

Object Recognition and Pose Estimation using Laser scans For Advanced Underwater Manipulation

1st Khadidja Himri

Computer Vision and Robotics Institute
Universitat de Girona
Girona, Spain
khadidja.himri@udg.edu

2st Roger Pi

Computer Vision and Robotics Institute
Universitat de Girona
Girona, Spain
roger.pi@udg.edu

3nd Pere Ridao

Computer Vision and Robotics Institute
Universitat de Girona
Girona, Spain
pere.ridao@udg.edu

4rd Nuno Gracias

Computer Vision and Robotics Institute
Universitat de Girona
Girona, Spain
nuno.gracias@udg.edu

5th Albert Palomer

Computer Vision and Robotics Institute
Universitat de Girona
Girona, Spain
albert.palomer@udg.edu

6th Narcis Palomeras

Computer Vision and Robotics Institute
Universitat de Girona
Girona, Spain
narcis.palomeras@udg.edu

Abstract—3D object recognition is an active research area in computer vision and robotics. The integration of spatial information with semantic knowledge has become an important task for robots in order to successfully perform autonomous intervention missions. This paper presents an approach for the recognition and pose estimation of underwater objects, with the goal of enabling autonomous underwater intervention in man-made structures. The methods are developed to be used with raw data consisting of 3D colorless point clouds collected by a fast laser scanner. The proposed approach contains two main phases: Object recognition from range data, and feature-based semantic SLAM. The first goal consists of recognizing different objects present in the scene. For this purpose, a recognition and pose estimation pipeline was developed enclosing different steps such as segmentation, identification, and estimation of the position and orientation for each targeted object. The second goal aims at improving the AUV navigation in an underwater environment by using the result of the recognition and pose estimation pipeline to feed a feature based SLAM algorithm. As the AUV moves along the trajectory, the SLAM algorithm builds a map, recognizes targeted objects and integrates them into this map, and localizes its position with respect to it. Compared to previous experimental results performed in a water tank, this paper emphasizes the importance of estimating the pose of the objects (namely the orientation), as a way of promoting the accuracy of the robot localization.

Index Terms—3D object recognition, 3D global descriptor, simultaneous localization and mapping (SLAM), Ensemble of Shape Functions (ESF), laser scanner, autonomous underwater vehicles (AUVs).

I. INTRODUCTION

The current state of the art in autonomous mobile manipulation is dominated by applications in indoor settings, such as home environments [1]–[3]. Robots have been demonstrated to be able to identify, locate and grasp different kitchen utensils [4].

In spite of such advances, there is still a sharp contrast of capability when compared to underwater settings, where autonomous manipulation is still performed at a very low level

of automation. Remotely Operated Vehicles (ROVs) equipped with manipulators (robot arms), are the main tool for intervention in different applications such as maintenance of offshore oil and gas structures, military and security operations, and archaeology and geology exploration [5]. In this context, a number of research centers have started pursuing the concept of an Intervention Autonomous Underwater Vehicle (I-AUV) [6] aiming at improving the automation level of underwater mobile manipulators.

Motivated by the behavior of robots in indoor environments, AUVs equipped with robot arms have been shown to perform tasks of similar complexity as indoor robots in kitchen settings. Examples of such are the grasping of objects laying on the seabed, the turning of valves on panels, or the plugging/unplugging of connectors [7].

The capability to simultaneously navigate and map the environment at the object level is now prevalent on many indoor applications [8]–[10]. For the underwater counterpart, this capability is not yet well established, in spite of the contributions that have been made towards AUV navigation autonomy [11]–[13].

In this paper, we present a navigation and mapping system for an AUV operating near a man-made environment. We proposed a pipeline for a semantic mapping system, based on colourless 3D point clouds extracted from a laser scanner.

Our goal consists of recognizing and estimating the pose (both position and orientation) of different objects in the scene using the laser scanner data. The identified objects, together with their pose, are used afterward as landmarks for simultaneous localization and mapping (SLAM).

To achieve our goal, two problems have to be addressed. Firstly, it is necessary to use a real-time scanner able to provide dense 3D point clouds in real-time with an adequately high update frequency (ideally more than 3 scans per second). Secondly, it is required to adapt the object recognition and location methods reported in state of the art in an indoor area

to the complexities of the underwater environment.

