Integration of Motion Cues in Optical and Sonar Videos for 3-D Positioning

S. Negahdaripour, H. Pirsiavash, H. Sekkati
Electrical and Computer Engineering Department
University of Miami
Coral Gables, FL 33124-0640
shahriar(h.pirsiavash)@(umiami)miami.edu

Abstract

Target-based positioning and 3-D target reconstruction are critical capabilities in deploying submersible platforms for a range of underwater applications, e.g., search and inspection missions. While optical cameras provide high-resolution and target details, they are constrained by limited visibility range. In highly turbid waters, target at up to distances of 10s of meters can be recorded by high-frequency (MHz) 2-D sonar imaging systems that have become introduced to the commercial market in recent years. Because of lower resolution and SNR level and inferior target details compared to optical camera in favorable visibility conditions, the integration of both sensing modalities can enable operation in a wider range of conditions with generally better performance compared to deploying either system alone.

In this paper, estimate of the 3-D motion of the integrated system and the 3-D reconstruction of scene features are addressed. We do not require establishing matches between optical and sonar features, referred to as opti-acoustic correspondences, but rather matches in either the sonar or optical motion sequences. In addition to improving the motion estimation accuracy, advantages of the system comprise overcoming certain inherent ambiguities of monocular vision, e.g., the scale-factor ambiguity, and dual interpretation of planar scenes. We discuss how the proposed solution provides an effective strategy to address the rather complex opti-acoustic stereo matching problem. Experiment with real data demonstrate our technical contribution.

1. Introduction

Underwater search and inspection of manmade structures is part of routine maintenance as well as homeland security operations. Cost, efficiency and eliminating risk to human divers calls for automated technologies that rely on the deployment of unmanned submersible vehicles equipped with imaging systems. While optical cameras give high-detailed images of target surfaces in good visibility conditions, these sensors can become ineffective in many harbor waters and other highly turbid environments.

Recent physics-based imaging systems and technique targeted for fog and haze offer some improvement in turbid underwater conditions [20, 21]. Also, laser scanning systems provide increased operational range [15], and integration with traditional close-range photogrammetry methods previously explored for terrestrial application may address certain challenges due to resolution, accuracy, speed and other operational requirements [5]. However, optical systems do not match the performance and range of *2-D high-frequency imaging sonar systems* that readily penetrate silt and mud [4, 11, 22]. The strategy to fuse visual cues from optical and sonar images can potentially provide enhanced 3-D reconstruction performance in comparison to the utilization of each sensing modality alone, and would generally extend the operational range. This strategy has been applied to *3-D sonar* and optical cameras, to register 3-D sonar data with known 3-D object models [8, 6], and to track seafloor natural contours for navigation [3]. More recently, deployment of 2-D sonar and optical cameras in stereo configuration has been proposed as a new paradigm in 3-D reconstruction, however, only a theoretical treatment of the epipolar geometry has been presented [16].

Estimation of 3-D motion and target structure is highly sought in target-based positioning and 3-D object reconstruction and (or) recognition in search and inspection of manmade structures (pipelines, ships, dams, bridges, etc.). Monocular motion problem has been studies extensively in the computer vision literature, and application for sonar video cameras is a natural extension. While either sensing modality can be deployed individually under favorable environmental [12], *integration of visual motion cues in optical and sonar images* - as proposed in this paper – can provide a solution for a larger range of medium conditions.

Deployment of an opti-acoustic system in stereo configuration for 3-D reconstruction requires establishing matches

in optical and stereo images, referred to as the *opti-acoustic correspondence* problem [16]. This undoubtedly is the most difficult issue in developing a suitable 3-D target reconstruction technique, as the image formation processes of the two sensing modalities follow different physical principles. Our method based on visual motion cues circumvent this problem by exploiting the epipolar geometries of the optical and sonar sequences, dealing with the correspondence problem in each of the optical and sonar motion pairs, separately. In addition to improved estimation, the integration of motion cues from the two sensing modalities allows us to overcome certain inherent ambiguities of monocular vision, e.g., up to scale 3-D reconstruction, and dual interpretation of planar surfaces. We have devised an efficient method based on the hierarchical estimation of 3-D motion for planar surfaces. As it becomes evident, the assumption of scene planarity can be readily relaxed by computing the essential matrix instead, with more matches. While not a topic of this paper, we discuss how the solution proposed here can provide an effective approach to address the opti-acoustic correspondence problem by propagating certain seed matches along epipolar lines [7, 13]. Effectiveness of the proposed approach is demonstrated based on results from experiments with real data.

1.1. Overview

As stated, proposal to utilize an opti-acoustic stereo imaging for 3-D scene reconstruction requires solving the correspondence problem, which is yet to be solved [16]. As explained below, our solution here can establish the preliminary step in addressing this difficult correspondence problem, by the integration of visual motion cues in sonar and optical video imaging systems. This paper is concerned with all but the last step of the following process:

1. Features are more abundant in optical images, and we can generally identify scene points that (roughly) lie on some dominant plane Π_g. With a minimum of 4 correspondences of co-planar points, we can fix the underlying projective transformation, which can be decomposed into the rigid body motion $\{\mathbf{R}_o, \mathbf{t}_o\}$ of the camera and the normal vector $\mathbf{n}_o$ of Π_g, up to the well-known scale factor ambiguity of monocular video. We are not restricted to planar surfaces, as we can alternatively deploy *5 points* in arbitrary configurations [19]. In this case, we make use of the essential matrix, rather than the projective homography of planar points, and decomposition to $\{\mathbf{R}_o, \mathbf{t}_o\}$ up to scale-factor ambiguity;

2. We utilize information from the sonar sequence to resolve the scale factor ambiguity. This does not require explicit matching between optical and sonar images, only a minimum of one match in the sonar sequence;

3. We can now determine the 3-D positions of points on the plane Π_g, in both views. In fact, having established the epipolar geometry of the motion sequences, we can construct any 3-D point from 2 projections in either the sonar or optical views. However, the estimated motion is not optimal, as we have not fully utilized all available information.

