Sensor Fusion of Structure-From-Motion, Bathymetric 3D, and Beacon-Based Navigation Modalities.

Hanumant Singh[1], Garbis Salgian[2], Ryan Eustice[1], Robert Mandelbaum[2]

[1]Woods Hole Oceanographic Institution,
Woods Hole MA

[2]Sarnoff Corporation
Princeton, NJ

Abstract

This paper describes an approach for the fusion of 3D data underwater obtained from multiple sensing modalities. In particular, we examine the combination of image-based Structure-From-Motion (SFM) data with bathymetric data obtained using pencil-beam underwater sonar, in order to recover the shape of the seabed terrain. We also combine image-based egomotion estimation with acoustic-based and inertial navigation data on board the underwater vehicle.

We examine multiple types of fusion. When fusion is performed at the data *level, each modality is used to extract 3D information* independently. *The 3D representations are then aligned and compared. In this case, we use the bathymetric data as* ground truth *to measure the accuracy and drift of the SFM approach. Similarly we use the navigation data as ground truth against which we measure the accuracy of the image-based ego-motion estimation. To our knowledge, this is the first quantitative evaluation of image-based SFM and egomotion accuracy in a large-scale outdoor environment.*

Fusion at the signal *level uses the raw signals from multiple sensors to produce a single coherent 3D representation which takes optimal advantage of the sensors' complementary strengths. In this paper, we examine how low-resolution bathymetric data can be used to seed the higher-resolution SFM algorithm, improving convergence rates, and reducing drift error. Similarly, acoustic-based and inertial navigation data improves the convergence and drift properties of egomotion estimation.*

Keywords: bathymetry, pencil-beam sonar, underwater sensing, sensor fusion, structure from motion, ego-motion, shape recovery, multi-resolution

1 Introduction

1.1 Motivation

We consider the scenario of an underwater robotic vehicle traveling through an unknown environment. The requirements for typical archaeological, biological, forensic and geological applications often call for high resolution quantitative mapping of such previously unsurveyed sites. We note that the limited view associated with optical and acoustic sensors underwater implies collecting a large number of sensor readings, and corresponding navigation readings, which are composited into a global perspective.

Thus the requirement for high resolution mapping in turn necessitates a methodology for high resolution navigation information. There are several sensing modalities which are traditionally used for this purpose. For underwater vehicles, the most common method of measuring terrain structure uses bathymetric data from acoustic sensors. For navigation, vehicles typically use acoustic transponders in combination with inertial navigation measurements.

These modalities, for both terrain mapping and navigation, have their strengths and weaknesses:

- **Bathymetry:** Bathymetric sensors underwater utilize time of flight for range-sensing while focussing the beam using an array of transducers (for transmit, receive or both) into a very tight cone. The resolution of such sensors is a function of the frequency. Moreover, there is a tradeoff between resolution (frequency) and the range as higher frequencies are attenuated much faster due to absorption in sea water. In a typical deployment the bathymetric beam is scanned over the terrain while the vehicle translates perpendicular to the direction of scanning. As pointed out earlier, accumulation of data over time requires navigation data,

acquired independently. In reality, the resolution of the bathymetric map is usually limited by the resolution of the navigation (Section 2.1 discusses this at length and includes a discussion of when this does not hold true).

- **Acoustic Navigation - Long baseline / Inertial:** Acoustic long baseline (LBL) navigation utilizes fixed transponder beacons on the ocean floor that can be interrogated from the vehicle. By calculating the time of flight from the vehicle to several beacons one can triangulate the position of the vehicle. Here too there is a fundamental tradeoff between the range of the navigation system and the resolution due to the more rapid attenuation of higher versus lower frequencies. We note also that besides lower range resolution, the larger ranges associated with lower frequencies also limit our update rate between fixes due to longer travel times associated with acoustic energy travelling between the vehicle and transponder. LBL navigation does however provide a bounded error over the entire range of operation. In addition to LBL, acoustic doppler current profilers are often used to obtain vehicle velocity information that can be integrated with vehicle attitude information to obtain high resolution navigation at higher (than LBL) update rates. However, by itself, the navigation error grows as a function of distance travelled. Typically a complementary filter is used to blend LBL and inertial navigation.