This paper follows recent paper by the same authors [14], on the use of semantic SLAM with objects recognition from point clouds. The present paper focuses on the object recognition and registration part, with emphasis on the pose estimation.

II. METHODOLOGY

The proposed approach is composed of two components: the 3D object recognition pipeline and the feature-based semantic SLAM. The following section describes to the main lines of the proposed 3D object recognition pipeline, which are summarized in the block diagram of Fig. 1.

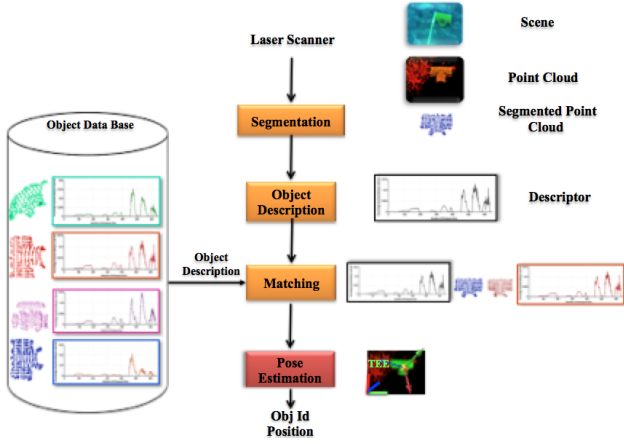


Fig. 1. Block diagram of the proposed method.

A. Object recognition and pose estimation

The proposed pipeline begins with the collection of colourless 3D point clouds of an underwater scene using a laser scanner. Target objects are then recognized among a set of possible objects available in a database. The database contains 3D models available *a priori*, which were previously scanned with Microsoft Kinect.

The different modules and their methods now are briefly described.

1) *Real time laser scanner*: In land and aerial domains, it is common for autonomous robots to sense the objects in the environment using either depth cameras or stereo cameras for 3D perception. These sensors have very limited range in the underwater environment. Infrared depth sensors are typically restricted to 20 cm. Stereo cameras can sense considerably further than infrared light, but require the presence of texture-based features to compute the 3D points, which normally implies the use of strong artificial illumination and the presence of natural texture in the environment. To overcome these limitations a real-time laser scanner was designed and developed in our lab (see [15]). The laser scanner, as shown in the Fig. 2, is composed of a laser line projector, a mirror steered with a galvanometer which projects the line at different parts of the scene through a flat viewport, and a camera with its own flat viewport. The control of the sensor, together with the laser detection implemented internally by

the camera hardware, allows the sensor to work at higher acquisition rates than commercially available laser scanners. Moreover, the sensor uses the results of [16] to increase the 3D triangulation speed by using an elliptical cone to represent the light projected underwater, as the laser line is distorted because of the refraction of the flat viewport.

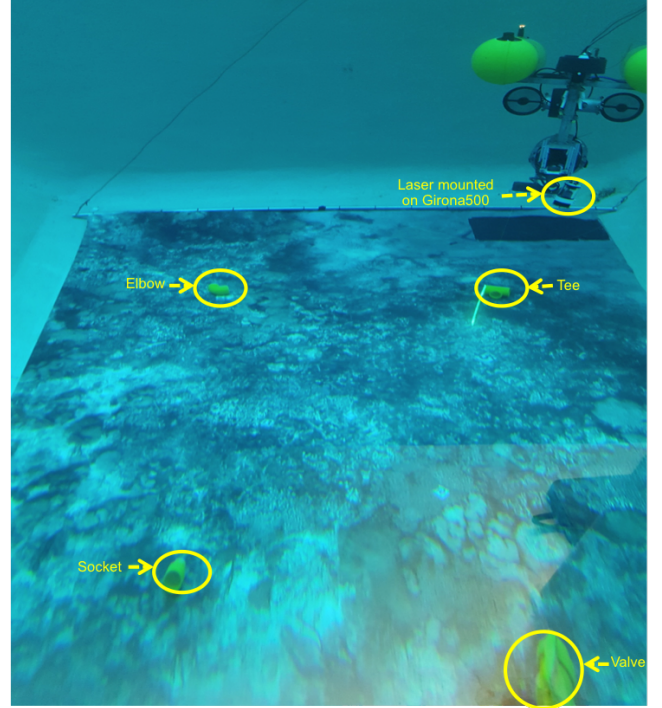


Fig. 2. Girona 500 AUV with the laser scanner mounted on it

2) *Segmentation*: The recognition pipeline implemented in this paper requires a non-trivial step of separating the 3D points belonging to the objects of interest from the rest of the scene. This step is called segmentation and consists of clustering the points representing the object in one homogeneous cluster based on similar characteristics following the approach proposed in [17].

