Large Area 3-D Reconstructions From Underwater Optical Surveys

Oscar Pizarro, *Member, IEEE*, Ryan Michael Eustice, *Member, IEEE*, and Hanumant Singh, *Member, IEEE*

***Abstract*—Robotic underwater vehicles are regularly performing vast optical surveys of the ocean floor. Scientists value these surveys since optical images offer high levels of detail and are easily interpreted by humans. Unfortunately, the coverage of a single image is limited by absorption and backscatter while what is generally desired is an overall view of the survey area. Recent works on underwater mosaics assume planar scenes and are applicable only to situations without much relief. We present a complete and validated system for processing optical images acquired from an underwater robotic vehicle to form a 3-D reconstruction of the ocean floor. Our approach is designed for the most general conditions of wide-baseline imagery (low overlap and presence of significant 3-D structure) and scales to hundreds or thousands of images. We only assume a calibrated camera system and a vehicle with uncertain and possibly drifting pose information (e.g., a compass, depth sensor, and a Doppler velocity log). Our approach is based on a combination of techniques from computer vision, photogrammetry, and robotics. We use a local to global approach to structure from motion, aided by the navigation sensors on the vehicle to generate 3-D submaps. These submaps are then placed in a common reference frame that is refined by matching overlapping submaps. The final stage of processing is a bundle adjustment that provides the 3-D structure, camera poses, and uncertainty estimates in a consistent reference frame. We present results with ground truth for structure as well as results from an oceanographic survey over a coral reef.**

***Index Terms*—Computer vision, structure from motion, 3-D reconstruction, underwater vehicles.**

I. Introduction

A. Context

OPTICAL imaging of the ocean floor offers scientists high levels of detail and ease of interpretation. However, light underwater suffers from significant attenuation and backscatter, limiting the practical coverage of a single image to only a few square meters [1]. For many scientific surveys the area of interest is much larger, and can only be covered by hundreds or thousands of images acquired from a robotic vehicle or towed sled. Such surveys are required to study hydrothermal vents and spreading ridges in geology [2], ancient shipwrecks and settlements in archeology [3], forensic studies of modern shipwrecks and airplane accidents [4], [5], and surveys of benthic ecosystems and species in biology [6], [7].

The visible spectrum in water has attenuation lengths of the order of meters, thus most underwater vehicles carry out optical imaging surveys using their own light source. Apart from casting shadows that move across the scene as the vehicle moves, power and/or size limitations lead to lighting patterns that are far from uniform. Also with the advent of autonomous underwater vehicles (AUVs) for imaging surveys [2], [7] additional constraints are imposed by the limited energy budget of an AUV. AUV surveys are typically performed with strobed light sources rather than continuous lighting, and acquire low overlap imagery to preserve power and cover larger areas.

Generating a composite view by exploiting the redundancy in multiple overlapping images is usually the most practical and flexible way around this limitation. Recent years have seen significant advances in mosaicing [8], [9] and full 3-D reconstruction [10, ch. 9], [11] though most of these results are land based and do not address issues particular to underwater imaging. Underwater mosaicing has been motivated largely by vision-based navigation and station keeping close to the sea floor [12]–[15]. The large area mosaicing problem with low overlap under the assumption of planarity is addressed in [16]. Mosaicing assumes that images come from an ideal camera (with compensated lens distortion) and that the scene is planar [17]. Under these assumptions, the camera motion will not induce parallax. Thus, no 3-D effects are involved and the transformation between views can then be correctly described by a 2-D homography. These assumptions often do not hold in underwater applications since light attenuation and backscatter rule out the traditional land-based approach of acquiring distant, nearly orthographic imagery. Underwater mosaics of scenes exhibiting significant 3-D structure usually contain significant distortions.

In contrast to mosaicing, the information from multiple underwater views can be used to extract structure and motion estimates using techniques from structure from motion (SFM) and photogrammetry [18]. We propose that when dealing with a translating camera over nonplanar surfaces, recovering 3-D structure is the proper approach to providing a composite global view of an area of interest. The same challenges seen in mosaicing underwater apply to SFM underwater with the added

Manuscript received June 17, 2008; revised November 29, 2008; accepted February 06, 2009. Current version published May 13, 2009. This work was supported in part by the Center for Subsurface Sensing and Imaging Systems (CenSSIS) Engineering Research Center of the National Science Foundation under Grant EEC-9986821, and in part by the MIT Presidential Fellowship. This paper was presented in part at the IEEE OCEANS Conference, Kobe, Japan, November 2004.

Associate Editor: B. R. Calder.

O. Pizarro was with the Joint Program in Oceanographic Engineering of the Massachusetts Institute of Technology, Cambridge, MA 02139 USA and the Woods Hole Oceanographic Institution, Woods Hole, MA 02543 USA. He is now with the Australian Centre for Field Robotics, The University of Sydney, Sydney, N.S.W. 2006, Australia (e-mail: o.pizarro@cas.edu.au).

R. M. Eustice was with the Joint Program in Oceanographic Engineering of the Massachusetts Institute of Technology, Cambridge, MA 02139 USA and the Woods Hole Oceanographic Institution, Woods Hole, MA 02543 USA. He is now with the Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: eustice@umich.edu).

H. Singh is with the Department of Applied Ocean Physics and Engineering, Woods Hole Oceanographic Institution, Woods Hole, MA 02543 USA (e-mail: hanu@whoi.edu).

Digital Object Identifier 10.1109/JOE.2009.2016071

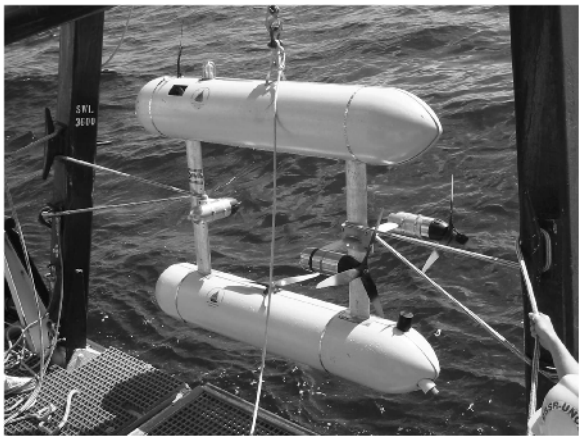


Fig. 1. SeaBED vehicle ready for deployment in Bermuda.

requirement that all scene points must be imaged at least twice to produce a roughly uniform distribution of reconstructed feature points through triangulation (50% overlap in the temporal image sequence). These techniques are considerably more complex than mosaicing: even for land-based applications (with high overlap, structured motion, and uniform lighting) consistency at large scales cannot be guaranteed unless other sensors are available. Some promising work has gone into 3-D image reconstruction underwater [19] using a stereo-rig with high overlap imagery in a controlled environment.

Underwater vehicles for scientific surveys use navigation sensors that provide pose estimates. This information can be used to constrain and regularize the underwater structure from motion problem. In previous work [20], [21], we showed in detail how to improve the search for corresponding features between images using pose estimates. In addition, we used navigation sensors to provide estimates of baseline magnitude and to select a unique solution in cases where there is ambiguity in the image-based solution.

B. Imaging Platform

The SeaBED AUV acquired the field data used in this paper (Fig. 1). The vehicle was designed as a calibrated and pose-instrumented platform for underwater imaging. SeaBED is capable of maneuvering at slow speed and is passively stable in pitch and roll. The vehicle specifications are summarized in Table I. SeaBED collected the field data used in this paper following survey patterns preprogrammed as a mission and executed in dead-reckoning mode. The vehicle makes acoustic measurements of both velocity and altitude relative to the bottom. Absolute orientation is measured within a few degrees using a magneto–inductive compass and inclinometers, while depth is obtained from a pressure sensor.

C. Outline

Our methodology (Fig. 2) takes a local-to-global approach inspired by mosaicing [9] and SFM [11], [22] but takes advantage of navigation and attitude information. The 3-D structure of local subsequences is derived independently and then registered in a global frame for bundle adjustment. Our approach is more suitable than purely sequential methods [23] because in a typical underwater survey each 3-D feature appears only in few images and each image only contains a small fraction of all features making the global solution approach a series of weakly correlated local solutions.

TABLE I
SUMMARY OF THE SEABED AUV SPECIFICATIONS

Vehicle	
Depth rating	2000 meters
Size	2.0 m (L) × 1.5 m (H) × 1.5 m (W)
Mass	200 kg
Maximum Speed	1.0 m/s
Batteries	2 kWh Li-ion pack
Propulsion	3 150 W brushless DC thrusters
Navigation	
Attitude+Heading	Tilt ± 0.5°, Compass ± 2°
Depth	Paroscientific pressure sensor, 0.01%
Velocity	RDI Navigator ADCP $\pm 1 - 2mm/s$
Angular rates	Crossbow 3-axis gyro
Altitude	RDI Navigator
Optical Im.	
Camera	Pixelfly 12bit 1280 × 1024 CCD
Lighting	one 200 Ws strobe
Separation	1m between camera and light
Acoustic Im.	
Sidescan sonar	MST 300 kHz (300 m depth rating)
Pencilbeam sonar	Imagenex 881 675 kHz
Other Sensors	
CT	Seabird 37SBI

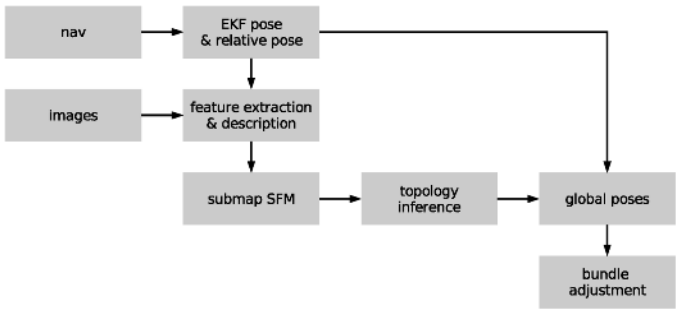


Fig. 2. Flowchart of structure and motion recovery from underwater imagery. An image sequence is processed into short submaps of structure and motion aided by navigation information. Submaps are then matched to infer and refine additional spatial constraints (such as loop closures and overlapping parallel tracklines). An initial estimate of poses and structure in a global frame is then used to perform a final bundle adjustment.

The following sections briefly describe our approach focusing on feature extraction and description, submap generation based on two and three view processing, topology exploration, and local-to-global registration. The last section presents results from a coral reef survey and validation of the proposed framework by tank experiments with associated ground truth.

II. FEATURE EXTRACTION AND DESCRIPTION

We calculate the relative pose between images using a feature-based approach under wide-baseline imaging conditions with changing illumination and unknown scene structure. A

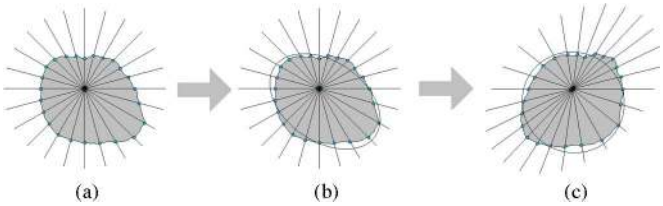


Fig. 3. Steps in determining an affine invariant region. (a) boundary points determined along rays. (b) An ellipse approximates the boundary points using the method of moments. (c) The elliptical region is mapped onto a unit disc.

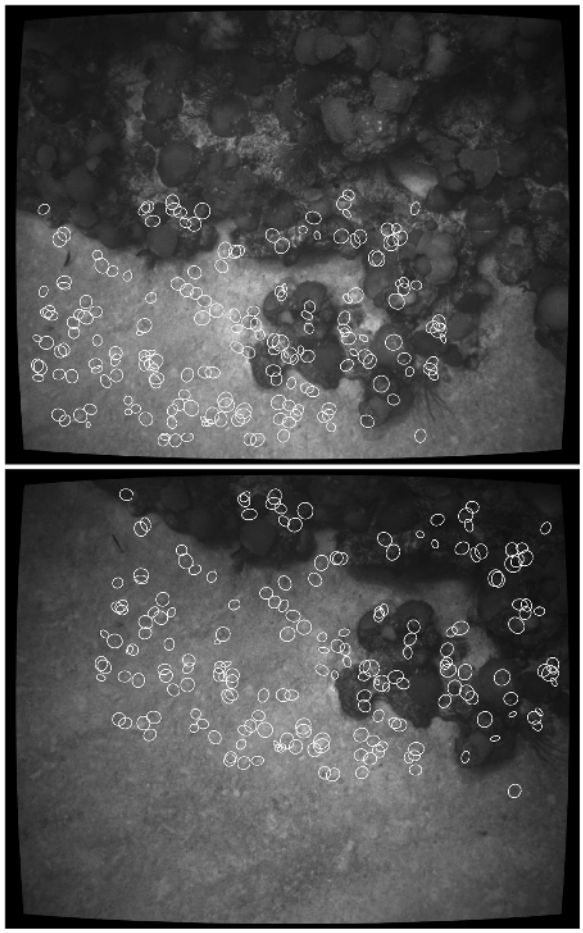


Fig. 4. Extracted affine invariant regions. Only regions that are found in correspondence are shown. Scene dimensions $\approx$ 2.5 m $\times$ 2 m.

modified Harris corner detector [24] yields well-localized, repeatable interest points by selecting local maxima of the smaller eigenvalue of the second moment matrix. To describe each interest point we determine a neighborhood around it that is invariant to affine geometric transformations using a modified version of the method proposed by Tuytelaars [25], [26] (Fig. 3). The original method defines an affine invariant region around an intensity extreme point by determining affine invariant points along rays radiating from the intensity extremum. The boundary point associated with a ray corresponds to the extremum of an affine invariant function that can be related to the presence of a boundary (Figs. 4 and 5). The boundary points along the rays $r_{\text{invariant}}(\theta)$ are given by

$$r_{\text{invariant}}(\theta) = \arg_r \max |f(r,\theta) - f_o| \qquad (1)$$

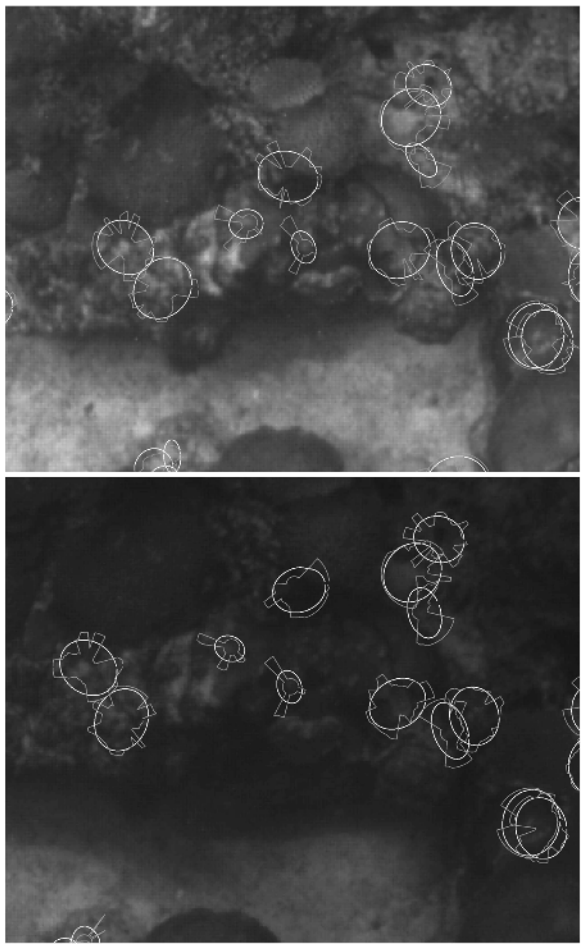


Fig. 5. Detail of some extracted regions from the images in Fig. 4. The actual border samples outline jagged regions. The elliptical regions that approximate the border samples are also shown. Scene dimensions $\approx$ 0.6 m $\times$ 0.5 m.

where f_o is the extremum of intensity and $f(r,\theta)$ are the image values considered in polar coordinates. This region is extracted in an affine invariant manner in the sense that an affine transformation will "stretch" the individual rays but the boundary points should remain recognizable since points that form a ray remain in a ray when affinely transformed (colinearity is preserved and any translation should be accounted for by the repeatable interest point detector).

