AN EXPERIMENTAL STUDY OF AERIAL STEREO VISUAL ODOMETRY

Jonathan Kelly and Gaurav S. Sukhatme

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

Abstract: Unmanned aerial vehicles normally rely on GPS to provide pose information for navigation. In this work, we examine *stereo visual odometry* (SVO) as an alternative pose estimation method for situations in which GPS in unavailable. SVO is an incremental procedure that determines ego-motion by identifying and tracking visual landmarks in the environment, using cameras mounted on-board the vehicle. We present experiments demonstrating how SVO performance varies with camera pointing angle, for a robotic helicopter platform. Our results show that an oblique camera pointing angle produces better motion estimates than a nadir view angle, and that reliable navigation over distances of more than 200 meters is possible using visual information alone.

Keywords: Gaussian distributions, mobile robots, position estimation, stereo vision, unmanned aerial vehicles.

1. INTRODUCTION

Accurate pose estimation is critical for many mobile robotics tasks, such as navigation and mapping. The most straightforward pose estimation technique, dead reckoning, simply integrates robot velocity measurements. A drawback of this approach is that the resulting estimates are subject to sensor-dependent bias and drift, and accumulate error over time. For example, when wheel odometry is used, wheel slip will often cause a robot's pose estimate to rapidly diverge from the true pose. At the opposite extreme, global localization methods, including GPS, can provides absolute pose information with bounded error in some limited spatial volume. However, global localization depends on the existence of, and access to signals from, calibrated external beacons (e.g. GPS satellites).

Visual odometry (VO), or visual ego-motion estimation, describes the process of incrementally estimating changes in robot pose by identifying and tracking visual landmarks in the environment. Although VO pose estimates are produced by integration and accumulate unbounded error, previous results for ground robots have shown that positioning accuracy to within a few percent of measured ground truth is possible, over

Fig. 1. The AVATAR, USC's robotic helicopter.

distances of several hundred meters. More importantly, this type of exteroceptive sensing can enable a robot to navigate in unknown environments where global localization is impossible, for example in terrestrial urban canyons (where GPS signals are blocked), or on remote planetary surfaces. Visual ego-motion estimation can also be combined with other navigation algorithms, including vision-based 3D obstacle avoidance (Hrabar *et al.*, 2005), without the need for any additional sensors.

To date, the most prominent use of visual odometry has been on the NASA Mars Exploration Rover (MER) robots (Cheng *et al.*, 2005). The rovers use stereo VO to correct pose estimates produced from wheel odometry when driving over high-slip terrain. Other related work includes a variety of visual odometry systems for ground navigation (Nistér *et al.*, 2004; Corke *et al.*, 2004), autonomous helicopter control (Amidi *et al.*, 1999), and low-altitude terrain mapping (Jung and Lacroix, 2003).

While visual odometry has been successfully employed in several ground and aerial robot applications, relatively little work has been done to characterize its performance for aerial platforms over longer flight distances. In this paper, we study how the performance of *stereo visual odometry* varies with camera pointing angle, for a robotic helicopter platform. Unlike ground robots, helicopters are able to rapidly change all of their 6-DOF pose parameters simultaneously. Also unlike ground robots, the camera pointing angle on a helicopter is not restricted by proximity to the terrain surface – camera angles from horizontal (zero degrees down) to nadir view (90 degrees down) are possible. Our work complements the study in (Olson *et al.*, 2003) of the optimal camera field of view for planetary rover visual odometry.

The helicopter trajectory, altitude, velocity and the camera pointing angle all influence the rate at which landmarks move across the left and right camera image planes. The amount of displacement between frames, in turn, determines how reliably the landmarks can be tracked. During forward flight and for a nadir view camera angle (looking directly down at the terrain surface), landmarks will move into and out of the image frame comparatively quickly, limiting the amount of time during which a particular landmark is in view. An oblique camera angle allows tracked landmarks to remain in the cameras' field of view for a longer period of time. Our approach, however, uses triangulated 3D landmark positions to estimate pose change – for oblique pointing angles, many of the landmarks will lie at a significant distance from the camera, increasing the uncertainty in their estimated positions. Our goal is to characterize these tradeoffs and to determine what camera angle provides the best VO performance.

