Underwater Mosaicing and Trajectory Reconstruction using Global Alignment

Nuno Gracias and José Santos-Victor
Instituto Superior Técnico & Instituto de Sistemas e Robótica
Av. Rovisco Pais, 1049–001 Lisboa Codex, Portugal

Abstract— **This paper deals with the problem of constructing high quality mosaics of the sea bed. It focuses on the use of long image sequences with time-distant superpositions, such as the ones arising from loop trajectories or zig-zag scanning patterns. An algorithm is presented for the simultaneous creation of mosaics and the estimation of the camera trajectory.**

The method comprises three major stages. The first stage consists of the sequential estimation of the image motion, using a reduced image motion model. The set of resulting consecutive homographies is cascaded, in order to infer the approximate topology of the camera movement. The topology information is then used to predict the areas where there is image overlap resulting from non-consecutive images.

Secondly, a motion refinement is performed, by iteratively executing the following two main steps. (1) Point correspondences are established between non-adjacent pairs of images that present enough overlap. (2) The topology is refined, by searching for the set of homographies that minimizes the overall sum of distances in the point matches.

The final stage of the algorithm consists of estimating the set of homographies and a world plane description that best fit the observation data. As the main concern is attaining high registration accuracy, a general parameterization of the homographies with 6 DOF for the pose is used, which is capable of modelling the effects of wave–induced general rotation and translation.

The overall method is fully automatic in the sense it does not require human intervention at any of the stages, apart from the beforehand specification of the most adequate motion model for the first stage.

We present results obtained from shallow water image sequences acquired by a ROV. The images present some of the common difficulties of underwater mosaicing, such as non planar sea-bottom, moving objects and severe illumination changes. This sequence also serves to illustrate the robustness and good performance of the presented algorithm.

I. Introduction

Over the last decade, the topic of video mosaicing has recieved considerable attention from the underwater vision community, mainly due to its large application in site exploration and autonomous navigation. A contributing factor has been the generalized availability of video cameras onboard modern underwater vehicles.

The work described here forms part of the research done by the European Project NARVAL[1]. The main scientific goal of this project is the design and implementation of reliable navigation systems for mobile robots in unstructured environments. Also, strong emphasis is put on the ability to navigate without resorting to global positioning methods. The approach followed in this paper, of constructing large accurate mosaics for posterior navigation reference, constitutes an important step towards that goal.

The method comprises three major stages. Firstly, the image motion is computed in a sequential manner, using a simple image motion model, in order to create a set of consecutive homographies. These homographies are cascaded in order to infer the approximate topology of the camera movement. The topology information is then used to predict the areas where there is image overlap resulting from non-consecutive images. This overlap is valuable in the sense it can be used to further refine the motion estimation and the final mosaic.

Secondly, a motion refinement is performed, by iteratively executing the following two main steps. (1) Point correspondences are established between non-adjacent pairs of images that present enough overlap. This is a time consuming operation but is alleviated by the fact that a prior information exists on the location of the image correspondences, computed at the first stage. (2) The topology is refined, by searching for the set of homographies that minimize the overall sum of distances in the point matches.

Finally, a global minimization is carried out, using the most general 6-degree of freedom motion model and a cost function based on the errors of the point matches between all the images. The minimization process searches for the best set of pose parameters (describing the 3D positions and orientations of the camera) and for the best fitting description of the world plane.

The basic assumptions behind this method are that the sea bottom is essentially flat, is static and exhibits no illumination change. This is seldom the case in underwater mapping applications, especially in shallow waters. However, the use of robust estimation over point feature matching greatly alleviates these assumptions and allows for the correct recovery of image motion. As an illustrative example, Figure 1 contains two consecutive frames of an image sequence used in this work, which were successfully matched.

This paper builds upon previous work on mosaic and pose estimation [8], [9]. The main contribution lies in the fact that, by incorporating the single world plane description on the overall estimation problem, the trajectory reconstruction becomes possible without the need to explicitly define a world referential associated with the mosaic. Another contribution is on the formulation of the mosaic creation as an estimation problem where all the

Email: {ngracias,jasv}@isr.ist.utl.pt. The work described in this paper has been supported by the Portuguese Foundation for the Science and Technology PRAXIS XXI BD/13772/97, and NARVAL Esprit-LTR Proj. 30185.

