The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-2/W8, 2017
5th International Workshop LowCost 3D – Sensors, Algorithms, Applications, 28–29 November 2017, Hamburg, Germany

LOW COST EMBEDDED STEREO SYSTEM FOR UNDERWATER SURVEYS

Mohamad Motasem Nawaf*, Jean-Marc Boï, Djamel Merad, Jean-Philip Royer, Pierre Drap

Aix-Marseille Université, CNRS, ENSAM, Université De Toulon, LSIS UMR 7296,
Domaine Universitaire de Saint-Jérôme, Bâtiment Polytech, Avenue Escadrille Normandie-Niemen, 13397, Marseille, France.
mohamad-motasem.nawaf @univ-amu.fr

Commission II

KEY WORDS: Image processing, underwater imaging, embedded systems, stereo vision, visual odometry, 3D reconstruction.

ABSTRACT:

This paper provides details of both hardware and software conception and realization of a hand-held stereo embedded system for underwater imaging. The designed system can run most image processing techniques smoothly in real-time. The developed functions provide direct visual feedback on the quality of the taken images which helps taking appropriate actions accordingly in terms of movement speed and lighting conditions. The proposed functionalities can be easily customized or upgraded whereas new functions can be easily added thanks to the available supported libraries. Furthermore, by connecting the designed system to a more powerful computer, a real-time visual odometry can run on the captured images to have live navigation and site coverage map. We use a visual odometry method adapted to low computational resources systems and long autonomy. The system is tested in a real context and showed its robustness and promising further perspectives.

1. INTRODUCTION

Mobile systems nowadays undergo a growing need for self localization to accurately determine its absolute/relative position over time. Despite the existence of very efficient technologies that can be used on-ground (indoor/outdoor) and in-air such as Global Positioning System (GPS), optical, radio beacons, etc. However, in the underwater context most of these signals are jammed so that the corresponding techniques cannot be used. On the other side, solutions based on active acoustics such as imaging sonars and Doppler Velocity Logs (DVL) devices remain expensive and require high technical skills for deployment and operation. Moreover, their size specifications prevent their integration within small mobile systems or even being hand held. The research for an alternative is ongoing, notably, the recent advances in embedded systems outcome relatively small, powerful and cheap devices. This opens interesting perspectives to adapt a light visual odometry approach that provides relative path in real-time, this describes our main research direction. The developed solution is integrated within underwater archaeological site survey where it plays an important role to facilitate image acquisition.

In underwater survey tasks, mobile underwater vehicles (or divers) navigate over the target site to capture images. The obtained images are treated in a later phase to obtain various information and also to form a realistic 3D model using photogrammetry techniques (Drap, 2012). In such a situation, the main problem is to totally cover the underwater site before ending the mission. Otherwise, we may obtain incomplete 3D models and the mission cost will raise significantly as further exploitation is needed. However, the absence of an overall view of the site especially under bad lighting conditions makes the scanning operation blind. In practice, this yields to over-scanning the site which is a waste of time and cost. Moreover, the quality of the taken images may go below an acceptable limit. This mainly happens in terms of lightness and sharpness, which is often hard to quantify visually on the fly. In this work, we propose solutions for the aforementioned problems. Most importantly, we propose to guide the survey based on a visual odometry approach that runs on a distributed embedded system in real-time. The output egomotion helps to guide the site scanning task by showing approximate scanned areas. Moreover, an overall subjective lightness and sharpness indicators are computed for each image to help the operator to control the image quality. Overall, we provide a complete hardware and software solution for the problem through the conception and realization of a hand-held stereo embedded system for underwater imaging. See Figures 1 and 3. The system equipped with two high definition cameras can take and store hardware synchronized stereo images while having very long autonomy. In contrary to other commercially available off-the-shelf products where the system role ends with image storage, the designed system is based on distributed and embedded systems with ARM processors and Linux operating system and is capable of running most image processing techniques smoothly in real-time. The available optimized open source libraries such as OpenCV (Itseez, 2015) and OpenCL (Stone et al., 2010) allows an easy extension of the provided functions and fully customize the system to suite different contexts.

In common approaches of visual odometry, a significant part of the overall processing time is spent on feature points detection, description and matching. In the tested baseline algorithm, the aforementioned operations represent $\sim$65% of processing time in case of local/relative bundle adjustment (BA) approach, which occupies in return the majority of the time left. This would overload the available computing resources. Hence, in our proposed method we rely on low level Harris based detection and template matching procedure which speed up significantly the feature matching speed. Further, whereas in traditional stereo matching the search for correspondence is done along the epipolar line within certain fixed range, in our method we proceed first by computing *a priori* rough depth belief based on image lightness and

*Corresponding author

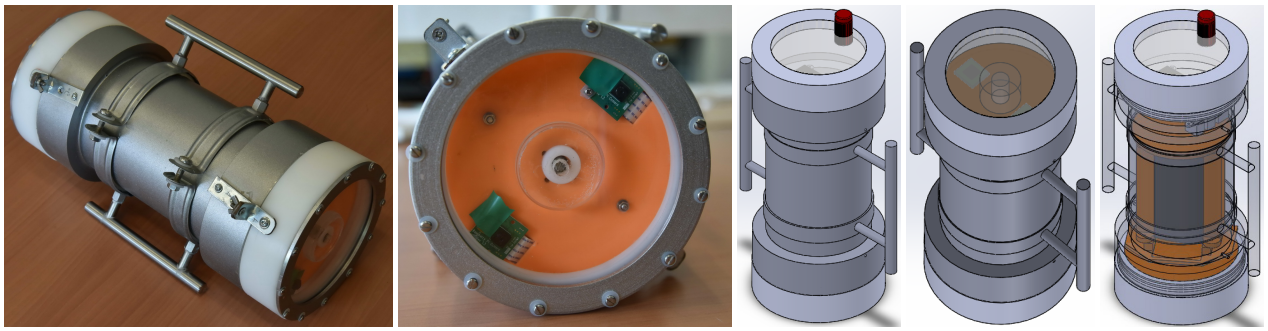

Figure 1. The built system prototype and design sketches.

following the law of light divergence over distance. This is only valid for a configuration where the light source is fixed to the system, which is the case here. Hence, our first contribution is that we benefit from the rough depth estimation to limit points correspondence search zone to reduce processing time.

