Combined Visual and Inertial Navigation for an Unmanned Aerial Vehicle

Jonathan Kelly, Srikanth Saripalli, Gaurav Sukhatme

Combined Visual and Inertial Navigation for an Unmanned Aerial Vehicle

Jonathan Kelly, Srikanth Saripalli and Gaurav S. Sukhatme

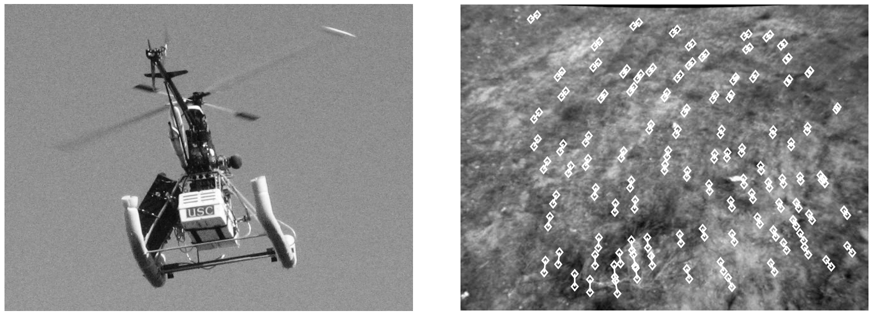
Robotic Embedded Systems Laboratory
University of Southern California
Los Angeles, California, USA 90089-0781
{jonathsk,srik,gaurav}@usc.edu

Summary. We describe an UAV navigation system which combines stereo visual odometry with inertial measurements from an IMU. Our approach fuses the motion estimates from both sensors in an extended Kalman filter to determine vehicle position and attitude. We present results using data from a robotic helicopter, in which the visual and inertial system produced a final position estimate within 1% of the measured GPS position, over a flight distance of more than 400 meters. Our results show that the combination of visual and inertial sensing reduced overall positioning error by nearly an order of magnitude compared to visual odometry alone.

1 Introduction

Unmanned aerial vehicles (UAVs) typically depend on GPS to provide absolute positioning information for navigation. There are situations, however, in which GPS may be unavailable, for example when flying through urban canyons. In this paper, we describe a passive system that allows an UAV to navigate reliably without GPS or other wide-area positioning information. The system combines stereo *visual odometry* with measurements from an inertial measurement unit (IMU) to continuously estimate the pose (position and attitude) of the vehicle.

For the work presented here, we use a robotic helicopter as our UAV platform. Helicopters are able to rapidly change all of their 6-DOF pose parameters simultaneously, making this a particularly challenging estimation problem. Our solution is designed to take advantage of complimentary IMU and camera sensor characteristics. The IMU is able to accurately measure rapid changes in angular rotation rates and linear accelerations, but is subject to unbounded low-frequency drift. Visual motion estimates, in contrast, are generally more accurate when the cameras' field of view changes relatively slowly. By fusing their output, each sensor is able to compensate for the weaknesses inherent in the other.



(a) (b)

Fig. 1. (a) Helicopter during a data acquisition flight. The avionics box (white) is located beneath the main helicopter chassis; the vision computer (black) is mounted on left-hand side of the vehicle. Both stereo cameras are also visible near the front of the landing skids. (b) Two-frame feature tracks overlaid on a left stereo image.

We formulate visual odometry as a maximum-likelihood estimation (MLE) problem, where point landmarks are tracked across sequential stereo image pairs; the changes in triangulated landmark positions are used to determine robot ego-motion. These visual measurements are then optimally combined with data from the on-board IMU in an extended Kalman filter (EKF). Although the resulting trajectory estimates are produced by integration, we demonstrate that an absolute positioning accuracy to within 1% of the measured GPS position can be achieved, over flight distances of more than 400 meters. As a percentage, this difference is almost an order of magnitude lower than the result using visual odometry alone.

2 Related Work

Visual odometry (VO) describes the process of incrementally estimating changes in robot pose by identifying and tracking visual landmarks in the environment. Many VO implementations use stereo cameras, as stereo allows the depth of landmarks to be calculated directly from known camera geometry. However, monocular, omnidirectional [1] and spherical cameras [2] have been employed as well. In particular, Nistér et al. have developed a VO system that operates in real-time with either monocular or stereo imagery [3]. VO is also used on-board the NASA Mars Exploration Rover (MER) robots [4] – the rovers rely on visual estimates during periods when wheel odometry is unreliable, such as when driving over high-slip terrain.

