VISION-BASED SLAM FOR ROVS: PRELIMINARY EXPERIMENTAL RESULTS

M. Caccia *

* *CNR-ISSIA Genova, Via De Marini 6, 16149 Genova, Italy*

Abstract: This paper focuses on preliminary results of a vision-based simultaneous localisation and mapping (SLAM) system for ROVs operating in the proximity of the seabed. In particular, the experiments carried out with the Romeo ROV in typical operating conditions demonstrate the possibility of defining suitable *unambiguous* features, i.e. image templates, to be extracted automatically by the system and used as landmarks. The overall system is based on and integrated with the already developed *laser triangulation optical correlation* (LTOC) sensor, for the estimate of the camera/vehicle linear motion on the basis of monocular video data.

Keywords: SLAM, underwater vision, ROVs.

1. INTRODUCTION

The ability for a robotic vehicle to move in an unknown environment building a map, using only relative observations of the environment, and simultaneously using this map to navigate has been defined as a "Holy Grail" of autonomous vehicle research community (Dissanayake *et al.*, 2001). The problem has been theoretically formalized and solved using extended Kalman filter, and a number of more efficient solutions, from the point of view of algorithmic complexity, computing time, resource allocation, and data association management have been proposed (Guivant and Nebot, 2001)(Williams, 2001)(Julier and Uhlmann, 2001)(Stachniss *et al.*, 2005).
SLAM techniques could be of great help for unmanned underwater vehicles to compensate biases of dead-reckoning positioning systems when working in a constrained operating area. In this case, however, the main problem is to identify *unambiguous* features, to be used as landmarks, in a natural unstructured environment.
In the field of underwater robotics, theoretical research has been carried out leading to the development of constant-time SLAM through decoupled stochastic mapping. Preliminary experimental results have been presented relying on data acquired by a mechanically scanned sonar in a tank (Leonard and Feder, 2001), and by a sixteen element synthetic aperture sonar (SAS) carried on the nose of the AUV "Caribou" (Newman *et al.*, 2003). In the latter case, data gathered from a 40 minute survey were processed by the constant-time SLAM algorithm and the results satisfactorily compared to both a ground truth and the quadratic time full covariance SLAM algorithm. Anyway, data association and measurement clustering were performed manually. The difficulties in automatically identifying suitable features from sonar data in typical operating environment motivated the adoption of a 3-D evidence grid in (Fairfield *et al.*, 2005), where a particle filter based SLAM algorithm was applied to data collected by an array of pencil beam sonars mounted on the DEPTHX autonomous hovering vehicle during the exploration of a flooded cenote.
When working near the seabed in clear water the possibility of using underwater optical vision for motion estimation and localization has been explored in the last years. In (Williams and Mahon, 2004), an interesting SLAM algorithm based on the fusion of sonar (range and bearing) and vision (bearing) information for feature initialization and tracking respectively is presented. Preliminary results, obtained processing data collected by the Oberon ROV on the Australian great barrier reef, point out the need of

developing methods capable of tracking and identifying large numbers of points within the SLAM framework. Mosaic-based concurrent mapping and localization has been performed using Phantom class ROVs by (Negahdaripour and Xu, 2002) and (Gracias *et al.*, 2003) demonstrating the high accuracy of vision-based techniques in underwater motion estimation, even through relevant tests at sea as in the case of the autonomous navigation of a Phantom 500SP ROV over previously acquired mosaics (Gracias *et al.*, 2003).

In this context, the main contribution of this paper is the demonstration, through extended tests at sea in operating conditions carried out with the Romeo ROV (Caccia *et al.*, 2000), that *unambiguous* image templates can be detected and tracked in an underwater natural environment and used as landmarks for SLAM. The proposed approach, based on the identification of high local variance image templates, as originally discussed in (Misu *et al.*, 1999), was already satisfactorily applied to the development of a laser-triangulation optical-correlation (LTOC) sensor for the estimate of ROV altitude and horizontal speed in the proximity of the seabed.(Caccia, 2005).

