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 an effective constraint, 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 Mourtarlier 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-
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.

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\(^\text{1}\) 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 notionally as \( x_{uv} \). 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.
We define the local-level frame, denoted subscript $ℓ$, as the V AN state vector consists entirely of historical trajectory associated with image acquisition. Thus, unlike the typical camera measurements and navigation prior. The global poses are used as constraints in a recursive estimation modulo scale (via calculation of the Essential matrix). These frames. In this approach, the camera provides observation of pairwise measurements from spatially neighboring image sequences. 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 registered imagery results in observation of relative robot motion with respect to a place it has previously visited.

![Diagram of 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.](image)

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 denoted $x_{wv}$. 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 $x_{ℓv}$.

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 $x_{ik} = x_{ij} \oplus x_{jk}$, and $x_{ji} = \ominus x_{ij}$, 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 samples sampled at image acquisition. In our nomenclature these samples are referred 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 sensor 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]. 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, $x_v$, contains both pose, $x_p$, and kinematic terms, $x_κ$, and is defined as

$$x_v = \begin{bmatrix} x_p^T & x_κ^T \end{bmatrix}^T.$$  

Here, $x_p$ 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., $x_p \equiv x_ℓv \equiv [x, y, z, \phi, \theta, \psi]^T$), and $x_κ$ represents any kinematic state elements that are required for propagation of 6\(^2\) 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, \( x_v \sim \mathcal{N}(\mu_v, \Sigma_{vv}) \), with mean and covariance given by

\[
\mu_v = \begin{bmatrix} \mu_v^T, \mu_v^2 \end{bmatrix}^T \quad \text{and} \quad \Sigma_{vv} = \begin{bmatrix} \Sigma_{pp} & \Sigma_{ppc} \\ \Sigma_{pcp} & \Sigma_{cc} \end{bmatrix}.
\]

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

\[
\begin{align*}
\dot{x}_v(t) &= f(x_v(t), u(t), t) + w(t) \\
\dot{z}[t_k] &= h(x_v[t_k], t_k) + \nu[t_k].
\end{align*}
\]

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

**Prediction**

\[
\begin{align*}
\dot{\mu}_v &= f(\mu_v(t), u(t), t) \\
\dot{\Sigma}_{vv}(t) &= F_x \Sigma_{vv}(t) + \Sigma_{vv}(t) F_x^T + Q(t).
\end{align*}
\]

**Update**

\[
\begin{align*}
\mu_v &= \bar{\mu}_v + K(z[t_k] - h(\bar{\mu}_v, t_k)) \\
\Sigma_{vv} &= (I - KH_x) \Sigma_{vv} (I - KH_x)^T + KR_k K^T
\end{align*}
\]

where \( F_x = \frac{\partial f}{\partial x_v} |_{\mu_v} \) and \( H_x = \frac{\partial h}{\partial x_v} |_{\mu_v} \) are the process and observation model Jacobians, respectively. In this formulation, the predicted vehicle distribution, \( x_v \sim \mathcal{N}(\mu_v, \Sigma_{vv}) \), 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, \( x_v \), 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, \( x_p \). 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 \( p(\xi_t | z^t, u^t) \) where \( z^t \) represents all measurements up to time \( t \) (including camera and navigation sensors), \( u^t \) is the set of all control inputs, and \( \xi_t \) 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, \( \xi_t \), to include the vehicle’s pose of where it was when it acquired that image (i.e., \( \xi_t = [x_v^T, x_p^T]^T \)). Therefore, at this time instance the augmented state distribution, \( \xi_t \sim \mathcal{N}(\mu_t, \Sigma_t) \), is given by

\[
\begin{align*}
\mu_t &= [\mu_v[t_1]^T, \mu_p[t_1]^T]^T = [\mu_v, \mu_p]^T \\
\Sigma_t &= \begin{bmatrix} \Sigma_{vv} & \Sigma_{vp}^T \\ \Sigma_{pv} & \Sigma_{pp} \end{bmatrix} = \begin{bmatrix} \Sigma_{vv} & \Sigma_{vp1} & \cdots & \Sigma_{vpn} \\ \Sigma_{pv1} & \Sigma_{pp1} & \cdots & \Sigma_{ppn} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{vpn} & \Sigma_{pvn1} & \cdots & \Sigma_{pvn} \end{bmatrix},
\end{align*}
\]

