Noname manuscript No.
(will be inserted by the editor)

Selective Visual Odometry for Accurate AUV Localization

Fabio Bellavia · Marco Fanfani · Carlo Colombo

Received: date / Accepted: date

Abstract In this paper we present a stereo Visual Odometry (VO) system developed for autonomous underwater vehicle localization tasks. The main idea is to make use of only highly reliable data in the estimation process, employing a robust keypoint tracking approach and an effective keyframe selection strategy, so that camera movements are estimated with high accuracy even for long paths. Furthermore, in order to limit the drift error, camera pose estimation is referred to the last keyframe, selected by analyzing the feature temporal flow.

The proposed system was tested on the KITTI evaluation framework and on the New Tsukuba stereo dataset to assess its effectiveness on long tracks and different illumination conditions. Results of a live archaeological campaign in the Mediterranean Sea, on an AUV equipped with a stereo camera pair, show that our solution can effectively work in underwater environments.

Keywords Visual Odometry · Stereo · Underwater · AUV · RANSAC · Feature Matching · Keyframe Selection

1 Introduction

The exploration of underwater environments is a relevant topic both for the industry and the academic world. Typically inspection operations in the sea require an expensive setup that could include support ships, high qualified personnel and remotely operated vehicles (ROVs). While this setup could be affordable for industrial targets, academic research campaigns, e.g. for archaeological, biological and ecological purposes, need more user-friendly and economic alternatives.

Autonomous Underwater Vehicles (AUVs) can be easily deployed in order to survey underwater areas in close proximity to the seabed without human supervision. Moreover, they can be equipped with variable payload encompassing acoustic devices and high resolution cameras. AUVs must be able of self-localizing (Paull et al 2014), to reach predefined way-points and to navigate on the basis of cues extracted from on-line measurements. This can be achieved using dead-reckoning techniques based on inertial measurements or employing fixed acoustic transponder deployed into the sea before mission (Long BaseLine system, LBL) (Whitcomb et al 1999). More recently, dynamic transponders (Ultra-Short BaseLine system, USBL) have been proposed, which require a support ship and are sensitive to noise.

An alternative approach employs Simultaneous Localization and Mapping (SLAM) techniques (Durrant-Whyte and T.Bailey 2006), globally exploiting the optical (Kim and Eustice 2013) or acoustic (Mallios et al 2014) input data. Furthermore, Visual Odometry (VO) approaches aim at estimating the vehicle trajectory incrementally with only local information provided by the current acquired images, using for instance the Kanade Lucas Tomasi (KLT) tracker (Shi and Tomasi 1993) as done in (Badino et al 2013).

1.1 Related Work

VO systems are mainly defined in two subsequent, continuously iterated steps: i) feature point extraction from

F. Bellavia · M. Fanfani · C. Colombo
Computational Vision Group (CVG)
University of Florence, Via S. Marta, 3, Florence, Italy
E-mail: {name}.{surname}@unifi.it
Tel: +39 055 275 8509 - Fax: +39 055 275 8570

images and computation of correspondences; ii) local motion estimation by registration of independent 3D maps (Horn 1987) or minimization of the reprojection error over a set of 2D/3D matches (Garro et al 2012). In (Scaramuzza and Fraundorfer 2011; Fraundorfer and Scaramuzza 2012) an in deep review on VO is given.

Nistér et al (2004) proposed a VO framework for terrestrial applications that uses monocular and stereo camera setups in order to accurately compute the navigated path. Since VO systems tend to accumulate a drift error as the estimation proceeds, the authors put *firewalls* at fixed times to localize the estimation and avoid error growth.

VO approaches have been presented also for the underwater domain. Using pipelines that include feature tracking and motion estimation, those systems are able to compute planar (Botelho et al 2009) or full six degrees-of-freedom (Corke et al 2007; Wirth et al 2013) incremental transformations of the camera pose. Inertial measurements can also be included to improve performance (Hildebrandt and Kirchner 2010).

Underwater environments are typically characterized by unstructured, noisy and highly textured images, with repetitive patterns and poor local illumination conditions (vignetting effects and other artifacts). Consequently, image keypoints are hard to track and match correctly underwater, even if stable and robust feature detectors/descriptors are employed, such as the Scale Invariant Feature Transform (SIFT) (Lowe 2004), the Speeded-Up Robust Features (SURF) (Bay et al 2008) or feature descriptors based on the Zernike moments (Eustice et al 2008; Kim and Eustice 2009).

The majority of proposed solutions prefer stereo setups instead of monocular setups to improve the robustness and accuracy of the output, avoiding issues such as the delayed 3D feature initialization (Montiel et al 2006), i.e. when a point is seen for the first time, and the scale factor uncertainty (Strasdat et al 2010).

1.2 Our Contribution

In this paper a stereo VO system, named SSLAM, is introduced. The main idea is to make use of only highly reliable data in the estimation process. This is reflected mainly in the feature matching scheme and the selection of good video frames.

The feature matching process is the main source of noise in a VO system, especially in underwater environments, since wrong matches can lead to very erroneous estimates. In order to limit as much as possible the occurrence of errors in early processing stages, we choose to employ an accurate matching strategy based on the detection of high repeatable corner keypoints (Bellavia et al 2011), matched by an accurate SIFT-like descriptor (Bellavia et al 2010). Moreover, a robust loop chain matching scheme is adopted, improving upon VISO2-S (Geiger et al 2011).

The other aspect mainly characterizing our approach is the selection of the keyframes used as base references for estimating the camera trajectory. Since errors propagate from the uncertainty in the 3D map, higher for distant points that correspond to matches with low temporal flow, the proposed approach picks a input frame as keyframe when enough features with strong temporal flow are detected.

Our keyframe choice is similar to the firewall concept of (Nistér et al 2004), but instead of a constant time selection our approach works adaptively according to the input sequence. Moreover, the proposed approach is more robust than other methods which evaluate 3D flow (Geiger et al 2011) or use predefined thresholds over the average flow disparity (Lee et al 2011).

This paper extends our previous work (Bellavia et al 2013), by providing a detailed description of the proposed method in Sect. 2, followed in Sects. 3-4 by a preliminary evaluation on the KITTI framework (Geiger et al 2012), commonly the main reference for VO comparison, and the New Tsukuba dataset (Martull et al 2012), relevant for illumination issues. Finally, in Sect. 5 we present results of an open sea test with cameras on board of an AUV. Conclusions and final remarks are given in Sect. 6.

2 System Overview

We assume to work with intrinsic and extrinsic calibrated stereo input data: A stereo frame $f_t = (I_t^l, I_t^r)$, composed by the left I_t^l and I_t^r right images at time t, is rectified exploiting the relative positions of the left and right cameras (i.e. the extrinsic calibration), and the radial distortion is corrected. The intrinsics camera parameters (focal length, screw, aspect ratio and principal point) are assumed known for all the frames of the input sequence.