From another side, traditional visual odometry methods based on local BA suffers from rotation and translation drifts that grows with time (Mouragnon et al., 2009). In contrary, the solutions based on using features from the entire image set, such as global BA (Triggs et al., 2000), require more computational resources which are very limited in our case. Similarly, the simultaneous localization and mapping (SLAM) approaches (Thrun et al., 2005), which are known to perform good loop closure, are computationally intensive especially when complex particle filters are used (Montemerlo and Thrun, 2007), and they can only operate in moderate size environments if real-time processing is needed. In our method, we adopt a semi-global approach (Nawaf et al., 2016), which proceed in the same way as local method in optimizing a subset of image frames. However, it differs in the way of selecting the frames subset, as local methods use Euclidean distance and deterministic pose representation to select frames, our represents the poses in a probabilistic manner, and uses a divergence measure to select such sub set.

The rest of the paper is organized as follows: We survey related works in Section 2. In Section 3 we describe the designed hardware platform that we used to implement our solution. Our proposed visual odometry method is explained in Section 4. The analytical results are verified through simulation experiments presented in Section 5. Finally, we present a summary and conclusions. We note that parts of this work have been presented in (Nawaf et al., 2016) and (Nawaf et al., 2017).

2. RELATED WORKS

2.1 Feature Points Matching

Common ego-motion estimation methods rely on feature points matching between several poses (Nistér et al., 2004). The choice of the used approach for matching feature points depends on the context. For instance, features matching between freely taken images (6 degrees of freedom), has to be invariant to scale and rotation changes. Scale invariant feature descriptors (SIFT) (Lowe, 2004) and the Speeded Up Robust Features (SURF) (Bay et al., 2006) are well used in this context (Nawaf and Trémeau, 2014). In this case, the search for a point's correspondence is done w.r.t. all points in the destination image.

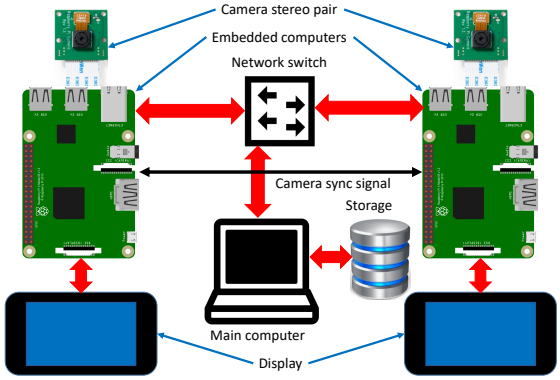

Figure 2. The built system internal design, it is composed mainly of (1) stereo camera pair, (2) Raspberry Pi © computers and (3) monitors.

In certain situations, some constraints can be imposed to facilitate the matching procedure. In particular, limiting the correspondence search zone. For instance, in case of pure forward motion, the focus of expansion (FOE) being a single point in the image, the search for the correspondence for a given point is limited to the epipolar line (Yamaguchi et al., 2013). Similarly, in case of sparse stereo matching the correspondence point lies on the same horizontal line in case of rectified stereo or on the epipolar line otherwise. This speeds up the matching procedure first by having less comparisons to perform, and second low-level features can be used (Geiger et al., 2011). According to our knowledge there is no method that proposes an adaptive search range following a rough depth estimation from lightness in underwater imaging.

2.2 Ego-Motion Estimation

Estimating the ego-motion of a mobile system is an old problem in computer vision. Two main categories of methods are developed in parallel, namely; simultaneous localization and mapping (SLAM) (Davison, 2003), and visual odometry (Nistér et al., 2004). In the following we highlight the main characteristics for both approaches.

SLAM family of methods uses probabilistic model to handle vehicle pose, although this kind of methods is developed to handle motion sensors and map landmarks, they work efficiently with visual information solely. In this case, a map of the environment is built and at the same time it is used to deduce the relative pose, which is represented using probabilistic models. Several solutions to SLAM involve finding an appropriate representation for the observation model and motion model while preserving effi-

Figure 3. The build hand-held stereo system in action.

cient and consistent computation time. Most methods use additive Gaussian noise to handle the uncertainty which imposes using extended Kalman Filter (EKF) to solve the SLAM problem (Davison, 2003). In case of using visual features, computation time and used resources grows significantly for large environments. We refer to (Bailey and Durrant-Whyte, 2006) for a comprehensive review of SLAM methods.