- **Stereo:** Traditional stereo systems rely on establishing correspondence between two camera images taken simultaneously. They can generate high resolution depth maps, of the order of the pixel resolution of the cameras. Accuracy can be high, but it diminishes with distance to the target object; it is also dependent on the FOV, and the baseline distance between the cameras. On the other hand, stereo is computationally expensive, and traditional real-time approaches have difficulty with regions of low image texture, and near occlusion boundaries. Also, stereo computation provides only an instantaneous range map from each location. In order to combine this stream of range data into a coherent swath of range data as the vehicle moves, the navigation of the vehicle must be taken into account in order to align the range images. Unfortunately we note that the use of LBL / inertial navigation is by itself does not provide enough accuracy to combine the range data seamlessly.

- **Structure-from-Motion:** SFM techniques [2], [5], [4] generally estimate both terrain structure and egomotion of the host vehicle simultaneously, using sequences of camera images as input. This is both good and bad. On the positive side, the estimates of structure and egomotion are self-consistent. On the other hand, this enforced self- consistency means that any errors in the egomotion estimate will have corresponding errors in the structure estimate, in order to keep consistent with the image stream. A typical example of this involves the well-known ”ambiguity” between small rotations and small translations of a camera system: it is easy to confuse the two based on image data alone. SFM methods provide high resolution range data, which is also aligned over long sequences of images. Accuracy can be high, although it is impacted by errors in egomotion and grows over time, as described above. Similarly, SFM methods theoretically offer pixel-level precision in egomotion estimates, though this is impacted by errors in shape estimation. SFM is also generally very computationally expensive.

Ideally, one would like to use a combination of these complementary sensors. However, the fusion of data from multiple modalities is a difficult problem. The first step involves alignment between the sensors, which gather data at asynchronous times, and have differing characteristics. Once the data has been aligned, the problem of providing the most accurate estimate of range and egomotion based on the multiple, possibly conflicting, inputs is another unsolved area of research (estimation theory).

This paper addresses the problem of fusing a computer-vision SFM algorithm with other sensor modalities to recover both egomotion and structure of the environment. For this purpose, we selected an iterative, multi-resolution SFM algorithm described in the literature [3]. The multi-resolution aspect allowed combination with sensing modalities of differing resolution. The iterative nature of the algorithm allowed us to fuse in other modalities by injecting information from these other sensors at each iteration. In this way, we ”guided” the SFM algorithm to converge on the consistent solution which best matched with the other modalities.

Our focus is on underwater vehicles, and in particular the modalities of bathymetry (for shape information), high frequency LBL navigation data (for egomotion) and SFM structure and egomotion.

1.2 Levels of Sensor Fusion

In this paper we examine multiple levels of sensor fusion.

1.2.1 Data-level fusion

The simplest form of fusion involves allowing each sensor modality to operate independently, and then to combine the data produced as a final step. We term this *fusion at the data level*. In our case, an example algorithm would be:

1. Run SFM on the image sequence. Use iterative, multi-resolution approach to converge on a consistent solution to both both egomotion and structure in the scene.

2. Gather navigation data about the vehicle's position over time using LBL data.

3. Gather bathymetric acoustic data using a combination of the LBL navigation data (position) and pencil-beam bathymetric sonar giving a "scanline" of data per cycle.

4. Align the egomotion from SFM with the navigation data.

5. Align the SFM shape data with the bathymetry.

6. Compare or combine the results.

We have implemented this approach, as described in section 3.1. In our case, step 6 of the above approach consisted of comparing the SFM with the bathymetry results in order to evaluate the performance of the SFM approach. Hence, we assumed the acoustic data to be ground truth. This is a reasonable assumption, since though sonar produces low resolution data, the accuracy of the range information is independent of the distance to the terrain. Further, the sensor error does not accumulate over time, since the beacon-based navigation data as pointed out earlier is bounded over the entire site.

1.2.2 Signal-level fusion

A more complex version of sensor fusion combines information from multiple sensors *on an ongoing basis* to provide a single coherent representation. We term this *fusion at the signal level*.

Signal-level fusion is particularly useful in the case of SFM. Consider, for instance, SFM without external information from other sensing modalities: In general, the problems of estimating egomotion and structure from image sequences are mutually dependent. Prior accurate knowledge of egomotion allows structure to be computed by triangulation from corresponding image points. This is the principle behind standard parallel-axis stereo algorithms, where the baseline is known accurately from calibration. In this case, knowledge of the epi-polar geometry provides for efficient search for corresponding points.

On the other hand, if prior information is available regarding the structure of the scene, then egomotion can be computed directly. Essentially, one considers the space of all possible poses of the camera. One then searches for the pose for which the perspective projection of the environment onto the image plane most closely matches the actual image obtained.

