ICT Tools for Preservation of Underwater Environment: a Vision-based Posidonia Oceanica Monitoring

Francesco Ruscio[1,2], Giovanni Peralta[1,2], Lorenzo Pollini[1,2], and Riccardo Costanzi[1,2]

[1] Dipartimento di Ingegneria dell'Informazione (DII), Università di Pisa, Pisa, Italy

[2] Interuniversity Center of Integrated Systems for the Marine Environment (ISME), Italy

ABSTRACT

Underwater monitoring activities are crucial for the preservation of marine ecosystems. Currently, scuba divers are involved in data collection campaigns that are repetitive, dangerous and expensive. This article describes the application of ICT tools to underwater visual data for monitoring purposes. The data refer to a Posidonia Oceanica survey mission carried out by a scuba diver using a Smart Dive Scooter equipped with visual acquisition and acoustic localization systems. An acoustic-based strategy for the geo-referencing of the optical dataset is reported. It exploits the synchronization between the audio track extracted from a camera and the transponder pings adopted for the acoustic positioning. The positioning measurements are employed within an Extended Kalman Filter to estimate the diver's path during the mission. A Visual Odometry algorithm is implemented within the filter to refine the navigation state estimation of the diver with respect to the acoustic positioning only. Moreover, a smoothing step based on the Rauch-Tung-Striebel smoother is applied to further improve the estimated diver's positions. Finally, the paper reports the results of two different data processing for monitoring applications. The first one is an image mosaicking obtained by concatenating subsequent frames, whereas the second one refers to

a qualitative distribution of the Posidonia Oceanica over the mission area accomplished through an image segmentation process. The two outcomes are plotted over a satellite image of the surveyed area, showing that the proposed process is an effective tool capable of facilitating divers in their monitoring and inspection activities.
Keywords – underwater vision, underwater data geo-referencing, visual odometry, posidonia oceanica monitoring

I. Introduction

Climate change and rapid urbanization are negatively affecting water quality as well as endangering underwater ecological habitats (Pörtner et al., 2014). The monitoring of underwater environments is of utmost importance to ensure the protection of marine ecosystems. Among the various marine species, Posidonia Oceanica (PO) is considered as one of the most important indicators of the biodiversity and water quality in Mediterranean coastal areas (Buia, Gambi, & Dappiano, 2004). As reported in the European Commission's directive 92/43/CEE, PO represents a priority natural habitat, and the regular monitoring activities play an crucial role in its preservation. Currently, the monitoring of PO is mainly performed by scuba divers, who survey the area of interest to assess the health state and the extension of PO meadows. In the Italian context, this activity is handled by the Regional Agencies for the Environmental Protection (Agenzia Regionale per la Protezione dell'Ambiente, ARPA), employing professional divers who spend several days per year in visual surveys which are repetitive, costly and, most importantly, dangerous for the human operator. Ligurian ARPA (Agenzia Regionale per la Protezione dell'Ambiente Ligure, ARPAL) and the Department of Information Engineering of the University of Pisa are collaborating towards the integration of classical methodologies with Information

Communication Technology (ICT) tools to carry out the underwater campaigns. This collaboration aims at reducing the necessary time at sea for scuba divers with consequent benefits in terms of safety and costs as well as an improvement of the quality of acquired data and of the precision of their geo-referencing. Looking ahead, the employment of remote sensing and data processing could lead to a gradual shift from diver-based to unmanned surveys.

This research work concerns the use of ICT tools on visual data for underwater monitoring applications. In particular, this paper deals with an implementation of a geo-referencing strategy based on acoustic synchronization applied to underwater images. Furthermore, it shows how the geo-referenced images can be exploited for both navigation and monitoring purposes. The visual data used within this research activity refer to an underwater survey carried out over a PO meadow, whose aim was a performance assessment of the system proposed by Costanzi et al. (2019). The survey, which took place in front of the Ligurian Coast in March 2019, involved the use of a Smart Dive Scooter (SDS). The SDS is a modified version of a standard dive scooter equipped with a visual acquisition system and a modem in transponder configuration. The visual acquisition system is in charge of acquiring images of the inspected area, whereas the acoustic modem is used together with a compatible Ultra Short BaseLine (USBL) device to track the diver during the mission. Differently from other approaches, which are usually based on well recognizable events in the video (Nocerino et al., 2018), the geo-referencing strategy adopted here exploits the synchronization between the audio track recorded by a camera and the acoustic transponder pings. Filtering the signal at proper frequencies to delete environmental noise, the transponder pings are easily identifiable within the audio track. This suggests that it is possible to accomplish a synchronization between the video streams and the position measurements. The time alignment between the two signals consequently results in the geo-referencing of the data acquired by the camera. However, this approach suffers where few acoustic measurements are available. Therefore, a filtering process is carried out with the aim of refining the positions occupied by the diver throughout the mission and consequently improving the geo-referencing of the visual data. In particular, two different navigation filters are described. Given the non-linearity of the system, both filters are based on an Extended Kalman Filter (EKF). The first filter exploits only the acoustic measurements provided by the USBL, whereas the second one combines also information extracted from images. Such information comes from a Visual Odometry (VO) algorithm tailored for the specific dataset. The algorithm extracts the salient points in a frame and then tracks them in the following images. According to the changes that motion induces on the images of the onboard cameras, the VO algorithm incrementally estimates the motion of the diver. The combination of acoustic positioning measurements and visual information allows to improve the navigation state estimation with respect to the acoustic positioning only. As an additional step, the Rauch-Tung-Striebel (RTS) smoother (Rauch, Tung, & Striebel, 1965) is used to further improve the accuracy achieved during filtering process. At this stage, the geo-referenced frames can be elaborated to extract useful information about the mission area. In the paper, data processing is used for two different monitoring applications. The first one concerns a sea bottom reconstruction through image mosaicking. The second one involves a description of the PO presence over the mission area using image segmentation process. Both the outcomes are plotted over a satellite image of the surveyed area, thus providing an effective visual representation.

The proposed strategy represents a valuable approach to address the geo-referencing issue of independent data and shows how the results of the synchronization process can be exploited for both monitoring and reconstruction of the sea bottom. Moreover, it is worth mentioning that this approach is not limited to optical data, but can also be applied to acoustic images (Franchi, Ridolfi, and Zacchini, 2018). According to the goal of the mission, the achieved results represent an efficient way to evaluate the details of the seabed. Within the context of this research work, biologists can take advantage of this type of information to assess spread and health condition of PO.

This work represents an extension of the research activity presented in the IEEE/MTS Oceans 2020 Global Conference (Ruscio, Costanzi, Pollini, Manzari, Barbieri, & Gaino, 2020), where the feasibility of the audio-based geo-referencing strategy of visual data was proved. In this work, the results of that strategy applied to the same visual data are exploited for navigation and monitoring purposes.