Towards this end, we present VO results using stereo imagery acquired from several flights of a robotic helicopter, and compare these results with corresponding GPS and inertial measurement unit (IMU) data. The results demonstrate that an oblique pointing angle provides superior performance when compared to a nadir pointing angle, and that accurate navigation over distances of more than 200 meters is possible using visual information only.

2. STEREO VISUAL ODOMETRY

Our visual odometry algorithm is based on the approach presented in (Olson *et al.*, 2003) and originally described in (Matthies, 1989). We track point landmarks through sequential stereo image pairs, triangulating the 3D positions of the landmarks at each time step from their projections into the left and right camera images. We then find the rotation and translation parameters that best align the corresponding 3D point clouds. This process is repeated for each new stereo pair to produce an estimate of the change in vehicle pose over time. Each step is described below.

2.1 Landmark Tracking and Triangulation

Given an initial stereo pair, we begin by identifying image features that can be reliably tracked from frame to frame. We use the KLT algorithm (Shi and Tomasi, 1994) to select and track salient feature points in the left stereo image. For each selected left image point $\boldsymbol{p}_l = [x_l, y_l]$, we then search for a matching point $\boldsymbol{p}_r = [x_r, y_r]$ 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 search to a narrow region centered on the right epipolar line. Only matching points with a correlation score above a fixed threshold (0.75 in our implementation) are included in further processing. For points that are included, we estimate the sub-pixel disparity by fitting a biquadratic polynomial to the correlation values in the 3×3 region around the right integer correlation peak.

The 3D positions of the landmarks are found by stereo triangulation. For each pair of left-right image points, we back-project a ray $\boldsymbol{r}_l$ from the left camera optical center $\boldsymbol{C}_l$ through $\boldsymbol{p}_l$, and a ray $\boldsymbol{r}_r$ from the right camera optical center $\boldsymbol{C}_r$ through $\boldsymbol{p}_r$. In the ideal, error-free case, these rays would intersect in a single point, however noise and matching errors inevitably cause the rays to diverge. Instead, we find the midpoint

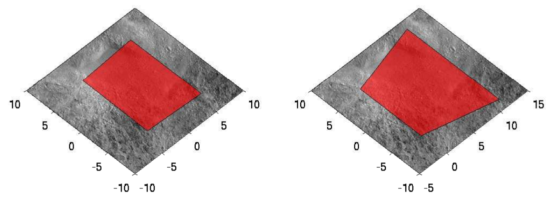

Fig. 2. Difference in visible terrain surface area for a single camera at an 85 degree downward pointing angle (left) and a 66.5 degree downward pointing angle (right). The camera has a horizontal field of view of 50 degrees and is positioned at a height of 10 meters.

of the shortest perpendicular segment connecting the rays. The endpoints of this segment are:

$$\boldsymbol{P}_l = \boldsymbol{C}_l + \lambda_l \boldsymbol{r}_l \quad (1)$$

$$\boldsymbol{P}_r = \boldsymbol{C}_r + \lambda_r \boldsymbol{r}_r \quad (2)$$

where the scalar constants λ_l and λ_r are the distances from $\boldsymbol{C}_l$ and $\boldsymbol{C}_r$ to the endpoints. Solving for λ_l and λ_r, we obtain:

$$\lambda_l = \frac{\boldsymbol{B}\cdot\boldsymbol{r}_l - (\boldsymbol{B}\cdot\boldsymbol{r}_r)(\boldsymbol{r}_l\cdot\boldsymbol{r}_r)}{1-(\boldsymbol{r}_l\cdot\boldsymbol{r}_r)^2} \quad (3)$$

$$\lambda_r = \frac{\boldsymbol{B}\cdot\boldsymbol{r}_r - (\boldsymbol{B}\cdot\boldsymbol{r}_l)(\boldsymbol{r}_l\cdot\boldsymbol{r}_r)}{(\boldsymbol{r}_l\cdot\boldsymbol{r}_r)^2-1} \quad (4)$$

