Diver tracking in open waters: A low-cost approach based on visual and acoustic sensor fusion

Mohamed Walid Remmas, Ahmed Chemori, Maarja Kruusmaa

Diver Tracking in Open Waters: A Low-Cost Approach Based on Visual and Acoustic Sensor Fusion

Walid Remmas*

Department of Computer Systems

Tallinn University Of Technology

Ehitajate tee 5, 12616 Tallinn, Estonia

LIRMM, University of Montpellier, CNRS

161 Rue Ada, 34095 Montpellier, France

walid.remmas@taltech.ee

Ahmed Chemori

LIRMM

University of Montpellier, CNRS

161 Rue Ada, 34095 Montpellier, France

ahmed.chemori@lirmm.fr

Maarja Kruusmaa

Department of Computer Systems

Tallinn University Of Technology

Ehitajate tee 5, 12616 Tallinn, Estonia

maarja.kruusmaa@taltech.ee

Abstract

The design of a robust perception method is a substantial component towards achieving underwater human-robot collaboration. However, in complex environments such as the oceans, perception is still a challenging issue. Data-fusion of different sensing modalities can improve perception in dynamic and unstructured ocean environments. This work addresses the control of a highly-maneuverable autonomous underwater vehicle for diver tracking based on visual and acoustic signals data fusion measured by low-cost sensors. The underwater vehicle U-CAT tracks a diver using a 3-DOF fuzzy logic Mamdani controller. The proposed tracking approach was validated through open water real-time experiments. Combining acoustic and visual signals for underwater target tracking provides several advantages compared to previously done related research. The obtained results suggest that the proposed

*Corresponding author

solution ensures effective detection and tracking in poor visibility operating conditions.

Keywords: Underwater Robotics, Diver Tracking, Data-Fusion, Computer Vision, Collaborative Robotics

1 Introduction

Research on collaborative robots (Colgate et al., 1996), has brought humans and robots to share the same workspace. Combining the high accuracy and speed of robots with the expertise of humans, cobots are used to reduce high-risk and/or laborious work requiring human intervention, thus, reducing work-related accidents. Collaborative robots are used in various applications, such as in industry (Hentout et al., 2019) for manufacturing and assembling, in robotics for rehabilitation (Aggogeri et al., 2019) nursing (Robinson et al., 2014), and in space exploration (Bernard et al., 2018).

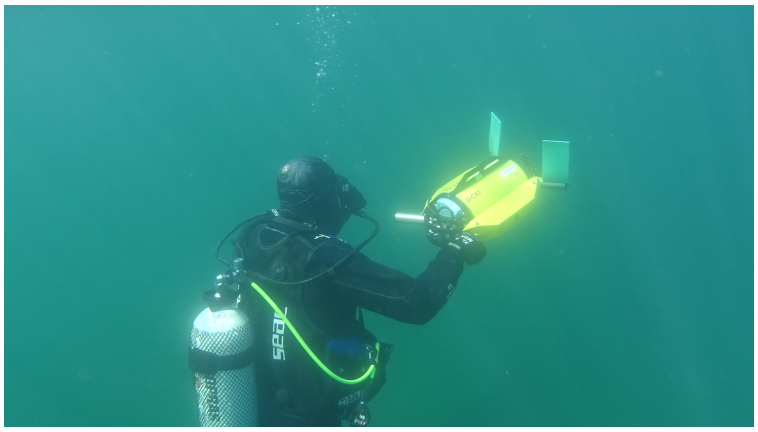

Figure 1: Picture of the U-CAT robot with a diver during lake inspection.

Underwater human-robot collaboration is another potential application in this field of research (Islam et al., 2019; Mišković et al., 2015; Gomez Chavez et al., 2019). On one side, the autonomy of underwater robots is still limited, due to the fact that most of the communication and localization technology developed for on-land applications is impractical under water. On the other side, humans have relatively limited payload capacity, limited diving time, and cannot risk to go in confined spaces. Therefore, Autonomous Underwater Vehicles (AUVs) could be used in underwater missions to help divers, carrying extra payload, collecting data and samples, taking footage of the inspected area, performing photogrammetry, and so on.

A substantial component to achieve underwater human-robot collaboration, is to detect and track a diver. Accurately tracking a diver using small underwater vehicles would be a big step into developing underwater

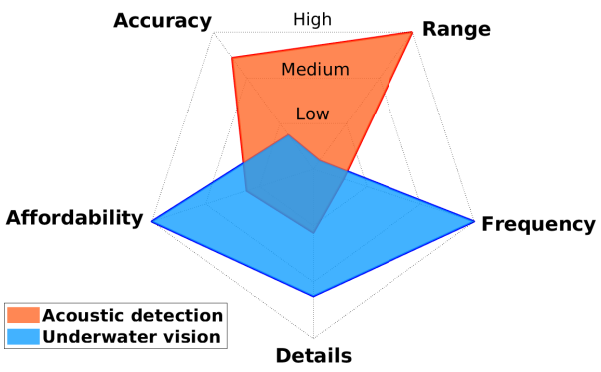

Figure 2: **Comparison of two methods of underwater perception:** Acoustic detection is more accurate and works in a wider range, while underwater vision is more affordable, has a higher sampling frequency and higher resolution.

companion robots that can be used for underwater archaeology and off-shore structures inspection (Mišković et al., 2015) (cf. Figure 1).

The research of target detection and tracking is quite mature in land applications (Ren et al., 2017; Kakinuma et al., 2012; Lekkala and Mittal, 2016; Yagimli and Varol, 2009; Zou and Tseng, 2012); however, it is still a challenging topic in underwater environments. The poor propagation of electromagnetic waves under water restricts high-bandwidth communication, which makes most of the communication and localization technology developed for on-land applications impractical in sea. Furthermore, the particular optical properties of light propagation in water (absorption and diffusion phenomena , presence of sediments, high turbidity, etc.) limits the use of vision based methods for underwater target tracking (Duntley, 1963).

A lot of research has been done in this context, using either vision-based, and/or acoustic-based methods. In spite of the limited detection range and the difficulty of feature extraction using digital cameras, there are situations where the visibility conditions are good, for example when the object is close, or when the AUV operates in calm or shallow waters. A wide variety of vision-based research has been conducted, and different features can be used to detect an object. The color as a feature was used by (Yu et al., 2001), and later by (Dudek et al., 2005) to visually guide an amphibious legged underwater robot. This technique is robust against scale and rotational variations, and partial occlusions, however, the phenomenon of color absorption in water limits the use of this technique to only clear water and close-range applications.

Shape as a feature for underwater object detection was used in (Han and Choi, 2011) and (Lee et al., 2003). However, this technique is only robust when the target has an a priori known, and invariant shape. A study

comparing other methods such as Template Matching, Weighted Template Matching, and Mean-Shift based techniques were conducted in (Kim et al., 2012). Yet, template matching techniques require good visibility to extract relevant features from the image. Many other techniques exist, for instance, optical flow can be used for underwater pipeline tracking (Cheng and Jiang, 2012), however, optical flow techniques are not well-suited for diver tracking applications, as using this method leads to the detection of all moving objects in a scene. Background subtraction techniques can also be used for detecting a moving object (Prabowo et al., 2017); nevertheless, this technique works better when having a static camera. (Sattar and Dudek, 2007; Sattar and Dudek, 2009) proposed an algorithm based on the periodic motion of a diver's flippers. However, this method requires good visibility, and close distance to the flippers. (Buelow and Birk, 2011) proposed an algorithm based on the Fourier Mellin Invariant to detect moving objects, such as divers, on a moving scene. (Chavez et al., 2015) proposed a variation of the Nearest Class-Mean Forests to detect and track divers visually. (DeMarco et al., 2013) showed the potential of sonar imagery in detecting divers based on computer vision segmentation techniques to detect moving objects in the sonar image, and processing the identified blobs using cluster classification.

Recently, enabled by the increase of computing power of GPUs' parallel architectures, neural network modeling is widely used in image processing (LeCun et al., 2015). Therefore, state of the art algorithms for on-land object detection and classification are based on deep neural networks (DNN) architectures (Redmon and Farhadi, 2018; Li et al., 2019; Cao et al., 2019; Duan et al., 2019). Many researchers are now using DNN techniques either with digital cameras for underwater target tracking (Islam et al., 2019; Shkurti et al., 2017) and multi-target tracking (Xia and Sattar, 2019; de Langis and Sattar, 2019), or with sonar imagery (Cardozo et al., 2017; Kamal et al., 2013; Lee, 2017; Song et al., 2017). Depending on the data-set on which this method is trained on, it can be robust against partial occlusions, rotations, and scale-variations. Furthermore, down-scaled DNN models have been recently proposed in the literature (Redmon and Farhadi, 2018; Howard et al., 2017). Such models can be implemented on vehicles with limited computing capabilities, for fast and reliable target (or multi-target) detection. In spite of the work and research developed for underwater target tracking, vision-based techniques still require good visibility, being in a close range to the target, and having the target inside the camera's FOV.

