2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Robust Scale Initialization for Long-Range Stereo Visual Odometry

Michael Warren[1] and Ben Upcroft[1]

Abstract— **Achieving a robust, accurately scaled pose estimate in long-range stereo presents significant challenges. For large scene depths, triangulation from a single stereo pair is inadequate and noisy. Additionally, vibration and flexible rigs in airborne applications mean accurate calibrations are often compromised. This paper presents a technique for accurately initializing a long-range stereo VO algorithm at large scene depth, with accurate scale, without explicitly computing structure from rigidly fixed camera pairs. By performing a monocular pose estimate over a window of frames from a single camera, followed by adding the secondary camera frames in a modified bundle adjustment, an accurate, metrically scaled pose estimate can be found. To achieve this the scale of the stereo pair is included in the optimization as an additional parameter. Results are presented both on simulated and field gathered data from a fixed-wing UAV flying at significant altitude, where the epipolar geometry is inaccurate due to structural deformation and triangulation from a single pair is insufficient. Comparisons are made with more conventional VO techniques where the scale is not explicitly optimized, and demonstrated over repeated trials to indicate robustness.**

I. INTRODUCTION

Visual Odometry (VO) is a well established field, with a large sum of literature on the topic in recent years [1], [2], [3], [4]. However, stereo VO using rigidly fixed camera pairs has received little investigation as a pose estimator in long-range (or ultra-short baseline) applications, mostly from a lack of accuracy and the nonlinear effects due to extremely small disparities at this range. It is well known that with increasing distance of the scene from a camera pair, the quantization inherent from tracking features via single pixels means that depth error grows quadratically [5] with distance (Fig. 1). Most applications of visual odometry avoid the problems inherent in long distance sensing by either performing VO in environments where the camera system is in close proximity to scene structure [2], [6], [7] or increasing the baseline of the stereo pair [8].

In these more conventional scenarios, 3D scene structure is initially triangulated from the calibrated stereo pair, meaning no special initialization step is needed. Additionally, scale is implicitly defined by the stereo baseline. However, at large depths such triangulation is noisy and structural deformation from vibration or other factors may render accurate triangulation from the rigid pair impossible. This paper presents a technique for accurately initializing the pose of a stereo camera pair from 8-10 sequential frames, at accurate metric scale, for long-range applications where triangulation from a single pair is neither reliable nor accurate. In previous work [18] we use a triplet of cameras to estimate initial scale, but this initialization routine often suffers from degeneracies caused by linear motion and the poor observability of scale from a limited set of camera frames. To address these shortcomings, this paper presents a novel initialization routine where modification is made to bundle adjustment that includes the scale of the rigid-stereo transform as an optimizable parameter, in addition to optimizing the stereo transform within fixed bounds (an inequality constrained optimization) to account for structural deformations. By performing the initialization over a large window (∼8-10) of frames, poor scale observability is addressed and degeneracies are avoided.

With the technological advancement of multi-rotor and fixed-wing Unmanned Aerial Vehicles (UAVs), vision has a unique position to fill as a full 6DOF sensor in areas where the Global Positioning System (GPS) is unreliable due to urban and natural canyons, and maintains accurate estimates for far greater distances than inertial sensing alone [9], [10], [11]. In pure vision based pose estimation, stereo has been shown to estimate pose at altitudes of $40m$ [8], [12]. Additionally, it has applications in other schemes such as riverine environments [13] or even pose estimation on other planets such as Mars, where localization infrastructure is non-existent.

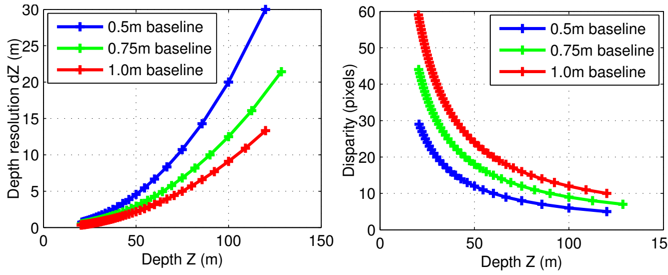

Fig. 1. Depth versus depth resolution and disparity for selected baselines of a stereo pair with 1024×768 image resolution, focal length of f =∼6mm and a $3.75\mu m$ pixel size.

While the obvious technique in long-range sensing is to increase the baseline-to-depth ratio, in many cases this is neither beneficial nor practical. With increasing baseline the stereo calibration becomes increasingly unreliable due to flex and vibration induced deformation, and the apparatus becomes unwieldly when placed on smaller vehicles. Structural deformation can be counteracted by engineering but this also means larger, heavier rigs, which are specifically unsuitable for flying vehicles where size and weight restrictions are paramount. In many long-range stereo applications, a short

[1] The authors are with the CyPhy Lab, School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia michael.warren, ben.upcroft at qut.edu.au

978-1-4673-6357-0/13/$31.00 ©2013 IEEE

baseline remains the only feasible implementation when seen from an engineering perspective.

