Monocular-Based 3-D Seafloor Reconstruction and Ortho-Mosaicing by Piecewise Planar Representation

Tudor Nicosevici*, Shahriar Negahdaripour**, Rafael Garcia*

* Computer Vision and Robotics Group, University of Girona, Girona, Spain
Email: {tudor,rafa}@eia.udg.es

** Electrical and Computer Engineering Department, University of Miami, Miami, FL USA
Email: shahriar@miami.edu

Abstract— **Photo-mosaicing techniques have become popular for seafloor mapping in various marine science applications. However, the common methods cannot accurately map regions with high relief and topographical variations. Ortho-mosaicing borrowed from photogrammetry is an alternative technique that enables taking into account the 3-D shape of the terrain. A serious bottleneck is the volume of elevation information that needs to be estimated from the video data, fused, and processed for the generation of a composite ortho-photo that covers a relatively large seafloor area.**

We present a framework that combines the advantages of dense depth-map and 3-D feature estimation techniques based on visual motion cues. The main goal is to identify and reconstruct certain key terrain feature points that adequately represent the surface with minimal complexity in the form of piecewise planar patches. The proposed implementation utilizes local depth maps for feature selection, while tracking over several views enables 3-D reconstruction by bundle adjustment. Experimental results with synthetic and real data validate the effectiveness of the proposed approach.

I. Introduction

Visual surveys have become an important component of seafloor mapping for scientific studies; e.g., [1], [2]. Developments in HDTV and very-high resolution digital systems have enabled the imaging of benthic habitats with unprecedented details, thus offering tremendous potential for exploration and new discoveries in various domains of marine sciences, including biology, geology and archeology. Coupled with recent advances in automatic and autonomous navigation, submersible imaging platforms provide mapping capabilities far surpassing those achieved from traditional scientific diver-based surveys. At the same time, these go hand in hand with tremendous processing requirements and the need for technical/algorithmic developments to generate large-area composite maps that match the resolution of individual frames (or exceed it by the employment of super-resolution techniques [3]).

Mapping in the underwater environment is inherently a complex problem. Light attenuation and backscattering drastically limit the range and coverage area of optical sensors; at best no more than a few meters in each dimension. For this reason alone, extended effort has to be devoted merely to align partially overlapping frames seamlessly in order to provide a larger coverage one that may otherwise be available in a single frame in the absence of limited visibility. Furthermore, unstructured clutter in most benthic environments demand more complex algorithms to process the image data. For example, underwater mosaicing systems have been developed based on the traditional photogrammetry mapping techniques applied to satellite and aerial imagery, assuming the planarity of the mapped scene; e.g., [4]. This enables the registration of image frames using simple transformations with only a small number of parameters, known as planar homographies; e.g., [5]. Unfortunately, most regions and (or) objects of interest for scientistic studies are hardly planar; hydro-thermal vents, coral reefs, and shipwrecks to name a few. This holds even more true in close-range imaging, targeted for recording the very fine-scale target details. In such cases, the parallax effects induce image deformations that strongly violate the planar homography model. However, there is sufficient information within overlapping regions to estimate the 3-D relief of the mapped area based on multiple-view geometrical constraints [6]. This can then be used to generate a so-called ortho-rectified mosaic [7], [8], [9].

In recent years, some work have explored the application of stereo imaging for underwater 3-D terrain reconstruction [10], [11]. This involves the use of two cameras (or generally more) in order to obtain local 3-D maps from disparity cues. The incremental local maps, generated as the stereo system moves, may be merged into a global 3-D reconstruction of the surveyed area [12]. Another approach is the application of structure/depth from motion (SFM/DFM) methods based on monocular images [6], [13]. SFM involves the extraction and tracking of a sparse set of features in a sequence, and the estimation of their 3-D positions using multiple views. This can be achieved rather robustly based on a bundle adjustment technique [14]. In theory, a 3-D dense map may then be generated by surface interpolation [15]. However, the 3-D dense reconstruction accuracy is highly dependent on the

terrain complexity within interpolated areas. The high relief and unstructured nature of most cluttered benthic environments of interest significantly limit the utility of "unguided" feature-based methods. In contrast, the DFM techniques provide dense local maps by exploiting the information redundancy over the entire overlapping areas. However, the merging of (somewhat noisy) dense local depth maps to generate an *accurate* large-area global map is not trivial, and in fact remains a challenging problem. Furthermore, although 3-D dense maps would provide rich and detailed information, large-area reconstruction becomes prohibitive due to very high computational costs.

We propose a framework for the integration of SFM and DFM paradigms that exploits the unique advantages each offers. Here, dense depth maps serve to establish the topological characteristics of the terrain locally, and thus to guide the selection of a small set of sparse surface features that characterize the terrain complexity with a desired level of accuracy. These features define the vertices of planar patches that are fit to the local depth maps by Delaunay triangulation. The 3-D piecewise planar representation is iteratively updated by verifying consistency with the dense local maps.

Our surface modeling strategy is borrowed from various other earlier applications, including synthetic generated images and video in computer graphics. It is also motivated to achieve significant computational savings by utilizing a minimal yet suitable 3-D model in representing and processing a very large volume of surface data. However, our method differs from those in earlier applications for the necessity to estimate critical 3-D information directly from monocular motion cues. For example, we enforce global estimation consistency by tracking features over several views, and recomputing their 3-D positions by bundle adjustment. As a result, the need arises for local adjustment in the selection of surface features to ensure they are visually trackable.