When neither accurate egomotion nor structure information is available, a classical chicken-and-egg problem exists: We need egomotion to estimate good structure, and we need shape information to estimate good egomotion. To solve this problem, we selected a correlation-based algorithm in the literature which assumes a very coarse starting-point for both egomotion and structure, and then alternatively and iteratively refines the estimates of both. The updated estimate of egomotion is used to obtain an improved estimate of structure, which in turn is used to refine the estimate of egomotion. The algorithm converges on a solution which provides consistency between egomotion, recovered shape, and the image sequence.

Non-uniqueness: Note that the solution obtained by such a convergent approach is guaranteed to be consistent, but not necessarily unique. For example, a well-known ambiguity in SFM exists between small rotations and small translations of the camera. When imaging a distant scene, a small translation to the right induces a uniform flow-field, with flow vectors pointing leftwards, and all parallel to each other. On the other hand, a small pan of the camera to the right causes a very similar flow field. The only difference between the two flow fields occurs at the top and bottom of the fields: in the case of panning, the flow vectors are not quite parallel, but rather lie on slightly curved lines. However, for small motions, and regular FOV, the difference is very difficult to measure.

In the case of SFM, the problem is reversed. Given the image sequences, one can compute the flow fields. One is then left with the problem of inferring the changing camera pose, i.e. whether the camera indeed panned or translated. Since the difference between the two flow fields is below the noise level of the flow-field estimation process, the two egomotions cannot be distinguished accurately.

Such errors in egomotion propagate into the estimation of shape, since at all times consistency between shape, egomotion, and imagery must be maintained. Thus, multiple consistent solutions are possible. In our above example, one solution may describe a a camera which is translating and slowly panning over a curved surface, while another solution describes a non-rotating camera panning over a flat surface. Both solutions would be consistent with the input imagery.

Obtaining the "correct" solution:

In this paper, we use other sensor modalities to force SFM to converge to the solution which best matches the other sensor data. As a result, the SFM solution for shape becomes consistent with the bathymetric solution, and the resolution of the recovered surface is vastly enhanced.

In summary, the bathymetry is used to constrain the SFM algorithm to converge on a solution which matches actual real-world shape, and the SFM results are then used to greatly enhance the resolution of the recovered terrain. The fusion of the two complementary modalities is better

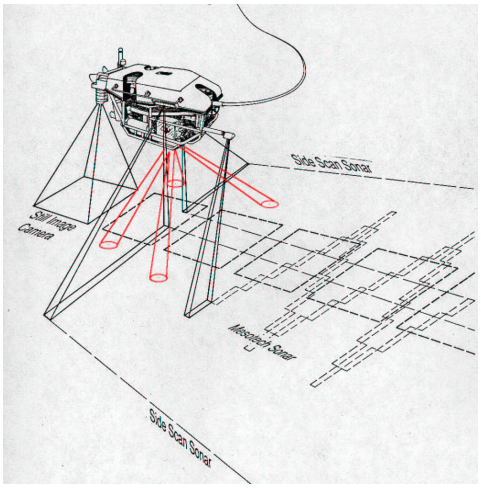

Figure 1: Vehicle and sensor footprint of the bathymetric sonar. During survey operations the vehicle is driven very slowly so that the consecutive scans alongtrack are closely spaced for maximum redundancy

than either one could produce individually.

1.3 Contributions

The contributions of this paper are:

1. To our knowledge, this paper represents the first quantitative evaluation of image-based SFM and egomotion accuracy in a large-scale outdoor environment.

2. The paper illustrates how SFM can be directed to converge on a consistent solution which is "close" to the correct solution, by incorporating data from complementary sensing modalities.

3. A signal-level fusion of data from complementary modalities on a large- scale outdoor real-world scene.

2 Background

The configuration of the vehicle, camera, and side-scan sonar is shown in figure 1.

2.1 High resolution 3D bathymetric mapping

Sonar sensors capable of cm level resolution in underwater applications have existed for decades, but our ability to generate self-consistent maps at these resolutions has until recently been limited by the lack of comparable navigation accuracy.

The Imagenex 675 Khz pencil beam sonar [1] used in collecting the data for this paper, for example, has been available commercially for over 15 years and has a range resolution of 1 cm and a beam angle to the 3dB down point of 1.5 degrees. The navigation resolution of acoustic long baseline (LBL) systems which are typically used for XY navigation underwater on the other hand, is of the order of 1-10m. Further, the update rates for the sonar versus the navigation are also significantly different (10 Hz as opposed to 0.1 Hz respectively).

