Ocean Engineering 133 (2017) 197–214

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng

Stereovision-based target tracking system for USV operations

Armando J. Sinisterra, Manhar R. Dhanak*, Karl Von Ellenrieder

Department of Ocean and Mechanical Engineering, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA

ARTICLE INFO

Keywords:
Stereovision-based navigation
Target tracking
Unmanned surface vehicle
Extended Kalman filter
Computer vision
Object detection

ABSTRACT

A stereovision based methodology to estimate the position, speed and heading of a moving marine vehicle from a pursuing unmanned surface vehicle (USV) is considered, in support of enabling a USV to follow a target vehicle in motion. The methodology involves stereovision ranging, object detection and tracking, and minimization of tracking error due to image quantization limitations and pixel miscorrespondences in the stereo pixel-matching process. The method consists of combining a simple stereovision-matching algorithm, together with a predictive-corrective approach based on an extended Kalman filter (EKF), and use of suitable choices of probabilistic models representing the motion of the target vehicle and the stereovision measurements. Simple matching algorithms perform faster at the expense of potential errors in depth measurement. The approach considered aims to minimize the tracking errors related to such errors in stereovision measurements, thereby improving the accuracy of the state estimation of the vehicle. Results from simulations and a real-time implementation reveal the effectiveness of the system to compute accurate estimates of the state of the target vehicle over non-compliant trajectories subjected to a variety of motion conditions.

1. Introduction

Applications based on stereovision systems are gaining popularity in areas such as automatic pedestrian detection (Bhowmick et al., 2011; Bertozzi et al., 2004; Sinharay et al., 2011), and obstacle avoidance in general for both indoor and outdoor environments, mostly from static cameras or ground vehicles (Goldberg et al., 2002). The strength of such systems reside in the fact that they provide accurate 3-D reconstruction of the surrounding scenarios, while combining a number of computer vision techniques, in support of extracting relevant data from the image, such as colors, edges, corners, and key feature points among others (Harris and Stephens, 1988), (Lowe, 1999) as well as categorizing different components in the scene, such as pedestrians, background, foreground, roads, cars, etc (Dalal and Triggs, 2005; Papageorgiou and Poggio, 2000). The technique used has been ported to the area of marine vehicle navigation and its dynamic environment (Huntsberger et al., 2011; Larson et al., 2006; Wang et al., 2011). The system has to be modified to adjust to a number of situations such as constant motion associated with the sea state, changes in illumination due to the position of the sun or the presence of clouds, glare, reflections, partial or total occlusions due to rain or another vehicle crossing through, among other conditions that could impact its performance (Heo et al., 2008). Although navigation of Unmanned Surface Vehicles (USVs), is still in an early stage of development, there are many functionalities that can be inherited from the more mature navigation of Unmanned Ground Vehicles (UGVs). In the specific case of stereovision odometry, there have been many successful implementations for UGVs (Mathies, 1989; Goldberg et al., 2002), highlighted in important autonomous navigation projects such as the Mars Exploration Rovers (MER). However, the complexities associated with a dynamic marine environment require special attention when addressing the problem of autonomous maritime navigation.

Different approaches regarding marine hazard detection have made important contributions in the field of autonomous maritime navigation. The differences mainly lie in the use of different sensors and the way the data are processed for each to obtain equivalent information. In Larson et al. (2007), for example, the use of monocular camera and radar-based systems are described. While the monocular system has limitations in operating under a wide range of illumination conditions, the radar system generates reliable detections but with a relatively slow scanning rate, making it inadequate for high-speed vessels.

A Laser Imaging Detection and Ranging (LIDAR) system is described in Ruiz and Granja (2009), and is used to assist human pilots in inland waterways. It converts a commercial LIDAR to a two-axis laser scan and process the resultant 2D range image to detect discrete objects with an effective range between 300 and 500 m. Some difficulties using the system were found by the authors on moving objects, due to the long scanning times. An interesting approach is also taken in Kohlbrecher et al. (2011), where a portable system using

* Corresponding author.
E-mail addresses: asiniste@fau.edu (A.J. Sinisterra), dhanak@fau.edu (M.R. Dhanak), ellenrie@fau.edu (K. Von Ellenrieder).

http://dx.doi.org/10.1016/j.oceaneng.2017.01.024
Received 1 April 2016; Received in revised form 2 January 2017; Accepted 22 January 2017
Available online 16 February 2017
0029-8018/

occupancy grid mapping at multiple scales is used in conjunction with a LIDAR sensor. The same platform was tested on both UGVs and USVs given satisfactory results for small-scale scenarios.

Sonar systems, on the other hand, have been used in Autonomous Underwater Vehicles (AUVs), as the one described in Folkesson and Leonard (2011), for which an estimated local map of the position of the vessel, and the detected features are matched with an *a priori* map of the scene to get the absolute measurements. The *a priori* map is made using an AUV equipped with a sophisticated suite of sensors and a side-scan sonar, while the AUV used for the operation is a low-cost platform equipped with a depth sensor, altimeter, GPS, compass and a forward looking sonar.

One of the most complete autonomous maritime platforms is reported in Larson et al. (2006), where a behavior-based hazard avoidance system for USVs is defined by the combination of a deliberative obstacle avoidance scheme using a path planner (making use of digital nautical charts, DNC) with an occupancy grid map, for a far-field implementation, and a reactive avoidance scheme for a fast response to obstacles in close proximity, using a suite of near-field sensors, among which the stereovision camera is included. The objective here is to follow the original path as much as possible, while avoiding moving and stationary obstacles.

A similar approach is described in Huntsberger et al. (2011), where more emphasis is placed on implementation of the stereovision sensor. The Hammerhead system is described as a left and right pairs of directional cameras covering a combined horizontal field of view of approximately 100°. There are also some image processing improvements with respect to the work in Larson et al. (2006) for handling noise, reflections and low lightening conditions, among other issues. The software for the vision system is divided in two main tasks; one is the stereovision ranging which returns a 3D image cloud of the scene. The second task refers to the classifier which returns a contact lists of detected objects in the scene such as boats, channel markers, buoys, etc. The combination of these two routines has resulted in a robust vision system platform which is capable not only to measure distances to all the components in the scene but also to recognize some important categories within a typical marine environment. One of the main operations involved following a target boat at a slow speed implementing object detection and tracking routines based on a supervised learning algorithm where the classifier is trained offline with hand-labeled data. The speed and the heading are estimated by changes in position, and thus, their accuracy is limited by the uncertainty of the stereovision sensor which grows with distance.

Another different approach regarding stereovision is described in (Wang, et al., 2011) and (Wang et al., 2012). Here, monocular and stereovision methods are combined in order for the system to detect and track the objects of interest within the scene while performing stereovision ranging over these particular regions. This approach makes use of some characteristics of the marine environment such as textures and colors where obstacles over the sea surface are visual pop-outs that can be targeted by using different monocular techniques. They start first by computing the sea-sky line (horizon) to distinguish the actual navigating objects encountered over the sea surface. The horizon is also used to compute the normal to the sea surface, and used to align the coordinate system of the image plane of the left camera for further projection from 3D to 2D; that is, the projection of the world coordinates onto the image plane of the camera. Next, in the saliency detection stage, a binary image is examined to address potential obstacles that are subject to the appearance of false positives caused by different sources of noise such as, for example, light reflections from the water surface. In order for the system to get rid of these false positives, feature points (Harris corners) (Harris and Stephens, 1988) are computed over the potential obstacles and are tracked using the KLT tracker (Bouguet, 1999) over a specific number of frames under the criterion that actual navigating objects will have a continuous motion and will be easily traceable as opposed to noise which tend to disappear after a few number of frames. The result of this process is the segmentation of the objects of interest from the scene, which are marked by a boundary box that encloses the region of interest for further stereovision processing. This approach has yielded interesting results such as the capacity of the system to detect and locate multiple objects over a wide range (from 30 to 300 m) from a USV in real-time. However, all of the results presented over field testing describe simplistic maritime scenarios, in which a single object is detected from among an extensive surrounding sea surface with no other components in the scene that may be encountered in a typical marine environment, such as, for example, nearby shorelines, inland waters, harbors, vegetation or any other onshore structured components that can be close to the navigation area. Furthermore, as can also be inferred in Huntsberger et al. (2011) and all stereovision systems in general, the accuracy of the system radically drops after a certain distance, sometimes even within what is considered its effective range, and thus the measurements are no longer accurate. This represents a major problem when performing following-behavior routines in the context of USV operations, where the entire state of the target vehicle (position, speed and heading) needs to be defined solely from stereovision measurements.

The work described here is a study of an application implemented on a Unmanned Surface Vehicle (USV) which will be referred to as the chase vehicle (CH), for following a target vehicle (TG), possibly in the presence of other marine vehicles, shore lines, buoys, etc. This requires a robust tracking algorithm, that accounts for all the intrinsic sources of error of the stereovision system, such as limitations in image quantization of the camera as well as pixel miscorrespondences in the stereovision pixel-matching algorithm (Matthies and Shafer, 1987). To this end, a predictive-corrective methodology based on an extended Kalman filter (EKF) (Thrun et al., 2006) is used to minimize the errors in the stereovision measurements associated with the position of the target vehicle. This approach is applied in the time domain rather than the spatial domain of the stereovision image (disparity image of the stereovision matching algorithm) (Wang et al., 2012), which results in less algorithm complexity and reduced computational time.

The computer vision system has been implemented on a 16 ft Wave-Adaptive Modular Vessel (WAM-V) USV in support of unmanned applications of the vehicle being considered, depicted in Fig. 1. This vessel is referred to as the chase vehicle (CH) in the discussion below. The WAM-V offers a convenient set of characteristics in the context o autonomous marine navigation such as:

- a small size, which makes it easy to handle in field tests,
- a high payload capacity that accommodate the suite of sensors and controllers needed in USV applications,
- two electric motors, which makes it easier to control,
- an independent suspension system per (inflatable) pontoon, that isolates the center tray of the catamaran vehicle, which carries the payload, from the motions induced by incident waves.

The last of these characteristics mechanically enables partial motion compensation for the stereovision system, allowing a more stable image acquisition process, so that it is less prone to errors.

Fig. 1. The WAM-V 16', denoted as the chase vehicle.

Table 1
WAM-V specifications.

Length (m)	4.88
Beam (m)	2.44
Payload (kg)	113.40
Propulsion	2× electric motors (outboards)
Batteries	2× 105 A h Li NMC
Max. speed (m/s)	Up to 5.66
Platform weight (kg)	181.44
Draft (m)	0.1524–0.4064

Table 1 describes the main specifications of the WAM-V, (Marine Advanced Research, 2015):

The software is implemented using C/C++ programming language, and makes use of the camera's API and the OpenCV library. The data flow of the algorithm is depicted in Fig. 2, and is comprised mainly of three components, namely, stereovision processing, object detection and tracking, and an EKF routine, each of them enclosed by a dashed box. The stereovision processing first acquires a pair of raw images from a commercial Bumblebee2 stereovision camera (from Point Grey) denoted as L and R in the figure, for left and right cameras, respectively. It then rectifies these images, and carries out a pixel matching process using the Sum of Absolute Differences (SAD) (Bradski and Kaehler, 2008). The pixel-matching process continues for each of the pixels in the original images leading to what is known as the disparity map. The disparity values are then used along with the intrinsic parameters of the camera (computed in an offline camera calibration process) in order to infer the physical 3D-coordinates (depth image) associated with each pixel. The object detection and tracking tasks were initially implemented using a simple color tracker, which allowed us to evaluate a standalone performance of the rest of the system. This detector returns the pixel coordinates corresponding to the centroid of a particular region of interest (identified as a 'yellow blob'), to be used along with the disparity map in order to get the specific position of the target boat in physical coordinates, relative to the position of the camera. In a similar way, a more robust object detection and tracking algorithm was incorporated later on into the system, namely the TLD tracker (Tracking-Learning-Detector) (Kalal et al., 2012). This new feature added a higher level of versatility to the system, adapting its response to changes in the appearance of the target vehicle due to on-plane and off-plane rotations, translations, scaling and changes in illumination.

These relative coordinates of the target vehicle (TG) are then provided to the EKF along with the absolute position of the chase boat (CH), in order to weigh (in a probabilistic way) a predictive model of the motion of TG and its last measurement update, leading to an estimate of the true absolute position of TG at each time step, minimizing the tracking errors, measured with respect to its reference position given by the GPS as North-East (N-E) absolute coordinates.

2. Extended Kalman Filter (EKF)

The EKF estimates a Gaussian approximation of the true belief $bel(\mathbf{x}_t)$ (at time t, represented by a mean $\boldsymbol{\mu}_t$ and a covariance $\boldsymbol{\Sigma}_t$ of the state transition ($\mathbf{x}_t$)), and measurement probabilities ($\mathbf{z}_t$), governed by nonlinear functions. The main idea is to linearize these nonlinear functions around the mean value of their corresponding original Gaussians, using a first order Taylor expansion. The projection of the Gaussian variables over the linear approximation results in a Gaussian distribution, similar to the case of the regular Kalman filter implementation.

