

Vision Only Pose Estimation and Scene Reconstruction on Airborne Platforms

Michael Warren¹, David McKinnon², Toby Gifford², He Hu¹, Michael Shiel¹, Dawid Preller¹ & Ben Upcroft¹. Paper ID 3

I. INTRODUCTION

We aim to demonstrate unaided visual 3D pose estimation and map reconstruction using both monocular and stereo vision techniques. To date, our work has focused on collecting data from Unmanned Aerial Vehicles, which generates a number of significant issues specific to the application. Such issues include scene reconstruction degeneracy from planar data, poor structure initialisation for monocular schemes and difficult 3D reconstruction due to high feature covariance.

Most modern Visual Odometry (VO) and related SLAM systems make use of a number of sensors to inform pose and map generation, including laser range-finders, radar, inertial units and vision [1]. By fusing sensor inputs, the advantages and deficiencies of each sensor type can be handled in an efficient manner. However, many of these sensors are costly and each adds to the complexity of such robotic systems. With continual advances in the abilities, small size, passivity and low cost of visual sensors along with the dense, information rich data that they provide our research focuses on the use of unaided vision to generate pose estimates and maps from robotic platforms. We propose that highly accurate ($\sim 5\text{cm}$) dense 3D reconstructions of large scale environments can be obtained in addition to the localisation of the platform described in other work [2].

Using images taken from cameras, our algorithm simultaneously generates an initial visual odometry estimate and scene reconstruction from visible features, then passes this estimate to a bundle-adjustment routine to optimise the solution. From this optimised scene structure and the original images, we aim to create a detailed, textured reconstruction of the scene.

By applying such techniques to a unique airborne scenario, we hope to expose new robotic applications of SLAM techniques. The ability to obtain highly accurate 3D measurements of an environment at a low cost is critical in a number of agricultural and urban monitoring situations. We focus on cameras as such sensors are small, cheap and light-weight and can therefore be deployed in smaller aerial vehicles. This, coupled with the ability of small aerial vehicles to fly near to the ground in a controlled fashion, will assist in increasing the effective resolution of the reconstructed maps.

A. Related Work

A significant amount of work has been conducted in the realm of pose estimation and scene reconstruction in computer vision.

¹Department of Mechanical Engineering, University of Queensland, St Lucia, Queensland, Australia {m.warren1, h.hu2, ben.upcroft}@uq.edu.au, {michael.shiel, dawid.preller}@uqconnect.edu.au

²Australasian CRC for Interaction Design (ACID) CIRAC, Queensland University of Technology, Kelvin Grove, Queensland, Australia dave@acid.net.au

The capabilities to accurately derive the egomotion of single [3] or multiple [4, 5, 6] moving cameras on robotic platforms has been demonstrated using visual cues, particularly for ground-based vehicles. Much previous work in visual Smoothing and Mapping (SAM) and pose-only estimation has depended on the fusion of GPS and/or IMU data with monocular bundle-adjustment to generate an optimal solution [7, 8, 2]. Use of visual information for aiding UAV pose estimation was recently demonstrated by Bryson *et al.*, where 3D stereo features were integrated with information from an IMU and GPS receiver using non-linear least squares smoothing to compute vehicle pose. Clark *et al.* uses IMU and GPS information with visual bundle adjustment to compute dense maps of the environment. However, only qualitative results are shown, whereas we intend to compare our results to both INS and GPS for a quantitative comparison. Additionally, our work differs from Bryson and Clarke in that our pose estimate will be unaided by such inertial or GPS measurements.

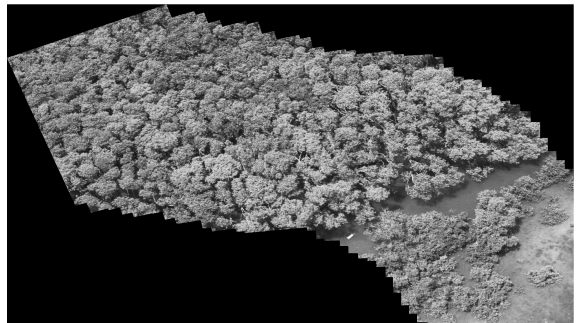


Fig. 1. A densely reconstructed 2D ground plane using only camera pose to inform map generation over a subsequence with approximately 10cm accuracy.

In more advanced SLAM applications, Konolige *et al.* [4] demonstrated frameSLAM, a visual SLAM method using key frames to reduce the size of the nonlinear system to solve [3]. Our work is most closely related to the first core steps taken in frameSLAM: 1) precise, real-time visual odometry for incremental pose estimation, and 2) nonlinear least squares estimation for local registration. In our work, loop closure is not yet considered. However, aerial platforms introduce a number of demanding scenarios unique to the nature of the platform. Unlike unaided visual pose estimation on ground vehicles, features of interest are near planar in structure, and we introduce 3-point calibrated resectioning techniques [9] to overcome these specific problems. Furthermore, we argue that stereo cameras are required to robustly overcome these issues. This is in contrast to conventional belief where the baseline on a UAV is often regarded as too small to account for range determination.