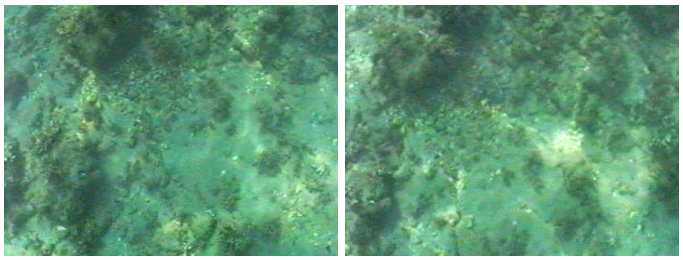

Fig. 1. Two sequential frames, illustrating the difficulty of the matching process for images of very shallow waters, where the lighting conditions change rapidly.

degrees of freedom arising from the mosaic geometry are taken into account. These are conveniently parametrized as geometrically meaningful entities: pose parameters and world plane description.

Some of the topics presented here relate to the work of a large group of authors. One of the early references to the idea of using mosaics as visual maps is the work of Zheng[21], where panoramic representations were applied to route recognition and outdoor navigation. Recently, Kelly[13] has addressed the feasibility and implementation issues of using large mosaics for robot guidance, predicting a large impact of these techniques on industrial environments. In the context of underwater robotics, the use of mosaicing techniques for navigation is a topic of increasing research interest [15], [5], [6]. Xu[20] investigated the use of seafloor mosaics, constructed using temporal image gradients, in the context of concurrent mapping and localization and for real-time applications. Comparative results on vehicle positioning and mosaicing for long image sequences are reported by Negahdaripour in [14], where calibrated testing of the algorithm presented in this paper is included.

Considering the topic of global registration, several approaches have been proposed using topology inference of neighboring frames [17], [12], and restricted parameterizations for the projection matrices [2]. The main differences of our approach are twofold: (1) the parameterization of the homographies with complete and meaningful 3D pose parameters and, (2) the inclusion of the unknown single world plane condition.

This paper is organized as follows. Section II gives a brief presentation of some of the geometric concepts used throughout the text. The initial estimation of the homographies is described in Section III. Next, in Section IV, the iterative scheme of matching / optimization is presented, where point correspondences between new pairs of images are searched for, along with optimization to refine the overall topology. The following section describes the final trajectory estimation, using the most general motion model, which allows for the precise registration of images over a common world plane. Some selected results are presented in Section VI, in the form of final mosaics and VRML renderings of the camera path, which testify to the good performance of the overall algorithm. Finally Section VII presents some discussion and conclusions.

II. Geometrical Background

In this section we will assume the reader to be familiar with the basic concepts and properties of projective geometry[3]. The camera model used in this paper is the standard pin-hole model, which performs a linear projective mapping of the 3D world into the image frame. We also assume that the camera calibration has been performed beforehand, and that the 3×3 matrix K containing the intrinsic parameters has been estimated[19], [11].

A collineation in the projective space of dimension 2 is commonly referred to as a planar transformation or homography, and is represented by a 3×3 matrix defined up to scale. It establishes a one-to-one relation between corresponding points over two images of the same 3-D plane. The computation of a planar transformation requires at least four pairs of corresponding points. In the case of more than four correspondences, a straight-forward least-squares linear estimation can be accomplished[7].

The final stage of our algorithm, described in Section V., deals with the estimation of the camera path using all of the available degrees of freedom. In this stage, the minimization algorithm requires an initialization with a rough estimate of the 3-D camera path. In order to provide this, we have made use of the decomposition described in [4], which relates the homography matrix H with the camera rotation, translation and the world plane which induces the homography. The decomposition is the following

$$H_{21} = K \left(R_{21} + n_1 \frac{t^T}{d_1} \right) K^{-1} \qquad (1)$$

where R_{21} and t are, respectively, the 3×3 rotation matrix and the 3×1 translation vector relating the two 3-D camera frames. The world plane is accounted for through the unitary vector n_1, containing the outward plane normal expressed in the camera 1 coordinates, and the distance d_1 of the plane to the first camera center, measured along the optical axis.

The problem of recovering the motion parameters from an homography for an intrinsically calibrated camera is discussed in-depth by Faugeras[4]. In the most general case there are eight different sets of solutions. However, only two are feasible if one considers the world plane to be non-transparent. These two solutions can conveniently found by means of the SVD decomposition of $M_{21} = K^{-1} H_{21} K$, as presented by Triggs[18].

III. Initial Motion Estimation

The first part of the algorithm consists on the sequential estimation of inter-frame homographies.

