High Altitude Stereo Visual Odometry

Michael Warren
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Queensland, Australia
Email: michael.warren@qut.edu.au

Ben Upcroft
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Queensland, Australia
Email: ben.upcroft@qut.edu.au

***Abstract*—Stereo visual odometry has received little investigation in high altitude applications due to the generally poor performance of rigid stereo rigs at extremely small baseline-to-depth ratios. Without additional sensing, metric scale is considered lost and odometry is seen as effective only for monocular perspectives. This paper presents a novel modification to stereo based visual odometry that allows accurate, metric pose estimation from high altitudes, even in the presence of poor calibration and without additional sensor inputs. By relaxing the (typically fixed) stereo transform during bundle adjustment and reducing the dependence on the fixed geometry for triangulation, metrically scaled visual odometry can be obtained in situations where high altitude and structural deformation from vibration would cause traditional algorithms to fail. This is achieved through the use of a novel constrained bundle adjustment routine and accurately scaled pose initializer. We present visual odometry results demonstrating the technique on a short-baseline stereo pair inside a fixed-wing UAV flying at significant height (∼30-100m).**

I. Introduction

Stereo-based Visual Odometry (VO) has received significant attention in recent years as a robotic pose estimator [1, 2, 3]. Having been demonstrated over trajectories exceeding 50km with and without loop closure, stereo VO is a well studied problem. However, adequate performance in a number of applications is prevented by two specific limitations:

- A need for a relatively large baseline-to-depth ratio to achieve accurate triangulation
- A strict dependence on accurate calibration for epipolar and rectified image feature matching

With increasing distance of the scene from a stereo pair, accuracy in triangulation decreases, and the geometry can be considered to approach a monocular approximation as the baseline-to-depth ratio becomes smaller (Fig. 1). This has two effects: fast error build up due to poorly triangulated structure, and a weakly observable scale that is typically constrained by the stereo baseline. In addition, pose initialization is almost impossible, as the triangulation of scene from a single pair is inadequate and highly error prone. For these reasons, most stereo-based visual odometry solutions restrict themselves to ground vehicles and very low altitude multi-rotor applications [4, 5].

Moreover, structural deformations between cameras can cause serious issues for field robotics applications [6] and must be accounted for. Knocks, pressure changes and vibration can cause a stereo calibration to degenerate such that epipolar matching and scene point triangulation is seriously affected. To counteract this, applications have been typically restricted to short baseline pairs and/or low-impact environments. While engineering has a role to play in this situation, and deformation can be typically ignored on many indoor robots and ground vehicles, it must be considered in environments where structural changes from internal sources of the environment can cause misalignment. In airborne applications calibrations are most significantly affected by vibration, while pressure changes underwater can cause similar effects. In order to succeed in applications outside the realm of indoor and short-term demonstration, reducing this strict dependence is essential.

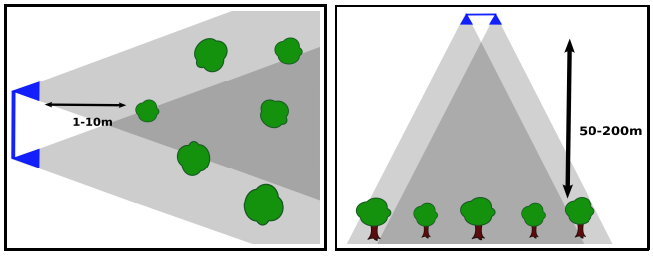

Fig. 1. The typical configuration of a stereo pair in ground based VO (left), and airborne VO (right), showing the dramatically reduced observability in the airborne case.

As a result of both deformation and poor triangulation issues, metric visual odometry for longer range stereo remains an open problem in robotics.

We propose a solution that relaxes the dependence on triangulation from geometrically fixed stereo pairs in addition to an accurate stereo calibration, while still maintaining metrically accurate visual odometry. This paper presents three major changes:

- A metrically scaled pose initializer for high altitude (30-100m) stereo
- A constrained bundle adjustment implementation for online calibration to counteract deformation
- A modified visual odometry algorithm for distant sensing while maintaining metricity

Together, these three methods provide the ability to perform metrically scaled, accurate visual odometry at high altitude without the need for additional sensors. We demonstrate the methodology in both a simulated experiment and on stereo visual data from a fixed-wing airborne vehicle flying at signif-

icant altitude, where vibration significantly deforms the stereo calibration and scale is weakly observable due to the extremely small baseline-to-depth ratio. We note here that a standard VO algorithm with fixed stereo geometry will fail rapidly given both scale observability and structural deformation, meaning a quantifiable comparison to the presented algorithm is impossible.

The ability to perform metrically scaled visual odometry at high altitude has a number of niche, but significant applications. Importantly, on a high altitude aircraft it serves as a redundant sensing mode when others may fail: flying through urban and natural canyons can mean GPS failure and a fall-back to dead reckoning that is typically handled by inertial sensing. Stereo VO provides a viable alternative that is not subject to the bias drifts inherent in inertial sensors, and can be considered in a number of applications to complement or potentially replace existing sensors. Further, high altitude stereo is a viable sensing mode where access to accurate global positioning is limited and costly, e.g. for lighter-than-air craft on planets such as Mars.

II. Related Work

As processing power has increased and camera cost reduced, visual odometry has seen applications on ground vehicles [3, 7, 8, 9], airborne vehicles [5, 10, 11, 12] and underwater [13, 14]. Many implementations have been described, with some dependent on an Extended Kalman or Information Filter (EKF or EIF) backend, a single camera integrated with an IMU or INS to maintain metricity [15], and the pure stereo case [7, 9]. Of note, most applications covering distances greater than a few tens of metres are ground based, applied over periods of hours and rarely exceed a minimum scene depth of 40m.