From another side, visual odometry methods use structure from motion methodology to estimate the relative motion (Nistér et al., 2004). Based on multiple view geometry fundamentals (Hartley and Zisserman, 2003), approximate relative pose can be estimated, this is followed by a BA procedure to minimize re-projection errors, which yields in improving the estimated structure. Fast and efficient BA approaches are proposed simultaneously to handle larger number of images (Lourakis and Argyros, 2009). However, in case of long time navigation, the number of images increases dramatically and prevent applying global BA if real time performance is needed. Hence, several local BA approaches have been proposed to handle this problem. In local BA, a sliding window copes with motion and select a fixed number of frames to be considered for BA (Mouragnon et al., 2009). This approach does not suit S-Type motion commonly used in surveys since the last n frames to the current frame are not necessarily the closest. Another local approach is the relative BA proposed in (Sibley et al., 2009). Here, the map is represented as Riemannian manifold based graph with edges representing the potential connections between frames. The method selects the part of the graph where the BA will be applied by forming two regions, an active region that contains the frames with an average re-projection error changes by more than a threshold, and a static region that contains the frames that have common measurements with frames in active region. When performing BA, the static region frames are fixed whereas active region frames are optimized. The main problem with this method is that distances between frames are metric, whereas the uncertainty is not considered when computing inter-frames distances.

3. HARDWARE PLATFORM

As mentioned earlier, we use an embedded system platform for our implementation. Being increasingly available and cheap, we choose the popular Raspberry Pi ©(RPi) [1] as main processing unit of our platform. This allows to run smoothly most of image processing and computer vision techniques. A description of the built system is shown in Figure 2, which is composed of two RPi's computers each is connected to one camera module to form a stereo pair. The cameras are synchronized using a hardware trigger. Both computers are connected to one more powerful computer that can be either within the same enclosure or on-board in our case. Using this configuration, the embedded computers are responsible for image acquisition. The captured stereo images are first partially treated on the fly to provide image quality information as will be details in Section 4.1. images are then transferred to the main computer which handles the ego-motion computation that the system undergoes. For visualization purposes, we use two monitors connected to the embedded computers to show live navigation and image quality information (See Figure 3).

[1] A credit-card size ARM architecture based computer with 1.2 GHz 64-bit quad-core CPU and 1GB of memory, running Rasbain ©, a Linux based operating system.

4. VISUAL ODOMETRY

Starting by computing and displaying image quality measures, the images are transfered over the network to a third computer as shown in Figure 2. This computer is responsible for hosting the visual odometry process, which will be explained in this section. We start first by introducing the used feature matching approach and then we present the ego-motion estimation, finally we explain the semi-global BA approach.

4.1 Image Quality Estimation

Real-time image quality estimation provides two benefits, first, it can alert the visual odometry process of having bad image quality, two reactions can be taken in this case, either pausing the process until taken image quality is recovered, or predicting position estimation based on previous poses and speed. We go for the first case while leaving the second for further development in future. Second, image quality indicators provide direct information to the operator to avoid going too fast in case of blur, or changing the distance to the captured scene when going under or over-exposed.

The first indicator is the image sharpness, we rely on image gradient measure that detects high frequencies often associated with sharp images, hence, we use a Sobel kernel based filtering which computes the gradient with smoothing effect. This removes the effect of dust commonly present in underwater imaging. We consider the sharpness measure to be the mean value of the computed gradient magnitude image. The threshold can be easily learned from images by fixing a minimum number of matched feature points needed to estimate correctly the ego-motion. Similarly, an image lightness indicator is estimate as the average of L channel in CIE-LAB color space.

4.2 Sparse Stereo Matching

Matching feature points between stereo images is essential to estimate the ego-motion. As the two cameras alignment is not perfect, we start by calibrating the camera pair. Hence, for a given point on the right image we can compute the epipolar line containing the corresponding point in the left image. However, based on the known fixed geometry, the corresponding point position is

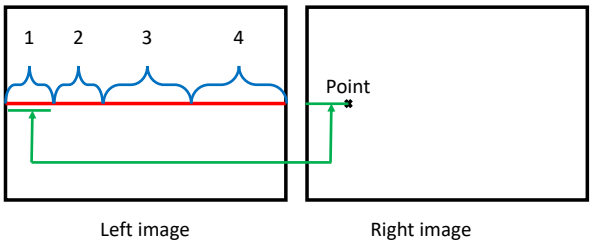

Figure 4. Illustration of stereo matching search ranges. (1) Impossible (2) Impossible in deep underwater imaging due to light's fading at far distances(3) Possible disparity (4) The point is very close so it becomes overexposed and undetectable.

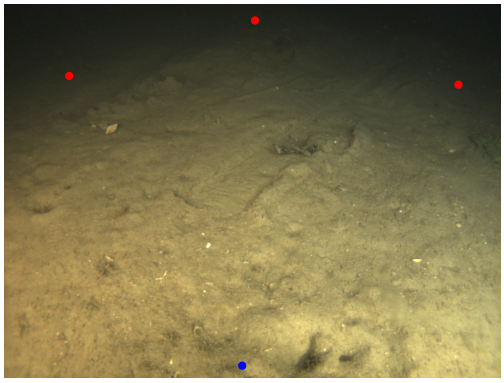

Figure 5. An example of underwater image showing minimum disparity (red dots, ~ 140 pixels) and maximum disparity (blue dot, ~ 430 pixels). Photograph COMEX ©.

constrained by a positive disparity. Moreover, given that at deep water the only light source is the one used in our system, the most far distance that feature points can be detected is limited, see Figure 5 for illustration. This means that there is a minimum disparity value that is greater than zero. Furthermore, when going too close to the scene, parts of the image will become overexposed, similar to the previous case, this imposes a limited maximum disparity. Figure 4 illustrates the aforementioned constraints by dividing the epipolar line into 4 zones in which only one is an acceptable disparity range. This range can be straightforward identified by learning from a set of captured images (oriented at 30 degrees for better coverage).

