# Optimal motion estimation from visual and inertial measurements

Dennis Strelow and Sanjiv Singh
*Carnegie Mellon University*
{dstrelow, ssingh}@cs.cmu.edu

## Abstract

*Cameras and inertial sensors are good candidates to be deployed together for autonomous vehicle motion estimation, since each can be used to resolve the ambiguities in the estimated motion that results from using the other modality alone. We present an algorithm that computes optimal vehicle motion estimates by considering all of the measurements from a camera, rate gyro, and accelerometer simultaneously. Such optimal estimates are useful in their own right, and as a gold standard for the comparison of online algorithms.*

*By comparing the motions estimated using visual and inertial measurements, visual measurements only, and inertial measurements only against ground truth, we show that using image and inertial data together can produce highly accurate estimates even when the results produced by each modality alone are very poor. Our test datasets include both conventional and omnidirectional image sequences, and an image sequence with a high percentage of missing data.*

## 1. Introduction

Cameras and inertial sensors are each good candidates for autonomous vehicle navigation because they do not project any detectable energy into the environment, estimate six degree of freedom motion, are not subject to outages or jamming, and are not limited in range. In addition, cameras and inertial sensors are good candidates to be deployed together, since in addition to the obvious advantage of redundant measurements, each can be used to resolve the ambiguities in the estimated motion that results from using the other modality alone. For instance, the image measurements can counteract the error that accumulates when integrating inertial readings, and can be used to distinguish between the effects of gravity and vehicle acceleration in accelerometer readings. Conversely, inertial data can resolve the ambiguities in motion estimated by a camera that sees a degenerate scene, such as one containing too few features, features infinitely far away, or features in an accidental geometric configuration; to remove the discontinuities in estimated motion that can result from features entering or leaving the camera's field of view; and to make motion estimation more robust to mistracked image features.

Of course, estimating motion using each modality separately and combining the individual estimates, e.g., by averaging them, may simply combine two qualitatively incorrect estimates to produce a third incorrect estimate. In this paper, we present an algorithm that instead considers all of the measurements from images, a rate gyro, and an accelerometer simultaneously to produce an optimal estimate of the vehicle motion and motion error covariances. In many applications, this optimal estimate is of interest in its own right. In others, the optimal estimate is important in understanding the best quality we can expect given a particular sensor configuration, vehicle motion, environment, and set of observations, and the inherent sensitivity of the estimate with respect to random observation errors. In particular, optimal estimates are useful as a gold standard for the comparison of online algorithms which, given sufficient computing power, produce real-time but suboptimal estimates of the vehicle's motion.

By comparing the motions estimated using visual and inertial measurements, visual measurements only, and inertial measurements only against ground truth, we show that using image and inertial data together can produce highly accurate estimates even when the results produced by each modality alone are very poor. Our test datasets include both conventional and omnidirectional image sequences, and an image sequence with a high percentage of missing data, i.e., where each point is visible in only a small fraction of the images.

## 2. Related work

Most of the existing algorithms for motion estimation using both visual and inertial data are online rather than optimal methods. Huster and Rock[4] describe an online method for estimating the motion of an autonomous underwater vehicle relative to a single point. Similarly, Kaminer, *et al.*[7] describe an online method for estimating the position of an aircraft relative to a distant aircraft carrier, which is treated as a point. Qian, *et al.*[12] describe a more general method for simultaneously estimating the motion of a camera and the sparse structure of the environment in which the camera moves. In addition to the basic difference between optimal and online methods, each of these methods differs from ours in that they implicitly assume that the points are visible throughout the entire sequence.

Deans and Hebert[3] consider online, batch, and online-batch hybrid methods for bearings only simultaneous localization and mapping (SLAM). In this work, the planar motion of a vehicle and the location of landmarks observed by the vehicle's omnidirectional camera are simultaneously estimated from the landmarks' vehicle coordinate system bearings and the vehicle's odometry. This method is based on the variable state dimension filter[10], which naturally handles cases where points are not visible in every image, but the incorporation of vehicle motion models into this framework is problematic.

