specifications and can be tuned with human interpretation. Figure 21 depicts an example of our image processing methods.

In order to detect colored objects, time-stamped RGB images are fed into our object detector. The detector undistorts the images and converts its pixel values from RGB to CIE L*a*b* color space. For each specified object color, we apply thresholds on all three image channels to get the single channel binary image. After smoothing out the binary images using morphological operators, the detections are returned as the thresholded image regions.

For each detection, we compute geometrical shape features such as length and width in pixel values, convexity, solidity, ellipse variance, and eccentricity from the contour points. We use these features to classify the shapes into small circular objects, large objects, and outliers. A small set of intuitively tunable features and manually set thresholds serve to perform the classification.

On the first day of preparation, we set the shape parameters and tuned the coarse color thresholding parameters. Before each trial, we adjusted the camera white balance and refined the color thresholds.

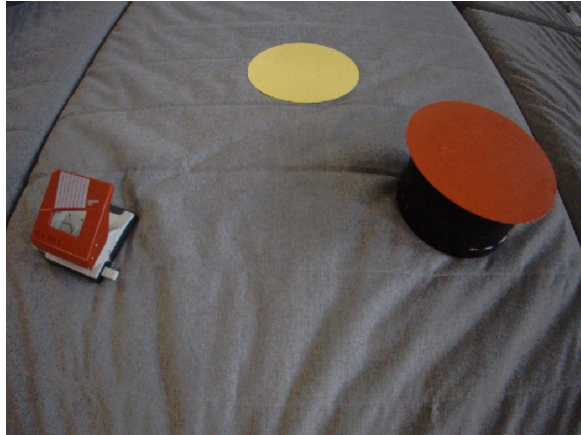
Given the 2D object detections, our aim is to find the 3D object pose for tracking. Figure 22 displays the problem of calculating the position of two points ${}^C\mathbf{p}_1$ and ${}^C\mathbf{p}_2$ on an object with respect to the *camera* coordinate frame C . Assuming that the objects lie on a plane perpendicular to the gravity aligned odometry frame z -axis ${}^O\mathbf{z}$, and that the metric distance m between the two points is known, we formulate the constraints:

$$\|{}^C\mathbf{p}_1 - {}^C\mathbf{p}_2\| = m, \quad (6)$$

$${}^C\mathbf{n}_T^\top ({}^C\mathbf{p}_1 - {}^C\mathbf{p}_2) = 0, \quad (7)$$

where ${}^C\mathbf{n}_T = \mathbf{R}_{CO}{}^O\mathbf{z}$ is the object normal expressed in the *camera* coordinate frame and \mathbf{R}_{CO} is the rotation matrix from the odometry coordinate frame O to the *camera* coordinate frame C .

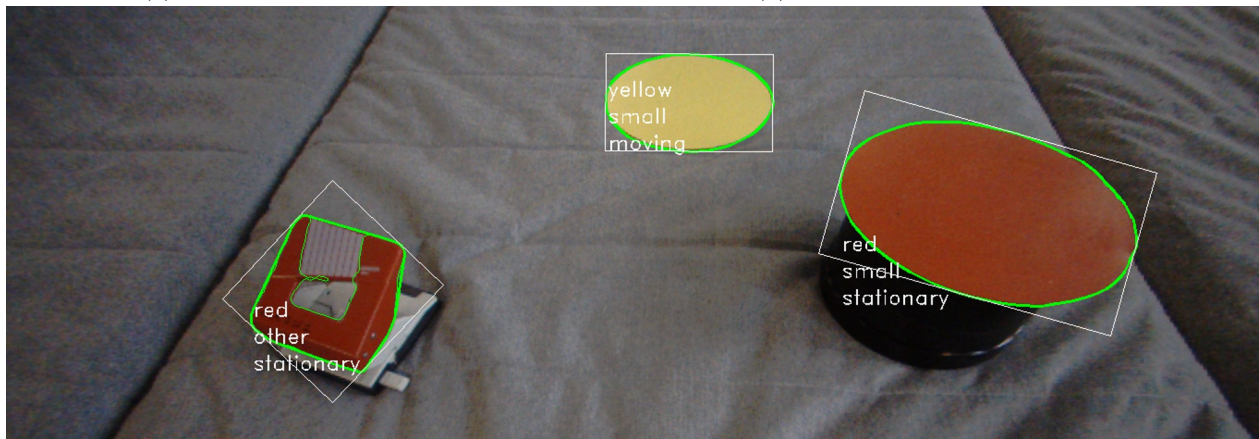
The relation between the mapped points \mathbf{u}_1 and \mathbf{u}_2 in the image and the corresponding points ${}^C\mathbf{p}_1$ and ${}^C\mathbf{p}_2$



(a) A distorted input image.

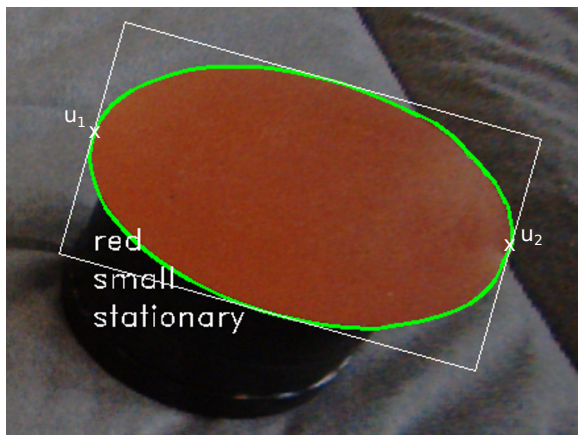


(b) The binary image for the red color.

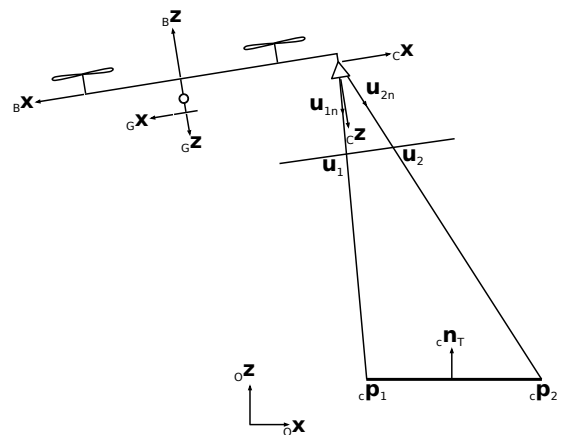


(c) Detections with classifications.

Figure 21: The object detection pipeline. We use blob-detection and hand-crafted shape classification.



(a) An object detection with image points.



(b) The camera/object configuration.

Figure 22: The problem of calculating the 3D positions $c\mathbf{p}_1$ and $c\mathbf{p}_2$ from a single monocular object detection \mathbf{u}_1 and \mathbf{u}_2 . Assuming that the object lies on flat ground and that its physical size is known, the inverse projection problem can be solved.