The paper is organised as it follows. The architecture of the upgraded LTOC system for SLAM purposes is presented in section 2, where the introduction of new modules and their relationship with the existing system is discussed as well as the basic function of each component. A brief overview of problem formulation, vehicle, landmark and observation models, and landmark initialization, data association and estimation process is given is section 3. Experimental set-up and results are discussed in section 4.

2. SYSTEM DESIGN

The proposed vision-based SLAM system for ROVs consists of a laser-triangulation optical-correlation sensor for surge and sway estimation, integrated with a conventional SLAM algorithm, and a detector and tracker of *high confidence* image templates used as landmarks.

As in the case of the laser-triangulation optical-correlation sensor presented in (Caccia, 2005), the main idea is that, in a natural environment, image templates are characterised by high local variance, which can also be seen as a measurement of their observability. In particular, a method for automatic extraction of areas of interest, originally proposed for the case of autonomous spacecraft landing (Misu *et al.*, 1999), has been adopted: after a 2-D band-pass filtering to enhance specific spatial wavelengths, local variances are computed to evaluate contrast, and high-local variance areas are extracted as templates. Both SLAM visual landmarks and image templates used for inter-frame motion estimation are extracted with the same algorithm, the only difference is that, to reduce ambiguity, SLAM areas of interest are larger and have a higher threshold on the value of the local variance.

It is worth noting that, as in (Caccia, 2005), only one video camera is used for image depth, i.e. vehicle altitude, estimation through laser spot tracking, speed estimation and SLAM through image template tracking.

As shown in Figure 1, the *red* component of the image is processed by the *laser-triangulation altimeter* which computes the image depth Z, i.e. range from the seabed, on the basis of the image coordinates of the laser spots. The vision-based estimate of the scene depth can be integrated with altitude measurements h_S supplied by acoustic altimeters mounted on the vehicle to increase system reliability and field of work.

Since the laser spots induce artificial high local variance regions in the scene, the detection of natural image templates is inhibited in the proximity of their image coordinates m_L. Natural image templates, defined as discussed before, are managed by the *speed* and *SLAM optical template detector and tracker* modules. Since the ROV is assumed to work at constant heading and altitude, template tracking is performed through the computation of the normalised correlation coefficient between the image and the template. Token tracking fails when the correlation gets lower than a suitable threshold. In order to reduce computational workload, only a neighborhood of the predicted location, computed according to the estimated motion, is considered in the case of image templates used for inter-frame motion estimation. It is worth noting that, for the sake of coding simplicity, at each step a set of new image templates for speed estimation are detected and tracked. In order to have independent measurements of speed and position, *speed* templates cannot overlap *SLAM* templates.

The vehicle horizontal speed is computed by the *Motion from tokens speed estimator* in the camera-fixed reference frame from token displacements in consecutive images assuming that the image depth is supplied by the *laser triangulation altimeter*. The *SLAM* module implements a classical SLAM algorithm, updating the system augmented state without any care to computing and memory optimisation problems since the goal, at this stage, is to verify the possibility of using natural visual landmark for positioning.

3. PROBLEM FORMULATION

3.1 Vehicle and landmark models

Denoting with x, y and ψ, the vehicle horizontal position and orientation in an earth-fixed reference frame and with u and v its surge and sway speed (with respect to the ground) in a body-fixed reference frame, ROV kinematics is modelled by:

$$\begin{bmatrix} x(k+1) \\ y(k+1) \end{bmatrix} = \begin{bmatrix} x(k) \\ y(k) \end{bmatrix} + \mathbf{R}\left(\psi(k)\right) \begin{bmatrix} u(k)+\nu_u(k) \\ v(k)+\nu_v(k) \end{bmatrix} \Delta t \quad (1)$$

