2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Identification of Suitable Interest Points Using Geometric and Photometric Cues in Motion Video for Efficient 3-D Environmental Modeling

T. Nicosevici, R. Garcia, S. Negahdaripour, M. Kudzinava and J. Ferrer

Abstract— **Many applications in mobile and underwater robotics employ 3D vision techniques for navigation and mapping. These techniques usually involve the extraction and 3D reconstruction of scene interest points. Nevertheless, in large environments the huge volume of acquired information could pose serious problems to real-time data processing. Moreover, In order to minimize the drift, these techniques use data association to close trajectory loops, decreasing the uncertainties in estimating the position of the robot and increasing the precision of the resulting 3D models. When faced to large amounts of features, the efficiency of data association decreases drastically, affecting the global performance.**

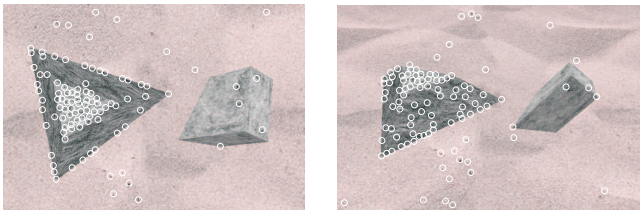
This paper proposes a framework that highly reduces the number of extracted features with minimum impact on the precision of the 3D scene model. This is achieved by minimizing the representation redundancy by analyzing the geometry of the environment and extracting only those features that are both photometrically and geometrically significant.

I. Introduction

Vision-based navigation and mapping algorithms use visual features to create maps of the environment. As the robot navigates the map increases in size and complexity to a point where the computational costs become too high for real-time processing. Moreover the efficiency of data association, a crucial part of of the systems, decreases as the complexity of the map augments. Therefore, it is essential for these systems to extract few but representative environment features.

Most of the 3D vision proposals found in the literature [1], [2], [3], [4], [5], [6] are based on the reconstruction of sets of scene key points. These points are matched in various views of the scene, either supplied by multiple cameras or by a single moving camera. Provided the position of the camera(s) within a reference frame (either by pre-calibration on the case of multiple cameras or on-the-fly auto-calibration in case of moving cameras), the 3D position of the key points is estimated. The result is a cloud of 3D points with respect to the chosen reference frame that can be interpreted as a set of discrete measurements of the viewed region. However, the vast majority of natural scenes are hardly discrete and a sparse model of them could be hard to interpret either by humans or by machines. In order to overcome this problem, the key points are interpolated using linear or quadratic techniques resulting in a continuous model.

It is obvious that the precision of the 3D model is highly dependent on the accuracy of the estimation of the 3D points position. Therefore, the key points have to be reliably tracked over multiple views.



(a) (b)

Fig. 1. Using Harris corner detector on two views of a synthetic scene top view (a) and perspective view (b). Most extracted features (represented by white circles) have little geometric significance

The problem of extracting key image features has represented a topic of intensive research in the last decade, resulting in the development of a variety of interest point detectors: Shi and Tomasi [7], SIFT [8], SURF [9], affine covariant [10], etc. All of these proposals use a similar approach based on extracting points that represent regions with high image intensity gradient. Practice has proven that these regions are highly discriminative and they are more robust to image noise, changes in illumination, camera point of view, etc.

However, even an accurate reconstruction of the 3D points obtained by an interest point detector cannot guarantee a consistent 3D reconstruction of the scene. Changes in image intensities could be a result of rich textures, shadow/light changes and do not necessarily represent edges/corners of scene objects. Hence, as image features are extracted using image intensity measurements, they do not necessarily have a geometric meaning. A representative example is illustrated in Fig. 1. Harris corner detector was applied on two views of a synthetic scene. The features are cluttered in the regions with high texture and very few on the actual corners/edges of the object. The 3D model based on these features would represent a poor approximation of the real geometry of the 3D objects. In order to overcome this problem, the 3D reconstruction algorithms found in the literature extract dense sets of points from the scene. By increasing the number of measurements, the accuracy of the resulting model increases. Although this is an acceptable solution for 3D reconstruction of small areas, in the case of robot navigation it imposes a set of drawbacks mostly related to the incremental complexity of the problem. Therefore, it is important to devise solutions to extract a minimum number of features that could efficiently model the environment. To the best of our knowledge no previous work has addressed the problem of reducing the