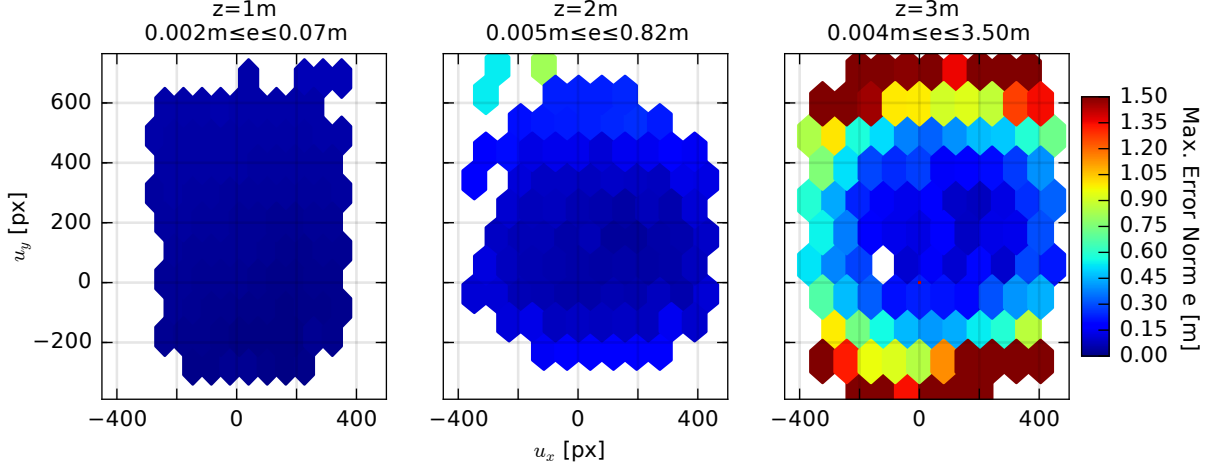


Figure 23: A Vicon motion capture validation of the detection error on the image plane. Best accuracy is obtained for centrally positioned, close objects, while detections at the image boundary and high altitudes degrade.

is expressed with two scaling factors λ_1 and λ_2 as:

$$C\mathbf{p}_1 = \lambda_1 \mathbf{u}_{1n}, \quad (8)$$

$$C\mathbf{p}_2 = \lambda_2 \mathbf{u}_{2n}, \quad (9)$$

where \mathbf{u}_{1n} and \mathbf{u}_{2n} lie on the normalized image plane pointing from the *focal point* to the points $C\mathbf{p}_1$ and $C\mathbf{p}_2$ computed through the perspective projection of the camera. For a pinhole model, the perspective projection from image coordinates $(u_x \ u_y)^\top$ to image vector $(u_{nx} \ u_{ny} \ 1)^\top$ is:

$$u_{nx} = \frac{1}{f_x}(u_x - p_x), \quad u_{ny} = \frac{1}{f_y}(u_y - p_y), \quad (10)$$

where p_x, p_y is the principal point and f_x, f_y is the focal length obtained from an intrinsic calibration procedure (Furgale et al., 2013; Kalibr Github, 2017).

Inserting Equation 8 and Equation 9 into Equation 6 and Equation 7 and solving for λ_1 and λ_2 yields the scaling factors which allow the inverse projection from 2D to 3D:

$$\lambda_1 = m \frac{|C\mathbf{n}_T^\top \mathbf{u}_{2n}|}{\kappa}, \quad \lambda_2 = m \frac{|C\mathbf{n}_T^\top \mathbf{u}_{1n}|}{\kappa}, \quad \kappa = \|(C\mathbf{n}_T^\top \mathbf{u}_{2n}) \mathbf{u}_{1n} - (C\mathbf{n}_T^\top \mathbf{u}_{1n}) \mathbf{u}_{2n}\|. \quad (11)$$

We consider the mean of $C\mathbf{p}_1$ and $C\mathbf{p}_2$ as the object center point in 3D. Figure 23 shows a ground truth validation of the detection error on the image plane for different camera poses. Due to image boundary effects, image smoothing, and resolution changes, the detections at the pose boundary tend to be less accurate than those near the center. We compensate for this in the tracking and servoing pipeline.

The object's 3D position (and velocity) is tracked in a multi-target hybrid Kalman Filter (KF). A hybrid KF was chosen for task since a constant sampling time for the arrival of the detections cannot be guaranteed. The tracker first removes all outlier detections based on their classification, shape-color inconsistency, and flight altitude-size inconsistency. It then computes the inverse projection of the detections from 2D to 3D in camera coordinate frame using Equation 11. Using the MAV's pose estimate and the extrinsic calibration of the camera to the MAV IMU (Furgale et al., 2013; Kalibr Github, 2017), it transforms the object position from the camera coordinate frame to the *odom* coordinate frame.

For each observed object, a KF is initialized to track its position and velocity. The assignment of detections to already initialized KFs is performed in an optimal way using the Hungarian algorithm (Kuhn, 1955) with

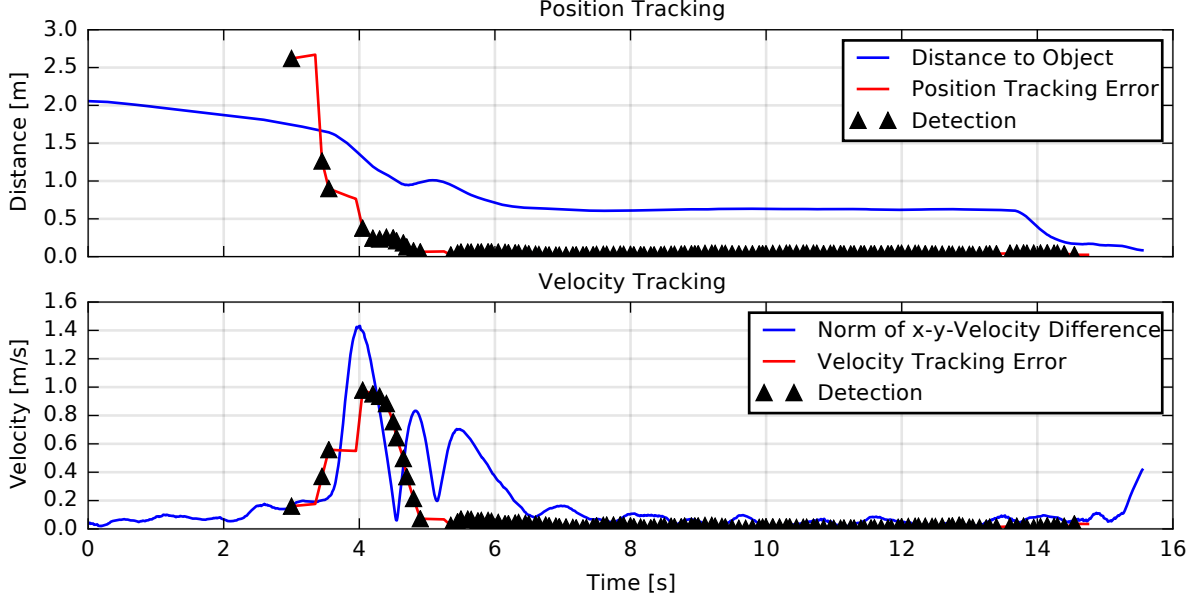


Figure 24: A Vicon motion capture validation of tracking and picking up a moving object at 1 km/h. Large initial tracking errors quickly converge when centering above the object.

Euclidean distance between the estimated and measured position as the cost. The filter uses a constant 2D velocity motion model for moving objects, and a constant position model for static ones. The differential equation governing the constant position model can be written as:

$$\dot{\mathbf{x}}_s(t) = \begin{pmatrix} \dot{p}_x(t) \\ \dot{p}_y(t) \\ \dot{p}_z(t) \end{pmatrix} = \mathbf{v}_s(t), \quad (12)$$

where $\mathbf{x}_s(t)$ is the position of the object with components $p_x(t)$, $p_y(t)$ and $p_z(t)$ and $\mathbf{v}_s(t)$ is a zero-mean Gaussian random vector with independent components. This model was chosen since it provides additional robustness to position estimation.