In the air, visual odometry has received some attention in recent years, both in monocular [16, 17] and stereo formats [18, 19]. While monocular methods have been demonstrated at altitudes above 80m [20] and distances exceeding 1.6km, stereo methods have been restricted to altitudes below 40m [21], and have typically not exceeded distances of more than 230m. Some monocular methods such as PTAM [22] and derivatives have been applied to hovering vehicles both indoors and outdoors [23] with success, but have often included inertial sensing to constrain scale and assist motion estimation. Weiss *et al.* have noted the deficiencies of stereo at small baseline-to-depth ratios, where the utility of stereo can be considered to reduce to an effectively monocular scenario, and hence includes an Inertial Measurement Unit to assist in scale and pose recovery. Clearly, a metrically scaled purely visual odometry algorithm has not been demonstrated over a large trajectory in situations of large scene depth, common in many UAV applications.

In its generic form, bundle adjustment [24, 25] optimizes over feature projections only to reduce error build-up and optimize both camera pose and scene structure. Some attempt, however, has been made in recent years to integrate additional *objectives* such as sensor readings or scale terms [26] into the optimization. By determining appropriate weights for each error metric and observation (e.g. compass bearing, inertial readings etc. in addition to projective observations), a unified framework can be made that finds the optimum of two or more separate minimization objectives. This is in contrast to a filter based solution that often discards the information-rich feature projections and assumes visual odometry as a black box, similar to an inertial sensor, integrating the output with other sensors to inform a probabilistically accurate pose. With increasing compute power and efficient sparsification of the bundle adjustment problem, online operation of a solution that incorporates feature projections with additional sensors and reduces dependence on a full-featured sensor fusion is becoming feasible.

In contrast to an objective based bundle adjustment, attempts have been made to apply constraints generated from other sensors [27, 28]. These *constraints* are subtly different to the aforementioned *objectives*, and are extensively used in other optimization applications [29]. Put simply; an objective based solution weights re-projection objectives with other sensors in a unified framework, a constraint based solution will apply bounds on the space in which estimated parameters can move. Lhuillier [27] has attempted to incorporate these constraints with VO by using a standard projection-only bundle adjustment step then adding GPS based pose constraints and re-optimizing the solution.

We use a different formulation to the above methods: instead of using an additional sensor such as GPS to constrain scale and pose, use is made of an additional camera forming a stereo pair to constrain scale. Results are presented over a larger trajectory and compared to ground truth, unlike [27], where no ground truth is presented. To differentiate the algorithm from other stereo based visual odometry the stereo transform is allowed freedom to move but is restricted by applying a soft log-barrier constraint [29] within predefined bounds to reduce dependence on accurate calibration. By reducing any dependence on rigid-stereo triangulation we additionally avoid issues of weak triangulation given by small baselines.

III. Methodology

We describe the methodology in three major sections:

- A modified stereo-aware bundle adjustment that utilizes constraints to maintain a metrically scaled stereo pair
- A metrically scaled pose initialization for short-baseline stereo
- An original visual odometry algorithm for very short-baseline stereo

A. Constrained Bundle Adjustment

1) Stereo Bundle Adjustment: Given m ($j \in \{1, \ldots, m\}$) scene points observed at n unique time points/locations ($i \in \{1, \ldots, n\}$) by a single camera, the traditional model used for the projection of point j in space ($\mathbf{X}_j \in \mathbb{P}^3$) into its location in image i ($\mathbf{x}_{i,j} \in \mathbb{P}^2$) is straightforward [30]:

$$\mathbf{x}_{i,j} \simeq \mathbf{K}[\mathbf{R}_i|\mathbf{t}_i]\mathbf{X}_j \tag{1}$$

where $\mathbf{K}$ encodes the internal properties of the camera, and $\mathbf{R}_i, \mathbf{t}_i$ denote the pose of the camera at time i. Here we extend the single camera case to multiple unique rigidly linked cameras ($k \in \{0, \ldots, l\}$) (up to structural deformation) and express additional cameras in terms of the base camera via the stereo transform $\mathbf{T}_0^k = [\mathbf{R}^k|\mathbf{t}^k]$. This can be expressed in a modified general projection equation:

$$\mathbf{x}_{i,j}^k \simeq \mathbf{K}^k \left[\mathbf{R}_i|\mathbf{t}_i\right] \mathbf{T}_0^k \mathbf{X}_j \tag{2}$$

In this paper we only consider the case of two cameras, where $\mathbf{T}_0^0 = [\mathbf{I}|\mathbf{0}]$ and $\mathbf{T}_0^1 = \left[\mathbf{R}^1|\mathbf{t}^1\right]$. Intrinsics $\mathbf{K}^k$ are considered to be unique to each physical camera. For traditional visual odometry with two cameras the transform $\mathbf{T}_0^1$ would typically remain fixed. However, in this paper we include the parameters that make up this transform as additional optimizable variables. This allows the algorithm to counteract any deformation caused by external factors. This leads to a total of $6n+6+3m$ parameters with which to optimize: 6 for each base camera, 6 for the stereo transform of the secondary camera and 3 for each scene point. We leave the details to a separate paper [31].

2) Optimization Constraints: The stereo transform is the effective scale constraint in most visual odometry algorithms. By allowing the parameters of the stereo transform $\mathbf{T}_0^1$ freedom to move, this scale constraint is potentially lost. Alternatively, a prior calibration provides a very strong constraint on the allowable motion range of the stereo transform. In the case of small baseline-to-depth ratios this constraint becomes important due to the high error in recovering camera poses. We make further additions to the bundle adjustment methodology described above by constraining the allowable motion of certain parameters of the transform, ensuring scale and the important geometry of the calibration is maintained.