Examples of combined visual and inertial sensing include a system described by Roumeliotis et al. in [5], designed to safely and precisely land a spacecraft on the surface of a planetary body. An EKF is used to fuse monocular camera, laser altimeter, and IMU data. The laser altimeter measurements provide an absolute scale factor for the single-camera motion estimates. Unlike [5], we determine scale and depth directly from stereo correspondences.

In [6], Amidi et al. present a visual odometer designed specifically for an autonomous helicopter. They estimate vehicle attitude using gyroscopes and vehicle position by tracking ground targets with stereo cameras. To reduce computational demands, the targets are assumed to be planar and the camera view is assumed to change slowly with time. Our approach allows for operation over varied terrain, as long as the captured images have sufficient texture.

Finally, the work discussed here is closely related to aerial simultaneous localization and mapping (see, for example, [7]), although we are primarily concerned with point-to-point navigation, and so do not maintain a map of landmark positions.

3 Helicopter Platform and Vision System

AVATAR (Autonomous Vehicle Aerial Tracking and Reconnaissance), our experimental testbed, is a robotic helicopter built on a modified Bergen RC chassis. The helicopter is equipped with a suite of avionics hardware including a Novatel RT-2 GPS receiver, Inertial Science ISIS inertial measurement unit, PNI TCM2-50 electronic compass, and a PC-104 Linux computer for autonomous control. Our ISIS IMU has roll, pitch and yaw rate gyros and three independent single-axis accelerometers. A full description of the control system is given in [8].

For our visual navigation studies, the helicopter carries an additional Mini-ITX Linux computer and two color FireWire cameras from Videre Design. Each camera has a resolution of 640×480 pixels and a field of view of 50 degrees horizontally. The cameras are mounted on a stereo bench with a 50 cm baseline.

4 Visual Odometry

Our stereo visual odometry algorithm is based on the approach described by Matthies and Shafer in [9] and refined by Matthies in [10]. We track point landmarks across sequential stereo image pairs, and find the incremental change in camera pose by aligning corresponding sets of triangulated landmark positions. The steps in the algorithm are described in more detail below.

4.1 Feature Tracking and Stereo Triangulation

Given an initial stereo pair, the first step in the motion estimation process is to identify landmarks that can be tracked reliably. We use the KLT algorithm [11] to find salient feature points in the left stereo image. For each selected left image point, we then search for a corresponding point in the right image using normalized cross-correlation with a 15-pixel square correlation window. Our stereo cameras are accurately calibrated before each flight, allowing us to limit this correspondence search to a narrow region centered on the right epipolar line. If the correlation score for the best match in the right image is below a fixed threshold (0.75 in our implementation), the point pair is not included in further processing. For matches above the threshold, we determine the sub-pixel disparity by fitting a biquadratic polynomial to the correlation values in a 3×3 region around the integer correlation peak.

The 3D positions of the landmarks are found by stereo triangulation. To model error in the position estimates, we consider the image coordinates as point vectors corrupted by zero-mean, white Gaussian noise. The noise covariance is approximated directly from the curvature of the biquadratic polynomial used for the sub-pixel disparity calculation. A 3D covariance ellipsoid is computed for each landmark using standard Gaussian error propagation [4].

As each new stereo pair is acquired, the KLT tracker updates the 2D positions of all tracked landmarks in the most recent left camera image. The 3D positions of the landmarks are then re-triangulated with respect to the current stereo bench coordinate frame. When a landmark is no longer visible in either the left or right camera image, it is replaced by a newly-initialized landmark, in order to track an approximately constant number of landmarks over time. In our present implementation, we attempt to track 200 landmarks from frame to frame.

4.2 Robust Visual Motion Estimation

At each time step, the triangulation procedure above yields two sets of corresponding 3D landmark observations, before and after the helicopter has undergone an unknown rotation R and translation T. The relationship between each pair of observations is:

$$P_{i,c} = RP_{i,p} + T + e_i \tag{1}$$

where $P = [p_x, p_y, p_z]^T$ is a 3D position vector, the first subscript indexes the specific landmark, and the second subscript denotes whether the observation was made at the current or previous time step. The vector e_i is a zero-mean, white Gaussian noise term that models the combined errors in both the current and previous position estimates. There is one such equation for each observed landmark.

