Real-Time Localization and Dense Mapping in Underwater Environments from a Monocular Sequence

Alejo Concha*, Paulo Drews-Jr†‡, Mario Campos‡ and Javier Civera*
*I3A - Universidad de Zaragoza, Spain alejocb, jcivera@unizar.es
†NAUTEC - Universidade Federal do Rio Grande, Brazil paulodrews@furg.br
‡VeRLab - Universidade Federal de Minas Gerais, Brazil mario@dcc.ufmg.br

***Abstract*—In this paper we present an algorithm that estimates in real-time a 3D dense reconstruction of an underwater scene and the vehicle pose, being the *only* input a monocular image sequence. Our algorithm selects a set of keyframes from a seabed sequence and estimates a depth *for every pixel* from the information contained in the images using direct mapping methods. The procedure does not require extra sensing input or assumptions about the scene. Our experimental results in a pool and a seabed sequence show that such minimal sensing configuration can achieve a high degree of accuracy.**

I. Introduction

The accurate tridimensional reconstruction of underwater scenes is an active area of research having three main important areas of application; namely the autonomous navigation of underwater robots [1], the registration of seabed natural environments for its posterior study [2], [3], and the inspection of underwater structures (e.g., marinas, ship hulls or pipelines) for assessment and maintenance [4], [5]. The basic algorithms come from the robotic field called SLAM [6], standing for Simultaneous Localization and Mapping, that aims to estimate the robot pose and a 3D geometric map of the scene from sensor data.

The rapid attenuation of the electromagnetic signals in aquatic medium constraints the sensing possibilities for marine robots, e.g., neither GPS nor LIDAR can be used. [7], [8] are two recent surveys on the topic of underwater localization, mapping and navigation. Sonar has been successfully used in structured marina-like environments [1], [9]; but it captures limited information of the environment with low accuracy.

In underwater images, the light suffers absorption and scattering by the medium before reaches the camera. It generated an effect called *haze*. Basically, haze becomes a serious issue since it reduces the overall contrast in images and causes color shift, directly impacting on the reduction in the visibility. Besides these limitation, vision stands out as an important alternative in most applications due to its low cost, rich information in short range and high frame rate.

The existing research in underwater visual SLAM has used predominantly stereo cameras; from the early approach of [10] to more recent ones showing large reconstructions, e.g. [11], [12], [13]. In most of the cases, visual sensors are fused with inertial measurements, Doppler velocities or depth pressure sensors [14], [15]. Such requirement limits the applicability of the algorithms, as all of these sensors are only available in large and expensive vehicles.

Another key limitation of these traditional methods is that they use *feature-based* reconstruction methods [16], [17], meaning that they can only reconstruct a sparse set of salient image points. These methods are able to estimate the camera pose very accurately; but the sparseness of the estimated maps make them inappropriate for autonomous robotic navigation.

Dense reconstructions can be built on top of these sparse point clouds via triangulation [13]. The assumption there is that low-gradient areas between salient points are planar; leading to inaccurate results if the density of salient points in the image is low. The recent work [18] uses a region-growing algorithm to expand a feature-based reconstruction into a more dense one. Typically both methods are computationally expensive, making them unsuitable for online robot navigation.

Our main contribution is the use of *direct* monocular SLAM methods [19], [20] that achieve real-time and dense –one point per image pixel– 3D reconstructions from the only input of a monocular sequence. Notice that our proposal overcomes the two limitations mentioned in the above paragraphs. We use a minimal low-cost sensor configuration of one camera, suitable for small vehicles. And we achieve dense, one-point-per-pixel 3D reconstructions without relying on any extra assumptions. This technique opens new opportunities to exploration of the benthic areas using cheap and small vehicles.

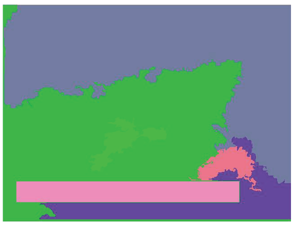
The rest of the paper is organized as follows. Section II describes a classification algorithm that rejects hazy image regions. Section III describes the direct SLAM algorithm. Section IV shows the experimental results and section V gives the conclusions.