II. THEORY

The pose estimation can be broken into two main parts:

- Using visual data association to give a motion estimate between camera frames
- Optimising the motion estimation by minimising the re-projection error of camera and observed feature positions

In addition, we include the development of high density, high accuracy textured 3D maps as an outcome of the observed motion. The preliminary stage of our pose estimation routine consists of an accurate Visual Odometry (VO) system that tracks SIFT features [10] across multiples frames and uses these to simultaneously determine the camera pose and develop scene structure. The core of our pose estimation routine consists of a multi-camera (monocular or stereo) bundle adjustment routine that optimises the VO estimate over a window of several recent frames. Bundle adjustment is an application of the well known non-linear least squares solving routine for large estimation problems, and optimises both frame and feature positions to reach a refined estimate with minimal re-projection error [11, 12]. We currently use the bundle adjusted camera and scene-feature positions to generate high density depth maps from each camera frame of the same resolution as the original image. We will show re-projection of these depth maps into a global co-ordinate system, fitting of a polygonal model to the data and texturing of the map using the original camera images. A sample 2D reprojection is shown in Fig. 1, where camera pose is used to reproject the images onto a ground-plane fitted to the reconstructed scene.

A. Dealing with Airborne Data

The use of data from an airborne platform introduces a number of critical issues and difficulties in our algorithms that are not as significant in ground-vehicle based applications such as that presented by Konolige *et al.* [2]. There are two main difficulties

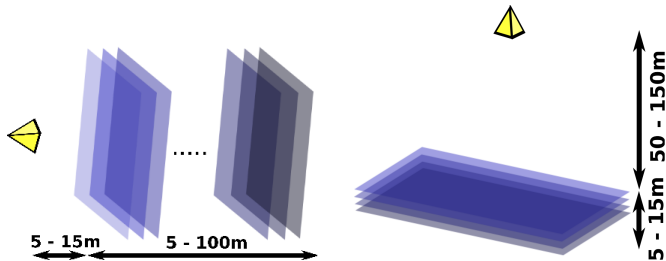


Fig. 2. The typical planes of interesting features in a) Ground based scenarios and b) Aircraft based scenarios

in dealing with such data: near-planar structure and distant scene features. In generating the initial scene structure, planar data introduces difficulties with degenerate solutions using monocular cameras. Generating the initial essential matrix between camera frames and selecting the correct structure solution is often ambiguous, and means that the wrong scene initialisation is chosen. We will show through experiment with a stereo rig that adding a second camera completely negates the structure-from-planes issue present in previous monocular schema. Although the baseline of a stereo pair on a UAV does not provide accurate range estimates, it constrains the geometry of the solution. Additionally, the existence of a small range of interest at a significant distance introduces a high variance in depth estimates, which can adversely affect the generation of an accurate reconstruction. Again, stereo-cameras assist this scenario by providing more depth information, reducing the covariance estimate and improving the solution.

III. CURRENT STATE OF RESEARCH

A. Airborne Platform

Experimental data for our analysis is gathered from an airborne robotic platform consisting of a 1/3 scale Piper Cub radio-controlled aircraft with sensors that include two digital colour cameras in a stereo configuration, 1.6Ghz computer running Ubuntu 9.10, XSens Inertial Navigation System and generic NMEA GPS. A number of datasets have been gathered using the robotic platform for use with the aforementioned algorithms. These include both monocular and stereo sets with GPS and INS ground truth, and at differing rural locations.

B. Experimental Results

We have generated consistent pose estimates for a number of datasets, including those generated from our airborne platform and ground-based data publicly available on the internet. Using data from our airborne platform at an altitude between 50 and 150m we have successfully managed to generate camera pose on monocular sequences up to 200 frames, covering distances of approximately 50m. This limited distance is due to the aforementioned difficulties with planar structure and high feature covariance. From the generated results, dense 2D maps with approximately 10cm accuracy of the ground surface have been constructed by projecting the original images to a ground plane extracted from the pose of the cameras (Figure 1). The aforementioned difficulties can be overcome by using stereo information, which will enable unique initialisation and pose estimates over several thousand frames. We will present results on stereo vision-only pose estimates over large trajectories, dense 3D reconstructions of the observed scenes and provide quantitative analysis against GPS and INS ground truth.

REFERENCES

- [1] S. Thrun and J. Leonard, *Simultaneous localization and mapping*. Springer, 2008, ch. 37, p. 871.
- [2] K. Konolige and M. Agrawal, "FrameSLAM: from Bundle Adjustment to Realtime Visual Mapping," *East*, pp. 1–11.
- [3] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 756777, 2004.
- [4] K. Konolige, M. Agrawal, and J. Sola, "Large scale visual odometry for rough terrain," in *Proc. International Symposium on Robotics Research*, 2007.
- [5] R. Koch, M. Pollefeys, and L. Van Gool, "Multi Viewpoint Stereo from Uncalibrated Video Sequences," pp. 55–, 1998.
- [6] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual modeling with a hand-held camera," *International Journal of Computer Vision*, vol. 59, no. 3, p. 207232, 2004.
- [7] M. Bryson, M. Johnson-Roberson, and S. Sukkarieh, "Airborne smoothing and mapping using vision and inertial sensors," in *2009 IEEE International Conference on Robotics and Automation*. Ieee, May 2009, p. 31433148.
- [8] R. Clark, M. Lin, and C. Taylor, "3D environment capture from monocular video and inertial data," *Three-dimensional image capture and applications VII: 16-17 January, 2006, San Jose, California, USA*, 2006.
- [9] O. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *INT. J. PATTERN RECOG. ARTIF. INTELL.*, 1988.
- [10] D. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, vol. 2, 1999, p. 11501157.
- [11] C. Engels, H. Stewénius, and D. Nistér, "Bundle adjustment rules," *Photogrammetric Computer Vision*, vol. 2, 2006.
- [12] B. Triggs, P. F. Mclauchlan, R. I. Hartley, and A. W. Fitzibbon, "Bundle adjustment - a modern synthesis," in *Vision Algorithms: Theory and Practice*, LNCS, vol. pages, pp. 298–375.