The main advantages of the proposed approach include the ability 1) to model relatively complex 3-D surfaces by a small number of features; 2) to adaptively sample the terrain surface by a non-uniform mesh as the local shape complexity dictates; 3) to maintain global consistency by bundle adjustment; 4) to scale up the mapped terrain by the number of features, rather than the region size, which provides tremendous computational savings. In the balance of the paper, we provide a detailed description of our system, present illustrative examples and experimental results, and finally conclude with some guidelines describing the further work.

II. Algorithm description

Fig. 1 outlines the main modules of the system, designed to process data incrementally as it is acquired. This calls for both local and global computations, where extracted local information guides the global estimation. The first few steps are applied to pairs of consecutive images to compute frame-to-frame (F2F) disparities (optical flow); obtain local depth-maps; estimate F2F camera motions; and extract the initial set of image features. The second part deals with global computations: To estimate the 3-D positions of the feature

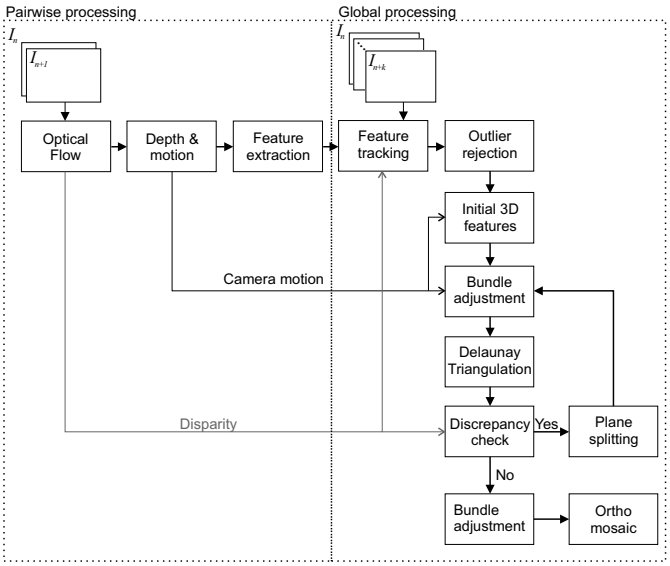


Fig. 1. Flowchart of proposed algorithm.

points and to utilize them in generating a 3-D model with nonuniform surface meshes. The map is then adjusted iteratively to improve consistency with the dense data. Finally, an ortho-mosaic is constructed. The detailed computations for each module of the proposed method is described hereafter.

A. Optical Flow, 3-D Motion and Depth Map Computation

The first step is the computation of the 2-D optical flow $\mathbf{v} = [u\ v]^T$ from pairs of images. The adopted GDIM-based method was proposed by Negahdaripour [16], and later generalized to take advantage of color in addition to intensity information for improved robustness and estimation accuracy [17]. The computed optical flow for each pair $\{I_n, I_{n+1}\}$ of consecutive images provides an estimate of local disparities for feature tracking and depth computation.

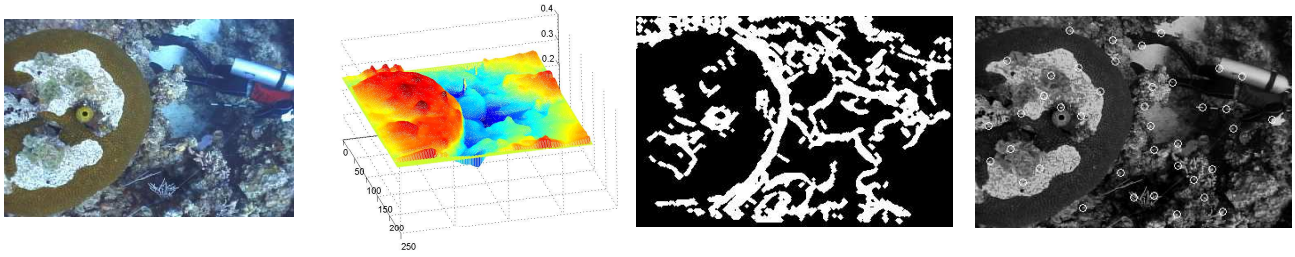
The differential image motion model of Longuet-Higgins is the basis of the motion and depth estimation module [18]:

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

Here, ω and t are the camera rotation and translation velocities respectively, and Z is the distance to a scene point along the optical axis. Utilizing image motion model in (1), F2F 3-D motions and depth maps are computed iteratively from the optical flow [6]. One should note that both the 3-D motion and depth maps are computed up to scale (due to the well-known scale-factor ambiguity of monocular vision). The correct scaling can be determined from a single distance (depth) measurement, or knowledge of motion magnitude. Fig. 2 illustrates an underwater image and its estimated dense depth map.

B. Feature Extraction and Tracking

The image features that initially characterize the scene are extracted using the Harris corner detector [19], also taking into account the topological information. The Harris corner



(a) (b) (c) (d)

Fig. 2. Depth map computation: Sample image of an underwater scene (a); depth map computed from two overlapping frames (b); boundary points of regions with high depth gradient magnitude (c) to extract surface features according to intensity cornerness measure (d).