4. We subsequently solve an optimization problem, utilizing optical and sonar matches $\{\mathbf{p}_o, \mathbf{p}'_o\}$ and $\{\mathbf{p}_s, \mathbf{p}'_s\}$, respectively, to update the motion and plane parameters;

5. Reconstruction of 3-D points can now be carried out by triangulation with matched features in the optical and sonar views. Furthermore, we can readily establish optic-acoustic matches by reprojection onto the 4 views. While not the scope of this paper, various strategies can be adopted to propagate these sparse matches. For example: 1) Regions within sets of nearby 3-D triplets $\{\mathbf{P}_i, \mathbf{P}_j, \mathbf{P}_k\}$ can be approximated by planar patches and reprojected onto the four views, initializing a local search with small regions to refine the match. The epipolar geometry of the stereo system, derived in [16], can be applied to restrict the problem to a 1-D search.

In this work, the scope has been limited to planar scenes since we can carry out experiments that allow us to readily assess accuracy; i.e., position/pose of the reconstructed plane.

2. Preliminaries

2.1. Notation

$\mathbf{P} = (X, Y, Z)^T$ and $\tilde{\mathbf{P}} = (\mathbf{P}^T, 1)^T$ denote a 3-D point P and its homogeneous coordinates. Coordinates in the sonar and optical coordinate systems are denoted $\mathbf{P}_s$ and $\mathbf{P}_o$. Perspective projection of P onto the optical image is $\mathbf{p}_o = (x, y)$, with homogeneous coordinate $\tilde{\mathbf{p}}_s = \lambda(\mathbf{p}_o, 1)^T$ ($\|\lambda\| > 0$):

$$\tilde{\mathbf{p}}^i = \mathbf{C}^i \tilde{\mathbf{P}}_o \tag{1}$$

where $\mathbf{C}^i$ is the 3×4 camera matrix of view i. For the sonar view, we employ the spherical coordinates $[\theta, \phi, \Re]^T$, comprising range, azimuth angle and elevation angle:

$$\mathbf{P}_s = \begin{bmatrix} X_s \\ Y_s \\ Z_s \end{bmatrix} = \Re \begin{bmatrix} \cos\phi \sin\theta \\ \cos\phi \cos\theta \\ \sin\phi \end{bmatrix}. \tag{2}$$

Inverse of this transformation is

$$\theta = \tan^{-1}\left(\frac{X_s}{Y_s}\right), \quad \phi = \tan^{-1}\left(\frac{Z_s}{\sqrt{X_s^2 + Y_s^2}}\right), \tag{3}$$

and $\Re = \sqrt{X_s^2 + Y_s^2 + Z_s^2}$.

2.2. Sonar to Optical Transformation

The transformation between the coordinate systems of the sonar and optical cameras are defined in terms of 3×3 rotation matrix $\mathbf{R}$ and 3-D translation vector $\mathbf{t}$:

$$\mathbf{P}_s = \mathbf{R}\mathbf{P}_o + \mathbf{t} \tag{4}$$

The transformation parameters, namely, rotation matrix and translation vector can be determined by external calibration of the opti-acoustic system [17].

2.3. Motion Transformation

Let $\{\mathbf{R}_o, \mathbf{t}_o\}$ and $\{\mathbf{R}_s, \mathbf{t}_s\}$ denote the transformation between the coordinate systems of the optical and sonar cameras in two viewpoints of the opti-acoustic system. These allows us to write

$$\mathbf{P}'_k = \mathbf{R}_k \mathbf{P}_k + \mathbf{t}_k \quad k = \{s, o\} \tag{5}$$

where $(.)'$ denotes coordinates in the second view.

Clearly, $\{\mathbf{R}_o, \mathbf{t}_o\}$ and $\{\mathbf{R}_s, \mathbf{t}_s\}$ are not independent, as the motion of the sonar and optical cameras are related by the rigidity constraint. We can readily show that

$$\{\mathbf{R}_s, \mathbf{t}_s\} = \left\{\mathbf{R}\mathbf{R}_o\mathbf{R}^T, (\mathbf{I} - \mathbf{R}\mathbf{R}_o\mathbf{R}^T)\mathbf{t} + \mathbf{R}\mathbf{t}_o\right\} \tag{6}$$

2.4. Plane Representation

We utilize the projective transformation of points on a single plane Π_g. A point P on Π_g with normal $\mathbf{n}$ satisfies the equation $\mathbf{P}_x \cdot \mathbf{n}_x = -1$, where $x = \{s, o\}$ allows us to use either the optical or sonar coordinate system as the reference frame. We can readily show that

$$\mathbf{n}_s = \left(\frac{1}{1 - \mathbf{t}^T \mathbf{R}\mathbf{n}_o}\right) \mathbf{R}\mathbf{n}_o \tag{7}$$

2.5. Image Measurements

Denoting 3×4 camera matrices $\{\mathbf{C}, \mathbf{C}'\}$ of two optical views, perspective projections of a 3-D scene point P

$$\begin{aligned} \tilde{\mathbf{p}} &= \mathbf{C}\tilde{\mathbf{P}}_o \\ \tilde{\mathbf{p}}' &= \mathbf{C}'\tilde{\mathbf{P}}'_o \end{aligned} \tag{8}$$

give the optical matches in two views. The corresponding sonar measurements comprise the range and azimuth angles, given in (2). It is more suitable to work with 2-D point $\mathbf{p}_s = (x_s, y_s) = \Re(\sin\theta, \cos\theta)$ as sonar measurements. Therefore, quadruplet set $\{\mathbf{p}_o, \mathbf{p}'_o, \mathbf{p}_s, \mathbf{p}'_s\}$ comprises the opti-acoustic matches of a 3-D point P in two positions of the integrated system.