Dean and Hebert's batch method provides optimal estimates of the vehicle's motion, and is closely related to the optimal method we describe. However, estimating six degree of freedom motion from image measurements and inertial sensors introduces some difficulties that do not arise in estimating planar motion from bearings and odometry. In particular, using image measurements for six degree of freedom motion requires careful modeling and calibration of the camera, especially in the omnidirectional case, whereas camera modeling and calibration are not required if only bearing will be taken from the images. In addition, the use of accelerometer observations for six degree of freedom motion requires estimation of the vehicle's velocity and orientation relative to gravity, which odometry does not require.

A second batch method is described by Jung and Taylor[6]. This method applies shape-from-motion to a set of widely spaced keyframes from the image sequence, then interpolates the keyframe positions by a spline that best matches the inertial observations. The resulting algorithm provides a continuous estimate of the sensor motion, and only requires that feature correspondences be established between the keyframes, rather than between every image in the sequence. However, this algorithm is not optimal in the same sense as ours, since the image and inertial measurements are not used simultaneously. In particular, the interpolation phase will propagate rather than fix errors in the motion estimated in the shape-from-motion phase.

## 3. Method

Our method is a batch algorithm that uses all of the observations from an image sequence, rate gyro, and accelerometer to produce an optimal estimate of the sensor motion and the motion error covariances. More specifically, this algorithm uses Levenberg-Marquardt to minimize a total error with respect to the sensor rotation, linear translation, and linear velocity at the time of each image, with respect to the world coordinate system gravity vector, and with respect to the three-dimensional world coordinate system positions of the tracked points. We assume that sparse point feature correspondences are provided.

Here, we describe our error function, and refer the reader to [11] for a discussion of Levenberg-Marquardt, which is widely used.

### 3.1 Error function

The overall error function is:

$$E_{\text{combined}} = E_{\text{visual}} + E_{\text{inertial}} \tag{1}$$

The visual error term is:

$$E_{\text{visual}} = \sum_{i,j} D(\pi(C_{\rho_i,t_i}(X_j)), x_{ij}) \tag{2}$$

$E_{\text{visual}}$ specifies an image reprojection error given the six degree of freedom camera positions and three-dimensional point positions. In this error, the sum is over $i$ and $j$, such that point $j$ was observed in image $i$. $x_{ij}$ is the observed projection of point $j$ in image $i$. $\rho_i$ and $t_i$ are the camera-to-world rotation Euler angles and camera-to-world translation, respectively, at the time of image $i$, and $C_{\rho_i,t_i}$ is the world-to-camera transformation specified by $\rho_i$ and $t_i$. $X_j$ is the world coordinate system location of point $j$, so that $C_{\rho_i,t_i}(X_j)$ is location of point $j$ in camera coordinate system $i$. $\pi$ gives the image projection of a three-dimensional point specified in the camera coordinate system. In our current implementation, $\pi$ may be either a conventional (i.e., perspective or orthographic) or an omnidirectional projection.

All of the individual distance functions $D$ are Mahalanobis distances. The covariances can be isotropic, or directional covariances found from the image texture[8][2].

The inertial error term is:

$$\begin{aligned} E_{\text{inertial}} &= \sum_{i=1}^{f-1} D\left(\rho_i, I_\rho(\tau_{i-1}, \tau_i, \rho_{i-1})\right) \\ &+ \sum_{i=1}^{f-1} D\left(v_i, I_v(\tau_{i-1}, \tau_i, \rho_{i-1}, v_{i-1}, g)\right) \\ &+ \sum_{i=1}^{f-1} D\left(t_i, I_t(\tau_{i-1}, \tau_i, \rho_{i-1}, v_{i-1}, g, t_{i-1})\right) \end{aligned} \tag{3}$$

$E_{\text{inertial}}$ gives an error between the estimated positions and velocities and the incremental positions and velocities predicted by the inertial data. Here, $f$ is the number of images, and $\tau_i$ is the time image $i$ was captured. $\rho_i$ and $t_i$ are the camera rotation and translation at time $\tau_i$, just as in the equation for $E_{\text{visual}}$ above. $v_i$ gives the camera's linear velocity at time $\tau_i$, and $g$ is the world coordinate system gravity vector.