Due to the rigidity of a well-engineered stereo pair, even under deformation, any movement between the cameras is physically restricted to at most a few degrees or millimetres. We encode this in the algorithm by implementing a *strictly feasible region* for some of the parameters that represent this deformation. From a known calibration, we implement the feasible region based on the initial value p of a parameter plus a bound $\pm q$ (See Fig. 2). In this paper, the initial parameter values p are chosen from an initial calibration performed on the ground, and the values q empirically evaluated. They could alternatively, for example, be estimated via an analysis of material expansion based on temperature or elasticity of the material under load. It is assumed that the magnitude of the feasible region defined by $(p-q) \leftrightarrow (p+q)$ is sufficient to account for the maximum possible deformation of the rig, without being large enough to lose the effectiveness of this strict initial value p as a strong prior.

In a general optimization problem, implementing a strictly feasible region of a parameter x can be easily expressed as an inequality constraint: $c_t(x) > 0$, $t \in I$ (the set of inequality constraints), $x \in \mathbf{x}$ (the set of parameters), where the *constraint equations* are of the form:

$$c_t(x) = x - b \tag{3}$$

where b is the barrier value. In the above bundle adjustment implementation, the set $\mathbf{x}$ includes the 6 parameters of the stereo transform, and hence yields 12 constraints, 2 per parameter representing an upper and lower bound.

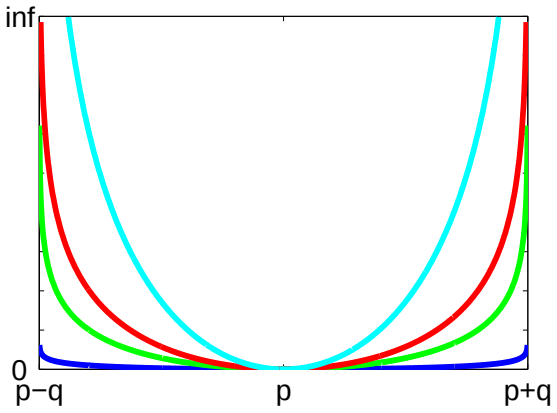

Fig. 2. The log barrier cost for varying values of μ_t, within the barriers $p-q$ and $p+q$.

Bundle adjustment attempts to minimize the sum of squares objective function by modifying camera poses, structure and the stereo transform:

$$\mathbf{P}^*, \mathbf{X}^*, \mathbf{T}^* := \underset{\hat{\mathbf{P}}, \hat{\mathbf{X}}, \hat{\mathbf{T}}}{\operatorname{argmin}} \sum_{i,j}^{k} \| \epsilon_{ij}^k \|^2 \tag{4}$$

with cost function:

$$f(\epsilon) = \sum_{i,j}^{k} \| \epsilon_{ij}^k \|^2 \tag{5}$$

where $\epsilon_{ij}^k = \mathbf{x}_{ij}^k - \hat{\mathbf{x}}_{ij}^k$ indicates the re-projection error between the observed feature and its current estimation for all cameras in a bundle adjustment optimization. By the introduction of additional terms, a logarithmic barrier function can be integrated as a soft constraint to constrain the optimization of specific parameters. i.e.

$$P(\epsilon, x) = f(\epsilon) - \sum_{t \in I} \mu_t \log c_t(x) \tag{6}$$

where μ_t is termed the *barrier parameter* and is used to tune the cost as the parameter approaches the barrier (See Fig. 2)[1].

The log barrier cost is also integrated into the bundle adjustment Jacobian and used to augment the relevant parameters $\mathbf{x}$. By splitting the parameters into *shared* parameters θ_S, *independent* parameters θ_I and scene points θ_P, the normal equations for an update step becomes:

$$\begin{bmatrix} \mathbf{S} & \mathbf{M} & \mathbf{N} \\ \mathbf{M}^\top & \mathbf{I} & \mathbf{O} \\ \mathbf{N}^\top & \mathbf{O}^\top & \mathbf{P} \end{bmatrix} \begin{bmatrix} \Delta\hat{\theta}_S \\ \Delta\hat{\theta}_I \\ \Delta\hat{\theta}_P \end{bmatrix} = \begin{bmatrix} \mathbf{e}_S \\ \mathbf{e}_I \\ \mathbf{e}_P \end{bmatrix} \tag{7}$$

[1]Alternative 'hard' constraints can be applied to the problem, such as the commonly termed *gradient-projection* method [29], but these methods result in greater implementation complexity.

where the matrices $\mathbf{S}$, $\mathbf{I}$ and $\mathbf{P}$ included in this expression are block diagonal defined according to the concatenation of the sub-matrices,

$$\begin{aligned}\mathbf{S}^k &= \textstyle\sum_{i,j} \mathbf{A}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \mathbf{A}_{ij}^k \\ \mathbf{I}_i &= \textstyle\sum_{j,k} \mathbf{B}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \mathbf{B}_{ij}^k \\ \mathbf{P}_j &= \textstyle\sum_{i,k} \mathbf{C}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \mathbf{C}_{ij}^k\end{aligned} \tag{8}$$

and

$$\begin{aligned}\mathbf{M}_i^k &= \textstyle\sum_j \mathbf{A}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \mathbf{B}_{ij}^k & \mathbf{N}_j^k &= \textstyle\sum_i \mathbf{A}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \mathbf{C}_{ij}^k \\ \mathbf{O}_{ij} &= \textstyle\sum_k \mathbf{B}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \mathbf{C}_{ij}^k & \mathbf{e}_{S^k} &= \textstyle\sum_{i,j} \mathbf{A}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \epsilon_{ij}^k \\ \mathbf{e}_{I_i} &= \textstyle\sum_{j,k} \mathbf{B}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \epsilon_{ij}^k & \mathbf{e}_{P_j} &= \textstyle\sum_{i,k} \mathbf{C}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \epsilon_{ij}^k\end{aligned} \tag{9}$$