Considering the prior knowledge that the objects can only move on a plane parallel to the ground, we chose a 2D constant velocity model for moving objects where the estimated velocity is constrained to the xy -plane. The differential equation governing this motion can be written as:

$$\dot{\mathbf{x}}_m(t) = \begin{pmatrix} \dot{p}_x(t) \\ \dot{p}_y(t) \\ \dot{p}_z(t) \\ \dot{v}_x(t) \\ \dot{v}_y(t) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{x}_m(t) + \mathbf{v}_m(t), \quad (13)$$

where $\mathbf{x}_m(t)$ is the state vector with components $p_x(t)$, $p_y(t)$ and $p_z(t)$ as the position and v_x and v_y as the velocity. The vector $\mathbf{v}_m(t)$ was again a zero-mean Gaussian random vector with independent components.

Figure 24 exemplifies a ground truthed tracking evaluation. Since an object is usually first detected on the image boundary, it is strongly biased. We weight measurements strongly in the filter, such that the initial tracking error converges quickly once more accurate central detections occur.

As shown in Figure 23 and Figure 24, centering the target in the image reduces the tracking error and increases the probability of a successful pickup. Figure 25 summarizes our servoing approach. The algorithm

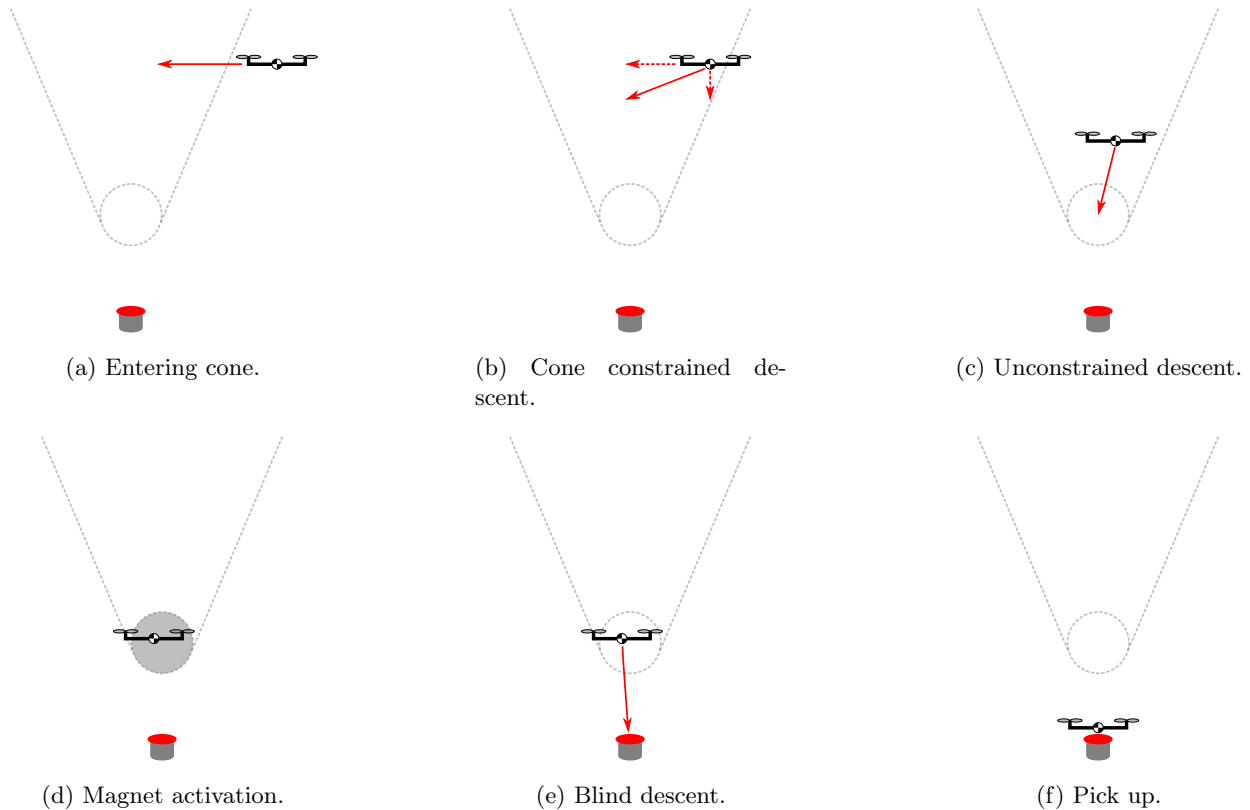
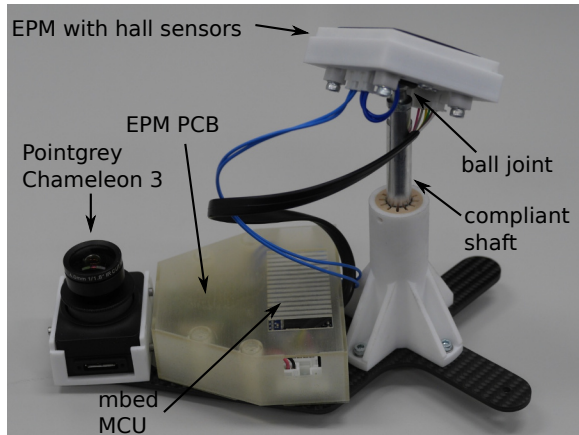


Figure 25: The six stages of visual servoing. The cone limited descent rate ensures continuous object detections.

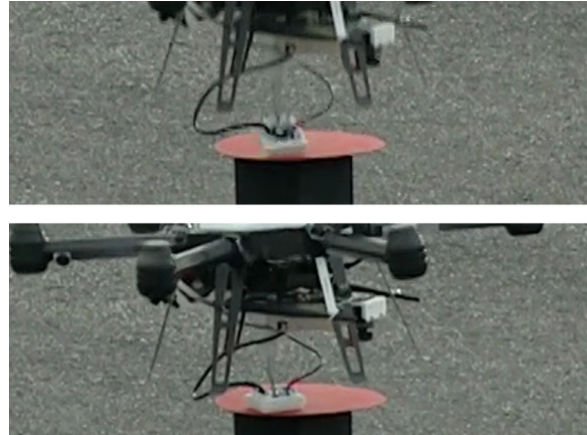
directly sends tracked object positions and velocities to the controller relative to the gripper frame G . Given its prediction horizon, the NMPC automatically plans a feasible trajectory towards the target. In order to maintain the object within FoV and to limit descending motions, z -position is constrained such that the MAV remains in a cone above the object. When the MAV is centered in a ball above the object, it activates the magnet and approaches the surface using the current track as the target position. The set point height is manually tuned such that the vehicle touches the target but does not crash into it. If the gripper does not sense target contact upon descent and the servoing times out, or the tracker loses the target during descent, the MAV reverts to exploration.

6.4 Gripper

The aerial gripping of the ferromagnetic objects is complemented by an energy-saving, compliant EPM gripper design. Unlike regular electromagnets, EPMS only draw electric current while transitioning between the states. The gripper-camera combination is depicted in Figure 26a. The gripper is designed to be lightweight, durable, simple, and energy efficient. Its core module is a NicaDrone EPM with a typical maximum holding force of 150 N on plain ferrous surfaces. The EPM is mounted compliantly on a ball joint on a passively retractable shaft (Figure 26b). The gripper has four Hall sensors placed around the magnet to indicate contact with ferrous objects. The change in magnetic flux density indicates contact with a ferrous object. The total weight of the setup is 250 g.



(a) The module.



(b) The gripper retraction on impact.

Figure 26: The gripper-camera combination used to detect and pick-up objects. The gripper is compliant and has Hall sensors to perceive contact.

7 Preparation and Development

Our preparation before the competition was an iterative process involving simulations, followed by indoor and outdoor field tests. New features were first extensively tested in simulation with different initial conditions. This step served to validate our methods in controlled, predictable conditions. Then, indoor tests were conducted in a small and controlled environment using the MAV platforms to establish physical interfaces. Finally, upon completing the previous two steps, one or more outdoor tests were executed. The data collected in these tests were then analyzed in post-processing to identify undesirable behaviors or bugs arising in real-world scenarios. This procedure was continuously repeated until the competition day.

