Towards High-resolution Imaging from Underwater Vehicles

Hanumant Singh
Chris Roman
Dept. of Applied Ocean Physics and Engineering
Woods Hole Oceanographic Institution
Woods Hole, MA 02543
hsingh@whoi.edu
croman@whoi.edu

Oscar Pizarro
Australian Center for Field Robotics
University of Sydney
Sydney 2006, Australia
o.pizarro@acfr.usyd.edu.au

Ryan Eustice
Department of Naval Architecture & Marine Engineering
University of Michigan
Ann Arbor, MI 48109
eustice@umich.edu

Ali Can
GE Global Research
Niskayuna, NY 12309
can@research.ge.com

Abstract

Large area mapping at high resolution underwater continues to be constrained by sensor-level environmental constraints and the mismatch between available navigation and sensor accuracy. In this paper, advances are presented that exploit aspects of the sensing modality, and consistency and redundancy within local sensor measurements to build high-resolution optical and acoustic maps that are a consistent representation of the environment. This work is presented in the context of real-world data acquired using autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs) working in diverse applications including shallow water coral reef surveys with the Seabed AUV, a forensic survey of the RMS Titanic in the North Atlantic at a depth of 4100 m using the Hercules ROV, and a survey of the TAG hydrothermal vent area in the mid-Atlantic at a depth of 3600 m using the Jason II ROV. Specifically, the focus is on the related problems of structure from motion from underwater optical imagery assuming pose instrumented calibrated cameras. General wide baseline solutions are presented for these problems based on the extension of techniques from the simultaneous localization and mapping (SLAM), photogrammetric and the computer vision communities. It is also examined how such techniques can be extended for the very different sensing modality and scale associated with multi-beam bathymetric mapping. For both the optical and acoustic mapping cases it is also shown how the consistency in mapping can be used not only for better global mapping, but also to refine navigation estimates.

KEY WORDS—underwater vehicles, structure from motion, bathymetric mapping, computer vision

1. Introduction

A number of oceanographic applications require large area site surveys from underwater imaging platforms. Such surveys are typically required to study hydrothermal vents and spreading ridges in geology (Yoerger et al. 2000), ancient shipwrecks and settlements in archaeology (Ballard et al. 2002), forensic studies of modern shipwrecks and airplane accidents

The International Journal of Robotics Research
Vol. 26, No. 1, January 2007, pp. 55-74
DOI: 10.1177/0278364907074473

Figures appear in color online: http://ijr.sagepub.com

(Howland 1999; NTSB 2002), and surveys of benthic ecosystems and species in biology (Singh et al. 2004a). Scientific users in these disciplines often rely on multiscalar, multisensor measurements to best characterize the environment.

At finer scales, for resolutions down to millimeters, optical imaging of the seafloor offers scientists a high level of detail and ease of interpretation. However, light underwater suffers from significant attenuation and backscatter, limiting the practical coverage of a single image to a few square meters. To cover larger areas of interest, hundreds or thousands of images may be required. The rapid attenuation of the visible spectrum in water implies that a composite view of a large area (or photomosaic) can only be obtained by exploiting the redundancy in multiple overlapping images distributed over the scene. Although there has been considerable effort in this regard for land-based applications, the constraints on imaging underwater are different and are far more difficult to deal with. Mosaicing assumes that images come from an ideal camera (with compensated lens distortion) and that either the scene is planar or the camera is undergoing purely rotational motions. Under these assumptions the camera motion will not induce parallax and therefore no 3D effects are involved and the transformation between views can be correctly described by a 2D 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 3D 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 ideas from structure from motion (SFM) and photogrammetry.

For coarser resolutions O(10 cm), but covering far greater O(10–100 m) swaths, acoustic sensing centered at several hundred kilohertz is the modality of choice. Multibeam sensors mounted on underwater platforms can provide high-resolution three-dimensional scans of the environment that can be transformed into bathymetric maps.

Unfortunately, for both optical and acoustic sensors, the fundamental limitation in converting high-resolution sensor measurements into quantitative maps is the mismatch between sensor accuracy and navigation. Due to the rapid attenuation of the electromagnetic spectrum GPS signals are not available underwater. Instead underwater imaging platforms typically rely on a combination of acoustic transponders and inertial navigation systems. Acoustic transponders (Milne 1983), like sonar systems, must trade off range for resolution. Although transponders have been built to work at frequencies as high as 300 kHz, providing centimeter-level accuracy over an area of 100 square meters, typical large area surveys utilize lower frequency (8–13 kHz) long-baseline transponders that provide meter-level accuracy across several kilometers. The deployment of such systems is nontrivial and usually requires significant time and effort as each individual transponder must be deployed and its position independently calibrated from the surface.

Inertial navigation sensors such as Doppler velocity logs used in combination with fiber optic or ring laser gyros can provide navigation estimates underwater (Whitcomb et al. 1999) that grow as a function of time (distance traveled). However, such systems inherently provide an estimate whose error characteristic grows without bound over time (distance). Although expensive, from a cost, power and size standpoint, these systems are far easier to use as they are integral to the underwater vehicle and as such do not require any extra effort for deployment and use.

In this paper we examine the role of quantitative mapping with optical and sonar sensors for underwater applications within the bounds of current sensing and navigation methodologies. We present the results of a complete system methodology that begins with sensor-specific algorithms for image enhancement in the underwater environment. We then look at the applications of structure from motion algorithms, microbathymetric mapping and visually based navigation to show that sensor consistency and registration across multiple swaths when used in combination with underwater positioning and attitude information can be used to provide superior mapping while simultaneously improving the quality of our navigation. In the rest of this paper, Section 2 looks at the "structure from motion" problem underwater, examining in detail the problem of two view image registration of optical imagery from underwater robotic vehicles. Section 3 examines the case for microbathymetric sonar mapping. Section 4 briefly discusses sensor and terrain based navigation, while we offer some concluding thoughts in Section 5.

2. Large Area Optical Imaging Underwater

The fundamental problem of obtaining a large area perspective of an underwater scene is that due to constraints associated with the attenuation and backscatter of light, which imply that one can only image a small section of the seafloor with a single image. Thus one must register multiple image frames into a single composite. The highly unstructured nature of underwater terrain, and issues associated with moving lighting on underwater robotic vehicles all combine to make image registration a difficult problem.

Considerable efforts in this regard have been made with respect to black and white imagery as well as two-dimensional photomosaicing (Marks et al. 1995; Pizarro and Singh 2003; Singh et al. 2004b). In this paper we choose to focus instead on the more challenging tasks associated with quantitative mapping underwater with color imagery.

Figure 1 outlines the basic problems that we address in this paper. Underwater lighting and the effects of large terrain (compared with the field of view) dominate the task associated with image registration underwater.

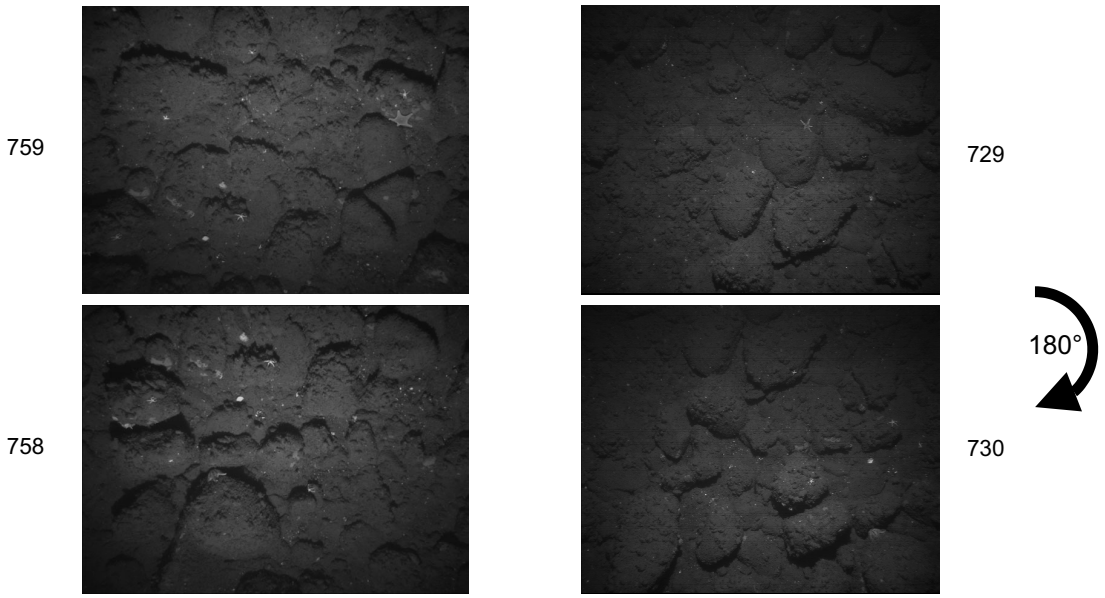

Fig. 1. Underwater lighting and large relief are the primary difficulties associated with image registration underwater. The two image pairs on the left and right have typical overlap between each other and across the pairs for the purposes of large scale optical mapping underwater. These images are as they appear when recorded from an AUV. It is extremely difficult to register such images temporally or spatially.

2.1. Color Imaging Underwater

Underwater lighting is one of the primary difficulties associated with optical imaging underwater. Seawater serves to nonlinearly attenuate different parts of the color spectrum as a function of distance (Mobley 1994; Mertens 1970).

The attenuation of visible light underwater at a particular wavelength may be modeled by a simple exponential function with an attenuation coefficient that is a function of absorption and scattering. However, this attenuation coefficient is a complex function of the wavelength. For each wavelength, the value of the attenuation coefficient depends on the depth and salinity of the water. The changes due to depth are significant in the upper layers of the water, with the dominant effects occurring in the top 200 m, while the salinity level as well as organic/inorganic materials and organisms also significantly affect the attenuation coefficient (Mobley 1994; Mertens 1970).

In addition to the attenuation, the lighting pattern of a light source affects the spatial variations of the light on the seafloor Jaffe 1990). The light source is usually a single strobe for AUVs due to cost, size and power limitations. A combination of different light sources, such as halogen-gas light, incandescent light and strobes can be used to generate a more uniform lighting pattern for ROVs and manned submersibles but even so there is a strong dependence on altitude that affects such geometries, which leads to a non-uniform lighting pattern over the entire domain of application.

