2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Visual-Inertial Telepresence for Aerial Manipulation

Jongseok Lee[1], Ribin Balachandran[1], Yuri S. Sarkisov[1,2], Marco De Stefano[1], Andre Coelho[1]
Kashmira Shinde[1], Min Jun Kim[1], Rudolph Triebel[1,3] and Konstantin Kondak[1]

***Abstract*— This paper presents a novel telepresence system for enhancing aerial manipulation capabilities. It involves not only a haptic device, but also a virtual reality that provides a 3D visual feedback to a remotely-located teleoperator in real-time. We achieve this by utilizing onboard visual and inertial sensors, an object tracking algorithm and a pre-generated object database. As the virtual reality has to closely match the real remote scene, we propose an extension of a marker tracking algorithm with visual-inertial odometry. Both indoor and outdoor experiments show benefits of our proposed system in achieving advanced aerial manipulation tasks, namely grasping, placing, force exertion and peg-in-hole insertion.**

I. Introduction

Aerial manipulators exploit the manipulation capabilities of robotic arms located on a flying platform [1]. These systems can be deployed for tasks that are unsafe and costly for humans. Some notable examples are repairing rotor blades of wind turbines and inspecting oil and gas pipelines in refineries. However, building an autonomous aerial manipulator [2]–[4] poses several challenges to the current state-of-the-art robotic technologies. To this end, existing and close-to-market aerial manipulators are often tailored to a specific task such as contact inspection [5]–[7].

An alternative is the remote control of an aerial manipulator (namely, aerial tele-manipulation). Aerial tele-manipulation, by having a human-in-the-loop, has an advantage that several demands on robot's cognitive modules can be replaced by its teleoperator. Furthermore, recent studies show promising results that indicate a possibility for deployment of such systems under an imperfect communication between the robot and the operator. For example, bilateral teleoperation with force feedback has been demonstrated in Kontur-2 mission [8] where a cosmonaut from the International Space Station successfully operated a robot on Earth. In aerial tele-manipulation, the works on force feedback [9] and shared control [10] can be notably found.

Additionally, 3D visual feedback is an another important aspect of aerial tele-manipulation systems for enhancing their manipulation capabilities. During our field experiments with such platforms, we experienced that a 2D visual feedback solely based on the live video streams is not sufficient to achieve precise manipulation tasks. Thus, we deduced that aerial telepresence systems must involve both real-time force and 3D visual feedback, which accurately displays the interactions of the robotic arm with the objects. Several studies confirm that a virtual environment where one can change its sight-of-view and provide haptic guidance (e.g. virtual fixtures) improves the system capabilities [11]–[13].

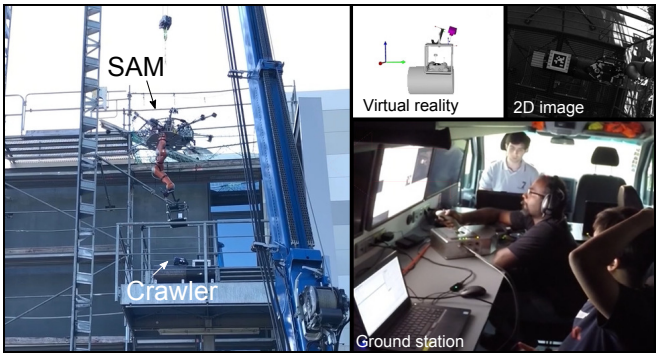

Fig. 1. An illustration of the proposed concept. Our aerial robot SAM [14] is designed to achieve a manipulation task in a remote location where humans find it difficult to reach (see left side of the figure). Consequently teleoperator from a ground station does not have any visual contact to the scene. Therefore, the robot's onboard perception system must provide a visual feedback to the operator with both 2D and 3D information which overall enhance its manipulation capabilities (depicted in the right side).

Therefore, we propose an advanced visual-inertial telepresence system, which utilizes visual and inertial sensors to provide 3D visual feedback to the operator. The resulting system is equipped with a haptic feedback and a virtual reality with virtual fixtures. In particular, for creating the 3D display of a remote scene, we consider an object localization approach where an object database and a marker tracking algorithm are used. As existing marker tracking methods did not suffice our requirements in terms of robustness and run-time, we propose a new object tracking algorithm by extending ARToolKitPlus [15] with onboard visual-inertial odometry (VIO). Lastly, an extension of the framework to multiple objects is also addressed for pick-and-place tasks.

The proposed concept is tightly integrated to a collision-safe aerial manipulator called cable-Suspended Aerial Manipulator (SAM [14]). In particular, the main scenario of interest is to deploy and retrieve an inspection robotic crawler (as illustrated in Fig. 1). This scenario, which was designed under the scope of EU project AEROARMS, is relevant to inspection and maintenance of gas and oil pipelines in refineries [16]. It involves grasping, placing and pressing tasks which need to be performed by a remotely located operator. The proposed algorithm is validated indoors and a peg-in-hole task with a margin of error less than 1cm is studied,

[1] Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Wessling, Germany. email: jongseok.lee@dlr.de
[2] Space CREI, Skolkovo Institute of Science and Technology (Skoltech), Moscow, Russia.
[3] Computer Vision Group, Technical University of Munich, Garching, Germany

which further displays SAM's advanced manipulation skills.

In summary, our main contributions are as follows.

- A visual-inertial telepresence system for aerial manipulation where a new object localization approach is proposed for creating virtual reality of the remote scene.
- An extend ARToolKitPlus [15] with onboard VIO for improving its run-time and robustness.
- Experimental validations showing advanced manipulation skills with SAM for the first time. In particular, our field experiments indicate overall system as a viable option for inspection and maintenance applications.

Experiments can be seen in the video: `https://www.youtube.com/watch?v=onOc05Ymxzs`.

