Reckless motion estimation from omnidirectional image and inertial measurements

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

Abstract

Two approaches to improving the accuracy of camera motion estimation from image sequences are the use of omnidirectional cameras, which combine a conventional camera with a convex mirror that magnifies the field of view, and the use of both image and inertial measurements, which are highly complementary. In this paper, we describe optimal batch algorithms for estimating motion and scene structure from either conventional or omnidirectional images, with or without inertial data.

We also present a method for motion estimation from inertial data and the tangential components of image projections. Tangential components are identical across a wide range of conventional and omnidirectional projection models, so the resulting method does not require any accurate projection model. Because this method discards half of the projection data (i.e., the radial components) and can operate with a projection model that may grossly mismodel the actual camera behavior, we call the method "reckless" motion estimation, but we show that the camera positions and scene structure estimated using this method can be quite accurate.

1. Introduction

Recent omnidirectional cameras combine a conventional camera with a convex mirror that multiplies the mirror's field of view, typically to 360 degrees in azimuth and 90-140 degrees in elevation. Omnidirectional images are likely to increase the accuracy of camera motion estimation, because features viewed over the omnidirectional field of view constrain the camera's position from a wider range of directions, and because each tracked feature is likely to be seen through a larger percentage of the image sequence. A second approach to increasing the accuracy of camera motion estimation is to use both image and inertial measurements, which are highly complementary. For instance, inertial measurements can resolve ambiguities in image-only motion estimation that result from viewing a degenerate scene, too few features, of features in an accidental geometric configuration; reduce the discontinuity in the estimated motion that results from features entering and leave the field of view; establish the global scale; and make motion estimation more robust to mistracked features. On the other hand, image measurements can counteract error accumulation that results from integrating inertial data alone, and can be used to separate the effects of acceleration, gravity, and accelerometer bias in accelerometer readings.

We have developed optimal batch algorithms for estimating sensor motion and scene structure from both conventional and omnidirectional images, with or without inertial measurements. In this paper, we present these algorithms and experimental results that show the relative advantages of omnidirectional images, inertial measurements, and both omnidirectional images and inertial measurements in combination. The algorithm for estimating sensor motion from image and inertial data is particularly strong in that it can produce good estimates even if the estimates produced by image or inertial measurements alone are poor, and typically converges in just a few iterations even if the initial estimate is poor.

We also present a more radical method for motion estimation from image and inertial measurements. This method uses only the tangential components of the image projections, which are identical across a wide range of conventional and omnidirectional projection models, and can therefore operate without any accurate projection model. Because this method discards half of the projection data (i.e., the radial components) and can operate with a projection model that may grossly mismodel the actual camera behavior, we call the method "reckless" motion estimation, but the camera positions and scene structure recovered using this method were quite accurate in our initial test, which we also describe.

For brevity, we have excluded a discussion of related work. Please see [5] and [6] for reviews of the existing work most closely related to our approach to motion from omnidirectional cameras, and to our method for motion estimation

from image and inertial measurements, respectively.

2. Method

2.1 Optimal motion from image data

Our method for estimating camera motion and scene structure from images alone is similar to nonlinear shape-from-motion and bundle adjustment, and uses Levenberg-Marquardt to minimize:

$$E_{\text{visual}} = \sum_{i,j} D(\pi(C_{\rho_i,t_i}(X_j)) - x_{ij}) \tag{1}$$

E_{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. ρ_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_{ρ_i,t_i} is the world-to-camera transformation specified by ρ_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.

π gives the image projection of a three-dimensional point specified in the camera coordinate system, and can be either conventional (e.g., orthographic or perspective) or omnidirectional. Our omnidirectional projection model is described in [5]. This projection model handles noncentral omnidirectional cameras, in which the camera-mirror combination does not have a single effective viewpoint, and omnidirectional cameras in which the camera and mirror are misaligned by a general but known (i.e., calibration) 6 DOF transformation.

All of the individual distance functions D are Mahalanobis distances. Common choices for the covariances defining the distances are uniform isotropic covariances, or directional covariances determined using image texture in the vicinity of each feature[3][2].