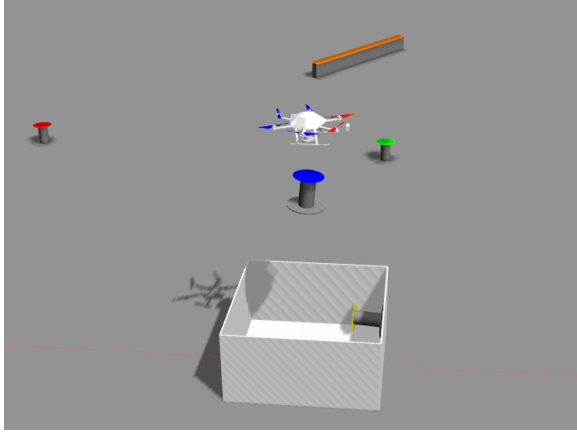
7.1 Gazebo

Gazebo was used as the simulation environment to test our algorithms before the flight tests (Gazebo, 2017). An example screenshot is shown in Figure 27a. We opted for this tool as it is already integrated within ROS, providing an easily accessible interface. Moreover, ASL previously developed a Gazebo-based simulator for MAVs, known as RotorS (Furrer et al., 2016). It provides multirotor MAV models, such as the AscTec Firefly, as well as mountable simulated sensors, such as generic odometry sensors and cameras. As the components were designed with the similar specifications as their physical counterparts, we were able to test each module of our pipeline in simulation before transferring features to the real platforms. Furthermore Software-in-the-Loop (SIL) simulations decouple software-related issues from problems arising in real-world experiments, thus allowing for controlled debugging during development.

7.2 Testing and Data Collection

In the months preceding the final competition, high priority was given to testing each newly developed feature outdoors. A specific, available location shown in Figure 27b served as our testing area. Through the precise accuracy of our RTK receiver, it was possible to define and mark every relevant point in the testing area to replicate the setup of the actual arena in Abu Dhabi. For instance, we drew the eight-shaped path and placed the take off and drop off zones.

To simplify flight trials and validate the functionalities of our MAVs in a physical environment, several extra tools were developed. Node manager (Node Manager, 2017) is a Graphical User Interface (GUI) used to



(a) Gazebo-based simulation.



(b) Field test setup.

Figure 27: Preparation before the competition. SIL simulations and field testing.

Trial	Challenge 1			Challenge 3					Points
	Landed	Intact	Time	Red	Green	Blue	Yellow	Orange	
ICT 1	No	No	-	0	1	0	1	0	6
ICT 2	No	No	-	1	1	0	0	0	3
GCT 1	Yes	Yes	124 s	2	2	0	0	0	6
GCT 2	Yes	Yes	54 s	0	1	1	1	0	10

Table 2: Summary of the results obtained in Challenge 1 and Challenge 3 in the Individual Challenge Trials (ICTs) during the first two days and the Grand Challenge Trials (GCTs) on the third day. The columns for Challenge 3 show the objects delivered into the bin and the resulting points.

manage the running and configured ROS nodes on different hosts, such as a ground control stations and the MAVs. Moreover, a Python GUI was implemented to provide visual feedback to an operator checking the MAV states. This interface displays relevant information, such as, the status of RTK fix, navigation waypoints or position and velocity in arena frame.

On-field data was collected using rosbags: a set of tools for recording from and playing back ROS topics. During the tests, raw sensor data, such as IMU messages, camera images, Lidar measurements, and RTK positions, as well as the estimated global state, were recorded on-board the MAVs. This information could then be analyzed and post-processed off-board after the field test.

8 Competition Results

We successfully deployed our MAV system in the MBZIRC 2017 in Abu Dhabi. Competing against 24 other teams for a total prize money of almost USD 2M, we placed second in the individual Challenge 3 and second in the Grand Challenge. Table 2 summarizes the results achieved during the competition. The following sections analyze each challenge separately and summarize the lessons learned from developing a competition-based system.³

³See also the video: <https://youtu.be/DXYFAkjHeho>.

8.1 Challenge 1

In Challenge 1 we gradually improved during the competition days to an autonomous landing time comparable to the top three teams of the individual challenge trials. We showed start-to-stop landing within 56 s and landing at 15 km/h within approximately 4 s after the first detection of the landing platform from a hovering state.

8.1.1 Individual Challenge Trials

During the two individual challenge trials, on the 16th and 17th March 2017, the MAV could not land on the target. This was due to different reasons. Before the first day, the on-board autopilot was replaced with an older, previously untested version due to hardware problems arising during the rehearsals before the first challenge run. This caused the MAV to flip during take off, since some controller parameters were not tuned properly. Calibration and tuning of the spare autopilot could not be done on-site, as test facilities were not available.

During the second day of Challenge 1, the MAV could take off and chase the platform but crashed during a high-velocity flight, as its thrust limits were not yet tuned correctly. This led to the loss of altitude during fast and steep turn maneuvers and subsequent ground contact. After analyzing this behavior over the first two days, the on-board parameters were properly configured and we were ready to compete in the two Grand Challenge trials.

8.1.2 Grand Challenge Trials

In both landing trials of the Grand Challenge, the MAV landed on the moving platform while it was driving at the maximum speed of 15 km/h (see Figure 1a). In the first run, the MAV performed three attempts before being able to land. The first two attempts were automatically aborted due to weak tracker convergence and the consequential noncompliance of the Lidar safety check (see Section 5.4).

Figure 28 shows the data collected on-board the MAV during the first trial of the Grand Challenge. Time starts when the starting signal was given by the judges. The first 60 s are cut since the platform was not in the MAV FoV. The attempts were executed in the following time intervals: [64.5, 69.1] s, [91.2, 96.5] s and [117.6, 125.0] s. Figure 28b shows that every time the tracker converged, the MAV could reduce the position error below 1 m in about 4 s. The first two attempts were aborted approximately at 52 s and at 79 s, due to weak convergence of the tracker module. The third attempt started at 117.6 s and after 4 s the MAV touched the landing platform, as can be seen in the spike of the velocity in Figure 28c, at 121 s. The motors were not immediately turned off, as explained in Section 5.4, for safety reasons. The “switch-off motors” command was triggered at 124 s, successfully concluding the Challenge 1 part of the Grand Challenge. During this first run, a total of 397 individual platform detections were performed, which are marked in Figure 28a. No outliers reached the tracker. The tracker and both detectors run at the same frequency as the camera output, ~ 50 Hz. When the platform was clearly visible, it could be detected in almost every frame, as evidenced in Figure 28a. Depending on flight direction, lighting and relative motion, the total number of detections per landing attempt varies significantly. Interestingly, the first landing attempt had a very high number of detections of both detectors, but was aborted due to a combination of a strict security criterion and a non-ideally tuned speed of the platform. The cross detector (black marks) could detect the platform after the abortion, thus resetting the filter because the MAV performed a fast maneuver to move back to the center of the field, effectively gaining height for a better camera viewpoint. It would have been possible to directly re-use these measurements to let the filter converge and start chasing the platform again before moving to a safe hovering state. However, we disabled immediate re-convergence to minimize flight conditions which might result in crash with the platform. The second and third landing attempts prove the capability of the tracker to accurately predict the platform position - even with sparse (second) and no detection (third) during the final approach, the MAV was able to chase respectively land safely on the platform. Figure 29

shows a 3D plot of the successful landing approach and its steep descent directly after tracker convergence.

In the second trial the MAV landed in the first attempt, in 56 s. Hence, we successfully demonstrated our landing approach and contributed to winning the silver medal in the Grand Challenge.