For natural scenes few interest points correspond to sharp corners of planar surfaces. Instead they are generally blob-like features at different scales. By using rays radiating from the interest point instead of an intensity extremum, the matching procedure is simplified since the feature is well localized. In essence, we sample the neighborhood along lines radiating from the interest point. Our current implementation uses a diameter of 51 pixels (i.e., rays that are 25 pixels in length without sampling the central pixel) and samples every 6° (for a total of 60 lines). For each line, the boundary point corresponds to the maximum difference in intensities between the intensity extremum nearest to the interest point and points along the ray.

The set of maximal points is approximated with an elliptical neighborhood by using the method of moments where the samples along the boundary are placed on an ellipse that has the

same second moment matrix as the original samples. This elliptical region is then mapped onto the unit circle W. In practice, the polar representation used to determine the boundary is resampled so that the boundary points have the same radius instead of applying a 2-D affine transformation to the region. The canonical form of the region is stored as a polar representation with resampled radii. This representation is particularly convenient when the description of the region is based on Zernike moments since the basis functions are presented more compactly in polar form (Section II-B).

To increase robustness to changes in lighting, before calculating descriptors of the patch W the resampled intensities $f(x,y)$ are demeaned and normalized by the energy content in the patch

$$N\left(f(x,y)\right) = \frac{\left(f(x,y) - \bar{f}_W\right)}{\sqrt{\sum_{i,j}\left(f(x+i,y+j) - \bar{f}_W\right)^2}} \tag{2}$$

where $\bar{f}_W$ is the mean of $f(x,y)$ over the patch W. The normalized patch satisfies

$$N\left(f(x,y)\right) = N\left(af(x,y) + b\right) \tag{3}$$

effectively providing invariance to affine changes in intensity. Fig. 5 illustrates several matches despite significant lighting changes between extracted regions.

A. Orientation Normalization

Navigation instruments provide attitude information that can simplify the description and matching of features. For example, normalized correlation as a similarity metric fails in the presence of modest rotations (more than a few degrees) between an image pair I and I'. It is possible to use descriptors that are invariant to rotations at the price of less discrimination. However, knowledge of 3-D orientation for camera frames c and c' in a fixed reference frame w allows for normalization of orientation viewpoint effects via a homography.

The infinite homography H_∞ defined as [10, ch. 12]

$$\mathrm{H}_\infty = \mathrm{K}\,{}^b_a\mathrm{R}\,\mathrm{K}^{-1} \tag{4}$$

where ${}^b_a\mathrm{R}$ is the orthonormal rotation matrix from frame a to frame b and K is the camera calibration matrix [10, ch. 7] (containing intrinsic parameters for focal length, principal point coordinates, and skew in pixel shape), warps an image taken from camera orientation a into an image taken from camera orientation b. This warp is exact and independent of scene structure; there is no scene-induced parallax between viewpoints a and b, because a and b share the same projective center.

Given 3-D camera rotation matrices ${}^w_c\mathrm{R}$ and ${}^w_{c'}\mathrm{R}$ generated from vehicle orientation measurements, we can warp images I and I' each into a canonical viewpoint coordinate frame oriented parallel with frame w (e.g., the warped images correspond to a camera coordinate frame x, y, z oriented with north, east, down).

B. Description by Zernike Moments

We chose to use Zernike moments as descriptors as they are compact (generated from a set of orthogonal complex polynomials) and highly discriminating [16], [27]. Typical applications only use the magnitude of Zernike moments as this provides rotational invariance, but we can precompensate for orientation using attitude sensors, and therefore, utilize the full complex moments.

Zernike moments are derived from Zernike polynomials, which form an orthogonal basis over the interior of the unit circle, i.e., $x^2 + y^2 = 1$ [28]. If we denote the set of polynomials of order n and repetition m by $V_{nm}(x,y)$, then these polynomials are complex, and their form is usually expressed as

$$V_{nm}(x,y) = V_{nm}(\rho,\theta) = R_{nm}(\rho)e^{jm\theta} \tag{5}$$

with n a positive or zero integer, m an integer such that $n - |m|$ is even, and $|m| \leq n$. We have also defined polar coordinates $\rho = \sqrt{x^2+y^2}$, $\theta = \arctan(y/x)$. Note that $V^*_{nm}(\rho,\theta) = V_{n,-m}(\rho,\theta)$.

The radial polynomial $R_{nm}(\rho)$ is real and of degree $n \geq 0$, with no power of ρ less than $|m|$

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^s(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!}\rho^{n-2s}. \tag{6}$$

The Zernike moment of order n with repetition m corresponding to the projection of an image function $f(x,y)$ (in the unit circle) is given by

$$A_{nm} = \frac{n+1}{\pi}\int\!\!\int_{x^2+y^2\leq 1} f(x,y)V^*_{nm}(x,y)dxdy. \tag{7}$$

Note that A_{nm} is complex and $A^*_{nm} = A_{n,-m}$. In the case of a discrete image $f[x,y]$, the moments can be approximated for points inside the unit circle as

$$A_{nm} = \frac{n+1}{\pi}\sum_x\sum_y f[x,y]V^*_{nm}(x,y), \qquad x^2+y^2 \leq 1. \tag{8}$$

The orthogonality relation for V_{nm} permits reconstruction of an image from its Zernike moments

$$\int\!\!\int_{x^2+y^2\leq 1} V_{nm}(x,y)V^*_{pq}(x,y)dxdy = \frac{\pi}{n+1}\delta_{np}\delta_{mq} \tag{9}$$

so that

$$\hat{f}(x,y) = \sum_{n=0}^{\infty}\sum_m A_{nm}V_{nm}(x,y), \qquad x^2+y^2 \leq 1. \tag{10}$$

C. Similarity Measure

A vector of moments can be used directly as the descriptor for an image feature. Similarity between features can then be expressed as a distance between vectors. The problem with this

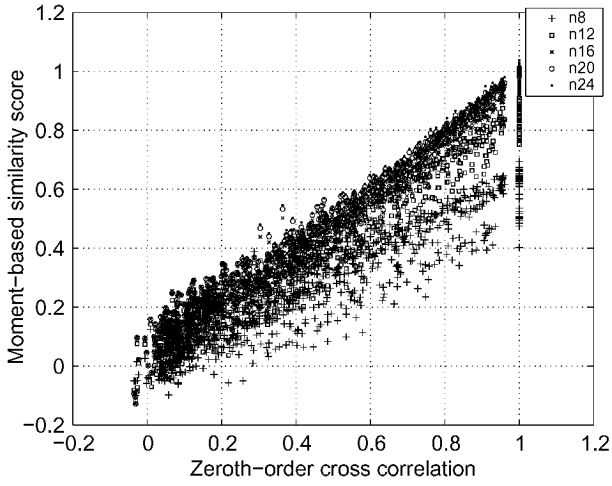


Fig. 6. Similarity score versus actual correlation score for varying number of coefficients. The approximation improves as more terms are added, in particular, for high correlations.

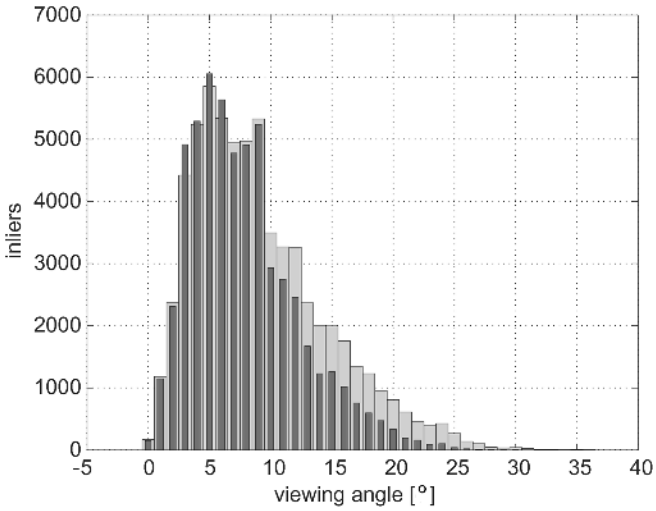


Fig. 7. For the matches classified as inliers it is possible to calculate the viewing angle change between cameras viewing the feature. For all matches, across all pairs in the trial, we show the number of inliers as a function of viewing angle. For narrow-baseline conditions (angles of 10° or less) both regions behave similarly. For larger viewing angles the affine invariant region (light gray wide bars) outperforms the fixed window method (dark gray narrow bars).

approach is that the distances between vectors of moments do not necessarily have an obvious meaning. Using training data it is possible to derive a distance metric [29] but this requires relearning the metric if the training set no longer represents the imagery. Instead, we determine that the cross correlation between image patches can be expressed conveniently by weighted Zernike moments and can form a feature descriptor from appropriately weighted moments.

We express the zeroth-order cross-correlation coefficient between image patches f and g in terms of their moments

$$S(f,g) = \int\int_{x^2+y^2\leq 1} f(x,y)g(x,y)dxdy \tag{11}$$

and by replacing $f(x,y)$ and $g(x,y)$ by their expansions in terms of Zernike moments [as in (10)], rearranging the sums and integrals, and using orthogonality of Zernike polynomials [(9)], we have

$$S(f,g) = \sum_n \sum_m A_{nm}(f)A^*_{nm}(g)\frac{\pi}{n+1} \tag{12}$$

where * denotes the complex conjugate.

This result suggests that we construct a vector of descriptors from all Zernike moments up to order n and repetition m by concatenating the coefficients $\sqrt{\pi/(n+1)}A_{nm}$ for all considered n and m into a vector $\mathbf{s}$. We can then define the similarity score $d_{f,g}$ (based on Zernike moments of up to order n and repetition m) for the preliminary match as

$$d_{f,g} = \mathbf{s}(f)^\top \mathbf{s}(g)^* = \sum_{nm} \sqrt{\frac{\pi}{n+1}}A_{nm}(f)\sqrt{\frac{\pi}{n+1}}A^*_{nm}(g). \tag{13}$$

To obtain the exact correlation score requires evaluating an infinite sum. In practice, only a few coefficients suffice to approximate image patches reasonably well. The quality of the reconstruction depends on the number of terms used and the frequency content of the image patch. To determine the number of coefficients required we conducted a simple test based on the self-similarity of the descriptors for multiple (over 18 000) patches from typical imagery. To test the performance of the descriptors for other values of correlation score we generated a synthetic sequence of image patches where each image is a small random perturbation of the previous one. This yields patches that are highly correlated with nearby patches in the sequence but uncorrelated with those that are distant (the frequency content of the synthetic patches was adjusted so that the autocorrelation scores approximated those observed in typical underwater imagery). Results are summarized as curves of similarity score versus true correlation for different order of descriptors in Fig. 6, and show that the reconstruction quality improves as the order (n) is increased from 8 to 24. Overall, we chose to use all moments up to order $n = 16$ as a compromise between quality of approximation and compactness. In addition, the use of moments results in significant computational savings when calculating similarity between multiple features. For example, the 51-pixel diameter patch used in our implementation requires multiplying 2041 $(\pi D^2/4)$ pixel values in the disc to calculate the correlation directly while the similarity measure that approximates the cross correlation requires multiplying 153 ($n \leq 16$ and all valid repetitions m) weighted moments.

To evaluate the performance of our method, the affine invariant region extraction and moment-based descriptor was compared to a fixed-window correlation-based match on a sequence of underwater imagery. We conducted our test for a diverse range of baseline magnitudes by matching each of 67 images to the next six images in a test sequence (for a total of 351 two-view matches). The details of the robust two-view matching technique we used are described in Section III. We used it here as a means to compare similarity-based measures over many correspondences by determining which proposed matches are inliers, i.e., consistent with the epipolar geometry.

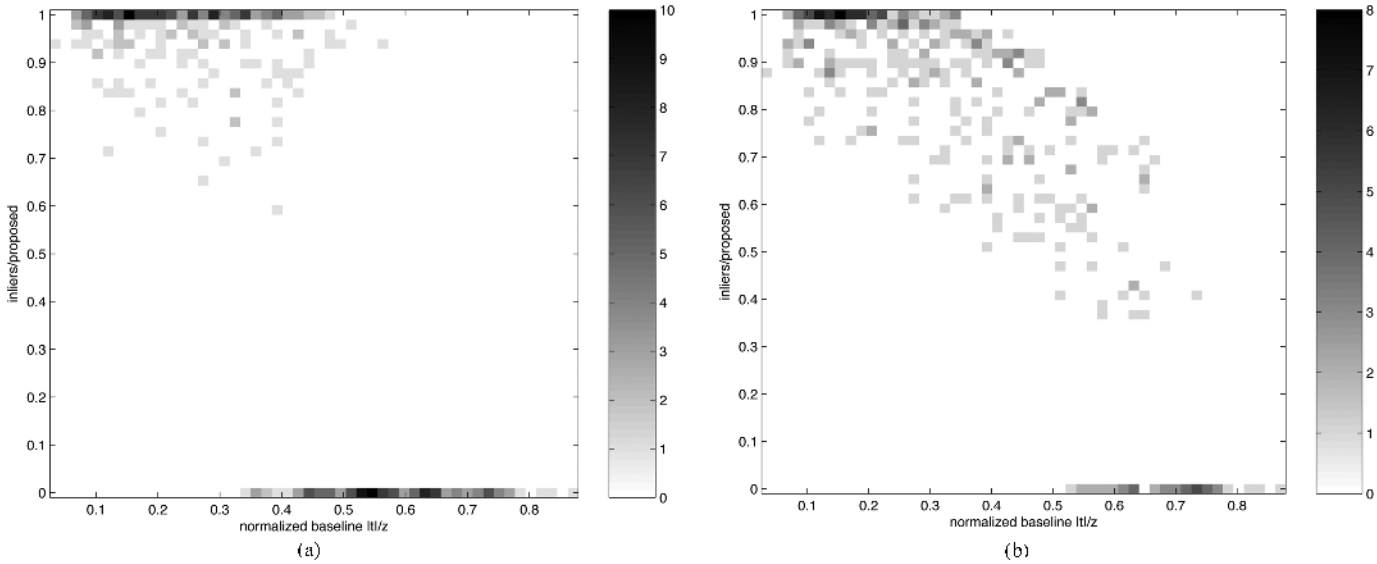


Fig. 8. (a) The distribution of the ratio of inliers to proposed matches against baseline magnitude for the 351 test pairs under fixed-window matching. For narrow baseline, most proposals are inliers, but for normalized baseline larger than 0.4, this abruptly changes to a low ratio. (b) For the affine-invariant region, the degradation is gradual and inliers are detected for wider baselines. Color bar values correspond to number of occurrences.

Navigation sensors provide an image-independent estimate of baseline magnitude $|\mathbf{t}|$ and altitude z, which allows us to formulate a normalized baseline magnitude $|\mathbf{t}|/z$. This is the relevant quantity for induced parallax. For pairs that could be matched reliably and for which the camera pose could be calculated accurately, the change in viewing angle to a feature can be calculated from the camera poses and from the rays in correspondence (Fig. 7).

The fixed-window feature method failed to match 122 of the 351 pairs, typically for large baselines. This can be seen in Fig. 8 for normalized baseline magnitudes above 0.45. The affine-invariant regions failed on only 44 pairs, with a gradual degradation in matching performance for increasing baseline.