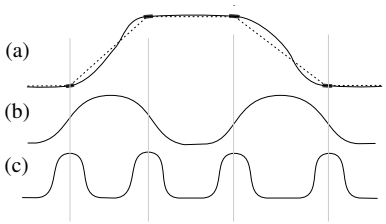


Fig. 2. Simple 2D example of ideal features extraction from topological point of view: (a) 4 feature points provide a good initial piece-wise linear approximation of the curved profile; (b) absolute value of first derivative; (c) the 4 features correspond to the maxima of the response of the second derivative.

amount of the extracted features by jointly analyzing the geometrical and photometrical characteristics of the environment.

We propose an algorithm that extracts image features that are consistent with the 3D structure of the scene. The features can be robustly tracked over multiple views and serve as vertices of planar patches that suitably represent scene surfaces, while reducing the redundancy in the description od 3D shapes. In other words, the extracted features will offer good tracking properties while providing the basis for 3D reconstruction with minimum model complexity.

In order to better understand the concept, consider the simple example in Fig. 2a, which illustrates a 2-D profile as the cross section of a 3-D relief. By extracting features around the edges of the slopes (marked in dark grey) and applying linear interpolation (dotted lines), a good initial approximation of the shape is obtained.

The following section provides a detailed description of the approach along illustrative example, followed by a presentation of a set of experimental results validating the proposal. The paper concludes with a brief presentation of what we have accomplished and what is still to be done.

II. Algorithm Description

The proposed algorithm was developed for Autonomous Underwater Vehicle (AUV) navigation based on monocular vision systems. This particular application imposes a series of problems in addition to those mentioned earlier:

- underwater environments are cluttered with very few well defined geometrical characteristics (i.e. edges and corners);
- illumination changes, back-scattering and light attenuation increase the difficulty of feature tracking;
- 3D reconstruction is based on a single moving camera, using no external information regarding the camera motion.

Fig. 3 outlines the main modules of the proposal, which is designed to process the data as it is acquired. There are two parallel modules of processing: (*i*) geometric features processing and (*ii*) photometric features processing. The information of the two modules is merged in order to generate features that are both geometrically representative and robustly trackable. The obtained features are 3D reconstructed and interpolated in order to obtain the 3D model.

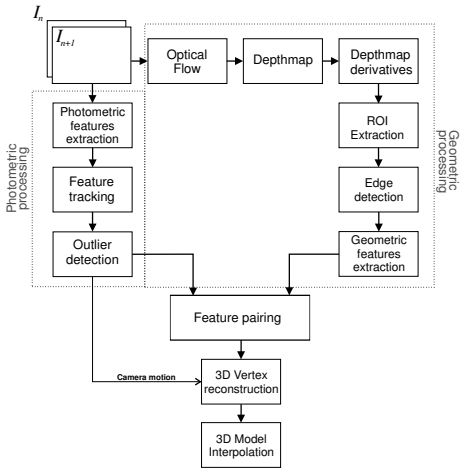


Fig. 3. Flowchart of the feature extraction algorithm.

A detailed description of each module of the proposal in provided hereafter.

A. Optical Flow and Depth Map Computation

The first step of the geometrical features extraction is the computation of the 2-D optical flow $\mathbf{v} = [u\ v]^T$ from pairs of images. The adopted generalized dynamic image model (GDIM) based method was proposed by Negahdaripour [11], and later generalized to take advantage of color in addition to intensity information for improved robustness and estimation accuracy [12]. The computed optical flow for each pair $\{I_n, I_{n+1}\}$ of consecutive images provides an estimate of local disparities for depth computation.

The Longuet-Higgins differential image motion model is the basis of the depth estimation module:

$$\mathbf{v} = \mathbf{A}_\omega \omega + \frac{1}{Z}\mathbf{A}_t t \tag{1}$$

Here, ω and t are camera rotation and translation velocities respectively, and Z is the distance to a scene point along the optical axis. Based on (1), pairwise 3-D motions and depth maps are computed iteratively from the optical flow [13]. It should be noted that both the depth maps are computed up to scale (due to the well-known scale-factor ambiguity of monocular vision). The correct scaling can be determined with a single distance (depth) measurement, or knowledge of motion magnitude. Fig. 4b illustrates the depth map estimation for the synthetic scene shown in Fig. 4a.