A. Related Works

Several researchers aimed to provide 3D information of the remote scene for tele-manipulation. For this, 3D reconstruction techniques have been notably applied so far [17]–[21] where they aimed to create 3D visualization of an unknown environment. However, their applicability to our use-case is limited as the scene has to be mapped first, and then pre-processed for coping with the noisy 3D vision data. Unlike these methods, our approach differs as we use object localization algorithms. Two benefits are: (i) a real-time display is possible, and (ii) the framework can also be extended to a pick-and-place task, which requires the visualization of both the hand-held object and the target of placement. The later is difficult with the existing methods when the hand-held object is not rigidly fixed to a gripper. A recent work AeroVR [22] uses a similar concept to ours. While the system demonstrates an inspiring way to also include tactile feedback, the scope differs as AeroVR uses VICON system for indoor usage.

For object localization, learning-based [23]–[25] and geometry-based [15], [26] approaches can be found. Recent learning-based methods with deep neural networks can be broadly formulated with either explicit [23] or implicit [25] representations. However, we do not consider machine learning approaches as the assumption that the test data distribution to come from training distribution is routinely violated in the context of field robotics. Within the geometric methods, Fidicual marker systems (based on creating artificial features on the scene) are widely used in robotics for ground truths [26], applications where environments are known [27], simplifying the perception problem in lieu of sophistication [28] and calibration [29]. However, as we aim for creating the real-time virtual reality, our use-case provides stringent requirements on their limitations in run-time and inherent time-delays. Note that authors [30] show that coping with time delays in the display improves the performance of the tele-operation. Furthermore, as we use hand-eye cameras, our localization method should be robustness to loss-of-sight as the camera is not guaranteed to see the markers during the operations. Robustness is important when using haptic guidance or virtual fixtures for example, where inaccurate haptic feedback can cause negative effects in terms of the manipulation performance [31], [32].

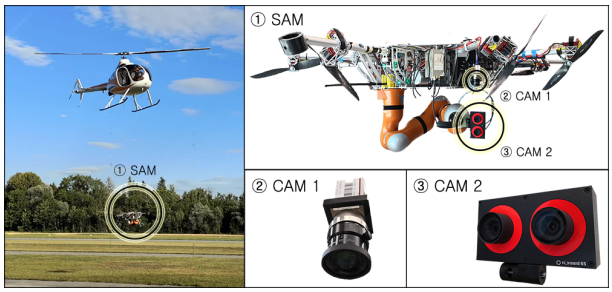

Fig. 2. Illustration of our collision-safe aerial manipulation concept; SAM with helicopter as an aerial carrier (left). Both hand-eye and eye-to-hand cameras are now integrated (right). We denote CAM1 as mako and CAM2 as hand-eye camera (hc for brevity).

II. Cable Suspended Aerial Manipulator

1) Robot hardware: An aerial manipulator SAM [14] is a complex flying robot composed of an aerial carrier, a cable-Suspended platform and a 7 degrees of freedom (DoF) industrial robotic arm KUKA LWR [33]. An aerial carrier (e.g. crane, manned/unmanned helicopter[1]) provides means to transport the robotic platform to a location (see Fig. 2). Then, a platform suspended to the carrier performs balancing act by autonomously damping out the disturbances induced by the carrier and the manipulator. This oscillation damping control is performed using eight omni-directional propellers and three winches as its actuators. Design and control aspects of SAM have been presented previously in [14].

2) Sensors choices and integration: Relevant sensors for realizing our vision-based telepresence system are as follows. KUKA LWR [33] is equipped with torque and position sensors as its *proprioceptive* sensors. Each joint contains a torque sensor, incremental and absolute position sensors which measure its joint torques and angles. Furthermore, SAM is equipped with optical devices as its *exteroceptive* sensors. As shown in Fig. 2, a monocular camera (Allied-vision Mako) is installed on the frame of the platform to display the overall operational space of the robotic arm. This is because the operator prefers eye-to-hand view which is more natural to a human. The camera provides high resolution images of 1292 by 964 px at 30Hz. Additionally, a stereo camera is integrated near the tool-center-point (tcp). Accuracy of the ficidual marker systems depends on the distance and its size which justifies the integration of a hand-eye camera [26]. We use a commercial 3D vision sensor that provides built-in VIO. Rcvisard provides 1280 by 960 px images at 25Hz and VIO estimates can be acquired at 200Hz. Details on VIO algorithm can be found in [34]

3) Haptic device: A portable and space-qualified haptic device, the Space Joystick [8] has been integrated to teleoperate the LWR located on SAM remotely.

[1]The purpose of the aerial carrier is to transport the system and hover. We use a crane in this study which also provides better safety, versatility, robustness and applicability for our considered application scenario.

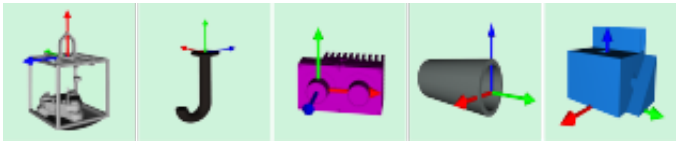

Fig. 3. An example of pre-generated object database.

Fig. 4. Failure modes of fidicual marker system in the field experiments. The figure shows a nominal case (left), and failure modes namely lost of sight and shadow occlusion (others).

III. Vision-Inertial Aerial Telepresence

A. 3D Visual Feedback with Object Localization

The aim is virtually displaying the robot and the objects so that an operator can tele-manipulate remotely. If done in real-time, the operator can *see* the virtual remote scene and perform the tasks. Here, accuracy is crucial as the virtual world has to closely match the real remote scene. In our approach, we realize such 3D visual feedback using cameras, object localization algorithms and known object database (see Fig. 3). Once objects to be actively manipulated are known a-priori, the essence of the problem simplifies to computing relative transformation of an objects with respect to the camera $\mathbf{T}^{\text{hc}}_{\text{object}}(t)$ and robot's tcp $\mathbf{T}^{\text{tcp}}_{\text{object}}(t)$. Here, t denotes time. A fixed transformation $\mathbf{T}^{\text{tcp}}_{\text{hc}}$ can be precisely estimated from CAD models or hand-eye calibration [35].

$$\mathbf{T}^{\text{tcp}}_{\text{object}}(t) = \mathbf{T}^{\text{tcp}}_{\text{hc}}\mathbf{T}^{\text{hc}}_{\text{object}}(t) \tag{1}$$

In this way, one can exploit object localization methods based on fiducial markers systems. These systems are widely adopted in robotics community and have been used as ground truths for its accuracy [26]. While learning-based pose estimation methods [25] can be leveraged under the same framework (for several applications where markers are not readily available), we limit our scope to validating the virtual reality concept in lieu of sophisticated object localization methods. Note that we use Instant Player [36] for creating the display as it supports various hierarchies of a scene graph. Using a nested hierarchy, relative transformation between an object and tools can be routed to display the scene, while a flat hierarchy can be used to extend the framework in order to display multiple objects and tools.