In most stereo-based VO algorithms, scene structure is directly triangulated from physical stereo pairs and used to initialize a new pose update in an iterative fashion [14], dropping the structure and re-triangulating at every step. However, at long range the triangulation from a single pair is inadequate (See Fig. 1, reflective of the setup utilized in this paper) and noisy, meaning a poor pose update and ultimately poor reliability. By triangulating from a single camera moved spatially through time, a large pseudo-baseline can be generated more akin to monocular VO, and the integration of multiple observations from several images (rather than just two) ensures a high accuracy estimate of structure can be found. Here lies the dichotomy: long-range stereo necessitates observations from wide baselines (beyond the range of a single stereo pair) to maintain accuracy and reliability, but an accurately scaled initialization is dependent on the geometry of the stereo baseline.

Also inherent in long-range sensing is a range bias from short baseline triangulation that grows with increasing depth. At disparities on the order of 10 pixels or less, triangulation from a single pair follows a non-Gaussian curve with a long tail [15], that tends to overestimate scene depth and underestimate camera motion [16], adversely affecting scale. Such non-Gaussian effects must be considered in filtered applications [17], where individual observations are marginalized out quickly. However, these effects are less relevant in a bundle-adjusted context where a least-squares multi-view approach quickly approaches a Gaussian error, due to wide baseline observations over time.

In order to adequately initialize an accurately scaled pose estimate at long range, previous techniques have utilized other sensors [10] to approximate scale, but only in a monocular perspective. In contrast, we are interested in a vision-only initialization for rigid-stereo VO, without assuming known structure and ensuring robustness against poor initializations. Our previous work has attempted to initialize scale at long-range from only a triplet of images [18], using the stereo baseline to achieve a linear solution to the scale error, but this approach is unreliable due to degeneracies from linear trajectories and the lack of sufficient information to accurately estimate the true scale. Failure is frequent and means repeated attempts at initialization. We address these shortcomings by proposing a new initialization that uses a larger set of cameras over a fixed 'window' of frames.

A simple but naïve implementation of the proposed initialization technique is to perform a fully monocular pose estimate using only a single physical ("base") camera, then include the secondary camera in a batch bundle adjustment to recover scale. However, the introduction of a set of images from this second camera at an incorrect scale causes bundle adjustment to be poorly initialized, with the obvious consequence of unreliable and poor convergence performance [19]. The introduction of the second camera's images causes a high re-projection error that will force the base cameras and scene structure to move significantly with the scale error. With the introduction of a specific scale term into the bundle adjustment routine that encodes the scale of the entire scene via the stereo baseline, convergence can then proceed without significantly modifying camera positions or scene depth.

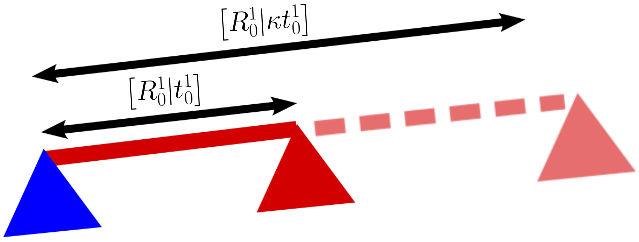

Fig. 2. Scaling of the stereo transform via the scale term κ

This paper presents a technique for accurately initializing the pose of a stereo camera pair over multiple frames, at accurate metric scale, for long-range applications where triangulation from a single pair is neither reliable or accurate, and the information from a small set of cameras is insufficient to estimate scale accurately at long range. In standard rigid-stereo VO scale is implicitly defined by the stereo baseline. In long-range applications, however, the stereo baseline is inadequate to accurately triangulate structure. Therefore, we initialize the stereo VO algorithm from a set of monocular poses, re-add the secondary cameras via the stereo baseline, then explicitly optimize for the scale to recover metricity. By performing an initialization in this way, issues of poor triangulation and pose updates are avoided. An accurate scale can be recovered in just the first few frames.

In addition, we note that the technique is applicable to the unreliable calibration often present in these situations. A high vibration environment, coupled with flimsy rigs to reduce weight, means that stereo calibrations are difficult to maintain and present an additional challenge to a robust initialization in more conventional algorithms. As allowed by the proposed technique, including the parameters of the stereo transform as optimizable variables means that vibration induced deformation can be alleviated while still maintaining a viable, scaled VO output. We demonstrate the performance of the proposed algorithm in the presence of poor epipolar geometry as a result of this deformation. Results are presented in the context of a fixed-wing UAV at significant altitude ($100 - 300ft$), where the baseline-to-depth ratio is small and relative disparity is small.

The rest of this paper is outlined as follows: Section II describes the novel initilialization, VO method and modified bundle adjustment methodology. Section III demonstrates results of the proposed algorithm on both simulated and field-gathered data, and finally the paper is concluded in Section IV.

