Visually Augmented Navigation for Autonomous Underwater Vehicles

Ryan M. Eustice, *Member, IEEE,* Oscar Pizarro, *Member, IEEE,* and Hanumant Singh, *Member, IEEE*

***Abstract*— As autonomous underwater vehicles (AUVs) are becoming routinely used in an exploratory context for ocean science, the goal of visually augmented navigation (VAN) is to improve the near-seafloor navigation precision of such vehicles without imposing the burden of having to deploy additional infrastructure. This is in contrast to traditional acoustic long baseline navigation techniques, which require the deployment, calibration, and eventual recovery of a transponder network. To achieve this goal, VAN is formulated within a vision-based simultaneous localization and mapping (SLAM) framework that exploits the systems-level complementary aspects of a camera and strap-down sensor suite. The result is an environmentally-based navigation technique robust to the peculiarities of low-overlap underwater imagery. The method employs a view-based representation where camera-derived relative-pose measurements provide spatial constraints, which enforce trajectory consistency and also serve as a mechanism for loop-closure, allowing for error growth to be independent of time for revisited imagery. This article outlines the multi-sensor VAN framework and demonstrates it to have compelling advantages over a purely vision-only approach by: *(i)* improving the robustness of low-overlap underwater image registration, *(ii)* setting the free gauge scale, and *(iii)* allowing for a disconnected camera-constraint topology.**

***Index Terms*— Computer vision, navigation, mobile robotics, underwater vehicles, SLAM, and robotic perception.**

I. Introduction

FROM exploring abandoned mines in Pennsylvania [1], to exploring other planets in our solar-system [2], robots extend our reach to areas where human investigation is considered too dangerous, too technically challenging, or both. While high profile missions like the 2004 Mars rovers epitomize the lengths that we will go to in search of new origins of life, it cannot be overstated that exploring the deep-abyss of our own oceans can be nearly as alien and offer just as startling discoveries about early life. Though manned vehicles like Alvin [3], [4] have been responsible for many of the most important deep-science discoveries [5], [6], the extreme design requirements, operational costs, risk of life, and limited availability preclude its ubiquitous use. Therefore, out of necessity the deep-sea has become an arena where the presence of mobile robotics is pervasive and their scientific utility revolutionary [7]–[9].

While underwater mobile robotics have made significant inroads into mainstream science over the past two decades, a limiting technological issue to their widespread utility, especially for exploration, is the lack of *easily* obtainable precision navigation [10]. With the advent of the global positioning system (GPS) [11] many surface and air vehicle applications are able to easily obtain their position anywhere on the globe with precision of a few meters via the triangulation of satellite transmitted radio signals. Unfortunately, these radio signals do not penetrate sub-sea [12], [13] (nor underground [1], nor even indoors [14]). Hence, traditional underwater navigation strategies use acoustic ranging systems whereby seafloor-tethered beacons relay time-of-flight range measurements for triangulated positioning [13], [15]. The cost, complexity, and limitations of this infrastructure dependent solution, however, leave much to be desired, which is further complicated by the fact that alternative strap-down methods suffer from a position drift that grows unbounded with time [13].

Over the past decade, a big research push within the terrestrial mobile robotics community has been to develop environmentally-based navigation algorithms, which eliminate the need for additional infrastructure and bound position error growth to the size of the environment — a key prerequisite for truly autonomous navigation. The basis of this work has been to exploit the perceptual sensing capabilities of robots to "beat-down" accumulated odometric error by localizing the robot with respect to landmarks in the environment. The question of how to use such a methodology for navigation and mapping was first theoretically addressed in a probabilistic framework in the mid 1980's with seminal papers by Smith, Self, and Cheeseman [16] and Moutarlier and Chatila [17], which have since then become the cornerstone of the research field known as simultaneous localization and mapping (SLAM).

One of the major challenges of a SLAM methodology is that defining what constitutes a feature from raw sensor data can be nontrivial. In man-made environments, typically composed of planes, lines and corners primitives, features can be more eas-

Manuscript received December 22, 2005; revised January 18, 2006. 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, in part by the Woods Hole Oceanographic Institution through a grant from the Penzance Foundation, and in part by a National Defense Science and Engineering Graduate (NDSEG) Fellowship awarded through the Department of Defense. This paper was presented in part at the IEEE International Conference on Robotics and Automation, New Orleans, USA, April 2004.

R. Eustice was with the Joint Program in Oceanographic Engineering of the Massachusetts Institute of Technology, Cambridge, MA, and the Woods Hole Oceanographic Institution, Woods Hole, MA during the tenure of this work; presently he is with the Department of Naval Architecture and Marine Engineering at the University of Michigan, Ann Arbor, MI 48109 USA (email: eustice@umich.edu).

O. Pizarro was also with the Joint Program in Oceanographic Engineering of the Massachusetts Institute of Technology and the Woods Hole Oceanographic Institution. He is presently with the Australian Centre for Field Robotics at The University of Sydney, Sydney, Australia (email: o.pizarro@acfr.usyd.edu.au).

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

ily defined [18]. However, unstructured outdoor environments can pose a more challenging task for feature extraction and matching, which has lead to scan-matching based approaches that do not require an explicit representation of features [19], [20]. These view-based, data-driven techniques have traditionally been used with accurate perceptual sensors such as laser range finders where raw data can be matched directly (e.g., in an iterative closest point sense [21]). Along these lines, our underwater approach is to use a camera as an accurate and inexpensive perceptual sensor to collect near-seafloor imagery that can also be matched directly. Motivation for such an approach comes from the fact that, in practice, autonomous underwater vehicles (AUVs) typically collect imagery using a digital-still camera and not video (to minimize the amount of power consumption spent on illumination [22]). This results in a temporally low-overlap image sequence with the implication that 3D features in the environment are not observed for more than a couple of frames. Such a low-overlap constraint implies that a view-based representation is particularly suitable for this type of data, since overlapping image pairs from a calibrated camera can be registered in a pairwise fashion to extract "zero-drift", relative-pose modulo scale measurements, without explicitly representing 3D feature points. In this way, registering an image taken from time t_i to an image taken at time t_j provides a spatial constraint whose error is bounded regardless of time or the trajectory followed between the two views.

In the rest of this article we present our framework and methodology for incorporating camera-derived relative-pose measurements with vehicle navigation data in a view-based SLAM context (Fig. 1). In particular, §II and §III describe our assumptions and coordinate frame conventions, respectively. §IV presents a delayed-state SLAM framework for fusing camera measurements that also serves as a foundation for probabilistic link hypothesis. In §V we explain how to actually make the pairwise camera measurement using a systems-level, feature-based, image registration approach. We show that a multi-sensor approach has compelling advantages over a camera-only navigation system and, in particular, that it improves registration robustness via a novel pose-constrained correspondence search. Results are presented in the context of a real-world data set collected by an AUV in a rugged undersea environment, and for tank data collected by a remotely operated vehicle (ROV) for which position ground-truth is available.

II. Assumptions

Our application is based upon using a pose instrumented AUV equipped with a single downward-looking digital-still calibrated camera to perform underwater imaging and mapping. We assume that the vehicle can make acoustic measurements of both velocity and altitude relative to the seafloor, that absolute orientation is measured to within a few degrees over the entire survey area via inclinometers and a magnetic compass, and that bounded positional estimates of depth, Z, are provided by a pressure sensor. A detailed discussion of our particular AUV platform can be found in [23], [24]. For convenience, Table I provides a short summary of assumed sensor characteristics. In brief we assume:

- An ideal (i.e., lens distortion compensated) calibrated camera.
- A pose-instrumented platform.
- Known reference frames (e.g., extrinsic camera to vehicle coordinate transform).
- Pairwise image registration using a six degree of freedom (DOF) motion model to accommodate low-temporal overlap.

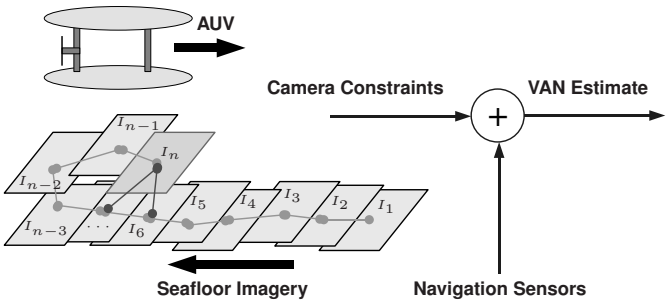


Fig. 1. The objective of visually augmented navigation (VAN) is the real-time fusion of "zero-drift" camera measurements with navigation sensor data to close-the-loop on dead-reckoned error. For this purpose VAN adopts a top-down systems-level approach to visual navigation. At its core, VAN is founded upon registering raw imagery to generate pairwise camera constraints that are then fused with navigation sensor data in a view-based SLAM framework.

TABLE I
Typical pose sensor characteristics for underwater platforms.

Measurement	Sensor	Precision
Roll/Pitch	Tilt Sensor	$\pm 0.5°$
Heading	Magnetic Compass	$\pm 2.0°$
3-Axis Angular Rate	AHRS	$\pm 1.0°/s$
Body Frame Velocities	Acoustic Doppler	$\pm$1–2 mm/s
Depth	Pressure Sensor	$\pm 0.01\%$
Altitude	Acoustic Altimeter	± 0.1 m

III. 6-DOF Coordinate Frame Relationships

This section describes the reference frames used in vehicle navigation and their 6-DOF coordinate frame relationships as illustrated in Fig. 2. We follow standard SNAME[1] convention [25] and define the vehicle frame, denoted subscript v, to be coincident with a fixed point on the vehicle and oriented such that the positive X_v-axis is aligned with the bow, positive Y_v-axis to starboard, and Z_v-axis down, thus completing a right handed coordinate frame.

Additionally, we must consider each onboard sensors' internal coordinate frame (in which measurements are expressed) and its subsequent relationship to the vehicle. The sensor frame, denoted subscript s, is assumed to be static and known with respect to the vehicle frame (i.e., calibrated beforehand); we denote this sensor to vehicle coordinate frame relationship notationally as $\mathbf{x}_{vs}$. Also, two navigation frames are defined and used for expressing vehicle pose. The first is the world

[1]The Society of Naval Architects and Marine Engineers.

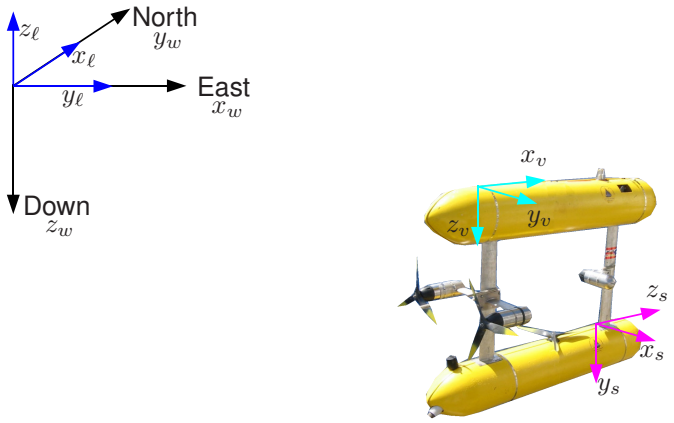


Fig. 2. Illustration of the different reference frames used within VAN. Frames w and ℓ represent the world and local-level frames, respectively. Frame v represents the vehicle reference frame while frame s represents an arbitrary sensor frame. The sensor and vehicle frames are attached to the same rigid body and therefore are static with respect to each other.

frame, denoted subscript w, which is a static reference frame located at the water surface oriented with X_w-East, Y_w-North, and Z_w-Up. It is useful for displaying results since it follows standard map convention; vehicle position with respect to this frame is denote $\mathbf{x}_{wv}$. The second navigation frame that we define is the local-level frame, denoted subscript ℓ. This frame is coincident with the world frame, however, it is oriented with X_ℓ-North, Y_ℓ-East, Z_ℓ-Down and corresponds to a zero-orientation (i.e., local-level) version of the vehicle frame. This frame is useful for navigation because standard compass-measurements are consistent with the right-hand rule convention about the z-axis. Vehicle position in this frame is denoted $\mathbf{x}_{\ell v}$.

Throughout this article, we adopt the Smith, Self, and Cheeseman coordinate frame convention [16]. Standard coordinate transformation operations are the compounding and inverse coordinate frame relationships, which are denoted as $\mathbf{x}_{ik} = \mathbf{x}_{ij} \oplus \mathbf{x}_{jk}$, and $\mathbf{x}_{ji} = \ominus \mathbf{x}_{ij}$, respectively.

IV. View-Based SLAM Estimation Framework

Typical structure-from-motion (SFM) approaches estimate both camera motion and 3D scene structure from a sequence of video frames. In our application, however, the low degree of temporal image overlap (typically on the order of 35% or less with digital-still imagery) motivates us to focus on recovering pairwise measurements from spatially neighboring image frames. In this approach, the camera provides observation of the 6-DOF relative coordinate transformation between poses modulo scale (via calculation of the Essential matrix). These measurements are used as constraints in a recursive estimation framework that determines the global poses consistent with the camera measurements and navigation prior. The global poses correspond to samples from the robot's trajectory at the times associated with image acquisition. Thus, unlike the typical feature-based SLAM estimation problem, which keeps track of the current robot pose and an associated landmark map [16], the VAN state vector consists entirely of historical trajectory

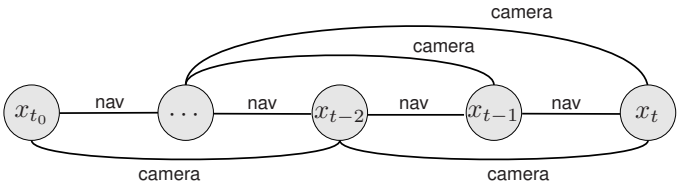


Fig. 3. A view-based representation consists of a network of navigation and camera constraints over a collection of delayed-state vehicle poses.

samples sampled at image acquisition. In our nomenclature these samples are refereed to as delayed-states.

The delayed-state approach corresponds to a view-based representation of the environment where dead-reckoned sensor navigation provides temporal (Markov) observations while overlapping imagery provides both temporal and non-temporal (i.e., spatial image overlap) pose constraints (Fig. 3). This view-based approach can be traced through the literature to a batch scan-matching method by Lu and Milios [19] using laser range data, a delayed decision making framework by Leonard and Rikoski [26] for feature initialization with sonar data, and the hybrid batch/recursive formulations by Fleischer [27] and McLauchlan [28] using camera images. In this context, pairwise registered imagery results in observation of relative robot motion with respect to a place it has previously visited.