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 \( \xi_t = [x_v^T, x_p^T, \ldots, x_p^n]^T \) with

\[
\begin{align*}
\mu_t &= \begin{bmatrix} \mu_v \\ \mu_p \end{bmatrix} \\
\Sigma_t &= \begin{bmatrix} \Sigma_{vv} & \Sigma_{vp1} & \cdots & \Sigma_{vpn} \\ \Sigma_{pv1} & \Sigma_{pp1} & \cdots & \Sigma_{ppn} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{vpn} & \Sigma_{pvn1} & \cdots & \Sigma_{pvn} \end{bmatrix}
\end{align*}
\]

Note that in (4) the vehicle’s current pose, \( x_p \), is fully correlated with \( x_v \) by definition. Therefore, when the \( n \)-th delayed-state, \( x_{pn} \), is augmented in (5), its cross-correlation with the other delayed-states in \( \Sigma_t \) 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 \( x_v \) continue to evolve through the vehicle dynamic model, \( f(\cdot, t) \), while the delayed-state entries do not:

\[
\dot{\xi}_t = \frac{d}{dt} \begin{bmatrix} x_v \\ x_{p1} \\ \vdots \\ x_{pn} \end{bmatrix} = \begin{bmatrix} f(x_v(t), u(t), t) + w(t) \\ \vdots \\ \vdots \end{bmatrix} \quad 0_{m \times 1}
\]

Similarly, navigation sensor observation models continue to remain a function of only the current vehicle state, \( x_v \), which results in sparse Jacobians of the form

\[
H_x = \begin{bmatrix} H_{x_v} & 0_{m \times 6} & \cdots & 0_{m \times 6} \end{bmatrix},
\]

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 \( x_{pi} \), \( x_{pj} \), 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, \( x_{vc} \), is known.

1) Camera Relative Pose: The delayed-state entries \( x_{pi} \) and \( x_{pj} \) correspond to vehicle poses \( x_{vc_i} \) and \( x_{vc_j} \), respectively, as represented in the local-level navigation frame defined in \( \S III \). Hence, using the extrinsic camera to vehicle pose, \( x_{vc} \), we can express the transformation from camera frame \( i \) to \( j \) using the tail-to-tail operation as

\[
\begin{align*}
x_{c_jc_i} &= x_{c_jc} + x_{c_i} \\
&= (x_{c_jc} + x_{vc}) + (x_{c_ic} + x_{vc}), \quad (6a)
\end{align*}
\]
from a pressure sensor, and (ii) error measurements of depth variations ($z$) are implicitly fixed within the filter by two sources: pose observation means in terms of reducing navigation error. 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.

with Jacobian
\[
J_{c_j} = \begin{bmatrix}
\frac{\partial x_{c_j}}{\partial (x_{t_{ij}}, x_{t_{ij}})} \\
\frac{\partial (x_{c_j}, x_{t_{ij}})}{\partial (x_{t_{ij}}, x_{t_{ij}})}
\end{bmatrix}
\]

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 $\alpha_{ji}$ and $\beta_{ji}$, respectively, is recoverable from image space. Realizing that the relative-pose measurement $x_{c_j}$ is parameterized by

\[
x_{c_j} = \begin{bmatrix}
\alpha_{ji} \\
\beta_{ji} \\
n_j
\end{bmatrix}
\]

we can express the bearing-only baseline measurement of $\alpha_{ji}$ and $\beta_{ji}$ as

\[
\alpha_{ji} = \tan^{-1} \left( \frac{y_{ji}}{x_{ji}} \right), \\
\beta_{ji} = \tan^{-1} \left( \frac{z_{ji}}{\sqrt{x_{ji}^2 + y_{ji}^2}} \right),
\]

with Jacobian

\[
J_{\alpha\beta} = \begin{bmatrix}
\frac{\partial (\alpha_{ji}, \beta_{ji})}{\partial (x_{c_j}, x_{c_j})} \\
\frac{\partial (x_{c_j}, x_{c_j})}{\partial (x_{c_j}, x_{c_j})}
\end{bmatrix}
\]

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 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-validation in bearing-only SLAM applications [36].

C. Link Hypothesis

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, $j t_{ji}$, 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.