For each image I_k, a set of point features, corresponding to textured areas, is extracted using the Harris corner detector[10]. The features are then matched directly on the following image I_{k+1}, using correlation-based procedure, from which two lists of corresponding points are obtained.

Due to the error prone nature of the matching process, it is likely that a number of points will be mismatched. Therefore, a robust estimation technique is required to

filter out matching outliers, and estimate the homography $H_{k,k+1}$ that relates the coordinate frames of I_k and I_{k+1}. In this paper, a variant of LMedS with random sampling[7] was used for minimizing the median of sum of the square distances,

$$\begin{aligned}\underset{i}{\text{med}}(&d^2({}^{(k)}x_i, T_{k,k+1} \cdot {}^{(k+1)}x_i) \\ &+d^2({}^{(k+1)}x_i, T_{k,k+1}^{-1} \cdot {}^{(k)}x_i))\end{aligned} \tag{2}$$

where $d(\cdot,\cdot)$ stands for the point-to-point Euclidean distance, and ${}^{(k)}x_i$ is the location of the i^{th} feature extracted from image I_k and matched with ${}^{(k+1)}x_i$ on I_{k+1}. The minimizing algorithm works by randomly sampling sets of points with the minimum number of matches required for the linear computation of H. The set that minimizes the cost function is selected and the homography is re-estimated using simple least-squares with all matches whose distance are below a specified error limit. The image matching is considered successful if the number of matches used in the final least-squares estimation is sufficiently high.

In order to speed up the initial matching process, the computed homography for the current pair of images is used to restrict the correlation search over the next pair. If, after the random sampling LMedS, the image matching is not successful then the process is repeated with larger correlation areas.

In underwater vision applications it is very common for the image acquisition rate to be high when compared to the camera motion. This results in very high overlapping between consecutive frames that convey redundant information. In the work presented, a selection criteria was used to selected a subset of frames, thus reducing the memory and processing requirements for the next stages. The frames are selected such that their superposition is the smallest above a given minimum acceptable overlap percentage. This threshold insures the ability of the selected images to be correctly matched, and is chosen based on the results of preliminary matching trials.

IV. Iterative Motion Refinement

After the initial motion estimation step, every image in the reduced sequence can be spatially related with any other image, by appropriately cascading the homographies. Possible overlap between non-consecutive images can be predicted, and used for searching new image matches.

In this stage, the topology is refined by performing iterative steps of image matching and global optimization. The image matching part is conducted over overlapping frames, and is similar to what was described above. If new matches are found, then the topology is re-estimated by means of a global optimization procedure. This procedure uses a reduced representation for the camera motion, based on 3 parameters per image (2D translation and rotation), that implicitly assumes the camera is facing the ground and keeps a constant distance. The reason behind the choice of a simpler motion model for the first two stages of the algorithm, has to do with the effectiveness of the topology inference. Alternatively, one could have resorted to the use of the most general 8-parameter homographies, as this is the only model that can cope with general perspective distortion *and* allow for fast linear estimation. However, it has more degrees of freedom than required. Consequently, small errors in the initial inter-frame motion estimation tend to quickly accumulate, and make it impossible to infer the neighboring relations among non-consecutive frames.

The cost function to be minimized is the sum of distances between each correctly matched point and its corresponding point after being projected onto the same image frame, *i.e.*,

$$\begin{aligned}F(X,\Theta) = \sum_{i,j}\sum_{n=1}^{N_{i,j}} [&d^2\left(x_n^i, H(\Theta_i,\Theta_j)\cdot x_n^j\right) \\ &+d^2\left(x_n^j, H^{-1}(\Theta_i,\Theta_j)\cdot x_n^i\right)]\end{aligned}$$

where $N_{i,j}$ is the number of correct matches between frame i and j, and $H(\Theta_i,\Theta_j)$ is the homography constructed using the motion parameter vectors Θ_i and Θ_j. The minimization is carried out using a non-linear least squares algorithm[16]. The cycle of matching and topology refinement is executed until no new image pairs can be matched.

In order to speed-up the optimization procedure (and, thus, the motion refinement cycle time), a sub-mosaic aggregation scheme was implemented and tested. Under this scheme the complete sequence is initially divided into sets of consecutive images to form small rigid sub-mosaics. Inside each sub-mosaic the homographies are considered static and only the inter-mosaic homographies are taken into account in the optimization algorithm. This reduced parameter scheme significantly improves the speed of evaluating the cost function and does not affect the capability of inferring the appropriate trajectory topology.