Our method alternates between two main steps: i) computation of feature correspondences between the last detected keyframe f_i and the current frame f_j; ii) estimation of the relative six degrees-of-freedom motion from f_i to f_j as $\mathrm{P}_{i,j} = [\mathrm{R}_{i,j}|\mathbf{t}_{i,j}] \in \mathbb{R}^{3\times4}$, where $\mathrm{R}_{i,j} \in \mathbb{R}^{3\times3}$ is the rotation matrix and $\mathbf{t}_{i,j} \in \mathbb{R}^3$ is the translation vector, see Fig. 1.

The absolute position P_n for a general camera frame f_n w.r.t. the first frame of the sequence is obtained by concatenating the incremental transformations $\mathrm{P}_{0,0}$,

$\mathrm{P}_{0,k}, \ldots, \mathrm{P}_{i,j}, \mathrm{P}_{j,n}$, where $\mathrm{P}_{0,0}$ is composed by the identity rotation matrix and the null translation vector and f_0, f_k, f_i, f_j are selected keyframes.

After a new pose is successfully estimated, the current keyframe is updated accordingly to temporal flow.

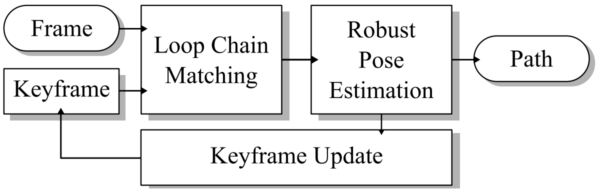


Fig. 1: The SSLAM system pipeline

2.1 Loop Chain Matching

Given the last detected keyframe $f_i = (I_i^l, I_i^r)$ and a new frame $f_j = (I_j^l, I_j^r)$, stable corner features are extracted using the HarrisZ detector (Bellavia et al 2011). Correspondences in the four image pair (I_i^l, I_i^r), (I_i^l, I_j^l), (I_i^r, I_j^r) and (I_j^l, I_j^r) are found exploiting the sGLOH descriptor with the sCOr Nearest Neighbour matching (Bellavia et al 2014) on the L_1 distance. In order to reduce computation times and to improve the matching accuracy we introduce both an epipolar *spatial* constraint exploiting the stereo calibration and a *temporal* flow limit under the hypothesis that image points are subject to a limited motion between subsequent frames, see Fig. 2.

In more detail, let $\mathbf{x}_s^d = [x_s^d, y_s^d]^\mathrm{T} \in \mathbb{R}^2$, $d \in \{l, r\}$, $s \in \{i, j\}$ be a point in the image I_s^d. Then, a spatial correspondence is searched for into a rectangular window of size $2\delta_x \times 2\delta_y$, with $\delta_x >> \delta_y$ since the stereo pair is rectified, centered on the old feature position, i.e.

$$|x_s^l - x_s^r| < \delta_x \tag{1}$$

$$|y_s^l - y_s^r| < \delta_y \tag{2}$$

while for temporal matches we use a circular search region of radius δ_r, as

$$\| \mathbf{x}_i^d - \mathbf{x}_j^d \| < \delta_r \tag{3}$$

Then RANSAC (Fischler and Bolles 1981) iterations are executed to eliminate wrong correspondences. Finally, only matches that are consistent among the four image pairs, $\{(\mathbf{x}_i^l, \mathbf{x}_i^r), (\mathbf{x}_i^l, \mathbf{x}_j^l), (\mathbf{x}_j^l, \mathbf{x}_j^r), (\mathbf{x}_i^r, \mathbf{x}_j^r)\}$, are retained and collected into the set $C_{i,j}$.

The proposed loop chain matching draws inspiration from the *circle match* of VISO2-S (Geiger et al 2011). However we choose to employ a robust detector and descriptor pair, avoiding the two-step matching strategy employed by VISO2-S, and achieving longer and more stable keypoint tracks, crucial for the pose estimation, especially for the underwater environment.

In particular, the HarrisZ detector (Bellavia et al 2011), with results comparable to other state-of-the-art detectors, is used to extract robust and stable corner features in the affine scale-space instead of the simpler corner and blob masks used in VISO2-S. To obtain the candidate correspondences, the sGLOH descriptor with the sCOr Nearest Neighbour matching (Bellavia et al 2014) replaces the concatenation of Sobel filter responses employed by VISO2-S.

2.2 Robust Pose Estimation

To estimate the relative pose $\mathrm{P}_{i,j}$ between the last keyframe f_i and the current frame f_j, at first a local 3D map is computed exploiting the stereo matches $(\mathbf{x}_i^l, \mathbf{x}_i^r)$ of f_i. The corresponding 3D point $\mathbf{X}_{i,j}$ can be computed using the iterative linear triangulation method described in (Hartley and Sturm 1997), since both intrinsic and extrinsic parameters are known.

After initializing $\mathrm{P}_{i,j}$ with the last estimated transformation, the system projects $\mathbf{X}_{i,j}$ onto f_j: The obtained projections $\widetilde{\mathbf{x}}_j^l$ and $\widetilde{\mathbf{x}}_j^r$, and the previously matched keypoints $\mathbf{x}_j^l$ and $\mathbf{x}_j^r$, are used to compute on both the right and the left image the reprojection errors that have to be minimized in order to obtain an accurate localization for f_j, see again Fig. 2. More formally, we want to minimize

$$\mathcal{D}(\mathrm{P}_{i,j}) = \sum_{C_{i,j}} \| \widetilde{\mathbf{x}}_j^l - \mathbf{x}_j^l \| + \sum_{C_{i,j}} \| \widetilde{\mathbf{x}}_j^r - \mathbf{x}_j^r \| \tag{4}$$

Since outliers could be still present among the matches in $C_{i,j}$, this minimization is wrapped in a RANSAC framework. At every RANSAC iteration, three candidate loop matches are randomly extracted and used to compute a candidate transformation $\hat{\mathrm{P}}_{i,j}$. Then using the whole matching set, inlier correspondences for $\hat{\mathrm{P}}_{i,j}$ are found, bounding the reprojection error to a threshold δ_t. RANSAC ends when the maximum inlier set $C_{i,j}^* \subseteq C_{i,j}$ is found. A final refinement is carried out using all the loop matches in $C_{i,j}^*$.

If the estimation fails due to wrong matches or high noisy data, which practically leads to a final small RANSAC consensus set $C_{i,j}^*$, the frame f_j is discarded and the next frame f_{j+1} is tested.