However, fiducial markers systems and their extensions [15], [26]–[28] have also significant drawbacks. It arises as we consider floating base manipulation outdoors. For example, shadows are inevitable for outdoor experiments and once it destroys certain shapes of the tags, the methods would naturally fail as its assumptions on the artificial visual features are violated. Similarly, the hand-eye camera (hc) can lose the view on the marker as the manipulator and the base can move rapidly. These failure modes (reported in Fig. 4) have consequences on the mission success rates. This is because it is difficult for the operator to remotely perform precise manipulation with live streams of 2D images. Eye-to-Hand views typically suffer from the occlusions of the grasping points by the robotic arm (also found in humanoid robots) and lacks depth information. Lastly, time delays that are inherent in these systems must be corrected in order to create a real-time virtual display of the scene.

For tackling these problems we propose Algorithm 1 for which multiple tags are placed on an object with a target tag x. The algorithm initializes by detecting all the tags (we denote multiART+ which is based on ArtoolKitPlus [15]), and saving their relative poses to the target (tag_init). While the process is running, k detected tags and their IDs are counted (counter_multiART+). If all the tags are detected, n+1 pose estimates of the target tag x can be computed by transforming pose estimates of non-target tags T^{hc}_{y} and their relative transformation to the target tag T^{y}_{x} (trafo3d). Then, RANSAC [37] is applied to these estimates to remove outliers, and then averaging to reduce variance (ransac_avg). Then, relative transformations are updated by applying RANSAC for the saved estimates, and averaging. In case atleast one tag is detected, the same step is applied to estimate the target tag x. These steps have advantages that (1) accuracy and orientation ambiguity of ArtoolKitPlus can be improved with RANSAC, and (2) the algorithm is robust to loss-of-sight of a target (similar to [27], [28], [38]).

However, the algorithm must be robust to loss-of-sight on all the tags, as we consider object tracking for floating base manipulators. Algorithm 1 addresses this problem by integrating VIO estimates of camera motion with respect to its inertial coordinate $\mathbf{T}^{\text{w}}_{\text{hc}}(t)$. If no tags are detected, (2) can be used to still estimate the target $\mathbf{T}^{\text{hc}}_{\text{x,avg}}(t)$ (vio_integrate). In (2), $\mathbf{T}^{\text{hc}}_{\text{w}}(t)\mathbf{T}^{\text{w}}_{\text{hc}}(t-1)$ is a relative transformation of camera motion from time t-1 to t and assumes a static object.

$$\mathbf{T}^{\text{hc}}_{\text{x,avg}}(t) = \mathbf{T}^{\text{hc}}_{\text{w}}(t)\mathbf{T}^{\text{w}}_{\text{hc}}(t-1)\mathbf{T}^{\text{hc}}_{\text{x,avg}}(t-1) \tag{2}$$

In a similar fashion, the delay of the system t_d can be computed (delay_computation) and corrected with VIO algorithm by using (3) (vio_delay_compensator). The herein delay is present in any perception system (e.g. rectifying an image) and fiducial marker systems (they are not real-time). In (3), $\mathbf{T}^{\text{w}}_{\text{hc}}(t)$ and $\mathbf{T}^{\text{hc}}_{\text{x,avg}}(t)$ are computed using VIO and multi-tag tracking. On the other hand, $\mathbf{T}^{\text{hc}}_{\text{w}}(t+t_d)$ can be computed using linear and angular velocity estimates of VIO, multiplied by the delay time t_d.

$$\mathbf{T}^{\text{hc}}_{\text{x,avg}}(t+t_d) = \mathbf{T}^{\text{hc}}_{\text{w}}(t+t_d)\mathbf{T}^{\text{w}}_{\text{hc}}(t)\mathbf{T}^{\text{hc}}_{\text{x,avg}}(t) \tag{3}$$

These two steps have several advantages. The algorithm is robust to failure modes of fidicual marker systems (see Fig. 4) as it copes with missing tag detection, and time delays are incorporated by using velocity signals and computed delay time. Furthermore, maximum run-time of the algorithm can be pushed to that of VIO data. The algorithm deals also with drifts of VIO estimates by using relative motion estimates

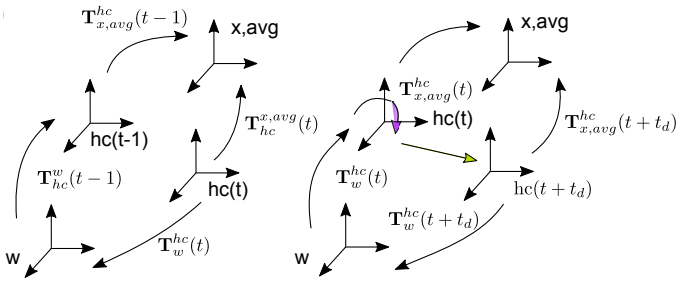

Fig. 5. Illustration of (2) and (3). Left: (2) uses VIO position and orientation estimates of camera motion to still estimate the object (denoted x,avg) when marker is not detected. Right: (3) uses linear (yellow arrow) and angular velocity (blue arrow), and computed time delay t_d to predict the motion of the camera in $t+t_d$ seconds.

only when the tag detection is lost. Note that the method is one way to use commodity vision sensors with VIO modules in order to further improve performance. Illustration of these two steps are found in Fig. 5.

