Visual SLAM for Underwater Vehicles using Video Velocity Log and Natural Landmarks

Dr. Joaquim Salvi
University of Girona
Girona, Spain
Email: qsalvi@eia.udg.es

Dr. Yvan Petillot
Heriot-Watt University
Edinburgh, Scotland
Email: Y.R.Petillot@hw.ac.uk

Stephen Thomas
Heriot-Watt University
Edinburgh, Scotland
Email: stho121@gmail.com

Josep Aulinas
University of Girona
Girona, Spain
Email: josepaulinas@gmail.com

Abstract— A visual SLAM system has been implemented and optimised for real-time deployment on an AUV equipped with calibrated stereo cameras. The system incorporates a novel approach to landmark description in which landmarks are local submaps that consist of a cloud of 3D points and their associated SIFT/SURF descriptors. Landmarks are also sparsely distributed which simplifies and accelerates data association and map updates. In addition to landmark-based localisation the system utilises visual odometry to estimate the pose of the vehicle in 6 degrees of freedom by identifying temporal matches between consecutive local submaps and computing the motion. Both the extended Kalman filter and unscented Kalman filter have been considered for filtering the observations. The output of the filter is also smoothed using the Rauch-Tung-Striebel (RTS) method to obtain a better alignment of the sequence of local submaps and to deliver a large-scale 3D acquisition of the surveyed area. Synthetic experiments have been performed using a simulation environment in which ray tracing is used to generate synthetic images for the stereo system.

I. INTRODUCTION

Simultaneous localisation and mapping (SLAM) is an approach to localisation in which the vehicle uses relative observations of environment landmarks to incrementally construct an environment map and simultaneously localise itself within this map. Embedding this joint mapping and localisation approach in a probabilistic framework it is observed that as more observations are made the correlation between estimated landmark positions increases. Consequently, in time the relative position of landmarks will be known with high accuracy, and the localisation uncertainty of the vehicle relative to map will be bounded by the map quality and sensor accuracy. This approach to localisation is necessary in many environments including underground and underwater as satellite-based GPS is unavailable due to attenuation of the signals.

Almost any type of sensor can be incorporated into the SLAM framework but vision-based systems are rapidly increasing in popularity as they are passive, cheap, light-weight and have a long range, high resolution, low power requirements, excellent object recognition capabilities and can also provide motion estimates. However, until recently the strong attenuation and scattering of light underwater has limited the use of video and underwater vehicles have primarily utilised acoustic technologies to sense their environment. Generally, a Doppler Velocity Log (DVL) is used in conjunction with an Inertial Navigation Unit (INU) to estimate the velocity

of the vehicle with respect to the seabed, Side Scan Sonar (SSS) is used to acquire images of the environment and acoustic cameras are used for monitoring. The most substantial drawback of acoustic technologies is their cost, although their size, weight and power consumption cannot be neglected. Video cameras are significantly cheaper and provide high-resolution images that are ideal for scientific exploration and offshore structures inspection. Consequently, underwater vehicles have been developed that are able to acquire video images from close range. These vehicles have excelled in applications including docking, inspection and generating maps of flat-terrain by aligning hundreds of images using the photo-mosaicing technique. However, photo-mosaicing is unsuitable for mapping environments of interest such as coral reefs, hydrothermal vents, archaeologic sites and man-made structures as they have significant 3D relief which results in misalignments and artefacts that deteriorate the mapping [1].

Simultaneous localisation and mapping has been extensively researched as a solution to this problem for land-based vehicles and excellent results have been obtained both indoors and outdoors. In the case of vision-based SLAM, most approaches use a dense distribution of landmarks, each described by a single 2D image feature. In an underwater scenario images are generally corrupted by a more significant level of noise and distortion, features can be sparsely distributed, and appearance alone is often not discriminant. Therefore, a successful underwater vision-based SLAM system must employ much more robust features so that data association is possible in these harsh conditions, and also take into account the likely sparseness of maps. Unsurprisingly, only a few papers have tackled vision-based SLAM underwater. Eustice et. al. [2] proposed a system based on the Information filter with measurements provided by inertial sensors and monocular video. Mahon et. al. [3] proposed a system based on the extended Kalman filter with measurements provided by a pencil beam scanning sonar and monocular video. Finally, Saez et. al. [4] proposed an offline system based on entropy minimisation with measurements provided by a dense 3D stereo-vision system.