III. Submap Generation

The core of our algorithm for SFM is based on robust estimation of the essential matrix (Fig. 9) [20] from a set of candidate correspondences between two views. These potential correspondences are proposed between features that have descriptor vectors with high similarity scores. To prevent calculating the similarity between all features in both images, the navigation-based estimates of interimage motion and vehicle altitude are used to limit the search space (Fig. 10) by propagating pose and altitude uncertainties through the two-view point-transfer equation [21].

A modified version of RANdom SAmple Consensus (RANSAC) [30] determines the putative correspondences that are consistent with an essential matrix (Fig. 11). In cases of multiple valid solutions, we select the one closest (in the Mahalanobis distance sense) to the navigation-based prior. The inliers and the essential matrix estimate are used to produce a maximum *a posteriori* (MAP) estimate of relative pose with the navigation-based estimates as a prior [31]. The solution includes the triangulated 3-D features (Fig. 12).

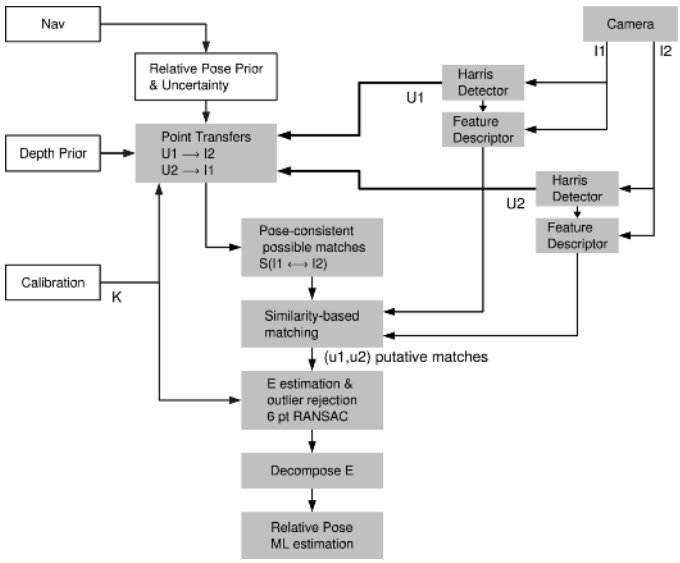


Fig. 9. Overview of our approach to relative pose estimation from instrumented and calibrated platforms. Unshaded blocks represent additional information compared to the uninstrumented/uncalibrated case. Given two images, we detect features using the Harris interest point detector. For each feature, we then determine the search region in the other image by using sensor-based pose and altitude information. Putative matches are proposed based on similarity. We then use RANSAC and the proposed six-point algorithm to robustly estimate the essential matrix, which is then decomposed into motion parameters. The pose is then refined by minimizing the reprojection error over all matches considered inliers.

A. Essential Matrix Estimation

Relative pose from calibrated cameras is a five-degrees-of-freedom (5 DOF) problem (3 DOF for rotation and 2 DOF for direction of motion between cameras) because of loss of scale. Minimal five-point algorithms [32]–[34] tend to be ill-posed, have complex implementations, and can present up to 20 solutions that then have to be tested. We use a six-point method presented by the authors [20], which is simpler than five-point

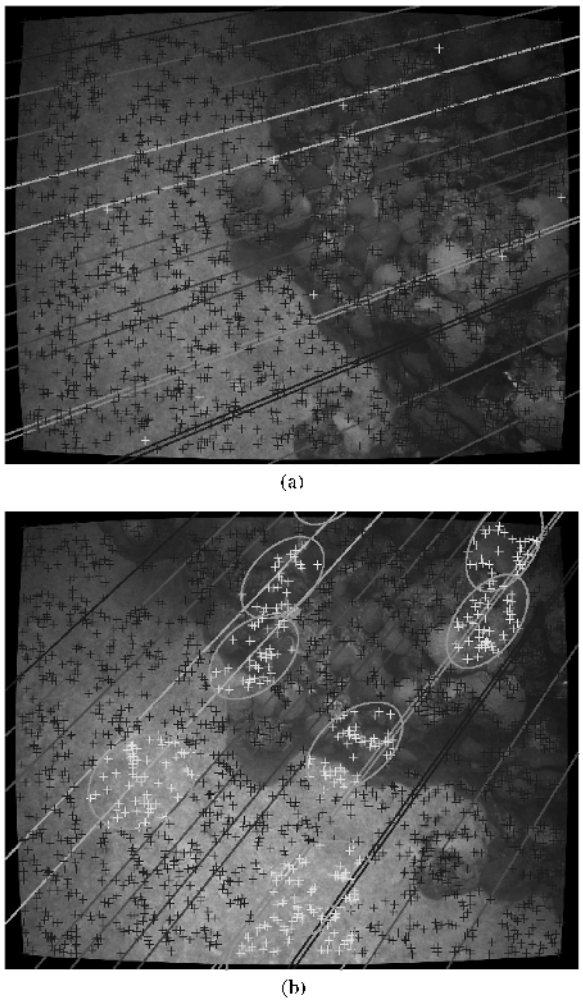


Fig. 10. Prior pose restricted correspondence search on a pair of underwater coral reef images. (a) Interest points are shown as crosses. A sampling of interest points (light crosses) is transferred to the right image. (b) The 99% confidence regions for the transferred points based on the pose prior and depth standard deviation of 0.75 m. The candidate interests points that fall within these regions are highlighted. Scene dimensions $\approx$ 2.5 m $\times$ 2 m.

implementations and overcomes the failure of the classic linear six-point algorithm in the presence of planar scenes [35]. Our proposed algorithm will produce up to six possibly valid essential matrices. Using synthetic data sets (generated for both planar and nonplanar scenes) and random relative motion, we have determined that one of the essential matrices produced by this six-point algorithm always corresponds to the true camera motion for perfect (noise-free) measurements. We have also observed that for perfect measurements of points in a planar configuration, the proposed six-point algorithm always produces two essential matrices, one which corresponds to the true essential matrix, and one which corresponds to the (incorrect) output of the linear six-point algorithm.

B. Robust Essential Matrix Estimation

The following two statements must hold for the proposed six-point algorithm to be useful in the context of estimating the essential matrix from a large set of putative correspondences. First, the quality of the solution must degrade gracefully in the presence of measurement noise. Second, we must be able to select the correct solution from up to six essential matrices. We address these issues in the next subsections.

1) Effects of Noise: To test how the performance of this algorithm degrades in the presence of noise, we performed 1000 trials with randomly generated scenes and motions. For each trial, the essential matrices computed by the six-point algorithm were decomposed into their respective rotation and translation representation. Even though the proposed six-point algorithm degrades in the presence of noise, we show in [20] that a large number of estimates will be close to the true motion. This suggests that the algorithm can be used effectively in a RANSAC context where it is reasonable to expect that there will be point combinations yielding an essential matrix close to the true one and will explain a large fraction of the correctly matched points.

2) Outlier Rejection (RANSAC): To eliminate outliers (correspondences inconsistent with the motion model) an essential matrix between the two images is estimated using RANSAC [30]. The basic steps for outlier rejection based on RANSAC are augmented to include checking for physically realizable point configurations. The added robustness comes at the expense of additional computation, though this is incurred only when a proposed essential matrix seems superior to the current "best" estimate. To be physically realizable, a configuration of points and relative pose must do the following:

- place all points in front of both cameras (cheirality constraint) [10, ch. 20];
- the scene points lie only a few meters in front of the camera because the attenuation lengths underwater for the visible spectrum are in the range of 5–25 m [1];
- the 3-D points must not lie between both cameras since the ocean floor is a "solid surface" and both cameras must be on the same side of it.

Enforcing these constraints resolves many cases of ambiguities but does not resolve all ambiguous pairs. It is important to bear in mind that during the RANSAC stage we are mainly interested in determining matches that are consistent with an essential matrix. If the inliers support multiple motion interpretations, the ambiguity is resolved when determining the final motion estimate, as described in Section III-B6.

3) Reprojection Error: Given a set of n_{in} measured correspondences $S_{in} = \{\mathbf{u}_i \leftrightarrow \mathbf{u}'_i\}$, under the assumption of isotropic Gaussian noise corrupting the interest point locations, it can be shown [10, ch. 10] that the maximum-likelihood estimate (MLE) for the fundamental matrix $F = K^{-\top}EK$ minimizes the sum of squared reprojection errors

$$D\left(F, \hat{\mathbf{u}}_i, \hat{\mathbf{u}}'_i\right) = \sum_i d\left(\mathbf{u}_i, \hat{\mathbf{u}}'_i\right)^2 + d\left(\mathbf{u}'_i, \hat{\mathbf{u}}_i\right)^2 \qquad (14)$$

where $d(\cdot,\cdot)$ represents the Euclidean distance and $\hat{\mathbf{u}}_i$ and $\hat{\mathbf{u}}'_i$ are the estimated ideal correspondences (i.e., before corruption with Gaussian noise) that exactly satisfy $\hat{\underline{\mathbf{u}}}'_i \mathrm{F} \hat{\underline{\mathbf{u}}}_i$.

The reprojection errors are used to rank the quality of the essential matrices proposed in the RANSAC loop. The number of inliers for a proposed essential matrix is determined by the number of correspondences with reprojection errors below a threshold t. This threshold is set based on the expected noise in

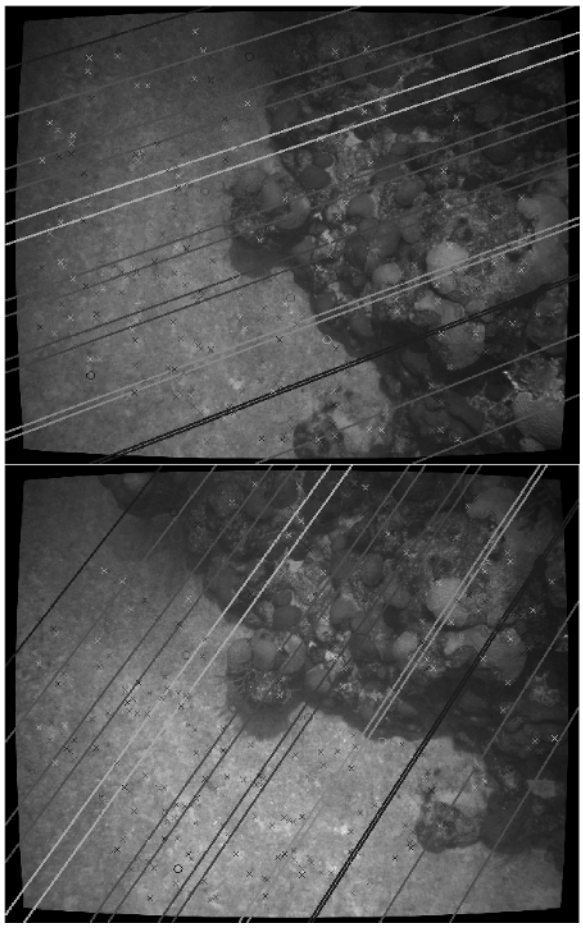


Fig. 11. Epipolar geometry and correspondences. The given image pair illustrates the MAP refined image-based epipolar geometry. RANSAC determined 398 consistent inliers designated "x," from the putative set of 405 matches. The rejected outliers are designated "o." Scene dimensions $\approx$ 2.5 m $\times$ 2 m.

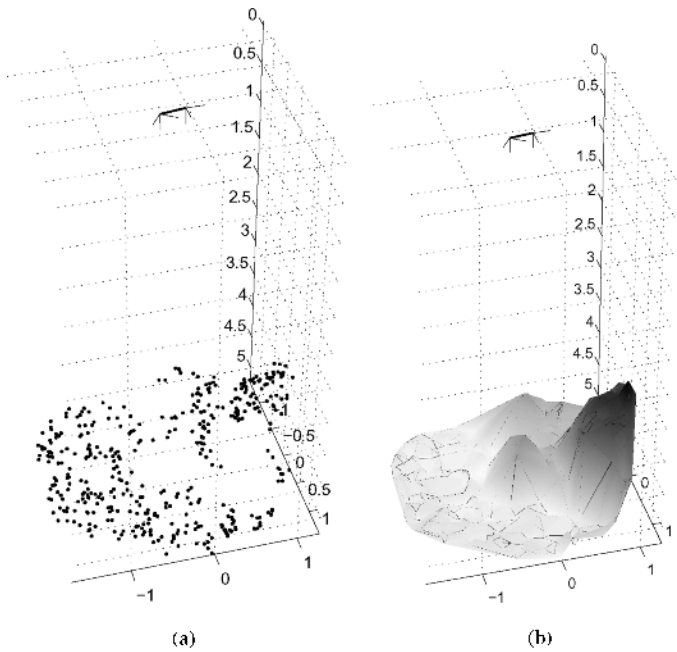


Fig. 12. Triangulated inliers for the pair in Fig. 11. Coordinates in meters, in the reference frame of the first camera. (a) 3-D feature locations. (b) Interpolated surface, shaded by depth from the first camera. The camera frames are as a three-axis frame connected by the baseline (wider line).

feature locations and with some testing on actual images. Calculating the reprojection error requires triangulating the ideal feature points [36]. We use Kanatani's fast iterative method [37]. Fig. 11 shows the resulting image-based points considered inliers by RANSAC. The epipolar geometry in the figure is a refinement by MAP estimation from the RANSAC inliers (Section III-B6). Fig. 12 illustrates the triangulated correspondences and the cameras in the frame of the first camera.

4) From the Essential Matrix to Motion Estimates: The essential matrix that best explains the data according to RANSAC is decomposed into a rotation and translation using singular value decomposition (SVD) [10, ch. 10]. This approach has a fourfold ambiguity on relative pose. To determine which is the correct solution we check that triangulated points are in front of both cameras.

5) Two-View Critical Configurations: Planar or nearly planar scenes are frequently encountered in surveys of the ocean floor. For the uncalibrated case, there is a continuum of fundamental matrices consistent with the data. In the case of a calibrated camera, two views of an unknown plane will have at most two valid essential matrices [38]. The ambiguity can be resolved by requiring all points to be in front of both cameras except in the case where all points are closer to one camera than the other. This situation can happen when the vehicle motion has a significant component toward or away from the bottom.

Planar scenes are a particular case where scene points and the camera centers fall on a ruled quadric [39], [40]. In the general case of ruled quadrics, there will be up to a threefold ambiguity in motion and structure for the uncalibrated case. For the calibrated case, the number of interpretations is two. Each interpretation will place the scene points and camera centers on distinct ruled quadrics. A dense set of points (hundreds) from a natural scene is unlikely to fall on a ruled quadric, but in cases of low overlap (tens of points), this could happen. In Section III-B6, we use the motion prior from navigation instruments to disambiguate image-based solutions.

6) Final Motion Estimate: The previous section recognizes that the output of the RANSAC stage is a set of inliers associated with one of possibly several interpretations of motion. The six-point algorithm can be used with more than six points and in fact we use it with all inliers to generate possible essential matrices. In cases of multiple interpretations, we must rely on additional information. We choose the relative pose encoded in the essential matrix that is closest to the relative pose prior from navigation sensors. More specifically, the image-based relative pose with the smallest Mahalanobis distance $||\mathbf{p}_i - \mathbf{p}_{\text{nav}}||_{\Sigma_{\text{nav}}}$, with Σ_{nav} the covariance of the prior, is selected as the initial estimate

$$||\mathbf{p}_i - \mathbf{p}_{\text{nav}}||_{\Sigma_{\text{nav}}} = \sqrt{(\mathbf{p}_i - \mathbf{p}_{\text{nav}})^\top \Sigma_{\text{nav}}^{-1} (\mathbf{p}_i - \mathbf{p}_{\text{nav}})} \quad (15)$$

where $\mathbf{p}_i = [\mathbf{t}_i^\top, \boldsymbol{\Theta}(\mathbf{R}_\mathbf{i})^\top]^\top$ are the translation and orientation parameters (as Euler angles) for the ith essential matrix, and $\mathbf{p}_{\text{nav}}$ is the similarly defined relative pose from the navigation sensors. Since relative pose is recovered only up to scale from

images, the baseline magnitude is normalized to unit length and the covariance is constrained to be zero in the direction of motion. The baseline of the image-based solution is then scaled according to the projection of the prior baseline

$$\mathbf{t} = \frac{\mathbf{t}_i^\top \mathbf{t}_{\text{nav}}}{||\mathbf{t}_i||}\mathbf{t}_i. \tag{16}$$

7) Bundle Adjustment: The final relative pose estimate is based on a bundle adjustment of pose and 2-D feature locations. From Bayes rule, we have

$$p(\mathbf{x}|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \tag{17}$$

which in terms of parameter estimation states that the posterior distribution $p(\mathbf{x}|\mathbf{z})$ of a vector of parameters (associated with a model) $\mathbf{x}$, given observations $\mathbf{z}$, is proportional to the likelihood of the data, given the parameters $p(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{x})$. The MAP estimate $\mathbf{x}^*$ maximizes the total posterior probability of the model given the observations and prior knowledge. We choose to refer to the MLE when using only image-based measurements (uniform prior) and MAP estimation when including navigation sensor measurements, though in practice, the navigation information is included as additional observations.

