Noname manuscript No.
(will be inserted by the editor)

Tracking of Ocean Surface Objects from Unmanned Aerial Vehicles with a Pan/Tilt Unit using a Thermal Camera

Håkon Hagen Helgesen · Frederik Stendahl Leira · Thor I. Fossen · Tor Arne Johansen

Received: date / Accepted: date

Abstract This paper presents four vision-based tracking system architectures for marine surface objects using a fixed-wing unmanned aerial vehicle (UAV) with a thermal camera mounted in a pan/tilt gimbal. The tracking systems estimate the position and velocity of an object in the North-East (NE) plane, and differ in how the measurement models are defined. The first tracking system measures the position and velocity of the target with georeferencing and optical flow. The states are estimated in a Kalman filter. A Kalman filter is also utilized in the second architecture, but only the georeferenced position is used as a measurement. A bearing-only measurement model is the foundation for the third tracking system, and because the measurement model is nonlinear, an extended Kalman filter is used for state estimation. The fourth tracking system extends the bearing-only tracking system to let navigation uncertainty in the UAV position affect the target estimates in a Schmidt-Kalman filter. All tracking architectures are evaluated on data gathered at a flight experiment near the Azores islands outside of Portugal. The results show that various marine vessels can be tracked quite accurately.

Keywords Unmanned Aerial Vehicle · Machine Vision · Target Tracking · Optical Flow · Georeferencing

Håkon Hagen Helgesen · Frederik Stendahl Leira · Thor I. Fossen · Tor Arne Johansen

NTNU Centre for Autonomous Marine Operations and Systems, Department of Engineering Cybernetics, Norwegian University of Science and Technology, O. S. Bragstads plass 2D, 7491 Trondheim, Norway

E-mail: hakon.helgesen@itk.ntnu.no

1 INTRODUCTION

The use of unmanned aerial vehicles is increasing rapidly and a lot of research is directed towards UAVs. Visual sensors, such as infrared and visual spectrum cameras, are often a part of UAV operations today, and can be useful for navigation [1–5], search and rescue applications [6, 7], sense and avoid technology [8], horizon detection [9, 10], inspection [11], and obviously also in many other applications.

An application where UAVs equipped with a visual sensor can be of use is autonomous ship control. Autonomous ships need to obey the International Regulations for Preventing Collisions at Sea (COLREGS) [12]. The main challenge related to COLREGS is collision avoidance. Therefore, a system for detecting and keeping track of obstacles near the planned path of the ship is required. On-board sensors such as a radar, LIDAR and camera can be used to detect objects in the environment of the ship [13–15]. However, in order to have a robust system it might not be sufficient to only place sensors on-board the ship. Limited range and resolution as well as objects floating in the water-line can make it challenging to detect objects. UAVs can be used to overcome some of the challenges and make a robust system in combination with sensors on-board the ship. By monitoring the planned path of the ship with a UAV, objects that are difficult to find with ship sensors can be located [16]. It is necessary to keep track of the objects for a certain time period to make decisions that obey COLREGS. This is where object detection and tracking become important.

Object detection is the process of detecting objects of importance with respect to some predefined criterion. Tracking is the task of generating a time-dependent position (often also velocity) trajectory for objects detected in a sequence of images. Object detection and tracking have been studied thoroughly and the research on the topics is mature [17]. However, the focus has traditionally been directed towards applications where the sensor is at rest or moving slowly. This is especially the case for segmentation techniques used to find moving objects. Fixed-wing UAVs operate at relatively high velocity, which causes the images captured onboard to be more contaminated by blur than images captured at rest. Moreover, the scene changes rapidly, which makes many conventional detection methods inappropriate for UAVs. Since the UAV may move significantly faster than tracked objects, the accuracy of the UAV navigation system must be carefully considered. Therefore, suitable tracking systems utilizing images captured from a fixed-wing UAV operating at high speed is an interesting research area.

A tracking strategy for vision-based applications onboard UAVs is described in [18], and a correlation method for ship detection with a visual sensor mounted in a UAV is presented in [19]. However, a visual spectrum sensor might not necessarily be the best option for object detection at sea. A thermal camera is a more attractive option since the temperature or emissivity difference between the sea surface and surface objects (such as a marine vessel) is often significant. A thermal camera has successfully been utilized to detect and track objects at sea in [20].

Inspired by the relationship between optical flow and the velocity of a camera mounted in a fixed-wing UAV [1, 21], a vision-based tracking system for marine surface objects utilizing thermal images was presented in [22]. The navigation states of the UAV were used to acquire the NE positions of marine surface objects with a georeferencing technique [23]. Moreover, optical flow was used with the navigation data to recover the NE velocities of the objects. A Kalman filter was utilized to track the objects in the NE plane.

A drawback of the strategy in [22] is that the tracking system depends on accurate measurements (or estimates) of the UAV pose without looking into the navigation uncertainty of the UAV. Thus, the navigation data from the autopilot were assumed to reflect the true states perfectly. This will obviously affect the performance and robustness of the tracking system in situations where the navigation data are unreliable. Another issue is the fact that visual sensors only have bearing measurements and no range measurement. This originates from the fact that only two coordinates in the North-East-Down (NED) reference frame can be acquired with two image coordinates. Therefore, [22] solves this by making the flat-earth assumption, which theoretically makes it possible to compute the range as a function of the UAV attitude and altitude. However, the calculated range is fragile for errors in the navigation states. A tracking system without the range calculation is a viable alternative and leads to the use of a nonlinear tracking filter, such as the extended Kalman filter (EKF).

The complete solution for handling navigation uncertainty and track objects at the same time is equivalent to simultaneous localization and mapping (SLAM) with moving landmarks. Airborne SLAM is discussed in [24] and bearing-only airborne SLAM is the topic in [25] and [26]. In the SLAM context, the target position and velocity are used to correct the navigation states of the UAV. This is useful if absolute sensors, such as GPS, are unavailable. However, erroneous data association or landmark motion that deviates from the motion model (maneuvering targets) can wrongly adjust the UAV navigation states. This is obviously not desired in situations where the navigation states are reliable. Thus, it can be beneficial to only let the uncertainty of the UAV pose affect the target and not the other way around. This mindset leads to the Schmidt-Kalman filter. Target tracking with a Schmidt-Kalman filter is described in [27–29].

1.1 Main Contribution of this Paper

This paper looks into the problem of tracking a single target at sea in thermal images captured with a fixed-wing UAV with a pan/tilt gimbal. The tracking system in [22] is compared with three other alternatives. The first alternative removes the velocity measurement from the tracking system in [22] to investigate its usefulness. A bearing-only measurement model is the foundation for the second alternative. The relative position between the target and the UAV is used in the measurement model, which removes the need to calculate the range. The third alternative extends this system to let navigation uncertainty in the UAV NED positions affect the target estimates in a Schmidt-Kalman filter. The methods are compared thoroughly on data gathered at a flight experiment near the Azores outside of Portugal.

1.2 Organization of this Paper

The remainder of this paper is divided into six sections. Section 2 defines the notations for the derivations in the rest of this paper. Section 3 derives the relationship between optical flow and the NE velocities of a moving target at sea and is based on the work in [22]. The tracking systems are presented in Section 4. Section 5 describes the experiments carried out to gather data. The results are presented in Section 6, before the paper is concluded in Section 7.

2 Notation and Preliminaries

Vectors and matrices are represented by lowercase and uppercase bold letters, respectively. $\mathbf{X}^{-1}$ denotes the inverse of a matrix and $\mathbf{X}^\top$ the transpose of a matrix or vector. A vector $\mathbf{x} = [x_1, x_2, x_3]^\top$ is represented in homogeneous coordinates as $\underline{\mathbf{x}} = [x_1, x_2, x_3, 1]^\top$. The operator $\mathbf{S}(\mathbf{x})$ transforms the vector $\mathbf{x}$ into the skew-symmetric matrix

$$\mathbf{S}(\mathbf{x}) = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix}$$

and $\mathbf{0}_{m\times n}$ is a matrix of zeros with dimension $m \times n$.

Several reference frames are considered in this paper, but the three most important are: the body-fixed frame {B}, the North-East-Down frame {N} (Earth-fixed, considered inertial) and the camera-fixed frame {C}. The rotation from {N} to {B} is represented by the matrix $\mathbf{R}_b^n \in SO(3)$, with $SO(3)$ representing the Special Orthogonal group. Similar transformations exist between the other reference frames.

A vector decomposed in {B}, {N} and {C} has superscript ^{b}, ^{n} and ^{c}, respectively. A point in the environment decomposed in {N} is $\mathbf{t}^n = [x^n, y^n, z^n]^\top$: note that a point located at sea level corresponds to $z^n = 0$. The same point decomposed in {B} is $\mathbf{t}^b = [x^b, y^b, z^b]^\top$.

The Greek letters ϕ, θ, and ψ represent the roll, pitch, and yaw angles, respectively, defined according to the zyx convention for principal rotations [30]. ψ_{gb} and θ_{gb} are the gimbal pan and tilt angles, which correspond to a rotation about the body z- and y-axis, respectively. A 2-dimensional camera image has coordinates (r, s) in the image plane. The derivative $[\dot{r}, \dot{s}]^\top$ of the image plane coordinates is called optical flow. $s\theta$ and $c\theta$ denote the sine and cosine functions with θ as input. The subscript $_f$ is used to indicate that the corresponding parameter is related to a feature (landmark) detected in the image. It should not be mixed with the letter f, which will be used for the focal length of the lens.

3 Machine Vision

This section presents the machine vision system necessary for detecting objects at the sea surface and obtaining measurements that can be used in a tracking system. The first part focuses on optical flow (OF) and how objects are detected in the images. The second part explains how the NED positions of a pixel in the image can be recovered by georeferencing. The third part derives the relationship between OF and the NE velocities of detected objects.

3.1 Optical Flow and Object Detection

Optical flow can be defined as a velocity field that transforms one image into the next in a sequence of images [21] [31]. A single OF vector can be understood as the 2-dimensional displacement (in the image plane) of a feature detected in two consecutive images.

SIFT [32] is a method that can be used to calculate OF by locating scale and rotation invariant features (keypoints) within an image. In practice, it means that features, which change in size and/or orientation with respect to the camera (between two images), can be detected in both images. This is a significant advantage in images captured from a UAV since the scale and rotation of objects change rapidly with the attitude and altitude of the UAV. Another advantage with SIFT (and other point detectors), is the fact that only the current image is used to find features. Thus, a change in background, which must be expected to occur in images captured from a UAV, will not necessarily affect the detection rate. This is not the case for methods relying on some sort of background subtraction/modeling.

Each detected feature gets a descriptor, which is a vector consisting of properties related to the feature. The descriptors are used to find common features in different images through a FLANN nearest neighbor search [33]. OF vectors are calculated as the displacement of common features in two consecutive images.

In this paper, it is assumed that features are only located on the target, such that the mean position and velocity of the features are measures of the position and velocity of the target. This assumption is rarely violated for the images captured in the experiment because objects at sea usually have a strong thermal signature. Moreover, since the sea temperature is constant (homogeneous) it is not likely that features will appear on the sea surface in thermal images, unless an object is present. It is also important to emphasize that the issue is less salient in situations where the number of features on the target significantly exceeds the number of features at other locations. Nevertheless, it is obviously something to be aware of in cases where features are located at other locations or when multiple targets can be present in the images. In these situations, it is necessary to combine SIFT with a method that can locate the area of a target so that only the features of interest are used. An example of a method for extracting the area of a target is presented in [20].