Suppose that the state transition and measurement distributions are governed by nonlinear functions g and h, as given by

$$\mathbf{x}_t = g(\mathbf{u}_t, \mathbf{x}_{t-1}) + \boldsymbol{\varepsilon}_t(\mathbf{R}_t) \tag{1}$$

$$z_t = h(\mathrm{x}_t) + \delta(\mathrm{Q}_t) \tag{2}$$

where $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\delta}_t$ are zero-mean Gaussian random vectors with corresponding covariance $\mathbf{R}_t$ and $\mathbf{Q}_t$, which model the uncertainty introduced by the state transition and measurement distributions respectively. Since g and h are nonlinear functions, the belief will no longer be a Gaussian distribution. Therefore, a first order Taylor approximation to the nonlinear functions is obtained around the mean value of the original Gaussian, as follows:

$$g(\mathbf{u_t}, \mathbf{x}_{t-1}) \approx g(\mathbf{u_t}, \boldsymbol{\mu}_{t-1}) + g'(\mathbf{u}_t, \boldsymbol{\mu}_{t-1})(\mathbf{x}_{t-1}-\boldsymbol{\mu}_{t-1}) \text{ , where}$$
$$g'(\mathbf{u}_t, \boldsymbol{\mu}_{t-1}) = \frac{\partial g(\mathbf{u}_t, \boldsymbol{\mu}_{t-1})}{\partial \mathbf{x}_{t-1}} = \mathbf{G}_t \tag{3}$$

and,

$$h(\mathbf{x}_t) \approx h(\overline{\boldsymbol{\mu}}_t) + h'(\overline{\boldsymbol{\mu}}_t)(\mathbf{x}_t - \overline{\boldsymbol{\mu}}_t), \text{ where}$$
$$h'(\overline{\boldsymbol{\mu}}_t) = \frac{\partial h(\mathbf{x}_t)}{\partial \mathbf{x}_t} = \mathbf{H}_t \tag{4}$$

Here, $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$represent the current and previous state vectors and $\boldsymbol{u}_t$ is a known input to the system. $g(\boldsymbol{u}_t, \boldsymbol{\mu}_{t-1})$ is the value of the function $g(.)$ evaluated around the mean of the previous state vector ($\boldsymbol{\mu}_{t-1}$), whereas $g'(u_t, \boldsymbol{\mu}_{t-1})$ is the Jacobian of $g(.)$ with respect to the previous state vector, evaluated around the corresponding mean value. $h(\overline{\boldsymbol{\mu}}_t)$ is the value of function $h(.)$ evaluated around $\overline{\boldsymbol{\mu}}_t$ corresponding to the current predicted state vector, and finally $h'(\overline{\boldsymbol{\mu}}_t)$ is the Jacobian of $h(.)$ also with respect to the current predicted state vector.

Once this linearization process has taken place, one can proceed to apply the EKF algorithm, which is based on the one for regular Kalman filters, as shown in Table 2, (Thrun et al., 2006). Where $\overline{\boldsymbol{\mu}}_t$ and $\overline{\boldsymbol{\Sigma}}_t$ represent the control update stage (i.e., state prediction), described as $\overline{bel}(\mathbf{x}_t)$, and $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ represent the measurement update stage described as the belief $bel(\mathbf{x}_t)$.

2.1. Motion model

The basis of the EKF model is that the only available information regarding the target vehicle (TG) is that provided by the stereovision sensor; no GPS or compass data are available to update the estimate of the true state of the TG, nor any inputs to the TG systems (e.g., thrust forces or rpm of the thrusters) are known. Therefore, use of a uniformly accelerated kinematic model is proposed, with a state vector formed by the absolute position and velocity (in the XY-plane, as shown in Fig. 3):

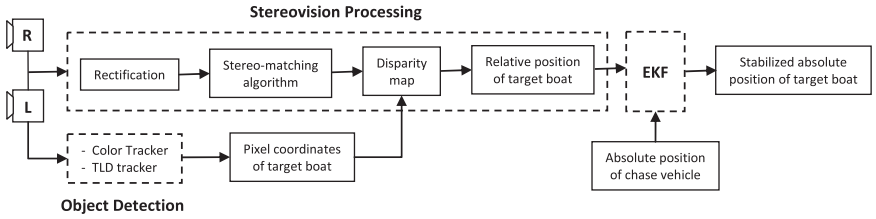


Fig. 2. Data flow of the stereovision-based tracking system.

Table 2
The Extended Kalman Filter Algorithm.

1: Algorithm ExtendedKalmanFilter($\boldsymbol{\mu}_{t-1}$, $\boldsymbol{\Sigma}_{t-1}$, u_t, $\mathbf{z}_t$)
2: $\bar{\boldsymbol{\mu}}_t = g(u_t, \boldsymbol{\mu}_{t-1})$
3: $\bar{\boldsymbol{\Sigma}}_t = \mathbf{G}_t\boldsymbol{\Sigma}_t\mathbf{G}_t + \mathbf{R}_t$
4: $\mathbf{K}_t = \bar{\boldsymbol{\Sigma}}_t\mathbf{H}_t^T(\mathbf{H}_t\bar{\boldsymbol{\Sigma}}_t\mathbf{H}_t^t + \mathbf{Q}_t)^{-1}$
5: $\boldsymbol{\mu}_t = \bar{\boldsymbol{\mu}}_t + \mathbf{K}_t(\mathbf{z}_t - h(\bar{\boldsymbol{\mu}}_t))$
6: $\boldsymbol{\Sigma}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\bar{\boldsymbol{\Sigma}}_t$
7: return$\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$

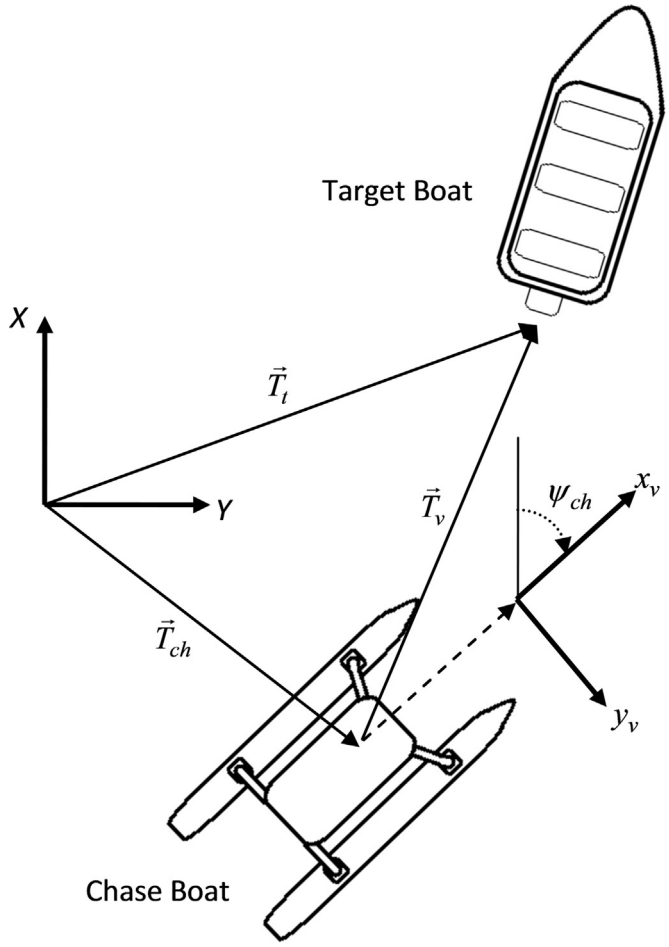


Fig. 3. Schematics representation of the tracking problem (top view).

$$\begin{bmatrix} X_t \\ Y_t \\ \dot{X}_t \\ \dot{Y}_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \\ \dot{X}_{t-1} \\ \dot{Y}_{t-1} \end{bmatrix} + \begin{bmatrix} T^2/2 \\ T^2/2 \\ T \\ T \end{bmatrix} a + \boldsymbol{\varepsilon}(\mathbf{R}_t) \tag{5}$$

Eq. (5) represents the state transition model, as defined in Eq. (1), and since it is already linear, it can be expressed in the regular Kalman filter form:

$$\mathbf{x}_t = \mathbf{A}_t\mathbf{x}_{t-1} + \mathbf{B}_t\mathbf{u}_t + \boldsymbol{\varepsilon}_t(\mathbf{R}_t) \tag{6}$$

where, $\mathbf{x}_t = [X_t \;\; Y_t \;\; \dot{X}_t \;\; \dot{Y}_t]^T$, is the state vector of TG, T is the time interval between measurements, and

$$\mathbf{A}_t = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \; \mathbf{B}_t = \begin{bmatrix} T^2/2 \\ T^2/2 \\ T \\ T \end{bmatrix}, \; \mathbf{u}_t = a$$

corresponds to the state transition model, the control-input model, and the control input, respectively.

The idea is to generate a motion model with a small and constant (Gaussian) acceleration with mean value of a and a standard deviation of σ_a, under the assumption that the target boat moves at speeds whose values lie in a narrow range (ideally at constant velocity). Although this is a strong assumption, the system adjusts to changes in speed in a recursive manner given that it also incorporates the value of the last measurement in the estimation of the true state of the boat, and thus, updating the position and the velocity in the *XY*-plane (Fig. 3). It may also be noticed that although the mean value of the acceleration is small, it has a relatively high standard deviation (uncertainty of the motion model) that allows the system to adjust its output accordingly, giving the appropriate weight (in the EKF process) to the predictive (motion) model and the measurement update, depending on the uncertainty associated with the current measure of the position of the target boat.

The noise covariance of the motion model is given by:

$$\mathbf{R}_t = E[\mathbf{x}\mathbf{x}^T] = E[[X_t \;\; Y_t \;\; \dot{X}_t \;\; \dot{Y}_t]^T [X_t \;\; Y_t \;\; \dot{X}_t \;\; \dot{Y}_t]] \tag{7}$$

where E[.] is the expected value. In order to determine the noise covariance $\boldsymbol{R}_t$, the acceleration a is replaced by σ_a in the second term of the right hand side of Eq. (5) to get the standard deviations corresponding to the position and velocity of the target boat:

$$\sigma_X = \frac{T^2}{2}\sigma_a, \;\; \sigma_Y = \frac{T^2}{2}\sigma_a, \;\; \sigma_{\dot{X}} = T\sigma_a, \;\; \sigma_{\dot{Y}} = T\sigma_a \tag{8}$$

Using (7) and (8), and neglecting the cross-correlation terms between X and Y coordinates (independent random variables assumption), we obtain:

$$\mathbf{R}_t = \begin{bmatrix} T^4/4 & 0 & T^3/2 & 0 \\ 0 & T^4/4 & 0 & T^3/2 \\ T^3/2 & 0 & T^2 & 0 \\ 0 & T^3/2 & 0 & T^2 \end{bmatrix} \sigma_a^2 \tag{9}$$

2.2. Measurement model

According to Eq. (2), the measurement model has to be defined in terms of the state vector. For this, one can express the absolute position of TG in terms of a rotation matrix **Rot**, the relative position of TG with respect to CH $\mathbf{T}_v$ (provided by the stereovision system), and the absolute position of CH $\mathbf{T}_{ch}$, as described by Eq. (10), and depicted in Fig. 3:

$$\mathbf{T}_t = \mathbf{RotT}_v + \mathbf{T}_{ch}$$
or
$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \cos\psi_{ch} & -\sin\psi_{ch} \\ \sin\psi_{ch} & \cos\psi_{ch} \end{bmatrix} \begin{bmatrix} x_v \\ y_v \end{bmatrix} + \begin{bmatrix} X_{ch} \\ Y_{ch} \end{bmatrix} \tag{10}$$

where ψ_{ch} and (X_{ch}, Y_{ch}), is the heading angle and the N-E absolute coordinates of the chase vehicle (from GPS on-board), respectively; (x_v, y_v) is the relative position of TG with respect to CH (from stereovision camera), and (X_t, Y_t) represents the state vector (without the velocity components, since it is not part of the measurements) and corresponds to the absolute position of TG.