All the vision techniques based either on digital camera or acoustic imaging mentioned above suffer from the same issue; the diver needs to be within the sensor's field of view. This limits the diver's mobility and cognitive load as the diver has to always stay in the camera's field of view. Moreover, during underwater missions, underwater vehicles can be subject to external perturbations which might cause losing the diver

from the camera's field of view. Recovering from such a scenario has so far not been addressed.

To address this issue, we have investigated a solution which combines visual and acoustic signals for diver detection. Acoustic signals enable a wider range for underwater target detection. This solution allows to detect and track divers that can be far from the robot, or outside its camera's field of view, allowing the diver to move freely without being constrained to stay within a limited distance to the robot, and allows to recover from eventual cases where the robot loses the diver from its camera's field of view. This robust solution is also low-cost compared to previous studies, where a multi-modal diver detection scheme was used (Mandić et al., 2016; Chavez et al., 2017).

In underwater environments, the acoustic modality is generally more suitable, and can be used in various ways, such as sonar imaging, or accurately localizing a beacon. Many techniques and applications using sonar imagery have been developed in the literature. A sonar-based real-time underwater object detection using the AdaBoost method was proposed by (Kim and Yu, 2017) which uses Haar-like (Viola and Jones, 2001) features. (Zhao et al., 2009) presented a method to detect spherical shaped objects using sonar images. (Petillot et al., 2002) proposed a robust pipeline detection and tracking technique using side-scan sonars and a multi-beam echo-sounder. However, devices for sonar imaging are quite expensive, and cannot be implemented on small size AUV's. Acoustic signals can also be used for self-localization (Costanzi et al., 2017) or to localize an acoustic pinger using an array of hydrophones (Choi et al., 2017; Kasetkasem et al., 2017). Still, The drawback of using acoustic sensors is influenced by several factors, such as reflections, low operating frequencies, aquatic life disturbance, and their efficiency is highly affected when placed near the robot's mechanisms. Furthermore, the physics of sound propagation differ from one underwater environment to another. Indeed, depending on the type of the environment where the robot is operating, several factors may impact the quality and accuracy of acoustics signals. In shallow waters, reflections on the sea bottom and on the sea surface occur more often than in deep waters. Moreover, the movement of the sea surface also affects the signal and the reflections' delays. The ambient noise, such as breaking waves, rain, bubbles, and biological sources also affects the quality of the acoustic communication (Preisig, 2007; Jensen et al., 2011). All these factors need to be taken into account, and an appropriate acoustic filter needs to be used depending on the type of the underwater environment.

Acoustic perception can be augmented with machine vision techniques (Chavez et al., 2017). This work is motivated by the complementarity of acoustic and vision based object detection methods, as illustrated in Figure 2. The fusion of visual and acoustic signals removes their respective limitations, and provides advantages for diver following compared to aforementioned approaches (Mandić et al., 2016). In this context,

the goal of this work is to track a diver (or any other dynamic target) by fusing acoustic and visual signals acquired from low-cost sensors, to have a more robust, more accurate and more complete detection in open water conditions, while taking into account the underwater robot hardware capabilities. The advantage of the presented work, compared to what has already been done in the aforementioned studies, is an increased robustness towards diver tracking, where the tracking is successful, whether the diver is on the camera's field of view or not. Moreover, the proposed solution involves the use of relatively low-cost sensors, where the diver is detected visually by a digital camera using a down-scaled embedded DNN, and acoustically, by tracking a pinger carried by the diver using an array of low-cost hydrophones. The proposed diver tracking scheme is implemented on the omnidirectional underwater robot U-CAT (Salumäe et al., 2016), and validated in an open water environment.

2 U-CAT bio-mimetic AUV

Inspired by the swimming abilities of marine animals, U-CAT (Salumäe et al., 2016; Chemori et al., 2016) is an autonomous underwater vehicle (AUV) actuated by four oscillating fins. It was developed, as part of the ARROWS European project, at the Center for Biorobotics (Tallinn University of Technology) in Estonia.

U-CAT is a low-cost, resource constrained robot, that has a small size allowing it to manoeuver in confined environments (Preston et al., 2018). Its design is user-friendly, and its weight does not exceed 19 kilograms, which allows for quick and easy deployment. U-CAT can operate in depths up to 100 meters, and has a battery life of at least six hours.

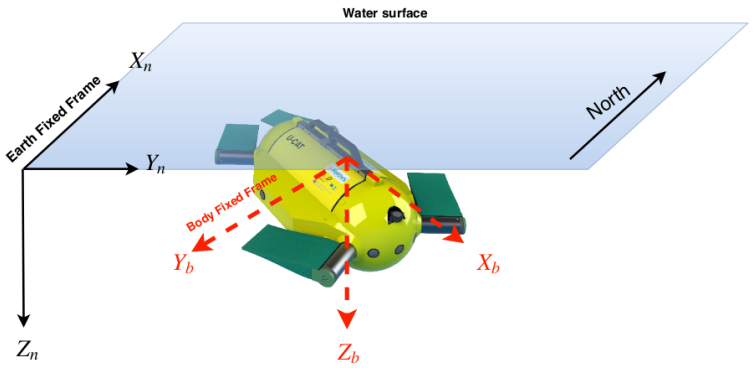

Figure 3: Illustration of the Earth Fixed Frame R_n (North East Down convention) and the robot's Body Fixed Frame R_b.

Using its four flexible fins, the robot can easily move and manoeuver in six degrees of freedom (6 DOF).

U-CAT is equipped with an MPU-9150 Inertial Measurement Unit (IMU), and low-cost perception sensors, such as an array of three *Aquarian Audio H1c* hydrophones for establishing the heading relative to an acoustic beacon, and a *PointGrey Chameleon2* camera running at 15 frames per second.

3 Proposed tracking control scheme

This section describes the proposed control scheme for diver detection and tracking using visual and acoustic signals (as illustrated in Figure 4). The visual detection of the diver is described in section 3.1.

The down-scaled DNN model tiny-YOLOv3 (Redmon and Farhadi, 2018) is used for detecting the diver visually using the on-board digital camera. This approach is chosen because it enables fast ($\approx$ 20 fps) and reliable detection on resource constrained computers. The algorithm is implemented on U-CAT's Jetson TX2 embedded computer using ROS (Bjelonic, 2018). The chosen DNN structure is a trainable model that achieves state-of-the-art performances for object detection on resource-constrained embedded systems. The data-set for the training of the model was a collection of diver images from previous experiments with U-CAT in different environments (lakes, oceans). 3000 diver selected images were manually annotated using YOLO-Annotation-tool.

The annotated images were then augmented to 8000 images by performing scale and rotation variations. Another no-diver 2000 images (also collected by U-CAT's camera in previous experiments) were added to the data-set. The performance of the model is evaluated and summarized in Table 1 based on mean Average Precision (mAP) and Intersection over Union (IoU) scores.

Table 1 shows the evaluation of the detection accuracy. Indeed, the performance of this scaled-down DNN model is low compared to other methods containing more hidden neural layers (Islam et al., 2019). The used DNN model was not compared to relative works in the literature due to the lack of a data-base that includes diver images in different water types and different visibility conditions. However, experimental results show that combining visual measurements using the implemented DNN model with acoustic measurements is enough to achieve an efficient and accurate diver tracking.

Figure 5 shows an illustration of the vision detection result. The object of interest is wrapped inside a pink bounding box. Its center's pixel coordinates are denoted (X_m, Y_m) and its width W_m. As the goal is to track the object, the objective here is to fit the detected object inside a virtual bounding box located at the center of the image whose coordinates are (X_c, Y_c), and its width is W_c.