3.2 Recovering the NED Positions of a Pixel

This section seeks to explain how it is possible to acquire NED coordinates of a pixel in the image plane, normally referred to as georeferencing in the literature. Georeferencing is described in [34–36]. The method derived in this paper will be based on [22, 23] because a similar payload setup is utilized.

The pinhole camera model [37] relates a point in the image plane with coordinates decomposed in a camera-fixed coordinate frame {C}. The relationship between the frames is displayed in Fig. 1 and can be described mathematically as

$$\begin{bmatrix} r \\ s \\ 1 \end{bmatrix} = \frac{f}{z^c}\begin{bmatrix} x^c \\ y^c \\ \frac{z^c}{f} \end{bmatrix}, \quad z^c \neq 0 \tag{1}$$

Equation (1) describes the connection between the pixel (r, s) and the camera-fixed coordinates (x^c, y^c, z^c). z^c is the distance between the lens aperture and the plane the captured pixel is located in (range), and f is the focal length of the lens. Equation (1) can be expressed in matrix form as

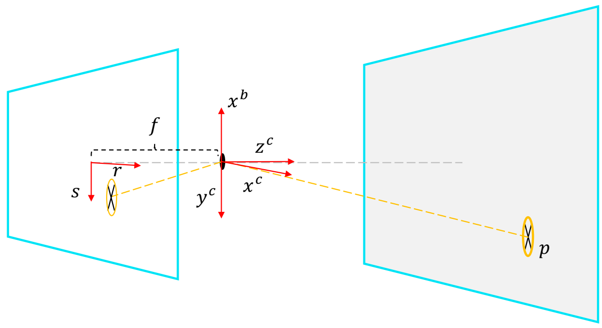

Fig. 1 Illustration of the pinhole camera model. The letter p marks the feature position in NED.

$$z^c \begin{bmatrix} r \\ s \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} x^c \\ y^c \\ z^c \end{bmatrix}}_{\mathbf{p}^c} = \mathbf{A}\mathbf{p}^c \tag{2}$$

where the pixel position (r, s) should be represented in meters for (2) to be valid. It is more useful to decompose $\mathbf{p}^c$ in {N} since the origin of {C} moves with the UAV. It can be achieved by utilizing a transformation $\mathbf{G}^c_n$ between {C} and {N} [23]

$$z^c \begin{bmatrix} r \\ s \\ 1 \end{bmatrix} = \mathbf{A}\mathbf{G}^c_n \underline{\mathbf{p}}^n \tag{3}$$

where $\underline{\mathbf{p}}^n$ is the homogeneous coordinate vector of the pixel decomposed in {N}. $\mathbf{G}^c_n$ is defined as

$$\mathbf{G}^c_n := \begin{bmatrix} \mathbf{R}^c_n & -\mathbf{R}^c_n \mathbf{r}^n_{nc} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1 \; \mathbf{r}_2 \; \mathbf{r}_3 & -\mathbf{R}^c_n \mathbf{r}^n_{nc} \end{bmatrix}$$

where $\mathbf{R}^c_n$ is the rotation matrix between {C} and {N}, with column vectors $\mathbf{r_1}$, $\mathbf{r_2}$ and $\mathbf{r_3}$, and $\mathbf{r}^n_{nc}$ is the the position of the origin of {C} relative to {N} decomposed in {N}.

The rotation matrix $\mathbf{R}^c_n$ can be expressed as

$$\mathbf{R}^c_n = (\mathbf{R}^n_b \mathbf{R}^b_c)^{-1} = (\mathbf{R}^n_b (\mathbf{R}^c_m \mathbf{R}^m_b)^{-1})^{-1} \tag{4}$$

where {m} is referred to as the mounted frame. $\mathbf{R}^n_b$ is the well known rotation matrix between {N} and {B}, defined according to the zyx convention and specified in terms of the Euler angles (roll (ϕ), pitch (θ), yaw (ψ)) [30]. The rotation between {B} and {m} is given by the gimbal orientation. {B} is aligned with {m} when the gimbal has zero pan (ψ_{gb}) and tilt (θ_{gb}). In the body-fixed frame, the pan and tilt movement correspond to a rotation along the body z- and y-axis, respectively. Hence, the rotation is defined as

$$\begin{aligned} \mathbf{R}^m_b &= (\mathbf{R}_z(\psi_{gb})\mathbf{R}_y(\theta_{gb}))^\top = \mathbf{R}_y^\top(\theta_{gb})\mathbf{R}_z^\top(\psi_{gb}) \\ &= \begin{bmatrix} \cos\psi_{gb}\cos\theta_{gb} & \sin\psi_{gb}\cos\theta_{gb} & -\sin\theta_{gb} \\ -\sin\psi_{gb} & \cos\psi_{gb} & 0 \\ \cos\psi_{gb}\sin\theta_{gb} & \sin\psi_{gb}\sin\theta_{gb} & \cos\theta_{gb} \end{bmatrix} \end{aligned} \tag{5}$$

where $\mathbf{R}_z(\alpha)$ and $\mathbf{R}_y(\alpha)$ are principle rotations about the z- and y-axis (by an angle α), respectively [30]. Since the x-axis of {C} should be aligned with the horizontal direction in the image plane (r) and not the body x-axis (Fig. 1), the rotation from {C} to {m} is a rotation of -90 degrees about the camera z-axis:

$$\mathbf{R}^c_m = \mathbf{R}_z(-90^o) = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6}$$

By assuming that the origin of {C} coincides with the origin of {B}, $\mathbf{r}^n_{nc}$ can be simplified as the NED coordinates of the UAV. In practice, for the experiment described in Section 5, the origin of {C} is located within centimeters of the origin of {B}. Therefore, the assumption is reasonable for a UAV at more than tens of meters altitude.

Only two coordinates of the NED positions can be recovered by the pixel coordinates (r, s). However, since objects at the sea surface is of interest, the down position of pixels in the image is close to zero as long as the origin of {N} is placed at sea level. Consequently, one can identify the NE coordinates with the two image coordinates and use zero as the down position. For this to be valid, it is necessary to assume that all pixels in the image are located at sea level (unless a digital elevation map is available) and have a limited height compared to the altitude of the UAV. This is normally referred to as the flat-earth assumption in the literature. In practice, an object height of 10 meters did not degrade the results significantly when the UAV operated at an altitude of 100 meters, but the accuracy obviously decreases with the height of the object.

The NE coordinates of the pixel (r, s) are given by (3) as

$$\begin{bmatrix} N_{obj} \\ E_{obj} \\ 1 \end{bmatrix} = z^c \mathbf{G}_{NE}^{-1} \mathbf{A}^{-1} \begin{bmatrix} r \\ s \\ 1 \end{bmatrix} \tag{7}$$

where $\mathbf{G}_{NE}$ is defined as

$$\mathbf{G}_{NE} := \begin{bmatrix} \mathbf{r}_1 \; \mathbf{r}_2 \; -\mathbf{R}_n^c \mathbf{r}_{nc}^n \end{bmatrix}$$

In order to find the NE coordinates, the range z^c needs to be computed with an altitude measurement. When deriving the relationship between OF and velocity, it will also be necessary to calculate x^c and y^c. These coordinates can be expressed as a function of the UAV navigation states and gimbal orientation. This is explained in Appendix A.

The georeferencing algorithm depends on measurements (or estimates) of the UAV NED positions, the Euler angles (roll, pitch and yaw), the gimbal orientation (pan and tilt angles), the focal length of the lens and the pixel position in the image plane. The accuracy depends on the sensors used to measure or estimate these parameters. The NED positions of the UAV can be measured by GPS, but the down position is not very accurate with single frequency GPS receivers without differential correction. Therefore, an altimeter might be useful in low-altitude applications. The Euler angles can be estimated with an inertial measurement unit (IMU) and some heading reference. The gimbal orientation cannot necessarily be measured, but a set-point should be available. The focal length of the lens is given in the lens specification, but a more accurate estimate of the focal length is obtained with camera calibration [23]. The pixel position is known from the feature extraction.

3.3 Transformation Between Optical Flow and Velocity

This section derives the relationship between OF and velocity. Assume that a feature at pixel position (r, s) is of interest. Differentiation of the pinhole camera model (1) yields

$$\begin{bmatrix} \dot{r} \\ \dot{s} \end{bmatrix} = \frac{1}{z_f^c} \begin{bmatrix} f & 0 & -f\frac{x_f^c}{z_f^c} \\ 0 & f & -f\frac{y_f^c}{z_f^c} \end{bmatrix} \begin{bmatrix} \dot{x}_f^c \\ \dot{y}_f^c \\ \dot{z}_f^c \end{bmatrix} \tag{8}$$

where $[\dot{r}, \dot{s}]^\top$ is the OF vector of the feature. The vector $[\dot{x}_f^c, \dot{y}_f^c, \dot{z}_f^c]^\top$ on the right-hand side is recognized as [21]

$$\dot{\mathbf{p}}_f^c = \begin{bmatrix} \dot{x}_f^c \\ \dot{y}_f^c \\ \dot{z}_f^c \end{bmatrix} = \mathbf{v}_{f/c}^c + \boldsymbol{\omega}_{f/c}^c \times (\mathbf{p}_f^c - \mathbf{o}_f^c) \tag{9}$$

where $\mathbf{v}_{f/c}^c$ and $\boldsymbol{\omega}_{f/c}^c$ are the linear and angular velocities of the feature with respect to {C} decomposed in {C}, respectively. $\mathbf{p}_f^c = [x_f^c, y_f^c, z_f^c]^\top$ is the position of the feature decomposed in {C}. $\mathbf{o}_f^c$ is the feature point of rotation decomposed in {C} such that $(\mathbf{p}_f^c - \mathbf{o}_f^c)$ is the arm of rotation. All rotations seen in the image are rotations about the camera center, hence the rotation point $\mathbf{o}_f^c$ coincides with the origin of {C}. Since it is also assumed that the origin of {C} coincides with {B}, the rotation of features caused by the UAV motion will be about the camera center. Thus, $\mathbf{o}_f^c$ is simply the zero vector.

The assumption of {C} coinciding with {B} has been tested experimentally. It was not possible to find an increase in accuracy when the distance between the origins was accounted for (when the true distance is limited to a few centimeters). Therefore, since the following derivation is simplified with the assumption, it is not accounted for in this paper. In situations where the origin of {C} is far from the origin of {B} one should be aware of the simplification.