In our approach, we propose to constraint the so-called acceptable disparity range further, which corresponds to the third range in Figure 4. Given the used lighting system, we can assume a light diffuse reflection model where the light reflects equally in all directions. Based on inverse-square law that relates light intensity over distance, image pixels intensities are roughly proportional to their squared disparities. Based on such an assumption we could use pixels intensity to constraint the disparity and hence limiting the range of searching for a correspondence. In order to do so, we are based on a dataset of stereo images. For each pair we perform feature points matches. Each point match (x_i, y_i) and (x'_i, y'_i), x being the coordinate in the horizontal axis, we compute the squared disparity $d_i^2 = (x_i - x'_i)^2$. Next, we associate each d_i^2 to the mean lightness value of a window centered at the given point computed from L channel in CIE-LAB color space. We assign a large window size (≈ 12) to compensate for using Harris operator that promotes local minimum intensity pixels as salient feature points. The computed $(\bar{l}_{x_i,y_i}, d_i^2)$ pair shows the linear relationship between the squared disparity and the average

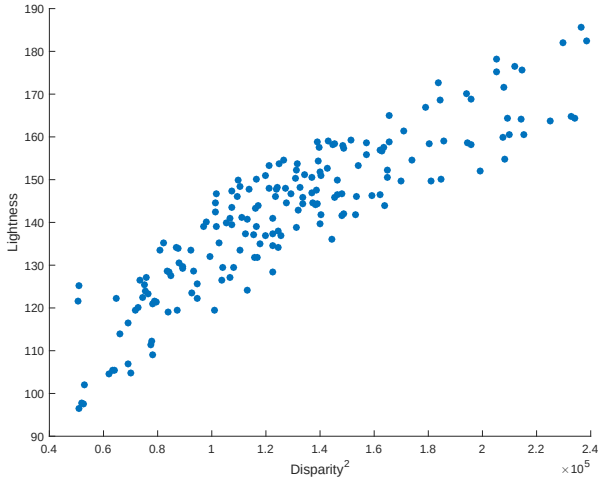

Figure 6. A subset of matched points squared disparity plotted against average pixel lightness.

lightness. A subset of such pairs is plotted in Figure 6.

In addition to finding the linear relationship between both variables, it is also necessary to capture the covariance that represents how rough is our approximation. More specifically, given the diagram shown in Figure 7, we aim at defining a tolerance t associated to each disparity as a function of lightness l. In our method, we rely on Principal Component Analysis (PCA) technique to obtain this information. In details, for a given lightness l_i, we first compute the corresponding squared disparity d_i^2 using a linear regression approach as follows:

$$d_i^2 = -\alpha l_i - \beta \tag{1}$$

$$\alpha = \frac{Cov(L, D^2)}{Var(L)} \tag{2}$$

$$\beta = \bar{l} - \alpha \bar{d^2} \tag{3}$$

where D and L are the disparity and lightness training set, $\bar{d}$ and $\bar{l}$ are their respective means

Second, let $\mathbf{V}_2 = (v_{2,x}, v_{2,y})$ be the computed eigenvector that correspondences to the smallest eigenvalue λ_2. Based on the illustration shown in Figure 7, the tolerance t associated to d_i^2 can be written as:

$$t = \sqrt{\lambda_2^2 (\frac{v_{2,x}^2}{v_{2,y}^2} + 1)} \tag{4}$$

By considering a normal error distribution of the estimated rough depth, and based on the fact that t is equal to one variance of D^2, we define the effective disparity range as:

$$d_i \pm \gamma \sqrt[4]{t} \tag{5}$$

where γ represents the number of standard deviations. It is trivial that γ is a trade-off between computation time and the probability of having points correspondences within the chosen tolerance range. We set $\gamma = 2$ which means there is 95% probability to cover the data.

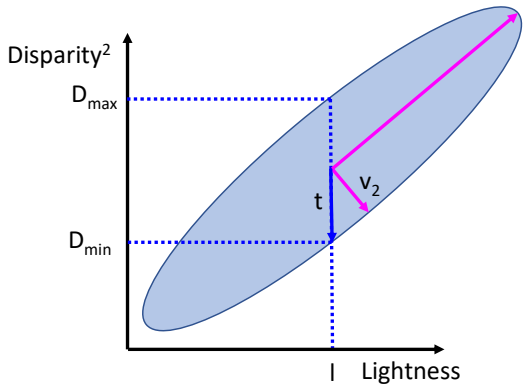

Figure 7. Illustration of disparity tolerance t given a lightness value l.

4.3 Initial Ego-Motion Estimation

Given left and right frames at time t (we call them previous frames), our visual odometry pipeline consists of four stages (An illustration is shown in Figure 8):

- Feature points matching for every new stereo pair $t+1$. As described in Subsection 4.2.
- 3D reconstruction of the matched feature points using triangulation as described in (Hartley and Zisserman, 2003). Two displaced point clouds are obtained at this step
- Relative motion computation using adaptation between the point clouds for the frames at t and $t+1$. The procedure is detailed in (Nawaf et al., 2017).
- Semi-Global BA procedure (Nawaf et al., 2016) is applied to minimize re-projection errors; to be explained in the following subsections.

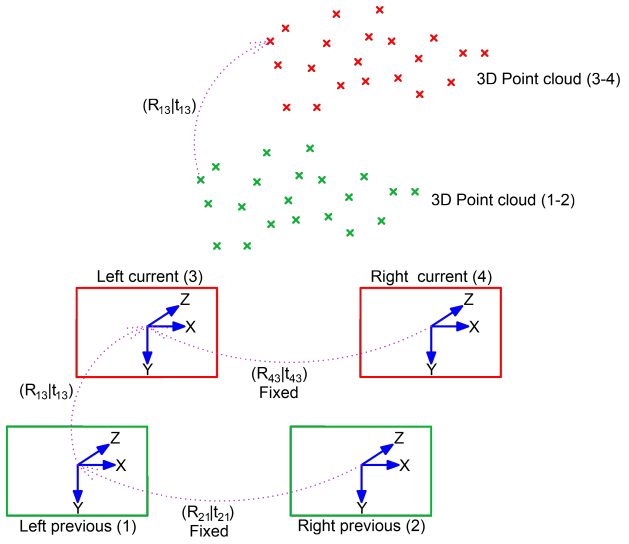

Figure 8. Image quadruplet, current (left and right) and previous (left and right) frames are used to compute two 3D point clouds. The transformation between the two points clouds is equal to the relative motion between the two camera positions.

