CAMERA-AIDED HUMAN NAVIGATION: ADVANCES AND CHALLENGES

Clark F. Olson
University of Washington, Bothell
18115 Campus Way NE, Campus Box 358534
Bothell, WA 98011-8246
cfolson@uw.edu

Andreas O. Robinson
Primordial, Inc.
1021 Bandana Boulevard East, Suite 225
Saint Paul, MN 55108
andreas.robinson@primordial.com

ABSTRACT

We examine the use of visual navigation techniques for improving human navigation and localization in GPS-denied environments. While many advances have been made in navigation techniques using stereo and monocular motion estimation, significant hurdles exist for a practical system, including size, weight, power, and cost limitations. Humans often move and rotate faster and with more complex motions than robots, therefore requiring increased processing speed and robustness and the use of specialized algorithms. Systems for long-range navigation in previously unmapped environments must deal with error drift. Monocular systems face the additional issue of scale drift, since the relative scale must be estimated repeatedly. Furthermore, the orientation of the system relative to the frame-of-reference must be initialized accurately. We have evaluated these issues using a simulator and discuss possible solutions.

Index Terms— Navigation, camera, motion estimation

1. INTRODUCTION

We investigate the development of a camera-based system to facilitate human navigation and localization in the GPS-denied environments. Accurate location information is crucial for performing tracking, providing situational awareness, and for providing accurate navigational guidance. However, there are numerous contexts in which GPS information is not available, including indoors, in caves, and in the presence of large buildings or GPS jamming technology.

Recent advances in computer vision technology have opened up the possibility of using camera-based systems to accurately perform localization in GPS-denied areas. Such systems can track the motion of the surroundings in the field of view of the camera(s), offering the promise of a compact, low-cost, low-power, light-weight solution. In order to be a practical solution, the vision-based system must not only satisfy size, weight, power, and cost considerations, but it must be accurate and capable of running in real-time. It must also be capable of operating in the presence of realistic human motions, and be robust to changes in lighting conditions or the presence of dynamic moving objects in the field of view.

We examine vision-based autonomous navigation systems using human-mounted cameras in both monocular and stereo scenarios. A monocular version offers substantially lower size, weight, power, and cost. On the other hand, a stereo version provides higher accuracy. Our simulations indicate that monocular odometry can be highly effective when combined with algorithms for extracting even relatively noisy scale information.

In order to test the efficacy of such a system, we performed simulations comparing monocular and stereo odometry errors in various configurations. Simulations were also performed to assess the feasibility of addressing the difficult problem of scale drift in monocular odometry, as well as to determine the impact of incorporating an orientation sensor into an otherwise purely vision-based system. For the stereo simulation experiments, visual odometry code was applied to randomly generated simulated visual features over various test trajectories to assess localization errors.

For the monocular simulations, we combined the simulator code for randomly generating and sampling simulated visual features with routines from the OpenCV library [1]. Despite issues with scale-drift, the monocular approach remains quite promising because our results indicate that with periodic scale updates, the monocular errors can be brought on par with stereo errors, even when the scale estimates are noisy and infrequent.

2. STEREO NAVIGATION

In stereo range estimation, the error varies with the distance to the landmark [2]. The crossrange (lateral) error is linear in the distance, while the downrange (forward) error varies quadratically in the distance. While this implies that nearby features are the best for instantaneous localization, such features are also the most difficult to match owing to the difference in viewpoint. It is important that the uncertainty in the location estimate is used in any motion estimation procedure.

Previous work has resulted in a method that is capable of accurate rover navigation over long distances using incremental stereo ego-motion [3], [4]. The use of stereo information in this method was crucial in both outlier rejection and reducing random errors that occur due to feature localization and drift in each frame. A maximum-likelihood formulation of motion estimation was used that

