Journal of Intelligent & Robotic Systems (2020) 100:1545–1557
https://doi.org/10.1007/s10846-019-01060-8

Guidance for Autonomous Aerial Manipulator Using Stereo Vision

Christoforos Kanellakis[1] · **George Nikolakopoulos**[1]

Received: 4 September 2018 / Accepted: 2 July 2019 / Published online: 16 August 2019

Abstract

Combining the agility of Micro Aerial Vehicles (MAV) with the dexterity of robotic arms leads to a new era of Aerial Robotic Workers (ARW) targeting infrastructure inspection and maintenance tasks. Towards this vision, this work focuses on the autonomous guidance of the aerial end-effector to either reach or keep desired distance from areas/objects of interest. The proposed system: 1) is structured around a real-time object tracker, 2) employs stereo depth perception to extract the target location within the surrounding scene, and finally 3) generates feasible poses for both the arm and the MAV relative to the target. The performance of the proposed scheme is experimentally demonstrated in multiple scenarios of increasing complexity.

Keywords Vision based guidance · Aerial manipulator · MAV

1 Introduction

MAVs are platforms that embody a significant active research effort within the robotics community, since they are characterized by simple mechanical design and versatile movement. These capabilities are suitable for the execution of complex tasks which are impossible or dangerous for a human operator to perform. These platforms, so far, have been integrated in the photography-filming industry, but, more and more resources are invested towards remote inspection applications. Some examples of up-to-date efforts to employ MAVs include infrastructure inspection [1, 2], public safety such as surveillance [3] and search and rescue missions [4].

A new trend that is currently emerging with fast pace includes the interaction capabilities of such platforms. Instead of carrying only sensors, ARWs could be endowed with lightweight dexterous robotic arms as depicted in Fig. 1, expanding their operational workspace [5, 6].

✉ Christoforos Kanellakis
chrkan@ltu.se

George Nikolakopoulos
geonik@ltu.se

[1] Robotics Group, Department of Computer, Electrical and Space Engineering, Luleå University of Technology, Luleå SE-97187, Sweden

Generally, the vision of integrating aerial robotic platforms in the industrial process is an emerging research movement in its infancy, with quite a few open challenges. Advanced localization, physical interaction, navigation and perception are capabilities that an ARW should possess when employed for the infrastructure inspection and maintenance tasks. Among these topics, the scope of this article is to propose a system with advanced perception capabilities, as the middle step before the manipulation task. These capabilities are primarily expressed by augmenting the environmental awareness of the robotic vehicle with detection modules. The detection modules are developed to identify targets with specific characteristics like shape, color, texture. The target recognition is coupled with the stability of the multirotor vehicle, since the control modules process the information of the image processing step. An industrial environment can be harsh and pose various challenges in the visual part, like illumination changes, occlusions, and target losses. Therefore the combination of visual processing with machine learning could be one of the most robust approaches in terms of object tracking.

Only a limited number of works have considered the visual guidance system as a means to assist the manipulation task. More specifically in [7], a vision-based guidance system for a 3 degrees of freedom (DoF) manipulator has been developed. This work presented an image-based visual servoing (IBVS) scheme using image moments to derive the velocity references for commanding the coupled system (MAV and manipulator), while the object detection was based on color thresholding. An adaptive controller was

Fig. 1 An aerial robotic worker

designed to switch between position and IBVS control, while the authors of [7] extended their work on manipulation in [8], by proposing a guidance system for cylindrical objects, where the detection has been performed using random and sampling consensus (RANSAC) ellipse detection. In this work a stochastic Model Predictive Control (MPC) has been employed to handle x and y rotational velocities as stochastic variables. In [9] an aerial manipulator guidance system has been presented, where the novelty of this work stems from the designed hierarchical control law that prioritizes tasks like collision avoidance, visual servoing, center of gravity compensation and joint limit avoidance during a flight. In [10] a tree cavity inspection system has been presented based on depth image analysis and image processing, while the overall goal was to drive the end-effector inside the cavity. In [11], a stereo vision system for object grasping has been proposed with a detection algorithm to learn a feature-based model in an off line stage and then use it online to detect the targeted object and estimate its relative pose. Finally, in [12] a hybrid visual servoing with a hierarchical task-priority control framework for MAVs has been presented. In this work a hybrid control framework has been developed combining image-based as well as position-based visual servoing for the target approaching.

The main aim of the current manuscript is to extend the state of the art of visual processing on guidance for aerial manipulation, by proposing an experimentally evaluated guidance system with two major merits. Firstly the ability to detect and track generic objects, without focusing on specific characteristics (geometry, shape, motion, color) compared to [7–10, 12]. Secondly by combining the robust object tracking with the stereo vision, the system is applicable to textured and planar targets compared to [11]. In this work the guidance system is limited to approaching the target without performing any interaction with the target. More specifically, the stereo guidance module is introduced to bring the target in the active workspace of the ARW. Additionally, in this work the implemented object tracker is based on the Kernelized Correlation Filter (KCF) [13]. This tracker provides a high speed performance and robust tracking efficiency, while it works for generic type of targets. Finally, this work is among the few that report experimental trials, considering target monitoring tasks, depicting the performance of the proposed guidance system.

The rest of this article is organized as follows. In Section 2 the hardware and software components of the experimental system are discussed, while in Section 3 the kinematic modeling and control of the robotic platform is presented. In Section 4 the vision guidance framework for aerial manipulation is established, including the object tracker and the stereo processing parts and in Section 5 multiple experimental results that prove the efficacy of the proposed scheme are presented. Finally in Section 6, the conclusions are drawn.

2 System Description

2.1 AscTec NEO Hexacopter

This work employs the aerial research platform from Ascending Technologies, the NEO hexacopter, depicted in Fig. 1. The platform specifications are summarized in Table 1. It is also equipped with an onboard flight controller with a tuned low-level attitude controller. The onboard computer communicates with the flight controller at 100 Hz through a serial port, while the state estimation is performed by combining pose measurements with the onboard IMU.

2.2 The CARMA Aerial Manipulator

The robotic arm introduces manipulation capabilities to the multirotor and it is a planar robotic arm with 4 revolute joints mounted underneath the aerial platform, as shown in Fig. 2. The manipulator weight 500 gr, while it is capable of holding various types of end-effectors like a grasper, a brusher, a camera holder, or even an electromagnet for lifting heavy objects.

