IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 March 2023, accepted 3 May 2023, date of publication 9 May 2023, date of current version 15 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274597

RESEARCH ARTICLE

Development and Experimental Validation of Visual-Inertial Navigation for Auto-Docking of Unmanned Surface Vehicles

ØYSTEIN VOLDEN, ANNETTE STAHL, AND THOR I. FOSSEN, (Fellow, IEEE)
Department of Engineering Cybernetics, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Corresponding author: Øystein Volden (oystein.volden@ntnu.no)

This work was funded by the Norwegian Research Council (project no. 223254) through the NTNU Center of Autonomous Marine Operations and Systems (AMOS) at the Norwegian University of Science and Technology.

ABSTRACT Docking is a safety-critical operation for autonomous surface vehicles and requires highly accurate navigation signals. Since Global Navigation Satellite Systems (GNSS) can be unreliable and inaccurate in urban environments, other sensors should be considered for increased redundancy and reliability. To this end, we present a low-cost visual-inertial navigation system that we can use for automatic docking of small vehicles. The proposed system produces state estimates of the vehicle, including position, velocity, and attitude, based on raw image and inertial data. To simplify the navigation task, we use easily identifiable tags as a reference on the dockside. When the vehicle approaches the dock, a visual fiducial system recognizes the tags and estimates the relative pose between the camera onboard the vehicle and the tags at the dockside. The camera-tag pose and inertial data are then fused using an error-state Kalman filter for robust state estimation of the vehicle. For benchmarking, we use an unmanned surface vehicle equipped with a dual-antenna real-time kinematic GNSS receiver for accurate positioning and heading. We show that the proposed method performs well on regular and adverse weather data. Finally, we demonstrate that the proposed method performs well in feedback control through field experiments and can supplement traditional navigation systems for docking operations.

INDEX TERMS Autonomous docking, fiducial tags, unmanned surface vehicles, visual-inertial state estimation.

I. INTRODUCTION

Lately, the maritime industry has embraced autonomy for its cost-effectiveness and safety [1]. In the years to come, autonomous vehicles are expected to advance and play an important role in industries such as shipping, public transportation, and remote surveillance that are currently undergoing extensive digital transformations [2]. Nevertheless, several challenges remain before fully autonomous vehicles are ready for the commercial market. In particular, autonomous vehicles must provide highly resilient navigation systems to operate well at all times, including in safety-critical operations. This is particularly important for widespread acceptance among authorities, classification societies, and the general public, thus advancing commercial autonomy in the maritime sector [3].

The docking of a ship involves low-speed maneuvers in constrained urban environments and requires high-precision navigation signals to operate reliably. In this context, Det Norske Veritas, an international ship classification company, requires autonomous ships to obtain 0.1 m absolute position accuracy with 95% probability for automatic docking operations [4]. Unfortunately, commercial GNSS outputs positioning errors in the orders of meters [5], and Differential GNSS typically gives 1 m global accuracy [6]. Real-time kinematic (RTK) GNSS can be used to calculate position with centimeter-level accuracy. However, RTK GNSS is an expensive solution. Moreover, RTK GNSS is limited by the

The associate editor coordinating the review of this manuscript and approving it for publication was Li Zhang.

radio range and is minimally tolerant to datalink dropouts [7]. This motivates the use of alternative sensors to supplement the satellite-based navigation system.

Researchers are increasingly focusing on visual-based localization systems because they are low-cost, more robust, and more reliable than other sensor-based localization systems [5]. In this context, there are mainly two approaches for estimating the camera pose: The first is based on natural features [8], [9], [10], for example, key points and textures, while the second is based on artificial landmarks [11], [12], [13]. The first approach requires no intervention in the operating environment, thus proving to be a flexible choice when exploring unknown environments. However, the performance typically degrades in textureless environments or because of motion blur. The second approach with artificial landmarks does not suffer from these drawbacks and is therefore preferred if accuracy, robustness, and speed are essential. [14]. In robotic navigation, artificial tags such as the ARToolkit [15], ARTags [16], ArUco [13], AprilTag [12], and AprilTag2 [17] are most frequently used to obtain accurate camera-tag pose estimates. A key advantage of artificial tags is that the estimated camera-tag pose coming from the visual fiducial system does not drift. Landmark-based localization can therefore work as a full-fledged absolute positioning alternative to GNSS if the target point is known in a global frame.

Unfortunately, the visual measurements can be inaccurate or completely absent, for example, due to degraded perception level or large camera-tag distance. Moreover, visual-based localization systems require demanding image processing, usually performed on an embedded device with limited computational power. This typically leads to a lower measurement frequency than required for USVs in closed-loop control. We would therefore like to maintain an estimate of the pose of the vehicle by integrating the acceleration, and angular rate measurements, before the drift errors induced by the inertial measurement unit (IMU) are corrected with new visual measurements. High-rate IMU measurements and low-rate camera measurements are usually fused using a Kalman filter [18], similar to standard GNSS + inertial navigation system (INS) solutions [19]. To capture the nonlinearities in the process and measurement model, nonlinear formulations of the Kalman filter, e.g., the *Extended Kalman filter* (EKF) and the *unscented Kalman filter* (UKF) [20], are often used. While the UKF has advantages for highly nonlinear systems, it is computationally expensive compared to the EKF [21]. Therefore, the EKF is considered to be the workhorse for real-time state estimation applied to navigation systems. The accuracy of the EKF is, however, highly dependent on how well the nonlinearities of the models can be captured by linearizations about operating points. For example, marine surface vehicles may exhibit strong nonlinear behavior. As such, local linearization of the states in the EKF will not represent a sufficiently accurate approximation, potentially causing filter divergence. Therefore, the *error-state Kalman filter* (ESKF) was developed to improve the linear approximation. By estimating the error state instead of the true state, the ESKF allows for better linearization, where higher-order products become negligible since the error state tends to be linear [22].

For marine applications such as ship maneuvering, robust estimation of the attitude is of high importance. For example, a standard ESKF with attitude parametrized using Euler angles is not preferred because of singular points. Instead, the four-component quaternion, which has the lowest dimensionality possible for a singularity-free attitude representation, is favored. However, it has one superfluous degree of freedom. Thus we face the dilemma of using an attitude representation that is either singular or redundant. To evade this dilemma, we use the multiplicative extended Kalman filter (MEKF) formulation: An error-state EKF where attitude is parametrized using a four-dimensional unit quaternion [23]. However, the unit quaternion error is parametrized using a three-parameter attitude representation. This is beneficial since the three-parameter representation avoids singularities due to small attitude errors, and it represents the attitude with a minimal number of degrees [24]. Additionally, MEKF can handle biases in the sensors, which is important in attitude estimation applications [25, p.471]. After estimating the error state, it is injected into the nominal state, thus predicting the true state using a four-dimensional unit quaternion. Finally, a reset strategy is used in which the error-state vector is set to zero to prevent the state estimates from growing to large values for long-endurance applications [25, p.472].

To cope with the raised concerns, we extend the error-state attitude filter, i.e., the MEKF, to a complete navigation solution, which also includes the translational motions (position, linear velocity, specific force biases). As such, the proposed filter is able to estimate the full state of the vehicle with sufficient linearization properties and avoid gimbal lock situations due to a nonsingular attitude representation. By fusing drift-free, visual measurements with high-quality inertial data, we obtain robust and accurate state estimates in a local area close to the dockside at a high frequency. Hence, we contribute to the development of an independent, GNSS-free navigation system to increase navigation accuracy and redundancy in safety-critical maritime operations, which are rarely demonstrated in practice. For experimental validation, we use the Norwegian University of Science and Technology (NTNU) Otter USV, shown in Fig. 1.

A. RELATED WORK

One of the fundamental challenges for fully autonomous vehicle systems is to develop robust navigation systems that precisely localize the vehicle in its environment. Concerning vision-based localization systems, a lot of work has focused on feature-based methods using VO or simultaneous localization and mapping systems. While working well for certain indoor robotic applications, the performance of these techniques is usually poor in outdoor environments due to textureless areas (e.g., the sea surface) and challenging lighting conditions [26]. In contrast, artificial tags have shown advantages over feature-based methods due to accurate, robust, and

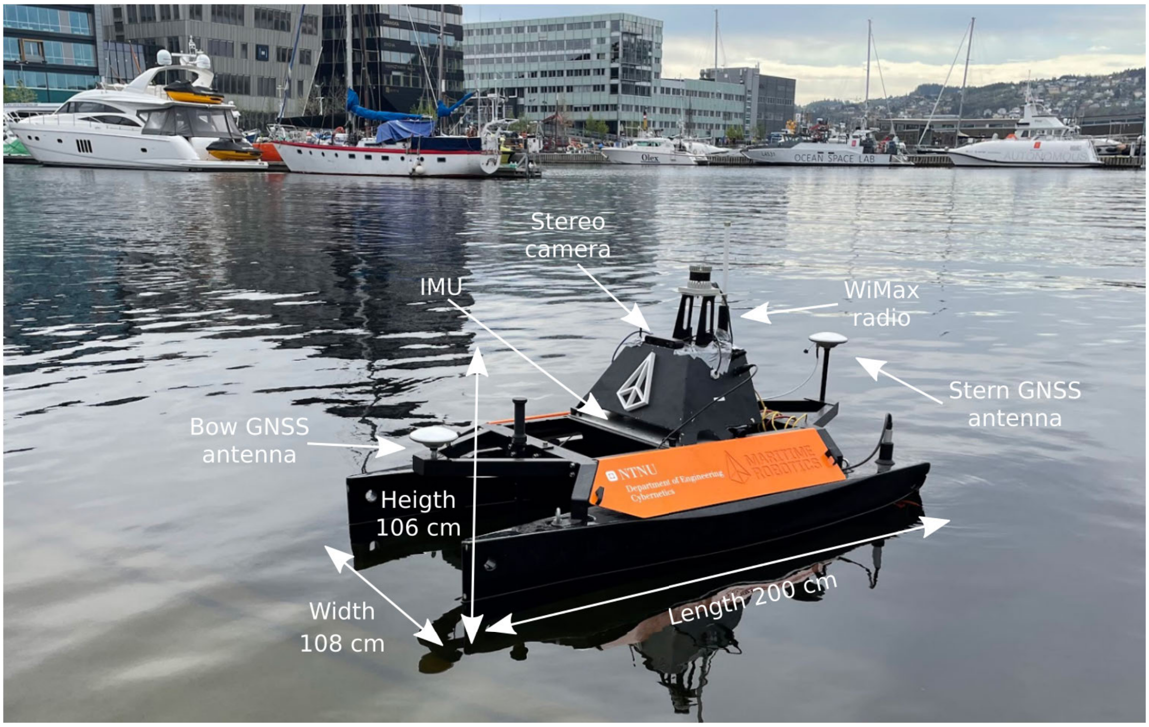


FIGURE 1. **Overview of the NTNU Otter unmanned surface vehicle.**

drift-free localization updates [14]. Moreover, the docking station is assumed to be fixed for surface vehicles. Hence, only a few tags are necessary to aid the vehicle for each dockside, which makes the solution cheap and practical. As such, the following literature review describes the work on vision-based navigation systems aided by landmarks, with a particular focus on docking operations in outdoor maritime environments.

The use of artificial tags in underwater operations has been investigated by many researchers lately because they function in GNSS-denied environments. In this context, Myint et al. [27] and Hsu et al. [28] developed vision-based docking and recharging systems for autonomous underwater vehicles using stereo vision. Both used a 3-D model-based matching method combined with custom 3-D markers to estimate the pose of the vehicle. Furthermore, Chen et al. [29] used a collection of Apriltags to estimate the camera-tag pose of a remotely operated vehicle (ROV) underwater before employing an EKF based on the camera-tag pose. The experimental results were promising, both with and without physical disturbances, but the dock was limited to a small indoor pool environment for the aforementioned work. Trslic et al. [30] tested a vision-based docking system for work-class ROVs in the North Atlantic Ocean. They combined monocular vision with a customized light marker (i.e., light beacons) to estimate the relative pose between the ROV and the docking station fixed to the mother ship. Through field experiments, they demonstrated that the vehicle was able to dock up to 8 m from the dynamic docking station. They employed a monocular camera with a single-tag configuration using a Perspective-n-Point (PnP) solver with four image-point correspondences. Unfortunately, such a configuration is prone to rotational ambiguity when in weak-perspective conditions [31].

Although vision-based docking in environments above the surface shares many commonalities with underwater environments, they also pose different challenges. In particular, visual sight above the surface is sensitive to adsverse weather such as rain and sunlight. To address the challenge of robust detection and pose estimation in outdoor environments, Volden et al. [32] suggested a complementary modular approach by combining a learning-based object detection model with traditional computer vision techniques based on monocular and stereo vision. They showed that the proposed vision-based docking system for USVs aided by ArUco markers performed well in the harbor environment in sunny and cloudy weather, despite being negatively affected by non-uniform light and water reflections. However, system performance under other adverse conditions was not validated, and the proposed method was not implemented in closed-loop control.

Also, the unmanned aerial vehicle (UAV) industry has shown increased attention to vision-based localization aided by artificial markers, especially for precision landing. For example, Malyuta et al. [33] designed a UAV system assisted by AprilTags for precise landing and automatic charging and demonstrated a 4-hour long mission outdoors without human intervention. Similar to our work, they use PnP with multiple tags to extract a single pose measurement rather than computing individual tag poses for each tag. However, instead of fusing visual and inertial data, they use a recursive least square filter to estimate the camera-tag pose based on visual data only. Further, Kayhani et al. [34] use a tag-based visual-inertial localization method that fuses inertial data with AprilTag measurements using an on-manifold EKF. The method was tested onboard a UAV with promising results but limited to indoor environments. Conceptually, Song et al. [35] propose a similar solution. However, they simplify the design of the inner codification of the tags to speed up the detection frequency. Unfortunately, these tags have very simple geometric complexity and are, therefore, likely to occur in natural scenes and provoke false positives.

B. MAIN CONTRIBUTIONS

In this paper, we demonstrate how visual-inertial sensor fusion aided by visual fiducial tags can be used to accurately estimate the complete state of the vehicle during the terminal docking phase. The novel contribution is the extension of the error-state attitude filter, i.e., the MEKF, to a complete navigation solution, which includes the translational motions (position, linear velocity, specific force biases). In addition, the paper also implements the error-state Kalman filter as a feedback algorithm by using reset functionality, which is necessary for long-endurance USV missions. We argue that our study is unique, as we provide a complete description of how USVs can dock based on a tag-based visual-inertial sensor fusion scheme. For experimental validation, we provide a synchronized dataset, including regular and adverse weather data, with visual and inertial data and highly accurate RTK GNSS for benchmarking. The robustness and performance of the proposed method are tested and evaluated in normal and adverse weather conditions, followed by recommendations on system capabilities and limitations. Results show that the proposed method performs well in normal conditions and conditions with degraded visibility due to sunlight, darkness, fog, and rain. Moreover, we show that our application-specific implementation of an ESKF performs well in feedback control through field experiments. We have made source code, data set, and instructions to run the algorithms in the Robot Operating System (ROS) [36] available in a public Github repository [37]. The core implementation is based on the MEKF algorithm described in the Marine Systems Simulator toolbox [38] but further interfaced and adapted to the USV sensor suite. In summary, the following are considered the main contributions of this study:

- We derive and implement a complete navigation solution aided by visual fiducial tags for robust estimation of the state of the vehicle.
- We carefully describe the design of the tag system and show that a multi-tag configuration with pre-specified pose offsets is beneficial to avoid flip ambiguity and achieve more robust and accurate tag measurements.
- We experimentally validate the proposed visual-inertial system in normal and adverse weather conditions and use dual-antenna RTK GNSS to benchmark the accuracy.
- We demonstrate that landmark-based navigation can be used for high-precision docking of USVs in feedback control through field experiments in the harbor.

C. OUTLINE

The paper is structured as follows. Section II introduces the proposed visual-inertial system, including the visual tag system, coordinate transformations, and the Kalman filter design. In Section III, the experimental setup is described, followed by a description of how the experiments were conducted. Then, in Section IV, we present and discuss the experimental results. Finally, we present the conclusion and discuss relevant issues for future work in Section V.

II. TAG-BASED VISUAL-INERTIAL SENSOR FUSION

A. SYSTEM OVERVIEW

The proposed method consists of two main components: A visual fiducial system to estimate the relative camera-tag pose and a sensor fusion scheme to estimate the full state of the vehicle, including position, velocity, and attitude. The visual fiducial system is used to recognize and uniquely identify the tags based on their inner coding, detect the four corners of each tag and compute the camera-relative position and orientation of the tags. The sensor fusion scheme utilizes an ESKF to fuse IMU data with relative camera-tag poses coming from the visual fiducial system, as seen in Fig. 2. The ESKF is also implemented to receive GNSS measurements when landmark updates are out of range. All coordinate systems necessary for the sensor fusion scheme, including the camera, the tag, the body, and the North-East-Down (NED) frame, are illustrated from the top view in Fig. 3.