To solve (1), we linearize using the first-order Taylor expansion of the rotation matrix R with respect to roll, pitch and yaw angles, $\Theta = [\alpha, \beta, \gamma]^T$:

$$P_{i,c} \approx R_0 P_{i,p} + J_{i,p}(\Theta - \Theta_0) + T + e_i \tag{2}$$

Here, $J_{i,p}$ is the Jacobian for landmark i with respect to Θ, evaluated at an initial rotation Θ_0. The noise vector e_i has covariance $\Sigma_i = \Sigma_{i,c} + R_0 \Sigma_{i,p} R_0$, where the point covariance matrices $\Sigma_{i,c}$, $\Sigma_{i,p}$ are determined during the stereo triangulation step. Using our Gaussian error model, the MLE values for Θ and T are found by minimizing the objective function:

$$M(\Theta, T) = \sum_i r_i W_i r_i^T \tag{3}$$

$$r_i = P_{i,c} - R_0 P_{i,p} - J_{i,p}(\Theta - \Theta_0) - T \tag{4}$$

where W_i is the inverse covariance matrix for e_i. After differentiating (3) with respect to Θ and T and setting the result to zero, we obtain:

$$\left[\sum_{i=0}^{n} H_i^T W_i H_i\right] \begin{bmatrix} \Theta \\ T \end{bmatrix} = \left[\sum_{i=0}^{n} H_i^T W_i L_i\right] \tag{5}$$

with $H = [J_{i,p} \; I]$ and $L_i = P_{i,c} - R_0 P_{i,p} + J_{i,p}\Theta_0$; a detailed derivation of this result can be found in [10].

The final motion estimate is obtained by iteratively computing (5), using the previous estimate of R the as the value of R_0 for the following iteration. The covariance matrix for the result is:

$$\Sigma_{\Theta,T} = \left[\sum_i H_i^T W_i H_i\right]^{-1} \tag{6}$$

To ensure that the estimation algorithm is robust against incorrect data associations and tracking errors, we embed the computation of (5) in a RANSAC procedure [12].

5 Extended Kalman Filter-Based State Estimation

Visual and inertial measurements are fused in an extended Kalman filter (EKF) to produce an estimate of the vehicle state. We use the continuous-discrete formulation of the EKF, in which the state estimate is propagated according to the underlying continuous-time non-linear system dynamics, while measurement updates are made at discrete time steps. Our state vector includes the position of the helicopter in the global NED frame, the velocity of the helicopter in the body frame, the attitude of the helicopter, gyroscope biases, accelerometer biases and the magnitude of the gravity vector. Further details on the EKF implementation are available in [13].

The filter process model is driven by the IMU linear accelerations and angular velocities, which substitute for control inputs in the system dynamics equations. Our use of the continuous-discrete form of the EKF is particularly important in this case, since measurements from the IMU arrive at approximately 100 Hz, at least three times faster than measurements from the vision system. We incorporate the relative (frame-to-frame) visual motion estimates into the filter as linear and angular velocities, after transforming from the stereo bench coordinate frame to the helicopter body coordinate frame.