A. Delayed-State Filtering

We begin by describing our representation of vehicle state and a general system model for state evolution and observation. We show how this representation can be used to incorporate camera-derived relative-pose measurements by augmenting our state representation to include historical trajectory samples (i.e., delayed-states). For the sake of conceptual clarity, we outline the procedure of delayed-state filtering within the context of an extended Kalman filter (EKF) [29], which is a well-known inference approach to SLAM [16].[2] While this work follows that of Garcia [34] and Fleischer [27], it substantially differs by extending the motion and camera models to deal with 6-DOF movement in a fully 3D environment.

1) Fixed-Size State Description: The vehicle state vector, $\mathbf{x}_v$, contains both pose, $\mathbf{x}_p$, and kinematic terms, $\mathbf{x}_\kappa$, and is defined as

$$\mathbf{x}_v \equiv \left[\mathbf{x}_p^\top, \quad \mathbf{x}_\kappa^\top\right]^\top.$$

Here, $\mathbf{x}_p$ is a 6-vector of vehicle pose in the local-level navigation frame where XYZ roll, pitch, and heading Euler angles are used to represent orientation [25] (i.e., $\mathbf{x}_p \equiv \mathbf{x}_{\ell v} \equiv \left[x, y, z, \phi, \theta, \psi\right]^\top$), and $\mathbf{x}_\kappa$ represents any kinematic state elements that are required for propagation of

[2]In practice, the EKF SLAM framework does not scale well to large-environments due to the quadratic complexity in maintaining the joint-correlations. In separate publications [30]–[33], we have developed a novel scalable framework based upon exploiting natural exact sparsity in the EKF's dual — an extended information filter (EIF). Since the EIF is the dual of the EKF, the methods and results we present in this article equally apply to the EIF framework. For conceptual clarity, however, we instead present the more standard EKF-based formulation so that we can focus on the contributions of our systems-level VAN methodology.

the vehicle process model (e.g., body-frame velocities, accelerations, angular rates). In addition, we assume that the vehicle state can be modeled as being normally distributed, $\mathbf{x}_v \sim \mathcal{N}\big(\boldsymbol{\mu}_v, \Sigma_{vv}\big)$, with mean and covariance given by

$$\boldsymbol{\mu}_v = \begin{bmatrix}\boldsymbol{\mu}_p^\top, & \boldsymbol{\mu}_\kappa^\top\end{bmatrix}^\top \quad \text{and} \quad \Sigma_{vv} = \begin{bmatrix}\Sigma_{pp} & \Sigma_{p\kappa} \\ \Sigma_{\kappa p} & \Sigma_{\kappa\kappa}\end{bmatrix}.$$

The vehicle state evolves through a time-varying continuous time process model, $\mathbf{f}(\,\cdot\,, t)$, driven by additive white noise, $\mathbf{w}(t) \sim \mathcal{N}\big(\mathbf{0}, \mathrm{Q}(t)\big)$, and deterministic control inputs, $\mathbf{u}(t)$, while discrete time measurements, $\mathbf{z}[t_k]$, of elements in the vehicle state are observed through an observation model, $\mathbf{h}(\,\cdot\,, t_k)$, corrupted by additive time independent Gaussian noise, $\mathbf{v}[t_k] \sim \mathcal{N}\big(\mathbf{0}, \mathrm{R}_k\big)$, with $E\big[\mathbf{w}\mathbf{v}^\top\big] = 0$. The resulting system model is:

$$\begin{aligned}\dot{\mathbf{x}}_v(t) &= \mathbf{f}(\mathbf{x}_v(t), \mathbf{u}(t), t) + \mathbf{w}(t) \\ \mathbf{z}[t_k] &= \mathbf{h}(\mathbf{x}_v[t_k], t_k) + \mathbf{v}[t_k].\end{aligned} \tag{1}$$

As is typical in the navigation literature, the vehicle state distribution is approximately maintained using a continuous-discrete EKF [29]:

$$\boxed{\textbf{Prediction}} \quad \begin{aligned}\dot{\boldsymbol{\mu}}_v(t) &= \mathbf{f}(\boldsymbol{\mu}_v(t), \mathbf{u}(t), t) \\ \dot{\Sigma}_{vv}(t) &= \mathrm{F_x}\Sigma_{vv}(t) + \Sigma_{vv}(t)\mathrm{F_x}^\top + \mathrm{Q}(t)\end{aligned} \tag{2}$$

$$\boxed{\textbf{Update}} \quad \begin{aligned}\mathrm{K} &= \bar{\Sigma}_{vv}\mathrm{H_x}^\top\big(\mathrm{H_x}\bar{\Sigma}_{vv}\mathrm{H_x}^\top + \mathrm{R}_k\big)^{-1} \\ \boldsymbol{\mu}_v &= \bar{\boldsymbol{\mu}}_v + \mathrm{K}\big(\mathbf{z}[t_k] - \mathbf{h}(\bar{\boldsymbol{\mu}}_v, t_k)\big) \\ \Sigma_{vv} &= \big(\mathrm{I} - \mathrm{K}\mathrm{H_x}\big)\bar{\Sigma}_{vv}\big(\mathrm{I} - \mathrm{K}\mathrm{H_x}\big)^\top + \mathrm{K}\mathrm{R}_k\mathrm{K}^\top\end{aligned} \tag{3}$$

where $\mathrm{F_x} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}_v}\big|_{\boldsymbol{\mu}_v}$ and $\mathrm{H_x} = \frac{\partial \mathbf{h}}{\partial \mathbf{x}_v}\big|_{\bar{\boldsymbol{\mu}}_v}$ are the process and observation model Jacobians, respectively. In this formulation, the predicted vehicle distribution, $\bar{\mathbf{x}}_v \sim \mathcal{N}\big(\bar{\boldsymbol{\mu}}_v, \bar{\Sigma}_{vv}\big)$, is computed between asynchronous sensor measurements by solving (2) via a fourth-order Runge-Kutta numerical integration approach [35].

Unfortunately, the fixed-size state description, $\mathbf{x}_v$, does not allow us to represent our pairwise camera constraints. This is because registration of an image pair results in a relative-pose measurement modulo scale, and not an absolute observation of elements in vehicle pose, $\mathbf{x}_p$. Therefore, before we can incorporate pairwise camera constraints, we have to first augment our state representation to include a history of vehicle poses where each delayed-state entry corresponds to an image in our view-based map. Under this representation, the distribution we are trying to estimate is $p\big(\boldsymbol{\xi}_t \big| \mathbf{z}^t, \mathbf{u}^t\big)$ where $\mathbf{z}^t$ represents all measurements up to time t (including camera and navigation sensors), $\mathbf{u}^t$ is the set of all control inputs, and $\boldsymbol{\xi}_t$ is our view-based SLAM state vector. Next, we describe the process of *how* delayed-states are added to our view-based map.

2) Augmenting our State Description with Delayed-States: At time t_1, corresponding to when the first image frame, I_1, is acquired, we augment our state description, $\boldsymbol{\xi}_t$, to include the vehicle's pose of where it was when it acquired that image (i.e., $\boldsymbol{\xi}_t = [\mathbf{x}_v^\top,\ \mathbf{x}_{p_1}^\top]^\top$). Therefore, at this time instance the augmented state distribution, $\boldsymbol{\xi}_t \sim \mathcal{N}\big(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\big)$, is given by

$$\begin{aligned}\boldsymbol{\mu}_t &= \big[\boldsymbol{\mu}_v[t_1]^\top,\ \boldsymbol{\mu}_p[t_1]^\top\big]^\top &&\equiv \big[\boldsymbol{\mu}_v^\top,\ \boldsymbol{\mu}_{p_1}^\top\big]^\top \\ \Sigma_t &= \begin{bmatrix}\Sigma_{vv}[t_1] & \Sigma_{vp}[t_1] \\ \Sigma_{vp}^\top[t_1] & \Sigma_{pp}[t_1]\end{bmatrix} &&\equiv \begin{bmatrix}\Sigma_{vv} & \Sigma_{vp_1} \\ \Sigma_{p_1 v} & \Sigma_{p_1 p_1}\end{bmatrix}.\end{aligned} \tag{4}$$

This process is repeated for each camera frame that we wish to include in our view-based map so that after augmenting n delayed states (one for each retained camera frame) we have $\boldsymbol{\xi}_t = \big[\mathbf{x}_v^\top,\ \mathbf{x}_{p_1}^\top,\ \cdots,\ \mathbf{x}_{p_n}^\top\big]^\top$ with

$$\boldsymbol{\mu}_t = \begin{bmatrix}\boldsymbol{\mu}_v \\ \boldsymbol{\mu}_{p_1} \\ \vdots \\ \boldsymbol{\mu}_{p_n}\end{bmatrix} \quad \text{and} \quad \Sigma_t = \begin{bmatrix}\Sigma_{vv} & \Sigma_{vp_1} & \cdots & \Sigma_{vp_n} \\ \Sigma_{p_1 v} & \Sigma_{p_1 p_1} & \cdots & \Sigma_{p_1 p_n} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{p_n v} & \Sigma_{p_n p_1} & \cdots & \Sigma_{p_n p_n}\end{bmatrix}. \tag{5}$$

Note that in (4) the vehicle's current pose, $\mathbf{x}_p$, is fully correlated with $\mathbf{x}_{p_1}$ by definition. Therefore, when the n^{th} delayed-state, $\mathbf{x}_{p_n}$, is augmented in (5), its cross-correlation with the other delayed-states in Σ_t is non-zero since the current vehicle state has correlation with each delayed-state.

The system model (1) must be also be extended to incorporate the augmented state representation. For the process model the only required change is that $\mathbf{x}_v$ continue to evolve through the vehicle dynamic model, $\mathbf{f}(\,\cdot\,, t)$, while the delayed-state entries do not:

$$\dot{\boldsymbol{\xi}}_t = \frac{d}{dt}\begin{bmatrix}\mathbf{x}_v \\ \mathbf{x}_{p_1} \\ \vdots \\ \mathbf{x}_{p_n}\end{bmatrix} = \begin{bmatrix}\mathbf{f}(\mathbf{x}_v(t), \mathbf{u}(t), t) + \mathbf{w}(t) \\ 0_{6\times 1} \\ \vdots \\ 0_{6\times 1}\end{bmatrix}.$$

Similarly, navigation sensor observation models continue to remain a function of only the current vehicle state, $\mathbf{x}_v$, which results in sparse Jacobians of the form

$$\mathrm{H}_\xi = \begin{bmatrix}\mathrm{H_x}, & 0_{m\times 6}, & \cdots, & 0_{m\times 6}\end{bmatrix},$$

where m is the dimension of the measurement. In the case of camera-derived measurements, however, the observation model is a function of delayed-states entries as we discuss next.

B. Pairwise Camera Observation Model

Pairwise image registration from a calibrated camera has the ability to provide a measurement of relative-pose modulo scale between delayed-state elements $\mathbf{x}_{p_i}$ and $\mathbf{x}_{p_j}$, provided images I_i and I_j have common overlap. In deriving the camera observation model we use the familiar Smith, Self, and Cheeseman coordinate transformation operations (i.e., head-to-tail, tail-to-tail, and inverse) [16], and assume that the extrinsic camera to vehicle pose, $\mathbf{x}_{vc}$, is known.

1) Camera Relative Pose: The delayed-state entries $\mathbf{x}_{p_i}$ and $\mathbf{x}_{p_j}$ correspond to vehicle poses $\mathbf{x}_{\ell v_i}$ and $\mathbf{x}_{\ell v_j}$, respectively, as represented in the local-level navigation frame defined in §III. Hence, using the extrinsic camera to vehicle pose, $\mathbf{x}_{vc}$, we can express the transformation from camera frame i to j using the tail-to-tail operation as

$$\mathbf{x}_{c_j c_i} = \ominus\mathbf{x}_{\ell c_j} \oplus \mathbf{x}_{\ell c_i} \tag{6a}$$

$$= \ominus(\mathbf{x}_{\ell v_j} \oplus \mathbf{x}_{vc}) \oplus (\mathbf{x}_{\ell v_i} \oplus \mathbf{x}_{vc}), \tag{6b}$$

with Jacobian

$$\mathrm{J}_{c_jc_i} = \frac{\partial \mathbf{x}_{c_jc_i}}{\partial(\mathbf{x}_{\ell v_j},\, \mathbf{x}_{\ell v_i})} = \underbrace{\frac{\partial \mathbf{x}_{c_jc_i}}{\partial(\mathbf{x}_{\ell c_j},\, \mathbf{x}_{\ell c_i})} \frac{\partial(\mathbf{x}_{\ell c_j},\, \mathbf{x}_{\ell c_i})}{\partial(\mathbf{x}_{\ell v_j},\, \mathbf{x}_{\ell v_i})}}_{\text{chain-rule}} \tag{7a}$$

$$= \ominus \mathrm{J}_{\oplus}\big|_{(\mathbf{x}_{\ell c_j},\, \mathbf{x}_{\ell c_i})} \begin{bmatrix} \mathrm{J}_{1\oplus}\big|_{(\mathbf{x}_{\ell v_j},\, \mathbf{x}_{vc})} & 0_{6\times 6} \\ 0_{6\times 6} & \mathrm{J}_{1\oplus}\big|_{(\mathbf{x}_{\ell v_i},\, \mathbf{x}_{vc})} \end{bmatrix}. \tag{7b}$$

2) 5-DOF Camera Measurement: What the camera actually measures, however, is not the 6-DOF relative-pose measurement (6), but instead only a 5-DOF measurement due to loss of scale in the image formation process. This loss of scale implies that only the baseline direction, as represented by azimuth and elevation angles α_{ji} and β_{ji}, respectively, is recoverable from image space. Realizing that the relative-pose measurement $\mathbf{x}_{c_jc_i}$ is parameterized by

$$\mathbf{x}_{c_jc_i} = [{}^{c_j}\mathbf{t}_{c_jc_i}^{\top}, \boldsymbol{\Theta}_{c_jc_i}^{\top}]^{\top} = [x_{ji}, y_{ji}, z_{ji}, \phi_{ji}, \theta_{ji}, \psi_{ji}]^{\top},$$

we can express the bearing-only baseline measurement of α_{ji} and β_{ji} as

$$\alpha_{ji} = \tan^{-1}(y_{ji}/x_{ji})$$
$$\beta_{ji} = \tan^{-1}\big(z_{ji}/\sqrt{x_{ji}^2 + y_{ji}^2}\big).$$

with Jacobian

$$\mathrm{J}_{\alpha\beta} = \frac{\partial(\alpha_{ji}, \beta_{ji})}{\partial {}^{c_j}\mathbf{t}_{c_jc_i}} = \begin{bmatrix} \frac{-y_{ji}}{x_{ji}^2+y_{ji}^2} & \frac{x_{ji}}{x_{ji}^2+y_{ji}^2} & 0 \\ \frac{-z_{ji}x_{ji}}{t^2\sqrt{x_{ji}^2+y_{ji}^2}} & \frac{-z_{ji}y_{ji}}{t^2\sqrt{x_{ji}^2+y_{ji}^2}} & \frac{x_{ji}^2+y_{ji}^2}{t^2} \end{bmatrix},$$

where

$$t = \big\|{}^{c_j}\mathbf{t}_{c_jc_i}\big\| = \sqrt{(x_{ji}^2 + y_{ji}^2 + z_{ji}^2)}.$$