The error is minimized with respect to the six degree of freedom camera position ρ_i, t_i at the time of each image, and with respect to the three-dimensional position X_j of each tracked point.

Motion estimation from image data alone does not recover the overall scale of the estimated motion and scene. That is, scaling the recovered camera translations and three-dimensional points by any factor produces an estimate that explains the observations as well as the original estimate. Motion estimation from image and inertial data, which we describe in subsection 2.2 below, does recover the scale factor.

2.2 Estimation from image and inertial data

Image and inertial data are highly complementary modalities for sensor motion estimation. Our algorithm for estimating sensor motion from image and inertial data extends the image only algorithm described in section 2.1, and uses Levenberg-Marquardt to minimize:

$$E_{\text{combined}} = E_{\text{visual}} + E_{\text{inertial}} + E_{\text{prior}} \tag{2}$$

The visual error term E_{visual} is the same as that described in section 2.1. The inertial error term is:

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

E_{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 τ_i is the time image i was captured. ρ_i and t_i are the camera rotation and translation at time τ_i, just as in the equation for E_{visual} above. v_i gives the camera's linear velocity at time τ_i. g and b are the world coordinate system gravity vector and accelerometer bias, respectively.

I_ρ, I_v, and I_t integrate the inertial observations to produce estimates of ρ_i, v_i, and t_i from initial values ρ_{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_ρ 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 Θ, $\Theta(\rho)$ gives the rotation matrix corresponding to the Euler angles ρ, 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

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

where a is the world coordinate system acceleration

$$a = \Theta(\rho) \cdot (a' + b) + g \quad (8)$$

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

The bias prior term E_{prior} is:

$$E_{\text{prior}} = f \cdot b^T C_b^{-1} b \quad (9)$$

The bias prior error term (9) is small if the bias is near zero, and reflects our expectation that the accelerometer voltage corresponding zero acceleration is close to the precalibrated value. As above, f is the number of images and b is the accelerometer bias. C_b is the accelerometer bias prior covariance, which we take to be isotropic with standard deviations 0.5 m/s^2.

The combined error function is minimized with respect to the six degree of freedom camera position ρ_i, t_i at the time of each image; the camera linear velocity v_i at the time of each image; the three-dimensional point positions of each tracked points X_j; the gravity direction with respect to the world coordinate system g; and optionally, the accelerometer bias b.

2.3 Reckless motion estimation

The normalized, or unitless, perspective projection is the most common projection model in shape-from-motion:

$$\pi_{\text{perspective}}(x, y, z) = (x/z, y/z) \quad (10)$$

Under this projection model, the image projection lies on a ray from the image center specified by the angle:

$$\theta = \text{atan2}(y/z, x/z) = \text{atan2}(y, x) \quad (11)$$

If camera focal length is modeled, the projection becomes:

$$\pi_{\text{perspective}}(f, x, y, z) = (fx/z, fy/z) \quad (12)$$

The projection then lays on the ray specified by the angle:

$$\theta = \text{atan2}(fy/z, fx/z) = \text{atan2}(y, x) \quad (13)$$

That is, the ray from the image center on which the projection lies is unchanged if focal length is modeled, and is independent of the specific focal length. Similarly, it is easy to see that the ray on which the projection lies is unchanged if radial distortion is modeled and is independent of the specific radial distortion values.

The normalized orthographic and weak perspective projection models are:

$$\pi_{\text{orthographic}}(x, y, z) = (x, y) \quad (14)$$

$$\pi_{\text{weak-perspective}}(x, y, z) = (x/z_o, y\ z_o) \quad (15)$$

where z_o is the distance to the origin in the camera coordinate system. In these two cases, the projections lie on the rays specified by

$$\theta = \text{atan2}(y, x) \quad (16)$$

$$\theta = \text{atan2}(y/z_o, x/z_o) = \text{atan2}(y, x) \quad (17)$$

respectively. So, the ray on which the projection lies is unchanged whether perspective, weak perspective, or orthographic projection is assumed. In fact, the same is true for any reasonable single viewpoint or noncentral omnidirectional projection model, as long as the camera's optical axis and the mirror's axis of revolution are assumed to be aligned.