II. System Description

The monitoring data used within this research activity were collected by a diver using a Smart Dive Scooter (SDS). The SDS was proposed by Costanzi et al. (2019) as a support tool for monitoring application of different marine species. The device is a modified version of the SUEX XJ VR dive scooter, which is one among the standard equipment commonly used by ARPAL professional divers to carry out their underwater surveys. It comes with a rear propeller and an ergonomic handle which eases divers underwater mobility. The main goal of the authors was to provide the scooter with sensors for environment monitoring. To this end, the scooter is equipped with a visual acquisition system with the aim of easily collecting optical data of the area to be inspected. The visual system consists of four GoPro Hero 5 cameras mounted on a support aluminum bar that is rigidly installed on the scooter front. The position of the cameras is designed to create two stereo-vision systems: the first two cameras are mounted along the vertical direction pointing the sea bottom, while the other two are mounted 45° with respect to the vertical direction pointing forward. This solution allows the ARPAL divers to use the SDS to quickly acquire images for different kinds of marine biological monitoring activities. Due to the lack of the Global Navigation Satellite System (GNSS) in underwater environments, together with the visual acquisition system, the SDS is equipped with an acoustic localization system. This allows to solve the problem of the underwater localization of the diver during the mission. Moreover, the acoustic signals can be exploited to achieve a geo-referencing of the collected data, as it will be explained in Section 3. The acoustic localization system consists of an EvoLogics 18/34 modem in transponder configuration installed below the scooter. The acoustic positioning is accomplished using a compatible Ultra Short BaseLine (USBL) device deployed from a support boat and controlled by a management software installed on a PC. The USBL emits an acoustic signal and the transponder is programmed to answer with an acknowledgement (ACK) message when interrogated. According to the distance and the direction of arrival of the acoustic transponder pings, the USBL computes the transponder relative position. Then, it is converted in an absolute position exploiting the USBL attitude estimation provided by an integrated Attitude and Heading Reference System (AHRS) and its global position measured by a GPS device on-board the support boat (Paull, Saeedi, Seto, & Li, 2014). Figure 1 depicts the overall SDS during its deployment.

Figure 1 - SDS employed during the mission, acoustic transponder below the scooter, cameras on the left (Costanzi et al., 2019).

Both visual and localization systems are stand-alone modules and are fixed to the dive scooter to make the three subsystems a unique rigid body, thus allowing for a quick and simple integration of the additional devices with the scooter. This solution does not require a single board computer or passing cables, which would have increased the cost and the complexity of the overall system. However, the hardware integration easiness results in an asynchronism between the cameras and the positioning system, since data are stored in physically independent devices. In particular, video streams acquired by cameras are archived on the integrated Micro SD (one per camera), whereas the positioning measurements are collected on the mission monitoring computer on-board the support boat. This requires a data synchronization process to achieve a correct geo-referencing of the collected images.

III. Data Synchronization

The collected images need to be correctly geo-referenced to be used for monitoring purposes. The geo-referencing of optical data was achieved applying the synchronization strategy proposed by Costanzi et al. (2019), which involves the exploitation in post-processing of the camera audio tracks. The approach is based on the identification of the transponder positioning events within the audio tracks acquired by the cameras. This idea derives from the fact that the operating frequencies of the transponder (*18-34kHz*) are within the camera audio channel frequency range. The spectrogram reported in Figure 2 shows about *30s* of the audio signal extracted from the audio channel of one of the four cameras. The *1Hz* signal peaks refer to the pings emitted by the transponder during the USBL acoustic positioning interrogations. In particular, it can be observed that such peaks are dominant at high frequencies compared to environmental noise, mainly due to the scooter's propeller. This suggests the possibility of processing the signal to retain only the

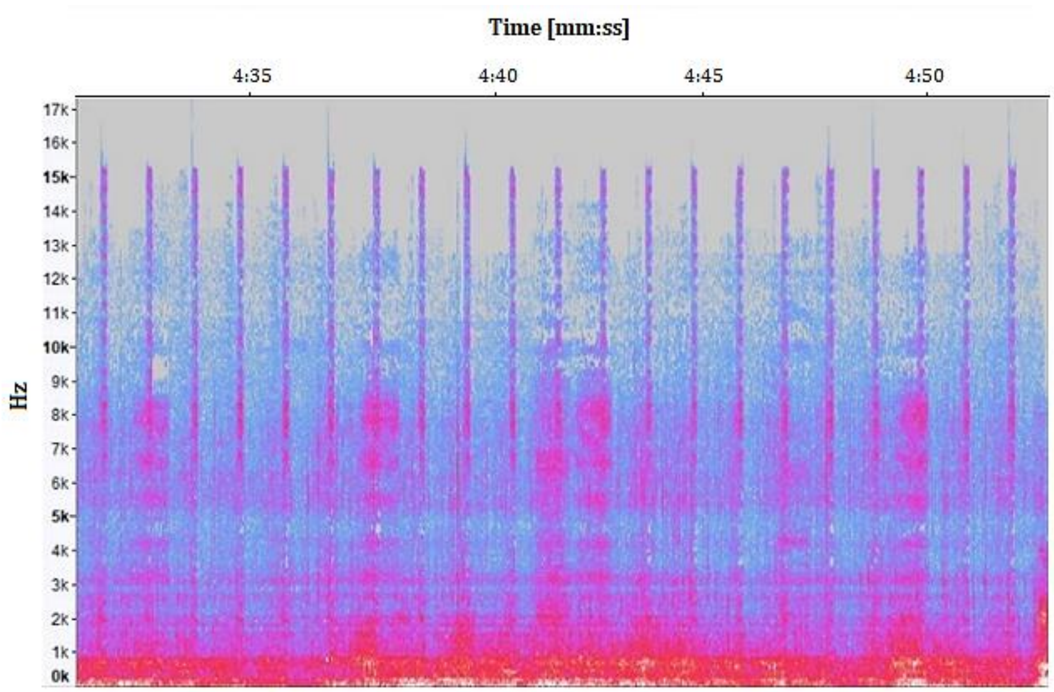

Figure 2 - Spectrogram of the audio signal extracted from one of the four videos.

information of interest. Consequently, a band pass filter in the range of *13kHz-15kHz* was applied to the signal, thus removing most of the environmental noise. This way, the transponder pings are clearly visible within the filtered audio track and it can be further exploited for synchronization purposes. The synchronization process was successfully addressed by exploiting pattern alignment between the filtered audio track and the recorded time series of the transponder pings. Throughout the mission, the transponder was interrogated at a frequency of *1Hz* to achieve a frequent positioning of the SDS. However, a constant interrogation rate would lead to an uncertainty in the synchronization process. For this reason, every *60s* the interrogation was interrupted for *1.30s*. This resulted in a well-defined timing pattern for transponder interrogation to avoid any potential ambiguity due to the constant interrogation frequency. Technically speaking, the pattern alignment between the two signals was achieved by computing the maximum of the cross-correlation function, resulting in a delay of *139s* between the video timestamp and the corresponding position

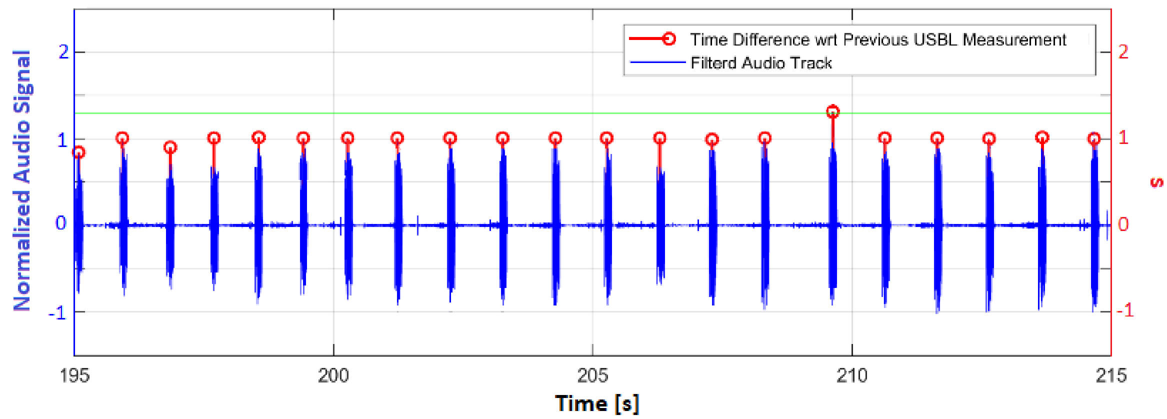

Figure 3 - Synchronization between transponder pings (red) and filtered audio track (blue)