Equation (8) might be rewritten by inserting (9):

$$\begin{bmatrix} \dot{r} \\ \dot{s} \end{bmatrix} = \frac{1}{z_f^c} \left[\mathbf{B} \,\middle|\, \mathbf{B} \right] \begin{bmatrix} \mathbf{v}_{f/c}^c \\ \boldsymbol{\omega}_{f/c}^c \times \mathbf{p}_f^c \end{bmatrix} \tag{10}$$

$$\mathbf{B} = \begin{bmatrix} f & 0 & -f\frac{x_f^c}{z_f^c} \\ 0 & f & -f\frac{y_f^c}{z_f^c} \end{bmatrix}$$

By the properties of the crossproduct [22], and using the skew-symmetric matrix $\mathbf{S}$ (defined in Section 2), it is possible to rewrite (10) and establish the relationship between OF and the linear and angular velocities as

$$\begin{bmatrix} \dot{r} \\ \dot{s} \end{bmatrix} = \underbrace{\frac{1}{z_f^c} \left[\mathbf{B} \,\middle|\, -\mathbf{B} \cdot \mathbf{S}(\mathbf{p}_f^c) \right]}_{\mathbf{M}(f, \mathbf{p}_f^c)} \begin{bmatrix} \mathbf{v}_{f/c}^c \\ \boldsymbol{\omega}_{f/c}^c \end{bmatrix} \tag{11}$$

where

$$\mathbf{M}(f, \mathbf{p}_f^c) = \frac{1}{z_f^c} \cdot$$
$$\begin{bmatrix} f, \; 0, \; -f\frac{x_f^c}{z_f^c}, & -f\frac{x_f^c}{z_f^c} y_f^c, & f z_f^c + f\frac{x_f^c}{z_f^c} x_f^c, & -f y_f^c \\ 0, \; f, \; -f\frac{y_f^c}{z_f^c}, & -f z_f^c - f\frac{y_f^c}{z_f^c} y_f^c, & f\frac{y_f^c}{z_f^c} x_f^c, & f x_f^c \end{bmatrix}$$

If the velocities of the UAV decomposed in {B} is known, it is possible to find the OF caused by the camera (UAV) motion. It will from now on be referred to as the theoretical flow $[\dot{r}_t, \dot{s}_t]^\top$, which is defined as

$$\begin{bmatrix} \dot{r}_t \\ \dot{s}_t \end{bmatrix} := \mathbf{M}(f, \mathbf{p}_f^c) \begin{bmatrix} \mathbf{v}_{n/c}^c \\ \boldsymbol{\omega}_{n/c}^c \end{bmatrix} \tag{12}$$

where $\mathbf{v}_{n/c}^c$ and $\boldsymbol{\omega}_{n/c}^c$ are the linear and angular velocities of the sea surface (NED) with respect to {C} decomposed in {C}. Waves are not considered as a part of the velocity and act like a disturbance to the system. Since the origin of {C} coincides with {B}, both {B} and {C} have the same linear velocity with respect to {N}. Therefore, it can be rewritten as

$$\mathbf{v}_{n/c}^c = \mathbf{R}_b^c \mathbf{v}_{n/c}^b = \mathbf{R}_b^c \mathbf{v}_{n/b}^b = -\mathbf{R}_b^c \mathbf{v}_{b/n}^b$$

where $\mathbf{v}_{b/n}^b$ is the body-fixed linear velocity of the UAV with respect to {N} decomposed in {B}.

The angular velocity can be rewritten as

$$\begin{aligned} \boldsymbol{\omega}_{n/c}^c &= \boldsymbol{\omega}_{n/b}^c + \boldsymbol{\omega}_{b/c}^c \\ &= \mathbf{R}_b^c(\boldsymbol{\omega}_{n/b}^b + \boldsymbol{\omega}_{b/c}^b) \\ &= -\mathbf{R}_b^c(\boldsymbol{\omega}_{b/n}^b + \boldsymbol{\omega}_{c/b}^b) \end{aligned}$$

where $\boldsymbol{\omega}_{b/n}^b$ is the angular velocity of {B} with respect to {N} decomposed in {B}. $\boldsymbol{\omega}_{c/b}^b$ is the angular velocity of {C} with respect to {B} decomposed in {B}. It is given by the gimbal motion and should be accounted for. A pan/tilt gimbal can only rotate about the body z- and y-axis. Thus, $\boldsymbol{\omega}_{c/b}^b$ can be approximated as [38]

$$\begin{aligned} \boldsymbol{\omega}_{c/b}^b &= \boldsymbol{\omega}_z(\dot{\psi}_{gb}) + \mathbf{R}_z(\psi_{gb})\boldsymbol{\omega}_y(\dot{\theta}_{gb}) \\ &= \begin{bmatrix} 0 \\ 0 \\ \dot{\psi}_{gb} \end{bmatrix} + \mathbf{R}_z(\psi_{gb}) \begin{bmatrix} 0 \\ \dot{\theta}_{gb} \\ 0 \end{bmatrix} \end{aligned}$$

where $\dot{\psi}_{gb}$ and $\dot{\theta}_{gb}$ are the derivatives of the pan and tilt angles, respectively. They need to be measured or approximated by e.g. a Taylor-series approximation. A first-order Taylor-series approximation is utilized in this paper.

The theoretical flow can now be calculated with (12). It is still, however, some work needed before the velocity of the feature itself is identified. The OF is a sum of the camera and feature motion with respect to {N}:

$$\begin{bmatrix} \mathbf{v}_{f/c}^c \\ \boldsymbol{\omega}_{f/c}^c \end{bmatrix} = \begin{bmatrix} \mathbf{v}_{f/n}^c + \mathbf{v}_{n/c}^c \\ \boldsymbol{\omega}_{f/n}^c + \boldsymbol{\omega}_{n/c}^c \end{bmatrix} \tag{13}$$

Equation (13) can be inserted into (11) where $\dot{r}_m$ and $\dot{s}_m$ now are defined as the measured OF, obtained with e.g. SIFT. Consequently,

$$\begin{bmatrix} \dot{r}_m \\ \dot{s}_m \end{bmatrix} = \mathbf{M}(f, \mathbf{p}_f^c) \begin{bmatrix} \mathbf{v}_{f/n}^c \\ \boldsymbol{\omega}_{f/n}^c \end{bmatrix} + \mathbf{M}(f, \mathbf{p}_f^c) \begin{bmatrix} \mathbf{v}_{n/c}^c \\ \boldsymbol{\omega}_{n/c}^c \end{bmatrix} \tag{14}$$

where the second term is recognized as the theoretical flow. Thus, it is possible to rewrite (14) as

$$\begin{bmatrix} \dot{r}_m - \dot{r}_t \\ \dot{s}_m - \dot{s}_t \end{bmatrix} = \mathbf{M}(f, \mathbf{p}_f^c) \begin{bmatrix} \mathbf{R}_n^c \mathbf{v}_{f/n}^n \\ \mathbf{R}_n^c \boldsymbol{\omega}_{f/n}^n \end{bmatrix} \tag{15}$$

Equation (15) only has two terms on the left side and six unknown velocity parameters on the right side. However, since the main motivation is to locate surface objects at sea, the angular velocity of the features located on the objects is assumed to be zero (constant object heading between successive images). Therefore, the final three columns of $\mathbf{M}(f, \mathbf{p}_f^c)$ disappears and (15) can be further simplified. Consequently,

$$\begin{bmatrix} \dot{r}_m - \dot{r}_t \\ \dot{s}_m - \dot{s}_t \end{bmatrix} = \frac{1}{z_f^c} \begin{bmatrix} f & 0 & -f\frac{x_f^c}{z_f^c} \\ 0 & f & -f\frac{y_f^c}{z_f^c} \end{bmatrix} \mathbf{R}_n^c \mathbf{v}_{f/n}^n \tag{16}$$

In addition, since the down velocity is expected to be zero, the third column of $\mathbf{R}_n^c$ can be discarded in (16) and the NE velocities ($v_{f/n}^N$ and $v_{f/n}^E$) of the feature can be calculated as

$$\begin{bmatrix} v_{f/n}^N \\ v_{f/n}^E \end{bmatrix} = \left(\frac{1}{z_f^c} \begin{bmatrix} f & 0 & -f\frac{x_f^c}{z_f^c} \\ 0 & f & -f\frac{y_f^c}{z_f^c} \end{bmatrix} [\mathbf{r}_1 \; \mathbf{r}_2] \right)^{-1} \begin{bmatrix} \dot{r}_m - \dot{r}_t \\ \dot{s}_m - \dot{s}_t \end{bmatrix} \tag{17}$$

The NE velocities of the feature are now identified. The velocity calculation depends on measurements of the parameters described at the end of Section 3.2 (for georeferencing). Additionally, the body-fixed linear and angular velocities of the UAV and an approximation of the pan and tilt angle derivatives need to be obtained. The linear velocity can be measured by GPS (when the attitude is known) and the angular velocity can be measured by gyros. The pan and tilt angles are approximated with a first-order Taylor-series approximation.

4 Target Tracking

This section presents four different architectures for tracking of marine surface objects. A single target is of interest and problems related to multiple target tracking, such as data association, are not within the scope of this paper. Measurements of the object position in the image plane and OF vectors are assumed available, and this section looks into how these measurements can be utilized in a tracking system. The first part presents the motion model. The rest of this section describes the tracking architectures.

4.1 Target Motion Model

The goal in target tracking is to estimate the position and velocity of an object of interest. A motion model for the target is required in order to use e.g. a Kalman filter for state estimation. How to choose a motion model is described in [39]. In this paper, a constant velocity model (white noise acceleration) is chosen. This is because the dynamics of typical surface objects are assumed to be slow. The position and velocity in the NE plane are of interest. The discrete time constant velocity motion model at time step $[k]$ is defined as

$$\mathbf{x}^t[k+1] = \mathbf{F}^t\mathbf{x}^t[k] + \mathbf{E}^t\mathbf{v}^t[k] \tag{18}$$

where $\mathbf{x}^t = [p_t^N, p_t^E, v_t^N, v_t^E]^\top$ is the state vector consisting of the target position and velocity, and $\mathbf{v}^t = [v_v^N, v_v^E]^\top$ is assumed to be zero-mean Gaussian white noise with covariance $\mathbf{Q}$. $\mathbf{F}^t$ and $\mathbf{E}^t$ are the system matrices defined as

$$\mathbf{F}^t = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{E}^t = \begin{bmatrix} \frac{1}{2}T^2 & 0 \\ 0 & \frac{1}{2}T^2 \\ T & 0 \\ 0 & T \end{bmatrix}$$

where T is the sampling period of the camera. The down position is zero for surface objects at sea (given that the origin of NED is placed at sea level) and not a part of the state vector. Note that the motion model is linear.

4.2 Tracking System based on Georeferencing and Optical Flow

The first tracking architecture is based on the work conducted in Section 3 and [22]. Georeferencing is used to obtain measurements of the NE positions for features detected on the object. Moreover, OF is used to obtain measurements of the NE velocities for the features. The mean position and velocity of every feature on the target are used as measurements in the tracking system. A Kalman filter can be used since the measurements are equal to the states. The measurement model is

$$\mathbf{z}^t[k] = \mathbf{x}^t[k] + \mathbf{w}^t[k] \tag{19}$$

where $\mathbf{w}^t$ is zero-mean Gaussian white noise with covariance $\mathbf{R}$.

The main advantage with this architecture is that linearization is avoided and that velocity information can be acquired directly. This is usually not the case for tracking systems with a single camera. Furthermore, the approach leads to a completely observable system.

The main drawback is the complexity of the transformation for extracting position and velocity measurements from the pixel position and OF. Additionally, the NED positions, attitude and gimbal orientation have to be accurately known in order to get reliable measurements for position and velocity. An error of just a couple of degrees in roll or pitch will increase the error in position and velocity significantly, especially at larger altitudes. This is because the calculated position and range strongly depend on the attitude. Moreover, it is not straightforward to describe the noise related to the measurements since the ”real” measurements (pixel position and OF) are used in a nonlinear transformation to obtain NE positions and velocities. The nonlinear transformation depends on states that are assumed perfectly known, but in practice all of these quantities will be somewhat uncertain. Thus, it is necessary to make a qualified guess for the uncertainty of the measurements since little is known about the real uncertainty of the parameters in the nonlinear transformation. In other words, this approach sacrifices some robustness in order to make the system linear.