One can use Eq. (10) replacing (x_v, y_v) for $\mathbf{z}_t$, to express the measurement model in terms of the state vector of the target boat, as in Eq. (2), to obtain:

$$\mathbf{z}_t = \underbrace{\begin{bmatrix} \cos\psi_{ch} & \sin\psi_{ch} \\ -\sin\psi_{ch} & \cos\psi_{ch} \end{bmatrix} \begin{bmatrix} X_t - X_{ch} \\ Y_t - Y_{ch} \end{bmatrix}}_{h(\mathbf{x}_t)} + \boldsymbol{\delta}(\mathbf{Q}_t) \tag{11}$$

According to Eq. (4), Jacobian $\mathbf{H}_t$ can be computed as follows:

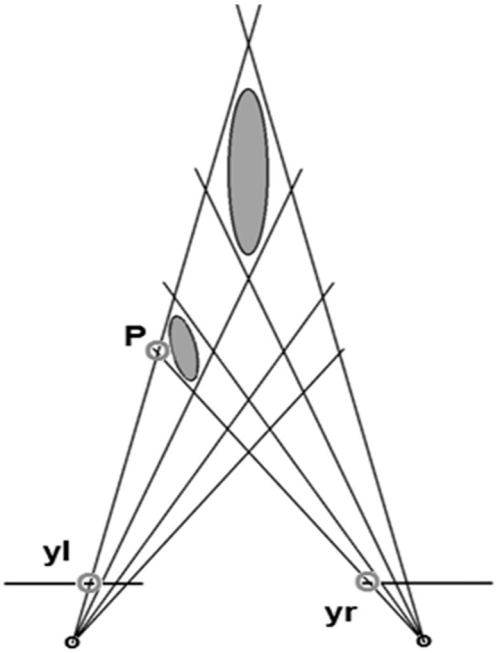


Fig. 4. Projection error of some point P in space.

$$\mathbf{H}_t = \frac{\partial h(\mathbf{x}_t = \bar{\boldsymbol{\mu}}_t)}{\partial \mathbf{x}_t} = \begin{bmatrix} \frac{\partial x_v}{\partial X_t} & \frac{\partial x_v}{\partial Y_t} & \frac{\partial x_v}{\partial \dot{X}_t} & \frac{\partial x_v}{\partial \dot{Y}_t} \\ \frac{\partial y_v}{\partial X_t} & \frac{\partial y_v}{\partial X_t} & \frac{\partial y_v}{\partial \dot{X}_t} & \frac{\partial y_v}{\partial \dot{Y}_t} \end{bmatrix} = \begin{bmatrix} \cos\psi_{ch} & \sin\psi_{ch} & 0 & 0 \\ -\sin\psi_{ch} & \cos\psi_{ch} & 0 & 0 \end{bmatrix} \tag{12}$$

In order to determine the uncertainty of the physical measurements computed by the stereo process (represented by the covariance matrix $\mathbf{Q}_t$), we first have to model the uncertainty of the pixel correspondences between the left and right images acquired by the stereo camera, or in other words, the error implicit in computing the disparity map. These miscorrespondences between pixels representing the same point in space, are modeled as a zero-mean Gaussian error distribution, with covariance matrix $\mathbf{N}_t$, as shown in Eq. (13), which is then used to derive the error distribution of the physical coordinates of the target boat. For example, suppose a physical point P is projected in both cameras at y_l and y_r, as shown in Fig. 4; due to intrinsic errors in the measurement, the stereovision system determines pixel coordinates y_l and y_r with some inaccuracies. There can be multiple sources of error at the pixel level, including image quantization, photometric and geometric distortions in the camera, and the effects of perspective distortion of the stereovision correspondence algorithm (Matthies and Shafer, 1987). This implicit error causes the true physical coordinates of a point P to be interpreted with some error located in the region adjacent to point P. In order to account for this uncertainty in the stereo-measurement, one could approximate this area to a multivariate normal distribution, such as the one represented by the elliptical iso-contour at the nominal standard deviation, in Fig. 4.

This would imply, however, that the relationship between the physical coordinates of point P and that of its projection in the image domain (in pixel coordinates), and vice versa, would be governed by a linear equation, whereas the stereovision triangulation equations for an ideal geometry are non-linear. Instead, we consider some point P projected in the right and left images represented as:

$$\mathbf{p} = [x_l \quad y_l \quad x_r \quad y_r]^T + N(0, \mathbf{N}_t) \tag{13}$$

where, (x_r, y_r) and (x_l, y_l) represent the pixel coordinates for the right and left images, respectively; and $N(0, \mathbf{N}_t)$ represents a Gaussian noise with zero mean and covariance matrix $\mathbf{N}_t$, defined as:

$$\mathbf{N}_t = \begin{bmatrix} \mathbf{V_r} & \mathbf{0}_{2x2} \\ \mathbf{0}_{2x2} & \mathbf{V_l} \end{bmatrix}, \text{ where } \mathbf{V_r} = \mathbf{V_l} = Q_f \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{14}$$

where a variance Q_f of 1 $pixel^2$ is used as the initial assumption.

In order to project the uncertainty of the measurements from the image domain to the physical domain, we make use of the stereo-triangulation equations for the horizontal plane, used in Bouguet's algorithm (Bradski and Kaehler, 2008):

$$S(\mathbf{p}) = \begin{bmatrix} x_{tg} \\ y_{tg} \end{bmatrix} = \begin{bmatrix} \frac{fT_y}{C_y - C_y^* - d} \\ \frac{T_y(y_l - C_y)}{C_y - C_y^* - d} \end{bmatrix} + N(0, \mathbf{Q}_t) \tag{15}$$

where x_{tg} and y_{tg} correspond to the horizontal relative coordinates of TG with respect to CH, (actually, it is with respect to the stereovision camera, but we have done this generalization in order to simplify the computations), $\mathbf{Q}_t$ is the same as in (11), f is the focal length of the camera in pixel units, T_y is the baseline between the cameras in centimeters (cm), d is the disparity value in pixel units, measured as $d = y_l - y_r$, C_y, and C_y^* represent the y coordinate of the principal point in the left and right images, respectively. Notice that the image coordinate y corresponds to the horizontal direction in the image plane, consistent with the physical coordinate y_v attached to the camera (Fig. 3).

Since triangulation is a nonlinear operation, the true distribution of the inferred physical coordinates is not Gaussian (Matthies and Shafer, 1987). Therefore, first order Taylor approximation is performed in order to get a linear representation of $S(\mathbf{p})$ in Eq. (15), which leads to the following approximation for $\mathbf{Q}_t$:

$$\mathbf{Q}_t = Q_f \mathbf{J} \mathbf{N}_t \mathbf{J}^T \tag{16}$$

where covariance matrix $\mathbf{N}_t$ has been approximately mapped to $\mathbf{Q}_t$, according to the notion of linear transformation of multivariate Gaussian variables, and the pixel variance Q_f has been factorized out. Also, $\mathbf{J}$ represents the Jacobian which can be expressed as:

$$\mathbf{J} = \frac{\partial S}{\partial \mathbf{p}} = \begin{bmatrix} \frac{\partial x_v}{\partial y_l} & \frac{\partial x_v}{\partial y_l} & \frac{\partial x_v}{\partial x_r} & \frac{\partial x_v}{\partial y_r} \\ \frac{\partial y_v}{\partial x_l} & \frac{\partial y_v}{\partial y_l} & \frac{\partial y_v}{\partial x_r} & \frac{\partial y_v}{\partial y_r} \end{bmatrix}$$

Solving for J, we obtain:

$$\mathbf{J} = \begin{bmatrix} 0 & \frac{fT_y}{D} & 0 & -\frac{fT_y}{D} \\ 0 & \frac{T_y(y_r - C_y^*)}{D} & 0 & -\frac{T_y(y_l - C_y)}{D} \end{bmatrix} \tag{17}$$

where $D = (C_y - C_y^* + y_r - y_l)^2$.

Replacing Eq. (17) in (16) and simplifying, we obtain:

$$\mathbf{Q}_t = Q_f \begin{bmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{bmatrix} \tag{18}$$

where:

$$Q_1 = \frac{2f^2 T_y^2}{D^2} \qquad Q_2 = \frac{fT_y^2}{D^2}(y_r - C_y^* + y_l - C_y)$$

$$Q_3 = Q_2 \quad Q_4 = \frac{T_y^2}{D_2}((y_r - C_y^*)^2 + (y_l - C_y)^2)$$

2.2.1. Accuracy of stereovision camera

One common characteristic of stereovision cameras is the decreasing level of accuracy with distance, which can be verified from the accuracy plot in Fig. 5, and Table 3, provided as a webpage Excel document by the manufacturer (Point Grey, 2016) as a guideline. These results depend on a set of parameters as the ones shown on Table 4, which define the values of our system.

From Fig. 5, one can see how the error in the stereo measurements tend to increase rapidly after 15 m, ending up with an error of 3.4 m at a distance of 16.9 m, corresponding to a disparity value of 1 pixel, according to Table 3. This represents an image quantization limitation that is reflected in the ranging capacity of the stereo camera, and thus,

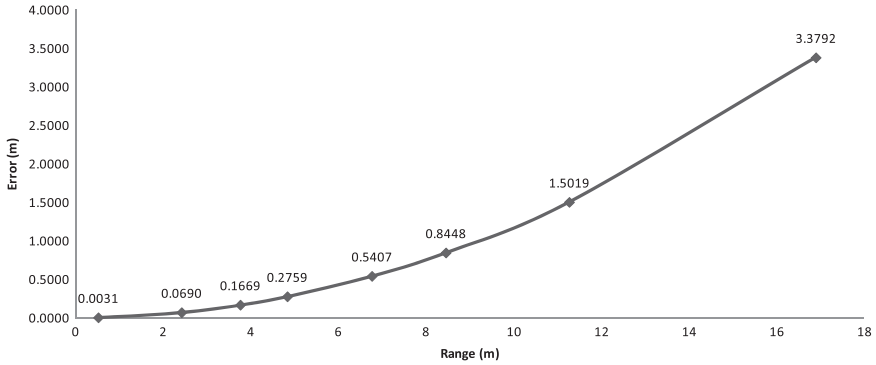


Fig. 5. Accuracy plot for Bumblebee 2 stereovision camera.

Table 3
Accuracy values of Bumblebee2 stereo camera.

Disparity	Z (m)	Z accuracy (m)
1	16.89600000	3.3792
1.5	11.26400000	1.5019
2	8.44800000	0.8448
2.5	6.75840000	0.5407
3.5	4.82742857	0.2759
4.5	3.75466667	0.1669
7	2.41371429	0.0690
33	0.51200000	0.0031

Table 4
Stereo accuracy parameters.

Stereo accuracy parameters	Values	Description
Baseline	12 cm	Distance between cameras of the Bumblebee2
Lens	2.5 mm	Nominal focal length for both cameras
Stereo resolution	320 pixels	Width of the image for stereo calculation
Correlation accuracy	0.2 pixels	Equivalent to a two standard deviations accuracy in the computation of disparity values.

an operational ranging limit of 16 m was set as the maximum acceptable stereo measurement in our system.

Recalling Section 2.2, the error modeling regarding the pixel matching accuracy, was assumed to have a variance Q_f of 1 $pixel^2$. This however, seemed to be a strong assumption and further analysis had to be made.

In order to work around this issue, a more precise discussion of the accuracy of this particular stereovision camera is described in the Technical Application Note (Point Grey, 2015), from the manufacturer of the stereovision camera. According to this, for a 320×240 stereovision resolution, using a stereovision mask of size 11 pixels, the standard deviation in disparity values is of 0.11 pixels. This implies that two standard deviations (0.22 pixels) captures 95% of a Gaussian probability density function. This value is referenced as the Correlation Accuracy (m).

Knowing that the variance of the disparity value (m^2) corresponds to the sum of the variances (Q_f) associated to each horizontal pixel coordinate (y_l and y_r), given by the stereovision correspondence process, and assuming these variances are equal, one can compute their value as:

$$Q_f = \frac{m^2}{2} \tag{19}$$

3. Simulations

Here, the results of the tracking system under three different simulated scenarios are described. Each scenario comprises of a different parameterized trajectory, for both, CH and TG, along with their corresponding motion conditions, which imply a series of speed changes govern by a constant acceleration motion model. This approach allows us to evaluate the response of the system to changes in the relative distance between both vehicles under different motion conditions.

Having generated the trajectory of both CH and TG, the relative and absolute position of the vehicles in motion are known, and serve as the reference values in the discussion below. These values can also be re-projected into the image domain in order to compute the pixel coordinates y_l and y_r as well as the disparity value, using the stereo-triangulation Eq. (15). These quantities, however, correspond to an ideal situation, and therefore a Gaussian noise is added in order to resemble actual values. This noise is parameterized with zero mean and a variance of Q_f according to Eq. (19). Subsequently, the relative coordinates between the vehicles are recomputed using these perturbed values of y_l and y_r, and Eq. (15).

At this point, the relative distance between both vehicles is known. Since the position of TG is referenced with respect to the absolute coordinates of CH, it is also required to include Gaussian noise to what emulates the GPS North-East coordinates of CH. This is done according to the actual accuracy specification of this particular sensor.