B. VISUAL FIDUCIAL SYSTEM

Fiducial markers are artificial landmarks of known size and shape that feature a specific pattern that is used to identify them. They can be used to establish a visual reference, thus assisting in applications such as camera calibration, localization, and mapping. In our work, we employ a ROS wrapper [39] of *AprilTag* [12]: A black-and-white square fiducial marker system. The visual fiducial system consists of three main components: A detector, a coding system, and a pose estimation algorithm.

1) DETECTOR, CODING SYSTEM AND POSE ESTIMATION

The detection process begins by searching for four-sided regions, known as a *quad*. By intersecting the line segments forming the quad, the detector obtains the four corners of the tag. Subsequently, a digital coding system assigns a unique ID to each tag based on its inner codification, also referred to as the *codeword*. The available codewords are carefully selected to be robust and error-correcting, allowing them to be used for longer ranges and in conditions with degraded visibility. We use 6 × 6 tag size for a balanced tradeoff between speed and accuracy. Once the tag is detected and uniquely classified,

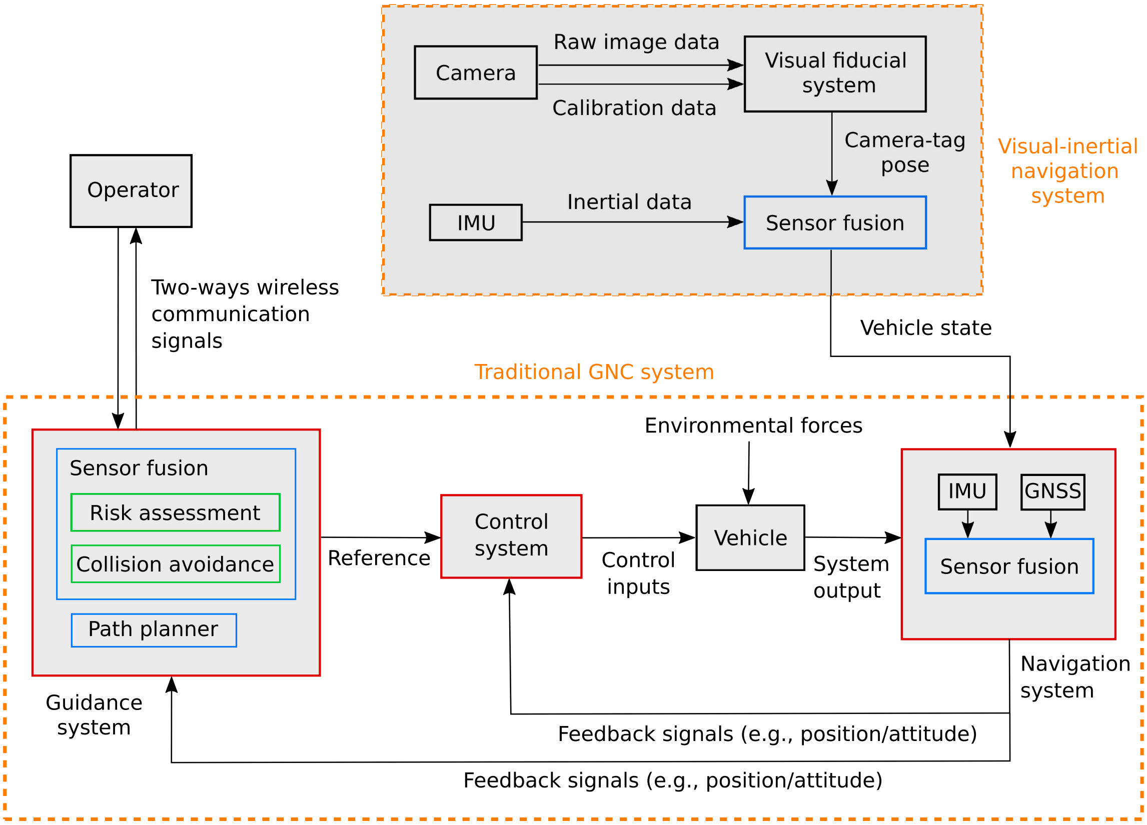
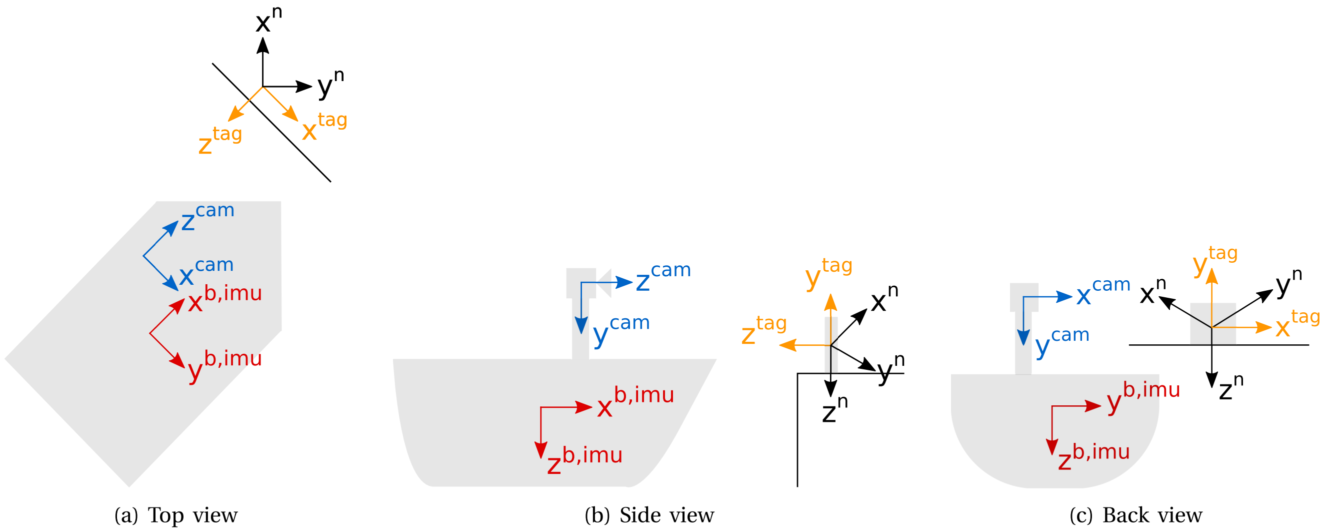


FIGURE 2. Overview of the proposed tag-based visual-inertial navigation system and its interaction with the traditional guidance, navigation, and control (GNC) system.



(a) Top view (b) Side view (c) Back view

FIGURE 3. Overview of the coordinate systems onboard the vehicle and at the dockside.

the image is undistorted using radial and tangential distortion coefficients before image-to-point correspondences are extracted from the associated tag corners. These correspondences, combined with the camera calibration matrix and the physical size of the tag, are used by the PnP solver to estimate the camera-tag pose. Note that the physical size of the tag resolves the scale ambiguity and enables the reconstruction of the absolute pose from single-view geometry. We refer to Olson [12] for a more detailed description of the AprilTag system.

C. COORDINATE TRANSFORMATIONS

For strapdown INS, it is common to express the measurements from the aiding sensors in the NED frame, with IMU measurements in the body frame. As such, we transform the measurements coming from the visual fiducial system and GNSS to the NED frame. We also transform the INS estimates from the NED frame to the global longitude-latitude representation when they are used by other components in the feedback loop. Since the IMU is located at the center of the vehicle with orientation aligned with the body frame, no lever-arm compensation is necessary to transform from the IMU frame to the body frame.

1) CAMERA: TAG TO NED

Let $\boldsymbol{T}^{\text{tag}}_{\text{cam}}$ be the transformation matrix containing the rotation and translation of the camera in the tag frame, computed by AprilTag. Furthermore, let $\boldsymbol{T}^{\text{cam}}_{b}$ be the transformation matrix that includes the fixed translation from the camera to the center of the vehicle in the camera frame. Finally, let $\boldsymbol{T}^{n}_{\text{tag}}$ be the transformation matrix that includes the rotation matrix to rotate from the tag frame to the NED frame. For convenience, the NED and the tag frame have identical origins located in the middle of the reference tag, as seen in Fig. 3. By chaining the transformations, we get

$$\begin{aligned}\boldsymbol{T}^n_b &= \boldsymbol{T}^n_{\text{tag}}\boldsymbol{T}^{\text{tag}}_{\text{cam}}\boldsymbol{T}^{\text{cam}}_{b} \\ &= \begin{bmatrix}\boldsymbol{R}_{z',-\psi_{\text{off}}}\boldsymbol{R}_{x,\pi/2} & 0_{3\times 1} \\ 0 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{R}^{\text{tag}}_{\text{cam}} & \boldsymbol{t}^{\text{tag}}_{\text{cam}} \\ 0 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_3 & \boldsymbol{t}^{\text{cam}}_{b} \\ 0 & 1\end{bmatrix},\end{aligned} \tag{1}$$

where $\boldsymbol{R}_{x,\pi/2} \in \text{SO(3)}$ rotates the x-axis of the tag frame by $\pi/2$, and $\boldsymbol{R}_{z',-\psi_{\text{off}}} \in \text{SO(3)}$ rotates the z-axis of the subsequent frame by $-\psi_{\text{off}}$. The last rotation, $\boldsymbol{R}_{z',-\psi_{\text{off}}}$, finalizes the transformation to the NED frame and is shown in Fig. 4. As such, both $\boldsymbol{T}^{n}_{\text{tag}}$ and $\boldsymbol{T}^{\text{cam}}_{\text{b}}$ are assumed to be known. Further, $\boldsymbol{T}^{\text{tag}}_{\text{cam}}$ can be reconstructed since AprilTag computes the translation $\boldsymbol{t}^{\text{tag}}_{\text{cam}}$ directly, and the unit quaternion $\boldsymbol{q}^{\text{tag}}_{\text{cam}}$ can be mapped to the rotation matrix $\boldsymbol{R}^{\text{tag}}_{\text{cam}} \in \text{SO(3)}$. Finally, the attitude of the vehicle in the NED frame can be reconstructed by extracting $\boldsymbol{R}^n_b \in \text{SO(3)}$ from (1) before transforming it into the unit quaternion $\boldsymbol{q}^n_b$. Also, the center position of the vehicle in the NED frame is computed as

$$\boldsymbol{p}^n_{nb} = \boldsymbol{R}^n_b\boldsymbol{p}^b_{nb}, \tag{2}$$

where $\boldsymbol{p}^b_{nb}$ is the relative position between the center of the vehicle and the NED origin expressed in the body frame.

2) GNSS: WGS-84 TO NED

Since strapdown INS uses flat-Earth coordinates, we transform the GNSS position of the vehicle from the World Geodetic System 1984 (WGS-84) ellipsoid [40] to NED coordinates. For convenience, the origin of the NED frame, expressed in longitude-latitude coordinates (l_0, μ_0), corresponds to the midpoint of the reference tag, as shown in Fig. 3. Because of the static-world assumption, (l_0, μ_0) is assumed to be known. Given (l_0,μ_0) and the radius of curvature in the meridian R_M [41], the North-East position (x^n, y^n) can be computed from (l,μ) by

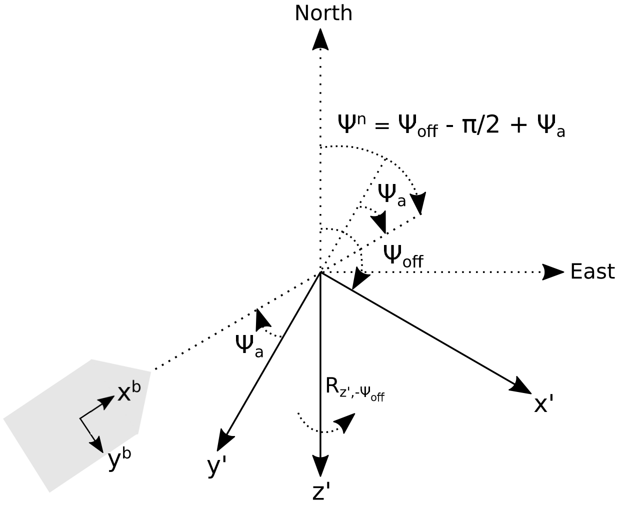


FIGURE 4. The heading of the vehicle in the NED frame depends on the angle of the vehicle relative to the axis pointing out of the tag (ψ_a) and the fixed yaw offset between true north and the x-axis of the tag frame (ψ_{off}).

$$x^n = \frac{\Delta\mu}{\text{atan2}(1, R_M)} \tag{3}$$

$$y^n = \frac{\Delta l}{\text{atan2}(1, \cos(\mu_0))}, \tag{4}$$

where $(\Delta l, \Delta\mu) := (l - l_0, \mu - \mu_0)$. We neglect the height since the vehicle operates in 2-D, i.e., on the surface of the geoid.

3) INS: NED TO WGS-84

Since the guidance and control system uses the longitude-latitude representation, we also need to transform the position of the vehicle from flat-Earth coordinates to WGS-84. Given the NED position (x^n, y^n) and the radius of curvature in the meridian R_M, the longitude-latitude error $(\Delta l, \Delta\mu)$ can be described as

$$\Delta l = y^n \text{atan2}(l, R_M \cos(\mu_0)) \tag{5}$$

$$\Delta\mu = x^n \text{atan2}(l, R_M), \tag{6}$$

where atan2(y,x) is the four-quadrant inverse tangent enclosing the result to $(-\pi, \pi]$. Longitude and latitude can then be computed as

$$l = \text{ssa}(l_0 + \Delta l) \tag{7}$$

$$\mu = \text{ssa}(l_0 + \Delta\mu), \tag{8}$$

where $\text{ssa}(\cdot)$ is the *smallest signed angle* confining the argument to the interval $[-\pi, \pi)$.

D. THE ERROR-STATE KALMAN FILTER

The ESKF is an *indirect* filter technique in which the Kalman filter is formulated as an error-state filter. The error state $\delta\boldsymbol{x}$ includes position, velocity, attitude, and IMU bias errors.

Since the IMU is strapped to the vehicle, the INS state $\boldsymbol{x}_{\text{ins}}$ is based on the integration of *strapdown navigation equations* describing the motion of the vehicle. These equations are driven by IMU specific force and attitude rate sensor (ARS) measurements. The IMU measurements are integrated to obtain position and attitude, which results in drift due to sensor biases, misalignments, and temperature variations [25, p.476]. To compensate for the sensor biases, the filter is implemented as a feedback filter where the error estimates are used to update the INS estimates directly, as seen in Fig. 5. In particular, a reset strategy is employed by setting the error state to zero after each incoming aiding measurement, i.e., from the camera or GNSS. As such, we ensure that $\hat{\boldsymbol{x}}_{\text{ins}} \rightarrow \boldsymbol{x}$ when the error-state estimate is fed back to the strapdown navigation equations, thus preventing the INS errors from accumulating.