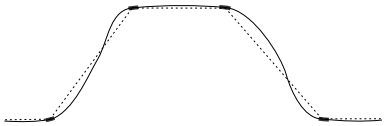


Fig. 3. Simple 2D example of ideal features extraction from topological point of view: 4 feature points provide a good initial piece-wise linear approximation of the curved profile.

detector basically yields points with high eigenvalues of the second moment matrix of image intensities. Use of solely this information provides a good basis for feature tracking and matching in most 2D mosaicing applications. However, we seek image features that additionally have the following properties:

- serve as vertices of planar patches that suitably represent the local surface;
- minimize characterization redundancy.

In order to better understand the concept, consider the simple example illustrated in fig. 3, 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. To locate the sloped edges, we use the first derivatives of the depth map, selecting only the regions with high depth gradient magnitude; see fig. 2(c). The edge of the segmented regions are then extracted, providing a mask for the Harris corner detector operator. The goal is to locate a number of features that have optimal "photometric"[1] and topological properties; they can be readily tracked visually, and also serve as appropriate initial nodes in splitting the surface into planar patches (see fig. 2(d) and also section *Surface Splitting*).

In our approach we take advantage of the optical flow to confine the search for the match of a point $p = (x, y)$ in the second view by normalized correlation: The search is centered at $p' = [x+u, y+v]$ in a $(C_u+1)\times(C_v+1)$ window; (C_u, C_v) presents the pixel position uncertainty determined from the optical flow estimation error covariance. By using a suitable search window that is adjusted locally based on the optical flow and its uncertainty, the probability of outliers is reduced while using small correlation windows. We also avoid the need for higher-order transformations, e.g., affine or projective, that may often be necessary within regions of high surface relief.

We next check for outliers, to ensure robust tracking as each feature point position p^{n+1} is determined in image I_{n+1}. This is carried out by evaluating the first order approximation of the geometric error d (Sampson distance) [5]:

$$d = \frac{(p^{n+1})^T F p^n}{|Fp^n|^2 + |Fp^{n+1}|^2} \tag{2}$$

where F is the fundamental matrix:

$$F = (K^{-1})^T S R K^{-1} \tag{3}$$

Here, K is the camera intrinsic matrix, S is the translation skew-symmetric matrix ($Sx = t \times x$ for any vector x), and R is the rotation matrix of the camera. (R, t) is computed in the motion and depth estimation module.

C. 3-D Feature Estimation

The 3-D features reconstruction is carried out by means of bundle adjustment [14]. In our case, the problem can be stated as a large-scale non-linear estimation having $N = 6N_C+3N_P$ unknowns where N_C and N_P are the numbers of camera and 3-D point positions, respectively, with $M = 2\sum_{i=1}^{N_c} N_{pi} > N$ redundant observations (N_{pi} is the number of features in camera view i). The cost function is defined based on the projection error of the features:

$$e_p = |\tilde{p} - C\tilde{P}| \tag{4}$$

where C is the camera matrix and $\{\tilde{P}, \tilde{p}\}$ denotes the homogenous coordinates of $\{3\text{-D},2\text{-D}\}$ points $\{P, p\}$.

In order to achieve convergence and accurate results, the bundle adjustment has to be provided with a good initial estimate of the unknown parameters. We can use the estimated 3-D motions and local depth maps. For a better estimate of each 3-D point, we triangulate using the matched 2-D projections from two views with the largest disparity.

D. 3-D Surface Construction and Adjustment

The previous module generates sparse 3-D points. The surface model is generated using the Delaunay technique with the 3-D points as the vertices of triangular patches. The next step is to examine how well the estimated model agrees with the dense depth maps and to adjust it suitably.

[1]We use this term loosely to imply strong local texture.

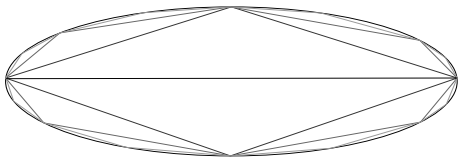


Fig. 4. Ellipse contour approximation

The surface modelling is inspired by the 2-D curved contour representation problem, demonstrated with the ellipse in fig. 4. Starting as the initial estimate with the line connecting the two points where the ellipse intersects its major axes, the next contour point is selected where the distance between the ellipse and the line segments (discrepancy) is maximum. At each iteration, exactly one line segment is replaced by two new lines connected at the point of maximum distance. The process continues until no distance between the line and the ellipse is larger than a pre-specified threshold. The final result is an effective representation of the curve by a chain of vertices, where the lines join. In our application, the use of a depth map for discrepancy test presents an important disadvantage: Being a local map, there exists viewpoint-dependent ambiguities due to scale and local surface orientation. A more effective approach is to assess the discrepancy in the image domain, by comparing the dense disparities predicted by the estimated planar model with the measured optical flow.

Each planar patch $L(P_A, P_B, P_C)$, defined by the vertices P_A, P_B, P_C, is projected in all nearby camera views. The best view I_k is chosen where the patch projection $l_k(p_A, p_B, p_C)$ has maximal area, offering the highest resolution for discrepancy computation. For each pixel $p = (x, y)$ within l_k, the discrepancy is defined by

$$e = \left| \begin{array}{c} u_k - \widehat{u}_k \\ v_k - \widehat{v}_k \end{array} \right| \tag{5}$$

(u_k, v_k) represents the estimated optical flow at p in image I_k, and $(\widehat{u}_k(x,y), \widehat{v}_k(x,y)$ is the predicted disparity of the planar patch:

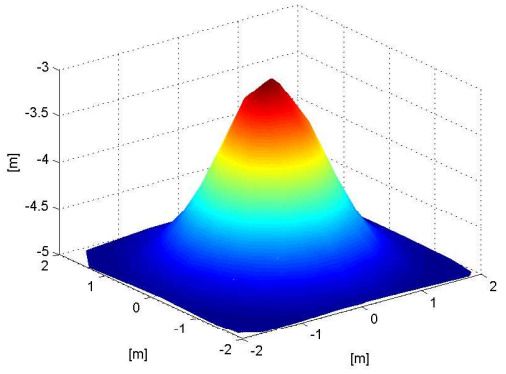
$$\widehat{v}_k = \left[\begin{array}{c} \frac{h_1 \cdot p}{h_3 \cdot p} \\ \frac{h_2 \cdot p}{h_3 \cdot p} \end{array} \right] - p \tag{6}$$

where h_i's are the rows of plane homography:

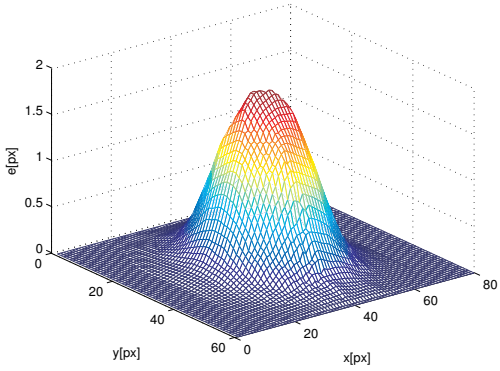
$${}^{k+1}H_k^L = R - tn_L^T \tag{7}$$

R and t are respectively the relative rotation matrix and translation vector between camera positions k and $k+1$ and n_L is the normal to the plane L. Moving parallel to the surface in fig. 5(a), we obtain the discrepancy map in (b), if the flat plane at the base of the true surface is assumed as the 3-D model.

The function e depends on the 3-D distance between the real surface and the estimated plane along camera principal axis, and is camera rotation invariant. As a result, an ideal point p_s to split the surface is where e is maximum. However, this measure alone does not necessarily yield a locally distinct point that can be readily tracked in order to establish its 3-D



(a)



(b)

Fig. 5. Synthetically generated surface (a), and Computed discrepancy e based on a flat surface model for an arbitrary camera motion (b).

position and the camera trajectory by bundle adjustment. For this reason, a revised measure is introduced:

$$m = e' + \mu c \tag{8}$$

where e' is obtained by Gaussian smoothing and normalization of e, and c is the normalization of the cornerness measure from the Harris feature detector [19]. The parameter μ may be set based on desired characteristics, as in a suitable function of distance from the peak-point of e. Here we have selected μ empirically. To avoid splitting the planes indefinitely, a threshold t_h is applied: There is no further splitting *if* $max(e') < t_h$. To summarize, the discrepancy of each plane from the surface is computed and thus split accordingly, the surface representation in updated with the newly extracted 3-D points and the fit of the new model to the depth data is checked. The algorithm iterates until the model stabilizes, *i.e.*, no new 3-D points are added to the model.

E. Ortho-mosaic Construction

The construction of the ortho-mosaic can be summarized in two main steps:

1) Selection of the projection plane O for the 2-D ortho-view;
2) Rendering of the ortho-mosaic.

TABLE I

OUTLINE OF THE SURFACE CONSTRUCTION AND ADJUSTMENT ALGORITHM

```
while adding new points
    perform Delaunay triangulation
    for each plane L
        find best view I_k
        compute e
        if max(e') > t_h
            compute measure m
            find new p_s
            track p_s in the sequence
            compute P_s by bundle adjustment
        end
    end
end
```

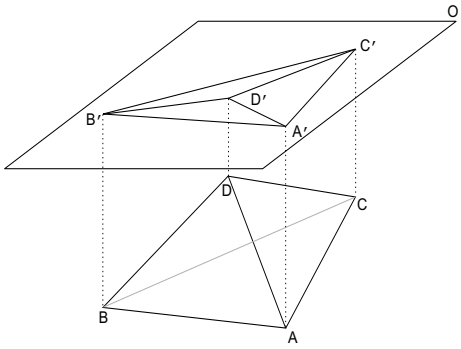


Fig. 6. Example: Ortho-projection of tetrahedron $ABCD$ onto plane O parallel to the base of tetrahedron.

The plane O is chosen to have the same tilt as the 3-D reconstructed surface. This maximizes the projection area, providing the highest level of mosaic detail. All the planar patches forming the 3-D model are mapped onto the destination plane along the projection rays parallel to the normal vector; see the example in fig. 6 for the ortho-projection of a tetrahedron onto the plane O. Note that the points $[ABCD]$ are projected along rays perpendicular to O.