Finally, we add a pose smoothing constraint between frames, so that the current relative pose estimation $\mathrm{P}_{i,j}$ cannot abruptly vary from the previous $\mathrm{P}_{z,i}$,

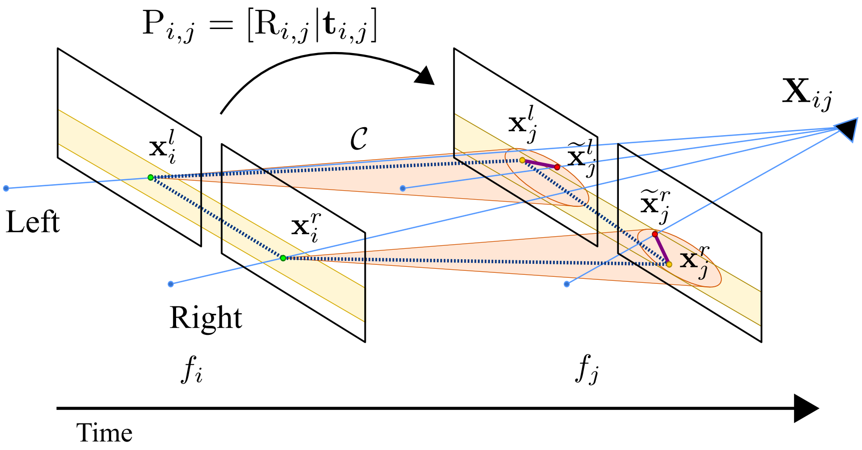


Fig. 2: (Best viewed in color) Keypoint matches between the keyframe f_i and the new frame f_j must satisfy the spatial constraint imposed by the epipolar rectification (yellow band) as well as the temporal flow motion restriction (orange cone). Furthermore, the four matching points must form a loop chain $\mathcal{C}$ (dotted line). In the ideal case, points $\mathbf{x}_j^l$, $\mathbf{x}_j^r$ in frame f_j must coincide with the projections $\widetilde{\mathbf{x}}_j^l$, $\widetilde{\mathbf{x}}_j^r$ of $\mathbf{X}_{i,j}$ obtained by triangulation of $\mathbf{x}_i^l$, $\mathbf{x}_i^r$ in f_i in order for the chain $\mathcal{C}$ to be consistent with the pose $\mathrm{P}_{i,j}$. However, due to data noise, in the real case it is required that the distances $\| \widetilde{\mathbf{x}}_j^l - \mathbf{x}_j^l \|$ and $\| \widetilde{\mathbf{x}}_j^r - \mathbf{x}_j^r \|$ are minimal

$z < i < j$. This is achieved by imposing that the relative rotation around the origin between the two incremental rotations $\mathrm{R}_{z,i}$ and $\mathrm{R}_{i,j}$ is bounded

$$|\mathbf{r}_{i,j}^{k}{}^{\mathrm{T}} \mathbf{r}_{z,i}^{k}| < \delta_{\theta_1} \tag{5}$$

where $\mathbf{r}_{a,b}^k$ is any k-th column of the rotation matrix $\mathrm{R}_{a,b}$. In the case of highly constrained motions, like those of a car, a further criterion can be added to let the estimate of the direction between two incremental translations $\mathbf{t}_{z,i}$ and $\mathbf{t}_{i,j}$ to change smoothly over time:

$$\frac{|\mathbf{t}_{i,j}{}^{\mathrm{T}} \mathbf{t}_{z,i}|}{\| \mathbf{t}_{i,j} \| \| \mathbf{t}_{z,i} \|} < \delta_{\theta_2} \tag{6}$$

This last constraint can also resolve issues in the case of no camera movement or when moving objects crossing the camera path cover the scene.

2.3 Keyframe Selection

Keyframes are selected according to the observation that 3D points related to low temporal flow disparity matches have a higher uncertainty when compared to 3D points with greater temporal disparities. As shown in Fig. 3, high temporal disparities can be found in distant frames. Moreover, only points with sufficient displacement can give information about both the translational and rotational motions. This idea is a straight generalization of the well-known baseline length issues related to the trade-off between reliable correspondence matching and accurate point triangulation (Hartley and Zisserman 2004).

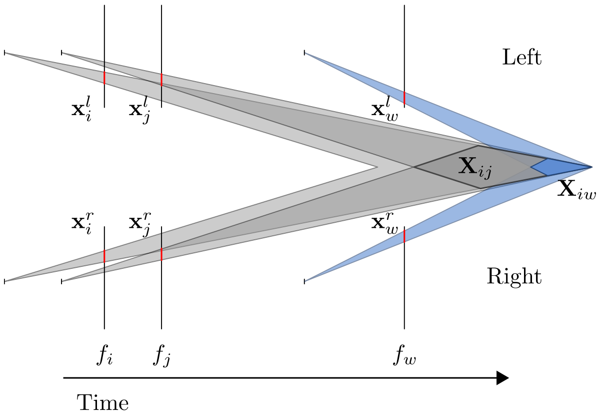


Fig. 3: (Best viewed in color) The uncertainty of matches in the image planes is lower bounded by the image resolution (red) and it is propagated to the 3D points. In order to estimate the 3D point $\mathbf{X}_{i,j}$, by using close frames f_i and f_j, a low temporal disparity flow is present in the image planes, and the 3D point location $\mathbf{X}_{i,j}$ can assume a higher range of values (dark gray quadrilateral). In the case of distant frames f_i and f_w, the possible locations of $\mathbf{X}_{i,w}$ are more circumscribed (blue quadrilateral), for the same resolution limits

Under this observation, two disjoint subsets $F_{i,j}$ and $\bar{F}_{i,j}$ are defined over the set of chain matches $C_{i,j}$ such that $C_{i,j} = F_{i,j} \cup \bar{F}_{i,j}$. The sets $F_{i,j}$ and $\bar{F}_{i,j}$ respectively include *fixed* and *non-fixed* points, i.e. matching points with low and high temporal disparities according to a given threshold δ_f

$$F_{i,j} = \{\mathcal{C} \in C_{i,j} | T_d(\| \mathbf{x}_i^d - \mathbf{x}_j^d \| \leq \delta_f)\} \text{ and,} \tag{7}$$

$$\bar{F}_{i,j} = C_{i,j} \setminus F_{i,j} \ , \tag{8}$$

where $d \in \{l, r\}$ and $T_d(*)$ is an indicator function that outputs 1 if its predicate is true for all admissible values of d or 0 otherwise. A high number of non-fixed points implies a high temporal flow, so that frame f_j is accepted as new keyframe if the number of non-fixed points between frames f_i and f_j is greater than a defined threshold δ_m

$$1 - \frac{|F_{i,j}|}{|C_{i,j}|} > \delta_m \quad . \tag{9}$$

The fixed and non-fixed point subsets are only used in the keyframe selection criterion. Indeed, no improvements were experimented using only non-fixed matches as input for RANSAC.

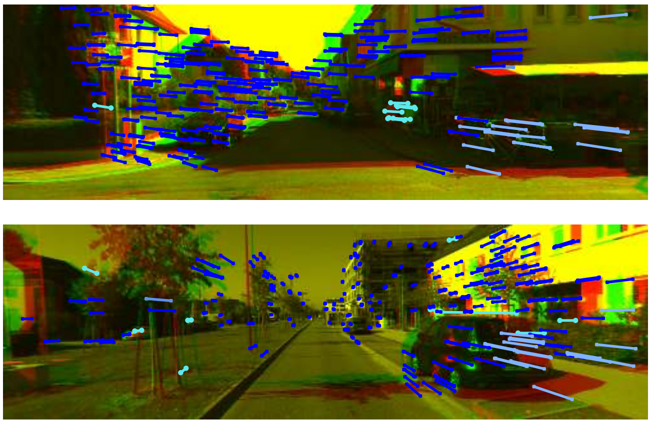


Fig. 4: (Best viewed in color) Examples of the temporal flow for successive keyframes of two different sequences of the KITTI dataset. Involved keyframes are superimposed as for anaglyphs, only images for the left cameras are shown. Good fixed and non-fixed point matches are shown in dark and light blue, respectively, while wrong correspondences are reported in cyan

Examples of fixed and non-fixed points estimates are shown in Fig. 4 (dark and light blue lines, respectively). Our strategy is adaptive and thus it can better handle keyframe drops with respect to the average flow threshold commonly employed by other systems such as (Lee et al 2011). As an example, referring to Fig. 4, the average flow in the top configuration is considerably higher than that of the bottom one, which would be discarded. Lowering the average flow threshold, so as to also accept the bottom frame, would also include very low disparity frames (just consider to replace in the bottom frame the non-fixed light blue matches by twice the matches with half disparity). This does not hold for the proposed frame selection which resembles RANSAC, since it minimizes the number of matches below a disparity threshold. On the other hand, the selection based on the average flow is close to the less robust least-square approach, which just minimizes the average error.