Hence, the pairwise 5-DOF camera observation model becomes (Fig. 4)

$$\mathbf{z}_{ji} = \mathbf{h}_{ji}(\boldsymbol{\xi}_t) = \mathbf{h}_{ji}(\mathbf{x}_{p_j}, \mathbf{x}_{p_i}) = [\alpha_{ji}, \beta_{ji}, \phi_{ji}, \theta_{ji}, \psi_{ji}]^{\top}, \tag{8}$$

with Jacobian

$$\mathrm{H}_{\xi} = \begin{bmatrix} 0 \cdots & \frac{\partial \mathbf{h}_{ji}}{\partial \mathbf{x}_{p_j}} & \cdots 0 \cdots & \frac{\partial \mathbf{h}_{ji}}{\partial \mathbf{x}_{p_i}} & \cdots 0 \end{bmatrix},$$

where

$$\frac{\partial \mathbf{h}_{ji}}{\partial(\mathbf{x}_{p_j}, \mathbf{x}_{p_i})} = \frac{\partial \mathbf{h}_{ji}}{\partial \mathbf{x}_{c_jc_i}} \frac{\partial \mathbf{x}_{c_jc_i}}{\partial(\mathbf{x}_{p_j}, \mathbf{x}_{p_i})} = \begin{bmatrix} \mathrm{J}_{\alpha\beta} & 0_{2\times 3} \\ 0_{3\times 3} & \mathrm{I}_{3\times 3} \end{bmatrix} \mathrm{J}_{c_jc_i}.$$

3) What do pairwise camera measurements tell us?: Now that we have derived *how* to model pairwise camera measurements, it's worth intuitively describing what a 5-DOF relative-pose observation means in terms of reducing navigation error. First of all, pairwise camera measurements (8) provide us with a bearing-only measurement of the baseline between poses — hence, we are dependent upon our navigation sensors to set the free-gauge scale. In our application this scale is implicitly fixed within the filter by two sources: *(i)* bounded-error measurements of depth variations (Z direction) coming from a pressure sensor, and *(ii)* Doppler velocities that provide an integrated measurement of XYZ position.

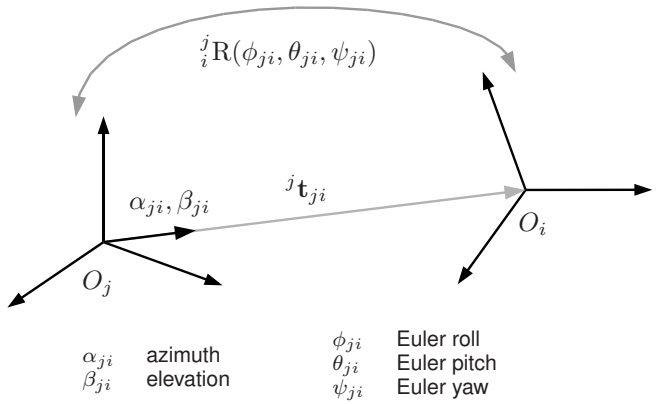


Fig. 4. An illustration of the pairwise 5-DOF camera measurement (i.e., relative-pose modulo scale).

Secondly, (8) tells us that camera measurements can only reduce relative positional error components that are *orthogonal* to the baseline motion. Referring to Fig. 4 we see that frame O_i can slide anywhere along the baseline, ${}^{j}\mathbf{t}_{ji}$, without effecting the measure of azimuth/elevation. This suggests that temporal camera measurements do very little to reduce *along-track* error growth (though, they still refine the direction of motion). Hence, long linear surveys will benefit far less from camera constraints than surveys incorporating cross-over points, where "loops" in the trajectory result in ample spatial constraints.

Finally, the nonlinear bearing-only constraints of (8) imply that linearization errors in the observation model will be less significant if we can maintain good map contact (e.g., typical boustrophedon surveys achieve this) to prevent our linearization point from drifting too far from the truth. This also suggests that when closing large loops, where the linearization point may be far from the true state, that we should incorporate the pairwise camera constraints in aggregate via some form of triangulation — a technique commonly used for feature-initialization in bearing-only SLAM applications [36].

C. Link Hypothesis

An essential task in a view-based representation is hypothesizing *probable* overlapping image pairs. Because image registration is arguably the slowest component in the VAN framework, it is to our advantage to feed the registration module only likely candidate pairs so as to not waste time attempting registration on images that have a low likelihood of overlap. Since our hovering AUV flies in a closed-loop bottom-following mode for camera surveys, it maintains an approximately constant altitude above the seafloor. For simplicity, our link hypothesis strategy is based upon a grossly-simplified 1D model for image overlap (i.e., analogous to a circular field of view assumption) that uses our state estimate and altimeter measured scene altitude to project image footprints onto a horizontal plane as illustrated in Fig. 5. When computing pairwise overlap we assume the larger altitude of the camera pair in our calculations (Fig. 5(b)).

Assuming the above mentioned configuration, image per-

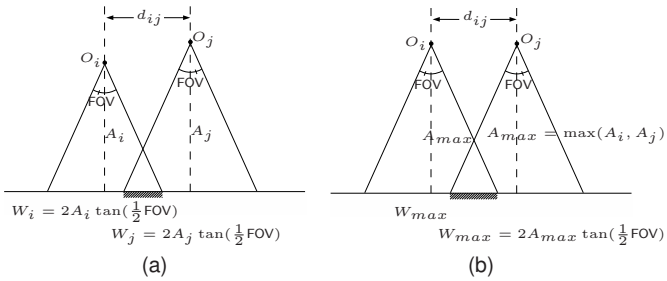


Fig. 5. Calculation of pairwise overlap for link hypothesis. (a) To simplify the calculation of image overlap, we reduce it to a 1D case on a horizontal plane. In the illustrations above, O_i and O_j are the camera centers, FOV is the field of view, A_i and A_j are the altimeter measured altitudes, W_i and W_j are the computed 1D image widths, and d_{ij} is the Euclidean baseline distance derived from our state estimate. (b) Assuming the vehicle's closed-loop control approximately maintains a constant altitude above the seafloor then $A_i \approx A_j$. Therefore, we further simplify the calculation by assuming the larger altitude for both cameras.

cent overlap, ϵ, can be defined as

$$\epsilon = \begin{cases} 1 - \frac{d_{ij}}{W_{max}} & 0 \leq d_{ij} \leq W_{max} \\ 0 & \text{otherwise} \end{cases}.$$

Here, d_{ij} is the Euclidean distance between the camera centers, $W_{max} = 2A_{max}\tan(\frac{1}{2}\text{FOV})$ is the 1D image width, $A_{\max}$ is the larger altitude of the pair, and FOV is the camera field of view. Under this scheme, we can set thresholds for minimum and maximum percent image overlap to obtain constraints on camera distance. We can then compute a first-order probability associated with whether or not the distance between the camera pair falls within these constraints. This calculation serves as the basis of our automatic link hypothesis algorithm, outlined in Algorithm 1, where all frames in our view-based map are checked to see whether or not they could overlap with the current robot view (i.e., linear complexity in the number of views). The k most likely candidates ($k = 5$ in our application) are then sent to our image registration module for comparison. While simple, we have obtained good results with this approximation over multiple distinct data sets, and it has been the basis for the work presented in this article using automatically proposed links.

1: define: k {maximum number of candidates to return}
2: define: $\epsilon_{min} \in [0,1]$ {minimum percent overlap}
3: define: $\epsilon_{max} \in [0,1]$ {maximum percent overlap}
4: define: $\alpha \in [0,1]$ {confidence-level}
5: **for all** I_i **do**
6: $A_{max} \leftarrow \max(A_i, A_r)$
7: $W_{max} \leftarrow 2A_{max}\tan(\frac{1}{2}\text{FOV})$
8: $d_{min} \leftarrow (1-\epsilon_{max}) \cdot W_{max}$
9: $d_{max} \leftarrow (1-\epsilon_{min}) \cdot W_{max}$
10: extract from our state, $\boldsymbol{\xi}_t$, the joint-marginal: $\begin{bmatrix} \mathbf{x}_{p_i} \\ \mathbf{x}_{p_r} \end{bmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} \boldsymbol{\mu}_{p_i} \\ \boldsymbol{\mu}_{p_r} \end{bmatrix}, \begin{bmatrix} \Sigma_{p_i p_i} & \Sigma_{p_i p_r} \\ \Sigma_{p_r p_i} & \Sigma_{p_r p_r} \end{bmatrix}\Big)$
11: compute the relative camera pose, $\mathbf{x}_{c_r c_i}$, and its first-order statistics (6),(7)
12: using $\mathbf{x}_{c_r c_i}$ compute the Euclidean distance d_{ri} and its first-order statistics: $d_{ri} \sim \mathcal{N}\big(\mu_{d_{ri}}, \sigma^2_{d_{ri}}\big)$ where $d_{ri} \leftarrow \|^{c_r}\mathbf{t}_{c_r c_i}\|$
13: compute the probability P_i that $d_{min} < d_{ri} < d_{max}$: $P_i \leftarrow \int_{d_{min}}^{d_{max}} \mathcal{N}\big(\tau; \mu_{d_{ri}}, \sigma^2_{d_{ri}}\big) d\tau$
14: **if** $P_i > \alpha$ **then**
15: add I_i to the candidate set S
16: **end if**
17: **end for**
18: sort candidate set S by P_i and return up to the k most probable candidates

Algorithm 1: View-based link hypothesis. Hypothesize which images, I_i, in our view-based map have a high probability of overlapping with the current robot view, I_r.

V. Generating the 5-DOF Camera Measurement

Having presented a view-based estimation framework capable of incorporating 5-DOF relative-pose measurements, we now turn our attention to explaining *how* we actually make the pairwise camera measurement. At its core is a feature-based image registration engine whose purpose is to generate pairwise measurements of relative-pose. Essential to this goal is the capability to cope with low-overlap image registration for two main reasons.

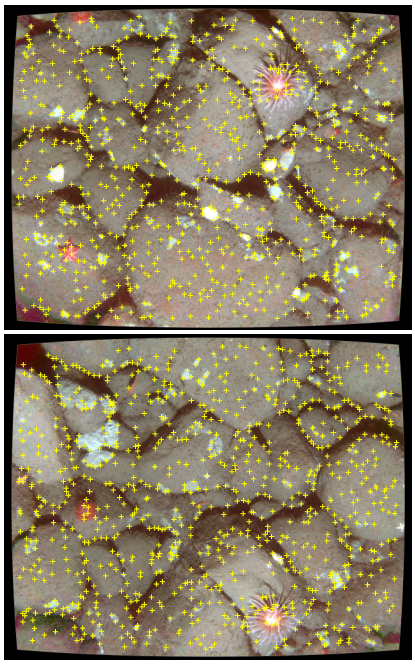
1) Low-overlap digital-still imagery is common in our temporal image sequences due to the nature of our underwater application. Therefore, we must be able to accommodate images in the temporal sequence having 35% or less sequential overlap.
2) Loop-closing and cross-track spatial image constraints are the greatest strength of a VAN methodology. It is these measurements that help to correct dead-reckoned drift error and enforce recovery of a consistent trajectory. Since low-overlap viewpoints are typical in this scenario, this condition would arise even if temporal overlap were much higher as with video-frame rates.

Thus, in order to be able to successfully handle low-overlap image registration, our approach has been to extend a typical state-of-the-art feature-based image registration framework to judiciously exploit our navigation prior wherever possible. For example, in §V-B we show how we can exploit absolute orientation sensor measurements to reduce viewpoint variability in our feature encoding, and also obtain a good initialization for pairwise maximum likelihood refinement. We also show in §V-C how we can use our pose prior and altitude measurements to improve the robustness of correspondence establishment via a novel pose-constrained correspondence search.

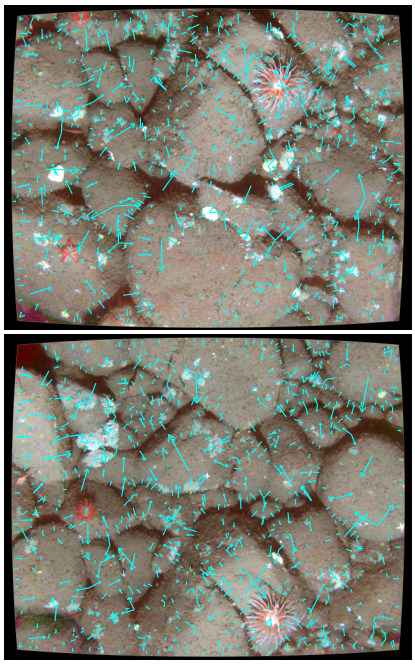
A. Pairwise Feature-Based Image Registration

1) Geometric Feature-Based Algorithm: Our feature-based registration algorithm generally follows a state-of-the-art geometrical computer vision approach as described by Hartley and Zisserman [37] and Faugeras, Luong, and Papadopoulo [38]. Figures 6 and 7 illustrate the overall hierarchy of our feature-based algorithm founded on:

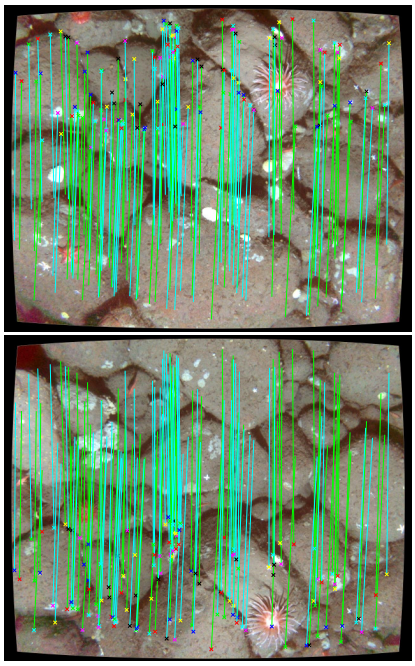
- Extract a combination of both Harris [40] and SIFT [41] interest points from each image. It has been our



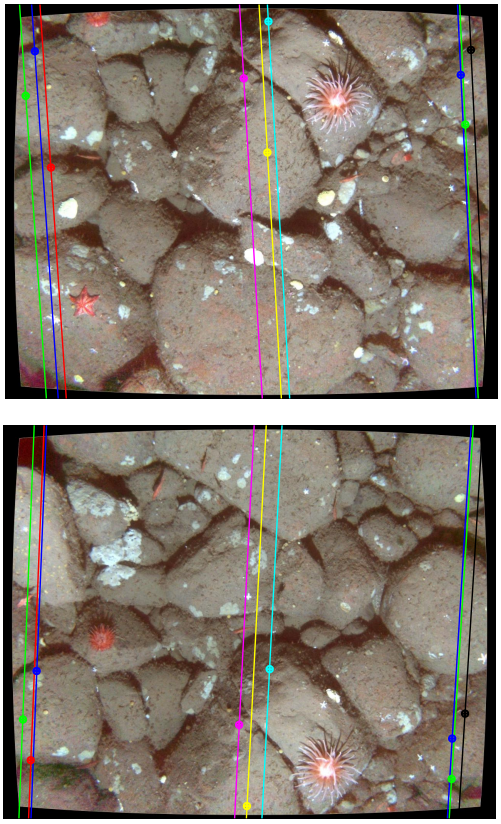
(a) Harris interest points.