The epipolar geometry of the opti-acoustic system is fixed by extrinsic calibration of the two cameras [16]. We next address how to establish the epipolar geometries for the two views of each of the sonar and optical cameras.

2.6. Epipolar Geometry of Opti-Acoustic Stereo System

As emphasized, our theoretical results requires neither the overlapping views of the integrated system, nor exploiting the stereo epipolar geometry. More precisely, we do not assume or take advantage of opti-acoustic correspondences between features in the optical and sonar views. Rather we utilize matches in two consecutive views of each camera. However, our method depends on the exterior calibration of the opti-acoustic system, and thus it is useful to borrow some results from the epipolar geometry of an opti-acoustic stereo system [16].

For a feature $\mathbf{p}_o$ in the optical image, its match lies on the epipolar contour given by

$$\Re^2 D(\theta) - N(\theta) = 0; \tag{9}$$

$$\begin{aligned} D(\theta) &= ((u_{31}u_{12} - u_{32}u_{11})\sin\theta + (u_{31}u_{22} - u_{32}u_{21})\cos\theta)^2 \\ N(\theta) &= (u_{31}\sigma_2 - u_{32}\sigma_1)^2 + \\ &\quad ((u_{12}\sigma_1 - u_{11}\sigma_2)\sin\theta + (u_{22}\sigma_1 - u_{21}\sigma_2)\cos\theta)^2 \end{aligned}$$

$$\begin{gathered} u_{k1} = y\, r_{k3} - r_{k2}, \quad u_{k2} = x\, r_{k3} - r_{k1} \quad (k = 1, 2, 3) \\ \sigma_i = t_x u_{1i} + t_y u_{2i} + t_z u_{3i} \quad (i = 1, 2) \end{gathered}$$

This can be readily transformed to the rectangular form $\mathbf{p}_s = (x_s, y_s) = \Re(\sin\theta, \cos\theta)$.

2.7. Epipolar Geometry of Motion Sequences

2.8. Optical Views

Assume N_o correspondences $\{\mathbf{p}_o, \mathbf{p}'_o\}$ in the optical motion sequence, where $N_{op} > 4$ are the projections of non-collinear co-planar scene points. It is well-known that each correspondence satisfies an up-to-scale projective transformation

$$\mathbf{p}'_o \approxeq \mathbf{Q}_o \mathbf{p}_o \tag{10}$$

The up-to-scale 3×3 projective transformation matrix $\mathbf{Q}_o$ is fixed with the minimum of 4 correspondences [14]. As stated, our results can be readily generalized to non-planar points, utilizing a minimum of 5 points that fix the underlying essential matrix [19].

2.9. Sonar Views

The transformation of scene points in the sonar reference frame are given by:

$$\mathbf{P}'_s = \mathbf{Q}_s \mathbf{P}_s; \quad \mathbf{Q}_s = (\mathbf{R}_s + \mathbf{t}_s \mathbf{n}_s^T) \tag{11}$$

It can be readily shown that sonar image points $\mathbf{p}_s = (x_s, y_s)$ satisfy the transformation:

$$\mathbf{p}_s = \mathbf{H}_s \mathbf{p}'_s \quad \mathbf{H}_s = \begin{bmatrix} \boldsymbol{\alpha} q_{11} & \boldsymbol{\alpha} q_{12} & \beta q_{13} \\ \boldsymbol{\alpha} q_{21} & \boldsymbol{\alpha} q_{22} & \beta q_{23} \end{bmatrix} \tag{12}$$

where q_{ij} denote the elements of $\mathbf{Q}_s$, $\boldsymbol{\alpha} = \cos\phi / \cos\phi'$ and $\beta = \Re \sin\phi / \cos\phi'$.

3. Computing Epipolar Geometry of Motion Sequences

3.1. Up to Scale 3-D Reconstruction

It is well-known that $\mathbf{Q}$ is fixed by the motion of the camera and the orientation of the scene plane: $\mathbf{Q}_o \approxeq \mathbf{R}_o + \mathbf{t}_o\mathbf{n}_o$, where $\approxeq$ denotes the up to scale equality. Furthermore, it has been shown that up-to-scale $\mathbf{Q}_o$ is generally decomposable into two interpretations, the true and dual solutions, in the form $\{\mathbf{R}, k^{-1}\mathbf{t}_o, k\mathbf{n}_o\}$ [14]; here, k is any non-zero constant that is associated with the well-known inherent scale factor ambiguity of monocular vision. (Again, we emphasize that we can generally utilize the essential matrix to achieve the sought after decomposition up to the same scale factor ambiguity, but require a minimum of 5 matches [19]).

Without loss of generality, the surface normal $\mathbf{n}_o$ may be fixed to a unit vector: $k_o = (\|\mathbf{n}_o\|)^{-1}$. Thus, optical camera translation $\mathbf{t}_o$ and planar surface normal $\mathbf{n}_o$ are known up to unknown scale k_o: $\hat{\mathbf{t}}_o = k_o^{-1}\mathbf{t}_o$ and $\hat{\mathbf{n}}_o = k_o\mathbf{n}_o$.