We assume conditionally independent measurements. The MLE is then

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \prod_i p(\mathbf{z}_i|\mathbf{x}). \tag{18}$$

For image-based measurements $\mathbf{z} = \mathbf{u}$ given the relative pose and structure $\mathbf{x} = [\mathbf{p}^\top, \mathbf{X}^\top]^\top$, the measurements can be considered to have Gaussian distributions centered around the true reprojections

$$p(\mathbf{u}|\mathbf{x}) \propto e^{[\mathbf{u}-\phi(\mathbf{x})]^\top[\mathbf{u}-\phi(\mathbf{x})]}. \tag{19}$$

Taking the negative log-likelihood, we express the MLE problem as a minimization of the cost function

$$[\mathbf{u} - \phi(\mathbf{x})]^\top [\mathbf{u} - \phi(\mathbf{x})]. \tag{20}$$

Since the measurements are assumed to be independent, the measurement covariance is diagonal and the cost function can be expressed as

$$\sum_i ||{}^c\mathbf{u}_i - \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_i)||^2 \tag{21}$$

where ${}^c\mathbf{u}_i$ is the measurement on camera c for feature i, and $\mathbf{p}_c$ is the relative pose estimate from imagery and $\mathbf{X}_i$ is the estimate of the position of the ith 3-D feature point. For MAP estimation, the pose sensors provide a relative pose prior. The initial estimate close to the navigation-based pose together with the extra cost term that penalizes large deviations from the navigation-prior provide a robust two-view relative pose estimate. The cost function being minimized then takes the form

$$\sum_i ||{}^c\mathbf{u}_i - \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_i)||^2 + ||\mathbf{p}_c - \mathbf{p}_{\text{nav}}||_{\Sigma_{\text{nav}}} \tag{22}$$

with the additional term accounting for the relative pose prior, which has the form of a Mahalanobis distance similar to (15) with $\mathbf{p}_c$ being the relative pose vector estimate from imagery.

8) Robust Estimation: The minimization of squared residuals is optimal in the maximum-likelihood sense for zero mean Gaussian noise. A Gaussian noise model has a distribution with small tails, reflecting that large errors are unlikely. But in practice large errors occur more often than the Gaussian distribution suggests (i.e., from poor feature localization or from incorrect correspondences that satisfy two-view but not multiview constraints). When this is ignored (and noise is assumed Gaussian), the minimization of squared residuals is strongly affected by outliers.

Least squares minimizes

$$E_{LS}(\mathbf{x}) = \sum_i (r_i(\mathbf{x}))^2 \tag{23}$$

where $r_i(\mathbf{x}) = (z_i - h_i(\mathbf{x}))/\sigma_i$ is the weighted residual for the ith measurement.

M-estimators [41] reduce the sensitivity to outliers by replacing the r_i^2 with a $\rho(r_i)$ that grows more slowly for large r_i while remaining symmetric, positive definite, and having a minimum at zero

$$E_M(\mathbf{x}) = \sum_i \rho\left(r_i(\mathbf{x})\right). \tag{24}$$

Several choices of $\rho(r)$ have been proposed. The Cauchy M-estimator [42] weighs the residuals in a manner that assumes a Cauchy distribution rather than a Gaussian, which allows for a larger proportion of large errors

$$\rho_C(r) = \frac{c^2}{2}\log\left(1 + (r/c)^2\right). \tag{25}$$

We use this estimator in all bundle adjustment calculations throughout this paper. The soft outlier threshold $c = 2.3849$ achieves 95% asymptotic efficiency on the standard normal distribution [41].

C. Growing Submaps in the Three-View Case

In cases where scene elements are viewed in three (or more) views the algorithm attempts to obtain the pose of the third view by a modified version of robust resection [30] (Fig. 13), otherwise the two-view essential matrix estimation is used. The resection stage produces the approximate pose of the camera that is most consistent with the proposed correspondences between image points and 3-D structure. The approach in [22] considers the bundle adjustment problem of the latest three views while reducing the free parameters to the latest camera pose and the feature points it views. It takes advantage of points seen in the three views as well as those in the last two views. Though efficient, this technique does not handle uncertainty and prior information in a consistent fashion. We have prior information of the relative pose between the first and second cameras as well as between the second and third cameras. We choose to fix the origin on the frame of the first camera and leave the second and third cameras to be adjusted. In essence, we solve the MAP estimate of the trifocal tensor as a way to produce an estimate of the latest pose and the uncertainty in pose and structure.

Given three views 0, 1, and 2 and the measured (noisy) correspondences between the views $\{{}^0\mathbf{u}_i \leftrightarrow {}^1\mathbf{u}_i \leftrightarrow {}^2\mathbf{u}_i\}$, and the correspondences between pairs of views $\{{}^1\mathbf{u}_i \leftrightarrow {}^2\mathbf{u}_i\}$, $\{{}^0\mathbf{u}_i \leftrightarrow {}^2\mathbf{u}_i\}$, $\{{}^0\mathbf{u}_i \leftrightarrow {}^1\mathbf{u}_i\}$, under the assumption of isotropic

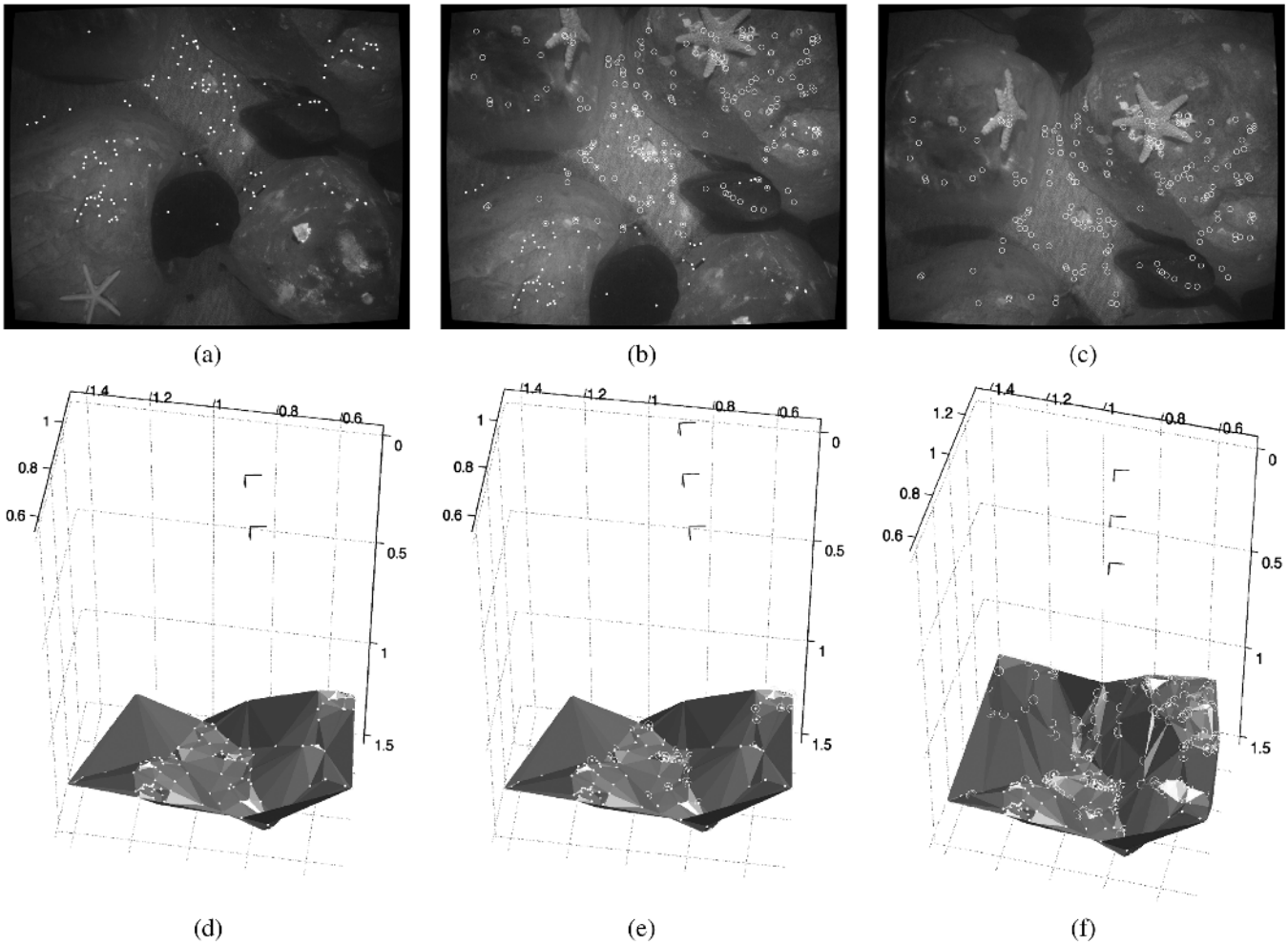


Fig. 13. Illustration of growth of a submap based on resection. Images (a) and (b) have corresponding features marked by dots. The structure and motion implied by those correspondences is illustrated in (d) with units in meters. Images (b) and (c) have correspondences marked by circles. The features viewed by the three images are marked by both a dot and a concentric circle. (e) These features are used in resection to initialize the pose of the third camera. (f) Then, the additional correspondences between (b) and (c) are triangulated and the poses refined. Scene dimensions $\approx$ 1 m $\times$ 0.8 m.

Gaussian noise corrupting the interest point locations, the MLE for the poses and structure minimizes the sum of squared reprojection errors

$$\sum_{c=0}^{2}\sum_{i} d\left({}^{c}\mathbf{u}_i, {}^{c}\hat{\mathbf{u}}_i\right)^2 + \sum_{c=1,2}\sum_{j} d\left({}^{c}\mathbf{u}_j, {}^{c}\hat{\mathbf{u}}_j\right)^2 + \sum_{c=0,2}\sum_{k} d\left({}^{c}l\mathbf{u}_k, {}^{c}\hat{\mathbf{u}}_k\right)^2 + \sum_{c=0,1}\sum_{l} d\left({}^{c}l\mathbf{u}_l, {}^{c}\hat{\mathbf{u}}_l\right)^2 \quad (26)$$

where $d(\cdot,\cdot)$ represents the Euclidean distance, ${}^{c}\hat{\mathbf{u}}_m$ are the estimated ideal correspondences (i.e., before corruption with Gaussian noise) for camera c, and m is the index into the correspondence set. The role of the structure is implicit in (26). More explicitly, we have that the projection of a 3-D point $\mathbf{X}_i$ onto a camera c with pose $\mathbf{p}_c$ is ${}^{c}\hat{\mathbf{u}}_i$

$${}^{c}\hat{\mathbf{u}}_i = \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_i). \quad (27)$$

Using the camera projection (27), we expand (26)

$$\begin{aligned}\min_{\mathbf{p}_1,\mathbf{p}_2,\{\mathbf{X}_i\},\{\mathbf{X}_j\},\{\mathbf{X}_k\},\{\mathbf{X}_l\}} & \sum_{c=0}^{2}\sum_{i=1}^{N_{012}} \left\|{}^{c}\mathbf{u}_i - \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_i)\right\|^2 \\ & + \sum_{c=1,2}\sum_{j=1}^{N_{12}} \left\|{}^{c}\mathbf{u}_j - \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_j)\right\|^2 \\ & + \sum_{c=0,2}\sum_{k=1}^{N_{02}} \left\|{}^{c}\mathbf{u}_k - \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_k)\right\|^2 \\ & + \sum_{c=0,1}\sum_{l=1}^{N_{01}} \left\|{}^{c}\mathbf{u}_l - \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_l)\right\|^2. \end{aligned} \quad (28)$$

The MAP estimate adds cost terms based on relative pose prior (from pose sensors) similar to the ones used in the relative pose MAP estimation, which biases the solution to the scale implied by the navigation sensors

$$\|\mathbf{e}_{01}\|_{\Sigma_{\text{nav}}} + \|\mathbf{e}_{12}\|_{\Sigma_{\text{nav}}} \quad (29)$$

where using the composition notation from [43] the discrepancy between vision- and navigation-based relative pose is given by

$$\mathbf{e}_{ij} = \ominus\hat{\mathbf{x}}_{ij} \oplus \mathbf{x}_{ij}^{\text{nav}} = \ominus\hat{\mathbf{x}}_j \oplus \hat{\mathbf{x}}_i \oplus \mathbf{x}_{ij}^{\text{nav}} \tag{30}$$

where $\oplus$ is the head-to-tail frame composition and $\ominus$ is the inverse frame composition. The weighted error is $\|\mathbf{e}_{ij}\|_{\Sigma_{\text{nav}}} = e_{ij}^{\top}\Sigma_{ij}^{-1}e_{ij}$, where Σ_{ij} corresponds to the estimated covariance of e_{ij} propagated from the covariance of $\mathbf{x}_{ij}^{\text{nav}}$.

D. Submap Size

We have proposed using reconstructions of subsequences as the basic unit from which to derive the network of images and feature points for a final global adjustment. An important issue in this approach is setting the size of submaps. The size (or number) of submaps affects the complexity of multiple bundle adjustments, the reliability of matching submaps, and the complexity of the resulting network of submaps. We discuss these points and suggest that it suffices to close submaps based on the number of features they contain, with improved performance arising from smaller submaps. Thus, we choose to create submaps with at least three images and up to 2000 3-D features.

1) Bundle Adjustment Complexity: Each step in a sparse bundle adjustment of N features and M views has complexity $\mathcal{O}((N+M)M^2)$ associated with the inversion of the sparse normal equations [44]. If we break down the problem into S submaps with no overlap, then each submap bundle adjusts with complexity $\mathcal{O}((1/S)(N+M)(M/S)^2)$ assuming that the features and the views are evenly distributed in each submap. The complexity for the total sequence (the bundle adjustment of S submaps) is $\mathcal{O}((N+M)M^2/S^2)$. Therefore, S smaller bundle adjustments reduce the overall complexity in proportion to S^2. In the presence of overlap between submaps, the complexity grows linearly with the overlap fraction v and number of submaps $S_v = S/(1-v)$. The complexity of processing one submap does not change but the overall complexity is $\mathcal{O}((1/(1-v))((N+M)M^2/S^2))$. If a sequence is to be split into submaps and each submap bundle adjusted, then there are significant computational savings to be had by using smaller maps.

