3D Reconstruction of Natural Underwater Scenes Using the Stereovision System IRIS

V. Brandou, A. G. Allais, M. Perrier, E. Malis, P. Rives, J. Sarrazin, P. M. Sarradin

***Abstract*—The aim of this study is to propose a 3-dimension reconstruction method of small-scale scenes improved by a new image acquisition method for quantitative measurements. A stereovision system is used to acquire images in order to obtain several shots of an object, at regular intervals according to a predefined trajectory. A complete methodology of 3D reconstruction is exposed to perform a dense 3D model with texture mapping. A first result on natural images collected with the stereovision system during sea trials has been obtained.**

***Index Terms*—Underwater imaging; stereovision system; visual servoing; dense disparity map; 3D reconstruction.**

I. Introduction

AFTER its revolutionary contribution in many fields such as medical practice, 3D imagery finds more and more applications in other domains, like video surveillance, due to the improvement of acquisition techniques, computer's performance and exploration of new methods of calculation.

In the field of deep-sea study, the aim of 3D imagery is to generate the 3D reconstruction of underwater natural small-scale scenes and thus complete sample analyzes and physicochemical measurements by a 3D visual observation that enables quantitative 3D measurements.

To determine the 3D model of an object for metric measures, we have to know camera parameters used to capture the image sequence. Moreover, if images can be collected at regular intervals, the number of unknown variables in the 3D reconstruction procedure is reduced, since extrinsic camera parameters (position and orientation) are known. Then the accuracy of the final 3D reconstruction is increased, and the processing can be faster. Consequently, the images cannot be collected by freely moving a camera around the object. Our experimental conditions are very specific; the main constraint is that the system used to collect images must be manipulated at very important depths, up to 6000 meters by an underwater vehicle (ROV) positioned on the sea floor. So, an array of cameras cannot be used since the system must be compact to be transported in the vehicle basket. Finally, the image acquisition is performed with a stereovision system operated by a manipulator arm. The image acquisition method that we propose enables us to know extrinsic camera parameters by following a specific trajectory defined by the geometry of a stereo rig. Indeed, the trajectory is generated by the displacement of one camera onto the position of the other one by visual servoing. With this method, we can register images at regular intervals directly linked to the geometry of the stereo rig.

Then, the 3D model of the underwater object is calculated with the collected images. Robust interest points are extracted and matched allowing to estimate fundamental matrix, which is used to rectify images in order to obtain a dense disparity map. The final result is a dense 3D reconstruction with texture mapping that enables metric measures.

The first part of this paper focuses on image acquisition method and results obtained during see trials with the practical implementation of the stereovision system. In the second part, a 3D reconstruction methodology are presented and first results on stereo images of underwater scene.

II. Image Acquisition Method

Image acquisition is of great importance in the process of 3D reconstruction since the method used to collect images affect directly the final results of reconstruction. This is why we have developed, implemented and tested a new method to collect stereo images.

In this part we expose two different ways to generate specific trajectories with the stereovision system which depend on the capabilities offered by the underwater robot equipped with a manipulator arm: visual servoing or pre-programmed trajectories of the robotic arm. Then we will present the implementation and the validation of the method, and the final test in deep water.

Manuscript received March 30, 2007.

V. Brandou is a research student in the Underwater Systems Department at IFREMER, centre de Toulon, France (phone: +33 (0)493304418 e-mail: vincent.brandou@ifremer.fr).

A. G. Allais is an engineer in the Underwater Systems Department at IFREMER, Toulon, France (e-mail: anne.gaelle.allais@ifremer.fr).

M. Perrier is a research scientist and the director of the *Positioning, Acoustics, Vision and Robotics* division of IFREMER's Underwater Dept., Toulon, France (e-mail: michel.perrrier@ifremer.fr).

E. Malis is a research scientist in the ICARE Project-team, INRIA, Sophia-Antipolis, France (e-mail: ezio.malis@sophia.inria.fr).

P. Rives is a research director in the ICARE Project-team, INRIA, Sophia-Antipolis, France (e-mail: patrick.rives@sophia.inria.fr).

J. Sarrazin is a research scientist in the Deep Ecosystem Study Department at IFREMER, centre de Brest, France (e-mail: jozee.sarrazin@ifremer.fr@ifremer.fr).