It goes without saying that a RANSAC-based implementation allows us to identify the outliers, e.g., mainly matches corresponding to non-coplanar scene points.

3.2. Resolving the Scale Factor Ambiguity

Taking note of (7), the surface normal in the sonar coordinate system is in the form

$$\mathbf{n}_s = \|k_s\|\hat{\mathbf{n}}_s; k_s = \frac{k_o^{-1}}{1 - k_o^{-1}\mathbf{t}^T\mathbf{R}\hat{\mathbf{n}}_o} \quad \hat{\mathbf{n}}_s = \text{sign}(k_s)\mathbf{R}\hat{\mathbf{n}}_o \tag{13}$$

Note that we have determined $\mathbf{R}$ from the decomposition of the projective transformation $\mathbf{Q}_o$. Thus, the scale factor ambiguity is resolved if we determine the magnitude of the surface normal k_s in the sonar coordinate system.

Expressing the surface normal in terms of magnitude k_s and unit vector $\hat{\mathbf{n}}_s = (\hat{\mathbf{n}}_x, \hat{\mathbf{n}}_y, \hat{\mathbf{n}}_z)^T$, we can write the plane equation $\mathbf{P}_s \cdot \mathbf{n}_s = -1$ in the form

$$k_s\left((\hat{n_x}\sin\theta + \hat{n_y}\cos\theta)\cos\phi + \hat{n_z}\sin\phi\right) = -1/\Re \tag{14}$$

It follows that

$$\sin(\phi + \gamma) = \frac{-1}{k_s\Re\sqrt{(\hat{n_x}\sin\theta + \hat{n_y}\cos\theta)^2 + \hat{n_z}^2}} \tag{15}$$

$$\text{where } \gamma = \tan^{-1}\left(\frac{\hat{n_x}\sin\theta + \hat{n_y}\cos\theta}{\hat{n_z}}\right) \tag{16}$$

Knowing the surface normal magnitude k_s, this equation gives the elevation angle of each planar point:

$$\phi = \{\phi_s, \pi - \phi_s\}$$

$$\phi_s = -\gamma + \sin^{-1}\left(\frac{-1}{{}_{k_s}\Re\sqrt{(\hat{n_x}\sin\theta + \hat{n_y}\cos\theta)^2 + \hat{n_z}^2}}\right) \tag{17}$$

The correct solution is chosen based on imaging constraint $-\delta\phi \geq \phi \leq \delta\phi$, since 2-D sonar camera projection rays have limited width $2\delta\phi$ in the elevation direction. In the limit as $\delta\phi \longrightarrow 0$, we obtain the pencil beam of 3-D sonar imaging systems; e.g., [1]. As an example, $\delta\phi = 7[deg]$ in a DIDSON camera [2].

Alternatively, if the elevation angle is known, we can determine the scale factor from

$$k_s = \frac{-1}{\Re\sin(\gamma + \phi)\sqrt{(\hat{n_x}\sin\theta + \hat{n_y}\cos\theta)^2 + \hat{n_z}^2}} \tag{18}$$

From these results, it can be readily established that the motion transformation of sonar image points in (12) can be expressed in terms of the unknown scale factor k_s: $\mathbf{H}_s$ depends on the unknown elevation angles ϕ and ϕ', which in turn are given in terms of the normal scaling according to (17); Note that ϕ', the elevation angles in the second sonar view are fixed by the sonar transformation (11), except for the unknown scaling. Equivalently, we can express these results in terms of unknown scaling in the optical coordinate system through (14). Despite the complexities of these equations, a one-parameter nonlinear optimizing problem can be readily solved. Furthermore, we can readily distinguish the correct solution from the dual solution as the dual solution is not satisfied by the sonar motion transformation. Theoretically, a single match in the sonar view that lies on Π_g gives two equations for fixing the unknown scaling, though many more are used in practice. We do not address how we identify the matches that lie of the sought after plane, however, a RANAC-based implementation can establish a potential approach.

To summarize, the up-to-scale estimates of motion and plane normal are determined from a minimum of 4 optical image correspondences of co-planar scene points. The scale factor is recovered from a minimum of one sonar image correspondence. In practice, this solution is sub-optimal, as it does not fully exploit the visual cues in the sonar views. In particular, 8 of the 9 unknowns are determined solely from the motion cues in the optical sequence. We can utilize this solution as an initial guess for a non-linear optimization method that accounts for all available measurements.

3.3. MLE Formulation

We model the measurement noise in the positions of features in the optical and sonar views, $\mathbf{p}_o = (x, y)$ and $\mathbf{p}_s = (x_s, y_s)$, by additive Gaussian distribution. This assumption has been utilized in a large body of relevant work in motion vision research with satisfactory results; see [9]. Furthermore, independent Gaussian model of the uncertainties in sonar image feature positions x_s and y_s translates to Rayleigh distribution for sonar range measurements, which agrees with the speckle noise model for sonar imaging sys-

tems [23]. This allows us to minimize the Mahalanobis distance between the measurements and reprojections as the ML estimate [1].