Here, the vector $\boldsymbol{B} = \boldsymbol{C}_r - \boldsymbol{C}_l$ is the stereo baseline. The midpoint of the segment is selected as the estimated landmark position:

$$\boldsymbol{P} = \frac{\boldsymbol{P}_l + \boldsymbol{P}_r}{2} \quad (5)$$

To model errors in the position estimates, we consider the left and right image coordinates as normally-distributed random vectors with covariance matrices $\boldsymbol{\Sigma}_l$ and $\boldsymbol{\Sigma}_r$, respectively. The covariance matrices can be estimated directly from the curvature of the biquadratic polynomial used for the sub-pixel disparity computation. The resulting covariance matrix for $\boldsymbol{P}$ is:

$$\boldsymbol{\Sigma}_{\boldsymbol{P}} = \boldsymbol{J}\begin{bmatrix}\boldsymbol{\Sigma}_l & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_r\end{bmatrix}\boldsymbol{J}^T \quad (6)$$

where $\boldsymbol{J}$ is the 3×4 Jacobian of $\boldsymbol{P}$ with respect to the left and right image coordinates (Cheng *et al.*, 2005).

When a landmark moves out of view in either the left or right image plane, it is replaced by a newly-initialized landmark, in order to track an approximately constant number of landmarks over time. We attempt to track 200 landmarks from frame to frame.

2.2 Maximum-Likelihood Motion Estimation

The tracking and stereo triangulation steps above produce two sets of corresponding 3D landmark positions, before and after the helicopter has undergone an unknown rotation $\boldsymbol{R}$ and translation $\boldsymbol{T}$. The relationship between the landmark positions can be written as:

$$\boldsymbol{P}_a^i = \boldsymbol{R}\boldsymbol{P}_b^i + \boldsymbol{T} + \boldsymbol{e}_i \quad (7)$$

where the subscript b indicates the landmark position in the pre-move (*before*) coordinate frame, the subscript a indicates the position in the post-move (*after*) coordinate frame, and the superscript i indexes the specific landmark. To solve equation (7), which is non-linear due to the rotation, we linearize by taking the first-order Taylor expansion of $\boldsymbol{R}$ with respect to roll, pitch and yaw rotation angles $\boldsymbol{\Theta} = [\alpha, \beta, \gamma]$:

$$\boldsymbol{P}_a^i \approx \boldsymbol{R}_0\boldsymbol{P}_b^i + \boldsymbol{J}_i(\boldsymbol{\Theta}-\boldsymbol{\Theta}_0) + \boldsymbol{T} + \boldsymbol{e}_i \quad (8)$$

where $\boldsymbol{J}_i$ is the Jacobian for landmark i with respect to $\boldsymbol{\Theta}$, evaluated at an initial rotation $\boldsymbol{\Theta}_0$. The Gaussian noise vector $\boldsymbol{e}_i$ has covariance $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_a^i + \boldsymbol{R}_0\boldsymbol{\Sigma}_b^i\boldsymbol{R}_0^T$. The maximum-likelihood estimate for $\boldsymbol{\Theta}$ and $\boldsymbol{T}$ is found by minimizing the objective function:

$$\boldsymbol{M}(\boldsymbol{\Theta},\boldsymbol{T}) = \sum_i \boldsymbol{r}_i\boldsymbol{W}_i\boldsymbol{r}_i^T \quad (9)$$

$$\boldsymbol{r}_i = \boldsymbol{P}_a^i - \boldsymbol{R}_0\boldsymbol{P}_b^i - \boldsymbol{J}_i(\boldsymbol{\Theta}-\boldsymbol{\Theta}_0) - \boldsymbol{T} \quad (10)$$