4.4 Uncertainty In Visual Odometery

Like any visual odometry estimation, the estimated trajectory using the method mentioned in the previous section is exposed to a computational error, which translates to some uncertainty that grows in time. A global BA may handle this error accumulation, however it is time consuming. From another side, a local BA is a tradeoff for precision and computational time. The selection of n closest frames is done using standard Euclidean distance. Loop closure may occur when overlapping with already visited areas, which in turn enhances the precision. This approach remains valid as soon as the uncertainty is equal in all directions. However, as uncertainty varies across dimensions, the selection of the closest frames based on Euclidean distance is not suitable. In the following, we are going to prove that it is the case in any visual odometry method. Also, we will provide more formal definition of the uncertainty.

Most visual odometry and 3D reconstruction methods rely on matched feature points to estimate relative motion between two frames. The error of matched features is resulting from several accumulated errors. These errors are due, non-exclusively, to the following reasons; the discretization of 3D points projection to image pixels, image distortion, the camera internal noise, salient points detection and matching. By performing image undistortion, and constraining the points matching with the fundamental matrix. The aforementioned errors are considered to follow a Gaussian distribution; so as their accumulation. This is actually implicitly considered in most computer vision fundamentals. Based on this assumption, we can prove that the error distribution of the estimated relative pose is unequal among dimensions. Indeed, it can be fitted to a multivariate Gaussian whose covariance matrix has non equal Eigen values as we will see later. Formally, given a pair of matched points between two frames $\mathbf{m} \leftrightarrow \mathbf{m}'$. They can be represented by a multivariate Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma}) \leftrightarrow \mathcal{N}(\mathbf{m}', \mathbf{\Sigma})$, where $\mathbf{\Sigma} = diag(\sigma^2, \sigma^2)$. The pose estimation procedure relies on the fundamental matrix that satisfies $\mathbf{m}'\mathbf{F}\mathbf{m} = 0$. Writing $\mathbf{m} = [x \quad y \quad 1]^\top$ and $\mathbf{m}' = [x' \quad y' \quad 1]^\top$ in homogeneous coordinates. The fundamental matrix constraint for this pair of points can be written as:

$$x'xf_{11}+x'yf_{12}+x'f_{13}+y'xf_{21}+y'yf_{22}+ y'f_{23}+xf_{31}+yf_{32}+f_{3}3=0 \tag{6}$$

where f_{ij} is the element at row i and column j of $\mathbf{F}$. To show the estimated pose error distribution, without loss of generality, we consider one example of configuration; identity camera intrinsic matrix $K = diag(1\ 1\ 1)$. Let us now take the case of pure translational motion between the two camera frames, $\mathbf{T} = [T_x\ T_y\ T_z]^\top$, and $\boldsymbol{\theta} = [\theta_x\ \theta_y\ \theta_z]^\top = [0\ 0\ 0]$, where $\mathbf{T}$ and $\boldsymbol{\theta}$ being translation vector and rotation angles respectively. and fundamental matrix in this case is given as a skew-symmetric matrix of $\mathbf{T}$, denoted $[\mathbf{T}]_\times$. In this case Equation 6 simplifies to:

$$-x'yT_Z + x'T_Y + y'xT_Z - y'T_X - xT_Y + yT_X = 0 \tag{7}$$

By using enough matched points we can recover the translation vector $\mathbf{T}$ by solving a linear system. However, the Gaussian error associated to $x, y, x'$and y' will propagate equally to the variables T_X and T_Y, in contrary to T_Z where the error distribution is different due to the product of two variables, each follows a Gaussian distribution. So their combined covariance is equal to $\Sigma/2$. Moreover, due to the usage of least square approach though an SVD decomposition. The error distributions associated to recovered pose parameters are correlated (even though the observations are uncorrelated) as explained in (Strutz, 2010), this is also demonstrated experimentally as we will see later.

4.5 Pose Uncertainty Modeling

Pose uncertainty is difficult to estimate straightforward. This is due to the complexity of the pose estimation procedure and the number of variables. In particular, noise propagation through two consecutive SVDs (used for Fundamental matrix computation and Essential matrix decomposition). Instead, inspired by the unscented Kalman filter approach proposed in (Wan and Van Der Merwe, 2000), we proceed similarly by simulating noisy input and trying to characterize the output error distribution in this case. This process is illustrated in Figure 9. In our work, we propose to learn the error distribution based on finite pose samples. This is done using a Neural Network approach which fits well to our problem as it produces soft output.