The segmentation is divided into two stages, as proposed in [17]. For the experiments of this paper, we follow a simplifying assumption that the objects lie on a flat surface. The plan that best fits the environment is estimated using the randomized M-estimator sample consensus (RMSAC), the points in the 3D scan that belong to this plane are identified and then removed. The remaining points are clustered using the Euclidean cluster extraction [17].

3) *3D Global descriptors*: The next step uses the result of the previous step, and computes a 3D global descriptor for each segmented cluster. The advantage of using these descriptors lies on fact that they can encode the shape or geometry information in a very compact way.

Following the classification in [18], 3D Object recognition methods are split into two groups: local and global methods. Local methods are based on local descriptors and are

suitable in cluttered environments where occlusions prevent unobstructed views of the objects. These methods have an important computational cost and are therefore not suitable for real-time operation. Conversely, global methods use global descriptors and require unobstructed views. However they are adequate in real-time systems due to a lower computational costs, and have the advantage of representing the entire object in one set of features. For this reason, global methods were chosen for the results of this paper.

In our previous paper([14]), we experimented two different global descriptors: The Viewpoint Features Histogram (VFH) and the Ensemble of Shape Functions (ESF), where ESF([19]) descriptors achieved a reliable result of recognition compared to using the VFH descriptor. Thus for the current study we opted for the Ensemble of Shape Functions (ESF).

4) *Matching*: The matching consists of identifying a newly scanned object among the objects included in the database. This is done based on the comparison of the result of ESF global descriptor of the newly acquired object against descriptors of objects that have been scanned previously. After computing and sorting these distances, the smallest distance is used as the criterion to identify the newly scanned object among those in the database. Such objects form a database of objects that are recognizable.

Our database integrates four objects: a reducing tee, elbow, a reducing socket and ball valve. Each object is represented by a set of eight partial views. The acquisition of these views was performed offline and using a Microsoft Kinect sensor.

In [14] a comparison was performed between the matching using Chi-square and based on Support Vector Machines, which led to similar result. In the present work, the matching was based on the Chi-square distance, as proposed in [20] [21].

5) *Pose estimation*: After the corresponding model is identified, the last step in the object recognition is the pose estimation, which involves determining the object position and orientation. The goal is to determine the pose with respect to the canonical pose of that object in the database. This is done by aligning the scanned view with a view of the recognized object in the database. This view is referred to as the matched view or the matched point cloud.

This step is divided into two parts:

a) *Translation between two point clouds*: After the object is recognized among a set of the four objects in the database, the difference between the centroids of the point clouds of the scanned object and the object in the database is computed.

b) *Alignment of the matched point cloud*: In order to align the matched point cloud to the same orientation as the scanned target, the following steps are executed:

- Roughly alignment - Consists of identifying the position and orientation of the input scan with the matched object from the database, by following the registration pipeline proposed in [22]. It consists of computing an initial alignment using Features based registration algorithm. The Fast Points Features Histogram (FPFH) local descriptor

was used to align the points in the matched point cloud to the points in the scanned point cloud.

- Fine alignment - In the second step, an iterative registration algorithm (ICP) [23] is applied to refine the result of the initial alignment. It uses the previous result as the initial starting point to align the matched point clouds and scanned point clouds.

The final output of the pose estimation step is the relative position and orientation of the object with respect to the laser scanner.

B. Simultaneous Localization And Mapping (SLAM)

SLAM deals with the problem of building a map of an unknown environment and using it to localize the vehicle simultaneously. The navigation filter of the Girona 500 AUV uses an extended Kalman filter to combine the navigation data from a pressure sensor, a Doppler velocity log (DVL) and an attitude and heading reference system (AHRS). This filter provides dead-reckoning navigation, which drifts over time and needs absolute measurements to correct it. Those measurements can come from either a global positioning system (GPS) when the vehicle is at the surface, or through the detection of visual landmarks in a map.