B. Extension of 3D Visualization to Multiple Objects

For tasks such as placing, virtually displaying multiple objects and their relative pose is required. For example, if an operator would like to place a cage (with inspection robot inside) on a pipe which have roughly the same dimension, the virtual reality should reflect it by displaying the pipe, the cage, and the orientation changes of the cage with respect to TCP (e.g. a hook). With 3D reconstruction methods, this is difficult as one explores the environment for mapping and process the noisy data points for displaying. In our system, we tackle this challenge by using the hand-eye camera to estimate the orientation of the held object, while the eye-to-hand camera estimates the pose of other objects (e.g. a pipe). Then, the forward kinematics are leveraged as given below.

$$\mathbf{T}^{\text{object,1}}_{\text{object,2}}(t) = \mathbf{T}^{\text{object,1}}_{\text{mako}}(t)\mathbf{T}^{\text{mako}}_{\text{base}}\mathbf{T}^{\text{base}}_{\text{tcp}}(t)\mathbf{T}^{\text{tcp}}_{\text{hc}}\mathbf{T}^{\text{hc}}_{\text{object,2}}(t) \quad (4)$$

In (4), transformation from the base to eye-to-hand camera (mako) $\mathbf{T}^{\text{mako}}_{\text{base}}$ and tcp to hand-eye camera $\mathbf{T}^{\text{tcp}}_{\text{hc}}$ can be computed using hand-eye calibration [35]. $\mathbf{T}^{\text{object,1}}_{\text{mako}}$ is essentially updating the local base frame, and the forward kinematics of the robotic arm $\mathbf{T}^{\text{base}}_{\text{tcp}}$ is typically accurate. $\mathbf{T}^{\text{hc}}_{\text{object,2}}$ displays the pose of the held object. For this, one can use only multi-marker tracking without linear velocity integration. This is because markers can always made visible when the objects are held by the robot.

C. Force Feedback with Space Joystick and LWR

The controller design must ensure a stable bilateral tele-manipulation with force feedback. The main technical challenge is to deal with communication time delays, packet loss and jitters, which can cause instability of the system. For tackling this, a four channel architecture with time-domain passivity approach (proposed in [8]) has been used. A schematic of the system is shown in Fig. 5 and it is briefly explained as follows. The human operator sends both position (velocity analogously) and force signals from the master device (Space Joystick) to the slave (KUKA LWR

Algorithm 1: Robust marker localization

Input: Image I, target marker ID x, n multi marker IDs y and mapping to object $\mathbf{T}^{\text{x}}_{\text{object}}$.
Output: Pose of the object $\mathbf{T}^{\text{stereo}}_{\text{object}}$(t).
Algorithm:
$\mathbf{T}^{\text{hc}}_{\text{x}}(0)$, $\mathbf{T}^{\text{hc}}_{\text{y}_1}(0)$, ..., $\mathbf{T}^{\text{hc}}_{\text{y}_n}(0) \leftarrow$ multiART+(I);
$\mathbf{T}^{\text{y}_1}_{\text{x}}$, $\mathbf{T}^{\text{y}_2}_{\text{x}}$, ..., $\mathbf{T}^{\text{y}_n}_{\text{x}} \leftarrow$ tag_init($\mathbf{T}^{\text{hc}}_{\text{x}}(0)$,$\mathbf{T}^{\text{hc}}_{\text{y}_1}(0)$, ..., $\mathbf{T}^{\text{hc}}_{\text{y}_n}(0)$)
while object_localization == True **do**
  k, id $\leftarrow$ counter_multiART+(I);
  if k == n+1 **then**
    $\mathbf{T}^{\text{hc}}_{\text{x}}$(t), ..., $\mathbf{T}^{\text{hc}}_{\text{x,yn}}$(t) $\leftarrow$ trafo3d($\mathbf{T}^{\text{hc}}_{\text{x}}$(t), $\mathbf{T}^{\text{hc}}_{\text{y}}$(t), $\mathbf{T}^{\text{y}}_{\text{x}}$);
    $\mathbf{T}^{\text{hc}}_{\text{x,avg}}$(t) $\leftarrow$ ransac_avg($\mathbf{T}^{\text{hc}}_{\text{x}}$(t), ..., $\mathbf{T}^{\text{hc}}_{\text{x,yn}}$(t));
    $\mathbf{T}^{\text{y}_1}_{\text{x}}$, $\mathbf{T}^{\text{y}_2}_{\text{x}}$, ..., $\mathbf{T}^{\text{y}_n}_{\text{x}} \leftarrow$ tag_init_update($\mathbf{T}^{\text{y}}_{\text{x,pre}}$, $\mathbf{T}^{\text{y}}_{\text{x}}$);
  else if 0 < k < n+1 **then**
    if x $\in$ id == False **then**
      $\mathbf{T}^{\text{hc}}_{\text{x,y1}}$(t), ..., $\mathbf{T}^{\text{hc}}_{\text{x,yn}}$(t) $\leftarrow$ trafo3d($\mathbf{T}^{\text{hc}}_{\text{y}}$(t), $\mathbf{T}^{\text{y}}_{\text{x}}$);
      $\mathbf{T}^{\text{hc}}_{\text{x,avg}}$(t) $\leftarrow$ ransac_avg($\mathbf{T}^{\text{hc}}_{\text{x,y1}}$(t), ..., $\mathbf{T}^{\text{hc}}_{\text{x,yn}}$(t));
    else
      $\mathbf{T}^{\text{hc}}_{\text{x}}$(t), ..., $\mathbf{T}^{\text{hc}}_{\text{x,yn}}$(t) $\leftarrow$ trafo3d($\mathbf{T}^{\text{hc}}_{\text{y}}(t)$, $\mathbf{T}^{\text{y}}_{\text{x}}$);
      $\mathbf{T}^{\text{hc}}_{\text{x,avg}}$(t) $\leftarrow$ ransac_avg($\mathbf{T}^{\text{hc}}_{\text{x}}$(t), ..., $\mathbf{T}^{\text{hc}}_{\text{x,yn}}$(t));
    end
  else
    $\mathbf{T}^{\text{hc}}_{\text{x,avg}}$(t) $\leftarrow$ Eq. (2);
  end
  $t_d \leftarrow$ delay_computation()
  $\mathbf{T}^{\text{hc}}_{\text{x,avg}}(t+t_d) \leftarrow$ Eq. (3)
  $\mathbf{T}^{\text{hc}}_{\text{object}}(t) = \mathbf{T}^{\text{hc}}_{\text{x,avg}}(t+t_d)\mathbf{T}^{\text{x}}_{\text{object}}$
end