For the imaging model, we use a common assumption that decomposes an image F into a reflectance image R, and an illumination image I (Highnam and Brady 1997; Chen 1990; Yoon and Ro 2002; Rzhanov et al. 2000)[14][15][16][17],

$$F(x, y, \lambda) = I(x, y, \lambda)\, R(x, y, \lambda) \tag{1}$$

where x, y represent the spatial pixel coordinates, and λ represents the color channel (i.e., red, green, or blue). Since the illumination component is formed by a slowly varying light field over a smooth surface, it is characterized by low-frequency fields. The high-frequency components are those associated with the reflectance component, which varies rapidly, and forms the local contrast.

We utilize the logarithm of the input image to estimate the reflectance image, by fitting a parametric surface to estimate the illumination image. For single-strobe underwater vehicle imagery, in the log domain, the slowly varying characteristic of the illumination can be captured by a parametric surface. The advantage of this method over more classical homomorphic filtering is that it is less sensitive to local intensity variations such as shadows and backscatter.

We start by noting that the variance of the reflectance component is small relative to its unit mean ($R \equiv N(\mu = 1, \sigma^2(\lambda))$ and $\sigma^2(\lambda) << 1$), and thus a logarithmic transformation yields approximately a Gaussian process with zero mean. This can be justified by approximating the logarithmic

transformation as a unit transformation around one,

$$\log R \approx R - 1 \equiv N(\mu = 0, \sigma'^2(\lambda)). \tag{2}$$

For a particular frame, and a color channel let $F_L(x, y) = \log F(x, y, \lambda)$ denote the logarithm of the input image. If we assume that the illumination component can be represented by a polynomial surface (experimentally we have found that a fourth order polynomial closely represents the light pattern formed by a single strobe in an attenuating media), then the parametric surface fitting equation can be formulated as

$$F_L \equiv \begin{pmatrix} F_L(0,0) \\ F_L(0,1) \\ \vdots \\ F_L(x,y) \\ \vdots \\ F_L(N,M) \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 1 \\ \vdots & & & & \vdots \\ x^4 & x^3 y & & y & 1 \\ \vdots & & & & \vdots \\ N^4 & N^3 M & \cdots & M & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{14} \\ p_{15} \end{pmatrix} \equiv \mathbf{S}P \tag{3}$$

where S refers to surface fitting parameters for each pixel, P is the parameter vector and N and M refer to the size of the image.

The least squares estimate for the parameter vector is

$$P = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T F_L. \tag{4}$$

Note that the $(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T$ term in the above equation is independent of the input image, and hence needs only to be computed once. Then for each frame and color channel, an illumination image can be generated by using the estimated polynomial surface,

$$\log I(x, y, \lambda) = \mathbf{S}P \tag{5}$$

The block diagram for this method is shown in Figure 2(a). Note that, this is similar to the well known process of homomorphic filtering except that the log of the illumination component is estimated using a parametric surface rather than a low pass filter. The imagery in Figure 2(b) and 2(c) shows the original image, the illumination component and the color compensated image computed by this method. We note that an implicit assumption in our formulation is that there is sufficient dynamic range in our pixel values to support logarithmic analysis. Given the nature of the underwater medium it is thus imperative to utilize high dynamic range cameras. All the optical imagery shown in this paper was acquired with cameras providing twelve bits of dynamic range.

We apply this methodology to all our images prior to image registration.

2.2. Structure from Motion with Pose Instrumented Cameras

After dealing with the effects of color, we turn our attention to the "structure from motion" (SFM) problem underwater. Our methodology takes a local to global approach inspired by mosaicing and other land-based applications of SFM (Hartley and Zisserman 2000; Pollefeys et al. 1999; Triggs et al. 2000; Fitzgibbon and Zisserman 1998; Slama 1980), but differs in that it takes advantage of navigation and attitude information (Pizarro et al. 2003). Local sequences are derived independently (Fitzgibbon and Zisserman 19980 and then registered in a global frame for bundle adjustment (Triggs et al. 20000. Our approach seems more suitable than pure sequential methods (Pollefeys et al. 1999; Fitzgibbon and Zisserman 1998) because in an underwater survey each 3D feature appears only in a few images making the global solution look more like a series of weakly correlated local solutions.

We assume a standard calibrated pin-hole camera model which implies that the homogeneous mapping from the world to the image plane can be described by the projection matrix P defined as

$$P = K[{}^c_w R | {}^c t_{cw}] \tag{6}$$

where ${}^c_w R$ and ${}^c t_{cw}$ encode the the coordinate transformation from world w, to camera centered coordinate frame c and K is the intrinsic camera calibration matrix.

Under this representation the interest point with pixel coordinates (u, v) in image I is imaged as

$$u = PX \tag{7}$$

where $u = [u, v]P^{TP}$, is the vector description of (u, v), $u = [uP^{TP}, 1]$ its normalized homogeneous representation, $X = [X, Y, Z]P^{TP}$ is the imaged 3D scene point, and $UXU = [XP^{TP}, 1]$ its normalized homogeneous representation. We note that for all homogeneous quantities, equality as defined by (7), is implicitly defined up to scale. The benefit of having a calibrated camera is that we can work with normalized image plane coordinates and thus describe the geometry in terms of the Essential matrix (Rzhanov et al. 2000) and recover the 5-degree-of-freedom camera pose from correspondences as described below.

Our feature-based approach to image registration is outlined in Figure 4. A combination of Harris and SIFT interest points are extracted for each image. For the Harris points we apply an orientation normalization, based upon the underwater robot's pose, before encoding the surrounding region with Zernike moments (Pizarro and Singh 2003). Correspondences are then posited based upon similarity and a pose constrained correspondence search. These are then fed into a robust least median of squares registration methodology with regularized sampling to extract a consistent inlier correspondence set. We then solve for the relative pose estimate using the inlier set and

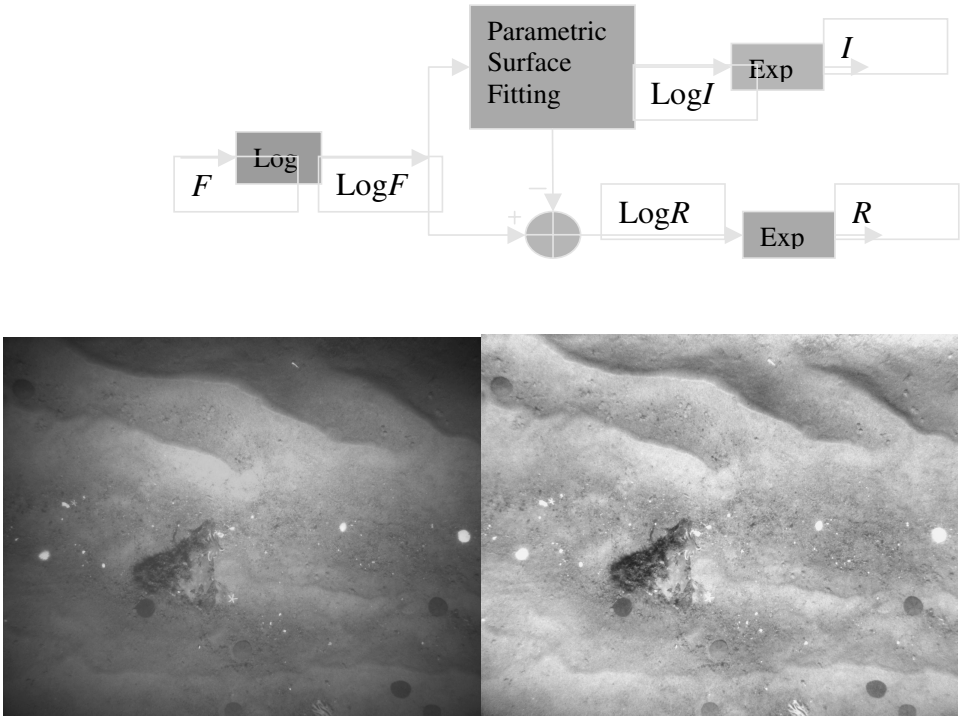

Fig. 2. High dynamic range cameras can be used with parametric surface fitting in the homomorphic domain as outlined (a) to color correct typical underwater imagery (b) to obtain an image (c) that is independent of the nonlinear attenuation of color underwater.

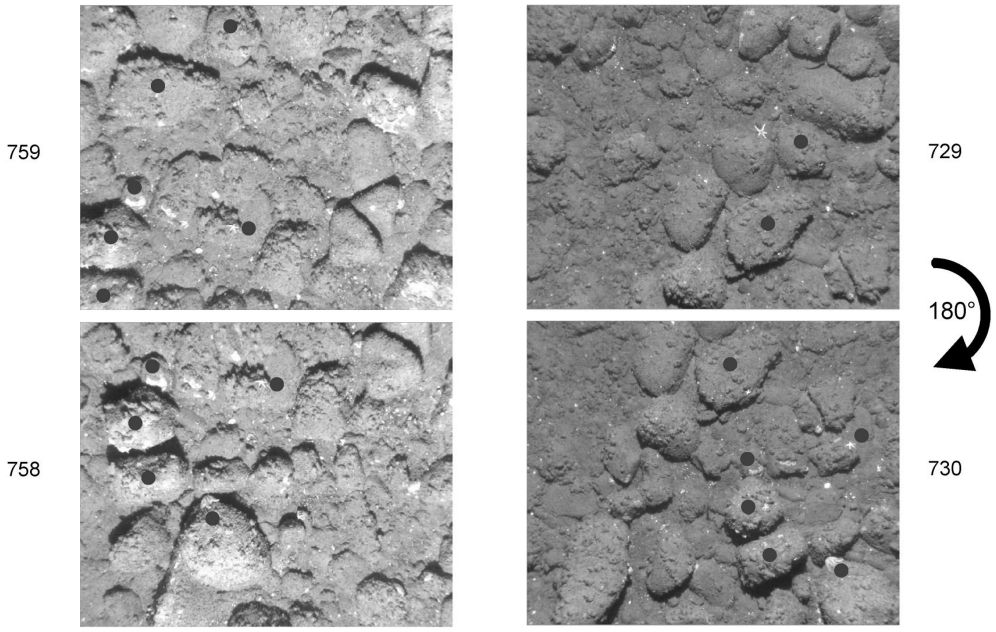

Fig. 3. The use of high dynamic range cameras (as shown in Figure 2) can help compensate for lighting differences between the two sets of images. However, it is still difficult to disambiguate the effects due to large relief with respect to the field of view of the camera. Even with the colored dots to show corresponding points across two sets of images it is difficult to register such images using conventional computer vision methodologies.