1) ERROR-STATE DYNAMICS

The main idea of the MEKF is that the unit quaternion error

$$\delta\boldsymbol{q}^n_b = \begin{bmatrix} \delta\eta \\ \delta\boldsymbol{\epsilon} \end{bmatrix} \tag{9}$$

can be parametrized using a three-parameter attitude representation where $\delta\eta$ and $\delta\boldsymbol{\epsilon}$ are the real and imaginary components of the unit quaternion error $\delta\boldsymbol{q}^n_b$. We will use the Gibbs vector $\boldsymbol{a}_g = [g_1, g_2, g_3]^T$ scaled by a factor of two for the mapping between the three-parameter attitude representation and the unit quaternion error:

$$\delta\boldsymbol{a}_g = \frac{\delta\boldsymbol{\epsilon}}{\delta\eta}, \quad \delta\boldsymbol{a} := 2\delta\boldsymbol{a}_g, \quad \delta\boldsymbol{q}^n_b = \frac{1}{\sqrt{4 + \delta\boldsymbol{a}^T\delta\boldsymbol{a}}}\begin{bmatrix} 2 \\ \delta\boldsymbol{a} \end{bmatrix}. \tag{10}$$

By scaling the Gibbs vector by a factor of two, the Kalman filter covariance estimates are given in radians squared, equivalent to angular errors using a first-order approximation [25, p.481]. With the three-parameter attitude error introduced using the Gibbs vector representation, we define the error state as

$$\delta\boldsymbol{x} = [(\delta\boldsymbol{p}^n_{nb})^T, (\delta\boldsymbol{v}^n_{nb})^T, (\delta\boldsymbol{b}^b_{\text{acc}})^T, \delta\boldsymbol{a}^T, (\delta\boldsymbol{b}^b_{\text{ars}})^T]^T, \tag{11}$$

where $\delta\boldsymbol{b}^b_{acc}$ and $\delta\boldsymbol{b}^b_{ars}$ denote the accelerometer and ARS bias error, respectively. Note that the $\delta\boldsymbol{a}$ vector replaces the unit quaternion error $\delta\boldsymbol{q}^n_b$, thus reducing the number of states. Further, the differential equations describing the error-state dynamics must be linearized such that they fit into the discrete-time system matrices. Hence, the resulting error-state dynamics are approximated by the following first-order linear differential equations:

$$\delta\dot{\boldsymbol{p}}^n_{nb} = \delta\boldsymbol{v}^n_{nb} \tag{12}$$

$$\delta\dot{\boldsymbol{v}}^n_{nb} \approx -\boldsymbol{R}(\hat{\boldsymbol{q}}_{\text{ins}})\boldsymbol{S}(\boldsymbol{f}^b_{\text{nb}} - \hat{\boldsymbol{b}}^b_{\text{acc}})\delta\boldsymbol{a} - \boldsymbol{R}(\hat{\boldsymbol{q}}_{\text{ins}})(\delta\boldsymbol{b}^b_{\text{acc}} + \boldsymbol{w}^b_{\text{acc}}) \tag{13}$$

$$\delta\dot{\boldsymbol{b}}^b_{\text{acc}} = -\frac{1}{T_{\text{acc}}}\delta\boldsymbol{b}^b_{\text{acc}} + \boldsymbol{w}^b_{b,\text{acc}} \tag{14}$$

$$\delta\dot{\boldsymbol{a}} \approx -\boldsymbol{S}(\boldsymbol{\omega}^b_{\text{nb}} - \hat{\boldsymbol{b}}^b_{\text{ars}})\delta\boldsymbol{a} - \delta\boldsymbol{b}^b_{\text{ars}} - \boldsymbol{w}^b_{\text{ars}} \tag{15}$$

$$\delta\dot{\boldsymbol{b}}^b_{\text{ars}} = -\frac{1}{T_{\text{ars}}}\delta\boldsymbol{b}^b_{\text{ars}} + \boldsymbol{w}^b_{b,\text{ars}}, \tag{16}$$

where $\boldsymbol{f}^b_{\text{nb}}$ is the specific force vector, and $\boldsymbol{\omega}^b_{\text{nb}}$ is the angular velocity vector, both expressed in the body frame. Further, T_{ars} and T_{acc} are time constants that ensure that the bias errors go exponentially to zero during dead reckoning, and the additive zero-mean Gaussian white noise terms $\boldsymbol{w}^b_{\text{acc}}$, $\boldsymbol{w}^b_{\text{ars}}$, $\boldsymbol{w}^b_{b,\text{acc}}$, and $\boldsymbol{w}^b_{b,\text{ars}}$, are used to model the measurement and bias noise, respectively. Also note that $\hat{\boldsymbol{q}}_{\text{ins}} \equiv \hat{\boldsymbol{q}}^n_b$. For further details on the error-state dynamics, we refer to Fossen [25, p.481-483].

2) KALMAN FILTER MEASUREMENTS

For positioning, GNSS is usually the primary sensor for aiding of surface vehicles. We will, however, use the relative position between the tag and the camera, estimated by AprilTag, if the landmark updates are accurate. By using the chain of homogeneous transformations in (1), we express the position of the vehicle in the NED frame. Furthermore, the AprilTag framework does not provide velocity directly. Hence, the error-measurement equations will include position but not velocity measurements:

$$\delta\boldsymbol{y}_p = (\boldsymbol{p}^n_{nb} + \boldsymbol{\epsilon}_p) - \hat{\boldsymbol{p}}^n_{nb} = \delta\boldsymbol{p}^n_{nb} + \boldsymbol{\epsilon}_p, \tag{17}$$

where $\boldsymbol{\epsilon}_p$ is assumed to be Gaussian white measurement noise.

In order to successfully estimate the unit quaternion for attitude determination, a heading reference is needed to guarantee observability. Commercial ships usually combine the gravity vector with a high-quality gyrocompass for this purpose [25, p.484]. We will, however, replace the gyrocompass with a camera, which computes the relative orientation between the tag and the camera. We express the attitude of the vehicle coming from AprilTag as the rotation between body and NED using the homogeneous transformations in (1). Regarding the gravity vector, we start by defining the normalized specific force vector $\boldsymbol{v}^b_1$ as

$$\boldsymbol{v}^b_1 := -\frac{\boldsymbol{f}^b}{g(\mu)}, \tag{18}$$

where the WGS-84 ellipsoidal gravity formula [40] is used to compute $g(\mu)$ based on latitude μ and $\boldsymbol{f}^b$ is the unbiased specific force vector expressed in the body frame. Then, the estimated vector is $\hat{\boldsymbol{v}}^b_1 = \boldsymbol{R}^T(\hat{\boldsymbol{q}}^n_b)\boldsymbol{v}^n_{01}$, where $\boldsymbol{v}^n_{01} = [0, 0, 1]^T$ is chosen as the *gravity reference vector*, pointing downwards

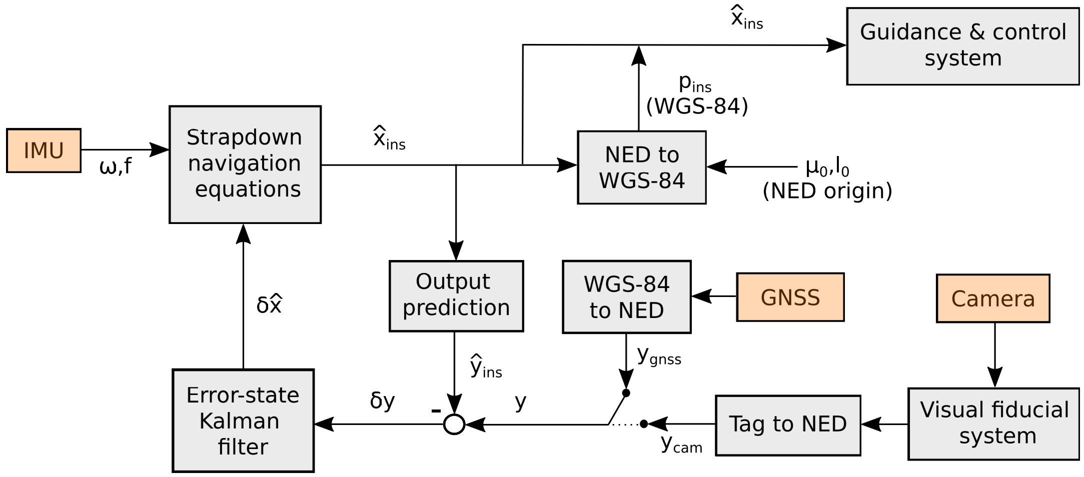


FIGURE 5. An overview of the error-state filter aided by drift-free camera and GNSS measurements. The filter receives either GNSS or camera measurements, depending on the camera-tag range.

with respect to the NED frame, and $\boldsymbol{R}^T(\boldsymbol{q}_b^n) \in SO(3)$ represents the unit quaternion rotation matrix from the NED frame to the body frame. The following error-measurement equation, describing the error gravity reference vector, is approximated as

$$\delta \boldsymbol{v}_1 = (\boldsymbol{v}_1^b + \boldsymbol{\epsilon}_1) - \boldsymbol{R}^T(\hat{\boldsymbol{q}}_b^n)\boldsymbol{v}_{01}^n \approx \boldsymbol{S}(\boldsymbol{R}^T(\hat{\boldsymbol{q}}_{\text{ins}})\boldsymbol{v}_{01}^n)\delta\boldsymbol{a} + \boldsymbol{\epsilon}_1, \tag{19}$$

where $\boldsymbol{\epsilon}_1$ is assumed to be Gaussian white measurement noise. As described by Fossen [25, p.484], we can use the scaled Gibbs vector $\boldsymbol{a}$ to express a nonsingular solution for the heading measurement

$$\psi = h(\boldsymbol{a}) = \tan^{-1}\left(\frac{2(a_1 a_2 + 2a_3)}{4 + a_1^2 - a_2^2 - a_3^2}\right), \tag{20}$$

where $\boldsymbol{a} = [a_1, a_2, a_3]^T$. By linearization about $\boldsymbol{a} = \hat{\boldsymbol{a}}$, we get

$$\delta y_\psi = \psi - h(\hat{\boldsymbol{a}}) \approx \frac{\partial h(\boldsymbol{a})}{\partial \boldsymbol{a}}^T_{\boldsymbol{a}=\hat{\boldsymbol{a}}} \delta\hat{\boldsymbol{a}}, \tag{21}$$

where the gradient can be computed using the chain rule. Note that the gradient used in the Kalman filter measurement matrix can also be computed from the unit quaternion:

$$\boldsymbol{c}_\psi(\hat{\boldsymbol{q}}_{\text{ins}}) := \frac{\delta h(\boldsymbol{a})}{\delta \boldsymbol{a}}\Big|_{\delta\boldsymbol{a}=\delta\hat{\boldsymbol{a}}}. \tag{22}$$

The error-measurement equations are summarized by (17), (19) and (21). Following this order, the measurement vector can be defined as

$$\delta \boldsymbol{y} = [(\delta \boldsymbol{y}_p)^T, (\delta \boldsymbol{v}_1)^T, \delta y_\psi]^T, \tag{23}$$

where $\delta \boldsymbol{y}_p$ denotes the measured position error, $\delta \boldsymbol{v}_1$ denotes the error gravity reference vector, and δy_ψ is the measured heading error. We refer to Fossen [25, p.483-486] for further details regarding the error-measurement equations.

3) MATHEMATICAL MODELLING OF THE ERROR-STATE FILTER

Given the differential equations describing the error state as well as the error-measurement equations, we proceed by presenting the mathematical models used for INS state propagation. The INS estimates are obtained by integrating the strapdown navigation equations with high-rate IMU measurements, i.e., specific force $\boldsymbol{f}_{nb}^b$ and angular velocity $\boldsymbol{\omega}_{nb}^b$. We emphasize that the strapdown navigation equations use the unit quaternion for attitude representation. It is only the MEKF that employs the Gibbs vector. The INS state also includes the estimated sensor biases $\hat{\boldsymbol{b}}_{\text{ins,acc}}^b$ and $\hat{\boldsymbol{b}}_{\text{ins,ars}}^b$ for online bias compensation since accelerometer and ARS biases will grow over time. Hence, it is necessary to estimate them during operation, especially for long-endurance applications. This results in the following system of differential equations describing the INS estimates

$$\dot{\hat{\boldsymbol{p}}}_{\text{ins}}^n = \hat{\boldsymbol{v}}_{\text{ins}}^n \tag{24}$$

$$\dot{\hat{\boldsymbol{v}}}_{\text{ins}}^n = \boldsymbol{R}(\hat{\boldsymbol{q}}_{\text{ins}})\boldsymbol{f}_{\text{ins}}^b + \boldsymbol{g}^n \tag{25}$$

$$\dot{\hat{\boldsymbol{b}}}_{\text{ins,acc}}^b = \boldsymbol{0} \tag{26}$$

$$\dot{\hat{\boldsymbol{q}}}_{\text{ins}} = \boldsymbol{T}(\hat{\boldsymbol{q}}_{\text{ins}})\boldsymbol{\omega}_{\text{ins}}^b \tag{27}$$

$$\dot{\hat{\boldsymbol{b}}}_{\text{ins,ars}}^b = \boldsymbol{0}, \tag{28}$$

where $\boldsymbol{f}_{\text{ins}}^b := \boldsymbol{f}_{\text{nb}}^b - \hat{\boldsymbol{b}}_{\text{ins,acc}}^b$ and $\boldsymbol{\omega}_{\text{ins}}^b := \boldsymbol{\omega}_{\text{nb}}^b - \hat{\boldsymbol{b}}_{\text{ins,ars}}^b$ are the bias-compensated IMU measurements. Further, $\boldsymbol{T}(\hat{\boldsymbol{q}}_{\text{ins}})$ is a 4×4 quaternion transformation matrix from the body frame to the NED frame since $\hat{\boldsymbol{q}}_{\text{ins}} \equiv \hat{\boldsymbol{q}}_b^n$, and $\boldsymbol{g}^n = [0, 0, g(\mu)]^T$ is the WGS-84 ellipsoidal gravity vector.

A key point is that the INS estimates are corrected by setting the estimated error-state vector to zero for each new measurement coming from the aiding sensor. This is mathematically equivalent to $\hat{\boldsymbol{x}}_{\text{ins}} \leftarrow \hat{\boldsymbol{x}}_{\text{ins}} + \delta\hat{\boldsymbol{x}}$. Hence, the estimated error state $\delta\hat{\boldsymbol{x}}$ is computed by the filter before it is injected into the INS state $\hat{\boldsymbol{x}}_{\text{ins}}$. Note that the unit quaternion $\hat{\boldsymbol{q}}_{\text{ins}}$ is rotated by the estimated unit quaternion error $\delta\hat{\boldsymbol{q}}_b^n$ using the Hamiltonian product, as shown in Algorithm 1. The error-state dynamics (12)-(16) can be represented by a 15-states model

$$\delta\dot{\boldsymbol{x}} = \boldsymbol{A}\delta\boldsymbol{x} + \boldsymbol{E}\boldsymbol{w} := \boldsymbol{f}(\delta\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w}) \tag{29}$$

$$\delta\boldsymbol{y} = \boldsymbol{C}\delta\boldsymbol{x} + \boldsymbol{\epsilon} := \boldsymbol{h}(\delta\boldsymbol{x}, \boldsymbol{u}) + \boldsymbol{\epsilon}, \tag{30}$$

where $\boldsymbol{w} = [(\boldsymbol{w}_{\text{acc}}^b)^T, (\boldsymbol{w}_{b,\text{acc}}^b)^T, (\boldsymbol{w}_{\text{ars}}^b)^T, (\boldsymbol{w}_{b,\text{ars}}^b)^T]^T$ and $\boldsymbol{u} = [(\boldsymbol{f}_{nb}^b)^T, (\boldsymbol{\omega}_{nb}^b)^T, (\boldsymbol{g}^n)^T]^T$. After initialization, the nonlinear error-state model (29) and (30) is linearized about $\delta\boldsymbol{x}[k] = \boldsymbol{0}$ and $\delta\boldsymbol{w}[k] = \boldsymbol{0}$ for each time step k using Euler's integration method. Hence, we obtain the discretized error-state model

$$\delta\boldsymbol{x}[\text{k+1}] = \boldsymbol{A}_d[\text{k}]\delta\boldsymbol{x}[\text{k}] + \boldsymbol{E}_d[\text{k}]\boldsymbol{w}[\text{k}] \tag{31}$$

$$\delta\boldsymbol{y}[\text{k}] = \boldsymbol{C}_d[\text{k}]\delta\boldsymbol{x}[\text{k}] + \boldsymbol{\epsilon}[\text{k}] \tag{32}$$

with discrete-time system matrices as in (33)–(35), shown at the bottom of the page, where h is the IMU sampling time. Finally, we discretize the INS estimates for the next time step $k+1$. We refer to Algorithm 1 for further details regarding the implementation of the filter.

4) FILTER TUNING AND VALIDATION

The process and measurement noise covariance matrices, $\boldsymbol{Q}$ and $\boldsymbol{R}$, respectively, are the main tuning components for the Kalman filter. The diagonal entries of $\boldsymbol{Q}$, describing the uncertainty of the IMU's accelerometer and gyro and their associated biases, can be determined from the random walk and in-run stability in the IMU datasheet. Then, $\boldsymbol{Q}$ is discretized using IMU sampling time h. The measurement noise matrix $\boldsymbol{R}$, incorporating the uncertainty of the measurement variables, is usually tuned after $\boldsymbol{Q}$ is determined. Fortunately, we have a synchronized experimental dataset with dual-antenna RTK GNSS available. Hence, we can use the RTK GNSS measurements as the true state for comparison with the state estimates and evaluate the performance using root mean square error (RMSE). By evaluating different values of the discretized measurement noise matrix $\boldsymbol{R}_d$, we can obtain a low RMSE value. Finally, we test $\boldsymbol{Q}_d$ and $\boldsymbol{R}_d$ on different experimental datasets to verify filter consistency. For convenience, the INS state estimate $\hat{\boldsymbol{x}}_{\text{ins}}$ is initialized using RTK GNSS position, velocity, and attitude measurements. Finally, we initialize the initial covariance matrix $\hat{\boldsymbol{P}}^-[0]$ with relatively low values on its diagonal entries because the estimated state is initialized with accurate measurements.