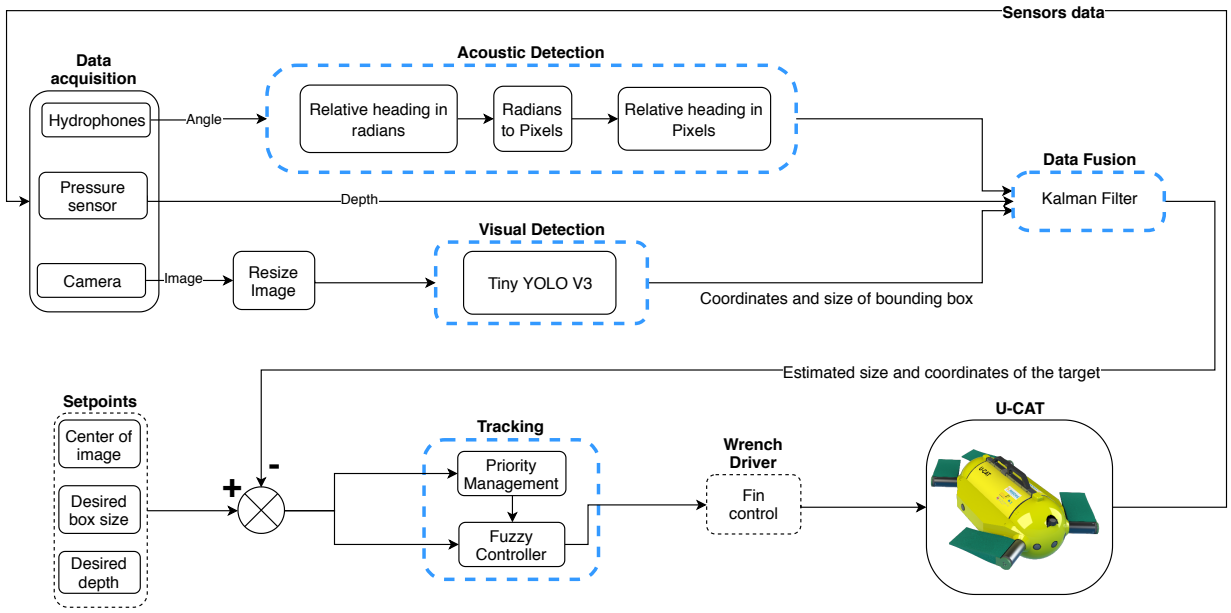

Figure 4: **Black diagram of the proposed tracking control scheme:** Dashed-blue blocks are described in section 3. The proposed tracking scheme is implemented on U-CAT using Robot Operating System (ROS), and all of the steps in this figure run online on U-CAT's embedded computer.

The acoustic detection approach is described in section 3.2. A quaternion based scheme using a low-cost array of hydrophones to detect the acoustic pinger is presented. Later on we describe the data-fusion scheme which is based on a model-free Kalman Filter (Kalman, 1960), since we do not have a mathematical model in disposal that captures the diver's free motion. Moreover, this data-fusion approach is chosen for its implementation simplicity, and its low computational cost.

Lastly, an adaptive fuzzy logic Mamdani controller for the three DOF (surge, depth and yaw) control of U-CAT is described. This approach is selected based on a previous study in which various control laws (PID, non-linear PID, sliding mode, adaptive state feedback, and fuzzy control) were tested and compared on U-CAT (Remmas, 2017). The obtained results showed that the fuzzy logic controller gives the best trajectory tracking results for heave and yaw control. The results are justified by the inaccuracies in U-CAT's dynamic model that limits the use of model-based control schemes, and the fact that fuzzy "expertise" based controller that gives the best results.

3.1 Visual detection

The errors E_C and E_W between the two boxes centers and widths are given by:

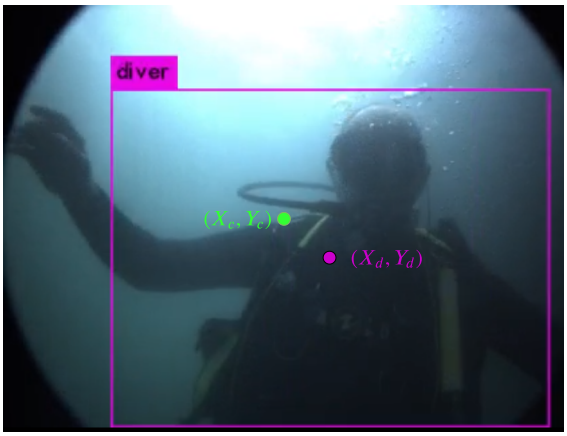

Figure 5: **Visual detection:** The goal is to center the detected box at the center of the image.

$$E_C = \begin{bmatrix} X_c \\ Y_c \end{bmatrix} - \begin{bmatrix} X_m \\ Y_m \end{bmatrix} \tag{1}$$

$$E_W = W_c - W_m \tag{2}$$

This concept allows a three DOF tracking control, such that the difference between the two boxes' centers E_C defines the desired depth and heading orientation of the robot, and the difference between the two boxes' widths E_W defines the desired distance between the robot and the target. The virtual box's width W_c is customizable and allows to define how close the robot should be to the object. In all our experiments, this parameter is chosen as $W_c = 60$. This value is selected so that the AUV keeps the diver within the camera's FOV without necessarily going too close to him.

3.2 Acoustic detection

A *Sonotronics EMT-01-3* Pinger operating at 9.6kHz is used as a beacon. It transmits a short burst signal every second. The received signal by each hydrophone is amplified and filtered. The phase shift for each hydrophone is then used to compute the relative yaw angle to the robot. The relative yaw angle between the pinger and the robot in the Earth Fixed Frame R_n (cf. Figure 3) is denoted by Ψ_p.

Considering only the yaw angle of the pinger with respect to the robot, let Q_Ψ be the quaternion representation of Ψ (yaw angle of robot in R_n) measured by the on-board IMU, and Q_{Ψ_p} the quaternion representation

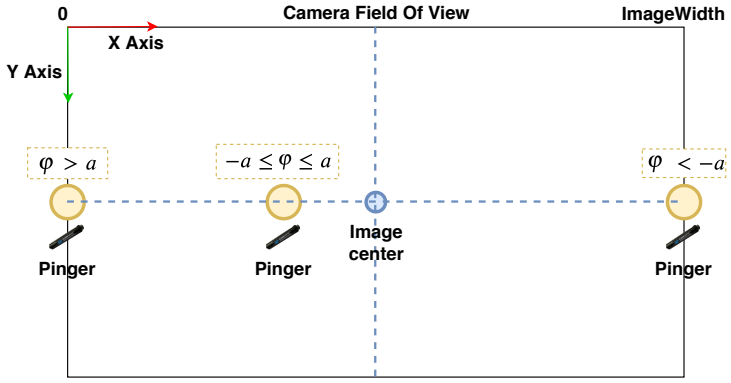

Figure 6: **Acoustic detection:** Conversion scheme from yaw angle error in radians to error in pixel coordinates in the camera frame.

of Ψ_p such that:

$$Q_\Psi = \left[cos(\tfrac{\Psi}{2}) \quad 0 \quad 0 \quad sin(\tfrac{\Psi}{2})\right]^T \tag{3}$$

$$Q_{\Psi_p} = \left[cos(\tfrac{\Psi_p}{2}) \quad 0 \quad 0 \quad sin(\tfrac{\Psi_p}{2})\right]^T \tag{4}$$

The shortest angle between these two quaternions is given by:

$$\Delta Q = [q_0, q_1, q_2, q_3]^T = Q_{\Psi_p} \otimes Q_\Psi^{-1} \tag{5}$$

where $\otimes$ denotes the quaternion multiplication, and Q^{-1} is the quaternion inverse of Q.

The quaternion angle error ΔQ is converted back to the desired yaw angle φ, that is:

$$\varphi = atan2(2q_0q_3, 1 - 2q_3^2) \tag{6}$$

Since the heading orientation to the pinger will be fused with camera measurements, the heading angle error is converted to pixel coordinates error noted $E_P = \Phi(\varphi, a)$ such that:

$$\Phi(x,a) = \begin{cases} 0 & ,\text{if} \quad x < a \\ ImageWidth & ,\text{if} \quad x > -a \\ \frac{ImageWidth}{2}(\frac{-x}{a}+1) & ,\text{otherwise} \quad ,(a \neq 0) \end{cases} \tag{7}$$

The function Φ remaps the measured pinger orientation into an image's X-coordinate pixel as illustrated in Figure 6. In our study, we consider $a = \pi/3$.

3.3 Proposed data-fusion control-scheme

One of the most popular data-fusion and estimation techniques is the Kalman Filter. It was originally proposed by Kalman (Kalman, 1960) and has been widely studied and applied since then. The used state and measurement vectors are defined as follows:

$$X_t = [x_t, y_t, w_t, z_t]^T \tag{8}$$

$$Y_t = [X_m, Y_m, W_m, Z_m, X_p]^T \tag{9}$$

where X_t is the state vector at time t. The variables x_t and y_t represent the estimated pixel coordinates of the target and w_t is the estimated width of a bounding box containing the target. The variable z_t represents the estimated depth of the robot. Y_t is the measurement vector. The variables (X_m, Y_m) and W_m represent the position coordinates and width of the detected bounding box by the camera, Z_m is the measured depth of the robot using the on-board pressure sensor, and X_p is the image coordinates of the detected heading measured by the hydrophones array.