3 Evaluation setup

3.1 General Overview

The KITTI vision benchmark suite (Geiger et al 2012) and the New Tsukuba stereo dataset (Martull et al 2012) were used to obtain a *dry* evaluation of our system, given that to the best of our knowledge no public underwater stereo datasets are available.

It's worth noting that even if these tests are conducted on terrain images, they can give an insight into the general performance of SSLAM, also for the underwater environment. Indeed, while tests on the KITTI dataset, composed by long trajectories, can provide results to assess the robustness of our method w.r.t. drift errors, the New Tsukuba sequence can provide evidence about its reliability in challenging illumination scenarios.

Unless otherwise specified, the parameter triplet $(\delta_r, \delta_x, \delta_y)$ for the spatial and temporal constraints (see Sect. 2.1) is set to $(500, 300, 12)$ px in the case of the KITTI dataset and to $(100, 100, 12)$ px for the New Tsukuba dataset, since images are taken at lower resolution and shorter baseline. For pose estimation (see Sect. 2.2) we set $\delta_{\theta_1} = 15°$ while the additional translation constraint $\delta_{\theta_2} = 10°$ is only employed for the KITTI dataset. About the flow constraints (see Sect. 2.3), we set $\delta_f = 55$ px and $\delta_m = 5\%$.

We tested SSLAM using keypoints detected at full and half resolution videos. In the latter case, the notation SSLAM† is used. For SSLAM†, keypoints localization is less accurate and bigger (normalized) keypoint patches are found, which are more sensitive to fast camera movements. Note also that more keypoints are found in full resolution SSLAM implementation than with SSLAM†. Nevertheless, different image resolutions do not affect the other parameters of the methods since keypoint positions are rescaled at the full resolution before the constrained matching in both cases.

3.2 The KITTI Dataset

Recently, the KITTI dataset has become a reference benchmark for VO systems. The dataset provides sequences recorded from car driving sessions on highways, rural areas and inside cities up to 80 km/h. The benchmark consists of 22 rectified stereo sequences from about 500 m to 5 km, taken at 10 fps with a resolution of 1241×376 pixels. Ground truth trajectories are available to users only for the first 11 sequences to train the parameters of the methods, while results should be submitted to the authors page for the remaining sequences to get a final ranking. Translation and rotation errors normalized with respect to the path lengths and speeds are computed in order to rank the methods.

3.3 The New Tsukuba Dataset

The New Tsukuba dataset is a virtual sequence that navigates into a laboratory reconstructed manually by computer graphics. Images with a resolution of 640×480 pixels are recorded at 30 fps for one minute while accurate ground truth positions are registered and provided to the users. The sequence is rendered with four different illuminations from the more classical *fluorescent* or *daylight* to the more challenging *flashlight* and *lamps*, see Fig. 5.

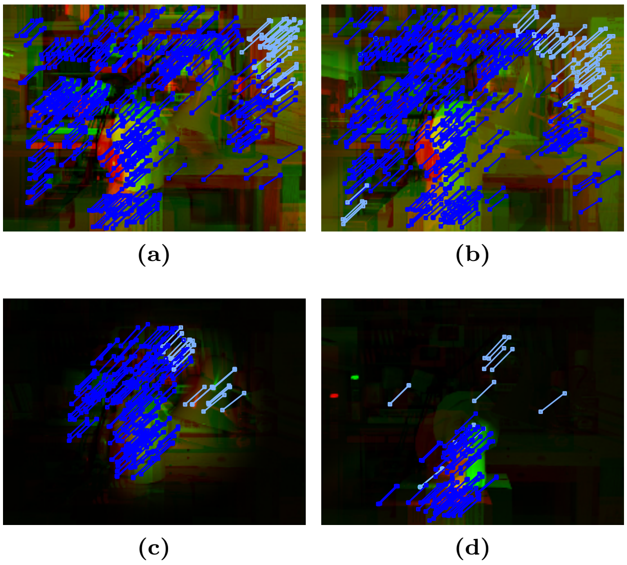


Fig. 5: Example keyframes of the New Tsukuba stereo dataset: (a) *fluorescent*, (b) *daylight*, (c) *flashlight*, (d) *lamps*. Keyframes are superimposed as for anaglyphs, only images for the left cameras are shown. Good fixed and non-fixed point matches are shown in dark and light blue, respectively, while wrong correspondences are reported in cyan

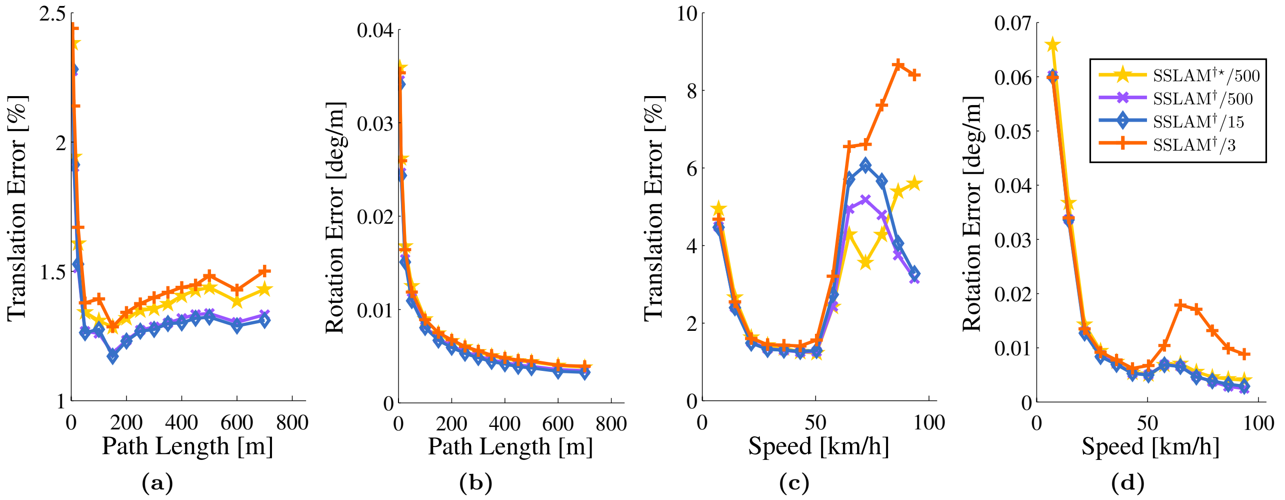
4 Evaluation Results

4.1 The KITTI Dataset

In order to show the benefits provided by our keyframe selection strategy, our full SSLAM† pipeline is compared against the case when keyframe selection is not employed, indicated by SSLAM†*. Furthermore, we present the results obtained with different numbers of RANSAC iterations to underline the stability of the method. In particular, results of SSLAM† with 500, 15 (set as default) and 3 RANSAC iterations, and SSLAM†* with 500 iterations are presented, indicated respectively by SSLAM†/500, SSLAM†/15, SSLAM†/3 and SSLAM†*/500.

