Stereovision Imaging on Submersible Platforms for 3-D Mapping of Benthic Habitats and Sea-Floor Structures

Shahriar Negahdaripour, *Senior Member, IEEE*, and Hossein Madjidi

***Abstract*—We investigate the deployment of a submersible platform with stereovision imaging capability for three-dimensional (3–D) mapping of benthic habitats and other sea-floor structures over local areas. A complete framework is studied, comprising 1) suitable trajectories to be executed for data collection; 2) data processing for positioning and trajectory followed by online frame-to-frame and frame-to-mosaic registration of images, as well as recursive global realignment of positions along the path; and 3) 3-D mapping by the fusion of various visual cues, including motion and stereo within a Kalman filter. The computational requirements of the system are evaluated, formalizing how processing may be achieved in real time.**

The proposed scenario is simulated for testing with known ground truth to assess the system performance, to quantify various errors, and to identify how performance may be improved. Experiments with underwater images are also presented to verify the performance of various components and the overall scheme.

***Index Terms*—Automated ROV visual navigation, stereovision, underwater 3-D mapping.**

I. Introduction

EQUIPPED with acoustic, laser, and/or optical sensors, unmanned submersible systems—remotely operated vehicles (ROV) and autonomous underwater vehicles (AUV)—are currently the most advanced platforms for automated monitoring and surveying of deep-sea environments. While acoustic systems including side-scan and forward-sector-scan, are widely used for evaluating the large-scale natural benthic habitat characteristics over extended regions, optical imaging is better suited for examining fine-scale community characteristics over local areas in ocean exploration and environmental studies, for monitoring and surveying, as well as for inspection. In particular, two-dimensional (2-D) and three-dimensional (3-D) high-resolution mapping, preferably based on online processing, are two key capabilities to benefit from the high data rate and resolution of optical imaging systems. Just as for various terrestrial applications, photo mosaicking to construct composite images of the sea-floor environments and/or 2-D visual maps for navigation has been investigated extensively in recent years, e.g., [17], [18], [21], [22], [36], [45], [52], [65], and [71]. Furthermore, there has been growing interest in exploring the use of video, potentially with other sensors, in the automatic control, positioning, or navigation of an unmanned underwater vehicle [18], [21], [22], [30], [34], [37], [41]–[43], [46], [48], [73].

Stereo imaging is a classic technique that has been utilized commonly for 3-D reconstruction in industrial robotics manufacturing systems [14], [19], [24], [67], [69], as well as for obstacle avoidance and 3-D mapping in the deployment of autonomous terrestrial and space mobile platforms [2], [12], [26], [38], [50], [59], [60], [62], [74]. However, only a small fraction of submersibles are equipped with stereo cameras, and those are utilized solely for qualitative 3-D visualization with no quantitative 3-D mapping capability (e.g., [63] and [70]). For example, stereo cameras of NASA's Phantom S2 ROV[1] have been mounted on a pan-and-tilt platform that can be controlled by human head motions via a head tracker. This provides the operator/scientist with the feeling of immersion in the remote undersea environment when visualizing the images on a stereoscopic display [63].

The capabilities of an underwater stereovision 3-D mapping system and assessment of performance varies with the range and/or coverage area, as is evident from the related work in the terrestrial domain. At one extreme, the existing solutions for industrial settings provide the natural model of a system to be developed for 3-D reconstruction of single objects and/or inspection and repair of undersea structures. At the other end rests the long-to-medium-distance navigation capability, with comparable specifications and performance to vision-based airborne systems and long-range autonomous outdoor mobile platforms (e.g., [8], [10], [20], [56], and [64]), where low-to-medium-range (20–200 KHz) and medium-to-high-range (0.2–2 MHz) acoustic-based positioning and mapping systems will undoubtedly continue to provide far superior performance. However, a wide variety of scientific work in marine geology, archeology, fisheries, underwater forensics, search and rescue, and commercial ROV/AUV deployment (e.g., microbathymetric mapping of thermal vents, coral reefs and fish habitats, and reconstruction of shipwrecks and crashed airplanes) fall within the intermediate range and resolution requirements, most closely comparable to the desired capabilities of autonomous planetary rovers, e.g., [31] and [50]. Roughly, it would be desired to automatically

Manuscript received March 2002; revised December 2002. This work was supported by the National Science Foundation under Grant BES-9711528. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The authors are with the Electrical and Computer Engineering Department, University of Miami, Coral Gables, FL 33146 USA (e-mail: shahriar@miami.edu; hossein@miami.edu).

Digital Object Identifier 10.1109/JOE.2003.819313

[1]Trademark of Deep Ocean Engineering, San Leandro, CA.

construct 3-D maps with geo-referenced spatial precision in the order of decimeter(s) (or better) and accuracy of centimeter(s) (or better) in constructing smaller objects and targets of interest, within a coverage area of 10s of meters in each dimension.[2] The main objective of this paper is to explore stereovision imaging for the development of *online* mid-range 3-D mapping capability, which covers the need for these applications of interest.

In the absence of existing underwater stereovision mapping systems and any previous standards, we use a coverage area of about 100×100 ft as a baseline to evaluate performance within the proposed framework, thus allowing assessment and comparison with both alternative technologies and future performance enhancements. We address a number of issues for both mapping and positioning. In particular, we investigate the realization of the desired capability in totality, comprising the planning of multitrack trajectories to be executed for data acquisition, incremental position estimation along each track followed by the online application of a recursive global-alignment strategy to overcome position drift errors upon completion of a track, and the reconstruction of a 3-D topographical map at the end of the trajectory, by utilizing a Kalman filter-based framework proposed in earlier work [29]. Last but not least, it is shown how online processing performance may be achieved by studying the computational requirements of the system.

A. *Relevant Work and Overview*

Stereovision systems have been extensively studied for various terrestrial and planetary applications, including object reconstruction in industrial robotics systems [14], [24]; indoor and outdoor mobile robotic navigation [2], [12], [60], [62], [74]; autonomous vehicles for highway and offroad navigation [6], [11], [55]; and planetary rovers [26], [38], [50], [59]. Most outdoor robotics and autonomous systems employ stereovision for collision avoidance and obstacle mapping, although the Mars rover is being targeted for planetary mapping. The arrangement of the cameras is commonly in a forward-look, instead of a down-look, for most underwater applications, a primary distinction that leads to some different problems in terms of mapping strategies, processing, and achievable accuracy over different parts of the image. In recent years, the development of panoramic imaging systems has opened up interesting applications for stereovision and visualization [49], [75]. In particular, it has been shown how stereo data may be constructed from the images obtained by a single rotating camera, enabling 3-D visualization and mapping [51].

Although only one camera is used, photogrammetry techniques for terrain mapping comprise the most relevant work to the application of stereovision systems underwater [3], [40]. Here, images taken from an aircraft at different positions along a flyover path are registered by matching targets at known coordinates on the ground. This allows the construction of photo mosaics, computation of the camera position between consecutive frames (equivalent to stereo baseline), and computation of terrain height (range) from the image disparity. Mimicking this scenario, the state of the art in marine archeology and biology studies, small grids of a few feet in dimension are laid on the ocean floor by divers. Natural objects within the grid are mapped [15], [53]. The grid is then moved and the process is repeated. A more advanced and efficient process is to lay down a network of "opti-acoustic" transponders over a larger area, say the proposed 100×100 ft section. Once the system is calibrated by acoustic positioning methods [7], [8], [10], [20], they can serve as visual beacons at known positions. Clearly, the deployment of optical or, similarly, laser scanning systems [27] for mapping is primarily to enable exploiting very-high-resolution data that is currently unavailable by acoustic imaging [61]. The main disadvantage in employing the proposed technology is the complexity of the task, which although practical for some applications, certainly prohibits day-to-day operations. Deployment of a stereovision system and reducing the number of necessary beacons down to a few (optimally zero, as an ultimate goal) for high-precision 3-D mapping is the primary motivation of this investigation. In particular, our results enable performance quantification in terms of 3-D mapping accuracy and vehicle positioning and path estimation; more generally, they highlight various system bottlenecks. These findings provide valuable information toward the realization of a new technology that would address the critical needs in a large variety of ocean-science and exploration applications.

In acquiring stereo images to map a sea-floor section, a suitable vehicle trajectory over the terrain should be chosen. During mission execution, it is important to ensure that the sensor platform maintains the desired path as closely as possible. High-frequency acoustic positioning, including ultra-short-baseline systems, can be utilized to achieve this objective. Once the images are acquired, they can be processed offline to reconstruct the terrain. The accuracy of current acoustic positioning systems is inadequate to register (mosaic) images at video resolution or to align individual 3-D maps from stereo pairs in order to construct a composite topographical map over the entire coverage area (without substantially compromising the registration precision and the reconstruction quality). However, the proposed strategy is still suitable, where one exploits the achievable precision of high-frequency acoustic positioning systems to ensure path following during data acquisition stage, to roughly register the images or 3-D maps and to potentially integrate with information based on stereo video processing to improve the final trajectory accuracy. A key engineering issue in combining information from video and acoustic data is to calibrate the two systems accurately enough so that the two measurements can be fused and integrated properly in a common coordinate system. In our proposed framework, we instead explore the direct use of video information online for trajectory following, although it becomes evident how the inevitable shortcomings can be overcome by utilizing acoustic-based systems.

In earlier underwater vision work based on a single down-look camera, simplistic assumptions have been made about the sea-floor structure, namely that it is relatively flat (e.g., [22] and [46]), in order to estimate the vehicle motion and trajectory. This assumption typically yields reasonable estimates over a small area and/or where variations in surface relief are negligible relative to the distance to the terrain.

[2]One cannot dispute the fact that "larger is better." However, there is clearly a tradeoff between the size of the coverage area and the achievable accuracy.

Furthermore, it typically works well for the construction of 2-D photo mosaics of the sea floor, which may potentially provide much of the sought-after information for scientific studies. Furthermore, the scale-factor ambiguity of monocular motion vision can be overcome by knowing the initial altitude or by deploying laser-mounted cameras [5], [13]. However, the desired levels of accuracy over the proposed coverage area of 100×100 ft cannot be achieved where the sea floor is not sufficiently flat. The stereo disparity provides strong cues not only for 3-D depth reconstruction, but also for the estimation of vehicle motion and trajectory. This enables us to overcome a serious difficulty in visual motion estimation, namely the nonlinear decoupling between the unknown 3-D motion and depth (which can alternatively be addressed by assuming a flat terrain).

Online processing demands near-real-time performance in computing the vehicle trajectory. Sequential causal processing based on frame-to-frame (F2F) registration are the simplest to implement and most suitable for real-time performance [71], but are also most sensitive to error accumulation. Aside from the integration with other navigation sensors or the placement and deployment of visual targets at known positions (in analogy with acoustic beacons), a mosaic-based positioning (MBP) paradigm is a strategy for bounding the drift error [48], [54]. Here, the trajectory estimation and 2-D photo mosaicking are carried out in a coupled process in the same spirit as the concurrent mapping and localization paradigm [9], [57], [58]. To explain briefly, integration of estimated motions based on F2F alignment is used to determine the new vehicle position after several, say N, frames. The last image is then transformed to the mosaic-coordinate system based on the estimated position. Registration with the mosaic is carried out to compute the misalignment error, the vehicle position is revised, and, consequently, the mosaic is updated with the new image. Noncausal processing, utilizing global constraints and data redundancy over a large number of viewing positions, enables devising more robust techniques, as utilized for photo mosaicking [54] and for photo mosaicking and trajectory estimation in sea-floor 2-D mapping and map-based local navigation [22]. In these earlier works, planar projection or homography models have been utilized to simplify the problem when dealing with monocular cameras and solely exploiting visual motion cues [23] (see Section II-A). As it becomes more clear, information from stereo disparity allows us to simplify the motion-estimation process without any simplistic assumptions about the seafloor topography. Furthermore, we are able to carry out a recursive alignment process to improve the trajectory-estimation process for any arbitrary scene topography. Finally, the most effective approach to reset the drift error is to plan looping trajectories, crossing over at the start point [17]. Here, each loop may be treated and processed independently of others, with no performance degradation from drift errors in other loops.

In the balance of the paper, we review some preliminary background in Section II. We describe the proposed stereo vision-imaging and mapping framework is Section III and we cover the computational requirements in Section IV. Experiments with ground truth and two controlled water-tank experiments are given in Section V and variations in the implementation of the stereovision mapping system to realize modular vision-based capabilities are covered in Section VI. A summary of contributions and future work is given in Section VII.

II. Background

A. Image Motion

Longuet–Higgins [32] formulates the displacement $\boldsymbol{v} = [u, v]$ of image points $[x, y]$ from frame k to $k+1$, induced by the rigid-body motion of a camera relative to a stationary scene

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} xy & -(1+x^2) & y \\ (1+y^2) & -xy & -x \end{bmatrix} \boldsymbol{\omega} + \frac{1}{Z} \begin{bmatrix} -1 & 0 & x \\ 0 & -1 & y \end{bmatrix}$$
$$\boldsymbol{v} = \boldsymbol{A}_w \boldsymbol{\omega} + \frac{1}{Z} \boldsymbol{A}_t \boldsymbol{t} \qquad (1)$$

where $\boldsymbol{\omega}$ and $\boldsymbol{t}$ are the 3-D vectors of the rotational and translational velocities of the camera and *depth* Z is the distance along the optical axis to a point on the scene surface (e.g., altitude of the camera from the sea floor). For small motions (about one pixel at the processing resolution), this constraint can theoretically be utilized to determine the motion $\boldsymbol{m} = [\boldsymbol{\omega}, \boldsymbol{t}]$ and/or depth map Z from the knowledge of image motion $\boldsymbol{v}$ over (the relevant part of) the image. In practice, the general motion problem—determining the unknown camera motion $\boldsymbol{m}$ and depth Z of each surface point—is difficult to solve accurately for a number of reasons, including the ill-conditioned nature of the problem [1]. If the motion is known (measured), the image motion $\boldsymbol{v}$ provides two constraints on depth Z, which is the motion cue (observation) for 3-D depth recovery exploited in our Kalman filter-based mapping system [29]. If the scene is relatively flat, Z may be assumed constant, which is absorbed in $\boldsymbol{t}$, and the problem simplifies drastically; only six unknown motion parameters have to be estimated from the image-motion measurements over the image. (In practice, the well-know inherent translation-rotation ambiguity in the computation of all 6 degrees of freedom is most severe when dealing with a relatively flat terrain [1].) When the surface is a tilted plane at orientation $\boldsymbol{n}$ relative to the camera, the problem can also be solved in closed form [33]; variation of this closed-form solution, when the motion can be larger than a few pixels [23], [66], has been utilized for application to vehicle-trajectory estimation and sea-floor mosaicking [21], [22].

If stereo images are acquired at frame time k, then the scene depth map Z can be determined from disparity cues. Consequently, knowledge of optical flow $\boldsymbol{v}$ and depth Z can be exploited to determine the motion $\boldsymbol{m}$ and simplistic assumptions about the scene structure are no longer necessary [4]. For example, a closed-form solution based on a least-square formulation can be written as

$$\begin{bmatrix} \boldsymbol{\omega} \\ \boldsymbol{t} \end{bmatrix} = \sum_{\text{points}} \begin{bmatrix} \boldsymbol{A}_w^T \boldsymbol{A}_w & \frac{1}{Z} \boldsymbol{A}_w^T \boldsymbol{A}_t \\ \frac{1}{Z} \boldsymbol{A}_t^T \boldsymbol{A}_w & \frac{1}{Z^2} \boldsymbol{A}_t^T \boldsymbol{A}_t \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{A}_w^T \boldsymbol{v} \\ \frac{1}{Z} \boldsymbol{A}_t^T \boldsymbol{v} \end{bmatrix}. \qquad (2)$$

As a result, disparity cues from stereo pairs not only provide information to determine the scene structure for 3-D mapping, but also significantly simplify an otherwise difficult motion

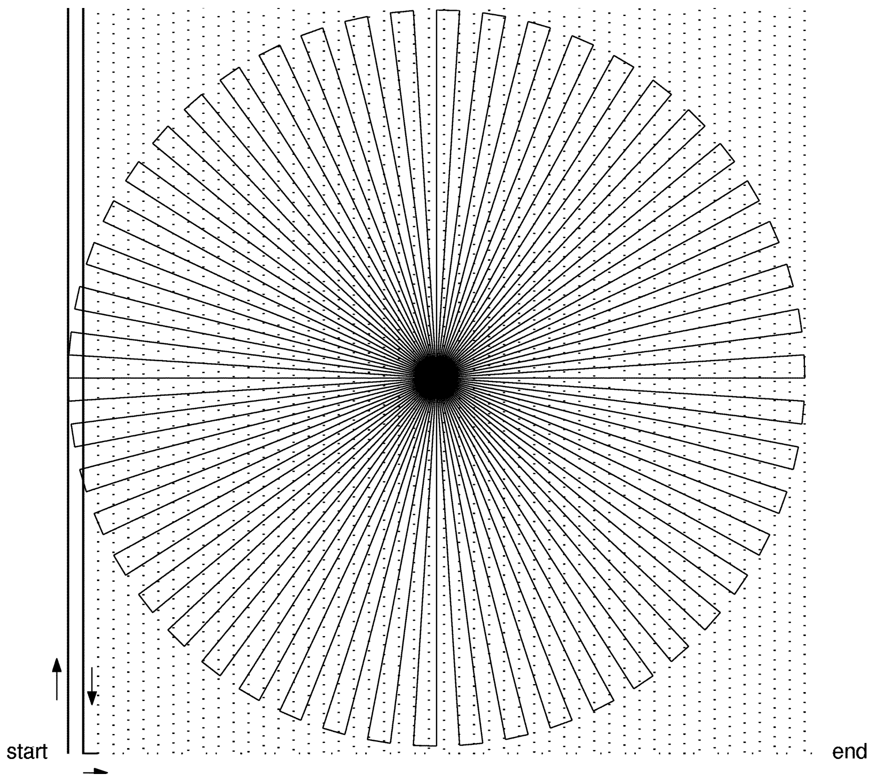


Fig. 1. Two possible trajectories to map an area, with parallel swaths or crossing loops.

problem when dealing with complex topographies. The two advantages are 1) to enhance motion-estimation accuracy and to deal with any type of scene topography and 2) to integrate the stereo and motion cues (within a Kalman filter). These are exploited in developing a stereovision-based 3-D positioning and mapping system, as discussed in details in the balance of the paper. We also explain how we would utilize and extend the existing capabilities of our vision-based ROV—namely, optical station keeping and mosaic-based positioning—for navigation and trajectory following [43], [46], [73].