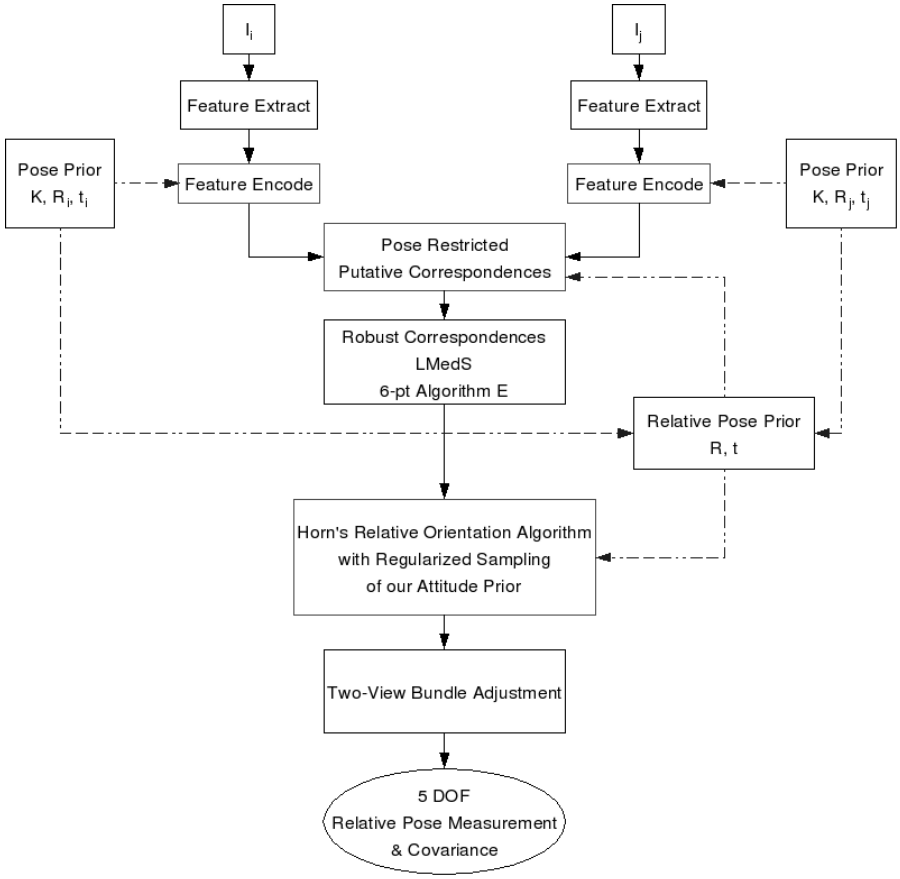

Fig. 4. An overview of the pairwise registration methodology.

Horn's relative orientation algorithm. Finally we refine our estimate utilizing a two view bundle adjustment step based upon minimizing the reprojection error over all the inliers.

The problem of initial feature correspondence is arguably the most difficult and challenging task for most feature-based registration. The pose prior (available on our underwater robot) can be used to relax the demands on the complexity of the feature descriptor. We use the epipolar constraint expressed as a two view point transfer model to restrict the correspondence search to probable regions. These regions are determined by our pose prior and are used to confine the interest point matching to a small subset of candidate correspondences.

In deriving the point transfer mapping we assume the projective camera matrices $P = K[I|0]$ and $P\prime = K[R|t]$, where for notational convenience we drop the explicit subscript/superscript notation and simply write the relative pose parameters as R, t. We begin by noting that the scene point X is projected through camera P as

$$u = PX = KX \tag{8}$$

which implies that explicitly accounting for scale we have

$$X \equiv ZK^{-1}u \tag{9}$$

The back projected scene point, X, can subsequently be reprojected into image I' as

$$u' = P'X = K(RX + t) \tag{10}$$

By substituting (9) into (10) and recognizing that the following relation is up to scale, we obtain the homogeneous point transfer mapping [17],

$$u' = KRK^{-1}u + Kt/Z \tag{11}$$

Finally, by explicitly normalizing (11) we recover the non-homogeneous point transfer mapping,

$$u' = \frac{H_\infty u + Kt/Z}{H_\infty^{3T} u + t_z/Z} \tag{12}$$

where $HB_{\infty B} = KRKP^{-1P}$, H_∞^{3T} refers to the third row of $HB_{\infty B}$, and tB_{zB} is the third element of t. When the depth of the scene point Z, is known in camera frame c, then (12)

describes a functional relation on Z (i.e., $u\prime = f(u, Z)$) that traces out the corresponding epipolar line in $I\prime$.

The two view point transfer mapping can be used to constrain the correspondence search between image pair (IB_{iB}, IB_{jB}) by using the pose prior for each of the camera vehicle poses. We define the parameter pose vector γ as

$$\gamma = [x_{p_i}^T, x_{p_j}^T, Z, u, v]^T \tag{13}$$

with mean, μ_γ, and covariance Σ_γ, given by

$$\mu_\gamma = \begin{bmatrix} \mu_{p_i} \\ \mu_{p_j} \\ Z \\ u \\ v \end{bmatrix}$$

and

$$\Sigma_\gamma = \begin{bmatrix} \Sigma_{p_i p_i} & \Sigma_{p_i p_j} & 0 & 0 & 0 \\ \Sigma_{p_j p_i} & \Sigma_{p_j p_j} & 0 & 0 & 0 \\ 0 & 0 & \sigma_z^2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where x_{p_i}, x_{p_j} refer to the vehicle pose at the moment the camera images i and j were acquired, Z and σB_{zB} represent the scene depth parameters as measured in camera frame I and (u,v) describe the feature location in pixels in image IB_{iB}.

To obtain a first-order estimate of the uncertainty in the point transfer mapping between IB_{iB} and IB_{jB} we compute

$$\mu_{u'} \approx (12)|_{\mu_\gamma} \tag{14}$$

$$\Sigma_{u'} \approx J \Sigma_\gamma J^T \tag{15}$$

where $\mu_{u'}$ is the predicted point location of u in IB_{jB}, $\Sigma_{u'}$ its covariance, and $J = \frac{\partial u'}{\partial \gamma}$ is the point transfer Jacobian (which we compute numerically).

We use the Gaussian distribution as an analytical tool to compute the first-order search bounds in $(u\prime, v\prime)$ space by noting that

$$(u' - \mu_{u'})^T \Sigma_{u'}^{-1} (u' - \mu_{u'}) = k^2 \tag{16}$$

defines an ellipse where k^2 follows a χ^2 distribution. Hence we can choose an appropriate k^2 such that with high probability α, the true mapping $u_{o'}$ falls within this region. Under this scheme we test all feature points in I_j to see if they fall within the ellipse, and if they do, then they are considered to be candidate matches for u. Since relative pose uncertainty depends on the reference frame in which it is expressed, we apply the two view search constraint both forwards and backwards to obtain a consistent candidate correspondence set. In other words, candidate matches in I_j that correspond to interest points in I_i are checked to see if they map back to the generating interest point in I_i. Based upon this set of consistent candidate matches, feature similarity is then used to establish the one-to-one putative correspondence set.

2.3. Large Area Structure from Motion

The temporal sequence of images is processed into a set of 3D submaps with estimates of coordinate transformations between temporally adjacent submaps. Our algorithm attempts to establish additional spatial relationships between submaps corresponding to overlap from parallel tracklines or loop closures.

While the sparse set of 3D points contained in the submaps do not consistently offer discriminating structure, the very fact that they exist as 3D points implies that their appearance in multiple views is characteristic enough to effectively establish correspondences and be reconstructed by the SFM algorithm. We therefore extend the feature description and similarity based matching between images to matching submaps by relying on the appearance of 3D points to propose corresponding features between submaps. The average of the descriptors of the 2D neighborhoods on all views is used as the appearance of the 3D point. The underlying assumption is that a similarity measure which was effective to match 3D points along track will also be effective when matching across submaps. Corresponding 3D points are proposed based on appearance and a robust registration using RANSAC with Horn's (1987) algorithm is used to determine which points are in correspondence and the transformation parameters.

We continue to check possible submap correspondences based on our navigation until all possible correspondences are exhausted or an upper limit is reached (we use eight times the number of submaps for the sparse structure associated with local area mapping). The submaps are then placed in a global frame by minimizing the descrepancies between composed global estimates and the transformations between submaps. Additional cost terms consider the navigation prior.

Once submaps are in a global frame, camera poses within submaps can also be placed in the global frame. These camera poses are then used to triangulate the location of 3D features. Sparse bundle adjustment then refines both camera poses and 3D feature locations.

Figures 8 and 9 illustrate this process. The results are from a survey performed in the Johns Hopkins University (JHU) Hydrodynamics Test Facility using the JHU ROV. As shown in the figure the results are highly consistent with ground-truth obtained by draining the test tank and laser scanning the scene geometry. We have also obtained similar results from a survey using the Seabed AUV at a coral reef off Bermuda (Singh et al. 2004a).

3. Self-consistent Bathymetric Mapping

Another application of our techniques is to multibeam mapping (Singh et al. 2000) where the areas of interest encompass several square kilometers that are typically mapped with a sonar with 10 cm sensor accuracy but where the navigation from the standard combination of long baseline transpon-