B. Depth map derivatives

In order to extract the geometric features, the system focuses its search on two types of regions of interest: (*i*) object edges and (*ii*) surface inflexions. Practically, these types of regions correspond to high responses of the second derivative of the depth map and will be called edges hereafter. Analyzing Fig. 2, it can be observed that the 4 ideal feature points correspond to local maxima of the second derivative (Fig. 2c).

The second derivative of the depth map is obtained by:

$$D''_m(x,y) = \frac{1}{N}\Sigma_{i=1}^{N} D_m(x,y) * LoG(\sigma_i) \tag{2}$$

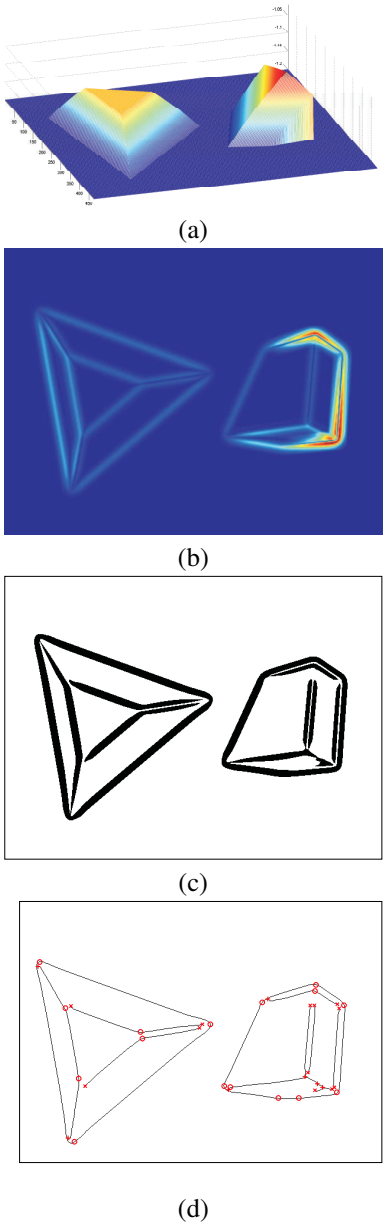


Fig. 4. Main steps of the proposal: (a) depth map of the scene, (b) computation of second derivative, (c) normalization and binarization and (d) edge traces ($\times$ represent line ends, $+$ represent line junctions and $\circ$ represent high curvature points)

where $*$ is the convolution operator, $LoG(\sigma_i)$ is the Laplacian of Gaussian with standard deviation $\sigma_i = m \cdot i$, m being a predefined constant. By using this approach, D''_m becomes less sensitive to noise, while having high responses on the edges of the surfaces(see Fig. 4b).

C. Extraction of Regions-of-Interest (ROI)

As mentioned earlier, the regions of interest correspond to those areas where D''_m has high values. In order to extract these regions a simple binarization would suffice. However the steepness and area of the slopes influence the magnitude and width of the peaks in D''_m. In this case, applying a binarization would either not detect certain edges or would generate false edges due to image noise. In order to obtain a constant binarization, D''_m is locally normalized using:

$$\widehat{D''_m}(x,y) = \frac{D''_m(x,y) - \overline{w}_n(x,y)}{\sqrt{\overline{v}_n(x,y) - \overline{w}_n^2(x,y)}} \quad (3)$$

where

$$\overline{w}_n(x,y) = \frac{\sum_{i=x-n}^{x+n}\sum_{j=y-n}^{y+n} D''_m(i,j)}{(2n+1)^2} \quad (4)$$

and

$$\overline{v}_n(x,y) = \frac{\sum_{i=x-n}^{x+n}\sum_{j=y-n}^{y+n} (D''_m(i,j))^2}{(2n+1)^2} \quad (5)$$

(Fig. 4c) shows the result after normalization and binarization using a preestablished threshold.