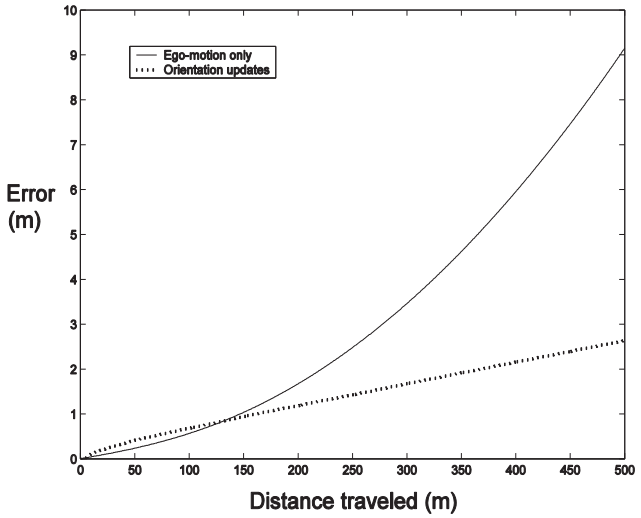

Fig. 1. Errors have superlinear growth unless orientation updates are provided.

models the error in the positions more accurately than a least-squares formulation and, thus, yields better results [5].

For long-range navigation, we must examine the rate of error growth as a function of distance travelled. The cumulative position error grows as the sum of terms corresponding to accumulating individual position errors and accumulating orientation errors. The first term grows slowly (on average with the square root of the distance travelled). However, the second term grows with the integral of the orientation error (which is also growing with the square root of the distance). Overall, the term corresponding to the orientation error dominates the error and yields an expected growth rate that is $O(d^{1.5})$ asymptotic growth, where d is the distance travelled [3]. This is illustrated in Fig. 1, which uses results from our simulator.

In order to eliminate the super-linear error growth, we have examined the use of an absolute orientation sensor to provide periodic updates to the orientation estimate. This information can be obtained from sensors such as compasses, gyros, or accelerometers. Orientation updates can greatly improve the long-range performance, reducing the accumulated error to a linear function of the distance traveled. However, magnetic interference may disrupt an orientation sensor. Closing the loop with previously encountered landmarks can also provide data that can be used to update the orientation.

The super-linear error growth is potentially a disadvantage with respect to techniques for simultaneous localization and mapping (SLAM) [6], [7]. However, SLAM techniques also suffer from such error unless they can close the loop by recognizing landmarks seen previously. Note that tracking a landmark through a few frames (as is also done with visual odometry) is not sufficient if the cameras are traveling on a path in which landmarks go out of view. They must return and recognize the landmarks to correct the accumulating error to the level when the re-imaged landmarks were first seen.

In contrast to robotic applications, human motions are complex and significant camera roll can be present in some cases. This would defeat a straightforward template matching approach for feature tracking, such as used on the Mars Exploration Rovers [4]. The use of a rotation invariant method for feature matching, such as SIFT [8], may be necessary in this case at the cost of additional computation time.

3. SIMULATIONS

In analyzing these techniques, we have built upon a previously constructed simulator that is able to predict how changes in the system and the algorithm are likely to affect the navigation performance [9]. For most experiments, purely vision-based errors were assessed, without augmenting sensors. Additional runs were also performed to assess the impact of adding an orientation sensor to the system.

The stereo simulation generates motion estimates by performing an iterative, non-linear optimization to improve on an initial closed-form solution [5]. For purposes of the simulation, 3d visual features are generated randomly along the selected trajectory, and a Gaussian model is used to add noise to the 2d projected locations, as well as the feature tracking estimates, both tracking over time and between the left and right cameras. These noisy 2d feature positions are then fed into the odometry code to generate motion estimates. Due to random variation, 100 runs over a given trajectory are averaged to produce the error estimates. There are, of course, also real-world issues that are not accounted for. For instance, in a human application the user may move/turn extremely rapidly causing motion blur and outpacing the ability of the processor to keep up with sufficient frames, and these issues are not currently accounted for by the simulation. The simulation also does not account for true outliers in the matching process. Typically, these are removed using a sampling process such as RANSAC [10].

Simulation suggests that without additional information the error in the stereo ego-motion technique is likely to rise beyond desired levels. With an orientation sensor, low error growths can be achieved (perhaps on the order of 0.5% growth with distance traveled, although this will vary according to system parameters). The simulations also indicate that:

- With respect to the camera field-of-view, there is a trade-off between inter-frame localization accuracy and the possibility of failure owing to insufficient features being tracked for large rotational motions.
- A larger baseline between the stereo cameras will generally yield improved accuracy. Of course, there is also a tradeoff with the size that is feasible for a human-mounted system.
- Frequent incremental updates are not necessarily optimal, since each update can introduce additional error.