II. METHODOLOGY

We describe the methodology in two sections:

- A modified stereo bundle adjustment that includes an explicit scale term κ
- A pose initializer for long-range stereo that ensures metric scale

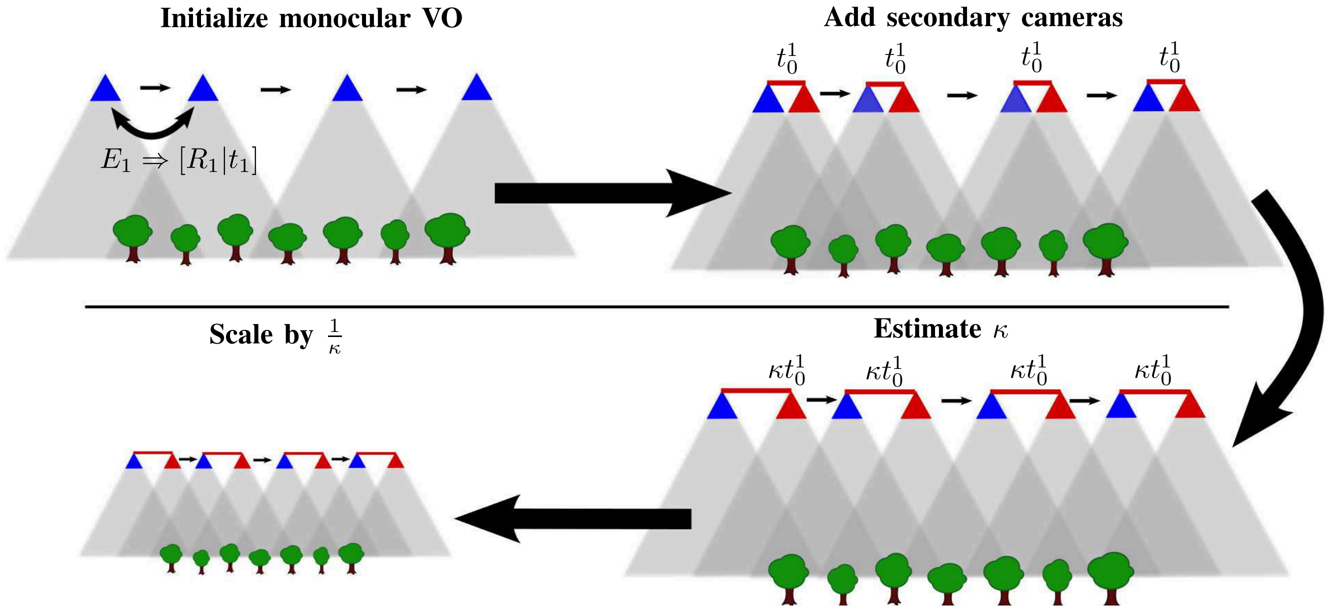

Fig. 3. The long-range stereo initialization routine. A monocular pose estimate on a set of base camera images is followed by adding the secondary camera and optimizing for the scale constant κ. This is followed by the application of $\frac{1}{\kappa}$ to recover metric scale, ready to perform a long-range stereo VO.

Following this, a modified stereo VO algorithm suited to long-range sensing is presented.

A. *Modified Stereo Bundle Adjustment*

Here we explicitly define the physical cameras as two unique rigidly linked cameras ($k \in \{0, 1\}$), with $k = 0$ as the base camera that determines the origin of the local co-ordinate system of the camera pair, and $k = 1$ as the secondary camera, that lies at some transform $\mathbf{T}^1_0$ from the origin or base camera.

Given m ($j \in \{1, \ldots, m\}$) 3D scene points $\mathbf{X}_j$ observed at n unique time points/locations ($i \in \{1, \ldots, n\}$), we define the general projection equation:

$$\mathbf{x}^k_{i,j} = \mathbf{K}^k \mathbf{M}_i \mathbf{T}^k_0 \mathbf{X}_j \qquad (1)$$

where $\mathbf{K}^k$ encodes the intrinsic properties of each physical camera, $\mathbf{M}_i = [\mathbf{R}_i|\mathbf{t}_i]$ denotes the extrinsics or pose of the base camera at time i and $\mathbf{T}^k_0 = [\mathbf{R}^k_0|\mathbf{t}^k_0]$ denotes the stereo transform between the base and other rigidly fixed cameras. In this paper we only consider the case of two cameras, where $\mathbf{T}^0_0 = [\mathbf{I}|\mathbf{0}]$ and $\mathbf{T}^1_0 = \left[\mathbf{R}^1_0|\mathbf{t}^1_0\right]$, but the algorithm can be easily extended to more than two. For standard visual odometry with two cameras, the transform $\mathbf{T}^1_0$ would typically remain fixed. However, we explicitly include the parameters that make up this transform as additional optimizable variables, and apply a boundary on the allowable space in order to maintain an accurate transform, the subject of a separate paper [18]. This allows the algorithm to alleviate small deformations caused by external factors such as vibration.