2) Uncertainty in Structure and Motion: An incremental reconstruction can drift relative to the "true" structure because the imaging process only relates features that are spatially close to each other. We choose to use the estimate of covariance in 3-D feature positions as an indication of possible drift, given that ground truth is not typically available. Our local bundle adjustment procedure fixes part of the gauge (scale) implicitly through the relative pose prior provided by navigation sensors. The reference frame origin and orientation are coincident with the first camera [45]. But for registration purposes the uncertainty (and weight) of reconstructed 3-D points should reflect the quality of the triangulation rather than an arbitrary choice of reference frame. Therefore, we choose to express the uncertainty of 3-D points with six gauge freedoms (orientation and translation). This is achieved by simply eliminating the rows in the Jacobian corresponding to the equations that fix the origin to the first camera before calculating the covariance of the poses and structure by using the pseudoinverse (zeroing the six smallest singular values) [46].

3) Submap Matching and Network Complexity: To propose putative matches based on similarity between submaps i and j takes time $\mathcal{O}(N_iN_j)$ where N_i and N_j are the number of features in each submap. Since $N_i = \mathcal{O}(N_j)$, we realize that registering submaps by similarity is $\mathcal{O}(N_i^2) = \mathcal{O}((N/S)^2)$. But matching all submaps to all submaps is $\mathcal{O}(S^2)$ with the lower cost of matching smaller maps offset by the need to match more maps. However, for a sparse network where most nodes have edges to a few adjacent nodes, as in a survey with a moving vehicle, we can expect that $\mathcal{O}(S)$ edges exist and that a reasonable matching technique will also perform $\mathcal{O}(S)$ matches. The overall complexity of matching for the sparse network case is $\mathcal{O}(N^2/S)$ with more (smaller) submaps saving effort at the submap matching stage.

E. Submap Closing

Once a submap contains enough 3-D features, it is closed and a new submap is started. The structure associated with the most recent half of the cameras in the map being closed is used to start the new submap. This provides a good balance between number of maps and improved matching.

We perform a final bundle adjustment using all poses and prior pose information on the submap before closing it. A sparse bundle adjustment routine [10, Appendix 4], [42] is used to minimize the cost function

$$\sum_c \sum_i \|^c\mathbf{u}_i - \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_i)\|^2 + \|\mathbf{e}_{c,c+1}\|_{\Sigma_{\text{nav}}} \tag{31}$$

where $\mathbf{p}_c$ is the pose estimate from imagery for the cth camera, $\mathbf{e}_{c,c+1}$ is the residual vector between the relative pose estimate from navigation sensors and imagery (30), and $\mathbf{X}_i$ the estimate of the position of the ith 3-D feature point. This is the same procedure used on the triplets (after resection) but it considers all views. The initial guess is provided by the incremental submap. Fig. 14 contains two views of the 3-D structure of a submap at this stage.

The relative pose between the new submap and the previous submap corresponds to the pose (in the reference frame of the submap being closed) of the camera that becomes the origin of the new submap.

IV. Global Representation

The temporal sequence of images is processed into a set of 3-D submaps with estimates of coordinate transformations between temporally adjacent submaps. This can be viewed as a graph where each node is the origin of a submap and the edges in the graph are the coordinate transformations between submaps (Fig. 15). Our algorithm attempts to establish additional spatial relationships between submaps (corresponding to overlap from parallel tracklines or loop closures).

A. Submap Matching

While additional edges in the graph could be determined at the image level using the two-view algorithm, we propose that spatial overlap is more easily resolved at the submap level (Fig. 17). Submaps must be matched to establish new

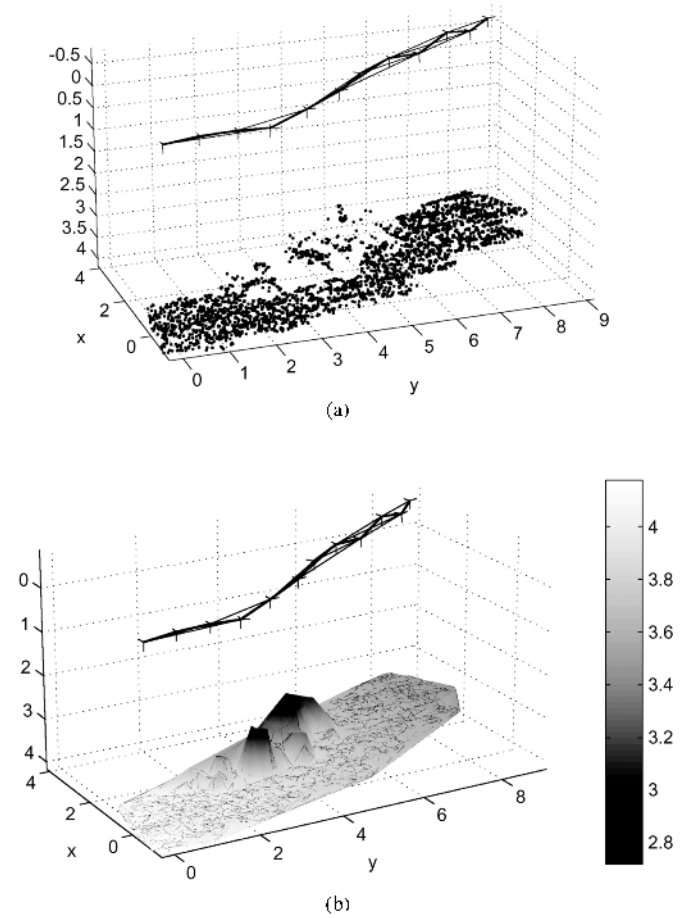


Fig. 14. Two views of the MAP bundle adjustment of an example of an incremental reconstruction consisting of 12 images and close to 3000 points. Cameras are represented as three-axes frames. The temporal sequence is from left to right. Temporally adjacent frames are connected by a wider line. Spatially adjacent frames (determined through resection) are linked with lines. (a) The dots represent the estimated position of 3-D points in the reference frame defined by the first camera. (b) For ease of interpretation, a surface has been fit through the points using a Delaunay triangulation. The surface is shaded according to the Z coordinate. Axes and color bar units are in meters.

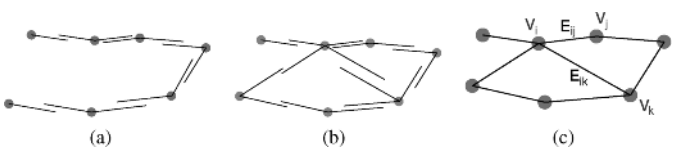


Fig. 15. Placing nodes (gray circles) in a globally consistent frame. From relative transformations (black links) in a temporal sequence (a), to proposing and verifying new additional links (b) to a network with nodes consistent with the relative transformations (c).

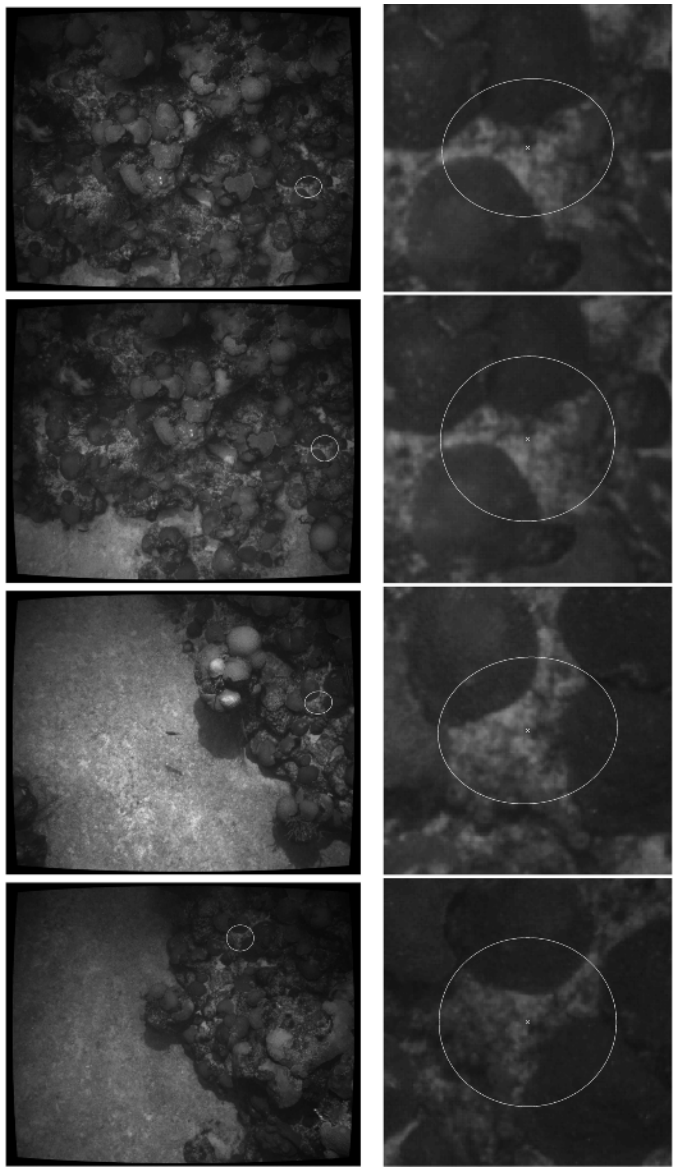


Fig. 16. Multiple views of a 3-D feature: (left column) the image and the feature neighborhood extracted as described in Section II-C and (right column) a detail of around the feature point. The top two rows correspond to images that belong to a submap on the first trackline of the survey. The bottom two rows are from a submap from the second trackline. Left-hand-side column scene dimensions $\approx$ 2.5 m $\times$ 2 m. Right-hand-side column scene dimensions $\approx$ 0.32 m $\times$ 0.25 m.

edges in the graph. Registering two sets of 3-D points with unknown correspondences is traditionally performed with iterative closest point (ICP) techniques [47]. In its strictest sense, ICP is only a refinement of the transformation between two sets of 3-D points that are already relatively well aligned and in which all points in one set have a match in the other. Given the fairly sparse set of 3-D points that make up a submap and the potentially poor initial alignment, ICP is not adequate for our application, and therefore, it is not used. However, the very fact that 3-D points are created from visual information implies that their appearance in multiple views (Fig. 16) is characteristic enough to effectively establish correspondences and be triangulated. Therefore, we extend the feature description and similarity-based matching between images to matching submaps by relying on the appearance of 3-D points to propose corresponding features between submaps. The underlying assumption is that a similarity measure that was effective to match 3-D points along track will also be effective when matching across submaps. Corresponding 3-D points are proposed based on appearance and a robust registration using RANSAC with Horn's algorithm [48] is used to determine which points are in correspondence and the transformation parameters (Fig. 17).

1) 3-D Feature Descriptors: For similarity-based matching purposes, we propose to describe a 3-D feature by the average of

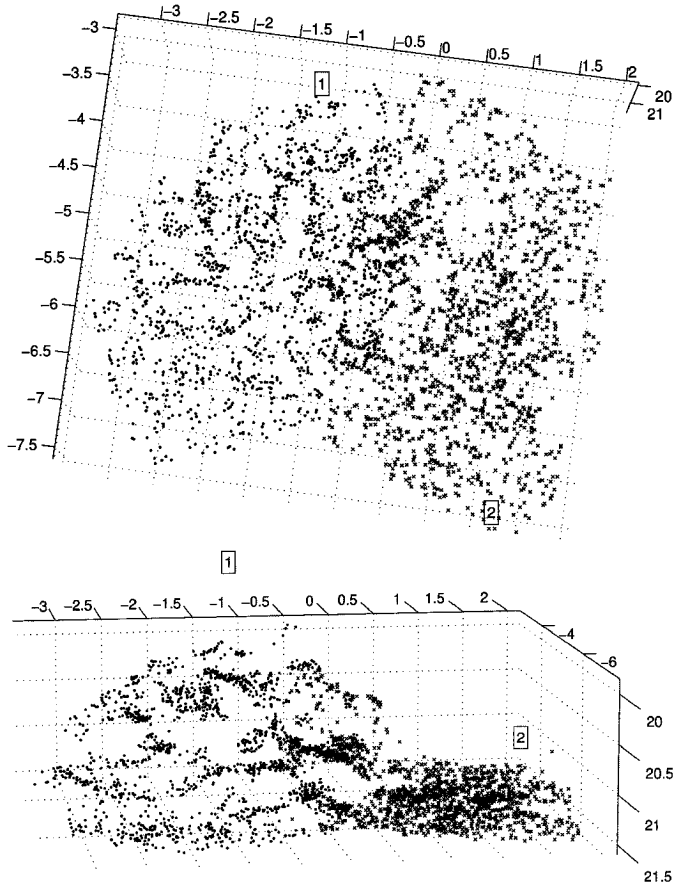


Fig. 17. Views of the registered low overlap submaps that contain the images in Fig. 16. The dots correspond to 3-D features of the submap on the first trackline of Fig. 16. The "x" symbols correspond to a submap on the second trackline of Fig. 16.

all acquired 2-D views of the neighborhood around the feature. We assume that for each view the neighborhood is represented in a canonical frame as described in Section II (i.e., an affine invariant region mapped onto a circle with orientation known to a few degrees from navigation).

Due to superposition and linearity the moment of an average image patch corresponds to the average of the moments. Thus, for a 3-D feature $\mathbf{X}$ viewed by N cameras, with an extracted 2-D region f_k from the kth camera, and associated feature descriptor $\mathbf{s}(f_k)$ (Section II-C), we construct a descriptor for the 3-D feature as the average of all 2-D descriptors

$$\mathbf{s}(\mathbf{X}) = \frac{1}{N}\sum_{k=1}^{N}\mathbf{s}(f_k). \tag{32}$$

2) Similarity Measure: Putative 3-D feature correspondences between different submaps are proposed based upon similarity of descriptors. The measure of Section II-C (which approximates the cross correlation between patches in the invariant frame) is used to propose matches.

3) Robust 3-D to 3-D Matching: Given putative correspondences between 3-D points from two submaps, we seek to register the two sets of 3-D points. The goal is to find the similarity transformation (translation, rotation, and scale) that aligns the 3-D points ${}^s\mathbf{X}_i$ from source submap s onto ${}^t\mathbf{X}_i$, the corresponding points on the target submap t.

To support robust outlier rejection, we utilize RANSAC based on a minimal set of three points (with Horn's algorithm [48]). This determines the inliers in the putative correspondence set and an initial approximation to the transformation. A second pass with a limited search range based upon the estimate from the first pass typically produces more proposals and correct matches. The RANSAC loop is modified to include prior knowledge regarding the transformation scale between submaps. As the scale of the submaps is derived from the same instruments, registered submaps should have a similarity transformation with a scale close to unity. This helps speed up the RANSAC loop by allowing us to only evaluate the support of transformations with scale c such that $0.9 \leq c \leq 1.1$.

For simplicity, we ignore the estimated covariance of 3-D points in the RANSAC loop. In this case, the solution from Horn's algorithm is equivalent to an unweighted least squares. Then, we refine this solution using the uncertainties of all corresponding structure points, which corresponds to minimizing the sum of Mahalanobis distances

$$\mathbf{d}_k = {}^t\mathbf{X}_k - {}^t_s\hat{\mathrm{T}}\,{}^s\mathbf{X}_k \tag{33}$$

$${}^t_s\mathrm{T}^* = \arg\min_{{}^t_s\mathrm{T}} \sum_k \mathbf{d}_k{}^\top \Sigma_k^{-1}\mathbf{d}_k \tag{34}$$

with the covariance of the error approximated by the first-order propagation of the covariance of the points being registered

$$\Sigma_k \approx \frac{\partial \mathbf{d}_k}{\partial {}^t\mathbf{X}_k}\Sigma_{{}^t\mathbf{X}_k}\frac{\partial \mathbf{d}_k}{\partial {}^t\mathbf{X}_k}^\top + \frac{\partial \mathbf{d}_k}{\partial {}^s\mathbf{X}_k}\Sigma_{{}^s\mathbf{X}_k}\frac{\partial \mathbf{d}_k}{\partial {}^s\mathbf{X}_k}^\top. \tag{35}$$

We assume that the estimates of structure points between submaps are uncorrelated, which is a valid assumption for submaps that do not share any cameras (e.g., across-track submaps). The covariance of the transformation parameters can be estimated to first order from the Jacobian of the cost function being minimized in (34) evaluated at the optimum.