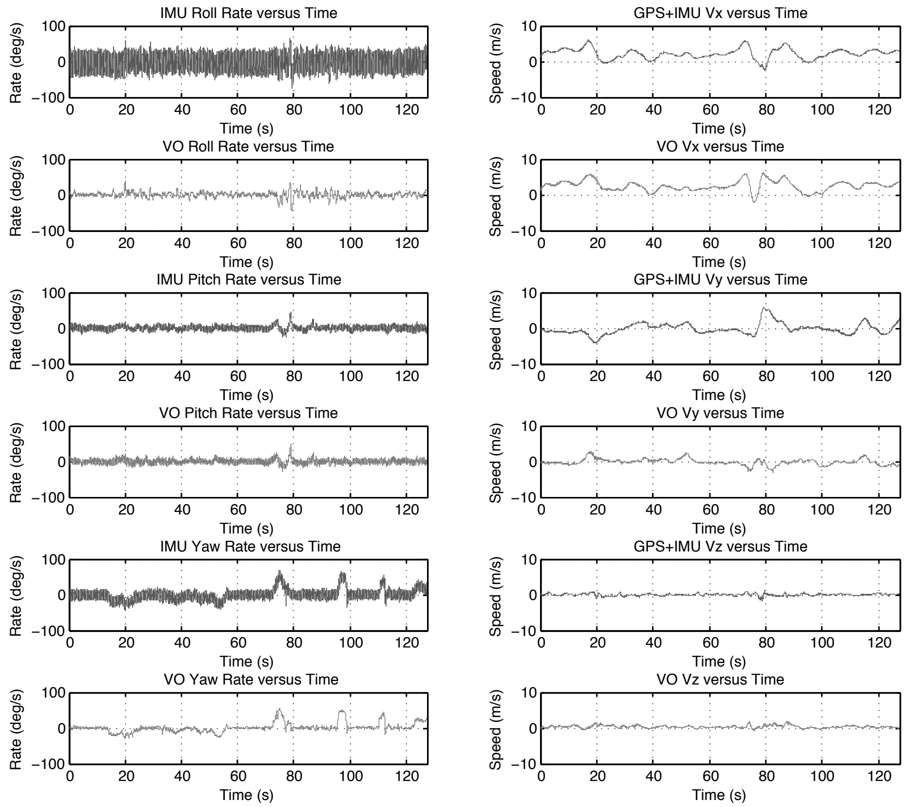


Fig. 2. Comparison of angular rates measured by the IMU and by VO (left column), and linear velocities measured by GPS+IMU and by VO (right column). All values are expressed in the helicopter body coordinate frame.

6 Experiments

To determine the performance of the proposed navigation system, we ran several experiments with the AVATAR helicopter at our test site in Downey, California. During each experiment, the helicopter was flown manually while we logged data from the on-board GPS receiver, IMU and the stereo cameras. The data were analyzed offline.

For the flight trials, our stereo cameras were mounted above the landing skids at the front of the vehicle, with the camera optical axes pointed at an angle of 60 degrees down from horizontal. We selected this oblique pointing angle, instead of a nadir pointing angle, so that we could perform several additional tracking and obstacle detection experiments using the same data. Images were captured from both cameras at 30 frames per second.

In the remainder of the paper we present results from our longest experiment, with a duration of 127.4 seconds and a ground track of 405.5 meters,

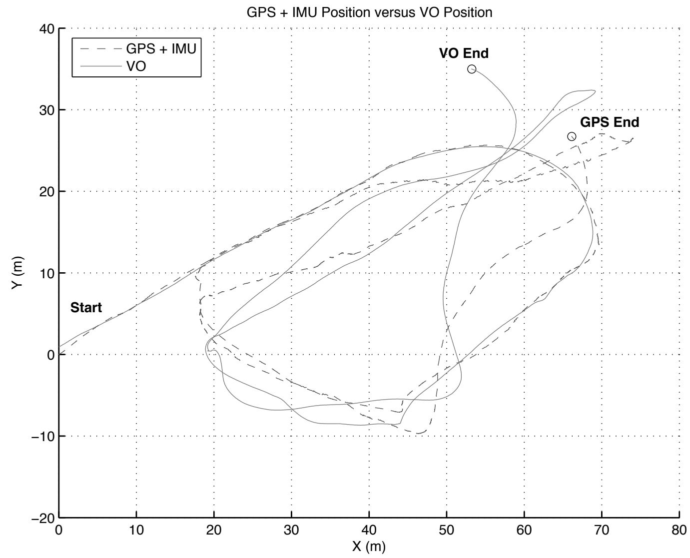


Fig. 3. Filtered GPS+IMU position estimate versus VO position estimate. The VO trajectory exhibits a characteristic misalignment due to the integration of small orientation errors over time. The difference between the final GPS+IMU position and the final VO position is 14.7 meters (after a 405.5 meter flight).