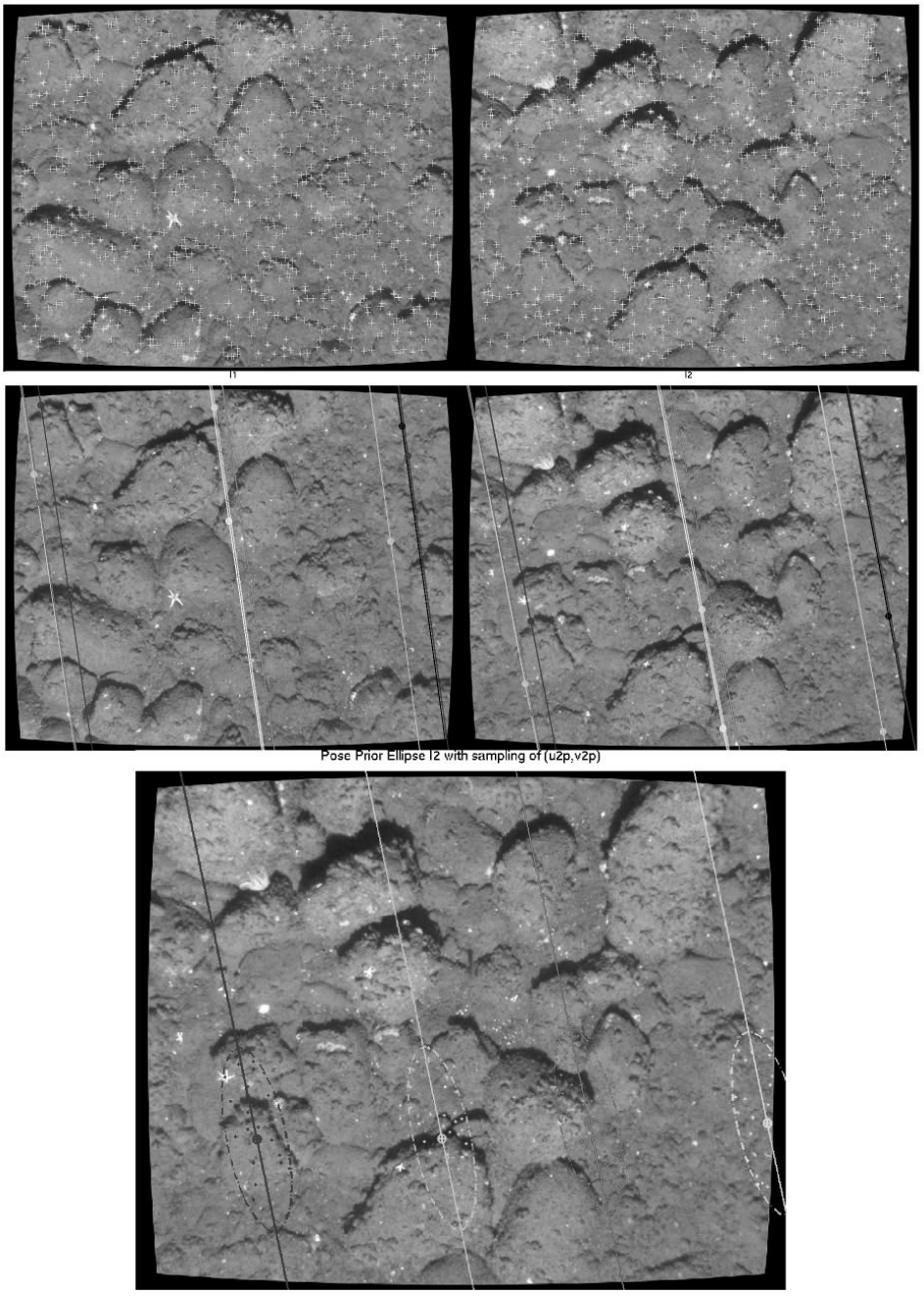

Fig. 5. Pose-restricted correspondence search on a pair of underwater reef images. A sampling of interest points are shown for the image pair in the top row. The middle row shows the color coded sensor instantiated epipolar lines for a small number of sample points. The bottom image shows the constrained search regions for the interest points in the right image based on our two view point transfer mapping.

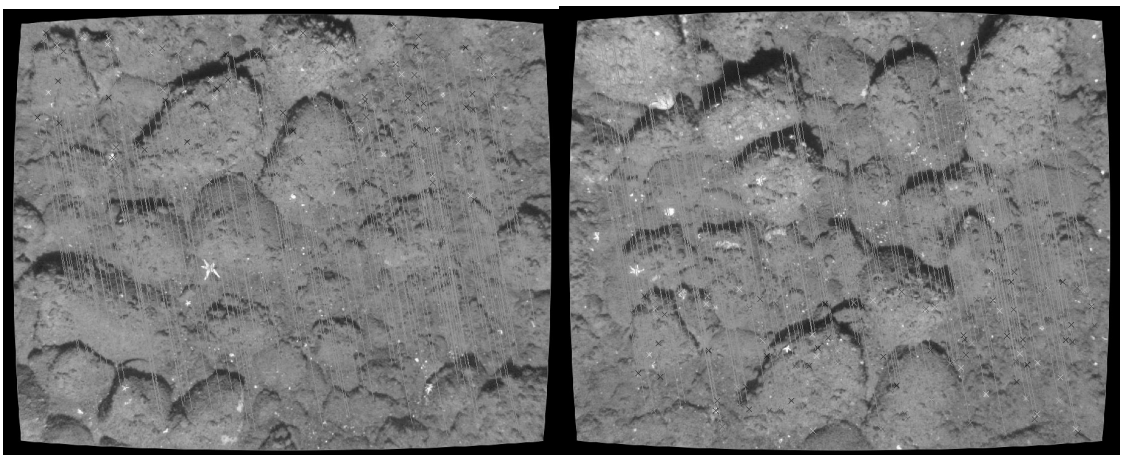

Fig. 6. By utilizing the pose constraints, we can efficiently obtain consistent correspondences between the images. The x's denote individual features in each image that have been registered to the other image. The lines show where each successful correspondence is registered on the other image.

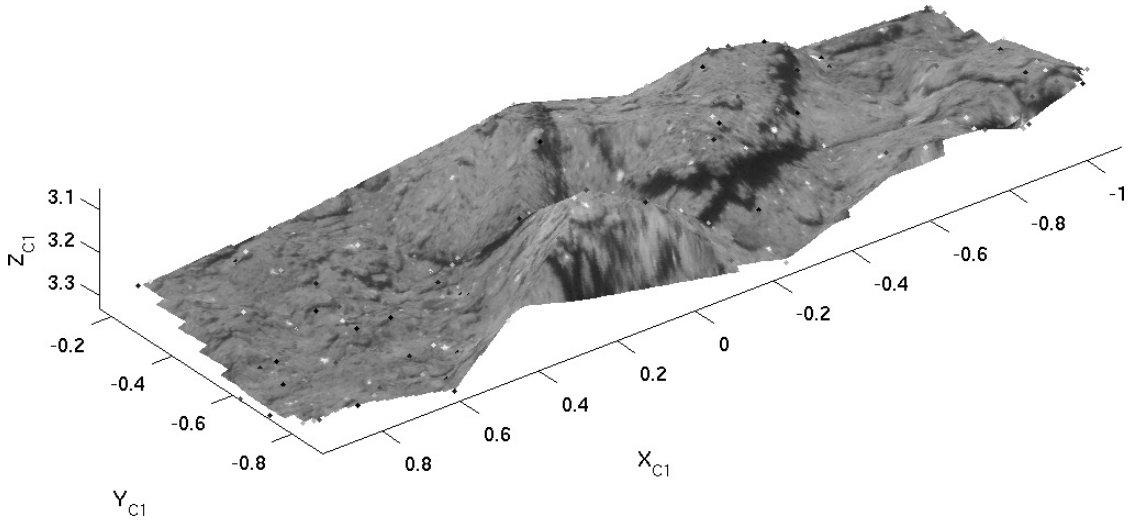

Fig. 7. The set of inliers in the correspondence set can be used to derive a coarse 3D grid of structure. Texture mapping the imagery corresponding to the structure provides a vivid quantitative representation of the scene.

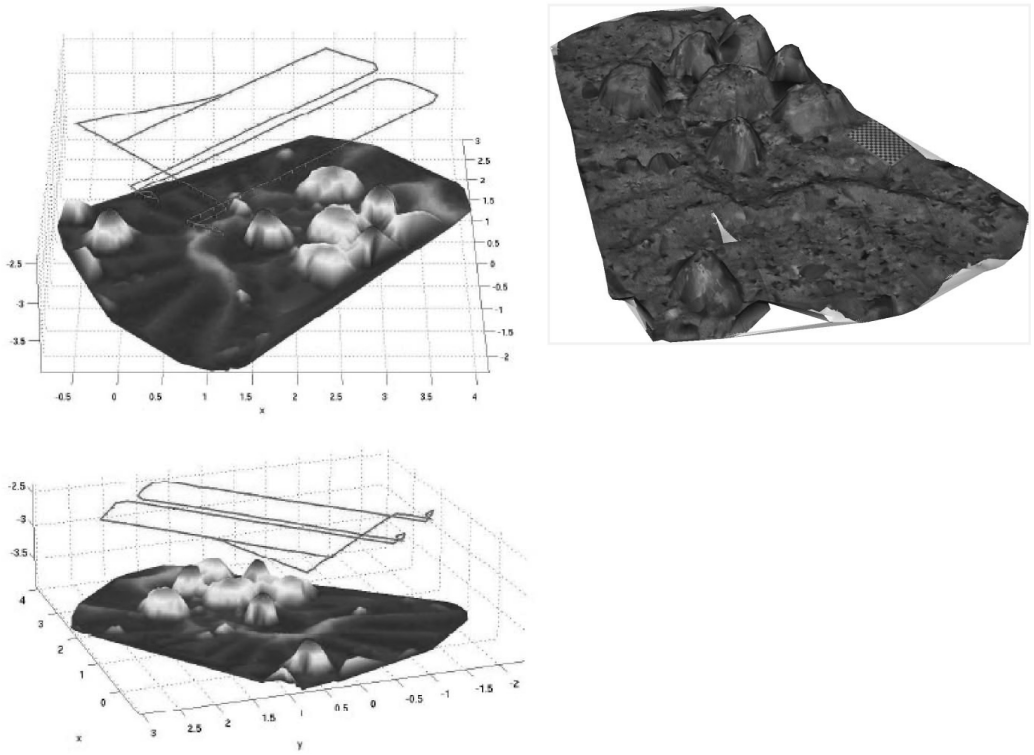

Fig. 8. Two views of the reconstruction of poses and structure for the JHU tank dataset (left). The scene consists of rocks placed on carpet. The camera poses are connected by a red line. A Delaunay triangulation interpolates a surface between 3D feature points. The structure is color-coded according to height. Units are in meters. A view of the texture mapped imagery on the structure is shown for comparison at right.

ders and inertial navigation is only good to a meter. To avoid this navigation limitation we break the total mapping problem down into small pieces, each of which contains internal errors typical of the mapping sonar rather than the navigation (Roman 2005). This is accomplished by assembling small bathymetry *sub-maps* using only the short-term dead reckoning information provided by the vehicle navigation sensors. Algorithmically this is accomplished using a delayed state extended Kalman filter (EKF) and a simple constant velocity dynamic model of the vehicle motion, as illustrated in Figure 10. This simple model is sufficient, given the slow dynamics typical of underwater survey vehicles. The current estimate of the filter state vector contains the position and velocity information required for a 6 degree-of-freedom (DOF) state estimate.

The delayed portion of the state vector is used to archive historical poses of the vehicle which serve as local 6-DOF origins for the small sub-maps.

After accounting for issues specific to acoustic sensors such as possible errors associated with weak returns, beam patterns effects resulting in the acoustic pulse not striking the bottom, and other false detections, we can approximate the sonar as a three-dimensional line scanner. These line scans are assembled into sub-maps using the range data and the vehicle position estimates extracted from the state vector at the time each sonar ping is taken as illustrated in Figure 11.