The estimated state vector $X_t = [x_t,\ y_t,\ w_t,\ z_t]^T$ is used to compute tracking errors as follows:

$$E_x = X_c - x_t \tag{10}$$

$$E_y = Y_c - y_t \tag{11}$$

$$E_w = W_c - w_t \tag{12}$$

$$E_z = Z_{desired} - z_t \tag{13}$$

E_x is the horizontal error between the center of the frame and the estimated bounding box's center, E_y is the vertical error between the center of the frame and the estimated bounding box's center. E_w is the error

between the virtual box's width and the estimated one, and E_z is the error between a desired depth and the estimated depth of the robot.

3.4 Fuzzy logic Controller

Based on the tracking errors defined in (10)-(13), a fuzzy logic Mamdani controller (Mamdani and Baaklini, 1975) was designed to control the robot, for the aim of tracking the diver as shown in Figure 7. In this study, the controller's inputs are the tracking errors and their variations, and the output is the force vector $\tau \in \mathbb{R}^{3\times 1}$ to be applied to the robot, with $\tau = [\tau_x,\ \tau_z,\ \tau_\Psi]$.

The inputs are fuzzified using trapezoidal membership functions (Figure 7). Such membership functions may lead to a steady state error which depends on the choice of a_1 and a_2. Furthermore, this type of membership functions prevents oscillations around the target (since perfectly fitting the detected bounding box at the center of the image is rather challenging).

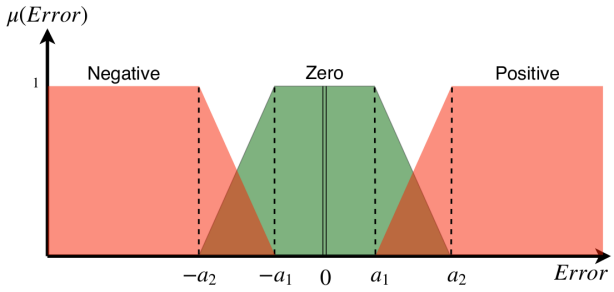

Figure 7: **Fuzzification membership functions:** a_1 and a_2 are positive constants defining the universe of discourse for each DOF.

The Rule base for our controller is defined in Table 2, where E_i and $\dot{E}_i$ $(i = \{x, y, w, z\})$ are the tracking error vector and its first-time derivative:

The output is computed using the Center Of Gravity defuzzification method, based on the membership function of Figure 8, and the forces are converted to fin oscillation directions and amplitudes by the wrench driver (Salumäe et al., 2019).

3.5 Priority management

The priority management was proposed in (Salumäe et al., 2016) as a solution to compensate for the high coupling in the actuation of the different DOFs of U-CAT. The proposed technique is based on Gaussian

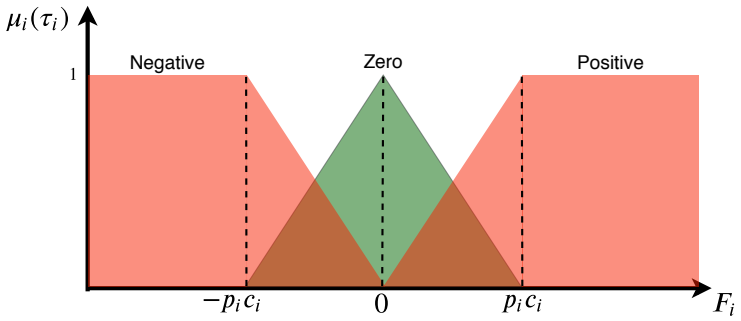

Figure 8: **Defuzzification membership functions:** c_i is the maximum achievable force or torque along i axis. p_i are weights for the control priority management (more details are given in subsection 3.5). The defuzzification process is done by using the Center Of Gravity method (COG).

membership functions used to moderate the control of each DOF depending on the control objective.

To prioritize the DOFs, the universe of discourse for the defuzzification process is modified through the multiplication of the constants c_i by the priority weights p_x, p_z, and p_Ψ (cf. Figure 8). The priority is given to yaw control, then to depth control and finally to surge control as follows:

$$p_x = e^{-\frac{E_x^2+E_y^2}{2\sigma^2}} \tag{14}$$

$$p_y = e^{-\frac{E_x^2}{2\sigma^2}} \tag{15}$$

$$p_\Psi = 1 - e^{-\frac{E_x^2}{2\sigma^2}} \tag{16}$$

Where σ^2 is the variance of the Gaussian functions.

Since U-CAT is relatively stable in roll and pitch, the control of those DOFs was not taken into consideration at this stage.

4 Experimental results in open waters

To validate the proposed control scheme, various experiments were conducted out near a small harbour in Banyuls-Sur-Mer, France (42° 28' 52.0"N, 3° 08' 10.0"E). Four experimental scenarios were tested to show the need of sensors' complementary for diver tracking, and to highlight the efficiency of the proposed tracking control scheme. In this section, a description of the experimental setup for the different scenarios will be

presented, followed by a presentation and discussion of the obtained results.

- *Visual tracking (Scenario 1):* In this scenario, U-CAT is only visually guided to track the diver. The goal of this test was to assess the robustness of the proposed vision detection and tracking algorithm in field operating conditions.
- *Acoustic tracking (Scenario 2):* In this test, U-CAT has to track the diver using acoustic sensing only. This experiment was carried out to evaluate the acoustic sensing based on an array of three hydrophones and a pinger for diver tracking.
- *Data-fusion based static tracking (Scenario 3):* In this test, the proposed data-fusion tracking scheme was tested for tracking a static target. The robot had to autonomously detect and track a pinger fixed to a colored waterproof source of light that was initially positioned far from the robot and out of its camera's field of view. A simple color segmentation detector was used to visually detect the colored source of light in this case. The same scenario was conducted three times where U-CAT started in different initial conditions; either facing the target, placed to its left, or its right.
- *Data-fusion based dynamic tracking (Scenario 4):* In this last scenario, the proposed diver tracking algorithm's efficiency was evaluated. A diver carrying the pinger was asked to move freely in open water at a known depth. The diver's depth was chosen to be two meters to ensure more reliable top view monitoring of the diver and the AUV from a static camera. This scenario was conducted three times, where the diver started at different initial locations, either close and within the robot's camera's field of view, or away and out of the embedded camera's field of view.

4.1 Visual tracking (Scenario 1)

Initially, the diver was in U-CAT's camera's FOV so that he can be detected from the beginning of the experiment. The diver then moved freely at a constant depth. The obtained results are shown in Figure 10.

Figure 10 shows that the AUV was tracking and centering the detected target on the image. After 8 seconds of tracking, the diver left the camera's field of view. The robot was tracking the last detected value; however, it failed to find the target. Furthermore, in field experiments, where the water is often turbid, and the visibility conditions are poor due to light scattering and absorption, visual detection often fails (as illustrated in Figure 9), which causes the tracking to fail. This shows the motivation behind using another sensor to complement the vision, and ensure detecting the diver when this last one's detection fails.

(a) (b)

Figure 9: **Underwater images from U-CAT's camera in open water conditions** (a) Frame example where visual detection is not successful due to poor visibility conditions. (b) Frame example where the diver is successfully detected when close enough.

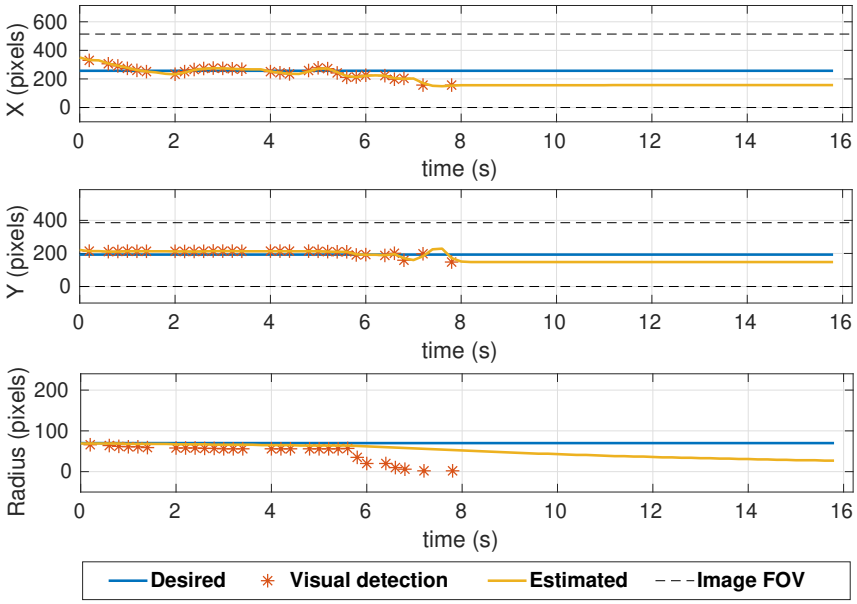

Figure 10: **Visual tracking (Scenario 1):** experimental results describing visual detection and estimation using the camera to track the diver.