III. EXPERIMENTAL SETUP

The experimental setup consists of the NTNU Otter USV, a land station, and visual fiducial tags. The land station includes an RTK base station that transmits correction data to the navigation system and a remote computer for the operator to upload missions or control the USV directly. The land station communicates with the NTNU Otter using point-to-point radio communication. The visual fiducial tags are located at the dock to aid the vehicle under the docking operation. Fig. 6 shows an overview of the experimental scene, including the land station, the tags, and the NTNU Otter in the harbor environment.

A. THE NTNU OTTER

The NTNU Otter is an underactuated vehicle produced by Maritime Robotics AS with two thrusters mounted at the stern. The software and hardware design were developed

$$\begin{aligned}
\boldsymbol{A}_d[k] &\approx \boldsymbol{I}_{15} + h\frac{\partial \boldsymbol{f}(\delta\boldsymbol{x}[k], \boldsymbol{u}[k], \boldsymbol{w}[k])}{\partial \delta\boldsymbol{x}[k]}\Big|_{\delta\boldsymbol{x}[k]=\boldsymbol{0}, \delta\boldsymbol{w}[k]=\boldsymbol{0}} \approx \boldsymbol{I}_{15} \\
&+ h\begin{bmatrix}
\boldsymbol{0}_{3\times3} & \boldsymbol{I}_3 & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} \\
\boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & -\boldsymbol{R}(\hat{\boldsymbol{q}}_{\text{ins}}[k]) & -\boldsymbol{R}(\hat{\boldsymbol{q}}_{\text{ins}}[k])\boldsymbol{S}(\boldsymbol{f}_{\text{ins}}^b[k]) & \boldsymbol{0}_{3\times3} \\
\boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & -\frac{1}{T_{\text{acc}}}\boldsymbol{I}_3 & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} \\
\boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{S}(\boldsymbol{\omega}_{\text{ins}}^b[k]) & -\boldsymbol{I}_3 \\
\boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & -\frac{1}{T_{\text{ars}}}\boldsymbol{I}_3
\end{bmatrix}
\end{aligned} \tag{33}$$

$$\begin{aligned}
\boldsymbol{C}_d[k] &\approx \frac{\partial \boldsymbol{h}(\delta\boldsymbol{x}[k], \boldsymbol{u}[k])}{\partial \delta\boldsymbol{x}[k]}\Big|_{\delta\boldsymbol{x}[k]=\boldsymbol{0}} \\
&\approx \begin{bmatrix}
\boldsymbol{I}_3 & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} \\
\boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{S}(\boldsymbol{R}^T(\hat{\boldsymbol{q}}_{\text{ins}}[k])\boldsymbol{v}_{01}^n) & \boldsymbol{0}_{3\times3} \\
\boldsymbol{0}_{1\times3} & \boldsymbol{0}_{1\times3} & \boldsymbol{0}_{1\times3} & \boldsymbol{c}_{\psi}^T(\hat{\boldsymbol{q}}_{\text{ins}}[k]) & \boldsymbol{0}_{1\times3}
\end{bmatrix}
\end{aligned} \tag{34}$$

$$\begin{aligned}
\boldsymbol{E}_d[k] &\approx h\frac{\partial \boldsymbol{f}(\delta\boldsymbol{x}[k], \boldsymbol{u}[k])}{\partial \delta\boldsymbol{w}[k]}\Big|_{\delta\boldsymbol{x}[k]=\boldsymbol{0}, \delta\boldsymbol{w}[k]=\boldsymbol{0}} \\
&\approx h\begin{bmatrix}
\boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} \\
-\boldsymbol{R}(\hat{\boldsymbol{q}}_{\text{ins}}[k]) & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} \\
\boldsymbol{0}_{3\times3} & \boldsymbol{I}_3 & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} \\
\boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & -\boldsymbol{I}_3 & \boldsymbol{0}_{3\times3} \\
\boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{I}_3
\end{bmatrix},
\end{aligned} \tag{35}$$

Algorithm 1 Error-State Kalman Filter With Attitude Estimation

```
    $\boldsymbol{Q}_d, \boldsymbol{R}_d$: Covariance matrices for the process and measurement noises
    $\hat{\boldsymbol{P}}^-, \hat{\boldsymbol{P}}$: A priori and posterior covariance matrices
    $\boldsymbol{K}$: Kalman gain
    $h$: IMU sampling time
    $\boldsymbol{g}^n$: WGS-84 ellipsoidal gravity vector as a function of latitude $\mu$

    $\delta\hat{\boldsymbol{x}}[k] = [\delta\hat{\boldsymbol{p}}^n_{nb}[k]^T, \delta\hat{\boldsymbol{v}}^n_{nb}[k]^T, \delta\hat{\boldsymbol{b}}^b_{\text{acc}}[k]^T, \delta\hat{\boldsymbol{a}}[k]^T, \delta\hat{\boldsymbol{b}}^b_{\text{ars}}[k]^T]^T$    (Error state)
    $\hat{\boldsymbol{x}}_{\text{ins}}[k] = [\hat{\boldsymbol{p}}^n_{\text{ins}}[k]^T, \hat{\boldsymbol{v}}^n_{\text{ins}}[k]^T, \hat{\boldsymbol{b}}^b_{\text{acc,ins}}[k]^T, \hat{\boldsymbol{q}}_{\text{ins}}[k]^T, \hat{\boldsymbol{b}}^b_{\text{ars,ins}}[k]^T]^T$    (INS state)

    Input: $(h, \mu, \boldsymbol{f}^b_{\text{nb}}, \boldsymbol{\omega}^b_{\text{nb}}, \boldsymbol{y})$
    Output: $\hat{\boldsymbol{x}}_{\text{ins}}$

$\hat{\boldsymbol{x}}_{\text{ins}}[0] = \boldsymbol{x}_0,\ \hat{\boldsymbol{P}}^-[0] = \boldsymbol{P}_0,\ \boldsymbol{Q}_d \leftarrow h\boldsymbol{Q}, \boldsymbol{R}_d \leftarrow h\boldsymbol{R},\ \boldsymbol{g}^n \leftarrow [0, 0, g(\mu)]^T$    (Initialization)

for each new IMU message do
    $\boldsymbol{f}^b_{\text{ins}}[k] \leftarrow \boldsymbol{f}^b_{\text{nb}}[k] - \hat{\boldsymbol{b}}^b_{\text{acc,ins}}[k],\ \boldsymbol{\omega}^b_{\text{ins}}[k] \leftarrow \boldsymbol{\omega}^b_{\text{nb}}[k] - \hat{\boldsymbol{b}}^b_{\text{ars,ins}}[k]$    (Bias compensation)
    $\boldsymbol{A}_d \leftarrow \boldsymbol{I}_{15} + h\boldsymbol{A} + \frac{1}{2}(h\boldsymbol{A})^2,\ \boldsymbol{C}_d \leftarrow \boldsymbol{C},\ \boldsymbol{E}_d \leftarrow h\boldsymbol{E}$    (Discrete-time system matrices)
    if new measurement then
        $\boldsymbol{K}[k] \leftarrow \hat{\boldsymbol{P}}^-[k]\boldsymbol{C}_d^T[k](\boldsymbol{C}_d[k]\hat{\boldsymbol{P}}^-[k]\boldsymbol{C}_d^T[k] + \boldsymbol{R}_d[k])^{-1}$    (Kalman gain)
        $\delta\hat{\boldsymbol{x}}[k] \leftarrow \boldsymbol{K}[k](\boldsymbol{y}[k] - \boldsymbol{C}_d[k]\hat{\boldsymbol{x}}_{\text{ins}}[k])$    (Estimation error)
        $\delta\hat{\boldsymbol{q}}^n_b[k] \leftarrow \frac{1}{\sqrt{4+\delta\hat{\boldsymbol{a}}[k]^T\delta\hat{\boldsymbol{a}}[k]}}\begin{bmatrix} 2 \\ \delta\hat{\boldsymbol{a}}[k] \end{bmatrix}$    (2 × Gibbs vector)
        $\hat{\boldsymbol{P}}[k] \leftarrow (\boldsymbol{I}_{15} - \boldsymbol{K}[k]\boldsymbol{C}_d[k])\hat{\boldsymbol{P}}^-[k](\boldsymbol{I}_{15} - \boldsymbol{K}[k]\boldsymbol{C}_d[k])^T + \boldsymbol{K}[k]\boldsymbol{R}_d[k]\boldsymbol{K}^T[k]$    (Corrector)

        // INS reset
        $\hat{\boldsymbol{p}}^n_{\text{ins}}[k] \leftarrow \hat{\boldsymbol{p}}^n_{\text{ins}}[k] + \delta\hat{\boldsymbol{p}}^n_{nb}[k]$
        $\hat{\boldsymbol{v}}^n_{\text{ins}}[k] \leftarrow \hat{\boldsymbol{v}}^n_{\text{ins}}[k] + \delta\hat{\boldsymbol{v}}^n_{nb}[k]$
        $\hat{\boldsymbol{b}}^b_{\text{acc,ins}}[k] \leftarrow \hat{\boldsymbol{b}}^b_{\text{acc,ins}}[k] + \delta\hat{\boldsymbol{b}}^b_{\text{acc}}[k]$
        $\hat{\boldsymbol{b}}^b_{\text{ars,ins}}[k] \leftarrow \hat{\boldsymbol{b}}^b_{\text{ars,ins}}[k] + \delta\hat{\boldsymbol{b}}^b_{\text{ars}}[k]$
        $\hat{\boldsymbol{q}}_{\text{ins}}[k] \leftarrow \hat{\boldsymbol{q}}_{\text{ins}}[k] \otimes \delta\hat{\boldsymbol{q}}^n_{\text{nb}}[k]$    (Schur product)
        $\hat{\boldsymbol{q}}_{\text{ins}}[k] \leftarrow \hat{\boldsymbol{q}}_{\text{ins}}[k]/||\hat{\boldsymbol{q}}_{ins}[k]||$    (Normalization)
    else
        $\hat{\boldsymbol{P}}[k] \leftarrow \hat{\boldsymbol{P}}^-[k]$    (No aiding)
    end if
    $\hat{\boldsymbol{P}}^-[k+1] \leftarrow \boldsymbol{A}_d[k]\hat{\boldsymbol{P}}[k]\boldsymbol{A}_d^T[k] + \boldsymbol{E}_d[k]\boldsymbol{Q}_d[k]\boldsymbol{E}_d^T[k]$    (Predictor)

    // INS propagation
    $\hat{\boldsymbol{p}}^n_{\text{ins}}[k+1] \leftarrow \hat{\boldsymbol{p}}^n_{\text{ins}}[k] + h\hat{\boldsymbol{v}}^n_{\text{ins}}[k]$
    $\hat{\boldsymbol{v}}^n_{\text{ins}}[k+1] \leftarrow \hat{\boldsymbol{v}}^n_{\text{ins}}[k] + h(\boldsymbol{R}^n_b(\hat{\boldsymbol{q}}_{\text{ins}}[k])\boldsymbol{f}^b_{\text{ins}}[k] + \boldsymbol{g}^n)$
    $\hat{\boldsymbol{q}}_{\text{ins}}[k+1] \leftarrow \hat{\boldsymbol{q}}_{\text{ins}}[k] + e^{\boldsymbol{T}^n_b(h(\boldsymbol{\omega}^b_{\text{ins}}[k]))}\hat{\boldsymbol{q}}_{\text{ins}}[k]$    (Exact discretization)
    $\hat{\boldsymbol{q}}_{\text{ins}}[k+1] \leftarrow \hat{\boldsymbol{q}}_{\text{ins}}[k+1]/||\hat{\boldsymbol{q}}_{\text{ins}}[k+1]||$    (Normalization)
end for
```

at the Department of Engineering Cybernetics, NTNU. The upper part of Fig. 15 in the appendix shows a hardware schematic of the NTNU Otter. Regarding in-vehicle software, we use the Underwater Systems and Technology Laboratory (LSTS) toolchain [42] for guidance and control and ROS for navigation. These middlewares run on individual computers and communicate over Ethernet. The LSTS toolchain consists of DUNE, the InterModule Communication (IMC) protocol, and the Neptus Graphical User Interface (GUI). We use DUNE for guidance and control and to interface with the hardware devices, while we use the IMC protocol to exchange data between individual DUNE tasks. The Neptus GUI is used to interact with the vehicle by setting waypoints or controlling it remotely. We use ROS to interface and fuse sensor data, i.e., from GNSS, IMU, and camera, and output the estimated state of the vehicle. Finally, we bridge the estimated state from ROS to IMC such that DUNE can use it in feedback control.

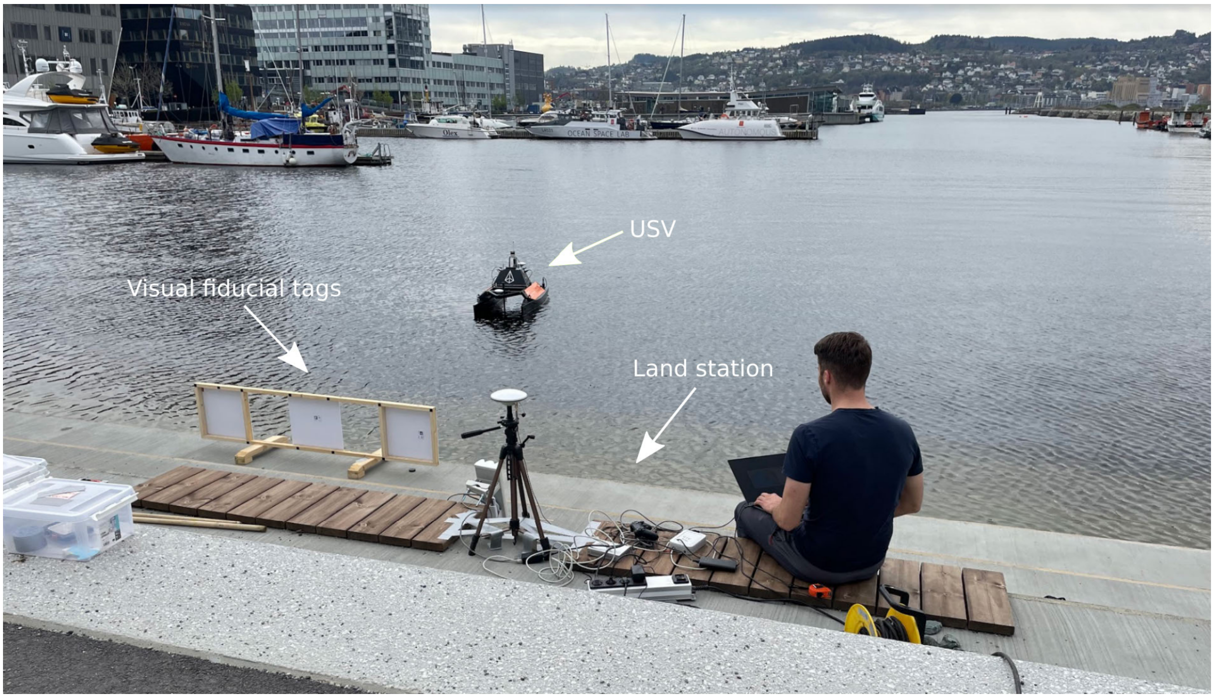


FIGURE 6. Overview of the experimental scene showing the land station, the NTNU Otter unmanned surface vehicle (USV), and the visual fiducial tags.