B. Homogeneous Transformations

Given the motion between two nearby camera positions $\boldsymbol{P}_k$ and $\boldsymbol{P}_{k+1}$ in terms of the rotational and translational components $\boldsymbol{m} = [\boldsymbol{\omega}, \boldsymbol{t}]$, the camera-coordinate systems at these positions are related by the 4×4 homogenous-transformation matrix $\mathbf{H}_{k,k+1}$

$$\mathbf{H}_{k,k+1} = \begin{bmatrix} \boldsymbol{R}_{3\times 3} & \boldsymbol{t} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix} \tag{3}$$

where the 3×3 rotation matrix $\boldsymbol{R}$ is determined from $\boldsymbol{\omega}$, according to the Rodriguez formula. Homogeneous transformation at camera position $\boldsymbol{P}_{k+1}$ with respect to the reference frame at position $\boldsymbol{P}_0$ can be obtained from

$$\mathbf{H}_{0,k+1} = \mathbf{H}_{0,k}\mathbf{H}_{k,k+1} \tag{4}$$

allowing us to determine the motion from reference position $\boldsymbol{P}_0$ to $\boldsymbol{P}_{k+1}$

$$\mathbf{H}_{0,k+1} = \begin{bmatrix} \boldsymbol{R}_{0,k+1} & \boldsymbol{t}_{0,k+1} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}. \tag{5}$$

These also show how interframe motions are integrated to determine the current position.

III. Proposed Framework

The specific mapping framework described here attempts to exploit the strength of various strategies. First, to follow a path that consists of several tracks, the discrepancy between the planned and executed trajectories along each track is minimized by the application of the MBP paradigm [48], which utilizes the F2F and frame-to-mosaic (F2M) registration [44], [54]. Similar in concept to the application for photo mosaicking [22], [54], but much more general for dealing with arbitrary topographies, a global-alignment technique is then applied at the end of each track to correct the integrated residual errors. Here, we employ redundant constraints and independent measurements among neighboring key positions on a predefined network, referred to as the *key-points network*, which is formally defined in Section III-B. The primary advantage is to achieve accurate vehicle localization and to closely follow the planned trajectory, leading to the acquisition of data with the desired overlaps among the images from various swaths. Finally, the estimated trajectory, the optical flow information for pairs of consecutive key frames, stereo disparity, and error covariance of various estimates are integrated to reconstruct a composite 3-D map by utilizing the Kalman filter formulation proposed in earlier work [29].

The process can be summarized in the following algorithm.

1. Plan a trajectory with L loops or tracks, each comprising of P *key points*

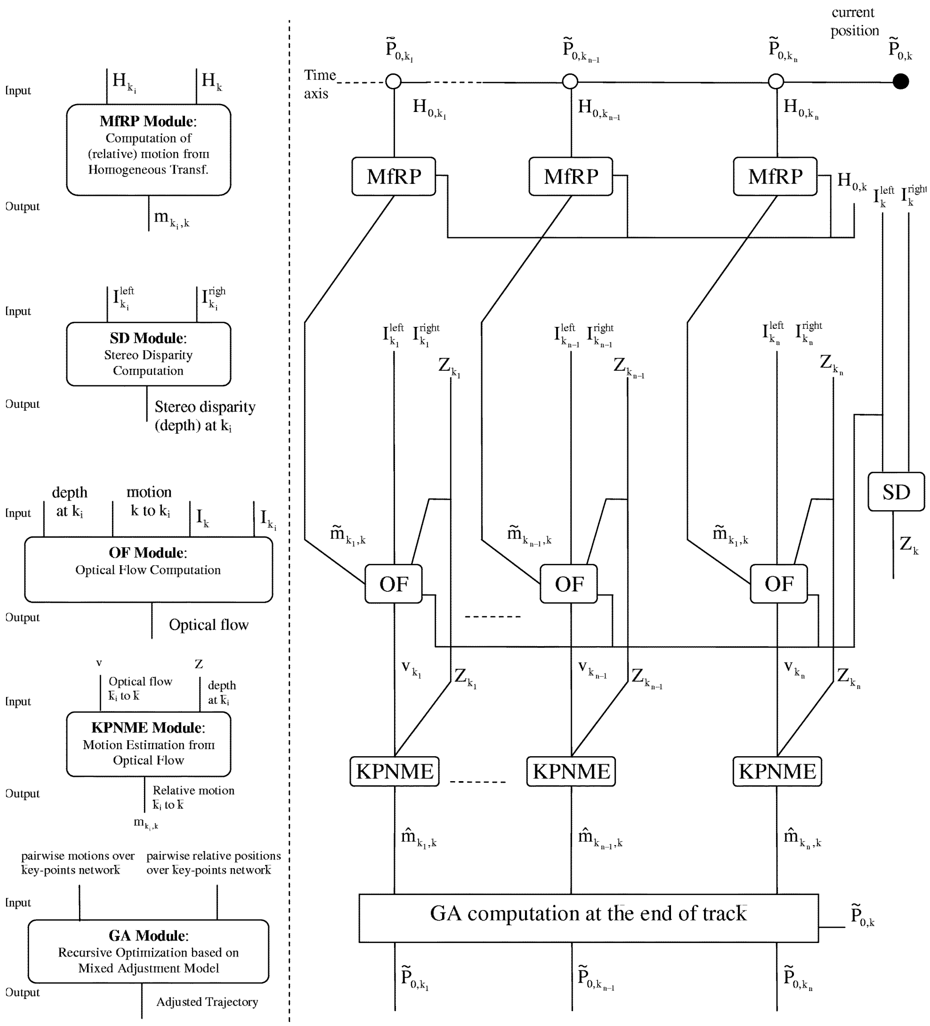


Fig. 2. Left: Input and output of each computational module. Right: Various computations that are carried out by processing units motion from relative position (MfRP), stereo disparity (SD), optical flow (OF), and key-points network motion estimation (KPNME), once a key point P_k is visited.

or *positions* at regular distances, depending on the desired altitude, vehicle speed, camera viewing angle and resolution, and the size of the mapping area.
2. For each track $i = 1 : L$.
- Do trajectory execution and following, as well as stereo data recording (and subsequently) processing at key positions.
- Between two consecutive key positions, integrate N motions computed from monocular frames I_i and I_{i+1} (e.g., the left or right sequence).
- At the next key position, do position correction by registering the current frame I_{i+1} with mosaic M_{i+1} and then updating the mosaic (2-D mosaicking and MBP with every N frames between two consecutive key points).
- Continue if not at the end of track i. Otherwise, go into station-keeping mode while performing a recursive global alignment for trajectory adjustment.

3. If not at the end of trajectory, reset the start point to go to the next track. Otherwise, stop, quit, station keep, etc.
4. Do (online or offline) 3-D mapping utilizing optical flow, stereo disparity, and final estimated trajectory.

Next, we provide the details.

A. Trajectory Planning

Image-based positioning relies on the integration of estimated incremental displacements over consecutive frames (see Section II-B). Thus, small errors typically integrate to significant drifts in position over an extended time, as in the case of a velocity Doppler system. In the absence of an independent mechanism to verify the vehicle's absolute position—e.g., an (ultra)short baseline acoustic system in combination with global positioning system (GPS)/differential global positioning system (DGPS) on the surface ship to track the ROV and establish absolute position—a desirable strategy to reduce or overcome the error is to accommodate some

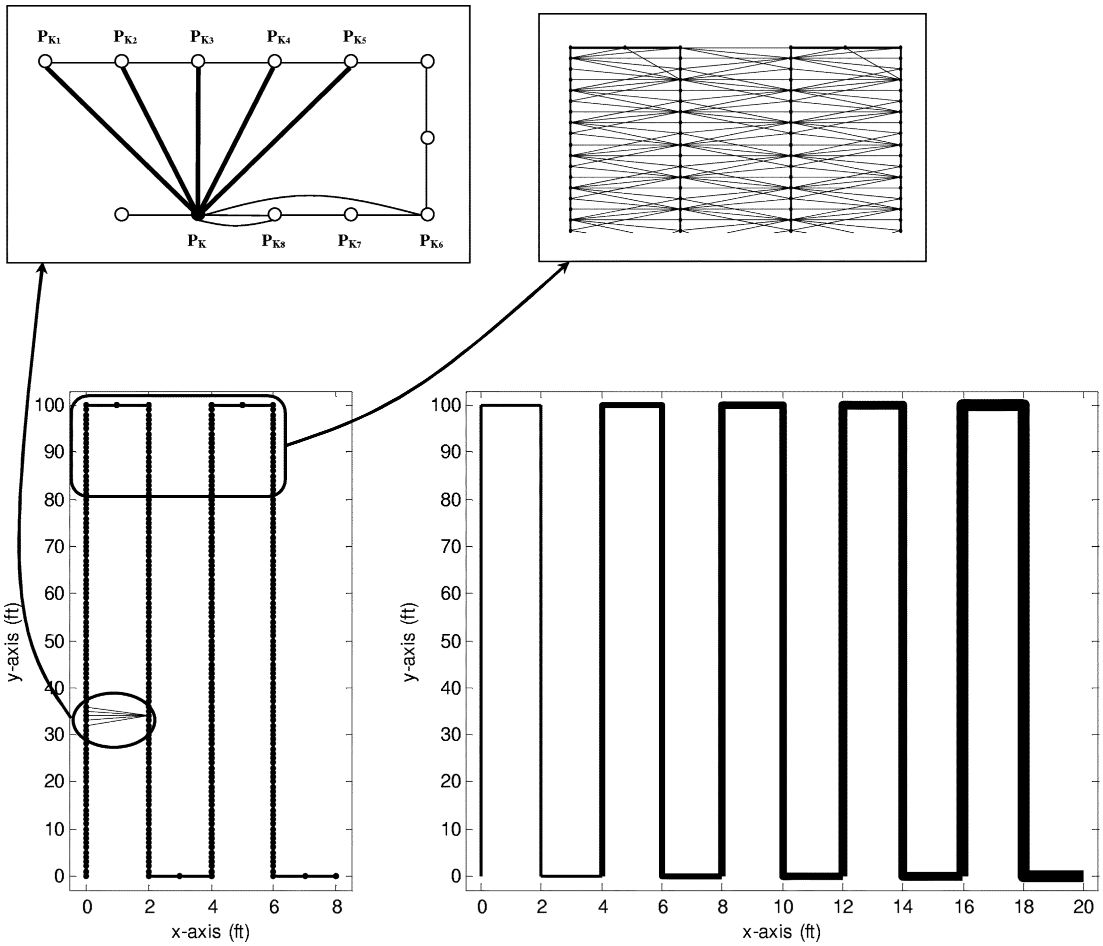
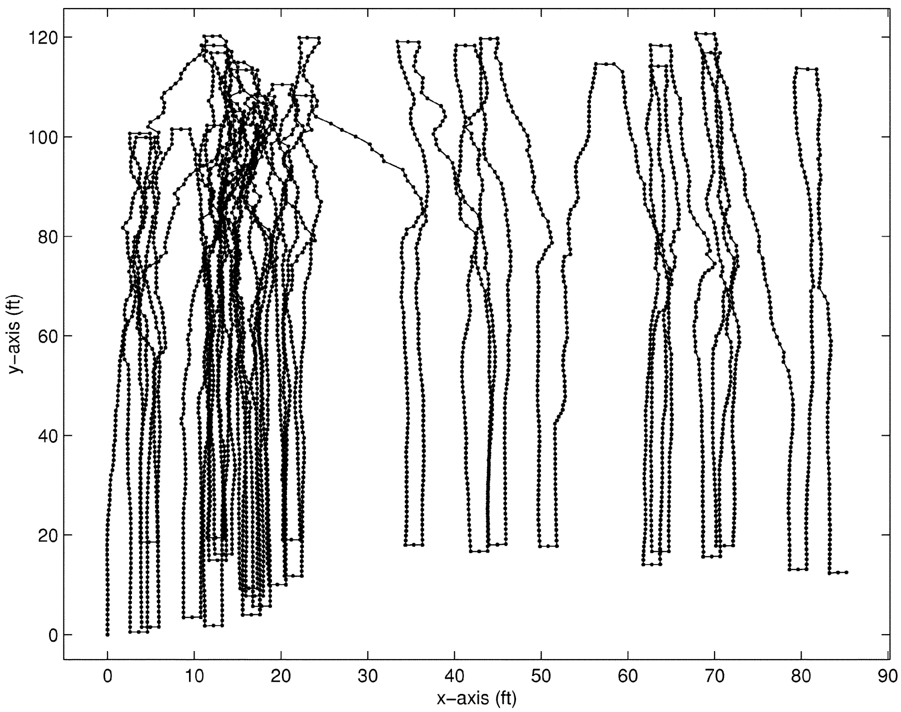


Fig. 3. Bottom left: Two sections from two consecutive tracks with parallel swaths have been magnified to show (top left) the connections among neighboring key points. Bottom right: First five tracks, shown with growing line thickness, highlight the sequential nature of trajectory updating in the GA process. Accuracy increases as new tracks provide additional observations.

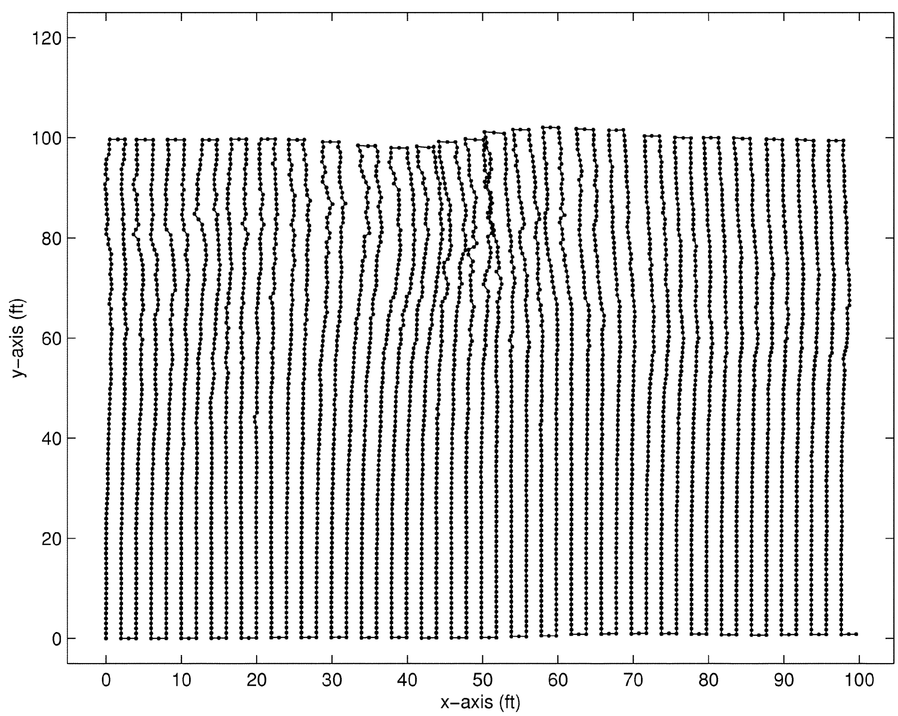
degree of overlap between the images at selected points along the trajectory and a reference image at some position known with good accuracy (e.g., the start point of the trajectory). An ideal scenario involves looping trajectories that cross at or near the known point [17]. Fig. 1 (path shown in solid) depicts the utilization of this idea to cover an area of some diameter D. In this scenario, the integration error is maintained bounded and independent for each loop, although the positioning accuracy decreases with distance from the reference point. Furthermore, image overlaps at the far end of each loop allow us to improve the trajectory estimation by exploiting independent constraints among neighboring points on different sections of the path near the perimeter; more precisely, by considering several independent paths between selected pairs of points (see below).