D. Geometrical features extraction

The extraction of interest regions along surface edges greatly decreases the size of the area where features are extracted, reducing drastically the complexity of the model. Nevertheless, in order to minimize even further the redundancy, the system extracts only key edge points. First, the edges are recovered by applying a thinning algorithm to the regions of interest [14]. The result is a pixel wide trace line following the edge (hereafter called traces), with each pixel corresponding to the local maxima of D''_m along the direction perpendicular to the edge (hence corresponding to points of maximum surface inflexion) (Fig. 4d).

In order to extract the geometrical interest points, three types of features are defined along the trace: (*i*) line end points, (*ii*) lines junction points and (*iii*) high curvature points. The trace image is a binary image with 0's corresponding to background and 1's corresponding to line traces (Fig. 4d). Line end points and line junction are obtained by convolving the binary image with specific kernels and extracting points with local maxima. The curvature of the trace line along each point p is obtained by computing C_p within a $2n+1$ 1D window along the line [15], with:

$$C_p = \frac{1}{(2n+1)} \sum_{i=p-n}^{p+n} exp(-d_{ip}^2)(1 - cos(\phi_p - \phi_i)) \quad (6)$$

where ϕ_p and ϕ_i represent the angles of the line normals at points p and i respectively; d_{ip} represents the euclidean distance between p and i.

High curvature points are extracted by locating local maximum of C_p where $C_p > t_c$. The threshold t_c is imposed in order to avoid false positives due to image aliasing.

Fig. 4d illustrates the extracted geometric features: line junctions are represented by a cross (+),line ends are represented with a diagonal cross ($\times$) and the circles ($\circ$) denote high curvature points.

E. Photometric features extraction and matching

As outlined earlier, in order to recover the 3D position of the interest points, the system has to track them over multiple images. However the neighboring areas of the geometric features might not provide sufficient information for reliably tracking the features. This drawback becomes

even more evident in the case of underwater scenes, where light attenuation and back-scattering effects dim the textures, limiting the efficiency of feature tracking. The solution to this problem consists in substituting the geometric features with neighboring photometric features. Among the wide set of alternatives present in the literature, the proposed approach makes use of Scale Invariant Features (SIFT) [8], as it presents a series of advantages in the context of the proposal:

- it generates dense sets of image features – increasing the chances of having neighboring photometric and geometric features;
- allows matching under a wide range of image transformations (i.e. rotation, scale, perspective) – an important aspect when imaging complex 3D scenes at close range as in the case of underwater vision;
- the image descriptors are highly discriminative – providing bases for data association (loop closing, SLAM, etc.).

As a first step the SIFT algorithm generates a scale space $L(x, y, \sigma)$ by convolving repeatedly an input image $I(x, y)$ using a variable-scale Gaussian, $G(x, y, \sigma)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{7}$$

In order to detect scale-invariable image locations, the algorithm analyzes the images at different scales and extracts the key points. These points represent scale-space extrema in the difference-of-Gaussian function $D(x, y, \sigma)$ convolved with the image:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \tag{8}$$

where k is a constant multiplicative factor.

Once extracted, each feature is assigned with a scale and an orientation vector (major direction of the local image gradient at the scale where the feature was extracted). The feature descriptor is calculated after aligning (rotating) the nearby area of the feature according to the assigned orientation, thus achieving invariance of the descriptor to rotation. Each set of feature descriptors is represented by a 128 elements vector obtained by analyzing image gradients in 4×4 windows around the feature. For each window a local orientation histogram with 8 bins is constructed. This way, every feature can be represented as a point in a 128-dimension descriptor space. The matching is carried out by computing point to point distances between features in the descriptors space. Each two closest points are considered matches if the distance between them is lower than a predefined threshold t_m, that allows to eliminate features that do not have any proper match and avoid specific instances of feature ambiguity.

F. Outlier rejection and camera motion estimation

To ensure robust tracking as each image feature position p^{n+1} is determined in image I_{n+1}, an outlier rejecting process is carried out. This process uses a RANSAC approach that evaluates the first order approximation of the geometric error d (Sampson distance) using the fundamental matrix (F) [1]. Once F is obtained from the set of correct matches, the camera motion is recovered for later use by the 3D reconstruction algorithm:

$$F = (K^{-1})^T S R K^{-1} \tag{9}$$

where K is the matrix encoding the intrinsic camera parameters (obtained by pre-calibration), S is the skew-symmetric translation matrix ($Sx = t \times x$ for any vector x), and R is the rotation matrix of the camera.