measurement. The plot in Figure 3 reports the result of the synchronization process, showing the alignment of the two acoustic signals in a time interval of about *20s*. The blue signal refers to the transponder pings within the filtered audio track. The red circles represent instants of the transponder answers recorded by the USBL in terms of time difference with respect to the previous measurement. It is possible to notice the timing interrogation pattern adopted to avoid the ambiguity due to a constant interrogation rate. In particular, the maximum peak around *210s*, highlighted by the green line, refers to an interruption of interrogation of *1.30s*, as indicated by the time difference with respect to the previous USBL measurement. The time alignment between the audio track extracted from the video and the recorded transponder pings entails the match between the video stream and the position measurements. Therefore, each image collected by the camera is associated with a specific timestamp of the localization system.

IV. Visual Odometry Algorithm

Thanks to the time alignment achieved with the synchronization process, it was possible to exploit the visual dataset in a VO framework with the aim of refining the navigation state estimation of the SDS with respect to the acoustic positioning only. VO is a technique that exploits the information extracted from a sequence of images to estimate the camera motion (Fraundorfer & Scaramuzza, 2011). Thanks to recent technological advances, optical sensors are widely employed in many robotic applications (Aquel, Marhaban, Saripan, & Ismail, 2016), thus increasing the scientific community's interest in VO. Cameras are passive sensors that allow robots to acquire vast information about the surrounding environment, which can be used for both navigation and monitoring purposes. Furthermore, VO can be effective in all those contexts where the GNSS is not available, such as underwater applications. Nonetheless, since VO works by computing the camera path incrementally, the errors introduced by each motion estimation step accumulate over time, thus generating a drift of the estimated trajectory from the actual one.

The typical flowchart involves the identification of important features within an image. Features are salient points in images, such as corners and blobs, which are identified by dedicated *detector* algorithms. Each feature is described using a vector called *descriptor*, which contains local information from the neighboring pixels. It assigns a distinctive identity to the feature to be distinguishable from the others. Several feature detectors are available (Siegwart, Nourbakhsh, & Scaramuzza, 2011) and it is important to choose the one that best suits the characteristics of the specific dataset. Considering the marine environment, a comparison between multiple feature detectors in underwater images can be found in (Codevilla, Gaya, Filho, & Botelho, 2015). The aim of the authors was to evaluate the robustness to turbidity and image degradation of different detectors. According to their results, features detectors based on Difference of Gaussian (DoG) (Lowe, 2004) and Hessian-Laplace (Mikolajczyk & Schmid, 2004) exhibited the best results for localization accuracy. Once the features are detected, a correspondence step is performed to identify corresponding features in different images of the same scene. It can be addressed in two different ways: *feature matching* and *feature tracking*. In the first approach features are detected independently in each frame and then matched according to the similarity between their descriptors. Conversely, the second one finds features in one image and tracks them in the following frames using local search techniques. This kind of approach is preferable when images are taken at nearby locations so that the amount of motion between consecutive frames is small. In the underwater scenario, the high variability of the environment precludes a complete generalization. Nevertheless, feature tracking approaches have shown better performance than methods based on feature matching in case of low quality underwater images (Ferrera, Moras, Trouvé-Peloux, & Creuze, 2019). This is mainly due to the lack of textures typical of underwater scenarios that can lead to erroneous matches. In contrast, tracking approaches are less prone to feature correspondence ambiguities as they are based on minimizing the photometric cost. From the set of point correspondences is then possible to estimate the motion of the camera between two consecutive frames. The motion is defined in terms of relative rotation and translation of the current camera frame with respect to the previous one. Different methods can be adopted to carry out this task, according to whether the correspondent features are defined in two or three dimensions (Huang & Netravali, 1994). This is usually related to the number of cameras, typically one (Monocular VO) or two (Stereo VO). Since correspondent features are usually contaminated by outliers, it is crucial to remove wrong data association in order to achieve an accurate camera motion estimation. The standard solution adopted in VO is represented by RANSAC (Fischler & Bolles, 1981), which is a stochastic algorithm for model estimation in the presence of outliers.

In the context of this research work, the main goal of the monitoring activity was to assess the health state and the spread of the PO meadow. Taking into account that PO has a distribution that is mainly horizontal, only the optical data collected by the two downward-facing cameras were considered. However, analyzing the videos extracted from these two cameras, it was found that for most of the mission there is not enough overlap between corresponding frames. This is primarily due to the baseline between the cameras being too large in relation to the average distance from the seabed. Consequently, it was chosen to use a Monocular VO approach, in which only the images acquired by one of the two cameras downward-facing were analyzed. The optical data collected during the monitoring activity are composed of 2704x1520 RGB images registered with an acquisition rate of *30fps*. Given the high computational cost due to the size of the images, a subsampling at a size of 640x360 was performed. In addition, taking into account the wide overlap between consecutive frames due to the slow dynamics typical of underwater missions, the framerate was decreased through a subsampling at *10fps*. Despite the aforementioned changes, the images obtained still contain sufficient information and enough overlap to justify the implementation of a VO algorithm. Prior to the actual mission, a planar checkerboard grid had been deployed on the seabed for camera calibration purposes. The diver performed several maneuvers above the calibration pattern to acquire various frames from different point of views. The calibration of the camera was accomplished using the perspective projection model (Zhang, 2000). Moreover, in the underwater environment the refractive effects induced by the water should be taken into account. The approach adopted here is to correct the refraction effects by including them in radial and tangential distortion parameters (Shortis, 2015). These are additional parameters encompassing non-linear phenomena, typically lens distortion, which are recovered within a linear least-squares solution during the calibration process.

In conclusion, the Monocular VO flowchart implemented in this work is as follows:

- *Image Undistortion*: it represents an indispensable step towards achieving a correct motion estimation. Each image is initially processed to account for both refraction effects

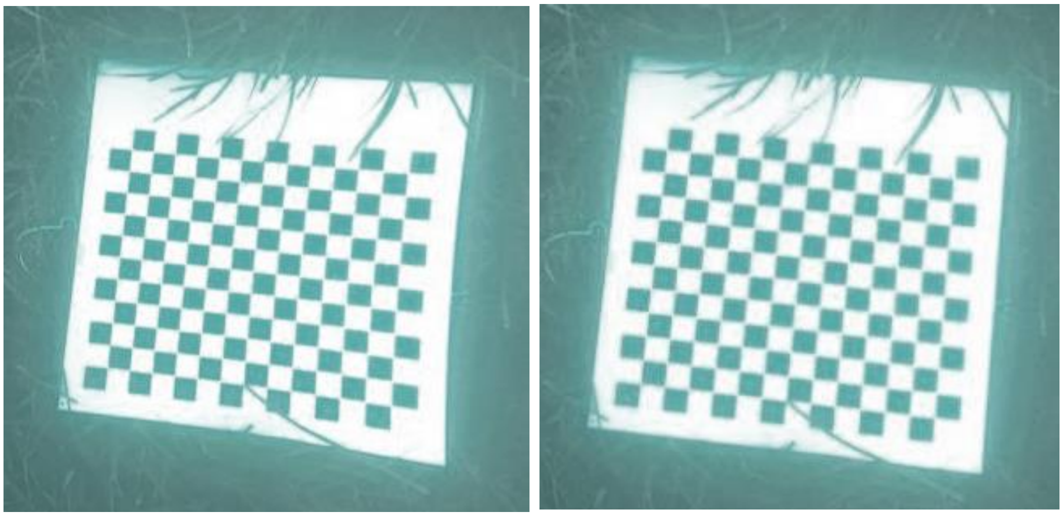

Figure 4 - Image undistortion applied to one of the images used for calibration. Original image on the left; undistorted image on the right.