Recent advances in navigation [8] - the use of higher frequency (300 kHz) LBL systems that can provide cm level precision taken in combination with bottom lock acoustic doppler current profilers (ADCPs) that provide velocity estimates at high update rates - have yielded XY estimates of the order of sensor precision and at comparable update rates.

Theoretically, the construction of a bathymetric map is the simple process of compensating for the coordinate transformations that convert data in a sensor frame (range and angle from the sensor) to a vehicle coordinate frame and then to a world referenced coordinate system.

Under the assumption of perfect information the bathymetric survey can be expressed by the equations

$$p_v = S \cdot p_s \tag{1}$$

$$p_w = V \cdot p_v = V \cdot S \cdot p_s \tag{2}$$

where p_s, p_v, and p_w are the individual sonar sensor readings (ping) coordinates in the sensor, vehicle and world coordinate frames respectively as expressed in homogeneous $[4 \times 1]$ coordinates.

S is the $[4 \times 4]$ homogeneous coordinate transformation matrix which relates the sensor to vehicle frame and V is the $[4 \times 4]$ homogeneous coordinate transformation matrix which relates the vehicle to world coordinate frame.

The results of simply applying these equations to data acquired from five overlapping passes (Figure 2) is shown in Figure 3. These results are seen to be inconsistent over the different passes. The inconsistency of these results has been shown to be a function of the small calibration biases that occur due to the distributed nature of the attitude sensors across the vehicle [6].

If we consider the inexact estimates of the S and V transforms, the transformation of pings to world coordinates has errors

$$\hat{p}_w = \hat{V} \cdot \hat{S} \cdot p_s \tag{3}$$

Since each transform has 6 degrees of freedom, it would seem that there are 12 parameters to determine. However,

$$p_w = V \cdot S \cdot p_s \tag{4}$$

can also be expressed as

$$p_w = \hat{V} \cdot (\hat{V}^{-1} \cdot V) \cdot (S \cdot \hat{S}^{-1}) \cdot \hat{S} \cdot p_s \tag{5}$$

where $(\hat{V}^{-1} \cdot V)$ is the transform from the world coordinate frame to the approximate vehicle frame and $(S \cdot \hat{S}^{-1})$ is the transform from approximate vehicle frame to ideal vehicle frame.

Defining Δ,

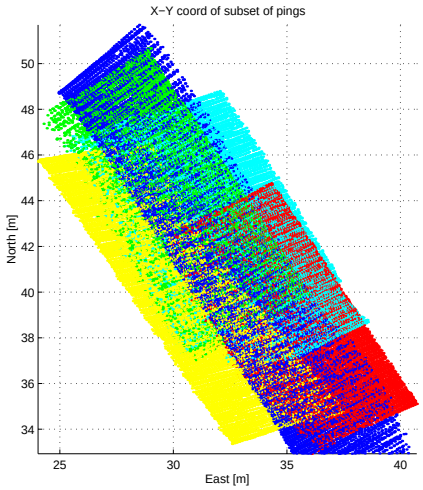

Figure 2: The footprints of five overlapping bathymetric swaths. Overlapping redundant data is the most powerful technique utilized underwater to examine self consitency and thus the accuracy of a map.

$$\Delta = (\hat{V}^{-1} \cdot V) \cdot (S \cdot \hat{S}^{-1}) \tag{6}$$

we point out that the real world coordinates are given by:

$$p_w = \hat{V} \cdot \Delta \cdot \hat{S} \cdot p_s \tag{7}$$

Moreover, since Δ is a transformation matrix, it has only 6 DOF and not 12 as might have seemed originally. Δ takes the form:

$$\Delta = \begin{bmatrix} & \Delta\alpha & & \Delta x \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

where $\Delta\alpha[3 \times 3]$ is the attitude bias of the sensor frame with respect to the vehicle frame and $\Delta x[3 \times 1]$ is the position bias of the sensor frame relative to the vehicle frame.

It has been shown [6] that detailed survey maneuvers can be performed that allow the estimation of Δ. The results of estimating and compensating for Δ are shown in Figure 3 which are seen to be consistent to the limits (5cm) at which this data was gridded.

2.2 Structure-From-Motion

We use the described in [3] to recover terrain shape and camera motion from a video sequence. The input to the algorithm consists of the camera-to-world transformation for the first reference frame H_{0w}, the focal length and some coarse ego-motion and shape estimates. Note that that these could be *very* coarse (such as a fronto-parallel plane at infinity for shape or zero ego-motion). The output is a list L of 3D points in world coordinates (initially empty) and the camera-to-world transformations H_{iw} for every frame i.