The plane O is digitized based on a predefined resolution; each point m on the grid is a pixel in the ortho-mosaic. In order to render the mosaic, the following transformation relating each point m to a corresponding point p from the original images is defined:

$$p = C_k T_n p_m \tag{9}$$

where T_n is the ortho-projection transformation of the patch L_n and C_k is the camera projection matrix corresponding to the k-th view; see fig. 7. The remaining problem is to determine which view I_k to use for rendering the patch L'_n. The decision criteria has been set such that we minimize the distortions introduced by projecting from the original images onto the ortho-mosaic. An affine homography is computed for each camera view k where a surface patch L_n is visible. This homography describes the transformation between the projection in camera view and ortho-projection L'_n according

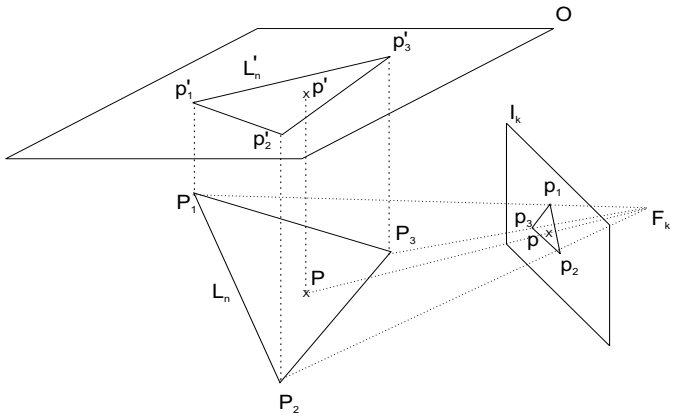


Fig. 7. Ortho-mosaic rendering process.

to

$$[p_k] =^k A_n [m_n]; \tag{10}$$

Using Singular Value Decomposition [5], ${}^kA_n$ can be expressed as

$${}^kA_n = E \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos(\alpha) & -sin(\alpha) & 0 \\ sin(\alpha) & cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{11}$$

where E is the homography defining an Euclidian transformation, α represents the affine rotation, and s_x and s_y represent the anisotropic scaling factors in x and y directions respectively. The distortion measure is defined as

$$d_k^n = 1 - \frac{min(s_x, s_y)}{max(s_x, s_y)} \tag{12}$$

The camera view with the lowest value of d_k^n, corresponding to the least rendering distortion, is chosen.

III. EXPERIMENTAL RESULTS

The testing of the ortho-mosaicing system is carried out in two ways:

- synthetic data, mainly focused on quantifying the efficiency of the surface splitting module;
- real data, assessing the proposed approach in entirety.

A. Synthetic Data

The objective of these experiments are to investigate the efficiency of the surface splitting algorithm and the discrepancy evolution using ground truth data. The surface topography, given in fig. 8a, resembles a natural terrain. In every case, we start with a flat-plane surface model and iteratively adjust it with new vertices based on the discrepancy measure, as described in section II-D. The termination criteria is established by threshold t_h, which is chosen to maintain a balance between the final discrepancy and model complexity.

We first show the results when the splitting is based merely on the surface shape properties. This corresponds to the ideal case where any selected terrain feature has sufficiently strong texture to be accurately tracked and matched in multiple views. The final model comprising 74 vertices is obtained after

11 iterations, corresponding to a "normalized" discrepancy measure (NDM) at 2.16% (with respect to the initial flat surface); see fig. 8b. This can be further reduced at the cost of increased model complexity. For example, the NDM is marginally reduced to 1.47% in 20 iterations, with a total of 212 vertices.

The second set of experiments focuses on examining how the intensity gradient measure may influence the behavior of the algorithm, utilizing 2 different texture maps. For a relatively rich and homogeneous texture, the algorithm generates an uniformly adjusted model with a final NDM at 3.29%; see fig. 8(c,d). The second case involves a weaker texture with various low-contrast regions. The final result is somewhat less accurate, with NDM=5.47%; see fig. 8(e,f). The variation in NDM performance can be observed in fig. 9. In the two cases, the algorithm reaches the converged values in 10 and 7 iterations, respectively, without reaching the threshold (fixed as in the ideal case). This is mainly because no new points for surface splitting can be identified with sufficient texture to be tracked robustly in the sequence.

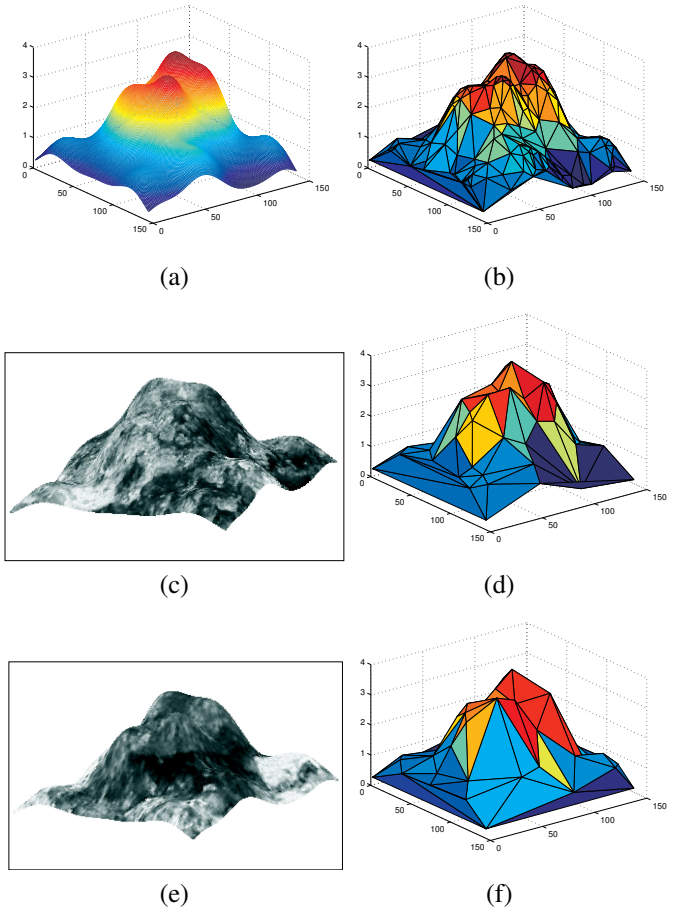


Fig. 8. Experimental results with synthetic data: (a) True surface; (b) Computed model with 74 vertices after 10 iterations of surface splitting algorithm, and selection of surface feature points (vertices) based solely on topographical measure; see discrepancy function in (5). Same surface rendered with two different texture maps (c,e), and computed surface models based on combined topographical and radiometric measure (d,f); see (8).