The basic theory of this modification is similar to that presented by Lhuillier *et al.* [20], where constraints are generated by other sensors in a more generalized bundle adjustment algorithm to constrain pose drift. However, in contrast to the use of GPS poses as boundary or inequality constraints on camera pose, our methodology applies inequality constraints on the rotational and translational components of the stereo transform. Deformations can be accounted for by including these parameters in a bundle adjusted solution, but the inclusion of strict boundaries ensures that the transform remains in a feasible parameter space as is limited by the physical baseline.

As a component of the key novelty in this paper, we introduce a scaling term κ to Eq. 1, that allows the translational component of the stereo transform $\mathbf{t}^k_0$ to scale in the vector defined by its units (Fig. 2):

$$\mathbf{x}^k_{i,j} = \mathbf{K}^k \left[\mathbf{R}_i|\mathbf{t}_i\right] \left[\mathbf{R}^k_0|\kappa\mathbf{t}^k_0\right] \mathbf{X}_j \qquad (2)$$

By optimizing this variable in addition to the scene points, base camera positions and stereo transform, any discrepancy in scale of the scene defined by adding the secondary cameras is handled efficiently without rendering the bundle adjustment problem as poorly initialized.

The addition of the scale term κ does not significantly affect computational performance of the bundle adjustment algorithm, but requires changes to the analytical derivatives of some optimization parameters. The analytical Jacobian for the optimized variables is shown in Table I, expressed in terms of a number of simpler components:

Assigning from (2), the normalized homogeneous pixel co-ordinates, $x_\beta, y_\beta, w_\beta$ can be expressed as:

$$\mathbf{x}_\beta = \begin{bmatrix} x_\beta \\ y_\beta \\ w_\beta \end{bmatrix} \equiv [\mathbf{R}_i\mathbf{R}^k_0|\mathbf{R}_i\kappa\mathbf{t}^k_0 + \mathbf{t}_i]$$

therefore, the projected pixel co-ordinates $\hat{\mathbf{x}}^k_{ij}$ can be stated as:

$$\hat{\mathbf{x}}^k_{ij} = \begin{bmatrix} x_\alpha \\ y_\alpha \\ 1 \end{bmatrix} \equiv \overline{f(\mathbf{K}^k\mathbf{x}_\beta)}$$

where $\overline{f(\mathbf{K}^k\mathbf{x}_\beta)}$ results in the affine part of $f(\mathbf{K}^k\mathbf{x}_\beta)$. Additionally, by making the assignment:

$$\mathbf{G} = \left[\begin{array}{cc} fs_x & \gamma \\ 0 & f \end{array}\right] \left[\begin{array}{ccc} \frac{1}{w_\beta} & 0 & \frac{-x_\beta}{w_\beta^2} \\ 0 & \frac{1}{w_\beta} & \frac{-y_\beta}{w_\beta^2} \end{array}\right]$$

(where f, s_x and γ are the intrinsic components focal length, aspect ratio and skew respectively) the components of the Jacobian can then be expressed simply as in Table I. As can be seen, the addition of the scale term does not significantly impact on the complexity of the Jacobian entries, allowing a computationally efficient solution.

TABLE I

THE ANALYTICAL DERIVATIVES

Stereo Transform	
$\frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \mathbf{t}_0^k} = \mathbf{G}\mathbf{R}_i\kappa$	$\frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \mathbf{r}^k} = \mathbf{G}\mathbf{R}_i \left[\bar{\mathbf{R}}^k\mathbf{X}_j\right]_\times$
$\frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \kappa} = \mathbf{G}\mathbf{R}_i\mathbf{t}_0^k$	
Extrinsics	
$\frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \mathbf{t}_i} = \mathbf{G}$	$\frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \mathbf{r}_i} = \mathbf{G}\left[\bar{\mathbf{R}}_i(\mathbf{R}^k\mathbf{X}_j + \kappa\mathbf{t}_0^k)\right]_\times$
Scene	
$\frac{\partial \hat{\mathbf{x}}_{ij}^k}{\partial \mathbf{X}_j} = \mathbf{G}\mathbf{R}_i\mathbf{R}^k$	

B. Pose Initialization

As has been described, in more traditional scenarios 3D scene structure is initially triangulated from the calibrated stereo pair, hence there is no need for a special initialization step. At large depths such triangulation 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 VO.

The novel components of the initialization procedure performed in this paper are described in Fig. 3. Initially, an essential matrix $\mathbf{E}_1$ between the base camera at two adjacent time-steps is recovered via the five-point algorithm [21], and relative pose (up to scale) extracted from this transform. To avoid degeneracies caused by near-planar structure, essential matrices pass an additional 'scene-spread' test as in [11]. This boot-strapping procedure ensures that accurate triangulation is achieved from a wide-baseline pair and is not dependent on the geometric stereo transform.