P. M. Sarradin is a research scientist in the Deep Ecosystem Study Department at IFREMER, centre de Brest, France (e-mail: pierre.marie.sarradin@ifremer.fr).

A. Trajectories Induced by the Geometry of the Stereo Rig

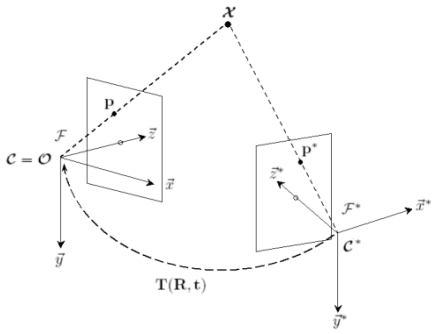

Fig. 1. Stereovision principle

The principle of our method is to generate and follow a trajectory which allows us to know the different positions of the cameras to compute an accurate 3D model of the scene. The stereovision system is composed of two cameras hung from the tip of an instrumented arm of an underwater robot. The trajectory is described by the repetition of the displacement of one camera at a start position C onto the start position C^* of the second camera (Fig. 1). We have shown in [3], that the trajectory of the controlled camera is defined by the geometry of the stereo rig which is unchanged during displacement. So, the adjustment of the stereo rig is chosen according to the dimensions of the object under study. Using a fixed geometry of the stereovision system, the trajectories generated by the cameras is inscribed on the surface of a cylinder [3]. Fig. 2 shows an example of a circular trajectory obtained with a pan angle α on the right camera and a distance *l* between the cameras.

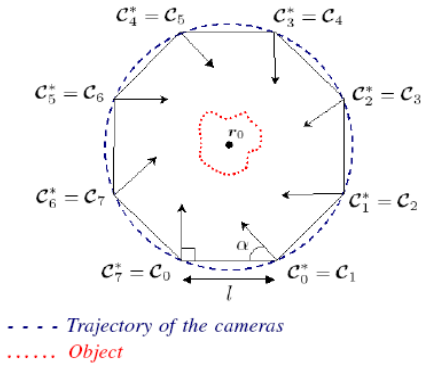

Fig. 2. Circular trajectory induced by the geometry of the stereo rig

B. Visual Servoing Invariant to Camera Intrinsic Parameters

The aim of visual servoing is to control robotic system movements by exploiting image sensor information [4]. A typical task consists in moving a camera fixed to the end-effector of a manipulator arm on a reference position according to an object. To perform this task two approaches can be used: model-based visual servoing if the 3D model of the observed scene is known, or model-free visual servoing which requires a preliminary learning step of the reference image at a reference position. In our case the 3D model being unknown, the second technique is more appropriate and teaching by showing approach can be used. Generally this technique is "camera-dependent", i.e. the same camera must be used to acquire reference image and to perform visual servoing. However, with different intrinsic camera parameters and without calibration, the current image can converge towards the reference image, though the camera positions do not coincide.

Owing to technical constraints and systems available on the shelf, our stereovision system called IRIS [1] is composed of two different cameras (one of them is mounted on pan and tilt). Moreover, the intrinsic parameters are influenced by temperature, salinity, pressure or wavelength in aquatic environment. Consequently, a visual servoing method independent to intrinsic parameters would be more appropriate, because it calculates an error function in a space invariant to camera intrinsic parameters [10]. This method is applied to our system and is described in [3].

C. Pre-Planned Trajectories

The same trajectory performed by visual servoing can be generated by "robot programs", which consist in programming positions of arm's rotation axes. So, the robot carries out position servoing thanks to information provided by its sensors, the servovalves. Thus this method does not require image information. The biggest inconvenient of this method is that there are as many "robot programs" as arm positions. However, using visual servoing method, only the initial position of the arm to begin the trajectory is required. Nevertheless, the advantage of pre-planned trajectories relies on the robustness of the method since the accuracy of the trajectory depends only on precision and adjustment of the manipulator arm, and not on camera images. Therefore, this method could be carried out in some specific cases and in order to compare 3D reconstruction results with the two image acquisition methods.

D. System Validation

The first experiments carried out in air under laboratory conditions, shown in [3], allowed us to validate the results obtained by simulation, to test pre-planned trajectories and visual servoing loop.