where $\boldsymbol{W}_i$ is the inverse covariance matrix for $\boldsymbol{e}_i$. After differentiating (9) with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{T}$ and setting the result to zero, we obtain:

$$\left[\sum_{i=0}^{n}\boldsymbol{H}_i^T\boldsymbol{W}_i\boldsymbol{H}_i\right]\begin{bmatrix}\boldsymbol{\Theta}\\ \boldsymbol{T}\end{bmatrix} = \left[\sum_{i=0}^{n}\boldsymbol{H}_i^T\boldsymbol{W}_i\boldsymbol{L}_i\right] \quad (11)$$

with $\boldsymbol{H} = [\boldsymbol{J}_i\ \boldsymbol{I}]$ and $\boldsymbol{L}_i = \boldsymbol{P}_a^i - \boldsymbol{R}_0\boldsymbol{P}_b^i + \boldsymbol{J}_i\boldsymbol{\Theta}_0$. The final motion estimate is produced by iteratively computing (11), using the previous $\boldsymbol{\Theta}$ as $\boldsymbol{\Theta}_0$ for the following iteration. To remove outliers, we embed the computation of (11) in a RANSAC procedure (Fischler and Bolles, 1981).

3. HELICOPTER AND VISION SYSTEM

Our test platform, the AVATAR (Autonomous Vehicle Aerial Tracking and Reconnaissance), is a robotic helicopter built on a modified Bergen RC chassis. The robot carries an array of avionics equipment including a Novatel RT-2 GPS receiver, Inertial Science ISIS inertial measurement unit, PNI ECM 2-50 electronic compass, and a PC-104 Linux computer for autonomous control; a complete description of the control system is provided in (Saripalli *et al.*, 2003).

For our visual odometry experiments, 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

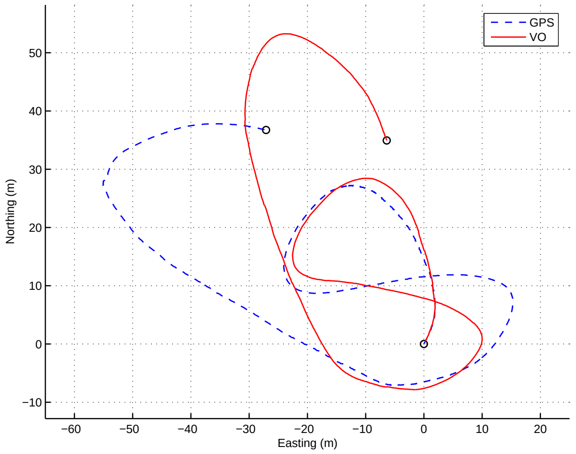

Fig. 3. VO ground track versus GPS ground track for flight 1. The cameras were pointed at an angle of 85.0 degrees down from horizontal. Total flight distance was 231.4 meters as measured by GPS.

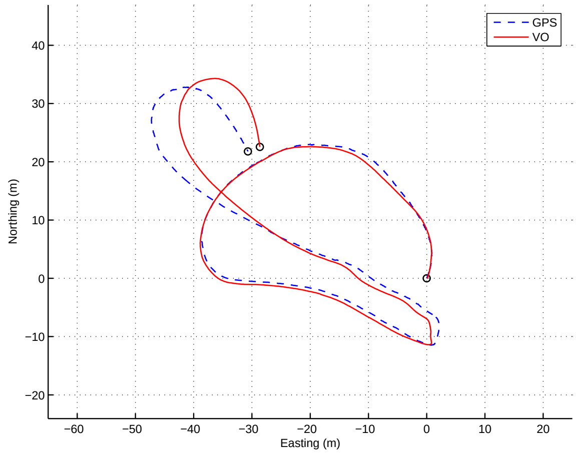

Fig. 4. VO ground track versus GPS ground track for flight 2. The cameras were pointed at an angle of 66.5 degrees down from horizontal. Total flight distance was 207.8 meters as measured by GPS.

pixels and a field of view of 50 degrees horizontally. The cameras are mounted on a stereo bench with a 50 centimeter baseline. During flight, we capture images synchronously from both cameras at 30 frames per second.