and lens distortion. This is obtained using distortion parameters estimated during the calibration process within a sixth order radial and second order tangential polynomial models. Figure 4 depicts an example of distortion removal to one of the images used for the calibration phase;

- *Feature Detection*: features are extracted employing the SIFT detection and description algorithm (Lowe, 2004), which uses DoG for detection and Histogram of Oriented Gradients (HOG) as descriptor. It allows to detect features invariants to scale changes, 2D rotations and reasonably invariant to viewpoint changes and illumination. The SIFT algorithm was selected for our VO flowchart because of its property to be invariant to changes in the surrounding environment. Figure 5 shows four images extracted from visual dataset. It is possible to notice the high variability within the image texture. The main problem with this algorithm is its high computational cost, which makes it unsuitable for real-time applications. Nevertheless, this is not a limitation for our application, as it runs in post-processing;

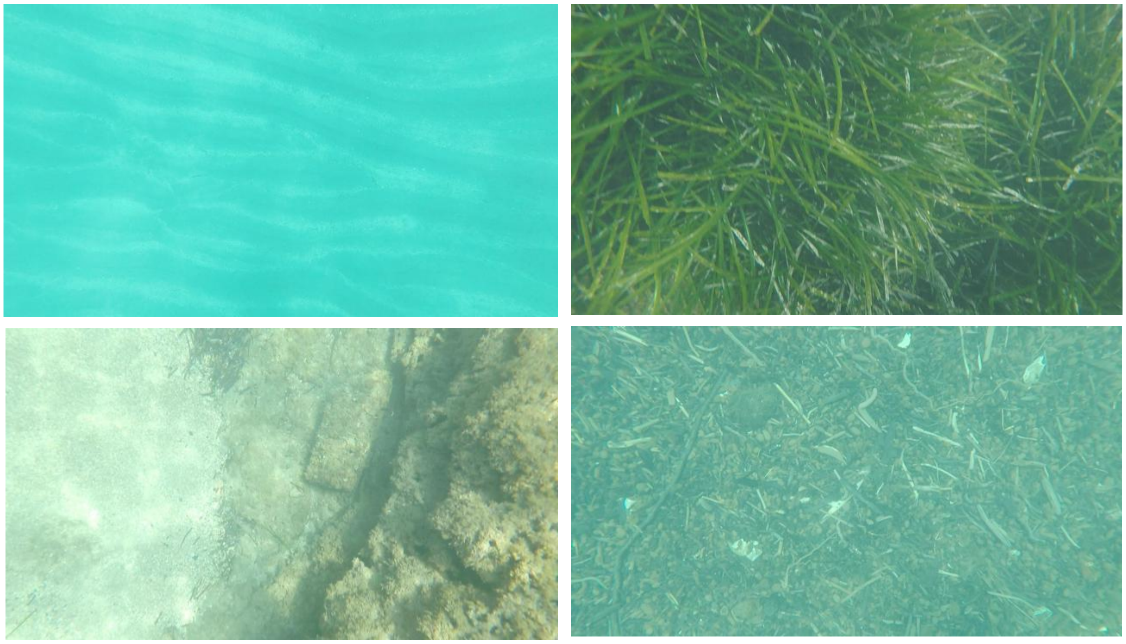

Figure 5 - Frames from visual dataset

- *Feature Tracking*: the feature correspondence in subsequent frames was addressed using a feature tracking approach. The algorithm adopted here is the Kanade-Lucas-Tomasi (KLT) tracker (Shi & Tomasi, 1994), whose goal is to align a sub-region of an image with the subsequent image. This is done by exploiting spatial intensity gradient to retrieve the position that yields the best match. As the tracker algorithm progresses over time, initial features decrease due to predominantly rectilinear motion. Furthermore, possible features may be occluded during motion. For this reason, new features are re-acquired whenever their number falls below a certain threshold. This ensures a sufficient amount of features for correct motion estimation;

- *Motion Estimation*: the motion between each pair of consecutive images is computed exploiting the *essential matrix*. It contains the motion information in terms of rotation and translation of a calibrated camera up to a scale factor for the translation. The matrix is computed from 2D-2D feature correspondence according to the epipolar constraints (Longuet-Higgins, 1981). Within this work, the essential matrix is calculated exploiting the 5-point algorithm (Nister, 2004), which performs better compared to the 8-point one when the scene is predominantly planar. It has been implemented in MATLAB environment using the Computer Vision Toolbox libraries. The motion parameters are then extracted using the Single Value Decomposition (SVD) technique. For a robust implementation, the essential matrix estimation is integrated with MLESAC (Torr and Zisserman, 2000), a generalization of the RANSAC estimator, which chooses the solution that maximizes the likelihood rather than just the number of inliers. This is done with the aim of making the motion estimation more reliable in the presence of possible outliers;

- *Estimation Failure*: it is not always possible to determine the essential matrix and, therefore, an estimate of the motion between two consecutive images. This occurs whenever there are not enough corresponding features from which to extract the essential matrix. To overcome this shortcoming, the motion is assumed to be analogous to the one estimated at the previous step. This assumption is reasonable for the specific mission, given the slow dynamics involved.

This VO approach is applied sequentially to each frame within the visual dataset to extract the motion of the SDS to which the camera is attached, in terms of linear and angular velocity. Such information are exploited together with the acoustic positioning measurements within a navigation filter to estimate the path followed by the diver throughout the mission.

V. Navigation

The aim of the research activity was to evaluate the conditions of the PO over the area of interest. This is achieved exploiting optical data acquired by the visual acquisition system on board the SDS. The acquisition campaign was carried out by a professional ARPAL scuba diver without any technological support for navigation. The diver tried to maintain a constant depth and a constant speed during the entire mission. In addition, the path followed by the diver was supposed to cover an area characterized by the border of Posidonia Oceanica. A good navigation estimation is crucial within this research activity, as it allows a geo-referencing of images even in portions of the trajectory where less frequent USBL measurements are available. In addition, it improves the geo-referencing of the visual data, leading to better monitoring results. Since the diver was supposed to move at constant depth and speed, it was assumed the problem to be 2D, considering the diver's position in horizontal plane only. Such a position is referred to a North – East reference frame whose origin is chosen correspondent to the position of the first GPS measurement.

In the following, two different navigation approaches will be presented. Due to the non-linearity of the navigation problem addressed, both of them are based on an Extended Kalman Filter (EKF), but they use models that exploit different kind of measurements. The first one exploits only USBL measurements to estimate the diver's position during the mission. Conversely, the second one combines USBL measurements with velocity information extracted from the VO estimation.