II. Haze Classification

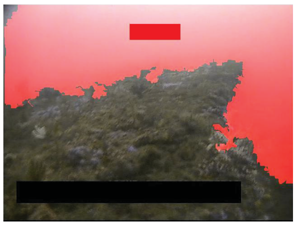
As we use a forward-looking camera that might be imaging scenes at a large depth, part of the image might be hazy and useless for a reconstruction algorithm. We use a SVM-classification scheme in order to identify such hazy areas.

First, we segment the image $\mathbf{I}$ into a set of superpixels $\Omega = \{\mathcal{S}_1, \ldots, \mathcal{S}_i, \ldots\}$. Superpixels are image regions of homogeneous color. In this work, we use the superpixel segmentation proposed in [21]. See figures 1(a) and 1(b) for an example of an underwater image and its segmentation into superpixels.

(a) Underwater image



(b) Superpixels



(c) Haze removed

Fig. 1: Haze classification example. (a) shows an underwater image taken by a robot with a frontal camera. Notice the haze effect in the upper part of the image. (b) shows the superpixel segmentation. Notice how the segmentation captures the main areas of the image. (c) shows the results of the SVM-based haze classification. Notice how the upper superpixel, containing the haze, has been correctly classified and will not be used for the reconstruction. Figure best seen in color.

For every superpixel $\mathcal{S}_i$ we extract a feature vector composed of six parameters. The first three parameters refer to the color of the superpixel, and are specifically the median of each color channel in the superpixel minus the median of the color channel in the entire image. The fourth and fifth parameters are the image coordinates of the superpixel centroid. The sixth parameter is the superpixel size. We normalized all the features.

As the two classes, haze and non-haze, are clearly separable, we obtained perfect classification results in all of our images. In our example, notice in figure 1(c) how the hazy area has been correctly classified. Image pixels classified as haze will not be used for localization and mapping.

III. Direct Localization and Mapping

A. Overview

Direct mapping uses the photometric intensity $\mathbf{I}(\mathbf{u})$ of a pixel $\mathbf{u}$ in a frame $\mathbf{I}$ to estimate its inverse depth ρ –better conditioned that the depth. The dense reconstruction is composed of one point $\mathbf{X}(\mathbf{u}, \rho, \mathbf{T})$ per pixel, that can be extracted from its image coordinates $\mathbf{u}$, inverse depth ρ and the camera pose transformation $\mathbf{T}$ for the image $\mathbf{I}$. Camera pose transformations $\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ are composed of a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. As the frames are highly overlapping, only a subset of reference keyframes $\mathbf{I_r}$ summarizing the sequence are used for mapping. The camera internal calibration is represented with the matrix $\mathbf{K}$. In order to reduce the effects of the refraction, the camera was calibrated in the underwater medium (*i.e.* the images of the classic chessboard are acquired in the underwater environment).

Our algorithm is divided into three separate threads. Although our aim is to produce a dense reconstruction $\mathcal{M} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$, the first two threads operate with a semidense reconstruction $\mathcal{M}^* \in \mathcal{M}$ of high-gradient image pixels $\mathbf{u}^* \in \mathbf{u}$. Low-gradient image pixels produce noisy reconstructions that are regularized under the assumption that real surfaces are smooth. Regularizing is the most demanding process in our SLAM pipeline and might compromise the real-time constraints of the camera pose tracking thread.

Our approach overcomes this problem estimating first a semidense map $\mathcal{M}^*$ of high-gradient pixels $\mathbf{u}^*$ from a set of keyframes $\mathbf{I_r}$ of the image sequence (see the details on section III-C). This semidense map is very accurate without the regularization step and hence will need a low computational cost. The camera pose $\mathbf{T}_k$ at time k is tracked using this semidense low-delay map (section III-B). The dense reconstruction $\mathcal{M}$ is estimated in a third lower-frequency thread not used by the tracking thread (section III-D).