G. Feature pairing

In order to obtain reliable key points for 3D reconstruction, the algorithm attempts to substitute geometric features with nearby photometric features that can be reliably tracked.

The substitution of each geometric feature with a photometric feature is carried out using criteria based on two measurements: the quality of the photometric feature and the distance between the geometric and photometric features. In the case of pairwise 3D reconstruction, the quality of the photometric feature is given by $\overline{D}_{ss}(i)$, which represents the distance between the feature and its match in the 128-dimension descriptor space, scaled by t_m. The decision criteria is defined as:

$$ps(k, i) = (1 - \overline{D}_{ss}(i)) \cdot cos(-\frac{\pi}{2} \cdot \frac{D_G(k, i)}{max_{DG}}) \tag{10}$$

where $D_G(k, i)$ is the euclidean distance between geometric feature k and photometric feature i and max_{DG} represents the maximal accepted distance between i and k. The use of the cosine function in (10) applies a nonlinear weight that rewards features which are closer to the geometric feature and penalizes those towards the outer radius max_{DG}.

For each each geometric feature k, ps is computed for all image features that fall within a radius of max_{DG}. The photometric feature with the highest score ps is considered the pair of k. This approach creates a tradeoff between feature tracking reliability and geometric precision.

III. Experimental Results

The testing of the technique was carried out in two steps:

- synthetic data, focused on testing the efficiency of the geometrical features in ideal cases;
- real data to assess the proposed approach when faced to real underwater scenes.

A. Synthetic Data

The objective of these experiments is to test the efficiency of the extractor of geometric features. This was carried out by reconstructing the scene using image pairs. In this first case it was assumed that the geometric features are matched in absence of noise. As the camera motion is known, the 3D points corresponding to the geometric features are computed and interpolated resulting in a 3D model of the scene.

In order to quantify the efficiency, the obtained 3D model is compared with the ground truth. The error is computed on

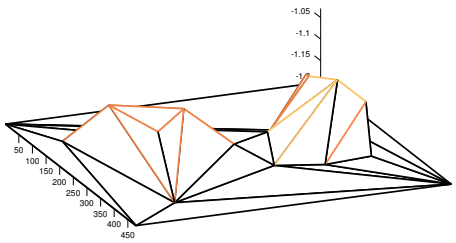


Fig. 5. 3D reconstruction of the synthetic scene. The model was obtained using 30 geometrical features

a point-to-point basis using a grid with the same resolution as the input images:

$$e_s[\%] = \frac{1}{n \cdot m} \sum_{x=1}^{m} \sum_{y=1}^{n} \frac{|Z_E(x,y) - Z_G(x,y)|}{Z_G(x,y)} \cdot 100 \quad (11)$$

Here e_s is mean error, $Z_E(x,y)$ is the estimated depth value at point (x,y), $Z_G(x,y)$ is the ground truth depth at point (x,y) and m,n is image width and height.

Depending on the complexity of the scene, the error e_s was found to vary between 0.6% and 1.6%. In the case of the scene presented in Fig. 1a, the system extracted 30 geometrical features resulting in a 3D model illustrated in Fig. 5 with a reconstruction error of $e_s = 0.8\%$. Again, it should be taking into account that the tests using synthetic data were intended solely to validate the effectiveness of the 3D reconstruction based on geometrical key points.

B. Real Data

The proposal was tested using various underwater scenes with the main objective of examining the error between the model obtained by using the full set of photometric features and the model obtained using the proposed technique. In other words, how the precision of the 3D model is affected by the reduction of its complexity. This error was defined similarly as in (11):

$$e_r[\%] = \frac{1}{n \cdot m} \sum_{x=1}^{m} \sum_{y=1}^{n} \frac{|Z_{GP}(x,y) - Z_P(x,y)|}{Z_P(x,y)} \cdot 100 \quad (12)$$

where $Z_GP(x,y)$ is the depth at (x,y) corresponding to the 3D model obtained using the proposed approach and $Z_P(x,y)$ is the depth at (x,y) as estimated from the model computed using the full set of photometric features.