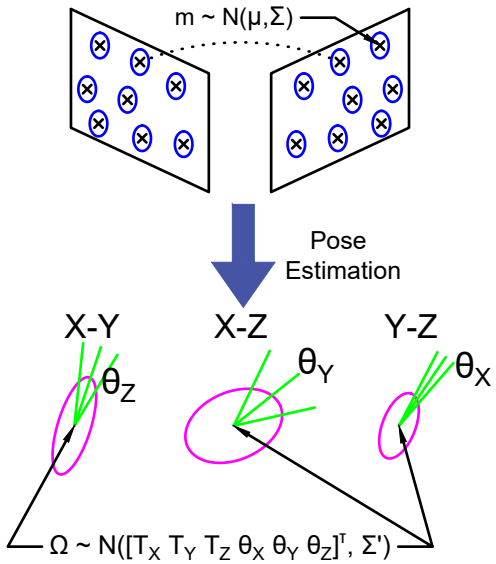

Figure 9. Illustration of error propagation through the pose estimation procedure. Estimated pose uncertainty is shown for each of the 6 DOF. Full covariance matrix can result from diagonal error distribution of matched 2D feature points

Formally, given a motion vector $\Omega_j = [\mathbf{T}_j \ \boldsymbol{\theta}_j]^\top$, ideally, we want to find the covariance matrix that express the associated error distribution. Being positive semi-definitive (PSD), such $n \times n$ covariance matrix has unique $(n^2 + n)/2$ entries, $n = 6$ in our case, this yields 21 DOF in which 6 are the variances. However, learning this number of parameters freely violates the PSD constraint. Whereas finding the nearest PSD in this case distorts largely the diagonal elements (being much fewer). At the same time, we found experimentally that the covariance between $\mathbf{T}$ and $\boldsymbol{\theta}$ variables is relatively small compared to such of inter $\mathbf{T}$ and inter $\boldsymbol{\theta}$. Thus, we propose to consider two covariance matrices Σ_T and Σ_θ. So in total we have 12 parameters to learn, in which 6 are the variances.

For the aim of learning Σ_T and Σ_θ, we have created a simulation of the pose estimation procedure. For a fixed well distributed 3D points $X_i \in \mathbb{R}^3 : i = 1..8$, we simulate two cameras with known intrinsics and extrinsic. The points are projected according to both cameras to 2D image points, let us say $\{\mathbf{x}_i \in \mathbb{R}^2\}$ and $\{\mathbf{x}'_i \in \mathbb{R}^2\}$. These points are disturbed with random Gaussian noise. Next, the 3D relative pose is estimated using the disturbed points. Let $\tilde{\Omega}_j = [\tilde{\mathbf{T}}_j \ \tilde{\boldsymbol{\theta}}_j]^\top$ be the estimated relative pose. Repeating the same procedure (with the same motion Ω_j) produces a point cloud of poses around the real one. Now, we compute the covariance matrices Σ_T and Σ_θ of the resulting pose cloud in order to obtain the uncertainty associated to the given motion Ω_j. Further, we repeat this procedure for a wide range of motion values[2]. Now, having the output covariance matrices (two for each motion vector Ω_j), we proceed to build a system which learns the established correspondences (motion $\Leftrightarrow$ uncertainty). So in case of new motion we will be able to predict the uncertainty. This soft output is offered by Neural Networks by nature, which is the reason we adopt this learning method. In our experiments, we found that a simple Neural Network with single hidden layer (Bishop, 1995) was sufficient to fit well the data. The input layer has six nodes that correspond to motion vector. The output layer has 12 nodes which corresponds to the unique entries in Σ_T and Σ_θ, hence, we form our output vector as:

$$O = [\Sigma_T^{11} \ \Sigma_T^{22} \ \Sigma_T^{33} \ \Sigma_T^{12} \ \Sigma_T^{13} \ \Sigma_T^{23} \ \Sigma_\theta^{11} \ \Sigma_\theta^{22} \ \Sigma_\theta^{33} \ \Sigma_\theta^{12} \ \Sigma_\theta^{13} \ \Sigma_\theta^{23}]^\top \tag{8}$$

where $\Sigma_\cdot^{ij}$ is the element of row i and column j of a covariance matrix.

In the learning phase, we use the Levenberg-Marquardt backpropagation which is a gradient-descent based as described in (Demuth et al., 2014). Further, by using the mean-squared error as a cost function we could achieve around 3% error rate. The obtained parameters are rearranged in symmetric matrices. In practice, the obtained matrix is not necessarily PSD. We proceed to find the closest PSD as $Q\Lambda_+Q^{-1}$, where Q is the eigenvector matrix of the estimated covariance, and Λ_+ is the diagonal matrix of Eigen values in which negative values are set to zero.

4.6 Semi-Global Bundle Adjustment

After initiating the visual odometry, the relative pose estimation at each frame is maintained with a table that contains all pose related information (18 parameters per pose, in which 6 for the position, and 12 for two covariance matrices). At any time, it is possible to get the observations in the neighborhood of the current pose being estimated in order to find potential overlaps to consider while performing BA. Since we are dealing with statistical representations of the observations, a divergence measure has to be considered. Here, we choose Bhattacharyya distance for being suitable to our problem (Modified metric variation can also be used (Comaniciu et al., 2003)) . Formally, the distance between two observations $\{\Omega^1, \Sigma_T^1, \Sigma_\theta^1\}$ and $\{\Omega^2, \Sigma_T^2, \Sigma_\theta^2\}$ is given as:

$$\begin{aligned} D = &\frac{1}{8}(\Omega^1 - \Omega^2)^\top \Sigma^{-1}(\Omega^1 - \Omega^2) + \\ &\frac{1}{2}\ln\left(\frac{\det \Sigma}{\sqrt{\det \Sigma^1 + \det \Sigma^2}}\right) \end{aligned} \tag{9}$$

where

$$\Sigma_\cdot = \begin{bmatrix} \Sigma_{T_\cdot} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\theta_\cdot} \end{bmatrix}, \ \Sigma = \frac{\Sigma^1 + \Sigma^2}{2} \tag{10}$$

Having selected the set of frames $\mathbb{F}$ in the neighborhood of the current pose statistically, we perform BA as follows; First, we divide $\mathbb{F}$ into two subsets similar to (Sibley et al., 2009), the first subset $\mathbb{F}_d$ contains the current and previous frames in time, whereas the other sub-set $\mathbb{F}_s$ contains the remaining frames, mostly resulting from overlapping with an already scanned area. Second, BA is performed on both subsets, however, the parameters related to $\mathbb{F}_s$ are masked as static so they are not optimized in contrary

[2] In the performed simulation, we use the range $[0 - 1]$ with 0.25 step size for each of the 6 dimensions, these values are in radians in case of rotation. This raises up to 15625 test cases.

to $\mathbb{F}_d$. This strategy is necessary in order to keep past trajectories consistent.