The partial derivatives are defined as $\mathbf{A}_{ij}^k = \frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \hat{\theta}_S}$, $\mathbf{B}_{ij}^k = \frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \theta_I}$ and $\mathbf{C}_{ij}^k = \frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \theta_P}$ respectively. Addition of the logarithmic barrier Jacobian components results in the augmented shared parameter matrix $\mathbf{S^k}$:

$$\mathbf{S^k} = \sum_{i,j} \mathbf{A}_{ij}^{k\top} \Sigma_{\mathbf{x}_{ij}^k}^{-1} \mathbf{A}_{ij}^k - \sum_{t \in I} \frac{\mu_t}{\log c_t\left(x\right)} \tag{10}$$

where the first term incorporates the Jacobian with respect to the shared stereo parameters, and the second term incorporates the additional Jacobian generated from the barrier function. As x approaches b the cost term $\log c_t\left(x\right)$ grows in a logarithmic fashion and the projective influence on the parameter reduces. A step that takes x beyond b will also yield an infinite cost and hence not be updated in a bundle adjustment step.

B. Pose Initialization

In order to set-up the iterative VO algorithm, an initial estimate of pose and 3D scene is required. In more traditional scenarios scene is initially triangulated from the calibrated stereo pair, hence there is no need for a special initialization step. At large depths triangulation from the rigid stereo pair is inaccurate and structural deformation may render triangulation impossible. Hence, a scaled solution is needed for camera pose without initially computing structure from a geometric pair, more akin to monocular pose initialization. Initially, the essential matrix E_1 between the base camera at two adjacent time-steps is recovered, and relative pose (up to scale) extracted from this transform ($s_1 t_1$) (Fig. 3). To avoid degeneracies caused by near-planar structure, essential matrices pass an additional 'scene-spread' test as in [20]. This boot-strapping procedure ensures that accurate triangulation is achieved from a wide-baseline pair and is not dependent on the geometric stereo transform.

To recover metric scale, an essential matrix E_2 is also computed between the second camera at the initial time-step and the base camera at the second time-step to give a second scaled transform ($s_2 t_2$). Through vector addition a linear solution to the scale terms is calculated:

$$\begin{bmatrix} t_1 & -R^{k\top} t_2 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} t^k \end{bmatrix} \tag{11}$$

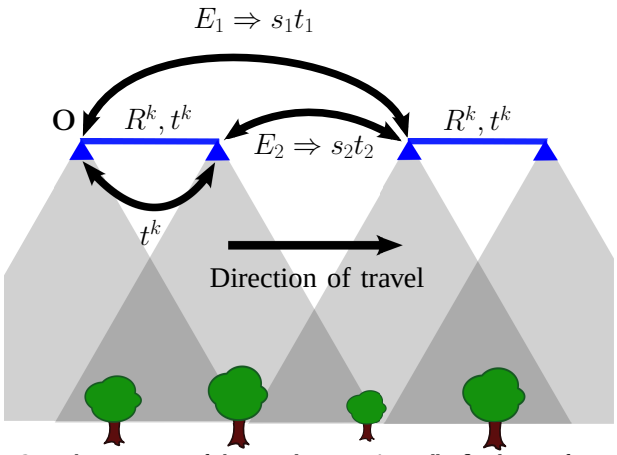

Fig. 3. The geometry of the initialization. A rigidly fixed pair of cameras at two time-steps. The transform t^k is already known from an approximate calculation

The relative poses are then scaled by the recovered terms to approximate metricity and then bundle adjusted with recoverd structure to optimize the initialization.

C. Short Baseline Visual Odometry

Following a pose initialization to set up the iterative pose estimation, visual odometry then follows 5 main repeating steps:

1) Image capture
2) Feature matching
3) Pose update
4) Structure triangulation
5) Constrained bundle adjustment

On a new set of images from a stereo pair, features are matched both between the pair and the previous base camera. From already triangulated structure and feature matches to the previous image, the new base camera pose $\mathbf{P}_i^0$ is found using calibrated 3-point pose estimation, performed inside a robust MLESAC estimator to ensure a reliable pose update. The secondary camera is initialized at the base camera and then moved via the initial stereo transform $\mathbf{T}_0^1$, derived from the initial calibration to avoid bias in the optimised solution.

New structure is then triangulated using only the base camera pairs to avoid dependence on fixed-stereo geometry, and then the constrained bundle adjustment algorithm is applied in a sliding window fashion to the last 12 frame-pairs and their associated structure. A Levenberg-Marquadt robust optimization routine is followed to ensure the estimation converges. At all times, the 6 parameters of the stereo transform are optimized subject to the afore-mentioned constraints.

IV. Experimental Results

To investigate the applicability of the algorithm, we present two separate experiments. First: a simulation that allows comparison of the recovered pose and stereo transform against a ground truth, and second: evaluation on field data gathered by a fixed-wing platform.

A. Simulated Experiment

The simulation consists of a stereo pair with 0.7m baseline flying at an altitude of 90m over a simulated ground environment. Scene features are projected into each camera with a 1 pixel variance σ. The stereo baseline is also given Gaussian noise on both the translational and rotational parameter to reflect vibration induced structural deformation, but this does not change throughout the experiment. Visual odometry is performed on the imagery generated from the pair for 400 frames, or approximately 2.4km of movement. To evaluate the effectiveness of the constrained optimization, two experiments are run:

- VO with constrained stereo optimization
- VO with unconstrained stereo optimization (see [31])

1) Results: Figures 4 and 5 show the simulated results. Figure 4 shows the variation of the stereo transform for the two separate experiments by subtracting the original parameter value from the estimate to show the difference. As expected, the stereo transform (green) is constrained within the pre-set bounds (red dashed lines) and hence scale is constrained to a known value. In contrast, the unconstrained optimization shows significant variation and scale is observed to drift over time.