8.2 Challenge 3

We finished second in Challenge 3 where we engaged up to three MAVs simultaneously. Overall, we showed steady performance over all trials and were even able to surpass the individual challenge winning score during the our last Grand Challenge trial (see Figure 30a). As Table 2 shows, we collected both moving and static small objects and at least two and at most four objects in each trial. Our system had a visual servoing success rate of over 90%.

8.2.1 Individual Challenge Trials

Two days before the challenge, we had two rehearsal slots of 20 min each to adjust our system for the arena specifications. On the first day, we measured the arena boundaries, set up the WiFi, and collected coarse object color thresholds. On the second day, first flight and servoing tests were performed with a single MAV. During these early runs, we noticed that the MAV missed some object detections and that the magnet could not grip the heavier objects with paint layers thicker than our test objects at home. We partially improved the gripping by extending the EPM activation cycle. Unfortunately, we were not able to find the detection issues due to missing debug data during the MBZIRC itself.

In the first individual trial we then collected a static green object and a moving yellow object securing second place already. In the second trial we were able to increase the accumulated flight time of our MAVs but did not improve on the number and score of objects (see Figure 30b and Figure 30d).

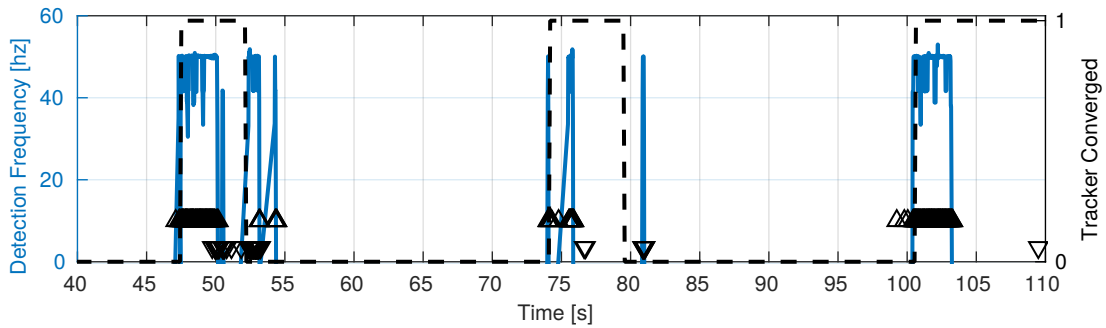
Despite achieving an accumulated flight time of less than 16%, our system still outperformed all but one team in the individual challenge. We believe that this was mainly due to our system’s accurate aerial gripping pipeline. Figure 31 shows a complete 30s gripping and dropping sequence from the second trial. When detecting an object in exploration, a waypoint above the estimated position of the object is set as reference position in the NMPC. In this way, the MAV quickly reaches the detected object and starts the pickup maneuver. The centering locks the object track. Precise state estimation through VIO and disturbance observance allows landing exactly on the disc even in windy conditions.

As shown in Figure 30c, in the two individual challenge trials, our MAVs detected 14 items, touched 13 items with the magnet, picked up 6 items, and delivered 4 objects. This corresponds to a servoing success rate of 93% but a gripping rate of only 46%. The gripping failures were mainly caused by the weak EPM mentioned above or erroneous contact sensing. The only missed servoing attempt resulted from a broken gripper connection (Figure 32e). The two object losses after successful gripping were once caused by erroneous contact sensing (Figure 32c) and once due to a crash after grabbing (Figure 32d).

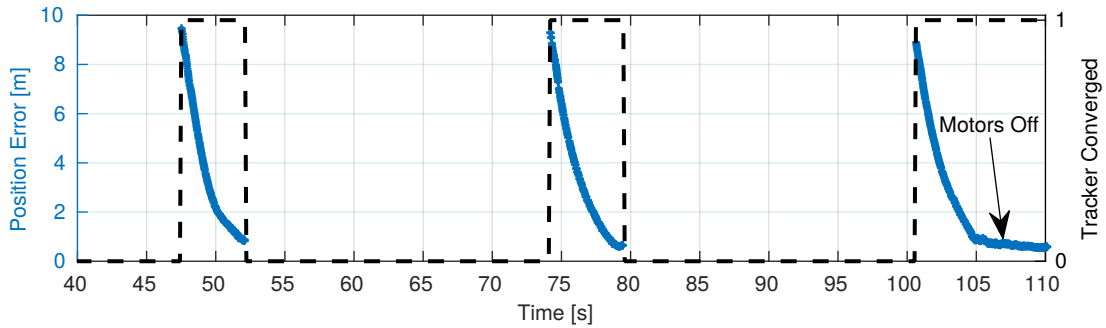
The limited flight time was mainly due to severe CPU overload through serialization of the high resolution nadir camera image. As the state estimation and control ran on the same processor, an overloaded CPU sometimes caused divergence of the MAV. Under these circumstances and given limited safety pilot capabilities, it was difficult to engage multiple MAVs simultaneously.

8.2.2 Grand Challenge Trials

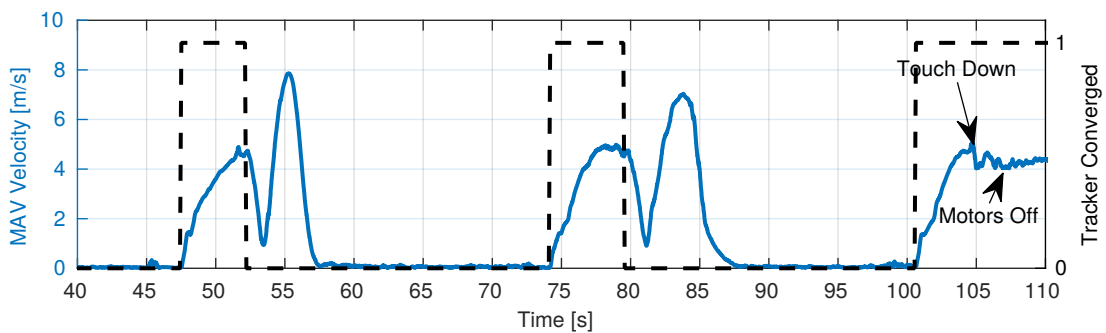
Once we gained more experience in the system setup, we were able to increase the flight time in the Grand Challenge. In the last trial, when we could also afford risk, we even used all three MAVs simultaneously. Furthermore, during the Grand Challenge we deactivated the Hall sensing as it has shown to be unreliable



(a) Detection rate during challenge run. A detection may come from any of the two detectors but must pass the outlier rejection. When the platform is clearly visible, the detection rate approaches the 50Hz output rate of the camera. Plotted rate is smoothed by a size 3 Gaussian kernel with $\sigma = 1$. Upwards facing triangles mark detections using the quadrilateral detector, downwards facing triangles mark detections using the cross detector.



(b) Position error of the MAV and the estimated position of the platform center. The fast reduction of distance between the MAV and the estimated target is recognizable in all three attempts.



(c) Magnitude of the MAV velocity vector. Periods with almost zero velocity indicate that the MAV was hovering above the center of the arena, waiting for the landing platform to enter its FoV.

Figure 28: Challenge 1: On-board data collected during the first Grand Challenge trial. Time starts when the starting signal was given by the judges. The first 60s are cut since the platform was not in the MAV FoV. The first two landing attempts were automatically aborted due to weak tracker convergence and the consequential noncompliance of the Lidar safety check, whereas in the last one the MAV landed successfully. In each plot, the dashed black line shows when the tracker converged.