4. EXPERIMENTS

To evaluate the performance of the visual odometry algorithm, we performed several experiments with the AVATAR at our test site in Downey, California. The terrain surface at the test site consists primarily of dirt, gravel and lose rocks, with some sparse grass cover and small hills (up to 1.5 meters in height). During the experiments, the helicopter was flown manually while we logged data from the on-board GPS, IMU and the stereo cameras. The GPS and IMU data and the stereo images were all accurately time stamped to facilitate post-flight analysis.

For each individual flight trial, we chose a different pointing angle for the stereo cameras, in the range from 45 degrees down from horizontal to 85 degrees down. We then ran the visual odometry algorithm offline with the captured image data. Pose estimates were computed using visual information only, and the results were compared against measurements from the on-board GPS and IMU. The GPS and IMU data were processed in an extended Kalman filter to produce estimates of the helicopter's absolute position in a northing-easting global coordinate frame; we use these values (denoted as 'GPS + IMU' hereinafter) as ground truth in the discussion that follows. Importantly, we did not artificially constrain the visual motion estimate (to lie in a plane, for example) – the full 6-DOF solution for the pose change was computed at each time step.

In the remainder of the paper we present results from two of our longest flight trials, which illustrate the variation in VO performance with camera angle. For the first flight, the stereo cameras were mounted at a pointing angle of 85.0 degrees down from horizontal; for the second flight, the cameras were mounted at a pointing angle of 66.5 degrees down. These angles were measured with helicopter resting on a flat surface. In forward flight, the helicopter's nose is typically pitched downward by an angle of three to eight degrees, giving an approximately nadir view of the terrain surface for the first camera configuration. Ground tracks for the flights are shown in Figure 3 and Figure 4, respectively. Before comparing the VO and GPS + IMU position measurements, we aligned the first 60 VO poses with the corresponding GPS + IMU poses using least squares. During both flights, the vehicle altitude varied between approximately 4 meters and 12 meters above the terrain surface, as measured by stereo triangulation. Results from six other trials indicated that the 66.5 degree pointing angle is close to optimal.

During flight 1, the helicopter covered a ground track of 231.4 meters as measured by GPS, at an average velocity of 4.3 meters per second. The visual motion estimate for the flight was computed using 1500 stereo pairs; the KLT tracker was able to track an average of 158.0 landmarks per frame, with each landmark tracked for an average of 4.7 frames. After triangulation and outlier removal, between 14 and 102 landmarks were used to compute the incremental motion estimates. The final position error (VO minus GPS + IMU) was 20.8 meters, or 9%, although the maximum error was significantly larger.

The ground track for flight 2 was 207.8 meters long, flown at an average velocity of 3.8 meters per

Table 1. Accuracy of visual odometry measurements compared to filtered GPS + IMU measurements. Column 2 indicates the total number of frames used by VO. Columns 3 and 4 are the integrated VO and GPS + IMU flight distances. The remaining three columns give average position error over the entire flight, and maximum and final position errors for VO relative to the GPS + IMU positions.

Flight	Camera Angle	Stereo Frames	GPS Flight Distance	VO Flight Distance	Average Position Error	Maximum Position Error	Final Position Error
1	85.0 deg	1500	231.4 m	212.2 m	11.1 m	37.7 m	20.8 m **(9.0%)**
2	66.5 deg	1600	207.8 m	198.0 m	2.0 m	5.0 m	2.18 m **(1.0%)**

second. The visual motion estimate was computed using 1600 stereo pairs; the KLT tracker was able to track an average of 161.8 landmarks per frame, with each landmark tracked for an average of 5.2 frames. Between 26 and 119 landmarks were used to compute the incremental motion estimates. The final position error in this case was 2.18 meters, or 1.0%.