Once the simulated information is determined, it is provided as input to the EKF process in order to estimate the state of TG, defined by its position and velocity. In the following sections, the context of each scenario and its most important results are described.

3.1. Modification of the covariance matrix of the measurement model

Fig. 6 represents the straight-line trajectory and corresponding motion of both the vehicles. The image on the left shows CH in black and TG in yellow. Also, the operational range of the stereovision camera is shown as a black dashed line. The image on the right describes the motion of CH in black and the motion of TG in green, in terms of the speed of the vehicles versus time, generated according to a parameterized trajectory and govern by a constant-acceleration/constant-speed -motion model.

3.1.1. EKF parameters

The parameters of the motion model of TG are based on the premise that the vehicle moves at nearly constant speed for the most part during the pursuit, and that changes in acceleration will be present over small periods of time. Although this is a strong assumption, a probabilistic interaction between the motion and measurement models, implicit in the EKF (as a predictive-corrective process), allows the system to adapt to changes in speed and relative positions between the

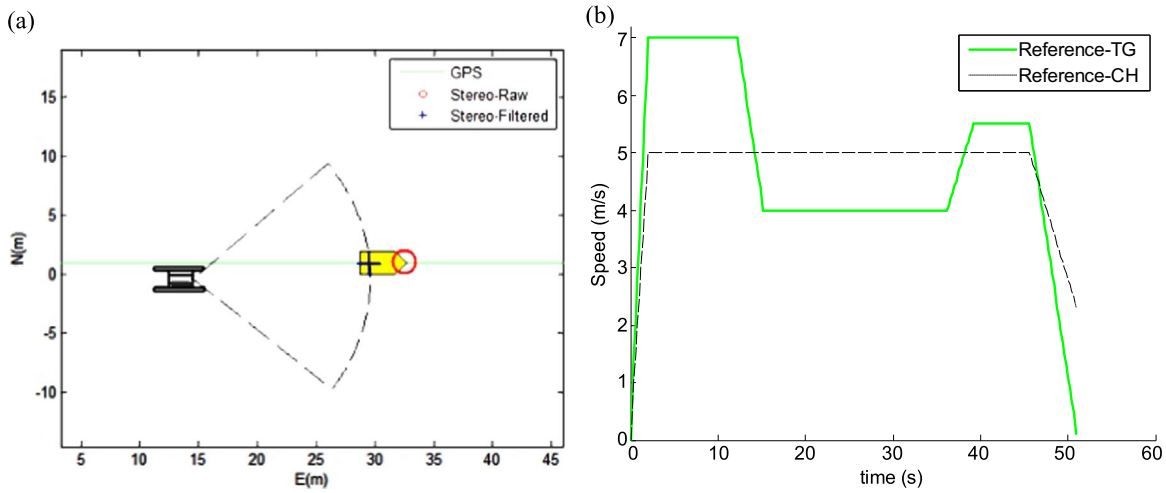


Fig. 6. (a) Straight-line trajectory for both vehicles. (b) Motion model of both vehicles. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

vehicles. The acceleration constant for the motion model is chosen to be small. Since the value of Q_f from the measurement model is already known, the value of the standard deviation of the acceleration is manually chosen in a way that optimizes the response of the system. The simulation was performed with both vehicles moving along a straight trajectory, according to the following set of parameters (Table 5):

Table 5
EKF parameters.

Parameters	Description	All speeds
dt(s)	Measurement cycle time	0.2
a(m/s^2)	Motion model acceleration	0.003
σ_a(m/s^2)	Std. deviation of acceleration	0.40
Q_f(pixel^2)	Noise variance of stereovision correspondence process	$0.1l^2/2$

In the absence of consideration of any noise, represented by N in Eq. (13), the response of the system using this set of parameters is optimal, as shown in Fig. 7(a). The estimate of the position of TG (in blue), closely matches the reference (in green), and represents a much better and more consistent result than using the raw stereovision measurements (in red), specially at larger distances. A similar satisfactory response of the system can be expected for the speed and heading estimations.

In reality, however, there are multiple sources of error that will impact the response of sensors, and therefore an estimate of the associated noise needs to be taken into account. Unfortunately, using a Gaussian estimate N for noise as in Eqs. (13) and (18) can lead to overcorrection of the raw measurements (Fig. 7(b)); the filtered output (in blue) does not match the reference (green) as well. The main problem appears to resides in the use of Eq. (18), which represents the noise-covariance matrix of the measurement model, and its incompatibility with the erratic nature of the sensor data.

In order to overcome this issue, we modify the matrix $\mathbf{Q_t}$. We eliminate the cross-correlation terms of the covariance matrix $\mathbf{Q_t}$, on the basis of independence between the x and y coordinates relative to the camera, and set the variance in the y direction to be constant and given by $k = 2m^2$, while continuing to allow the variance in the x direction to vary with position. This leads to a more stable response of the system, less sensitive to errors introduced by the sensors, using the same set of parameters and the perturbed sensor data. The modified version of $\mathbf{Q_t}$ is

$$\mathbf{Q_t} = Q_f * \begin{bmatrix} \left(\frac{2f^2 T_y^2}{D^2}\right) & 0 \\ 0 & k \end{bmatrix} \tag{20}$$

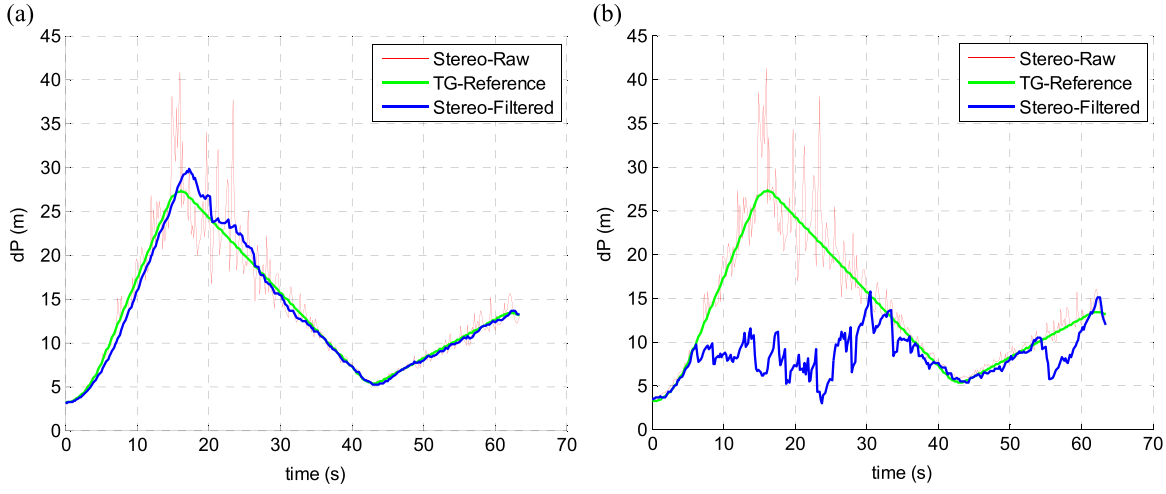
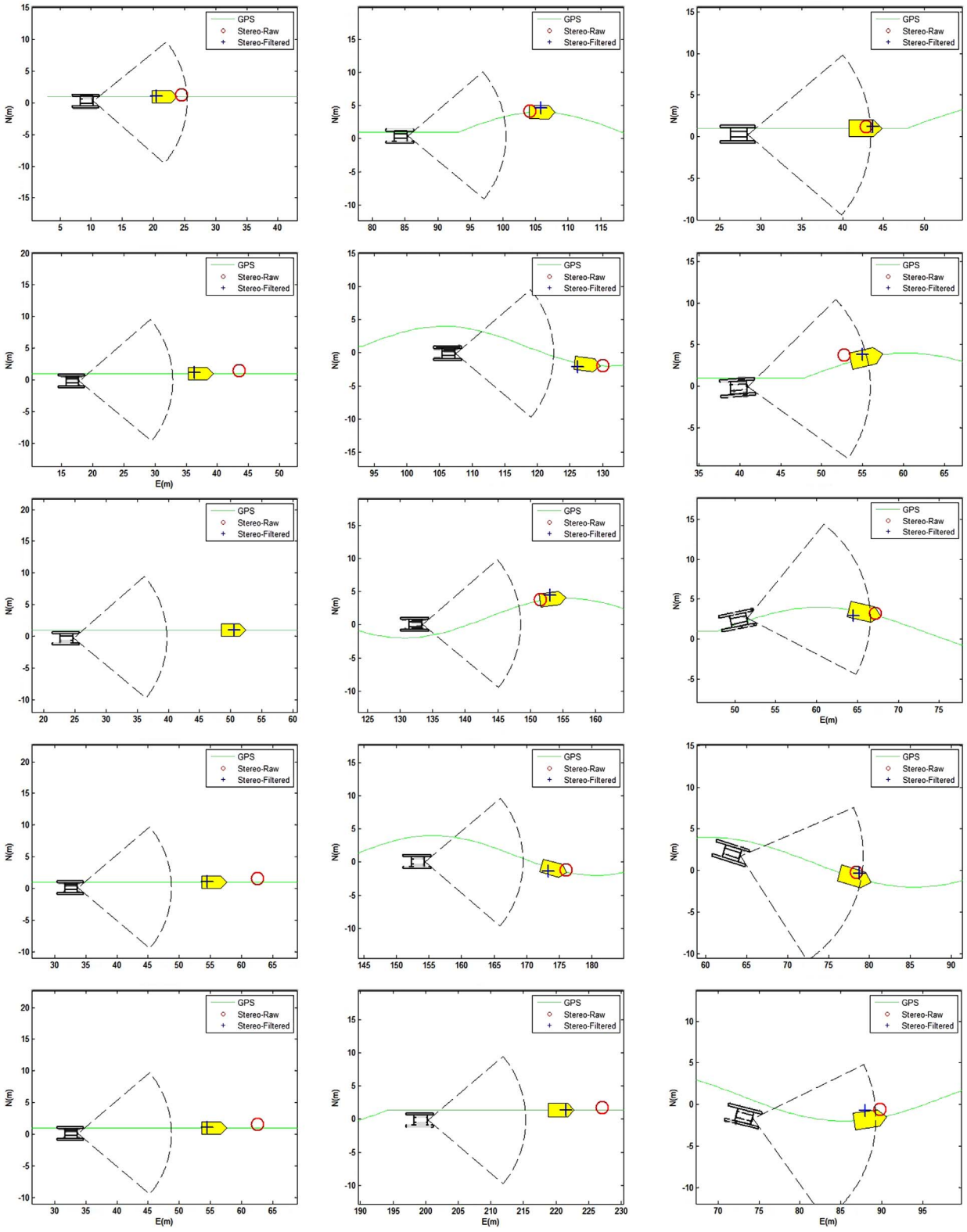


Fig. 7. Relative position of TG with respect to CH. (a) In the absence of any noise, (b) allowing for Gaussian estimate for system noise. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)



(a) Straight line trajectories for both vehicles.

(b) Straight line trajectory for CH, and sinusoidal trajectory for TG.

(c) Sinusoidal trajectory for both vehicles.

Fig. 8. Trajectories for the cases considered described in each column. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

The performance of the tracking system is analyzed in the following sections using this modified version of $\mathbf{Q}_t$, for a range of parameters in terms of the root mean square error (RMSE).