Alternatively, we can plan traverse swaths (a lawn-mowing pattern), which is a common strategy to cover an area over the sea floor in mapping applications (path with dotted lines). Misalignment of the images/depth maps from various tracks due to the imperfect (incomplete) knowledge of the sensor platform trajectory can be corrected by planning, and exploiting information from, image overlaps in adjacent swaths. This registration problem is more difficult than simple XY shifts, since general 6-degrees-of-freedom movements of the vehicle can induce complex image deformations that are to be resolved. The accuracy of the alignment is directly tied to the percentage overlap, which generally depends upon many factors, such as information content of the image over these regions, complexity of the surface relief, field of view of the sensor, etc.

Unlike looping trajectories, the path does not cross at the start position in subsequent swaths; thus, drift errors are still inevitable. However, the error would grow linearly with the distance from the start position, instead of the covered area; accuracy drops independently along X and Y directions as we move away from the reference position. Taking note of the tradeoff between the processing requirements and amount of image overlap to improve the precision, the emphasis on accuracy and cheaper cost of processing tilts the scale toward accommodating larger overlaps (chosen as 80% in this study) for adjacent swaths. Again, positioning accuracy drops with distance from the start (reference position) and is expected to be worse at the far end of each track. Although not investigated in this paper, a strategy to overcome this problem is to utilize sparse targets at known positions along the region perimeter. For example, the deployment of "opti-acoustic" transponders, which can be calibrated (to determine their positions) based on acoustic techniques (e.g., [7], [8], [10], and [20]), can be used as optical beacons during ROV data acquisition. Alternatively,



(a)



(b)

Fig. 4. Initial and GA-realigned trajectories.

optical beacons can be placed and calibrated utilizing skilled divers [15], [53].

B. Trajectory Following

We now describe the data-acquisition and online-processing strategies along the tracks of the trajectory. We consider a trajectory with parallel swaths, without loss of generality, as the process is similar for a path with looping tracks.

We process monocular images at video rate (FrameRate) of 30 frames/s, for F2F registration and motion integration, and record stereo data every N frames at the so-called key points.[3] We then do F2M position adjustment at key points (every N frames) by registering the current frame with the mosaic and,

[3]For video-rate processing and 2-D photo mosaicking, we use the left sequence of the stereo images, without loss of generality.

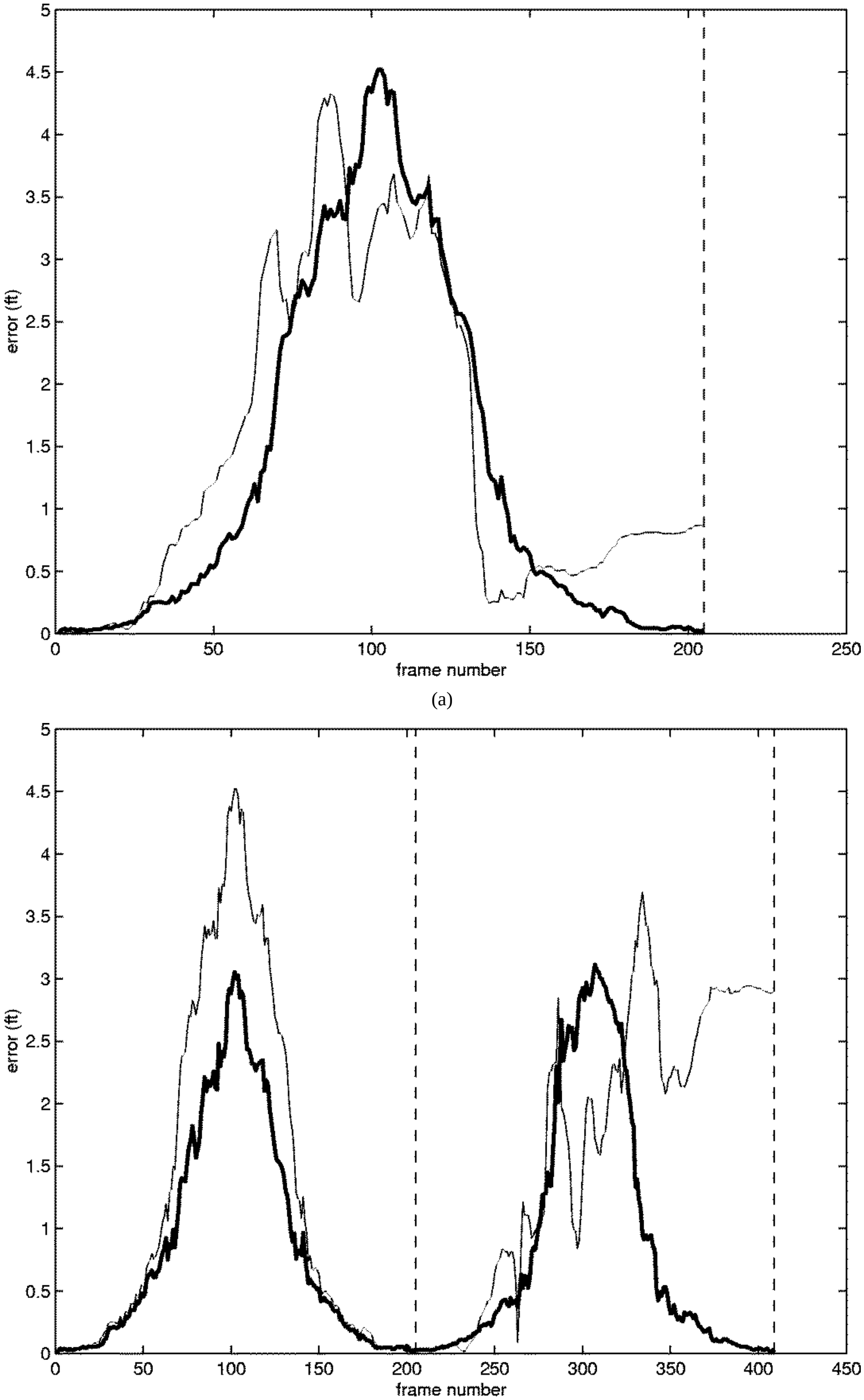
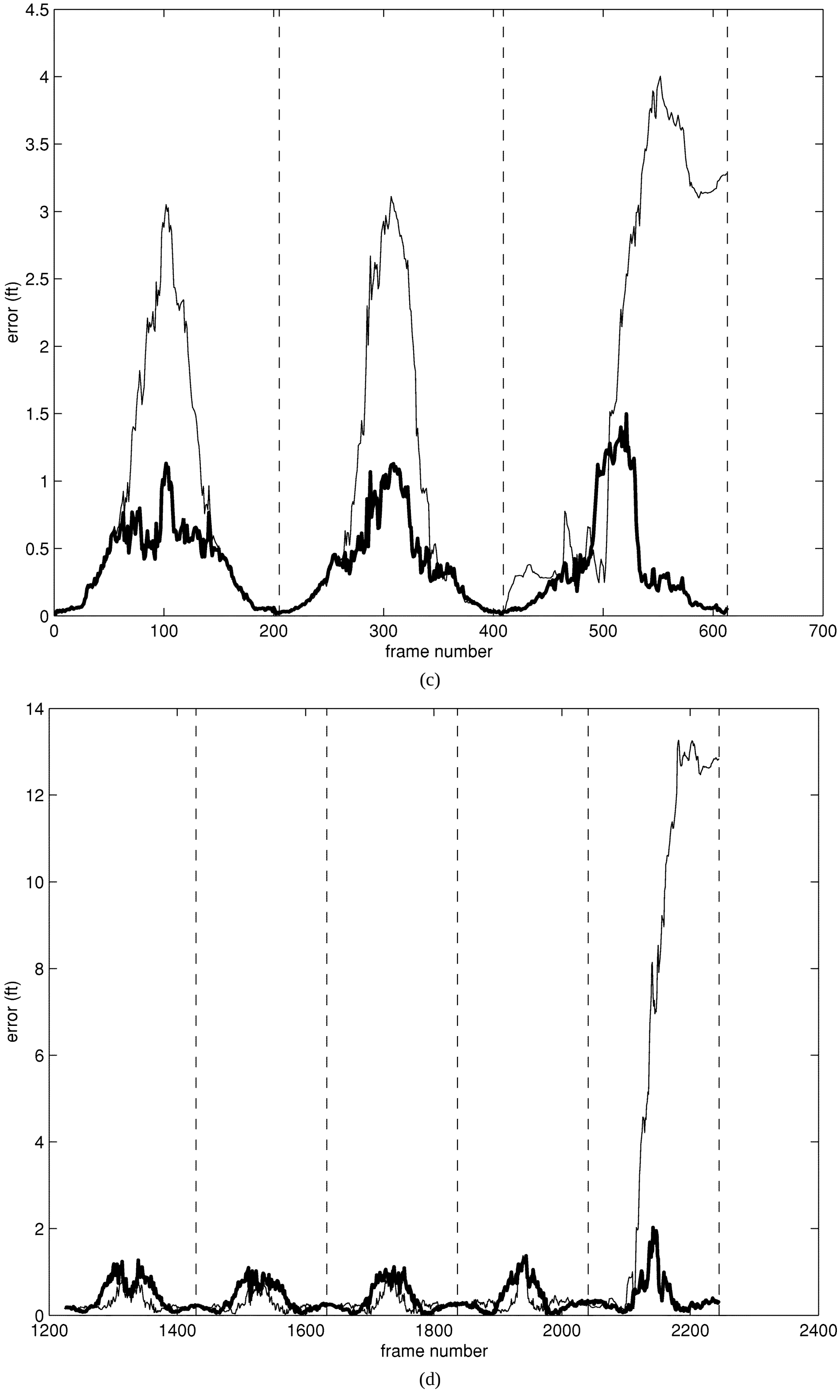


Fig. 5. Position errors of tracks 1–5, 11, 16, 21, and 25, at each track end, before and after GA correction, shown in solid and bold-solid, respectively. Only the current and previous four tracks are revised through the GA process.

consequently, update the mosaic with the latest frame [48]. We choose

$$N = \text{Round}\left\{\frac{(1\text{–OverLap})*\text{ImgSize}*\text{Altitude}*\text{FrameRate}}{(\text{FocalLength}*\text{Speed})}\right\}$$

with OverLap $= 90\%$ for the stereo sampling rate at consecutive key positions. These key points are depicted as nodes along a sample track in Fig. 3. As noted, each track i of the entire trajectory corresponds to two swaths (legs) along the X direction and two short segments along Y-axis to move from the forward to return swath, and to move to the start point of the next track.



(c)

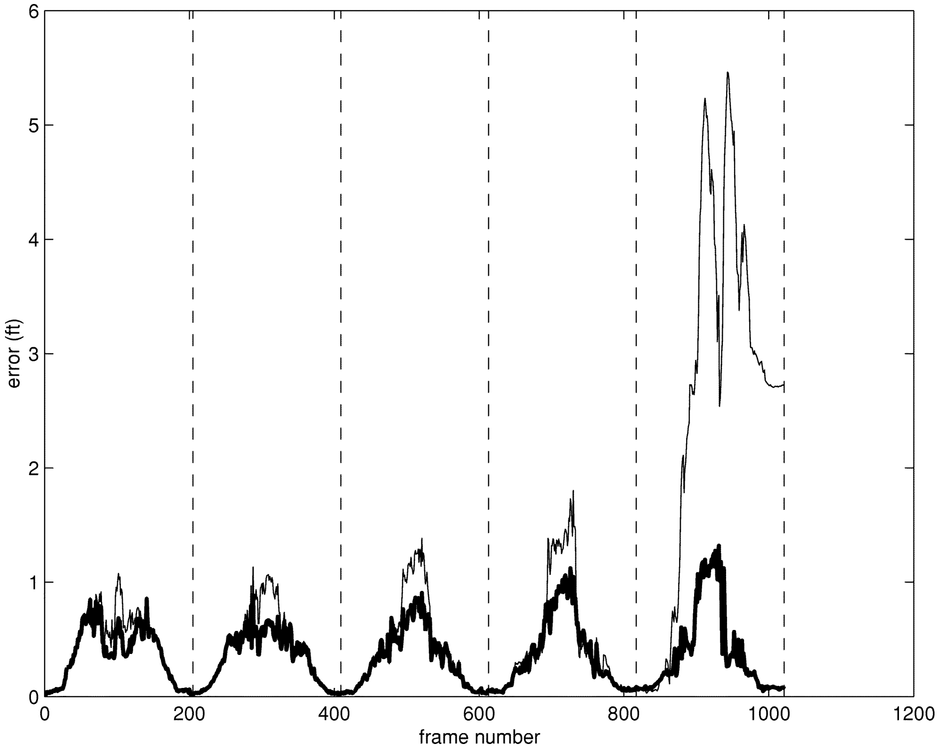
(d)

Fig. 5. *(Continued.)* Position errors of tracks 1–5, 11, 16, 21, and 25, at each track end, before and after GA correction shown in solid and bold-solid, respectively. Only the current and previous four tracks are revised through the GA process.

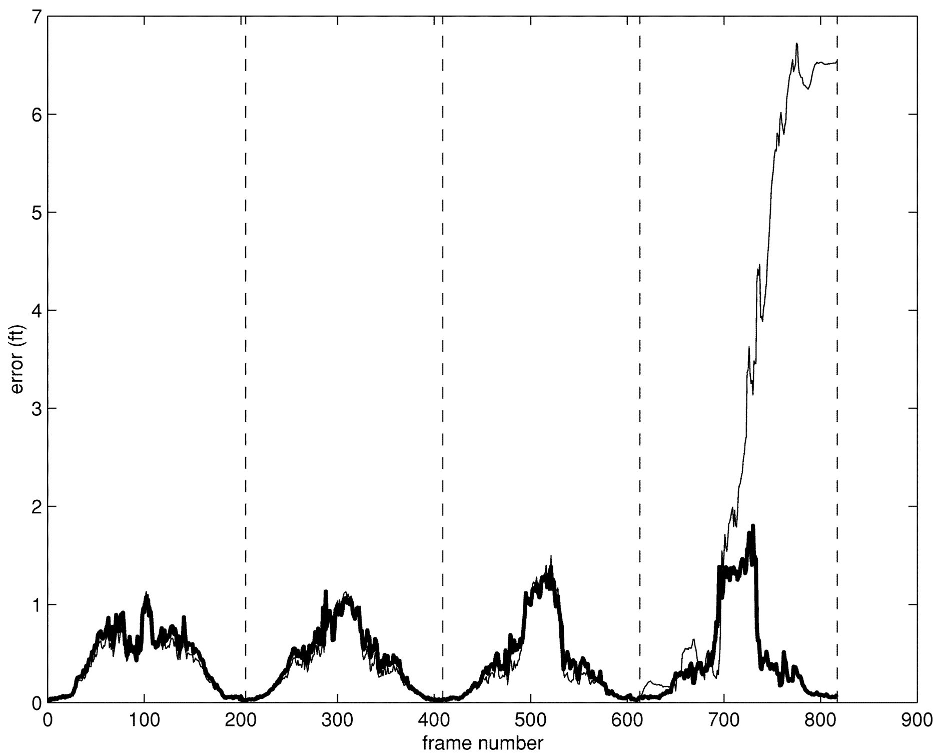
Points $\mathbf{P}\big((i-1)(2M+4)N\big)$ and $\mathbf{P}\big(i(2M+4)N-N\big)$ are the initial and end points of track i, where

$$M = \text{round}\left\{\text{AreaLength}\frac{\text{ImgSize} * \text{Altitude} * \text{FrameRate}}{\text{FocalLength} * N * \text{Speed}}\right\}$$

is the number of key points (mosaic updates) on each leg of the track. Points $\boldsymbol{P}(0)$ and $\mathbf{P}\big((2M+3)N\big)$ are the initial and end points of the first track. An overlap of 80% on two neighboring swaths gives three key points on the short segments along the Y-axis at positions $\mathbf{P}\big((i-1)(2M+4)N+MN\big)$, $\mathbf{P}\big((i-1)(2M+4)N+(M+1)N\big)$, and $\mathbf{P}\big((i-1)(2M+4)N+(M+$



(e)



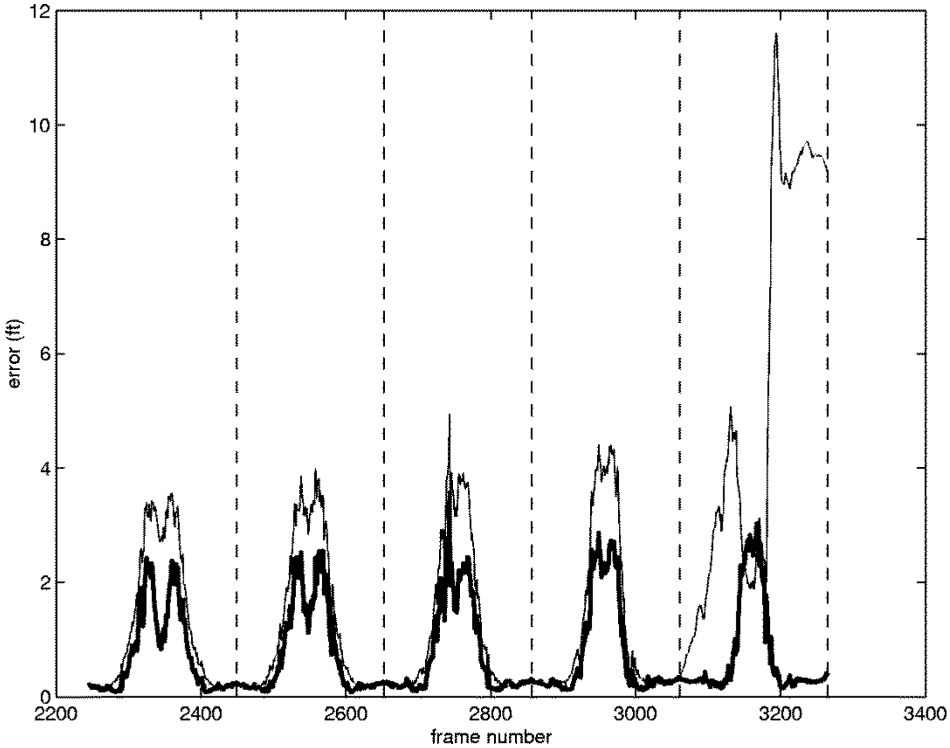
(f)

Fig. 5. *(Continued.)* Position errors of tracks 1–5, 11, 16, 21, and 25, at each track end, before and after GA correction shown in solid and bold-solid, respectively. Only the current and previous four tracks are revised through the GA process.