Figure 6 shows the relative average translation and rotation errors of the different SSLAM† variants for the first 11 sequences of the dataset, according to the KITTI error metric (Geiger et al 2012) for increasing path length and speed. We verified that similar results hold in the case of full resolution SSLAM. The chain loop matching scheme together with the chosen keypoint detector and descriptor can track long paths, without bundle adjustment or loop closure detection. SSLAM with keyframe selection exhibits less errors, confirming that the proposed keyframe selection strategy is effective. Moreover, results for SSLAM†/15 and SSLAM†/500 are equivalent, while SSLAM†/3 obtains inferior results but close to those obtained by SSLAM†*/500, underlining that our matching selection is robust, since the number of RANSAC outliers must be low to work with such few iterations.

We now report results on the KITTI odometry benchmark at submission time for only stereo methods that do not rely on laser data, more details are available online (Geiger et al 2012). Figure 7 shows the average translation and rotation errors of the different methods for increasing path length and speed. SSLAM and SSLAM†, ranked among the first positions of the KITTI benchmark, obtaining respectively a mean translation error of 1.57% and 2.15% w.r.t. the sequence length and a rotation error of 0.0042 and 0.0058 deg/m. These rank placements show the robustness of the proposed methodology. Note however that the benchmark provides partial results, since these error metrics cannot take into account all the properties of a VO system. In particular, referring to Fig. 8 where two sample tracks of the KITTI dataset are shown, it can be seen that while both Multi-frame Fea-



(a) (b) (c) (d)

Fig. 6: (Best viewed in color) Average relative errors on the first 11 sequences of the KITTI dataset. Plots (a-b) refer to the average relative translation and rotation error for increasing path length respectively, while plots (c-d) refer to increasing speed

ture Integration (MFI) (Badino et al 2013) and Visual odometry with Bundle adjustment (VoBa) (respectively ranked in 1^{st} and 4^{th} positions) provide slightly better results than SSLAM in term of KITTI metrics on long paths, SSLAM† (11^{th} ranked) clearly improves on the 7^{th} ranked efficient Visual Odometer (eVO) (Sanfourche et al 2013). This can also be observed in the relative translation error for an increasing path length in Fig. 7(a), where the SSLAM† plot remains stable when compared to the increasing error of eVO.

For the sake of completeness, we further added results with Stereo Structure from Motion (StereoSFM) (Badino and Kanade 2011), not available at submission time, based on the KLT tracker. As can be noted from Fig. 7, its results in terms of KITTI error metric are close to those of SSLAM, but by looking in detail Fig. 8a, SSLAM seems slightly better in tracking the path than StereoSFM.

Table 1 shows the input matches and the inliers found in the RANSAC pose estimation by SSLAM, SSLAM† and VISO2-S (Geiger et al 2011). Note that VISO2-S works similarly to SSLAM and its code is free available. As it can be noted, VISO2-S outputs a comparable number of initial input matches with SSLAM†, but a higher number of outlier (about 50%), being less robust and error prone than SSLAM†. Note also that the temporal and spatial flow constraints of VISO2-S ($\delta_r = 200$, $\delta_x = 200$ and $\delta_y = 3$) are more tight. These constraints would lead theoretically to a higher number of matches, with a lower probability to have an accidentally wrong match w.r.t. SSLAM and SSLAM†, employing relatively wider matching search areas. Yet, in practice, as clear from Table 1, VISO2-S outputs a lower number of inliers than SSLAM†. This is as a further evidence of the better behavior of the proposed methodology with respect to VISO2-S.

4.2 The New Tsukuba Dataset

To further demonstrate the robustness of our method, we tested SSLAM on the New Tsukuba sequence for all the available illuminations. Figure 9 shows the estimated trajectories together with the ground truth, for the *daylight* and *lamps* illuminations, results for *fluorescent* and *flashlight* illumination conditions are similar to the *daylight* and *lamps* illuminations and not reported. All the investigated methods track the sequence well, however results of SSLAM and StereoSFM are better in terms of the KITTI error metrics reported in Fig. 10. Note that these relative errors are higher than those obtained in the KITTI dataset for both methods.

Table 1: Average number of input matches before the RANSAC pose estimation and final inlier ratios

	pts	inl(%)
SSLAM	766	98
SSLAM†	222	96
VISO2-S	245	50

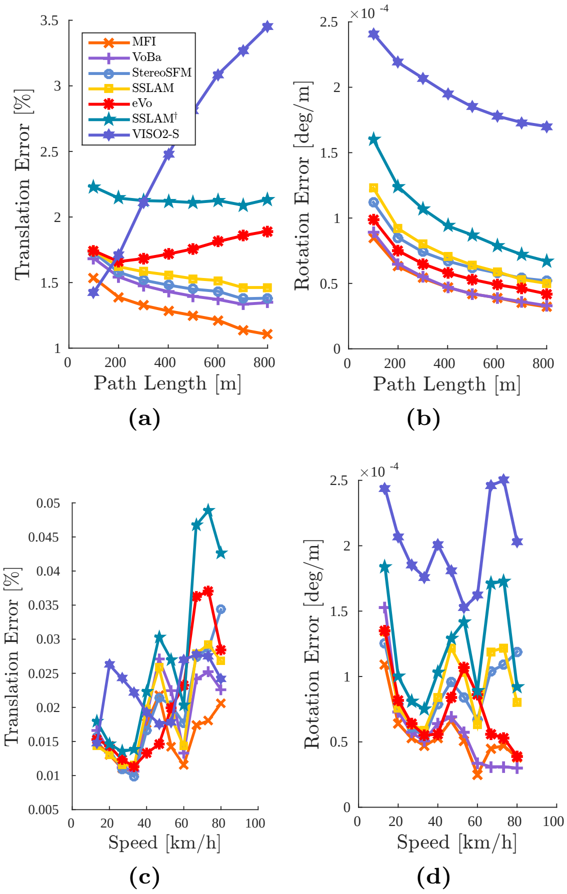


Fig. 7: (Best viewed in color) Average relative error on the KITTI benchmark. Plots (a-b) refer to the average relative translation and rotation error for increasing path length respectively, while plots (c-d) refer to increasing speed