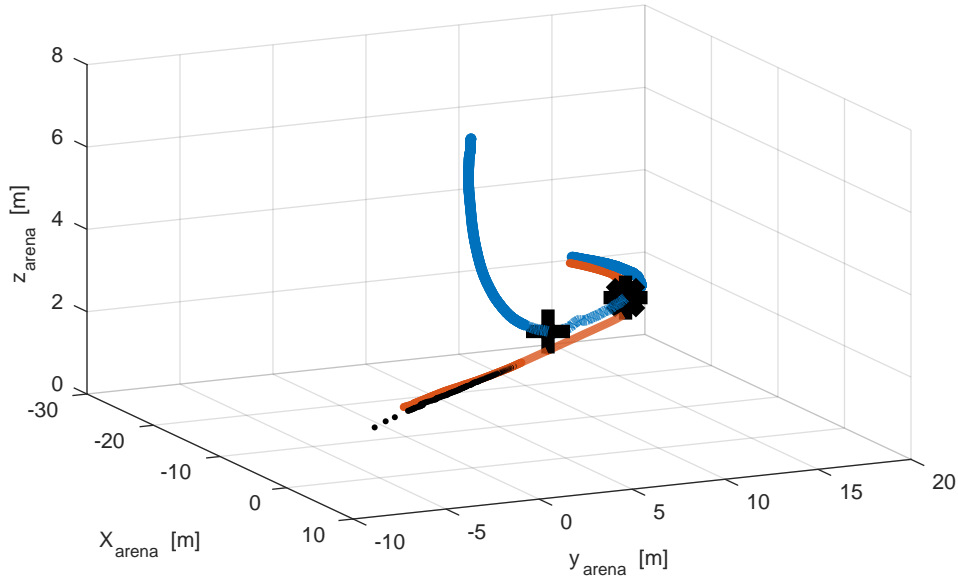


Figure 29: Final landing trajectory visualized in arena coordinates. The blue line corresponds to the MAV path, and the red line to the estimated platform path. The black cross marks the touchdown position, while the black asterisk indicates the position where the motors have been shut down. The plot begins at the time of tracker convergence. Each small black dot represents a visual detection of the target by any of the two detectors. Touchdown occurs 4.06 s after convergence and shutdown 6.33 s after convergence.

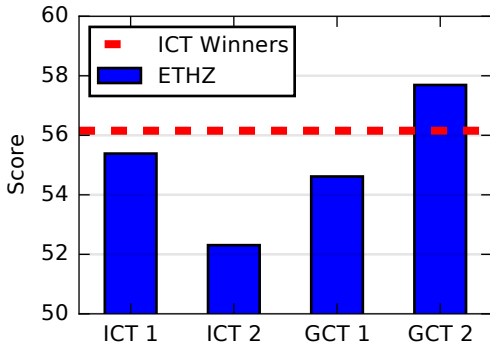
in the individual challenge trials. Instead we always inferred successful object pick up after servoing. This resulted in a maximum number of four delivered objects and a maximum score that surpassed the individual contest winner’s score (see Figure 30a and Figure 30b).

8.3 Lessons Learned

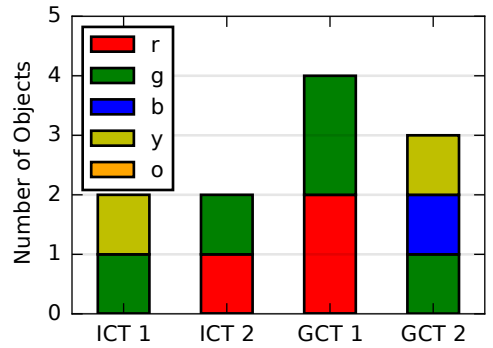
In general, we created two very successful robotic solutions from established techniques. Except for the hardware and low-level control system, our entire framework was developed and integrated in-house. We believe that especially the VIO-based state estimation and NMPC-pipeline differentiated our system from other competing systems which were mostly relying on Pixhawk or DJI GPS-IMU-fusion and geometric or PID tracking control (Loianno et al., 2018; Nieuwenhuisen et al., 2017; Beul et al., 2017; Baca et al., 2017). However, the advantage of a potentially larger flight-envelope came with the cost of integrating and testing a more complex system and greater computational load. Eventually, this resulted in less testing time of the actual challenge components and the complete setup in the preparation phase.

Obviously, this made SIL-testing even more valuable. The simulator described in Section 7.2 was a great tool to test new features before conducting expensive field tests. SIL also allowed changing parameters safely between challenge trials where actual flying was prohibited. Still, hardware and outdoor-tests were indispensable. On the one hand, they reveal unmodelled effects, e.g., wind, lightning conditions, delays, noise, mechanical robustness, and computational load. On the other hand, they lead to improvements in the system infrastructure, such as developing GUIs, simplifying tuning, and automatizing startup-procedures. A lack of testing time induced some severe consequences in the competition. In Challenge 1 a lot of testing was only conducted during the actual trials, where we first did not take off, then crashed, then aborted landing several times before finally showing a perfect landing (see Section 8.1). In Challenge 3 the detection and Hall sensors failures, as described in Section 8.2.1, could have been revealed in more extensive prior testing.

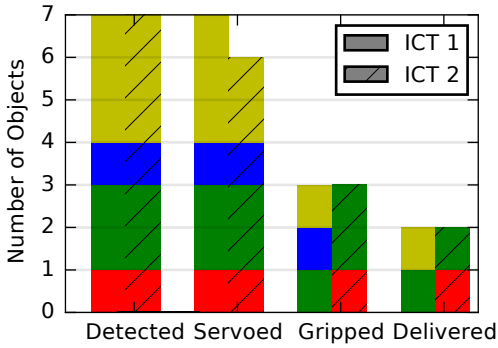
Also setting up a testing area was key in making a good transition between the home court and the compe-



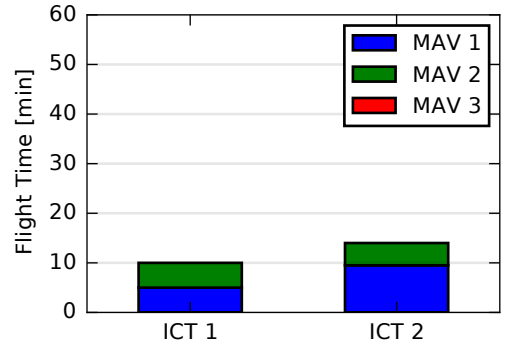
(a) Score as a combination of objects, time, and autonomy for all four trials. We continuously scored in each trial and were able to beat the Individual Challenge Trial (ICT) winner team's record with our last Grand Challenge Trial (GCT).



(b) Numbers and colors of items delivered to the bin. We collected up to four moving or static small objects per run.



(c) Aerial gripping statistics of the two ICTs. 13 out of 14 servoing attempts succeeded, but due to gripper failures, only 6 out of 13 objects were gripped and 4 of those delivered successfully.



(d) The accumulated flight time in the two ICTs. Due to state estimation errors only two MAVs were engaged simultaneously, reaching a flight time of 19 out of 120 min.

Figure 30: Statistics for Challenge 3 from the two individual and the two Grand Challenge trials.