No correlation between the target estimates and the navigation states of the UAV is maintained. In practice, this means that the tracking system works as a standalone system and trusts that the UAV pose is known accurately at all times. This is a major difference from the SLAM approach. A sketch of the system is displayed in Fig. 2.

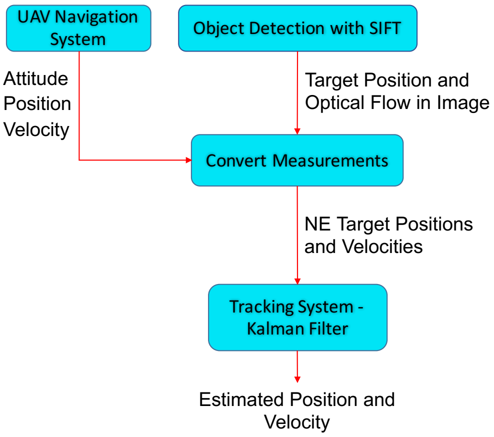
UAV Navigation System
Object Detection with SIFT
Attitude
Position
Velocity
Target Position and Optical Flow in Image
Convert Measurements
NE Target Positions and Velocities
Tracking System - Kalman Filter
Estimated Position and Velocity

Fig. 2 A sketch of the tracking architecture based on georeferencing and optical flow.

4.3 Tracking System based on Georeferencing

The second tracking system is almost equal to the tracking system based on georeferencing and OF. However, the velocity measurement is removed and only the position of the target is used in the Kalman filter. Therefore, the measurement model can be written as

$$\mathbf{z}^t[k] = \underbrace{\begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \end{bmatrix}}_{\mathbf{H}} \mathbf{x}^t[k] + \mathbf{w}^t[k] \tag{20}$$

where $\mathbf{w}^t$ is zero-mean Gaussian white noise with covariance $\mathbf{R}$. The main motivation behind this architecture is to evaluate the usefulness of the velocity measurement. This architecture shares the strengths and weaknesses with the tracking system in Section 4.2.

4.4 Tracking System based on Bearing-only Measurements

A camera makes relative bearing observations to the object in the image [25]. Therefore, it is possible to use the pixel location in the image directly instead of converting the pixel coordinates to NED. Remember that the pinhole camera model (1) relates pixel coordinates to coordinates in {C}. The position of the object decomposed in {C} ($\mathbf{p}^c_{f/c}$) is related to the UAV position ($\mathbf{p}^n_{uav}$) and object position ($\mathbf{p}^n_f$) decomposed in {N}, and the attitude of the UAV through the following model:

$$\mathbf{p}^c_{f/c} = \begin{bmatrix} x^c_f \\ y^c_f \\ z^c_f \end{bmatrix} = \mathbf{R}^c_n(\mathbf{p}^n_f - \mathbf{p}^n_{uav}) \tag{21}$$

Since the pixel coordinates of a feature can be measured by the object detection algorithm, the measurement model can be defined as

$$\begin{aligned} \mathbf{z}^t[k] &= \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} r \\ s \end{bmatrix} \\ &= \underbrace{\frac{f}{z^c_f} \begin{bmatrix} x^c_f \\ y^c_f \end{bmatrix}}_{\text{Insert eq. (21)}} \\ &\equiv \mathbf{h}^t(\mathbf{p}^n_{uav}[k], \mathbf{R}^c_n[k], \mathbf{p}^n_f[k], k) + \mathbf{w}[k] \end{aligned} \tag{22}$$

where $\mathbf{w}$ is zero-mean Gaussian white noise with covariance $\mathbf{R}$. Equation (21) is inserted to get a model depending on the UAV attitude, gimbal orientation, UAV NED positions, and the NED positions of the target. This is beneficial because the measurement model now directly depends on parameters of interest, and not the camera-fixed coordinates of the target.

The measurement model is nonlinear and the most common solution then is to use an Extended Kalman filter (EKF). In this tracking architecture, as for the tracking systems in Section 4.2 and 4.3, the UAV NED positions, attitude and gimbal orientation are assumed perfectly known. However, the need to calculate the range is removed and, thus, a weakness for the tracking system in Section 4.2 and 4.3 is eliminated. In order to use the EKF, it is necessary to find the Jacobian of $\mathbf{z}^t$ with respect to the states. The equations for the Jacobian gets the form

$$\frac{\partial \mathbf{h}^t}{\partial \mathbf{x}^t} = \begin{bmatrix} \frac{\partial z_1}{\partial x^n_f}\big|_{\hat{\mathbf{x}}^t_{k|k-1}} & \frac{\partial z_1}{\partial y^n_f}\big|_{\hat{\mathbf{x}}^t_{k|k-1}} & 0\;0 \\ \frac{\partial z_2}{\partial x^n_f}\big|_{\hat{\mathbf{x}}^t_{k|k-1}} & \frac{\partial z_2}{\partial y^n_f}\big|_{\hat{\mathbf{x}}^t_{k|k-1}} & 0\;0 \end{bmatrix} \tag{23}$$

where $\hat{\mathbf{x}}^t_{k|k-1}$ is the predicted state $\mathbf{x}$ at the current time step. The last two columns of the Jacobian are zero because the measurement model not depends on the target velocities. Note that the motion model for the target is still linear.

The tracking system based on bearings-only measurements are somewhat less simplistic than the tracking system based on georeferencing since it is nonlinear. The problems with linearization and initialization follow with the EKF and the measurements (pixel coordinates) might in many cases be less informative than the NE positions (for humans). However, [40] demonstrates a way to handle the lack of global stability for the EKF with a Double Kalman filter. This solution is especially interesting in applications where initialization and the stability of the EKF are troublesome.

The velocity of the object is not measured in this tracking system because it is impossible to calculate the NE velocities of the object without computing the range. Avoiding the range calculation is one motivation behind this tracking system, and, therefore, it is not an option to calculate the range in order to find the object velocity. Nevertheless, since the position estimates are expected to be more accurate without the range calculation, it is likely that the object velocity is estimated well without the velocity measurements from OF. This is also supported by the fact that the velocity measurement is a differentiation of the georeferenced position measurement, which means that you in practice do not provide the tracking filter with more information. Thus, one could argue that the velocity calculation is more interesting as a measure of the velocity at one time instant and not that useful for estimation.

This architecture is less computationally complicated since the georeferencing operation is not conducted, especially if the Jacobian is evaluated numerically. It may also be less affected by uncertainty in the UAV navigation system since the relative position between the

UAV and the target is used instead of the camera-fixed coordinates of the object, and this is why the range calculation (z_f^c) is avoided. Correlation between the navigation system of the UAV and the tracking system is not maintained in this architecture either. This might be problematic for the same reasons as described at the end of Section 4.2.

The covariance of the measurement noise is easier represented in this architecture. It can be designed as a diagonal matrix with a chosen pixel uncertainty related to each measurement. The tracking architecture is displayed in Fig. 3.

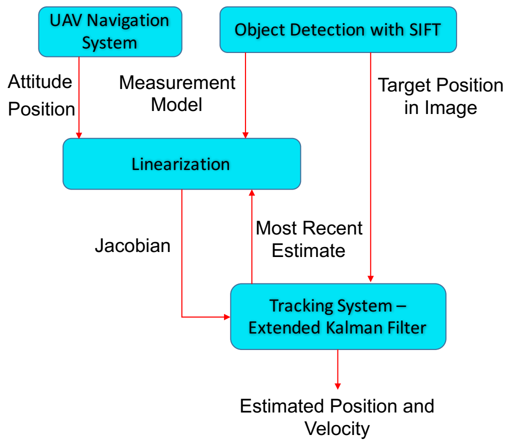

Fig. 3 A sketch of the tracking architecture based on bearing-only measurements.

4.5 Tracking System based on a Schmidt-Kalman Filter

The last target tracking architecture studied in this paper is based on a Schmidt-Kalman filter. A Schmidt-Kalman filter is used to maintain correlations between the target and the UAV without letting the target influence the UAV navigation system. In the SLAM framework, the target could influence the UAV pose, but this is not desired in situations where the UAV pose can be estimated relatively accurately by itself.

In this paper, a simplified version of the Schmidt-Kalman filter will be investigated. The bearing-only measurement model depends on the UAV attitude, gimbal orientation and NED positions. The attitude and the gimbal orientation will still be assumed perfectly known, but the error in the UAV position will be a part of the state-vector. The true NED positions $\mathbf{x}^o$ of the UAV can be written as

$$\mathbf{x}^o = \hat{\mathbf{x}}^o + \delta\mathbf{x}^o \tag{24}$$

where $\hat{\mathbf{x}}^o$ is the nominal state (given by the estimate from the navigation system) and $\delta\mathbf{x}^o$ is the error between the nominal state and the true state. In situations where the nominal state is unbiased, $\delta\mathbf{x}$ will be a zero-mean random variable with uncertainty equal to the uncertainty of the estimate $\hat{\mathbf{x}}^o$. $\hat{\mathbf{x}}^o$ is not considered as a random variable, but rather a true measure of the state $\mathbf{x}^o$. Correlation between the target and the UAV is achieved by augmenting the system (18) with the error state of the UAV NED positions:

$$\begin{bmatrix} \mathbf{x}^t[k+1] \\ \delta\mathbf{x}^o[k+1] \end{bmatrix} = \begin{bmatrix} \mathbf{F}^t & 0 \\ 0 & \mathbf{I}_{3\times3} \end{bmatrix} \begin{bmatrix} \mathbf{x}^t[k] \\ \delta\mathbf{x}^o[k] \end{bmatrix} + \begin{bmatrix} \mathbf{E}^t\mathbf{v}^t[k] \\ \mathbf{v}^o[k] \end{bmatrix} \tag{25}$$

where $\mathbf{I}_{3\times3}$ is the identity matrix of dimension 3×3 and $\mathbf{v}^o$ is white noise affecting the error state with known covariance. Ideally, one should estimate the error-state in an error-state Kalman filter as in [29] and use the corresponding state space model and estimated covariance in (25). However, since inertial sensor data (IMU) are unavailable, the error-state are in this case considered to be constant with a time-invariant known covariance. Therefore, $\mathbf{v}^o[k]$ is assumed to be zero. This is in correspondence with the case described in [28]. The main difference between the structure in this case and [29] is the fact that the covariance increases at each time update until a correction is available in [29]. This is more in line with navigation systems because the covariance of the position estimates is time-variant, and increases when the states are predicted with inertial sensors until the estimates are corrected by e.g. GPS measurements. Nevertheless, a constant covariance is also assessed to accentuate the advantage with the Schmidt-Kalman architecture.