Accordingly, we minimize

$$\begin{aligned}\mathcal{E}(\mathbf{R},\mathbf{T},\mathbf{n}) = &\Big(\textstyle\sum_{N_o}(\mathbf{p}_o-\widehat{\mathbf{p}}_o)^T{\Sigma^o}^{-1}(\mathbf{p}_o-\widehat{\mathbf{p}}_o)+\\ &\textstyle\sum_{N_o}(\mathbf{p}'_o-\widehat{\mathbf{p}}'_o)^T{\Sigma^o}^{-1}(\mathbf{p}'_o-\widehat{\mathbf{p}}'_o)\Big)\\ +\Lambda\,&\Big(\textstyle\sum_{N_s}(\mathbf{p}_s-\widehat{\mathbf{p}}_s)^T{\Sigma^s}^{-1}(\mathbf{p}_s-\widehat{\mathbf{p}}_s)+\\ &\textstyle\sum_{N_s}(\mathbf{p}'_s-\widehat{\mathbf{p}}'_s)^T{\Sigma^s}^{-1}(\mathbf{p}'_s-\widehat{\mathbf{p}}'_s)\Big)\end{aligned} \quad (19)$$

where N_o and N_s denote the number of matches in the optical and sonar views, and $\widehat{(.)}$ denotes any reprojection. In our results, the covariances matrix Σ^o and Σ^s of the optical and sonar measurement vectors have been set by assuming localization uncertainties of 1 [pix].

In addition to the relative scaling of the optical and sonar reprojection error magnitudes, the inclusion of Λ, provides for the interpretation of the optimization formulation as a regularization problem: Either sonar or optical measurements interchangeably can play the role of the regularizer. where adjustment of Λ is generally made in accordance with the medium turbidity condition and acoustic clutter level, in order to control how one camera's measurements regularize the solution based on measurements from the other. A theoretical foundation for the optimal selection of Λ, yet to be developed, will require verification through controlled experiments with a relatively large set of data under various medium conditions. Here, it is used mainly to normalize the average reprojection errors of the two sequences.

The unknown motion and surface parameters $\{\mathbf{R},\mathbf{T},\mathbf{n}\}$ may be defined in the coordinate system of either camera, and are transformed from one system to another through the relationship in (6).

We have applied the the Levenberg-Marquardt algorithm to solve this nonlinear optimization problem. Impact of outliers is minimized by removing points with reprojection errors that exceed an acceptable level. This has currently been set according to the distribution of the optical and sonar reprojections errors for the initial solution, which serves as the start point of the final iterative optimization process. Outliers typically include points with large measurement error, or relatively large distances from the plane Π_g (compared to their distances to the stereo system), as well as incorrect correspondences.

3.4. Application to Opti-Acoustic Correspondence Problem

While not the scope of this paper, the proposed solution serves as a preliminary step in establishing opti-acoustic matches for a denser reconstruction of 3-D objects: A set of 3-D points are reconstructed by triangulation from each pair of N_p and N_s optical and sonar matches. Reprojection of optical points onto the sonar image gives their sonar matches, and vice-versa. These give $N_p + N_s$ quadruplet matches in 4 views that serve as initial seeds to identify other matches by propagation. A particular approach is to define planar patches from sets of triplet 3-D points, say by Delaunay triangulation (not all 3-D points lie on Π_g). New matches are found by reprojecting each patch back onto both images and performing local search along corresponding epipolar lines. In addition to geometric constraints, physical models may be incorporated, where model parameters may be estimated from initial quadruplet matches.

[1] based on the estimated solution and the models in (10) and (12).

4. Experimental Results

We start with an example of the epipolar geometry of a calibrated opti-acoustic stereo imaging system. The exterior calibration is based on a nonlinear optimization method utilizing a set of matching features on a planar grid [18]. Fig. 1 shows the selected features, verifying that the matching points lie on corresponding epipolar curves. With a calibrated system, we can apply the proposed 3-D motion estimation technique.

Our first experiment is based on a data set acquired at the pool facility of Teledyne Benthos, N. Falmouth, MA. Collected in an indoor pool, the sonar grid images are corrupted by multiple returns due to reflections from the water surface and various pool walls. For the purposes of this paper, it suffices to use features matches manually, since we are assessing the accuracy of motion integration which is the main contribution. In practice, existing matching methods (e.g., SIFT [10]) can be applied since we require temporal correspondences among optical and among sonar features independently, and in particular sub-pixel accuracy can lead to improved performance in motion estimation.

Fig. 2 shows the stereo pairs, superimposed with the matched features (green circles). Crosses depict the projections of initial solution that is calculated from projective homography decomposition of the optical sequence (red), and scale determined from the sonar sequence. The final estimate from the MLE formulation is depicted by green crosses. While the initial solution matches the optical features well, there is larger discrepancies for the sonar features. Note that the scale is not necessary for the optical projections, and thus initial estimation of scale does not affect the location of reprojected points. Overall, the errors are decreased through the integration of motion cues. In the absence of ground truth to asses motion accuracy, we can compare the estimated normal of the plane from our motion cues, with that obtained in the calibration process. Estimated surface normals in the optical view from calibration, as well as initial and final processing are