$I_\rho$, $I_v$, and $I_t$ integrate the inertial observations to produce estimates of $\rho_i$, $v_i$, and $t_i$ from initial values $\rho_{i-1}$, $v_{i-1}$, and $t_{i-1}$, respectively. Over an interval $[\tau, \tau']$ where the camera coordinate system angular velocity is assumed constant, e.g., between the two inertial readings or between an inertial reading and an image time, $I_\rho$ is defined as follows:

$$I_\rho(\tau, \tau', \rho) = r(\Theta(\rho) \cdot \Delta\Theta(\tau' - \tau)) \tag{4}$$

where $r(\Theta)$ gives the Euler angles corresponding to the rotation matrix $\Theta$, $\Theta(\rho)$ gives the rotation matrix corresponding to the Euler angles $\rho$, and $\Delta\Theta(\Delta t)$ gives an incremental rotation matrix:

$$\Delta\Theta(\Delta t) = \exp\left(\Delta t \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}\right) \tag{5}$$

and where $\omega = (\omega_x, \omega_y, \omega_z)$ is the camera coordinate system angular velocity measurement from the rate gyro. Over an interval $[\tau, \tau']$ when the world coordinate system linear acceleration is assumed constant, $I_v$ and $I_t$ are given by the familiar equations:

$$I_v(\tau, \tau', \rho, v, g) = v + a(\tau' - \tau) \tag{6}$$

and

$$I_t(\tau, \tau', \rho, v, g, t) = t + v(\tau' - \tau) + \frac{1}{2}a(\tau' - \tau)^2 \tag{7}$$

where $a$ is the world coordinate system acceleration

$$a = \Theta(\rho) \cdot a' + g \tag{8}$$

and $a'$ is the camera coordinate system apparent acceleration given by the accelerometer.

Because inertial measurements are received at a higher rate than image measurements, the intervals $[\tau_{i-1}, \tau_i]$ between images span several inertial readings. To integrate the rotation, velocity, and translation over these larger intervals, $I_\rho$, $I_v$, and $I_t$ divide $[\tau_{i-1}, \tau_i]$ into the subintervals demarcated by the inertial measurement times and sequentially apply (4), (6), and (7) over each subinterval.

As with the distances in $E_{\text{visual}}$, all of the distances $D$ in $E_{\text{inertial}}$ are Mahalanobis distances. In the experiments described in Section 4, isotropic covariances have been used to specify the relative importance of the image, gyro, and accelerometer error terms. Specifically, we have used a standard deviation of 2.0 pixels, $10^{-4}$ radians, $10^{-3}$ m/s, and $10^{-3}$ m for the image, rotation, velocity, and translation error terms.

Two comments about $E_{\text{inertial}}$ are in order. First, note that $E_{\text{inertial}}$ does not make any assumptions restricting the relative timing between image and inertial readings. In particular, no synchronization between the image and inertial readings is required, and image and inertial readings can arrive at different rates, as long as each measurement is accurately timestamped. Second, a possible alternative formulation for $E_{\text{inertial}}$ is to use discrete differences to approximate the first and second derivatives of the estimated motion, and then require these derivatives to match the inertial measurements. But, this formulation requires that the durations between image times be small relative to the rate at which the derivatives change. Our formulation makes no such assumption, so our error function is suitable for cases where the duration between images is long.

Levenberg-Marquardt requires an initial estimate of the parameters. In the experiments described in Section 4, we have used estimates from an initial implementation of a visual-inertial online method that we have designed. We will describe this method in a future paper.