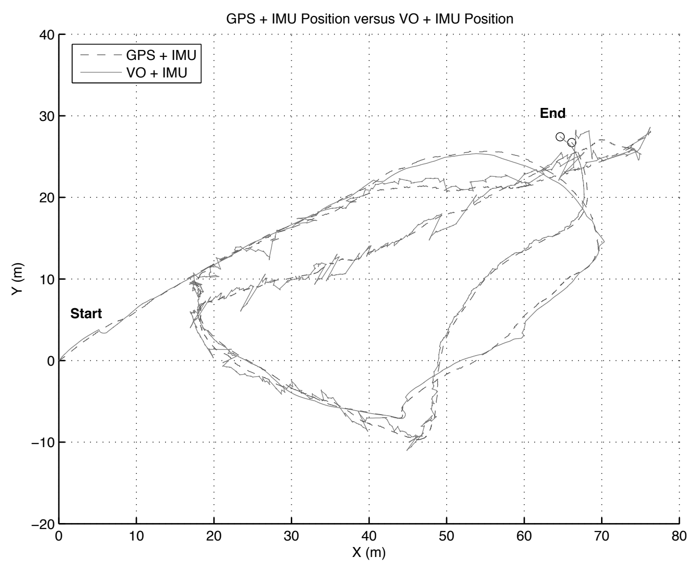


Fig. 4. Filtered GPS+IMU position estimate versus filtered VO+IMU position estimate. A single end label is shown, as the difference between the final GPS+IMU position and the final VO+IMU position is only 1.6 meters (after a 405.5 meter flight).

Table 1. Comparison of position estimates produced by GPS+IMU, VO alone and VO+IMU. Total flight distance, as measured by GPS, was 405.5 meters in *x, y*.

Sensors Used	Final Position (x, y) [m]	Final Position Error [m]	Average Position Error [m]	Maximum Position Error [m]
GPS+IMU	(66.13, 26.72)	–	–	–
VO alone	(53.39, 34.12)	15.14	4.66	15.14
VO+IMU	(64.63, 27.42)	1.65	1.27	5.58

as measured by the GPS receiver. The helicopter had a maximum forward velocity of 6.0 meters per second and a maximum yaw rate of 68 degrees per second. Vehicle altitude varied from approximately three meters to nine meters above ground level. Our shorter experiments followed similar flight profiles and produced comparable results (in terms of accuracy) to those described below.

Over the course of the longest flight a total of 3822 stereo image pairs were acquired. Figure 1b shows an image from the left stereo camera taken at $t = 74$ seconds, when the helicopter was five meters above the terrain surface. The KLT tracker was able to track an average of 163.6 features (out of 200 selected features) across sequential left image frames. After stereo triangulation and outlier removal, between 27 and 102 landmarks were used to compute the incremental visual motion estimates.

7 Discussion

We evaluated the accuracy of the system by comparing the output from the Kalman filter using GPS and IMU measurements with the integrated output from VO alone, and with the output from the Kalman filter using both VO and IMU measurements. This comparison is summarized in Table 1. In the text below, we use the terms 'GPS+IMU' to denote estimates produced using filtered GPS and IMU data, 'VO' to denote estimates produced using visual odometry alone, and 'VO+IMU' to denote estimates produced using filtered visual odometry and IMU data. Since we expect the GPS position estimate to be the most accurate over longer time intervals, we used the GPS+IMU flight path as our reference trajectory. Before plotting the ground tracks in Figures 3 and 4, we performed a least squares alignment of the first 30 frames (one second) of VO and VO+IMU data with the filtered GPS+IMU output.

Plots of the three-axis VO and IMU angular rate measurements are shown in the left column of Figure 2. There is strong agreement between the IMU pitch and yaw rates and the VO pitch and yaw rates, confirming that both sensors can correctly measure changes in vehicle orientation[1]. The measured

[1] The roll gyro on our IMU is particularly noisy, and so we did not compare the roll rates directly.

GPS+IMU and VO linear velocities are plotted in the right column of Figure 2. In this case, the x-axis and z-axis velocities measured by the GPS+IMU coincide approximately with the velocities measured by VO. However, there are portions of y-axis velocity curves that do not correlate well. A preliminary analysis indicates that the discrepancies may be due to poor VO output; over the regions in which the GPS+IMU and VO values differ, the number of landmarks used to compute the visual motion estimate dropped below the average for the flight. We are currently investigating this issue.