So, if we consider only the tangential component of each projection, it doesn't matter what intrinsics calibration, projection model, or camera type we assume. When coupled with inertial data, the tangential components provide enough information to perform motion estimation without any accurate camera model. We call motion estimation from the tangential components and inertial data "reckless" motion estimation, since it throws away half of the image data and uses a projection model that may grossly mismodel the camera's actual behavior, but as we'll show in section 3.3, motion estimated in this manner can be quite accurate.

Reckless motion estimation might be advantageous for conventional cameras when the camera's focal length or radial distortion are unknown, or when the camera's field of view is extreme (e.g., some lenses have a field of view of more than 180 degrees) and may not be well modeled by perspective projection with radial distortion. Reckless motion estimation might be advantageous for omnidirectional cameras when the mirror's profile is unknown, or when the distance between the camera and mirror is unknown.

The algorithm described in section 2.2 can easily be adapted to perform this estimation by elongating the observed projection covariances so that they lie along the image's radial lines, as illustrated in Figure 1. Choosing the projection covariances in this way specifies that reprojection error along a projection's radial component should be discounted, while reprojection error along the projection's tangential component should be weighted as normal. While the ideal projection distribution for reckless motion estimation is uniform along the radial component, we've found that approximating the ideal distributions by directional Gaussians is convenient and sufficient for proof of concept, as shown in section 3.3.

On the other hand, without inertial data the tangential components are not sufficient to recover the six degree of freedom camera positions and three-dimensional point positions: any estimate that aligns all of the camera optical axes and all of the three-dimensional points along a single common axis will reproject the points to the image centers, which are trivially on the same rays as the observed projections.

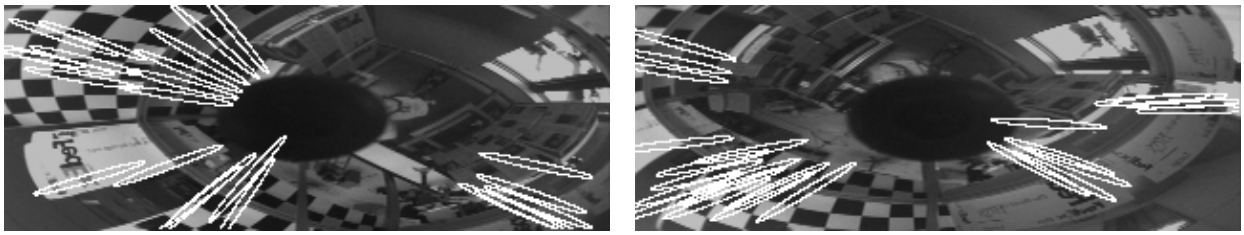

Figure 1. The algorithm in section 2.2 can be used to perform reckless motion estimation if the observed projection covariances are chosen to lie along the image's radial lines, as shown in these two example images. For clarity, the covariances have been shown here with a relatively small variance in the radial direction. The variances actually used in our tests are (2.0 pixels)^2 in the tangential direction and (10^8 pixels)^2 in the radial direction.

3. Results

3.1 Relative estimation accuracy

To evaluate the relative advantages of omnidirectional image and inertial measurements, we captured one conventional dataset and one omnidirectional dataset, both with corresponding inertial measurements, by mounting our sensor rig on a Yaskawa Performer-MK3 robotic arm. In this section we describe the preprogrammed motion and resulting observations, but for brevity we exclude a discussion of the sensor rig and its calibration. Please see [6] for a more detailed description of the sensor rig and its calibration.

The arm provides known and repeatable motions, and 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 pre-specified points, for a total distance traveled of about three 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[4][1], but because the sequences contain repetitive texture and large interframe motions, mistracking was common and was corrected manually.

As described in Section 2, our image-only method estimates the six degree of freedom position at the time of each image and the world coordinate system location of each tracked point. Our image-and-inertial 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, the world gravity vector, and the accelerometer bias. For the sake of brevity, we will concentrate here on the estimated (x, y) translation.