Table 1 AscTec NEO hardware and software specifications

AscTec NEO	
Diameter	0.59 m
Height	0.24 m
Propeller length	0.28 m
Payload	2 kg
Processing unit	Intel NUC i7-5557U
Flight time	Max 15 min
OS	Ubuntu Server 14.04

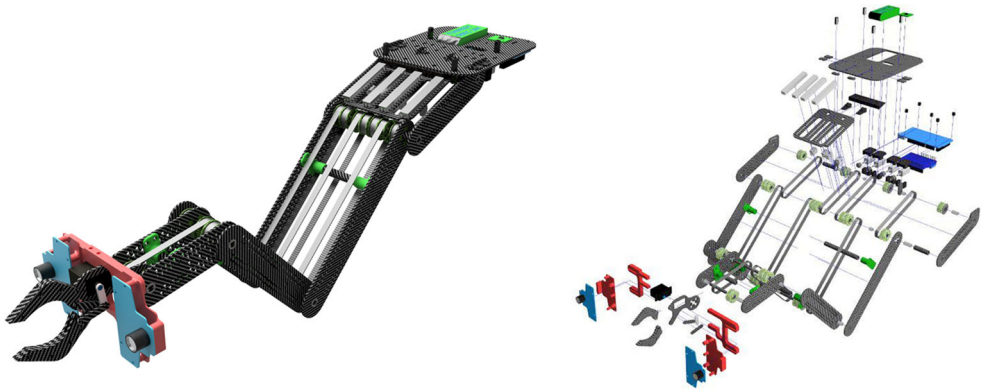


Fig. 2 Left - CAD design of the CARMA manipulator. Right - CARMA parts explosion view (Video Link at [14])

Some highlights on the design of the manipulator are the following:

- a robust and sturdy mechanism with belts for motion transmission
- linear potentiometers for joint angle feedback.
- multiple end-effector types

Compact AeRial MAnipulator (CARMA) is regulated using a cascaded position-velocity Proportional Integral Derivative (PID) control scheme. More specifically, the joint positions derived from the inverse kinematics consist the reference to four standalone PID controllers, one controller for every joint. A full description on the design and modeling of the manipulator was presented in [15].

2.3 Visual Sensor

The onboard system of sensors used, consists of a custom made stereo camera depicted in Fig. 3. The stereo camera is attached on the end-effector in an eye-in-hand configuration for the target detection and tracking tasks. The camera frame rate is set to 20fps at the resolution of 640x480 pixels. The baseline of the stereo sensor is 10 cm. All processing considers pre-calibrated visual sensor with known intrinsic and extrinsic parameters.

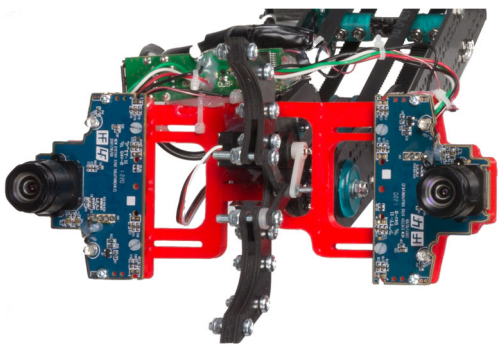


Fig. 3 Visual Sensor in an eye-in-hand configuration for aerial manipulation (Video Assembly at [14])

The software architecture of the complete vision system is modular and has the merit of integrating localization, control and guidance subsystems . A graphical overview of the proposed architecture and the utilized novel combination of the software components is provided in Fig. 4, where regarding the stereo module I_1, I_2 are camera frames, $\mathcal{P}$ and $\mathcal{P}_{bounded}$ are pointclouds, B is the bounding box and x^c, y^c, z^c are the centroid coordinates and waypoints $p^{\mathcal{W}}$, ϕ. From the perspective of the aerial vehicle the motor commands υ_v and υ_m for the hexarotor and the manipulator are generated using pose and twist measurements from the Motion Capture system (moCap) and IMU multi-sensor fusion. The software is implemented in C++, using ROS[1] framework and OpenCV[2] and PCL[3] libraries.

The guidance components consist of an integrated stereo based system (described in Section 4.2)). In both cases, the target is identified within the sequential frames (provides a bounding box), using the proposed robust detection scheme (described in Section 4.1). The former is used to extract the centroid of the manipulated object, compute its relative configuration with respect to the MAV, generate proper trajectory and align the end-effector properly with the grasping point, by processing the pointcloud generated from the stereo camera. Thus, the vision system is able to generate joint position commands for the manipulator and pose commands for the multirotor. All computations regarding the detection and tracking components are executed onboard the MAV, to avoid communication latency issues. The detection initialization is performed using an external station, allowing the user to select the object of interest, while communicating through a wireless link.

The multirotor includes three main subsystems to provide autonomous flight, namely the localization system based Vicon MoCap,[4] a Multi-Sensor-Fusion Extended Kalman Filter (MSF-EKF) [16] for state estimation and

[1] http://www.ros.org/

[2] https://opencv.org/

[3] http://pointclouds.org/

[4] https://www.vicon.com/

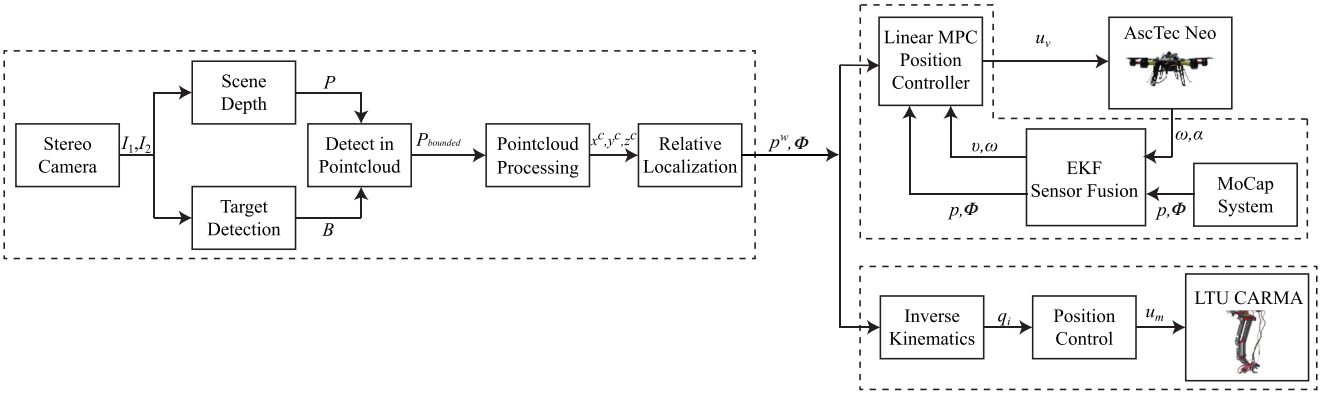


Fig. 4 Overall system software architecture of the guidance system

finaly the linear Model Predictive Control (MPC) position controller [17–19] for trajectory following.

The manipulator's forward and inverse kinematics are interfaced to support the guidance systems in setting/estimating the arm configuration. Moreover, the kinematics of the robotic arm consider and compensate the MAV pitch (from the odometry of the vehicle). The manipulator is endowed underneath the aerial platform and the manipulator base has a fixed position relative to the MAV center of mass. In the developed system the manipulator kinematics define the end-effector position relative to the manipulator base. In case the MAV does a pitch command the level of the manipulator base changes and the end-effector position is affected. A cascade joint position and velocity controller are implemented to control each joint, while the calculated joint variables are inserted in four independent cascade PID controllers.