3.2. Position vs. time

Fig. 8, shows a sequence of five snapshots (from top to bottom), regarding the positions of the vehicles in the North-East frame of

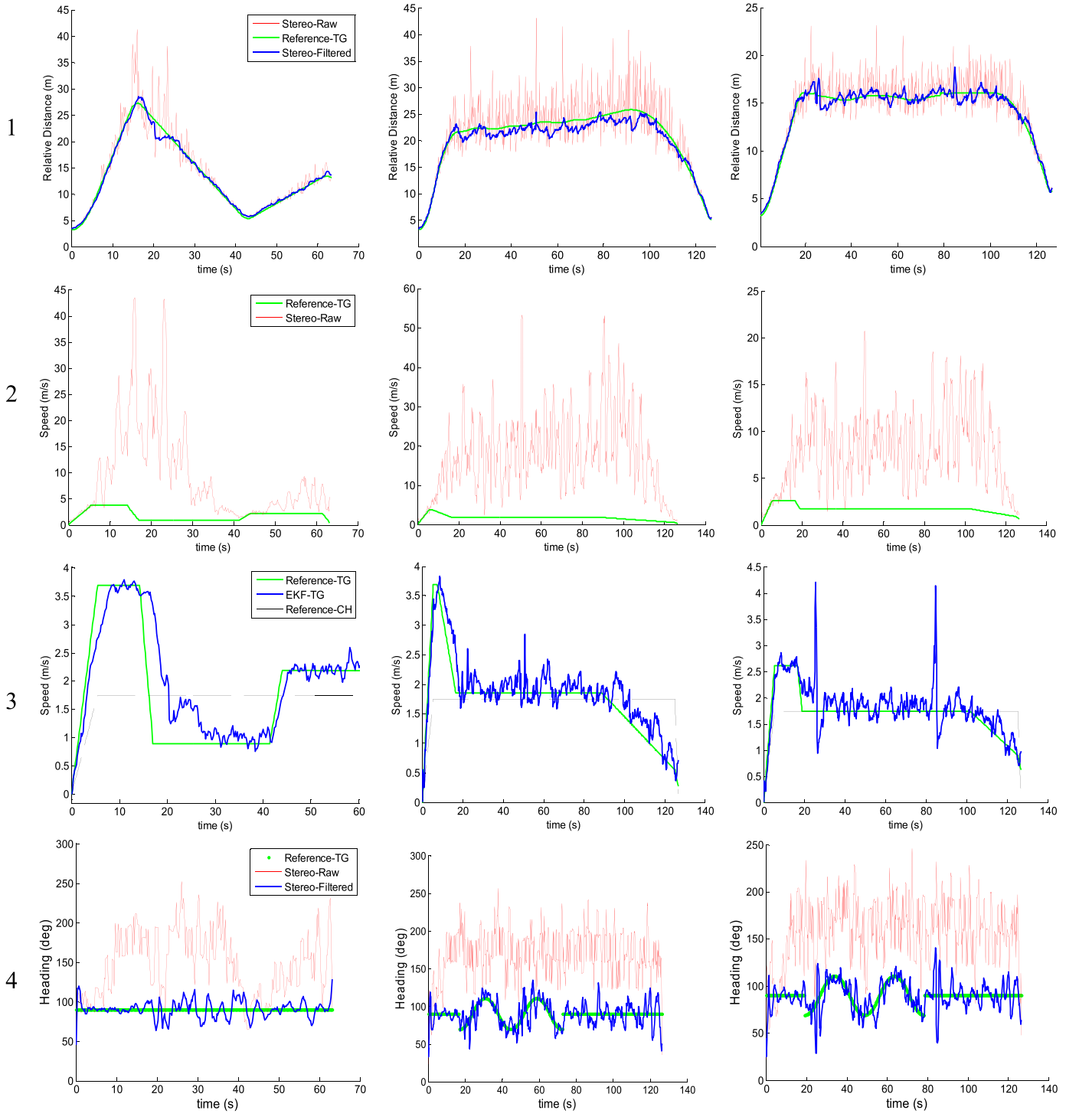


Fig. 9. Each row from top to bottom: Relative position of TG with respect to CH, speed vs. time computed with raw stereo-measurements, speed vs. time computed from filtered stereo-measurements, and heading vs. time. Every column describes each of the cases considered: a), b) and c). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

reference, for each of the three cases considered here: (a) straight line trajectories for both vehicles, (b) straight line trajectory for CH and sinusoidal trajectory for TG, and finally, (c) sinusoidal trajectories for both vehicles. CH and TG are represented in black and yellow respectively. The red circle and blue cross represent the stereovision raw and filtered measurements respectively. The sequence illustrates the performance of the tracking system for a common scenario in the context of following-behavior dynamics, where TG can be positioned in or out of the operational range of the stereovision camera (defined by the black dashed line). From the sequences, one can observe that even

within the operational range of the camera, for a distance close to the limit, the filtered output gives, in general, a more accurate estimation of the position of TG. This same response of the system holds even more strongly for the situation where the position of TG is far away relative to the operational limit of the camera.

Fig. 9, illustrates from top to bottom, the relative position of TG with respect to CH, the raw and filtered speeds of TG, and the heading of TG, for each of the three cases considered. Each case is addressed in a separate column.

The first row in Fig. 9, depicts the raw and filtered stereovision measurements of TG relative to CH for every case. Green points represent the ground truth values (generated by the simulated trajectory). The red line corresponds to the raw measurements, while the blue line represents the filtered measurements taken by the system. From this row, one can observe how the error of the raw stereovision measurements increases very quickly after a relative distance around 16 m, making this values no longer accurate. The filtered results, on the other hand, returns more accurate estimates of the position of TG, even when it is out of the operational range of the camera (for distances larger than 16 m). This is due to the fact that for large distances the motion model outweighs the measurement update in the EKF routine, in terms of the accuracy in the estimation of the position of TG.

3.3. Speed vs. time

The second row in Fig. 9, describe the speeds for both CH and TG as a function of time for each of the three cases considered. The speeds are computed from raw stereovision measurements (red line), as the magnitude of changes in the N-E position between consecutive time steps, using the raw stereovision measurements (after being smoothed with a moving average filter). Row 3, on the other hand, shows the speeds computed from filtered stereovision measurements. The green and black lines correspond to the reference speed profiles of TG and CH respectively, returned by the trajectory generator. The blue line represents the speed of TG computed using the velocity components returned by the EKF in the true-mean state vector. These changes in the speed of TG and CH, lead to different relative positions between both vehicles, allowing the assessment of the performance of the system at different positions, speeds and accelerations. It can be notice an important improvement in the accuracy of the speed estimation of TG when comparing rows 2 and 3.

One can also observe in the last column of row 3, two protrusive peaks as a result of a sudden (artificial) change in the heading of CH produced by the trajectory generator for the case (c) considered in this analysis, in order to transition from the straight to the sinusoidal trajectory. Therefore, this behavior must be discarded from the analysis.

3.4. Heading vs. time

Following the same notation through all the plots, last row in Fig. 9, portrays the reference heading of TG in green. The red line corresponds to the heading estimate using raw stereovision measurements (after being smoothed with a moving average filter). The blue line represents the heading of TG, estimated using the arctangent trigonometric function of the velocity components returned by the EKF in the true-mean state vector at every time step. The inadequate protrusions can be explained for the same reasons stated earlier for the corresponding speed estimate on last column in row 3. It can be evidenced the more accurate response of the blue line compared to the red line, with respect to the reference.

3.5. Root mean square errors (RMSE)

Tables 6–8, describe the RMSE for the three different trajectory cases considered in this analysis, each of them under three different speeds, with a maximum reference speed of TG of 3.6, 7 and 9 m/s, for each speed range, correspondingly (that is for slow, medium and fast speeds).

Table 6
Root mean square errors: Case 1.

Speeds Errors	Slow	Medium	Fast
Position EKF (m)	0.66	0.72	0.69
Position Raw (m)	2.42	2.79	2.44
Speed EKF (m/s)	0.57	0.79	0.79
Speed Raw (m/s)	11.27	12.16	9.67
Heading EKF (deg)	10.38	5.77	4.63
Heading Raw (deg)	71.18	54.87	45.45

Table 7
Root mean square error: Case 2.

Speeds Errors	Slow	Medium	Fast
Position EKF (m)	1.24	1.05	0.94
Position Raw (m)	3.50	2.54	2.42
Speed EKF (m/s)	0.27	0.44	0.53
Speed Raw (m/s)	18.66	11.43	8.91
Heading EKF (deg)	12.75	5.43	8.14
Heading Raw (deg)	86.17	69.72	61.07

Table 8
Root mean square error: Case 3.

Speeds Errors	Slow	Medium	Fast
Position EKF (m)	0.56	0.73	0.65
Position Raw (m)	1.51	1.47	1.35
Speed EKF (m/s)	0.32	0.66	0.71
Speed Raw (m/s)	7.28	5.29	3.8
Heading EKF (deg)	13.15	9.57	18.27
Heading Raw (deg)	77.06	53.40	44.37

Case 1 (Table 6) represents the RMSE when both vehicles travel along a straight line trajectory. Case 2 (Table 7), when CH travels straight while TG moves along a sinusoidal trajectory, and Case 3 (Table 8), when both vehicles move along a sinusoidal path.

Although most of this results were achieved with the same EKF parameters, for medium and fast speeds corresponding to Case 3, one of the parameters had to be adjusted in order to optimize the response of the system, namely $\sigma_a = 0.7\text{m/s}^2$. This implies a dependence of the system (and its optimal response) to the speed regime underlying the following behavior operation. This however did not represent a major concern in our USV test operations, given that the vehicles always run at slow speeds. However, this is definitely something important to consider when working with faster vehicles, where the corresponding adjustment has to be made to the system.

The simulations described in this section have been designed in order to resemble actual navigation conditions associated to following-behavior operations of USVs. The corresponding results suggest that the system may have a satisfactory response in terms of computing accurate estimates of the state of the target vehicle (TG), given by its position, speed and heading, using only the sensors onboard CH (the stereovision camera included).

4. Field experiments

The following set of experiments were performed using two different object detectors. The first set of experiments were based on using a color tracker, which provided a way to assess the stand alone

Fig. 10. Color tracker and disparity map output (top-right). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

operation of the system, and to track a consistent region in TG along the entire course of the experiments. The second set of experiments were performed using the TLD tracker, capable of detecting even drastic changes in the appearance of the target, thanks to an interaction between a median tracker and a window base object detector, which learns different appearances of the vehicle while storing the corresponding false and positive instances in a model that keeps growing over time, making the system more accurate every cycle.

4.1. Using the color tracker

The object tracking routine is depicted in Fig. 10, where a yellow blob on-board TG is used as the target to be tracked. Here, one can also observe the associated disparity map positioned at the top-right corner of the figure, as a result of the stereovision data processing executed simultaneously.

The modified version of the covariance matrix $\mathbf{Q}_t$, defined in Eq. (20), is used in the measurement model (just as in the simulations), where only the uncertainty in the measurement of the most critical dimension (along the x coordinate) is allowed to vary with respect to the distance of TG, relative to the position of CH.

The trajectory followed by TG is depicted in Fig. 11. Software improvements were made with respect to early implementations of the systems, leading to a stereovision measurement update rate of 15 Hz, and an overall update rate of 5 Hz, involving GPS and IMU data.

Due to the low speeds present in these tests, the acceleration parameter had to be decreased with respect to the configuration derived from the simulations in the previous section, in order to optimize the response of the system (although using the same parameters also improved the overall accuracy of the estimates). Table 9 describes the EKF parameters used for this set of tests.

Fig. 12 depicts four important periods during the operation, when both vehicles are in the limit or beyond the range of the stereovision camera (< 17 m), as can be noticed in the top-left and bottom-left images. For this experiment a threshold value (Th) in the body-fixed x direction relative to the camera, has been set to 17 m which corresponds to what is defined as the range of the stereovision camera. Raw

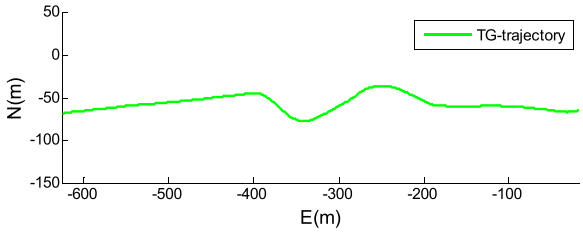


Fig. 11. Trajectory of target vehicle, test 1.

Table 9
EKF parameters.

Parameters	Description	All Speeds
dt(s)	Measurement cycle time	0.2
a(m/s^2)	Motion model acceleration	0.003
σ_a(m/s^2)	Std. deviation of acceleration	0.20
Q_f(pixel^2)	Noise variance of stereovision correspondence process	$0.11^2/2$

measurements above this value are considered highly inaccurate, and therefore the prediction of the mathematical motion model has a higher probabilistic weight in computing the estimate of the true state of TG as the output of the EKF. This can be seen from the top-right image, where even though the raw stereovision measurements (red line) significantly differ from the reference (in green), the EKF is able to compute a more accurate estimation (blue line) of the position of TG. The bottom-right image illustrates a situation where the tracker did not return any output, and thus, there is no raw measurement available of the position of TG. In this case, the system only uses the prediction of the state of the vehicle, given by the motion model, to estimate the true state of the TG.

Fig. 13 shows the relative position of TG with respect to CH along the entire trajectory. Here, each of the situations depicted in Fig. 12 is correspondingly viewed from a different perspective. An overall improvement in the estimate of the position, especially at regions near or beyond the operational range of the stereovision camera, can be discerned from Fig. 13.

Figs. 14 and 15, show the improvements in the estimates of the speed and heading of TG provided by the tracking system (in blue) using the EKF, compared to what it would be if only the raw stereovision measurements were used for this purpose (in red), even though the original (raw) data have been passed through a 15-elements moving average filter. The same smoothing filter has been applied to the heading estimates computed from raw data in Fig. 15.

The information in Table 10 reflects an overall improvement of the system in terms of the RMSE value with respect to the only reference available: the GPS (and IMU) data.