1) Monocular Visual Odometry: (**Fig. 3: Initialize Monocular VO**). From this initialization, a monocular VO is then performed using the imagery from the base camera only in 5 main repeating steps:

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

On a new image, upright SURF [22] descriptors are matched between the current and previous base camera. From already triangulated structure and feature matches to the previous image, the new base camera pose $\mathbf{P}_i^0 = \mathbf{K}^0\mathbf{M}_i$ is found using calibrated 3-point pose estimation, performed inside a robust MLESAC [23] estimator to ensure a reliable pose update. New structure is then triangulated using at least three observations, and then a Euclidean bundle adjustment is applied to the set of extracted poses and their associated structure. A Levenberg-Marquadt robust optimization routine is followed to ensure the estimation converges.

Importantly, the scale of this pose estimate is arbitrary, initialized so that the first camera pair has unity distance to ensure stable numerical precision, irrespective of the true distance between camera poses. In order to accurately recover scale, the stereo baseline must now be included (in lieu of other scale measurements from an IMU or external reference) by utilizing feature projections from the secondary camera to provide a geometric constraint on the scale term.

2) Scale Recovery: (**Fig. 3: Add secondary cameras, Estimate κ, Scale by $\frac{1}{\kappa}$**). Following an empirically derived number of initialization frames from the monocular VO, the secondary camera is introduced via the pre-calibrated stereo transform $\mathbf{T}_0^1$ to lie at the correctly scaled distance from each base camera. In this configuration, the re-projection error on the set of base cameras will be low due to the bundle adjusted set of poses and scene, while the re-projection error on the set of secondary cameras will be high due to the inaccurate scale. This will manifest in a large translational offset in feature projections in the second camera if the cameras are close to parallel in their x or y pixel co-ordinates, typical of most rigid-stereo configurations, whether vertical or horizontal.

The set of all base cameras and their corresponding secondary cameras are then optimized along with scene structure in a batch bundle adjustment that includes the scale term κ and the stereo transform subject to the previously mentioned boundaries. These boundaries allow small movements in the stereo transform to accommodate deformation induced error in the epipolar geometry without adversely impacting scale or camera pose.

Once the optimization routine has converged and the scale parameter κ is well estimated, the entire scene and camera geometry is scaled by the inverse $\frac{1}{\kappa}$ to recover metricity, forcing κ back to a unity value and the final solution as correctly scaled.

C. Stereo Visual Odometry

Following the monocular initialization and scale recovery, stereo VO can now perform as normal on the correctly scaled cameras and scene, closely following the 5 main steps as described in II-B.1. On a new set of images from a stereo pair, features are matched both between the pair and the previous base camera. The new base camera pose $\mathbf{P}_i^0$ is found using calibrated 3-point pose estimation, and the secondary camera initialized via the stereo transform $\mathbf{T}_0^1$. New structure is then triangulated using only the base camera set to avoid dependence on fixed-stereo geometry. Following this, the stereo bundle adjustment algorithm (including all cameras, the stereo transform with constraints, but not the scale term) is applied on a sliding window of 12 of the most recent frame pairs and their associated structure.

III. EXPERIMENTAL RESULTS

To investigate the applicability of the algorithm and its robustness over multiple trials and situations, we present results evaluated on both a simulated dataset and field gathered data. To show the performance of the proposed algorithm that includes both a monocular initialization and scale optimization, we compare it to a ‘standard’ stereo VO that triangulates structure from rigid stereo pairs and against a similar monocular intilizer that does not explicitly optimize scale.

A. Simulated Experiments

In the first simulated experiment, each algorithm is evaluated over multiple trials at a range of baseline-to-depth ratios. This is achieved by flying a pair of simulated cameras at varying altitudes over a simulated 3D scene generated from previously gathered LiDAR data[1] (Fig. 4). In all experiments the cameras have a $0.75m$ baseline and are flown at speeds ranging from $15m/s$ to $50m/s$ to ensure image overlap, mirroring the motion of a generalized fixed-wing UAV model. The total distance covered in each experiment is set to approximately $60m$, independent of flight speed and frame coverage. Additionally, the number of features tracked per frame is kept approximately equal independent of altitude. Each image has a resolution of 1024×768 pixels and each feature is projected with Gaussian noise of 1.0 pixels standard deviation. The simulation is performed for 20 trials at each altitude, and the final pose error compared to ground-truth recorded at each iteration.