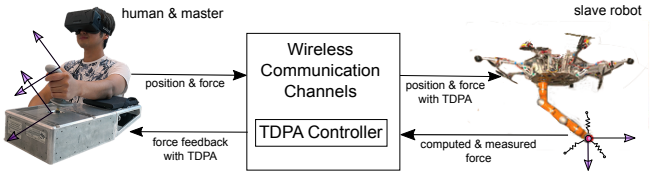

Fig. 6. Controller overview. Communication time delays, packet loss and jitter can cause instability of the overall system. For coping with this, TDPA is used for force feedback tele-manipulation.

mounted on the SAM). As these signals pass through communication channels (in the considered scenario, a wireless communication), they will get affected by time delay. To ensure stable tele-manipulation, we employ time domain passivity approach (TDPA [39]). Readers can refer to [8] for more details and implementations.

D. Haptic Guidance with Virtual Fixtures

On top of real-time virtual reality and haptic device, another aspect of our telepresence system is haptic guidance via virtual fixtures [12]. In this work, the virtual fixtures are implemented as artificial walls that guide the motion of the slave to the desired target point. If the teleoperator tries to move the slave device outside these walls, artificial forces are activated to limit the motion of tcp (slave) and also to provide haptic feedback to the teleoperator. The virtual fixtures in

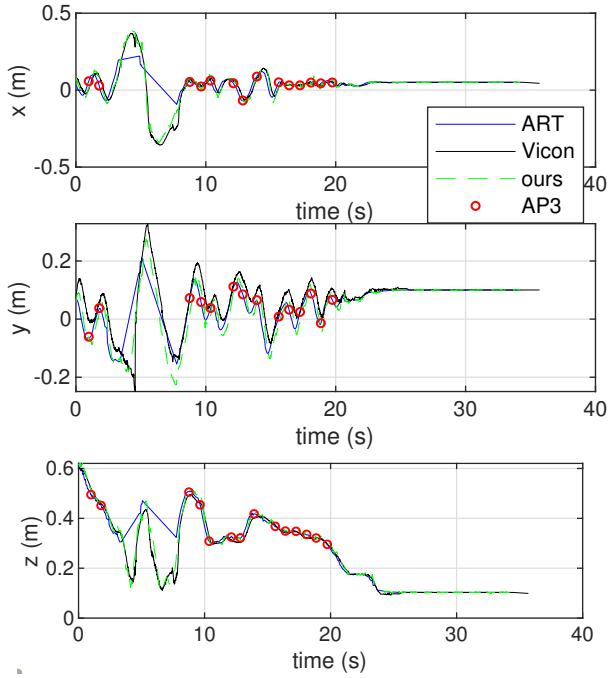

Fig. 7. Our proposed algorithm 1 for object tracking (denoted ours) is compared to ground truth (Vicon measurements). The algorithm is compared with two other popular fidicual detection frameworks namely AprilTag 2 (AP2) and ARToolKitPlus (multiART). Our proposed algorithm is robust to losing the fidicuals in an image, and compensates the delay.

this work are based on Voxmap-PointShell algorithm [36], [40] and more details on their implementation and parameter tuning can be found in [41].

IV. Experiments and Results

A. Robust Object Localization: Validation and Analysis

An object localization approach is taken for 3D visualization and thus, accuracy, run-time and robustness of the proposed algorithm is reported. These results are important as the created virtual reality should closely match the real remote scene. For this, we measure the ground truth of the relative poses between the object and the camera using Vicon tracking system and evaluate the performance on sequences that represent peg-in-hole insertion task (see video attachment). The algorithm is also compared to Apriltag2 [26] (AP2) and ARToolKitPlus [15] without (2) and (3) (multiART). In Fig. 7, estimated trajectories of relative poses are compared with Vicon measurements. As depicted, our proposed algorithm is robust against loss-of-sight problems of object localization with a hand-eye camera while AP2 and multiART produce jumps as no markers are detected (t=3s to t=8s as an example). This is due to the design of the algorithm where we utilize VIO estimates of the camera pose when the marker is not detected. Furthermore, multiART suffers from time delay, while AP2 has both the time delay, and slow run-time. On the other hand, our proposed algorithm compensates the time delay, resulting in accurate estimates. Five experiments have been conducted

TABLE I

Accuracy and run-time analysis

	AP2	multiART+	ours
$e_{x,rmse}$ [m]	0.1690	0.1124	**0.0252**
$e_{y,rmse}$ [m]	0.1265	0.0847	**0.0503**
$e_{z,rmse}$ [m]	0.1308	0.077	**0.0316**
$e_{\phi,rmse}$ [rad]	0.2843	0.1867	**0.1232**
$e_{\theta,rmse}$ [rad]	0.1955	0.1232	**0.0703**
$e_{\psi,rmse}$ [rad]	0.2565	0.1755	**0.1153**
t_{run} [s]	0.839 $\pm$ 0.0616	0.0525 $\pm$ 0.0218	**0.0049 $\pm$ 0.013**

to determine the accuracy of the selected methods with respect to the ground truth. Note that the trajectory selected includes loss-of-sight and time delay. The corresponding root mean squared errors (RMSE) have been reported in Table I. However, as seen in Table I, AP2 is slow while using high-resolution images, and this results in more errors as we compare the trajectories. In our approach, these trajectories are relevant as we aim for creating virtual reality with object localization methods. Within our experiments, the analysis of the accuracy, robustness and run-time further justifies the proposed algorithm and its additional complexity.