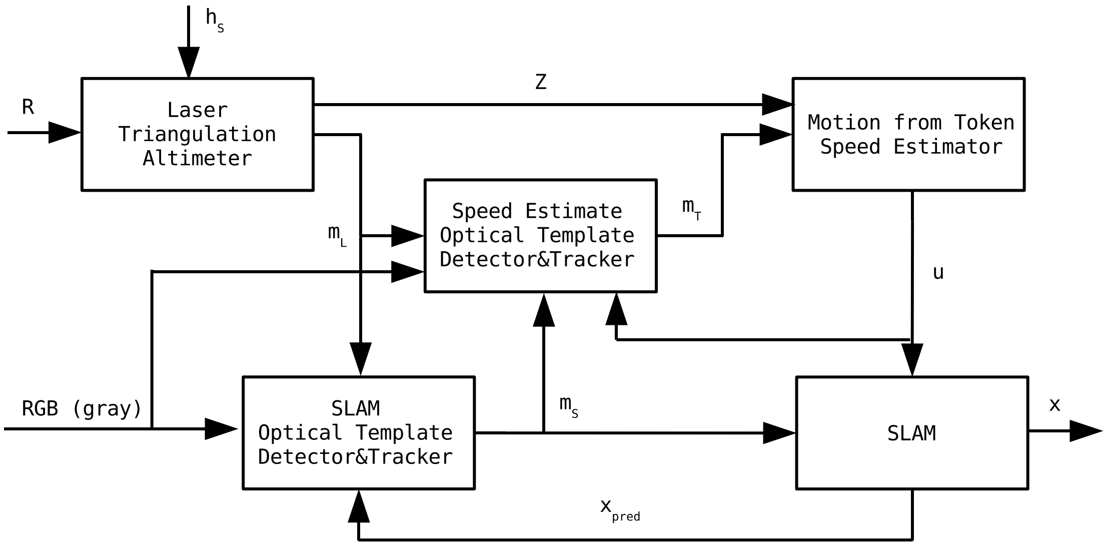


Fig. 1. Vision-based SLAM and speed estimate system design. Nomenclature: R, RGB: image components, h_S: sonar altitude, Z: image depth, m_L: laser spot coordinates, m_T: *speed* templates coordinates, m_S: *SLAM* template coordinates, u: surge and sway speed, x: vehicle and landmarks estimated position, x_{pred}: vehicle and landmarks predicted position

where $\mathbf{R} = \begin{bmatrix} cos\psi & -sin\psi \\ sin\psi & cos\psi \end{bmatrix}$, and $\nu_u(k)$ and $\nu_v(k)$ embed the control noise with covariance $Q(k)$. The explicit representation of the noise of the control action captures the structure of classical SLAM kinematic model where the vehicle state coincides with its position and typical control inputs are the robot horizontal speed and heading. The encapsulation of the uncertainty on these quantities in the control input allows its projection on the robot motion direction rather than a generic, overestimated, representation in terms of system noise on the vehicle's cartesian coordinates. Since the vehicle is assumed to operate at constant altitude Z from the seabed, the visual landmarks are assumed to have a fixed horizontal location $[X, Y]^T$. The resulting model for the i-th feature is

$$\begin{bmatrix} X_i(k+1) \\ Y_i(k+1) \end{bmatrix} = \begin{bmatrix} X_i(k) \\ Y_i(k) \end{bmatrix} + \begin{bmatrix} \nu_{X_i}(k) \\ \nu_{Y_i}(k) \end{bmatrix} \quad (2)$$

3.2 Observation model and landmark initialisation

The observation model of the i-th landmark is based on the camera perspective model and, assuming a downward-looking camera positioned in the origin of the vehicle-fixed frame, it gets:

$$\begin{bmatrix} m_i \\ n_i \end{bmatrix} = \frac{f}{Z}\mathbf{R}^T(\psi) \begin{bmatrix} X_i - x \\ Y_i - y \end{bmatrix} + \begin{bmatrix} \nu_m \\ \nu_n \end{bmatrix} \quad (3)$$

where m_i and n_i are the image coordinates in pixel of the i-th feature, f is the camera focal length, and ν_m and ν_n represent the measurement noise.

According to eqn. (3), the i-th landmark is initialized as

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \frac{Z}{f}\mathbf{R}(\psi) \begin{bmatrix} m_i \\ n_i \end{bmatrix} \quad (4)$$

and its covariance is given by $\mathbf{P}_{\underline{\hat{X}}_i} = \mathbf{P}_{\underline{\hat{x}}} + \frac{Z^2}{f^2}\mathbf{R}_m$, where $\mathbf{P}_{\underline{\hat{x}}}$ and $\mathbf{P}_{\underline{\hat{X}}_i}$ are the covariance matrices of the vehicle and i-th landmark location estimates, and $\mathbf{R}_m$. is the measurement noise covariance, which takes into account also the stochastic effects of small errors in the estimate of the feature altitude.