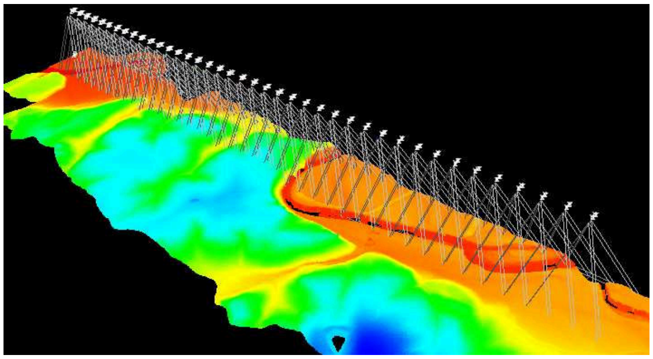

Fig. 4. The simulated scene

In the second experiment, the camera pair is flown at a $70m$ altitude for the same $60m$ distance, with a fixed flight speed of $30m/s$. However, in this case the proposed algorithm is initialized at a range of scales, defined by the Euclidean distance between the first monocular camera pair (see Fig. 3). By varying this value, the ability of the modified bundle adjustment algorithm to converge in the presence of a poor initial scale is examined. Without the optimization of κ as representative of the scale of the scene, bundle adjustment is rendered poorly initialized and should only converge if the initial scale closely approximates the truth. In contrast, with the addition of κ to the optimization, bundle adjustment should converge from a wide range of initialization scales. For this experiment, the algorithm with and without scale optimization is run for a range of initial scale errors, with 20 trials performed at each step to examine consistency of performance.

1) Results: The results of these experiments are shown in Figs. 5 and 6, where the final pose error of the novel initialization scheme is directly compared to the standard VO in the first experiment, and the relative scale before and after optimization with and without κ optimization is demonstrated in the second.

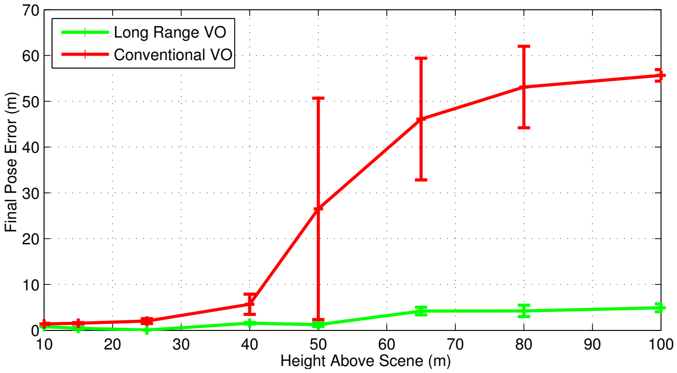

Fig. 5. Final pose error for varying altitudes calculated from the Euclidean distance between the final camera poses of the VO and ground truth. Conventional stereo VO in red, modified long-range VO in green.

In the first experiment (Fig. 5), the Euclidean distance between the final camera in the initialization and ground truth is used to define the final pose error. As reflected in the large variance and high average pose error of the conventional VO above altitudes of approximately $40m$, or a baseline-to-depth ratio of 1.9×10^{-2}, the standard stereo pose estimator fails to recover a sufficiently accurate pose. In many cases, the standard stereo estimator fails before the minimum number of frames is completed due to the poor triangulation and subsequent lack of viable feature tracks, in addition to forming degenerate pose updates that do not follow the known motion of the cameras. As the altitude approaches $100m$, the conventional estimator consistently fails, approaching the limit of error at $60m$, reflected in the high error, low variance result at the highest altitude.

In contrast, the long-range estimator with scale optimization shows robust performance even at altitudes up to $100m$, achieving an accurately scaled pose within $5m$ over the trajectory in all trials at the highest simulated altitude. In this configuration, according to Fig. 1, a 100m altitude corresponds to a disparity of ~9 pixels and a depth resolution of $12.5m$ from single pairs. As is clear from this analysis, triangulation from a stereo pair in a conventional VO is difficult and will generally fail in the first few frames at high altitudes.

In the second experiment (Fig. 6), the scale error (expressed as the ratio between ground truth and recovered distance) before and after optimization is demonstrated on the long-range VO algorithm. A direct comparison is made between convergence with explicit κ optimization,

[1] http://www.liblas.org/samples/

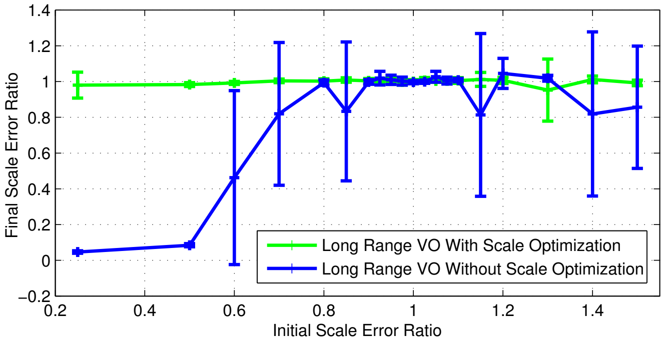

Fig. 6. Scale error comparison before and after optimization with and without explicit scale optimization. A scale of 1 is considered optimal. Bundle adjustment with κ in green, bundle adjustment without κ in blue.

and without. As can be seen, the addition of explicit scale optimization ensures the algorithm converges appropriately at a wide range of scale ratios $(0.25 \leftrightarrow 1.5)$. However, without the addition of the κ term the bundle adjustment algorithm fails to converge reliably except within a range of $\sim 10\%$ of the true value, demonstrating the necessity of the addition of the scale term. This addition ensures far more reliable convergence at a wide range of unknown initial scale values.