3 Reference Frames

In this established framework, several coordinate frames are used as depicted in Fig. 5. The world frame $\mathcal{W}$ is fixed inside the workspace of the robotic platform, the body frame of the vehicle $\mathcal{B}$ is attached on its base, while the manipulator's frame $\mathcal{M}$ is fixed on the base of the manipulator. Finally, the stereo camera frame $\mathcal{C}$ origins on the left camera and is firmly attached to the end-effector frame $\mathcal{E}$. The transformation of the point $p^{\mathcal{C}}$ to the frame $\mathcal{E}$ is expressed through the homogeneous transformation matrix $T_{\mathcal{C}}^{\mathcal{E}}$ ($p^{\mathcal{E}} = T_{\mathcal{C}}^{\mathcal{E}}\ p^{\mathcal{C}}$). For the rest of this article the superscript denotes the reference frame. Accordingly, $p^{\mathcal{E}}$ can be expressed in the manipulator's frame $\mathcal{M}$, using the forward kinematics. More specifically, $p^{\mathcal{M}} = T_{\mathcal{E}}^{\mathcal{M}}(q)\ p^{\mathcal{E}}$, where $T_{\mathcal{E}}^{\mathcal{M}}(q)$ is the homogeneous transformation matrix from the end effector's frame to the base frame, which depends on the current manipulator joint configuration $q = [q_1, \cdots, q_n]$. Finally, the manipulator is firmly attached to the MAV, thereafter the transformation matrix $T_{\mathcal{M}}^{\mathcal{B}}$ is constant, expressing the relative pose between the vehicle base and the manipulator base. The pose of the target $p^{\mathcal{B}}$, relative to the multirotor base frame, is calculated through $p^{\mathcal{B}} = T_{\mathcal{M}}^{\mathcal{B}}\ p^{\mathcal{M}}$.

4 Vision for Aerial Manipulation

4.1 Object Tracking

On of the baseline components for an ARW to fulfill autonomous guidance for aerial manipulation tasks is perception. More specifically, vision is considered a primary cue because of the rich information it can provide and is the key for a robust and reliable operation of the aerial platform. Within this work, the perception capabilities focus on target detection and tracking using the onboard camera, as well as the stereo processing module to extract the target waypoint. The object tracker forms the core module for a robust and stable aerial guidance system, to address challenges posed in complex environments, such as out-of-view events and

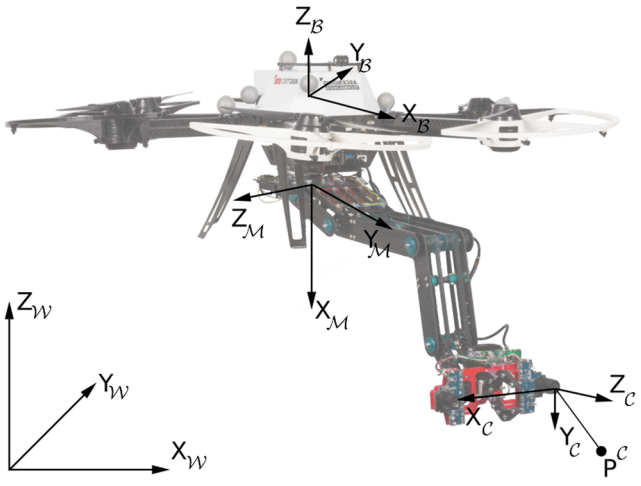


Fig. 5 Coordinate Frames of the aerial platform

background clutter [20] . During the years multiple efficient tracking algorithms [21] have been proposed, but many algorithms are not suitable for MAV applications, since they require high computational resources.

A tracking category that could address these challenges are the tracking-by-detection algorithms. Briefly, these tracking algorithms are treated as binary classification methods, since they constantly try to discriminate between the target and the background using decision boundaries. The tracking mechanism is online using patches of both target and background captured in recent and past frames.

In this article, the tracking-by-detection approach for robust tracking during manipulation guidance is based on the Kernelized Correlation Filter (KCF).

The outcome of this process results in a 2D bounding box, defined as a set B with x_b and y_b cooridnates (1).

$$B = \{(x_b, y_b) \in \mathbb{R}^2 | \, x_{min} < x_b < x_{max} \, , \; y_{min} < y_b < y_{max}\} \tag{1}$$

where $\{(x_{min}, y_{min}), (x_{min}, y_{max}), (x_{max}, y_{min}), (x_{max}, y_{max})\}$ are the four corners of the bounding box in the image plane.

4.2 Stereo Based Guidance

A major part of the proposed system includes the guidance layer based on stereo vision during the exploration phase of the MAV. This part is used when the target of interest lies within the depth range of the stereo camera. The goal is to bring the aerial platform in the proximity of the target by following a simple but efficient strategy.

The basis of the 3D perception of the system is structured around the reconstruction capabilities of the stereo sensor. The overall process is initiated by calculating the 3D structure of the area perceived from the stereo pair, using Semi Global Block Matching (SGBM) [22] method. The stereo mapping function $S(x, y)$ maps a point (x,y) from the image pixel coordinate frame to the camera frame as shown in Eq. 2.

$$S(x, y) = (X, Y, Z) \tag{2}$$

Thus, a pointcloud $\mathcal{P}$ is formulated as $\mathcal{P} = \{S(x, y)\}$.

A pointcloud filtering method is proposed to robustly isolate the region of interest, combining information from both the dense mapping and the object tracker presented in Section 4.1. More specifically, the points belonging to the 2D bounding box B are translated to a pointcloud $\mathcal{P}_{bounded}$ using the stereo mapping function as

$$\mathcal{P}_{bounded} = \{S(x, y) | \, x \in x_b \, , \; y \in y_b\} \tag{3}$$

In the proposed system the centroid extraction depends on the processed pointlcoud, therefore additional background parts in the model will downgrade the accuracy of the centroid. Therefore the clustering method Region Growing Segmentation [23], part of the pointcloud processing component (Fig. 4) is implemented using smooth constraints, to partition $\mathcal{P}_{bounded}$ into separate regions. The clustering of the bounded 3D points into groups is selected to remove parts of $\mathcal{P}_{bounded}$ that do not belong to the desired target and are directly passed from the object tracker. Usually, the extracted bounding box does not entirely enclose the target but also includes parts of the background.

The assumption in the proposed process is based on the concept that the target of interest covers the largest part of the bounding box and therefore the largest part of $\mathcal{P}_{bounded}$. The size of every cluster in $\mathcal{P}_{bounded}$ is verified by a heuristic threshold that has been designed to further merge neighboring clusters that do not meet size requirements. In this manner the 3D centroid of the target in $\mathcal{P}_{bounded}$ lies in the cluster with the maximum area. Finally, the centroid $[x^c, y^c, z^c]$ is extracted as the average position of the point in the cluster. Overall, there is no metric information of the target provided a-priori.

On top of the already described process, the pointcloud is filtered to remove invalid values with the aim of further refining the centroid position. It is also downsampled to reduce the number of points through Voxel Grid Filtering [24] for faster processing which is critical for the aerial platform. An extra step is considered for targets that are attached in planar surfaces, where the background plane is segmented using RANSAC [25]. Figure 6 provides a stepwise visualization of the pointcloud filtering process. In the clustered point cloud the points include only the circle and cross parts of the target, while the white background is merged after the final filtering step as shown in the right.