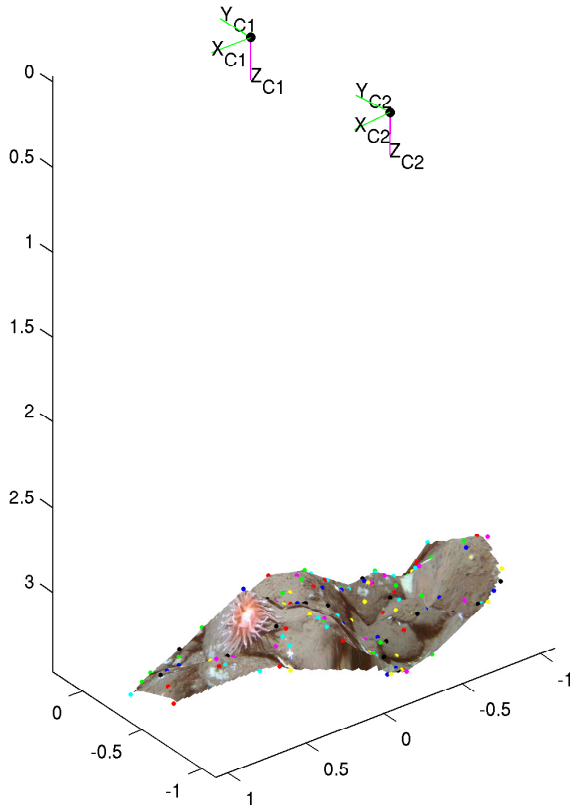
(b) SIFT feature points.



(c) Inlier correspondences.



(d) MLE epipolar geometry.



(e) MLE relative-pose and texture mapped scene (units in meters).

Fig. 6. Typical output from our pairwise feature-based image registration module for a temporally sequential pair of underwater images. To aid visualization, the images have been color corrected using the algorithm described in [39]. The pose and triangulated 3D feature points are the final product of a two-view MLE bundle adjustment step. The 3D triangulated feature points have been gridded in MATLAB to give a coarse surface approximation that has then been texture mapped with the common image overlap (the baseline magnitude is set to the navigation prior for visualization).

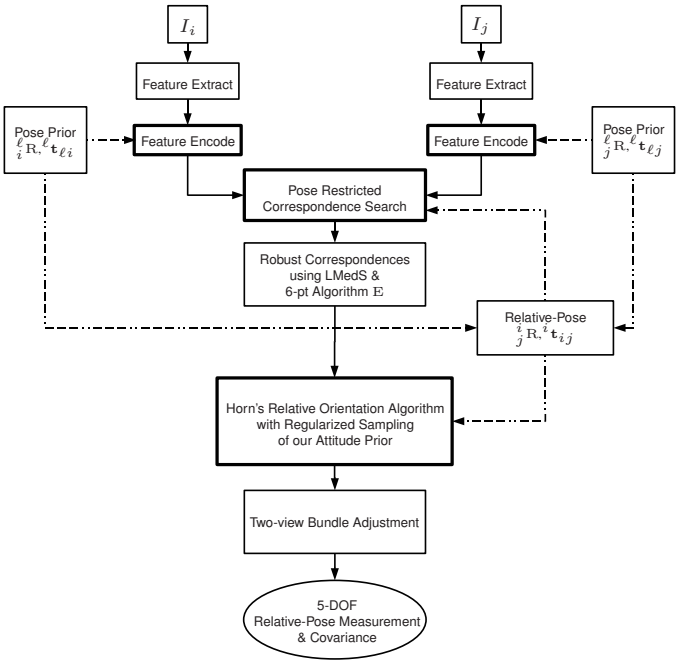


Fig. 7. An overview of the pairwise image registration engine. Dashed lines represent additional information provided by our state estimate, while bold boxes represent our systems-level extensions to a typical feature-based registration framework.

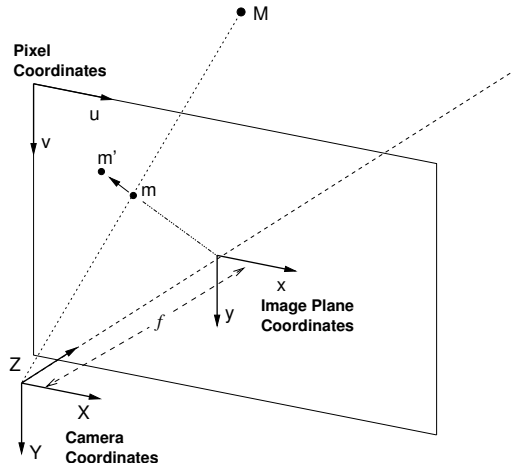


Fig. 8. Illustration of a pinhole camera model. An intrinsically calibrated camera implies that the mapping from Euclidean camera coordinates to image pixel coordinates is known. The pinhole projective mapping from scene point $\mathbf{M}$ to image point $\mathbf{m}$ is described in homogeneous coordinates in terms of a 3×4 projection matrix $\mathrm{P} = \mathrm{K}[\mathrm{R} \,|\, \mathbf{t}]$ where K is the 3×3 upper triangular intrinsic parameter matrix and R,$\mathbf{t}$ describe the extrinsic coordinate transformation from scene to camera centered coordinates [37]. In practice, we must also account for the lens distortion, which further maps $\mathbf{m}$ to $\mathbf{m}'$ [48].

experience that the Harris points provide a high density of temporal matches thereby yielding a high precision observation of along-track motion, while the SIFT's rotational and scale invariance adds cross-track robustness by providing a sufficient number of putative correspondences for loop-closing. For the Harris points, we first normalize the surrounding interest regions by exploiting our navigation prior to apply an orientation correction via the infinite homography [37] before compactly encoding using Zernike moments [42].

- Establish putative correspondences between overlapping candidate image pairs based upon similarity and a pose-constrained correspondence search (PCCS) [43].
- Employ a statistically robust least median of squares (LMedS) [44] registration methodology with regularized sampling [45] to extract a consistent inlier correspondence set. For this task we use a 6-point Essential matrix algorithm [46] as the motion-model constraint.
- Solve for a relative-pose estimate using the inlier set and Horn's relative orientation algorithm [47] initialized with samples from our orientation prior.
- Carry out a two-view maximum likelihood estimate (MLE) to extract the 5-DOF relative-pose constraint (i.e., azimuth, elevation, Euler roll, Euler pitch, Euler yaw) and first-order parameter covariance based upon minimizing the reprojection error over all inliers [37].

2) Calibrated Camera Model: Within our feature-based framework, we assume a standard calibrated pin-hole camera model [37] as illustrated in Fig. 8. This means that the homogeneous mapping from world to image plane can be described by a 3×4 projection matrix P defined as

$$\mathrm{P} = \mathrm{K}\big[{}^{c}_{w}\mathrm{R} \,\big|\, {}^{c}\mathbf{t}_{cw}\big].$$

Here, ${}^{c}_{w}\mathrm{R}$ and ${}^{c}\mathbf{t}_{cw}$ encode the coordinate transformation from world, w, to camera centered coordinate frame, c, and

$$\mathrm{K} = \begin{bmatrix} \alpha_u & s & u_o \\ 0 & \alpha_v & v_o \\ 0 & 0 & 1 \end{bmatrix}$$

is the *known* 3×3 upper triangular intrinsic camera calibration matrix with α_u, α_v the pixel focal lengths in the x, y directions, respectively, (u_o, v_o) is the principle point measured in pixels, and s is the pixel skew.

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

$$\underline{\mathbf{u}} = \mathrm{P}\underline{\mathbf{X}} \tag{9}$$

where $\mathbf{u} = [u, v]^\top$ is the vector description of (u, v), $\underline{\mathbf{u}} = [\mathbf{u}^\top, 1]^\top$ its normalized homogeneous representation, $\mathbf{X} = [X, Y, Z]^\top$ is the imaged 3D scene point, and $\underline{\mathbf{X}} = [\mathbf{X}^\top, 1]^\top$ its normalized homogeneous representation. Note that for all homogeneous quantities, equality in expressions such as (9) is implicitly defined up to scale. The benefit of having a calibrated camera is that we can "undo" the projective mapping in (9) and instead work with Euclidean rays:

$$\underline{\mathbf{x}} = \mathrm{K}^{-1}\underline{\mathbf{u}} = \big[{}^{c}_{w}\mathrm{R} \,\big|\, {}^{c}\mathbf{t}_{cw}\big]\underline{\mathbf{X}}.$$

The implication is that we can now describe the epipolar geometry in terms of the Essential matrix [37] and recover the 5-DOF camera pose from correspondences. For our application, we obtain the intrinsic calibration matrix, K, by calibrating in water using Zhang's planar method [49] and employ Heikkilä's radial/tangential distortion model [48] to compensate for both lens and index of refraction effects.

B. Exploiting Sensor-Measured Absolute Orientation

1) Infinite Homography View Normalization: Establishing feature correspondences is arguably the most difficult task in a feature-based registration approach — this is especially true for low-overlap image registration. Without any knowledge of extrinsic camera information, robust techniques must rely upon encoding features in a viewpoint invariant way. For example, rotational and scale differences between images render simple correlation-based similarity metrics useless. Therefore, to overcome these limitations, advanced techniques generally rely upon encoding some form of locally invariant feature descriptor such as differential invariants [50], generalized image moments [42], [51], [52], and affine invariant regions [53]–[55]. These higher-order descriptions, however, also tend to be computationally expensive.

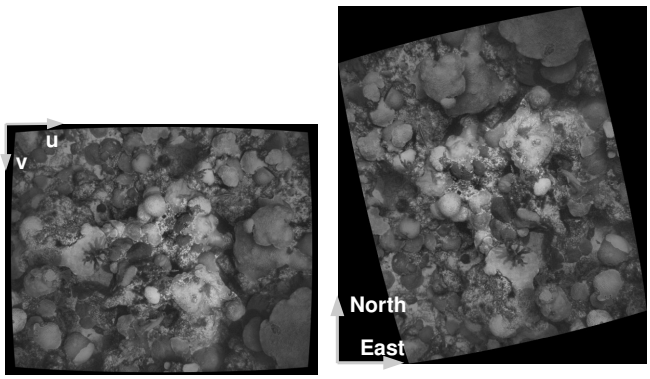
In the case of an instrumented platform with absolute measurements of orientation, we can use sensor-derived information to our advantage to relax the demands of the feature encoding while at the same time making it a more discriminatory metric. For example, attitude can be measured with bounded-error over the entire survey site. Therefore, in our application we use sensor-derived absolute orientation information on camera pose $\mathbf{x}_{\ell c_i}$ to normalize the feature regions around the Harris interest points in image I_i via the infinite homography:

$$\mathrm{H}_\infty = \mathrm{K}\,{}^{\ell}_{c_i}\mathrm{R}\,\mathrm{K}^{-1}.$$

This homography warps image I_i, taken from camera pose $\mathbf{x}_{\ell c_i}$, into a synthetic view I_ℓ, corresponding to a simulated view from a colocated frame at a canonical orientation. This viewpoint mapping is *exact* for points at *infinity* where $\underline{\mathbf{X}} = [X, Y, 0, 1]$, but otherwise can be used to compensate for viewpoint orientation (note that scene parallax is still present).

We compute H_∞ based upon our attitude estimate at image acquisition and apply it as an orientation correction to our images when encoding the Harris features. As demonstrated in Fig. 9, this warp effectively yields a synthetic view of the scene from a canonical camera coordinate-frame aligned North, East, Down. This allows normalized correlation to be used as a similarity metric between Harris points and tends to work well for temporally sequential image sequences by generating a high density of matches. This scheme in concert with SIFT features has proven to be successful for obtaining robust similarity matches over the entire survey site.

2) Sampling from our Orientation Prior: We can also take advantage of our absolute orientation prior by obtaining an initial relative-pose solution using Horn's algorithm [47]. Given a set of inlier feature correspondences and an initial orientation guess, Horn's algorithm iteratively calculates a relative-pose estimate based upon enforcing the co-planarity condition over all ray pairs (i.e., if a ray from the left and right camera are to intersect then they must lie in a plane that also contains the baseline). If the orientation guess is approximately close to the true orientation, Horn's algorithm quickly converges to a minimal co-planarity error solution. Since orientation can be measured with bounded precision over the *entire* survey site while the camera baseline cannot, we use Horn's algorithm to obtain our initial 5-DOF relative-pose solution, which is then refined in a two-view bundle adjustment step based upon minimizing the reprojection error [37].



(a) Original lens distortion compensated image, I_i. (b) Normalized image, I_ℓ, using H_∞ warp.

Fig. 9. A demonstration of synthetically normalizing for the camera orientation via the infinite homography. The imagery is of deep-water coral.

C. Pose-Constrained Correspondence Search (PCCS)

As previously mentioned, the problem of initial feature correspondence establishment is arguably the most difficult and challenging task of a feature-based registration methodology. As we show in this section, having a pose prior relaxes the demands on the complexity of the feature descriptor — instead of having to be globally unique within an image, it now is required to be only locally unique. We use the epipolar geometry 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 altitude, and are used to confine the interest point matching to a small subset of candidate correspondences. The benefit of this approach is that it simultaneously relaxes the demands of the feature descriptor while at the same time improves the robustness of similarity matching.