4.2. Using the TLD tracker

The following set of field tests make use of the state of the art TLD tracker (Kalal et al., 2012), and its C++ implementation (Nebehay, 2013) which provides a convenient method for detecting and tracking the target vehicle (Fig. 16). It provides the system with an instance of TG, and it adapts to changes in its appearance using a combination of computer vision and machine learning techniques, where a tracker based on Lucas-Kanade algorithm estimates the motion of the object of interest between consecutive frames and compares this output with the one from an online object detector.

As a result of this comparison and the implementation of a predefined criteria (similarity measure), the model of the object detector is continuously expanding (learning) with positive and negative examples of the object of interest, allowing the system to keep tracking TG even though its initial appearance has changed. Although the TLD tracker is capable of following TG for relatively long periods of time, it eventually incurred in some drift and lost the target. Another issue was the fact that having a bounding box representing the object of interest does not provide a consistent region to be tracked since a range of different distances can be found within the bounding box. In order to work around this issue, a predefined range around the median of disparity values corresponding to the upper half subset of the entire range of disparity values was used in order to define an area (in a binary image) representing the region enclosed by the bounding box. This happens to be closer to the camera, and thus segmenting the

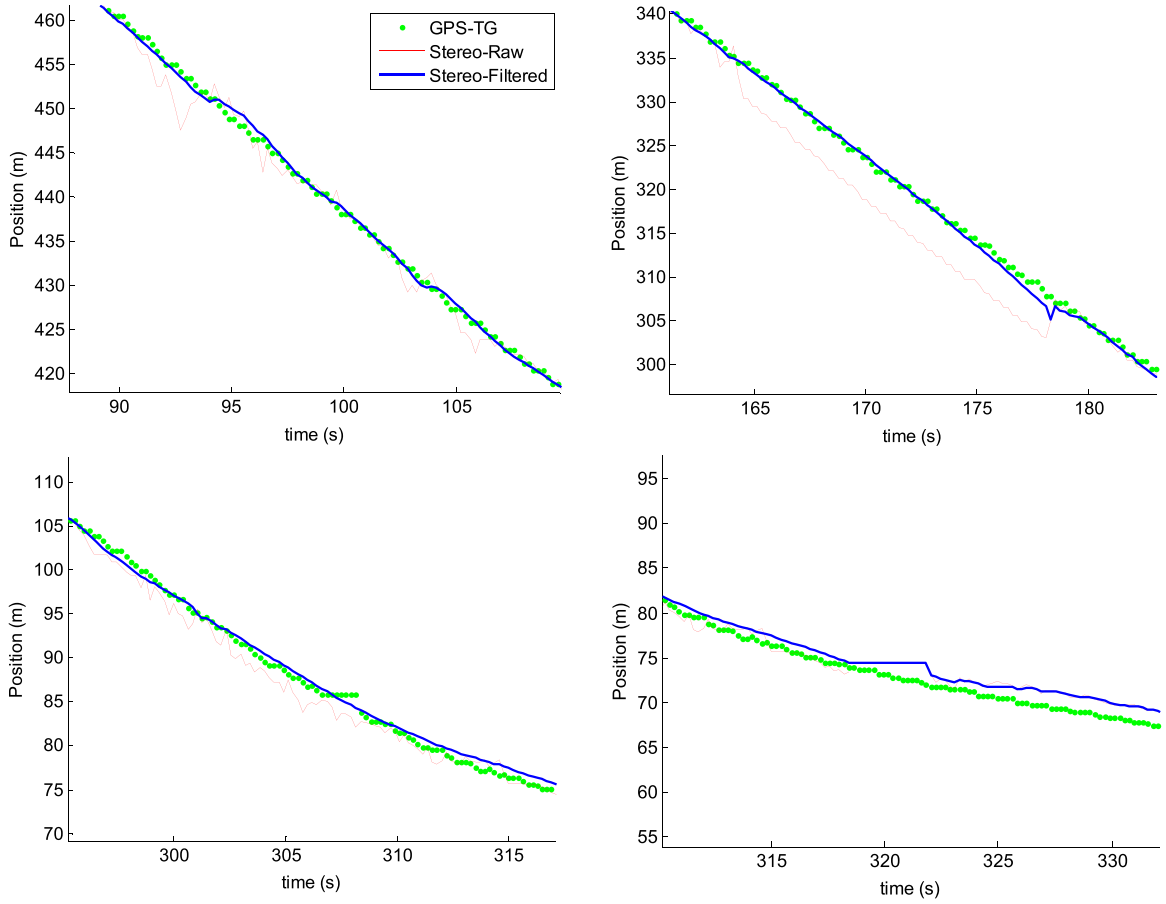


Fig. 12. Absolute position vs. time. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

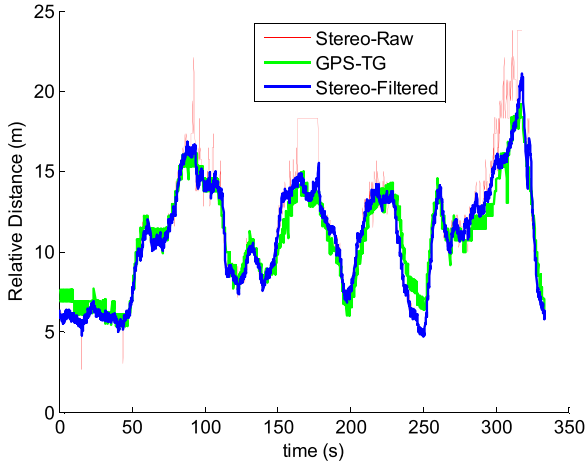


Fig. 13. Relative position of TG with respect to CH.

actual target from the background. The fact that the median of this subset of disparity values is used instead of the largest value (which corresponds to the pixel representing the physical point closer to the camera), makes this approach more robust to errors from the disparity map, and also leads to a more stable output (less jumps to large disparity values). The output from this stage is the centroid of this segmented area in pixel coordinates, representing the position of TG, as well as the median value of the upper-half subset of disparity values that define its corresponding depth (associated to the position of TG) with respect to the stereovision camera.

The trajectory followed by TG is depicted in Fig. 17, and is composed of small turns and a straight line path. The EKF parameters used in this experiment are as in Table 9. The performance of the tracking system in estimating the position, speed and heading of TG is depicted in Figs. 18–20.

The root mean square errors in estimating the state of the TG are provided in Table 11.

4.3. Other tests

The results of the root mean square errors (RMSE) computed for twelve different field tests are presented in Table 12. The main changing aspect during the experiments, was the variation in lighting conditions, not only due to the time of the day, but also due to the weather. The trajectories involved simple straight paths as well as short curves and zigzags, as shown in Fig. 23. The tests were performed in calm sea states and low speeds, based on the full capacity of the USV. The table is divided into two categories: Color Tracker and TLD tracker, each of which describes the corresponding RMSE associated with the true state of TG, namely, position, speed and heading.

Unfortunately, the position errors described in Table 12 are computed with respect to the GPS measurements on-board TG. Furthermore, the system estimates the N-E coordinates of TG relative to the corresponding coordinates of CH, which implies that the position estimation is also subject to the intrinsic error in the GPS onboard CH. Therefore, an accurate estimate of the position error requires an actual ground-truth reference. Nevertheless, an overall trend of improvement is apparent on inspection of the column (a) in Figs. 21 and 22, using the color tracker and the TLD respectively, where the distances between both vehicles in some of the tests are depicted. This can also be corroborated from the corresponding RMSE results in Table 12.

Using the TLD tracker over relative long periods of time to track an object that is constantly changing its appearance, such as the target

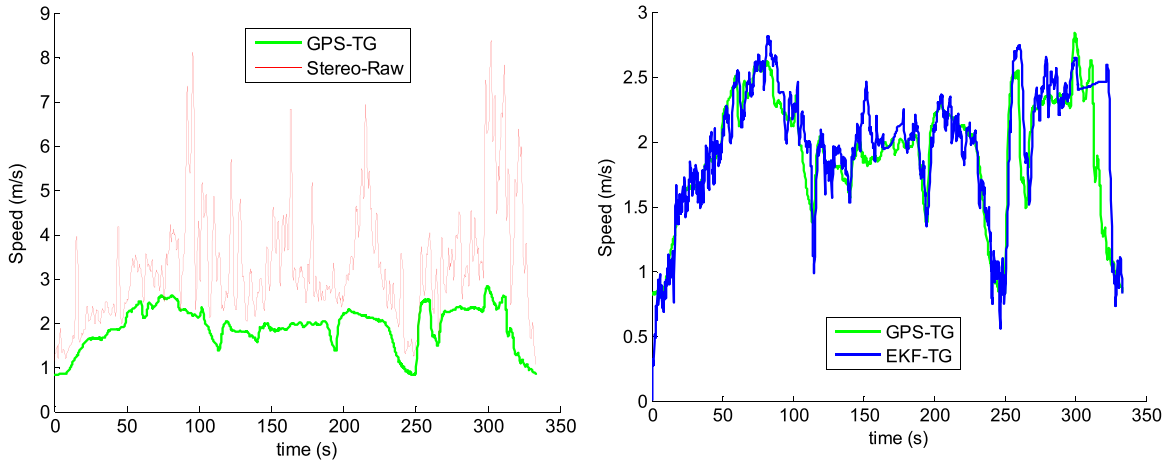


Fig. 14. Speed vs. time. raw output (left), and filtered output (right). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

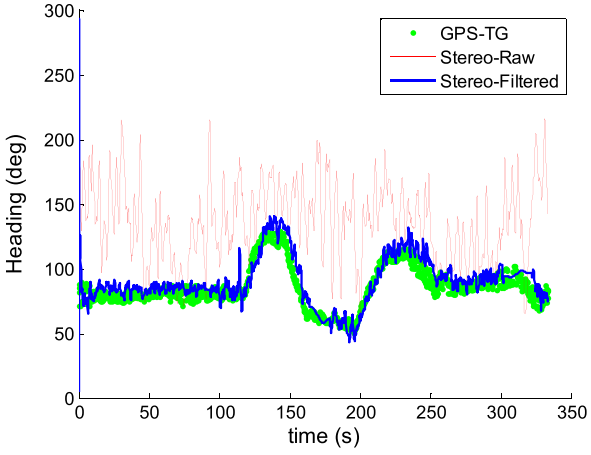


Fig. 15. Heading vs. time. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Table 10
Root mean square errors.

Position EKF (m)	0.79
Position Raw (m)	1.40
Speed EKF (m/s)	0.27
Speed Raw (m/s)	1.82
Heading EKF (deg)	8.65
Heading Raw (deg)	60.99

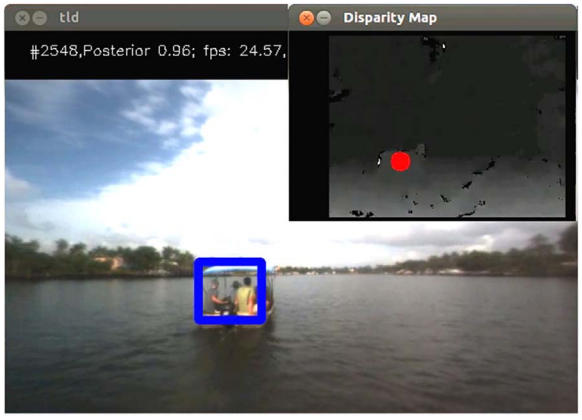


Fig. 16. TLD tracker and disparity map output (top-right).

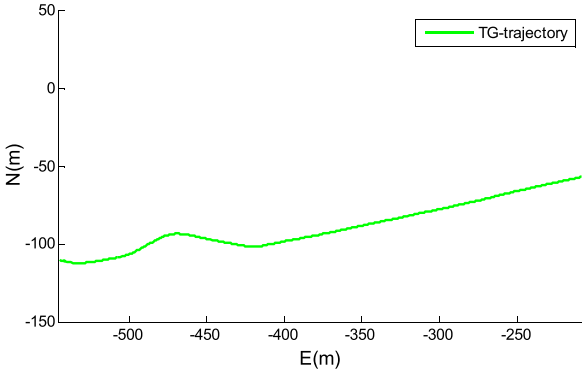


Fig. 17. Trajectory of TG.

vehicle, leads to progressive degradation of its tracking accuracy. As a consequence, the bounding box enclosing an instance of TG gets larger and eventually does not represent the region of interest anymore. Another similar problem is due to a drift in the bounding box to a region completely off the region of interest. This problem is sometimes related but not conditioned to the first problem.

These two problems were especially likely when specular reflections were present, and in general, when the camera was positioned facing the sun, as can be seeing in Fig. 24, where the left image depicts the initial position of the bounding box and the right image a subsequent (drifted) location.

This undesirable drift from the region of interest is noticeable by inspecting row 3, column (a) in Fig. 22, that is, the relative distance of TG with respect to CH plot, where the tracking accuracy progressively degrades starting at some point around 90 s.