B. Peg-in-Hole Insertion with Virtual Fixtures

A peg-in-hole insertion task with margins of error less than 1cm is considered in which operator does not have any direct visual contact to the real scene. The main challenge in this setting is on the fidelity of virtual reality and resulting virtual fixtures. With the fidelity provided by our proposed algorithm and resulting virtual fixtures, a peg-in-hole task has been performed (see the attached video material). The results are depicted in Fig. 9 and Fig. 10. Fig. 9 plots force signals acting on the slave end-effector which constitutes computed force from master's position commands, and force due to the virtual fixtures. Position tracking of tcp towards the target (hole) is shown in Fig. 10. As these position signals are expressed in LWR base frame (see Fig. 6 for definition), the target also moves due to the motion of SAM. This experiment shows the benefits of our proposed telepresence system, as SAM is able to perform a precise manipulation task. Note that the accuracy of object localization improves over reported values in Table I when the peg is near the hole (shown in Fig. 7) which makes the task feasible.

C. Field Experiments and Validation

A field experiment is conducted in order to demonstrate the applicability of SAM within a relevant industrial scenario for aerial manipulation. This scenario involves a maintenance and inspection task in which SAM has to deploy and retrieve a 6.4kg inspection robot to a remotely located pipe. To transport the inspection robot, a cage (approximately of the same size as the pipe and the inspection robot) has been designed. For this mission, SAM has to (a) grasp the cage with a hook at location A with a hook used as end-effector for the LWR, (b) move to location B where the pipe is located, (c) place the cage on the pipe, and (d) press the cage while the inspection robot moves out. The teleoperator is located in a ground station and thus, has no direct visual contact to the

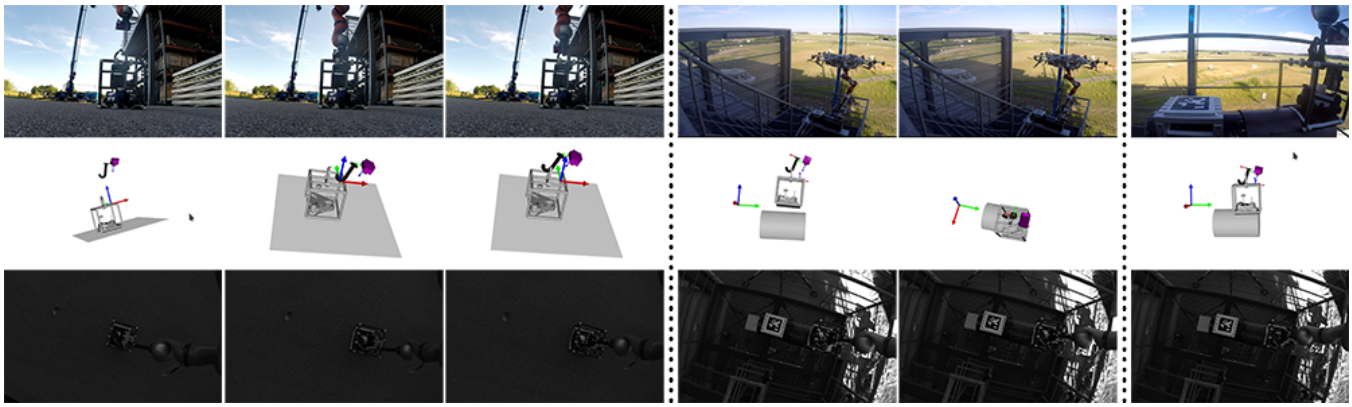

Fig. 8. Results of field experiments for AEROARMS [16] industrial scenario. SAM successfully deployed and retrieved a pipe inspection robot by performing grasping (left), placing (middle) and pressing (right). As we consider outdoor manipulation tasks with an industry relevancy, the system has to both address force feedback, and 3D visual feedback. 2D visual feedback (bottom row), as depicted above, is not sufficient as the depth information is missing and subject to under exposure. On the other hand, the virtual environment (middle row) does not suffer from these problems, and the operator can zoom-in and out, and change its sight-of-view. These experiments show SAM with telepresence as a viable option for future applications.

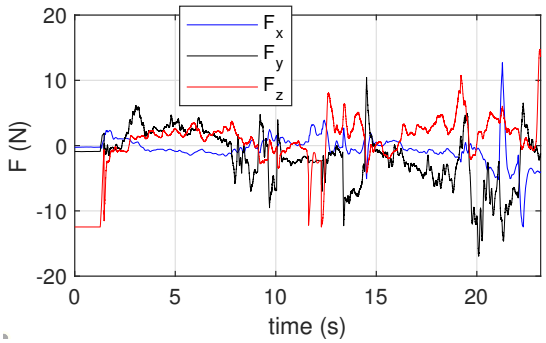

Fig. 9. Force signals on slave's end-effector expressed in LWR base frame. These forces compose of artificial force from a virtual fixture, and computed forces from master's commanded positions.

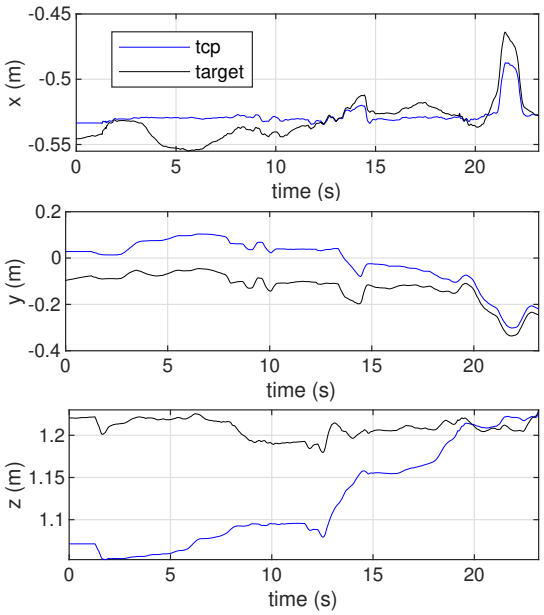

Fig. 10. TCP and target positions expressed in LWR base frame. For peg-in-hole insertion, tcp is commanded to follow the target. Note that the target position changes as SAM moves, and it is expressed i in LWR base frame.

scene. For this scenario, we tackle precision grasping, placing and pressing tele-manipulation tasks at a remote location, and the results are depicted in Fig. 8. In particular, 2D images alone do not show the depth information (placing task) and are often occluded (grasping and pressing phases). With only force feedback, a precise manipulation is difficult for this scenario. On the other hand, the virtual reality provides 3D information of the remote scene, and moreover, one can change the sight-of-view to avoid an occluded visual feedback. These results show the benefits of our telepresence system. By touching and seeing, the teleoperator is able to perform precise manipulation tasks for an industrial use-case.