1) Epipolar Uncertainty Representation: Zhang [45] first characterized epipolar geometry uncertainty in terms of the covariance of the fundamental matrix while Shen [56] used knowledge of the pose prior to restrict the correspondence search to bands along the epipolar line calculated by propagating pose uncertainty. However, a criticism of both of these characterizations is that the uncertainty representation is hard to interpret in terms of physical parameters — how does one interpret the covariance of a line? Our approach is to use a two-view point transfer mapping that benefits from a direct physical interpretation of the pose parameters and, in addition, can take advantage of scene range data if available. While similar to Lanser's technique [57], our approach does not assume nor require that an *a priori* CAD model of the environment exist.

2) Two-View Point Transfer Model: In deriving the point transfer mapping we assume projective camera matrices $\mathrm{P} = \mathrm{K}[\mathrm{I} \,|\, \mathbf{0}]$ and $\mathrm{P}' = \mathrm{K}[\mathrm{R} \,|\, \mathbf{t}]$, where for notational convenience we simply write the relative-pose parameters as R,$\mathbf{t}$.

We begin by noting that the scene point $\mathbf{X}$ is projected through camera P as

$$\underline{\mathbf{u}} = \mathrm{P}\underline{\mathbf{X}} = \mathrm{K}\mathbf{X},$$

which implies that explicitly accounting for scale we have

$$\mathbf{X} \equiv Z\mathrm{K}^{-1}\underline{\mathbf{u}}. \tag{10}$$

The back-projected scene point, $\mathbf{X}$, can subsequently be re-projected into image I' as

$$\underline{\mathbf{u}}' = \mathrm{P}'\underline{\mathbf{X}} = \mathrm{K}(\mathrm{R}\mathbf{X} + \mathbf{t}). \tag{11}$$

By substituting (10) into (11) and recognizing that the following relation is up to scale, we obtain the homogeneous point transfer mapping [37]:

$$\underline{\mathbf{u}}' = \mathrm{KRK}^{-1}\underline{\mathbf{u}} + \mathrm{K}\mathbf{t}/Z. \tag{12}$$

Finally, by explicitly normalizing (12) we recover the *non-homogeneous* point transfer mapping

$$\mathbf{u}' = \frac{\mathrm{H}_\infty\underline{\mathbf{u}} + \mathrm{K}\mathbf{t}/Z}{\mathbf{H}_\infty^{3\top}\underline{\mathbf{u}} + t_z/Z} \tag{13}$$

where $\mathrm{H}_\infty = \mathrm{KRK}^{-1}$, $\mathbf{H}_\infty^{3\top}$ refers to the third row of H_∞, and t_z is the third element of $\mathbf{t}$.

When the scene depth Z of the image point $\underline{\mathbf{u}}$ is known, then (13) describes the exact two-view point transfer mapping. When Z is unknown, however, then (13) describes a functional relationship on Z (i.e., $\mathbf{u}' = f(\mathbf{u}, Z)$) that traces out the corresponding epipolar line in I' [58].

3) Point Transfer Mapping with Uncertainty: Now that we have derived the two-view point transfer mapping (13), in this section we show how we can use it to constrain our correspondence search between image pair (I_i, I_j) by using our *a priori* pose knowledge from $\boldsymbol{\xi}_t$. We begin by defining the parameter vector, $\boldsymbol{\gamma}$, as

$$\boldsymbol{\gamma} = [\mathbf{x}_{p_i}^\top, \mathbf{x}_{p_j}^\top, Z, u, v]^\top \tag{14}$$

with mean, $\boldsymbol{\mu}_\gamma$, and covariance, Σ_γ, given by

$$\boldsymbol{\mu}_\gamma = \begin{bmatrix} \boldsymbol{\mu}_{p_i} \\ \boldsymbol{\mu}_{p_j} \\ Z \\ u \\ v \end{bmatrix} \qquad \Sigma_\gamma = \begin{bmatrix} \Sigma_{p_ip_i} & \Sigma_{p_ip_j} & 0 & 0 & 0 \\ \Sigma_{p_jp_i} & \Sigma_{p_jp_j} & 0 & 0 & 0 \\ 0 & 0 & \sigma_Z^2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Here, $\mathbf{x}_{p_i}$, $\mathbf{x}_{p_j}$ are the delayed-state vehicle poses extracted from $\boldsymbol{\xi}_t$ (used to calculate relative camera pose according to (6)), Z and σ_Z represent the scene depth parameters as measured in camera frame i, and (u, v) describe the feature location in pixels in image I_i. In defining Σ_γ we employ the standard assumption that features are extracted with isotropic, independent, unit variance noise [37] when defining the Σ_{uv} sub-block. To obtain a first-order estimate of the uncertainty in the point transfer mapping between I_i and I_j we compute

$$\boldsymbol{\mu}_{u'} \approx (13)\big|_{\boldsymbol{\mu}_\gamma} \tag{15}$$

$$\Sigma_{u'} \approx \mathrm{J}\Sigma_\gamma\mathrm{J}^\top \tag{16}$$

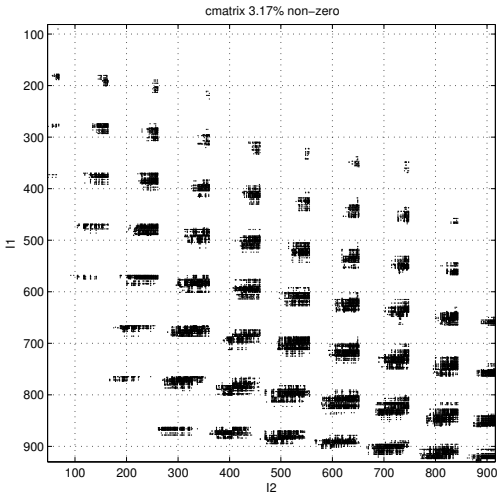


Fig. 10. The pose-constrained candidate correspondence matrix associated with Fig. 11(c). The rows/columns correspond to an ordering of the feature indices in I_i/I_j, respectively. Here, a nonzero entry indicates a potential match. Note that without any *a priori* pose knowledge this matrix would be full meaning that we would be forced to rely purely upon the discriminatory power of the feature similarity measure to establish correspondences. Instead, by applying the PCCS, we reduce the possible space of matches by over 97%.

where $\boldsymbol{\mu}_{u'}$ is the predicted point location of $\mathbf{u}$ in I_j, $\Sigma_{u'}$ its covariance, and $\mathrm{J} = \frac{\partial \mathbf{u}'}{\partial \boldsymbol{\gamma}}$ is the point transfer Jacobian.[3]

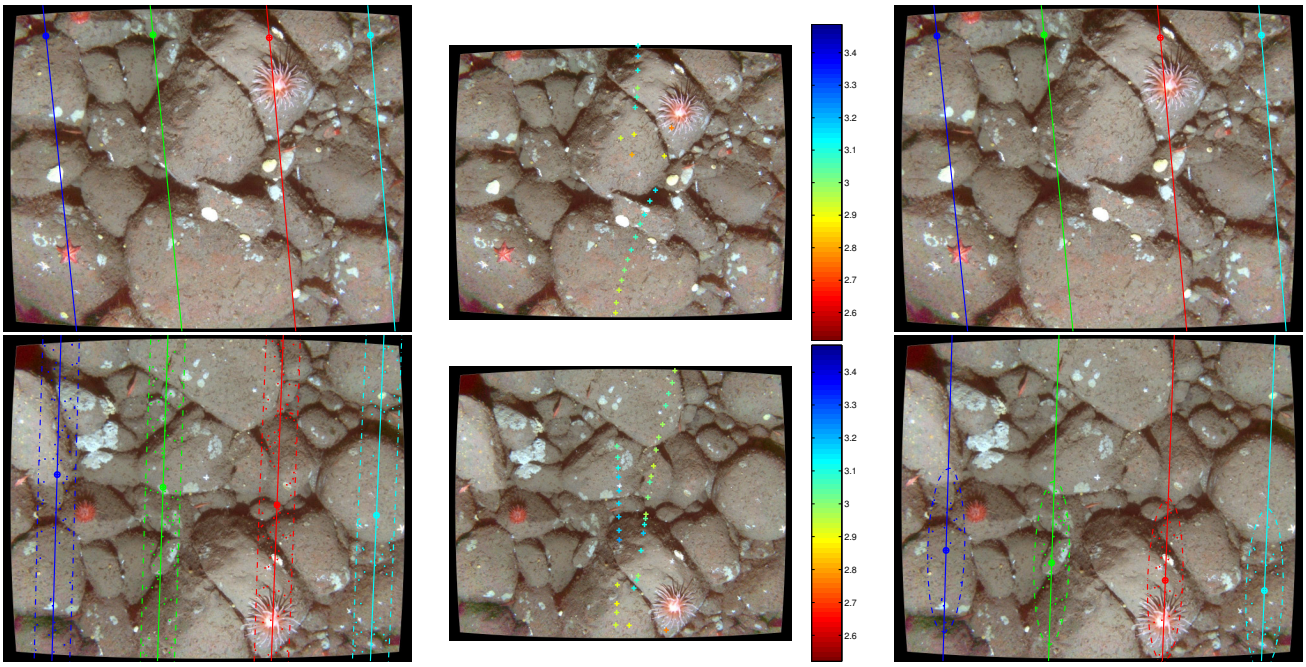
We use this knowledge to restrict our correspondence search using a Mahalanobis distance test:

$$\big(\mathbf{u}' - \boldsymbol{\mu}_{u'}\big)^\top \Sigma_{u'}^{-1} \big(\mathbf{u}' - \boldsymbol{\mu}_{u'}\big) = k^2 \tag{17}$$

where the threshold k^2 follows a χ_2^2 distribution. Under this scheme we test all feature points in I_j to see if they satisfy (17), and if they do, then they are considered to be candidate matches for $\mathbf{u}$. 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.

Algorithm 2 describes the PCCS in pseudo-code where we use scene depth, Z, and its uncertainty, σ_Z, as a convenient parameterization for controlling the size of the search regions in I_j through the definition of the parameter vector $\boldsymbol{\gamma}$. For example, in the case where no *a priori* knowledge of scene depth is available, choosing any finite value for Z and setting $\sigma_Z \to \infty$ recovers a search *band* along the epipolar line in I_j whose width corresponds to the uncertainty in relative camera pose, $\mathbf{x}_{c_jc_i}$ (Fig. 11(a)). On the other hand, when knowledge of an average scene depth, Z_{avg}, exists (e.g., from an altimeter), then it and an appropriately chosen σ_Z can be used to limit the search space to *ellipses* centered along the epipolar lines (Fig. 11(c)). Furthermore, in the case where dense scene range measurements are available (e.g., from a

[3]We compute this Jacobian numerically as described in [37, §A4.2].



(a) Search bands corresponding to no scene depth prior (i.e., $\sigma_Z \rightarrow \infty$).

(b) Altimeter measured scene depth projected into the image plane (units in meters).

(c) Search bands become regions based upon altimeter measured scene depth constraint.

Fig. 11. Demonstration of the PCCS for a temporal pair of underwater images. Images are arranged I_i above and I_j below; the two-view mapping is shown for $I_i \rightarrow I_j$. (a) Pose-prior instantiated epipolar lines are shown in both I_i and I_j. The search *bands* in I_j correspond to *no* knowledge of scene depth with width attributable to relative-pose uncertainty. (b) Altimeter measured scene depth projected into the image plane of each view (the altimetry is derived from the beam range measurements of the DVL). (c) The search regions now become *ellipses* based upon the altimetry constraint.

laser range finder or multibeam sonar), then scene depth, Z, can be assigned on a point-by-point basis with high precision. In any case, the PCCS greatly improves the reliability and robustness of feature similarity matching by reducing the candidate correspondence set to a relatively few number of options as demonstrated in Fig. 10.

D. Are Pairwise Camera Measurements Correlated?

We now address the question of whether or not pairwise camera measurements are correlated. Recall that a primary assumption in the system model (1) is that measurements are assumed to be corrupted by time independent noise. In our view-based framework, images are pairwise registered to produce a 5-DOF relative-pose measurement that is then fed to the filter as an observation between the two corresponding delayed-states. If an image is reused multiple times to make multiple pairwise measurements, for example $I_i \leftrightarrow I_j$ and $I_j \leftrightarrow I_k$, then this raises the possibility that camera measurements $\mathbf{z}_{ij}$ and $\mathbf{z}_{jk}$ could be statistically correlated. Neglecting such a correlation would put too much weight on the filter update step since it would treat observations $\mathbf{z}_{ij}$ and $\mathbf{z}_{jk}$ as being *independent* pieces of information. Unfortunately, actually computing all possible measurement correlations quickly becomes intractable in any scan-matching framework. Thus, like other scan-matching algorithms [19], [20], out of practicality we assume relative-pose measurements to be statistically independent.[4] We argue, however, that for our AUV application the low degree of temporal image overlap in our digital-still imagery renders the measurement independence assumption not particularly far from the truth.

To see this, we note that our camera-derived relative-pose measurement and covariance are generated as an end-product of a feature-based two-view maximum likelihood estimate based upon minimizing the reprojection error. As is standard in the vision community, the image feature locations are assumed to be corrupted by independent isotropic noise of unit variance [37]. Denoting the set of common features between $I_i \leftrightarrow I_j$ as F_{ij}, and the set between $I_j \leftrightarrow I_k$ as F_{jk}, the implication of this noise model is that for null pairwise feature intersection (i.e., $F_{ij} \cap F_{jk} = \varnothing$), the corresponding camera measurements $\mathbf{z}_{ij}$ and $\mathbf{z}_{jk}$ are *uncorrelated* [58]. Hence, pairwise independence holds for image sequences with less than 50% sequential image overlap, which is frequently the case for our along-track digital-still imagery and *approximately* the case for our cross-track imagery where the number of re-observed point correspondences is low.

VI. Results

In this section we present results demonstrating VAN's application to underwater trajectory estimation. The first set of results are for experimental validation of the VAN framework

[4]Recent work by Mourikis and Roumeliotis [59] reports an EKF filtering technique that can account for correlation in temporal relative-pose measurements, however, this technique also becomes intractable in the general case.