### 3.2 Visual-only and inertial-only estimates

In Section 4 we compare the estimates produced by our method to estimates produced using only visual or only inertial measurements. The algorithm used to find the visual-only estimates minimizes $E_{\text{visual}}$ with respect to the camera rotation and translation at the time of each image, and with respect to the three-dimensional point locations. This method is optimal and is essentially the same as bundle adjustment[14] or nonlinear shape-from-motion[13]. In our experiments, the initial estimates provided to this visual-only method are the same as the initial estimates provided to the visual-with-inertial algorithm, and are close to the correct solution. This ensures that the differences between the visual-inertial and visual-only estimates are not due to differences in the initial estimates.

To estimate motion from inertial measurements only, we simply integrate the inertial measurements forward from the first image time to the last image time. As initial conditions we use the initial position, initial linear velocity, and gravity vector estimated using our optimal algorithm.

## 4. Results

### 4.1 Overview

This section describes the results of running our algorithm on two datasets, one perspective and one omnidirectional, produced by identical motions obtained by mounting the sensor rig on a preprogrammed robotic arm. For each dataset, we compare the motions estimated using visual and inertial measurements, only visual measurements, and only inertial measurements.

### 4.2 Configuration

The sensor rig consists of a Sony XC-55 industrial vision camera, 3 orthogonally mounted CRS04 rate gyros from Silicon Sensing Systems, and a Crossbow CXL04LP3 3 degree of freedom accelerometer. The gyros and accelerometer measure motions of up to 150 degrees per second and 4 g, respectively. The camera exposure time is set to 1/200 second to reduce motion blur. To take conventional perspective images, the camera is paired with a 6 mm lens. To take omnidirectional images, the camera is paired with a 16 mm lens and a convex mirror.

Images were captured at 30 Hertz on a PC using a conventional frame grabber. To remove the effects of interlacing, only one field was used from each image, producing 640 × 240 pixel images. Voltages from the gyros and the accelerometer were simultaneously captured on the same PC at 200 Hertz with two separate Crossbow CXLDX analog-to-digital acquisition boards.

The camera intrinsic parameters (e.g., focal length and radial distortion) were calibrated using the method in [5]. This calibration also accounts for the reduced geometry of our one-field images. The accelerometer voltage-to-acceleration calibration was performed using a field calibration that accounts for non-orthogonality between the individual $x$, $y$, and $z$ accelerometers. The individual gyro voltage-to-rate calibrations were determined using a turntable with a known rotational rate. The fixed gyro-to-camera and accelerometer-to-camera rotations were assumed known from the mechanical specifications of the mount. For the omnidirectional images, we have assumed that the mirror is ideally positioned relative to the camera.

### 4.3 Observations

To perform experiments with known and repeatable motions, the rig was mounted on a Yaskawa Perfomer-MK3 robotic arm, which has a maximum speed of 3.33 meters per second and a payload of 2 kilograms. The programmed motion translates the camera $x$, $y$, and $z$ through seven prespecified points, for a total distance traveled of about two meters. Projected onto the $(x, y)$ plane, these points are located on a square, and the camera moves on a curved path between points, producing a clover-like pattern in $(x, y)$. The camera rotates through an angle of 270 degrees about the camera's optical axis during the course of the motion.

Each sequence consists of 152 images, approximately 860 gyro readings, and approximately 860 accelerometer readings. In the perspective sequence, 23 features were tracked, but only 5 or 6 appear in any one image. In the omnidirectional sequence, the wide field of view enabled tracking of 6 points throughout the entire sequence, although individual points sometimes temporarily left the camera's vertical field of view. In both sequences, the points were tracked using the Lucas-Kanade algorithm[9][1], but because the sequences contain repetitive texture and large interframe motions, mistracking was common and was corrected manually.