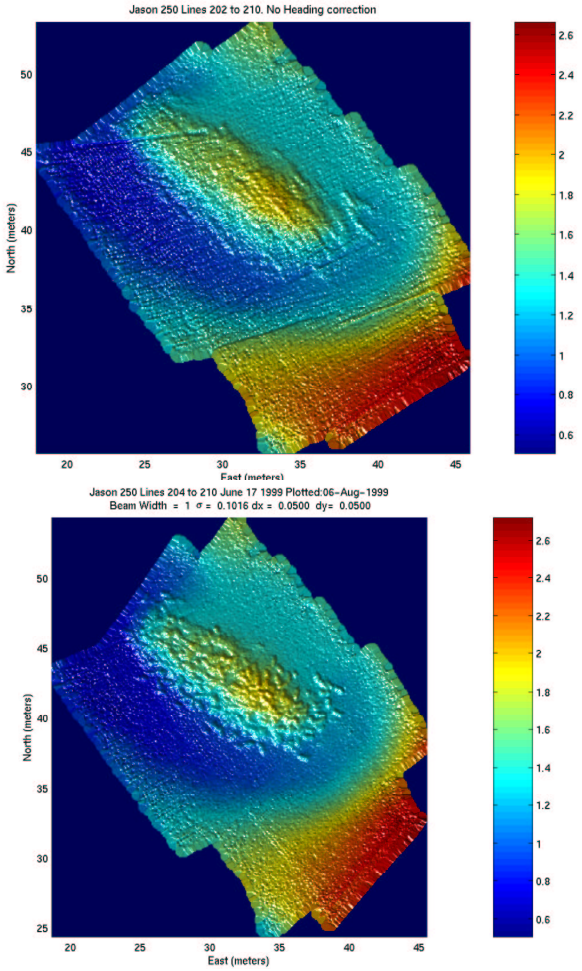

Figure 3: A comparison of bathymetric mapping before and after compensating for attitude biases. The heading bias tends to smear out individual features alongtrack while the roll bias and depth offsets introduce linear discontinuities and smearing perpendicular to the direction of travel. This site is a Phoenician shipwreck dating to 750 B.C. off of the coast of Israel in approximately 400 meters of water depth

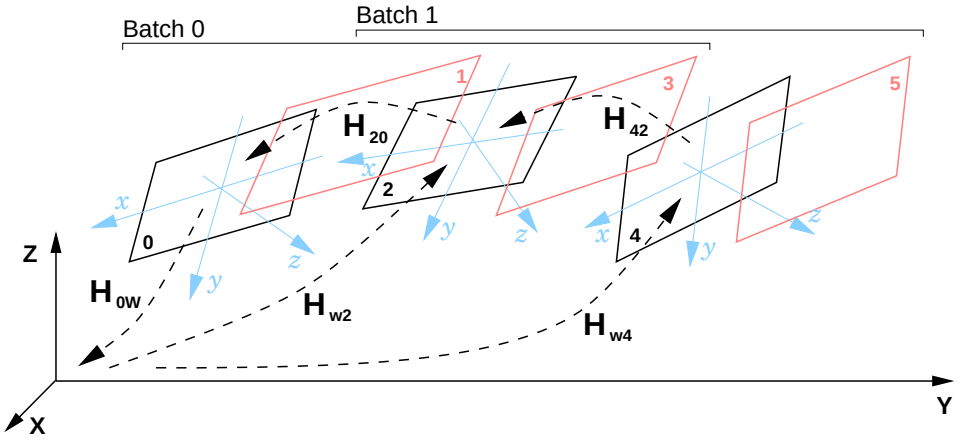

Figure 4: Several frames from a longer sequence, with their respective coordinate systems (xyz). The world coordinate system is denoted (XYZ). H_{ij} is the homogeneous transformation between coordinate systems i and j.