B. Semidense Tracking

The current camera pose $\mathbf{T}_k$ at time k is estimated by minimizing the photometric reprojection error $\Delta\mathbf{I}_k^*$ using the inverse compositional approach [22]. Such approach estimates an incremental transformation $\Delta\mathbf{T}_k$ with respect to a seed transformation $\mathbf{T}_{k,s}$, which has some computational advantages over a more straightforward optimization:

$$\Delta\hat{\mathbf{T}}_k = \arg\min_{\Delta\mathbf{T}_k} \Delta\mathbf{I}_k^*. \tag{1}$$

The photometric error $\Delta\mathbf{I}_k^*$ is as follows:

$$\Delta\mathbf{I}_k^* = I_r(\pi(\mathcal{M}^*, \mathbf{T}_r\Delta\mathbf{T}_k) - I_k(\pi(\mathcal{M}^*, \mathbf{T}_{k,s})), \tag{2}$$

where $\mathbf{u}^* = \pi(\mathcal{M}^*, \mathbf{T})$ stands for the projection model of a set of high gradient 3D points $\mathcal{M}^*$ in an image whose camera pose is $\mathbf{T}$. $\mathbf{T}_r$ refers to the transformation of the closest reference frame.

The current camera pose is related to the previous estimate as follows $\mathbf{T}_k = \mathbf{T}_{k,s}\Delta\mathbf{T}_k^{-1}$. For the optimization procedure, we adopted a minimal parametrization of the camera pose. The rigid body transformation $\mathbf{T}$ is mapped to the tangent space $\mathfrak{se}(3)$ of the euclidean space $SE(3)$ at the identity. The tangent space is named the *twist coordinates* ϵ and we denote it with $\epsilon = (w, v)^t \in \mathbb{R}^6$. $w \in \mathbb{R}^3$ is the angular velocity and $v \in \mathbb{R}^3$ is the linear velocity. ϵ is mapped into $SE(3)$ by the exponential map $\mathbf{T} = \exp_{\mathfrak{se}(3)}(\epsilon)$ and the inverse is done by the logarithmic map $\epsilon = \log_{SE(3)}(\mathbf{T})$.

C. Semidense Mapping

For each high-gradient pixel $\mathbf{u}^*$ in a reference keyframe $\mathbf{I_r}$, its inverse depth ρ is estimated by minimizing the photometric error in m overlapping views $\{\mathbf{I_1}, \ldots, \mathbf{I_o}, \ldots, \mathbf{I_m}\}$.

$$\hat{\rho} = \arg\min_{\rho} \Delta\mathbf{I}_r^*, \tag{3}$$

$$\Delta \mathbf{I}_r^* = \sum_{o=1, o \neq r}^{m} ||(\mathbf{I}_r(\mathbf{u}^*) - \mathbf{I}_o(\pi(\mathbf{T}_o, \pi^{-1}(\mathbf{u}^*, \mathbf{T}_r, \rho)))||, \quad (4)$$

where $\mathcal{M}^* = \pi^{-1}(\mathbf{u}^*, \mathbf{T}_r, \rho)$ stands for the function that backprojects the pixels $\mathbf{u}^*$ from each reference image $\mathbf{I}_r$ to a semidense map $\mathcal{M}^*$ at inverse depth ρ. This map is again projected on the overlapping images $\mathbf{I}_o$; and the photometric difference of the semidense map between $\mathbf{I}_r$ and $\mathbf{I}_o$ is minimized.

For the first overlapping view we perform an exhaustive search in the epipolar line for every possible inverse depth. In the rest of them, the search space is limited by the inverse depth ρ and an estimated variance σ_ρ from the first view.