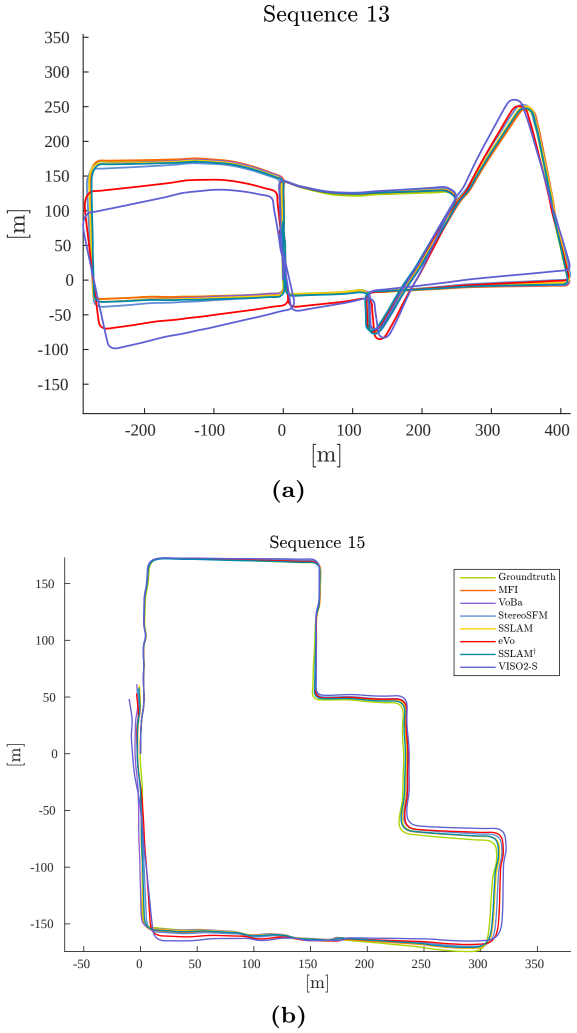


Fig. 8: (Best viewed in color) Trajectories on the sequences 13 (a) and 15 (b) of the KITTI dataset

Table 2: Average computational time for a single frame on a Intel-i7 3.50GHz CPU, 8 cores are used

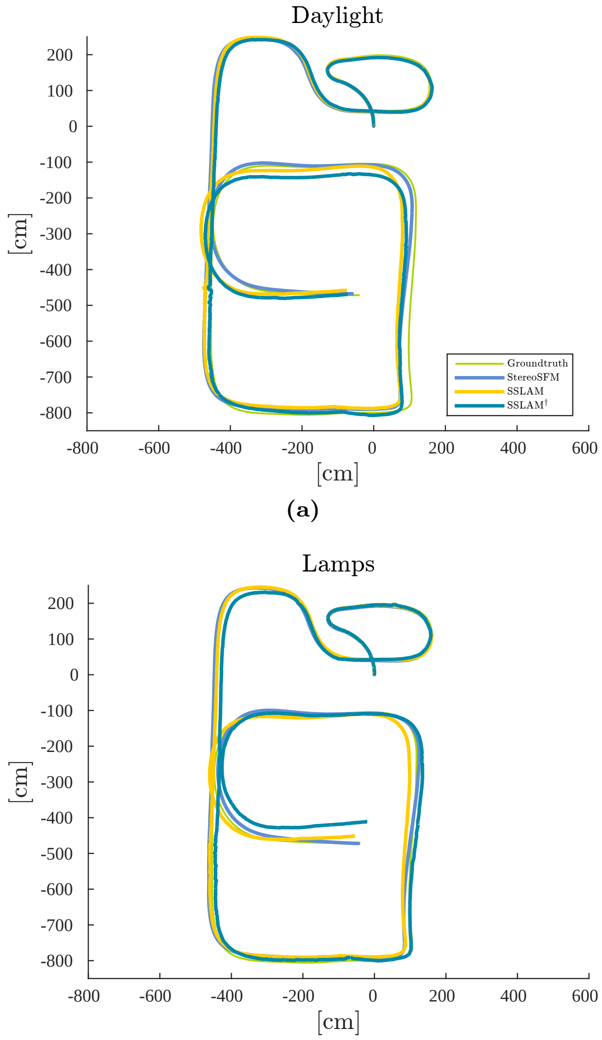
	SSLAM	SSLAM†
KITTI	3.85 s	0.55 s
New Tsukuba	0.87 s	0.24 s

4.3 Running Times

Table 2, reports the average running times of our SSLAM multithreaded C/C++ implementation, freely available[1], for a single frame computation. As it can be noted, SSLAM scales with the resolution. The feature detector is accurate but slow, since it requires large size kernel convolutions. However by taking into account that only keyframes are required by SSLAM, real-time performance is achieved when the keyframe computational time is less than f_k/f_v, where f_k is the keyframe rate and f_v is the frame rate of the video sequences. Table 3 shows the average number of frames between two consecutive keyframes and the corresponding standard deviations. The values are slightly higher for the New Tsukuba dataset with respect to KITTI, according to the different camera speeds. Note also that the keyframe distribution is not uniform but it is denser near camera turns and accelerations, see Fig. 11.

In conclusion the ratio f_k/f_v is equal to 0.20 s and 0.17 s respectively for the KITTI and New Tsukuba sequences. Although only SSLAM† can run almost in real-time on these datasets, real-time requirements could be fulfilled if the robot could be adapted to make slow turns and accelerations.

[1] `cvg.dsi.unifi.it`



(a)

(b)

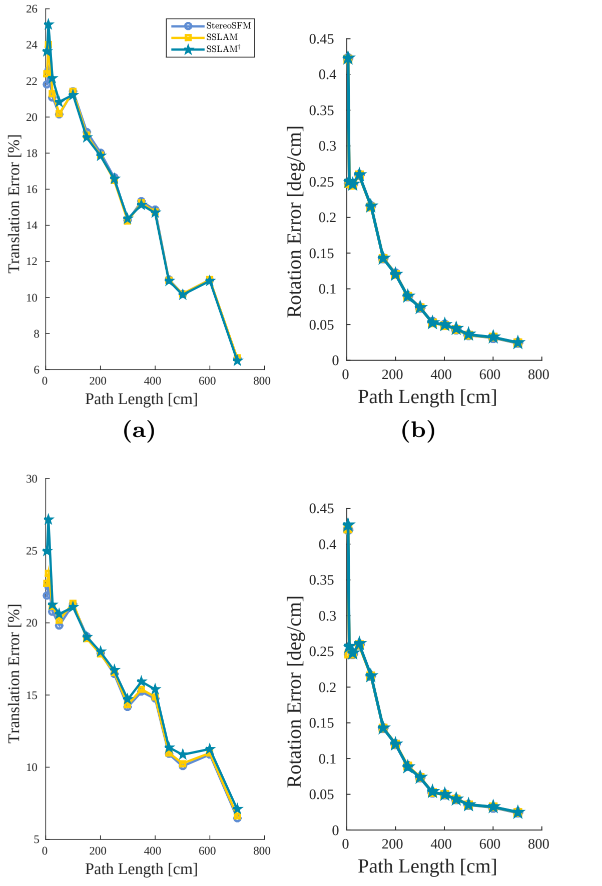
Fig. 9: (Best viewed in color) Estimated trajectories on the New Tsukuba sequence for the *daylight* (a) and *lamps*(b) illuminations, respectively

Table 3: Framerate (fps) and average (Avg) number of frames between two consecutive keyframes with the corresponding standard deviations (std) for the KITTI and New Tsukuba sequences

	fps	Avg	Std
KITTI	10	2	1
New Tsukuba	30	5	3

5 Test in the Sea

Underwater data were gathered during a test mission in the Israeli Sea using the MARTA AUV (Allotta et al 2014) developed within the ARROWS project (Allotta et al 2013). Tests were conducted during daylight at a maximum depth of 40 meters. Images of the sea bottom were acquired with a pair of Basler ACE 2040 cameras with GiGE connection, installed in waterproof housings on the side of the vehicle and pointing toward the sea floor at an angle of around 15 degrees w.r.t. the AUV vertical axis. No artificial illumination was needed; Sunlight passing through the water produces noticeable flickering effects (see Fig. 12) that increase the difficulty of finding correct feature matches.