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

\[
z_{ji} = h_{ji}(x_{c_j}) = [\alpha_{ji}, \beta_{ji}, \phi_{ji}, \theta_{ji}, \psi_{ji}]^T,
\]

with Jacobian

\[
H_{\xi} = \begin{bmatrix}
0 & \cdots & \frac{\partial h_{ji}}{\partial x_{c_j}} & \cdots & \frac{\partial h_{ji}}{\partial x_{c_j}} & \cdots & 0
\end{bmatrix},
\]

where

\[
\frac{\partial h_{ji}}{\partial (x_{c_j}, x_{c_j})} = \frac{\partial (x_{c_j}, x_{c_j})}{\partial (x_{c_j}, x_{c_j})} = \begin{bmatrix}
J_{\alpha\beta} & \theta_{ji} & \phi_{ji} & \psi_{ji} & \alpha_{ji} & \beta_{ji}
\end{bmatrix}
\]

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.
cent overlap, $\epsilon$, can be defined as

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

Here, $d_{ij}$ is the Euclidean distance between the camera centers, $W_{\text{max}} = 2A_{\text{max}} \tan(\frac{1}{2}\text{FOV})$ is the 1D image width, $A_{\text{max}}$ 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.

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.

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, $I_r$.

1: define: $k \{\text{maximum number of candidates to return}\}$
2: define: $\epsilon_{\text{min}} \in [0, 1] \{\text{minimum percent overlap}\}$
3: define: $\epsilon_{\text{max}} \in [0, 1] \{\text{maximum percent overlap}\}$
4: define: $\alpha \in [0, 1] \{\text{confidence-level}\}$
5: for all $I_i$ do
6: $A_{\text{max}} \leftarrow \max(A_i, A_r)$
7: $W_{\text{max}} \leftarrow 2A_{\text{max}} \tan(\frac{1}{2}\text{FOV})$
8: $d_{\text{min}} \leftarrow (1 - \epsilon_{\text{min}}) \cdot W_{\text{max}}$
9: $d_{\text{max}} \leftarrow (1 - \epsilon_{\text{max}}) \cdot W_{\text{max}}$
10: extract from our state, $\xi_t$, the joint-marginal:
11: compute the relative camera pose, $x_{r\rightarrow c}$, and its first-order statistics (6),(7)
12: using $x_{r\rightarrow c}$, compute the Euclidean distance $d_{ri}$ and its first-order statistics:
13: compute the probability $P_t$ that $d_{\text{min}} < d_{ri} < d_{\text{max}}$:
14: if $P_t > \alpha$ then
15: add $I_i$ to the candidate set $S$
16: end if
17: end for
18: sort candidate set $S$ by $P_t$ and return up to the $k$ most probable candidates

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 Papadopoulos [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
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).
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 $3 \times 4$ projection matrix $P$ defined as

$$P = [w]_{cu}R [cw]_{t_{cw}}.$$  

Here, $R$ and $[cw]_{t_{cw}}$ encode the coordinate transformation from world, $w$, to camera centered coordinate frame, $c$, and

$$K = \begin{bmatrix} c_{uu} & u_0 & 0 \\ 0 & c_{vv} & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

is the known $3 \times 3$ upper triangular intrinsic camera calibration matrix with $c_{uu}$, $c_{vv}$ the pixel focal lengths in the $x$, $y$ directions, respectively, $(u_0, v_0)$ 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

$$\mathbf{u} = P \mathbf{X}$$  \hspace{1cm} (9)$$

where $\mathbf{u} = [u, v]^T$ is the vector description of $(u, v)$, $\mathbf{u} = [u^T, 1]^T$ its normalized homogeneous representation, $\mathbf{X} = [x, y, z]^T$ is the imaged 3D scene point, and $\mathbf{X} = [x^T, 1]^T$ 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:

$$\mathbf{x} = K^{-1} \mathbf{u} = [c]_{wu}R [cw]_{t_{cw}} \mathbf{X}.$$  

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 Heikkila’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 $x_{cc_i}$ to normalize the feature regions around the Harris interest points in image $I_i$ via the infinite homography:

$$H_\infty = K_{ci}^\dagger R K^{-1}.$$  

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

We compute $H_\infty$ 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 correction 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].

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 Lanzar’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 $P = K[I\mid 0]$ and $P' = K[R\mid t]$, where for notational convenience we simply write the relative-pose parameters as $R, t$. 