Here and in the dense mapping of next section III-D we only optimize the scene depth ρ and assume that we have accurate enough camera poses from the semidense tracking described in section III-B. That makes our approach only valid for local mapping; and a pose graph optimization similarly to [23] would be needed for mapping large areas.

D. Dense Mapping

Finally, our dense mapping algorithm estimate the depth ρ of every pixel $\mathbf{u}$ in every reference image $\mathbf{I}_r$. The error function $\Delta \mathbf{E}$ to minimize is now composed of two terms, $\Delta \mathbf{I}_r$ and $\mathbf{R}$, with a relative weight λ

$$\Delta \mathbf{E} = \lambda \Delta \mathbf{I}_r + \mathbf{R} \quad (5)$$

The first term $\Delta \mathbf{I}_r$ stands for the photometric error between the reference image and a set of highly overlapping views. This term is similar to the one defined in equation 4; but this time for *every* pixel $\mathbf{u}$ instead of high-gradient ones

$$\Delta \mathbf{I}_r = \sum_{o=1, o \neq r}^{m} ||(\mathbf{I}_r(\mathbf{u}) - \mathbf{I}_o(\pi(\mathbf{T}_o, \pi^{-1}(\mathbf{u}, \mathbf{T}_r, \rho)))||, \quad (6)$$

The second term $\mathbf{R}$ is a regularization term that favours 3D reconstructions of low depth variations. This second term is of key importance to produce accurate and dense maps from visual sensors. The underlying assumption is that most of the surfaces in 3D scenes are smooth, which is quite reasonable in a wide array of cases. The depth of homogeneous-color image regions, difficult to estimate based on their photometric information, is then additionally constrained by this assumption and their neighboring highly-textured pixels.

The specific form of this term is

$$\mathbf{R} = g(\mathbf{u}) ||\nabla \rho||_\epsilon \quad (7)$$

where $\nabla \rho$ is the depth gradients and $g(\mathbf{u})$ a weighting term that allows larger depth gradients in regions of higher color gradients. This models the reasonable assumption that depth discontinuities are usually associated with high color gradients.

The global energy can be minimized very efficiently by decoupling its two terms and converting them to their primal-dual formulation via the Legendre-Fenchel transform. The reader is referred to [19], [20] for the specific details of this minimization and more information about direct monocular SLAM methods.

Fig. 2: Experimental platform used for the image sequence acquisition: an underwater vehicle Seabotix LBV300-5 equipped with a color camera.

IV. Experimental Results

We evaluated our system in two monocular sequences acquired from an underwater robot. The camera was positioned forward looking and calibrated underwater. The image sequences were obtained with an underwater vehicle, the Seabotix LBV300-5, equipped with a color camera (Fig. 2). Two sequences are presented here. Firstly, images were acquired in a pool whose floor simulates a rocky marine seafloor. The second sequence of images were acquired in Brazil's Southeast Coast (approximately 17 nmi from the coast) with depths ranging from 12m to 20m. The imaged area is known as *Parcel do Carpinteiro* – a reef which of irregular bottom topography formed by beach rocks, gravel and sand.

Figure 3 summarizes the results of the pool experiment. Figure 3(a) shows an illustrative keyframe of the sequence. Figure 3(b) shows in red the image area that has been classified as hazy. The bottom part shows the estimated depths of the pixels, where darker grey means larger depths. Notice how the closest scene parts, corresponding to the bottom of the image, are correctly estimated at small depths. Finally, figure 3(c) shows the estimated scene map, composed of the point clouds of every keyframe. Notice the accuracy of the reconstruction.

Figure 4 shows the results of the seafloor sequence. Figure 4(a) shows a sample frame of the sequence. Figure 4(b) shows in red the image area that has been classified as hazy. Figure 4(c) shows the estimated scene map from the whole sequence.

Method	Average cost [ms]	Average # points
Semidense mapping	449	24000
Dense mapping	3174	76000

TABLE I: Average computational cost for semidense and dense mapping.