In this paper we propose a solution to generate 3D maps underwater using an AUV equipped with only stereo cameras. The solution is based on the extended and unscented Kalman filters and incorporates a novel approach to landmark description, data association and motion estimation. The performance

of the system has been evaluated in synthetic environments and the results suggest the approach is particularly well suited to applications in tempered and tropical waters where visibility is less restricted.

II. PROPOSED SOLUTION

A. Image pre-processing

Identifying landmarks underwater is complicated by the highly dynamic lighting conditions, decreasing visibility with depth and turbidity, and image artifacts such as aquatic snow. Therefore, the video images are pre-processed to minimise the influence these effects have on the 3D reconstruction and data association. Firstly, a Homomorphic filter is used to normalise the brightness across the greyscale image and compensate for non uniform lighting patterns. Following this Contrast-Limited Adaptive Histogram Equalization (CLAHE) is applied to small regions of the image to enhance the contrast. A further bilinear interpolation is performed to remove artificially induced boundaries between regions. Finally, an Adaptive Noise-Removal Filtering is carried out to remove the noise produced by the equalization especially in those areas with small variance (constant brightness). The resulting images are brighter, better contrasted and normalised. This facilitates the comparison of two images acquired at different times and viewpoints, enabling the matching of image features. Applying this process the number of features detected in the images is increased by approximately ten times and features are much better distributed in the image. In order to extract metrics from the two images both cameras were calibrated to obtain the intrinsic matrices K_L and K_R and the relative transformation $^R T_L = [^R R_L \ ^R t_L]$, where $^R R_L$ and $^R t_L$ are the rotation and translation of camera L wrt camera R respectively. Lens distortion is removed and both images are rectified.

B. Landmark Characterisation

Based on several surveys of feature descriptors [5], [6] we choose to extract a mixture of SIFT and SURF features from each pair of synchronised, pre-processed images to obtain a dynamic trade-off between quantity, robustness and computational complexity. These features are invariant to image translation, scaling and rotation, and are not sensitive to changes in illumination, affine/perspective distortion, addition of noise, clutter and occlusion, which makes them ideal for wide-baseline stereo matching. Both algorithms construct a scale-space, Hessian for SURF and Gaussian for SIFT, locate maxima (keypoints) within this scale space and then generate a feature descriptor for each keypoint. On average the algorithm required 104ms to construct a 6 octave Gaussian scale-space and extract the maxima, compared to 31ms for the equivalent Hessian scale space. Moreover, on average each SIFT descriptor was generated in 1.9ms compared to 1.5ms for each SURF descriptor. Therefore, SURF is considerably faster, although SURF also tends to return a lower number of keypoints. Once features have been extracted from each image their putative correspondences are identified by computing the sum of squared differences (SSD) between

each pair of descriptors and applying gated nearest neighbour matching. The time required for this matching is negligible, however, these matches tend to include outliers. Because the stereo system is calibrated the fundamental matrix, F which describes the relative camera pose can be computed by $F = K_R^{-T} \ ^R R_L \ T \ K_L^{-1}$ where T is the skew matrix of the translation vector $^R t_L$. The fundamental matrix defines the bilinear constraint $m_R^T F m_L = 0$ between the $2D$ homogenous coordinates of corresponding image points m_R and m_L. In words, this constraint enables us to map any point in one camera to a corresponding line (epipolar line) in the opposite camera which represents all possible locations at which the same ray could be projected taking into account all possible scene depths. Therefore, for each point we can compute the deviation of its correspondence from the relevant epipolar line and eliminate the majority of outliers by applying a simple threshold. Note that it is preferable to be strict at this point and remove some correct matches rather than accept false matches as these will deteriorate the $3D$ reconstruction. Finally, we compute the disparity between the remaining $2D$ points and remove those whose disparity is larger than 3σ, where σ is the standard deviation of the disparity distribution. This process permits the removal of further outliers, since outliers generally have large disparity discrepancies.

