Stereo Self-Calibration for Seafloor Mapping using AUVs

Clayton Kunz and Hanumant Singh
Deep Submergence Laboratory
Woods Hole Oceanographic Institution
266 Woods Hole Rd., Mailstop 7
Woods Hole, MA 02543 USA

Abstract— Visual maps of the seafloor should ideally provide the ability to measure individual features of interest in real units. Two-dimensional photomosaics cannot provide this capability without making assumptions that often fail over 3-D terrain, and are generally used for visualization, but not for measurement. Full 3-D structure can be recovered using stereo vision, structure from motion (SFM), or simultaneous localization and mapping (SLAM). Of these techniques, only stereo vision is suitable for fully dense (i.e. a distance measurement for each imaged pixel) 3-D structure in the absence of significant frame-to-frame overlap. Stereo vision is notoriously dependent on camera calibration, however, which is difficult to compute and maintain in the field. The fewer dependencies an AUV mapping system has on camera calibration, the more reliably it will be able to produce useful maps of the seafloor. We present a system for recovering the 7-DOF relationship between the AUV’s estimation frame and the camera rig (Euclidean offsets plus scale), which reconciles the robot’s odometry-based pose estimate with stereo visual odometry. The combination of robust frame-to-frame visual feature matching, subpixel stereo correspondence estimation, and high-accuracy on-board vehicle navigation sensors enables us to self-calibrate the extrinsic parameters of the stereo rig including scale, and produce metric maps using only vehicle navigation and the computed camera calibration. Using data acquired in the Bering Sea by the SeaBED AUV in August of 2009, our initial results indicate that accumulated navigation drift is less than 0.5% of distance travelled, suggesting that a visual SLAM system for correcting drift and building a final map would only require the robot’s path to cross itself every few hundred meters. In addition to providing a large-scale metric 3-D map, the corrected stereo calibration enables scientists to measure the sizes of imaged objects without additional hardware such as laser points or acoustic ranging systems.

Introduction

Autonomous underwater vehicles (AUVs) capable of low altitude “flight” are increasingly being used to survey the sea floor photographically. Photographic surveys allow scientists to gather much more information about the terrain being imaged than do bathymetric surveys; in particular individual biological or geological specimens can be identified and counted. If 3-D visual data is available from an AUV, then the imaged specimens can also be measured, giving scientists additional insight into the sea floor biosphere. Any AUV building a large-area visually textured 3-D seafloor map can only do so by combining several images captured from different locations. There are several approaches one can take: monocular reconstruction using structure from motion (SFM), reconstruction using additional ranging sensors (e.g. multibeam sonar), and stereo vision. Any of these approaches can incorporate vehicle navigation information to improve overall map consistency, for example using techniques from simultaneous localization and mapping (SLAM).

Because the distance between the AUV and the sea floor is not large compared to the terrain relief, all of these approaches generally rely on the use of a perspective camera model, rather than the more restrictive (and numerically better-behaved) affine or orthographic models used by aerial or orbital photographic surveys. Perspective, or pin-hole camera models map points in space $X$ to pixels in the image $x$ via a $3 \times 4$ homogeneous projection matrix: $x = K X$, where $K$ is the $3 \times 3$ upper-triangular intrinsic calibration matrix, and $T$ is the $3 \times 4$ Euclidean transformation mapping points in a world coordinate frame to points in the camera-centric coordinate frame. Both $X$ and $x$ are represented using homogeneous coordinates, so vectors are considered equivalent if they have the same “real” or non-homogeneous coordinates are recovered by normalizing the vectors so that the last coordinate is one. Camera calibration is the process of recovering the elements of $K$, and possibly $T$ and additional non-linear lens distortion parameters, from image data.

In this paper, we are concerned with the problem of using images and navigation data acquired by an AUV during a real-world deployment to calibrate a pair of cameras. Stereo vision is of particular interest for users of AUVs because a well-calibrated stereo pair enables dense 3-D reconstruction without the need for large overlap between subsequently captured image pairs. There is currently wide interest in the use of stereo vision on board underwater vehicles (see [1], [2], [3] for example). The ability to self-calibrate a stereo pair (i.e. to calibrate without the need for special calibration targets) is especially desired, because it is often difficult to run through a calibration procedure at sea, and rough conditions can change the orientation of a pair of cameras relative to each other –
Fig. 1 shows conditions encountered while acquiring the data set used in this paper. Moreover, much historical imagery is available without accompanying camera calibration information, though often with associated vehicle navigation data.