- **USBL-Only Extended Kalman Filter: (USBL EKF)**

Considering that the USBL provides positioning measurements, the state used for this estimation filters is composed of:

- n, the diver's position along the North direction in a NED reference frame;
- e, the divers' position along the East direction in a NED reference frame;
- v, the diver's speed;
- ψ, the heading angle;

Thus, the state vector can be expressed as:

$$x = [n, e, v, \psi]^T \in \mathcal{R}^4 \quad (1)$$

The state-transition model represents a first-order kinematic model with slowly variable heading and speed along the course axis. It is defined by the following equations:

$$\begin{cases} n_{k+1} = n_k + v_k \, cos\psi_k \, dT \\ e_{k+1} = e_k + v_k \, sin\psi_k \, dT \\ v_{k+1} = v_k + \, w_v \, dT \\ \psi_{k+1} = \psi_k + \, w_\psi \, dT \end{cases} \quad (2)$$

where w_v and w_ψ represent the white noise which characterize the derivative with respect to the time of the state corresponding component. The time interval *dT* has a value of *0.1s*, as the filter loop rate is *10Hz*.

On the other hand, the observation model exploits the measurements provided by the USBL in order to refine the predicted diver's position. Since the acoustic localization system gives the scooter North and East coordinates, the measurement equation is defined as:

$$y_k = h(x_k) + w_y \quad (3)$$

where $h(x_k) = \begin{bmatrix} n_k \\ e_k \end{bmatrix}$ and $w_y \in \mathcal{R}^2$ represents the white noise vector associated with the measurements. For each available measurement, the correction step is applied to revise the prediction. However, possible outliers in the measurements are rejected on the basis of a threshold on the Mahalanobis distance associated to the measurements (De Maesschalck, Jouan-Rimbaud, & Massart, 2000).

- **VO-USBL Extended Kalman Filter (VO-USBL EKF)**

The purpose of this filtering is to estimate the diver's position by merging the USBL measurements with the information coming from the VO algorithm. In particular, the algorithm provides an estimate of the distance covered by the camera in the time interval between two consecutive images. This information can therefore be considered as an estimate in terms of linear and angular velocities. However, since the VO is carried out using a monocular approach, the linear velocity estimation is performed up to a scale factor. Therefore, the filter is in charge to estimate it, using the position measurements provided by the USBL. In addition, the USBL position measurements allow to limit the drift generated by the VO process. In view of this, the state vector used to describe the dynamic of the system is composed of:

- n, the diver's position along the North direction in a NED reference frame;
- e, the divers' position along the East direction in a NED reference frame;

- v, the diver's speed;
- ψ, the heading angle;
- ω, the angular velocity around the axis orthogonal to the motion plane;
- λ, the scale factor.

Hence, the vector state can be represented as:

$$x = [n, e, v, \psi, \omega, \lambda]^T \in \mathcal{R}^6 \tag{4}$$

The state dynamics can be written as:

$$\begin{cases} n_{k+1} = n_k + v_k \, cos\psi_k \, dT \\ e_{k+1} = e_k + v_k \, sin\psi_k \, dT \\ v_{k+1} = v_k + \, w_v \, dT \\ \psi_{k+1} = \psi_k + \, w_\psi \, dT \\ \omega_{k+1} = \omega_k + \, w_\omega \, dT \\ \lambda_{k+1} = \lambda_k + \, w_\lambda \, dT \end{cases} \tag{5}$$

where w_v, w_ψ, w_ω and w_λ are the speed, heading, angular velocity and scale factor white noise, respectively. As in the previous filter, the time interval dT has a value of *0.1s*, as the filter loop rate is *10Hz*.

As mentioned above, the correction phase exploits measurements provided by the USBL and VO estimation. These two information are used as correction independently. Therefore, two different observation models can be written:

$$y_k = h(x_k) + w_y \tag{6}$$

where, with USBL measurements, it is defined as:

$$h(x_k)_{USBL} = \begin{bmatrix} n_k \\ e_k \end{bmatrix} \tag{7}$$

while, with information provided by the visual odometry, it is described as:

$$h(x_k)_{VO} = \begin{bmatrix} v_k \ \lambda_k \\ \omega_k \end{bmatrix} \tag{8}$$

and $w_y \in \mathcal{R}^2$ represents the white noise vector associated with the USBL and VO measurements, respectively.

Also in this case, a check for rejecting possible outliers is applied using the Mahalanobis distance.

VI. Image Segmentation

In addition to using the images for navigation purposes, it is also possible to exploit them to extract useful information about the area being inspected. As an example of application, the synchronized video stream was used to describe the presence of the PO meadow over the mission area. The identification of the PO presence was carried out through an image segmentation process exploiting the Machine Learning (ML) technique proposed in (Burgera, Bonin-Font, Lisani, & Petro, 2016). The implemented algorithm is based on a Support Vector Machine (SVM) algorithm, which belongs to the supervised learning algorithms. It first needs a training phase to obtain a classification mode, which can be subsequently used to detect the presence of PO in the dataset. In the ML technique proposed by the authors, each image is divided into patches of the same size, which is editable by the user. The model carries out the classification of each independent patch according to the response of the patch to a series of Gabor filters. Finally, an iterative refinement step based on patch color is applied to achieve a pixel level classification.

The geo-referenced frames were processed using the aforementioned ML technique and the achieved results were then used to describe the coverage of PO over the mission area, as it will be explained in Section 7. Firstly, a training phase was carried out to achieve a classification model tailored for the specific scenario. The adopted training set corresponds to the 5% of the entire dataset. Images with different tonalities and textures were selected to make the model robust with respect to different illumination and environmental conditions. Each image used for training has a binary counterpart acting as ground truth, which was hand-labelled by ARPAL biologists according to the pixels covered with *PO*. Then, the model achieved with the training phase was applied to the remaining 95% part of the dataset to identify the distribution of PO over the surveyed

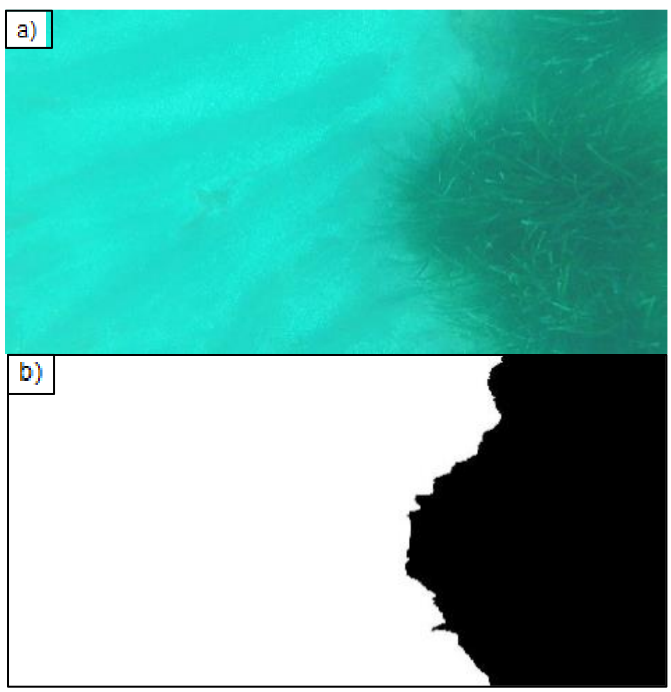

Figure 6 - Example of image segmentation: a) original image; b) segmented image. PO percentage = 40%

area. The model carries out the classification returning a binary image, where black pixels refer to PO presence, white otherwise. The percentage of black pixels in each frame was selected as an indicator to describe the presence of PO over the area covered by the image. Figure 6 depicts an example of image segmentation, where 40% of PO presence was computed.