3.3 Data association and estimation process

Estimates of the vehicle and landmark locations are provided by Kalman filter according to the theoretic formulation of the SLAM problem, summarised in (Dissanayake *et al.*, 2001) and omitted in the following for the sake of brevity. Here, it is sufficient to remember that the prediction of the augmented state, i.e. vehicle and landmark locations, and its covariance is performed according to equations (1) and (2), where the measurements of surge and sway speed are provided by the optical correlation-based *motion from token speed estimator*.

On the other hand, the landmark association process is crucial for the correct operation of the SLAM estimator. Since landmarks are recorded as image templates, the data association process basically consists in searching in the current image any landmark template that could be observed by the camera. In particular, given the expected vehicle position $\underline{\hat{x}}$, the i-th landmark, whose expected location is $\underline{\hat{X}}_i$, can be in the camera field of view if

$$\left(\underline{\hat{X}}_i - \underline{\hat{x}}\right)^T \left(\mathbf{P}_{\underline{\hat{X}}_i} + \mathbf{P}_{\underline{\hat{x}}} + \mathbf{D}\right)^{-1} \left(\underline{\hat{X}}_i - \underline{\hat{x}}\right) \leq \lambda \quad (5)$$

where $\mathbf{D} = \frac{Z^2}{f^2}\mathbf{R}(\psi) diag\,(H/2, W/2)$, being H and W the image height and width respectively, represents the size of the seabed observed by the camera. In the case, the i-th landmark satisfies the condition (5) and its image template is tracked in the acquired image, the observation $[m_i\ n_i]^T$ is associated with itself and used for updating the state estimate and covariance according to Kalman filter equations.

3.4 Independence of speed and position measurements

The estimation-theoretical formulation of SLAM relies on the assumption that the vector of control input, i.e. the vision-based estimated speed, and observations, i.e. landmark image coordinates, are independent. Since both templates for speed estimation and position updates are extracted by the same images, attention has to be paid to avoid any overlapping between them. To do that, the tracking and extraction of landmark image template is performed before of the speed measurements and SLAM estimation process. Thus, at the generic k-th step the following actions are executed:

(1) laser spots are tracked or extracted from the red component of the image;
(2) pixels in the neighborhood of the laser spots are inhibited for template extraction and tracking;
(3) *SLAM* templates are searched for and tracked, and, in the case no landmark is tracked, a new one is searched if the vehicle is sufficiently far from the existing ones;
(4) pixels in the neighborhood of the tracked/new *SLAM* template are inhibited for *speed* template extraction and tracking;
(5) *speed* templates are tracked and surge and sway speed are computed;
(6) SLAM prediction, observation and update are performed;
(7) SLAM state is augmented with the new landmark (if a new landmark has been extracted);
(8) *speed* templates are extracted.

It is worth noting that to reduce the number and to avoid spatial clusters of landmarks, new SLAM templates are searched for only if the vehicle is far enough by the existing one according to a criterium of the type:

$$\forall i \in [1, N]\ , \quad \left(\underline{\hat{X}}_i - \underline{\hat{x}}\right)^T \left(\mathbf{P}_{\underline{\hat{X}}_i} + \mathbf{P}_{\underline{\hat{x}}} + \mathbf{L}\right)^{-1} \left(\underline{\hat{X}}_i - \underline{\hat{x}}\right) \geq \mu \qquad (6)$$

where N is the number of existing landmarks, and $\mathbf{L}$ is a matrix fixing the minimum range between landmarks.