4. MONOCULAR NAVIGATION

Accurate monocular navigation is more difficult than stereo navigation. In particular, a calibrated stereo rig provides scale information that is important to metric navigation. A monocular application requires the relative scale between the frames to be estimated [11], [12]. Drift in the scale estimate causes super-linear error growth in the same manner that drift in the orientation estimate does.

We implemented a monocular algorithm simulation using the 8-point algorithm [13] to solve for the fundamental matrix based on the motion of tracked 2d simulated features. The camera motion is then extracted using the intrinsic camera parameters. Next, the 3d locations of the feature points are triangulated based on the estimated camera motion. This sparse 3d model can then be used to maintain consistent scale between frames by enforcing constant distance between pairs of features over time. The major challenge in the monocular case is that the scale of the environment cannot in general be directly assessed (without assumptions) from a monocular camera view. This means that even if the initial scale is known, the scale estimate will gradually drift over time. This is a problem even for systems that provide localization relative to a SLAM map, since it is still crucial in such a system to maintain internal consistency in the scale estimate over time.

In addition to the baseline monocular odometry simulation runs, further tests were performed in which periodic approximate scale estimates were provided to system, with varying degrees of noise. This was done to estimate the required quality of scale updates which would be necessary in order maintain reasonable monocular error drift. One approach for performing these scale estimates in a pure vision system would involve extracting the ground plane (when visible) [14] and estimating the vertical distance from the camera to the features on the ground plane. Also, in theory, scale could be periodically measured and updated by using object recognition techniques to detect objects of roughly predictable size in the field of view, though we see this as a higher risk approach than ground plane extraction.

Several experiments were performed with the simulator in both straight courses and those containing rotation. In these experiments, the relative scale was estimated using a simple method that forces the median distance between each pair of estimated features to remain the same (considering only pairs that are found in both images). In cases where scale and/or orientation updates were provided to the system, they were provided every 0.5m. Figure 2 shows an example track for the case of circular motion. When no scale or orientation updates were provided to the system, the estimated track diverges quickly from the correct path. Orientation updates, by themselves, were not sufficient to yield accurate results. However, even noisy scale estimates resulted in a significant improvement. When both updates were used the resulting error was less than 1% over the 500 meter course. Fig. 2 uses orientation updates with σ = 1 degree and scale updates with σ = 5%. (Increasing to σ = 10% yields 9.8% more error and σ = 20% yields 39.2% more error for scale updates only.)

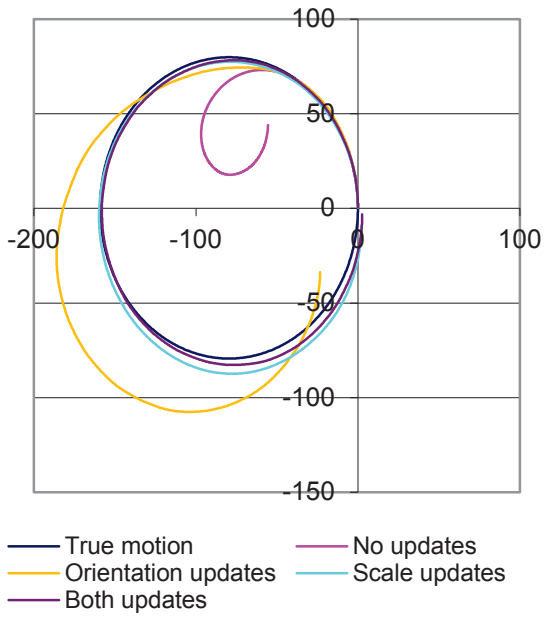

Fig. 2. Experiment comparing scale and orientation errors for monocular navigation. (Units are in meters.)

5. CHALLENGES

5.1. Scale updates

In our simulations, drift in the scale estimate over time is the largest source of error for monocular odometry. (Artificially eliminating this drift sharply reduces the error.) Periodic updates (even intermittently with considerable noise) yield significant improvements. However, current methods other than non-vision sensors are speculative.