The input sequence is processed in batches consisting of a few consecutive frames, with at least one frame overlap between consecutive batches. The algorithm repeats the following steps:

- Process the current batch, with reference frame r, as described in [3]. The result is a dense depth map D_r in a frame r centered coordinate system and ego-motion relative to the reference frame for every inspection image i in the current batch ($\mathbf{\Omega}_{ir}$, $\mathbf{T}_{ir}$).
- Project every point in the current depth map into the world coordinate system and add it to the list of 3D points:
$$L = L \cup \{H_{rw}\mathbf{d} \mid \mathbf{d} \in D_r\}$$
- If this is the last batch, stop.
- Let k denote the reference frame of the next batch. From H_{rw}, $\mathbf{\Omega}_{kr}$, $\mathbf{T}_{kr}$ compute camera-to-world and world-to camera transformations H_{kw}, H_{wk} for I_k.
- Project the points in L that are visible in I_k into D_k (the initial depth estimate for the next batch) and remove them from L:
$$L' = \{\mathbf{t} \in L \mid visible(\mathbf{t}, I_k, f)\}$$
$$D_k = \{H_{wk}\mathbf{t}' \mid \mathbf{t}' \in L'\}$$
$$L = L \setminus L'$$

After processing the last batch, L is a list of dense 3D points in the world coordinate system. For visualization, the points are projected on a plane and Delaunay triangulation is used to generate a mesh suitable for texture mapping. No additional parametric surface fitting is used.

One obvious issue in comparing the bathymetric and SFM range data is the choice of origin. Even a small offset between the position and the orientation of the origin in the two datasets could lead to errors that are of the same order as produced by the SFM algorithm itself. Thus to align the two independent datasets we chose to manually pick common features across the datasets and to fit a single affine transformation across these common features. This resulted in a common origin and orientation with respect to which we could make comparisons.

3 Experiments

3.1 Performance evaluation: Roman sequence

The first example is using data collected at the site of a Roman shipwreck. Figure 5 shows the height maps obtained from bathymetry (left) and SFM on a sequence of 11 frames. Figure 6 shows a texture-mapped rendering of the terrain. No navigational data was used in the SFM algorithm. Since individual amphorae are relatively isolated, this sequence allows us to hand-select alignment points between bathymetry and structure obtained using the SFM algorithm.

We selected a small number of points (12) in the two 3D structures and computed a rigid body transformation that brings the SFM data into alignment with the bathymetry. Figure 7 shows a cross section through the two surfaces (the SFM surface is sampled at the same resolution as the bathymetry). Note that the overall terrain configuration has been correctly recovered by SFM.

3.2 Performance evaluation: Phoenician sequence

The second example is for a longer sequence (56 frames) at the site of a Phoenician shipwreck. Figure 8 (bottom) shows the terrain structure obtained from bathymetry and the camera position over time, as reported by the navigation system. The terrain is presented as a 3D mesh with false-color coding of height. The left part shows a top-down view, and the right side a “side” view, which illustrates the terrain configuration. Unit axes are meters (the negative values for the Z axis represent depth re mean sea level).

The top row shows results from the SFM algorithm.

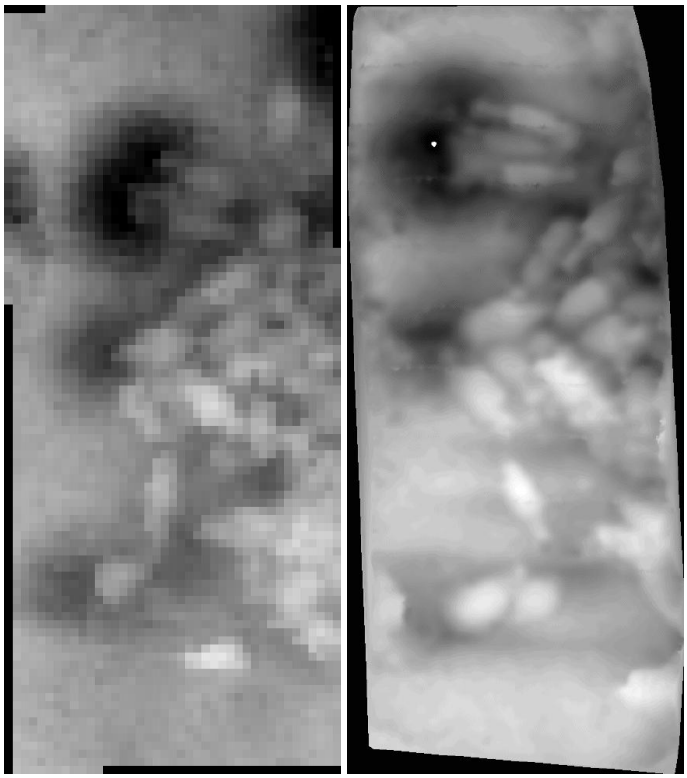

Figure 5: Height map recovered from bathymetry (left) and SFM (right). The difference in resolution is easily notable