V. Trajectory Estimation

The main objective of the final stage of the algorithm is attaining a highly accurate registration. A more general parameterization for the homographies is therefore required, capable of modelling the warping effects caused by wave-induced general camera rotation and changes on the distances to the sea floor. Bearing this in mind, a parameterization was chosen in which all the camera pose 6 degrees of freedom are explicitly taken into account. This has also the additional advantage of allowing the camera path to be recovered during the process.

Furthermore, the estimation of the homographies for this model does not impose, *per se*, the condition of a single world plane from which the homographies are induced. This condition can be imposed by augmenting the overall estimation problem with additional parameters that describe the position and orientation of the world plane. The world plane description must then be included on the parameterization of the homographies.

The adopted parameter scheme is the following. One of the camera frames is chosen (usually the first) as the origin for the 3-D referential, where the optical axis is coincident with the referential Z-axis. The world plane is parameterized with respect to this frame by 2 angular values that define its normal. As the trajectory and plane reconstruction can only be attained up to an overall scale factor, this ambiguity is removed by setting the plane distance to 1 metric unit[1], measured along the Z-axis. The homography relating frames i and j is

$$H(\Theta_p, \Theta_i, \Theta_j) = K \cdot \left(R\left(\Theta_i\right) + n\left(\Theta_p\right) \cdot t^T\left(\Theta_i\right)\right) \cdot$$

$$\cdot \left(R\left(\Theta_j\right) + n\left(\Theta_p\right) \cdot t^T\left(\Theta_j\right)\right)^{-1} \cdot K^{-1}$$

where Θ_i and Θ_j are pose vectors containing 3 rotation angles and 3 translation values with respect to the reference frame, $R\left(\Theta_i\right)$ and $R\left(\Theta_j\right)$ are rotation matrices, $t^T\left(\Theta_i\right)$ and $t^T\left(\Theta_j\right)$ are the translation components, and $n\left(\Theta_p\right)$ is the 3-vector with the outward plane normal. The pose vector for the reference camera is the null 6-vector.

The cost function is similar to the one previously used in the iterative motion refinement, where the distances between matched points are measure in their respective image frames, and summed over all pair of correctly matched images, *i.e.*,

$$F(X, \Theta) = \sum_{i,j} \sum_{n=1}^{N_{i,j}} [d^2\left(x_n^i, H(\Theta_p, \Theta_i, \Theta_j) \cdot x_n^j\right)$$

$$+ d^2\left(x_n^j, H^{-1}(\Theta_p, \Theta_i, \Theta_j) \cdot x_n^i\right)]$$

For a set of M images, the total number of parameters to be estimated is $(M-1) \times 6 + 2$.

The initialization values for the complete parameter set are computed using Equation (1). As there are two valid solutions for the decomposition of the homographies relating each frame with the reference frame, the solutions are chosen such that the variance of the world plane normals is minimized. The considered world plane normal is the average of the selected set.

As before, the cost function is minimized using non-linear least squares.

VI. Results

Extensive testing was conducted in order to evaluate the performance of the algorithms. The image sequences for the results shown in this paper were acquired by a Phantom ROV during a NARVAL Project sea trial, in Villefranche-sur-mer in France. The ROV is equipped with a Sony pan-and-tilt camera, facing the sea floor. It is mounted in the center of a spherical glass housing which induces very little image distortion.

[1]If additional information is available on the real distance to the sea floor (for example, from an altimeter), then it can be straightforwardly used here.

Fig. 5. Mosaic detail of the same region using two different rendering methods, the *median* (left) and the *closest* operator (right).

The camera calibration was performed under water using a standard calibration grid and the method described by Heikkilä in [11].

The first sequence refers to a flat sandy area, fully surrounded by algae. During the acquisition, the vehicle was manually driven to follow a zig-zag trajectory that covered most of the area. The sequence comprises 1000 images, corresponding to 400 seconds of video. After the initial matching, a set of 129 images was selected using the criterion of minimal overlap above 50%, which resulted in an average overlap of 54.4%. The mosaic obtained from the last stage of the algorithm, in shown in Figure 2. It was created by choosing the contributing points which were located the closest to the center of their frames. This rendering method is useful when creating the mosaics for navigation and mosaic-based localization. For the cases where the illumination changes are *not* strong, it compares favorably with other commonly used rendering methods, such as the average or the median. This is due to the fact that it better preserves the textures and minimizes the effects of barrel distortion, which tends to be larger near the image borders. A small section of the rendered mosaic is displayed in Figure 3 along with one of the original frames for the same area. The quality of the registration can be assessed from the fact that the visual features (such as small algae leafs) are not disrupted along the visible boundaries of the contributing images.