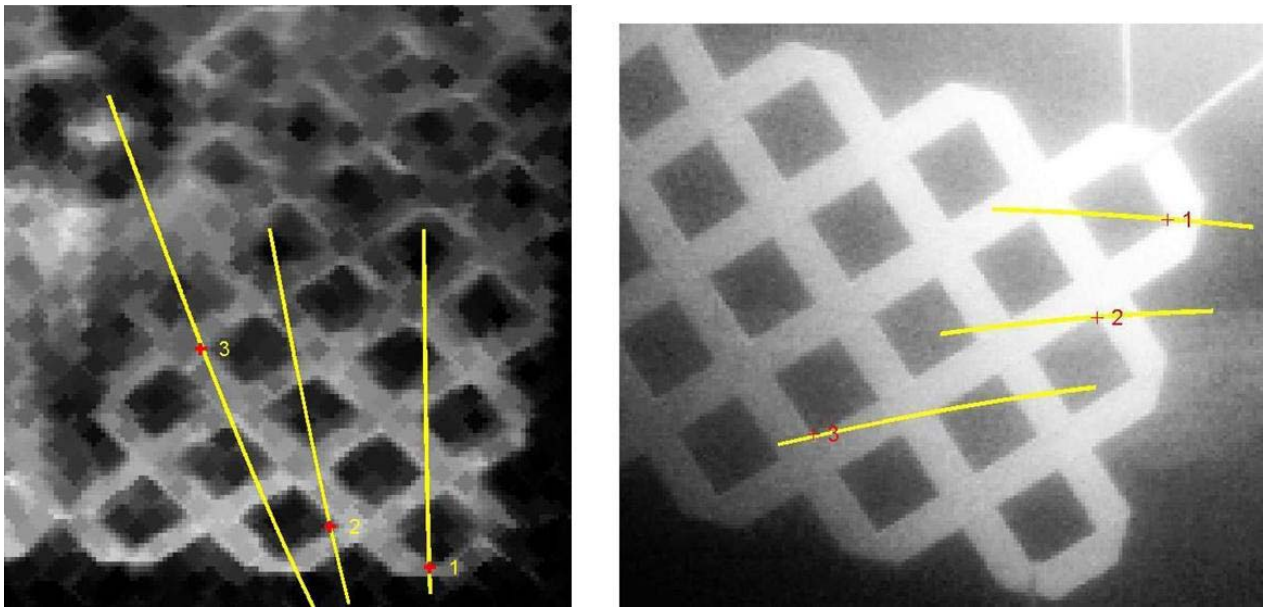

Figure 1. Epipolar geometry of corresponding features in a calibrated opti-acoustic stereo pair.

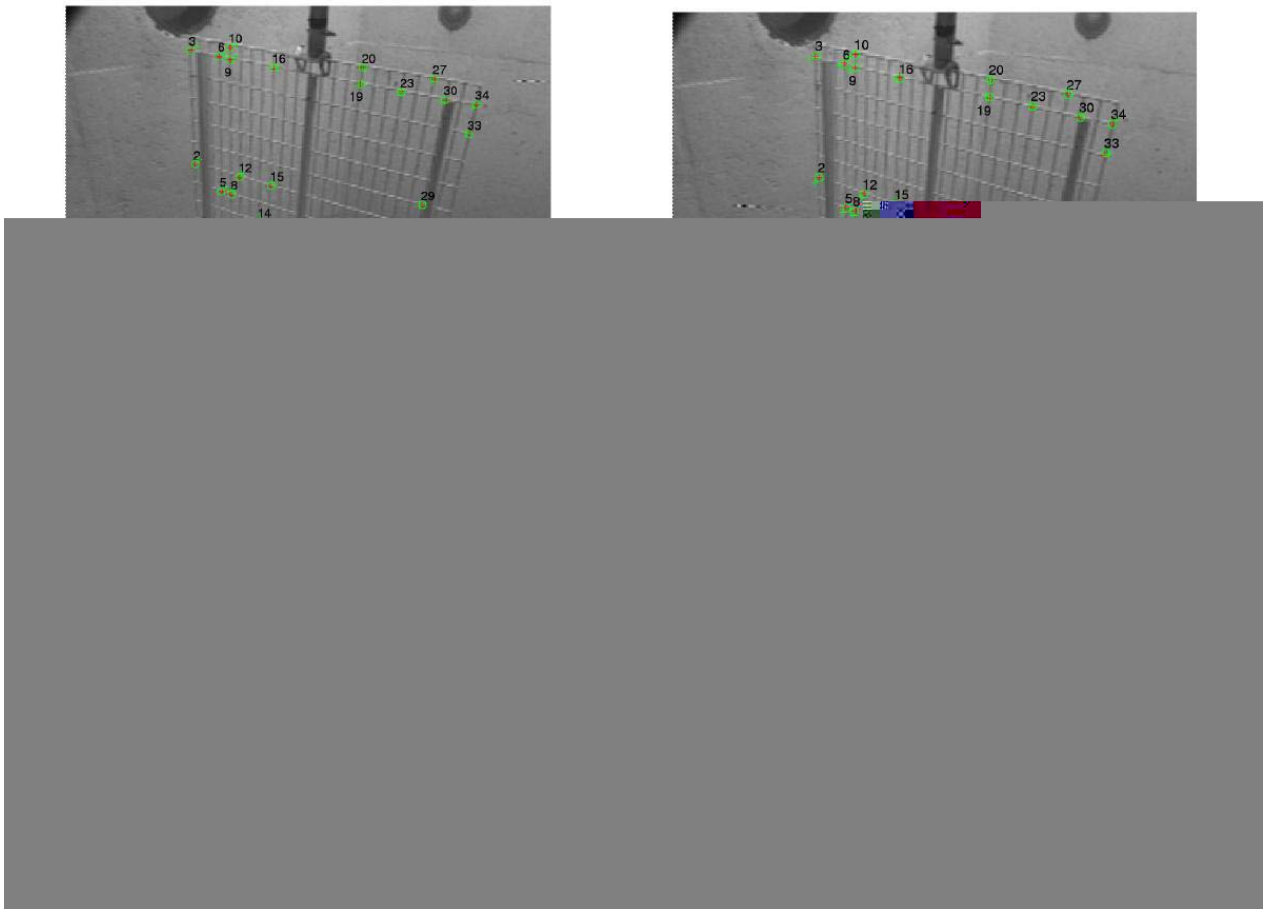

Figure 2. Consecutive optical (a-b) and sonar (a'-b') views of the grid with matching features for motion estimation (green circles). Crosses show reprojections based on initial solution from projective homography decomposition of the optical sequence (red) and final estimate from proposed solution by integration of optical and sonar motion cues (green).