Figure 5 highlights the average re-projection error at the conclusion of bundle adjustment on each new frame. While the unconstrained bundle adjustment shows a lower average re-projection error, it is clear that a smaller error does not necessarily translate to a better estimate of certain parameters (Fig. 4). In contrast, even with a higher average re-projection error, the constrained optimization shows a significantly improved estimation of the stereo transform.

Additionally, the constrained estimator shows a lower average number of bundle adjustment iterations (Fig. 5), as the logarithmic barrier will force a breakout earlier when the stereo parameters can no longer be optimized beyond the bounds.

B. Field Data Experiment

In a further demonstration of the algorithm, visual imagery was gathered from a fixed-wing airborne platform flown with a stereo pair of cameras. The stereo pair underwent significant deformation from vibration within the fuselage (See Fig. 12). The visual odometry algorithm is again run over the imagery, both with and without a set of stereo transform constraints.

1) Experimental Platform: The data-gathering platform is a large fixed-wing Unmanned Aerial Vehicle (UAV) with fuselage length of 2.3m (Fig. 6), remotely piloted within visual range from the ground. The aircraft includes an off-the-shelf computer system for logging both visual and inertial data, and a pair of IEEE1394B colour cameras, rigidly fixed to each other via an aluminium L-bar situated inside the fuselage of the aircraft. The cameras are placed facing down towards the terrain in the fuselage, as seen in Fig. 6. Each camera uses a 6mm lens with a field of view of approximately $42° \times 32°$. The cameras are calibrated before flight using a checker-board

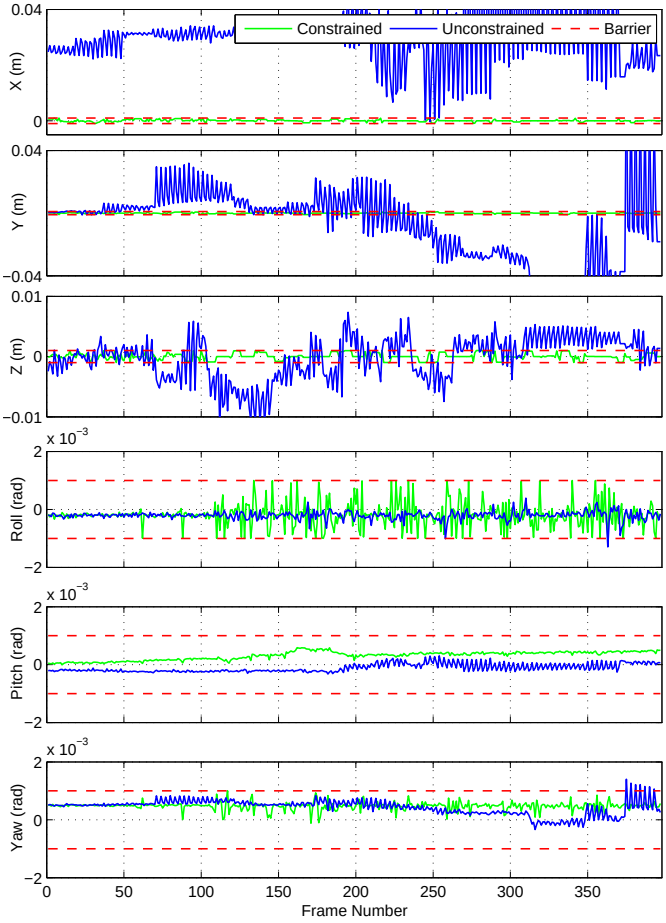

Fig. 4. The stereo transform values compared to the known calibration for the simulated experiment with constraints (green) and without constraints (blue), compared to the original calibration. The bounds of the constraints are shown by the red dotted lines.

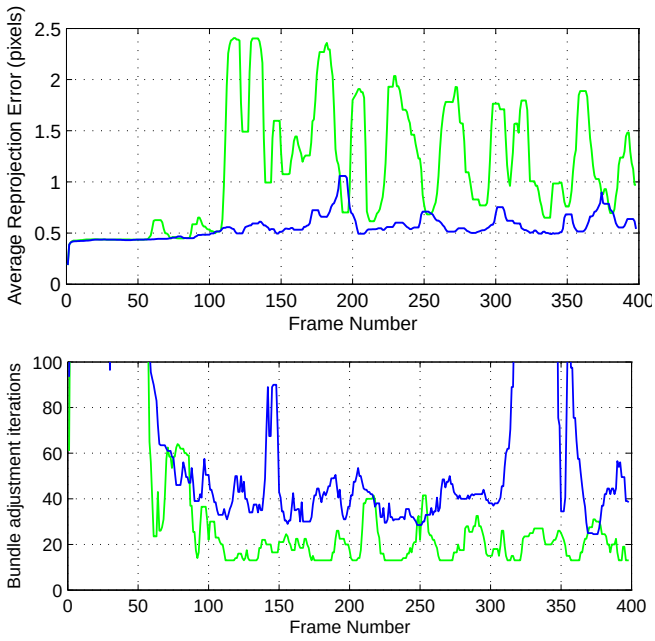

Fig. 5. Top: Average final re-projection error and Bottom: Bundle adjustment iterations per frame for the simulated experiment with constraints (green) and without constraints (blue).