The first data set presented in this paper was extracted from an image sequence of a coral reef in Bahamas. The images were acquired by the Underwater Vision Laboratory of University of Miami during a survey where the camera was located at an altitude of approximatively 2 meters. The resolution of the images is 360 by 240 pixels and the depth variance of the scene is around 1.5 meters.

After processing the dataset, the algorithm yielded a set of 56 geometrical features and 343 photometric features. One of the input images is illustrated in Fig. 6a. Once the depth map has been extracted (Fig. 6b), the system computes the edges of the scene surfaces (Fig. 6c). The final geometrical features are shown in 6d. Regarding the image

TABLE I

RESULTS IN CASE OF THE BAHAMAS DATA SET.

max_{DG}	Resulting features	Complexity [%]	e_r[%]
10	23	6.7	7.52
15	39	11.37	5.2
19	**51**	**14.87**	**4.5**
25	53	15.45	4.92
30	59	16.03	5.22

Model error e_r and complexity are affected by tuning max_{DG}. The complexity represents the percentage of final features out of the total number of photometric features.

processing and feature extraction, there are a few adjustable parameters: range of σ_i in (2), Dm'' binarization threshold, t_c for extracting curvature points, maximal scale-space match distance t_m and max_{DG} in (10). The optimal value of these parameters has been empirically established and proved to generate consistent results trough extensive testing using multiple datasets. However, among these parameters, adjusting max_{DG} has proven to have the greatest impact and it is discussed hereafter.

Table I shows that different values of max_{DG} influence both the complexity and precision of the 3D model. Using low values, few geometric features are paired with photometric features resulting in a higher e_r. As illustrated in Fig. 7, as max_{DG} is increased, more features are added (Fig 7b) decreasing the model error (Fig 7a) down to a minimum. Testing on different datasets showed that the minimum model error is achieved when $15 \leq max_{DG} \leq 25$. If max_{DG} is increased beyond this range, it decreases the influence of the euclidean distance $D_G(k,i)$ in (10), resulting in the extraction of feature points further from the ideal geometrical position. The obtained model with texture rendering is illustrated in Fig.8.

IV. CONCLUSIONS AND FURTHER WORK

A framework for optimizing the extraction and tracking of image features has been proposed. The presented technique is intended to reduce the computational costs for robot navigation and to improve data association efficiency in large scene reconstruction. The key aspect of the methodology is the extraction of geometrical representative regions and to associate them with image features that can be robustly tracked in multiple views. The experimental results have shown that this approach enables the reduction of 3D model complexity up to 90% with a precision cost of only 4-5%.

An important topic of ongoing research is to assess the reliability of the resulting features for data association and loop closure. This will be carried out by testing the behavior of the proposal under extreme image transformations.

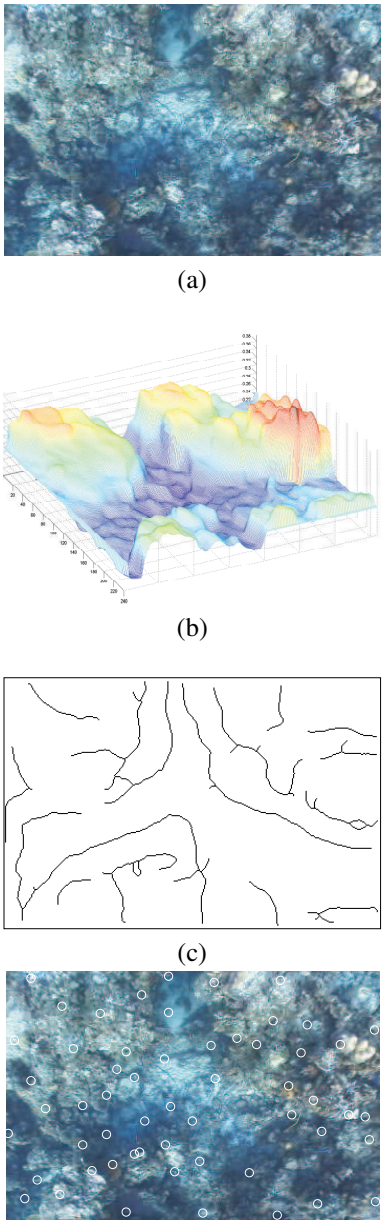


Fig. 6. Results on the Bahamas coral reef data: (a) image where feature extraction takes place, (b) computed depth map, (c) resulting edge traces and (d) geometrical features.