3.1. An EKF Formulation for Vehicle Navigation

The complete filter state vector x_{aug} is partitioned into a vehicle state x_v, which describes the current estimate of the vehicle pose, and a delayed portion containing the sub-map origins. The state vector in (17) shows the vehicle state and the N delayed sup-map origins.

$$x_{aug} = [x_v^\top, x_{s_1}^\top, \cdots, x_{s_N}^\top]^\top \tag{17}$$

This state vector will grow in length as new sub-maps are created and delayed states are added to the filter. The notation for the delayed states indicate that the delayed state x_{s_1}, marked by subscript s, points from the common origin o to

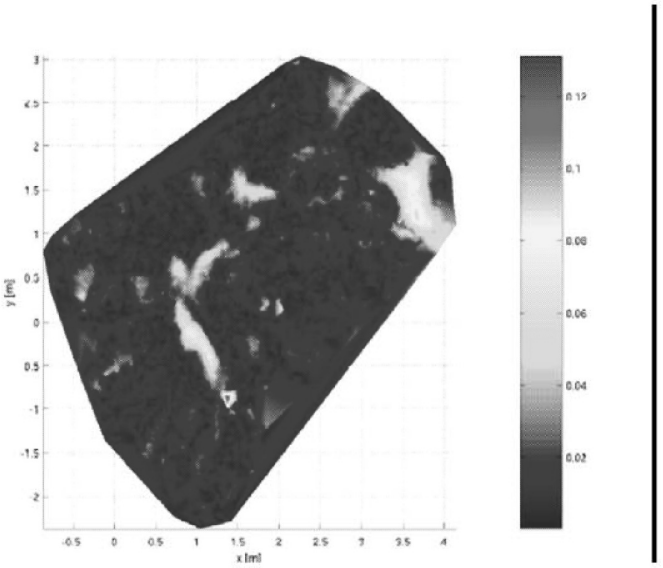

Fig. 9. The distance map from SFM 3D points to the ground truthed laser scan after ICP registration. Areas of large discrepancies tend to correspond to the carpet being buoyant for the visual survey. An outlier in the reconstruction produced the large error visible at approximately $x = 1.4$ m, $y = 0.8$ m.

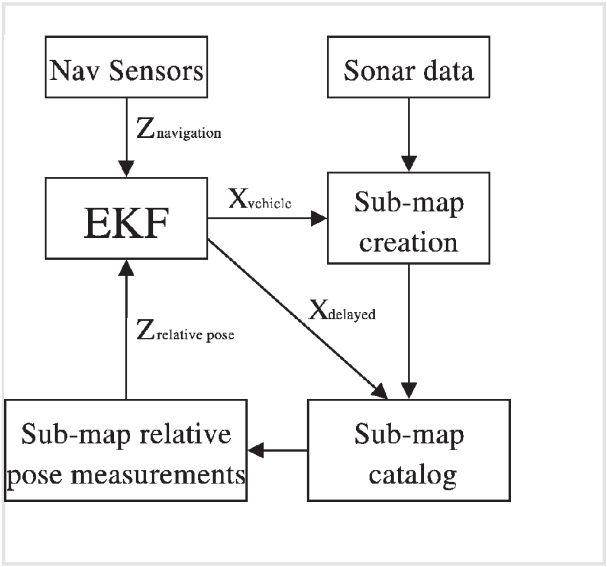

Fig. 10. The delayed state EKF block diagram. The sub-mapping algorithm utilizes vehicle navigation data to create small bathymetric sub-maps. The sub-map origins are held in a delayed state vector and used to create relative pose measurements that reduce navigation error.

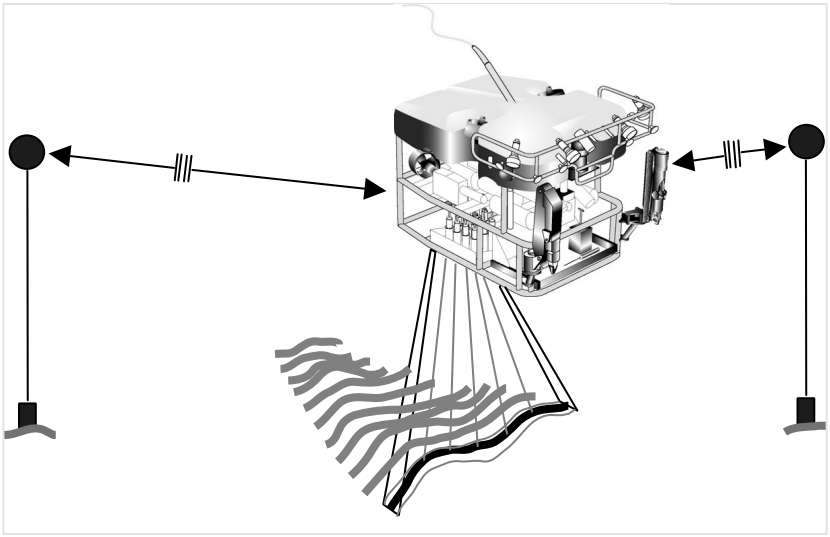

Fig. 11. A schematic representation of multibeam sonar mapping. The underwater vehicle projects a linear beam on the seafloor that is combined with vehicle navigation based on acoustic beacons to obtain a global map of the environment. Typically the navigation errors far exceed the intrinsic accuracy of the sonar sensor.

the origin of sub-map 1. The delayed state poses consist of 6-DOF frames defined by:

$$x_{s_i} = [x, y, z, \theta, \phi, \psi]^\top \tag{18}$$

where θ, ϕ, ψ are Euler angles.

The covariance matrix for the filter describes the covariance of the vehicle P_{vv}, the covariance of the sub-map origins $P_{s_i s_i}$, and all of the respective cross-correlations. As the new sub-maps are created additional rows and columns are added to the covariance matrix. These new elements are non-zero because the current state is correlated with all prior states. Intuitively, setting the covariance of a newly added delayed state $P_{s_N s_N}$, equal to the current vehicle covariance allows that sub-map origin to inherit the correct uncertainty of its position estimate.

$$P_{aug} = \begin{bmatrix} P_{vv} & P_{vs_1} & \cdots & P_{vs_N} \\ P_{s_1 v} & P_{s_1} & \cdots & P_{s_1 s_N} \\ \vdots & \vdots & \ddots & \vdots \\ P_{s_N v} & P_{s_N s_1} & \cdots & P_{s_N s_N} \end{bmatrix} \tag{19}$$

To describe the pose of the vehicle we use a 6-DOF parameterization with position and attitude variables measured in a local level reference frame. The additional states included in the vehicle state vector are the vehicle body frame velocities and angular rates.

$$x_v = [\underbrace{x, y, z, \theta, \phi, \psi}_{\text{position}}, \underbrace{u, v, w, \dot{\theta}, \dot{\phi}, \dot{\psi}}_{\text{velocity}}]^\top. \tag{20}$$

For a vehicle dynamic model we consider a constant velocity model $f(x_v(t))$ perturbed by white noise w with zero mean and diagonal covariance Q. The constant velocity assumption seems adequate for this application where the vehicle dynamics are relatively slow. This vehicle model relates the vehicle body frame velocities to local level frame velocities through the nonlinear rotation $R(\theta, \phi, \psi)$. The white noise w adds to the linear and angular accelerations.

$$\dot{x}_v(t) = f(x_v(t)) + w(t) \tag{21}$$

$$= \begin{bmatrix} R(\theta, \phi, \psi) \begin{bmatrix} u \\ v \\ w \end{bmatrix} \\ \begin{bmatrix} \dot{\theta} \\ \dot{\phi} \\ \dot{\psi} \end{bmatrix} \\ O_{[6\times 1]} \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ w_{[6\times 1]} \end{bmatrix} \tag{22}$$

where, $w_{[6\times 1]} = [w_1, w_2, w_3, w_4, w_5, w_6]^\top$. When the entire augmented state is considered, the delayed states are not affected by the vehicle model, that is, $\dot{x}_{aug} = \left[\dot{x}_v(t)^\top, 0^\top_{[6N\times 1]}\right]^\top$.

Navigation sensor measurements,$z[t_k]$, are handled using nonlinear measurement models of the form $h_n(x_v[t_k])$. These measurement models are implemented as *mixed-coordinate* functions that predict the sensor measurement in the individual sensor coordinate systems. The sensor measurements are assumed to be corrupted by a time-independent zero mean Gaussian noise v with covariance R, where $E[wv^\top] = 0$.

Measurements for the navigation sensors are available at discrete times t_k.

$$z[t_k] = h_n(x_v[t_k]) + v \tag{23}$$

For this application we incorporate navigation measurements of the body frame velocities, surface relative depth and three axis attitude. We use the LBL position estimates to evaluate the output of the simultaneous localization and mapping (SLAM) algorithm but do not incorporate them or any other ground reference position estimates in the filter. Thus, the vehicle model within this framework is dead reckoning.

To make measurements between delayed states we use the tail to tail operation as defined by Smith et al. (1990). This nonlinear operation predicts the relative 6-DOF transform $x_{s_{ij}}$ between the filter states x_{s_i} and x_{s_j}.

$$z_{s_{ij}} = h_s(x_{s_i}, x_{s_j}) = \ominus x_{s_i} \oplus x_{s_j} \tag{24}$$

$$H_{s_{ij}} = [0, \frac{\partial x_{s_{ij}}}{\partial x_{s_i}}, 0, \frac{\partial x_{s_{ij}}}{\partial x_{s_j}}, 0] \tag{25}$$

The sparse Jacobian $H_{s_{ij}}$ of the relative pose estimate is used in the update equations when the measurement is incorporated into the filter.

The EKF uses the continuous time prediction equations to propagate the state estimate forward in time until the next navigation or delayed state measurement can be made.

$$\dot{\hat{x}}_v(t) = f(x_v(t)) \tag{26}$$

$$\dot{P}_{aug}(t) = \begin{bmatrix} F_v(t) & 0 \\ 0 & 0 \end{bmatrix} P_{aug}(t) + P_{aug}(t) \begin{bmatrix} F_v(t) & 0 \\ 0 & 0 \end{bmatrix}^\top + \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \tag{27}$$

These equations and the update equations require the Jacobians of the process model and the navigation sensor measurements with respect to the vehicle state:

$$F_v = \frac{\partial f(x(t))}{\partial x_v(t)}_{x_v(t)} \qquad H_n = \frac{\partial h_n(x[t_k])}{\partial x_v[t_k]}_{x_v[t_k]} \tag{28}$$

The prediction equations are nonlinear and are integrated numerically using a Runge–Kutta approximation. The integration produces the prediction of the mean vehicle state $\bar{x}_v^-$ and covariance P_{aug}^- using all the measurements prior to time t_k.