Once the set of correct correspondences has been obtained their $3D$ structure can be determined using a linear triangulation. Firstly the points are converted to metric coordinates $\hat{m}_L = K_L^{-1} m_L$ and $\hat{m}_R = K_R^{-1} m_R$. Following this we compute matrix A_i for every correspondence i as follows:

$$A_i = \begin{pmatrix} 0 & -1 & \hat{y}_{Li} & 0 \\ -1 & 0 & \hat{x}_{Li} & 0 \\ (-R_2 + \hat{y}_{Ri} R_3) & -t_y + \hat{y}_{Ri} t_z \\ (-R_1 + \hat{x}_{Ri} R_3) & -t_x + \hat{x}_{Ri} t_z \end{pmatrix} \quad (1)$$

where $^R R_L = (R_1 \ R_2 \ R_3)^T$ and $^R t_L = (t_x, t_y, t_z)^T$. Finally, the singular value decomposition of matrix A_i is computed so that $A_i = U_i D_i V_i^T$. The $3D$ point M_i before normalisation and wrt camera L corresponds to the fourth column of V_i [7]. Finally, we remove isolated 3D points since they introduce large residues in the re-observations of landmarks. The whole process permits the acquisition of a local $3D$ surface of the imaged seabed measured wrt the current vehicle position.

These 3D points and their corresponding 2D points and feature descriptors are then stored together as a local submap. During this step the 3D points are transformed to the coordinate system of the vehicle using a fixed transformation so that each local submap can contain features observed from different sensors all referenced to a single coordinate system. In this implementation we represent landmarks as a single, complete local submap and use the centre of gravity of the cloud of 3D points as an anchor point in the global map. Therefore, each landmark has an arbitrary shape (non-geometric landmarks) and combines many robust SIFT or SURF features, which results in extremely distinctive and very robust landmark descriptions. To the best of our knowledge no existing SLAM

implementation utilises a similar approach, and almost all employ significantly less descriptive features such as edges, contours or individual geometric descriptors.

C. Data Association

Low overlap imagery is common for AUVs due to the fact that vehicle speeds are moderate, video frame rates are usually low, and attenuation, scattering and limited illumination restrict visibility and dictate that the vehicle must stay relatively close to the seabed which reduces the effective field of view for each image. Taking this into account and also the fact that landmarks are represented by large sparsely distributed submaps, it is reasonable to assume that usually only one landmark will be visible at any point in time. As a result the current approach to data association simply involves identifying the maximum likelihood match between the current observation and the set of map landmarks based on both the 3D points and 2D descriptors.

Firstly, the 2D descriptors of the current observation are compared to the set of map landmarks that are estimated to be in the vicinity of the vehicle. The vicinity is determined as a function of the camera field of view (range and aperture), and the uncertainty of the vehicle position and landmark positions. The 2D descriptor matching also utilises the gated nearest neighbour algorithm based on the SSD and generally produces outliers due to the changes in viewpoint and indistinct descriptors associated with surfaces such as sand, mud and rock. Consequently, the epipolar constraint is applied to remove outliers in the same manner as described previously. However, in this case the fundamental matrix is unknown and depends on the rotation and translation relative to the first observation of the landmark. Therefore, the algorithm attempts to robustly estimate the fundamental matrix using the normalised 8-point algorithm and least median squares (LMedS) method [8]. When the LMedS estimation completes the set of inliers according to the epipolar constraint is known and the algorithm registers the corresponding 3D points using the method proposed by Mian et. al. [9] to recover the relative transformation. Consider two $3 \times n$ matrices M and S which contain the corresponding $3D$ points and their associated gravity centers $\hat{m}$ and $\hat{s}$. The method computes $\hat{M}$ and $\hat{S}$ by subtracting $\hat{m}$ and $\hat{s}$ from every point in M and S respectively, which shifts the center of gravity to the origin. Following this $K = \hat{S}\hat{M}^T/n$ is computed and a singular value decomposition is performed to obtain $K = UAV^T$ and from this the rotation matrix $R_1 = VU^T$. If $det(R_1) > 0$ the desired rotation matrix $R = R_1$, otherwise:

$$R = V \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & det(VU^T) \end{pmatrix} U^T \qquad (2)$$

The translation vector t is $t = \hat{m} - R\hat{s}$. Using this rotation and translation the stored landmark gravity center can be transformed into the current vehicle frame. The compatibility of the landmark and observation can also be evaluated based on the least mean squared (LMS) registration error in addition to the number of inliers.