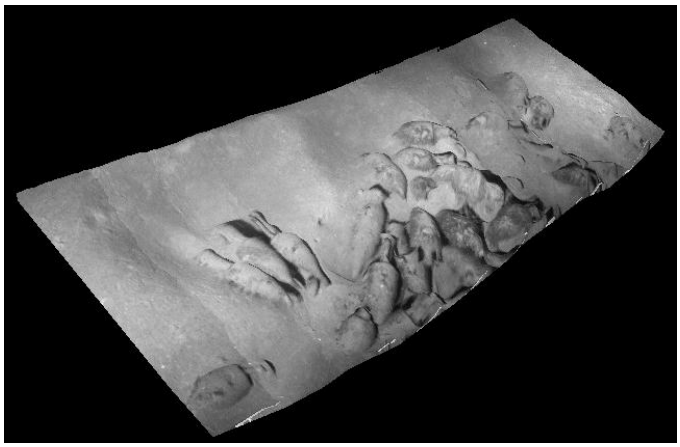

Figure 6: Texture-mapped rendering of the terrain recovered from SFM.

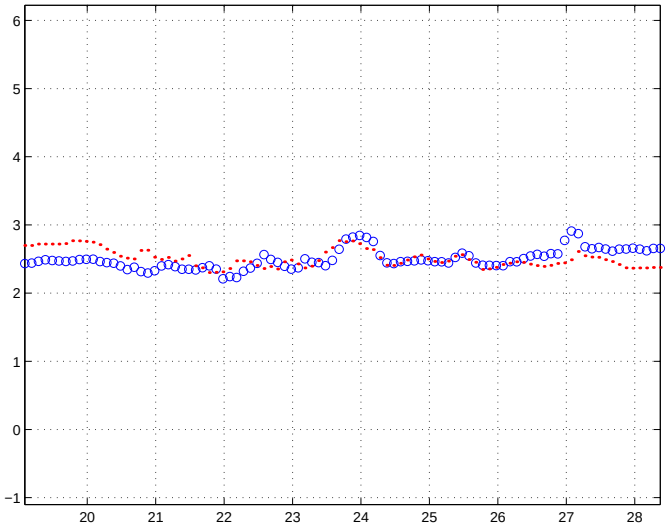

Figure 7: Comparison of 3D recovered from image-based SFM (red dots) and bathymetry (blue circles). The units are meters for both axes

The sequence of 56 frames was processed in consecutive batches (2 frames at a time). Information about camera motion was provided for the first batch, in order to get the right scale factor. One can see from the results that the algorithm "drifts" over time. While local structure is correctly recovered, the globa shape of the terrain is not. This is mainly due to the small rotation vs. small translation confusion (as discussed in the introduction).

The center row shows the results of the SFM algorithm, with navigational data provided as initial ego-motion estimates for every batch. The overal drift has been eliminated.

4 Conclusion

We presented examples of multi-modal data combination for recovering high-resolution terrain structure over extended areas. By using navigational data to constrain a structure from motion algorithm, we showed that the "drift" of the SFM algorithm over long video sequences can be greatly reduced.

A known solution for obtaining globally correct shape is to use bundle adjustment ([7]) as a final step in SFM. However, the bundle adjustment step is time consuming, and can only be applied after all the data has been processed. By using navigational data as initial estimates for SFM, results can be obtained continuously, as the vehicle moves through the environment.

References

[1] Imagenex Technology Corporation. http://www.imagenex.com, May 2001.

[2] E. Krotkov and R. Hoffman. Terrain mapping for a walking planetary rover. *RA*, 10:728–739, 1994.

[3] R. Mandelbaum, G. Salgian, and H. Sawhney. Correlation-based estimation of ego-motion and structure from motion