4. EXPERIMENTAL SET-UP AND RESULTS

Experiments have been carried out on data collected by the Romeo ROV equipped with a laser-triangulation optical-correlation sensor mounted downward-looking below the vehicle in the Portofino Park area, Ligurian Sea, in Summer 2005. During the experiments the vehicle worked in auto-altitude navigating through way-points on the basis of vision-based estimates of its horizontal speed and altitude (Caccia, 2004). The recorded images of 360 x 272 pixels have been post-processed for SLAM implementation and evaluation with a laptop equipped with an Intel Pentium 4 at 3.06 GHz CPU running the Linux operating system. In order to simplify the problem during the first demonstration phase, the ROV worked at constant heading, so matrix $\mathbf{R}(\psi)$ can be assumed equal to the identity
The size of image templates for speed estimate was 16x16 pixels, while landmark templates measure 32x32 pixels. The normalized correlation threshold for template tracking was 0.89, while the local variance minimum value for SLAM templates was 12.3. No more than eight templates are tracked at each step for estimating the vehicle speed.
During the trial discussed in the following, which took 44 minutes and 15 seconds and 13275 processed images, the vehicle moved along a two-dimensional grid at a constant altitude of 1.31 m. The ROV altitude and surge and sway speed measured by the LTOC sensor are shown in Figure 2. Figure 3 shows the vehicle

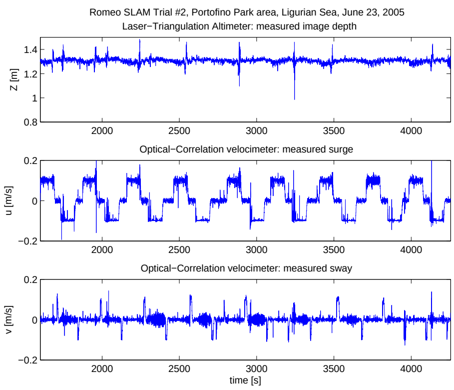
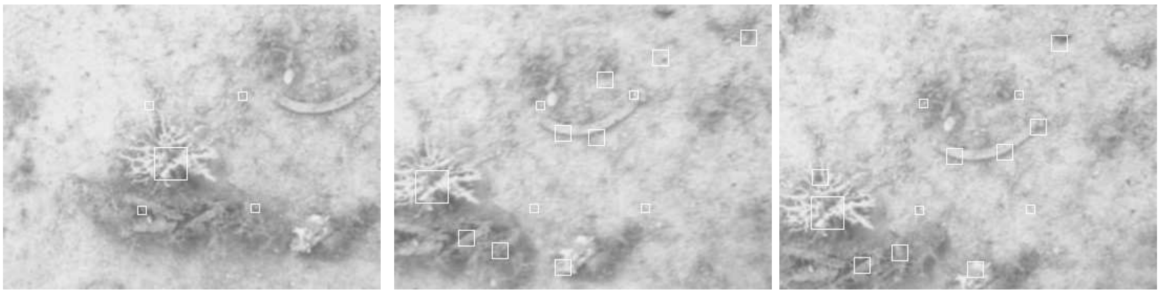
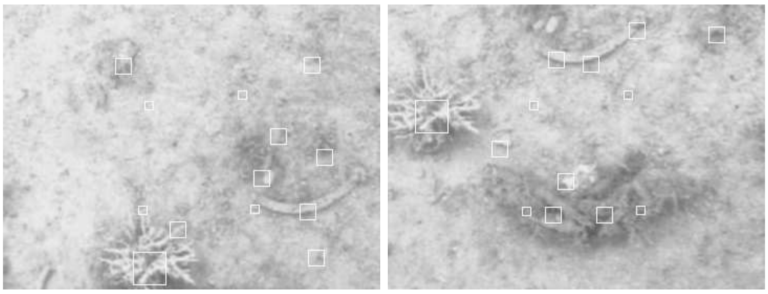


Fig. 2. ROV altitude and horizontal speed measured by the laser-correlation optical-correlation sensor.

estimated position together with the identifiers of the tracked SLAM landmarks. As shown in the first and second graph from the top, the corrections executed by the estimated ROV x and y position are always between the bounds given by the expected estimate standard deviations. The ROV estimated path and the estimated landmark positions are shown in Figure 4. As far as the quality of the automatically identified visual landmarks is concerned, as well as the tracking capability of the system, Figure 5 shows the tracked image template corresponding to landmark 0 at different passages of the ROV over its location. The tracked small image templates used for estimating the ROV speed are also displayed. The estimated vehicle position (denoted by hat) at the corresponding times are reported in Table 1, together with the displacement