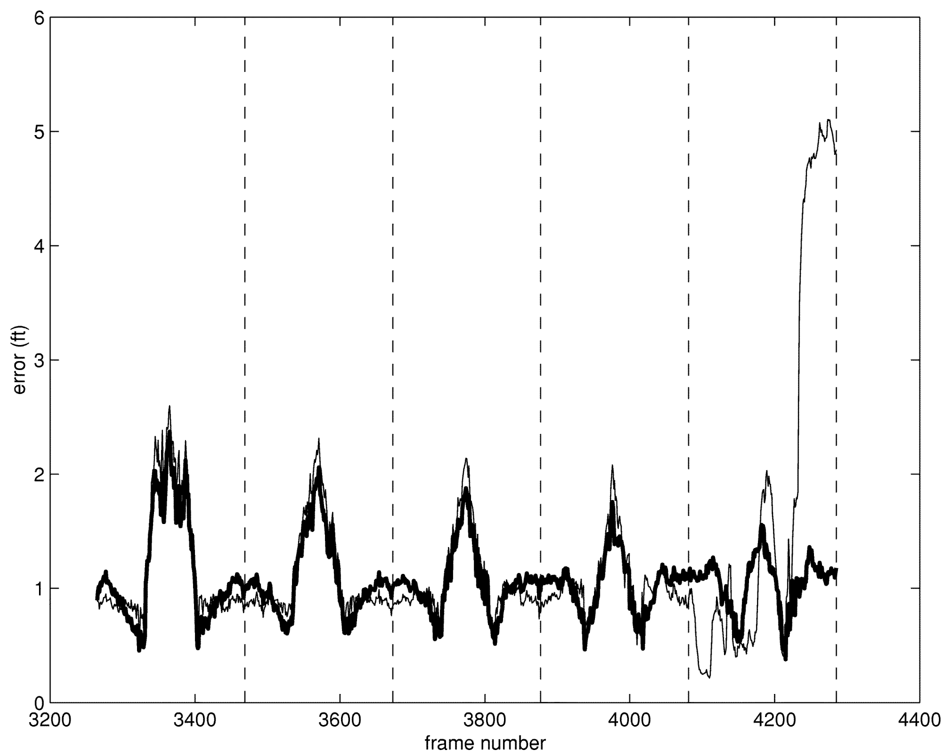
$2)N\big)$. Overall, approximately $L = M/4$ tracks are necessary to cover the entire area.

The alignment with the mosaic (MBP) at keypoints, although significantly reducing the drift errors, cannot totally eliminate them, as discussed in [48]. It primarily serves to maintain a reasonable level of positioning accuracy to achieve the desired large overlap over adjacent swaths of each track (ideally 80%, as planned). At the end of each track, a trajectory adjustment is again carried out through a global-realignment process, which is described in the next section.

As an example used in this study, we update the mosaic every $N = 15$ frames (once per 0.5 s) for a camera with nearly 50° in



(g)

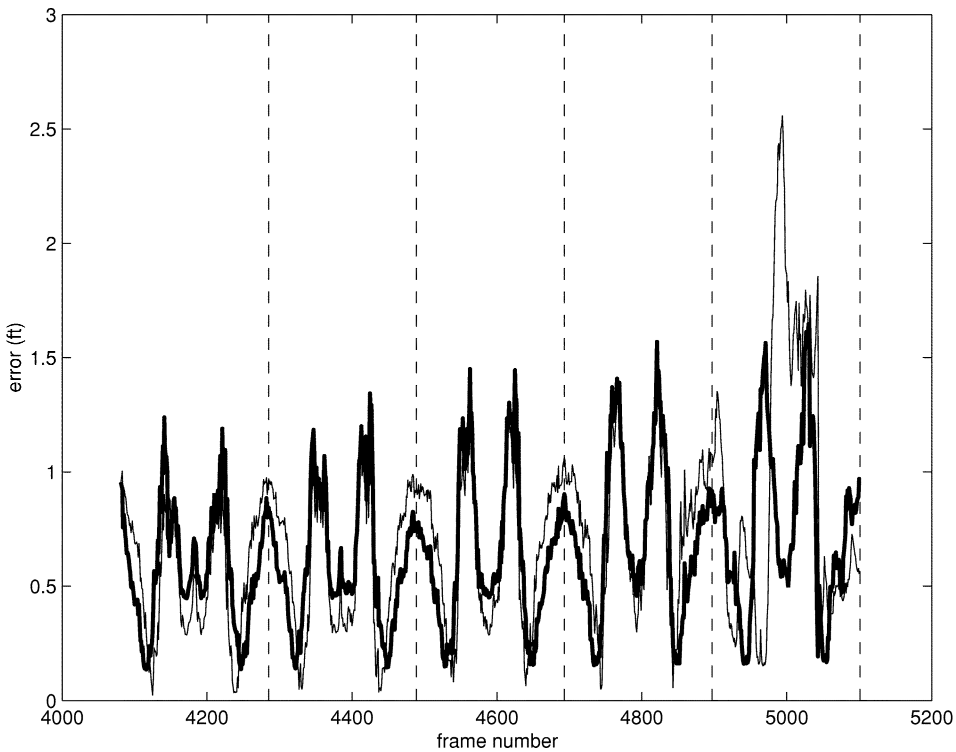


(h)

Fig. 5. *(Continued).* Position errors of tracks 1–5, 11, 16, 21, and 25, at each track end, before and after GA correction shown in solid and bold-solid, respectively. Only the current and previous four tracks are revised through the GA process.

the field of view, on a vehicle traveling at an altitude of roughly 10–11 ft from the bottom at 2 ft/s. This gives approximately $M = 90$ mosaic updates over a leg of roughly AreaLength $=$ 100 ft. We process in real time 1500 monocular frames on each leg of the track: roughly 75 000 monocular frames totally over the entire trajectory. In comparison, we record roughly $P = 200$ stereo pairs on each track, 5000 in total, which are utilized in global-alignment computations and 3-D mapping.

For offline trajectory estimation and 3-D mapping, we may choose to solely integrate F2F motions for positioning, which



(i)

Fig. 5. *(Continued).* Position errors of tracks 1–5, 11, 16, 21, and 25, at each track end, before and after GA correction shown in solid and bold-solid, respectively. Only the current and previous four tracks are revised through the GA process.

may be continued along subsequent tracks until we reach the end of the trajectory (without intermediate corrections). However, drift errors arising from the integration of such a large number of frames, even with corrections based on MBP, can lead to significant drifts. Next, we describe a trajectory-adjustment scheme that is carried out at the end of each track. Through this realignment process, we frequently revise the positions of key positions on the current and previous tracks, leading to a significant improvement in the overall trajectory. Since the outcome feeds back to the trajectory-following process, this naturally leads us to the acquisition of data at more or less the desired points along the preplanned trajectory.

C. Trajectory Adjustment

After the completion of each track, the trajectory is updated based on constraints from a network of selected key positions along the current and previous tracks. The motivation is to incrementally improve the trajectory as new data becomes available. By defining numerous connections (based on relative positions) among neighboring key points on adjacent swaths, we exploit the redundancy of data and constraints to improve the estimated trajectory. Consequently, we can maintain and more accurately follow the desired track.

The network connections, translating to rigid-body motion constraints in the trajectory-adjustment process, are defined between neighboring key points. To reduce the computations somewhat, we have selected every third key point in defining the network nodes. Also, a connection is set among these points where there is an image overlap of 70% or more; Fig. 3(b) depicts a sample of the nodes and connections over a small region of two adjacent swaths.

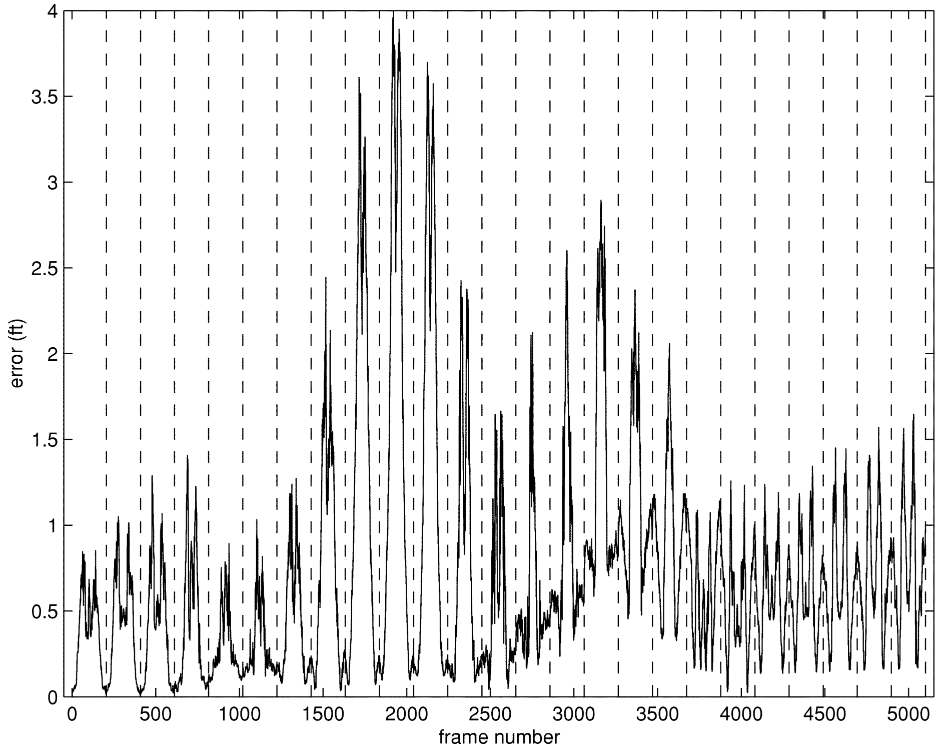
Noting that the impact of older observations and constraints diminish (nearly exponentially) as new tracks are visited and processed, the optimization for the adjustment is carried out over a moving window of the current and T previous tracks ($T = 4$ in our implementation). This avoids a significant computational complexity with the growth in the number of positions/tracks to be updated. In this case, the network remains exactly the same for tracks 5 to L (L =25, in our case).

The trajectory adjustment is an optimized "global" realignment of the key positions, which is consistent with the estimated relative motion between each pair of connected points on the network. In other words, the problem is to find the optimum positions of these trajectory points, knowing the pairwise motions between them. (It is noted that in our particular implementation, "global" implied the last $T+1=5$ frames and, therefore, the readjustment is only global within a local region!) The relative motion for each pair of connected key points on the network is determined from direct measurements of the optical flow using (2); see also Fig. 3 and Section IV for computational details.

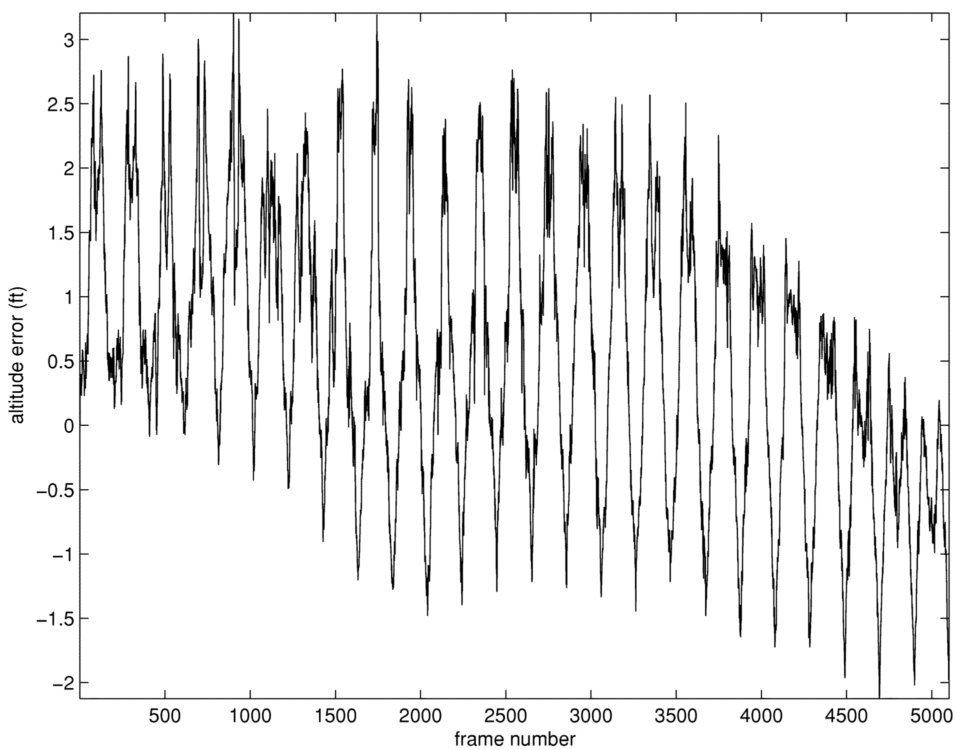
For global alignment, we need to carry out a constrained optimization over available key positions. Our constraint is the transformation between the camera-coordinate systems at any two key positions $\mathbf{P}_i$ and $\mathbf{P}_j$ (defined by the network connections) in terms of the motions between them

$$\mathbf{P}_j = \boldsymbol{R}\mathbf{P}_i + \boldsymbol{t} \tag{6}$$

where $\boldsymbol{R}$ is computed from the rotational vector $\boldsymbol{\omega}$ using the Rodriguez formula. This equation cannot be expressed in terms of



(a)

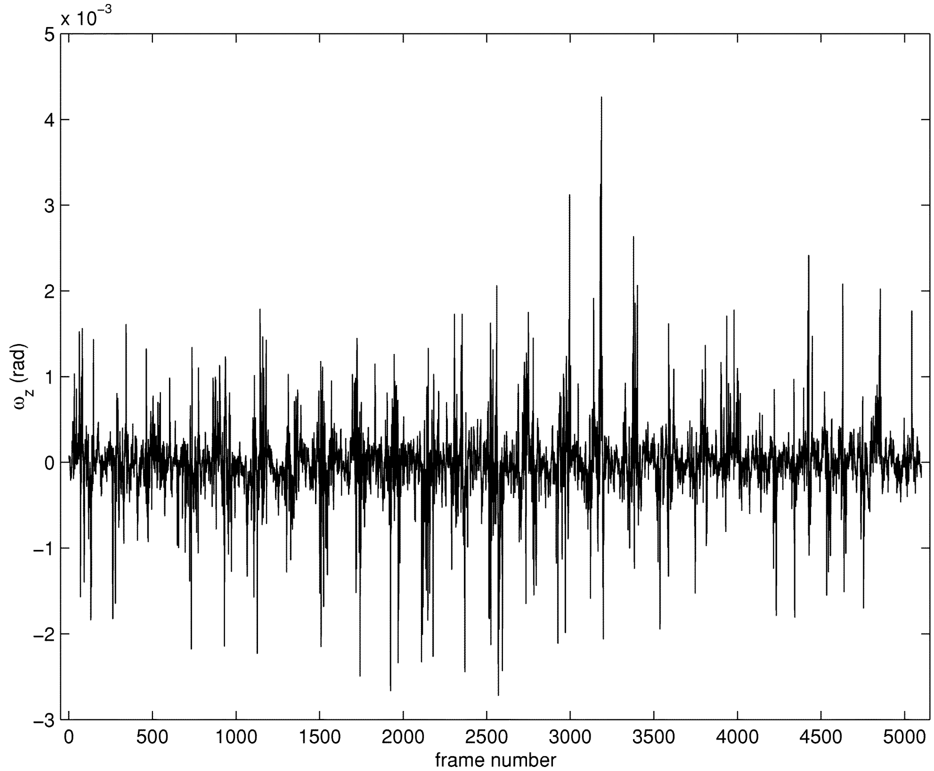


(b)

Fig. 6. (a) Position errors along the trajectory and (b) the Z component. Dashed lines separate sections within consecutive tracks 1–25.

the measurements (motion parameters $\boldsymbol{m}$) explicitly; that is, in the form $\boldsymbol{m} = \boldsymbol{f}(\mathbf{P}_1, \mathbf{P}_2)$. We instead use the *mixed-adjustment model* in the implicit form $\mathbf{P}_j - \boldsymbol{R}(\boldsymbol{\omega})\mathbf{P}_i - \boldsymbol{t} = \mathbf{0}$ [25], [39]. We have one such vector equation for each pair of connected key points. Expressing all the unknown position vectors as one large vector $\boldsymbol{X}$ and, similarly, $\boldsymbol{L}$ for corresponding motion vectors, we have a nonlinear system $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{L}) = \mathbf{0}$. We use this as our constraint in the optimization process, minimizing a quadratic cost function that penalizes deviations from the measured motion parameters. With the aid of Lagrange multiplier vector Λ, we incorporate our constraint into the cost function to finally minimize

$$\phi(\boldsymbol{X}, \hat{\boldsymbol{L}}, \Lambda) = (\hat{\boldsymbol{L}} - \boldsymbol{L})^T \boldsymbol{C}_L^{-1} (\hat{\boldsymbol{L}} - \boldsymbol{L}) + 2\Lambda^T \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{L}}) \quad (7)$$



(c)

Fig. 6. *(Continued).* (c) Heading direction error. Dashed lines separate sections within consecutive tracks 1–25.

where $\{\boldsymbol{L}, \boldsymbol{C}_L\}$ are the estimated and error covariance of the motion vectors computed from the optical flow between pairs of connected key points (see Section IV).