B. Field Data Experiment

In this experiment, stereo visual data gathered from a UAV is used to test the proposed algorithm and compare its output to a GPS/INS ground-truth. In this experiment, due to vibration induced deformation on the camera rig, the epipolar geometry is not well aligned causing standard methods of stereo VO to fail, even at low altitude, due to inaccurate triangulation. To counteract this, we implement optimization of the stereo transform during the bundle adjustment in addition to estimation of scale. Additionally, the altitudes at which the VO algorithm is expected to accurately initialize are below the minimum baseline-to-depth ratio demonstrated successfully for the conventional VO in the simulated experiment, meaning a conventional stereo VO method is incapable of performing adequately.

1) Experimental Platform: The platform used for data gathering is a remotely-piloted fixed-wing Unmanned Aerial Vehicle (UAV) (Fig. 7), flown within visual line of sight from the ground. The aircraft includes two Flea2 Firewire cameras, rigidly fixed to each other via an aluminium L-bar with an approximate baseline of 0.77m, placed looking down towards the terrain (see Fig. 7). An XSens Mti-G INS/GPS system is used as the ground truth measurement system, with the inertial measurement unit placed on the camera rig approximately 10cm behind the front camera, and the GPS receiver placed on the fuselage directly above the front camera. An off-the-shelf computer system is used to log image data, raw inertial and GPS data from the INS, along with a filtered pose estimate from the INS.

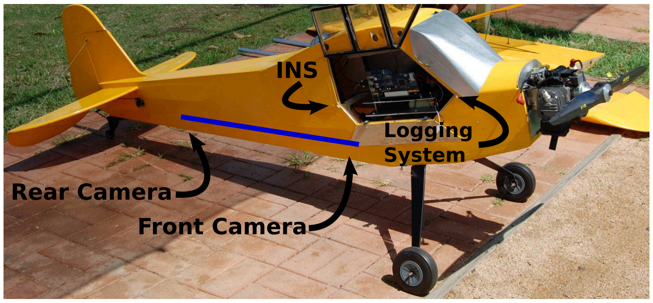

Fig. 7. The experimental platform showing component layout. Blue line indicates length and orientation of stereo baseline between on-board cameras.

Each camera uses a $6mm$ lens with a field of view of approximately $42° \times 32°$. The cameras are calibrated before flight using a calibration-pattern to achieve an intrinsic calibration for each camera and an approximated stereo transform between the cameras. In flight, however, vibration causes misalignment in the epipolar geometry on the order of around 15 pixels.

2) Dataset: Data was collected over an approximately 5 minute flight, at an altitude of $20 - 100m$ and a speed of $\sim 20m/s$. Bayer encoded colour images are logged at a resolution of 1280×960 pixels at $30Hz$ and later converted to color for processing (Fig. 8). Raw GPS and filtered INS poses were recorded at $4Hz$ and $120Hz$ respectively from the XSens MTi-G to give a ground truth position comparison. The area flown over by the aircraft consisted of rural farmland with relatively few trees, animals and buildings.

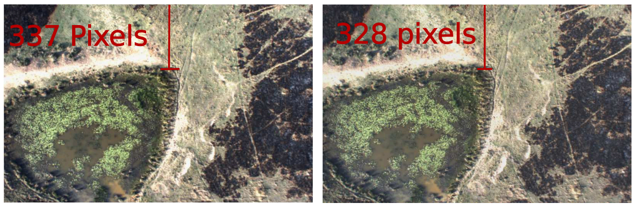

Fig. 8. An example image pair from the dataset, showing the small disparity between the stereo pair. Left: Front Camera, Right: Rear Camera.

C. Results

The monocular initialization routine and stereo VO algorithm were applied to a sequence of the trajectory at approximately $80 - 100m$ above ground level, as shown in Figures 9 and 10. With a final pose error between the modified VO and ground truth of $22m$ over the $2.70km$ trajectory, and a total distance covered within an error of $< 2\%$, the result is within the error bounds of most visual odometry algorithms in the typically simpler ground-based case. From this, it can be seen that the presented pose initializer is capable of estimating scale accurately even at high altitude on field-gathered data, and allowing a metrically scaled pose estimate at extremely small baseline-to-depth ratios.

As described, a conventional VO fails within a few frames and does not perform adequately in this scenario. This is due to both the inadequate structure triangulation from a single pair due to the distance of the scene, but also the poor epipolar geometry.

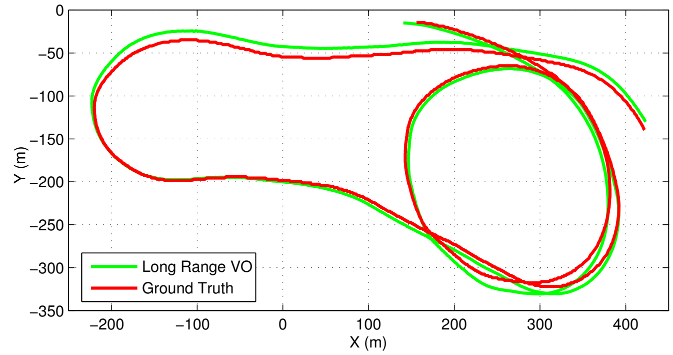

Fig. 9. Top view of comparison between visual odometry (green) and GPS/INS ground truth (red). The entire trajectory is 2698m, with a relative distance error of $\sim 2\%$. Note the accurate scale of the final trajectory.