Figure 5 and Figure 6 compare the incremental (frame-to-frame) translational and rotational (yaw only) measurements for flight 1 and flight 2, respectively. To generate the incremental values for the GPS + IMU plots, we used the data points whose time stamps most closely matched the VO image time stamps. The difference between the corresponding VO and GPS + IMU time stamps was normally less than 3×10^{-4} seconds. Plots of the incremental rotation change were produced by computing an orientation vector for each frame; this vector is the normalized difference between the current position and the integrated position ten frames ahead. We used the angle between the current orientation vector and the previous orientation vector as a measure of the change in rotation. This process was repeated for the GPS + IMU measurements as well.

Although the GPS position is accurate over distances of more than two to three meters, the individual measurements have a large variance. To better indicate the correlation between the VO and GPS + IMU results, the GPS + IMU plots in Figure 5 and Figure 6 were smoothed using a five-measurement sliding window averaging filter. For the quantitative results presented in Section 5, we used the raw data values. The sharp spikes appearing in the GPS + IMU plots are due to the temporary loss of the telemetry signal when the helicopter was flying far from our ground station.

5. DISCUSSION

Table 1 presents results from both flights. The flight distances measured by GPS + IMU were longer than the distances measured by VO. This is to be expected, since noise in the GPS + IMU measurement effectively increases the length of the integrated GPS flight path.

We begin by noting that the plots of incremental VO and GPS + IMU translation and rotation estimates are well-correlated in both cases, indicating that visual odometry can accurately estimate the motion of the vehicle. The mean translation error (by frame) for flight 1 was -0.01 meters, with a standard deviation of 0.03 meters, while the mean rotation error was -0.4 degrees, with a standard deviation of 4.2 degrees. The mean

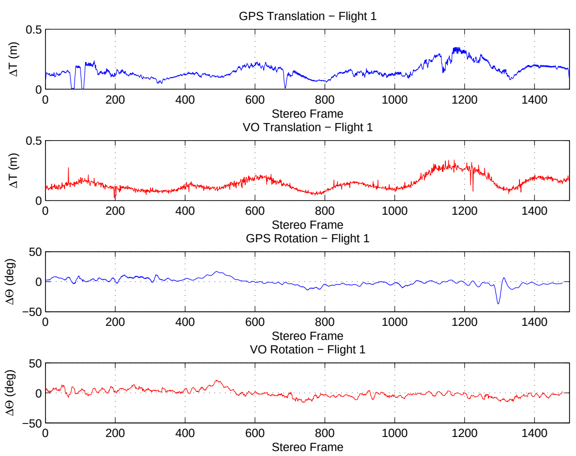

Fig. 5. Comparison of incremental translation and rotation estimates from VO and GPS + IMU for flight 1.

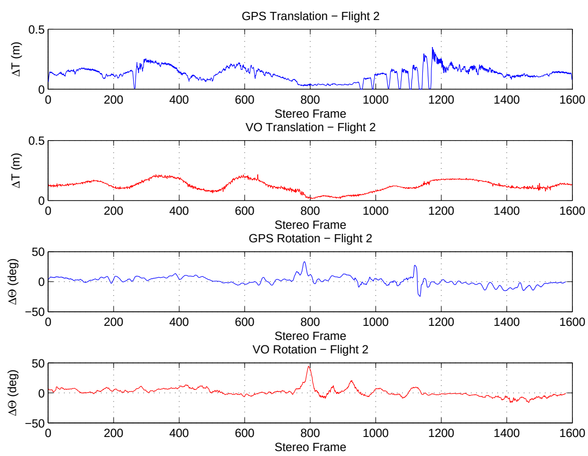

Fig. 6. Comparison of incremental translation and rotation estimates from VO and GPS + IMU for flight 2.

translation error for flight 2 was -0.006 meters, with a standard deviation of 0.04 meters, while the mean rotation error was -0.14 degrees with a standard deviation of 5.5 degrees. In all cases, the mean values are near zero, indicating that the VO estimates are most likely unbiased. However, for flight 1 (Figure 3) the VO result accumulates significant orientation error after only 400 frames – this, in turn, dramatically increases the average position error for the entire flight. In contrast, the VO estimates for flight 2 shows good agreement with the GPS + IMU data over the majority of the ground track.