To explain what is achieved here, we note that positions at keypoints are computed in a sequential fashion as we move along each swath (of a track). With drift error, any pair of nearby positions $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$ along adjacent swaths/tracks may and will not be consistent with the rigid body motion $\boldsymbol{m}_{ij} = [\boldsymbol{\omega}_{ij}, \boldsymbol{t}_{ij}]$ that is estimated from the direct measurements between them (or any other path different from that is executed by the vehicle). More precisely, the motion $\boldsymbol{m}_{ij}$ computed from a noisy optical flow for images I_i and I_j seldom satisfies the rigid-body motion constraint $\mathbf{P}_j - \boldsymbol{R}(\boldsymbol{\omega}_{ij})\mathbf{P}_i - \boldsymbol{t}_{ij} = \mathbf{0}$. The optimization process involves modifying the estimated motions to $\boldsymbol{m}'_{ij} = [\boldsymbol{\omega}'_{ij}, \boldsymbol{t}'_{ij}]$, by minimum adjustments $\delta\boldsymbol{m}_{ij} = \boldsymbol{m}'_{ij} - \boldsymbol{m}_{ij}$, and the corresponding positions to $\mathbf{P}_j$ are minimized in the least-square sense over all the keypoint connections over the entire network.

To solve the nonlinear optimization problem, a number of recursive methods can be utilized. The details of selected implementations and their comparisons are given in a separate paper, which provides a more comprehensive treatment of this global-alignment scheme [35].

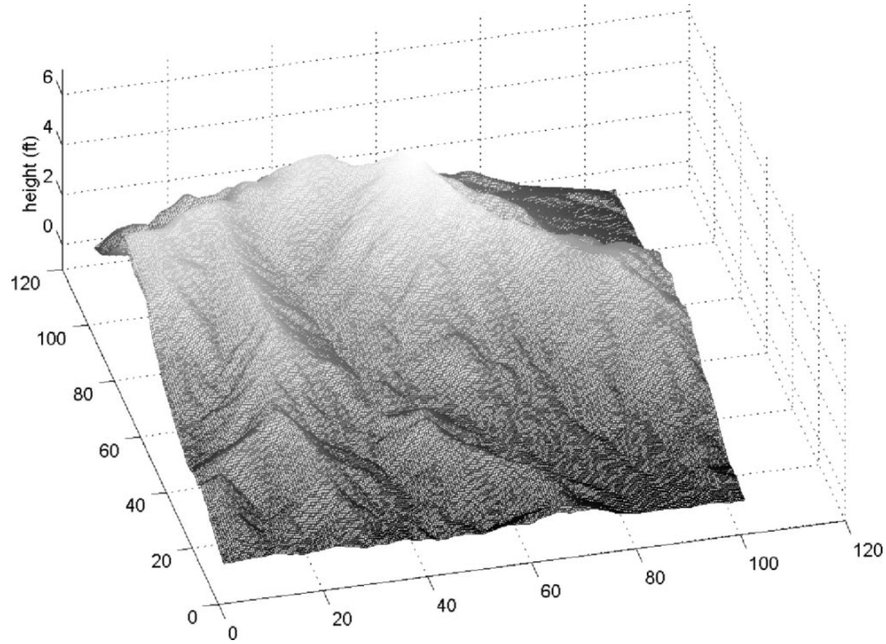
Finally, we should draw a comparison with the global-adjustment methods in some earlier work, primarily on photo mosaicking [22], [54]. Though these schemes are truly global and consider the entire data set, application to large areas, as proposed here, is prohibitively complex. This motivated the GA application to local regions, namely five consecutive tracks, of the trajectory at each update time. (In practice, a global alignment over the entire trajectory can be done in offline computations before the final 3-D map of the entire area is calculated.) More importantly, however, these other methods utilize a much simpler projective model, namely planar homography [23], which is valid for planar scenes. Our global-alignment method, based on the unconstrained rigid body motion model in (6), is applicable to any arbitrary scene, since we can exploit the 3-D position measurements from stereo imagery (or any other method that provides us with the estimates of the 3-D scene topography).

D. Three-Dimensional Mapping

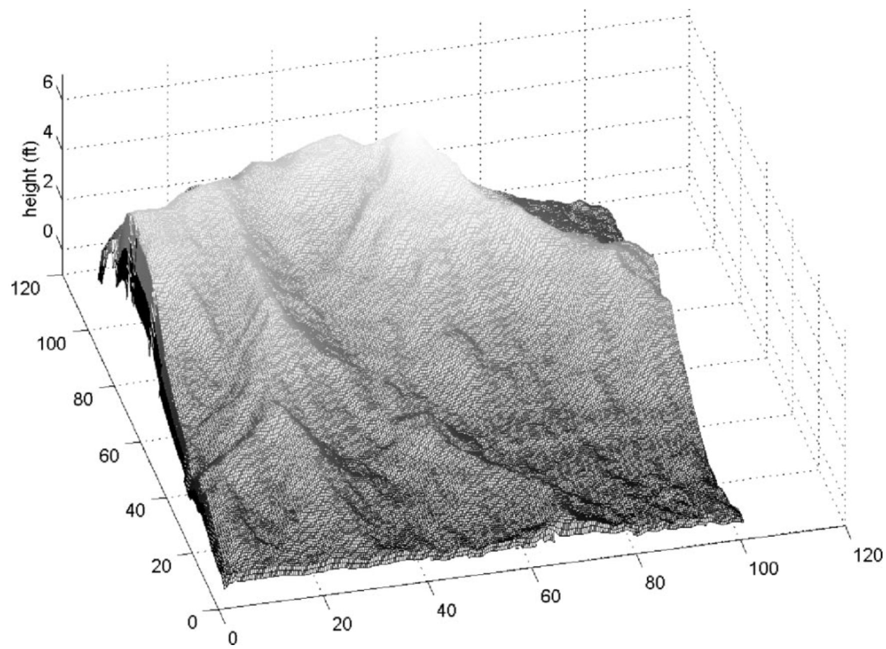
For 3-D mapping, we employ the Kalman filter-based framework described in [29] and [47], which is a recursive estimation method for the integration of stereo disparity and the motion/shading flow that are induced by the motion of the vehicle/source (installed on the vehicle). In addition, incorporating a regularization observation enables us to characterize the (locally) smooth structure of most physical surfaces, while formulating a scalar filter, one for each pixel. In our application, we utilize stereo and motion cues solely. These comprise the disparity computed from stereo pairs acquired at mosaic update times and optical flow calculated for the network of PL points; P points on each of L tracks ($PL = 5000$ in our example).

IV. Processing Requirements

Despite massive amount of stereo and video data, the processing fits very suitably within the framework of parallel and pipeline computing architectures and can be carried out online in (near) real time. We have depicted in Fig. 2 various modules with the corresponding input(s) and output(s), denoted MfRD, SD, OF, and KPNME, for carrying out distinct computations.



(a)

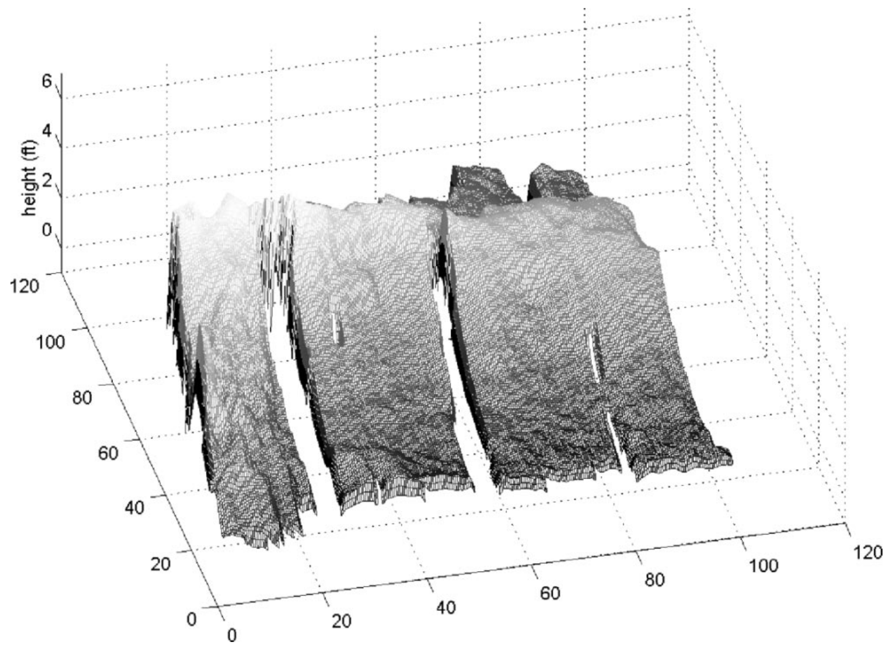


(b)

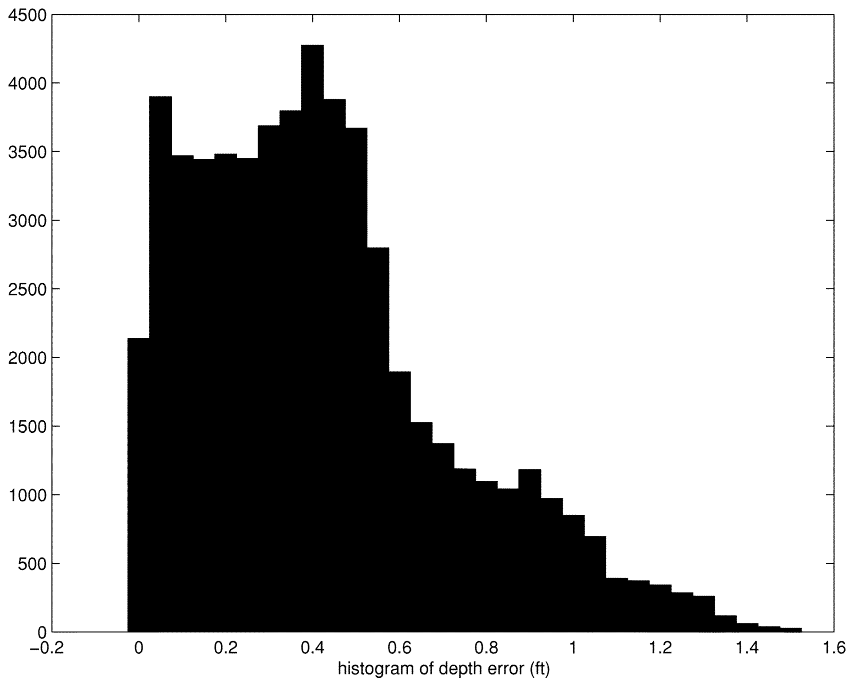
Fig. 7. (a), (b) Original terrain topography and the estimated 3-D map based on corrected trajectory through the global-alignment process.

For real-time performance, the entire processing for all of these modules has to be completed within the travel time between consecutive key points on the network (order of 0.5 s), in addition to video-rate F2F computations. We next examine the processing sequence and details.

1) *F2F and F2M MBP:* As the vehicle moves from a key point P_k to the next P_{k+N}, motion from consecutive frames I_i to I_{i+1} ($i = k, k+1, \ldots, k+N$) are computed and integrated, giving the initial estimate of position at P_{k+N}. Registration of I_{k+N} with the mosaic M_{k+N} (in the same coordinate system) allows us to make a correction, giving the new estimate of the position at P_{k+N}. These computations have been clocked at more than twice video rate on a 100 MHz dual-processor pentium processor.
2) *Motion from Relative Positions (MfRP):* This is a simple calculation from (5) to determine the motion between a key point and each nearby connected point on the network, based on their positions (with respect to the reference position), described in the form of a homogeneous transformation.
3) *Stereo Disparity (SD):* Computation of SD (depth) has to be done once per N frames, since stereo pairs are acquired at every key point. This can be done over five to six levels ($d_o - 2^5 \leq$ disparity $\leq d_o + 2^5$) at video rate, where d_o is the average disparity over the image [68].



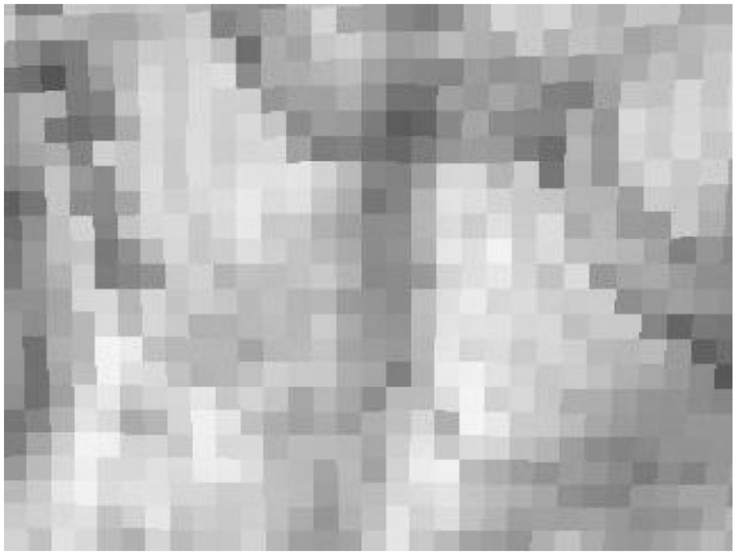
(c)



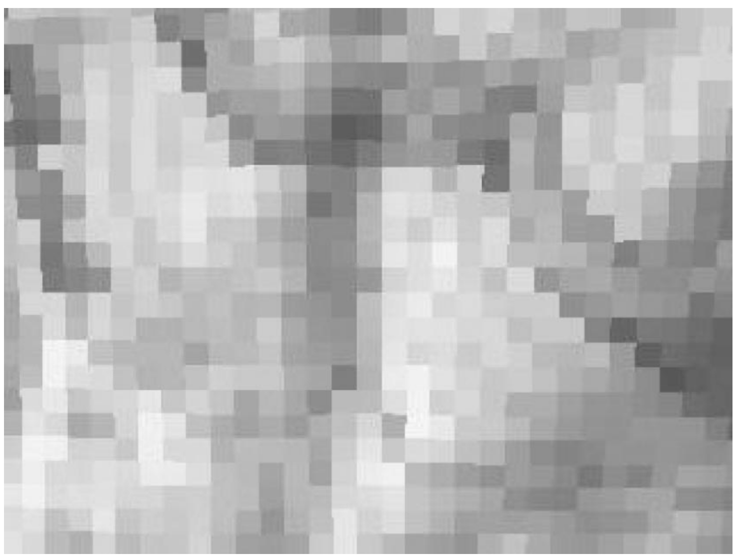
(d)

Fig. 7. *Continued.* (c), (d) Map from the trajectory with no global correction. This is reminiscent of a scenario where the ROV may follow an incorrectly estimated trajectory that does not cover the entire area (c), resulting in holes in the estimated map. (d) Histogram of the absolute depth error map based on GA correction, given in parts (a) and (b).

4) *Optical Flow (OF):* This involves the same type of computation as for SD, but is applied pairwise to the images at the key position and each point connected to it on the network. Each point P_k on the network is connected to no more than 7 previous key points P_{ki} ($i = 1, 2, \ldots n$; $n \leq 7$) on the same or previous leg of the track; see Fig. 3(a) and (b). Once a network node P_k is visited, images I_k, I_{ki} ($i = 1, 2, \ldots n; n \leq 7$) are the input to the OF computing module to determine pairwise optical flows between I_k and I_{ki}. However, by utilizing the relative position of the pair of key points and the 3-D depth map previously calculated at P_{ki}, we can significantly simplify the multiresolution OF computation process. More precisely, we theoretically know the optical flow based on the relative positions P_k and P_{ki} and range (from disparity) at position P_{ki} (previously calculated). Due to integration drift, there may be some estimation error, and this can be computed over one to two levels of a multiresolution OF algorithm (which is the same as SD calculations, but with fewer levels).