1) NAVIGATION SYSTEM

We use two types of navigation systems. The first navigation system is a traditional satellite-based navigation system based on GNSS and IMU. The second navigation system uses visual and inertial data to estimate the state of the vehicle, given that visual tags are visible to the camera view.

a: SATELLITE-BASED SENSOR CONFIGURATION

The NTNU Otter uses two independent satellite-based navigation systems. The first navigation system includes two U-blox F9P GNSS receivers [43] and an ADIS 16490 IMU [44] with synchronized data acquisition through a SentiBoard [45]. The first GNSS receiver configured as a ‘‘moving base’’ receives raw GNSS data from an antenna at the stern of the USV and correction data from the RTK base. The second GNSS receiver configured as a ‘‘rover’’ receives raw GNSS data from an antenna at the bow of the USV and correction data from the moving base. By using this dual-antenna configuration, the rover obtains the heading of the USV. The second navigation system, an SBG Ellipse 2D INS [46], is aided by raw GNSS data from stern and bow antennas and correction data from the RTK base. For our experiments, we mainly used the SBG Ellipse 2D INS. However, we did use the navigation data from the U-blox receivers as ground truth in the last experiment for validation. Since both navigation systems receive corrections from the same base with centimeter accuracy, the navigation data produced are almost identical. Both navigation systems express the heading of the vehicle in the NED frame and the position of the vehicle in global latitude-longitude coordinates.

b: VISUAL-INERTIAL SENSOR CONFIGURATION

The visual-inertial sensor suite consists of a ZED 2i stereo camera [47] and an ADIS 16490 IMU. We use the stereo camera with 2208 × 1242 pixel resolution (per camera) and stream the image data in monochrome pixel format at a frequency of 15 Hz. The monochrome pixel format is chosen to reduce bandwidth and processing while maintaining high image resolution, thus increasing the pose accuracy. The camera also includes built-in polarizing filters to reduce glare and reflections for increased image quality outdoors. We emphasize that only the left camera is used in the experiments since we employ monocular pose estimation. The remaining details of the camera are shown in uppermost part of Table 1. The IMU is rigidly mounted to the center of the vehicle and provides tri-axis angular rate measurements and tri-axis accelerometer-based specific force measurements at a sample rate of 250 Hz. However, the sample rate was downsampled to 50 Hz for the experiments. The lowermost part of Table 1 shows the IMU specifications necessary to tune the process noise covariance matrix $\boldsymbol{Q}$.

2) GUIDANCE AND CONTROL SYSTEM

The guidance system consists of a path planner and an integral line-of-sight (ILOS) guidance law [48]. The path planner computes the desired path based on a set of waypoints (WPs) and desired speeds from the operator. Then, the ILOS guidance law computes the desired yaw based on the desired path and the estimated state coming from the navigation system. The control system includes a proportional heading controller and a proportional-integral speed controller.

TABLE 1. Sensor specifications.

ZED 2i stereo camera	
Resolution	2208 × 1242
Pixel format	Monochrome
Sample rate	15 Hz
Sensor type	CMOS
Interface	USB 3.0
Focal length	2.12 mm (fixed)
Aperture	f/1.8
Field of View	110 deg (horizontal)
Circular polarizing filter	Yes
ADIS 16490 IMU	
Sample rate	250 Hz
In-run gyro rate bias stability	1.8 $\frac{deg}{h}$
Angular random walk	0.09 $\frac{deg}{\sqrt{h}}$
In-run accelerometer bias stability	3.6 μg
Velocity random walk	0.008 $\frac{m/s}{\sqrt{h}}$

It produces desired thrust based on the estimated state and the desired speed and yaw from the guidance system to act accordingly. The control system also allows manual control signals to be transmitted from a PlayStation 4 (PS4) controller connected to the remote control computer.

3) CLOCK SYNCHRONIZATION

To perform accurate sensor fusion, we use the Precision Time Protocol (PTP) to synchronize the sensor readings across the network of devices onboard the NTNU Otter. The hardware clocks are synchronized with sub-microsecond precision using a primary-secondary setup with PTP. We configure the Beaglebone Black computer [49] to be the primary clock and the Nvidia Jetson Xavier computer [50] to be the secondary clock. Then, the primary clock derives the time from a GNSS receiver using the National Marine Electronics Association ZDA message, as shown in Fig. 15 in the appendix.

B. LAND STATION

The land station includes an RTK base station that sends correction data to the navigation system and a remote control laptop running a GUI to interact with the vehicle. We used the remote control computer to upload predefined missions to the guidance system or to control the USV directly using a PS4 controller. The RTK base station includes a GNSS antenna, a U-blox F9P GNSS receiver, and a BeagleBone Black computer, as shown in the lower part of Fig. 15 in the appendix. The RTK survey procedure lasted for 18 hours before we conducted the experiments, resulting in an absolute precision of 5.5 cm.

C. TAG CONFIGURATION

When more accurate and precise pose estimates are required, a tag configuration consisting of multiple tags, commonly referred to as a *tag bundle*, is useful. A tag bundle is used to extract a single pose from multiple tags rather than the poses of the individual tags. Hence, the pose estimation algorithm uses 4×n tag corners, where n is the number of detected tags. Our specific tag configuration uses three coplanar AprilTags with fixed translations to each other, where the leftmost tag in Fig. 7 is the *primary tag*. The origin of the tag coordinate system is centered in the middle of the reference tag with axes defined according to the tag coordinate system in Fig. 3. We also specify the tag IDs of interest. As such, the visual fiducial system only searches for the specified tag IDs and reduces the number of false positives. The tags also have a matt surface to reduce the amount of reflection. Consequently, the visual fiducial system is more resistant to challenging illumination. Table 2 summarizes the remaining details of the tag bundle.

D. EXPERIMENTAL DESCRIPTION

We perform three experiments to demonstrate how visual-inertial state estimation can be used for automatic docking of USVs. In Experiment 1, we assess the pose accuracy of the proposed method under a regular weather scenario, as shown in Fig. 7. First, we compare the pose accuracy of AprilTag with different marker configurations, i.e., single-tag and multi-tag, against ground truth RTK GNSS heading and position. We then proceed with the multi-tag configuration and use the subsequent camera-tag pose with inertial data as input to the proposed filter, as shown in the upper part of Fig. 2. Finally, the heading and position produced by the filter are compared to ground truth RTK GNSS measurements for benchmarking.

In Experiment 2, we evaluate the robustness and performance of the proposed method in adverse weather. More specifically, we distinguish between partly degraded visibility and significantly degraded visibility and assess the following type of adverse weather: Sunlight, darkness, fog, and rain. Due to a notable reduction in performance under significantly degraded weather conditions, we only assess the proposed filter for partly degraded visibility scenarios, as seen in Fig. 9. In Experiments 1 and 2, the proposed filter is initialized with RTK GNSS measurements. We switch to visual measurements when the absolute error in heading and position between RTK GNSS and AprilTag is below 1 degree and 0.5 m, respectively. As such, we avoid sudden jumps in sensor measurements and reduce the chance for filter divergence.

In Experiment 3, we demonstrate how the proposed method performs in feedback control through field experiments in the harbor environment. Again, we use RTK GNSS measurements as input to the filter and switch to visual measurements when the error in heading and position between RTK GNSS and AprilTag is lower than a certain threshold over time. Moreover, we ensure that the estimated state used directly by the vehicle does not deviate too much from the RTK GNSS measurements. Because it is practically hard to measure the tag location with centimeter accuracy in a global frame and find the exact angle offset between the x-axis of the tag and true north, we increase the margins. Hence,

TABLE 2. Description of the tag configuration.

Tag configuration						
Tag type	**Dictionary family**	**Tag ID**	**Tag size [m]**	**Relative translation** [a] (x,y,z) **[m]**	**Relative orientation** [b] (q_w,q_x,q_y,q_z)	**Master tag**
AprilTag	36h11	227	0.412	(0,0,0)	(1,0,0,0)	Yes
AprilTag	36h11	252	0.412	(0.902,0,0)	(1,0,0,0)	No
AprilTag	36h11	546	0.412	(1.801,0,0)	(1,0,0,0)	No

[a] Translation relative to master tag in tag frame, where the translation is expressed in Euclidean space.
[b] Orientation relative to master tag in tag frame, where the orientation is expressed in unit quaternions.

we allow the estimated state to deviate more from the true heading and position of the vehicle. The experiment was conducted using the desired path with desired speed set to 0.5 m/s between the initial vehicle position and the target waypoint in front of the dock. We emphasize that Experiments 1 and 2 are performed offline on a high-performance laptop using experimental datasets, while Experiment 3 is conducted online on embedded devices onboard the USV. The field trial was conducted in weather conditions comparable to Experiment 1.

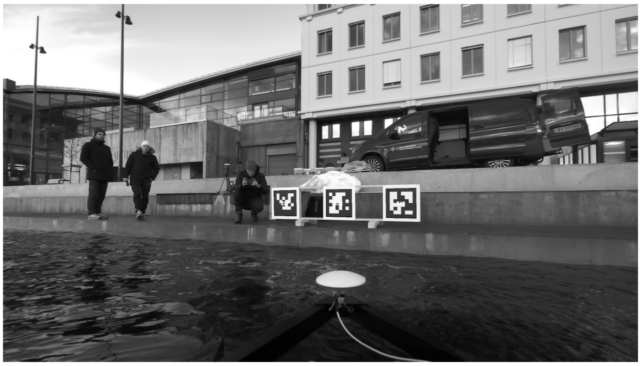


FIGURE 7. The vehicles' camera view under the docking operation in Experiment 1.

IV. EXPERIMENTAL RESULTS

We present the experimental results by plotting the USV position and heading, estimated by AprilTag, RTK GNSS + IMU (SBG INS), and the ESKF for different docking scenarios. The first two experiments express the vehicle position in a local NED frame, centered around the reference tag, while the last experiment expresses the vehicle position in geodetic coordinates. The heading of the USV is expressed relative to true north for all the experiments.

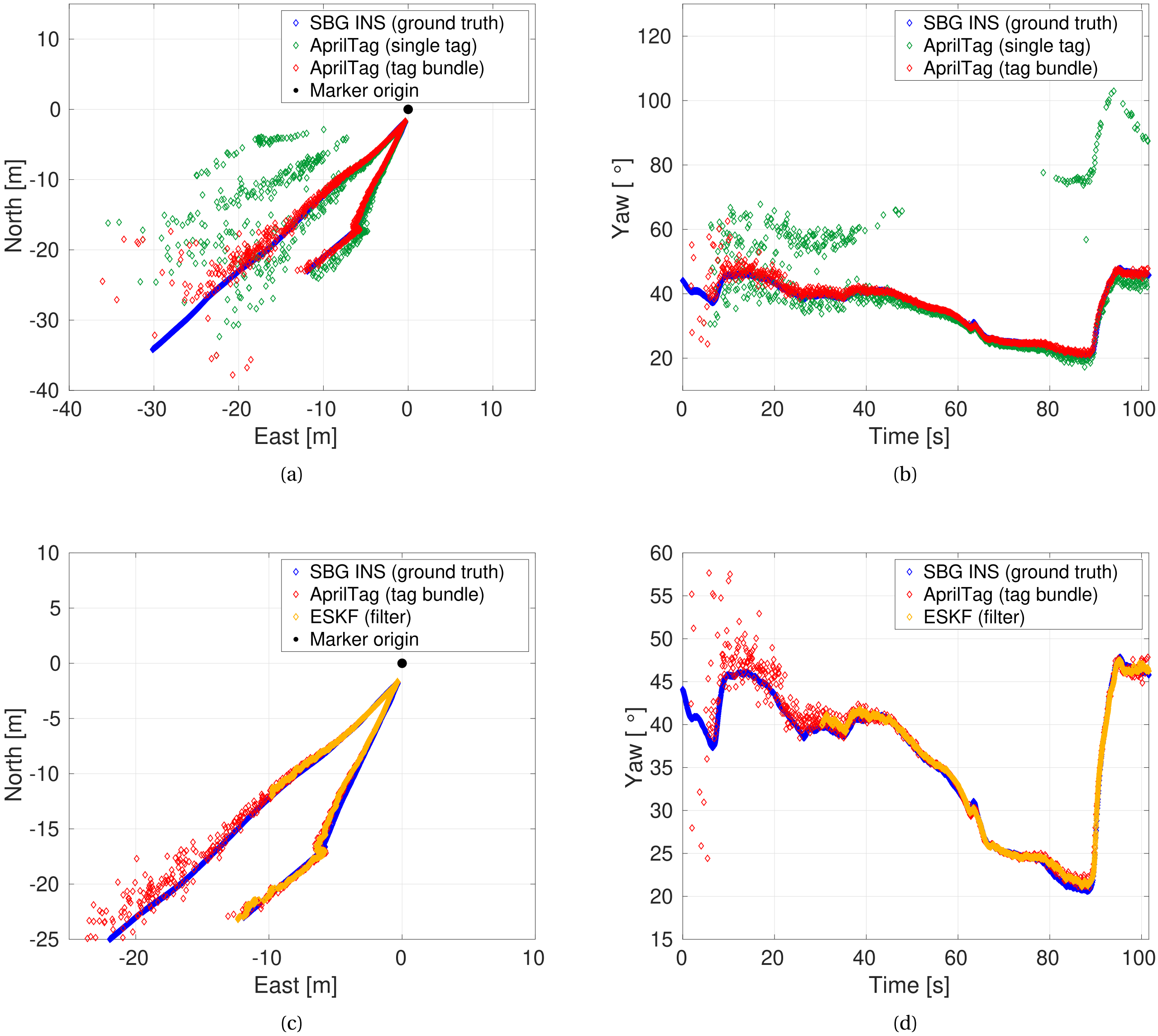
A. EXPERIMENT 1: SINGLE-TAG VS. MULTI-TAG CONFIGURATION

The results from Experiment 1 are shown in Fig. 8. Figs. 8a and 8b show the position and heading of the vehicle under a dock-then-undock sequence, respectively, estimated by AprilTag with single-tag and multi-tag configurations and compared to ground truth RTK GNSS. As seen, the single-tag configuration was particularly vulnerable to ambiguities (i.e., mirrored solutions) when the Euclidean camera-tag distance was more than 10 m. In contrast, the tag bundle configuration showed more accurate pose results, except for camera-tag distances above 30 m. In particular, the tag bundle configuration was more accurate and less noisy the closer the vehicle got to the landmarks. Figs. 8c and 8d revisit the same dock-then-undock scenario but estimate the position and heading using the ESKF and compare it to AprilTag with multiple tags and ground truth RTK GNSS. When the absolute error in heading and position between RTK GNSS and AprilTag is below 1 degree and 0.5 m, respectively, the ESKF start to output position and heading estimates. This happens at position (-12,-9.8) in the North-East frame, as seen in Fig. 8c. The ESKF continues to estimate until the dock-then-undock sequence is over, regardless of the absolute error in heading and position. The estimates tend to stay close to the true heading and position for the remaining part of the docking sequence.

B. EXPERIMENT 2: ADVERSE WEATHER CONDITIONS

1) 2.1: SUNLIGHT

The results from Experiment 2.1 is shown in Fig. 10. Figs. 10a and 10b show the position and heading of the USV, estimated by AprilTag and the ESKF, and compared to RTK GNSS in partly degraded conditions influenced by sunlight. As observed in Fig. 10a, the filter initially follows the outliers produced by the AprilTag system, thus deviating from the true path by several meters. Then, the filter converges to the true heading and position when the vehicle gets closer to the dock. Figs. 10c and 10d show the position and heading of the USV, estimated by AprilTag and compared to RTK GNSS in a new scenario influenced by significantly degraded conditions due to sunlight. We observe that the AprilTag measurements deviate heavily from ground truth position and heading and are even absent for a short



(a)

(b)

(c)

(d)

FIGURE 8. The results from Experiment 1. (a)-(b) Estimated position and heading of the vehicle using AprilTag with single-tag and tag bundle configurations. (c)-(d) Estimated position and heading of the vehicle using the error-state Kalman filter (ESKF).