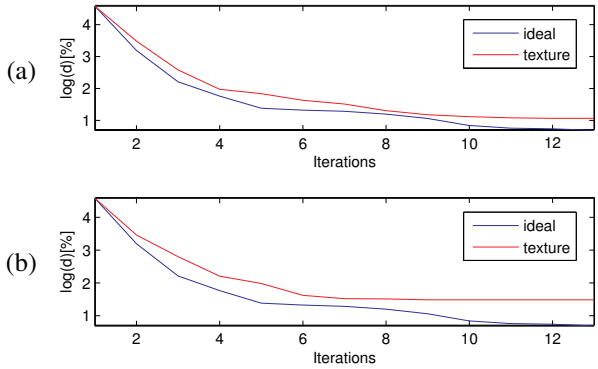
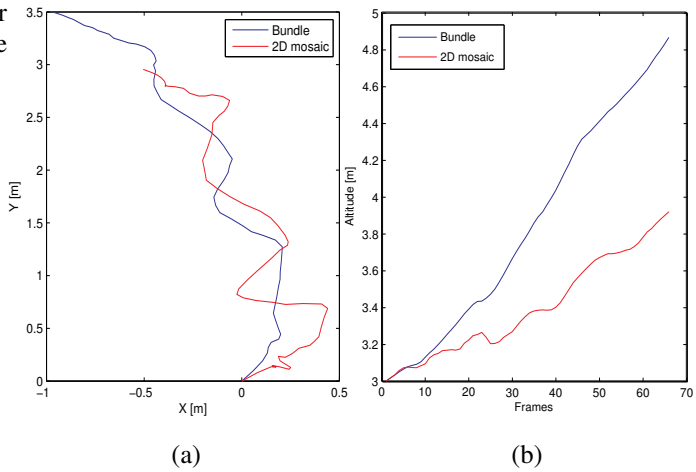


Fig. 9. Logarithmic plot of mean discrepancy measure as a percentage of initial flat surface discrepancy (NDM)-vs- number of surface splitting iterations. Comparison between use of geometric information solely (ideal) and two cases with different surface texture richness for computing the surface models given in fig. 8



(a) (b)

Fig. 10. Estimated camera trajectories for coral reef scene (2D mosaic -vs- 3D bundle adjustment): (a) projection onto XY plane; (b) Z component.

Two important issues should be highlighted:

- Poor texture results in sparse high gradient points and regions that reduce the algorithm efficiency;
- The regions with highest discrepancy in the final model correspond to quasi-null texture areas (low contrast regions, saturated regions, etc.), where the algorithm is not able to find reliable splitting points.

B. Real Data

We now present the experiment carried out using an image sequence of a coral reef site in the Bahamas acquired with a handheld camcorder. The sequence comprises 66 images of 360×240 pixels, covering an area of approximatively 4×7 meters.

The first result, depicted in fig. 13(a), is the photo-mosaic generated with a traditional 2-D mosaicing technique. Inaccurate estimate of image deformations are obtained due to inability to correctly model the terrain relief by the simplistic homography models, thus producing excessive distortions and various misalignments at the mosaic seams (some regions with distinct discontinuities have been highlighted). This drawback

becomes even more obvious when analyzing the erroneously estimated camera trajectory, in comparison to the estimate by bundle adjustment; see fig. 10.

Applying the method described in the paper to this image sequence, 76 features are initially identified and tracked. After 5 iterations of plane splitting, the resulting surface model comprise 184 vertices and 356 planar patches. The 3-D model is depicted in fig. 12(a,b) from two views, with the estimated camera trajectory superimposed on the first. Fig. 13(b) illustrates the ortho-mosaic, generated from the 3-D model. Noticeable distortions are present at the extremities of the mosaic, due to two reasons:

1) For some planes, we have $min(d_k^n) \gg 0$ meaning that none of the camera views can provide suitable information to render the surface (see section II-E);
2) Most of the 3-D points defining planes at the extremities of the ortho-mosaic are surface features that are viewed at low gazing angles, and are thus poorly reconstructed by triangulation from multiple views;

It is concluded that for this sequence comprising a single short transect, there is insufficient and less reliable observations near the boundaries to be able to project the regions accurately onto the ortho-mosaic. To map the entire region with uniform precision, one needs to plan transects over these areas to acquire several near upright views. This allows the selection and accurate 3-D estimation of suitable features to model the local terrain surfaces.

To reduce the distortion effect, some post-processing of the 3-D model can be done by assuming that these "inaccurate" 3-D points at the extremities of the model are (nearly) coplanar. This gives the modified 3-D model in fig. 12(c), which can be compared with the pre-processed model in (b). This gives the improved ortho-mosaic illustrated in fig. 13(c).