B. Edge Proposal or Topology Refinement

Starting from a temporal sequence, we wish to establish additional links between overlapping submaps (which will lead to establishing additional links between overlapping imagery). This can be viewed as a refinement of a graph where each node is a submap reference frame and each edge (or link) is a relative transformation. Since submaps can be linked only to spatially neighboring submaps, the graph is expected to be sparse. This would require verifying only $\mathcal{O}(S)$ links if the node positions were well known. Yet as links are added, we expect the spatial relationships between nodes to change, possibly requiring additional link checks. Verifying edges (Section IV-A) is computationally expensive, so our approach must concentrate on likely links by considering uncertainty in the node positions and by updating node position estimates as links are added.

Possible approaches to estimating links include the following:

- estimating relative transformations from current global estimates: $\hat{\mathbf{x}}_{ik} = \ominus\hat{\mathbf{x}}_{wi} \oplus \hat{\mathbf{x}}_{wk}$;
- estimating relative transformations from composition of relative transformations: $\hat{\mathbf{x}}_{ik} = \mathbf{x}_{ij} \oplus \mathbf{x}_{jk}$.

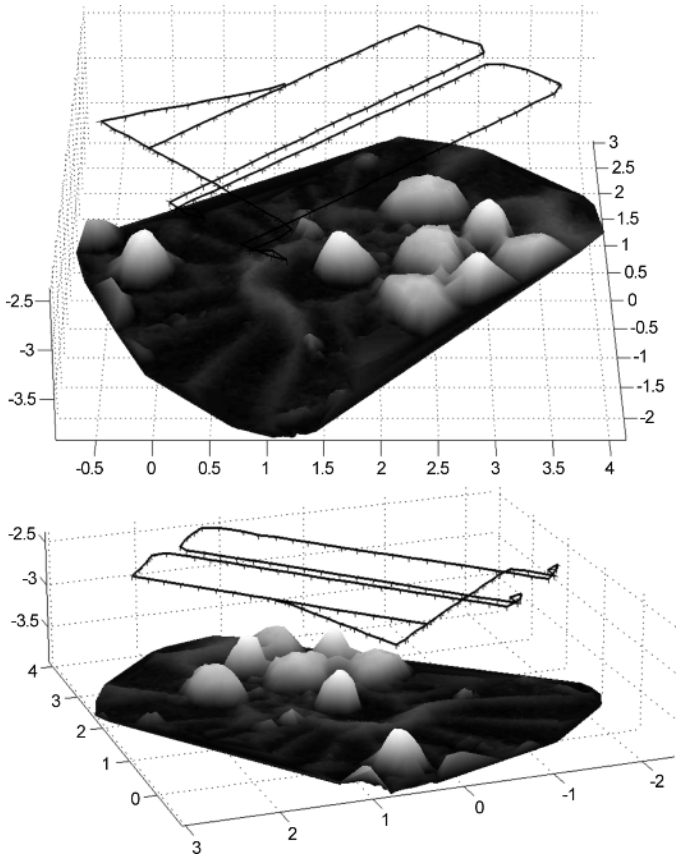


Fig. 18. Two views of the reconstruction of poses and structure for the Johns Hopkins University (JHU) tank. The camera poses are connected by a line to suggest the trajectory followed by the ROV. A Delaunay triangulation interpolates a surface between 3-D feature points. The structure is shaded according to height. Units are in meters.

If estimates of the node poses are maintained in a global reference frame, then additional links can be inferred by measuring distances and uncertainties between nodes. Though the proposal process is simple, maintaining nodes in a common frame requires enforcing consistency among the cycles that may form as additional edges are included in the graph. It should be noted that while consistency is important before attempting a bundle adjustment (Section IV-C) it is not essential when attempting to establish edges in a sparse graph.

The alternative approach is to remain in relative frame space and use composition of relative transformations to express the relative pose between nodes that do not have a direct link. Because there may be multiple paths between nodes, an approximate solution is to use a shortest path algorithm such as Dijkstra's algorithm. The Atlas framework advocates this approach for map matching [49]. In this case, the proposal process is more complex since it must place nodes relative to a base node by composition along the shortest path. As more edges become available more paths must implicitly be considered. We apply this approach to the 6-DOF problem of submap transformations.

After estimating relative transformations between a pair of submaps it is necessary to determine which submaps are likely to overlap. This depends on several factors such as camera field of view and altitude. A simple approach is to calculate the distance and uncertainty between the centroids of the structure of each submap according to the relative transformation and its uncertainty. A maximum allowable distance for overlap can be estimated based on the camera field of view and the altitude of the cameras. For overlap calculations, we model the submap as a disc with diameter equal to the width of the camera footprint. This is a simple and conservative measure since submaps tend to be longer than their width.

The proposal stage calculates a 99% confidence interval for the distance between submaps. If the maximum distance for overlap is within the interval (or greater), then overlap is considered possible. The most likely link is the one that has the highest proportion of the confidence interval within the maximum distance for overlap. By proposing the most likely link within the range, the graph tends to "zipper up" nodes, closing loops last. Alternatively, the least likely link within range of overlap could be verified first. Because there is a lower probability that the nodes actually do overlap this strategy can lead to a high proportion of unsuccessful matching attempts.

The proposal and verification steps are repeated until a user-defined maximum number of allowable links is achieved. A good choice is eight times the number of submaps which, on average, allows maps to connect to the previous and next maps in the temporal sequence and up to six other nearby maps.

C. Node Estimation: Global Poses From Relative Poses

Once relative poses are determined we must place nodes in a global frame such that they remain consistent with all the measured relative poses. This can be formulated directly as an optimization problem to yield batch or recursive nonlinear least squares solutions [50], [51]. These approaches suffer from requiring to maintain the cross covariances between submap frames. Sharp *et al.* [52] proposed a cycle consistency approach that operates in relative space but produces consistent global estimates without having to estimate or store cross covariances. The graph can be seen as a distributed network and consistent, conservative global estimates can be generated through fusion by covariance intersection [53].

1) Nonlinear Weighted Least Squares: We seek to determine the global poses that best explain all the relative pose measurements and consider the navigation-based prior. By defining a cost function associated with these discrepancies we can then optimize an initial guess.

We define e_{ij} as the disparity pose vector between the composition of the estimates of global transformations ${}^{w}_{i}\hat{\mathrm{T}}$ and ${}^{w}_{j}\hat{\mathrm{T}}$ and the measured relative transformation ${}^{j}_{i}\mathrm{T}$. Throughout this discussion, we use $\hat{\mathbf{x}}$ to represent an estimate of $\mathbf{x}$. In Smith, Self, and Cheeseman's (SSC) [43] notation, the relative pose vector implied by the estimates of pose is obtained from a tail-to-tail operation

$$\hat{\mathbf{x}}_{ij} = \ominus\hat{\mathbf{x}}_{wi} \oplus \hat{\mathbf{x}}_{wj} \tag{36}$$

where the transformation/pose parameters are related to the homogeneous transformation as $\mathbf{x}_{ik} = \rho({}^{i}_{k}\mathrm{T})$. The disparity between the relative pose measurement $\mathbf{x}_{ij}$ (the MAP estimate

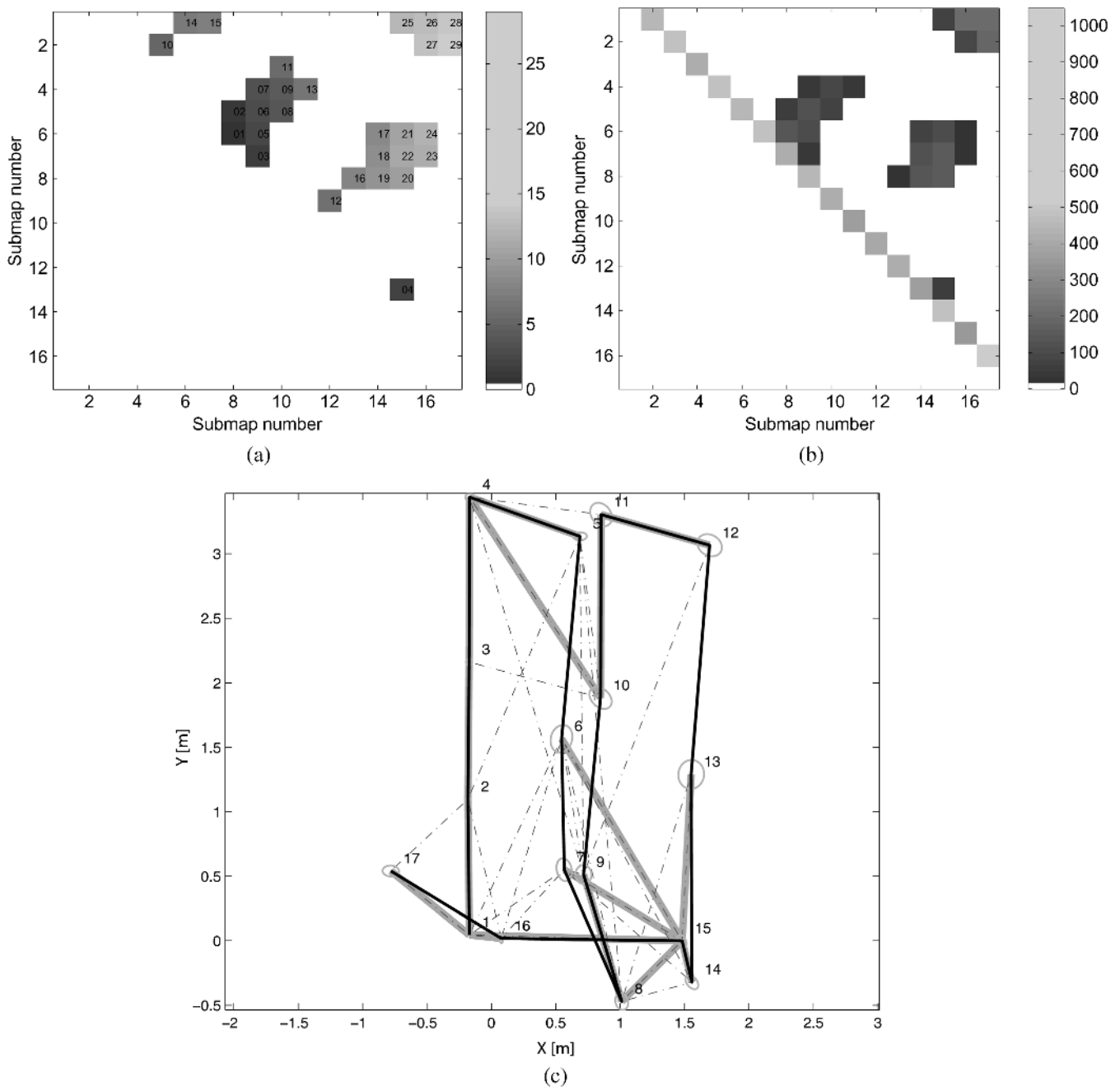


Fig. 19. (a) Color-coded order in which links across track were added to the graph. The "zipper" effect in parallel tracklines is apparent as links close in time are established before more distant ones. (b) Color-coded number of matching features between submaps. The loop closure can be seen in the relatively high number of common features between the first and last submaps. (c) The plane view of the submap origins according to the shortest path algorithm: the temporal sequence (fine black), the additional links (dotted–dashed), and the shortest uncertainty path from the origin node (wide gray).

from imagery and navigation) and the relative pose $\hat{\mathbf{x}}_{ij}$ from the tail-to-tail composition of estimates $\hat{\mathbf{x}}_j$ and $\hat{\mathbf{x}}_i$ is the error measure we seek to minimize

$$\mathbf{e}_{ij} = \ominus\hat{\mathbf{x}}_{ij} \oplus \mathbf{x}_{ij} = \ominus\hat{\mathbf{x}}_j \oplus \hat{\mathbf{x}}_i \oplus \mathbf{x}_{ij}. \tag{37}$$

$\mathbf{e}_{ij}$ can be thought of as the residual transformation in a short cycle formed by the tail-to-tail estimate of the transformation $\ominus\hat{\mathbf{x}}_j \oplus \hat{\mathbf{x}}_i$ and the measured transformation by map matching $\mathbf{x}_{ij}$. Ideally the residual transformation should be the identity (corresponding to no rotation and no translation). We use the rotation vector representation (where the direction of the vector specifies the axis of rotation and the magnitude of the vector corresponds to the angle of rotation) for the orientation parameters of the residual transformation [54]

$$\mathbf{e}_{ij} = \rho\left({}^{j}_{i}\mathrm{T}^{-1}\,{}^{j}_{w}\hat{\mathrm{T}}\,{}^{w}_{i}\hat{\mathrm{T}}\right). \tag{38}$$

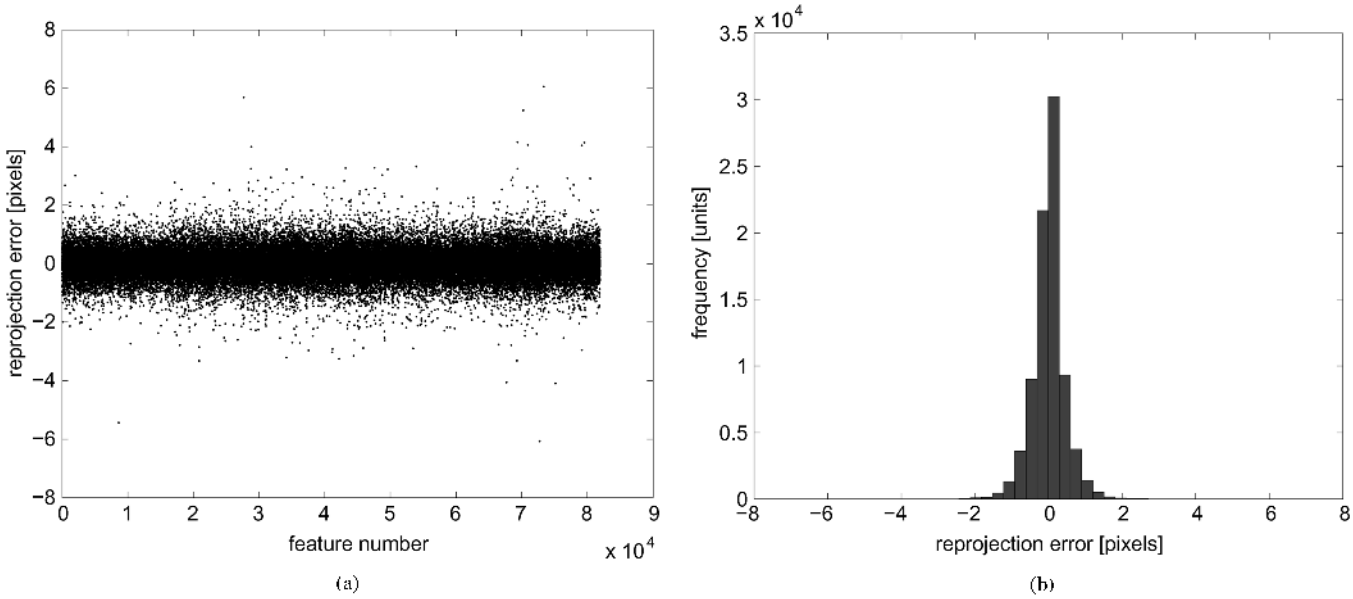


Fig. 20. (a) The reprojection errors (both x and y coordinates) for all reconstructed features. Some outliers are present though their effect is reduced by using an M-estimator in the bundle adjustment. (b) A histogram of the same errors. For visualization purposes, 95% of the features with lowest associated reprojection errors are displayed in the reconstructions of Fig. 18.