Require: U_i {the set of feature points in image I_i}
Require: U_j {the set of feature points in image I_j}
Require: $\begin{bmatrix}\boldsymbol{\mu}_{p_i}\\ \boldsymbol{\mu}_{p_j}\end{bmatrix}, \begin{bmatrix}\Sigma_{p_ip_i} & \Sigma_{p_ip_j}\\ \Sigma_{p_ip_j}^\top & \Sigma_{p_jp_j}\end{bmatrix}$ {*a priori* pose knowledge}
Require: Z, σ_Z^2 {scene depth prior}
1: $C_{ij} \leftarrow 0_{U_i \times U_j}$ {initialize the ij correspondence matrix}
2: **for all** $\mathbf{u}_i \in U_i$ **do** {forward mapping from I_i to I_j}
3: assemble $\boldsymbol{\gamma}$ as in (14)
4: do point transfer $\boldsymbol{\mu}_{u'_i} \leftarrow (15)\big|_{\boldsymbol{\mu}_\gamma}$ $\Sigma_{u'_i} \leftarrow (16)\big|_{\Sigma_\gamma}$
5: **for all** $\mathbf{u}_j \in U_j$ **do** {Mahalanobis test}
6: **if** $\left(\mathbf{u}_j - \boldsymbol{\mu}_{u'_i}\right)^\top \Sigma_{u'_i}^{-1} \left(\mathbf{u}_j - \boldsymbol{\mu}_{u'_i}\right) < k^2$ **then**
7: $C_{ij}(\mathbf{u}_i, \mathbf{u}_j) \leftarrow 1$ {flag $\mathbf{u}_i$, $\mathbf{u}_j$ as candidate match}
8: **end if**
9: **end for**
10: **end for**
11: repeat lines 1–10 with the role of U_i, U_j swapped to produce the ji correspondence matrix C_{ji}
12: $C \leftarrow C_{ij} \& C_{ji}^\top$ {compute the bitwise AND between C_{ij} and C_{ji} to find a consistent forwards/backwards mapping}
13: assign putative matches from the candidate correspondence matrix C using image feature similarity measures

Algorithm 2: Pose-constrained correspondence search.

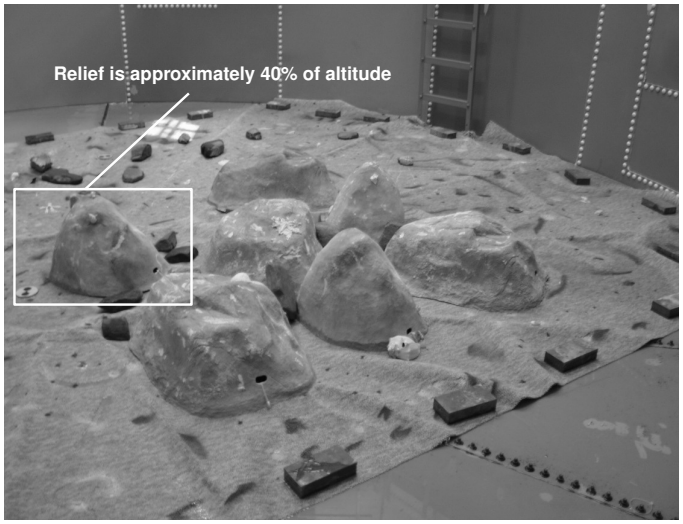


Fig. 12. A partial view of the JHU experimental setup. Low-pile carpet, artificial landscaping boulders, and riverbed rock were all placed on the tank floor to create a natural looking seafloor with extensive scene relief for a camera altitude of 1.5 m.

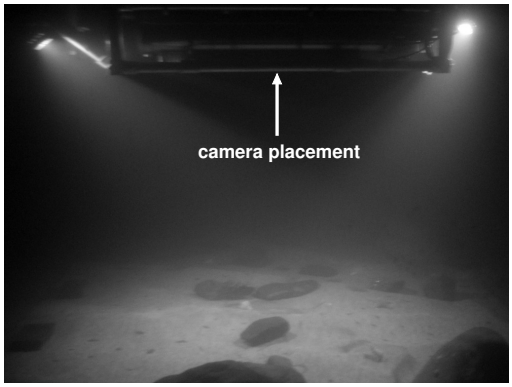
using a ROV at the Johns Hopkins University (JHU) Hydrodynamic Test Facility with ground-truth. The second set of results are for a real-world data set collected by the SeaBED AUV during a benthic habitat classification survey conducted at the Stellwagen Bank National Marine Sanctuary.

A. Experimental Validation: JHU Test Tank

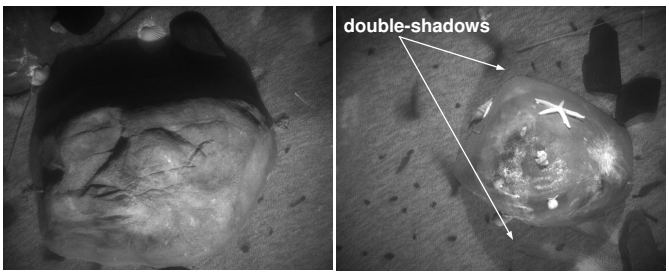
To better understand the error characteristics of VAN as compared to traditional dead-reckoning navigation, we collaborated with our colleagues at JHU to collect an in-tank ROV data set with ground-truth.

1) Experimental Setup: The experimental setup consisted of a single downward-looking digital-still camera mounted to a moving underwater pose instrumented ROV at the JHU Hydrodynamic Test Facility [60]. Their vehicle [61] is instrumented with a typical suite of oceanographic dead-reckoning navigation sensors capable of measuring heading, attitude, XYZ bottom-referenced Doppler velocities, and a pressure sensor for depth. The vehicle and test facility are also equipped with a high frequency acoustic long-baseline (LBL) system, which provides centimeter-level bounded error XY vehicle positions used for validation purposes only. A simulated seafloor environment (Fig. 12) was created by placing textured carpet, riverbed rocks, and landscaping boulders on the tank floor and was appropriately scaled to match a rugged seafloor environment with considerable 3D scene relief.

In addition, we also tested an innovative dual-light configuration consisting of fore and aft lights on the ROV with the camera mounted in the center as shown in Fig. 13. This dual-light configuration was meant to alleviate viewpoint illumination effects by improving the signal-to-nose ratio in shadowed regions so that fully automatic cross-track correspondence could be achieved.

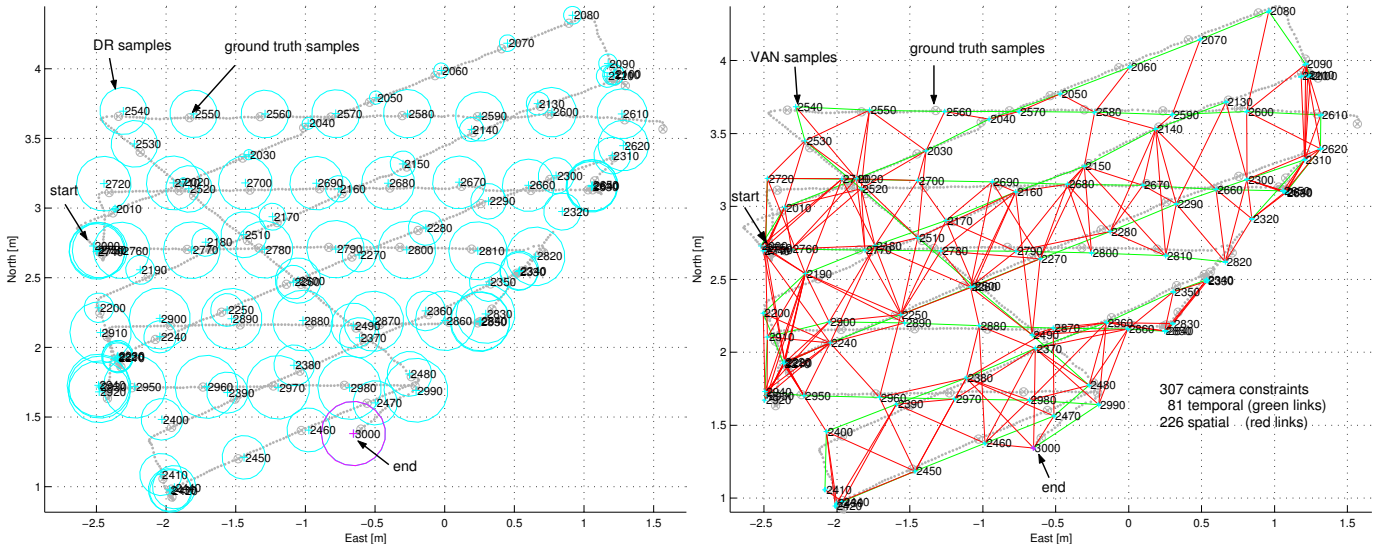


(a) Dual-light ROV configuration.



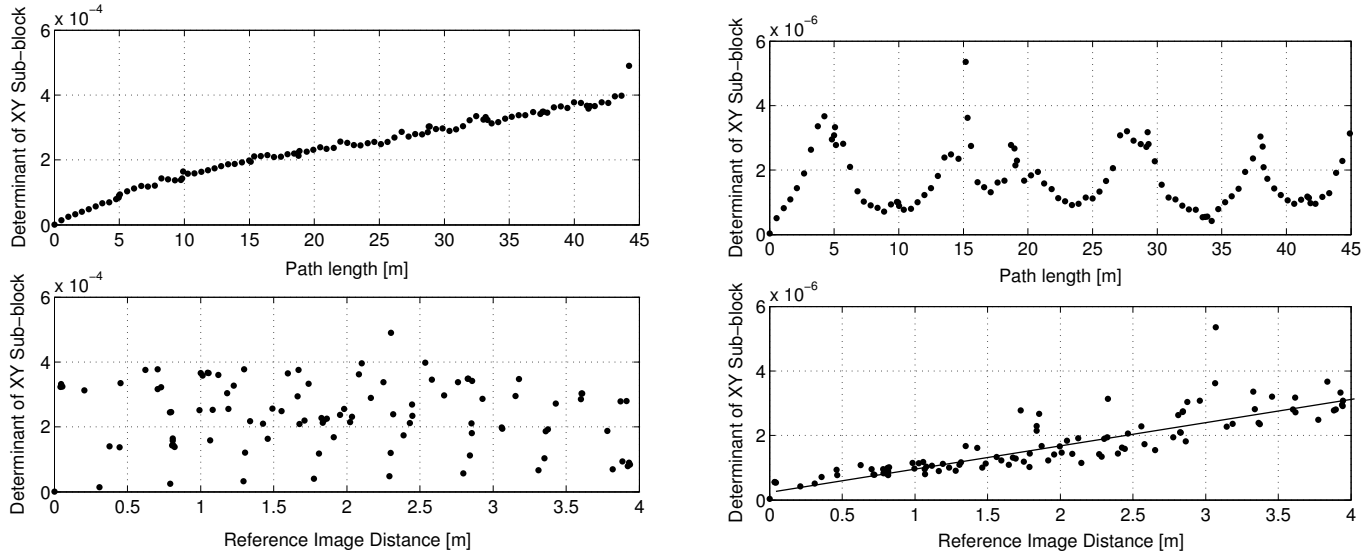
(b) Single-light illumination. (c) Dual-light illumination.

Fig. 13. The dual-light setup used on the JHU ROV. (a) This experimental dual-light configuration, with the camera mounted in the center, made fully automatic image registration robust to the effects of viewpoint variant scene illumination. (b) Traditional single-light configurations cast significant shadows and cause objects to look very different from differing vantage points making automatic correspondence establishment difficult. (c) In contrast, the innovative dual-light configuration increases image illumination invariance by creating double-shadowed regions, which are imaged with high fidelity.



(a) DR trajectory. (b) VAN trajectory.

Fig. 14. JHU tank results comparing DR and VAN trajectories to 300 kHz LBL ground-truth for a 101 image sequence. We subsampled the image sequence using only every 10^{th} frame to achieve roughly 25% temporal overlap (frame numbers start at 2000). The survey consisted of two overlapping grid trajectories, one oriented NE/SW and the other E/W. The vehicle pose samples and 3σ confidence ellipses are shown for all 101 views. The corresponding time samples from the ground truth trajectory are designated by the gray circles. The VAN result is end-to-end fully automatic including link hypothesis (Algorithm 1) and correspondence establishment. Notice that XY uncertainty grows monotonically in the DR trajectory estimate while for VAN it is constrained by the camera-constraint topology.



(a) DR uncertainty characterization. (b) VAN uncertainty characterization.

Fig. 15. Uncertainty characteristics of VAN versus DR for the JHU tank data set of Fig. 14. Plots (a) and (b) show the determinant of the XY covariance sub-block for each vehicle pose in the view-based map. The determinant is plotted versus both path length and Euclidean distance away from the first image in the view-based map. Notice that the DR uncertainty is clearly a monotonic function of path length whereas VAN uncertainty is related to the distance away from the reference image.

2) Experimental Results: Fig. 14 depicts the estimated XY trajectory for a 101 image sequence comprised of roughly 25% temporal image overlap. For this experiment, the vehicle started near the top-left corner of the plot at $(-2.5, 2.75)$ and then drove a course consisting of two grid-based surveys, one oriented SW to NE and the other W to E. Both plots show the spatial XY pose topology, 3σ confidence bounds, and network of camera constraints — note that the VAN result is end-to-end fully automatic. Again, green links correspond to registered sequential images while red links correspond to non-sequential pairs — in all there are 307 camera constraints (81 temporal / 226 spatial). Notice that the XY uncertainty in the dead-reckoned (DR) estimate grows monotonically with time while in the VAN estimate it is constrained by the camera-link topology.

Fig. 15 further corroborates the above observation and in particular shows that VAN exhibits a linear trend in error growth as a function of distance away from the reference node. Note that the spread of points away from this linear fit is due to inhomogeneity in the number of edges per node in the corresponding pose-constraint network. Nonetheless, this raises the interesting engineering question of how one might go about reducing the slope of the linear relationship exhibited in Fig. 15(b)? From a camera perspective, design criteria that could help improve this performance are:

- *Higher resolution images.* Increased resolution improves both the accuracy and precision with which 2D feature points can be extracted and localized within the viewable image plane. This in turn improves the accuracy and precision of the relative-pose camera measurement.
- *Wider field of view (FOV).* Increasing the camera's FOV improves the pairwise observability of camera motion and, hence, the overall precision of the camera-derived relative-pose measurement. However, increasing the FOV also results in lower spatial resolution, so a good balance between the two is required.
- *Better characterization of feature repeatability.* Recall that our image registration module employs the standard assumption that features are extracted with independent, isotropic, unit variance pixel noise. This noise model does not have any real physical basis, but rather is assumed merely for convenience. Hence, it would be worthwhile to setup a testbed of seafloor imagery for measuring the repeatability of our image feature extractors under different viewing, surface, and lighting conditions. This would provide a more accurate characterization of the feature extraction precision and, thus, a better description for the overall precision of our relative-pose camera measurements. The end effect of this characterization on navigation performance would be a more optimal blending of strap-down sensor versus camera-derived pose measurements.
- *Better camera calibration.* Our registration framework assumes that we are using a calibrated camera, which implies that the projective mapping from Euclidean ray space to image pixel space is known. A poor calibration could introduce a persistent bias into the camera-derived relative-pose measurements and, hence, effect the overall consistency of the state estimate. Therefore, obtaining an accurate calibration is important.