The update equations are then written in a discrete form.

$$W = P_{aug}^- H^\top \left[H P_{aug}^- H^\top + R \right]^{-1} \tag{29}$$

$$\bar{x}_v[t_k] = \bar{x}_v^- + W[z[t_k] - h(\bar{x}_v^-)] \tag{30}$$

$$P_{aug}[t_k] = [I - WH]P_{aug}^-[I - WH]^\top + WRW^\top \tag{31}$$

where, $H = \left[H_n, 0_{[m \times 6N]} \right]$, when the update is for a navigation sensor updating m states, and $H = H_{s_{ij}}$ when the update is for a delayed state measurement. The matrix R contains the appropriate measurement covariance for the navigation sensor or relative pose measurement.

3.2. Sub-map Creation and Matching

Prior to mapping the raw sonar data collected for each ping is beamformed and range detected. Range detection is accomplished by looking at each beam and choosing the range at which the peak amplitude was returned. To reduce possible errors associated with weak returns, beams not striking the bottom and other false detections so ranges are thrown out using amplitude thresholding and median filtering with neighboring beams.

Sub-maps are created using the range data and the vehicle position estimates extracted from $x_v(t)^-$ at the time each ping is taken. The beam ranges are projected and kept as points in 3D dot cloudsS_i using the delayed states $i = 1 \cdots N$ as local origins. Using the notation from Smith et al. (1990) the locations of an individual point in space can be written as

$$s_i = (\ominus x_{s_i} \oplus x_p(t)^-) \oplus x_r \tag{32}$$

where, x_r is the location of the range point in the vehicle coordinate frame and $x_p(t)^-$ is the vehicle pose extracted from $x_v(t)^-$.

Sub-maps are created by collecting pings in this manner over short time scales during which the vehicle position error associated with the dead reckoned navigation grows slowly. Determining the size at which to break a terrain map and begin another is a balance between creating a map small enough that it itself does not contain a significant amount of internal error or distortion and creating a map large enough to contain enough 3D information that it can be registered to another map.

Given these criteria there are a few obvious limitations in applying this technique. For starters, the dead reckoned navigation must be reasonable enough that sub-maps can be reasonably constructed. Also, the seafloor cannot be flat and featureless. Fortunately, there is little interest in mapping such areas.

In an effort to algorithmically break and initiate sub-maps we monitor the characteristics of the sub-maps as they are created. To estimate the amount of 3D information in a sub-map we monitor the eigenstructure of the principal component matrix

$$\frac{1}{N} \sum_{N}^{\Sigma} i = 1 \tag{33}$$

where, $s_i = [x, y, z]^\top$ is one of N points in the dot-cloud S_i. A minimum map size threshold is defined by the condition

number of this matrix. If this matrix is poorly conditioned the sub-map is dominantly planar or too small to contain sufficient 3D information.

The maximum map size is limited by comparing a measure of vehicle covariance relative to the current sub-map's origin to a threshold. This threshold is set proportional to the approximate accuracy of the sonar range detection and an estimate of the potential bias between the heading sensor and the velocity sensor. When the threshold is exceeded a new map is started, independent of the minimum map size condition. Intuitively this test breaks the sub-map before the dead reckoned navigation error dominates the mapping error. The navigation uncertainty for this comparison can be determined on line using the Jacobian H_s associated with the relative pose operation $h_s(x_{s_i}, x_p)$ and the current covariance matrix P_{aug} as

$$P_{s_i p} = H_s P_{aug} H_s^\top \tag{34}$$

As a measure of uncertainty we use the determinant of the upper left 3×3 partition of $P_{s_i p}$ as a volumetric error measurement.

Sub-map registration is accomplished using a two-step process. For coarse alignment overlapping sub-map dot clouds are transformed to a common coordinate frame using the predicted relative pose and then gridded and matched using a 2D correlation method. Experimentally, correlation has proven to be a robust method of providing an initial guess for a final registration based on the ICP algorithm. The 2D transform $\Delta_{corr} = [\delta_x, \delta_y]$ is determined from

$$\Delta_{corr} = \arg\min_{\Delta} \frac{1}{N} \sum_{I} (Z_1 - Z_2(\Delta)) \tag{35}$$

where, I is the number of common bins between the gridded height surface for the *base* map, Z_1, and the gridded height surface for the matching sub-map, Z_2, shifted by $\Delta = [\delta_x, \delta_y]^\top$. The gridded surfaces can be created using any gridding approach for non-uniformly sampled surfaces. We have used a Gaussian gridding technique that computes grid values based on a distance- dependent weighting of the dot cloud points.

Fine-scale sub-map registration with an ICP method is then performed using the sub-map dot clouds after they are transformed by the initial Δ_{corr}. We have implemented the typical point-to-point and point-to-plane methods for this alignment (Besl and McKay 1992; Rusinkiewic and Levoy 2001).

To reduce some computational costs the sub-map dot clouds are randomly down- sampled. The randomization helps limit the effect of the sampling pattern on the registration. If sampling effects such as striping, low point density and occlusion are present the registration can be easily biased to align artifacts in the sampling rather than the underlying terrain.

To evaluate the result of the registration we combine both sub-map dot clouds into a single dot cloud and use a local measurement of the surface variance. Without ground truth for the unstructured surface it is difficult to evaluate the absolute accuracy of the registration. By locally fitting planes to small sections of the surface that have been binned in x and y, and by calculating the standard deviation of the orthogonal projection errors to this plane, we get an estimate of the local error. Ideally, the dot cloud would represent the sampled surface and have zero thickness. Although this measure of surface variance removes the overall effect of terrain, it is still lower bounded by the sonar range accuracy and terrain shape within the individual bins. This measure is also upper-bounded by how varied the neighboring terrain is. Over rugged terrain mis-registration adds directly to the surface variance and over flat terrain mis-registration will not contribute to the surface variance. We have found that coarse correlation followed by ICP registration will reduce the surface variance indicating that the surfaces have been aligned correctly.

The results of our framework are illustrated in Figures 12 and 13 using data collected by the Jason ROV at the TAG Hydrothermal Vent Site located at a depth of 2600 m on the mid-ocean ridge in the Atlantic Ocean. One can see that the resulting map is a far better representation of the environment. We have also used consistency within the submaps to derive corrected navigation estimates for the vehicle trajectory over the course of the survey.

4. Pose-based Navigation

Note that the submap registration may also be used to infer navigation estimates for each submap origin, as illustrated in Figure 14. Such results fall out naturally from our algorithm for maintaining sensor consistency across the global framework. While these measurements are far sparser than can be traditionally obtained using classical inertial navigation systems, they trade- off errors that grow unbounded with time for errors that are are a function of spatial distance and topology.

We can further build upon the delayed-state EKF framework and two-view structure from motion case to formulate a vision-based SLAM approach to provide high precision, accurate navigation measurements on an underwater robotic vehicle. Similar to the bathymetry-based EKF submapping strategy, our methodology is to employ pairwise-generated image measurements as spatial constraints in a graph over a collection of historical vehicle poses. However, because we are able to generate camera measurements at the pairwise level, we choose instead to maintain all pose samples that are associated with image acquistion, as illustrated in Figure 15. This differs from the aggregate submapping strategy used for bathymetry-based navigation and implies that the EKF's scalability becomes a severe issue (due to the quadratic complexity of maintaining the covariance) as the image-based navigation uses orders of magnitude more delayed-states.

A well known and very attractive property of formulating SLAM in the information form is that the information matrix (as in Fisher information) has the direct interpretation as a

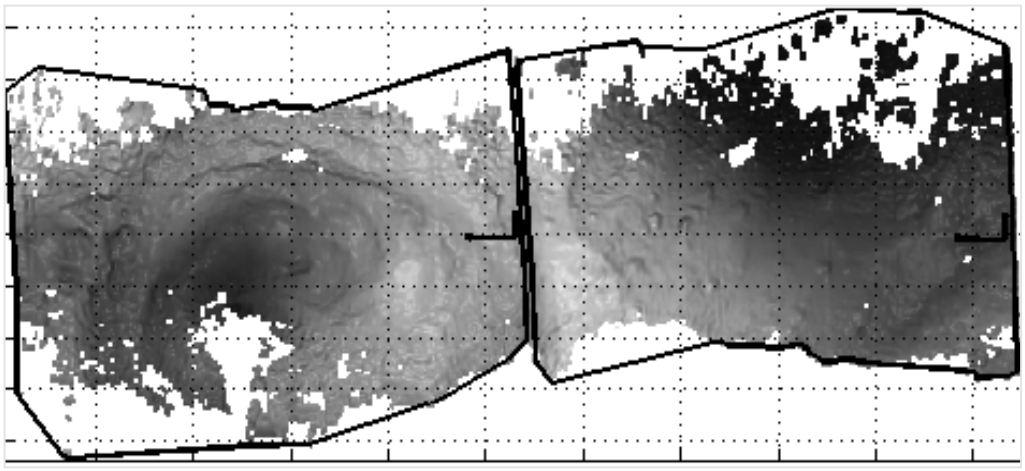

Fig. 12. Two sample sub-maps showing their outlines and local reference frames. Note the fine scale features that can be resolved individually within each sub-map. Normal smoothed navigation tends to blur these real and often significant features.