Some aspects of the (x, y) components of the estimated motion are shown graphically in Figure 2. The (x, y) translation estimated using both visual and inertial data is shown as a smooth dashed line in the left hand plot of Figure 2 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 2. 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 2 also shows the (x, y) components of the motion that results from integrating the inertial measurements only, as a diverging dash-dotted line. This divergence is due to noise in the inertial readings and small errors in the estimated initial velocity, gravity, and accelerometer bias used to integrate the data. The divergence differs slightly in the two plots because the accelerometer noise and

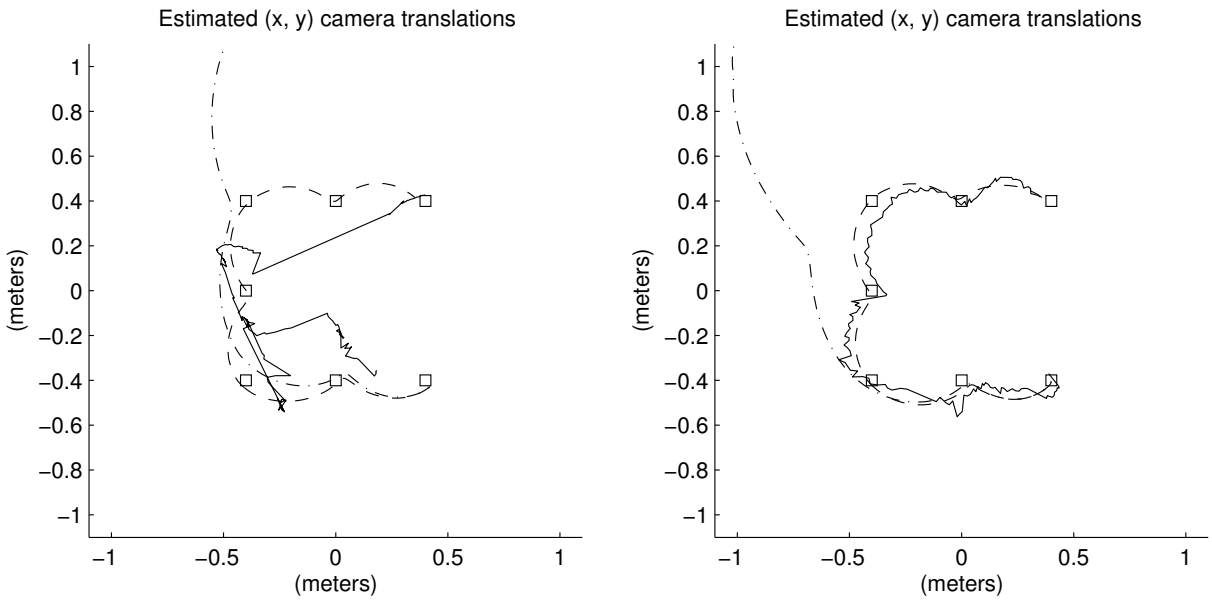

Figure 2. The estimated (x, y) camera translations for the perspective sequence (left) and the omnidirectional sequence (right). The visual-only, inertial-only, and visual-with-inertial translation estimates are shown as the solid, dashed, and dash-dotted lines, respectively. The boxes show the known (x, y) ground truth positions.

	rotation error (radians)	translation error (centimeters)	scale error
perspective, image only	0.252 / 0.372	23.5 / 33.1	N/A
omnidirectional, image only	0.094 / 0.148	8.54 / 12.9	N/A
perspective, image and inertial	0.108 / 0.136	4.03 / 6.60	-5.5%
omnidirectional, image and inertial	0.106 / 0.137	3.67 / 4.72	+2.2%

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.

estimate errors differ in the two datasets.

For these datasets, omnidirectional data is an improvement over conventional data in both the image-only and image-and-inertial cases, and image-and-inertial is an improvement over image-only in both the conventional and omnidirectional cases. As one might expect, of all four cases, the omnidirectional, image-and-inertial estimate is closest to ground truth.