A full stereo calibration involves recovering the parameters of $K$ for each of the two cameras, as well as the extrinsic relationship between the two cameras, and the relationship between the pair of cameras and the navigation frame of the AUV, which is roughly 20 degrees of freedom: 4 for each $K$ matrix, and 6 for each $T$ matrix. While our ultimate goal is complete self-calibration, in this paper we concentrate on the problem of using visual information together with vehicle navigation data to recover the location of one of the two cameras, as well as an overall scale factor (seven degrees of freedom); we explain why this simplification is a reasonable first step below. The paper is organized as follows: in the next section we examine previous work on the self-calibration problem. We then present an example scenario, describe our solution to determining the seven relevant degrees of freedom, and examine its implications. Finally, we turn to the broader problem of full self-calibration and address practical considerations toward its solution.

A RECENT HISTORY OF STEREO SELF-CALIBRATION

The geometric relationships between 3-D points imaged by perspective cameras have been studied in the photogrammetric community for over 100 years. More recently, interest in the uncalibrated case led to the idea of the fundamental matrix $F$, which captures the epipolar geometry between a pair of cameras via the homogeneous relation $x^T F x = 0$, where $x$ and $x'$ are the two images (one from each of two cameras) of a point in space. For a stereo pair with fixed optics, the $F$ relating the port and starboard cameras does not change, while for a single moving camera, $F$ will be different for each image pair. The surprising result from the early descriptions and analyses of $F$ [4], [5], is that the information contained in the fundamental matrix is sufficient to allow one to reconstruct a 3-D scene and the camera matrices used in imaging it up to a projective transformation of space. Moreover, $F$ can be computed from image correspondences alone, for example via the so-called “eight-point algorithm” [6], enabling a reconstruction up to a projective (15 parameter) ambiguity without any calibration. These two results are what make structure from motion possible.

While a projective reconstruction has uses, it is not sufficient for visualization or measurement of natural terrain. Most reconstructions therefore are “upgraded,” first to an affine (12 parameter) ambiguity, and then to a similarity (7 parameter) ambiguity of the cameras and the points imaged. Both of these upgrade steps can in theory be performed for a stereo rig using image-to-image matches together with some mild assumptions about the camera geometry (e.g. that the pixels are square) [7]. These upgrades require multiple stereo views of the same scene, with correspondences established between port and starboard cameras, as well as between images captured by the two cameras after they have moved. The upgrade to an affine ambiguity requires identifying the $4 \times 4$ projective transformation mapping the triangulated 3-D points from the first view to the second, and then computing the eigenvectors of the transformation, ultimately yielding the “infinite homography,” $H_\infty$. This is considered a rather delicate step, as projective spaces do not provide good distance measurements to minimize (though see [8] for several approaches), pixel reprojection error may not be sensitive enough, and the actual motion undergone by the rig has a large influence on the shape of the error surface [9], [10].

The upgrade from an affine ambiguity to a metric ambiguity again relies on $H_\infty$, and uses a linear function to estimate the image of the absolute conic $\omega$, from which one can compute $K$ via the relation $\omega = (KK^T)^{-1}$. Additional constraints on self-calibration can be had from controlled camera motions, however this is impractical on AUVs, where perfect control is not achievable. Alternative to this “bottom-up” approach is given in [11], which solves for all of the camera parameters simultaneously in a “top-down” fashion, minimizing only over the fundamental matrix equations induced by the four camera views.

It is important to note that without external information, it is not possible to further upgrade a reconstruction, say to a Euclidean ambiguity. This is why it makes sense to use the navigation instruments on an AUV to “finish” the self-
calibration: precisely those seven degrees of freedom which image information cannot address are measured by gyros and velocity logs. Even so, it is not always necessary to resolve these ambiguities. In particular, effective visualization generally only requires a reconstruction up to a similarity, as such a view provides a sense of the layout of features on the seafloor and their relative abundances. In addition, people implicitly add scale to what they see, applying their own experience and expertise by assigning sizes to animals, for example. This implicit scale is only approximate, of course. An alternative explanation, [10] suggests that the intrinsic parameters of the camera need only be known approximately for a good visualization; moreover he suggests that these parameters can only be approximately recovered from image data because of the flatness of the error surfaces. Both [10] and [11] suggest setting the principal points of the cameras to be equal to the image centers rather than solving for them, justifying their suggestions by showing that the resulting reconstructions do not depend greatly on the choice of the principal points.

**EXAMPLE TRACK / DATA**

As the ultimate goal of this work is to produce a texture-mapped Euclidean map of the sea floor, we must recover the geometric relationship between the stereo cameras and the

![Figure 2](image1.png)  
Figure 2. The track of the SeaBED AUV during a dive in the Bering Sea, August 2009, at depths between 200 and 250 meters. The image on the right is zoomed-in from the circled area on the left. On the right, only the south-bound portion of the transit contained image-to-image overlap.