### 4.4 Estimated motion

As described in Section 3, our method estimates the six degree of freedom position and linear velocity of the camera at the time of each image, the world coordinate system location of each tracked point, and the world gravity vector. In this subsection, we'll give a brief overview of the estimates resulting from our experiments. For the sake of brevity, we will concentrate on the estimated $(x, y)$ translation.

Some aspects of the $(x, y)$ components of the estimated motion are shown graphically in Figures 1 and 2. The $(x, y)$ translation estimated using both visual and inertial data is shown as a smooth dash-dotted line in the left hand plot of Figure 1 for the perspective sequence, and in the right hand plot for the omnidirectional sequence. In each plot the seven squares show the known $(x, y)$ positions of the camera's ground truth motion. A summary of the error in these estimates versus ground truth is given in Table 1.

Similarly, the $(x, y)$ translations estimated using visual measurements only for the perspective and omnidirectional sequences are shown as the erratic solid lines in the left and right plots, respectively, of Figure 1. The summary of the error in these estimates is also given in Table 1. For the perspective sequence, the poor estimate is due to a combination of few points visible in each frame, and the planarity of the points. This leads to a large ambiguity between each camera position's rotation and translation, which is resolved by the rotational rate observations in the visual-with-inertial estimate. In the omnidirectional sequence, the overall shape of the visual-only estimate is nearly correct because all of the points are seen throughout most of the sequence. Some large scale errors are present due to points temporarily leaving the camera's vertical field of view, and the small scale erratic motion in the estimate is due to vibration between the omnidirectional camera rig's two components, the camera and mirror.

Each plot in Figure 1 also shows the $(x, y)$ components of the motion that results from integrating the inertial measurements only, as a diverging dashed line. This divergence is due to noise in the inertial readings and small errors in the gravity and initial velocity estimates used to integrate the data.

The left of Figure 2 shows, for the perspective sequence, covariance ellipses describing the estimated error covari-