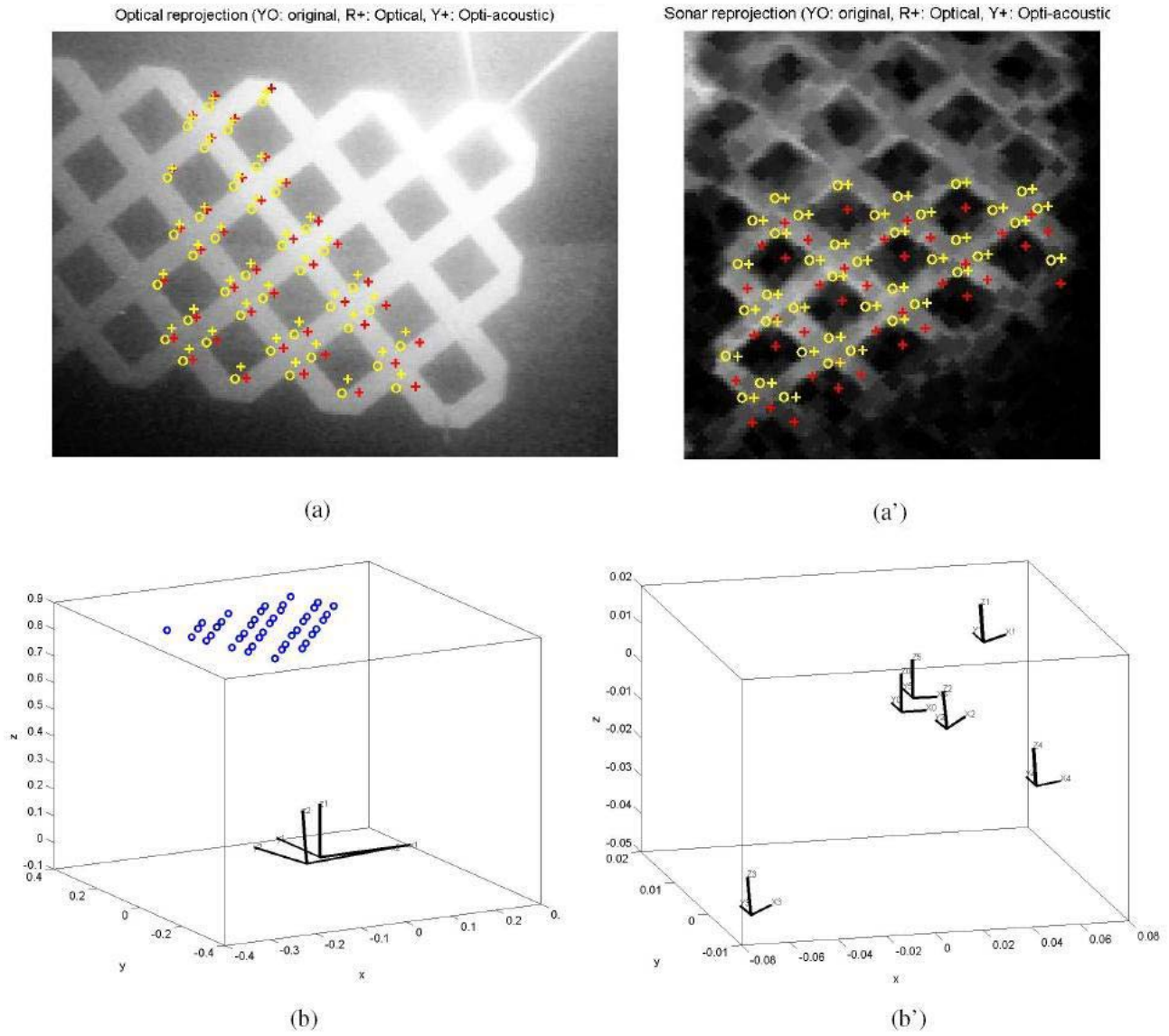

Figure 3. (a,a') Sample stereo pair with features used for motion estimation. (b) Estimated 3-D planar grid points relative to the two consecutive positions of the optical camera, and (b') 6 positions of the optical camera in the last experiment with the closed trajectory. Ideally, initial and final camera coordinate frames $X_0Y_0Z_0$ and $X_5Y_5Z_5$ should coincide.

$\mathbf{n}_{cal} = [-.18, .22, -0.67]$, $\mathbf{n}_{ini} = [-0.33, 0.47, -0.45]$ and $\mathbf{n}_{fin} = [-.26, .45, -0.58]$, respectively. There is a large improvement after the integration process.

A second experiment is based on data collected in our water tank with higher turbidity; see figs. 3 (a,a'). Here we have shown the first stereo pair superimposed by features in each camera used for motion estimation (yellow circles), and the 3-D points relative to the two positions of the optical cameras established by our motion estimation method. To quantify accuracy based on ground truth over a longer trajectory: 1) Motion estimation has been repeated for a sequence of 6 such opti-acoustic stereo images; 2) First pair have been added to the end of the sequence to close the trajectory. Therefore, using the initial pose as the reference, the trajectory should end at the origin with the same exact stereo pose. Table 1 summarizes the experimental results for the monocular optical sequence, in comparison to the proposed method based on integrated opti-acoustic motion estimation. For the unknown scale factor of monocular sequence, we have used the optic-acoustic estimate. In (a,a'), the red and yellow crosses are the reprojected features at the end of the trajectory, for the monocular optical and opti-acoustic estimates, respectively. In addition to the estimated trajectory length, the reprojection errors, while comparable in the optical camera coordinate frame, are noticeably improved in the coordinate frame of the sonar camera.