After determining the error distribution arising with a new pose, it has to be compounded with the error propagated from the previous pose. Similar to SLAM approaches, we propose to use a *Kalman filter* like gain which allows controllable error fusion and propagation. Given an accumulated previous pose estimation defined by $\{\Omega^p, \Sigma_T^p, \Sigma_\theta^p\}$ and a current one $\{\Omega^c, \Sigma_T^c, \Sigma_\theta^c\}$, an updated current pose is calculated as:

$$\Omega^u = \Omega^c \tag{11}$$

$$\Sigma_\theta^u = (I - \Sigma_\theta^p(\Sigma_\theta^p + \Sigma_\theta^c)^{-1})\Sigma_\theta^p \tag{12}$$

$$\Sigma_T^u = (I - \Sigma_T^p(\Sigma_T^p + \Sigma_T^c)^{-1})\Sigma_T^p \tag{13}$$

5. EVALUATION

The first experiments were carried out to test and enhance the hardware platform with the help of a diver, snapshots of the operation are shown in Figure 3. The taken images are processed using photogrammetry techniques to validate the quality of the taken images. The reconstructed 3D models, as illustrated in Figure 10, are of good quality. Stereo image synchronization is also validated by observing the relative pose estimation between each pair and comparing it with the calibration external parameters.

The proposed visual odometry method is desired to represent a trade-off between precision and computation time, the maximum precision being the case of global BA, whereas the fastest computation time is pure visual odometry. Moreover, a performance improvement is expected w.r.t local method due for better selection of neighboring observations. Therefore, we analyze the performance of our method from two points of view; computation time and precision.

5.1 Computation Time

We tested and compared the computation speed of our method compared to using high level feature descriptors, specifically SIFT and SURF. At the same time, we monitor the precision for each test. The evaluation is done using the same set of images.

We run our experiments using the speed optimized BA toolbox proposed in (Lourakis and Argyros, 2009). In the obtained results, the computation time when using the reduced matching search range as proposed in this work is $\sim 72\%$ compared [3] to the method using the whole search range (range 3 in Figure 4). Concerning SIFT and SURF, the computation time is 342% and 221% respectively compared to the proposed method. The precision of the obtained odometry is reasonable which is within the limit of 3% for the average translational error and $0.02[deg/m]$ for the average rotational error.

5.2 Simulation Using Orthophoto

Our work falls within a preliminary preparation for a real mission. All the experiments are tested within a simulated environment which uses images from previously reconstructed orthophoto (Drap et al., 2015). The advantage of using simulated environment is that we can define precisely the trajectory, and then, after running the visual odometry method we can evaluate the performance and tune different components. Especially, with the lack of real sequences provided with odometry ground truth. Hence, we created a dataset of images based on simulating stereo camera motion. We evaluate the proposed semi-global BA compared to three cases, using global BA, local BA and without using BA. As expected, the method that uses global BA performs best in this context. The translation error is 1.2% while the rotation error 0.009 [deg/m]. Followed by our method, with 2.44% of translation and 0.011 [deg/m] of rotation errors. This is fairly ahead of the local BA method that achieved 3.68% of translation and 0.012 [deg/m] of rotation errors. The optimization free visual odometry showed the largest divergence with a translation error of 6.8% and rotation error of 0.08 [deg/m].

[3] The time evaluation is shown in percentage because the evaluation is carried out on three platforms with different computational power, in which one is an embedded unit. The minimum computation time being 220 ms

6. CONCLUSIONS AND PERSPECTIVES

In this work, we introduced several improvements to the current traditional visual odometry approach in order to serve in the context of underwater surveys. The goal is to be adapted to embedded systems known for their lower resources. The sparse feature points matching guided with a rough depth estimation using lightness information is the main factor beyond most of the gain in computation time compared to sophisticated feature descriptors combined with brute-force matching. Also, using stochastic representation and selection of frames in the semi-global BA improved the precision compared to local BA methods while remaining within real-time limits.

Our future perspectives are mainly centered on reducing the overall system size, for instance, replacing the main computer in our architecture with a third embedded unit, which in turn does not keep evolving. This also allows to reduce the power consumption which increases the navigation time. On the other hand, dealing with visual odometry failure is an important challenge specially in the context of underwater imaging, which is mainly due to bad image quality. The ideas of failing scenarios discussed in this paper can be extended to deal with the problem of interruptions in the obtained trajectory.

ACKNOWLEDGEMENTS

We wish to thank David Scaradozzi and his group for building the designed system, Olivier Bianchimani for carrying out the tests underwater, and Mohamad Alshara for his drawings. This work has been partially supported by both a public grant overseen by the French National Research Agency (ANR) as part of the program *Contenus numériques et interactions (CONTINT) 2013* (reference: ANR-13-CORD-0014), GROPLAN project (Ontology and Photogrammetry; Generalizing Surveys in Underwater and Nautical Archaeology) [4], and by the French Armaments Procurement Agency (DGA), DGA RAPID LORI project (LOcalisation et Reconnaissance d'objets Immergés). Logistic support for underwater missions is provided by COMEX [5].