Our stereovision system IRIS integrates two underwater color cameras which are 6000 meter water depth-rated. One is a standard underwater camera; the other one includes a pan and tilt mechanism to adjust the angle between the two cameras. The distance between the camera is adjustable.

Stereo system IRIS was tested during the MoMARETO cruise, which was held from August 6 to September 6, 2006 on the French RV *Pourquoi Pas?* with the *victor6000* ROV. The main objective of the cruise [13] was to study the spatial and temporal dynamics of hydrothermal communities colonizing the MoMAR area located on the Azores Triple Junction. The first trials were conducted on the 850m deep Menez Gwen area and the second experiments were carried out on the Lucky Strike hydrothermal vent field, at a depth of 1750m.

The vehicle was positioned on the seafloor at a fixed and

stable attitude in front of a small scene of interest (approximately 1 m^3) that had to be reconstructed in 3D. Then, the stereovision system IRIS was hung from the tip of the arm manipulator (Fig. 3).

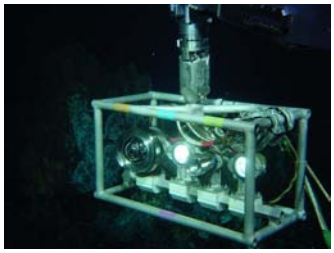

Fig. 3. Stereovision system IRIS hung from the tip of the instrumented arm of the *victor6000* ROV

A trajectory was generated around the object that had to be reconstructed using the whole arm workspace, so that the images were collected at regular spatial intervals (Fig. 4). During the second trials, the vehicle was deployed in two positions to collect more images of the scene.

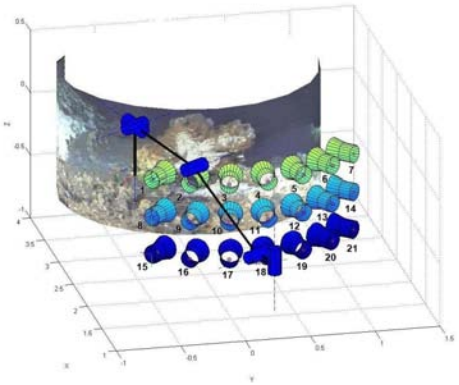

Fig. 4. Representation of the different camera positions that define the trajectory around the object

During MoMARETO cruise, trajectories were generated by visual servoing, and using pre-planned trajectories detailed in the previous section. Although, the second method is heavier to put into operation, the exploitation of the image sequences for the 3D reconstruction is the same for both methods. The only difference could concern the 3D reconstruction accuracy, but this comparison is beyond the scope of the paper.

Additional images were registered to calibrate the stereovision system in order to make a metric reconstruction. These images must be collected in situ since the intrinsic parameters of the cameras can change according to the deep sea environment. A calibration pattern was deployed on the sea floor by the arm of the underwater vehicle, and then a series of image pairs was collected from different viewpoints with the stereovision cameras.

In the following section, a 3D reconstruction method and the first result obtained on natural underwater images collected during the MoMARETO cruise are presented.

III. 3D Reconstruction

This part addresses the problem of reconstructing a 3D model, with texture, from a sequence of images of an underwater object registered with stereo cameras during the MoMARETO cruise. The 3D reconstruction is performed off-line since the whole process is very time-consuming. Each step that we will develop hereafter to obtain the 3D model are summarized in Fig. 5.

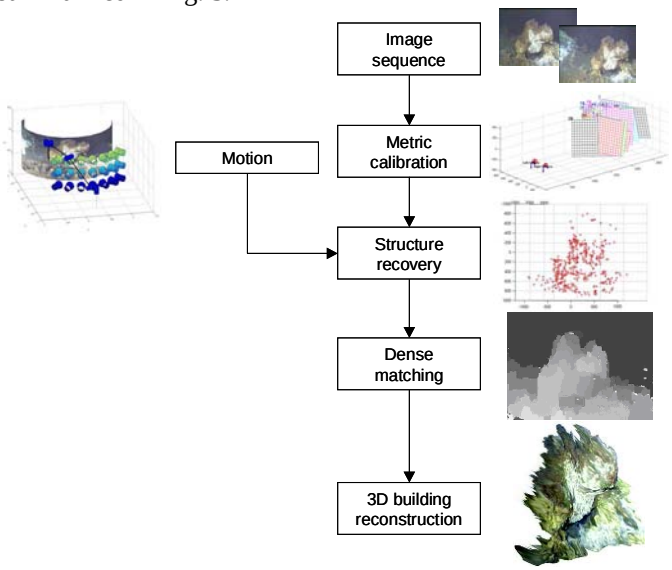

Fig. 5. 3D reconstruction overview

The first results of the method presented in this paper have been obtained using only one stereo image pair (Fig. 6). The future extension of this method to an image sequence is explained in the last part.