The centroid information is transfered to the body frame of the aerial vehicle $\mathcal{B}$ using the transformation from camera as well as the manipulators kinematics. The stereo guidance subsystem is finalized with the generation of the proper waypoint $Wp = [p^W, \Phi]$ using the extracted centroid location, where p^W represent the x, y, z positions in frame $\mathcal{W}$, while Φ the orientation of the MAV in frame $\mathcal{W}$. In this case the aerial manipulator is given a predefined joint configuration q_1, q_2, q_3, q_4 according to the task requirements. The MAV waypoint is converted into position-velocity-yaw trajectory, which is provided to the utilized linear model predictive controller. The trajectory generator takes into account the sampling time T_s of the position controller and the desired velocity along the path, denoted by $\vec{V}_d$. The trajectory points are obtained by linear interpolation between the waypoints, in such a way that the distance between two consecutive trajectory points equals the step size $h = T_s||\vec{V}_d||$. The velocities are then set parallel to each waypoint segment and the yaw angles are also linearly interpolated with respect to the position within the segment. The overall process is summarized in Algorithm 1.

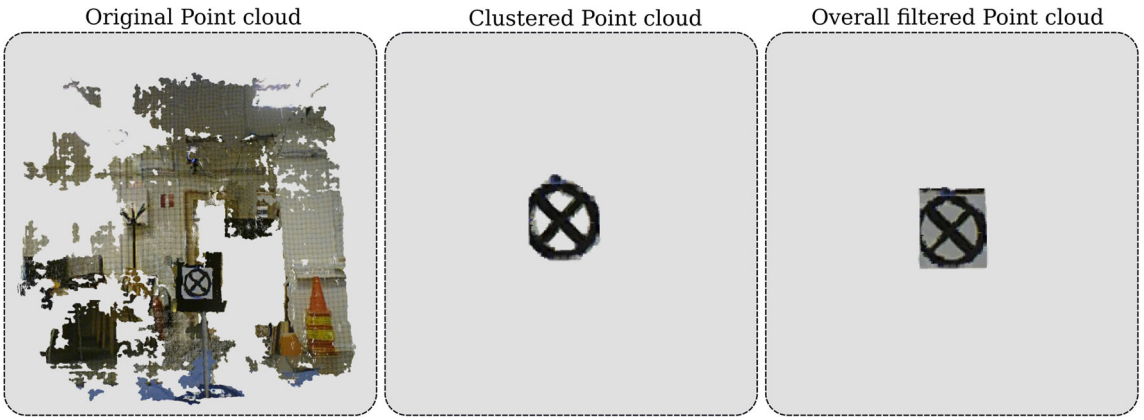


Fig. 6 Pointcloud filtering steps, on the left the original pointcloud, in the middle the clustered pointcloud and on the right the final filtered pointcloud including the whole target

Algorithm 1 Stereo guidance.

Select object to track
for $\{i : 1 : \#frames\}$ **do**
$\{x_b, y_b\} \leftarrow$ Object Tracker
$\mathcal{P}_{bounded} \leftarrow$ Stereo Mapping(x_b, y_b)
$x^c, y^c, z^c \leftarrow$ Pointcloud processing$(\mathcal{P}_{bounded})$
$q_1, q_2, q_3, q_4 \leftarrow$ Manipulator Joint configuration(x^c, y^c, z^c)
$p^W, \Phi \leftarrow$ Waypoint Extraction(x^c, y^c, z^c)
end for

5 Experimental Results

The developed guidance system has been extensively tested in real scale experimental trials. The evaluation was performed indoors in the Field Robotics lab flight arena located at Luleå University of Technology. The flight arena covers a volume of $5 \times 5 \times 3$ m^3. The validation process is two-fold, representing each part of the proposed system. More specifically, the tests were focused, initially, on the performance of the visual tracking standalone system. The second validation step considered the guidance submodule based on stereo processing for the case of target monitoring.

5.1 Visual Tracking

This experimental part is designed to demonstrate the performance of the tracker, while the MAV is flying close to the target of interest. These experiments include the manual navigation of the ARW in the frontal area of the object of interest following different paths, including hovering, longitudinal and lateral motions. The main goal is to provide an insight of the tracker capabilities to track targets with different characteristics (e.g. shape, color) during the deployment of the aerial manipulator, while on the other hand analyze the computation time of this module. To this end, the trials have been performed considering three different types of objects to track: 1) a planar pattern, 2) a custom 3D printed object with rectangular base housing a semicircle, and 3) a screwdriver tool, which are targets with incremental complexity. Moreover, the computational analysis considers the execution times of the aforementioned parts using the available hardware system (as presented in Section 2), while it has been realized through ROS.

Figure 7 demonstrates the use of KCF in the current guidance system. More specifically, the figure provides snapshots of different instances from the onboard visual sensor of the two objects, showing the ability to continuously monitor the target that lies within the field of view of the camera.

The system has undergone an analysis of the computation time for the most critical parts 1) the object tracker and 2) the point cloud processing part of the stereo module. The results consider the execution time for 100 executions of each part, which are visualized in following histograms (Fig. 8a, b). The results show that the stereo processing module is the most computational demanding process of the proposed system with an average performance of 0.4584 sec per run. On the other hand, the tracker execution time averages an 0.0121sec. Additional timing dependencies of the system depend on the internal communication architecture of ROS, on network latencies, as well as the camera frame rate.

5.2 Stereo Based Guidance

This section presents the validation of the proposed system for a target monitoring mission. More specifically, the end-effector of the aerial platform is autonomously guided to a desired position relative to the target in an initially unknown environment, without performing any physical interaction. The experimental trials examine the performance of the