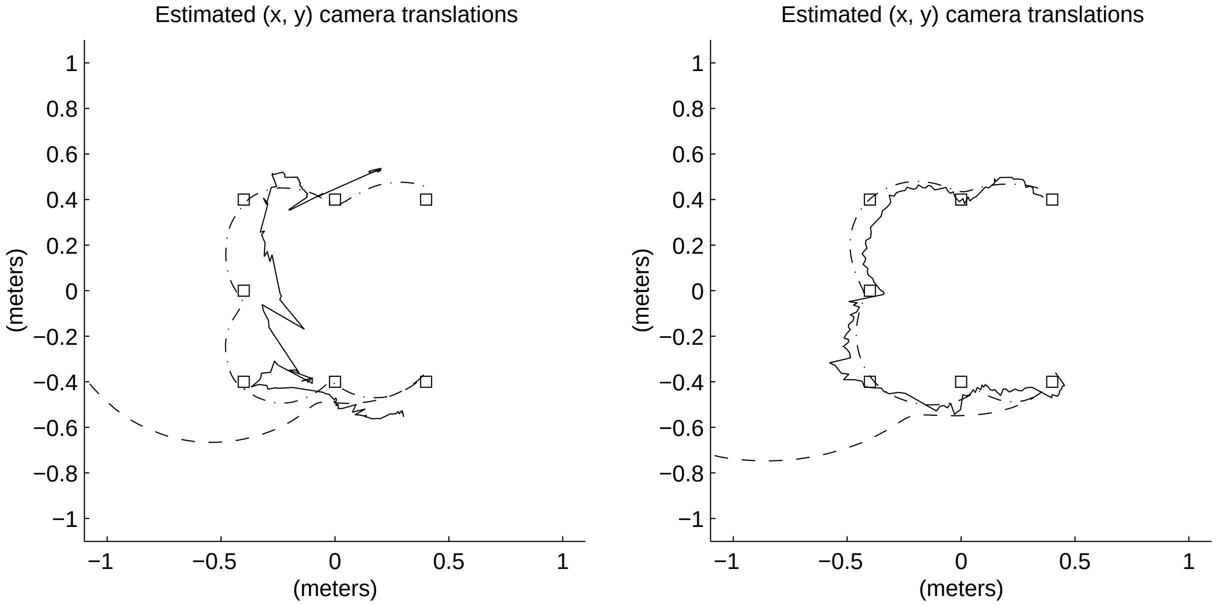

**Figure 1. The estimated ($x$, $y$) camera translations for the perspective sequence (left) and the omnidirectional sequence (right). The visual-only, inertial-only, and visual-with-inertial translation estimates are shown as the solid, dashed, and dash-dotted lines, respectively. The boxes show the known ($x$, $y$) ground truth positions.**

| | rotation error (radians) | translation error (centimeters) |
|---|---|---|
| perspective visual only | 0.45 / 0.56 | 15.1 / 25.1 |
| perspective visual and inertial | 0.07 / 0.10 | 4.3 / 6.3 |
| omni visual only | 0.09 / 0.15 | 8.5 / 12.8 |
| omni visual and inertial | 0.09 / 0.11 | 7.2 / 9.0 |

**Table 1. Errors versus ground truth for the four estimates. Each entry gives the average error before the slash and the maximum error after the slash.**

ances in the ($x$, $y$) translations for every fifteenth image of the sequence. The 1 standard deviation boundary resulting from using visual data only are shown as the large dotted ellipses. The solid ellipses show the error boundaries that result from using both visual and inertial data; in this case, 5 standard deviation boundaries are used for visibility. To provide a direct comparison, both the visual-only and visual-with-inertial parameter covariances are evaluated at the visual-with-inertial parameter estimate. This solution, shown as a dash-dotted line, is the same as in Figure 1. The right of Figure 2 shows the corresponding boundaries for the omnidirectional sequence; in this case, 5 $\sigma$ boundaries are shown for both the visual-only and visual-with-inertial estimates.

## 5. Conclusions and future work

We have presented an algorithm for finding optimal motion estimates using both visual and inertial data. Our experimental results show that this algorithm can eliminate the ambiguities in motion that result from using either modality alone, producing highly accurate motion estimates even when the estimates from either modality alone are very poor.

As mentioned in Section 3, we are currently developing an online method for estimating sensor motion from image and inertial measurements, which uses many of the same components described in this paper. To improve the accuracy of sensor motion estimation over longer time periods, we are investigating the incorporation of vehicle motion models and *a priori* scene information into both the optimal and online methods.