Fig. 6. Stereo image pair (positions 4 and 5 represented in Fig. 4)

A. *Relating Images*

A preliminary step in 3D reconstruction consists in extracting and matching some features in the stereo images in order to recover the accurate geometry linking these views. The SIFT method (Scale Invariant Feature Transform) [9] is very suitable in our case because the interest points are invariant to image scaling and rotation, and partially invariant to changes in illumination and 3D camera viewpoints. For every keypoint a descriptor is defined, which is a vector with 128 dimensions based on local image gradient directions in the keypoint neighborhood. Thus, the descriptors allow us to compare the points by using the RANSAC algorithm [5].

Few false matches remain after matching but can be removed thanks to additional constraints linked to the epipolar geometry. It corresponds to the intrinsic projective geometry

between two views and is represented by the fundamental matrix. Using (1), the fundamental matrix is calculated:

$$\mathbf{p}'^{\mathrm{T}}\mathbf{F}\mathbf{p} = 0 \quad (1)$$

where points $\mathbf{p}$ and $\mathbf{p}'$ are a pair of matching points in the two images. Thus, the search of a corresponding pixel in the second image can be restricted to a single line. The fundamental matrix will also be used later to rectify images in order to perform a dense matching. The fundamental matrix is estimated with the normalized 8-point algorithm shown in [6], combined with a RANSAC algorithm which is a very robust estimator capable to cope with a large proportion of outliers, using only a set of input point matches.

B. Camera Calibration

The camera parameters are calculated off-line from the image pairs representing the planar checkerboard using the camera calibration toolbox developed by Jean-Yves Bouguet [2]. First, the intrinsic parameters are estimated for each camera, and then the external parameters are estimated to determine the geometry of the stereo rig, in order to know the theoretical spatial distribution of the images around the object.

C. Structure and Motion Recovery

A basic triangulation of point pairs is worked out to obtain an estimation of the 3D structure, using the geometry of the stereo rig and the internal camera parameters. Although this triangulation is quite rough, it provides a first estimate which is used as an initialization of the 3D structure in the next section.

In case of visual servoing, some uncertainties are induced by noise in the input images, and can introduce a shift between the real and the theoretical camera positions. On the other hand, in case of pre-planned the arm trajectory, the accuracy depends on the precision and the adjustment of the manipulator arm. So, a basic triangulation of the projected point matches is not sufficient to improve the final reconstruction result quality. Therefore, we use an algorithm of minimization, such as a sparse bundle adjustment algorithm [15], which works out the best possible fit and corrects the relative camera pose of all views and the corresponding 3D features. This algorithm requires the theoretical camera positions, the camera intrinsic parameters, and the 3D structure estimated by triangulation.

This algorithm allows us to find the 3D points $\mathcal{X}_i$ and the parameters of the camera view $\mathbf{P}_k$ such that the mean squared distances between the observed image points $\mathbf{m}_{ki}$ and the reprojected image points $\mathbf{P}_k(\mathcal{X}_i)$ are minimized. The camera projection model we use takes also radial distortion into account. For m views and n points the following criterion should be minimized:

$$\min_{\mathbf{P}_k,\mathcal{X}_i} \sum_{k=1}^{m}\sum_{i=1}^{n} D(\mathbf{m}_{ki}, \mathbf{P}_k(\mathcal{X}_i))^2 \quad (2)$$

where $D(\mathbf{m},\hat{\mathbf{m}})$ represents the Euclidean image distance. Fig. 7 shows the results of the bundle adjustment on a pair of images, but the algorithm can easily be applied on the 21 images of the sequence. This sparse object gives the outlines of the object shape, even if there are no sufficient surface details for a good visual reconstruction.