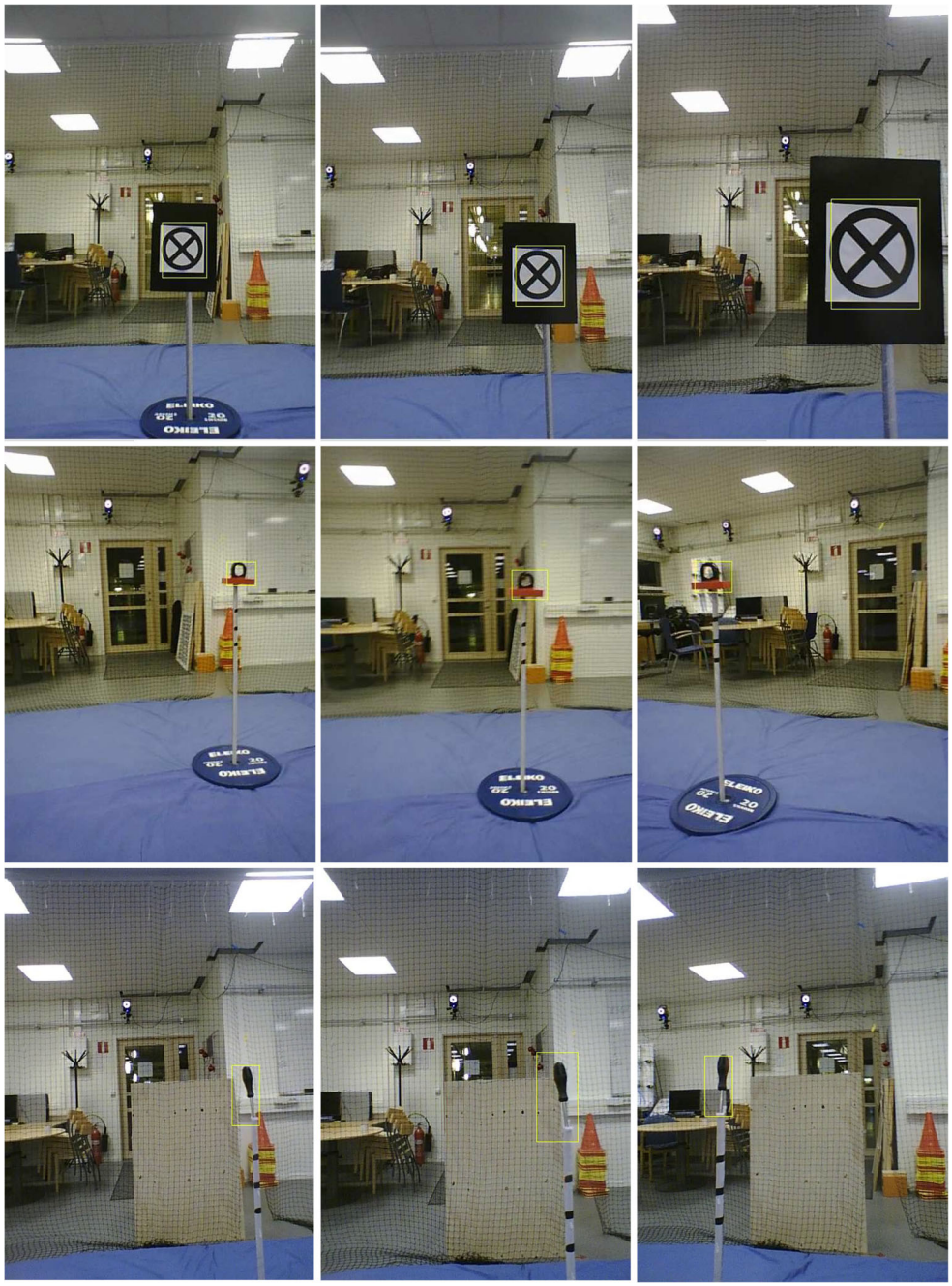
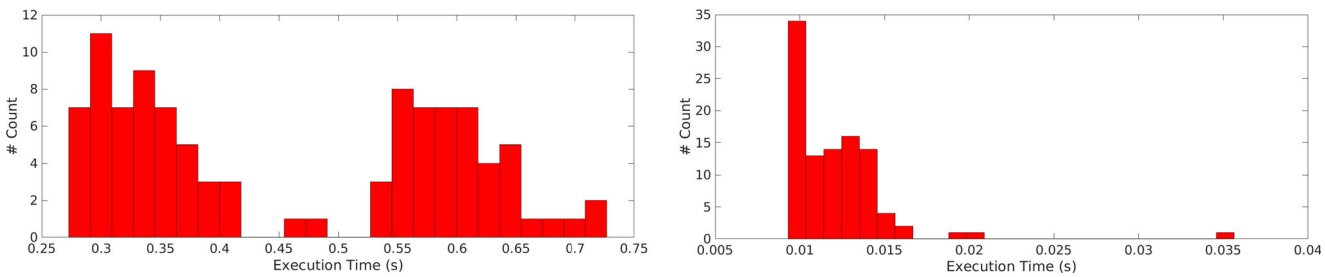


Fig. 7 Experimental tracking results for three different objects. In the first row, multiple snapshots of the tracking process for the object 1 have been extracted, in the second row, multiple snapshots of the tracking process for the object 2 have been extracted, while in the third row, multiple snapshots of the tracking process for the object 3 have been extracted (Video Link at: https://youtu.be/a7g_2Ip2VWE)



(a) Stereo module (b) object tracker

Fig. 8 Execution time versus number of encountered time delays after a set of 100 executions for the visual tracker

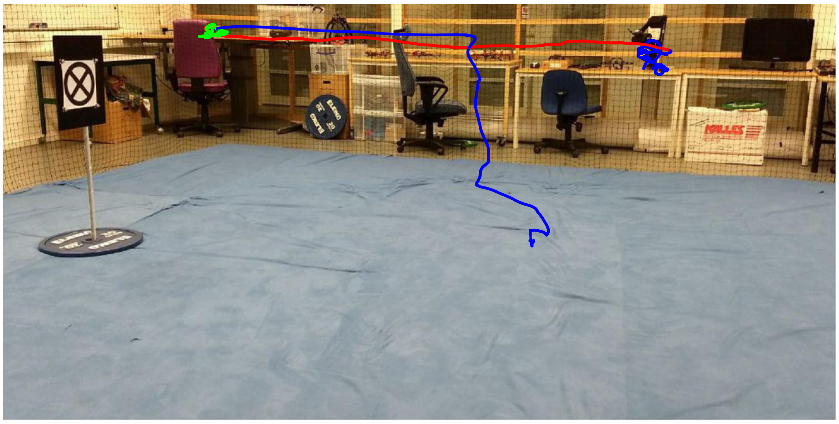


Fig. 9 MAV actual trajectory derived from the experimental trials of the stereo-based guidance. Case 1: relative distance with the target 25 cm. The developed guidance system contributes to the task with the red and green part of the overall trajectory. The red part is followed after the extraction of the centroid, while in the green part the robotic platform is hovering. The blue parts of the trajectory constitute the initialization (hovering on a fixed position) and termination phases (landing) of the experiment. (Video Link at:https://youtu.be/MObjUF1NI-8

system in terms of task execution and accuracy regarding the end-effector - target alignment. A merit of this approach is the depth information derived from the stereo system, which simplifies it's architecture. Nevertheless, it is crucial to mention that the performance depends on the stereo camera specifications.

Initially the aerial vehicle takes off and navigates to a user defined waypoint, using the high level position control. When the MAV reaches the waypoint, the target of interest lies within the field of view of the stereo camera. The next step for the operator is to select the bounding box for the desired target, so that the tracking algorithm can learn online the target for sequential detection, as discussed in the previous section. A generic object of interest is placed on top of a bar inside the flight arena. While the aerial platform hovers at the initial waypoint, the depth from the stereo camera is converted in a pointcloud and is processed using the refining methods to extract its 3D position from the rest of the background. In this manner the relative position between the MAV body frame and the target are calculated. In parallel, the current position of the manipulator is calculated from the forward kinematics to calculate the relative transformation between the end-effector and the MAV base. Afterwards, the end-effector is driven to the final grasping configuration, based on the application requirements, using its inverse kinematics. The joint configuration for the final grasping is predefined, but always considers the position of the object.

Within this work three experimental trials have been performed to showcase the performance of the system in various situations. More specifically, experiments one and two deal with the same target but different monitoring positions, while experiment three presents the system operation with a different target.

Figure 9 depicts the 3D trajectory followed by the aerial platform, while Fig. 10 depicts the path of the end-effector position versus the execution time of the experiment. x_{ee}, y_{ee} and z_{ee} correspond to the end-effector position measurements in the $\mathbb{W}$ frame. The Fig. 10 shows that the proposed approach was able to perform the task and drive the end-effector close to the target approaching the reference values in all axes. The object tracking process, detected and kept the object inside the cameras' field of view during all the phases of the experiment successfully.