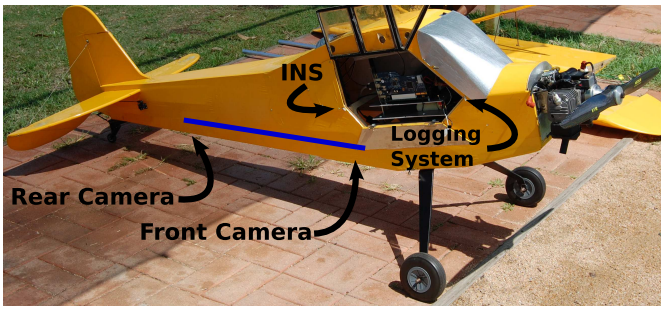

Fig. 6. The experimental platform showing component layout. Blue line indicates length and orientation of stereo baseline between on-board cameras.

pattern to achieve a standard intrinsic parameter calibration and approximated stereo transform between the cameras.

An XSens MTi-G INS/GPS system is used as the ground truth measurement system on the aircraft, with a manufacturer claimed positional accuracy of 2.5m Circular Error Probability (CEP). Size and weight restrictions prevent the use of more accurate DGPS systems, however, the MTi-G itself provides a reasonably accurate estimate of pose over broad scales. The MTi-G unit is rigidly attached to the onboard camera rig, while the GPS receiver is installed directly above the front camera.

2) Dataset: Data was collected over an approximately 5 minute flight, at an altitude of 20-100m and a speed of 20m/s. Bayer encoded colour images are logged at a resolution of 1280×960 pixels at 30Hz and later converted to color for processing. GPS, unfiltered IMU data and filtered INS pose were recorded at 120Hz from the XSens MTi-G to give ground truth position and orientation comparison. The area flown over consisted of rural farmland with relatively few trees, animals and buildings.

3) Results: Figures 8, 9 and 10 show the performance of the algorithms on 1450 stereo frames of the dataset (processed at a 3 image sub-sample from the original $30Hz$ data), covering a distance of 2.75km. Figure 10 shows the variation of the parameters in the stereo transform over the dataset. Without constraint, the stereo transform drifts significantly and shows repeated errors due to the poor observability. This is reflected in the pose comparison with ground truth (Fig. 8), where scale drift results in a poor trajectory in comparison to ground truth.

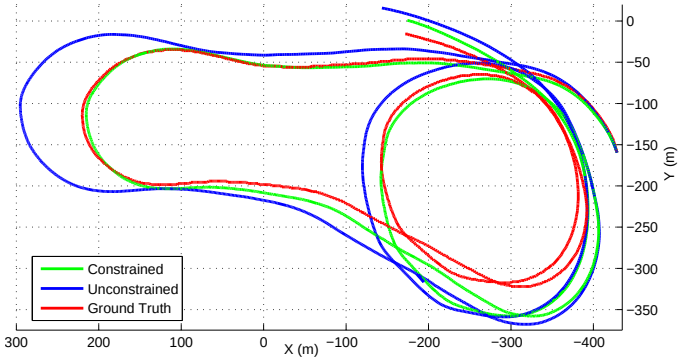

Fig. 8. Visual odometry results for the solution with stereo constraints (green) and without constraints (blue), compared to ground truth (red).

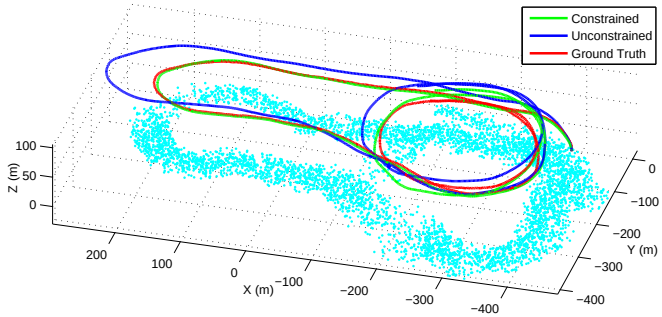

Fig. 9. Secondary view of visual odometry results for the solution with stereo constraints (green) and without constraints (blue), compared to ground truth (red). Observed structure shown in cyan.

In comparison, the constrained optimization shows significantly improved pose estimates over the trajectory, and this is reflected in the parameters of the stereo transform as shown in green in Figure 10, where the values are bounded by the constraints shown in red. A qualitative evaluation of the epipolar geometry for an example frame (Fig. 13) shows an improved alignment.