B. Real-World Results: Stellwagen Bank

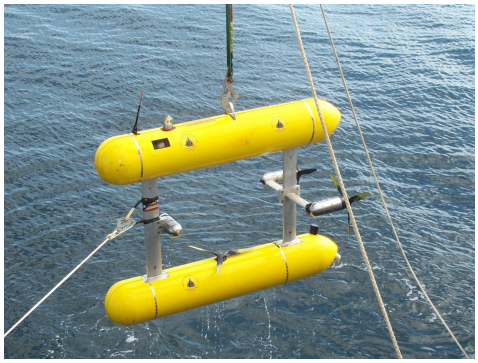
1) Experimental Setup: The SeaBED AUV [23], [24] conducted a grid-based survey for a portion of the Stellwagen Bank National Marine Sanctuary in March 2003. The vehicle was equipped with a single down-looking camera and was instrumented with the navigation sensor suite tabulated in Table I. As depicted in Fig. 16, SeaBED conducted the survey in a bottom-following mode where it tried to maintain constant altitude over a sloping, rocky, ocean seafloor. The intended survey pattern consisted of 15 North/South legs each $180\ \mathrm{m}$ long and spaced $1.5\ \mathrm{m}$ apart while maintaining an average altitude of $3.0\ \mathrm{m}$ above the seafloor at a forward velocity of $0.35\ \mathrm{m/s}$. Closed-loop feedback on the DR navigation estimate was used for real-time vehicle control.

We processed a small subset of the data set using 100 images from a South/North trackline pair, the results of which are shown in Fig. 17. Plot (b) depicts the VAN estimated camera trajectory and its 3σ confidence bounds. Successfully registered image pairs are indicated by the red and green links connecting the camera poses where green corresponds to temporally consecutive image frames and red to spatially neighboring image frames. For comparison purposes, plot (a) depicts the DR trajectory overlaid on top of the VAN estimated XY trajectory. Both plots are in meters where X is East and Y is North.

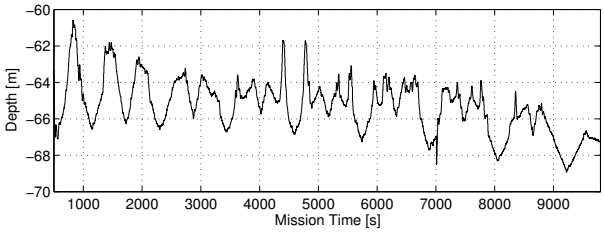
Our feature-based registration algorithm was successful in automatically establishing putative correspondences between sequential image pairs (green links), however, automatic cross-track image registration (red links) proved to be too difficult for this data set. The cause for this is due to significant variation in scene appearance when illuminated from reciprocal headings. The SeaBED AUV uses a single-camera / single-light geometry consisting of a down-looking digital-still camera in the nose and a flash strobe in the tail (not the dual-light configuration like in the previous tank data set). Hence, strong shadows are cast in opposite directions for parallel tracklines viewed from reciprocal headings. Therefore, for this data set cross-track putative correspondences were manually established for 19 image pairs, which are indicated by the red spatial links in Fig. 17.

2) Experimental Results: A number of important observations in Fig. 17 are worth pointing out. First, note that the VAN uncertainty ellipses are smaller for camera poses that are constrained by spatial constraints. Since spatial links provide a mechanism for relating past vehicle poses to the present, they also provide a means for correcting DR drift error. While trajectory uncertainty in a DR navigation system grows monotonically unbounded with time, in contrast VAN's error growth is essentially a function of network topology and distance away from the reference node (i.e., the first image) like we saw with the JHU ground-truth data set.

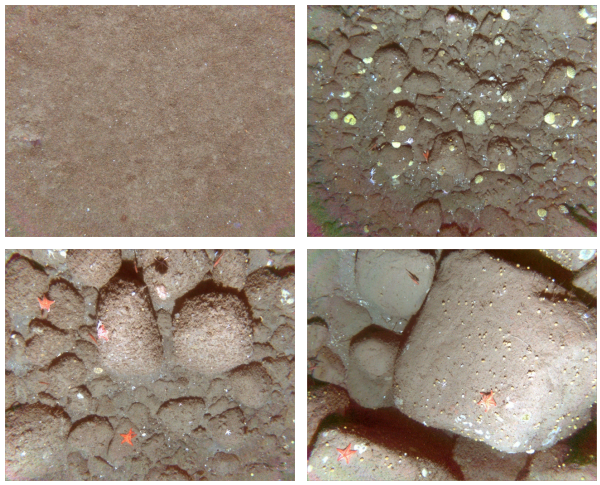
Secondly, note the delayed-state smoothing that occurs in the VAN framework. Spatial links not only decrease the



(a) SeaBED AUV.

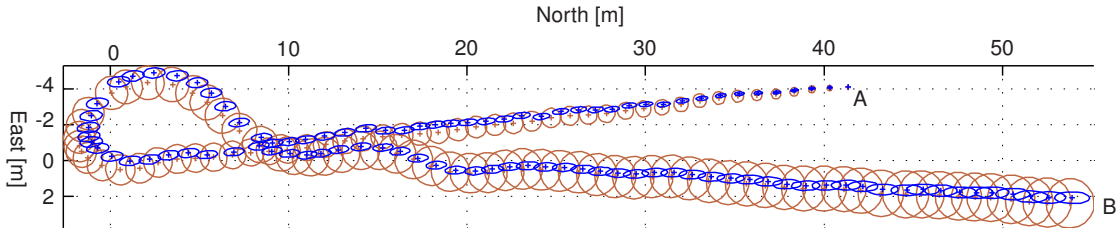


(b) Vehicle depth vs. time.

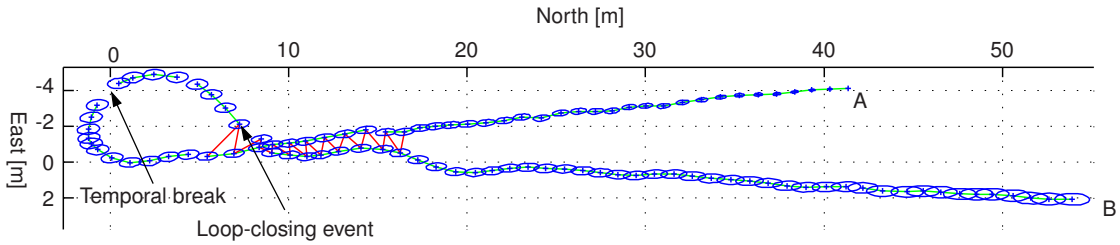


(c) Sample seafloor imagery.

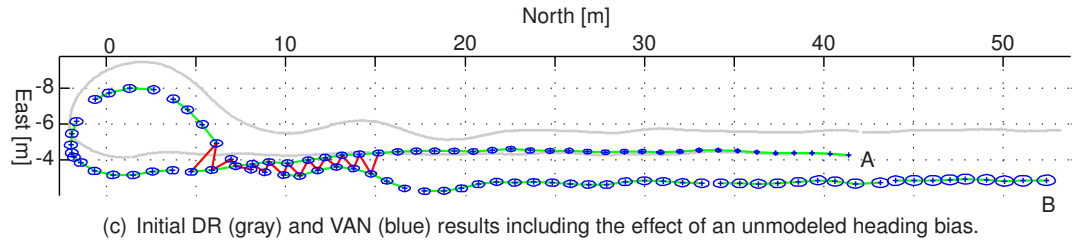
Fig. 16. A depiction of Stellwagen Bank data set. (a) The SeaBED AUV used for the experiment. (b) A plot of vehicle depth vs. time for this mission. Since the vehicle was trying to maintain constant-altitude, the depth plot serves as a proxy for terrain variation. Note that depth excursion are on the order of several meters. (c) A sampling of imagery collected during the survey. The seafloor topography ranges from pure sand (upper left) to large boulders (lower right).



(a) DR (brown) overlaid atop of VAN estimate (blue).



(b) VAN estimate with camera links shown.



(c) Initial DR (gray) and VAN (blue) results including the effect of an unmodeled heading bias.

Fig. 17. A comparison of the VAN and DR recovered trajectories; the survey started at A and ended at B. (a) Shown in blue is the XY plot of 100 estimated camera poses with 3σ confidence ellipses. For comparison, overlaid in brown is the DR estimated trajectory. Notice that the DR error monotonically increases while the VAN error is bounded for images in the vicinity of the cross-over point. (b) The same 100 estimated camera poses, but with image constraints superimposed. The green links indicate that a camera-derived measurement was made between temporally consecutive image pairs, while the red links indicate that a cross-track measurement was made. In all, there are 19 cross-track measurements. (c) The initial DR and VAN results, each of which includes an unmodeled heading bias. Notice that the DR trajectory (gray) does not lie within the 3σ VAN estimate (blue). This discrepancy comes from an unmodeled compass bias, which when accounted for, produces the results shown in (a).

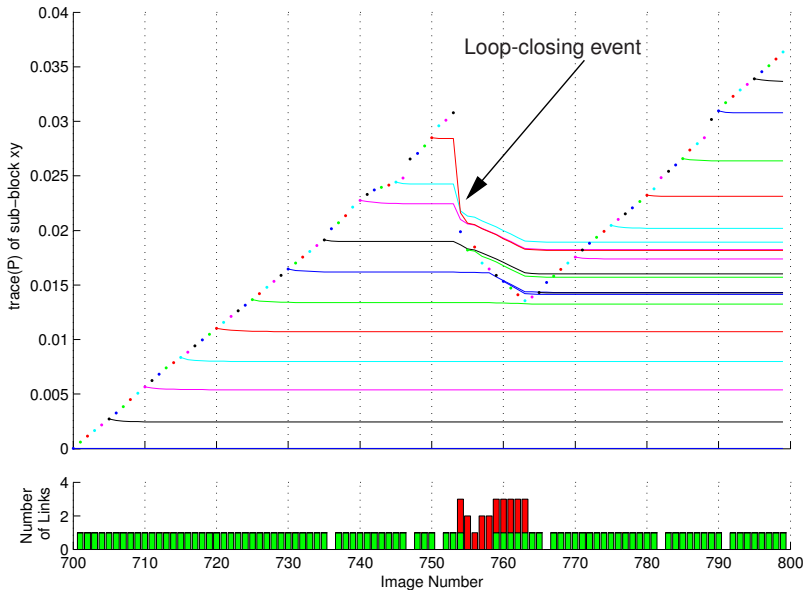


Fig. 18. A depiction of the time-evolution of the Stellwagen Bank pose-network uncertainty. (top) For each delayed-state entry in $\boldsymbol{\xi}_t$ (i.e., for each vehicle pose $\mathbf{x}_{p_i}$), the trace of its 2×2 XY covariance sub-block is plotted versus image frame number. The dots depict the pose uncertainty at image insertion and the lines show their time evolution for every 5^{th} delayed-state. A couple of key events are worth pointing out. First, notice the monotonically increasing uncertainty in XY position between frames 700–753. This initial period corresponds to only sequential pairwise camera measurements. Second, notice the regional smoothing and sharp decrease in uncertainty at frame number 754; this is the first cross-track camera measurement. The pose uncertainty continues to decrease as more cross-track measurements are made (frames 755-763). Finally, from frame 764 onward, uncertainty begins to increase again as no more spatial measurements can be made. (bottom) A bar graph of the number of successfully registered image pairs for each frame number. Sequential camera measurements are green and cross-track measurements are red. Notice that the decrease in XY uncertainty in the covariance plot coincides with the first cross-track measurement.

uncertainty of the image pair involved, but also decrease the uncertainty of other delayed-states that are correlated. In particular, Fig. 18 characterizes the time-evolution of the view-based map uncertainty by plotting the trace of the XY covariance sub-block for each delayed-state versus image frame number. Note the sudden decrease in uncertainty occurring at image frame 754 — this event coincides with the first cross-track link. Information from that spatial measurement is propagated along the network to other vehicle poses via the shared correlations in the covariance matrix. This result is consistent with the spatial error trend exhibited by Fig. 15(b).

Thirdly, referring back to Fig. 17, note that a temporal (green) link does not exist between consecutive image frames near XY location $(-4, 0)$. A break like this in the temporal image chain prevents concatenation of the relative camera measurements and in a purely vision-only approach could cause algorithms that depend on a connected topology to fail. It is a testament to the robustness of VAN that a disconnected camera topology does not present any significant issue since the Kalman filter continues to maintain correlations between the delayed-state entries despite the absence of camera measurements.

Finally, an additional point worth mentioning is that VAN results in a self-consistent estimate of the vehicle's trajectory. Referring to Fig. 17(c), initial processing of the image sequence resulted in a VAN trajectory estimate that did not lie within the 3σ confidence bounds predicted by DR. In particular, VAN recovered a crossing trajectory while the DR estimate consisted of two parallel South/North tracklines. Upon further investigation it became clear that the cause of this discrepancy was a significant nonlinear heading bias present in the AUV's magnetic compass. We used an independently collected data set to calculate a compass bias correction and then applied it to our heading data to produce the results shown in Fig. 17(a) where DR and VAN are now in agreement. Essentially, VAN camera-derived measurements had been good enough to compensate for the large heading bias and still recover a consistent vehicle trajectory despite the unmodeled compass error (recall that in a Kalman update the prior will essentially be ignored if the measurements are very precise).

VII. Conclusion

In conclusion this article presented a systems-level framework for visual navigation termed "visually augmented navigation." VAN's systems-level approach leads to a robust solution that exploits the complementary characteristics of a camera and strap-down sensor suite to overcome the peculiarities of low-overlap underwater imagery. Key strengths of the VAN framework were shown to be:

- *Self-consistency.* Camera measurements forced the VAN trajectory to cross-over despite the presence of an unmodeled compass bias (Fig. 17(c)).
- *Robustness.* Trajectory estimation gracefully handles having a disconnected image topology since the Kalman filter continues to build correlation between camera poses (Fig. 17(b)).
- *Smoothing.* The delayed-state EKF framework means that information from loop-closing events gets distributed throughout the entire map via the joint-correlations (Fig. 18).