D. Visual Odometry

The data association algorithm described previously involves a registration of the set of observed 3D points with a map landmark, which provides an estimate of the relative rotation and translation between the first and current observation of the landmark. Note that for small motions the rotation and translation estimate obtained will be quite noisy as we are estimating unconstrained 6DOF motion. However, by recording the estimated position and pose of the vehicle when the landmark is created we have demonstrated that it is possible to perform fairly robust inter-frame motion estimation with almost no additional computational burden. Due to the fact that at least 8 correspondences are required for the fundamental matrix estimation, continuous visual odometry that is independent of the Kalman filter state is not always attained. This limitation arises from the fact that when an insufficient number of features are matched it is not possible to recover the motion during this period. Consequently, while this motion estimation approach can significantly reduce the error in the estimation of the vehicle location and pose due to its high accuracy, estimates are overconfident due to the inherited error after the non-observation period.

E. Video Velocity Log

The motion of the vehicle relative to the scene is also estimated by analysing the spatial differences between related pixels in consecutive video frames. Consider a three dimensional point in the scene $P = [x, y, z]$, the velocity V of a relative motion between P and the camera is defined as:

$$V = -T - \omega P \qquad (3)$$

where T is the translational component and ω is the rotational component of the motion. The motion field v is defined as:

$$v = f\frac{ZV - V_zP}{Z^2} \qquad (4)$$

where f is the focal length of the camera. From these two equations we can compute the components of the motion field:

$$v_x = \frac{T_zx - T_xf}{Z} - \omega_yf + \omega_zy + \frac{\omega_xxy}{f} - \frac{\omega_yx^2}{f} \qquad (5)$$

$$v_y = \frac{T_zy - T_yf}{Z} + \omega_xf - \omega_zx - \frac{\omega_yxy}{f} - \frac{\omega_xy^2}{f} \qquad (6)$$

We remove the dependence on the rotational component by compensating for rotation using measurements from other sensors and hence only estimate the translational component. For considering the case when $T_z \neq 0$ we introduce the point $p_0 = [x_0, y_0]^T$ with $x_0 = \frac{fT_x}{T_z}$ and $y_0 = \frac{fT_y}{T_z}$. In this case we have:

$$v_x = (x - x_0)\frac{T_z}{Z} \qquad (7)$$

$$v_y = (y - y_0)\frac{T_z}{Z} \qquad (8)$$

Therefore, the motion vectors radiate from a common origin p_0 when the motion is a pure translation in the z direction. Note that when $T_z = 0$ the equations become:

$$v_x = -f\frac{T_x}{Z} \tag{9}$$

$$v_y = -f\frac{T_y}{Z} \tag{10}$$

Therefore, in this case all the motion vectors are parallel. The motion vectors for a large number of pixels (optical flow) are computed using the pyramidal implementation of the Lucas-Kanade feature tracker, which is available in the OpenCV-library. This method uses the spatial intensity gradient and a Newton-Raphson iteration to find matches between feature points. For the available frame rate and speed of the AUV the shift between consecutive images is small which means the algorithm is computationally efficient. Once the motion vectors are obtained they are converted to metrics using the intrinsic parameters of the camera estimated from the camera calibration. External estimates of the angular velocities ω_x, ω_y and ω_z are then used to compensate for the rotational components of Equations 5 and 6 and the rate of change of the depth T_z is provided by an altitude sensor. Finally, the translation velocity of each vector is computed by solving for T_x and T_y and the median is used for the final estimation.