The measurement model is still given by (22), but it is now necessary to evaluate the Jacobian with respect to the UAV NED positions in addition. The new measurement Jacobian gets the form

$$\frac{\partial \mathbf{h}^t}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial z_1}{\partial x_f^n} & \frac{\partial z_1}{\partial y_f^n} & 0 \; 0 & \frac{\partial z_1}{\partial x_{uav}^n} & \frac{\partial z_1}{\partial y_{uav}^n} & \frac{\partial z_1}{\partial z_{uav}^n} \\ \frac{\partial z_2}{\partial x_f^n} & \frac{\partial z_2}{\partial y_f^n} & 0 \; 0 & \frac{\partial z_2}{\partial y_{uav}^n} & \frac{\partial z_2}{\partial y_{uav}^n} & \frac{\partial z_2}{\partial z_{uav}^n} \end{bmatrix} \tag{26}$$

where the partial derivatives are evaluated at the current best estimate for $\mathbf{x}^t$ and $\hat{\mathbf{x}}^o$. The covariance matrix for the augmented system gets the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}^t & \mathbf{P}^{to} \\ (\mathbf{P}^{to})^\top & \mathbf{P}^o \end{bmatrix} \tag{27}$$

where $\mathbf{P}^{to}$ is the cross-covariance between the target and UAV NED positions, and $\mathbf{P}^o$ is the covariance for the UAV NED positions. The equations for the Schmidt-Kalman filter is described thoroughly in [28] and will not be explained further here. It is important, however, to point out that you don't want to estimate or predict the state $\delta\mathbf{x}^o$, but rather account for the uncertainty. Therefore, the Schmidt-Kalman filter forces the corresponding elements in the Kalman gain to zero. Both the Kalman gain and covariance for the target are influenced by $\mathbf{P}^{to}$ and $\mathbf{P}^o$. The tracking architecture is displayed in Fig. 4.

This architecture shares the strengths and weaknesses with the tracking system based on bearing-only measurements. However, the main advantage with this architecture is that the estimated covariance for the target states accounts for uncertainty in the UAV NED positions. Hence, the estimated covariance is expected to reflect the true uncertainty more accurately so that the estimates are more consistent. That can for example be crucial for data association purposes in multiple target tracking.

This approach is obviously not useful in situations where the sensor position and orientation are known perfectly. Moreover, it complicates the system slightly since a direct link between the target tracking system and the UAV navigation system is created. The Schmidt-Kalman filter is also a suboptimal approach since information only are allowed to flow from the UAV navigation system to the tracking system, and not the other way around (as in the SLAM approach).

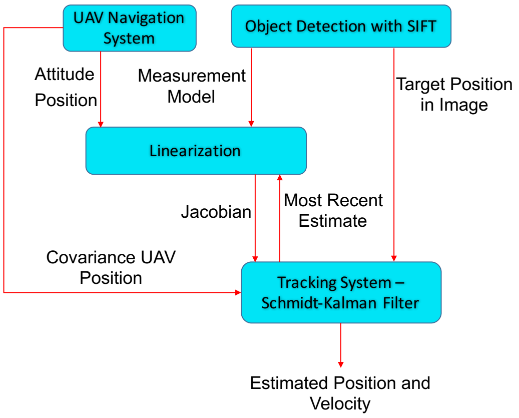

Fig. 4 A sketch of the tracking architecture based on a Schmidt-Kalman filter.

5 Experimental Setup

A flight experiment consisting of several flights has been conducted near the Azores outside of Portugal. The X8 Skywalker fixed-wing UAV interfaced with a retractable pan/tilt gimbal was used to gather data. A light-weight payload [23], with a FLIR Tau2 640 thermal camera with a focal length of 19mm and resolution of 640 × 480 pixels, was used to capture images from the flights. The thermal camera has a frame rate of 7.5 frames per second, and has been calibrated with the method proposed in [23] to increase the accuracy of the camera intrinsic matrix. The experiment was conducted at sea with marine vessels operating in the area. An image captured at the experiment is displayed in Fig. 5.

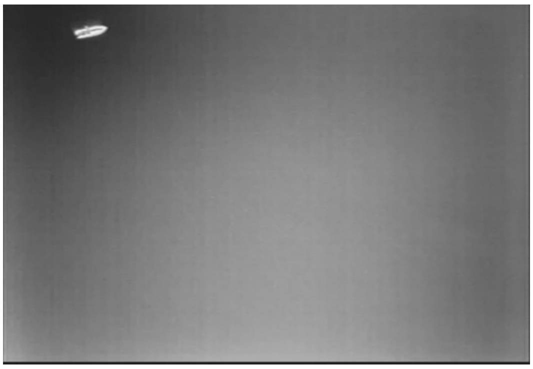

Fig. 5 A thermal image captured at the flight experiment. A small boat is present in the image.

The images and the navigation data gathered at the experiment have been processed off-line. The Open Source Computer Vision Library (OpenCV) [41] has been used to implement SIFT, which is used for feature extraction and OF calculation. Matched features between successive images are assigned a value indicating the uncertainty of the match. Matches with more than twice the uncertainty of the best match have been removed to increase the reliability of the OF vectors. Fig. 6 displays a single image captured in the experiment with OF vectors acquired by SIFT.

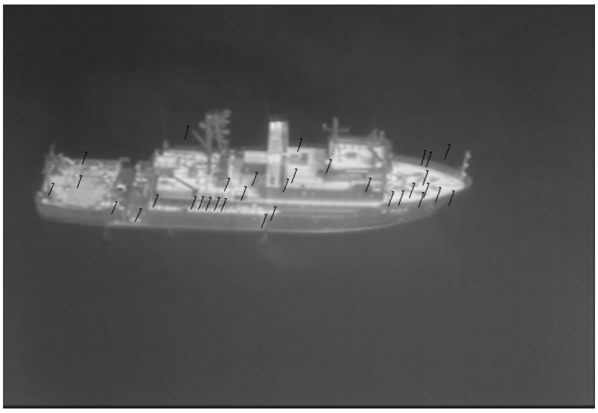

Fig. 6 Optical flow vectors acquired by SIFT on an image captured at the flight experiment.

The data used to evaluate the tracking systems consist of thermal images with the target (the vessels displayed in Fig. 5 and 6), GPS measured position and speed for the vessels (used as a reference for validation) and navigation data for the UAV (estimated by the autopilot). The navigation data are stored with a frequency of 10Hz. The GPS measurements for position and speed of the vessels (target) are stored with a frequency of 2Hz. The mean pixel position for the features are used as a measurement of the target position in the image plane for the tracking systems in Section 4.4 and 4.5. The mean pixel position for all features is used for georeferencing in the tracking system in 4.3. Each feature on the target is treated independently for the tracking system in Section 4.2 because the magnitude of optical and theoretical flow varies with the pixel position. The NE positions and velocities of each feature are calculated, and the mean position and velocity for all features are used as measurements in the tracking system.

The gimbal pan and tilt angles were both controlled manually and automatically with a gimbal controller [42] during the flight. The pan and tilt angles cannot be measured directly and only the commanded set-point is available. Therefore, a possible source of uncertainty in the results is the accuracy of the pan and tilt angles. How this is handled in practice is described more carefully in [22].

The measurement models require time-synchronized data. The images, navigation data and gimbal orientation are stored by the on-board computer. The data are not synchronized in hardware, and thus the time stamp is given by the on-board computer software. Hence, the time stamps can be somewhat uncertain when the on-board computer has a lot of tasks. This is because a delay will be added to when a sensor actually obtained the measurement. Moreover, the GPS receiver and serial communication have a typical delay of 100-200 ms [43]. To reduce the impact of uncertainty in the time stamps, the images have been synchronized off-line by adjusting the time stamp for images where the time between subsequent images differs substantially from the frame rate. Furthermore, the mean time between consecutive images (without any adjustment) was in accordance with the frame rate of the camera. This was also the case for the navigation and gimbal data. Moreover, since it is less time consuming to store navigation data (compared to an image), the time stamps for the navigation and gimbal data were accepted without adjustment between samples.

The tracking systems are implemented in Matlab. Prediction is performed for every received image. Measurements are used to correct the prediction whenever the target is detected in the images. One of the main goals with the tracking systems is to be able to predict the trajectory of the target when measurements are unavailable. Therefore, the experiment contains longer periods where the target is outside of the field of view of the camera. The tracking systems are able to run in real-time on a MacBook Pro (2015 version) with an Intel dual core i7 processor when images arrive at a frequency of 7.5 Hz. A non-optimized implementation of SIFT in OpenCV is used, and it can process more than 13 images each second.

6 Results

This section presents the results for the off-line processing of the data gathered at the flight experiment. The results are divided into seven parts. The first part seeks to verify the relationship between OF and velocity. The second part describes the flight used to evaluate the different tracking system architectures. The latter parts of this section present the results for the tracking systems and an evaluation of the consistency of the estimates.

6.1 Flight 1 - Finding the Position and Velocity of a Marine Vessel at Rest

The first test is based on images with the large marine vessel displayed in Fig. 6, which is located in the camera field of view for a short period on two separate occasions. The main motivation behind this test is to extract measurements for the position and velocity of the vessel, as described more closely in Section 3. The vessel is approximately 70 meters long and has a width of 13 meters. Fig. 7 shows the UAV path (estimated by the navigation system) and the path of the vessel (measured by GPS) for a period of 80 seconds. Fig. 8 shows the gimbal orientation in the same time span. The vessel has almost zero speed, but the system has no knowledge about the motion of the vessel.

The vessel is not in the camera field of view in the the time intervals [20, 50] and [72, 80]. Furthermore, SIFT is not able to find features on the ship in some images. 600 images were captured in the time period and features were detected on the vessel in 250 images. A part of the vessel is visible in approximately 400 images. However, 100 of these images only contain a very small part of the vessel. The whole vessel is visible in 200 images.

Fig. 9 and 10 show the theoretical flow and OF measured by SIFT in horizontal (r) and vertical (s) direction in the image plane. Since the vessel is at rest, the theoretical flow is expected to be equal to the measured OF. The noise level is fairly large in s and the accuracy is better in r, but you can clearly see that the theoretical flow and OF are correlated. Considering the

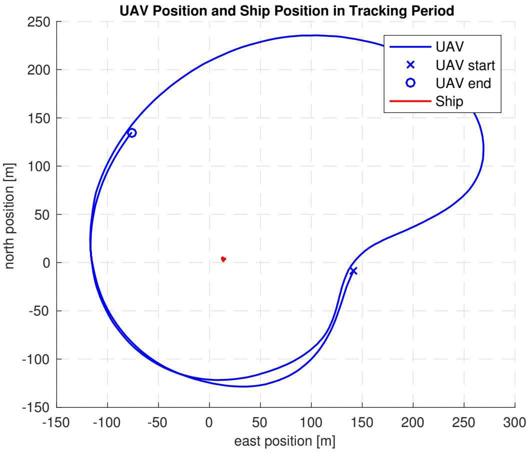

Fig. 7 Position in the NE plane for the UAV and the ship in the first test.

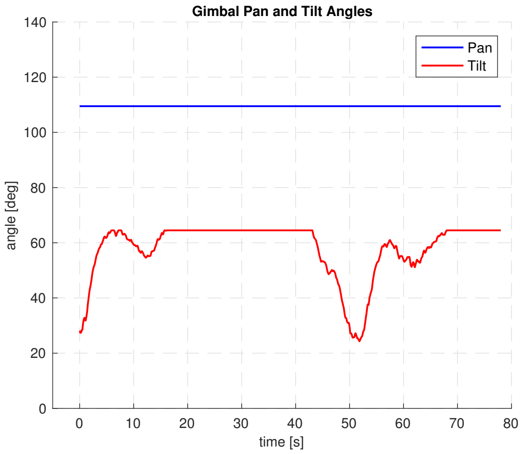

Fig. 8 Gimbal orientation in the first test.