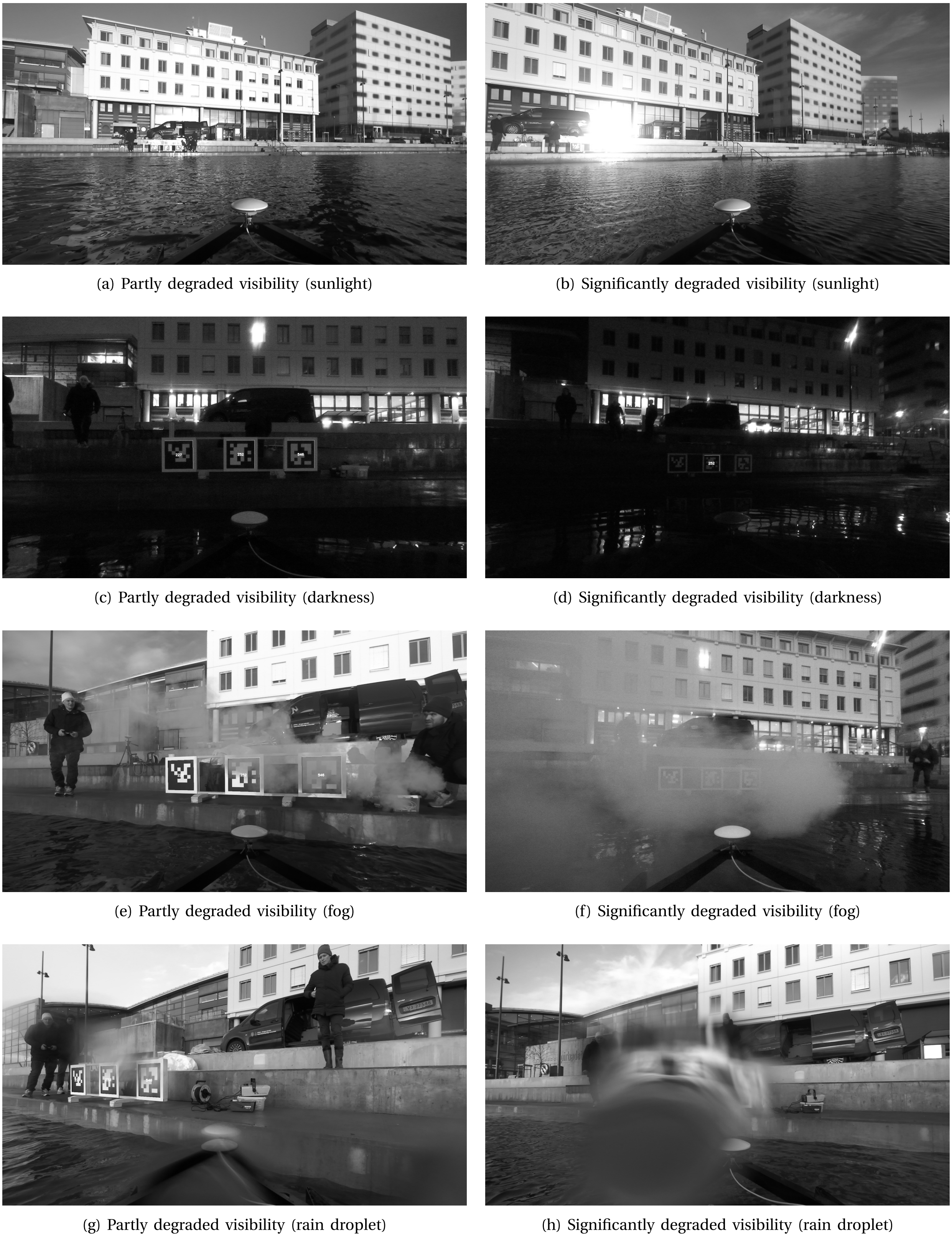
period. Since the camera was not pointing towards the tags in the last part of the docking sequence, we neglected the associated ground truth measurements thereafter. A visual representation of the scenes is shown in Figs. 9a and 9b, respectively.

2) 2.2: DARKNESS

The results from Experiment 2.2 is shown in Figs. 11. Figs. 11a and 11b show the position and heading of the USV, estimated by AprilTag and the ESKF, and compared to RTK GNSS in partly degraded conditions influenced by darkness. We observe that the filter is performing well once below the specified thresholds but produces increasingly noisy behavior as the vehicle start to reverse. Figs. 11c and 11d show the position and heading of the USV, estimated by AprilTag and compared to RTK GNSS in a new scenario influenced by significantly degraded conditions due to darkness. Notably, the AprilTag system performs decently when close to the dockside. Still, a non-negligible amount of the measurements are outliers, potentially degrading the navigation performance. A visual representation of the scene is shown in Figs. 9c and 9d.

3) 2.3: FOG

The results from Experiment 2.3 is shown in Fig. 12. Figs. 12a and 12b show the position and heading of the USV, estimated by AprilTag and the ESKF, and compared to RTK GNSS in partly degraded conditions influenced by fog. Note that we used a smoke machine to simulate foggy conditions. Hence, the produced fog is more concentrated and not as


(a) Partly degraded visibility (sunlight)

(b) Significantly degraded visibility (sunlight)

(c) Partly degraded visibility (darkness)

(d) Significantly degraded visibility (darkness)

(e) Partly degraded visibility (fog)

(f) Significantly degraded visibility (fog)

(g) Partly degraded visibility (rain droplet)

(h) Significantly degraded visibility (rain droplet)

FIGURE 9. **Adverse weather dataset.**

uniform as natural fog. As such, the camera was able to penetrate the fog to some extent, which again caused the visual fiducial system to recognize the markers and produce accurate pose estimates. Consequently, the filter produced

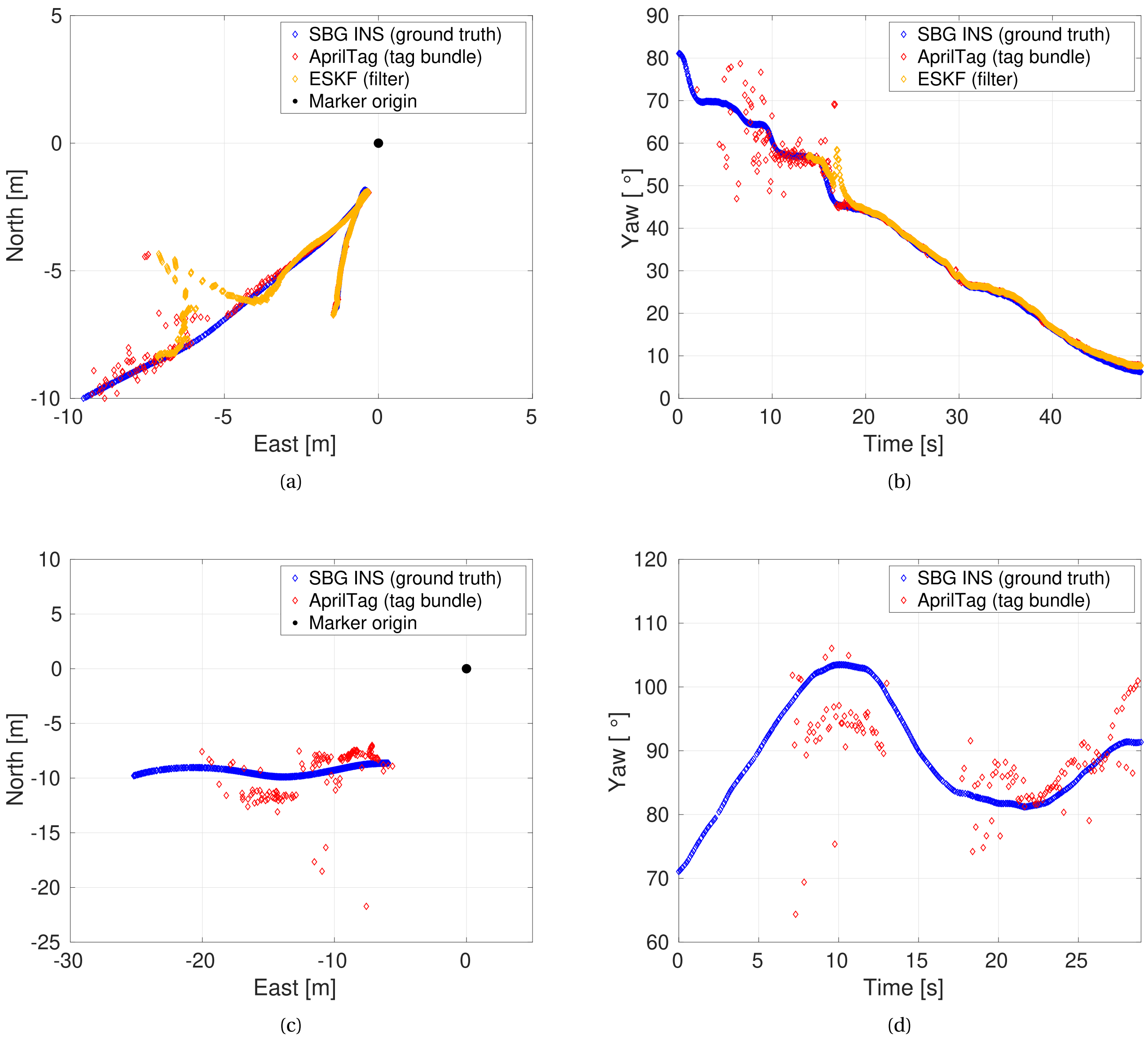


FIGURE 10. The results from Experiment 2.1. (a)-(b) Estimated position and heading of the vehicle using AprilTag and the error-state Kalman filter (ESKF) under partly degraded conditions (sunlight). (c)-(d) Estimated position and heading of the vehicle using AprilTag under significantly degraded conditions (sunlight).

accurate estimates throughout the entire dock-then-undock sequence. Figs. 12c and 12d show the position and heading of the USV, estimated by AprilTag and compared to RTK GNSS in a new scenario influenced by significantly degraded conditions due to fog. This time, it was less wind, resulting in a more concentrated amount of fog around the harbor, as seen in Fig. 9f. As a result, the visual fiducial system was less resilient in recognizing the tags covered by thick fog, resulting in downgraded performance by the AprilTag system.

4) 2.4: RAIN

The results from Experiment 2.4 are shown in Fig. 13. Figs. 13a and 13b show the position and heading of the USV, estimated by AprilTag and the ESKF, and compared to RTK GNSS in partly degraded conditions influenced by raindrops that cover the camera lens. This resulted in a blurry and slightly distorted camera view, as seen in Fig. 9g. The AprilTag system, however, produced very accurate measurements. Consequently, the filter also performed very well. Figs. 13c and 13d show the position and heading of the USV, estimated by AprilTag and compared to RTK GNSS in a new scenario influenced by significantly degraded conditions due to heavy raindrops occluding the camera view. This resulted in a heavily distorted and blurry camera view, making the tags non-recognizable, as seen in Fig. 9h. Remarkably, the AprilTag system did not produce any measurements except for a few measurements at the start and the end of the scenario.

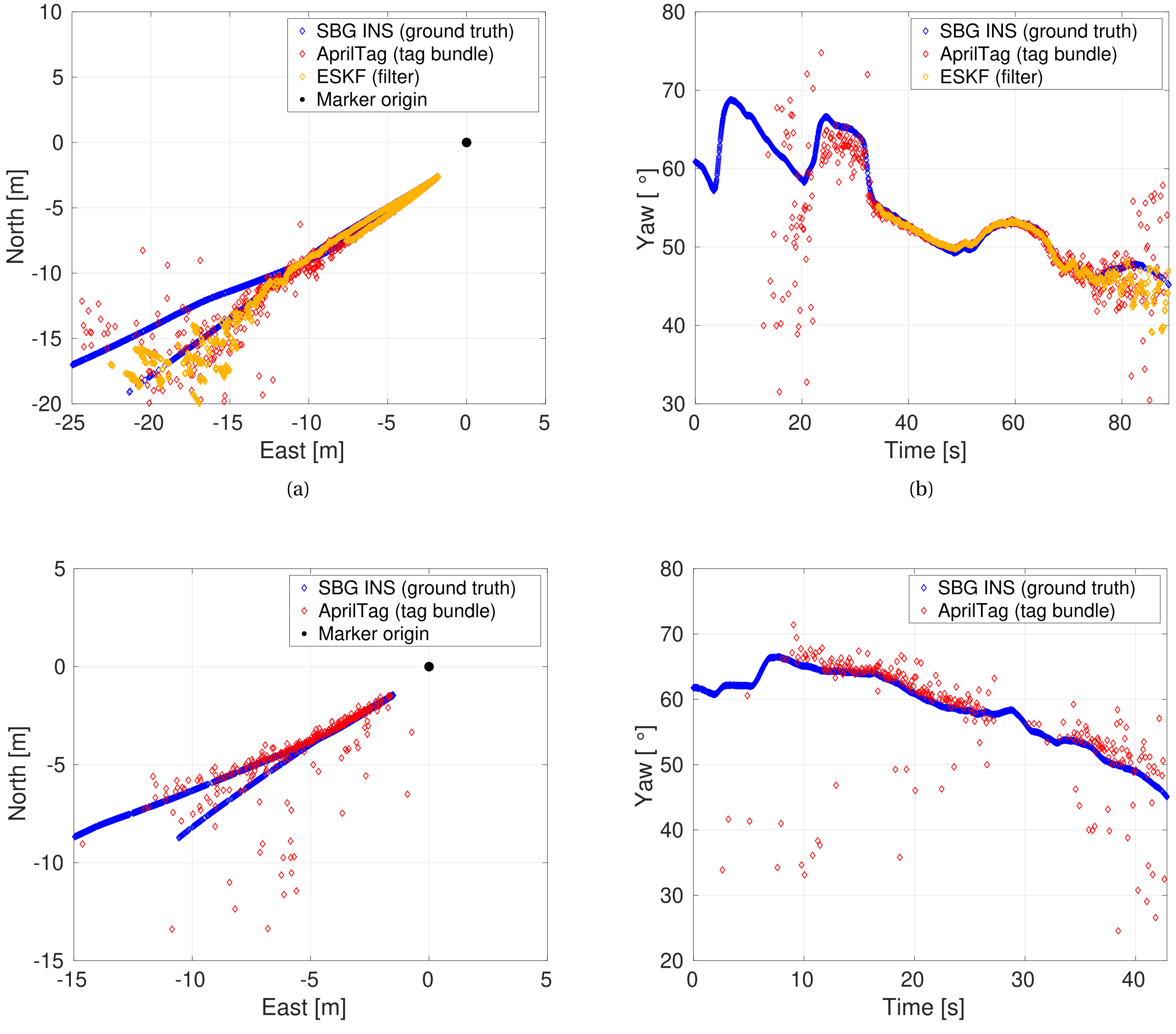


FIGURE 11. The results from Experiment 2.2. (a)-(b) Estimated position and heading of the vehicle using AprilTag and the error-state Kalman filter (ESKF) under partly degraded conditions (darkness). (c)-(d) Estimated position and heading of the vehicle using AprilTag under significantly degraded conditions (darkness).

C. EXPERIMENT 3: FIELD VERIFICATION

The results from Experiment 3 are shown in Fig. 14. Figs. 14a and 14b show the estimated position and heading of the USV, which is compared to RTK GNSS. The vehicle is heading from its initial position (WP1) to the target position (WP2), close to the reference tag. Initially, the filter use RTK GNSS signals from the SBG INS for safe initialization. The filter accepts the visual measurements if the estimated state and AprilTag do not deviate more than 2 degrees and 1.5 m in heading and position, respectively, compared to RTK GNSS. Moreover, five consecutive measurements have to fulfill this criterion before the vehicle switches navigation source. As a result, the USV is aided by visual measurements when it is less than 23 m from the reference tag. As seen in Fig. 14a, the estimated position starts to oscillate immediately. Additionally, the initial jump in position and heading put the vehicle slightly off course relative to the desired path. However, as the vehicle approaches the dock, the estimated position gets closer to the GNSS position and is less oscillating. The mission is succeeded when the vehicle is less than 2.5 m from WP2.