VII. Results

In this section, the results of the two adopted navigation filters are reported. Figure 7 compares the trajectory estimation carried out with the two filters detailed in Section 5. In particular, the estimated diver's path in terms of North and East coordinates is reported. In addition, the figure shows a classification of the positioning measurements acquired by the USBL in terms of outliers and inliers. Only the inliers were used for the correction phase of the two filters. The red path represents the trajectory estimated by the USBL EKF. In this case, only the acoustic measurements acquired by the USBL were exploited for the correction phase. Conversely, the blue path depicts the trajectory computed by the VO-USBL EKF. Differently from the previous filter, this one merges the acoustic measurements with velocity estimated by the VO algorithm. Despite its simplicity, the USBL EKF provides good results throughout the entire mission. However, in those areas where acoustic measurements are not available or are rejected, it exhibits some limitations. On the contrary, the additional information extracted from the camera within the VO framework allows the VO-USBL EKF to achieve a better performance even where less frequent USBL measurements are available. Indeed, even if the USBL positioning fails, the filter is still able to perform corrections according to the velocity estimations provided by the VO algorithm. The correction based on velocity information is subject to a drift on the position that increases with time. However, as soon as a new USBL measurement becomes available, the position is corrected

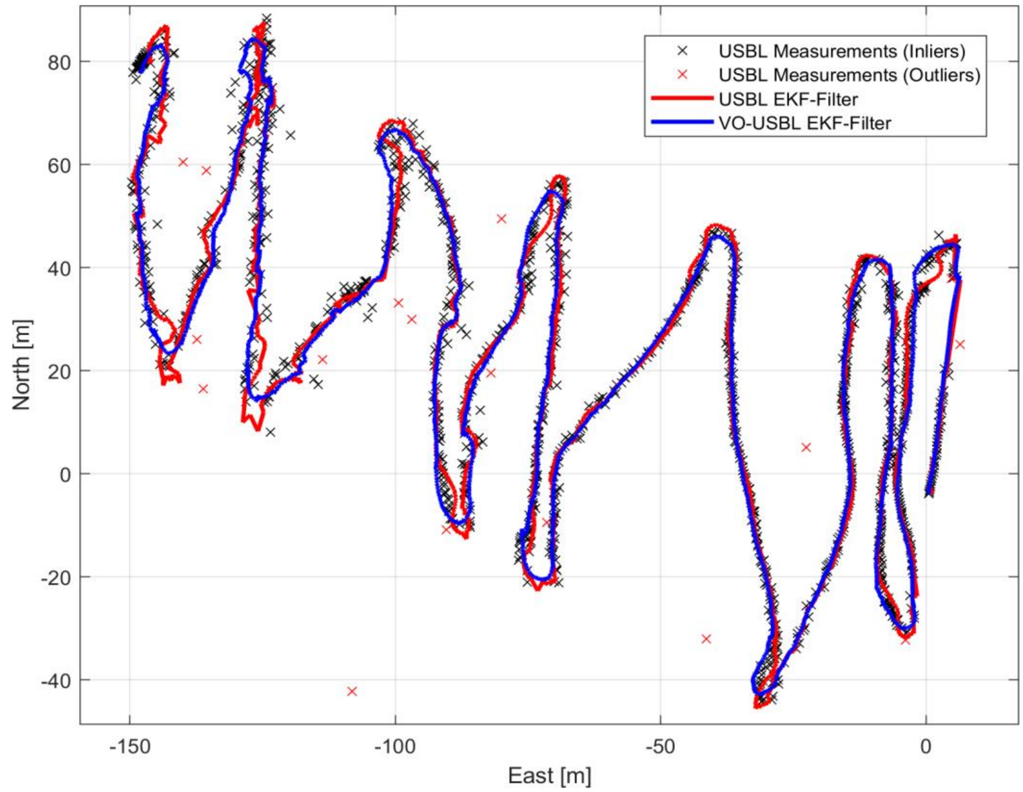

Figure 7 – Estimated diver's path using USBL-only EKF (red) and USBL-VO EKF (blue) together with USBL measurements (x)

accordingly. In addition, the filter is capable of estimating the scale factor resulting from the employment of a Monocular VO approach. This makes it possible to retrieve the actual displacement between two consecutive camera poses.

As a further step, the Rauch-Tung-Striebel (RTS) smoother was applied to the state estimated by the VO-USBL EKF. The RTS smoother is an effective two-pass algorithm for fixed interval smoothing (Rauch, Tung, & Striebel, 1965). Its general implementation consists of a forward pass for filtering followed by a backward recursive smoothing. The aim of employing the RTS smoothing algorithm is to refine the estimation obtained during the filtering process. Indeed, a smoother is based on a higher information content as it is a non-causal approach, exploiting the whole dataset of observations (both future and past data) to obtain the estimation. Figure 8 reports the results of the smoothing process applied to the VO-USBL EKF estimates. At this stage, the

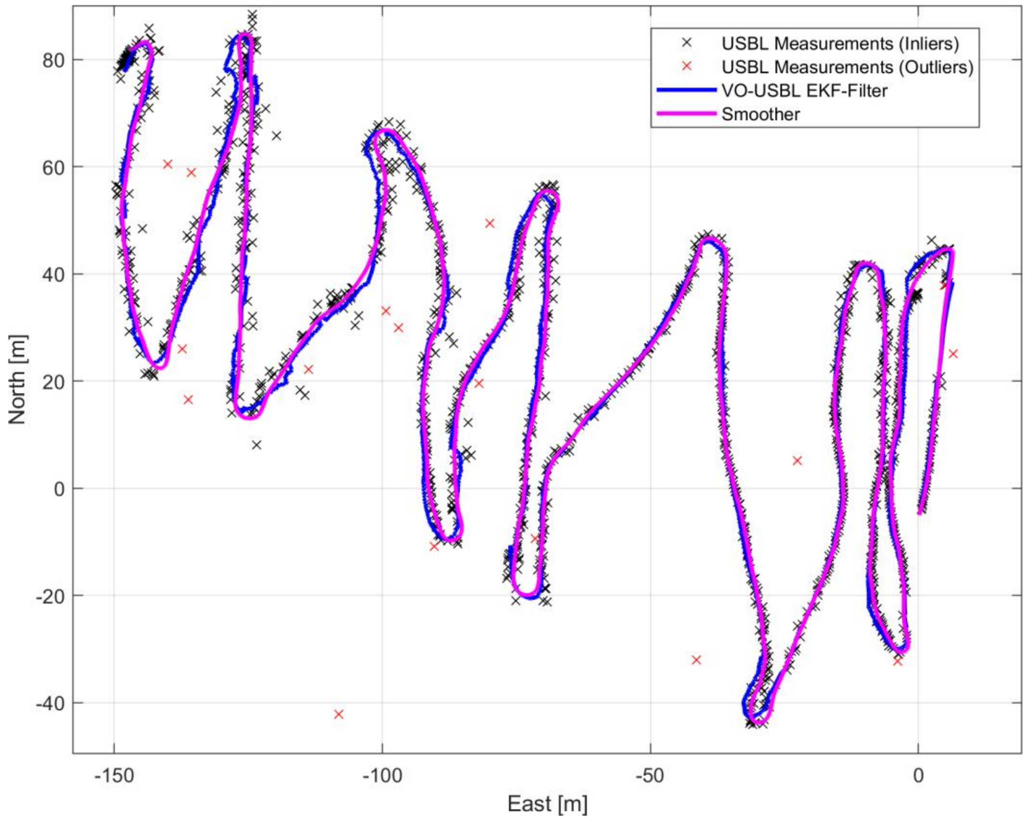

Figure 8 – The Rauch-Tung-Striebel (RTS) smoother applied to the VO-USBL EKF estimates

geo-referenced frames can be exploited to extract information about the environment. In the following, the results of two data processing used for different monitoring purposes are reported. The first one concerns a mosaic from images acquired during the mission. The *image mosaicking* (also known as *image alignment*) was achieved by concatenating subsequent images according to diver's motion. In particular, firstly each image is scaled based on the estimated scale factor value. This allows images to be conveyed from pixels to meters, so that they are consistent with the reference system used for the mission area. Then, a rigid transformation is applied to each image according to the estimated translation and rotation between consecutive frames. Such geometric transformations have been implemented in MATLAB environment without employing any dedicated computer vision library. The result of the mosaicking process is reported in Figure 9. It shows the geo-referenced images superimposed on a satellite image of the surveyed area. This