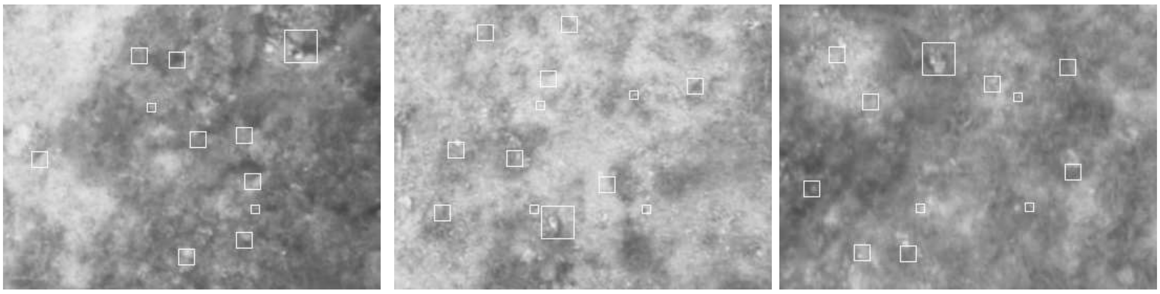


(a) t=1600.0 s (b) t=1848.4 s (c) t=2133.3 s



(d) t=2753.7 s (e) t=4234.4 s

Fig. 5. ROV camera views at times when landmark 0 is initialized or tracked.



(a) t=1681.8 s (b) t=2299.7 s (c) t=2316.7 s

Fig. 6. ROV camera views at times when landmarks 3, 6 and 7 are initialized.

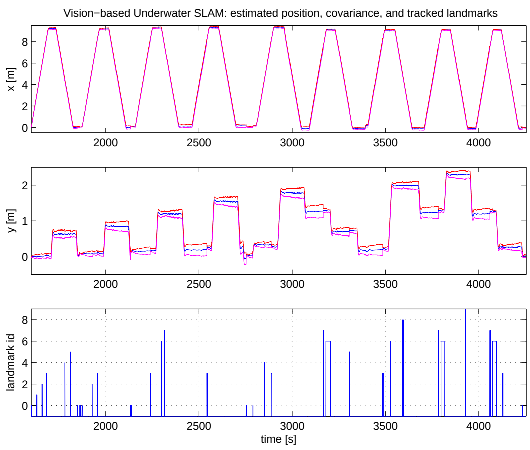


Fig. 3. ROV estimated position and tracked SLAM landmark identifiers.

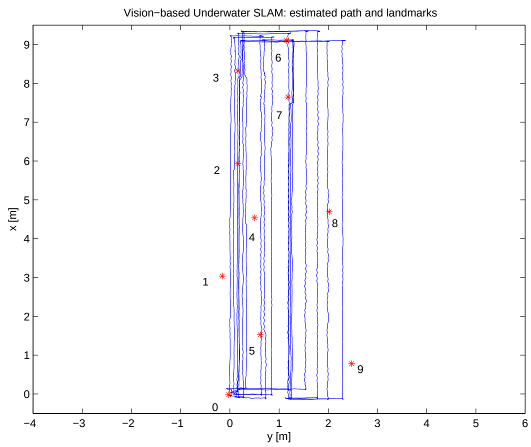


Fig. 4. ROV estimated path and SLAM landmark locations.

measurements (denoted by the subscript GT) obtained by tracking a set of 16 by 16 pixel templates in the first image and in any other one. It is worth noting that this was the only ground-truth verification compatible with the ROV configuration, which does not include high precision DVL and gyros or high frequency LBL. Considering that the x and y axes are directed upward and rightward respectively, the size of the area covered by the image measures about 33x44 cm, and to a positive displacement of the vehicle should correspond a mirrored motion of the image template, the reader can verify the precision of the SLAM system for positioning the ROV with respect to a point of the environment. In order to provide the reader a feeling of

Table 1. Estimated ROV position while tracking landmark 0

time [s]	$\hat{x}$ [m]	$\hat{y}$ [m]	x_{GT}	y_{GT}
1600.0	0.00	0.00	0.00	0.00
1848.4	0.03	0.15	0.03	0.15
2133.3	0.06	0.14	0.06	0.14
2753.7	0.12	0.02	0.12	0.02
4234.4	-0.06	0.14	-0.06	0.14

how the system identifies environmental features, the image templates corresponding to the landmarks 3, 6 and 7 selected during operations are shown in Figure 6.