A semantic EKF-based SLAM filter is proposed here where landmark measurements are given by the 3D object recognition pipeline as illustrated in Fig. 3 . The state vector for the implemented filter is the following:

$$\vec{x} = [x \ y \ z \ u \ v \ w \ l_1 \ \dots \ l_N], \quad (1)$$

where $[x \ y \ z]$ and $[u \ v \ w]$ are the position and linear velocity vectors of the AUV, and l_i is the landmark i pose vector defined as:

$$l_i = [lx_i \ ly_i \ lz_i \ l\phi_i \ l\theta_i \ l\psi_i]. \quad (2)$$

The navigation filter uses a constant velocity model with attitude input:

$$\hat{\vec{x}}_k = \begin{bmatrix} \begin{bmatrix} x_{k-1} \\ y_{k-1} \\ z_{k-1} \end{bmatrix} + \mathcal{R}(\phi_k \theta_k \psi_k) \left(\begin{bmatrix} u_{k-1} \\ v_{k-1} \\ w_{k-1} \end{bmatrix} t + \begin{bmatrix} n_{u_{k-1}} \\ n_{v_{k-1}} \\ n_{w_{k-1}} \end{bmatrix} \frac{t^2}{2} \right) \\ u_{k-1} + n_{u_{k-1}} \\ v_{k-1} + n_{v_{k-1}} \\ w_{k-1} + n_{w_{k-1}} \\ l_{1_{k-1}} \\ \vdots \\ l_{N_{k-1}} \end{bmatrix}, \quad (3)$$

where t is the sample time, $[n_u \ n_v \ n_w]$ is the noise vector and $[\phi_k \ \theta_k \ \psi_k]$ are the Euler angles used as the filter input u_k .

Object detection measurements, are integrated in the filter as linear updates being

$$z_k = [lx_i \ ly_i \ lz_i \ l\phi_i \ l\theta_i \ l\psi_i], \quad (4)$$

and

$$H = \begin{bmatrix} -\mathcal{R}(\phi_k \theta_k \psi_k)^T & 0_{3 \times 3} & \mathcal{R}(\phi_k \theta_k \psi_k)^T & 0_{3 \times 3} & \dots \\ 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & I_{3 \times 3} & \dots \end{bmatrix}, \quad (5)$$

where $[l_x \ l_y \ l_z]$ is the relative position of the landmark with respect to the vehicle, $[l\phi_i \ l\theta_i \ l\psi_i]$ is the landmark orientation with respect the inertial frame, and $\mathcal{R}(\phi_k \theta_k \psi_k)$ is the vehicle orientation rotation matrix at time k .

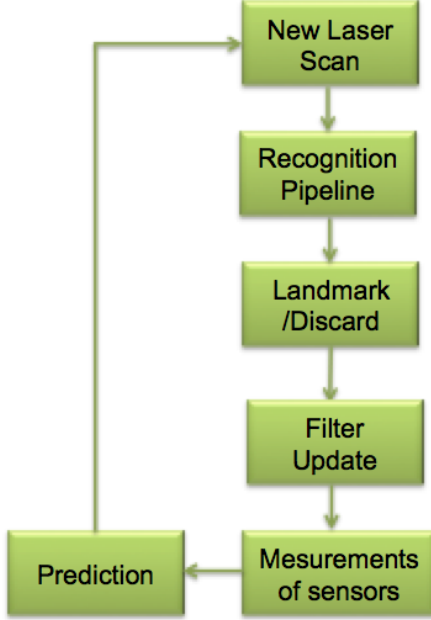


Fig. 3. Fundamental steps of EKF-SLAM .

III. EXPERIMENTAL RESULTS

This section presents the experimental results on using the laser scanner mounted on the AUV GIRONA500 and operating in CIRS water tank. A bottom laying poster was used for computing a trajectory ground truth. Four pressure PVC pipes where placed on the bottom of the tank, as seen in Fig. 2.

The object database includes the scanned point clouds of the four PVC objects, consisting of a reducing tee object of $107 \times 74 \times 63 \text{ mm}^3$, a reducing socket of $57 \times 43 \times 63 \text{ mm}^3$, an elbow of $76 \times 41 \times 63 \text{ mm}^3$ and ball valve of $85 \times 125 \times 63 \text{ mm}^3$.