The field experiments for AEROARMS industrial scenario did not use the haptic guidance using virtual fixtures and VIO compensations for achieving the basic teleoperation tasks. For further improving the inspection and maintenance scenario, we plan to perform a user-study to investigate the degree of improvements with this shared autonomy concept and further joint demonstration with recent developments on SAM [42], [43]. Lastly, robotic introspection [44] for object localization is another research direction that can support in industrial deployments of these systems.

V. Conclusion

This paper presents a vision-inertial telepresence concept in which onboard sensors, an object tracking algorithm and databases of objects were utilized to provide a 3D visualization of the scene in real-time. From our experiences in the field, we believe that providing a 3D visual feedback to the tele-operator is required in aerial manipulation applications at remote sites where a direct and close visual contact to the objects are genuinely difficult. Our demonstration of advanced aerial manipulation shows that SAM with telepresence is a viable concept for inspection and maintenance applications.

VI. Acknowledgements

Special thanks to Michael Vilzmann for the support on FCC, Thomas Hulin for the support on peg-in-hole experiments and Nari Song for the support on video editing.

REFERENCES

[1] F. Ruggiero, V. Lippiello, and A. Ollero, "Aerial manipulation: A literature review," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1957–1964, July 2018.

[2] K. Kondak, F. Huber, M. Schwarzbach, M. Laiacker, D. Sommer, M. Bejar, and A. Ollero, "Aerial manipulation robot composed of an autonomous helicopter and a 7 degrees of freedom industrial manipulator," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 2107–2112.

[3] F. Ruggiero, M. A. Trujillo, R. Cano, H. Ascorbe, A. Viguria, C. Perz, V. Lippiello, A. Ollero, and B. Siciliano, "A multilayer control for multirotor uavs equipped with a servo robot arm," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 4014–4020.

[4] M. J. Kim, R. Balachandran, M. De Stefano, K. Kondak, and C. Ott, "Passive compliance control of aerial manipulators," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 4177–4184.

[5] K. Bodie, M. Brunner, M. Pantic, S. Walser, P. Pfndler, U. Angst, R. Siegwart, and J. Nieto, "An omnidirectional aerial manipulation platform for contact-based inspection," in *Proceedings of Robotics: Science and Systems (RSS'19)*, Freiburg im Breisgau, Germany, Jun. 22-26 2019.

[6] M. A. Trujillo, J. R. Martínez-de Dios, C. Martín, A. Viguria, and A. Ollero, "Novel aerial manipulator for accurate and robust industrial ndt contact inspection: A new tool for the oil and gas inspection industry," *Sensors*, vol. 19, no. 6, 2019.

[7] P. J. Sanchez-Cuevas, P. Ramon-Soria, B. Arrue, A. Ollero, and G. Heredia, "Robotic system for inspection by contact of bridge beams using uavs," *Sensors*, vol. 19, no. 2, 2019.

[8] J. Artigas, R. Balachandran, C. Riecke, M. Stelzer, B. Weber, J. Ryu, and A. Albu-Schaeffer, "Kontur-2: Force-feedback teleoperation from the international space station," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 1166–1173.

[9] M. Mohammadi, A. Franchi, D. Barcelli, and D. Prattichizzo, "Cooperative aerial tele-manipulation with haptic feedback," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 5092–5098.

[10] A. Franchi, C. Secchi, M. Ryll, H. H. Bulthoff, and P. R. Giordano, "Shared control : Balancing autonomy and human assistance with a group of quadrotor uavs," *IEEE Robotics Automation Magazine*, vol. 19, no. 3, pp. 57–68, Sep. 2012.

[11] V. Falk, D. H. Mintz, J. G. Grunenfelder, J. I. Fann, and T. Burdon, "Influence of three-dimensional vision on surgical telemanipulator performance," *Surgical Endoscopy*, vol. 15, pp. 1282–1288, 2001.

[12] A. Bettini, P. Marayong, S. Lang, A. M. Okamura, and G. D. Hager, "Vision-assisted control for manipulation using virtual fixtures," *IEEE Transactions on Robotics*, vol. 20, no. 6, pp. 953–966, Dec 2004.

[13] K. Huang, D. Chitrakar, F. Rydén, and H. J. Chizeck, "Evaluation of haptic guidance virtual fixtures and 3D visualization methods in telemanipulation—a user study," *Intelligent Service Robotics*, Jul 2019.

[14] Y. S. Sarkisov, M. J. Kim, D. Bicego, D. Tsetserukou, C. Ott, A. Franchi, and K. Kondak, "Development of sam: Cable-suspended aerial manipulator," in *2019 IEEE International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 5323–5329.

[15] D. Wagner and D. Schmalstieg, "Artoolkitplus for pose tracking on mobile devices," in *Proceedings of 12th Computer Vision Winter Workshop (CVWW07)*, 01 2007, p. 139146.

[16] A. Ollero, G. Heredia, A. Franchi, G. Antonelli, K. Kondak, A. Sanfeliu, A. Viguria, J. R. Martinez-de Dios, F. Pierri, J. Cortes, A. Santamaria-Navarro, M. A. Trujillo Soto, R. Balachandran, J. Andrade-Cetto, and A. Rodriguez, "The aeroarms project: Aerial robots with advanced manipulation capabilities for inspection and maintenance," *IEEE Robotics Automation Magazine*, vol. 25, no. 4, pp. 12–23, Dec 2018.