Figure 5 shows the cost of the semidirect tracking process, as a function of the number of tracked points. The mean computational cost for our experiments was 19 milliseconds. Table I shows the cost of semidense and dense mapping process. Notice the high dense mapping cost, the average being higher than 3 seconds. This high cost might produce a tracking

(a) Sample image, pool sequence (b) Estimated depth (c) Dense 3D reconstruction

Fig. 3: Pool sequence results. (a) shows a sample image from the sequence. (b) shows the estimated depths (the darker the further). Red stands for regions classified as haze. (c) shows the reconstructed 3D map of the scene from a slightly different viewpoint of the sample image. Figure best seen in color.

(a) Sample image, seafloor sequence (b) Estimated depth (c) Dense 3D reconstruction

Fig. 4: Seafloor sequence results. (a) shows a sample image of the sequence. (b) shows the estimated depths (the darker the further). Red stands for regions classified as haze. (c) shows the reconstructed 3D map of the scene from a slightly different viewpoint of the sample image. Figure best seen in color.

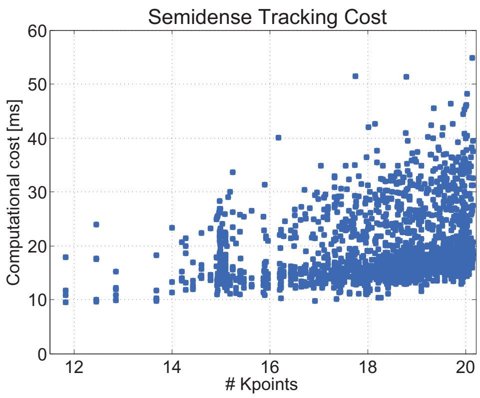


Fig. 5: The computational cost of the semidense tracking is linear with the number of tracked points. In our experiments the mean cost was 19 milliseconds.

failure if the vehicle moves fast; being a faster semidense mapping needed.

We run our experiments in a computer with a a 3.5 GHz Intel Core i7-3770K CPU and 8.0 GB of RAM memory. It is worth remarking that, as reported in [19], [20], dense mapping can be greatly speeded up with GPU programming.

V. Conclusion

In this paper we have presented a SLAM framework for the real-time estimation of the pose of an underwater vehicle/camera and a dense reconstruction of the seafloor. We address the challenging case of a forward looking monocular camera as the only sensor. We believe that this minimal setting is essential for the development of low-cost underwater robots able to navigate autonomously without collision.

Our method starts by classifying the image regions into hazy and non-hazy. Non-hazy image regions are used in a direct mapping framework that estimates in real-time the camera pose and a dense map of the scene. The processing is divided into three threads. A first thread estimates a semidense map of high-gradient pixels from a set of keyframes of the sequence. The goal is to produce a low-cost map to track the camera pose. The tracking thread estimates the pose for each frame assuming a semidense map. Finally, a dense mapping thread estimates a dense map by minimizing a photometric cost –the photometric difference between corresponding points in several views– and a regularization cost that favors smooth solutions. The experimental results show that our pipeline produces accurate and real-time reconstructions from the only input of a monocular sequence. Our future work will be focused in use of the information from hazy regions in order to improve the multiview reconstruction.

Acknowledgment

This research was partially funded by the Spanish government with the projects IPT-2012-1309-430000 and DPI2012-32168. This research is also partly supported by CNPq, CAPES and FAPEMIG. This paper represents a contribution of the INCT-Mar COI funded by CNPq Grant Number 610012/2011-8. We would like to thank our colleagues from NAUTEC-FURG and LOG-FURG for their help with the experimental data acquisition.

References

[1] D. Ribas, P. Ridao, J. D. Tardós, and J. Neira, "Underwater SLAM in man-made structured environments," *Journal of Field Robotics*, vol. 25, no. 11-12, pp. 898–921, 2008.

[2] A. Sedlazeck, K. Koser, and R. Koch, "3D reconstruction based on underwater video from ROV Kiel 6000 considering underwater imaging conditions," in *IEEE/OES OCEANS*, 2009, pp. 1–10.