D. DISCUSSION OF RESULTS

In Experiment 1, we found that a tag configuration consisting of three coplanar AprilTags outperforms a single-tag configuration. This was particularly revealing when the vehicle was far from the dock, in which the tag is small and low-resolution. In such situations, the projection of the

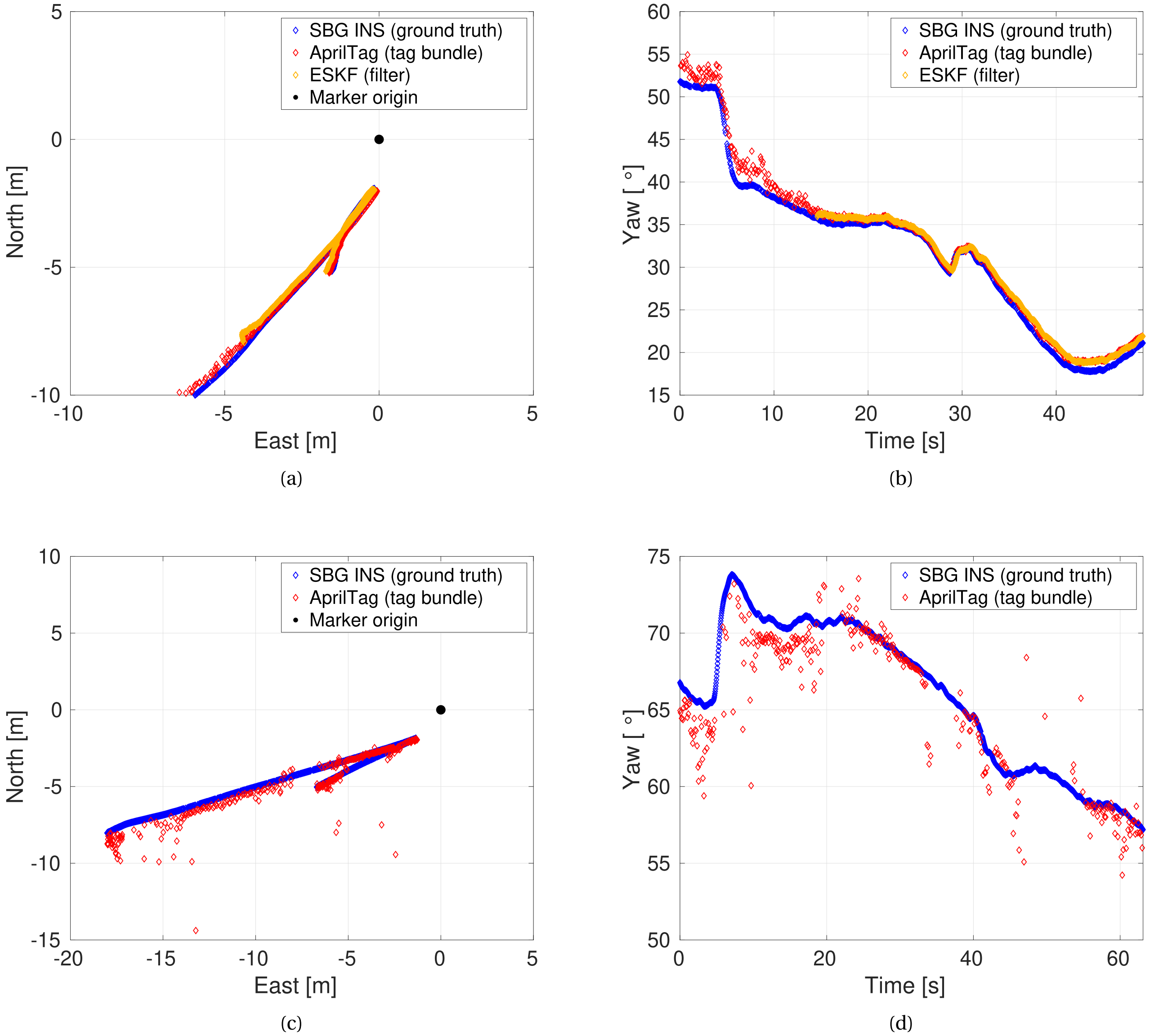


FIGURE 12. The results from Experiment 2.3. (a)-(b) Estimated position and heading of the vehicle using AprilTag and the error-state Kalman filter (ESKF) under partly degraded conditions (fog). (c)-(d) Estimated position and heading of the vehicle using AprilTag under significantly degraded conditions (fog).

object is close to affine, thus leading to a flip ambiguity if only one tag is used [31]. As a result, the PnP solver will return the wrong solution approximately 50% of the time, as seen in Figs. 8a and 8b. To overcome the flip ambiguity, we use a set of three tags that all lie on the same plane. Since we know the relative position and orientation offsets, we can obtain more image-point correspondences directly to extract a unique and robust solution from the PnP solver. However, the set of tags must span a sufficiently large region to prevent them from being flipped, thus producing two valid solutions. Therefore, the tags are 0.412 m × 0.412 m large and have a position offset of approximately 1 m to each other. We emphasize that the tag bundle configuration may suffer from flip ambiguity when multiple tags are not detected simultaneously in a single frame. This typically happens at large camera-tag distances or in weak-perspective conditions. Although the proposed tag configuration shows promising results, there are several possibilities for improvements. For example, non-coplanar markers can resolve the ambiguity directly since the PnP problem no longer involves a planar model. Furthermore, we can place the tags along multiple axes, e.g., in a triangle in the x-y tag coordinate system, to exploit other tag bundle geometries. Finally, it is possible to design a recursive tag system, i.e., a small tag inside a larger tag, to cover longer distances, as shown by Romero-Ramire et al. [51].

In Experiment 2, we experienced that each type of adverse weather poses different challenges. For example, the Electro-Optical (EO) camera was sensitive to sunlight, although supported by a circular polarizing filter to resist glare and

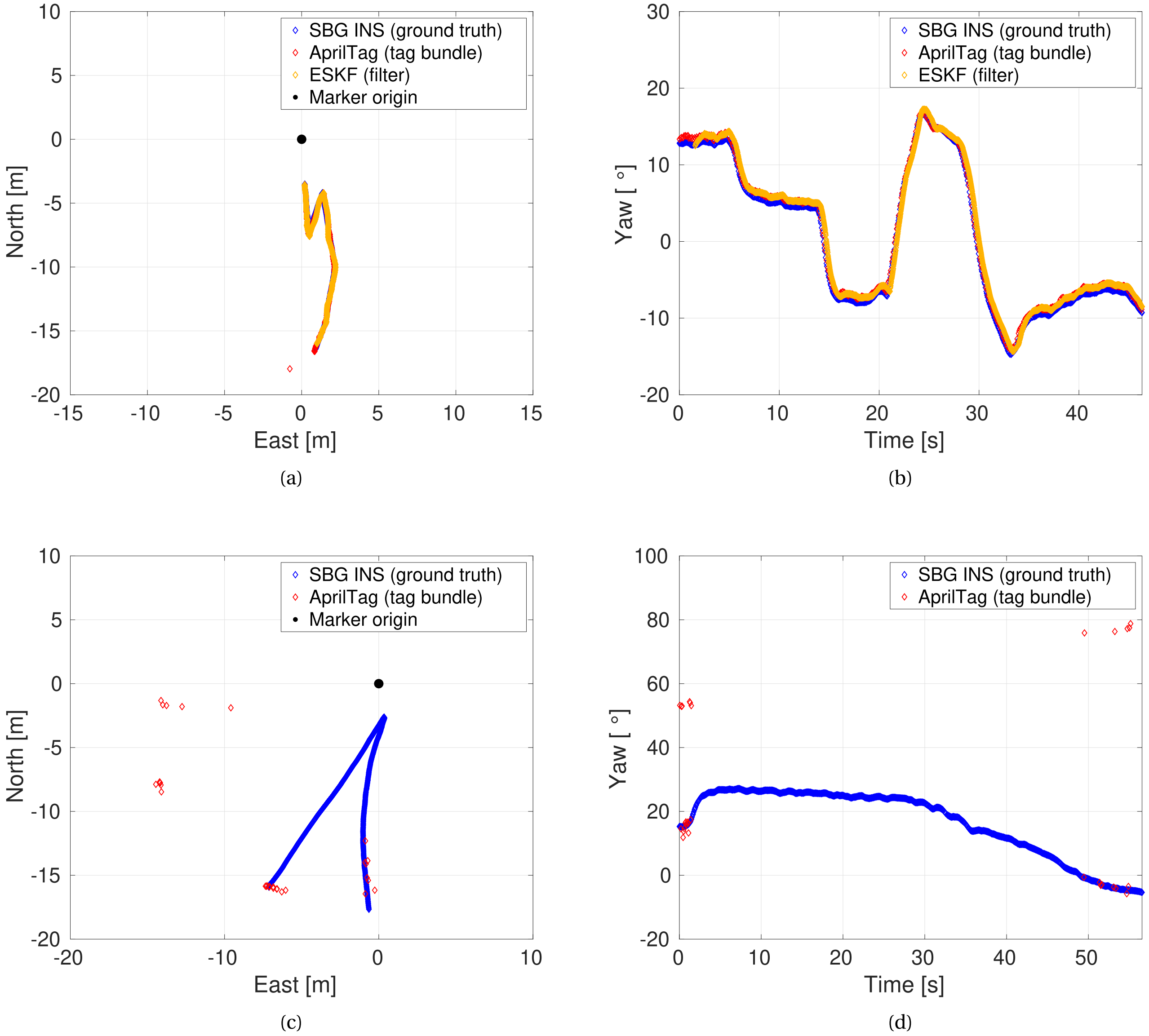
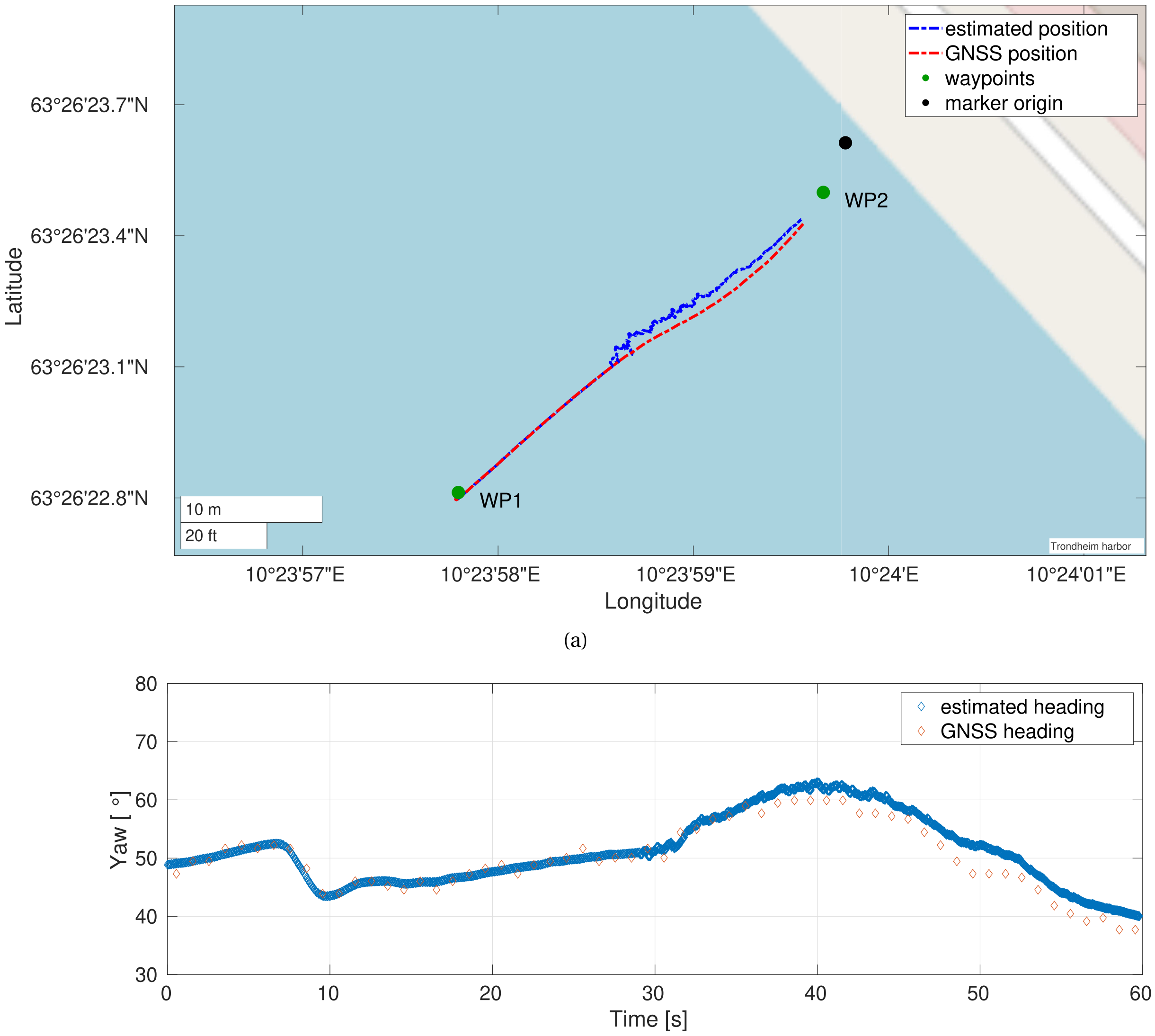


FIGURE 13. The results from Experiment 2.4. (a)-(b) Estimated position and heading of the vehicle using AprilTag and the error-state Kalman filter (ESKF) under partly degraded conditions (rain). (c)-(d) Estimated position and heading of the vehicle using AprilTag under significantly degraded conditions (rain).

reflections. As a result, the visual scene was changed rapidly and caused poor performance for short periods, as seen in Figs. 10a and 10b. Contrary to sunlight, darkness does not change the visual scene rapidly, and the performance is more predictable for a given docking sequence. Nevertheless, both types of visual degradation are challenging to deal with using EO cameras because of limiting contrast ratios. High dynamic range cameras are often applied to cope with such problems, thus capturing more details in both low-light and bright conditions. In complete darkness, infrared (IR) cameras can be used to recognize the tags. If the tags emit uniform thermal energy different from the temperature around, the tags can be detected by heat signatures in the IR spectrum of wavelengths. Moreover, thermal cameras in the mid-wave or longwave infrared band can penetrate fog. IR cameras are, however, expensive compared to proximity EO cameras. Furthermore, they typically have lower resolution, e.g., 640 × 512 resolution, which leads to low detection performance at longer distances. Regarding the rainy situations, we noted that the performance of the AprilTag system was drastically reduced when heavy raindrops covered the camera lens. We emphasize that these scenarios were provoked by physically touching the camera lens with wet fingertips, and rainy weather during other experimental testing did not cause similar degradation of the camera view. In addition, equipment such as wipers and air blowers can remove the raindrops completely. As such, we consider low-light and bright conditions more significant problems to face. In addition, fog could also be a significant problem in geographical areas exposed to that.



(a)

(b)

FIGURE 14. **The results from Experiment 3. (a)-(b) The estimated position and heading of the vehicle under the terminal docking phase, i.e., from the initial position (WP1) to the target position (WP2). The visual-inertial navigation system is activated when the vehicle is approximately 23 m from the reference marker. The true Global Navigation Satellite Systems (GNSS) position and heading from the redundant navigation system are also plotted for comparison.**

In Experiment 3, we demonstrated through field verification that the proposed filter, based on visual tags and inertial data, can be employed for automatic docking of a USV in feedback control. We experienced that measuring the tag location with centimeter accuracy in a global frame and finding the exact angle offset between the x-axis of the tag and true north was practically hard. As a result, we increased the margins to allow the estimated state to deviate more from the true heading and position of the vehicle. Not surprisingly, this resulted in a jump, followed by oscillating estimates of position and heading when switching from GNSS to the camera, as seen in Fig. 14. Nevertheless, the Kalman filter almost converged to the true state when it approached the dock. That being said, we believe the switching strategy can be improved, especially since we assume perfect GNSS measurements using RTK, to which the camera measurements are compared. Hence, if RTK is not available, giving less accurate measurements, the switching criterion will no longer work. To cope with situations where ground truth measurements are not available, we can instead identify outliers among the estimates by checking that the innovations of the filter are *consistent*. By monitoring the innovations of the filter, i.e., the difference between the measurement and its prediction, the consistency check assesses whether the measured innovations follow their expected statistical properties, such as zero mean and autocorrelation.

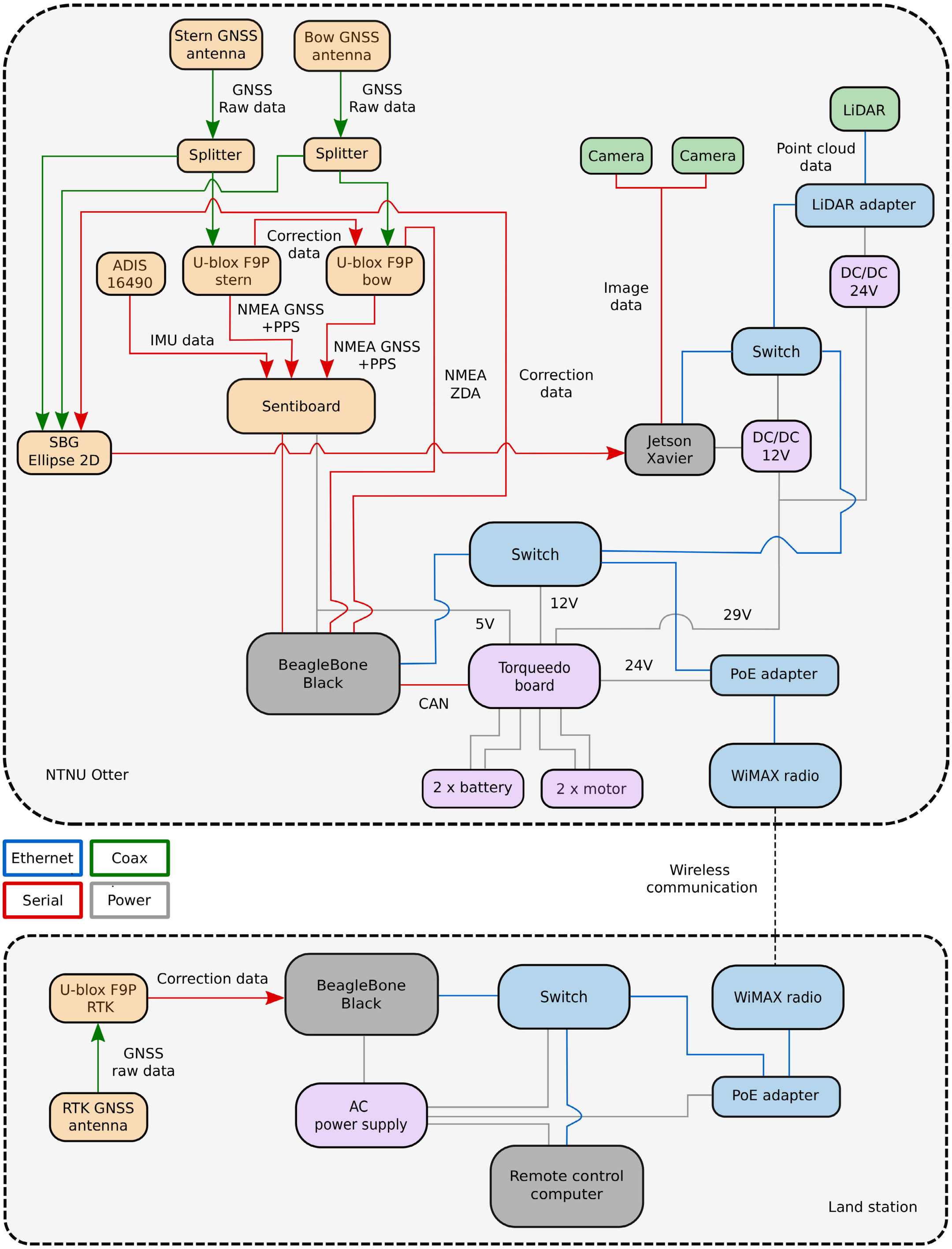


FIGURE 15. Hardware schematic of the NTNU Otter unmanned surface vehicle and the land station.