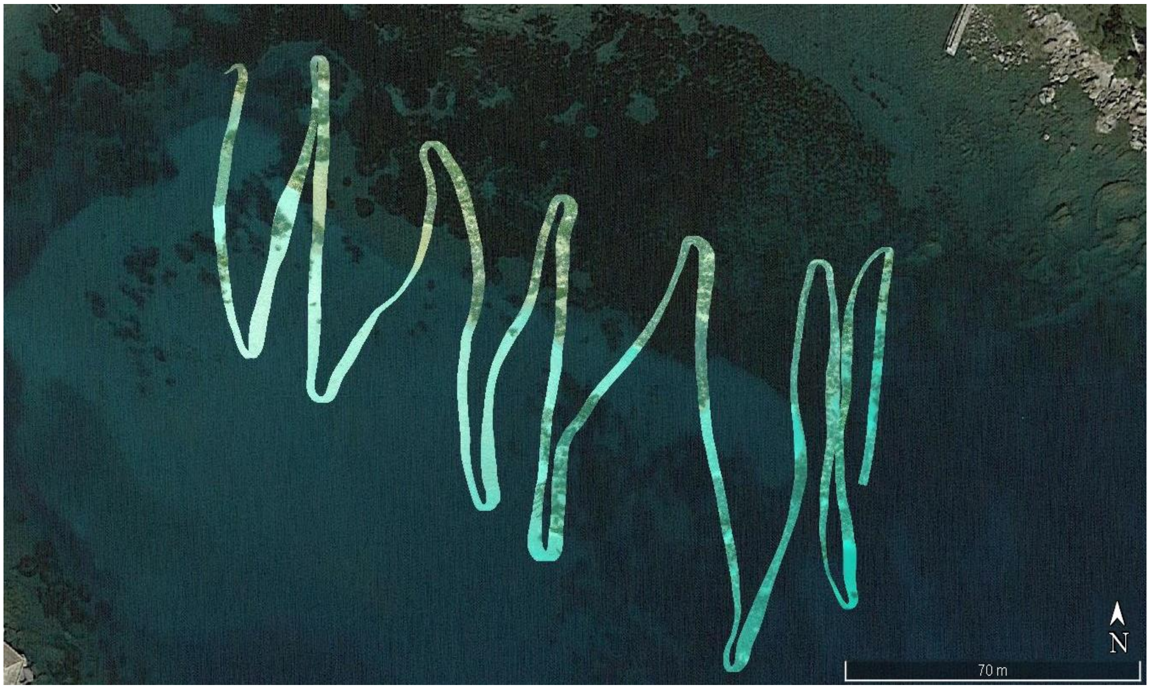

Figure 9 – Image mosaicking of the surveyed area

visual representation is an efficient way to evaluate the details of the seabed. Within the context of this research work, biologists can take advantage of this type of information to assess the health condition of PO. The second example of monitoring application concerns a qualitative description of the PO coverage over the mission area. In particular, the presence of PO was defined exploiting the segmentation process described in Section 6. Each estimated position was associated with a value that indicates the percentage of black pixels within the corresponding segmented image. Such value represents an indicator of the concentration of PO in that specific position. Figure 10 shows the distribution of PO along the estimated diver's path. Each estimated position together with its PO concentration value were plotted over a satellite image of the mission area. A chromatic scale was defined to distinguish the level of PO presence in every estimated position: the green color refers to high PO concentration values, whereas the yellow color describes areas with low PO concentration values. The two figures show that the outcomes of the two data processing are

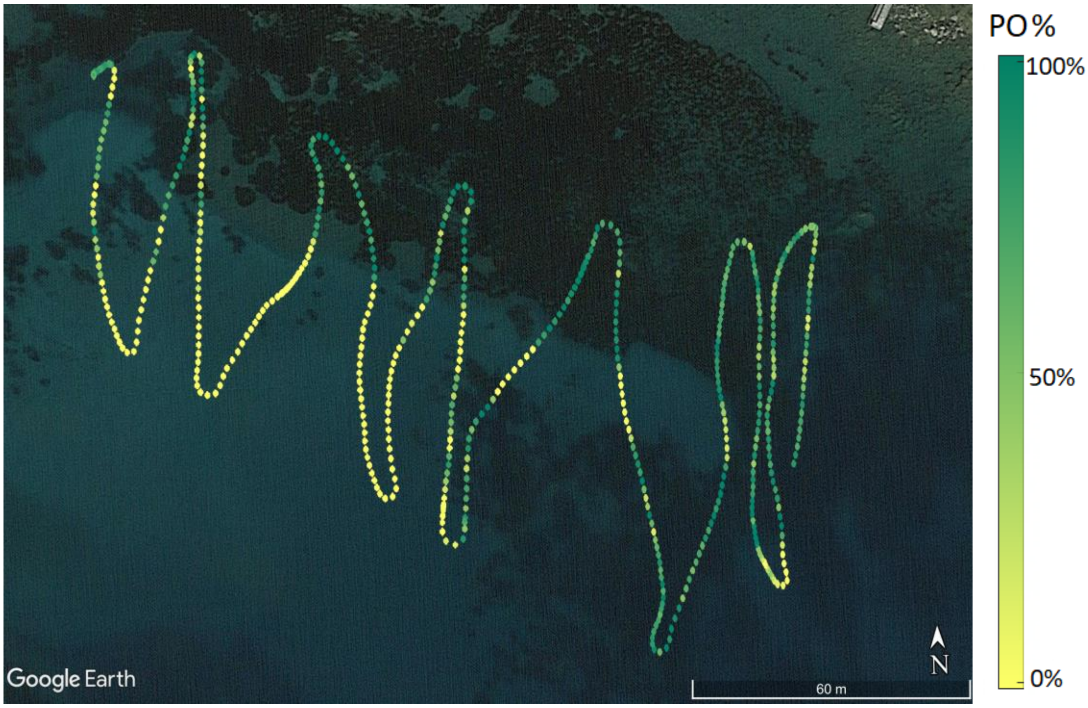

Figure 10 – Estimated diver's path with chromatic scale based on the percentage of PO detected by the ML algorithm

in accordance with the information retrievable form an aerial point of view, thus proving the effectiveness of the proposed geo-referencing strategy.