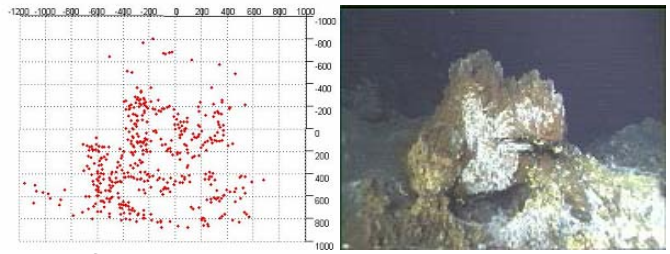

Fig. 7. Structure recovery

D. Dense Surface Estimation

The 3D structure obtained in the previous section contains only a sparse set of 3D points. In order to obtain a highly realistic 3D reconstruction, the 3D structure must be improved by a dense depth estimation. This step is composed of two main parts which are explained hereafter.

First, in order to simplify the dense matching procedure, the stereo images are rectified. It consists in transforming both images in a standard geometry so that both image planes are coplanar and the epipolar lines are projected to infinity: all the epipolar lines are parallel and horizontal. To rectify images, the general rectification method presented in [11] is used. It requires only two stereo images, the fundamental matrix previously estimated and the set of SIFT point matches used to calculate the fundamental matrix. The compatible homography that minimizes inter image distortion due to perspective effects is selected and applied to one image in order to make all matching epipolar lines coincide.

Given the fundamental matrix $\mathbf{F}$, the set of homographies which are consistent with the geometry of a particular image pair is:

$$\mathbf{H} = [\mathbf{e}']_{\times}\mathbf{F} - \mathbf{e}'\mathbf{v}^{\mathrm{T}} \quad (3)$$

where $\mathbf{v}$ is an arbitrary vector such that $\det\mathbf{H} \neq 0$.

Assuming the point matches follow a Gaussian distributed error, a compatible homography is estimated by minimizing the following least-square criterion for n points:

$$\min_{\mathbf{H}} \sum_{i=1}^{n} D^2(\mathbf{m}_i, ([\mathbf{e}']_{\times}\mathbf{F} - \mathbf{e}'\mathbf{v}^{\mathrm{T}})\mathbf{m}'_i) \quad (4)$$

where the compatible homography is parameterized as in (3).

Afterwards, both images are parameterized with polar coordinates centered on the epipole to make epipolar lines parallel to an image axis. Thus, given a point in an image, its corresponding point in the second image will be searched on the horizontal epipolar line [7] (Fig. 8).

In this paper, only two images are used for 3D reconstruction. They are chosen among all images collected during the cruise and can be chosen vertically or horizontally. Stereo images must be rectified before working out dense matching only if images are aligned horizontally (because there is an angle between the two cameras) (see Fig. 4).

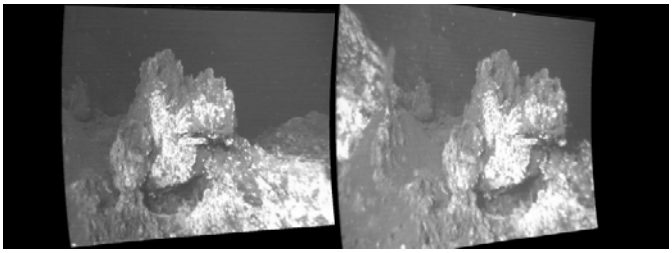

Fig. 8. Stereo images rectified

The second step is to perform dense matching on the rectified images. A large number of stereo matching algorithms exist and they can be classified in two main categories: local and global methods, according to the principle they are based upon. Other methods called cooperative algorithms use local and global approach at the same time. The difficulty is to choose an algorithm to perform a dense 3D reconstruction taking into account rendering, metrologic quality, computing speed and complexity of the scene. The taxonomy of [14] provides information about the overall performance of the principal algorithms (textureless regions, depth discontinuity regions, occluded regions). Finally, the graph-cut method gives excellent results, performing better in textureless areas and near discontinuities, and outperforming the other optimization methods. The major downsides are the relatively high computation time, and the need for precisely tuned parameters, whose values are often image-dependent. This algorithm remains however a very interesting choice for our application, since the quality of the rendering process is a higher priority compared to execution time.