The second set of images was captured while the vehicle followed a circular trajectory of several turns around a square shaped rock. It comprises 895 images from which 85 where selected using a 60% minimum overlap. The resulting mosaic for the complete algorithm is presented in Figure 4. This sequence was acquired in very shallow waters, of less than 2 meters in depth, where the effect of sunshine refracted from the surface is quite noticeable. Clearly, under these lighting conditions, a rendering operator such as the median is required, which is capable of removing transient data from the set of intensity contributions for the mosaic. For comparison purposes, Figure 5 illustrates the visual differences in applying the *median* and the *closest* operators.

The recovered 3D camera paths for the two sequences are illustrated in Figure 6.

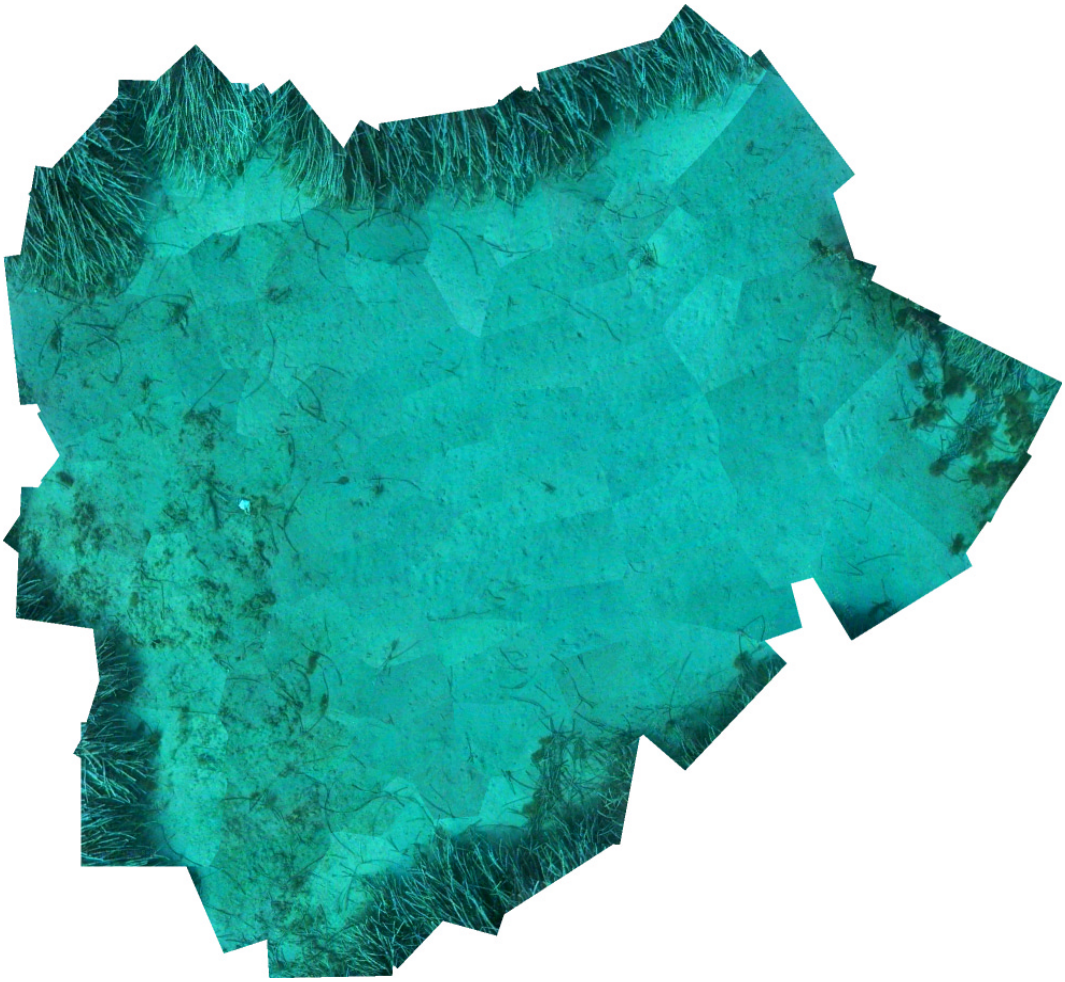

Fig. 2. Final mosaic for the first image sequence. It was created using 129 images selected from the original set of 1000 and rendered with the *closest* operator. The seafloor area covered is approximately 42 m^2.