(a)



(b)

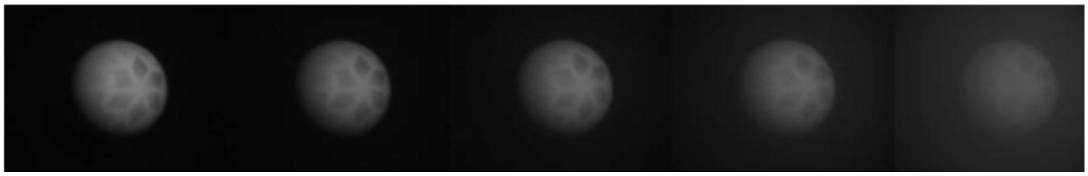
Fig. 8. (a), (b) Sample stereo pair from ground-truth data set showing the true resolution of the images.

5) *Key-Points Network Motion Estimation (KPNME):* The algorithm for global alignment at the end of each track requires the pairwise motions and the initial values of positions of the key points on the network. We can use the initial estimates based on the F2F/F2M computations (MBP). However, a better estimate can be obtained from the closed-form solution in (2), which utilizes pairwise optical flow and stereo-disparity measurements. As these become available at each position P_k, they are processed by the PKNME module. This computations can be done at several times the frame rate (the same as our MBP processing), since primary operations are the summations and products involving various image moments of depth Z and optical flow $\boldsymbol{v} = [u, v]$ over some region of the image.

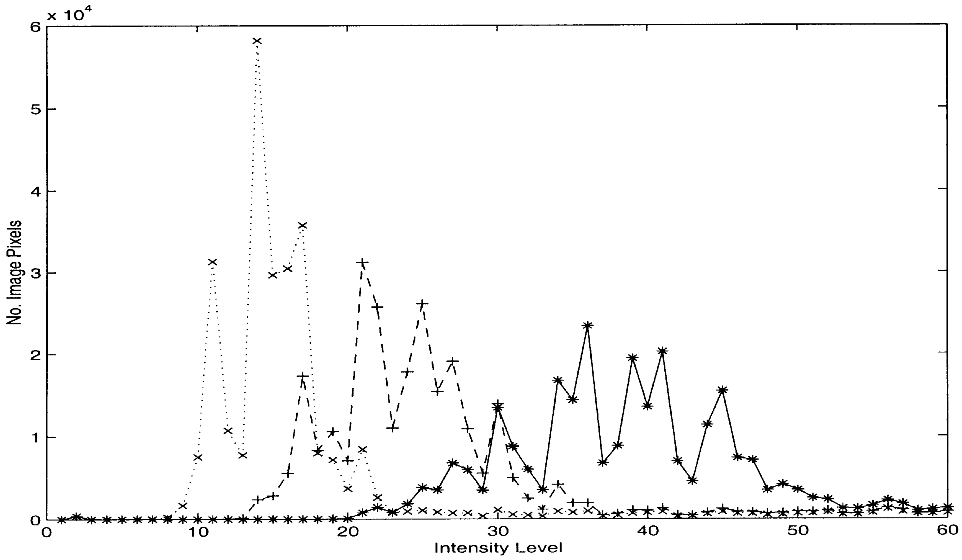
The above computations are to be completed by the time the vehicle moves from P_k to P_{k+N}; the process is restarted with stereo data at this next point. Once the vehicle is at the end of each track, it takes no more than $N/$FrameRate s (say, 0.5 s in our example) to process the various data at this position (as in any previous key position). What is left to do is the nonlinear optimization for GA; see also Fig. 2. Although recursive, the algorithm typically converges in less than a few iterations with good initial estimates. In fact, the main motivation for the processing at step 4, which can be done very quickly, is to obtain better position estimates that lead to a faster convergence. This is another factor that plays a role in the system's performance. Furthermore, it is emphasized that the ability to acquire video data and process in (near) real time is a significant advantage in the deployment of a vision-based integrated positioning and mapping system. Integration with other sensors (say, a high-frequency acoustic positioning system) can provide other benefits, but also requires proper integration and calibration of the "opti-acoustic" system.

V. Experiments

Three experiments are presented to demonstrate various issues discussed in the paper. The first is a simulation with



(c)



(d)

Fig. 8. *(Continued.)* (c), (d) Five images of the same object imaged in water at various turbidity levels. Histograms of three most turbid cases are given. The dynamic range clearly drops with increasing turbidity.

ground-truth data and the other two comprise real images recorded in an indoor water tank.

A. Ground-Truth Data

This experiment simulates an underwater mapping operation. The submersible platforms navigate at an altitude of 11 ft from the reference XY plane on the sea floor along parallel 100 ft swaths covering roughly a 100×100 ft area.[4] Sea-floor variations are assumed in the range of 0.5 to 6.5 ft above the reference plane. The vehicle moves in a horizontal plane with no rotation; i.e., components t_Z, ω_X, ω_Y, and ω_Z of the motion vector are zero. Stereo images are acquired at 204 key positions along each of 25 tracks. A 1-ft vehicle-travel distance between consecutive key points provides a 90% image overlap. A stereo baseline of 1 ft is assumed. The images are constructed (captured) by superimposing a prescribed texture map on a terrain map [see Fig. 4(a)], utilizing a 3-D solid modeling software package [76].

Fig. 4 shows the estimated trajectory based solely on the integration of F2F incremental motions along each track (no F2M correction is done), without (a) and with (b) GA correction process at the end of each track. Not having utilized the F2M process is primarily 1) to quantify the drift from the F2F integration, 2) to assess the impact on (e.g., convergence of) and improvement achieved through the GA process, and 3) to show the impact on depth estimation. To explain these factors more precisely, one notes that the network at each key point consists

[4]With a swath of 100 ft at 11-ft elevation from the ground plane, the mapped sea floor area exceeds 10 000 ft^2.

of several/two connections to several"adjacent-swath" and two previous "same-swath" points. The relative motions and positions of these points for the GA process are determined independently from their optical flows, while estimated positions from F2F/MBP solely serves to follow the planned trajectory during the mission (visual servoing). Thus, having foregone the MBP correction has no effect on the final outcome of the GA process. However, the primary impact is a slower convergence rate with the choice of a less accurate initial condition in the nonlinear GA optimization process.

Fig. 5 correspond to trajectories that are corrected by the GA process, for tracks 1–5, 11, 16, 21, and 25. The solid curves show the results right before and the boldface curve is immediately after the GA correction at the end of each loop. One notes that the primary difference is in the last track, in which the former case is determined by the F2F integration since the correction at the beginning of the track. This provides a measure of track-by-track error growth, while the entire trajectory without GA provides the impact at the global scale; (Fig. 4).

What these other plots also show is that up to track 5, each GA correction process affects the entire past trajectory and that new data allows us to improve the previous estimates. Since the alignment has an impact solely on the last five tracks beyond track 5, only these have been shown. Finally, one may conclude that although the GA process generally improves the results, this is not always true over the entire track (e.g., portions of the tracks 6 to 11). We have also shown in Fig. 6(a) the position error over the entire trajectory, (b) the altitude Z component, and (c) the heading; the other two components of the camera

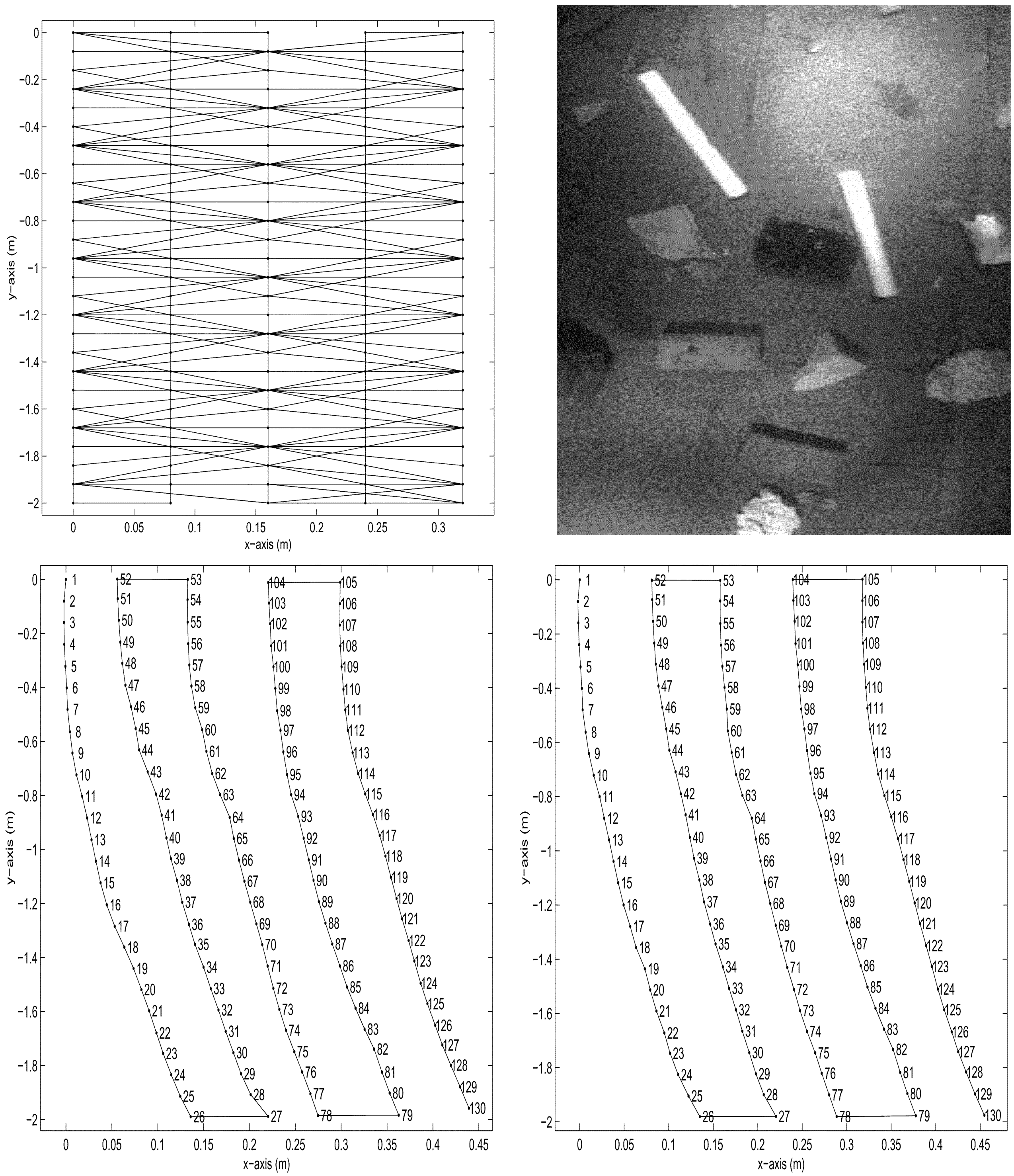
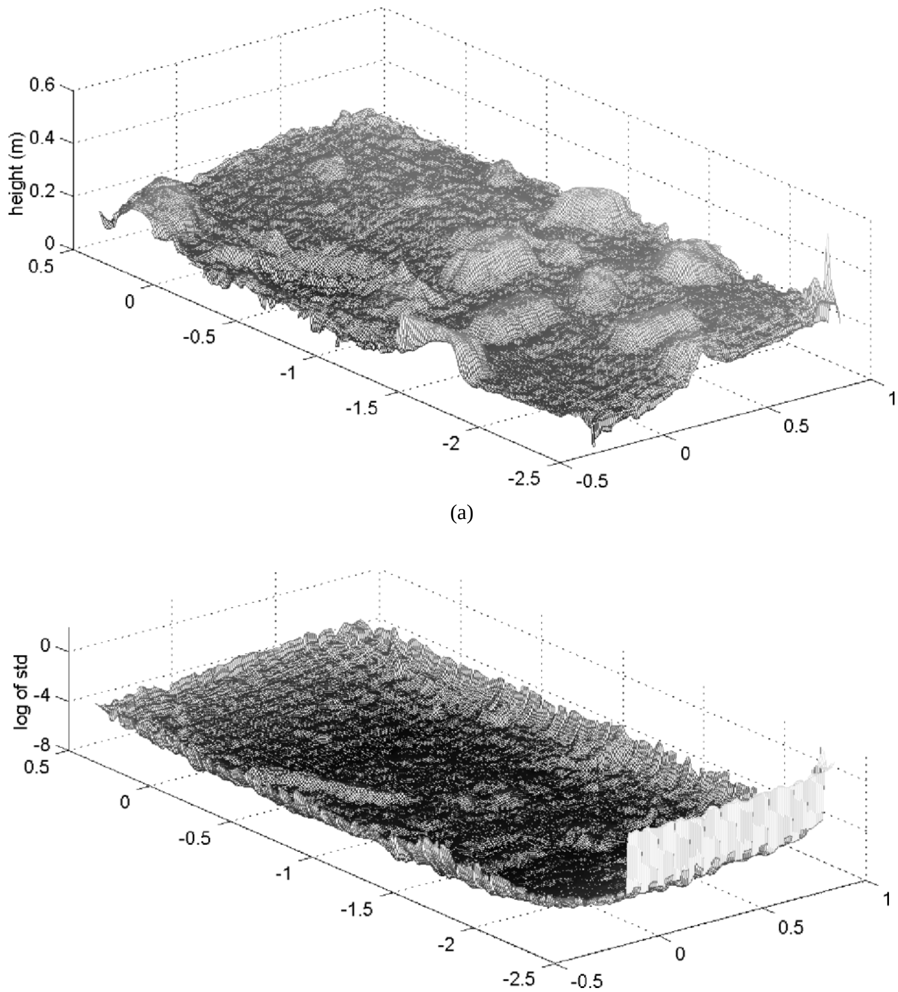


Fig. 9. (Top) First real data set; connections of key points for GA process and the photo-mosaic depicting the nonuniform lighting over the scene. (Bottom) Initial trajectory and (left) trajectory after the GA process.

pose are much smaller. As expected, the points farthest from the reference point have the largest errors.

The reconstructed 3-D depth map in Fig. 7(b) is based on the GA corrected path; this is in good agreement with the true topography. The histogram of the absolute errors in (d) shows an error of 0.6 ft or less for a large percentage of the points. Plot (c) is the 3-D map based on the erroneous trajectory in Fig. 4, obtained by pure integration. Although the poor quality of this result is not surprising, it is reminiscent of a scenario in which tracking an erroneous path results in stereo data that does not cover the entire terrain, therefore resulting in holes in the map. This also highlights the significance in utilizing the MBP scheme in visual servoing in order to reduce position drift along each track.

In evaluating these results, the performance includes a geo-referenced spatial accuracy of about 1 ft over the first five tracks,



(a)

(b)

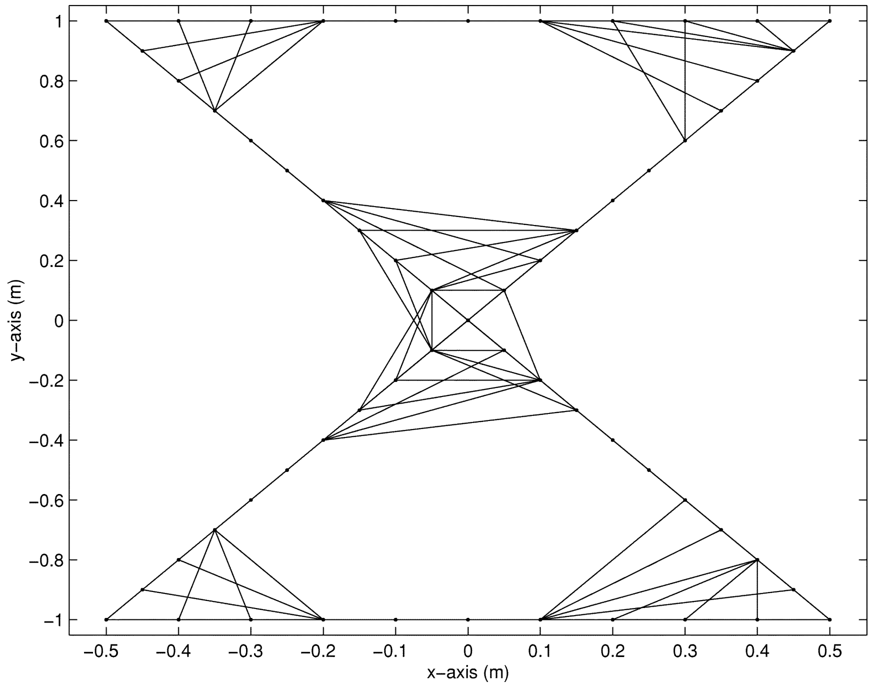
Fig. 10. Reconstructed depth and corresponding log-covariance (STD) maps for the first real data set.