3.2 Effect of bias estimation

During the minimization of (1) and (2) we usually choose to fix any problem parameters that can be calibrated beforehand, such as the camera intrinsics, gyro voltage-to-rate and accelerometer voltage-to-acceleration mappings, and omnidirectional camera camera-to-mirror transformation. However, the accelerometer voltage corresponding to zero acceleration typically changes between successive power ups and with temperature, which makes precalibrating the accelerometer problematic. This problem makes the estimation of the accelerometer bias, or difference between the calibrated zero acceleration voltage and actual zero acceleration voltage, desirable.

Table 2 summarizes the results of running the datasets described in section 3.1, using the image-with-inertial algorithm with and without bias estimation. The second half of this table is the same as the second half of Table 1, and is included for convenience. In both the conventional and omnidirectional cases, estimating the bias improves the estimate relative to ground truth, and in the omnidirectional dataset the effect is dramatic.

3.3 Reckless motion estimation

In this subsection we give the results from an initial test of reckless motion estimation. In this test, the sensor con-

	rotation error (radians)	translation error (centimeters)	scale error
perspective, bias assumed zero	0.171 / 0.271	4.44 / 7.70	-5.6%
omnidirectional, bias assumed zero	0.111 / 0.148	7.27 / 9.49	+24.6%
perspective, bias estimated	0.108 / 0.136	4.03 / 6.60	-5.5%
omnidirectional, bias estimated	0.106 / 0.137	3.67 / 4.72	+2.2%

Table 2. Errors versus ground truth for the two datasets, without and with bias estimation. Each entry gives the average error before the slash and the maximum error after the slash.

figuration and calibration are the same as those used in sections 3.1 and 3.2, and omnidirectional images are used, but the acquired dataset differs in three ways from the one used in sections 3.1 and 3.2:

- In the sequence in sections 3.1 and 3.2, the camera's optical axis always points in the same direction, but since tangential components alone give no information about the camera's translation along the optical axis, the visual data is not useful for reducing inertial integration drift in that direction. The sequence used in this section emphasizes rotation along all three sensor axes, which is required to reduce drift along the camera's optical axis when reckless motion estimation's tangential projection model is used. For the same reason, the initial velocity was fixed to (0, 0, 0) during minimization.

- The sequence includes 36 points tracked across 94 images, which is a large number of features compared to the Spartan datasets described in sections 3.1 and 3.2.

- Accurate ground truth is not available for the motion, so we have used camera and point positions estimated from the image and inertial algorithm described in section 2.2, using the correct (i.e., equiangular omnidirectional) projection model with isotropic projection covariances, as ground truth.

For the reckless motion estimation, we have assumed orthographic projection and projection covariances with variance $(2 \text{ pixels})^2$ in the tangential direction and variance $(10^8 \text{ pixels})^2$ in the radial direction. Images 37 and 68 from the sequence are shown in Figure 1, with the projection covariances overlaid. The camera intrinsics are assumed to be 1 pixel focal length and 0 radial distortion, whereas the calibrated values are 2210 pixels focal length and -0.157 radial distortion.

An initial estimate far from the correct estimate is used:

- All camera positions are placed at the origin

- The point positions are initialized by backprojecting the image position at which they first appear from the origin to a fixed distance in space, using the orthographic projection model. For points that appear in the first image, the resulting initial estimates are consistent with the tangential components observed in the first image.

- The velocities, gravity, and bias are all initialized to zero.

The average and maximum rotation errors in the reckless estimate, relative to the isotropic ground truth estimate, are 0.109 radians and 0.128 radians, respectively. The average and maximum translation errors are 4.05 cm and 9.56 cm, respectively, relative to a total distance traveled of about 2 meters. The average and maximum point errors are 7.83 cm and 26.1 cm, respectively. The scale error is +8.1%.

As an additional note, in our test we have used omnidirectional images and an orthographic projection model. But, the principle stays the same for combination of camera and projection model, and we would expect similar reckless estimation results from any combination.

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] 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.

[4] 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.

[5] D. Strelow, J. Mishler, S. Singh, and H. Herman. Extending shape-from-motion to noncentral omnidirectional cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2001)*, Wailea, Hawaii, October 2001.

[6] D. Strelow and S. Singh. Optimal motion estimation from visual and inertial data. In *IEEE Workshop on the Applications of Computer Vision*, Orlando, Florida, December 2002.