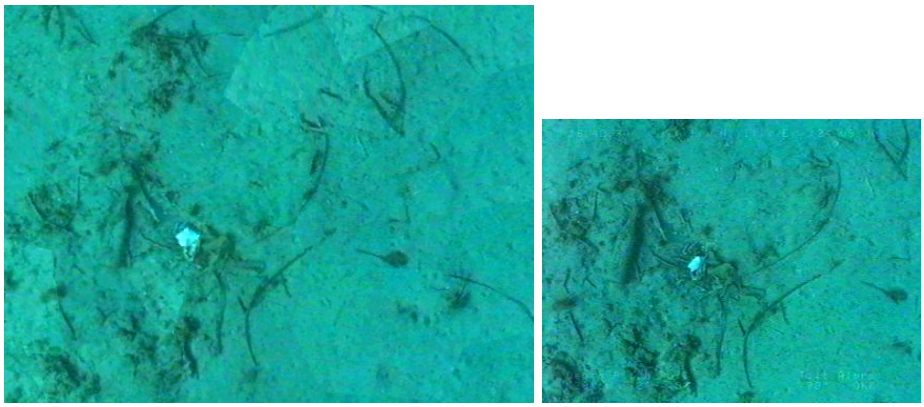

Fig. 3. Area detail of the mosaic for the first sequence (left), and one of the original images (right).