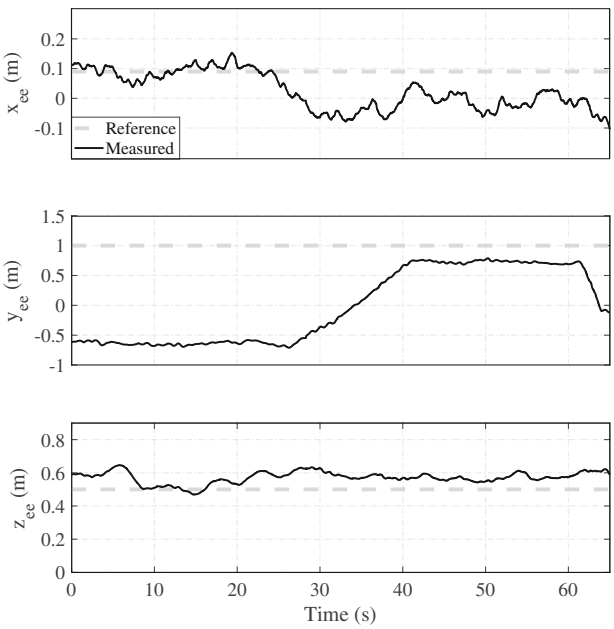


Fig. 10 End-effector setpoints vs the actual setpoints for the first experiment. The plots represent the initialization phase (centroid and waypoint calculation, the waypoint following, the hovering part relative to the target and finally the landing

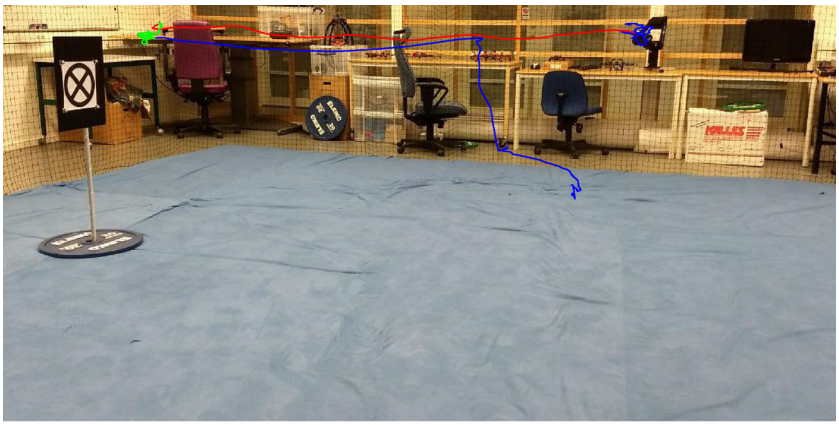


Fig. 11 MAV actual trajectory derived from the experimental trials of the stereo-based guidance. Case 2: the end-effector reaches the target. The developed guidance system contributes to the task with the red and green part of the overall trajectory. The red part is followed after the extraction of the centroid, while in the green part the robotic platform is hovering. The blue parts of the trajectory consist the initialization (hovering on a fixed position) and termination phases (landing) of the experiment. (Video Link at:https://youtu.be/MObjUF1NI-8

Moreover, the MAV is able to hover in front of the object at a desired distance.

Similarly, Fig. 11 depicts the 3D trajectory followed by the aerial platform, while Fig. 12 depicts the path of the end-effector position versus the execution time of the experiment. From the implemented tests the proposed approach was able to perform the task and drive the end-effector close to the target. The object tracking process, detected and followed the object during all the phases of the experiment successfully. Additionally, the method showed satisfactory performance for extracting the target centroid position.

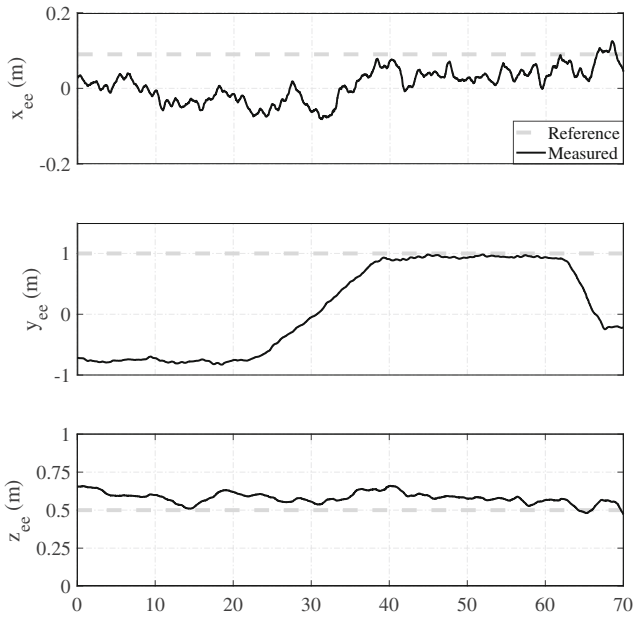


Fig. 12 End-effector setpoints vs the actual setpoints for the second experiment. The plots represent the initialization phase (centroid and waypoint calculation, the waypoint following, the hovering part relative to the target and finally the landing

Finally, experiment three presents the deployment of the system to approach a target with different shape and color. Figure 13 depicts the 3D trajectory followed by the aerial platform, while Fig. 14 depicts the path of the end-effector position versus the execution iterations of the experiment. In this case the object has been placed in another part of the flying arena and the main motion of the aerial vehicle was in the x axis. The plots depict the trajectory following and hovering parts of the manipulator guidance.

Those three experimental cases demonstrate the capabilities of the method, highlighting that the system can reach task-desired configurations. Table 2 summarizes the relative distance to the object as well as the Mean Absolute Error (MAE) for the real world experiments. The MAE values correspond to the hovering part in the relative position to the target and not the overall trajectory followed. The experimental trials show that the system is able to extract the depth with a substantial accuracy, while the other waypoints depend on the extracted bounding box.

The average time of execution from take-off till landing was about 2.5min, while the stereo module standalone takes around 1min. Nevertheless, experiment three demonstrates that the second object is more challenging to track and monitor, since it has smaller size inducing errors in the centroid extraction, which leads to greater deviations from the reference values.

Apart from the depth accuracy of the camera the bounding box selection from the object tracker is also critical for the centroid extraction. Figure 15 demonstrates a case where the extracted bounding box includes part of the background of the object on the right part, adding an offset on x axis in the centroid measurement. Overall, this system is able to guide the end-effector in close proximity with the target and can assist in the task of guidance as the initial step.