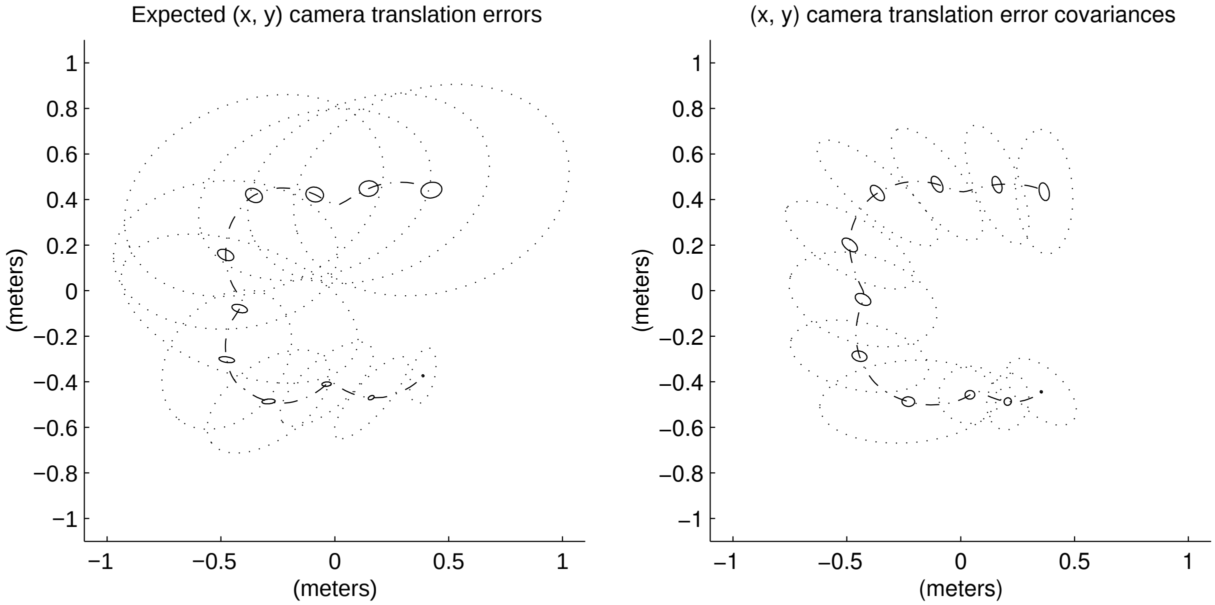

**Figure 2. The ($x$, $y$) camera translation error covariances for the perspective sequence (left) and the omnidirectional sequence (right). The dotted and solid ellipses give the visual-only and visual-with-inertial error boundaries, respectively. The dash-dotted curve is the visual-with-inertial motion estimate. For visibility, 5 $\sigma$ error boundaries are shown for all covariances, except in the perspective visual-only case, where the 1 $\sigma$ boundary is shown.**

## References

[1] S. Birchfield. KLT: An implementation of the Kanade-Lucas-Tomasi feature tracker. `http://vision.stanford.edu/~birch/klt/`.

[2] M. J. Brooks, W. Chojnacki, D. Gawley, and A. van den Hengel. What value covariance information in estimating vision parameters? In *Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, Vancouver, Canada, 2001.

[3] M. Deans and M. Hebert. Experimental comparison of techniques for localization and mapping using a bearing-only sensor. In D. Rus and S. Singh, editors, *Experimental Robotics VII*, pages 395–404. Springer-Verlag, Berlin, 2001.

[4] A. Huster and S. M. Rock. Relative position estimation for intervention-capable AUVs by fusing vision and inertial measurements. In *Twelfth International Symposium on Unmanned Untethered Submersible Technology*, Durham, New Hampshire, August 2001.

[5] Intel corporation open source computer vision library. `http://www.intel.com/research/mrl/research/opencv/`.

[6] S.-H. Jung and C. J. Taylor. Camera trajectory estimation using inertial sensor measurements and structure from motion results. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 2, pages 732–737, Kauai, Hawaii, December 2001.

[7] I. Kaminer, W. Kang, O. Yakimenko, and A. Pascoal. Application of nonlinear filtering to navigation system design using passive sensors. *IEEE Transactions on Aerospace and Electronic Systems*, 37(1):158–172, January 2001.

[8] Y. Kanazawa and K. Kanatani. Do we really have to consider covariance matrices for image features? In *Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, Vancouver, Canada, July 2001.

[9] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Seventh International Joint Conference on Artificial Intelligence*, volume 2, pages 674–679, Vancouver, Canada, August 1981.

[10] P. F. McLauchlan. The variable state dimension filter applied to surface-based structure from motion. Technical Report VSSP-TR-4/99, University of Surrey, Guildford, UK, 1999.

[11] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, United Kingdom, 1992.

[12] G. Qian, R. Chellappa, and Q. Zhang. Robust structure from motion estimation using inertial data. *Journal of the Optical Society of America A*, 18(12):2982–2997, December 2001.

[13] R. Szeliski and S. B. Kang. Recovering 3D shape and motion from image streams using nonlinear least squares. *Journal of Visual Communication and Image Representation*, 5(1):10–28, March 1994.

[14] P. R. Wolf. *Elements of Photogrammetry*. McGraw-Hill, New York, 1983.