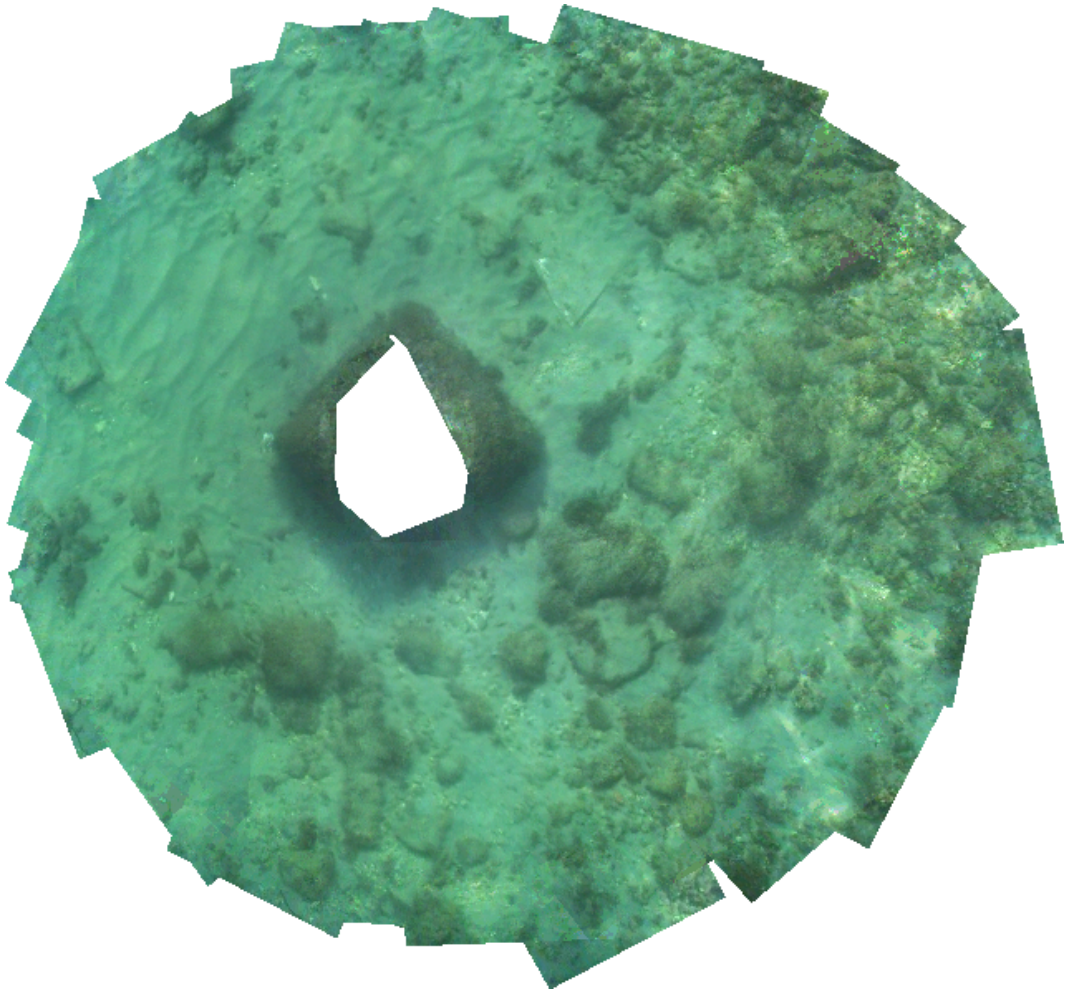

Fig. 4. Final mosaic for the second sequence, created using 85 frames selected from the original set of 895, and the *median* rendering operator.

VII. Conclusions

This paper contributes to the area of underwater mosaicing by presenting an integrated approach to the problem of pose estimation and mosaic construction. Also, it provides the means of finding a geometric description of sea-bottom plane, and is able to reconstruct the camera path without the specification of a world referential.

The importance of this work is emphasized by fact that it can be used to extend the navigation autonomy of camera-equipped AUVs, in two main aspects. (1) By making use of non-consecutive image overlaps, it provides a precise position and motion estimation when compared with other common sensing modalities such as sonar, compass and gyroscopes. (2) It enables the creation of high accuracy mosaics that can be used as reliable maps for posterior localization and servoing.

References

[1] NARVAL Consortium. NARVAL – Navigation of Autonomous Robots via Active Environmental Perception, Esprit–LTR Project 30185 http://gandalf.isr.ist.utl.pt/narval/index.htm.

[2] K. Duffin and W. Barrett. Globally optimal image mosaics. In *Graphics Interface*, pages 217–222, 1998.

[3] O. Faugeras. *Three Dimensional Computer Vision*. MIT Press, 1993.

[4] O. Faugeras and F. Lustman. Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 2(3):485–508, September 1988.

[5] Stephen Fleischer. *Bounded–Error Vision–Based Navigation of Autonomous Underwater Vehicles*. PhD thesis, Stanford University, California, USA, May 2000.

[6] R. Garcia, J. Batlle, X. Cufí, and J. Amat. Positioning an underwater vehicle through image mosaicking. In *Proc. International Conference on Robotics and Automation (ICRA)*, pages 2779–2784, Seoul, Korea, May 2001.

[7] N. Gracias. Application of robust estimation to computer vision: Video mosaics and 3–D reconstruction. Master's thesis, http:// www.isr.ist.utl.pt/labs/vislab/thesis, Lisbon, Portugal, April 1998.

[8] N. Gracias and J. Santos-Victor. Underwater video mosaics as visual navigation maps. *Computer Vision and Image Understanding*, 79(1):66–91, July 2000.

[9] N. Gracias and J. Santos-Victor. Trajectory reconstruction with uncertainty estimation using mosaic registration. *Robotics and Autonomous Systems*, 35:163–177, July 2001.

[10] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings Alvey Conference*, pages 189–192, Manchester, UK, August 1988.

[11] J. Heikkilä and Olli Silvén. A four–step camera calibration procedure with implicit image correction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997. IEEE Computer Society Press.