4.2 Acoustic tracking (Scenario 2)

For this scenario, the pinger was carried by the diver who was initially a few meters away from the robot. The diver was asked to stay at a stationary spot for this experiment. As there is no distance feedback to the diver using this acoustic method, U-CAT will move continuously even when reaching the diver. Figure 11 shows that the remapped detected relative heading was mostly on the left side of the image. This is because the robot kept swimming around the diver after reaching him. Figure 11 shows also noisy acoustic data due

to reflections. This shows clearly the need of a complementary visual sensor to track the diver efficiently.

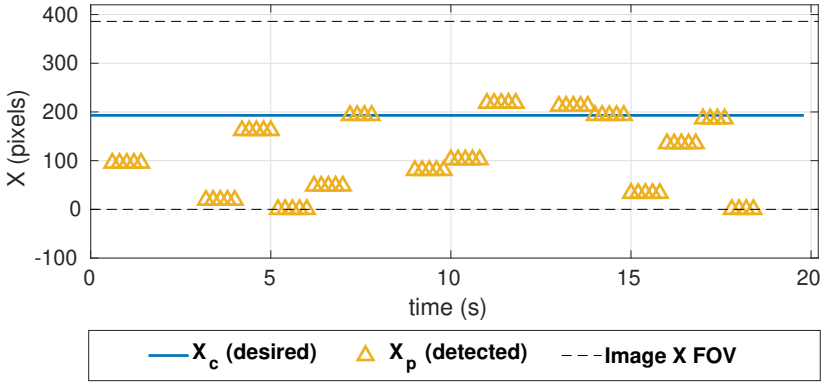

Figure 11: **Acoustic tracking (Scenario 2):** experimental results for tracking the diver based on acoustic signals only.

4.3 Data-fusion based tracking

As discussed previously, two different scenarios were conducted, namely static and dynamic tracking. The objective of scenario 3 is to evaluate the performance of the proposed control tracking scheme for the case of tracking a static object that was initially placed out of the camera's FOV. The goal of scenario 4 is to assess the performance of the proposed solution for dynamic diver tracking.

4.3.1 Static tracking (Scenario 3)

In this experiment, a target was mounted next to the pinger approximately 6 meters away from U-CAT (as illustrated in Figure 15). The robot moves towards the target, slightly oscillates left and right due to reflections in acoustic detection, and is facing the target after nearly $40s$. Figure 12 shows that the target was then centered on the camera's image based on the visual feedback. Since both the target and the AUV were operating at the same depth, the target was almost already centered in Y axis, and the slight vertical error to the camera's center is corrected based on visual feedback. It is worth to note that the proposed tracking scheme can successfully detect and track a static target that is initially out of the on-board camera's FOV. Figure 12 also shows that the desired width of the target within the camera's frame is retained, which translates into a maintained relative distance to the target. This clearly shows that combining both visual and acoustic measurements allows a more robust tracking of the object.

Figure 13 and Figure 14 show the obtained results when starting respectively placed to the right, and to the

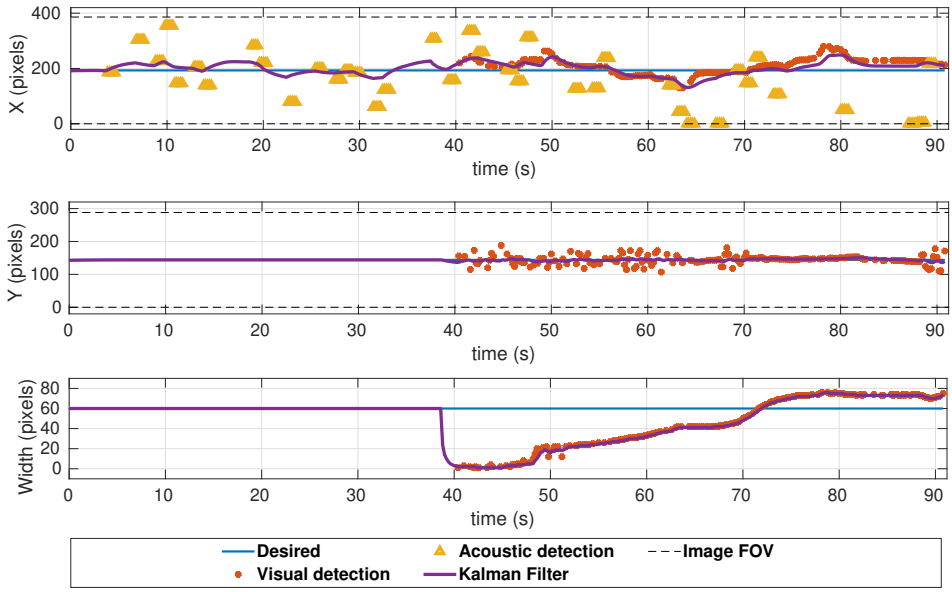

Figure 12: **Data fusion based static tracking (Scenario 3):** experimental results describing static target tracking based on the proposed data-fusion scheme where the AUV started facing the target.

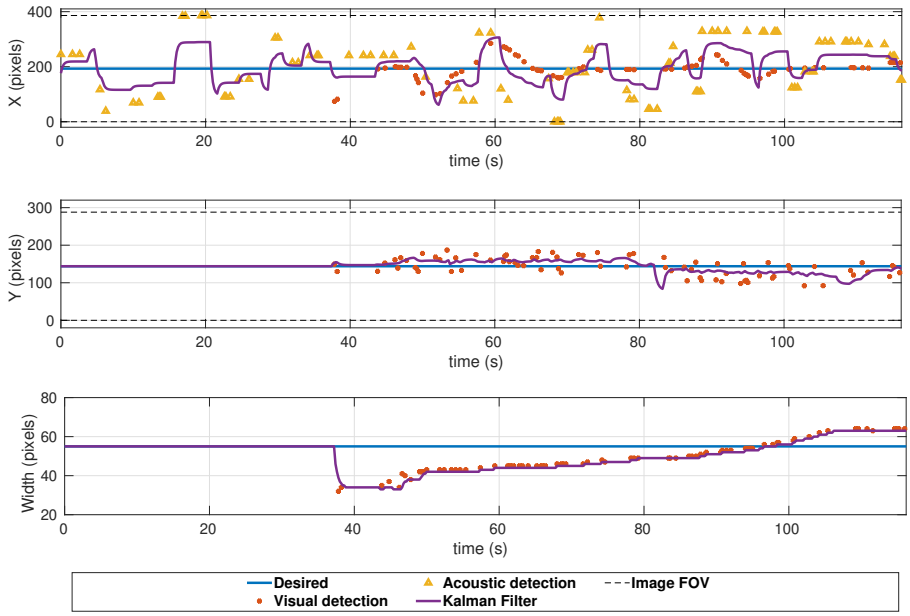

Figure 13: **Data fusion based static tracking (Scenario 3):** experimental results describing static target tracking based on the proposed data-fusion scheme where the AUV started to the right of the target.

left of the target. In the experiment illustrated by Figure 13, The robot relies on acoustics at the first 40-50 seconds when the target is out of the camera's visibility reach. When the target is detected visually, the

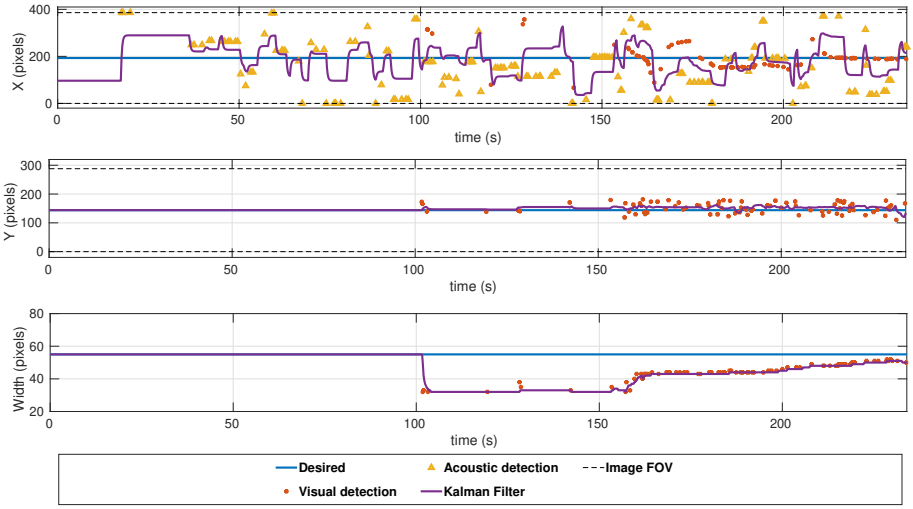

Figure 14: **Data fusion based static tracking (Scenario 3):** experimental results describing static target tracking based on the proposed data-fusion scheme where the AUV started to the left of the target.