III. EKF-SLAM AND UKF-SLAM

The SLAM algorithm, models, and customised filter functionality were implemented within the well defined hierarchical framework of the Bayes++ open source library as this enables rapid switching between fundamentally similar state based filters such as the extended Kalman filter and unscented Kalman filter. In addition, the Bayes++ library performs automatic checks to detect numerical failures and ill conditioned matrices and also maintains the symmetry of matrices which ensures the algorithm does not continue to trust state estimates that are undoubtedly incorrect. A Kalman filter is essentially composed of three steps: Prediction, Observation and Update. In this implementation the filter inputs are the visual odometry estimates and landmark re-observations provided by the stereo vision system. The full filter state x_k has the following form:

$$\hat{x}_k = \begin{bmatrix} \hat{v} & \hat{m}_0 & \dots & \hat{m}_{N-1x} \end{bmatrix}^T$$

where $\hat{v} = \begin{bmatrix} {}^E x & {}^E \Theta \end{bmatrix}$ and ${}^E x$ and ${}^E \Theta$ are 3 element vectors containing an estimate of the absolute position (x,y,z) and orientation (roll,pitch,yaw) of the vehicle with respect to Earth $\{E\}$. Once the ith landmark is observed the state is augmented with the 3 element vector m_i containing the absolute position of the landmark with respect to Earth. Note that for simplicity the Earth frame is located at the initial position of the vehicle. Accordingly, the full EKF covariance P_k has the form:

$$P_k = \begin{bmatrix} P_{vv} & P_{vm} \\ P_{vm}^T & P_{mm} \end{bmatrix} \tag{11}$$

Due to the structure, size and distribution of landmarks, the number of landmarks contained in the filter is significantly lower than typical SLAM implementations and provides considerable reductions in prediction and observation update times for comparatively sized maps.

A. Prediction Update

The predictions updates are based on a non-linear constant velocity prediction model, which provides a generic, platform independent solution. Only the vehicle position and orientation are affected by the prediction model so the update can be performed in linear time using the augmented state approach. In this case, the prediction update of the state is trivial and can be expressed as:

$$\hat{x}_{k|k-1} = \begin{bmatrix} f_v(\hat{v}_{k-1|k-1}, u_k) \\ \hat{m} \end{bmatrix} \tag{12}$$

where $\hat{v}_{k-1|k-1}$ is the current estimate of the 6 vehicle states contained within $\hat{x}_{k-1|k-1}$, the function $f_v(\hat{v}_{k-1|k-1}, u_k)$ is the augmented state function which models the motion kinematics and $u_k = 0$ as there are no control inputs. The augmented state version of the covariance prediction has linear complexity in the number of landmarks and has the form:

$$P_{k|k-1} = \begin{bmatrix} \nabla f_v P_{vv} \nabla f_v^T + Q_k & \nabla f_v P_{vm} \\ P_{vm}^T \nabla f_v^T & P_{mm} \end{bmatrix} \tag{13}$$

where ∇f_v is the Jacobian (linear approximation) of $f_v()$ evaluated at the estimate $\hat{v}_{k-1|k-1}$. The constant velocity parameter and the zero mean uncorrelated Gaussian process noises w_k which affect the motion observation and have covariance Q_k are fixed, determined offline and stored in a configuration file. Note that these process noises define the reaction of the filter to sudden changes of the ground truth orientation/velocity of the vehicle and require tuning for different vehicles and scenarios.

B. Observation Updates

The observation updates are based on a model which represents the geometry of the measurements obtained by the stereo-vision system at the real pose of the vehicle, together with the predicted measurements provided by the current filter state. To simplify the process and utilise a common model for all sensor measurements, all observations are pre-transformed to a common vehicle frame $\{V\}$ using fixed transformations. We consider that all landmarks are stationary and due to our data association process only a single landmark is observed at a given instant in time. The stereo camera noise has been experimentally demonstrated to be approximately Gaussian in the range of interest, which justifies the use of a Kalman filter.

Visual odometry observations are already relative to the Earth frame (Earth frame is fixed at the initial position of the vehicle) and consequently only a simple linear observation model is required for odometry updates. However, care must be taken to ensure that the roll, pitch and yaw angles remain in the range $[-\pi \ \pi]$ and that the innovation is computed correctly around the $\pm\pi$ boundary. Landmark observations are considerably more complicated due to the fact that landmarks

are observed in the camera frame but are stored in the Earth frame. From the current estimate of the six vehicle states $\hat{v}_{k|k-1}$ we compute the transformations ${}^{E}T_V$ and ${}^{V}T_E$ which describe the estimated transformation from the vehicle to world, and world to vehicle respectively. The predicted observation $\hat{z}_{i,k}$ for a single landmark m_i is then given by:

$$\hat{z}_{i,k} \quad = \quad h(\hat{v}_{k|k-1}, \hat{m}_{i,k-1}) = {}^{V}T_E \left[\begin{array}{c} \hat{m}_{i,k-1} \\ 1 \end{array} \right] \quad (14)$$