(a) (b)

(c) (d)

Fig. 10: (Best viewed in color) Average translation (a)-(c) and rotation (b)-(d) error for increasing path length for the New Tsukuba sequence with the *daylight* and *lamps* illuminations, respectively

The camera stereo pair was calibrated off-line directly in water to implicitly account for the water refraction index. The images acquired are rectified on-line and radial distortion is removed at the same time.

Figure 13 shows the computed trajectory and sparse 3D map. Moreover, with minor effort, the 3D point cloud can be upgraded to a meshed and textured model of the seabed, see Fig. 14. Although no ground-truth is available, the path and the 3D environment reconstruction are visually consistent with the input data.

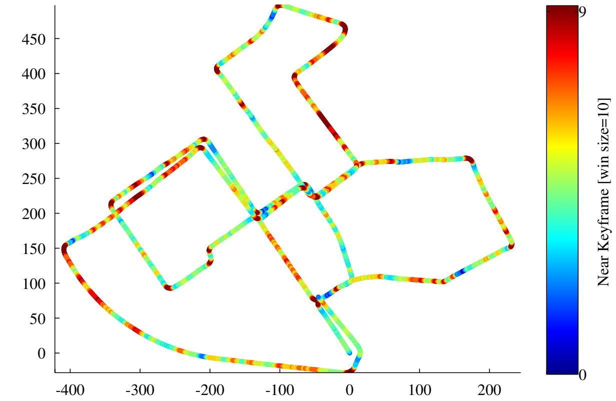


Fig. 11: (Best viewed in color) An example of keyframe distribution along the Sequence 00 of the KITTI dataset for SSLAM. At each estimated camera position the number of keyframes that fall inside a window of 10 frames centred at the camera location is shown according to the colorbar gradation

6 Conclusion

In this paper a practical stereo VO system was presented. The approach achieves a low drift error even for long paths and exploits only local information.

A robust loop chain matching scheme for tracking keypoints is provided, supported by a frame discarding system to improve pose estimation. According to the experimental results, the proposed keyframe selection strategy is effective to reduce the error of the estimated poses. Results from the KITTI and New Tsukuba datasets show the effectiveness of the system, which is robust even with an extremely small number of RANSAC iterations and able to properly estimate long path and to work under different illuminations. Underwater tests with an AUV equipped with stereo cameras, concretely demonstrate the goodness of the method.

7 Acknowledgments

This work has been supported by the European ARROWS project, founded by the European Union's Seventh Framework Programme for Research technological development and demonstration, under grant agreement no. 308724.

References

Allotta B, Colombo C, et al (2013) Teams of Robots for Underwater Archaeology: the ARROWS project. In: Proc. of the 6th International Congress "Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin"

Allotta B, Bartolini F, Conti R, Costanzi R, Gelli J, Monni N, Natalini M, Pugi L, Ridolfi A (2014) MARTA: an AUV for underwater cultural heritage. In: Proc. of the Underwater Acoustics 2014

Badino H, Kanade T (2011) A head-wearable short-baseline stereo system for the simultaneous estimation of structure and motion. In: IAPR Conference on Machine Vision Application, pp 185–189

Badino H, Yamamoto A, Kanade T (2013) Visual odometry by multi-frame feature integration. In: Proc of the International Workshop on Computer Vision for Autonomous Driving at ICCV

Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-Up Robust Features (SURF). Comput Vis Image Underst 110(3):346–359

Bellavia F, Tegolo D, Trucco E (2010) Improving SIFT-based descriptors stability to rotations. In: Proc. of International Conference on Pattern Recognition

Bellavia F, Tegolo D, Valenti C (2011) Improving Harris corner selection strategy. IET Computer Vision 5(2)

Bellavia F, Fanfani M, Pazzaglia F, Colombo C (2013) Robust selective stereo SLAM without loop closure and bundle adjustment. In: Proc. of 17th International Conference on Image Analysis and Processing, pp 462–471

Bellavia F, Tegolo D, Valenti C (2014) Keypoint descriptor matching with context-based orientation estimation. Image and Vision Computing

Botelho SC, Drews P, Oliveira G, da Silva Figueiredo M (2009) Visual odometry and mapping for Underwater Autonomous Vehicles. In: Proc. of the 2009 6th Latin American Robotics Symposium, pp 1–6

Corke P, Detweiler C, Dunbabin M, Hamilton M, Rus D, Vasilescu I (2007) Experiments with Underwater Robot Localization and Tracking. In: Proc. of the 2007 IEEE International Conference on Robotics and Automation, pp 4556–4561

Durrant-Whyte HF, TBailey (2006) Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms. IEEE Robotics and Automation Magazine 13:99–110

Eustice R, Pizarro O, Singh H (2008) Visually Augmented Navigation for Autonomous Underwater Vehicles. Oceanic Engineering, IEEE Journal of 33(2):103–122

Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24(6):381–395

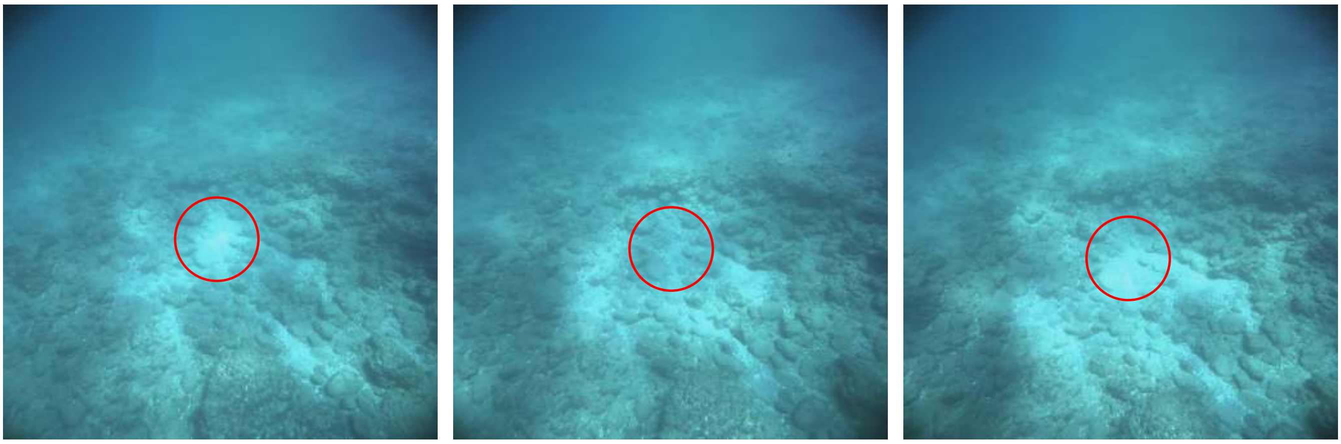
Fraundorfer F, Scaramuzza D (2012) Visual Odometry : Part II: Matching, Robustness, Optimization, and Applications. IEEE Robotics and Automation Magazine 19(2)