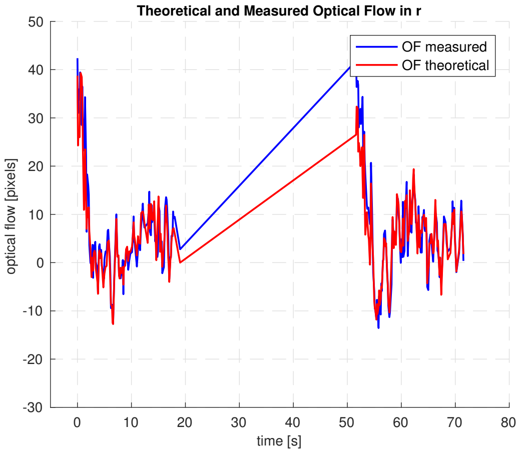

Fig. 9 Comparison of theoretical and measured optical flow in the horizontal direction (r). They should, in theory, be equal for objects at rest.

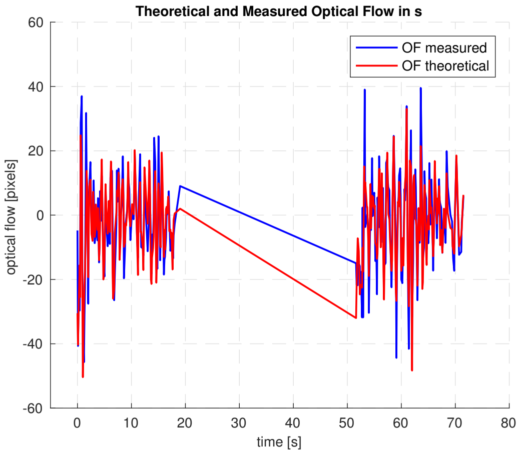

Fig. 10 Comparison of theoretical and measured optical flow in vertical direction (s). They should, in theory, be equal for objects at rest.

uncertainty related to synchronization of data and the accuracy of the sensors [23] into account, the results look reasonable.

Fig. 11 shows the georeferenced position of the vessel together with the UAV and vessel position. The georeferenced position does not vary significantly, which is the expected behavior for a target at rest. However, there seem to be a connection between the UAV navigation states and the obtained position since it varies with where the UAV is located on the path in Fig. 7. This is visible in the beginning and at 55 seconds since the georeferenced position is correlated at these time instants, and the UAV is located at approximately the same place on both occasions. Moreover, the georeferenced North and East position decrease as the UAV moves on the trajectory in Fig. 7. Ideally, the georeferenced position should be constant with small oscillations (because of measurement noise) about this value. Note that both the attitude, gimbal orientation and position can be the reason for the correlation between the georeferenced position and the UAV position on the path. Nevertheless, considering the issue with synchronization, the accuracy of the obtained position is not too troublesome. It should also be emphasized that, as shown in Fig. 6, features are not necessarily uniformly distributed on the target. Hence, the mean position of features can be quite far from the center of the vessel, which is were the GPS is located. This will obviously also affect the accuracy, especially with length of the vessel in mind.

Fig. 12 shows the velocity of the ship obtained by OF. The noise level is quite large, but the mean error is within $1m/s$ in both the North and East velocities. The calculated velocity is particularly vulnerable for the issues with synchronization of data since it is important

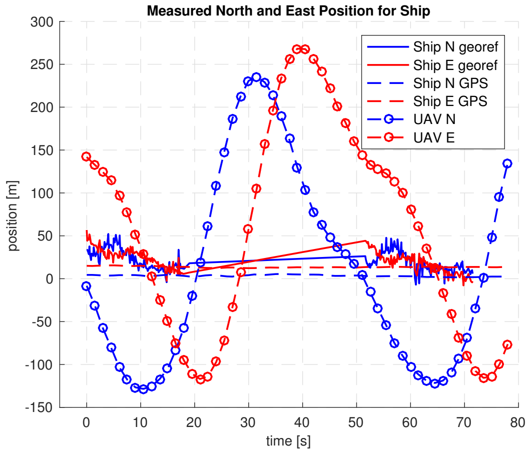

Fig. 11 Measured position of ship obtained with georeferencing together with UAV position and the ship position measured by GPS. The ship was not in the field of view of the camera in the time interval $[20, 50]$ and $[72, 80]$.

to know the exact attitude, velocity and position of the UAV when two consecutive images are captured. This is somewhat problematic since the sampling rate of the UAV navigation states just slightly exceeds the frame rate of the camera. Thus, it would be more beneficial to estimate the states of the UAV at a much higher frequency to minimize the impact of synchronization.

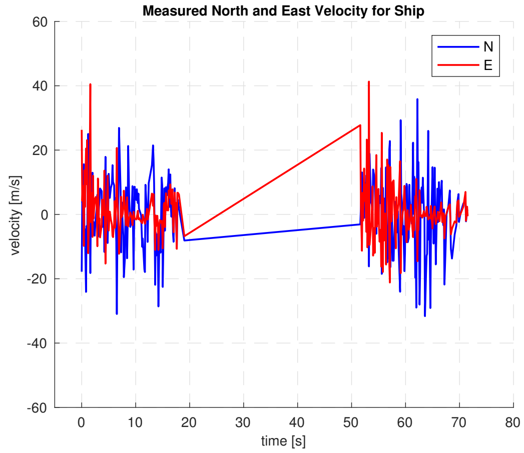

Fig. 12 Measured velocity of ship obtained with optical flow. The ship was not in the field of view of the camera in the time interval $[20, 50]$ and $[72, 80]$.

6.2 Flight 2 - Description

This section describes the flight data used to evaluate the tracking architectures in Section 4. The results are based on images with the small marine vessel displayed in Fig. 5. Fig. 13 shows the UAV path estimated by the navigation system together with the path of the vessel (measured by GPS) for approximately 55 seconds, which is the tracking period. Fig. 14 shows the gimbal orientation in the tracking period.

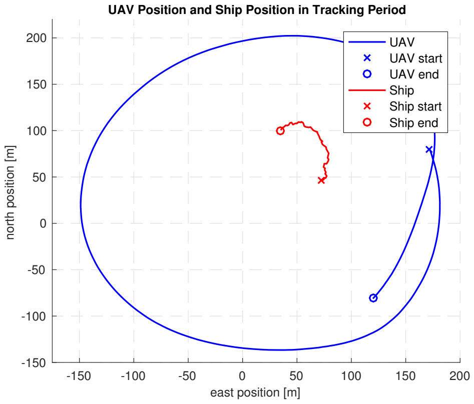

Fig. 13 Position in the NE plane for the UAV and the vessel in the second flight.

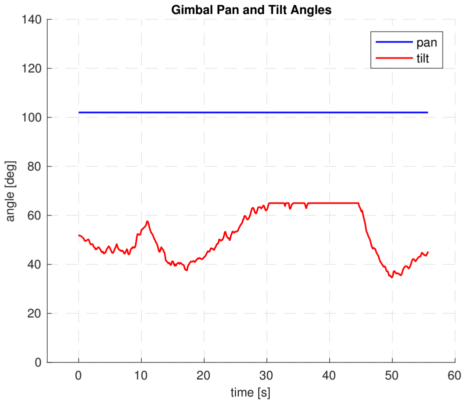

Fig. 14 Gimbal orientation during the tracking period in the second flight.

The vessel is only in the field of view of the camera in the time intervals [0, 5] and [37, 48]. Thus, the estimates are in a very large part of the tracking period solely based on prediction. 420 images were captured in the tracking period and features were detected on the vessel in 97 images. The initial covariance for the target states was chosen to be a diagonal matrix with a variance of $36m^2$ for the NE positions and $10(m/s)^2$ for the NE velocities. The process noise covariance (in continuous time) $\mathbf{Q}$ was designed as a diagonal matrix with a variance of $(3m/s)^2$, although smaller accelerations are expected in practice. The estimated position and velocity are initialized with the position obtained by georeferencing in the first image and zero, respectively. Details related to each tracking system is described more closely in the relevant section.

6.3 Flight 2 - Tracking System based on Georeferencing and Optical Flow

The tracking system based on georeferencing and OF (referred to as the first tracking system) uses measurements of both position and velocity. The covariance of the measurement noise was designed as a diagonal matrix with a variance (in continuous time) of $(12m)^2$ for the position measurements and $(6m/s)^2$ for the velocity measurements.

Fig. 15 and 16 display the estimated position and speed. The estimated position is quite close to the reference. Obviously, the estimates are slightly more accurate in the time intervals when measurements are available. Nevertheless, the predicted position is quite reasonable in both North and East when measurements are unavailable, especially since the target operates outside the field of view of the camera for 30 seconds and the vessel is maneuvering in that period. The estimated speed is also quite accurate. It is slightly above the GPS measured speed in the first part of the tracking period. This is most likely because the set of measurements are so limited in the beginning, and thus it is challenging for the estimates to converge before measurements are unavailable.

It is important to point out that the whiteness of the measurement noise is somewhat questionable. This is because the measurements strongly depend on the UAV navigation data, and thus, it is likely that subsequent measurement noise is correlated. If this is the case, it is also a violation of the conditions related to the optimality of the Kalman filter, which explains why the estimates are somewhat inaccurate at times. This is discussed more closely in Section 6.7.

6.4 Flight 2 - Tracking System based on Georeferencing

The tracking system based on georeferencing (second tracking system) uses measurements of position. The covariance of the measurement noise was designed as a diagonal matrix with a variance (in continuous time) of $(12m)^2$. This is in line with the chosen covariance for the measurement noise in the first tracking system.

Fig. 17 and 18 display the estimated position and speed. The estimated position is more accurate than for the tracking system based on georeferencing and OF. An increase in accuracy is especially visible in the time period where only prediction is used (measurements not available). This is mainly because the estimated speed is more accurate when the target moves outside the field of view of the camera (at approximately 5 seconds).

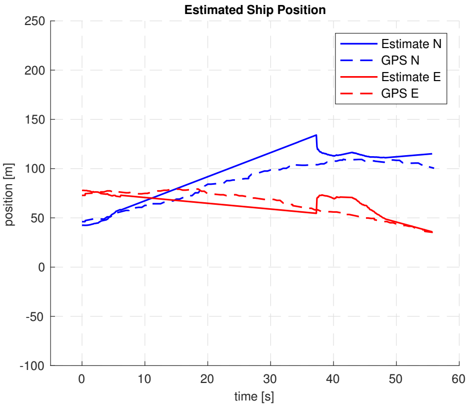

Fig. 15 Estimated position compared with the GPS measured position for the tracking system based on georeferencing and optical flow.

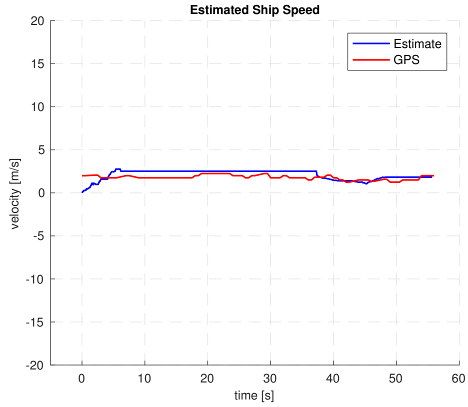

Fig. 16 Estimated speed compared with the GPS measured speed for the tracking system based on georeferencing and optical flow.

These results indicate that the velocity measurement actually leads to less accuracy for the estimates in the beginning of the tracking period. However, it is important to emphasize that the accuracy of the first tracking system is comparable to the second system near the end of the tracking period. Therefore, one cannot claim that the velocity measurement is useless. Moreover, the comparable accuracy in the end may indicate that the last couple of velocity measurements (before the target leaves the field of view in the beginning) are inaccurate, and perhaps the main reason for the error in estimated speed. Nevertheless, the results show that the reward for using the velocity measurements not compensates for the growth in complexity in this case. One cannot rule out the results could have been different in other scenarios, for example if the synchronization of data had been better.