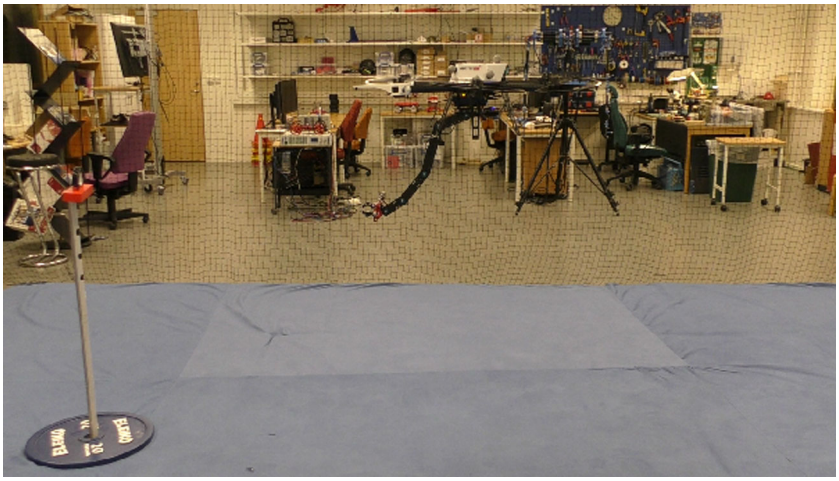


Fig. 13 Snapshot from the experimental trials of the stereo-based guidance, depicting the MAV hovering in front of the second object. (Video Link at:https://youtu.be/MObjUF1NI-8

5.3 Lessons Learned

Throughout the experimental trials many different experiences were gained that assisted in the development and tuning of the algorithms utilized. Based on this experiences, an overview of the lessons learned is provided including insights on the further developments in the field. This work, compared to the state of the art, tried to highlight the challenges of two major components that are critical for the guidance of the aerial manipulator, namely: 1) the object tracking and 2) the object localization. Overall, from a practical point of view, the aerial manipulator will be mainly utilized in cases that require interaction with the environment either with objects, surfaces or other generic regions of interest. In these cases the critical part is the sequential tracking of the object in multiple frames rather than the initial detection, since this role can be played by the operator. Moreover, once the object ha been identified in multiple frames it should be localized relative to the end-effector to generate the proper commands. Below are listed some challenges in different aspects of the end-effector guidance process.

Fast Tracking The ability to track the object in real time. In this work the utilized tracker was able to operate at the camera fps (20 fps) on an Intel NUC i7-5557U. The tracking speed depends on the application needs and there are other factors that can limit it except the tracking like the camera fps. This tracker is suitable and recommended for real time applications.

Generic Object Detection Ability to detect generic objects (without prior knowledge on geometry, shape, motion, color) depending on the application requirements. Section 5.1 presents experimental trials on the generic object tracking capabilities. The algorithm requires an

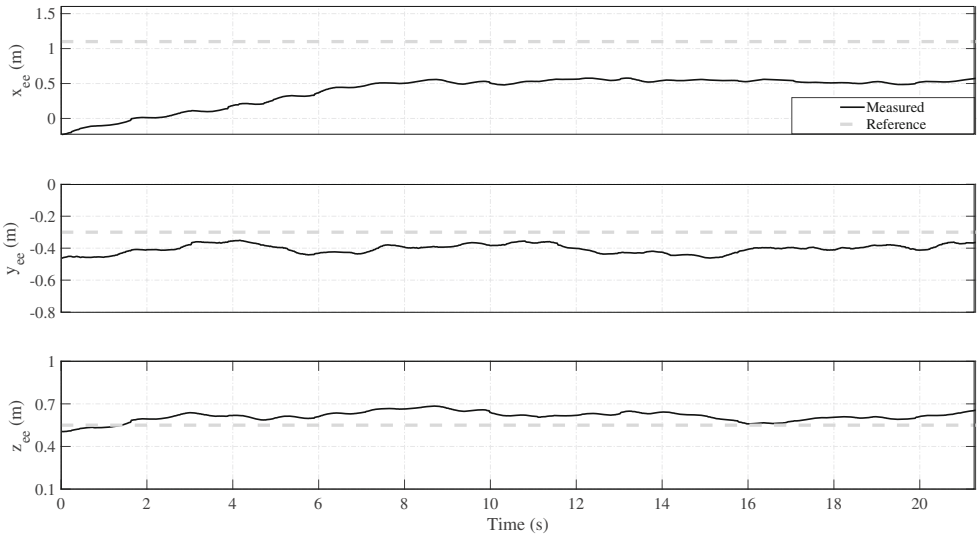


Fig. 14 End-effector setpoints vs the actual setpoints for the third experiment. The plots represent the waypoint following phase and the hovering part relative to the target

Table 2 Reference distance to the target and MAE

	Experiment 1	Experiment 2	Experiment 3
Distance to target	0.25m y-axis	0m y-axis	0.6m x-axis
Relative X MAE	0.098m	0.07m	0.02m
Relative Y MAE	0.025m	0.08m	0.08m
Relative Z MAE	0.07m	0.10m	0.10m

initial detection of the target, provided from an object detection algorithm or the operator and then is able to continue tracking it. The tracker shows substantial performance when tracking non identical and distinctive from their surroundings objects, regions/surfaces and is recommended for the respective application scenarios.

Object Re-Detection The ability to continue tracking the object after a loss event (target occlusion or target out-of-view) that often occur with abrupt motions. The current version of the tracker does not handle target re-detections, which is a major point for future work and improvements. Once the object is outside the field of view the tracking algorithm cannot recover.

Morphology Handling The ability to continue tracking the object when the morphology of the object changes due to different viewing angles/distances. The tracker is able to continue tracking the object up to an extent. There were cases where the MAV was flying around an object and the tracker was losing the object, while part of the object was still inside the field of view of the camera. The general experience gathered from the experimental trials showed that the tracker was able to handle 30-40% morphological angle distortions before losing track. On the other hand, the tracker shows substantial performance when varying relative distance to the target adapting the bounding box respectively. This tracker is recommended for cases when the guidance scenario aims to bring the end-effector close to the object without involving major angle distortions. Nevertheless, when the object is lost from the angle distortion the operator can re-initialize the tracking and continue the guidance.

Fig. 15 Pointcloud of the object having extracted the surrounding environment and the calculated centroid of the target depicted with the purple colored sphere

Complex Regions of Interest The ability to continue tracking the object of interest when the surrounding environment is complex and is difficult to distinguish them. Section 5.1 provides an example where the background and the object of interest have similar appearance and it is difficult for the tracker to operate without modifying the background. Case 3 of the object tracking is an example of the limitations and failure cases of the tracker. This tracker is not recommended in cases where the object is similar to its surroundings.

Depth Perception In realistic manipulation tasks, like cleaning tasks, it is imperative to have a dense and accurate estimation of the robot's workspace. In this work a custom made stereo camera has been employed as described in Section 4.2. The camera baseline was fixed at 10 cm. The stereo sensor was able to provide reasonable accuracy within a workspace of 2 m keeping the depth error with a mean value of 5 cm. Moreover, the camera intrinsic and extrinsic calibration is a fundamental process that affects its performance and should be repeated before every experiment. Overall, the performance of the specific hardware was substantial for experimental trials in the lab. Nevertheless, the depth perception plays an important role in the proposed guidance scheme and other alternatives could be also explored in future works to increase both accuracy and the range of the active workspace.

6 Conclusions

The aim of this article was to present a vision-based guidance system, structured around a robust object tracker, for aerial manipulation, while characterized by stereo processing for target monitoring tasks. The proposed system is considered the necessary tool to enable autonomous physical interaction tasks. Two different types of experiments have been presented to demonstrate the merits of the proposed method. Initially the object tracker has experimentally shown generic target tracking capabilities based on 3 different cases of objects. Additionally, the second experimental phase focused on the performance of the stereo-vision

guidance scheme. It should be stated that the system has been limited to approaching the target and not interacting with it, since during interaction, the MAV, the manipulator and the object are becoming a coupled system, that needs different overall control reconfiguration and it is considered as out of the scope for this article. Finally, lessons learned and limitations have been discussed, motivating future works in the field.