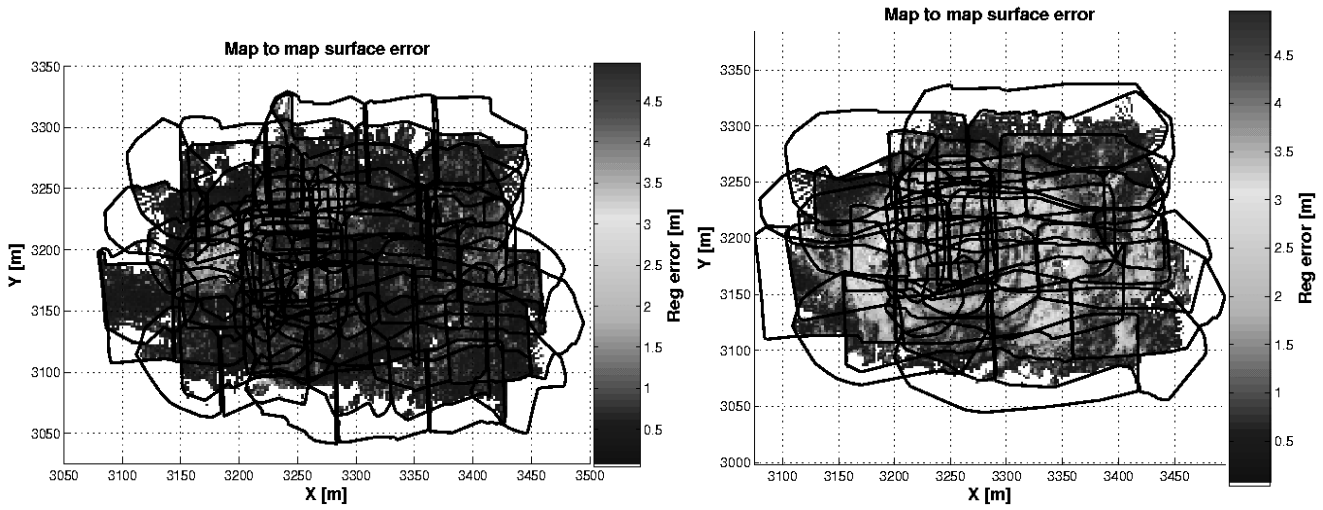

Fig. 13. Error in bathymetric mapping as measured by self consistency across multiple sub-maps. Map to map surface error for our algorithm (a) versus map to map surface error using a standard smoothing method (b).

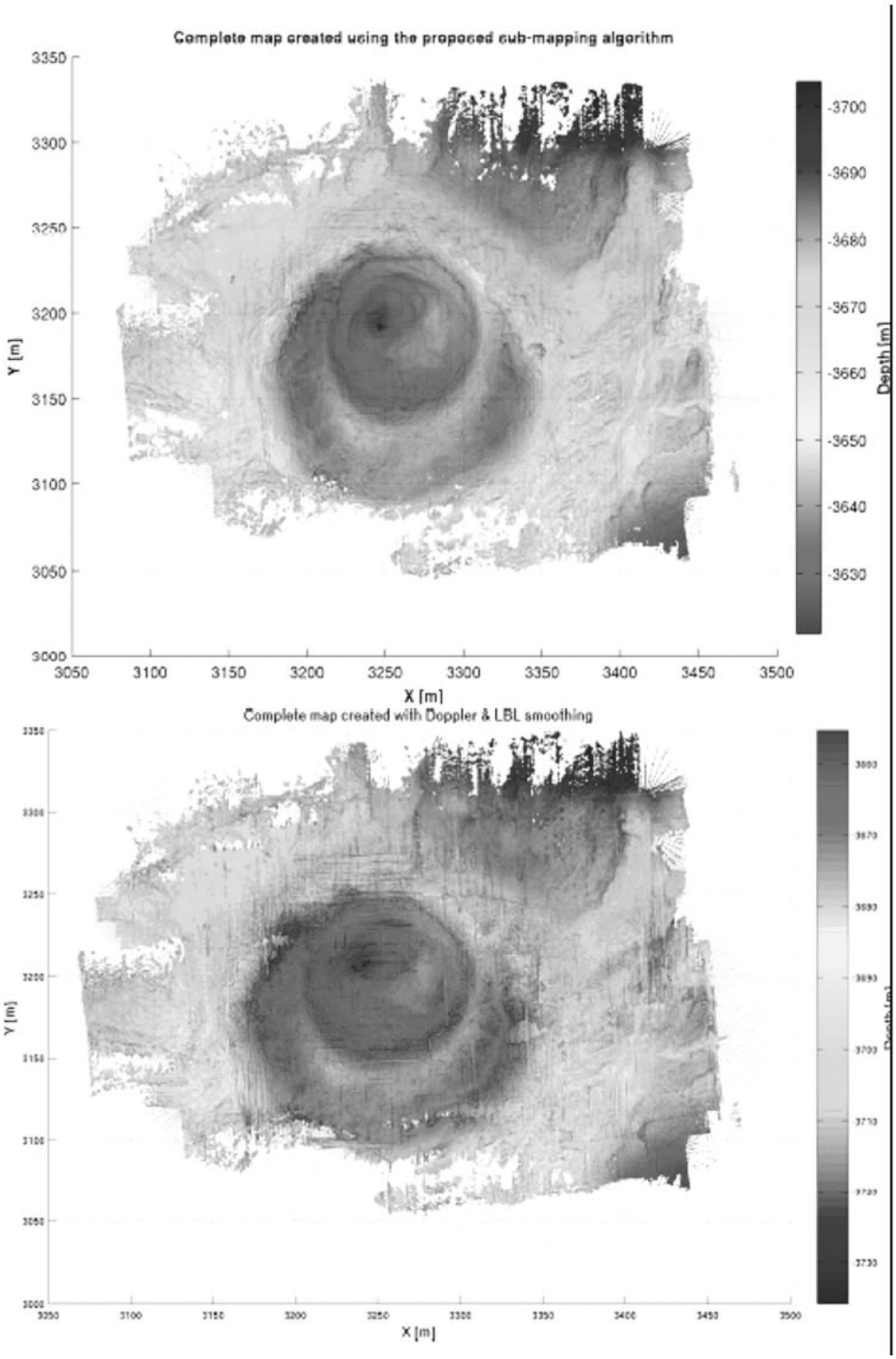

Fig. 14. A comparison between the terrain map created using sub-mapping (a), and a version of the map created using a standard smoothing method (b). The sub-map created map shows significantly less registration error and sonar scan patterning. The sub-mapped version also brings out details that were lost in the map created by smoothing, due to misregistration.

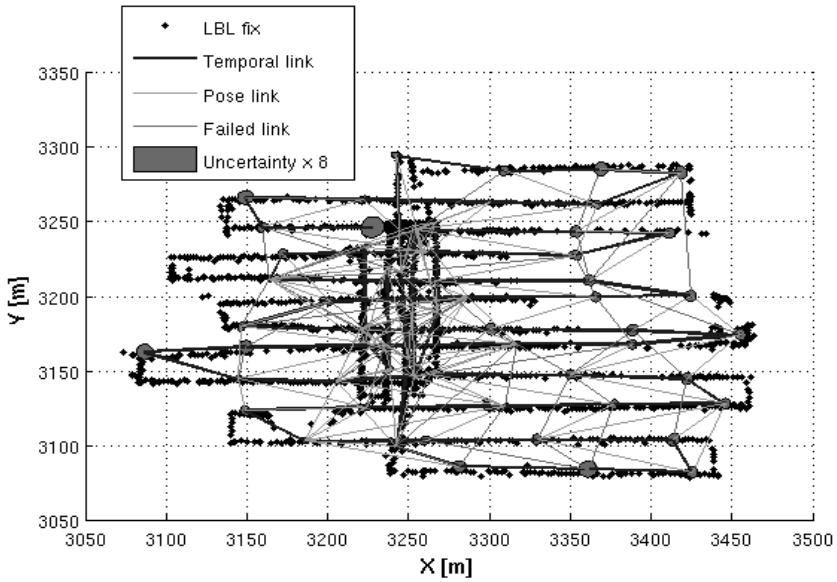

Fig. 15. Sub-mapping pose network. This pose network was established by the sub-mapping algorithm. Nodes indicate the location of the sub-map origins. Blue links indicate consecutive poses in time. Green links indicate where relative pose measurements were made. Magenta links indicate links that were tried but not established. The uncertainty ellipses have been scaled in size by a factor of 8 for visibility. Note that the poses fall into alignment with the LBL fix locations even though this algorithm did not use LBL measurements. This survey consisted of 62 sub maps and 92 established links.

Gaussian graphical model (Pearl 1988; Weiss and Freeman 2001). Sparsity in this model (i.e., missing edges) implies available conditional independencies in the joint-distribution, which can be exploited to realize efficient inference. Others have shown that the feature-based SLAM information matrix obeys a "close-to-sparse" structure when properly normalized (Thrun et al. 2004; Frese 2005). In our formulation of view-based SLAM (Eustice et al. 2005a) however, the information matrix is exactly sparse without having to make any sparse approximations. This implies that for a bounded graph structure, as is the case with typical underwater surveys, view-based SLAM systems comprise a sparse information parameterization without incurring any sparse approximation error.

Based upon this insight, we have implemented a view-based SLAM system for underwater applications built around fusing 6-DOF relative-pose camera measurements from monocular overlapping seafloor imagery with traditional underwater vehicle dead-reckon navigation sensors. Our state vector consists of samples from the robot's trajectory acquired at image acquisition and is maintained using a sparse extended information filter. We use our two view image registration engine to provide non-Markov edge constraints in the corresponding pose network. These "spatial" edges constrain the pose graph and enforce global consistency from local constraints. This system was tested with data collected using the *Hercules* ROV operating at a depth of 3750 m at the wreck of the *RMS Titanic*. The survey covered an area of about 3100 square meters on the seafloor with a accumulated survey path length over 3.4 kilometers. Although the exact formulation is detailed elsewhere (Eustice et al. 2005b) sample results are shown in Figure 15.

5. Conclusions

In this paper we have highlighted some of the fundamental issues associated with the lack of precise and accurate navigation and how they affect our ability to conduct high-resolution mapping efforts in the deep sea. We have presented three different applications of systems-level approaches for deep water mapping that exploit vehicle attitude and navigation information and enforce local and global consistency within sensor measurements to yield superior mapping results commensurate with sensor accuracy. While improving mapping fidelity, these methods also provide us with independent mechanisms for ground truthing, refining and bounding the coarse navigation estimates that are typical in the deep ocean.