EUSTICE et al.: VISUALLY AUGMENTED NAVIGATION FOR AUTONOMOUS UNDERWATER VEHICLES
We begin by noting that the scene point $X$ is projected through camera $P$ as

$$
u = P X = K X,$$

which implies that explicitly accounting for scale we have

$$X = Z K^{-1} \nu. \quad (10)$$

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

$$u' = P' X = K (RX + t). \quad (11)$$

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

$$u' = KRK^{-1} \nu + Kt/Z. \quad (12)$$

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

$$u' = \frac{H\nu + Kt/Z}{H^3\nu + tZ/Z} \quad (13)$$

where $H = KRK^{-1}$, $H^3$ refers to the third row of $H$, and $tZ$ is the third element of $t$.

When the scene depth $Z$ of the image point $\nu$ 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., $u' = f(\nu, Z)$) 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 $\xi$. We begin by defining the parameter vector, $\gamma$, as

$$\gamma = [x_{p_i}^T, x_{p_j}^T, Z, u, v]^T \quad (14)$$

with mean, $\mu_\gamma$, and covariance, $\Sigma_\gamma$, given by

$$\mu_\gamma = \begin{bmatrix} \mu_{p_i} \\ \mu_{p_j} \\ Z \\ u \\ v \end{bmatrix}, \quad \Sigma_\gamma = \begin{bmatrix} \Sigma_{p_i,p_i} & \Sigma_{p_i,p_j} & 0 & 0 & 0 \\ \Sigma_{p_j,p_i} & \Sigma_{p_j,p_j} & 0 & 0 & 0 \\ 0 & 0 & \sigma_Z^2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Here, $x_{p_i}$, $x_{p_j}$ are the delayed-state vehicle poses extracted from $\xi_i$ (used to calculate relative camera pose according to (6)), $Z$ and $\sigma_Z$ 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 $\Sigma_\gamma$ we employ the standard assumption that features are extracted with isotropic, independent, unit variance noise [37] when defining the $\Sigma_{uv}$ sub-block. To obtain a first-order estimate of the uncertainty in the point transfer mapping between $I_i$ and $I_j$ we compute

$$\begin{align*}
\mu_{\nu'} & \approx (13)|_{\mu_\gamma} \quad (15) \\
\Sigma_{\nu'} & \approx J_{\Sigma_\gamma} J_{\Sigma_\gamma}^T \quad (16)
\end{align*}$$

where $\mu_{\nu'}$ is the predicted point location of $\nu$ in $I_j$, $\Sigma_{\nu'}$ its covariance, and $J = \frac{\partial \gamma}{\partial \nu}$ is the point transfer Jacobian.

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

$$\left(\nu' - \mu_{\nu'}\right)^T \Sigma_{\nu'}^{-1} \left(\nu' - \mu_{\nu'}\right) = k^2 \quad (17)$$

where the threshold $k^2$ follows a $\chi^2_2$ 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 $\nu$. 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, $\sigma_Z$, as a convenient parameterization for controlling the size of the search regions in $I_j$ through the definition of the parameter vector $\gamma$. For example, in the case where no a priori knowledge of scene depth is available, choosing any finite value for $Z$ and setting $\sigma_Z \rightarrow \infty$ recovers a search band along the epipolar line in $I_j$ whose width corresponds to the uncertainty in relative camera pose, $x_{\gamma,c_i}$ (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 $\sigma_Z$ 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

---

3We compute this Jacobian numerically as described in [37, §A4.2].
out of practicality we assume relative-pose measurements framework. Thus, like other scan-matching algorithms [19], correlations quickly becomes intractable in any scan-matching. Unfortunately, actually computing all possible measurement z

I

and

measurements

z

I

↔

jk

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 \( z_{ij} \) and \( z_{jk} \) could be statistically correlated. Neglecting such a correlation would put too much weight on the filter update step since it would treat observations \( z_{ij} \) and \( z_{jk} \) 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., \( F_{ij} \cap F_{jk} = \emptyset \)), the corresponding camera measurements \( z_{ij} \) and \( z_{jk} \) 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.

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 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.

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.

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.
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-noise ratio in shadowed regions so that fully automatic cross-track correspondence could be achieved.

![Image](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.)

![Image](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.)
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 10th 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.

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 \( XY \) top pose, 3\( \sigma \) 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 m long and spaced 1.5 m apart while maintaining an average altitude of 3.0 m above the seafloor at a forward velocity of 0.35 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 3\( \sigma \) 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).

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 $3\sigma$ 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 $3\sigma$ VAN estimate (blue). This discrepancy comes from an unmodeled compass bias, which when accounted for, produces the results shown in (a).
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 in 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., $O(n^2)$) complexity per update to maintain the covariance matrix. Naively, 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**


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.