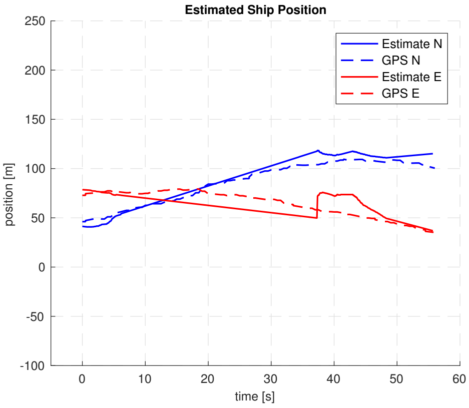

Fig. 17 Estimated position compared with the GPS measured position for the tracking system based on georeferencing.

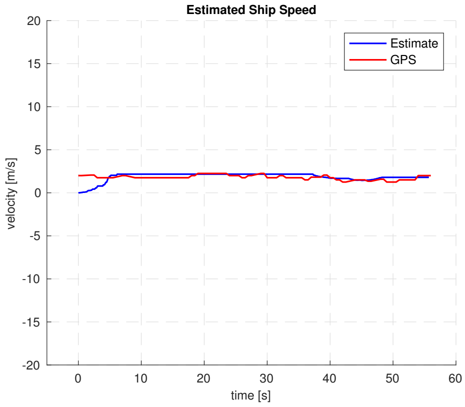

Fig. 18 Estimated speed compared with the GPS measured speed for the tracking system based on georeferencing.

6.5 Flight 2 - Tracking System based on Bearing-only Measurements

The tracking system based on bearing-only measurements (third tracking system) uses measurements of the target position in the image plane. The measurement noise was designed as a diagonal matrix with a variance of (75 pixels)^2 and (60 pixels)^2 for the horizontal and vertical dimension, respectively (converted to meters). Fig. 19 and 20 display the estimated position and speed. The estimated position is quite accurate. Moreover, it has slightly less drift than the tracking systems in Section 6.3 and 6.4 in the period without measurements. The estimated speed is also more accurate than for the tracking system based on georeferencing and OF. The increased accuracy is most likely because the system is less affected by uncertainties in synchronization of data than the system in Section 6.3. Additionally, it is not necessary to calculate the range explicitly.

The estimates from this tracking system are quite trustworthy, but it is very challenging to increase the accuracy when the set of available measurements are so limited. This is because measurements that are received just before the target disappears from the field of view, get a large influence on how the states are predicted when measurements are unavailable. This problem is enhanced when you have a small set of measurements because the tracking filter not necessarily converges with a small set of consecutive measurements. It is also worth noticing that the vessel is maneuvering (see Fig. 13), and not behaves as assumed in the constant-velocity motion model. Therefore, increased accuracy would have been expected if the set of measurements had been larger. This is obviously something that are relevant for the other tracking architectures as well.

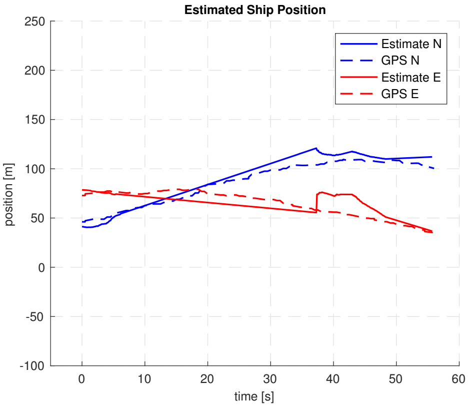

Fig. 19 Estimated position compared with the GPS measured position for the tracking system based on bearing-only measurements.

6.6 Flight 2 - Tracking System based on a Schmidt-Kalman Filter

The tracking system based on a Schmidt-Kalman filter (fourth tracking system) has the same measurements and measurement noise covariance as the bearing-only tracking system (Section 6.5). Because measurements of specific force were unavailable, the UAV NED positions were extracted from the autopilot and not estimated in an error-state Kalman filter. In order to use

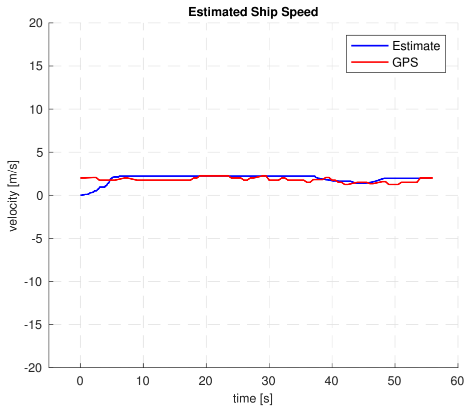

Fig. 20 Estimated speed compared with the GPS measured speed for the tracking system based on bearing-only measurements.

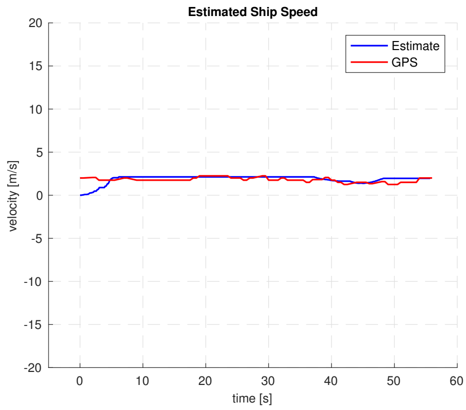

Fig. 22 Estimated speed compared with the GPS measured speed for the tracking system with a Schmidt-Kalman filter.

a Schmidt-Kalman filter, a constant covariance was designed to represent the effect of uncertainty in the UAV NED positions. The variance was chosen to be $10m^2$ for the North and East positions and $50m^2$ for the down position. Fig. 21 and 22 display the estimated position and speed, which are comparable to the results in Section 6.5. There is no obvious increase in accuracy for the estimates when accounting for uncertainty in the UAV NED positions. The factors discussed in Section 6.5 is also highly relevant for this tracking architecture. The benefit of using the Schmidt-Kalman filter is highlighted in the next section.

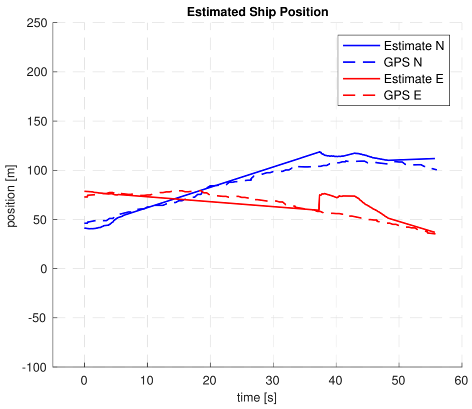

Fig. 21 Estimated position compared with the GPS measured position for the tracking system with a Schmidt-Kalman filter.

6.7 Flight 2 - Consistency Analysis

This section discusses the consistency [44] of the estimates from the different tracking systems. The normalized innovation squared (NIS) for the tracking systems is displayed in Fig. 23. The NIS is almost equal for the third and fourth tracking system, and thus only visible as one graph. All tracking systems have innovations within the 95 % confidence interval. The second, third and fourth tracking system have two measurements and, therefore, two degrees of freedom (DOF). The first tracking system has four measurements and 4 DOF. Fig. 23 clearly shows that the velocity measurement from OF increases the NIS significantly (compare the NIS for the first and second tracking system). This indicates that the velocity measurements are far away from the predicted measurements at several samples. Comparable NIS is achieved for the other tracking systems. Nevertheless, all tracking systems have NIS within the confidence bounds, and therefore, the first part of consistency is fulfilled.

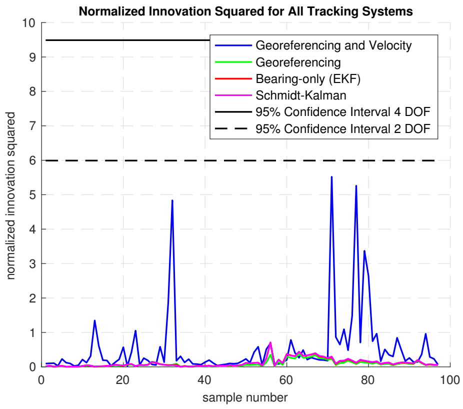

Fig. 23 Normalized innovation squared for all tracking systems. The black lines indicates the 95% confidence intervals with 2 and 4 degrees of freedom.

The second part of the consistency analysis is the whiteness test for the innovations [44]. The autocorrelation of the innovations shows that the innovations in velocity can be considered white (95% confidence interval) for the first tracking system. The innovations for the rest of the measurements are on the other hand not white for any tracking system. Hence, the measurement noise is correlated for consecutive time-steps and this violates the assumptions of the Kalman filter. Thus, one cannot conclude that either of the tracking systems are consistent, solely based on the definition of consistency [44]. This is suspected to be because the attitude of the UAV seems to influence the measurement noises more than the pixel and UAV position. Nevertheless, in a scenario where one only wants to track a single target, it is reasonable to claim that consistency is less important than the accuracy of the estimates.

In order to clarify the effect of the uncertainty in the UAV NED positions, the norm of the covariance matrix is investigated. Fig. 24 shows the norm of the covariance matrices in the time interval [38,52] for all tracking systems. The estimates in the Schmidt-Kalman filter have a more rapid increase in covariance when measurements are unavailable. This is in compliance with the expected behavior because the uncertainty related to the UAV NED positions is accounted for. Moreover, since the measurements are affected by the UAV position, the norm of the covariance in the Schmidt-Kalman filter decreases slower than for the other architectures when measurements are available. Notice that the covariance of the linear architectures have a slightly slower increase in covariance when measurements are unavailable (after 48 seconds).

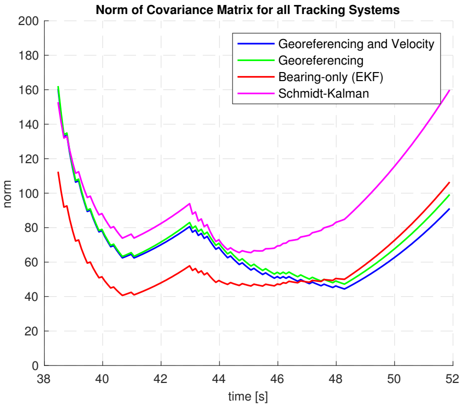

Fig. 24 Norm of covariance matrix for all tracking systems.

7 CONCLUSIONS

This paper presented four vision-based tracking systems for marine surface objects utilizing thermal images captured in a fixed-wing unmanned aerial vehicle with a retractable pan/tilt gimbal and a thermal camera. Experimental results show that it is possible to estimate the position and velocity of a vessel with thermal images captured from a fixed-wing UAV operating at high speed. More importantly, the systems are able to predict the position of the vessel quite well when it operates outside the field of view of the camera. The results indicate that measurement errors in the UAV NED positions do not influence the performance significantly, and it is suspected that measurement errors in attitude has a much larger influence.