These algorithms are applicable across the entire suite of imaging and robotic underwater vehicles—manned, towed, tethered and autonomous. Our work in these areas is continuing with emphasis on implementing a number of these algorithms in real time on board the vehicles to better help us

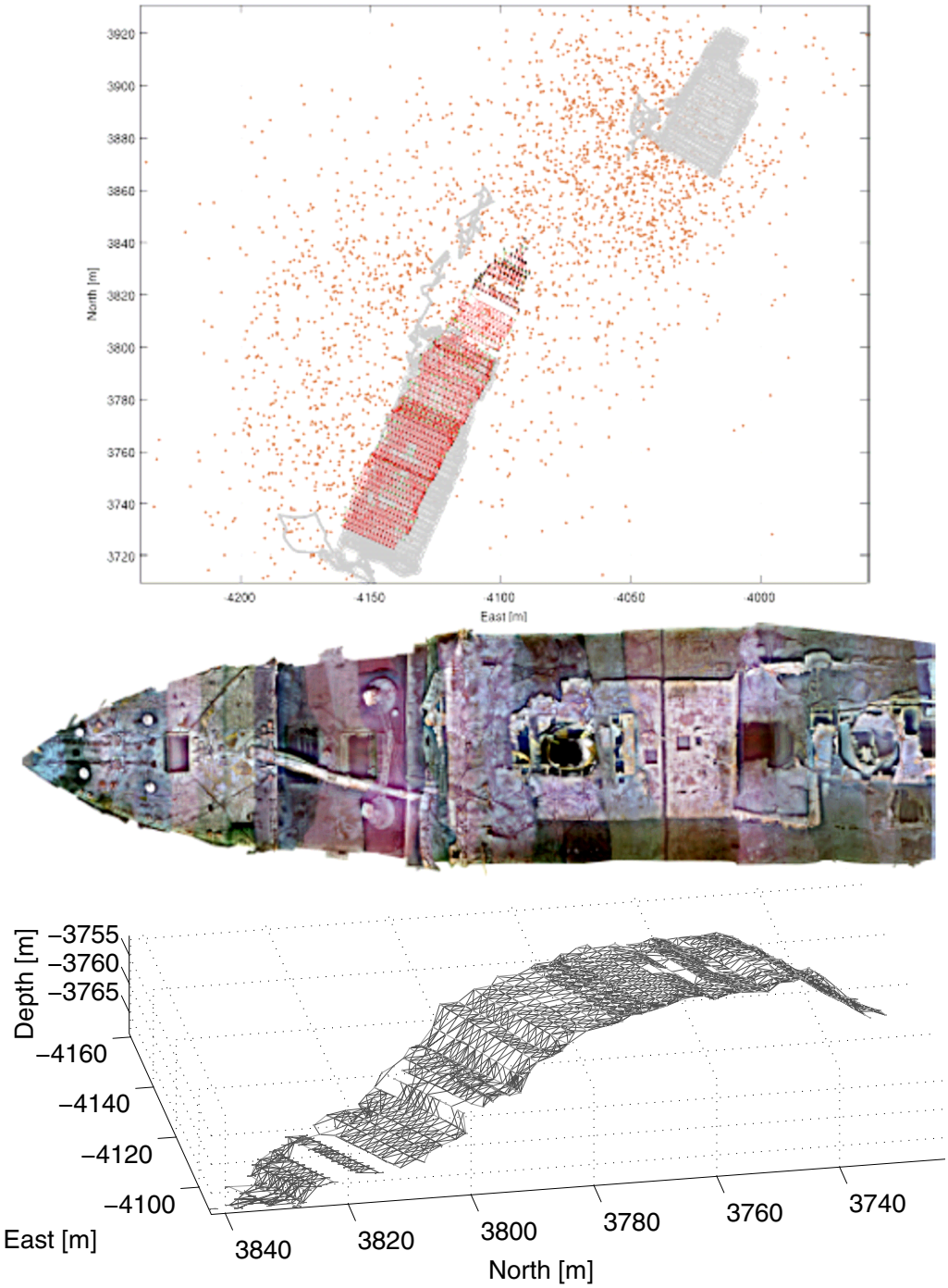

Fig. 16. Mapping the RMS Titanic. (a) An XY plot comparing the raw dead-reckon navigation data (gray), ship-board ultra-short baseline tracking (brown), and reconstructed survey trajectory from a vision-based 6 DOF SLAM information filter (red). (b) A photomosaic of the RMS Titanic constructed from the digital still images and (c) the complete 6-DOF visually based navigation results for the entire survey.

exploit our precision mapping algorithms for real-time adaptive surveys.

Acknowledgements

The authors gratefully acknowledge a multitude of collaborators on the following research expeditions from which data for this paper were gathered—Dr Rob Reves Sohn (Bermuda Coral and TAG), Professor Louis Whitcomb (JHU Hydrodynamics Facility), and Professor Robert Ballard (Titanic). This work was funded in part by a grant from the National Science Foundation's Censsis ERC under grant # EEC-9986821 and in part by NSF ITR grant number #ATM-0428122.

References

Ballard, R. D., Stager, L. E., Master, D., Yoerger, D. R., Mindell, D. A., Whitcomb, L. L., (2002). Iron age shipwrecks in deep water off Ashkelon, Israel. *American Journal of Archaeology* **106**(2): 151–168.

Chen, L. S. (1990). Solving the attenuation problem in emission computed tomography using homomorphic filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**(4): 714–717.

Eustice, R., Singh, H., and Leonard, J. (2005a). Exactly sparse delayed-state filters. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, Barcelona, Spain, 2005.

Eustice, R., Singh, H., Walter, M., Leonard, J., and Ballard, R. D. (2005b). Visually navigating the RMS Titanic with SLAM information filters. *Proceedings of 2005 Robotics Systems and Science Conference*, Cambridge, MA. June.

Fitzgibbon, A. W. and Zisserman, A. (1998). Automatic camera recovery for closed or open image sequences. *Proceedings of the 5th European Conference on Computer Vision*, Freiburg, Germany, June, pp. 311–326.

Frese, U. (2005). A proof for the approximate sparsity of slam information matrices. *Proceedings of IEEE International Conference on Robotics and Automation*, Barcelona, Spain, pp. 331–337.

Hartley, R. and Zisserman, A. (2000). *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Highnam, R. and Brady, M. (1997). Model-based image enhancement of far infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **19**(4): 410–415.

Horn, B. (1987). Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, **4**(4): 629–642.

Howland, J. (1999). Digital data logging and processing, Derbyshire Survey, 1997. Technical report, Woods Hole Oceanographic Institution, December.

Jaffe, J. S. (1990). Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering*, **15**(2): 101–111.

Marks, R., Rock, S., and Lee, M. (1995). Real-time video mosaicking of the ocean floor. *IEEE Journal of Ocean Engineering*, **20**(3): 229–241.

Mertens, L. E. (1970). *In-water Photography*. Photographic Science and Technology and the Graphic Arts Series, edited by W. Clark. John Wiley and Sons.

Milne, P. H. (1983). *Underwater Acoustic Positioning Systems*. Gulf Publishing Company, Houston, TX.

Mobley, C. D. (1994). *Light and Water, Radiative Transfer in Natural Waters*. Academic Press, San Diego, CA.

National Transportation Safety Board (2002). *Aircraft Accident Brief: EgyptAir Flight 990, Boeing 767-366ER, SU-GAP, 60 Miles South of Nantucket, Massachuesetts, October 31, 1999*, 2002. Aircraft Accident Brief NTSB/AAB-02/01, Washington, DC.

Pizarro, O. and Singh, H. (2003). Towards large area mosaicing for underwater scientific applications. *IEEE Journal of Oceanic Engineering, Special Issue on Underwater Image and Video Processing* **28**(4): 651–672.

Pizarro, O., Eustice, R., and Singh, H. (2003). Relative pose estimation for instrumented, calibrated imaging platforms. *Proceedings of Digital Image Computing Techniques and Applications*, Sydney, Australia, December, pp. 601–612.

Pollefeys, M., Koch, R., Vergauwen, M., and Van Gool, L. (1999). Hand-held acquisition of 3D models with a video camera. *Second InternationalConference on 3-D Digital Imaging and Modeling*, Los Alamitos, CA. IEEE Computer Society Press, pp. 14–23.

Roman, C. (2005). *Self-consistent Bathymetric Mapping from Robotic Vehicles in the Deep Ocean*. PhD. thesis, MIT/WHOI Joint Program in Oceanographic Engineering, May.

Rusinkiewic, S. and Levoy, M. (2001). Efficient variants of the ICP algorithm. *Proceedings IEEE 3DIM*, Canada, pp. 145–152.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman, San Mateo, CA.

Rzhanov, Y., Linnett, L. M., and Forbes, R. (2000). Underwater video mosaicing for seabed mapping. *Proceedings of the 2000 International Conference on Image Processing (ICIP 2000)*, Vancouver, BC, Canada, September, Vol. 1, pp. 224–227.

Singh, H., Eustice, R., Roman, C., Pizarro, O., Armstrong, R., Gilbes, F., et al. (2004a). Imaging coral I: Imaging coral habitats with the SeaBED AUV. *Subsurface Sensing Technologies and Applications*, **5**(1): 25–42.

Singh, H., Howland, J., and Pizarro, O. (2004b). Large area photomosaicking underwater. *IEEE Journal of Oceanic Engineering*, **29**(3): 872–886.

Singh, H., Whitcomb, L., Yoerger, D., and Pizarro, O. (2000). Microbathymetric mapping from underwater vehicles in

the deep ocean. *Computer Vision and Image Understanding*, **79**(1): 143–161.

Slama, C. C. (ed.) (1980). *Manual of Photogrammetry*, 4th edn. American Society of Photogrammetry, Bethesda, MD.

Smith, R., Self, M., and Cheeseman, P. (1990). Estimating uncertain spatial relationships in robotics. In *Autonomous Robot Vehicles*, Springer-Verlag, pp. 167–193.

Besl, P. J. and McKay, N. D. (1992). A method for registration of 3D shapes, *IEEE PAMI*, **14**: 239–256.

Thrun, S., Liu, Y., Koller, Z., Ghahramani, H., Durrant-Whyte, H., et al. (2004) Simultaneous mapping and localization with sparse extended information filters. *International Journal of Robotics Research*, **23**(7–8): 693–716.

Triggs, B., McLauchlan, P., Hartley, R., and Fitzgibbon, A. (2000). Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice*, Lecture Notes in Computer Science series (eds. W. Triggs, A. Zisserman and R. Szeliski), pp. 298–375, Springer-Verlag.

Weiss, Y. and Freeman, W. T. (2001). Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, **13**(10): 2173–2200.

Whitcomb, L., Yoerger, D., and Singh, H. (1999). Advances in Doppler-based navigation of underwater robotic vehicles. *Proceedings of the 1999 International Conference on Robotics and Automation*, Vol. 1, pp. 399–406.

Yoerger, D. R., Bradley, A. M., Cormier, M.-H., Ryan, W. B. F., and Walden, B. B. (2000). Fine-scale seafloor survey in rugged deep-ocean terrain with an autonomous robot. *IEEE International Conference on Robotics and Automation*, San Francisco, Vol. 2, pp. 1767–1774.

Yoon, J. H. and Ro, Y. M. (2002). Enhancement of the contrast in mammographic images using the homomorphic filter method. *IEICE Transactions on Information and Systems*, **E85-D**(1): 293–303.