VIII. Conclusions and Future Developments

The work presents the application of ICT tools on underwater visual data for Posidonia Oceanica monitoring purposes. The data refer to a survey carried out with a Smart Dive Scooter equipped with a vision and acoustic localization systems. A geo-referencing strategy for the optical data is reported. Firstly, a time alignment between the camera audio track and the transponder pings is accomplished. The synchronization resulted in a match between the video stream and the positioning measurement timestamps. Then, two different Extended Kalman Filters were implemented to estimate the diver's position throughout the mission. The first one exploits only measurements provided by the acoustic localization system. On the contrary, the second filter fuses acoustic measurements with information estimated by a Visual Odometry algorithm, showing good results even where less frequent measurements are available. A smoothing step based on the Rauch-Tung-Striebel smoother was then applied to further improve the estimated diver's positions. As a result, the achieved precise and frequent positioning allowed to associate a specific position with each frame within the video stream, which can then be exploited to extract useful information about the environment. To that effect, the paper reported two results of data processing for different monitoring applications. The first one is a mosaic composed of sequential images properly combined according to the estimated motion. The second result reports the percentage of the Posidonia Oceanica presence over the mission area, exploiting a segmentation process applied to the images. The two outcomes were plotted over a satellite image of the mission area, showing results that are compliant with the ones deductible from an aerial point of view.

The presented work demonstrates that an audio-based synchronization strategy followed by a navigation system can effectively address the geo-referencing issue of independent data. Moreover, it shows how the results of the synchronization process can be exploited for both monitoring purposes and reconstruction of the sea bottom. As far as the specific mission is concerned, ARPAL biologists can take advantage of the outcomes of the process to assess spread and health condition of PO. Future developments will involve the evaluation of sensitivity to different VO approaches (e.g. Monocular and Stereo) as well as the integration of other sensors, such as Inertial Measurement Unit, which could further refine the estimated motion. As a further work, the authors plan to derive a formal observability analysis also with the aim of identifying specific trajectories that may result in a loss of observability. Such a result would be of great interest when the proposed technique will be implemented for the navigation state estimation of inspection Unmanned Vehicles with the aim of improving the coverage performance as well as reducing the risk for human operators.

Acknowledgments

This work was carried out under the operational collaboration agreement between ARPAL and the Department of Information Engineering of the University of Pisa. The authors would like to thank all the SEALab members, the joint applied research laboratory between the Naval Support and Experimentation Centre (CSSN) of the Italian Navy and the Italian Interuniversity Research Center of Integrated Systems for Marine Environment (ISME), who supported the research team during the experimental campaign. The work was partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

References:

- Aqel, M. O. A., Marhaban, M. H., Saripan, M. I., & Ismail, N. B. 2016. Review of visual odometry: types, approaches, challenges, and applications. SpringerPlus. 5(1):1897. https://doi.org/10.1186/s40064-016-3573-7

- Buia, M. C., Gambi, M. C., & Dappiano, M. 2004. The seagrass ecosystems. In: Gambi, M.C., Dappiano, M., Eds., Mediterranean Marine Benthos: a manual for its sampling and study, Biologia Marina Mediterranea. 11(Suppl. 1):133–183.

- Burguera, A., Bonin-Font, F., Lisani, J.L., & Petro, A.B. 2016. Towards Automatic Visual Sea Grass Detection in Underwater Areas of Ecological Interest. In: IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA). pp. 1-4. Berlin, Germany. https://doi.org/10.1109/ETFA.2016.7733594

- Codevilla, F., Gaya, J. D. O., Filho, N. D., & Botelho, S. S. C. 2015. Achieving turbidity robustness on underwater images local feature detection. In: British Machine Vision Conference (BMVC). pp. 1–10. Swansea, UK. https://doi.org/10.13140/RG.2.1.3522.1842

- Costanzi, R. et al. 2019. Visual acquisition system for georeferenced monitoring and reconstruction of the sea bottom using audio for data synchronization. In: OCEANS 2019 – Marseille. pp. 1-6. Marseille, France. https://doi.org/10.1109/OCEANSE.2019.8867123

- De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D.L. 2000. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems. 50(1):1-18. https://doi.org/10.1016/S0169-7439(99)00047-7

- Ferrera, M., Moras, J., Trouve´-Peloux, P., & Creuze, V. 2019. Realtime monocular visual odometry for turbid and dynamic underwater environments. Sensors. 19(3):687. https://doi.org/10.3390/s19030687

- Fischler, M. A., & Bolles, R. C. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM. 24(6):381–395

- Franchi, M., Ridolfi, A. & Zacchini, L. 2018. A Forward-Looking Sonar-Based System for Underwater Mosaicing and Acoustic Odometry. In: 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV). pp. 1-6. Porto, Portugal. https://doi.org/10.1109/AUV.2018.8729795.

- Fraundorfer, F., & Scaramuzza, D. 2011. Visual odometry. Part I: The first 30 years and fundamentals. IEEE Robotics & Automation Magazine. 18(4):80-92. https://doi.org/10.1109/MRA.2011.943233.

- Huang, T., & Netravali, A. 1994. Motion and structure from feature correspondences: A review. Proc. IEEE. 82(2):252–268. https://doi.org/10.1142/9789812776952_0013

- Longuet-Higgins, H. 1981. A computer algorithm for reconstructing a scene from two projections. Nature. 293(10):133–135. https://doi.org/10.1038/293133a0

- Lowe, D.G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

- Mikolajczyk, K., & Schmid, C. 2004. Scale & affine invariant interest point detectors. International Journal of Computer Vision. 60(1):63–86. https://doi.org/10.1023/B:VISI.0000027790.02288.f2

- Nister, D. 2004. An Efficient Solution to the Five-Point Relative Pose Problem. IEEE Transactions on Pattern Analysis and Machine Intelligence. 26(6):756-770. https://doi.org/10.1109/TPAMI.2004.17

- Nocerino, E., Nawaf, M.M., Saccone, M., Ellefi, M.B., Pasquet, J., Royer, J.-P., & Drap, P. 2018. Multi-camera system calibration of a low-cost remotely operated vehicle for underwater cave exploration. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 42(1):329-337. https://doi.org/10.3929/ethz-b-000304138

- Paull, L., Saeedi, S., Seto, M., & Li, H. 2014. AUV navigation and localization: A review. IEEE Journal of Oceanic Engineering. 39(1):131-149. https://doi.org/10.1109/JOE.2013.2278891

- Pörtner, H.-O. et al. 2014. Ocean systems. Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. (Cambridge University Press, 2014). pp. 411–484. https://doi.org/10013/epic.45149

- Rauch, H.E., Tung, F., & Striebel, C.T. 1965. Maximum likelihood estimates of linear dynamic systems. AIAA Journal. 3(8):1445-1450. https://doi.org/10.2514/3.3166

- Ruscio, F., Costanzi, R., Pollini, L., Manzari, V., Barbieri, C., & Gaino, F. 2020. Geo-referenced visual data for Posidonia Oceanica coverage using audio for data synchronization. In: Global Oceans 2020:Singapore - U.S. Gulf Coast. pp:1-5, Gulf Coast. https://doi.org/10.1109/IEEECONF38699.2020.9389072

- Shortis, M. 2015. Calibration Techniques for Accurate Measurements by Underwater Camera Systems. Sensors. 15:30810-30826. https://doi.org/10.3390/s151229831

- Siegwart, R., Nourbakhsh, I., & Scaramuzza, D. 2011. Introduction to Autonomous Mobile Robots, 2nd ed. Cambridge, MA: The MIT Press.

- Shi, J., & Tomasi, C. 1994. Good features to track. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 593–600. https://doi.org/10.1109/CVPR.1994.323794

- Torr, P. H. S., & A. Zisserman. 2000. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. Computer Vision and Image Understanding. 78(1):138-156. https://doi.org/10.1006/cviu.1999.0832

- Zhang, Z. 2000. A Flexible New Technique for Camera Calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22(11):1330-1334. https://doi.org/10.1109/34.888718