A Calculating Camera-fixed Coordinates in terms of the UAV Navigation States

This appendix seeks to explain how the camera-fixed coordinates of a feature at pixel (r, s) can be computed as a function of UAV navigation states and gimbal orientation. Let the homogeneous coordinates of the feature be written as $\underline{\mathbf{t}}^n = [x^n_f, y^n_f, z^n_f, 1]^\top$ and $\underline{\mathbf{t}}^c = [x^c_f, y^c_f, z^c_f, 1]^\top$ decomposed in {N} and {C}, respectively. The relationship between the coordinates is

$$\underline{\mathbf{t}}^c = \mathbf{T}^c_n \underline{\mathbf{t}}^n \tag{28}$$

where $\mathbf{T}^c_n$ is the homogeneous transformation between {C} and {N} defined as

$$\mathbf{T}^c_n := \begin{bmatrix} \mathbf{R}^c_n & -\mathbf{R}^c_n \mathbf{r}^n_{nc} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}$$

By inserting (28) into the pinhole camera model (1) the equation can be solved with respect to x^n_f and y^n_f by assuming that z^n_f is known (z^n_f is zero in this case). The solution decomposed in {C} is calculated with (28) and given as

$$\mathbf{t}^c_* = \begin{bmatrix} x^c_f \\ y^c_f \\ z^c_f \end{bmatrix} = \frac{1}{r' s\psi_{gb} s\theta + f c\theta_{gb} c\phi c\theta + s' c\psi_{gb} c\theta_{gb} s\theta - f c\psi_{gb} s\theta_{gb} s\theta + r' c\psi_{gb} c\theta s\phi + s' c\phi c\theta s\theta_{gb} - s' c\theta_{gb} c\theta s\psi_{gb} s\phi + f c\theta s\psi_{gb} s\theta_{gb} s\phi} \begin{bmatrix} -r'(z^n_{uav} - z^n_{f)} \\ -s'(z^n_{uav} - z^n_f) \\ -f(z^n_{uav} - z^n_f) \end{bmatrix}$$

where r' and s' are the pixel coordinates and s and c are the sine and cosine functions. z^n_{uav} is the down position of the UAV. $\mathbf{t}^c_*$ only depends on known parameters, and thus the camera-fixed coordinates of features are known as long as all features are located at sea level, which is a sensible assumption in this case.

Acknowledgements This work has been carried out at the NTNU Centre for Autonomous Marine Operations and Systems. This work was supported by the Research Council of

Norway through the Centres of Excellence funding scheme, Project number 223254. The authors are grateful for the support from the University of Porto, the University of the Azores, UAV operators Lars Semb and Krzysztof Cisek, REP15 coordinators João Tasso Sousa and Kanna Rajan and valuable feedback from Edmund Førland Brekke.

References

1. L. Fusini, J. Hosen, H. H. Helgesen, T. A. Johansen, and T. I. Fossen, "Experimental validation of a uniformly semi-globally exponentially stable non-linear observer for gnss- and camera-aided inertial navigation for fixed-wing uavs," in *Proc. of the International Conference on Unmanned Aircraft Systems*, 2015, pp. 851–860.
2. J. Hosen, H. H. Helgesen, L. Fusini, T. I. Fossen, and T. A. Johansen, "A vision-aided nonlinear observer for fixed-wing uav navigation," in *Proc. of the AIAA Guidance, Navigation, and Control Conference*, 2016.
3. ——, "Vision-aided nonlinear observer for fixed-wing unmanned aerial vehicle navigation," *Journal of Guidance, Control, and Dynamics*, vol. 39, no. 8, pp. 1777–1789, 2016.
4. S. Zhao, F. Lin, K. Peng, X. Dong, B. M. Chen, and T. H. Lee, "Vision-aided estimation of attitude, velocity, and inertial measurement bias for uav stabilization," *Journal of Intelligent & Robotic Systems*, vol. 81, no. 3, pp. 531–549, 2016.
5. J. E. Gomez-Balderas, G. Flores, L. R. García Carrillo, and R. Lozano, "Tracking a ground moving target with a quadrotor using switching control," *Journal of Intelligent & Robotic Systems*, vol. 70, no. 1, pp. 65–78, 2013.
6. P. Doherty and P. Rudol, "A uav search and rescue scenario with human body detection and geolocalization," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4830, pp. 1–13, 2007.
7. P. Rudol and P. Doherty, "Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery," in *Proc. of the IEEE Aerospace Conference*, 2008.
8. X. Yu and Y. Zhang, "Sense and avoid technologies with applications to unmanned aircraft systems: Review and prospects," *Progress in Aerospace Sciences*, vol. 74, pp. 152 – 166, April 2015.
9. A. Hiba, T. Zsedrovits, P. Bauer, and A. Zarandy, "Fast horizon detection for airborne visual systems," in *Proc. of the International Conference on Unmanned Aircraft Systems*, 2016, pp. 886–891.
10. S. J. Dumble and P. W. Gibbens, "Efficient terrain-aided visual horizon based attitude estimation and localization," *Journal of Intelligent & Robotic Systems*, vol. 78, no. 2, pp. 205–221, 2015.
11. A. Ortiz, F. Bonnin-Pascual, and E. Garcia-Fidalgo, "Vessel inspection: A micro-aerial vehicle-based approach," *Journal of Intelligent & Robotic Systems*, vol. 76, no. 1, pp. 151–167, 2014.
12. "COLREGs - convention on the international regulations for preventing collisions at sea, international maritime organization (IMO)," 1972.
13. L. Elkins, D. Sellers, and W. R. Monach, "The autonomous maritime navigation (amn) project: Field tests, autonomous and cooperative behaviors, data fusion, sensors, and vehicles," *Journal of Field Robotics*, vol. 27, no. 6, pp. 790–818, 2010.
14. M. T. Wolf, C. Assad, Y. Kuwata, A. Howard, H. Aghazarian, D. Zhu, T. Lu, A. Trebi-Ollennu, and T. Huntsberger, "360-degree visual detection and target tracking on an autonomous surface vehicle," *Journal of Field Robotics*, vol. 27, no. 6, pp. 819–833, 2010.
15. T. Huntsberger, H. Aghazarian, A. Howard, and D. C. Trotz, "Stereo vision–based navigation for autonomous surface vessels," *Journal of Field Robotics*, vol. 28, no. 1, pp. 3–18, 2011.
16. T. A. Johansen and T. Perez, "Unmanned aerial surveillance system for hazard collision avoidance in autonomous shipping," in *Proc. of the International Conference on Unmanned Aircraft Systems*, 2016.
17. A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.
18. C. Martínez, I. F. Mondragón, P. Campoy, J. L. Sánchez-López, and M. A. Olivares-Méndez, "A hierarchical tracking strategy for vision-based applications on-board uavs," *Journal of Intelligent & Robotic Systems*, vol. 72, no. 3, pp. 517–539, 2013.
19. A. Kadyrov, H. Yu, and H. Liu, "Ship detection and segmentation using image correlation," in *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics*, Oct 2013, pp. 3119–3126.
20. F. S. Leira, T. A. Johansen, and T. I. Fossen, "Automatic detection, classification and tracking of objects in the ocean surface from uavs using a thermal camera," in *Proc. of the IEEE Aerospace Conference, Big Sky, US*, 2015.
21. M. Mammarella, G. Campa, M. L. Fravolini, and M. R. Napolitano, "Comparing optical flow algorithms using 6-dof motion of real-world rigid objects," *IEEE Transactions on systems, Man, and Cybernetics*, vol. 42, no. 6, pp. 1752–1762, 2012.
22. H. H. Helgesen, F. S. Leira, T. A. Johansen, and T. I. Fossen, "Tracking of marine surface objects from unmanned aerial vehicles with a pan/tilt unit using a thermal camera and optical flow," in *Proc. of the International Conference on Unmanned Aircraft Systems*, 2016, pp. 107–117.
23. F. S. Leira, K. Trnka, T. I. Fossen, and T. A. Johansen, "A ligth-weight thermal camera payload with georeferencing capabilities for small fixed-wing uavs," in *Proc. of the International Conference on Unmanned Aircraft Systems*, 2015, pp. 485–494.
24. J.-H. Kim and S. Sukkarieh, "Airborne simultaneous localisation and map building," in *Proc. of the IEEE International Conference on Robotics and Automation*, vol. 1, 2003.
25. M. Bryson and S. Sukkarieh, "Building a robust implementation of bearing-only inertial slam for a uav," *Journal of Field Robotics*, vol. 24, no. 1-2, pp. 113–143, 2007.
26. ——, "Bearing-only slam for an airborne vehicle," in *Proc. of the Australasian Conference on Robotics and Automation*, vol. 4, 2005.
27. C. Yang, E. Blasch, and P. Douville, "Design of schmidt-kalman filter for target tracking with navigation errors," in *Proc. of the IEEE Aerospace Conference*, 2010.
28. R. Y. Novoselov, S. M. Herman, S. M. Gadaleta, and A. B. Poore, "Mitigating the effects of residual biases with schmidt-kalman filtering," in *Proc. of the 8th International Conference on Information Fusion*, vol. 1, 2005.
29. E. F. Wilthil and E. F. Brekke, "Compensation of navigation uncertainty for target tracking on a moving platform," in *Proc. of the 19th International Conference on Information Fusion*, 2016, pp. 1616–1621.

30. T. Fossen, *Handbook of Marine Craft Hydrodynamics and Motion Control.* John Wiley & Sons, 2011.
31. B.K.P.Horn and B.G.Schunk, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–204, 1981.
32. D. Lowe, "Object recognition from local scale-invariant features," *Proc. of the International Conference on Computer Vision*, pp. 1150–1157, 1999.
33. M. Muja and D. G. Lowe, "Flann, fast library for approximate nearest neighbors," in *Proc. of the International Conference on Computer Vision Theory and Applications*, 2009.
34. G. Zhou, C. Li, and P. Cheng, "Unmanned aerial vehicle (uav) real-time video registration for forest fire monitoring," in *Proc. of IEEE International Geoscience and Remote Sensing Symposium.*, vol. 3, July 2005, pp. 1803–1806.
35. E. M. Hemerly, "Automatic georeferencing of images acquired by uav's," *International Journal of Automation and Computing*, vol. 11, no. 4, pp. 347–352, 2014.
36. D. B. Barber, J. D. Redding, T. W. McLain, R. W. Beard, and C. N. Taylor, "Vision-based target geo-location using a fixed-wing miniature air vehicle," *Journal of Intelligent and Robotic Systems*, vol. 47, no. 4, pp. 361–382, 2006.
37. S. Hutchinson, G. Hager, and P. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.
38. O. Egeland and J. T. Gravdahl, *Modeling and simulation for automatic control.* Marine Cybernetics Trondheim, Norway, 2002.
39. X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking: dynamic models," in *Proc. of SPIE*, vol. 4048, 2000, pp. 212–235.
40. T. Johansen and T. Fossen, "Nonlinear filtering with exogenous kalman filter and double kalman filter," in *Proc. of the European Control Conference, Aalborg*, 2016.
41. G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.
42. E. Skjong, S. Nundal, F. Leira, and T. Johansen, "Autonomous search and tracking of objects using model predictive control of unmanned aerial vehicle and gimbal: Hardware-in-the-loop simulation of payload and avionics," in *Proc. of the International Conference on Unmanned Aircraft Systems*, 2015, pp. 904–913.
43. J. Hansen, T. Fossen, and T. Johansen, "Nonlinear observer for ins aided by time-delayed gnss measurements: Implementation and uav experiments," in *Proc. of the International Conference on Unmanned Aircraft Systems*, 2015, pp. 157–166.
44. Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software.* John Wiley & Sons, 2004.