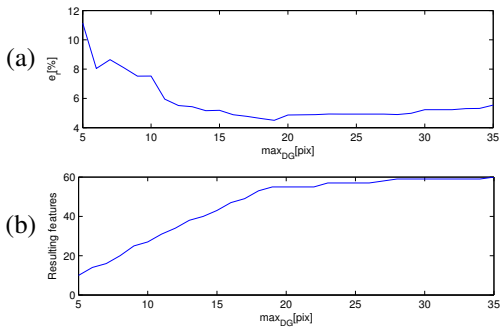


Fig. 7. Model evolution as function of max_{DG} for Bahamas dataset: (a) model error and (b) model complexity

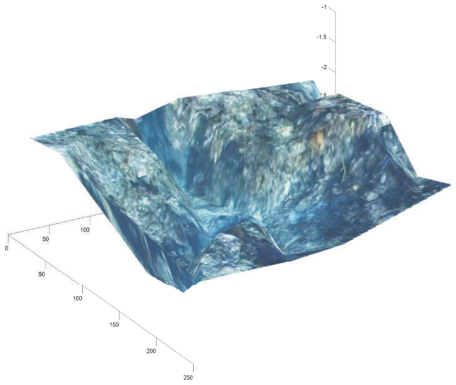


Fig. 8. Texture render of the obtained 3D model for Bahamas dataset.

Acknowledgement: This work has been partially funded through the MOMARNET EU project MRTN-CT-2004-505026, and in part by the Spanish Ministry of Education and Science under grant CTM2004-04205.

REFERENCES

[1] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[2] M. Brown, D. Burschka, and G. Hager, "Advances in computational stereo," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, p. 993, 2003.

[3] S. Negahdaripour, B. Hayashi, and Y. Aloimonos, "Direct motion stereo for passive navigation," *IEEE Trans. on Robotics and Automation*, vol. 11, no. 6, pp. 829–843, 1995.

[4] S. Negahdaripour and H. Madjidi, "Stereovision imaging on submersible platforms for 3-d mapping of benthic habitats and sea-floor structures," *IEEE Journal of Oceanic Engineering*, vol. 28, no. 4, pp. 625–650, October 2003.

[5] O. Pizzaro, R. Eustice, and H. Singh, "Large area 3d reconstructions from underwater surveys," in *MTS/IEEE Oceans*, Nov. 2004, pp. 678–687.

[6] H. Madjidi, "Three-dimensional global alignment of sensor positions from visual motion measurements," Master's thesis, University of Miami, May 2005.

[7] J. Shi and C. Tomasi, "Good features to track," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, Washington, USA, June 1994, pp. 593–600.

[8] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[9] H. Bay, T. Tuytelaars, and L. J. Van Gool, "Surf: Speeded up robust features," in *In proc. European Conference on Computer Vision*, 2006.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, pp. 43–72, 2005.

[11] S. Negahdaripour, X. Xu, A. Khamene, and Z. Awan, "3-d motion and depth estimation from sea-floor images for mosaic-based station-keeping and navigation of rovs/auvs and high-resolution sea-floor mapping," in *Workshop on Autonomous Underwater Vehicles AUV*, Aug. 1998, pp. 191–200.

[12] S. Negahdaripour and H. Madjidi, "Robust optical flow estimation using underwater color images," in *MTS/IEEE Oceans*, 22-26 Sept 2003, pp. 2309 – 2316 Vol.4.

[13] K. Kanatani, "Structure and motion from optical flow under perspective projection," in *Computer Vision, Graphics, and Image Processing*, 1987, pp. 122–146.

[14] L. Lam, S.-W. Lee, and Y. Suen, "Thinning methodologies-a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 9, p. 879, Sept. 1992.

[15] J. Deschênes and D. Ziou, "Detection of line junctions and line terminations using curvilinear features," *Pattern Recognition Letters*, vol. 21, pp. 637–649, 2000.