The difference between the modelled and predicted measurements $z_k - \hat{z}_{i,k}$ is known as the innovation vector, and is the basis of minimisation. Accordingly, the estimate of the state vector and its corresponding covariance matrix are updated by computing:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + W_k[z_k - \hat{z}_{i,k}] \quad (15)$$

$$P_{k|k} = P_{k|k-1} - W_k S_k W_k^T \quad (16)$$

where the innovation covariance matrix S_k is given by:

$$S_k = \nabla h P_{k|k-1} \nabla h^T + R_k \quad (17)$$

and the optimal Kalman gain W_k is given by:

$$W_k = P_{k|k-1} \nabla h^T S_k^{-1} \quad (18)$$

where ∇h is the Jacobian (linear approximation) of $h()$ evaluated at the estimate $\hat{x}_{k|k-1}$. The additive, zero mean uncorrelated Gaussian observation noises o_k which affect the observations and have covariance R_k are constant and are specified in a configuration file.

When a new landmark is observed the state and covariance is expanded using the augmented state approach, which requires the implementation of an inverse landmark observation model. In this model the estimated position of a new landmark m_j is given by a function $g(\hat{v}_{k|k}, z_k)$ which is essentially the inverse of $h(\hat{v}_{k|k-1}, \hat{m}_{i,k-1})$:

$$\hat{m}_{j,k} \quad = \quad g(\hat{v}_{k|k}, z_k) = {}^{E}T_V \left[\begin{array}{c} z_k \\ 1 \end{array} \right] \quad (19)$$

and the addition of the landmark to the state vector is trivial:

$$\hat{x}_{k|k} = \left[\begin{array}{ccc} \hat{v}_{k|k} & \hat{m} & g(\hat{v}_{k|k}, z_k) \end{array} \right]^T \quad (20)$$

Computing the Jacobian ∇g of g() at $\hat{v}_{k|k}$ the new landmark is added to the covariance matrix by computing:

$$P_{k|k} = \left[\begin{array}{cccc} P_{xx} & P_{x\Theta} & P_{xm} & P_{xx}\nabla g^T \\ P_{x\Theta}^T & P_{\Theta\Theta} & P_{\Theta m} & P_{x\Theta}\nabla g^T \\ P_{xm}^T & P_{\Theta m}^T & P_{mm} & P_{xm}\nabla g^T \\ \nabla g P_{xx} & \nabla g P_{x\Theta}^T & \nabla g P_{xm}^T & \nabla g P_{xx}\nabla g^T + R_k \end{array} \right] \quad (21)$$

where P_{xx} and $P_{\Theta\Theta}$ are the $3x3$ sub matrices of the covariance matrix which correspond to the x, y, z position, and roll, pitch, yaw angles respectively, and R_k is the covariance of the additive, zero mean uncorrelated Gaussian observation noises.

IV. RAUCH-TUNG-STRIEBEL SMOOTHER

The Kalman filter uses all measurements up to the current iteration to estimate the state at the current iteration. In contrast, the Rauch-Tung-Striebel (RTS) smoother is a post-processing filter which applies a forward and backward pass filter to the complete set of measurements. The output of the RTS has been shown to improve the accuracy of the stochastic map solution as well as providing smoother trajectories [10]. The RTS smoother was designed for fixed size state vectors and consequently the RTS fixed-interval smoother was adapted to work with the growing stochastic map by fixing the size of the state vector to the size of the stochastic map on the last iteration. Therefore, once the Kalman filter has finished, we set k equal to the last iteration $(n-1)$ and work backwards until we reach the starting point at $k = 1$. In each iteration the predicted smoother state is computed by:

$$\hat{\hat{x}}_{k+1|k} = f(\hat{x}_{k|k}, u_k) \quad (22)$$

and the predicted covariance matrix by:

$$\hat{\hat{P}}_{k+1|k} = \nabla f P_{k|k} \nabla f^T + Q_k \quad (23)$$

Then, the smoother gain matrix $J(k)$ is computed by:

$$J_k = P_{k|k} \nabla f^T \hat{\hat{P}}_{k+1|k}^{-1} \quad (24)$$

and, finally the filtered state and covariance are given by:

$$\tilde{x}_{k|k} = \hat{x}_{k|k} + J_k \left(\tilde{x}_{k+1|k+1} - \hat{\hat{x}}_{k+1|k} \right) \quad (25)$$

and

$$\tilde{P}_{k|k} = P_{k|k} + J_k \left(\tilde{P}_{k+1|k+1} - \hat{\hat{P}}_{k+1|k} \right) J_k^T \quad (26)$$

where the smoother is initialised so that $\tilde{x}(n|n) = \hat{x}(n|n)$ and $\tilde{P}(n|n) = P(n|n)$. Therefore, the final output of the system is a aligned sequence of partial reconstructions that together represent a large-scale 3D acquisition of the surveyed area.

V. EXPERIMENTAL RESULTS

In the experiment we have simulated a virtual $3D$ scenario of an underwater environment composed by a $3D$ surface which can be either introduced by an user or imported. The user can select a real underwater (or aerial) image which is stuck on the $3D$ surface conforming a virtual $3D$ scene but yet with a real texture. Note that the texture is deformed according to the shape of the surface. Then the user is asked to introduce the trajectory of the vehicle in 6 degrees of freedom. The algorithm interpolates the introduced trajectory generating the navigation data. At every vehicle position, the two virtual cameras render both images by means of ray tracing simulating image acquisition. Zero-mean Gaussian noise ($\sigma = 0.01 rad.$) has been added to vehicle orientation. Velocity may suffer error propagation and hence a biassed Gaussian noise ($\mu = 0.05 m/s, \sigma = 0.08 m/s$) has been considered.

The experiment shows how the SLAM approach is able to readjust vehicle trajectory even in the presence of large bias (see Fig. 1). Fig. 2 shows the interpolated and resampled
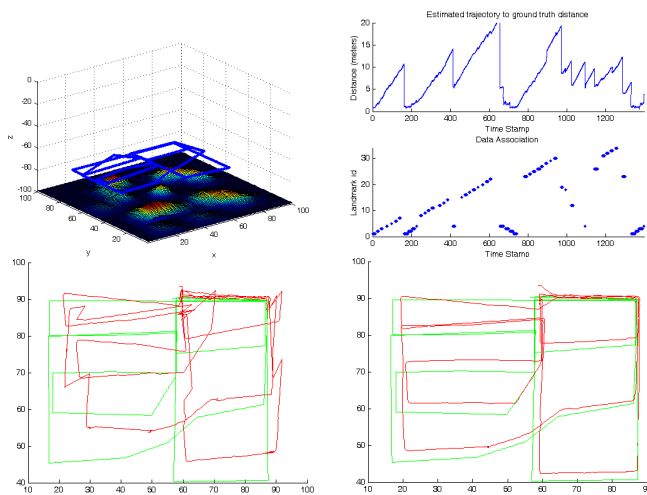
Fig. 1. SLAM Results (from left to right and up to left). Fig.a: Ground truth $3D$ surface and vehicle trajectory in 6-DOF; Fig.b: Discrepancy of the estimated vehicle trajectory with respect to ground truth. The figure shows how the discrepancy is reduced while any landmark is re-observed during the journey; Fig.c: the filtered trajectory (red) compared to ground truth (green) obtained by the EKF-SLAM algorithm (trajectory jumps are not due to filter inconsistency but to the fact that we are detecting few landmarks to simplify data association); and Fig.d: the smoothed trajectory (red) compared to ground truth (green) obtained by the RTS.

Fig. 2. Alignment of the 1398 local maps corresponding to 93,275 $3D$ points (from left to right and top to bottom): The surface aligned by the EKF-SLAM: 3D view (Fig.a) and Top view (Fig.b); The discrepancy to ground truth shown in Fig. 1 is $\mu = 4.28m$. and $\sigma = 2.80m$. The surface aligned by the RTS-smoother: 3D view (Fig.c) and Top view (Fig.d). The discrepancy to ground truth shown in Fig. 1 is $\mu = 0.84m$. and $\sigma = 0.78m$.

surface obtained by the EFK-SLAM algorithm and by the post-processing of the RTS smoother, demonstrating qualitatively and quantitatively that our approach obtains an accurate alignment of the $3D$ surfaces even in the presence of large noises and biases.