REFERENCES

Bailey, T. and Durrant-Whyte, H., 2006. Simultaneous localization and mapping (slam): Part ii. *IEEE Robotics & Automation Magazine* 13(3), pp. 108–117.

[4] `http://www.groplan.eu`
[5] `http://www.comex.fr/`

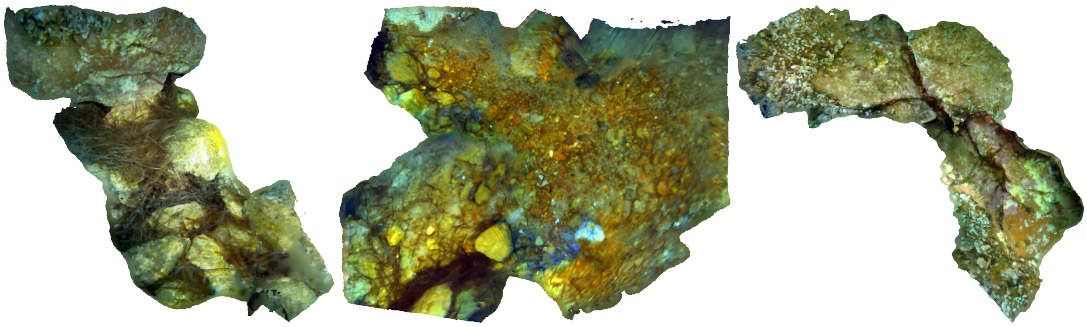

Figure 10. 3D reconstructed models using images captured with the built system.

Bay, H., Tuytelaars, T. and Van Gool, L., 2006. Surf: Speeded up robust features. *European Conference on Computer Vision* pp. 404–417.

Bishop, C. M., 1995. *Neural networks for pattern recognition*. Oxford university press.

Comaniciu, D., Ramesh, V. and Meer, P., 2003. Kernel-based object tracking. *IEEE Transactions on pattern analysis and machine intelligence* 25(5), pp. 564–577.

Davison, A. J., 2003. Real-time simultaneous localisation and mapping with a single camera. In: *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, IEEE, pp. 1403–1410.

Demuth, H. B., Beale, M. H., De Jess, O. and Hagan, M. T., 2014. *Neural network design*. Martin Hagan.

Drap, P., 2012. *Underwater photogrammetry for archaeology*. INTECH Open Access Publisher.

Drap, P., Merad, D., Hijazi, B., Gaoua, L., Nawaf, M. M., Saccone, M., Chemisky, B., Seinturier, J., Sourisseau, J.-C., Gambin, T. et al., 2015. Underwater photogrammetry and object modeling: A case study of xlendi wreck in malta. *Sensors* 15(12), pp. 30351–30384.

Geiger, A., Ziegler, J. and Stiller, C., 2011. Stereoscan: Dense 3d reconstruction in real-time. In: *IEEE Intelligent Vehicles Symposium*, pp. 963–968.

Hartley, R. and Zisserman, A., 2003. *Multiple view geometry in computer vision*. Cambridge university press.

Itseez, 2015. Open source computer vision library. `https://github.com/itseez/opencv`.

Lourakis, M. I. and Argyros, A. A., 2009. Sba: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software* 36(1), pp. 2.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), pp. 91–110.

Montemerlo, M. and Thrun, S., 2007. *FastSLAM: A scalable method for the simultaneous localization and mapping problem in robotics*. Vol. 27, Springer.

Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F. and Sayd, P., 2009. Generic and real-time structure from motion using local bundle adjustment. *Image and Vision Computing* 27(8), pp. 1178–1193.

Nawaf, M. M. and Trémeau, A., 2014. Monocular 3d structure estimation for urban scenes. In: *Image Processing (ICIP), 2014 IEEE International Conference on*, IEEE, pp. 3763–3767.

Nawaf, M. M., Drap, P., Royer, J. P., Merad, D. and Saccone, M., 2017. Towards guided underwater survey using light visual odometry. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W3, pp. 527–533.

Nawaf, M. M., Hijazi, B., Merad, D. and Drap, P., 2016. Guided underwater survey using semi-global visual odometry. *15th International Conference on Computer Applications and Information Technology in the Maritime Industries* pp. 288–301.

Nistér, D., Naroditsky, O. and Bergen, J., 2004. Visual odometry. In: *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, Vol. 1, IEEE, pp. I–652.

Sibley, D., Mei, C., Reid, I. and Newman, P., 2009. Adaptive relative bundle adjustment. In: *Robotics: science and systems*, Vol. 32, p. 33.

Stone, J. E., Gohara, D. and Shi, G., 2010. Opencl: A parallel programming standard for heterogeneous computing systems. *IEEE Des. Test* 12(3), pp. 66–73.

Strutz, T., 2010. *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*. Vieweg and Teubner.

Thrun, S., Burgard, W. and Fox, D., 2005. *Probabilistic robotics*. MIT press.

Triggs, B., McLauchlan, P., Hartley, R. and Fitzgibbon, A., 2000. Bundle adjustment: a modern synthesis. *Vision algorithms: theory and practice* pp. 153–177.

Wan, E. A. and Van Der Merwe, R., 2000. The unscented kalman filter for nonlinear estimation. In: *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, Ieee, pp. 153–158.

Yamaguchi, K., McAllester, D. and Urtasun, R., 2013. Robust monocular epipolar flow estimation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, IEEE Computer Society, Washington, DC, USA, pp. 1862–1869.