In order to close the loop, the GIRONA500 was teleoperated to follow an approximated square trajectory as shown in Fig. 2, starting from a position where the reducing socket was within the field of view of the laser scanner, and ending near the ball valve after passing through the elbow and the reducing tee.

As the robot measurement given by the compass and the IMU was nearly precise, an amount of noise was generated in the attitude and angular velocity estimation, with the purpose of emphasizing the improvement made by adding feature orientation to the filter updates.

In order to generate the ground truth a vision based localization method with an *a priori* known map was used. A poster image representing a seafloor (Fig. 5) was laid on the bottom of the water tank. An image registration algorithm was used

to register the images grabbed with the robot camera, to the image poster estimating the camera pose [24].

During the execution, each time a new object was recognized, a new feature had to be introduced in the map. To ensure consistency, several observations of the same feature were used (see Fig. 4) and a simple *ad hoc* consensus algorithm was used to reject outliers using the most promising observation for the feature initialization. For each pair of observations, the algorithm computed the position and attitude errors. The euclidean distance was used to compute the position error, while for the error in attitude the next formula was used:

$$d = 1 - \langle q_1, q_2 \rangle^2 \quad (6)$$

Two error thresholds were defined, one for the position and one for the attitude. Next, the observation which had more consensus in both, position and attitude was selected for the initialization.

In Fig. 4 the blue arrows represent the pose with highest consensus, the red ones are outliers and the green arrows represent poses in consensus with at least one observation. The cyan arrow is the average of the blue ones, being used for the initialization.

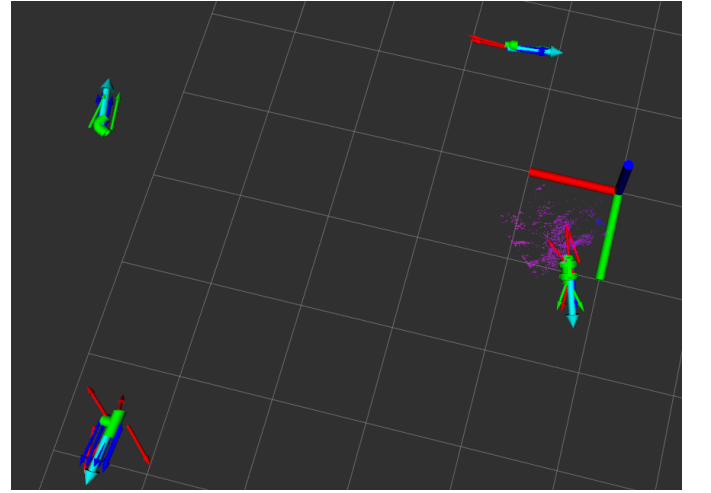


Fig. 4. Feature initialization based on Adhoc, the red arrow represents outliers, the blue one represents the consensus, the cyan the average of the consensus.

Whenever a feature was re-observed the corresponding object was identified and its pose was computed. Then, the Mahalanobis distance was used to test the compatibility of the pose. If the test was passed, an update using the observation pose was applied to the filter. Otherwise, the Mahalanobis distance involving only the position was re-computed performing a position only compatibility test. If successful, a position update was applied to the filter. In any other case no update was applied.

The results are shown in Fig. 6. The abscissa represents time-stamps and the ordinate the robot position $\pm 3\sigma$. The four colors: Black, red, blue, and purple represent respectively: the uncertainty of the EKF-SLAM filter without any update,

updated using only the position, updated using the pose and the ground truth. As expected, it can be appreciated that the dead reckoning uncertainty grows without limits. The position only SLAM is able to reduce the uncertainty at each feature observation bounding the uncertainty in the AUV position. Better performance was obtained with the SLAM using the feature position and attitude, which further reduces the robot uncertainty, and is more effective in the navigation estimation.