VI. CONCLUSIONS

This paper has presented a novel EKF-SLAM system which should be capable of operating in real-time on a slow moving AUV equipped with a calibrated stereo-vision system. The map constructed by the system contains a set of landmark submaps which are described by a set of observed 3D points and their associated SIFT/SURF descriptors. Consequently, the final global map represents a large-scale 3D reconstruction of the seabed consisting of hundreds of aligned partial reconstructions. By utilising such distinctive landmark descriptions the system is also able to employ maximum likelihood data association based on the 3D points and 2D descriptors with little risk of incorrectly associating landmarks. The set of consecutive 3D point correspondences are also registered to measure their compatibility and determine the relative rotation and translation between the first and current observation of the landmark. These relative motion estimates are then used to perform landmark-based visual odometry. To the best of our knowledge, this paper is the first that proposes EKF-SLAM to deal with the 3D reconstruction of the seabed using only vision. Analysis of the computational cost of the algorithm for the virtual scenarios suggests real-time operation should be achievable, however, it would be wise to optimise the feature extraction and incorporate submapping to bound computation time for larger scale maps. Currently we are
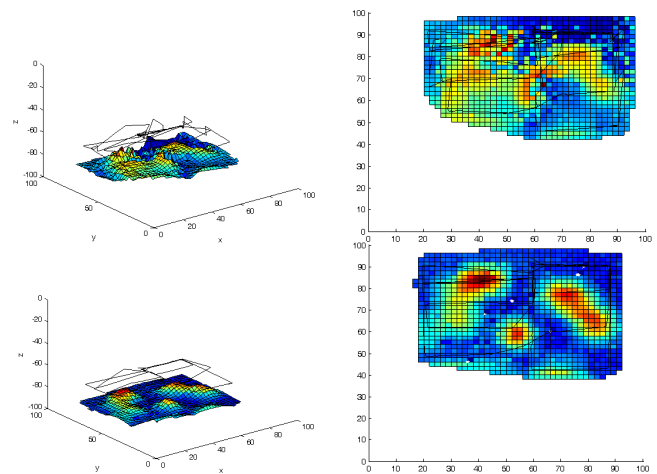
improving the visual odometry by introducing overlapping landmarks and active navigation techniques and investigating how this influences the robustness of the data association and map sparseness.

REFERENCES

[1] H. Singh, J. Howland, and O. Pizarro, "Advances in large-area photo-mosaicking underwater," *IEEE Journal of Oceanic Engineering*, vol. 29, no. 3, p. 872-886, 2004.

[2] R. M. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation for autonomous underwater vehicles," *IEEE J. Oceanic Engineering*, 2007.

[3] I. Mahon and S. Williams, "Slam using natural features in an underwater environment," in *IEEE Control, Automation, Robotics and Vision Conference*, vol. 3, December 2004, pp 2076–2081.

[4] J. M. Saez, A. Hogue, F. Escolano, and M. Jenkin, "Underwater 3d slam through entropy minimization," in *IEEE International Conference on Robotics and Automation*, May 2006, pp. 3562–3567.

[5] O. M. Mozos, A. Gil, M. Ballesta, and O. Reinoso, "Interest point detectors for visual slam," *LNAI 4788*, pp. 170–179, 2007.

[6] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR'03)*, June 2003, p. 257-263.

[7] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer-Verlag, 2003.

[8] X. Armangue and J. Salvi, "Overall view regarding fundamental matrix estimation," *Image and Vision Computing*, vol. 21, no. 2, pp. 205–220, February 2003.

[9] A. Mian, M. Bennamoun, and R. Owens, "A novel representation and feature matching algorithm for automatic pairwise registration of range images," *Journal of Computer Vision*, vol. 66, no. 1, pp. 19–40, 2006.

[10] I. Tena-Ruiz, S. Raucourt, Y. Petillot, and D. Lane, "Concurrent mapping and localization using sidescan sonar," *IEEE Journal of Oceanic Engineering*, vol. 29, no. 2, pp. 442–456, 2004.