Garro V, Crosilla F, Fusiello A (2012) Solving the PnP Problem with Anisotropic Orthogonal Procrustes Analysis. In: Second Joint 3DIM/3DPVT Conference: 3D Imaging, Modeling, Processing, Visualization and Transmission, pp 262–269

Geiger A, Ziegler J, Stiller C (2011) StereoScan: Dense 3D reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium

Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. of Computer Vision and Pattern Recognition, URL `http://www.cvlibs.net/datasets/kitti/eval_odometry.php`

Hartley R, Sturm P (1997) Triangulation. Computer Vision and Image Understanding 68(2):146–157



(a) (b) (c)

Fig. 12: Sampled sequential frames for the underwater sequence. Sunlight passing through water produces flickering artifacts: For example in the central area of images. In (a) a white blob is visible caused by color saturation, while in (b) the same area shows a group of small rocks. Again in (c) the area is saturated

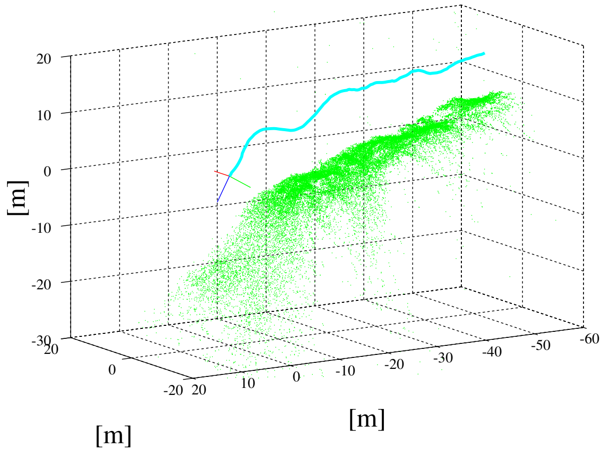


Fig. 13: Estimated trajectory (cyan) and 3D map (green point cloud) of the underwater test

Hartley RI, Zisserman A (2004) Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press

Hildebrandt M, Kirchner F (2010) IMU-aided stereo visual odometry for ground-tracking AUV applications. In: OCEANS 2010 IEEE - Sydney, pp 1–8

Horn BKP (1987) Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America A 4(4):629–642

Kim A, Eustice R (2009) Pose-graph visual SLAM with geometric model selection for autonomous underwater ship hull inspection . In: Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp 1559–1565

Kim A, Eustice RM (2013) Real-time visual SLAM for autonomous underwater hull inspection using visual saliency. IEEE Transactions on Robotics 29(3):719–733

Lee GH, Fraundorfer F, Pollefeys M (2011) RS-SLAM: RANSAC sampling for visual FastSLAM. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp 1655–1660

Lowe D (2004) Distinctive Image Features from Scale-Invariant Keypoints. Int J Comput Vision 60(2):91–110

Mallios A, Ridao P, Ribas D, Hernández E (2014) Scan matching SLAM in underwater environments. Autonomous Robots 36(3):181–198

Martull S, Martorell MP, Fukui K (2012) Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps. In: Proc. ofICPR2012 workshop TrakMark2012, pp 40–42, URL `http://www.cvlab.cs.tsukuba.ac.jp/dataset/tsukubastereo.php`

Montiel J, Civera J, Davison A (2006) Unified inverse depth parametrization for monocular SLAM. In: Proc. of Robotics: Science and Systems, IEEE Press

Nistér D, Naroditsky O, Bergen JR (2004) Visual odometry. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp 652–659

Paull L, Saeedi S, Seto M, Li H (2014) AUV Navigation and Localization: A Review. IEEE Journal of Oceanic Engineering 39(1):131–149

Sanfourche M, Vittori V, Besnerais GL (2013) eVO: A realtime embedded stereo odometry for MAV applications. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 2107–2114

Scaramuzza D, Fraundorfer F (2011) Visual Odometry: Part I - The First 30 Years and Fundamentals. IEEE Robotics and Automation Magazine 18(4)

Shi J, Tomasi C (1993) Good features to track. Tech. rep.

Strasdat H, Montiel JMM, Davison AJ (2010) Scale drift-aware large scale monocular SLAM. In: Proc. of Robotics: Science and Systems

Whitcomb L, Yoerger D, Singh H, Howland J (1999) Advances in Underwater Robot Vehicles for Deep Ocean Exploration: Navigation, Control, and Survey Operations. In: Proc. of the 9th International Symposium on Robotics Research, pp 346–353

Wirth S, Negre Carrasco P, Codina G (2013) Visual odometry for autonomous underwater vehicles. In: Proc. of 2013 MTS/IEEE OCEANS, pp 1–6

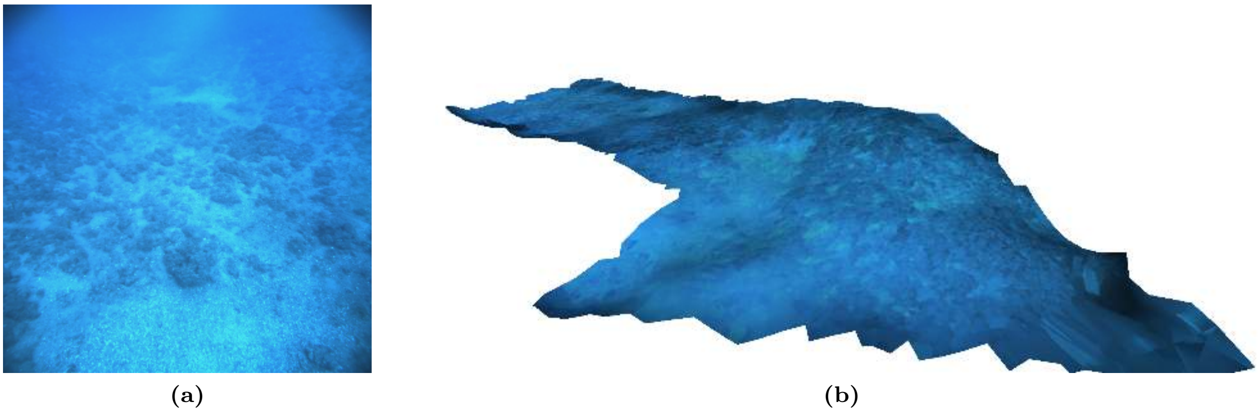


Fig. 14: (a): Frame of the underwater sequence; (b): 3D model of the seabed

View publication stats