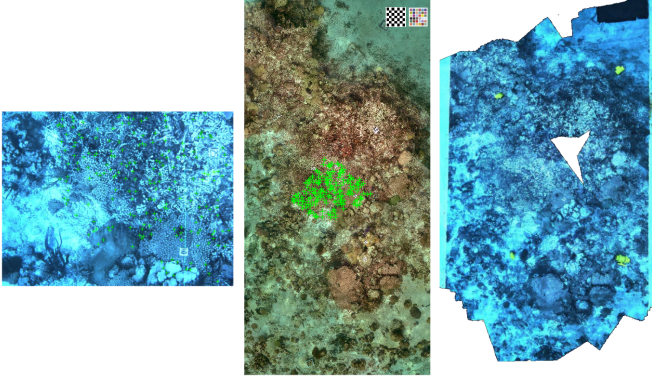


Fig. 5. Ground-truth creation using image registration. Original image acquired during the test (left), registration of this image against an image of the poster laying at the bottom of the test pool (center), and mosaic obtained from merging all images used for ground-truth (right).

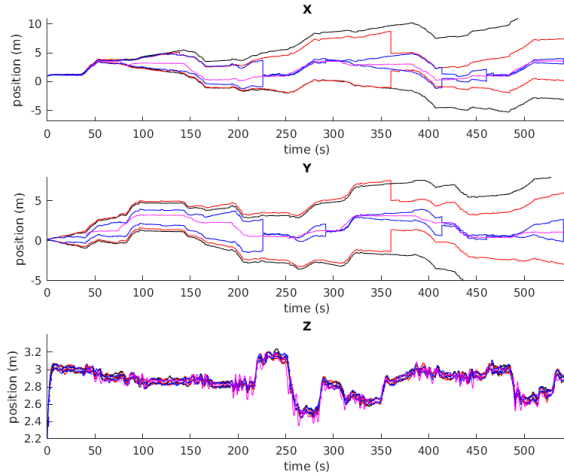


Fig. 6. Evolution of the covariance matrix in the three robot position (X, Y, Z).

IV. CONCLUSION

In this paper, we presented an object recognition and pose estimation pipeline with the ultimate objective of using it in a real-time localization and mapping system. A simple consensus algorithm has been used to estimate the initial pose of the features from several scans. Next, we have compared the estimation of the robot position using an SLAM method incorporating the object position and the object pose, obtaining

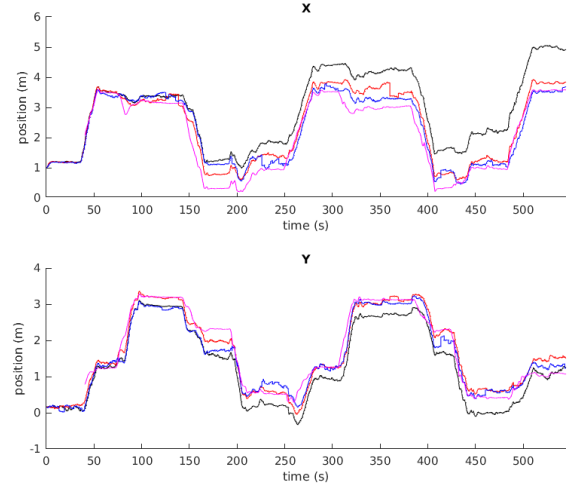


Fig. 7. Robot position belief.

better results in the second case as expected. Further work will focus on increasing the size of the object data base as well as allowing interconnection among the objects.

ACKNOWLEDGMENT

This work was supported by the Spanish Government through a FPI Ph.D. grant to K. Himri and projects TWIN-BOT (subproject: GIRONA1000) and STRONGMAR (with references: DPI2017-86372-C3-2-R and H2020-TWINN-2015 (CSA)-692427 respectively).