2–4-ft error at the far end of a few intermediate tracks, and no more than 1.5 ft (at the far end) for the remaining tracks. The larger errors can be tied directly to areas of high elevation on the topographical map, closest to the camera (see comments on image resolution next). The maximum elevation error is roughly 1.5 ft. Since the recovered depth based on stereo imaging is measured relative to the camera (11 ft from reference frame), this, with surface elevation in the range of 0.5 to 6.5 ft, corresponds to approximately $1.5/8 \approx 20\%$ maximum error (using average elevation of 8 ft). Ignoring roughly 15% of the points with the largest errors, this becomes $0.6/8 \approx 7.5\%$ for the rest. Finally, to put this into perspective, these results have been obtained by processing images constructed at half the typical 640×480 video resolution. Therefore, the immediate conclusion may be that the original desired specifications have not been achieved. However, let us reveal the last piece of the puzzle, which should provide us with a much rosier picture (and, thus, good prospects for the future of underwater stereo imaging).

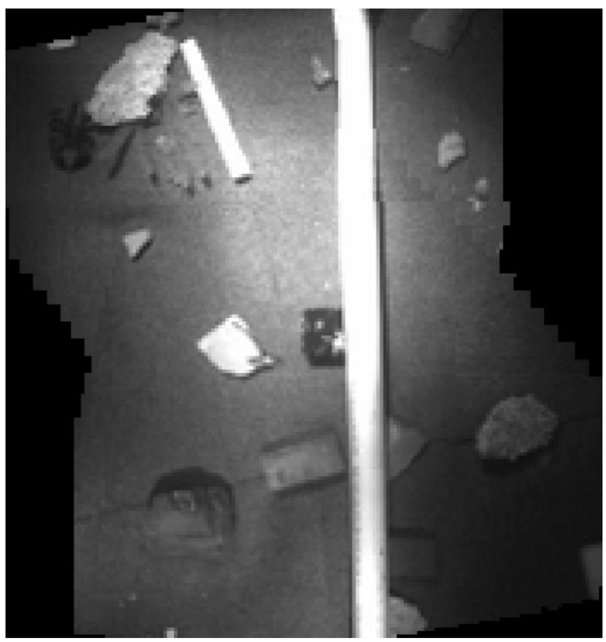
Fig. 8(a) and (b) is a typical 320×240 stereo pair processed by our system. The true resolution is about 1/10 of the image size. Based on the assumed camera parameters and average distance from the scene, each pixel represents a area of nearly 3×3 in over the surface, which is far less than for typical resolution in (underwater) video images. This resolution drops for regions closest to the camera, leading to larger maximum position errors.

Two factors influenced the image resolution when the data was generated and used. First, but less important, was the fact that our original experiments were done over a much smaller region of the terrain and that the texture map of the VR system was sufficient to produce high-resolution imagery. When the problem scale and the mapping area was expanded to fit the proposed 100×100 ft size, the same texture map was used, resulting in images with very low resolution. However, the primary motivation, maintaining the same texture map, was to simulate (what we sincerely believe to be) close to the "worst case" scenario for subsea applications. More precisely, unlike the terrestrial domain where the medium condition plays a negligible, if any, role in image quality (within the proposed problem boundaries), there can be large variations in image contrast and quality with water-turbidity conditions. To demonstrate, Fig. 8(c) are sample images, showing the same object under five different turbid water conditions. Fig. 8(d) shows the histograms of the last three images with the lowest image contrasts (most turbid cases), showing the decrease in dynamic range with increasing turbidity.

Noting the similarity of the true and estimated 3-D depth maps in Fig. 7, we conclude that poor resolution is responsible



(a)



(b)

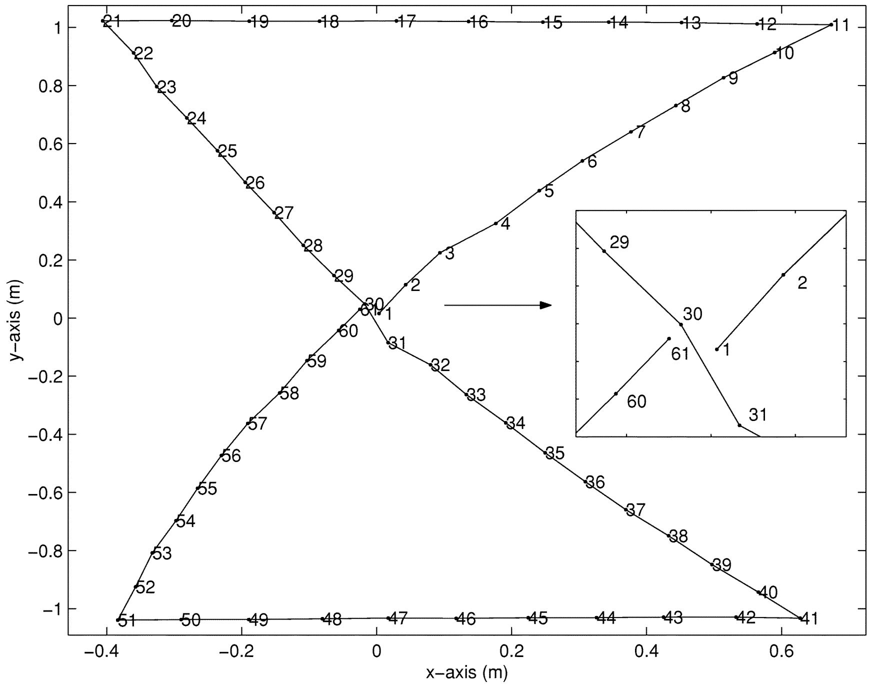
Fig. 11. Second real data set. Connections of key points for GA process and the photo mosaic, depicting the nonuniform lighting over the scene.

for larger than desired (depth) errors and that performance, both in trajectory and depth reconstruction, can significantly be improved with higher resolution images [47]. In particular, using as a basis the image quality in numerous ocean experiments to test other vision-based ROV[5] capabilities (e.g., [72]), we believe that the simulation results presented in this section should rate as the expected performance near the worse-case end of the spectrum. Accordingly, stereo imaging could provide performances that would meet or exceed our specifications. However, before engaging in immature ocean testing, detailed investigation based on a large number of controlled experiments are in order to truly characterize the system performance.

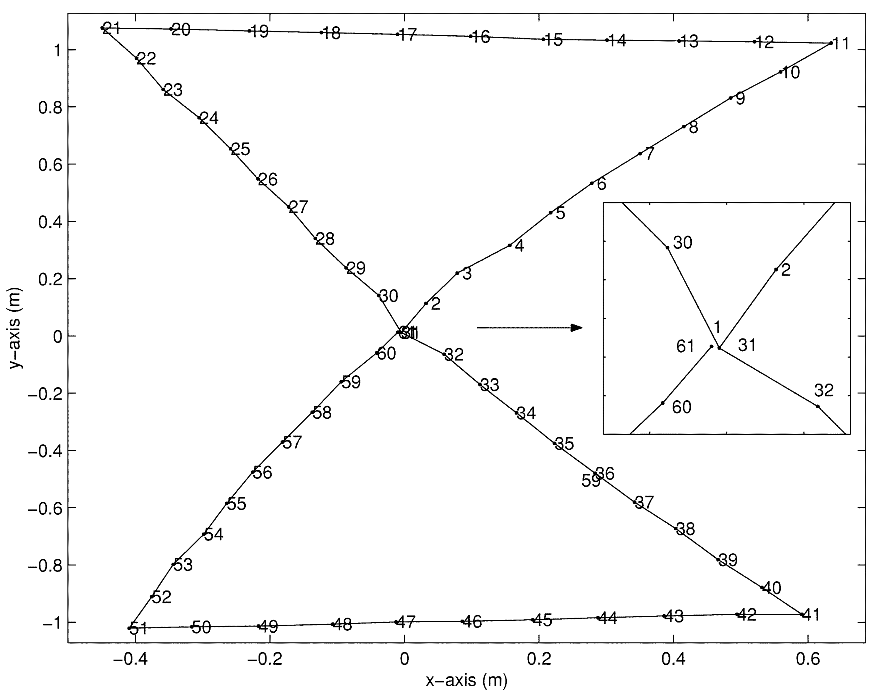
[5]Phantom XTL ROV; Trademark of Deep Ocean Engineering, San Leandro, CA.

B. Real Data

Two real data sets have been recorded in an indoor water tank in a down-look viewing configuration. These consist of 130 and 61 stereo pairs, respectively, from a distance of about 5 ft to the bottom, with objects of no more than a few inches in size. Unlike the first experiment, simulating the construction of the sea-floor topography, these experiments more closely represent a scenario



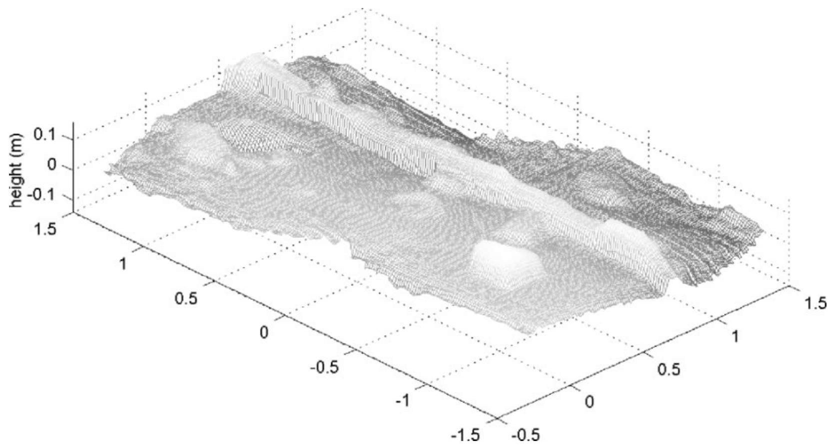
(c)



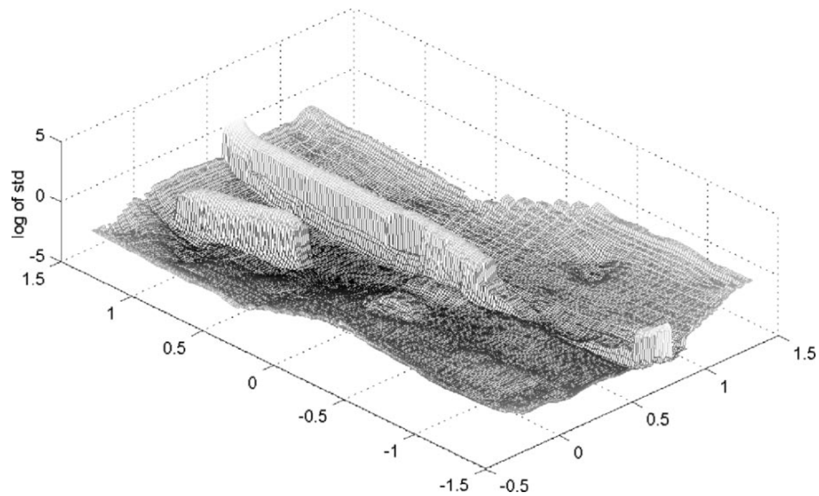
(d)

Fig. 11. *(Continued.)* (c) Initial trajectory and (d) trajectory after the GA process.

for the imaging and reconstruction of smaller objects over the sea floor. The trajectory of the stereo rig was controlled by an overhead track along the horizontal and vertical directions. The planned trajectory, up to the precision of the positioning system (see below), is given in Figs. 9(a) and 11(a), respectively, allowing us to test the two proposed data-collection strategies. The incremental motion between any two frames was measured (on the track system) to be 5 cm/frame. In the second data set, the first image was duplicated at positions 31 and 61 to ensure that the experiment is consistent with two loops that cross exactly at the reference position. This truly is the only ground-truth information in order to evaluate the results, as readings on the positioning system are subject to human error and the accuracy of the track system (see below). Since the overlapping area between consecutive frames satisfies the 10% overlap assumption, all the frames in the sequence are treated as key positions. The pair-wise connections of the key points with the overlap of 70% comprising the network in the application of the GA have been identified. One notes that there are far less overlaps in the second experiment; thus, a smaller number of constraints among neigh-



(a)



(b)

Fig. 12. Same as the previous figure, for the second real data set.

boring track points is available for the GA computations. Moreover, since the total number of images in these data sets is rather small, the global alignment is not applied track by track, but simultaneously on all the points that are pairwise connected on the network.

Both data sets were acquired under spatially varying lighting, simulating scenarios where portions of the scene may be poorly lit when imaged. The 2-D photo mosaics, constructed from the raw images with no radiometric adjustment, are given in panel (b) of these figures, characterizing the scene and depicting the imaging and lighting conditions. The nonuniform lighting is particularly significant in the second set, which has both fully bright (saturated over PVC pipe) and shaded areas. The total coverage area is clearly small as compared to the previous example, due to the water-tank-size limit, but controlled experiments with real images are critical in understanding, analyzing, and evaluating the performance of various system modules. Furthermore, the coverage area of roughly 1×2 m is quite common and realistic when imaging an area over the sea floor at high resolution, say a shipwreck winch.

The estimated trajectories before and after GA correction for the two experiments are given in Figs. 9, 11(c), and 11(d). The reconstructed scenes with the corresponding logarithmic plots of the covariance (STD) maps are given in Figs. 10 and 12.

The constructed trajectories clearly depict the imperfection of (bending in) our track system. This conclusion is readily drawn from the fact that independent trajectories bend at nearly the same area of the track in both experiments and, thus, cannot be due to drift error.[6] These trajectories also show the improvement due to the GA correction, although the discrepancies in the depth maps, before and after correction, are not apparent to the eye; thus, only the results after GA correction are used to demonstrate the 3-D reconstruction of various objects. To the best of our knowledge, based on size measurements, the objects have been constructed with very high precision and the pixel resolution is approximately 0.2 in (in comparison to 3 in for the previous experiment). However, regions of the map in the second set have been constructed incorrectly, primarily due to nonuniform lighting, leading to intensity saturation over the PVC pipes. The good news is that the inaccurate regions can be predicted based on the corresponding uncertainty (error-covariance) measures.

[6]Also, the two data sets were taken a few months apart. Based on these results, we now have concluded that the track assumed to be a high-precision positioning system is not quite perfect.

VI. Modular Capabilities

A number of modular capabilities can be realized, utilizing the stereovision-based system for 3-D mapping of the sea-floor topography and objects, which can be configured and implemented differently from, or to extend, the proposed framework. What ties these together includes some generic vision-based ROV capabilities that have been previously realized, namely the automatic optical station keeping and online trajectory planning [46], [72].

As an example, we may hypothesize a particular scenario in a human-supervised operation, where the system is utilized as proposed, but also allows the operator to interact with the ROV. This may include interrupting the mission temporarily, going into the station-keeping mode, planning another task, carrying out the task, and potentially resuming the mission.[7] One such task may be to acquire super-resolution images over a small region of the sea floor, for mapping certain manmade or natural objects of interest. A natural application arises in documenting a shipwreck in which certain parts and components are to be mapped at these higher resolutions (as compared to the entire area). In marine-science studies, another common scenario would involve the high-resolution mapping of unique or distinct natural objects or features, say a thermal vent or coral-reef mound, in addition to the surrounding area. Accordingly, appropriate data-acquisition tracks may be preprogrammed for such various capabilities and selected based on the application once the vehicle is put into the station-keeping mode over the sea-floor area of interest. Experiments with ground-truth and water-tank images could readily simulate such scenarios, but the data is expected to be at much higher resolutions than in either of these two cases. The final maps would then be registered in a common referenced coordinate system, where necessary.

VII. Summary

We have investigated a framework for the utilization of a stereovision imaging system in the deployment of ROVs for 3-D mapping of the sea floor and targets of interest within areas in the order of roughly 100×100 ft. This scale covers a large number of applications in ocean exploration and high-resolution mapping and sampling, e.g., coral reefs, fish habitats, thermal vents, shipwrecks, etc. We have addressed 1) path planning, execution, estimation, and following; 2) incremental trajectory correction as more data along each track of the path are acquired; as well as 3) 3-D mapping based on the images of the entire data set. We have proposed computational modules that enable the online processing of video images for trajectory following during data acquisition in real time. Preliminary experiments have been presented, suggesting that the framework, and in general stereo imaging, can provide the sought-after capabilities in sea-floor sampling and mapping. More precisely, the technology can fill a void at scales between that of high-frequency sonar and the current state of video imaging and mapping, which is limited to very small areas.

[7]Currently, all of these capabilities, aside from stereo-image acquisition, exist on our ROV vision system, utilizing a single down-look camera.

Current ongoing and future work involves improvements of individual modules and potential reconfigurations of various processing stages for higher accuracy and computational efficiency, as well as a detailed investigation of the framework under a variety of experiments with ground-truth data at various resolutions and controlled water-tank experiments under different turbidity conditions. This is a prerequisite to engagement in tedious controlled ocean testing, e.g., through diver-assisted experiments with known objects placed within a calibrated and mapped environment.

References

[1] G. Adiv, "Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-11, pp. 477–489, May 1986.

[2] A. A. Argyros and F. Bergholm, "Combining central and peripheral vision for reactive robot navigation," *Proc. IEEE Computer Vision Pattern Recognition*, pp. 646–651, June 1999.

[3] K. B. Atkinson, *Close Range Photogrammetry and Machine Vision*. Caithness, U.K.: Whittles, 1996.

[4] D. H. Ballard and O. A. Kimball, "Rigid body motion from depth and optical flow," *Comp. Vision, Graphics, Image Process.*, vol. 22, no. 1, pp. 95–115, 1983.