IV. Conclusions and Further Work

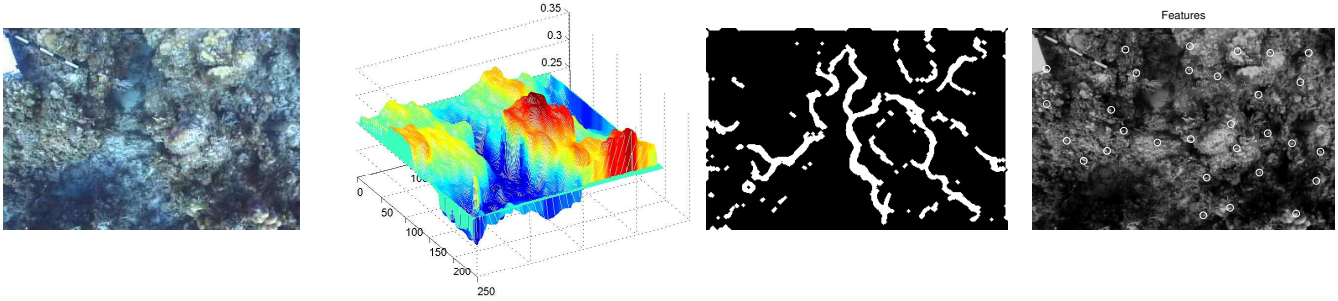
A framework based the integration of dense depth computation and 3-D feature reconstruction from monocular views has been proposed for seafloor ortho-mosaic construction over large survey areas. The key aspect of the methodology is to adequately model the terrain surface with minimal complexity. Our implementation builds on the use of dense local depth maps to select suitable features as initial vertices of piecewise planar meshes, tracking these features over the video sequence to accurate estimate their 3-D positions by bundle adjustment, and iteratively revising the model to obtain an accurate representation. The 3-D surface model provides the basis for the construction of the ortho-mosaic. Our experimental results validate the effectiveness of the proposed approach. Topics from ongoing work deal with improving the various modules of the algorithm:

- Incorporating more robust feature tracking techniques (*i*) in the presence of high affine transformations and (*ii*) to allow matching over surface regions with weaker texture, thus increasing efficiency in surface splitting module;
- Using higher-order surface models to allow continuity of surface gradients and a better surface representation;
- Implementation of a scene consistent Delaunay triangulation algorithm;
- Automated detection/correction of rendering distortions;
- Compensation of scene illumination changes to enhance the ortho-mosaic.

Acknowledgement: This work has been funded in part by the Spanish Ministry of Education and Science (MEC) under grant CTM2004-04205, the *Generalitat de Catalunya* under grant 2003PIV-B00032, the Office of Naval Research under Grant No. N000140310074-02, and the DoD Strategic Environmental Research and Development Program (SERDP) under grant CS-1333. Opinions, findings, conclusions or recommendations expressed in this manuscript are those of the author and do not necessarily reflect the views of any of the sponsors.