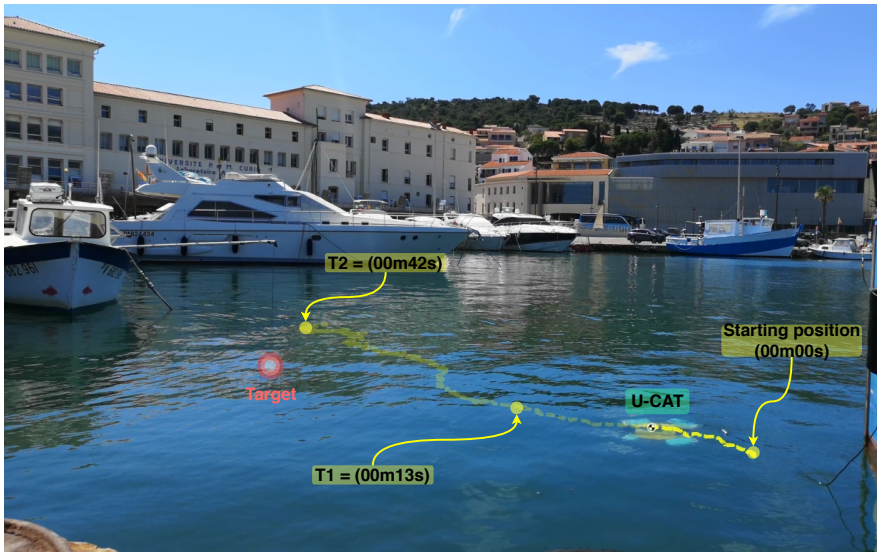

Figure 15: **U-CAT illustrative trajectory when homing towards a static target using the proposed data-fusion scheme:** At time T1, the AUV was heading towards the target based on acoustic detection. At time T2, U-CAT has found the target by combining both visual and acoustic sensors' data.

target estimated location within the camera's frame is improved, which allows the robot to reduce oscillations around the target, and center it in the camera's frame. In the last experiment for this scenario, where the robot started to the left of the target, it took longer (107s) to detect the target visually. This is mainly due to acoustic reflections, and motors noise since the robot was operating at a harbour where motorised vehicle were passing by. Nevertheless, the target was reached successfully, and after the object was detected visually, the robot managed to keep it in its camera's frame center.

As there is no ground truth to evaluate for the acoustic detections, we present the Table 3 that shows the error between the target's camera location, and the camera's center, for the three case scenarios. The results show that once the target was retrieved visually, it was mainly kept at the center of the camera's frame.

This clearly shows the robustness of the proposed method, but also its repeatability. This allows to detect and track an object initially far away from the robot, where no other method based on vision only could succeed. The results also show that the robot manages to do station keeping when tracking a static object.

4.3.2 Dynamic tracking (Scenario 4)

In this last scenario, we tested the proposed control approach to actively track a diver. The diver was carrying the pinger and moving freely in a 2D trajectory at a constant depth. As the robot can only swim at relatively slow velocity, the diver was asked to also move slowly.

When the diver starts within the camera's field of view, the robot has a better estimate of its location based on both visual and acoustic detections, as shown in Figure 16. When the diver starts out of the camera's field of view, as shown in Figure 17, it takes longer at the beginning to accurately detect and track the diver. But in both cases, the proposed solution allowed to track the moving diver for more than 8 minutes. The target left the camera's field of view several times, nevertheless, its location could always be recovered. The complementarity of the proposed approach allows to quickly head towards the diver, and keep him centered in the camera's image.

A shorter experiment was conducted for better data-clarity, and for illustrating the trajectory of both the robot and the diver graphically. As illustrated in Figure 19, U-CAT was accurately tracking the target throughout this experiment. Figure 18 shows clearly how the diver left the camera's field of view several times. The proposed tracking control approach allows to recover the diver location within the camera's FOV, and allows to center it within the camera's frame, and within a desired bounding box width. The presented results show the robustness of the proposed approach, where experimental results were validated throughout different scenarios, and where the AUV was operating in open-waters at poor visibility conditions, and subject to acoustic reflections and noise.

Table 4 shows the tracking performance in terms of error between desired and the fed-back position of the diver within the camera's frame. The higher errors are along the image X axis, since the diver was moving mainly horizontally. Both errors in Y and width are low considering the diver was keeping his depth constant, and moving relatively slowly. This shows that the diver was mainly centered in the camera, despite leaving

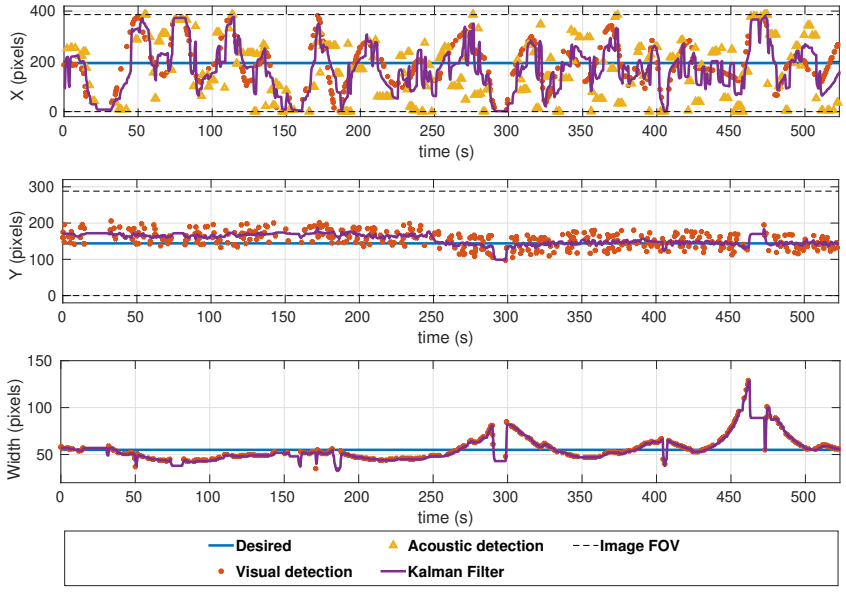

Figure 16: **Data-fusion based dynamic tracking (Scenario 4):** Diver starting close to the robot

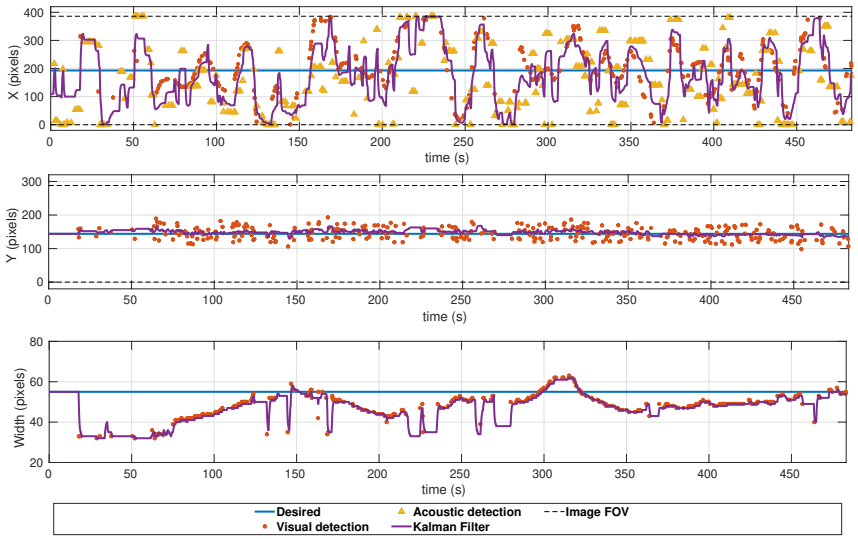

Figure 17: **Data-fusion based dynamic tracking (Scenario 4):** Diver starting away from the robot its FOV numerous times.

The results also illustrate that the proposed controller allows to achieve a 3D tracking motion, in spite of the high coupling in U-CAT actuators.