![Figure 3](image2.png)  
Figure 3. Unrectified images taken by the forward port camera before and after an 11 minute, 150 meter transit.
AUV’s navigation reference frame. We use the SeaBED AUV [12], which in this case was outfitted with four cameras arranged in two stereo pairs. SeaBED uses a Paroscientific depth sensor, an IXSEA Octans 3-axis fibre optic gyro for measuring heading, roll, and pitch as well as angular velocities, and an RDI ADCP for measuring translational velocity relative to the sea floor. The specific dive we focus on here is shown in Fig. 2, during which the AUV captured a set of images once every five seconds. On this dive we were particularly fortunate to recapture a recognizable feature on the sea floor after about 11 minutes and 150 meters of transit, as shown in Fig. 3. SeaBED uses the location of the DVL as its navigation reference frame; the cameras of interest are mounted about 1.4 meters ahead of the DVL and are pointed roughly straight down. The navigation data used here relies strictly on these internal sensors; USBL is only used together with GPS to globally place the dive as a whole.

During the first part of the dive, the AUV was attempting to drive to a specific location on the sea floor, and was being pushed westward by a current, resulting in the curved track. After SeaBED reached this first goal, it switched to line-following mode, crabbing against the current and producing a much straighter path. The result of this is that the robot captured images with significant overlap only before it reached the first goal. We use these images to determine the DVL to camera transformation and scale factor, and use the crossover point to validate the result below.

This work relies heavily on the establishment of pixel correspondences both between port and starboard cameras, and between image pairs captured sequentially in time. In both cases, we compute correspondences by the now-standard technique of comparing region descriptors computed around salient features – we use scaled Harris corners to detect features and moments of Zernike polynomials to describe them [13] – then using these matches to compute an F matrix. The F matrix in turn yields a pair of rectifying transformations that are applied to the images to allow a dense correlation process [14], ultimately producing a disparity map. Fig. 4 shows a pair of rectified images captured before and after a motion of the vehicle, with the computed disparity map. Once correspondence is established, triangulating the matching pixels is a linear process if the motion is Euclidean (i.e. the cameras are calibrated); otherwise [15] suggests a solution which is well-behaved in the projective case, involving finding the roots of a polynomial in one variable.

**RECOVERING CAMERA OFFSETS FROM NAVIGATION DATA**

In the following, let $T_B^A$ represent the $4 \times 4$ Euclidean transformation matrix mapping points in reference frame A to points in reference frame B, and denote frames after vehicle motion with a prime mark. Our frames of interest are W (the world), V (the vehicle), and C (the stereo rig, chosen arbitrarily to be the frame of the port camera). Points X are relative to the camera frame before motion; $X'$ relative to the camera frame after motion.

Our approach to recovering the camera offsets and scale from AUV navigation data is based on the idea of stereo visual odometry [16]. This technique recovers camera motion by matching triangulated points $X$ from one stereo pair to the same points $X'$ from the same pair after the rig has moved (we call such a set of images a “quad”): $X' = T_C^C X$. Recovering $T_C^C$ is the exterior orientation problem, and is solved using the method described in [17], wrapped in a RANSAC [18] loop to improve robustness. Note that the self-calibration
methods described above the first $T_C^{C'}$ estimated is a full projective transformation, and the methods implicitly or explicitly recover transformations to the cameras and structure such that $T_C^{C'}$ will ultimately be upgraded to a similarity transformation (Euclidean plus scale).

For each image quad we also compute a motion estimate for the vehicle, $T_V^{V'}$, based on the navigation sensors. The two transformations are related by

$$T_C^{C'} = T_W^{V'} T_C^{V} T_C^{C'} = T_W^{V'} T_V^{V'} T_C^{C'}.$$  

More simply, since the relationship between the camera and AUV does not change,

$$T_V^{V'} = T_C^{C'} (T_C^{V})^{-1}.$$  

The term on the left describes the AUV navigation estimate for one image-to-image time step, while the term on the right describes the AUV motion predicted by stereo visual odometry, given a particular hypothesized offset between the AUV reference frame and the camera reference frame. The task is to find the $T_C^{V}$ which minimizes the discrepancy between the two terms:

$$\sum_i (T_C^{V} (T_C^{V})^{-1} - T_V^{V'})$$  

where the sum is over a set of image quads.

There is also the unknown scale factor $\alpha$, which can be incorporated into the above equation by allowing the last row of $T_C^{V}$ to equal $[0 \ 0 \ 0 \ 1/\alpha]$ instead of $[0 \ 0 \ 0 \ 1]$. This scale factor renders the system weakly constrained: in the presence of vehicle rotation, any number of translation + scale combinations describe the camera motion roughly equally well, as shown in Fig. 5.