We also define the disparity between the global pose according to navigation and our estimate of global pose

$$\mathbf{e}_i = \rho\left({}^w_i\mathrm{T}^{-1}\,{}^w_i\hat{\mathrm{T}}\right) \tag{39}$$

or directly in SSC notation

$$\mathbf{e}_i = \ominus\hat{\mathbf{x}}_{wi} \oplus \mathbf{x}_{wi}. \tag{40}$$

In a similar fashion to [50], we seek a set of global transformations $\mathcal{T}^*$ of all N submaps $\mathcal{T} = \{\mathbf{x}_{w1} \cdots \mathbf{x}_{wN}\}$ that minimizes this error over all links. We formulate this as a weighted nonlinear least squares optimization

$$\mathcal{T}^* = \arg\min_{\mathcal{T}} \sum_{ij} e_{ij}^\top \Sigma_{ij}^{-1} e_{ij} + \sum_i e_i^\top \Sigma_i^{-1} e_i \tag{41}$$

where Σ_{ij} corresponds to the estimated covariance of e_{ij} propagated from the covariance of $\mathbf{x}_{ji}$, and Σ_i corresponds to the estimated covariance of e_i propagated from the covariance of $\mathbf{x}_{wi}$.

An alternative to minimizing the discrepancy between the composition of global poses is to directly minimize the 3-D distances between corresponding points of submaps, though it is computationally more intensive because the number of equations is proportional to the number of corresponding points instead of to the number of measured edges. However, this reduces the sensitivity to poorly triangulated networks [55] where the error in the frame transformations might appear small at the expense of large errors in the structure. The error measure becomes

$$\mathbf{d}_{ijk} = {}^w_i\hat{\mathrm{T}}\,{}^i\mathbf{X}_k - {}^w_j\hat{\mathrm{T}}\,{}^j\mathbf{X}_k \tag{42}$$

$$\mathcal{T}^* = \arg\min_{\mathcal{T}} \sum_{ij}\sum_k \mathbf{d}_{ijk}{}^\top \Sigma_{ijk}^{-1} \mathbf{d}_{ijk} + \sum_i e_i^\top \Sigma_i^{-1} e_i. \tag{43}$$

In cases where the frame-based refinement is unsatisfactory (i.e., the reprojection errors for the implied camera poses are large or have strong biases), we switch to this cost function.

2) Camera Poses From Submaps: Once submaps are placed in a global frame it is then possible to place the cameras that form the submaps in the same global frame. These camera poses are used as the initial estimate for the bundle adjustment of the complete data set. By construction the pose of each camera in a submap is in the frame of the first camera. The transformation from the node to the global frame can be composed with the transformation of the camera pose to the node origin. Since temporally adjacent submaps share cameras there is more than one way of mapping the cameras that are common between submaps. We use the geometric mean [56] of the pose estimates according to each submap (in the global frame) to obtain an initial estimate of the camera poses.

D. Bundle Adjustment

Once camera poses are in the global frame the same sparse bundle adjustment routine used to close the submaps is used on the entire data set. We obtain the MAP estimate by including cost terms associated with the navigation measurements, as described in Section III-E.

V. Results and Validation

A. JHU Tank Structure Ground Truth

To illustrate the submap matching process we present in Fig. 18 the resulting structure from a pose-instrumented remotely operated vehicle (ROV) survey (using the Seabed camera system) performed at the Johns Hopkins University (JHU) Hydrodynamics Test Facility (Baltimore, MD). We draped a carpet over the bottom of the tank and placed real and artificial rocks of varying sizes on the bottom to simulate an

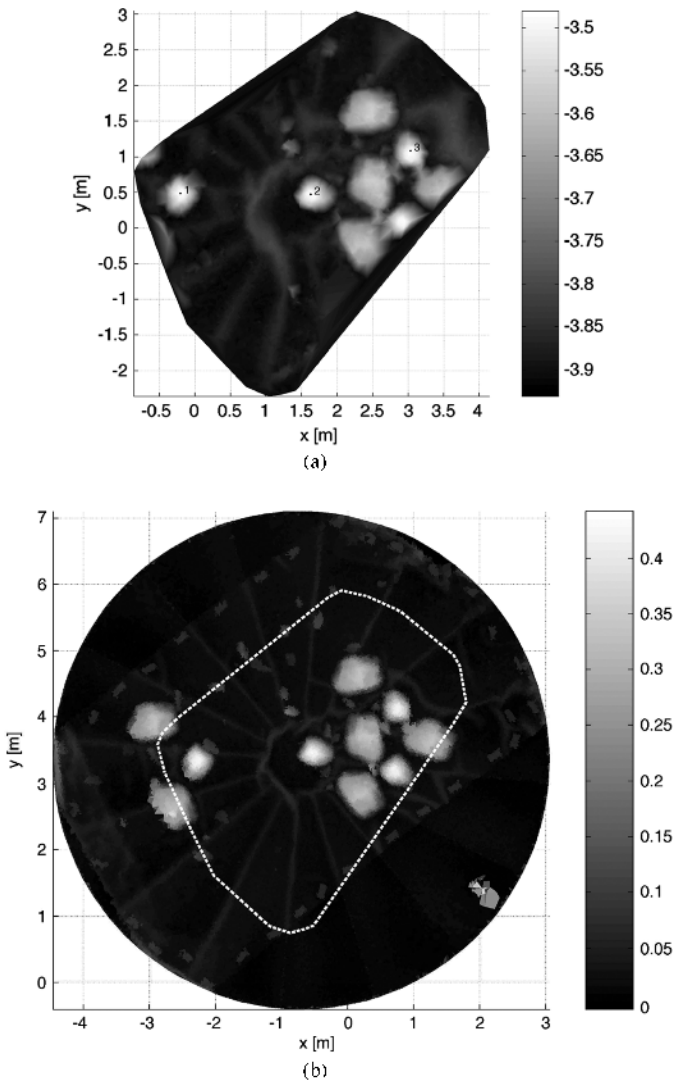


Fig. 21. (a) Height map from the SFM reconstruction. Surface based on a Delaunay triangulation. The labeled points were manually selected for the initial alignment with the laser scan. (b) Height map from the laser scan. The outline of the registered SFM reconstruction is shown as a segmented line. Color bar units are in meters.

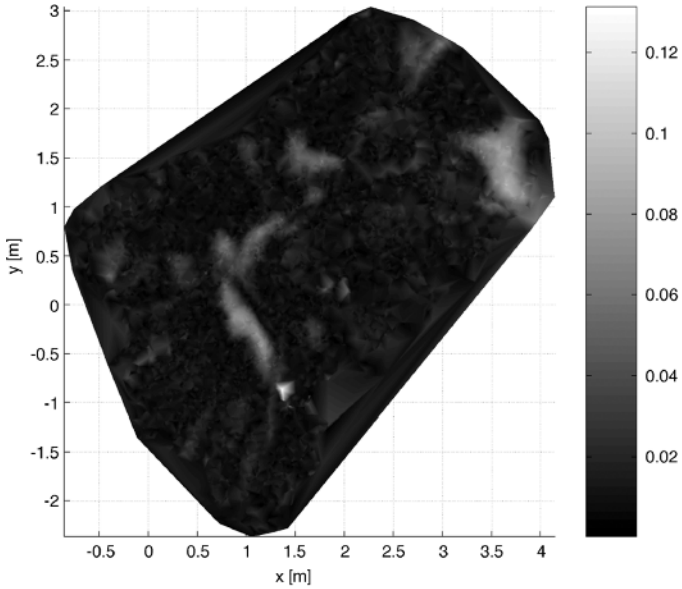


Fig. 22. Distance map from SFM 3-D points to the laser scan after ICP registration. Areas of large discrepancies tend to correspond to the carpet being buoyant for the visual survey. An outlier in the reconstruction produced the large error visible at approximate $x = 1.4$ m and $y = 0.8$ m. Color bar units are in meters.

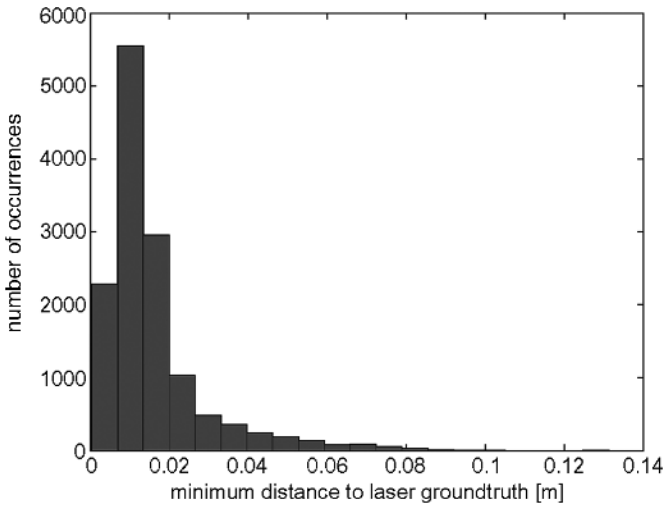


Fig. 23. Distribution of minimum distances to the laser scan from each recovered 3-D point. Because of the moving carpet only the points below the median error where used to calculate the registration transformation. The similarity-based registration results in a root mean square (RMS) distance of 3.6 cm. The scale is recovered to within 2%.

underwater scene with considerable 3-D structure. The evolution of the submap graph for that reconstruction is conveyed in Fig. 19 while the reprojection errors for the structure are presented in Fig. 20.

For validation purposes, the tank used in Fig. 18 was drained and scanned with an HDS2500 laser scanner (serial number P24, Leica Geosystems, St. Gallen, Switzerland). The registered model of the tank has more than 3.8 million points with an estimated accuracy of 1.2 mm. The surface area was approximately 41 m^2 resulting, on average, in nine range measurements for each cm^2.

We initially aligned SFM reconstruction with the laser data by selecting easily recognizable landmarks (Fig. 21) and then refined through ICP. Not all points could be used for registration since parts of the carpet moved (after the tank was drained the carpet settled under its own weight). We attempted two registration strategies to overcome the nonrigid transformation between surfaces: using points belonging only to rocks to register (segmenting by height under the assumption that the rocks in the scene did not move), and performing ICP based on the points with registration errors below the median error (under the assumption that at least half the points remained fixed). Results were very similar for both strategies and we present the median-based approach since it highlights regions where the carpet moved.

Figs. 22 and 23 indicate that the registration errors are of the order of centimeter level with a 2% change in scale. These results suggest that the approach is capable of delivering reasonable estimates of scene structure.

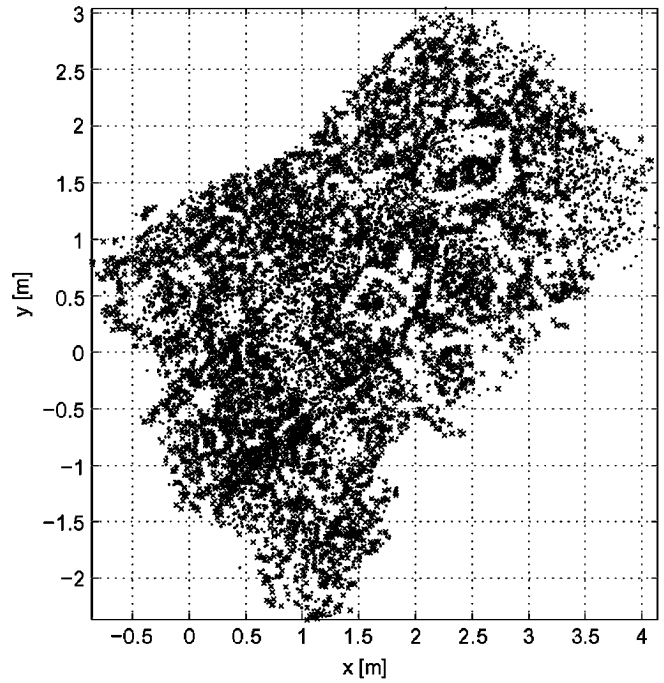


Fig. 24. Points below the median error (x) and above (dots). Registration parameters where calculated using points below the median error. By referring to Fig. 21, outliers tend to group around the smooth, raised folds of the carpet, which clearly do not correspond to the drained carpet surface.

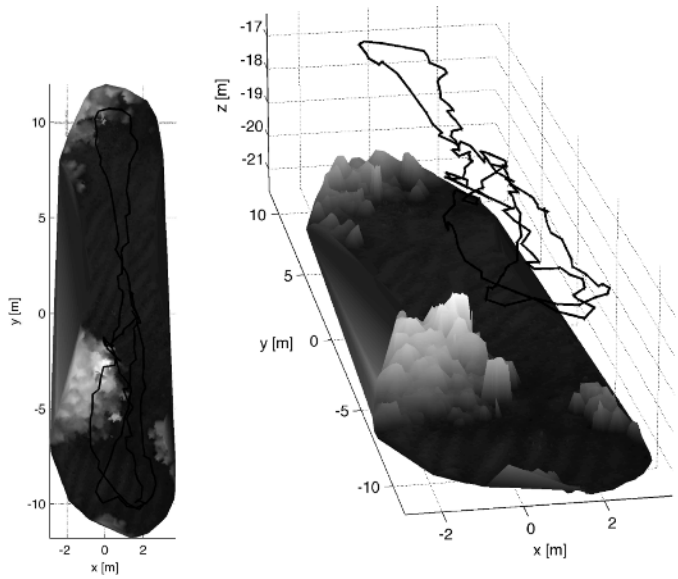


Fig. 25. Two views of the reconstruction as a surface through the recovered 3-D points. The camera trajectory is also presented as a black line (seen above the seafloor reconstruction on the right-hand side view). Strong swell significantly perturbed the vehicle trajectory yet the consistency of the reconstruction is apparent in features such as the sand ripples on the bottom.

By using points below the median error to calculate the similarity transformation to register the SFM and laser data we effectively segment the data into two halves, one of which was allowed to deform while the other was not. It is interesting to note from Fig. 24 that most of the outliers correspond to the broad carpet waves.

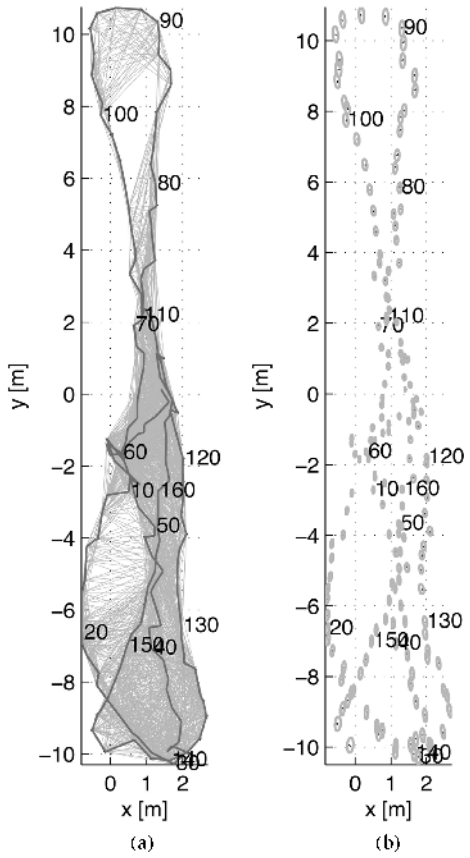


Fig. 26. (a) Plan view of the camera trajectory (dark gray) and common features between cameras (light gray links). (b) The 99% confidence ellipses for the xy position of the cameras. Every tenth camera is numbered on both figures to suggest the temporal sequence.

B. Bermuda Survey

In August 2002, the SeaBED AUV performed several transects on the Bermuda shelf as well as some shallow-water engineering trials. This section presents results from a shallow-water (20 m approximately) area survey programmed with several parallel tracklines for a total path length of approximately 200 m and intending to cover 200 m^2. Due to very strong swell and compass bias the actual path deviated significantly from the assumed path. This data set illustrates the capabilities to infer links in the graph of submaps to yield a consistent reconstruction.