Acknowledgements This work has received funding from the European Unions Horizon 2020 Research and Innovation Programs under the Grant Agreements No.644128, AEROWORKS and No.730302, SIMS

Funding Information Open access funding provided by Lulea University of Technology.

References

1. Mansouri, S.S., Kanellakis, C., Fresk, E., Kominiak, D., Nikolakopoulos, G.: Cooperative uavs as a tool for aerial inspection of the aging infrastructure. In: Field and Service Robotics, pp. 177–189. Springer (2018)
2. Lee, A.C., Dahan, M., Weinert, A.J., Amin, S.: Leveraging suas for infrastructure network exploration and failure isolation. J. Intell. Robot. Syst., 1–29 (2018)
3. Yuan, C., Liu, Z., Zhang, Y.: Learning-based smoke detection for unmanned aerial vehicles applied to forest fire surveillance. Journal of Intelligent & Robotic Systems. https://doi.org/10.1007/s10846-018-0803-y (2018)
4. Sampedro, C., Rodriguez-Ramos, A., Bavle, H., Carrio, A., de la Puente, P., Campoy, P.: A fully-autonomous aerial robot for search and rescue applications in indoor environments using learning-based techniques. Journal of Intelligent & Robotic Systems. https://doi.org/10.1007/s10846-018-0898-1 (2018)
5. Kondak, K., Ollero, A., Maza, I., Krieger, K., Albu-Schaeffer, A., Schwarzbach, M., Laiacker, M.: Unmanned aerial systems physically interacting with the environment: Load transportation, deployment, and aerial manipulation. In: Handbook of Unmanned Aerial Vehicles, pp. 2755–2785. Springer (2015)
6. Lindsey, Q., Mellinger, D., Kumar, V.: Construction of cubic structures with quadrotor teams. In: Proc. Robotics, Science & Systems VII (2011)
7. Kim, S., Seo, H., Choi, S., Kim, H.J.: Vision-guided aerial manipulation using a multirotor with a robotic arm. IEEE/ASME Trans. Mechatron. **21**(4), 1912 (2016)
8. Seo, H., Kim, S., Kim, H.J.: Aerial grasping of cylindrical object using visual servoing based on stochastic model predictive control. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 6362–6368. IEEE (2017)
9. Santamaria-Navarro, A., Grosch, P., Lippiello, V., Sola, J., Andrade-Cetto, J.: Uncalibrated visual servo for unmanned aerial manipulation. IEEE/ASME Transactions on Mechatronics (2017)
10. Steich, K., Kamel, M., Beardsley, P., Obrist, M.K., Siegwart, R., Lachat, T.: Tree cavity inspection using aerial robots. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4856–4862. IEEE (2016)
11. Ramon Soria, P., Arrue, B.C., Ollero, A.: Detection, location and grasping objects using a stereo sensor on uav in outdoor environments. Sensors **17**(1), 103 (2017)
12. Lippiello, V., Cacace, J., Santamaria-Navarro, A., Andrade-Cetto, J., Trujillo, M.A., Esteves, Y.R., Viguria, A.: Hybrid visual servoing with hierarchical task composition for aerial manipulation. IEEE Robot. Autom. Lett. **1**(1), 259 (2016)
13. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2014.2345390 (2015)
14. Compact AeRial MAnipulator (CARMA) Assembly overwiew. https://www.youtube.com/watch?v=rOV5K43cpho
15. Wuthier, D., Kominiak, D., Kanellakis, C., Andrikopoulos, G., Fumagalli, M., Schipper, G., Nikolakopoulos, G.: On the design, modeling and control of a novel compact aerial manipulator. In: 2016 24th Mediterranean Conference on Control and Automation (MED), pp. 665–670. IEEE (2016)
16. Lynen, S., Achtelik, M., Weiss, S., Chli, M., Siegwart, R.: A robust and modular multi-sensor fusion approach applied to mav navigation. In: Proc. of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS) (2013)
17. Alexis, K., Nikolakopoulos, G., Tzes, A.: Switching model predictive attitude control for a quadrotor helicopter subject to atmospheric disturbances. Control. Eng. Pract. **19**(10), 1195 (2011)
18. Alexis, K., Nikolakopoulos, G., Tzes, A.: Model predictive quadrotor control: Attitude, altitude and position experimental studies. IET Control Theory Appl. **6**(12), 1812 (2012)
19. Kamel, M., Stastny, T., Alexis, K., Siegwart, R.: Model Predictive Control for Trajectory Tracking of Unmanned Aerial Vehicles Using Robot Operating System, pp. 3–39. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-54927-9_1
20. Piccinini, P., Prati, A., Cucchiara, R.: Real-time object detection and localization with sift-based clustering. Image Vis. Comput. **30**(8), 573 (2012)
21. Smeulders, A.W., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1442 (2014)
22. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328 (2008)
23. Zhan, Q., Liang, Y., Xiao, Y.: Color-based segmentation of point clouds. Laser Scan. **38**(3), 155 (2009)
24. Rusu, R.B., Cousins, S.: 3d is here: Point cloud library (pcl). In: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–4. IEEE (2011)
25. Oehler, B., Stueckler, J., Welle, J., Schulz, D., Behnke, S.: Efficient multi-resolution plane segmentation of 3d point clouds. Intell Robot Appl, 145–156 (2011)

Christoforos Kanellakis is currently pursuing his Ph.D degree within the Control Engineering Group, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology (LTU), Luleå, Sweden. He received his Diploma from the Electrical & Computer Engineering Department of the University of Patras (UPAT), Greece in 2015. He currently works in the field of robotics, focusing on the combination of control and vision to enable robots perceive and interact with the environment.

George Nikolakopoulos is Professor on Robotics and Automation at the Department of Computer Science, Electrical and Space Engineering at Luleå University of Technology, Luleå, Sweden. His work is focusing in the area of Robotics, Control Applications, while he has a significantly large experience in Creating and Managing European and National Research Projects. In the past he has been working as project manager and principal investigator in Several R&D&I projects funded by the EU, ESA, Swedish and the Greek National Ministry of Research. George is the coordinator of H2020-ICT AEROWORKS project in the field of Aerial Collaborative UAVs and H2020-SPIRE project DISIRE in the field of Integrated Process Control. In 2013 he has established the bigger outdoors motion capture system in Sweden and most probably in Europe as part of the FROST Field Robotics Lab at Luleå University of Technology. In 2014, he has been nominated as LTU's Wallenberg candidate, one out of three nominations from the university and 16 in total engineering nominations in Sweden. In year 2003, George has received the Information Societies Technologies (IST) Prize Award for the best paper that Promotes the scopes of the European IST (currently known as ICT) sector. His publications in the field of UAVs have received top recognition from the related scientific community, while have been several times listed in the TOP 25 most popular publications in Control Engineering Practice from Elsevier. In 2014 he has received the 2014 Premium Award for Best Paper in IET Control Theory and Applications, Elsevier for the research work in the area of UAVs, His published scientific work includes more than 150 published International Journals and Conferences in the fields of his interest.