The speed output of the system is compared against the speed estimate from the GPS on-board TG, used as the reference value. A significant improvement over the raw speed estimates is apparent by an order of 10, according to Table 12. Some of these results are also portrayed on the column (b) in Figs. 21 and 22, where the raw speed has not been depicted due to its predictably erratic and inaccurate output, as it is been demonstrated in earlier discussions.

Heading errors also showed a major improvement when compared with the filtered estimates according to Table 12, as could be expected from earlier simulation results. Some of the experimental results are shown on column (c) in Figs. 21 and 22.

5. Conclusions and discussion

A stereovision-based target tracking system for a USV has been designed and developed and tested in the field using a WAM-V USV.

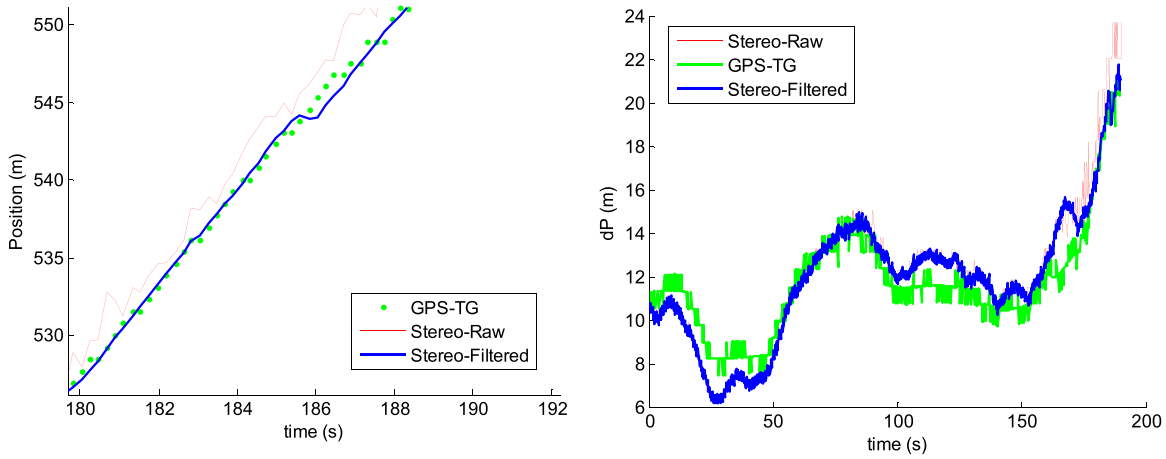


Fig. 18. Absolute position vs. time (left), and relative position of TG relative to CH vs. time (right).

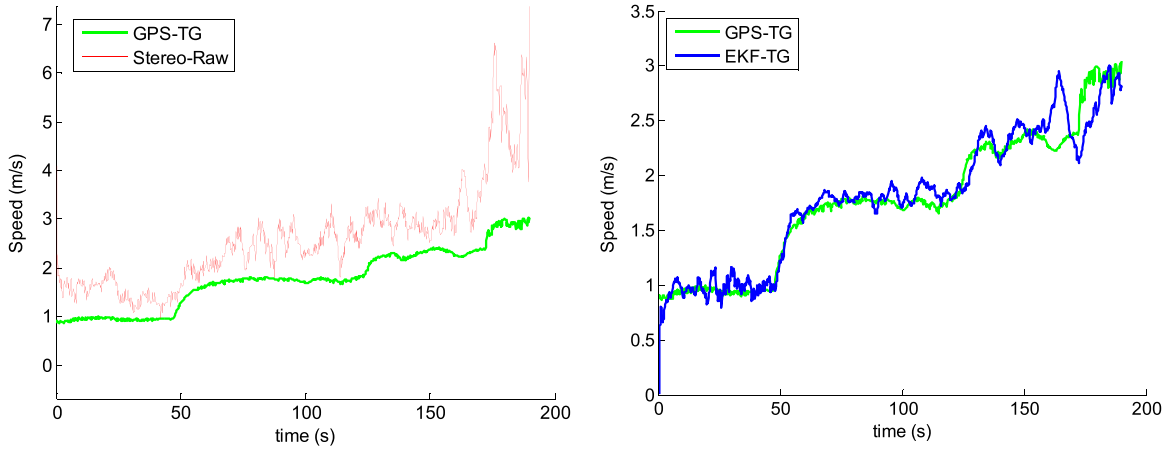


Fig. 19. Speed vs. time: raw (left), and filtered (right).

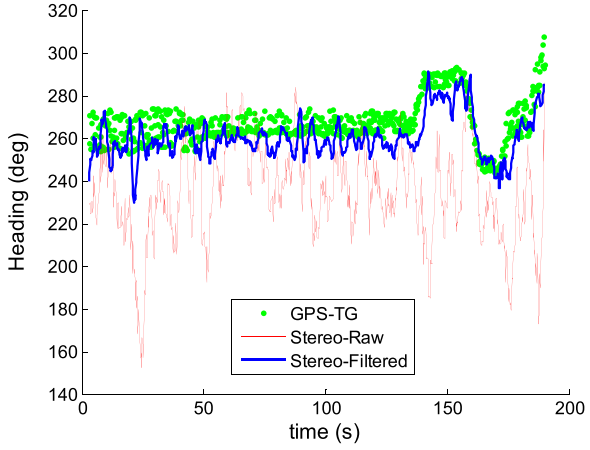


Fig. 20. Heading vs. time.

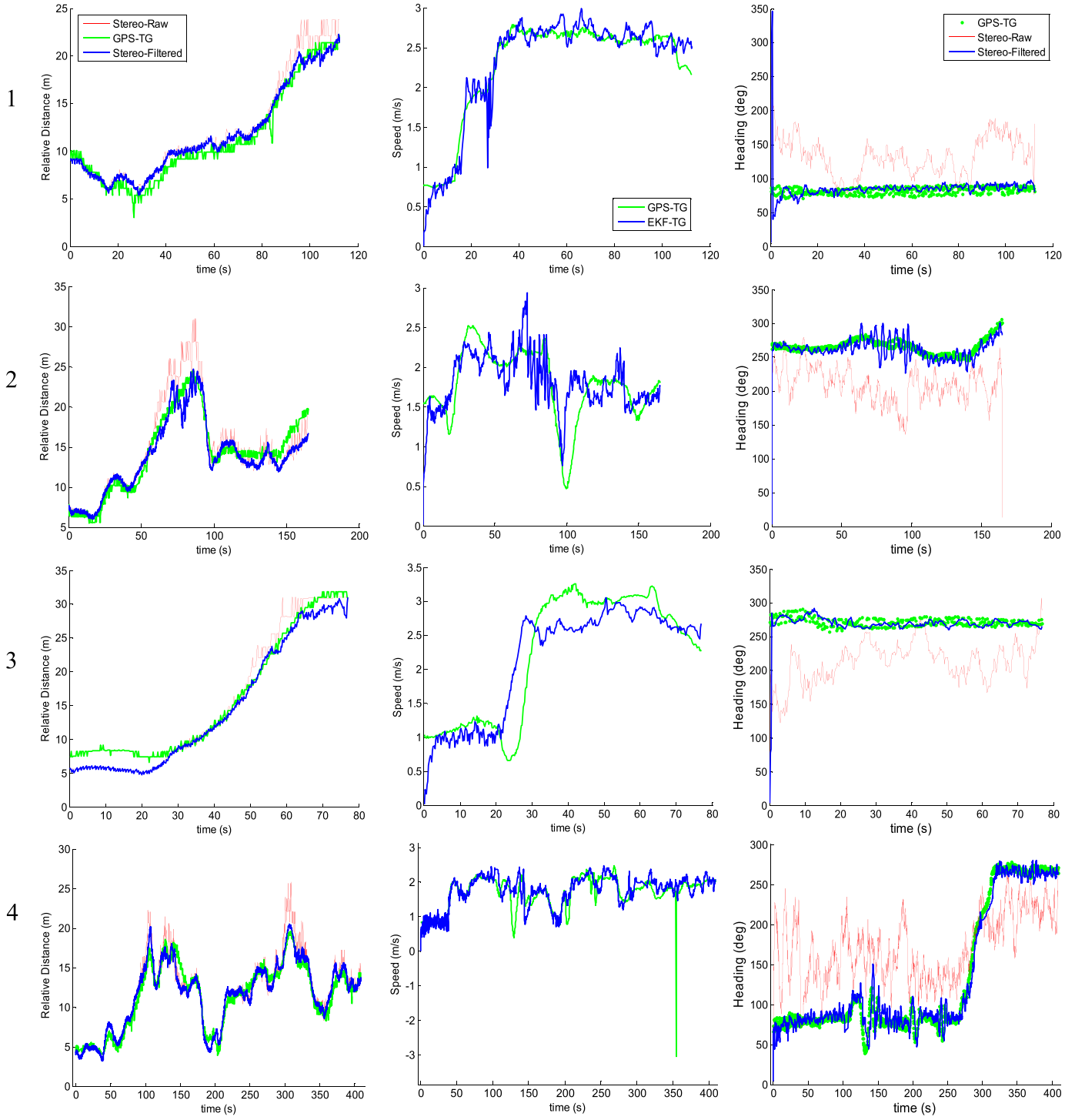
Table 11
Root mean square errors.

Position EKF (m)	1.03
Position Raw (m)	1.19
Speed EKF (m/s)	0.16
Speed Raw (m/s)	1.02
Heading EKF (deg)	10.39
Heading Raw (deg)	41.36

Table 12
Root mean square errors for different test cases.

Color tracker							
Test	Position (m)		Speed (m/s)		Heading (deg)		Time (s)
	EKF	Raw	EKF	Raw	EKF	Raw	
1	0.9	1.37	0.16	1.62	6.96	58.51	112
2	0.86	1.01	0.14	1.7	6.57	53.84	73
3	1.46	1.97	0.33	3.01	11.05	64.25	165
4	1.61	1.67	0.44	1.72	7.08	67.73	77
5	1.44	1.7	0.29	2.33	13.49	70.31	409
6	0.94	1.01	0.22	1.43	15.45	60.66	143
TLD tracker							
Test	Position (m)		Speed (m/s)		Heading (deg)		Time (s)
	EKF	Raw	EKF	Raw	EKF	Raw	
1	1.27	1.53	0.24	2.78	9.11	63.53	127
2	0.44	0.52	0.05	0.58	9.83	36.92	58
3	0.67	0.68	0.08	0.64	9.75	33.46	94
4	0.61	0.63	0.12	0.75	9.88	34.47	80
5	1.24	1.27	0.21	0.59	4.99	24.12	89
6	0.94	0.96	0.22	0.68	9.57	35.38	68

Extensive parametric simulations of target tracking have been conducted and the system has been field-tested in the marine environment using a stereovision camera mounted on a WAM-V USV in pursuit of a target boat. The tracking algorithm is based on real-time processing of stereovision data, supported by a custom extended Kalman filter for improved tracking of the target vehicle. For object detection and