A section of 169 images demonstrates matching and reconstruction along the temporal sequence and across track with multiple passes over the same area. Fig. 25 presents Delaunay triangulated surfaces through the reconstructed points and the camera trajectory. Plan views of the camera trajectory, the links (common 3-D features) between views, and the uncertainty in the xy position of the cameras are shown in Fig. 26.

Fig. 27 shows feature points and the convex hull of the submaps. Spatial overlap between temporally adjacent submaps is consistent while across-track overlap is a function of the trajectory followed by the vehicle.

VI. Concluding Remarks and Future Work

We have presented a brief overview of an underwater structure from motion algorithm that takes advantage of vehicle nav-

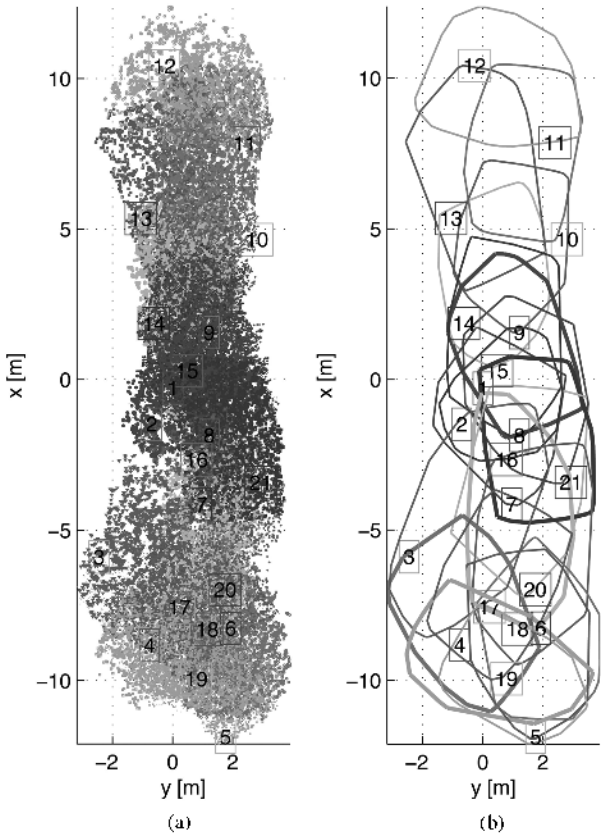


Fig. 27. (a) Plan view of the features for each submap. (b) Convex hull of the 3-D features of each submap. The varying degrees of spatial overlap between submaps are apparent in these figures.

igation estimates to constrain the image-based solution. This work will be extended to provide dense 3-D reconstructions of the ocean floor, which in turn can lead to improved imagery by range-based compensation of absorption [57]. Additional work will also exploit the resulting self-consistent pose-and-structure solution to detect and compensate for some navigation sensor biases [31].

ACKNOWLEDGMENT

The authors would like to thank the Captain and crew of the *R/V Weatherbird II* for their invaluable assistance in collecting the field data used in this paper. They would also like to thank J. Kinsey, Dr. L. Whitcomb, and C. Roman for pulling off the experiments at JHU.

REFERENCES

[1] S. Q. Duntley, "Light in the sea," *J. Opt. Soc. Amer.*, vol. 53, no. 2, pp. 214–233, 1963.
[2] D. R. Yoerger, A. M. Bradley, M.-H. Cormier, W. B. F. Ryan, and B. B. Walden, "Fine-scale seafloor survey in rugged deep-ocean terrain with an autonomous robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, San Francisco, CA, 2000, vol. 2, pp. 1767–1774.
[3] R. D. Ballard, L. E. Stager, D. Master, D. R. Yoerger, D. A. Mindell, L. L. Whitcomb, H. Singh, and D. Piechota, "Iron age shipwrecks in deep water off Ashkelon, Israel," *Amer. J. Archaeology*, vol. 106, no. 2, pp. 151–168, Apr. 2002.
[4] J. Howland, "Digital data logging and processing, Derbyshire survey, 1997," Woods Hole Oceanogr. Inst., Woods Hole, MA, Tech. Rep., Dec. 1999.
[5] National Transportation Safety Board, Washington, DC, EgyptAir Flight 990, Boeing 767-366ER, SU-GAP, 60 Miles South of Nantucket, MA, October 31, 1999, Aircraft Accident Brief NTSB/AAB-02/01, 2002.
[6] C. R. Smith, "Whale falls: Chemosynthesis at the deep-sea floor," *Oceanus*, vol. 35, no. 3, pp. 74–78, 1992.
[7] H. Singh, R. Eustice, C. Roman, O. Pizarro, R. Armstrong, F. Gilbes, and J. Torres, "Imaging coral I: Imaging coral habitats with the SeaBED AUV," *Subsurface Sens. Technol. Appl.*, vol. 5, no. 1, pp. 25–42, Jan. 2004.
[8] H. S. Sawhney and R. Kumar, "True multi-image alignment and its application to mosaicing and lens distortion correction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 3, pp. 235–243, Mar. 1999.
[9] H. S. Sawhney, S. C. Hsu, and R. Kumar, "Robust video mosaicing through topology inference and local to global alignment," in *Proc. Eur. Conf. Comput. Vis.*, Freiburg, Germany, 1998, pp. 103–119.
[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, MA: Cambridge Univ. Press, 2000.
[11] A. W. Fitzgibbon and A. Zisserman, "Automatic camera recovery for closed or open image sequences," in *Proc. 5th Eur. Conf. Comput. Vis.*, Freiburg, Germany, Jun. 1998, pp. 311–326.
[12] S. D. Fleischer, H. H. Wang, S. M. Rock, and M. J. Lee, "Video mosaicking along arbitrary vehicle paths," in *Proc. Symp. Autonom. Underwater Veh. Technol.*, Monterey, CA, Jun. 1996, pp. 293–299.
[13] S. Negahdaripour, X. Xu, and L. Jin, "Direct estimation of motion from sea floor images for automatic station-keeping of submersible platforms," *IEEE J. Ocean. Eng.*, vol. 24, no. 3, pp. 370–382, Jul. 1999.
[14] N. Gracias and J. Santos-Victor, "Underwater mosaicing and trajectory reconstruction using global alignment," in *Proc. OCEANS Conf.*, Honolulu, HI, 2001, pp. 2557–2563.
[15] S. Negahdaripour and X. Xun, "Mosaic-based positioning and improved motion-estimation methods for automatic navigation of submersible vehicles," *IEEE J. Ocean. Eng.*, vol. 27, no. 1, pp. 79–99, Jan. 2002.
[16] O. Pizarro and H. Singh, "Toward large-area underwater mosaicing for scientific applications," *IEEE J. Ocean. Eng.*, vol. 28, no. 4, pp. 651–672, Oct. 2003.
[17] R. Szeliski, "Image mosaicing for tele-reality applications," Cambridge Res. Lab., Cambridge, MA, Tech. Rep. CRL 94/2, May 1994.
[18] *"Manual of Photogrammetry,"* 4th ed. C. C. Slama, Ed., American Society of Photogrammetry, Bethesda, MD, 1980, ch. 2.
[19] S. Negahdaripour and H. Madjidi, "Stereovision imaging on submersible platforms for 3-D mapping of benthic habitats and sea-floor structures," *IEEE J. Ocean. Eng.*, vol. 28, no. 4, pp. 625–650, Oct. 2003.
[20] O. Pizarro, R. Eustice, and H. Singh, "Relative pose estimation for instrumented, calibrated imaging platforms," in *Proc. Conf. Digital Image Comput. Tech. Appl.*, Sydney, Australia, 2003, pp. 601–612.
[21] R. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation in an unstructured environment using a delayed state history," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2004, vol. 1, pp. 25–32.
[22] Z. Zhang and Y. Shan, "Incremental motion estimation through local bundle adjustment," Microsoft Research, Tech. Rep. MSR-TR-01-54, May 2001.
[23] P. A. Beardsley, A. Zisserman, and D. Murray, "Sequential updating of projective and affine structure from motion," *Int. J. Comput. Vis.*, vol. 23, no. 3, pp. 235–259, Jun. 1997.
[24] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vis. Conf.*, Manchester, U.K., 1988, pp. 147–151.
[25] T. Tuytelaars and L. Van Gool, "Wide baseline stereo matching based on local, affinely invariant regions," in *Proc. British Mach. Vis. Conf.*, Bristol, U.K., 2000, pp. 736–739.
[26] T. Tuytelaars and L. Van Gool, "Matching widely separated views baed on affine invariant regions," *Int. J. Comput. Vis.*, vol. 59, no. 1, pp. 61–85, 2004.
[27] W. Y. Kim and Y. S. Kim, "A region-based shape descriptor using Zernike moments," *Signal Process.: Image Commun.*, vol. 16, no. 1–2, pp. 95–102, Sep. 2000.
[28] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 489–497, May 1990.
[29] C. Schmid and R. Mohr, "Local greyvalue invariants for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 530–535, May 1997.
[30] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

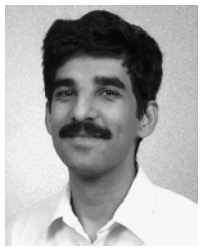
[31] O. Pizarro, "Large scale structure from motion for autonomous underwater vehicle surveys," Ph.D. dissertation, Massachusetts Inst. Technol./Woods Hole Oceanogr. Inst., Cambridge/Woods Hole, MA, Sep. 2004.
[32] B. K. P. Horn, "Relative orientation," *Int. J. Comput. Vis.*, vol. 4, no. 1, pp. 58–78, 1990.
[33] B. Triggs, "Routines for relative pose of two calibrated cameras from 5 points," INRIA, Montbonnot, France, Tech. Rep., 2000.
[34] O. Faugeras and S. Maybank, "Motion from point matches: Multiplicity of solutions," *Int. J. Comput. Vis.*, vol. 4, no. 3, pp. 225–246, 1990.
[35] J. Philip, "Critical point configurations of the 5-, 6-, 7-, and 8-point algorithms for relative orientation," KTH Royal Inst. Technol., Stockholm, Sweden, Tech. Rep. TRITA-MAT-1998-MA-13, 1998.
[36] R. I. Hartley and P. Sturm, "Triangulation," *Comput. Vis. Image Understanding*, vol. 68, no. 2, pp. 146–157, Nov. 1997.
[37] K. Kanatani, *Statistical Optimization for Geometric Computation: Theory and Practice*. Amsterdam, The Netherlands: Elsevier, 1996, ch. 5.
[38] H. C. Longuet-Higgins, "The reconstruction of a plane surface from two perspective projections," *Proc. R. Soc. Lond. B, Biol. Sci.*, vol. 227, no. 1249, pp. 399–410, May 1986.
[39] H. C. Longuet-Higgins, "Multiple interpretations of a pair of images of a surface," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 418, no. 1854, pp. 1–15, Jul. 1988.
[40] F. Kahl and R. I. Hartley, "Critical curves and surfaces for euclidean reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2002, vol. 2, pp. 447–462.
[41] Z. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," *Image Vis. Comput. J.*, vol. 15, no. 1, pp. 59–76, 1997.
[42] B. Triggs, P. F. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Vision Algorithms: Theory & Practice*, B. Triggs, A. Zisserman, and R. Szeliski, Eds. New York: Springer-Verlag, 2000, pp. 298–372.
[43] R. Smith, M. Self, and P. Cheeseman, *Estimating Uncertain Spatial Relationships in Robotics*, ser. Autonomous Robot Vehicles. New York: Springer-Verlag, 1990, pp. 167–193.
[44] P. F. McLauchlan and A. H. Jaenicke, "Image mosaicing using sequential bundle adjustment," in *Proc. British Mach. Vis. Conf.*, Bristol, U.K., 2000, pp. 616–625.
[45] P. F. McLauchlan, "Gauge independence in optimization algorithms for 3D vision," in *Proc. Workshop Vis. Algorithms*, 1999, pp. 183–199.
[46] D. Morris, "Gauge freedoms and uncertainty modeling for 3D computer vision," Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-15, Mar. 2001.
[47] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-14, no. 2, pp. 239–256, Feb. 1992.
[48] B. K. P. Horn, "Closed form solutions of absolute orientation using orthonormal matrices," *J. Opt. Soc. Amer. A*, vol. 5, no. 7, pp. 1127–1135, 1987.
[49] M. Bosse, P. Newman, J. Leonard, and S. Teller, "An Atlas framework for scalable mapping," MIT Marine Robotics Lab., Cambridge, MA, Tech. Memorandum 2002-04, 2002.
[50] F. Lu and E. Milios, "Globally consistent range scan alignment for environment mapping," *Autonom. Robots*, vol. 4, no. 4, pp. 333–349, 1997.
[51] M. C. Deans, "Bearing-only localization and mapping," Robotics Inst., Carnagie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-05-41, Sep. 2005.
[52] G. Sharp, S. Lee, and D. Wehe, "Multiview registration of 3D scenes by minimizing error between coordinate frames," in *Proc. 7th Eur. Conf. Comput. Vis.*, Copenhagen, Denmark, May 2002, pp. 587–597.
[53] C. Schlegel and T. Kämpke, "Filter design for simultaneous localization and map building (SLAM)," in *Proc. IEEE Int. Conf. Robot. Autom.*, Washington, DC, May 2002, pp. 2737–2742.
[54] X. Pennec and J.-P. Thirion, "A framework for uncertainty and validation of 3d registration methods based on points and frames," *Int. J. Comput. Vis.*, vol. 25, no. 3, pp. 203–229, 1997.
[55] M. Antone and S. Teller, "Scalable extrinsic calibration of omni-directional image networks," *Int. J. Comput. Vis.*, vol. 49, no. 2–3, pp. 143–174, 2002.
[56] X. Pennec, "Computing the mean of geometric features—Application to the mean rotation," INRIA, Sophia Antipolis, France, Res. Rep. RR-3371, Mar. 1998.
[57] H. Singh, C. Roman, O. Pizarro, R. M. Eustice, and A. Can, "Towards high-resolution imaging from underwater vehicles," *Int. J. Robot. Res.*, vol. 26, no. 1, pp. 55–74, Jan. 2007.

Oscar Pizarro (S'92–M'04) received the Engineer's degree in electronic engineering from the Universidad de Concepcion, Concepcion, Chile, in 1997, and the M.Sc. OE/EECS degree and the Ph.D. degree in oceanographic engineering from the Massachusetts Institute of Technology/Woods Hole Oceanographic Institution (WHOI) Joint Program, Cambridge/Woods Hole, MA, in 2003 and 2005, respectively.

His research is focused on underwater imaging and robotic underwater vehicles. He is currently working on robotic and diver-based optical imaging at the Australian Centre for Field Robotics, University of Sydney, Sydney, N.S.W., Australia.

Ryan Michael Eustice (S'00–M'05) received the B.S. degree in mechanical engineering from Michigan State University, East Lansing, in 1998 and the Ph.D. degree in ocean engineering from the Massachusetts Institute of Technology/Woods Hole Oceanographic Institution (WHOI) Joint Program, Cambridge/Woods Hole, MA, in 2005.

Currently, he is an Assistant Professor at the Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor. His research interests are in the areas of navigation and mapping, underwater computer vision and image processing, and autonomous underwater vehicles.



Hanumant Singh (S'87–M'95) received the B.S. degree as a distinguished graduate in computer science and electrical engineering from George Mason University, Fairfax, VA, in 1989 and the Ph.D. degree from the Massachusetts Institute of Technology/Woods Hole Oceanographic Institution (WHOI) Joint Program, Cambridge/Woods Hole, MA, in 1995.

He has been a member of the staff at WHOI since 1995, where his research interests include high-resolution imaging underwater and issues associated with docking, navigation, and the architecture of underwater vehicles.