An initial analysis of the flight 1 orientation error indicates that it is likely due to both the use of a smaller number of landmarks for the motion computations and to a higher landmark turnover rate. After outlier removal, fewer landmarks were used on average to compute the motion estimates for flight 1 compared to flight 2, and those landmarks were tracked through fewer frames. The selection of a large number of different landmark pairs at each time step de-correlates the frame to frame motion errors, which can result in increased drift (Matthies, 1989). We believe that this effect, and the more limited view of terrain surface for the nadir view cameras, leads to poorer performance. To confirm our hypothesis, we are currently performing additional simulation studies.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we presented experimental results for aerial stereo visual odometry using a robotic helicopter platform. Our results demonstrate that an oblique camera pointing angle can produce more accurate motion and pose estimates compared to a nadir pointing angle, although further experimental work is required to draw statistically meaningful conclusions. We also showed that in the oblique-angle case the visual estimate accumulates a maximum positioning error below three percent over a flight distance of more than 200 meters.

As future work, we plan to explore an alternative VO formulation in which pose change is estimated by minimizing landmark reprojection error in the image plane. This should enable VO to operate when the helicopter is flying at higher altitudes, where stereo depth estimates are less reliable. We are also implementing a modified version of the system 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, by a gift from the Okawa Foundation, and by a NSERC postgraduate scholarship from the Government of Canada. We would like to thank Srikanth Saripalli and our safety pilot Alan Butler for their advice and assistance.

REFERENCES

Amidi, Omead, Takeo Kanade and Keisuke Fujita (1999). A visual odometer for autonomous helicopter flight. *J. Robotics and Autonomous Systems* **28**, 185–193.

Cheng, Yang, Mark W. Maimone and Larry Matthies (2005). Visual odometry on the Mars exploration rovers. In: *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*. Vol. 1. Big Island, USA. pp. 903–910.

Corke, Peter Ian, Dennis Strelow and Sanjiv Singh (2004). Omnidirectional visual odometry for a planetary rover. In: *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS'04)*. Vol. 4. Sendai, Japan. pp. 4007–4012.

Fischler, Martin A. and Robert C. Bolles (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM* **24**(6), 381–395.

Hrabar, Stefan E., Peter I. Corke, Gaurav S. Sukhatme, Kane Usher and Jonathan M. Roberts (2005). Combined optic-flow and stereo-based navigation of urban canyons for a UAV. In: *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS'05)*. Edmonton, Canada. pp. 3309–3316.

Jung, Il-Kyun and Simon Lacroix (2003). High resolution terrain mapping using low altitude aerial stereo imagery. In: *Proc. IEEE Int'l Conf. Computer Vision (ICCV'03)*. Vol. 2. Nice, France. pp. 946–951.

Matthies, Larry (1989). Dyamic Stereo Vision. PhD thesis. Carnegie Mellon University.

Nistér, David, Oleg Naroditsky and James Bergen (2004). Visual odometry. In: *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR'04)*. Vol. 1. Washington D.C., USA. pp. 652–659.

Olson, Clark F., Larry Matthies, Marcel Schoppers
and Mark W. Maimone (2003). Rover navigation using stereo ego-motion. *J. Robotics and Autonomous Systems* **43**(4), 215–229.

Saripalli, Srikanth, James F. Montgomery and Gaurav S. Sukhatme (2003). Visually-guided landing of an unmanned aerial vehicle. *IEEE Trans. Robotics and Automation* **19**(3), 371–381.

Shi, Jianbo and Carlo Tomasi (1994). Good features to track. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'94)*. Seattle, USA. pp. 593–600.