The final position estimate computed using VO has an error of 3.7% relative to the GPS+IMU position. This is comparable to previous visual odometry results for ground robots [3], although ground robots usually move more slowly than the helicopter. The final position estimate computed using filtered VO+IMU data differs by only 0.4% from the GPS+IMU position – this difference is almost an order of magnitude less, as a percentage. More importantly, although the VO+IMU position estimate is noisy, it tracks the GPS+IMU estimate over the entire flight. The deviation in the VO position increases with time.

An important aspect of the VO algorithm is that the quality of the output depends upon accurate stereo triangulation. Our imagery was acquired at an average height of approximately 6 meters above the terrain surface, producing consistent triangulated landmark positions. However, we have found that above a height of 12 to 15 meters, it becomes very difficult to select a large set of inlying landmark points, although the majority of the image features are tracked reliably. The reason is that, as vehicle altitude increases and the left-right stereo disparity decreases, small errors in the disparity estimates lead to large changes in the triangulated landmark depths relative to the stereo bench. This, in turn, means that corresponding landmarks points often cannot be aligned along the depth axis without significantly increasing the error tolerance in the RANSAC procedure, which degrades the quality of the overall motion estimate.

8 Conclusions and Future Work

This paper presented a visual and inertial system that allows an UAV to navigate over significant distances without GPS or other wide-area positioning information. Fusing inertial measurements with visual odometry was shown to reduce positioning error by almost an order of magnitude compared to using visual odometry alone. We expect that with further analysis of the visual odometry output and tuning of the Kalman filter, we should be able to improve the system's overall performance for longer trajectories. Additionally, we are working to implement a version of the software which operates in real time on-board the helicopter.

Acknowledgments

This work was funded in part by the National Science Foundation under grants CNS-0325875, CNS-0540420, CNS-0520305 and CCF-0120778 and by a gift from the Okawa Foundation. Jonathan Kelly was supported by an NSERC postgraduate scholarship from the Government of Canada. We would like to thank our safety pilot Alan Butler for his assistance with the AVATAR flight trials.

References

1. P. I. Corke, D. Strelow, and S. Singh, "Omnidirectional visual odometry for a planetary rover," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS'04)*, vol. 4, Sept. 2004, pp. 4007–4012.
2. A. Levin and R. Szeliski, "Visual odometry and map correlation," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR'04)*, vol. 1, 2004, pp. 611–618.
3. D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR'04)*, vol. 1, June 2004, pp. 652–659.
4. Y. Cheng, M. W. Maimone, and L. Matthies, "Visual odometry on the Mars exploration rovers," in *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, vol. 1, Big Island, USA, Oct. 2005, pp. 903–910.
5. S. I. Roumeliotis, A. E. Johnson, and J. F. Montgomery, "Augmenting inertial navigation with image-based motion estimation," in *Proc. IEEE Int'l Conf. Robotics and Automation (ICRA'02)*, vol. 4, Washington D.C., USA, May 2002, pp. 4326–4333.
6. O. Amidi, T. Kanade, and K. Fujita, "A visual odometer for autonomous helicopter flight," *J. Robotics and Autonomous Systems*, vol. 28, pp. 185–193, Aug. 1999.
7. J. Langelaan and S. Rock, "Passive GPS-free navigation for small UAVs," in *Proc. IEEE Aerospace Conf.*, Big Sky, USA, Mar. 2005, pp. 1–9.
8. S. Saripalli, J. F. Montgomery, and G. S. Sukhatme, "Visually-guided landing of an unmanned aerial vehicle," *IEEE Trans. Robotics and Automation*, vol. 19, no. 3, pp. 371–381, June 2003.
9. L. Matthies and S. Shafer, "Error modeling in stereo navigation," *IEEE J. Robotics and Automation*, vol. RA-3, no. 3, pp. 239–250, June 1987.
10. L. Matthies, "Dyamic stereo vision," Ph.D. dissertation, Carnegie Mellon University, Oct. 1989.
11. J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, USA, June 1994, pp. 593–600.
12. M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Comm. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.
13. S. Saripalli, J. M. Roberts, P. I. Corke, G. Buskey, and G. S. Sukhatme, "A tale of two helicopters," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS'03)*, vol. 1, Las Vegas, USA, Oct. 2003, pp. 805–810.