We compensate for this weak constraint by seeding the minimization with an initial guess based on an AUV motion with minimal heading change. For a given image quad, if the heading change is close to zero, then we can directly estimate the rotation offset between the AUV and the camera, and the scale factor $\alpha$, only by examining the translational motion estimate of the AUV navigation and the visual odometry. If the estimated vehicle translation direction is given by $t_V$, and the estimated camera translation direction by $t_C$, then the axis of rotation between the vehicle and camera frames is given by $t_C \times t_V$, and the magnitude of the rotation is given by $\cos^{-1}(t_V^\top t_C)$. Finally the scale factor $\alpha$ is given by the ratio of the norms of the two translation vectors.

This estimate from a single AUV motion gives us a starting point for the full rotation offset of the cameras as well as the scale of the reconstruction. The only remaining parameters are translational, and can be seeded from an “eyeball” guess. Then we can minimize (1) using a standard algorithm such as Levenberg-Marquardt. To improve stability, we first minimize only over $x, y$ and heading, and then allow a full minimization over all degrees of freedom except $z$, which remains weakly constrained under typical AUV motion. Instead of attempting to find a reasonable norm for the matrix difference in (1), we reparameterize the equation using the six Euclidean degrees of freedom and minimize using Euclidean distance.

**RESULTS AND DISCUSSION**

Using the imagery from the SeaBED dive described above, we built a 3-D reconstruction with offsets recovered with this technique. Starting with the ballpark estimate of $(1.4, 0, 0)$ meters, the algorithm used 90 image quads, comprising most of the south bound track shown in Fig. 2. The final estimate of $(1.3708, -0.1598, 0)$ meters in translation, with roll, pitch, heading offsets of $-1.0697, 6.8798, 89.6580$ degrees and a scale factor of 1.2372 yielded the reconstruction partially shown in Fig. 6. Note that we fixed $z$ to zero because of the weak constraint on $z$ provided by our motion sequence. There are several interesting points to note:

- The final alignment is not perfect, but given that only pairwise comparisons are made, this is not surprising.
- The recovery of scale allows us to measure the size of the misalignment directly from the image: it is less than 60 cm, less than 0.5% percent of the distance travelled.
Since we did not use altimeter data in our computation, we have an independent measure of the accuracy of the recovered scale. In fact, the recovered ranges to the camera are within 5% of the measured DVL altitude over the entire loop, though these ranges vary considerably given the sheer number of ranges being compared: on the order of one million ranges per stereo pair, versus 4 ranges per DVL ping.

Our navigation error is consistent with other reports of error in pose estimates using DVL and fibre optic gyros [19]. This suggests that with a well-instrumented AUV, only occasional crossover points are necessary to tie a full dive together using SLAM. In particular, it is likely that the navigation instruments on board the AUV will yield better incremental position estimates than will a vision-based system, which in turn will be better at correcting for long-term drift.

Now that we can upgrade a stereo reconstruction from a similarity ambiguity to a full AUV-relative Euclidean model, the natural next question is whether or not it is possible to do a full self-calibration. While we have not yet answered this question, we do have a few insights. First of all, recovering the stereo $F$ matrix is not difficult, because over the course a given dive we have thousands of stereo pairs to work with, yielding millions of matching points. The computation of the $F$ matrix is thus very well constrained. The next step, recovering $H_\infty$ (and hence $K$ for each camera), is more problematic. In most AUV motion near the seafloor, the robot is attempting to maintain constant roll, pitch, and altitude. The vehicle (and camera) translation is thus orthogonal to its rotation, which is a degenerate motion for self-calibration [9]. The situation is made worse by a flat seafloor. Given that many AUVs designed for imaging are not controllable in roll and pitch, the best we can do to approximate a general motion is to program them to capture images over a rough seafloor while rotating in heading and translating vertically – this suggests specific motion sequences for self-calibration analogous to the sequences analyzed for multibeam calibration in [20]. We are actively pursuing work in this direction. Unfortunately these are not otherwise useful AUV motions, and are hard to find in archival data.

We have also not addressed the problem of lens distortion, which introduces strong nonlinearities into the recovered structure of a scene if it is ignored. Even with high quality lenses, distortion is an issue underwater as refraction through flat view ports renders the pinhole model incorrect [21]. As far as we are aware, other full self-calibration algorithms also ignore lens distortion. However, given that the optical characteristics of a camera are less likely to change over the course of a cruise than its physical placement, it should be possible to calibrate for lens distortion in a lab, and use self-calibration to recover extrinsic characteristics as we do here.

**ACKNOWLEDGEMENTS**

The authors thank Derya Akkaynak Yellin, John Bailey, Dr. Jerry Hoff, and the crew of the F/V Ocean Explorer.
REFERENCES