Roy and I. J. Cox [12] were the first ones to use this algorithm in the context of multi-camera stereovision. In order to explain the graph-cut method, we will concentrate on the case of graphs with only two terminals.

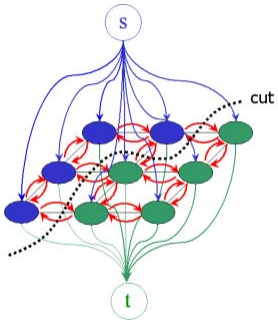

Fig. 15. Graph-cut example

Fig. 15 shows a simple example of a two terminal graph, which can be used to minimize an energy function on a 3×3 image with two labels. The two terminals are usually called source *s*, and sink *t*. They correspond to the set of labels (different depths) that can be assigned to pixels. The different nodes represent pixels of the image. In the general case of graph-cut theory, the goal is to find a cut that has a minimum cost among all cuts, by minimizing an energy function.

Let function f be the disparity function associated to each pixel of an image. We search labelling f that minimizes the energy. To define this energy function for f, a cost function is introduced, based on a photoconsistence criterion (similarity between intensities of a pixel $\mathbf{p}$ in the first image an the pixel $(\mathbf{p}+f_p)$ in the second image) called data term. A second term, called smoothness term, penalizes discontinuities between neighbourhood pixels. Thus, the energy can be written as:

$$E(f)=\sum_{p\in\mathcal{P}} D_p(f_p)+\sum_{\{p,q\}\in\mathcal{N}} V_{\{p,q\}}(f_p,f_q) \qquad (5)$$

where term D_p is the data term and $V_{\{p,q\}}$ is the smoothness term penalty between adjacent pixels.

In [8], the energy minimization considers the input images symmetrically, handles visibility properly, and imposes spatial smoothness while preserving discontinuities. Fig. 9 shows the dense disparity map obtained with the implementation of the graph-cut method presented in [8].

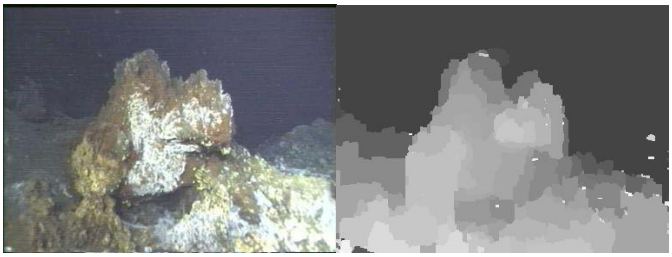

Fig. 9. Dense disparity map estimation with graph-cut method on MoMARETO images

The disparity map gives us a dense correspondence map between the stereo images. Thus, the depth map is computed by triangulation with matched point pairs and camera parameters. But each 3D point is considered independently. Therefore smoothing the surface is important to obtain a spatial coherence. In this paper, a spatial coherence is achieved by filtering the depth map, but in the future, a better choice would be to interpolate the depth map with a parametric surface model [7].

E. Visual Scene Representation

The dense map is converted to triangular surface meshes using the Delaunay triangulation algorithm. Thus, a texture mapping can be applied with relative easy and efficiency to the object. The triangular mesh makes possible to reduce the geometric complexity of a 3D surface representation. The resulting surface model obtained from the dense depth map is presented in Fig. 10.

The 3D representation is then visualized in a more realistic appearance by providing the wire-frame model with texture mapping. First, a reference image is chosen as the texture map in the image sequence, and then, each basic triangle primitive is easily mapped with texture, since the exact position of the reference image and of the 3D structure are known.

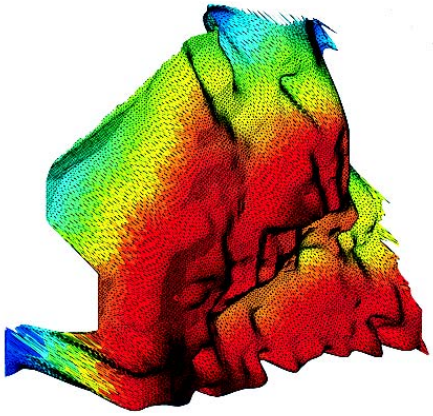

Fig. 10. Triangular mesh on a smooth dense depth map

Fig. 11. 3D model with texture

Fig. 11 shows the final reconstruction with texture mapping which can be used for 3D quantitative imaging. The dense depth map with texture mapping gives a 3D model with a good visual impression. The 3D model is stored in VRML format for easy visualization and exchange of information.