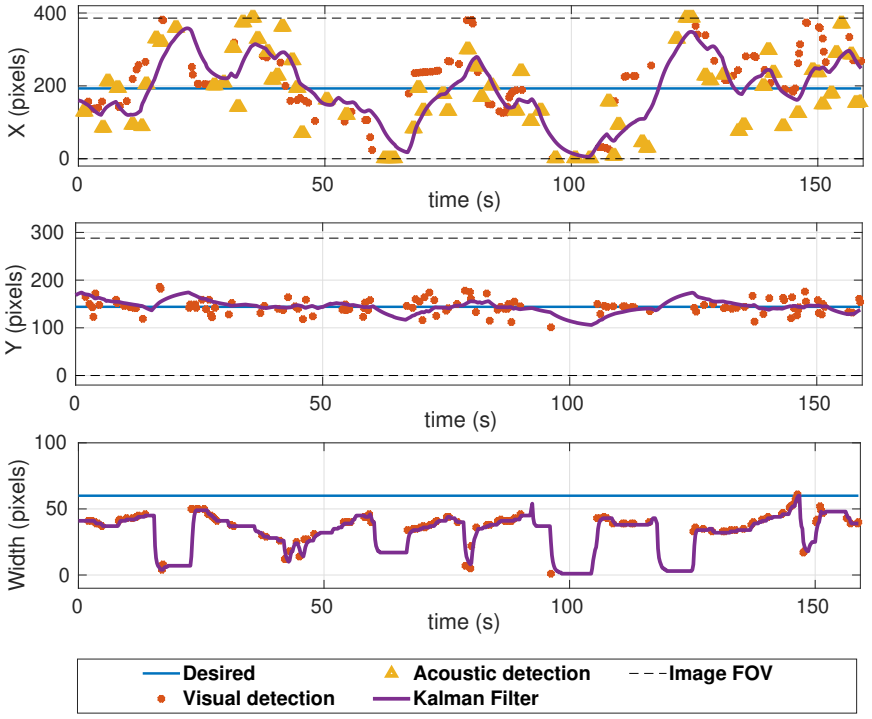

Figure 18: **Data-fusion based dynamic tracking (Scenario 4):** This shorter experiment is presented for plot clarity.

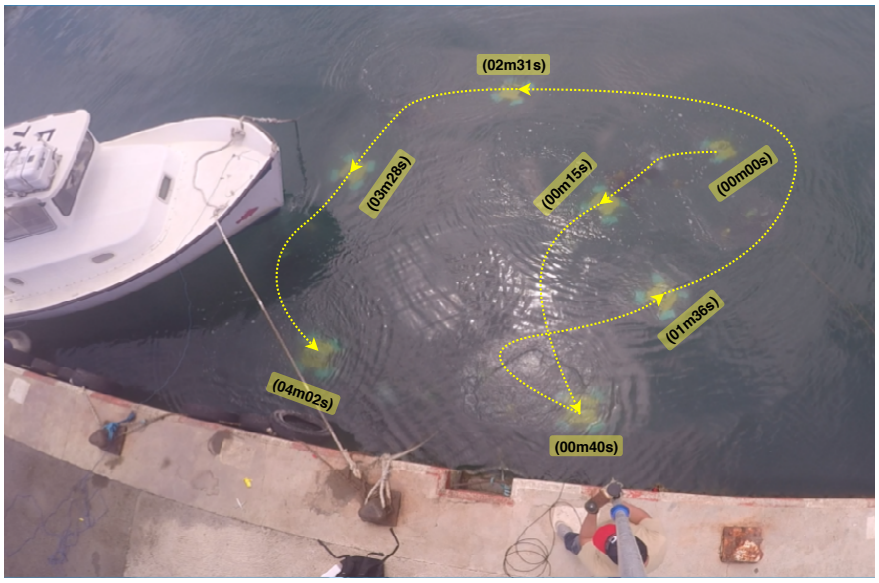

Figure 19: **U-CAT illustrative diver tracking trajectory:** description at different times.

5 Conclusion and future works

The aim of this work was to design and develop a low-cost data-fusion based diver tracking control scheme and to implement it on a highly manoeuvrable underwater robot U-CAT. Taking into account U-CAT's actuation

characteristics and hardware capabilities, a data-fusion based technique combining acoustic and visual signals was proposed. The proposed control strategy was validated through open water field experiments. Promising results were obtained, demonstrating the effectiveness and performance of the proposed data-fusion based control algorithm. The proposed solution demonstrates also that the target can be found, detected, and tracked, even if it is initially far away from a robot operating in poor visibility conditions.

Combining acoustic and visual signals for underwater detection and tracking shows a robust performance in harsh field conditions, and the ability to recover the diver's location when visual detection fails.

Further work will be carried out in future, which will include testing at larger trial areas, and improving of the acoustic detection accuracy, considering three DOF orientation detection with the hydrophones instead of heading orientation only, which will allow a diver tracking at any depth. The visual detection part will also be improved with the advancement of real-time object detection algorithms. Another interesting idea would be the implementation of a gesture recognition algorithm to allow a more efficient underwater human-robot collaboration.

Acknowledgments

This work is financed through Estonian Research Council Grant IUT-339. The authors would like to thank the "Observatoire Océanologique de Banyuls-sur-mer" for providing scientific divers and experimental setups. The authors would also like to thank Christian Meurer, Jaan Rebane, and Roza Gkliva for their valuable reviews and comments to improve this work.

ORCID

Walid Remmas orcid.org/0000-0001-8690-0496
Ahmed Chemori orcid.org/0000-0001-9739-9473
Maarja Kruusmaa orcid.org/0000-0001-5738-5421

References

Aggogeri, F., Mikolajczyk, T., and O'Kane, J. (2019). Robotics for rehabilitation of hand movement in stroke survivors. *Advances in Mechanical Engineering*, 11(4).

Bernard, T., Martusevich, K., Rolins, A. A., Spence, I., Troshchenko, A., and Chintalapati, S. (2018). A novel mars rover concept for astronaut operational support on surface eva missions. In *2018 AIAA SPACE and Astronautics Forum and Exposition*, page 5154.

Bjelonic, M. (2016–2018). YOLO ROS: Real-time object detection for ROS. `https://github.com/leggedrobotics/darknet_ros`.

Buelow, H. and Birk, A. (2011). Diver detection by motion-segmentation and shape-analysis from a moving vehicle. In *OCEANS'11 MTS/IEEE KONA*, pages 1–7. IEEE.

Cao, Y., Xu, J., Lin, S., Wei, F., and Hu, H. (2019). Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*.

Cardozo, P. O., dos Santos, M. M., Lilles, P., and Silva, S. (2017). Forward looking sonar scene matching using deep learning. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 574–579.

Chavez, A. G., Mueller, C. A., Birk, A., Babic, A., and Miskovic, N. (2017). Stereo-vision based diver pose estimation using lstm recurrent neural networks for auv navigation guidance. In *OCEANS 2017-Aberdeen*, pages 1–7. IEEE.

Chavez, A. G., Pfingsthorn, M., Birk, A., Rendulić, I., and Misković, N. (2015). Visual diver detection using multi-descriptor nearest-class-mean random forests in the context of underwater human robot interaction (hri). In *OCEANS 2015-Genova*, pages 1–7. IEEE.

Chemori, A., Kuusmik, K., Salumae, T., and Kruusmaa, M. (2016). Depth control of the biomimetic u-cat turtle-like auv with experiments in real operating conditions. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4750–4755.

Cheng, C. and Jiang, B.-T. (2012). A robust visual servo scheme for underwater pipeline following. In *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 456–459.

Choi, H., Woo, J., and Kim, N. (2017). Localization of an underwater acoustic source for acoustic pinger-based transit task in 2016 maritime robotx challenge. In *2017 IEEE Underwater Technology (UT)*, pages 1–7.

Colgate, E., Wannasuphoprasit, W., and Peshkin, M. (1996). Cobots: robots for collaboration with human operators. In Kwon, Y., Davis, D., and Chung, H., editors, *Proceedings of the ASME Dynamic Systems and Control Division*, volume 58, pages 433–439. ASME.

Costanzi, R., Monnini, N., Ridolfi, A., Allotta, B., and Caiti, A. (2017). On field experience on underwater acoustic localization through usbl modems. In *OCEANS 2017 - Aberdeen*, pages 1–5.

de Langis, K. and Sattar, J. (2019). Real-time multi-diver tracking and re-identification for underwater human-robot collaboration. *arXiv preprint arXiv:1910.09636*.

DeMarco, K. J., West, M. E., and Howard, A. M. (2013). Sonar-based detection and tracking of a diver for underwater human-robot interaction scenarios. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2378–2383. IEEE.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578.

Dudek, G., Jenkin, M., Prahacs, C., Hogue, A., Sattar, J., Giguere, P., German, A., Liu, H., Saunderson, S., Ripsman, A., Simhon, S., Torres, L. ., Milios, E., Zhang, P., and Rekletis, I. (2005). A visually guided swimming robot. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3604–3609.