[17] G. Hirzinger, M. Fischer, B. Brunner, R. Koeppe, M. Otter, M. Grebenstein, and I. Schäfer, "Advances in Robotics: The DLR Experience," *The International Journal of Robotics Research*, vol. 18, no. 11, pp. 1064–1087, 1999.

[18] D. Ni, A. Song, X. Xu, H. Li, C. Zhu, and H. Zeng, "3d-point-cloud registration and real-world dynamic modelling-based virtual environment building method for teleoperation," *Robotica*, vol. 35, no. 10, p. 19581974, 2017.

[19] D. Ni, A. Nee, S. Ong, H. Li, C. Zhu, and A. Song, "Point cloud augmented virtual reality environment with haptic constraints for teleoperation," *Transactions of the Institute of Measurement and Control*, vol. 40, no. 15, pp. 4091–4104, 2018.

[20] A. Leeper, S. Chan, and K. Salisbury, "Point clouds can be represented as implicit surfaces for constraint-based haptic rendering," in *2012 IEEE International Conference on Robotics and Automation*, May 2012, pp. 5000–5005.

[21] F. Rydn and H. J. Chizeck, "A method for constraint-based six degree-of-freedom haptic interaction with streaming point clouds," in *2013 IEEE International Conference on Robotics and Automation*, May 2013, pp. 2353–2359.

[22] G. A. Yashin, D. Trinitatova, R. T. Agishev, R. Ibrahimov, and D. Tsetserukou, "Aerovr: Virtual reality-based teleoperation with tactile feedback for aerial manipulation," in *IEEE International Conference on Advanced Robotics (ICAR)*, 2019.

[23] *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017.

[24] V. Lepetit, "Learning descriptors for object recognition and 3d pose estimation," in *Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–10.

[25] M. Sundermeyer, Z. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3d orientation learning for 6d object detection from RGB images," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, 2018, pp. 712–729.

[26] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, oct 2016, pp. 4193–4198.

[27] D. Malyuta, C. Brommer, D. Hentzen, T. Stastny, R. Siegwart, and R. Brockers, "Long-duration fully autonomous operation of rotorcraft unmanned aerial systems for remote-sensing data acquisition," *Journal of Field Robotics*, vol. 37, no. 1, pp. 137–157, 2020.

[28] M. Laiacker, F. Huber, and K. Kondak, "High accuracy visual servoing for aerial manipulation using a 7 degrees of freedom industrial manipulator," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 1631–1636.

[29] C. Nissler, M. Durner, Z.-C. Mrton, and R. Triebel, "Simultaneous calibration and mapping," in *International Symposium on Experimental Robotics (ISER)*, Buenos Aires, Argentina, Nov. 2018.

[30] F. Richter, Y. Zhang, Y. Zhi, R. K. Orosco, and M. C. Yip, "Augmented reality predictive displays to help mitigate the effects of delayed telesurgery," in *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, 2019, pp. 444–450.

[31] J. v. Oosterhout, J. G. W. Wildenbeest, H. Boessenkool, C. J. M. Heemskerk, M. R. d. Baar, F. C. T. v. d. Helm, and D. A. Abbink, "Haptic shared control in tele-manipulation: Effects of inaccuracies in guidance on task execution," *IEEE Transactions on Haptics*, vol. 8, no. 2, pp. 164–175, April 2015.

[32] H. Boessenkool, D. A. Abbink, C. J. M. Heemskerk, and F. C. T. van der Helm, "Haptic shared control improves tele-operated task performance towards performance in direct control," in *2011 IEEE World Haptics Conference*, June 2011, pp. 433–438.

[33] R. Bischoff, J. Kurth, G. Schreiber, R. Koeppe, A. Albu-Schaeffer, A. Beyer, O. Eiberger, S. Haddadin, A. Stemmer, G. Grunwald, and G. Hirzinger, "The kuka-dlr lightweight robot arm - a new reference platform for robotics research and manufacturing," in *ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*, June 2010, pp. 1–8.

[34] "Roboception rcvisard user manual," available at https://doc.rc-visard.com/latest/en/index.html.

[35] K. H. Strobl and G. Hirzinger, "Optimal hand-eye calibration," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2006, pp. 4647–4653.

[36] H. Thomas, H. Katharina, P. Carsten, S. Chun-Yi, R. Subhash, and L. Honghai, "Interactive features for robot viewers," in *Intelligent Robotics and Applications*, vol. 7508, ICIRA 2012.

[37] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[38] C. Nissler, S. Büttner, Z. Marton, L. Beckmann, and U. Thomasy, "Evaluation and improvement of global pose estimation with multiple apriltags for industrial manipulators," in *2016 IEEE 21st Interna-*

tional Conference on Emerging Technologies and Factory Automation (ETFA), Sep. 2016, pp. 1–8.

[39] B. Hannaford and J.-H. Ryu, "Time domain passivity control of haptic interfaces," *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*, vol. 2, pp. 1863–1869 vol.2, 2001.

[40] M. Sagardia and T. Hulin, "Multimodal evaluation of the differences between real and virtual assemblies," *IEEE Transactions on Haptics*, vol. 11, no. 1, pp. 107–118, Jan 2018.

[41] T. W. Martins, A. Pereira, T. Hulin, O. Ruf, S. Kugler, A. Giordano, R. Balachandran, F. Benedikt, J. Lewis, R. Anderl, K. Schilling, and A. Albu-Schäffer, "Space factory 4.0 - new processes for the robotic assembly of modular satellites on an in-orbit platform based on industrie 4.0 approach," in *69th International Astronautical Congress (IAC)*, October 2018.

[42] Y. S. Sarkisov, M. J. Kim, A. Coelho, D. Tsetserukou, C. Ott, and K. Kondak, "Optimal oscillation damping control of cable-suspended aerial manipulator with a single imu sensor," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, accepted, available online.

[43] A. Coelho, H. Singh, C. Ott, and K. Kondak, "Whole-body bilateral teleoperation of a redundant aerial manipulator," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, accepted, available online.

[44] H. Grimmett, R. Paul, R. Triebel, and I. Posner, "Knowing when we don't know: Introspective classification for mission-critical decision making," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 4531–4538.