5. CONCLUSIONS

This paper presents some encouraging experimental results in automatically extracting visual landmarks for simultaneous mapping and localization in underwater environment. On the basis of these preliminary results, current research focuses on developing a suitable SLAM-based navigation system able to allow an ROV to move with high accuracy to a set of points of interest specified by the user.

ACKNOWLEDGMENTS

The author is particularly grateful to Riccardo Bono, Gabriele Bruzzone, Giorgio Bruzzone and Edoardo Spirandelli for their fundamental support in software and hardware development and at sea operations.

REFERENCES

Caccia, M. (2004). Vision-based ROV horizontal motion control: experimental results. In: *Proc. of CAMS 2004*.

Caccia, M. (2005). Laser-triangulation optical-correlation sensor for ROV slow motion estimation. *IEEE Journal of Oceanic Engineering*.

Caccia, M., R. Bono, G. Bruzzone and G. Veruggio (2000). Unmanned underwater vehicles for scientific applications and robotics research: the ROMEO project. *Marine Technology Society Journal* **24**(2), 3–17.

Dissanayake, M.W.M., P. Newman, S. Clark, H.F. Durrant-Whyte and M. Csorba (2001). A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation* **17**(3), 229–241.

Fairfield, N., G.A. Kantor and D. Wettergreen (2005). Three dimensional evidence grids for SLAM in complex underwater environments. In: *Proc. of 14th International Symposium of Unmanned Untethered Submersible Technology (UUST)*.

Gracias, N.R., S. Van der Zwaan, A. Bernardino and J. Santos-Victor (2003). Mosaic-based navigation for autonomous underwater vehicles. *IEEE Journal of Oceanic Engineering* **28**(4), 609–624.

Guivant, J.E. and E.M. Nebot (2001). Optimization of the simultaneous localization and map-building algorithm for real-time mplementation. *IEEE Transactions on Robotics and Automation* **17**(3), 242–257.

Julier, S.J. and J.K. Uhlmann (2001). Simultaneous localization and map building using split covariance intersection. In: *Proc. of 2001 IEEE International Conference on Intelligent Robots and Systems*. Maui, USA. pp. 1257–1262.

Leonard, J.J. and H.J.S. Feder (2001). Decoupled stochastic mapping. *IEEE Journal of Oceanic Engineering* **26**(4), 561–571.

Misu, T., T. Hashimoto and K. Ninomiya (1999). Optical guidance for autonomous landing of spacecraft. *IEEE Transactions on aerospace and electronic systems* **35**(2), 459–473.

Negahdaripour, S. and X. Xu (2002). Mosaic-based positioning and improved motion-estimation methods for automatic navigation of submersible vehicles. *IEEE Journal of Oceanic Engineering* **27**(1), 79–99.

Newman, P.M., J.J. Leonard and R.J. Rikoski (2003). Towards constant-time SLAM on an autonomous underwater vehicle using synthetic aperture sonar. In: *Proc. of 11th International Symposium on Robotics Research*.

Stachniss, C., G. Grisetti and W. Burgard (2005). Recovering particle diversity in a rao-blackwellized particle filter for SLAM after actively closing loops. In: *Proc. of the 2005 IEEE International Conference on Robotics and Automation*. pp. 655 – 660.

Williams, S. and I. Mahon (2004). Simultaneous localisation and mapping on the great barrier reef. In: *Proc. ofIEEE International Conference on Robotics and Automation ICRA 2004*.

Williams, S.B. (2001). Efficient solutions to autonomous mapping and navigation problems. PhD thesis. University of Sidney, Australian Center for Field Robotics.