One approach for estimating the true scale would be to recognize the ground plane. With the assumption that the camera stays at roughly the same height, this would provide an estimate of the true scale. While the user may crouch or otherwise change the height of the camera, such cases can

be detected using the motion estimate and removed from consideration since the updates may be intermittent.

Another potential approach is to use object recognition to detect object of roughly known size (for example, doorways, people, and cars). The known size of the object could be used to reset the drift in the scale.

5.2. Pose initialization

For geo-referenced localization, it is necessary to initialize the pose upon entry into a GPS-denied environment. The position can be determined directly from GPS, but the camera orientation must be estimated. This can be performed, for example, by comparing the trajectory estimation by visual navigation with a GPS-derived trajectory. The visual trajectory would be rotated to the best fit of the GPS trajectory. The rotation would be ill-defined in the unlikely case that the trajectory is a perfectly straight line. Ground plane extraction could be used to resolve this ambiguity. The technique of comparing the estimated trajectory to a GPS-derived trajectory can also be used to initialize the scale for monocular navigation.

5.3. Fast rotation

One failure mode of this methodology occurs when the camera rotates (or translates) so quickly that there is motion blur or that the system is unable to process frames at a rate that keeps a significant portion of the previous frame in view. In this case, it is possible that no features are matched and the motion estimation fails. However, recovery from this situation is possible. If the camera rotates back to a previously viewed location, frames can be dropped to allow motion estimation to continue. Turning corners will present a more difficult case, since the camera would not return to the previous orientation. In this case, an orientation sensor would provide useful redundancy. Alternatively, the system may spend extra time processing frames during fast rotation and catch up after the rotation is completed or process frames in the background as the user continues to travel.

6. CONCLUSIONS

We have examined techniques for camera-aided human navigation. A successful system will need to combine robust feature matching with outlier rejection. An orientation sensor or gyro may need to be included to reduce localization error over long-distance navigation and provide redundancy. In addition, a monocular system may need to periodically update the scale using absolute, rather than relative scale estimation. The system should incorporate intelligence to recover from failures, such as might be caused by fast rotation of the camera. While challenges remain for a robust system, these initial results suggest that potential solutions show promise for addressing the key issues.

7. ACKNOWLEDGMENTS

We gratefully acknowledge funding of the work by the U.S. Army Communications-Electronics Research, Development and Engineering Center (CERDEC).

8. REFERENCES

[1] [Online]. http://opencv.willowgarage.com/wiki/

[2] L. Matthies and P. Grandjean, "Stochastic performance modeling and evaluation of obstacle detection with imaging range sensors," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 6, pp. 783-792, Dec. 1994.

[3] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone, "Rover navigation using stereo ego-motion," *Robotics and Autonomous Systems*, vol. 43, no. 4, pp. 215-229, June 2003.

[4] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the Mars exploration rovers," *Journal of Field Robotics*, vol. 24, no. 3, pp. 169-186, 2007.

[5] L. Matthies and S. A. Shafer, "Error modeling in stereo navigation," *IEEE Transactions on Robotics and Automation*, vol. 3, no. 3, pp. 239-248, June 1987.

[6] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052-1067, June 2007.

[7] L. M. Paz, P. Piniés, J. D. Tardós, and J. Neira, "Large-scale 6DOF SLAM with stereo-in-hand," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 946-957, Oct. 2008.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[9] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone, "Stereo ego-motion improvements for robust rover navigation.," in *Proceedings of the International Conference on Robotics and Automation.*, 2001, pp. 1099-1104.

[10] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381-396, June 1981.

[11] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. 652-659.

[12] F. Fraundorfer, D. Scaramuzza, and M. Pollefeys, "A constricted bundle adjustment parameterization for relative scale estimation in visual odometry," in *Proc. IEEE Intl. Conf. on Robotics and Automation*, 2010, pp. 1899-1904.

[13] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133-135, Sep. 1981.

[14] S. Se and M. Brady, "Ground plane estimation: error analysis and applications," *Robotics and Autonomous Systems*, vol. 39, no. 2, pp. 59-71, May 2002.