[3] N. Gracias, P. Ridao, R. Garcia, J. Escartin, M. L'Hour, F. Cibecchini, R. Campos, M. Carreras, D. Ribas, N. Palomeras *et al.*, "Mapping the moon: Using a lightweight AUV to survey the site of the 17th century ship La Lune," in *MTS/IEEE OCEANS*, 2013, pp. 1–8.

[4] Y. Petillot, S. Reed, and J. Bell, "Real time AUV pipeline detection and tracking using side scan sonar and multi-beam echo-sounder," in *MTS/IEEE OCEANS*, 2002, pp. 217–222.

[5] J. Vaganay, M. Elkins, D. Esposito, W. O'Halloran, F. Hover, and M. Kokko, "Ship hull inspection with the HAUV: US Navy and NATO demonstrations results," in *MTS/IEEE OCEANS*, 2006, pp. 1–6.

[6] H. Durrant-Whyte and T. Bailey, "Simultaneous localisation and mapping (SLAM): Part I the essential algorithms," *Robotics and Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[7] L. Paull, S. Saeedi, M. Seto, and H. Li, "AUV navigation and localization: A review," *IEEE Journal of Oceanic Engineering*, vol. 39, no. 1, pp. 131–149, 2014.

[8] F. Guth, L. Silveira, S. Botelho, P. Drews-Jr, and P. Ballester, "Underwater SLAM: Challenges, state of the art, algorithms and a new biologically-inspired approach," in *IEEE RAS & EMBS*, 2014, pp. 981–986.

[9] A. Mallios, P. Ridao, D. Ribas, and E. Hernández, "Scan matching SLAM in underwater environments," *Autonomous Robots*, vol. 36, no. 3, pp. 181–198, 2014.

[10] S. Negahdaripour and H. Madjidi, "Stereovision imaging on submersible platforms for 3-d mapping of benthic habitats and sea-floor structures," *IEEE Journal of Oceanic Engineering*, vol. 28, no. 4, pp. 625–650, 2003.

[11] R. M. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation for autonomous underwater vehicles," *IEEE Journal of Oceanic Engineering*, vol. 33, no. 2, pp. 103–122, 2008.

[12] M. Johnson-Roberson, O. Pizarro, S. B. Williams, and I. Mahon, "Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys," *Journal of Field Robotics*, vol. 27, no. 1, pp. 21–51, 2010.

[13] C. Beall, B. J. Lawrence, V. Ila, and F. Dellaert, "3D reconstruction of underwater structures," in *IEEE/RSJ IROS*, 2010, pp. 4418–4423.

[14] A. Kim and R. Eustice, "Real-time visual slam for autonomous underwater hull inspection using visual saliency," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 719–733, 2013.

[15] S. Pi, B. He, S. Zhang, R. Nian, Y. Shen, and T. Yan, "Stereo visual SLAM system in underwater environment," in *MTS/IEEE OCEANS*, 2014, pp. 1–5.

[16] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.

[17] F. Dellaert and M. Kaess, "Square Root SAM: Simultaneous localization and mapping via square root information smoothing," *The International Journal of Robotics Research*, vol. 25, no. 12, pp. 1181–1203, 2006.

[18] Y. Rzhanov, H. Hu, and T. Boyer, "Dense reconstruction of underwater scenes from monocular sequences of images," in *MTS/IEEE OCEANS*. IEEE, 2014, pp. 1–5.

[19] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *IEEE ICCV*, 2011, pp. 2320–2327.

[20] J. Stühmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *Pattern Recognition*, ser. Lecture Notes in Computer Science, M. Goesele, S. Roth, A. Kuijper, B. Schiele, and K. Schindler, Eds. Springer Berlin Heidelberg, 2010, vol. 6376, pp. 11–20.

[21] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[22] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework: Part 1," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[23] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular slam," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 834–849.