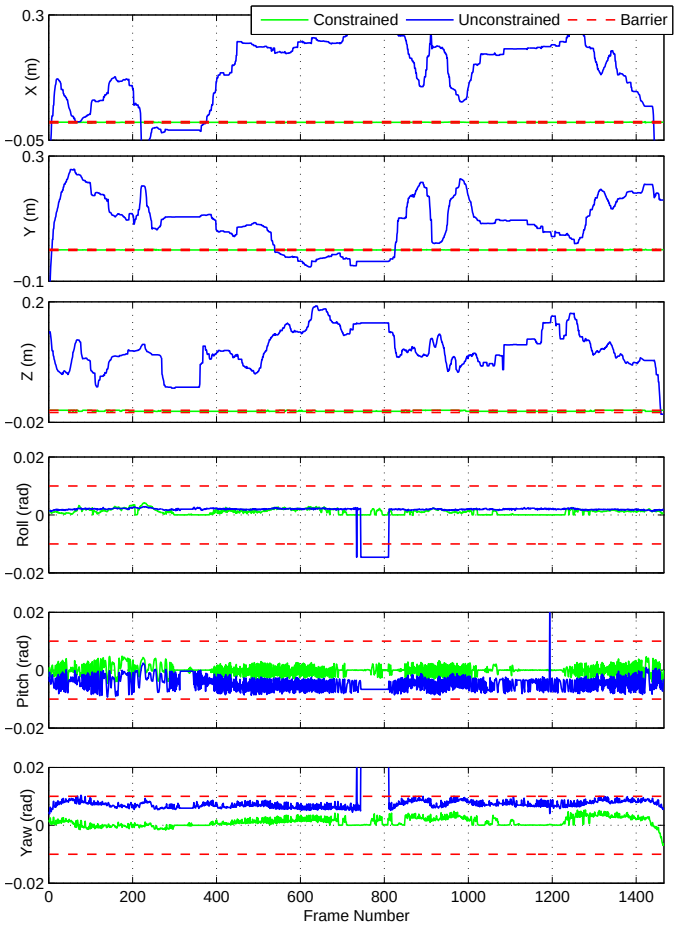

Fig. 10. The stereo transform values compared to the known calibration with constraints (green) and without constraints (blue), compared to the original calibration. The bounds of the constraints are shown by the red dotted lines.

Figure 11 shows a comparison of convergence between the constrained and unconstrained methods.

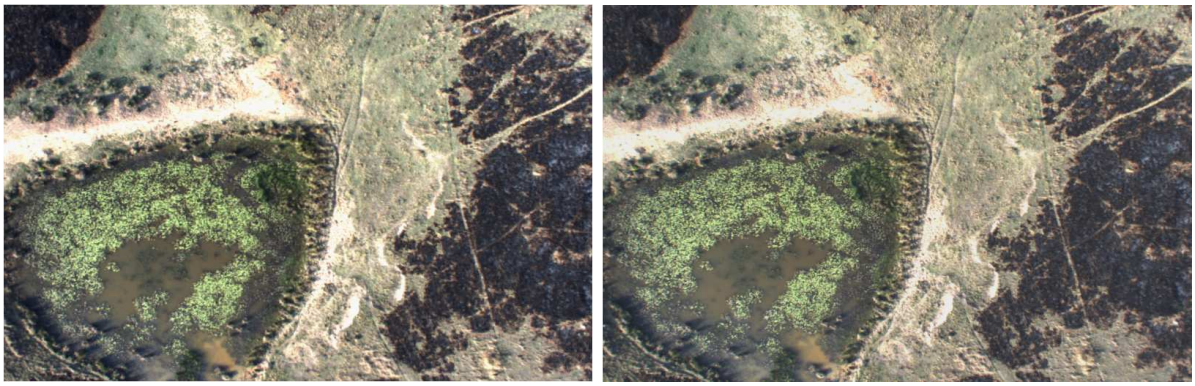

Fig. 7. An example image pair from the dataset, showing the small disparity between the stereo pair. Left: Front Camera, Right: Rear Camera.

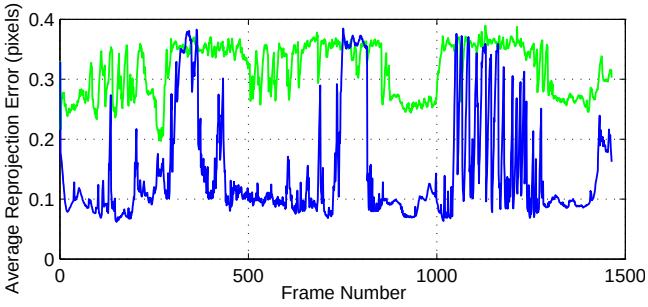

Fig. 11. Average final re-projection error for the visual odometry solution with stereo constraints (green) and without constraints (blue)

C. Discussion

Despite an overall smaller re-projection error, the unconstrained pose estimator shows poorer performance in generating an accurate pose because the important geometric information is lost. With the inclusion of constraints, convergence error is increased but shows better overall performance. This demonstrates that bundle adjustment need not rely on re-projection error alone as a metric of performance: the inclusion of constraints on the stereo transform in this case can yield better overall results.

Overall, these results demonstrate that stereo VO alone is inadequate to estimate pose with accurate scale in small baseline-to-depth applications. By applying constraints to the rigid geometry, scale can be retained even at the extremely small baseline-to-depth ratios exhibited in high altitude flight.

Fig. 12. The epipolar geometry of the original stereo calibration in flight. A selected pixel location in the front camera (left), and its corresponding epipolar line in the rear camera (right), showing the discrepancy caused by rig deformation.

It is important to note, however, that limitations apply for the technique: there is an upper limit to the altitude at which the algorithm can successfully work. With increased altitude scale becomes unobservable when the disparity of features tracked between a stereo pair drops below a single pixel, and will likely occur before this metric is reached. In this case, the altitude limit is likely to be beyond 200m, but would need to be experimentally evaluated as other effects are likely to affect the solution before this limit is reached.

Fig. 13. The epipolar geometry of the optimized stereo calibration in flight. A selected pixel location in the front camera (left), and its corresponding epipolar line in the rear camera (right), showing the correctly aligned epipolar geometry.

V. Conclusion

This paper has demonstrated the application of constrained optimization to the bundle adjustment problem to estimate the parameters of a stereo camera transform. By introducing a novel scaled pose estimator specific to the stereo problem and a modified visual odometry algorithm, the technique has been demonstrated on a difficult airborne dataset where traditional stereo algorithms will fail due to a small baseline-to-depth ratio and poor stereo calibration. Future work will examine the performance of the bundle adjustment algorithm by comparing the soft log-barrier constraint to harder constraints such as gradient-projection, which exhibits better convergence performance. Additionally, the algorithm will be evaluated in a number of different field-based datasets to show superior performance over long time periods.