REFERENCES

- [1] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927–941, 2008.
- [2] R. B. Rusu, B. Gerkey, and M. Beetz, "Robots in the kitchen: Exploiting ubiquitous sensing and actuation," *Robotics and Autonomous Systems*, vol. 56, no. 10, pp. 844–856, 2008.
- [3] N. Blodow, L. C. Goron, Z.-C. Marton, D. Pangercic, T. Rühr, M. Tenorth, and M. Beetz, "Autonomous semantic mapping for robots performing everyday manipulation tasks in kitchen environments," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pp. 4263–4270, IEEE, 2011.
- [4] R. B. Rusu, "Semantic 3d object maps for everyday manipulation in human living environments," *KI-Künstliche Intelligenz*, vol. 24, no. 4, pp. 345–348, 2010.
- [5] R. Capocci, G. Dooly, E. Omerdić, J. Coleman, T. Newe, and D. Toal, "Inspection-class remotely operated vehicles: a review," *Journal of Marine Science and Engineering*, vol. 5, no. 1, p. 13, 2017.
- [6] P. Ridao, M. Carreras, D. Ribas, P. J. Sanz, and G. Oliver, "Intervention auvs: the next challenge," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 12146–12159, 2014.
- [7] A. Carrera Viñas, N. Palomeras Rovira, N. Hurtós Vilarnau, P. Kormushev, and M. Carreras Pérez, "Cognitive system for autonomous underwater intervention," *Pattern Recognition Letters*, 2015, vol. 67, núm. 1, p. 91–99, 2015.
- [8] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 1352–1359, IEEE, 2013.
- [9] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 1722–1729, IEEE, 2017.

- [10] M. Dos Santos, P. O. Ribeiro, P. Núñez, P. Drews-Jr, and S. Botelho, "Object classification in semi structured environment using forward-looking sonar," *Sensors*, vol. 17, no. 10, p. 2235, 2017.
- [11] D. Ribas, P. Ridao, J. D. Tardós, and J. Neira, "Underwater slam in a marina environment," in *IEEE/RSJ International Conference on Intelligent Robots and Systems: 2007: IROS 2007, 2007*, p. 1455-1460, IEEE, 2007.
- [12] L. Chen, S. Wang, K. McDonald-Maier, and H. Hu, "Towards autonomous localization and mapping of auvs: a survey," *International Journal of Intelligent Unmanned Systems*, vol. 1, no. 2, pp. 97–120, 2013.
- [13] X. Yuan, J.-F. Martínez-Ortega, J. A. S. Fernández, and M. Eckert, "AEKF-SLAM: a new algorithm for robotic underwater navigation," *Sensors*, vol. 17, no. 5, p. 1174, 2017.
- [14] K. Himri, P. Ridao, N. Gracias, A. Palomer, N. Palomeras, and R. Pi, "Semantic slam for an auv using object recognition from point clouds," *IFAC-PapersOnLine*, vol. 51, no. 29, pp. 360–365, 2018.
- [15] A. Palomer, P. Ridao, J. Forest, and D. Ribas, "Underwater Laser Scanner: Ray-based Model and Calibration," *Manuscript submitted for publication*, pp. 1–11, 2018.
- [16] A. Palomer, P. Ridao, D. Ribas, and J. Forest, "Underwater 3D laser scanners: The deformation of the plane," in *Lecture Notes in Control and Information Sciences* (T. I. Fossen, K. Y. Pettersen, and H. Nijmeijer, eds.), vol. 474, pp. 73–88, Springer, 2017.
- [17] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in domestic environments," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pp. 1–6, IEEE, 2009.
- [18] Y. Guo, F. Sohel, M. Bennamoun, M. Lu, and J. Wan, "Rotational projection statistics for 3d local surface description and object recognition," *International journal of computer vision*, vol. 105, no. 1, pp. 63–86, 2013.
- [19] W. Wohlkinger and M. Vincze, "Ensemble of shape functions for 3d object classification," in *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, pp. 2987–2992, IEEE, 2011.
- [20] R. B. Rusu, Z. C. Marton, N. Blodow, and M. Beetz, "Persistent point feature histograms for 3d point clouds," in *Proc 10th Int Conf Intel Autonomous Syst (IAS-10), Baden-Baden, Germany*, pp. 119–128, 2008.
- [21] G. Hetszel, B. Leibe, P. Levi, and B. Schiele, "3d object recognition from range images using local feature histograms," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, pp. II–II, IEEE, 2001.
- [22] D. Holz, A. E. Ichim, F. Tombari, R. B. Rusu, and S. Behnke, "Registration with the point cloud library: A modular framework for aligning in 3-d," *IEEE Robotics & Automation Magazine*, vol. 22, no. 4, pp. 110–124, 2015.
- [23] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Robotics-DL tentative*, pp. 586–606, International Society for Optics and Photonics, 1992.
- [24] N. Gracias and J. Santos-Victor, "Trajectory reconstruction with uncertainty estimation using mosaic registration," *Robotics and Autonomous Systems*, vol. 35, pp. 163–177, July 2001.