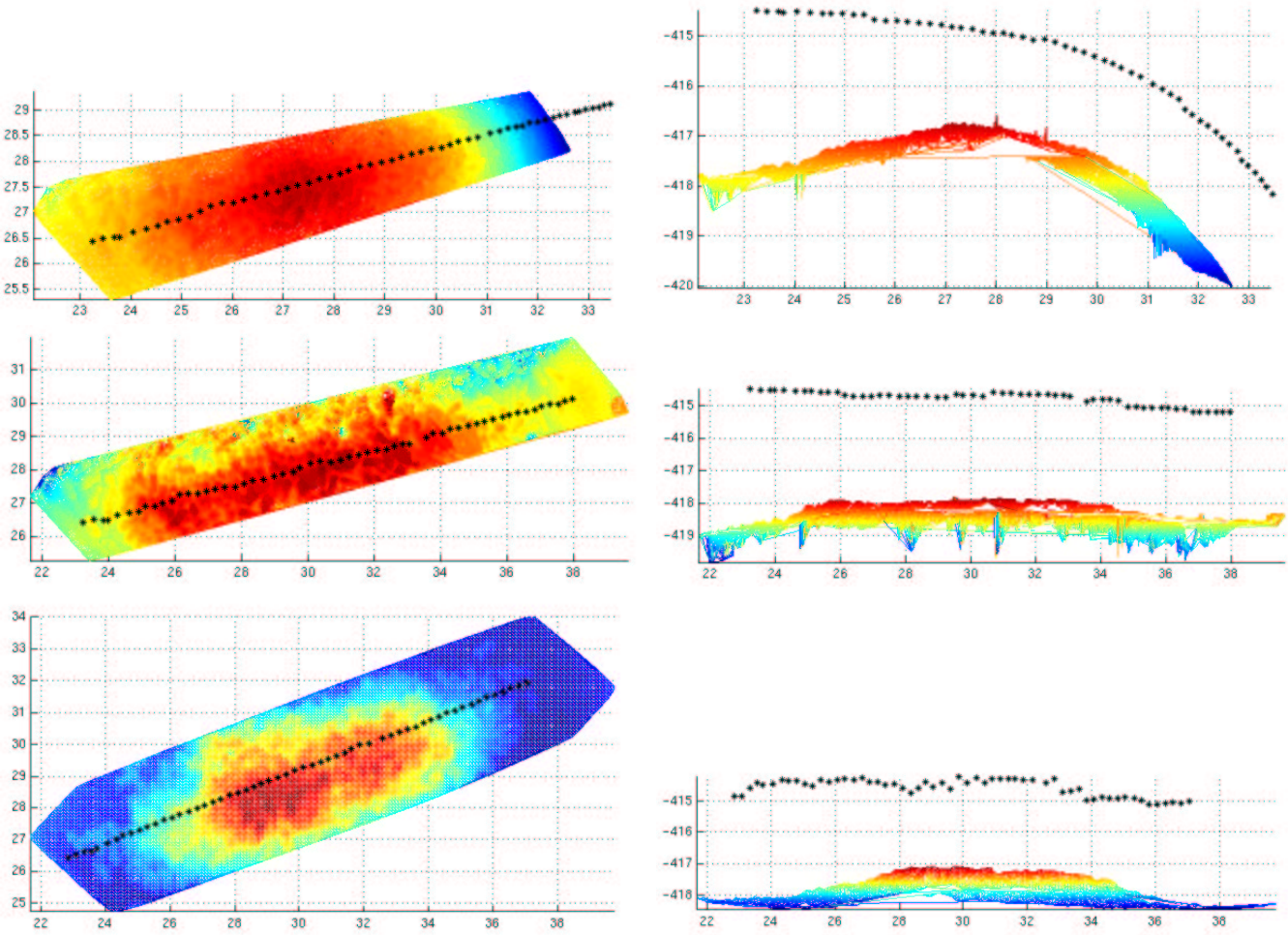

Figure 8: Comparison of 3D structure obtained from side-scan sonar and image-based SFM. Top and side view (*left* and *right*) of the reconstructed surface. *Top:* reconstruction fro SFM only. *Middle:* SFM using nav data as initial estimates. *Bottom:* Bathymetry. Camera positions are plotted on top.

and stereo. In *Proceedings of the 7th International Conference on Computer Vision*, volume 1, pages 544–550, Kerkyra, Greece, September 1999.

[4] R. Mandelbaum, G. Salgian, H. Sawhney, and M. Hansen. Terrain reconstruction for ground and underwater robots. In *International Conference on Robotics and Automation*, volume 1, San Francisco, USA, April 2000.

[5] C. Olson, L. Matthies, M. Schoppers, and M. Maimone. Robust stereo ego-motion for long distance navigation. In *CVPR00*, pages II:453–458, 2000.

[6] H. Singh, L. Whitcomb, D. Yoerger, and O. Pizarro. Microbathymetric mapping from underwater vehicles in the deep ocean. *Computer Vision and Image Understanding*, 79(1):143–161, July 2000.

[7] B. Triggs, P.McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Vision Algorithms: Theory and Practice*, 2000.

[8] L. Whitcomb, D. Yoerger, H. Singh, and J. Howland. Advances in underwater robot vehicles for deep ocean exploration: Navigation, control and survery operations. In *In Robotics Research - The Ninth International Symposium*, page to appear, Springer-Verlag, London, 2000.