[5] B. A. J. Barker, D. L. Davis, and G. P. Smith, "The calibration of laser-referenced underwater cameras For quantitative assessment of marine resources," in *Proc. OCEANS'01*, Honolulu, HI, Nov. 2001, pp. 1854–1859.

[6] S. Baten, M. Ltzeler, E. D. Dickmanns, R. Mandelbaum, and P. J. Burt, "Techniques for autonomous, off-road navigation," *IEEE Intell. Syst.*, vol. 13, pp. 57–65, Nov. 1998.

[7] M. Bernsten, "Principles and practice of acoustic systems," *Int. Underwater Syst. Design*, vol. 14, no. 1, 1992.

[8] H. V. Calcar, "What you need to know to improve LBL submersible tracking and performance," in *Proc. ROV'85*, 1985, pp. 43–51.

[9] R. N. Carpenter, "Concurrent mapping and localization with FLS," in *Proc. Workshop Autonomous Underwater Vehicles*, Cambridge, MA, 1998, pp. 133–148.

[10] D. J. Chen and E. Alvardo, "Ultrashort baseline navigation for the sea squirt AUV," in *Proc. UUST Symp.*, Durham, NH, 1991, pp. 644–649.

[11] M. Chen, T. Jochem, and D. Pomerleau, "AURORA: A vision-based roadway departure warning system," Carnegie Mellon Univ., Robotics Inst., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-97–21, 1997.

[12] I. Davis and A. Stentz, "Sensor fusion for autonomous outdoor navigation using neural networks," *Proc. IEEE/RSJ Int. Conf. Intelligent Robotic Syst. (IROS'95)*, vol. 3, pp. 338–343, Aug. 1995.

[13] D. L. Davis and R. F. Tusting, "Quantitative benthic photography using laser calibrations," in *Proc. UnderSea World*, San Diego, CA, 1991.

[14] G. Dudek, M. Jenkin, and E. Milios, *Robotic and Manufacturing Systems – Recent Results in Research, Development and Applications in Topological Exploration of Unknown Environments with Multiple Robots*, M. Jamshidi, F. Pierrot, and M. Kamel, Eds. Albuquerque, NM: TSI, 1996.

[15] S. English, C. Wilkinson, and V. Baker, *Survey Manual for Tropical Marine Resources*, 2 ed. Townsville, Australia: Australia Inst. Marine Sci., 1996, p. 402.

[16] R. Eustice, H. Singh, and J. Howland, "Image registration underwater for fluid flow measurements and photomosaicking," in *MTS/IEEE Oceans Conf.*, Providence, RI, 2000.

[17] S. D. Fleischer, H. W. Wang, S. M. Rock, and M. J. Lee, "Video mosaicking along arbitrary vehicle paths," in *Proc. Symp. AUV Technol.*, Monterey, CA, June 1996, pp. 293–299.

[18] S. D. Fleischer and S. M. Rock, "Global position determination and vehicle path estimation from a vision sensor for real-time mosaiking and navigation," in *Proc. Oceans*, Halifax, NS, Canada, Oct. 1997, pp. 641–647.

[19] R. Gershon, M. Benady, and T. Shalom, "3D reconstruction from multiple 2D views using tri-linear tensor technology," in *SPIE Photonics West Conf.*, San Jose, CA, Jan. 2001, pp. 32–39.

[20] E. M. Geyer, P. M. Creamer, J. A. D'Appolito, and R. G. Gains, "Characteristics and capabilities of navigation systems for unmanned untethered submersibles," in *Proc. UUST Symp.*, Durham, NH, 1987, pp. 320–347.

[21] N. Gracias and J. Santos-Victor, "Underwater video mosaics as visual navigation maps," *Comput. Vision Image Understanding*, vol. 79, pp. 66–91, July 2000.

[22] ——, "Underwater mosaicing and trajectory reconstruction using global alignment," in *Proc. OCEANS'01*, Honolulu, HI, Nov. 2001, pp. 2557–2563.

[23] R. Hartley and A. Zisserman, *Multiple View Geometry*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[24] T. Heimonen, J. Hannuksela, J. Heikkil, J. Leinonen, and M. Manninen, "Experiments in 3D measurements by using single camera and accurate motion," *Proc. IEEE Int. Symp. Assembly Task Planning (ISATP'01)*, pp. 356–361, May 2001.

[25] R. A. Hirvonen, "Adjustment by least squares in geodesy and photogrammetry," . New York: Frederick Ungar, 1971.

[26] A. Howard and H. Seraji, "An intelligent terrain-based navigation system for planetary rovers," *IEEE Robot. Automat. Mag.*, pp. 9–17, Dec. 2001.

[27] J. S. Jaffe and K. D. Moore, "Development and testing of a new underwater three dimensional imaging system: 3D sea scan," in *Proc. Mine Countermeasures Symp.*, Monterey, CA, Feb. 2000.

[28] A. Khamene, "Motion and Structure from Multiple Cues; Optical Flow, Shading Flow, and Stereo Disparity," Ph.D. Dissertation, Univ. Miami, ECE Dept., Coral Gables, FL, 2000.

[29] A. Khamene, H. Madjidi, and S. Negahdaripour, "3-D mapping of sea floor scenes by stereo imaging," in *Proc. Oceans'01*, Honolulu, HI, Nov. 2001, pp. 2577–2583.

[30] K. N. Leabourne, S. M. Rock, S. D. Fleischer, and R. L. Burton, "Station keeping of an ROV using vision technology," in *Proc. Oceans'97*, Halifax, NS, Canada, Oct. 1997, pp. 634–640.

[31] E. A. LeMaster, M. Matsuoka, and S. M. Rock, "Field demonstration of a Mars navigation system utilizing GPS pseudolite transceivers," in *Proc. IEEE Position, Location, Navigation Symp.*, Palm Springs, CA, Apr. 2002, pp. 150–155.

[32] H. C. Longuet-Higgins and K. Prazdny, "The interpretation of a moving retinal image," in *Proc. Royal Society of London*, vol. B-208, 1980, pp. 385–397.

[33] H. C. Longuet-Higgins, "The visual ambiguity of a moving plane," in *Proc. R. Soc. London, Series B*, vol. 223, 1984, pp. 165–175.

[34] J.-F. Lots, D. M. Lane, E. Trucco, and F. Chaumette, "A 2-D visual servoing for underwater vehicle station keeping," *Proc. IEEE Conf. Robotics Automat.*, pp. 2767–2772, May 2001.

[35] H. Madjidi and S. Negahdaripour, "Global alignment of sensor positions with noisy motion measurements," in *Proc. IEEE Int. Conf. Robotics and Automation (submitted for publication), to be held in New Orleans, LA, Apr. 26–May 1, 2004*.

[36] R. Marks, S. M. Rock, and M. J. Lee, "Real-time video mosaicking of the ocean floor," *IEEE J. Oceanic Eng.*, vol. 20, pp. 229–241, July 1995.

[37] R. Marks, H. H. Wang, M. Lee, and S. Rock, "Automatic visual station keeping of an underwater robot," in *Proc. OCEANS'94*, Brest, France, Sept. 1994.

[38] L. Matthies, E. Gat, R. Harrison, B. Wilcox, R. Volpe, and T. Litwin, "Mars microrover navigation: Performance evaluation and enhancement," *Autonomous Robots J.* , vol. 2, no. 4, pp. 291–311, 1995.

[39] E. M. Mikhail, *Observations and Least Squares*. New York: IEP, 1976.

[40] E. M. Mikhail, J. S. Bethel, J. C. McGlone, and C. McGlone, *Introduction to Modern Photogrammetry*. New York: Wiley, 2001.

[41] S. Negahdaripour, C. H. Yu, and A. Shokrollahi, "Recovering shape and motion from undersea images," *IEEE J. Oceanic Eng.*, vol. 15, pp. 189–198, July 1990.

[42] S. Negahdaripour and J. Fox, "Underwater optical stationkeeping: Improved methods," *J. Robot. Syst.*, pp. 319–338, Mar. 1991.

[43] S. Negahdaripour, L. Jin, X. Xun, C. Tsukamoto, and J. Yuh, "A real-time visionbased 3-D motion estimation system for positioning and trajectory following," *Proc. Third IEEE Workshop Applications Comp. Vision*, pp. 264–271, Dec. 1996.

[44] S. Negahdaripour, S. Zhang, X. Xun, and A. Khamene, "On shape and motion recovery from underwater imagery for 3D mapping and motion-based video compression," in *Proc. OCEANS'98*, Nice, France, Sept. 1998.

[45] S. Negahdaripour, X. Xun, A. Khamene, and Z. Awan, "3-D motion and depth estimation from sea-floor images for mosaic-based station-keeping and navigation of ROVs/AUV's and high-resolution sea-floor mapping," in *Proc. Workshop AUV Navigation*, Cambridge, MA, Aug. 1998, pp. 191–200.

[46] S. Negahdaripour, X. Xu, and L. Jin, "Direct estimation of motion from sea floor images for automatic station-keeping of submersible platforms," *IEEE J. Oceanic Eng.*, vol. 24, pp. 370–382, July 1999.

[47] S. Negahdaripour and P. Firoozfam, "Positioning and photo-mosaicking with long image sequences; comparison of selected methods," in *Proc. OCEANS'01*, Honolulu, HI, Nov. 2001, pp. 2584–2592.

[48] S. Negahdaripour and X. Xu, "Mosaic-based positioning and improved motion estimation methods for autonomous navigation," *IEEE J. Oceanic Eng.*, vol. 27, pp. 79–99, Jan. 2002.

[49] M. Ollis, H. Herman, and S. Singh, "Analysis and Design of Panoramic Stereo Vision Using Equi-Angular Pixel Cameras," Carnegie Mellon Univ., Robotics Inst., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-99–04, 1999.

[50] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone, "Stereo ego-motion improvements for robust rover navigation," *Proc. IEEE Int. Conf. Robotics Automation*, pp. 1099–1104, 2001.

[51] S. Peleg, M. Ben-Ezra, and Y. Pritch, "Omnistereo: Panoramic stereo imaging," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 3, pp. 279–290, March 2001.

[52] C. Roman and H. Singh, "Estimation of error in large area underwater photomosaics using vehicle navigation data," in *Proc. OCEANS'01*, Honolulu, HI, Nov. 2001, pp. 1849–1853.

[53] C. S. Rogers, G. Garrison, R. Grober, Z.-M. Hillis, and M. A. Franke, *Coral Reef Monitoring Manual for the Caribbean and Western Atlantic*. Virgin Islands National Park: St. John, USVI, 2001.

[54] H. S. Sawhney, S. Hsu, and R. Kumar, "Robust video mosaicing through topology inference and local to global alignment," in *Proc. Eur. Conf. Comp. Vision*, Berlin, Germany, June 1998, pp. 103–119.

[55] N. Schlegel, "Autonomous vehicle control using image processing," MSc. Thesis, Virginia Polytechnic Inst. State Univ., Blacksburg, VA, 1997.

[56] H. Singh, D. R. Yoerger, R. Bachmayer, A. M. Bradley, and W. K. Stewart, "Sonar mapping with the autonomous benthic explorer (ABE)," in *Proc. 9th Int. Symp. Unmanned Untethered Submersible Technology*, Durham, NH, 1995.

[57] R. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *Int. J. Robot. Res.*, vol. 5, no. 4, pp. 56–58, 1987.

[58] C. M. Smith, J. J. Leonard, A. A. Bennett, and C. Shaw, "Feature-based concurrent mapping and localization of autonomous underwater vehicles," in *Proc. OCEANS'97*, Halifax, NS, Canada, 1997, pp. 896–901.

[59] M. Snorrason, J. Norris, and P. Backes, "Vision based obstacle detection and path planning for planetary rovers," in *Proc. SPIE 13th Annu. AeroSense Conf.*, vol. 3693, Orlando, FL, Apr. 1999, pp. 44–54.

[60] B. Southall, T. Hague, J. A. Marchant, and B. F. Buxton, "Vision-aided outdoor navigation of an autonomous horticultural vehicle," in *Proc. 1st Int. Conf. Vision Systems (ICVS)*, Las Palmas de Gran Canaria, Spain, 1999, pp. 37–50.

[61] W. K. Stewart, "Remote-sensing issues for intelligent underwater systems," in *Proc. Conf. Computer Vision and Pattern Recognition*, Maui, HI, 1991, pp. 230–235.

[62] A. Stentz, "Robotic technologies for outdoor industrial vehicles," in *Proc. SPIE, Unmanned Ground Vehicle Technology III*, vol. 4364, G. R. G. and C. M. Shoemaker, Eds., 2001, pp. 192–199.

[63] C. R. Stoker, D. R. Barch, B. P. Hine III, and J. Barry, "Antactic undersea exploration using a robotic submarine with a telepresence user interface," *IEEE Intell. Syst.*, vol. 10, pp. 14–23, Dec. 1995.

[64] M. Tivey, D. Yoerger, A. Bradley, R. Catanach, A. Duester, S. Liberatore, and H. Singh, "Autonomous underwater vehicle maps seafloor," *EOS Trans. American Geophysical Union*, vol. 78, no. 22, pp. 229–230, June 1997.

[65] S. Tiwari, "Mosaicking of the ocean floor in the presence of three-dimensional occlusion in visual and side-scan sonar images," in *Proc. Symp. AUV Technol.*, Monterey, CA, June 1996, pp. 308–314.

[66] R. Tsai, T. S. Huang, and W. L. Zhu, "Estimating three-dimensional motion parameters of a rigid planar patch, II: Singular value decomposition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 525–534, Aug. 1992.

[67] V. Uffenkamp, "State of the art of high precision industrial photogrammetry," in *Proc. 3rd Int. Workshop Accelerator Alignment*, ch. II, Annecy, France, Sept.-Oct 1993, pp. 153–165.

[68] G. S. Van der wal, M. W. Hanson, and M. R. Piacentino, "ACADIA vision processor," *Proc. IEEE Int. Workshop Computer Architectures Machine Perception*, pp. 31–40, Sept. 2000.

[69] P. Wade, D. Moran, J. Graham, and C. B. Jackson. Robust and Accurate 3D Measurement of Formed Tube Using Trinocular Stereo Vision. [Online]. Available: http://www.bmva.ac.uk/bmvc/1997/papers/028/028.html

[70] I. M. Williams and J. H. J. Leach, "Measurements of benthic species using drop-video platforms– Comparative uses of stereo and single video systems," in *Proc. Australian Marine Science Assoc.*, Sydney, Australia, Nov. 1999, pp. 114–119.

[71] X. Xun and S. Negahdaripour, "Vision-based motion sensing for underwater navigation and mosaicing of ocean floor images," in *Proc. OCEANS'97*, Halifax, NS, Canada, Oct. 1997, pp. 1412–1417.

[72] X. Xu and S. Negahdaripour, "Automatic optical station keeping and navigation of an ROV; sea trial experiments," in *Proc. OCEANS'99*, Seattle, WA, Sept. 1999.

[73] X. Xu, "Vision-based ROV system," Ph.D. Dissertation, Univ. Miami, ECE Department, Coral Gables, FL, 2000.

[74] A. Yahja, S. Singh, and A. Stentz, "An efficient on-line path planner for outdoor mobile robots operating in vast environments," *Robot. Autonomous Syst.*, vol. 33, no. 2/3, pp. 129–143, 2000.

[75] Z. Zhu, K. D. Rajasekar, E. M. Riseman, and A. R. Hanson, "Panoramic virtual stereo vision of cooperative mobile robots for localizing 3D moving objects," in *Proc. IEEE Workshop Omnidirectional Vision*, June 2000, pp. 29–36.

[76] [Online]. Available: http://www.visualizationsoftware.com/3-Dem/vrml.html

Shahriar Negahdaripour (S'86–M'87–SM'95) received S.B., S.M., and Ph.D. degrees in 1979, 1980, and 1987 from Massachusetts Institute of Technology, Cambridge.

He joined the Electrical Engineering Department, University of Hawaii, Honolulu, in January 1987 as an Assistant Professor. In August 1991, he joined the University of Miami, Coral Gables, FL, where he is currently a Professor of electrical and computer engineering. Since the start of a project on automatic vision-based ROV station keeping in 1988, which was supported by the University of Hawaii Sea Grant College Program, he has been involved in a number of other projects on the development of various vision technologies for underwater applications, which were supported by the National Science Foundation, Office of Naval Research and the Naval Undersea Warfare Center, Newport, RI. In addition to numerous journal and conference papers in this subject area, he has presented a half-day tutorial at the IEEE OCEANS'98 Conference, Nice, France, on "Computer Vision for Underwater Applications."

Dr. Negahdaripour has organized and moderated panel sessions at IEEE OCEANS'01 and OCEANS'02 conferences on "Robust Video Mosaicking: Techniques, Standards, and Data Sets," and "Photo-Mosaicking of Underwater Imagery: Challenges and Technical Issues."

Hossein Madjidi received the B.Sc. and M.Sc. degrees in geomatics engineering from K. N. Tousi University of Technology, Iran, and is pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL.

His research interests include 3-D motion estimation and scene reconstruction from stereo imagery for sea-floor and benthic habitat mapping. He has also been involved in different areas of geomatics engineering and photogrammetry for several years.