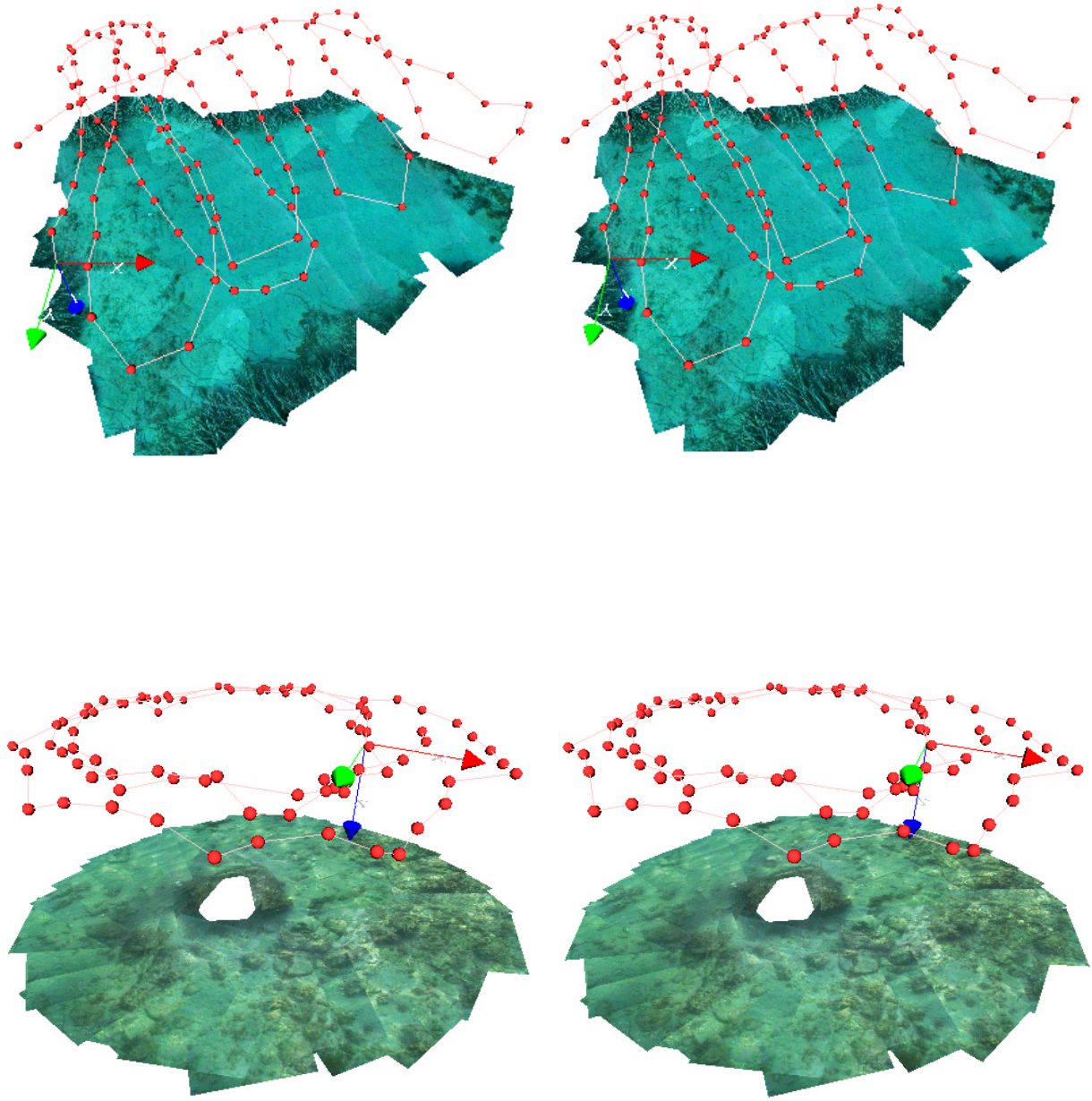

Fig. 6. VRML rendition of the camera path and mosaic for the two sequences The world referential is illustrated by the system of axis, which is coincident with the first camera frame. The views are arranged for crossed eye fusion.

[12] E. Kang, I. Cohen, and G. Medioni. A graph–based global registration for 2D mosaics. In *Proc. of the 15th International Conference on Pattern Recognition*, Barcelona, Spain, 2000.

[13] A. Kelly. Mobile robot localization from large scale appearance mosaics. *International Journal of Robotics Research (IJRR)*, 19(11), 2000.

[14] S. Negahdaripour and P. Firoozfam. Positioning and image mosaicing of long image sequences; Comparison of selected methods. In *Proc. of the IEEE OCEANS 2001*, Honolulu, Hawai, USA, November 2001.

[15] S. Negahdaripour, X. Xu, A. Khamene, and Z. Awan. 3D motion and depth estimation from sea-floor images for mosaic-based positioning, station keeping and navigation of rovs/auvs and high resolution sea–floor mapping. In *Proc. IEEE/OES Workshop on AUV Navigation*, Cambridge, MA, USA, August 1998.

[16] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1988.

[17] H. Sawhney, S. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *Proc. European Conference on Computer Vision*. Springer-Verlag, June 1998.

[18] B. Triggs. Autocalibration from planar scenes. In *Proc. of the European Conference on Computer Vision*, pages 89–105, Freiburg, Germany, June 1998.

[19] R. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV camera and lenses. *IEEE Journal of Robotics and Automation*, RA-3(4):323–344, 1987.

[20] X. Xu. *Vision–based ROV System*. PhD thesis, University of Miami, Coral Gables, Miami, May 2000.

[21] J. Zheng and S. Tsuji. Panoramic representation for route recognition by a mobile robot. *International Journal of Computer Vision*, 9(1):55–76, October 1992.