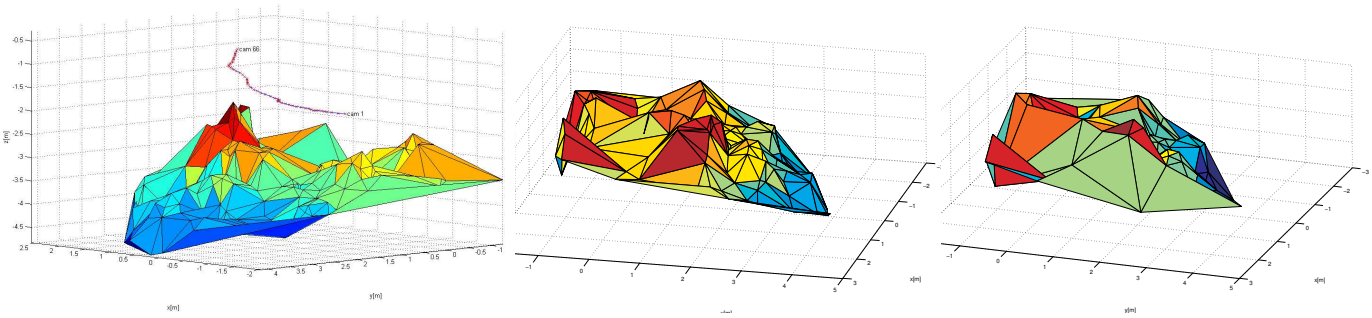
References

[1] J. Michel and R. Ballard, "The rms titanic 1985 discovery expedition," in *MTS/IEEE Oceans*, Sept. 1994, pp. 132–137.
[2] I. Williams and J. Leach, "Measurements of benthic species using drop-video platforms- comparative uses of stereo and single video systems," in *Proc. Australian Marine Science Association, Sydney, Australia*, 1999.
[3] D. Capel, "Image mosaicing and super-resolution," Ph.D. dissertation, University of Oxford, 2001.
[4] N. Gracias and J. Santos-Victor, "Underwater mosaicing and trajectory reconstruction unsing global alignment," in *MTS/IEEE Oceans*, 2001, pp. 2557–2563.
[5] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.
[6] K. Kanatani, "Structure and motion from optical flow under perspective projection," in *Computer Vision, Graphics, and Image Processing*, 1987, pp. 122–146.
[7] M. Holm, G. Denissoff, K. Juslin, M. Paljakka, R. M., and S. Rautakorpi, "Ortho-mosaics and digital elevation models from airborne video imagery using parallel global object reconstruction," *International Archives of Photogrammetry and Remote Sensing*, vol. 31(3), pp. 331–336, 1996.
[8] E. Mikhail, J. Bethel, and J. McGlone, *Introduction to Modern Photogrammetry*. John Wiley & Sons, New York, USA, 2001.
[9] Z. Qin, W. Li, M. Li, Z. Chen, and G. Zhou, "A methodology for true orthorectification of large-scale urban aerial images and automatic detection of building occlusions using digital surface model," in *IEEE Geoscience Remote Sensing Symposium*, 21-25 July 2003, pp. 417–438.
[10] C. Stoker, D. Barch, B. Hine III, and J. Barry, "Antactic undersea exploration using a robotic submarine with a telepresence user interface," in *IEEE Expert: Intel. Syst. and their Appl.*, vol. 10(6), 1995, pp. 14–23.
[11] S. Negahdaripour and H. Madjidi, "Stereovision imaging on submersible platforms for 3-d mapping of benthic habitats and sea-floor structures," *IEEE J. Oceanic Engineering*, vol. 28(4), October 2003.
[12] A. Khamene and S. Negahdaripour, "Motion and structure from multiple cues; image motion, shading flow, and stereo disparity," in *Computer Vision and Image Understanding*, vol. 90(1), May 2003, pp. 122–146.
[13] O. Pizzaro, "Large scale structure from motion for autonomous underwater vehicle surveys," Ph.D. dissertation, MIT and Woods Hole Oceanographic Institution, September 2004.
[14] B. Triggs, A. Zisserman, and R. Szeliski, Eds., *Bundle Adjustment: A Modern Synthesis*, 2000.
[15] D. Terzopoulos, "The computation of visible-surface representations," in *IEEE Trans. Pattern Analysis Machine Intel.*, July 1988, pp. 417–438.
[16] S. Negahdaripour, "Revised interpretation of optical flow; integration of radiometric and geometric cues," in *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 20, no. 9, September 1998.
[17] S. Negahdaripour and H. Madjidi, "Robust optical flow estimation using underwater color images," in *MTS/IEEE Oceans*, 22-26 Sept 2003, pp. 2309 – 2316.
[18] H. Longuet-Higgins and K. Prazdny, "The interpretation of a moving retinal image," in *Proc. Royal Society of London*, vol. B-208, 1980, pp. 385–397.
[19] C. Harris and M. Stephens, "A combined corner and edge detector," in *4th Alvey Vision Conference*. Matchester, U.K., 1988, pp. 147–151.



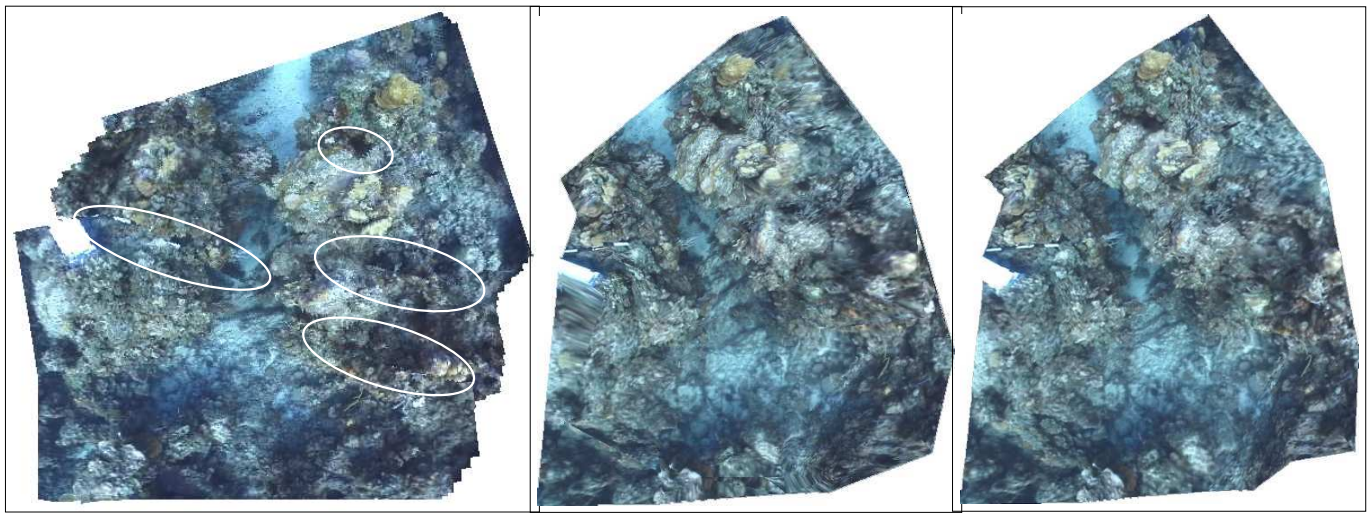
(a) (b) (c) (d)

Fig. 11. Selected results for real data: (a) A sample frame, and (B) corresponding local depth map. (c) Boundary points of regions with large depth gradient magnitude are used to select the initial set of features in (d) for this view.



(a) (b) (c)

Fig. 12. (a) 3-D model of surveyed area and estimated camera trajectory; (b) Same model from lateral view; (c) Model after post-processing, assuming extremity regions to be coplanar to minimize rendering distortions.



(a) (b) (c)

Fig. 13. 2-D mosaic of the coral reef sequence in surveyed area, with discontinuities in circled areas (a). Ortho-mosaic before (b) and after (c) post-processing.