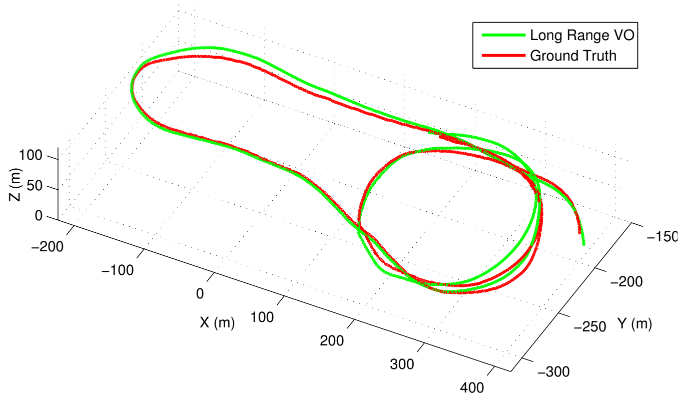

Fig. 10. Perspective view comparison between visual odometry (green) and GPS/INS ground truth (red), showing the approximate height of the trajectory from the ground plane.

IV. CONCLUSIONS

This paper has demonstrated an algorithm capable of accurately initializing a metrically scaled pose from purely visual data in a long-range stereo application, where triangulation from a rigid stereo pair is unreliable and standard VO methods of initialization will fail. By introducing a novel initialization that avoids rigid stereo triangulation and adding a new scale term into a stereo-aware bundle adjustment routine, an accurately scaled pose estimate can be generated at altitudes exceeding $80m$. Through the addition of a scale term, the optimization is well initialized and allows accurate estimation of the scale ratio between the estimate and truth defined by the previously calibrated stereo baseline. The technique has been shown on both simulated data and a difficult airborne dataset where standard stereo algorithms fail.

Future work will involve demonstrating the long-range stereo initialization and VO on a full dataset with loop-closure to constrain pose drift and an examination of the limits of the baseline-to-depth ratio on the performance of the algorithm.

REFERENCES

[1] A. Comport, E. Malis, and P. Rives, “Real-time Quadrifocal Visual Odometry,” *The International Journal of Robotics Research*, vol. 1, 2010.

[2] H. Lategahn and C. Stiller, “City GPS using Stereo Vision,” *2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012)*, pp. 1–6, July 2012.

[3] 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.

[4] 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.

[5] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, “Variable Baseline/Resolution Stereo,” *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

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

[7] B. Kitt, A. Geiger, and H. Lategahn, “Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme,” *2010 IEEE Intelligent Vehicles Symposium*, pp. 486–492, June 2010.

[8] 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.

[9] S. Weiss, M. Achtelik, and S. Lynen, “Real-time Onboard Visual-Inertial State Estimation and Self-Calibration of MAVs in Unknown Environments,” in *International Conference on Robotics and Automation*, vol. 231855, 2012.

[10] 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.

[11] 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.

[12] 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.

[13] A. Chambers, S. Achar, S. Nuske, J. Rehder, B. Kitt, L. Chamberlain, J. Haines, S. Scherer, and S. Singh, “Perception for a river mapping robot,” *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 227–234, Sept. 2011.

[14] A. Geiger, J. Ziegler, and C. Stiller, “StereoScan: Dense 3d reconstruction in real-time,” *2011 IEEE Intelligent Vehicles Symposium (IV)*, pp. 963–968, June 2011.

[15] G. Sibley, L. Matthies, and G. Sukhatme, “Bias reduction and filter convergence for long range stereo,” *International Journal of Robotics Research*, 2005.

[16] J. Rehder, K. Gupta, S. Nuske, and S. Singh, “Global pose estimation with limited GPS and long range visual odometry,” *2012 IEEE International Conference on Robotics and Automation*, pp. 627–633, May 2012.

[17] G. Sibley, G. Sukhatme, and L. Matthies, “The iterated sigma point kalman filter with applications to long range stereo,” in *Proceedings of Robotics: Science and Systems*, 2006.

[18] M. Warren and B. Upcroft, “High Altitude Stereo Visual Odometry,” in *Proceedings of Robotics: Science and Systems*, 2013.

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

[20] 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, July 2012.

[21] D. Nistér, “An efficient solution to the five-point relative pose problem.” *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 6, pp. 756–77, June 2004.

[22] H. Bay, T. Tuytelaars, and L. V. Gool, “Surf: Speeded up robust features,” in *European Conference on Computer Vision*, 2006.

[23] P. Torr, “MLESAC: A New Robust Estimator with Application to Estimating Image Geometry,” *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, Apr. 2000.