Travel distance over 6 frames=0.4291 [m]	**Optical**	**Opti-acoustic**
Trajectory error norm [m]	0.0346	0.0224
MSE of optical reprojections [pix]	4.9756	4.7741
MSE of sonar reprojections [m]	0.0324	0.0137
Pose angle error [deg]	1.0672	0.5271
MSE of 3D target points [m]	0.0387	0.0174

Table 1. Various errors comparing estimation results of monocular optical and opti-acoustic systems.

5. Summary

Integration of visual motion cues in the so-called opti-acoustic stereo imaging system – by utilization optical and sonar video cameras in stereo configuration – has been proposed in reconstructing 3-D motion and 3-D scene feature positions. No initial matches in the optical and sonar stereo pairs are necessary, only the correspondence between matches in sonar and optical sequences, individually. The proposed paradigm offers several advantages by providing a mechanism for: 1) 3-D estimation in a wider range of environmental conditions, that can be accomplished by either camera alone; 2) Computing more accurate 3-D estimates; 3) Overcoming the inherent ambiguities of monocular motion sequences; 4) Establishing opti-acoustic matches by utilizing 3-D estimates at a small number of features as seeds and propagation to nearby points along epipolar curves. Results have been given for real data collected in Benthos indoor pool and our tank facility in support of new theoretical findings. Pool conditions were much more favorable for optical imaging, both for good visibility and multiple acoustic reflections from various surfaces. The tank data was collected under more turbid conditions. Data to be collected under a wider range of conditions including open waters will enable more extended testing of the proposed method. Establishing opti-acoustic correspondences is the immediate problem of interest.

Acknowledgement: This work is based on research supported by ONR under grant N000140510717. Views, opinions and conclusions of the authors are not necessarily shared and endorsed by ONR.

References

[1] http://www.codaoctopus.com/3d_ac_im/index.asp

[2] http://www.soundmetrics.com

[3] C. Barat, M.J. Rendas, "Exploiting natural contours for automatic sonar-to-video calibration," *Proc. Oceans* Volume 1, Brest, France, June, 2005.

[4] E.O. Belcher, D.G. Gallagher, J.R. Barone, and R.E. Honaker, "Acoustic lens camera and underwater display combine to provide efficient and effective hull and berth inspections," *Proc. Oceans'03*, San Diego, CA, September, 2003.

[5] J.A. Beraldin, "Integration of laser scanning and close-range photography - The last decade and beyond," *Proc. XXth ISPRS Congress (Comm-V papers)*, Istanbul, 2004.

[6] U. Castellani, A. Fusiello, V. Murino, L. Papaleo, E. Puppo, M. Pittore, "A complete system for on-line 3D modelling from acoustic images," *Signal Proc. Image Comm.*, 20, 2005.

[7] Q. Chen and G. Medioni, "A volumetric stereo matching method: Application to image-based modeling," in *Proc. IEEE Conference Computer Vision & Pattern Recognition*, vol. 1, Fort Collins, CO, June 1999.

[8] A. Fusiello, and V. Murino, "Augmented scene modeling and visualization by optical and acoustic sensor integration," *IEEE Trans. Vis. & Comp. Graphics*, Vol 10(5), Nov-Dec, 2004.

[9] R.I Hartley, and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge Univ. Press, 2000.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[11] B. Kamgar-Parsi, L.J. Rosenblum, and E.O. Belcher, "Underwater imaging with a moving acoustic lens," *IEEE Trans. Image Proc.*, Vol 7(1), January, 1998.

[12] D.M Kocak and F.M. Caimi, "The current art of underwater imaging – With a glimpse of the past and vision of the future", *MTS Journal*, Vol 39(3), October, 2005.

[13] M. Lhuillier and L. Quan, "Match propagation for image-based modeling and rendering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24(8), 2002.

[14] H.C. Longuet-Higgins, "The visual ambiguity of a moving plane," *Proc. Royal Soc. London*, *Ser. B*, Vol **223**, 1984.

[15] K.D. Moore, J.S. Jaffe, and B.L. Ochoa, "Development of a new underwater bathymetric laser imaging system: L-Bath," *Journal of Atmospheric and Oceanic Technology*, Vol 17(8), August, 2000.

[16] S. Negahdaripour, "Epipolar geometry of opti-acoustic stereo imaging," to appear in *IEEE Trans. PAMI*.

[17] S. Negahdaripour, Calibration of DIDSON forward-scan acoustic video camera, *Proc. Oceans'05*, Washington D.C., August, 2005.

[18] S. Negahdaripour, H. Sekkati, and H. Pirsiavash "Opti-acoustic stereo imaging: On system calibration and 3-D target reconstruction," in review.

[19] D. Nister, "An efficient solution to the five-point relative pose problem," *IEEE Trans. PAMI*, Vol 26(6), June 2004.

[20] Y. Y. Schechner and N. Karpel, "Clear underwater vision," *Proc. CVPR*, vol. 1, 2004.

[21] S. G. Narasimhan and S. K. Nayar, "Removing weather effects from monochrome images," *Proc. CVPR*, Hawaii, December 2001

[22] R. Vesetas, and G. Manzie, "AMI: A 3-D imaging sonar for mine identification in turbid waters," *Proc. Oceans'01*, Honolulu, HI, November, 2001.

[23] R.F. Wagner, S.W. Smith, J.M. Sandrik, H. Lopez, "Statistics of speckle in ultrasound B-scans," *IEEE Trans. on Sonics and Ultras.*, vol 30, May 1983, pp. 156-163.