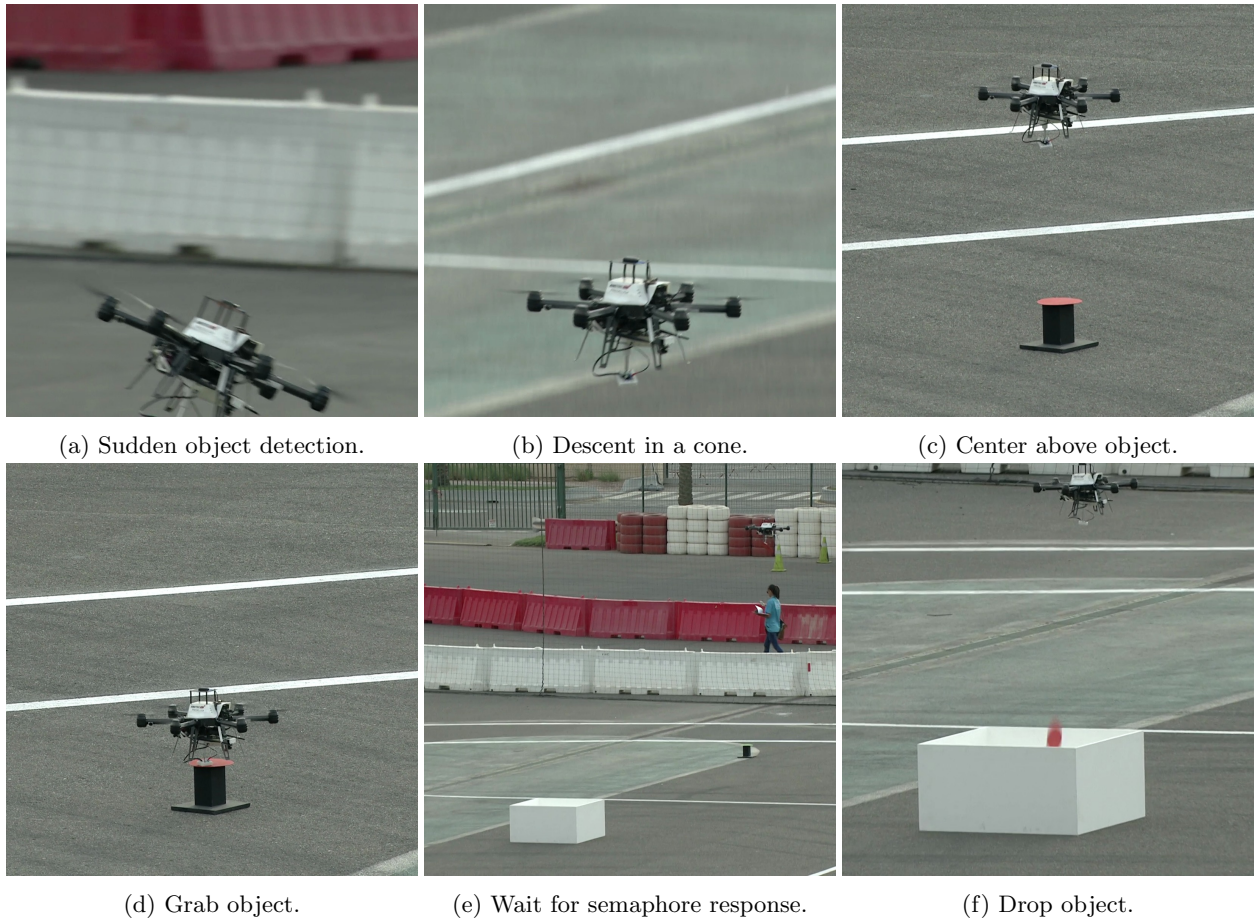


Figure 31: An example aerial gripping sequence from the second trial of Challenge 3. The total time between object detection and drop off is 30s. If an MAV detected an object, it usually servoed it accurately.

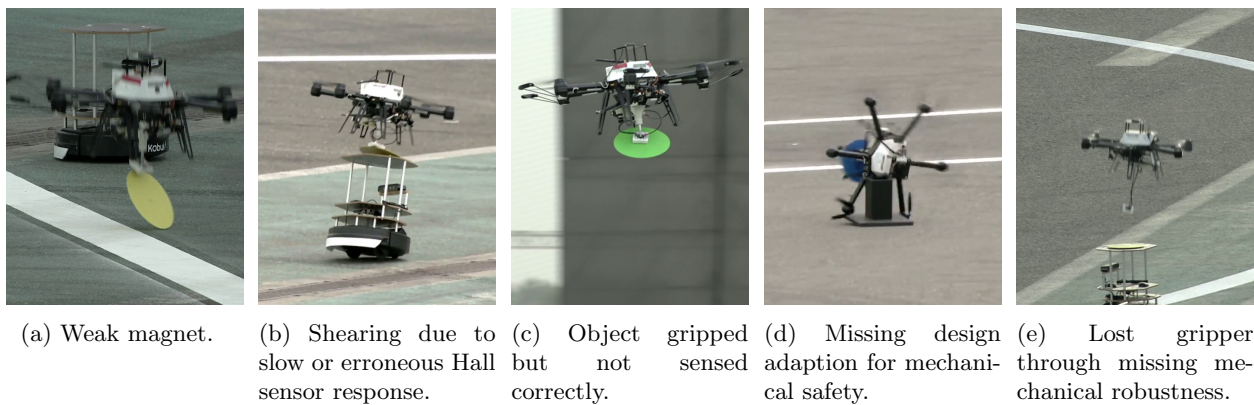


Figure 32: Common aerial gripping failures sorted by descending frequency.

tition arena. One should put great consideration into which parts to reproduce from the challenge specifications. On the one hand, one does not want to lose time on overfitting to the challenge, on the other hand challenge criteria has to be met. While we spent a critical amount of time on reproducing the eight-shaped testing environment for Challenge 1, as described in Section 7.2, we missed some critical characteristics, e.g., object weight, rostrum, and arbitrary paint thickness, in testing for Challenge 3 which eventually led to failed object gripping, and missing mechanical adaptation (see Section 8.2.1).

Furthermore, when developing the systems, we benefited from working with a modular architecture. Every module could be developed, tested, and debugged individually. However, this can also hide effects, that only occur when running the full system. For example, in Challenge 3 we faced state estimation errors and logging issues due to computational overload in a (too) late phase of the development (see Section 8.2.1). Also, whereas testing only one platform instead of all three reduced testing effort and led to a great aerial gripping pipeline, it did not improve multi-agent engagement.

Finally, having proprietary platforms from AscTec which were unfortunately not supported anymore, caused problems when either replacing hardware or facing software issues in the closed-source low level controller of the platform. Eventually, this led to the hardware failure in the first trial of Challenge 1 in Section 8.1.1 and made maintaining the three identical systems in Challenge 3 even more difficult.

9 Conclusion

In conclusion, our team demonstrated the autonomous capabilities of our different flying platforms by competing for multiple days both in Challenge 1, Challenge 3 and in the Grand Challenge. The MAVs were able to execute complex tasks while handling outdoor conditions, such as wind gusts and high temperatures. The team was able to operate multiple MAVs concurrently, owing to a modular and scalable software stack, and could gain significant experience in field robotics operations. This article presented the key considerations we addressed in designing a complex robotics infrastructure for outdoor applications. Our results convey valuable insights towards integrated system development and highlight the importance of experimental testing in real physical environments.

The modules developed for MBZIRC will be further developed and improved to become a fundamental part of the whole software framework of ASL. At the moment, work on manipulators attached to MAVs and cooperation strategies with a UGV is in progress, which will contribute towards dynamic interactions with the environment.

MBZIRC was a great opportunity not only to train and motivate young engineers entering the field of robotics, but also to cultivate scientific research in areas of practical application. We think our system is a good reference for future participants. ASL plans to join the next MBZIRC in 2019, in which we will try to include outdoor testing earlier in the development phase.

Acknowledgments

This work was supported by the Mohamed Bin Zayed International Robotics Challenge 2017, the European Community's Seventh Framework Programme (FP7) under grant agreement n.608849 (EuRoC), and the European Union's Horizon 2020 research and innovation programme under grant agreement n.644128 (Aeroworks) and grant agreement n.644227 (Flourish). The authors would like to thank Abel Gawel and Tonci Novkovic for their support in the hardware design of the gripper, Zachary Taylor for his help in field tests, Fadri Furrer, Helen Oleynikova and Michael Burri for their thoughtful coffee breaks and code reviews, Zeljko and Maja Popović for their warm welcome in Abu Dhabi and their filming during the challenge, and the students Andrea Tagliabue and Florian Braun for their active collaboration.