- *Time-independent error characteristics.* Uncertainty in a DR system grows monotonically time, while in a VAN approach it is a function of network topology. Essentially, VAN allows error to be a function of space and not time — space being distance away from the reference node in a connected topology (Fig. 15).

This article's goal was to outline our camera/navigation systems-level fusion methodology. We showed that by maintaining a collection of historical vehicles poses, we are able to recursively incorporate pairwise camera constraints derived from low-overlap imagery and fuse them with onboard navigation data. For this purpose, we demonstrated that tracking the mean and covariance statistics of this representation using a standard EKF SLAM approach allows us to exploit state information for image registration including pose-constrained correspondences, link hypothesis, and image-based feature-encoding. Furthermore, we showed that the EKF provides a mechanism for propagating camera information throughout the entire pose-network via the shared correlations.

Despite the advantages of this approach, a well-known point of contention with EKF-based SLAM inference is that it requires quadratic (i.e., $\mathcal{O}(n^2)$) complexity per update to maintain the covariance matrix. Naïvely, this would seem to limit the VAN framework to relatively small environments. In separate publications [30]–[33], we report how to achieve exactly the same state result as the EKF formulation, while alleviating the quadratic computational burden. This is accomplished by recasting the estimation problem within the context of an extended information filter (EIF) (i.e., the dual of the EKF). The implication of this is that we can retain VAN's desirable standalone navigation attributes while exploiting the EIF's sparse representation to achieve large-area scalability on the order of kilometers as demonstrated in [31]–[33].

ACKNOWLEDGMENTS

We graciously thank our collaborators, Prof. Louis Whitcomb and Dr. James Kinsey at Johns Hopkins University, for their help in collecting the JHU data set.

REFERENCES

[1] S. Thrun, D. Hähnel, D. Ferguson, M. Montemerlo, R. Triebel, W. Burgard, C. Baker, Z. Omohundro, S. Thayer, and W. Whittaker, "A system for volumetric robotic mapping of abandoned mines," in *Proc. IEEE Intl. Conf. Robot. Auto.*, vol. 3, Taipei, Taiwan, Sept. 2003, pp. 4270–4275.

[2] J. Crisp, M. Adler, J. Matijevic, S. Squyres, R. Arvidson, and D. Kass, "Mars exploration rover mission," *J. Geophysical Research*, vol. 108, no. E12, pp. ROV 2–1:17, 2003.

[3] E. Allen, "Research submarine ALVIN," in *Proceedings*. U.S. Naval Institute, 1964, pp. 138–140.

[4] J. Donnelly, "1967 — ALVIN's year of science," *Naval Research Reviews*, vol. 21, no. 1, pp. 18–26, 1968.

[5] R. Ballard, "Hydrothermal vent fields of the East Pacific Rise at 21 deg.N, and Galapagos Rift at 86 deg.W," *EOS, Trans. Amer. Geophysical Union*, vol. 60, no. 46, p. 863, 1979.

[6] J. B. Corliss, J. Dymond, L. I. Gordon, J. M. Edmond, R. P. von Herzen, R. D. Ballard, K. Green, D. Williams, A. Bainbridge, K. Crane, and T. H. van Andel, "Submarine thermal springs on the Galapagos Rift," *Science*, vol. 203, no. 4385, pp. 1073–1083, 1979.

[7] R. Ballard, D. Yoerger, W. Stewart, and A. Bowen, "ARGO/JASON: a remotely operated survey and sampling system for full-ocean depth," in *Proc. IEEE/MTS OCEANS Conf. Exhib.*, 1991, pp. 71–75.

[8] D. Yoerger, A. Bradley, B. Walden, M. Cormier, and W. Ryan, "Fine-scale seafloor survey in rugged deep-ocean terrain with an autonomous robot," in *Proc. IEEE Intl. Conf. Robot. Auto.*, vol. 2, San Francisco, CA, USA, Apr. 2000, pp. 1787–1792.

[9] T. Shank, D. Fornari, D. Yoerger, S. Humphris, A. Bradley, S. Hammond, J. Lupton, D. Scheirer, R. Collier, A. Reysenbach, K. Ding, W. Seyfreid, D. Butterfield, E. Olson, M. Lilley, N. Ward, and J. Eisen, "Deep submergence synergy: Alvin and ABE explore the Galapagos rift at 86°W," *EOS, Trans. Amer. Geophysical Union*, vol. 84, no. 41, pp. 425,432–433, Oct. 2003.

[10] J. Kinsey, R. Eustice, and L. Whitcomb, "Underwater vehicle navigation: recent advances and new challenges," in *IFAC Conf. on Manoeuvring and Control of Marine Craft*, Lisbon, Portugal, Sept. 2006, In Press.

[11] B. Hofman Wellenhof, H. Lichtenegger, and J. Collins, *Global positioning system (GPS): theory and practice*, 5th ed. New York: Springer-Verlag, 2001.

[12] W. Stewart, "Remote-sensing issues for intelligent underwater systems," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, Maui, HI, USA, 1991, pp. 230–235.

[13] L. Whitcomb, D. Yoerger, H. Singh, and J. Howland, "Advances in underwater robot vehicles for deep ocean exploration: navigation, control and survery operations," in *Proc. Intl. Symp. Robotics Research*, Snowbird, UT, USA, Oct. 1999, pp. 346–353.

[14] N. Bulusu, D. Estrin, L. Girod, and J. Heidemann, "Scalable coordination for wireless sensor networks: self-configuring localization systems," in *Proc. Intl. Symp. Comm. Theory Apps.*, Ambleside, UK, July 2001.

[15] M. Hunt, W. Marquet, D. Moller, K. Peal, W. Smith, and R. Spindel, "An acoustic navigation system," Woods Hole Oceanographic Institution, Tech. Rep. WHOI-74-6, Dec. 1974.

[16] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Autonomous Robot Vehicles*, I. Cox and G. Wilfong, Eds. Springer-Verlag, 1990, pp. 167–193.

[17] P. Moutarlier and R. Chatila, "An experimental system for incremental environment modeling by an autonomous mobile robot," in *Proc. Intl. Symp. Experimental Robotics*, Montreal, Canada, June 1989.

[18] J. Tardos, J. Neira, P. Newman, and J. Leonard, "Robust mapping and localization in indoor environments using sonar data," *Intl. J. Robotics Research*, vol. 21, no. 4, pp. 311–330, Apr. 2002.

[19] F. Lu and E. Milios, "Globally consistent range scan alignment for environment mapping," *Autonomous Robots*, vol. 4, pp. 333–349, Apr. 1997.

[20] J. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *Proc. IEEE Intl. Symp. Comp. Intell. Robot. Auto.*, Monterey, CA, Nov. 1999, pp. 318–325.

[21] Z. Zhang, "iterative point matching for registration of free-form curves and surfaces,," *Intl. J. Computer Vision*, vol. 13, no. 2, pp. 119–152, Oct. 1994.

[22] A. Bradley, M. Feezor, H. Singh, and F. Sorrell, "Power systems for autonomous underwater vehicles," *IEEE J. Oceanic Eng.*, vol. 26, no. 4, pp. 526–538, Oct. 2001.

[23] H. Singh, R. Eustice, C. Roman, and O. Pizarro, "The SeaBED AUV – a platform for high resolution imaging," in *Unmanned Underwater Vehicle Showcase*, Southampton Oceanography Centre, UK, Sept. 2002.

[24] H. Singh, R. Armstrong, F. Gilbes, R. Eustice, C. Roman, O. Pizarro, and J. Torres, "Imaging coral I: imaging coral habitats with the SeaBED AUV," *J. Subsurface Sensing Tech. Apps.*, vol. 5, no. 1, pp. 25–42, Jan. 2004.

[25] T. Fossen, *Guidance and control of ocean vehicles*. New York: John Wiley and Sons Ltd., 1994.

[26] J. Leonard and R. Rikoski, "Incorporation of delayed decision making into stochastic mapping," in *Experimental Robotics VII*, ser. Lecture Notes in Control and Information Sciences, vol. 271. Springer-Verlag, 2001, pp. 533–542.

[27] S. Fleischer, "Bounded-error vision-based navigation of autonomous underwater vehicles," Ph.D. dissertation, Stanford University, May 2000.

[28] P. McLauchlan, "A batch/recursive algorithm for 3D scene reconstruction," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, vol. 2, Hilton Head, SC, USA, 2000, pp. 738–743.

[29] A. Gelb, Ed., *Applied optimal estimation*. Cambridge, MA: MIT Press, 1982.

[30] R. Eustice, H. Singh, and J. Leonard, "Exactly sparse delayed-state filters," in *Proc. IEEE Intl. Conf. Robot. Auto.*, Barcelona, Spain, 2005, pp. 2417–2424.

[31] R. Eustice, H. Singh, J. Leonard, M. Walter, and R. Ballard, "Visually navigating the RMS Titanic with SLAM information filters," in *Proc. Robotics: Science & Systems*. Cambridge, MA: MIT Press, June 2005, pp. 57–64.

[32] R. M. Eustice, H. Singh, and J. J. Leonard, "Exactly sparse delayed-state filters for view-based SLAM," *IEEE Trans. Robot.*, vol. 22, no. 6, pp. 1100–1114, Dec. 2006.
[33] R. M. Eustice, H. Singh, J. J. Leonard, and M. R. Walter, "Visually mapping the RMS Titanic: conservative covariance estimates for SLAM information filters," *Intl. J. Robotics Research*, vol. 25, no. 12, pp. 1223–1242, 2006.
[34] R. Garcia, J. Puig, P. Ridao, and X. Cufi, "Augmented state Kalman filtering for AUV navigation," in *Proc. IEEE Intl. Conf. Robot. Auto.*, vol. 4, Washington, D.C., May 2002, pp. 4010–4015.
[35] P. V. O'Neil, *Advanced engineering mathematics*, 4th ed. Pacific Grove, CA: Brooks/Cole Publishing Company, 1995.
[36] J. Leonard, R. Rikoski, P. Newman, and M. Bosse, "Mapping partially observable features from multiple uncertain vantage points," *Intl. J. Robotics Research*, vol. 21, no. 10, pp. 943–975, Oct. 2002.
[37] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2000.
[38] O. Faugeras, Q. Luong, and T. Papadopoulu, *The geometry of multiple images*. MIT Press, 2001.
[39] H. Singh, C. Roman, O. Pizarro, R. Eustice, and A. Can, "Towards high-resolution imaging from underwater vehicles," *Intl. J. Robotics Research*, vol. 26, no. 1, pp. 55–74, Jan. 2007.
[40] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conf.*, Manchester, U.K., 1988, pp. 147–151.
[41] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[42] O. Pizarro, "Large scale structure from motion for autonomous underwater vehicle surveys," Ph.D. dissertation, Massachusetts Institute of Technology / Woods Hole Oceaonographic Institution Joint Program, September 2004.
[43] R. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation in an unstructured environment using a delayed state history," in *Proc. IEEE Intl. Conf. Robot. Auto.*, vol. 1, New Orleans, USA, Apr. 2004, pp. 25–32.
[44] P. Rousseeuw and A. Leroy, *Robust regression and outlier detection*. New York: John Wiley and Sons, 1987.
[45] Z. Zhang, "Determining the epipolar geometry and its uncertainty: a review," *Intl. J. Computer Vision*, vol. 27, no. 2, pp. 161–198, 1998.
[46] O. Pizarro, R. Eustice, and H. Singh, "Relative pose estimation for instrumented, calibrated imaging platforms," in *Proc. Digital Image Computing Apps.*, Sydney, Australia, Dec. 2003, pp. 601–612.
[47] B. Horn, "Relative orientation," *Intl. J. Computer Vision*, vol. 4, no. 1, pp. 59–78, Jan. 1990.
[48] J. Heikkilä and O. Silvén, "A four-step camera calibration procedure with implicit image correction," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, Puerto Rico, 1997, pp. 1106–1112.
[49] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
[50] C. Schmid and R. Mohr, "Matching by local invariants," INRIA, Technical Report 2644, Aug. 1995.
[51] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 5, pp. 489–497, May 1990.
[52] F. Badra, A. Qumsieh, and G. Dudek, "Rotation and zooming in image mosaicing," in *Proc. IEEE Workshop on Apps. of Computer Vision*, Princeton, NJ, Oct. 1998, pp. 50–55.
[53] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. European Conf. Computer Vision*, Copenhagen, Denmark, May 2002, pp. 0–7.
[54] F. Schaffalitzky and A. Zisserman, "Viewpoint invariant texture matching and wide baseline stereo," in *Proc. IEEE Intl. Conf. Computer Vision*, Vancouver, BC, July 2001, pp. 636–643.
[55] T. Tuytelaars and L. Van Gool, "Wide baseline stereo matching based on local, affinely invariant regions," in *Proc. British Machine Vision Conf.*, 2000, pp. 412–425.
[56] X. Shen, P. Palmer, P. McLauchlan, and A. Hilton, "Error propogation from camera motion to epipolar constraint," in *Proc. British Machine Vision Conf.*, Sept. 2000, pp. 546–555.
[57] S. Lanser and T. Lengauer, "On the selection of candidates for point and line correspondences," in *Proc. Intl. Symp. Computer Vision*. IEEE Computer Society Press, 1995, pp. 157–162.
[58] R. Eustice, "Large-area visually augmented navigation for autonomous underwater vehicles," Ph.D. dissertation, Massachusetts Institute of Technology / Woods Hole Oceaonographic Institution Joint Program, June 2005.
[59] A. I. Mourikis and S. I. Roumeliotis, "On the treatment of relative-pose measurements for robot localization," in *Proc. IEEE Intl. Conf. Robot. Auto.*, Orlando, FL, May 2006, pp. 2277–2284.
[60] J. Kinsey, D. Smallwood, and L. Whitcomb, "A new hydrodynamics test facility for UUV dynamics and control research," in *Proc. IEEE/MTS OCEANS Conf. Exhib.*, vol. 1, Sept. 2003, pp. 356–361.
[61] D. Smallwood, R. Bachmayer, and L. Whitcomb, "A new remotely operated underwater vehicle for dynamics and control research," in *Proc. Intl. Symp. Unmanned Unteth. Subm. Tech.*, Durham, NH, USA, 1999, pp. 370–377.

Ryan M. 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 Joint Program, Woods Hole, MA, in 2005.

Currently, he is an Assistant Professor with 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.

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

His research is focused on underwater imaging and robotic underwater vehicles. He is currently working on robotic and diver-based optical imaging of coral reef systems as an ARC Australian Postdoctoral Fellow at the Australian Centre for Field Robotics, University of Sydney, Australia.

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, Cambridge, MA/Woods Hole Oceanographic Institution (WHOI), Woods Hole, MA, joint program 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.