V. CONCLUSION

Researchers and classification societies have raised concerns regarding the need for an independent, GNSS-free navigation system to improve accuracy and redundancy in the terminal docking phase for maritime vehicles. To this end, we demonstrate how a visual-inertial navigation system aided by fiducial tags can be used for high-precision docking of USVs in this paper. Concerning the tag system design, we found the multi-tag configuration to outperform the single-tag configuration in terms of positioning and heading accuracy. However, the multi-tag configuration may still suffer from flip ambiguity if the camera-tag distance is high or if the visual environment is degraded due to adverse weather or motion blur. As a result, we only recommend using the tag measurements from the multi-tag configuration if two or more tags are detected simultaneously in a single frame. In adverse weather leading to partly degraded visibility, we found the visual fiducial system to perform satisfactorily in terms of positioning and heading accuracy. However, the AprilTag system performed much worse when the visual sight was significantly degraded. Hence, if EO cameras are used, we only suggest employing the proposed system for automatic docking under normal weather conditions or in partly degraded conditions comparable to the scenarios in Experiment 2. Through field verification, we also experienced that the conversion between local NED coordinates and global WGS-84 coordinates, either to express the tag measurements in global latitude-longitude coordinates or to express GNSS measurements in NED coordinates, can be problematic. For example, the AprilTag measurements were subject to an offset when transformed to global coordinates for feedback control purposes. Similarly, the GNSS measurements were slightly inaccurate compared to the AprilTag measurements when transformed to NED coordinates, thus affecting the switching strategy negatively. As such, we recommend measuring the reference tag and the associated angle offset accurately to minimize the errors induced by the coordinate transformations.

In this work, we have described how the proposed visual-inertial navigation system can be used for automatic docking, given that a predetermined path exists with the camera pointing towards the landmarks. However, the camera will no longer obtain navigation information if the landmarks are outside the camera's field of view (FOV). In future work, we plan to overcome this limitation by implementing path planning algorithms and guidance control laws that allow the vehicle to preserve landmarks inside the FOV while at the same time converging to the desired path, similar to Sans-Muntandas et al. [52]. We also plan to adaptively incorporate uncertainties from the landmark observations in the measurement model rather than assuming the same amount of additive gaussian noise for all aiding measurements, i.e., from GNSS and the camera. Finally, we plan to exploit new tag bundle geometries to improve the performance of the tag system and complement the sensor suite with other sensing technologies to enhance perception in adverse weather.

APPENDIX

See hardware schematic in Fig. 15.

ACKNOWLEDGMENT

The authors would like to thank Petter Solnør for his assistance during the experimental verification.

REFERENCES

[1] T. Relling, M. Lützhöft, R. Ostnes, and H. P. Hildre, "A human perspective on maritime autonomy," in *Augmented Cognition: Users and Contexts*, D. D. Schmorrow and C. M. Fidopiastis, Eds. Cham, Switzerland: Springer, 2018, pp. 350–362.

[2] L. Kretschmann, H.-C. Burmeister, and C. Jahn, "Analyzing the economic benefit of unmanned autonomous ships: An exploratory cost-comparison between an autonomous and a conventional bulk carrier," *Res. Transp. Bus. Manage.*, vol. 25, pp. 76–86, Dec. 2017.

[3] V. Bolbot, G. Theotokatos, E. Boulougouris, and D. Vassalos, "A novel cyber-risk assessment method for ship systems," *Saf. Sci.*, vol. 131, Nov. 2020, Art. no. 104908.

[4] *Autonomous and Remotely Operated Ships*, DNV, Bærum, Norway, 2018. Accessed: Jun. 3, 2022.

[5] M. O. A. Aqel, M. H. Marhaban, M. I. Saripan, and N. B. Ismail, "Review of visual odometry: Types, approaches, challenges, and applications," *SpringerPlus*, vol. 5, no. 1, pp. 1–26, Dec. 2016.

[6] L. S. Monteiro, T. Moore, and C. Hill, "What is the accuracy of DGPS?" *J. Navigat.*, vol. 58, no. 2, pp. 207–225, 2005.

[7] *Do You Need RTK for Your Geo-MMS UAV Mapping System?* Geodetics, San Diego, CA, USA, 2022. Accessed: May 25, 2022.

[8] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 834–849.

[9] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[10] S.-H. Zhong, Y. Liu, and Q.-C. Chen, "Visual orientation inhomogeneity based scale-invariant feature transform," *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5658–5667, Aug. 2015.

[11] F.-E. Ababsa and M. Mallem, "Robust camera pose estimation using 2D fiducials tracking for real-time augmented reality systems," in *Proc. ACM SIGGRAPH Int. Conf. Virtual Reality Continuum Appl. Ind. (VRCAI)*. New York, NY, USA: Association for Computing Machinery, 2004, pp. 431–435.

[12] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3400–3407.

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

[14] V. Mondéjar-Guerra, S. Garrido-Jurado, R. Muñoz-Salinas, M. J. Marín-Jiménez, and R. Medina-Carnicer, "Robust identification of fiducial markers in challenging conditions," *Expert Syst. Appl.*, vol. 93, pp. 336–345, Mar. 2018.

[15] M. Fiala, "ARTag, a fiducial marker system using digital techniques," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 590–596.

[16] H. Kato and M. Billinghurst, "Marker tracking and HMD calibration for a video-based augmented reality conferencing system," in *Proc. 2nd IEEE ACM Int. Workshop Augmented Reality (IWAR)*, Feb. 1999, pp. 85–94.

[17] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4193–4198.

[18] R. G. Brown and P. Y. C. Hwang," *Introduction to Random Signals and Applied Kalman Filtering: With MATLAB Exercises and Solutions*, 3rd ed. New York, NY, USA: Wiley, 1997.

[19] J. A. Farell and M. Barth, *The Global Positioning System and Inertial Navigation*. New York, NY, USA: McGraw-Hill, 1998.

[20] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[21] S. K. Biswas, L. Qiao, and A. G. Dempster, "Computationally efficient unscented Kalman filtering techniques for launch vehicle navigation using a space-borne GPS receiver," in *Proc. 29th Int. Tech. Meeting Satell. Division Inst. Navigat. (ION GNSS)*, Nov. 2016, pp. 186–194.

[22] J. Solà, "Quaternion kinematics for the error-state Kalman filter," 2017, *arXiv:1711.02508*.

[23] F. L. Markley and J. L. Crassidis, *Fundamentals of Spacecraft Attitude Determination and Control*, vol. 1286. New York, NY, USA: Springer, 2014.

[24] J. L. Crassidis, F. L. Markley, and Y. Cheng, "Survey of nonlinear attitude estimation methods," *J. Guid., Control, Dyn.*, vol. 30, no. 1, pp. 12–28, Jan. 2007.

[25] T. I. Fossen, *Handbook of Marine Craft Hydrodynamics and Motion Control*, 2nd ed. Hoboken, NJ, USA: Wiley, 2021.

[26] M. Kalaitzakis, B. Cain, S. Carroll, A. Ambrosi, C. Whitehead, and N. Vitzilaios, "Fiducial markers for pose estimation," *J. Intell. Robot. Syst.*, vol. 101, no. 4, pp. 1–26, Apr. 2021.

[27] M. Myint, K. Yonemori, K. N. Lwin, A. Yanou, and M. Minami, "Dual-eyes vision-based docking system for autonomous underwater vehicle: An approach and experiments," *J. Intell. Robot. Syst.*, vol. 92, no. 1, pp. 159–186, Sep. 2018.

[28] H. Y. Hsu, Y. Toda, K. Yamashita, K. Watanabe, M. Sasano, A. Okamoto, S. Inaba, and M. Minami, ''Stereo-vision-based AUV navigation system for resetting the inertial navigation system error,'' *Artif. Life Robot.*, vol. 27, pp. 1–14, Jan. 2022.

[29] J. Chen, C. Sun, and A. Zhang, ''Autonomous navigation for adaptive unmanned underwater vehicles using fiducial markers,'' in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 9298–9304.

[30] P. Trslic, M. Rossi, L. Robinson, C. W. O'Donnel, A. Weir, J. Coleman, J. Riordan, E. Omerdic, G. Dooly, and D. Toal, ''Vision based autonomous docking for work class ROVs,'' *Ocean Eng.*, vol. 196, Jan. 2020, Art. no. 106840.

[31] T. Collins and A. Bartoli, ''Infinitesimal plane-based pose estimation,'' *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 252–286, Sep. 2014.

[32] Ø. Volden, A. Stahl, and T. I. Fossen, ''Vision-based positioning system for auto-docking of unmanned surface vehicles (USVs),'' *Int. J. Intell. Robot. Appl.*, vol. 6, no. 1, pp. 86–103, Mar. 2022.

[33] 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,'' *J. Field Robot.*, vol. 37, no. 1, pp. 137–157, Jan. 2020, doi: 10.1002/rob.21898.

[34] N. Kayhani, W. Zhao, B. McCabe, and A. P. Schoellig, ''Tag-based visual-inertial localization of unmanned aerial vehicles in indoor construction environments using an on-manifold extended Kalman filter,'' *Autom. Construct.*, vol. 135, Mar. 2022, Art. no. 104112.

[35] J. Song, Z. Liu, X. Liu, and J. Guo, ''Tightly coupled visual inertial odometry based on artificial landmarks,'' in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Aug. 2018, pp. 63–70.

[36] ROS. (2022). *ROS—Robot Operating System*. [Online]. Available: https://ros.org/

[37] O. Volden. (2022). *AprilTag State Estimation*. [Online]. Available: https://github.com/oysteinvolden/apriltag_state_estimation

[38] T. I. Fossen and T. Perez. (2004). *Marine Systems Simulator (MSS)*. [Online]. Available: https://github.com/cybergalactic/MSS

[39] D. Malyuta and M. Wolfgang. (2022). *Apriltag_ros*. [Online]. Available: https://github.com/AprilRobotics/apriltag_ros

[40] W. Mularie, ''World geodetic system 1984–its definition and relationships with local geodetic systems,'' Dept. Defense, NIMA, USA, Tech. Rep., 2000. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA280358

[41] J. Farrell, *Aided Navigation: GPS With High Rate Sensors*. New York, NY, USA: McGraw-Hill, 2008.

[42] J. Pinto, P. S. Dias, R. Martins, J. Fortuna, E. Marques, and J. Sousa, ''The LSTS toolchain for networked vehicle systems,'' in *Proc. MTS/IEEE OCEANS*, Jun. 2013, pp. 1–9.

[43] U-Blox. *ZED-F9P Module, 2022*. Accessed: Jun. 4, 2022. [Online]. Available: https://www.u-blox.com/en/product/zed-f9p-module

[44] Analog Devices. (2022). *ADIS16490*. Accessed: Jun. 8, 2022. [Online]. Available: https://www.analog.com/media/en/technical-documentation/data-sheets/adis16490.pdf

[45] SentiSystems. (2021). *SentiPack*. Accessed: Jun. 8, 2022. [Online]. Available: https://sentisolution.com/wp-content/uploads/2020/08/datasheet.pdf

[46] SBG Systems. (2022). *Ellipse Series*. Accessed: Jun. 4, 2022. [Online]. Available: https://www.sbg-systems.com/products/ellipse-series/#ellipse-d_rtk_gnss_ins

[47] Stereolabs. (2022). *ZED 2I*. Accessed: Jul. 13, 2022. [Online]. Available: https://www.stereolabs.com/assets/datasheets/zed-2i-datasheet-feb2022.pdf

[48] W. Caharija, M. Candeloro, K. Y. Pettersen, and A. J. Sørensen, ''Relative velocity control and integral LOS for path following of underactuated surface vessels,'' *IFAC Proc. Volumes*, vol. 45, no. 27, pp. 380–385, 2012.

[49] J. Kridner, G. Coley, and P. J. R. Day. (2022). *Beaglebone Black System Reference Manual*. Accessed: Jun. 4, 2022. [Online]. Available: https://github.com/beagleboard/beaglebone-black/wiki/System-Reference-Manual

[50] Nvidia. (2022). *Jetson AGX Xavier Developer Kit*. Accessed: Jun. 4, 2022. [Online]. Available: https://developer.nvidia.com/embedded/jetson-agx-xavier-developer-kit

[51] F. J. Romero-Ramire, R. Muñoz-Salinas, and R. Medina-Carnicer, ''Fractal markers: A new approach for long-range marker pose estimation under occlusion,'' *IEEE Access*, vol. 7, pp. 169908–169919, 2019.

[52] A. Sans-Muntadas, E. Kelasidi, K. Y. Pettersen, and E. Brekke, ''Path planning and guidance for underactuated vehicles with limited field-of-view,'' *Ocean Eng.*, vol. 174, pp. 84–95, Feb. 2019.

ØYSTEIN VOLDEN received the M.Sc. degree in engineering cybernetics from the Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), in 2020, where he is currently pursuing the Ph.D. degree in engineering cybernetics. He is also affiliated with the NTNU Centre for Autonomous Marine Operations and Systems. He works with topics related to computer vision and state estimation for unmanned surface vehicles.

ANNETTE STAHL received the Ph.D. degree in applied mathematics from Heidelberg University, Germany, with a main focus on computer vision in relation to variational methods for motion estimation using physical prior knowledge. She is currently an Associate Professor and the Head of the Robotic Vision Group, Department of Engineering Cybernetics, Norwegian University of Science and Technology—NTNU, Norway. She is also an affiliated Scientist with the Center of Excellence for Autonomous Marine Operations and Systems—NTNU AMOS and a Scientist with the Centre for Research-based Innovation for Autonomous Ships—SFI AUTOSHIP. She spent two years as a Postdoctoral Researcher with the School of Computing, Dublin City University—DCU, Ireland, and three years with the Department of Mathematical Sciences, NTNU, where she worked on isogeometric analysis based methods for graphics and visualization. After this period, she was a Researcher with the High-Performance Computing Group, NTNU, and SINTEF Ocean, Norway, where she was concerned with computer vision-based aquaculture applications. She is currently working within the field of robotic vision targeting underwater, on the sea surface, on land, in air and space, and indoor and industrial related robotic applications. In 2016, she was awarded an Onsager Fellowship from NTNU's Research Excellence.

THOR I. FOSSEN (Fellow, IEEE) received the M.Sc. degree in marine technology and the Ph.D. degree in engineering cybernetics from the Norwegian University of Science and Technology (NTNU), Trondheim, in 1987 and 1991, respectively. He is currently a naval architect and a cyberneticist. He is also a Guidance, Navigation, and Control Professor with the Department of Engineering Cybernetics, NTNU. His academic background, besides cybernetics, is computer science, cybersecurity, aerospace engineering, marine technology, and inertial navigation systems. He has been a Fulbright Scholar in flight control with the Department of Aeronautics and Astronautics, University of Washington, Seattle. He has authored six textbooks. He is one of the co-founders and the former Vice President of Research and Development with Company Marine Cybernetics, which DNV acquired, in 2012. He is also the Co-Founder of ScoutDI, in 2017. He received the Automatica Prize Paper Award, in 2002, and the Arch T. Colwell Merit Award, in 2008, with the SAE World Congress. He was elected to the Norwegian Academy of Technological Sciences, in 1998, and the Norwegian Academy of Science and Letters, in 2022.