References

[1] D. Nistér, O. Naroditsky, and J. Bergen, "Visual Odometry for Ground Vehicle Applications," *Journal of Field Robotics*, vol. 23, no. 1, pp. 3–20, Jan. 2006.

[2] G. Sibley, C. Mei, I. Reid, and P. Newman, "Vast-scale Outdoor Navigation Using Adaptive Relative Bundle Adjustment," *The International Journal of Robotics Research*, vol. 29, no. 8, pp. 958–980, May 2010.

[3] K. Konolige and M. Agrawal, "FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1066–1077, 2008.

[4] A. D. Wu and E. N. Johnson, "Autonomous Flight in GPS-Denied Environments Using Monocular Vision and Inertial Sensors," in *Infotech@Aerospace*, 2010, pp. 1–19.

[5] A. Cherian, J. Andersh, V. Morellas, N. Papanikolopoulos, and B. Mettler, "Autonomous Altitude Estimation of a UAV Using a Aingle Onboard Camera," *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3900–3905, Oct. 2009.

[6] J. N. Maki, "Mars Exploration Rover Engineering Cameras," *Journal of Geophysical Research*, vol. 108, no. E12, p. 8071, 2003.

[7] K. Konolige, M. Agrawal, and J. Sola, "Large Scale Visual Odometry for Rough Terrain," in *Proc. International Symposium on Robotics Research*, 2007.

[8] D. Scaramuzza, "Performance Evaluation of 1 point RANSAC Visual Odometry," *Journal of Field Robotics*, vol. 28, no. 5, pp. 792–811, 2011.

[9] M. Warren, D. McKinnon, H. He, and B. Upcroft, "Unaided Stereo Vision Based Pose Estimation," in *Australasian Conference on Robotics and Automation*. ARAA, 2010.

[10] D. Eynard, P. Vasseur, and V. Fr, "UAV Altitude Estimation by Mixed Stereoscopic Vision," in *International Conference on Intelligent Robots and Systems*, no. Iros, Teipei, 2010, pp. 4–9.

[11] J. Min, Y. Jeong, and I. S. Kweon, "Robust Visual Lock-On and Simultaneous Localization for an Unmanned Aerial Vehicle," *Electrical Engineering*, pp. 93–100, 2010.

[12] A. Xu and G. Dudek, "A Vision-Based Boundary Following Framework for Aerial Vehicles," in *International Conference on Intelligent Robots and Systems*, Montreal, 2010, pp. 81–86.

[13] O. Pizarro, "Large Scale Structure from Motion for Autonomous Underwater Vehicle Surveys," *Ocean Science*, 2004.

[14] C. Beall, B. Lawrence, V. Ila, and F. Dellaert, "3D Reconstruction of Underwater Structures," in *International Conference on Intelligent Robots and System*, vol. 0448111, Taipei, 2010, pp. 4418–4423.

[15] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-Time Monocular Visual Odometry for On-Road Vehicles with 1-Point RANSAC," *Seminar*, 2011.

[16] F. Caballero, L. Merino, J. Ferruz, and a. Ollero, "Vision-Based Odometry and SLAM for Medium and High Altitude Flying UAVs," *Journal of Intelligent and Robotic Systems*, vol. 54, no. 1-3, pp. 137–161, Jul. 2008.

[17] V. Guizilini and F. Ramos, "Visual Odometry Learning for Unmanned Aerial Vehicles," in *International Conference on Robotics and Automation*, Shanghai, 2011, pp. 6213–6220.

[18] J. Kelly and G. Sukhatme, "An Experimental Study of Aerial Stereo Visual Odometry," *Proc. Symp. Intelligent Autonomous ...*, pp. 1–6, 2007.

[19] I.-k. Jung, S. Lacroix, C. Roche, and T. C. France, "High Resolution Terrain Mapping Using Low Altitude Aerial Stereo Imagery," *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp. 946–951 vol.2, 2003.

[20] M. Warren, D. McKinnon, H. He, A. Glover, and M. Shiel, "Large Scale Monocular Vision-only Mapping from a Fixed-Wing sUAS," in *Field and Service Robotics*, 2012, pp. 1–14.

[21] S. Lacroix, "Digital Elevation Map Building From Low Altitude Stereo Imagery," *Robotics and Autonomous Systems*, vol. 41, no. 2-3, pp. 119–127, Nov. 2002.

[22] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 1–10, Nov. 2007.

[23] S. Weiss, D. Scaramuzza, and R. Siegwart, "Monocular SLAM Based Navigation for Autonomous Micro Helicopters in GPS Denied Environments," *Journal of Field Robotics*, vol. 28, no. 6, pp. 854–874, 2011.

[24] C. Engels, H. Stewénius, and D. Nistér, "Bundle Adjustment Rules," *Photogrammetric Computer Vision*, vol. 2, 2006.

[25] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle Adjustment - a Modern Synthesis," *Vision algorithms: theory and practice*, pp. 153–177, 2000.

[26] J. Michot and A. Bartoli, "Bi-objective Bundle Adjustment with Application to Multi-Sensor SLAM," *3DPVT'10*, 2010.

[27] M. Lhuillier, "Incremental Fusion of Structure-from-Motion and GPS using Constrained Bundle Adjustments." *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 12, pp. 2489–2495, Jul. 2012.

[28] M. Lhuillier, "Fusion of GPS and Structure-from-Motion using Constrained Bundle Adjustments," *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[29] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, Aug. 2000.

[30] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[31] M. Warren, D. McKinnon, and B. Upcroft, "Online Calibration of Stereo Rigs for Long-Term Autonomy," in *International Conference on Robotics and Automation (ICRA)*, Karlsruhe, 2013.