Duntley, S. Q. (1963). Light in the sea∗. *J. Opt. Soc. Am.*, 53(2):214–233.

Gomez Chavez, A., Ranieri, A., Chiarella, D., Zereik, E., Babić, A., and Birk, A. (2019). Caddy underwater stereo-vision dataset for human–robot interaction (hri) in the context of diver activities. *Journal of Marine Science and Engineering*, 7(1):16.

Han, K. M. and Choi, H. T. (2011). Shape context based object recognition and tracking in structured underwater environment. In *2011 IEEE International Geoscience and Remote Sensing Symposium*, pages 617–620.

Hentout, A., Aouache, M., Maoudj, A., and Akli, I. (2019). Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017. *Advanced Robotics*, 33(15-16):764–799.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Islam, M. J., Fulton, M., and Sattar, J. (2019). Toward a generic diver-following algorithm: Balancing robustness and efficiency in deep visual detection. *IEEE Robotics and Automation Letters*, 4(1):113–120.

Islam, M. J., Ho, M., and Sattar, J. (2019). Understanding human motion and gestures for underwater human–robot collaboration. *Journal of Field Robotics*, 36(5):851–873.

Jensen, F. B., Kuperman, W. A., Porter, M. B., and Schmidt, H. (2011). *Computational ocean acoustics*. Springer Science & Business Media.

Kakinuma, K., Hashimoto, M., and Takahashi, K. (2012). Outdoor pedestrian tracking by multiple mobile robots based on slam and gps fusion. In *2012 IEEE/SICE International Symposium on System Integration (SII)*, pages 422–427.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35.

Kamal, S., Mohammed, S. K., Pillai, P. R. S., and Supriya, M. H. (2013). Deep learning architectures for underwater target recognition. In *2013 Ocean Electronics (SYMPOL)*, pages 48–54.

Kasetkasem, T., Worasawate, D., Tipsuwan, Y., Thiennviboon, P., and Hoonsuwan, P. (2017). A pinger localization algorithm using sparse representation for autonomous underwater vehicles. In *2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 533–536.

Kim, B. and Yu, S. (2017). Imaging sonar based real-time underwater object detection utilizing adaboost method. In *2017 IEEE Underwater Technology (UT)*, pages 1–5.

Kim, D., Lee, D., Myung, H., and Choi, H. (2012). Object detection and tracking for autonomous underwater robots using weighted template matching. In *2012 Oceans - Yeosu*, pages 1–5.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. In *Nature*, volume 521.7553, pages 436–444.

Lee, P.-M., Jeon, B.-H., and Kim, S.-M. (2003). Visual servoing for underwater docking of an autonomous underwater vehicle with one camera. In *Oceans 2003. Celebrating the Past ... Teaming Toward the Future (IEEE Cat. No.03CH37492)*, volume 2, pages 677–682 Vol.2.

Lee, S. (2017). Deep learning of submerged body images from 2d sonar sensor based on convolutional neural network. In *2017 IEEE Underwater Technology (UT)*, pages 1–3.

Lekkala, K. K. and Mittal, V. K. (2016). Simultaneous aerial vehicle localization and human tracking. In *2016 IEEE Region 10 Conference (TENCON)*, pages 379–383.

Li, Y., Chen, Y., Wang, N., and Zhang, Z. (2019). Scale-aware trident networks for object detection. *arXiv preprint arXiv:1901.01892*.

Mamdani, E. H. and Baaklini, N. (1975). Prescriptive method for deriving control policy in a fuzzy-logic controller. *Electronics Letters*, 11(25):625–626.

Mandić, F., Rendulić, I., Mišković, N., and Đula Nađ (2016). Underwater object tracking using sonar and usbl measurements. *Journal of Sensors*, 2016.

Mišković, N., Bibuli, M., Birk, A., Caccia, M., Egi, M., Grammer, K., Marroni, A., Neasham, J., Pascoal, A., Vasilijević, A., and Vukić, Z. (2015). Overview of the fp7 project "caddy — cognitive autonomous diving buddy". In *OCEANS 2015 - Genova*, pages 1–5.

Petillot, Y. R., Reed, S. R., and Bell, J. M. (2002). Real time auv pipeline detection and tracking using side scan sonar and multi-beam echo-sounder. In *OCEANS '02 MTS/IEEE*, volume 1, pages 217–222 vol.1.

Prabowo, M. R., Hudayani, N., Purwiyanti, S., Sulistiyanti, S. R., and Setyawan, F. X. A. (2017). A moving objects detection in underwater video using subtraction of the background model. In *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pages 1–4.

Preisig, J. (2007). Acoustic propagation considerations for underwater acoustic communications network development. *ACM SIGMOBILE Mobile Computing and Communications Review*, 11(4):2–10.

Preston, V., Salumäe, T., and Kruusmaa, M. (2018). Underwater confined space mapping by resource-constrained autonomous vehicle. *Journal of Field Robotics*, 35(7):1122–1148.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv*.

Remmas, W. (2017). Modélisation et commande du robot sous-marin bio-inspiré U-CAT. Master's thesis, National Polytechnic School of Constantine, Algeria.

Ren, P., Fang, W., and Djahel, S. (2017). A novel yolo-based real-time people counting approach. In *2017 International Smart Cities Conference (ISC2)*, pages 1–2.

Robinson, H., MacDonald, B., and Broadbent, E. (2014). The role of healthcare robots for older people at home: A review. *International Journal of Social Robotics*, 6(4):575–591.

Salumäe, T., Chemori, A., and Kruusmaa, M. (2016). Motion control architecture of a 4-fin u-cat auv using dof prioritization. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1321–1327.

Salumäe, T., Chemori, A., and Kruusmaa, M. (2019). Motion control of a hovering biomimetic four-fin underwater robot. *IEEE Journal of Oceanic Engineering*, 44(1):54–71.

Sattar, J. and Dudek, G. (2007). Where is your dive buddy: tracking humans underwater using spatio-temporal features. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3654–3659.

Sattar, J. and Dudek, G. (2009). Underwater human-robot interaction via biological motion identification. In *Robotics: Science and Systems*.

Shkurti, F., Chang, W., Henderson, P., Islam, M. J., Higuera, J. C. G., Li, J., Manderson, T., Xu, A., Dudek, G., and Sattar, J. (2017). Underwater multi-robot convoying using visual tracking by detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4189–4196.

Song, Y., Zhu, Y., Li, G., Feng, C., He, B., and Yan, T. (2017). Side scan sonar segmentation using deep convolutional neural network. In *OCEANS 2017 - Anchorage*, pages 1–4.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I.

Xia, Y. and Sattar, J. (2019). Visual diver recognition for underwater human-robot collaboration. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6839–6845. IEEE.

Yagimli, M. and Varol, H. S. (2009). A gps-based system design for the recognition and tracking of moving targets. In *2009 4th International Conference on Recent Advances in Space Technologies*, pages 6–12.

Yu, S.-C., Ura, T., Fujii, T., and Kondo, H. (2001). Navigation of autonomous underwater vehicles based on artificial underwater landmarks. In *MTS/IEEE Oceans 2001. An Ocean Odyssey. Conference Proceedings (IEEE Cat. No.01CH37295)*, volume 1, pages 409–416 vol.1.

Zhao, S., Lu, T.-F., and Anvar, A. (2009). Automatic object detection for auv navigation using imaging sonar within confined environments. In *2009 4th IEEE Conference on Industrial Electronics and Applications*, pages 3648–3653.

Zou, J. and Tseng, Y. (2012). Visual track system applied in quadrotor aerial robot. In *2012 Third International Conference on Digital Manufacturing Automation*, pages 1025–1028.

Table 1: tiny-YOLOv3 performance

mAP (%)	54.26
IoU (%)	58.41
Frames per second	20

Table 2: Fuzzy controller rule base

E_i \ $\dot{E}_i$	Negative	Zero	Positive
Negative	Negative	Negative	Zero
Zero	Negative	Zero	Positive
Positive	Zero	Positive	Positive

Table 3: Scenario 3: RMSE between the camera's center and desired width, and the object pixels location and its width.

RMSE (in pixels)	**Test 1 Facing the target**	**Test 2 right to the target**	**Test 3 left to the target**
RMSE X	42	67	72
RMSE Y	17	20	21
RMSE Width	12	15	8

Table 4: Scenario 4: RMSE between the camera's center and desired width, and the object pixels location and its width.

RMSE (in Pixels)	**Test1 Diver initially close from the robot**	**Test 2 Diver initially away from the robot**	**Test 3 Shorter experiment**
RMSE X	119	124	106
RMSE Y	37	28	33
RMSE W	24	19	31