References

- Achtelik, M., Achtelik, M., Weiss, S., and Siegwart, R. (2011). Onboard imu and monocular vision based control for mavs in unknown in- and outdoor environments. In *Conference on Robotics and Automation (ICRA)*. IEEE.
- Akinlar, C. and Topal, C. (2011). Edlines: A real-time line segment detector with a false detection control. *Pattern Recogn. Lett.*, 32(13):1633–1642.
- Baca, T., Stepan, P., and Saska, M. (2017). Autonomous landing on a moving car with unmanned aerial vehicle. In *Mobile Robots (ECMR), 2017 European Conference on*, pages 1–6. IEEE.
- Bähnemann, R., Schindler, D., Kamel, M., Siegwart, R., and Nieto, J. (2017). A decentralized multi-agent unmanned aerial system to search, pick up, and relocate objects. In *International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE.
- Beul, M., Houben, S., Nieuwenhuisen, M., and Behnke, S. (2017). Fast autonomous landing on a moving target at mbzirc. In *Mobile Robots (ECMR), 2017 European Conference on*, pages 1–6. IEEE.
- Blanco, J. L. (2014). nanoflann: a C++ header-only fork of FLANN, a library for nearest neighbor (NN) with kd-trees. <https://github.com/jlblancoc/nanoflann>.
- Bloesch, M., Omari, S., Hutter, M., and Siegwart, R. (2015). Robust visual inertial odometry using a direct ekf-based approach. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 298–304.
- Carius, J., Wermelinger, M., Rajasekaran, B., Holtmann, K., and Hutter, M. (2018). The ETH-UGV team in the MBZ International Robotics Challenge. *Journal of Field Robotics (JFR)*. (Under Review).
- Chameleon 2 Datasheet (2017). Retrieved July 12, 2017, from <https://eu.ptgrey.com/chameleon-usb2-cameras>.
- Chameleon 3 Datasheet (2017). Retrieved October 09, 2017, from <https://www.ptgrey.com/support/downloads/10578>.
- Furgale, P., Rehder, J., and Siegwart, R. (2013). Unified temporal and spatial calibration for multi-sensor systems. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE.
- Furrer, F., Burri, M., Achtelik, M., and Siegwart, R. (2016). *Robot Operating System (ROS): The Complete Reference (Volume 1)*, chapter RotorS—A Modular Gazebo MAV Simulator Framework, pages 595–625. Springer International Publishing, Cham.
- Gadeyne, K. (2001). BFL: Bayesian Filtering Library. <http://www.oroocos.org/bfl>.
- Gawel, A., Kamel, M., Novkovic, T., Widauer, J., Schindler, D., von Altshofen, B. P., Siegwart, R., and Nieto, J. I. (2017). Aerial picking and delivery of magnetic objects with mavs. In *Conference on Robotics and Automation (ICRA)*. IEEE.
- Gazebo (2017). Retrieved July 17, 2017, from <http://gazebo.org/>.
- Kalibr Github (2017). Retrieved July 17, 2017, from <https://github.com/ethz-asl/kalibr>.
- Kamel, M., Alonso-Mora, J., Siegwart, R., and Nieto, J. (2017a). Nonlinear model predictive control for multi-micro aerial vehicle robust collision avoidance. In *International Conference on Intelligent Robots and Systems (IROS)*.
- Kamel, M., Burri, M., and Siegwart, R. (2017b). Linear vs nonlinear mpc for trajectory tracking applied to rotary wing micro aerial vehicles. In *International Federation of Automatic Control (IFAC)*, volume 50, pages 3463–3469. Elsevier.
- Kaplan, E. (2005). *Understanding GPS - Principles and applications*. Artech House.

- Kuhn, H. W. (1955). The hungarian method for the assignment problem. In *Naval Research Logistics (NRL)*, volume 2. Wiley.
- Lidar Datasheet (2017). Retrieved July 12, 2017, from <http://buy.garmin.com/en-US/US/p/557294>.
- Loianno, G., Spurny, V., Baca, T., Thomas, J., Thakur, D., Hert, D., Penicka, R., Krajnik, T., Zhou, A., Cho, A., et al. (2018). Localization, grasping, and transportation of magnetic objects by a team of mavs in challenging desert like environments. *IEEE Robotics and Automation Letters*.
- Lynen, S., Achteik, M., Weiss, S., Chli, M., and Siegwart, R. (2013). A robust and modular multi-sensor fusion approach applied to mav navigation. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE.
- MAV Control Github (2017). Retrieved October 12, 2017, from https://github.com/ethz-asl/mav_control_rw.
- MSF Github (2017). Retrieved July 12, 2017, from https://github.com/ethz-asl/ethzasl_msf.
- Mueller, M. W., Hehn, M., and D’Andrea, R. (2015). A computationally efficient motion primitive for quadcopter trajectory generation. *IEEE Transactions on Robotics*, 31(6):1294–1310.
- NicaDrone OpenGrab EPM Datasheet v3 (2017). Retrieved October 09, 2017, from http://nicadrone.com/index.php?id_product=66&controller=product.
- Nieuwenhuisen, M., Beul, M., Rosu, R. A., Quenzel, J., Pavlichenko, D., Houben, S., and Behnke, S. (2017). Collaborative object picking and delivery with a team of micro aerial vehicles at mbzirc. In *Mobile Robots (ECMR), 2017 European Conference on*, pages 1–6. IEEE.
- Nikolic, J., Rehder, J., Burri, M., Gohl, P., Leutenegger, S., Furgale, P. T., and Siegwart, R. (2014). A synchronized visual-inertial sensor system with fpga pre-processing for accurate real-time slam. *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 431–437.
- Node Manager (2017). Retrieved July 17, 2017, from http://wiki.ros.org/multimaster_fkcie.
- Olson, E. (2011). AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3400–3407. IEEE.
- Piksi Accuracy (2017). Retrieved July 17, 2017, from <https://support.swiftnav.com/customer/en/portal/articles/2492803-understanding-gps-rtk-technology>.
- Piksi Datasheet V2 (2017). Retrieved July 12, 2017, from http://docs.swiftnav.com/pdfs/piksi_datasheet_v2.3.1.pdf.
- Piksi Github (2018). Retrieved February 02, 2018, from https://github.com/ethz-asl/ethz_piksi_ros.
- Rovio Github (2017). Retrieved July 12, 2017, from <https://github.com/ethz-asl/rovio>.
- Scaramuzza, D. and Fraundorfer, F. (2011). Visual odometry [tutorial]. *IEEE Robotics Automation Magazine*, pages 80–92.
- Smach (2017). Retrieved July 17, 2017, from <http://wiki.ros.org/smach>.
- Tagliabue, A., Kamel, M., Siegwart, R., and Nieto, J. (2017a). Robust collaborative object transportation using multiple mavs. *arXiv preprint arXiv:1711.08753*.
- Tagliabue, A., Kamel, M., Verling, S., Siegwart, R., and Nieto, J. (2017b). Collaborative object transportation using mavs via passive force control. In *Conference on Robotics and Automation (ICRA)*. IEEE.
- Thrun, S. (2002). Particle filters in robotics. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 511–518. Morgan Kaufmann Publishers Inc.
- USB 3.0 Interference (2017). Retrieved July 17, 2017, from <https://www.intel.com/content/www/us/en/io/universal-serial-bus/usb3-frequency-interference-paper.html>.