IV. Conclusion and Future Work

In this paper, we have presented a complete reconstruction method of small-scale natural underwater objects to supervised exploration of ocean floors. The first step concerns the acquisition of images, with the stereovision system IRIS which enables us to collect images, at regular spatial intervals that depend on the stereovision rig geometry. This constraint is used as a priori knowledge to optimize and improve the final 3D reconstruction which is carried out in a post-processing stage. The first results of 3D reconstruction obtained with only two natural images collected during the MoMARETO cruise turned out to be very efficient and promising. However, the accuracy of the reconstruction is greatly improved when more views are used. So, the natural extension of the current work would be to perform the 3D reconstruction from a large set of images. The exploitation of all collected images requires data fusion (multiple depth map combination and texture fusion) and a modification of the matching method to provide a more powerful constraint to identify mismatches (estimation of the trifocal tensor).

References

[1] Allais, A-G, Brandou, V, Hoge, U, Bergmann, M, Lévêque, J-P, Léon, P, Cadiou, J-F, Sarrazin, J, and Sarradin P-M (2005). “Design of optical instrumentation for 3D and temporal deep-sea observation”, *Proc. of the 1st international conference of optical complex systems*, OCS, Marseille, France.

[2] Bouguet, J-Y, and Perona, P. Camera calibration from points and lines in dual-space geometry. Technical Report, California Institute of Technology, 1998.

[3] Brandou, V, Malis, E, Rives, P, Allais, A-G, and Perrier, M (2006). “Active stereovision using invariant visual servoing”, *IROS*, Beijing, China.

[4] Espiau, B, Chaumette, F, and Rives, P. “A New Approach to Visual Servoing in Robotics”, IEEE *Trans. on Robotics and Automation*, pp. 313-326, 1992.

[5] Fischler, M-A, and Bolles, R-C (1981). “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”, *ACM*, Vol 24, pp 381-395.

[6] Hartley, R, and Zisserman, A (2000). “Multiple view geometry in computer vision”, Cambridge University Press, pp 265.

[7] Koch, R. Pollefeys, M. and L. Van Gool. “Multi viewpoint stereo from uncalibrated video sequences”. In *Proceedings of European Conference on Computer Vision*, ECCV, volume 1, pages 55-71, Fribourg, Allemagne, juin 1998.

[8] Kolmogorov, V, Zabih, R, and Gortler, S. "Generalized Multi-Camera scene Reconstruction Using Graph Cuts" In: *Fourth International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, July 2003.

[9] Lowe, D (2004). “Distinctive image features from scale-invariant keypoints”, *International Journal of Computer Vision*, Vol 60(2), pp 91-110.

[10] Malis, E. “Visual servoing invariant to changes in camera intrinsic parameters”, IEEE *Transaction on Robotics and Automation*, 20(1):72-81, February 2004.

[11] Oram, D. “Rectification for any epipolar geometry”, *12th British Machine Vision Conference* (BMVC 2001), September 2001.

[12] Roy, S. and Cox, I. J. “A maximun-flow formulation of the n-camera stereo correspondance probleme”. *Proceedings of 6th International Conference on Computer Vision*, pages 492–499, 1998.

[13] Sarrazin, J, Sarradin, P-M, and the Momareto cruise participants (2006). “Momareto: a cruise dedicated to the spatio-temporal dynamics and the adaptations of hydrothermal vent fauna on the Mid-Atlantic Ridge”, *InterRidge News*, Vol 15, pp 24-33.

[14] Scharstein, D, and Szeliski, R. A., “Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms”, *IJCV* 47(1/2/3):7-42, April-June 2002.

[15] Triggs, B, McLauchlan, P, Hartley, R, and Fiztgibbon, A (2000). “Bundle adjustment: a modern synthesis”, In B. Triggs, A. Zisserman, R. Szeliski (Eds.), Vision Algorithms: Theory and Practice, LNCS, Vol 1883, pp 298-372, Springer-Verlag.