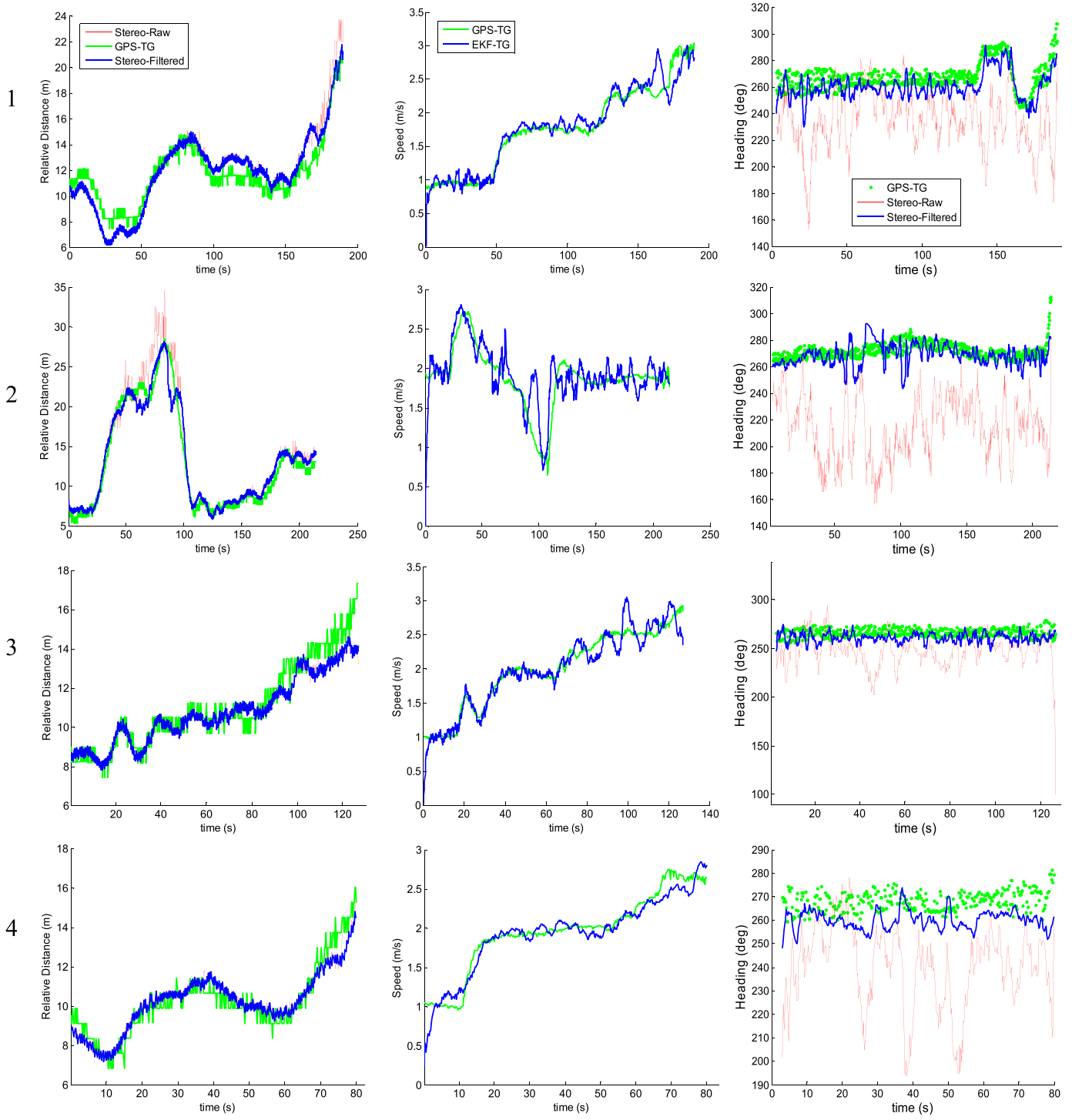
(a) Relative distance of TG with respect to CH (b) Filtered Speed Vs. Time (c) Heading Vs. Time

Fig. 21. Field tests using the color tracker.

tracking tasks, two different approaches are explored:

1. Use of a color tracker, which is a basic but effective method that allows assessment of the stand-alone response of the system.
2. Use of a TLD tracker, which is a state of the art method that combines computer vision and semi-supervised machine learning techniques, in order to track an object of interest, allowing for changes in its appearance and orientation while "learning" them.

The TLD tracker is a more versatile and robust method for tracking the target vehicle than the color tracker. However, it has limitations in tracking the target, specifically a) during extreme out-of-plane rotations of the target vehicle, and b) when the vehicle is moving directly towards the sun on the horizon, leading to specular reflections and an overall partial occlusion of the scene. These two main causes eventually lead to a drifted and sometimes over-enlarged bounding box, giving an inaccurate state of the target or a false positive.



(a) Relative distance of TG with respect to CH (b) Filtered Speed Vs. Time (c) Heading Vs. Time

Fig. 22. Field tests using the TLD tracker.

The system was implemented on a 14 ft Wave-Adaptive Modular Vessel (WAM-V) which offers a convenient set of characteristics for this application, including mechanical motion compensation through its independent suspension system that mitigates the motion of the stereovision camera during the image acquisition process.

The stereovision camera used in this work was the Bumblebee 2 from Point Grey. It is a high quality, off-the-shelf product, with a ranging distance of approximately 15 m. Higher resolution cameras may be used to increase the ranging distance.

The results obtained in this study, may be extended for a broader range of speeds and distances, as suggested by the simulation results through incorporating a faster USV and a longer-range stereovision camera. Also, important improvements can be made to the object detection and tracking tasks associated with marine applications that

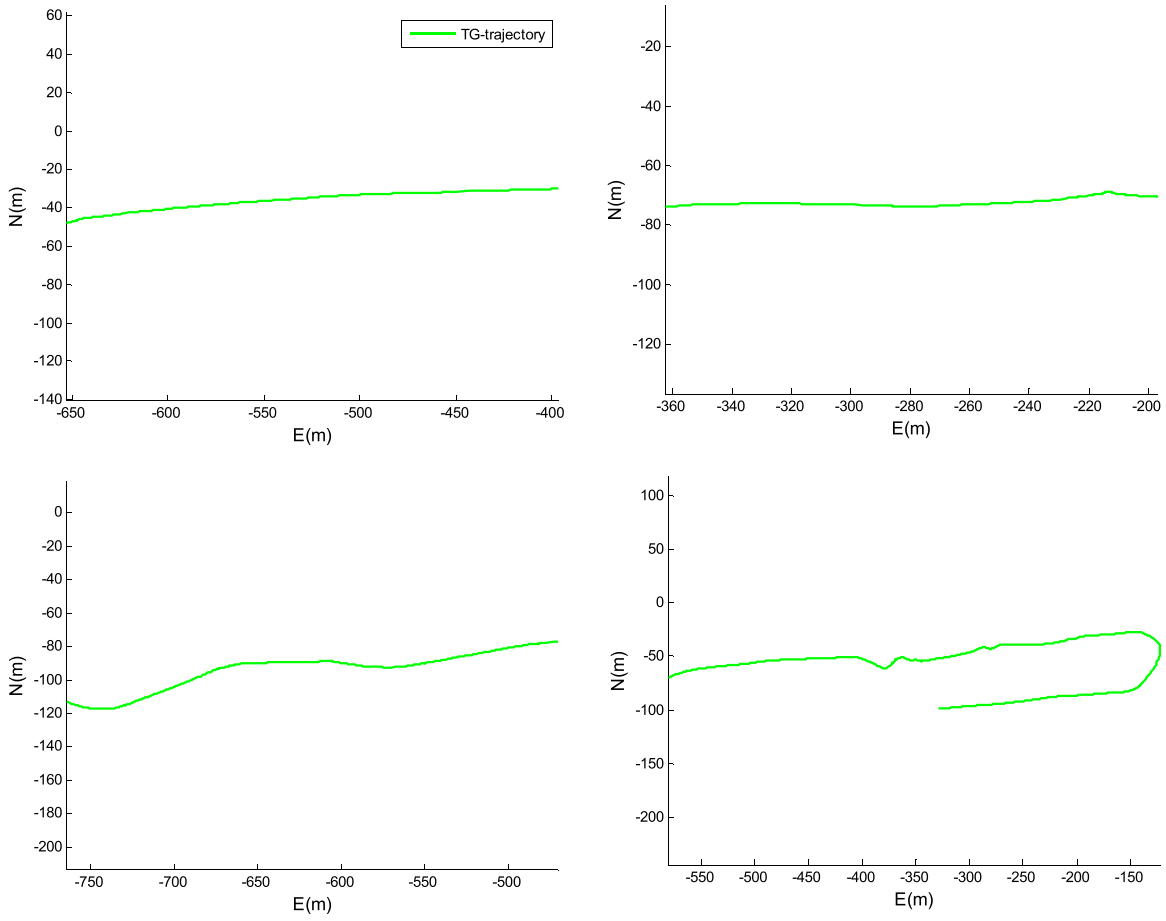


Fig. 23. Some of the trajectories followed by the target vehicle during testing.

Fig. 24. Degradation of tracking accuracy due to specular reflections.

entail dynamic environments and that are subject to high uncertainties, including partial and total occlusions, specular reflections and changes in illumination, among others.

Acknowledgements

This work was supported by the Office of Naval Research under Grant N000141110926 (Program Manager: Kelly Cooper). The authors want to give special thanks to Ivan Bertaska, Edoarda Sarda, John Frankenfield, and all the Marine Research Lab team at SeaTech, whose invaluable collaboration made the culmination of this project possible.

References

Bertozzi, M., Broggi, A., Fascioli, A., Tibaldi, A., Chapuis, R., Chausse, F., 2004. Pedestrian localization and tracking system with Kalman filtering. In: Proceedings of the IEEE Intelligent Vehicles Symposium.

Bhowmick, Brojeshwar, Bhadra, Sambit, Sinharay, Arijit, 2011. Stereo vision based pedestrians detection and distance measurement for automotive application. In: Proceedings of the Second International Conference on Intelligent Systems, Modelling and Simulation (ISMS). Kuala Lupmur.

Bouguet, J.Y., 1999. Pyramidal implementation of the Lucas Kanade Feature Tracker. Tech. Rep., Intel Corporation, Microprocessor Research Labs.

Bradski, Gary, Kaehler, Adrian, 2008. Learning OpenCV. O'Reilly Media, Inc., Sebastopol, California.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA, pp. 886–893.

Folkesson, J., Leonard, J., 2011. Autonomy through SLAM for an underwater robot.

Robot. Res., 55–70.
Goldberg, S.B., Maimone, M.W., Mathies, L., 2002. Stereo vision and rover navigation software for planetary exploration. In: Proceedings of the IEEE Aerospace Conference. Big Sky, MT.
Harris, C., Stephens, M., 1988. A combined corner and edge detector. In: Proceedings of the Alvey Vision Conference, pp. 147–151.
Heo, Y.S., Lee, K.M., Lee, S.U., 2008. Illumination and camera invariant stereo matching. Comput. Vis. Pattern Recognit. CVPR 2008.
Huntsberger, T., Aghazarian, H., Howard, A., Trotz, D.C., 2011. Stereo vision–based navigation for autonomous surface vessels. J. Field Robot. 28 (1), 3–18.
Kalal, Zdenek, Mikolajczyk, Krystian, Matas, Jiri, 2012. Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. 34 (7), 1409–1422.
Kohlbrecher, S., Von Styk, O., Meyer, J., Klingauf, U., 2011. A flexible and scalable SLAM system with full 3D motion estimation. In: Proceedings of the IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), pp. 155–160.
Larson, J., Bruch, M., Halterman, R., Rogers, J., Webster, R., 2007. Advances in autonomous obstacle avoidance for unmanned surface vehicles. AUVSI Unmanned Systems North America, Washington, DC.
Larson, J., Ebken, J., Bruch, M., 2006. Autonomous navigation and obstacle avoidance for unmanned surface vehicles. In: Proceedings of the SPIE Unmanned Systems Technology VIII, Defense Security Symposium, vol. 6230. Orlando, FL.
Lowe, D., 1999. Object recognition from local scale invariant features. In: Proceedings of the IEEE International Conference on Computer Vision.
Marine Advanced Research, 2015. 14 feet WAM-V USV. ⟨http://www.wam-v.com/14-wam-v-usv/⟩.
Mathies, L., 1989. Dynamic Stereo Vision (Ph.D. thesis). Department of Computer Science, Carnegie Mellon University, CMU-CS-89-195.
Matthies, L., Shafer, S.A., 1987. Error modeling in stereo navigation. IEEE J. Robot. Autom. 3 (3), 239–248.
Nebehay, Georg, 2013. OpenTLD. ⟨http://www.gnebehay.com/tld/⟩ (accessed 2014).
Papageorgiou, C., Poggio, T., 2000. A trainable system for object detection. IJCV 38 (1), 15–33.
Point Grey, 2015. Stereo Accuracy and Error Modeling, April 28, 2015. ⟨http://www.ptgrey.com/KB/10589⟩ (accessed 12.05.15).
Point Grey, 2016. Bumblebee2 1394a, Stereo Accuracy Chart. ⟨https://www.ptgrey.com/bumblebee2-firewire-stereo-vision-camera-systems⟩.
Ruiz, A.R.J., Granja, F.S., 2009. A short-range ship navigation system based on LADAR imaging and target tracking fro improved safety and efficiency. IEEE Trans. Intell. Transp. Syst. 10 (1), 186–197.
Sinharay, A., Pal, A., Bhowmick, B., 2011. A Kalman filter based approach to de-noise the stereo vision based pedestrian position estimation. In: Proceedings of the IEEE 13th International Conference on Computer Modelling and Simulation (UKSim), pp. 110–115.
Thrun, Sebastian, Burgard, Wolfram, Fox, Dieter, 2006. Probabilistic Robotics. The MIT Press, Cambridge, Massachusetts.
Wang, H., Wei, Z., Wang, Z., Ow, C.S., Ho, K.T., Feng, B., 2011. A vision-based obstacle detection system for unmanned surface vehicle. In: Proceedings of the IEEE Conference on Robotics, Automation and Mechatronics (RAM), September 2011, pp. 364–369.
Wang, Han, Wei, Zhuo, Ow, Chek Seng, Ho, Kah Tong, Feng, B., Huang, Junjie, 2012. Improvement in real-time obstacle detection system for USV. In: Proceedings of the 12th International Conference on Control Automation Robotics & Vision (ICARCV), pp.1317–1322.