A Robust Structure and Motion Replacement for Bundle Adjustment

Shane Mitchell
University of Queensland, Australia
shane.mitchell@uqconnect.edu.au

Michael Warren
University of Queensland, Australia
m.warren1@uq.edu.au

David McKinnon
Queensland University of Technology, Australia
david.mckinnon@qut.edu.au

Ben Upcroft
University of Queensland, Australia
ben.upcroft@uq.edu.au

Abstract

This paper demonstrates the application of a robust form of pose estimation and scene reconstruction using data from camera images. We demonstrate results that suggest the ability of the algorithm to rival methods of RANSAC-based pose estimation polished by bundle adjustment in terms of solution robustness, speed and accuracy, even when given poor initialisations. Our simulated results show the behaviour of the algorithm in a number of novel simulated scenarios reflective of real world cases that show the ability of the algorithm to handle large observation noise and difficult reconstruction scenes. These results have a number of implications for the vision and robotics community, and show that the application of visual motion estimation on robotic platforms in an online fashion is approaching real-world feasibility.

1 Introduction

This paper presents the application of a new robust structure and motion estimation technique to the application of visual motion estimation in a number of simulated scenarios.

Visual motion estimation is a key technique in photogrammetry, computer vision and robotics. The ability to obtain accurate motion estimates and scene structure from vision based techniques is a critical component of developing low-cost, automated control of robotic platforms and producing high-quality 3D reconstructions of observed scenes.

RANSAC-based essential matrix estimation polished by bundle adjustment is the current state of the art in visual pose estimation techniques and has been applied to the fields of robotic control, SLAM and visual scene reconstruction. Bundle adjustment is, however, fundamentally a least-squares optimisation technique and as such is subject to local minima that mean good solutions are not always guaranteed. Bundle adjustment is also prohibitively slow to use in low-cost, online robotics applications and is difficult to parallelise. The need for a high-speed, automated technique that is guaranteed to give accurate estimates in the face of poor scene features, bad matches and difficult scenarios such as planar scenes is essential for high-quality, low-cost robotic sensing and control tasks.

We apply a relatively new algorithm presented in the literature [Schweighofer and Pinz, 2006] that uses an alternative method of estimating structure and pose in an attempt to avoid the inherent local minima problems in current techniques. This algorithm is theoretically 8 times faster than sparse-aware versions of bundle adjustment and is largely parallelisable. We attempt to confirm the theory of the algorithm by applying our own implementation to several simulated pose estimation scenarios and a novel robotics-oriented simulated UAV scenario.

2 Related Work

2.1 Visual Pose Estimation

The estimation of pose of a calibrated camera given a number of image frames is a longstanding problem in photogrammetry, computer vision and robotics. By knowing the true pose of a camera, a number of applications are immediately apparent. Given corresponding scene points a reconstruction of the scene can be generated to give accurate 3D models without the use of other expensive sensors such as laser range-finders. In robotics, the ability to accurately estimate pose given a

low cost sensor such as a camera is an extremely important application [Konolige *et al.*, 2007; Nistér *et al.*, 2006; Konolige *et al.*, 2010]. Such an ability allows robots to safely traverse and interact in the real-world environment [Braid *et al.*, 2006], and given the low-cost of visual sensors presents the opportunity for low-cost robots.

Many techniques exist to solve the problem of camera pose [Haralick *et al.*, 1994; Kalantari *et al.*, 2009; Nistér and Stewénius, 2007; Comport *et al.*, 2010], and often generate a solution given corresponding scene points generated by a feature detector such as SIFT [Lowe, 1999] or SURF [Bay *et al.*, 2006]. Others may depend on other salient features such as line segments or fiducial based landmarks as mapped 3D references [Nuske *et al.*, 2009].

Many techniques take a small sample of points in a RANSAC [Fischler and Bolles, 1981] based fashion and attempt to extract the fundamental or essential matrix [Luong and Faugeras, 1996] through varying methods [Hartley and Zisserman, 2004; Nistér, 2004]. These techniques, however, are usually subject to a number of important metrics for accurate pose estimation, such as a necessary high number of feature correspondences, a low number of incorrect matches, or accurate estimation of 3D structure. For real-world data this is not an easy problem.

Optimisation of an initial pose estimate generated by RANSAC depends on an already high-quality estimate. Techniques such as bundle adjustment are subject to the issues common to all least-squares minimisation techniques, where the final solution can be a local, rather than a global, minimum.

2.2 General Camera Models

The General Camera model is an imaging model that attempts to create a generic representation for all imaging sensor types [Sturm, 2005; Grossberg and Nayar, 2001]. By implementing such a model, the type of camera used in visual sensing tasks is abstracted and useful algorithms become independent of camera type. This powerful concept can be applied to many visual pose and scene structure estimation tasks [Lu *et al.*, 2000].

By using a general camera model, the application of monocular, stereo or more cameras for a given task becomes significantly easier as the calibration and scene triangulation methods are abstracted away from any algorithm [Sturm and Ramalingam, 2004].

2.3 Bundle Adjustment

Bundle adjustment is an application of the well known non-linear least squares error minimisation routine for large visual pose and structure estimation problems [Triggs *et al.*, 2000; Engels *et al.*, 2006], and is used to optimise both frame and feature positions from estimates generated by methods described above to reach a refined estimate with minimal re-projection error. It is a widely accepted method for improving such pose and structure estimates and, when carefully implemented, can give impressive results [Konolige and Agrawal, 2008; Sibley *et al.*, 2009]. However, bundle adjustment has a number of inherent deficiencies which have the potential to limit its usefulness in pose estimation tasks. For example, bundle adjustment will often not converge to a correct solution if the initial estimate is poor. This can be attributed in part to the large number of parameters with which it is often required to optimise over, causing the algorithm to converge to a local minimum in the cost function which can often be of significant distance from the global minimum. Solutions to the problem include reducing the number of optimisation parameters or initialising the routine with a good initial estimate. In visual pose and scene reconstruction tasks, this is sometimes difficult to achieve with reliability, and is one of the last key blocks to implementation of a robust visual pose and structure algorithm for use in robotic and mapping tasks.

Additionally, bundle adjustment is a computationally expensive procedure. In the naive case, bundle adjustment requires large Jacobian matrix inversions that consume a significant amount of CPU time, but by intelligently taking into account the fact that the relevant matrices are often sparse and segmenting and inverting parts of the matrix that are symmetric the overall speed can be improved [Engels *et al.*, 2006]. However, even with such improvements, bundle adjustment remains a problem that often cannot be used in an online fashion for robotic tasks on low-cost hardware due to the computational processing involved.

2.4 Robust Structure and Motion Estimation

Schweighofer *et al.* [Schweighofer and Pinz, 2006] present a new method of structure and motion estimation that *attempts* to solve structure and motion problems in a fast, easily implemented and globally optimal fashion. The method attempts to solve pose and structure in a way that means bundle adjustment is not required or necessary. The algorithm can use an initial estimate of structure and camera pose to converge quickly, but Schweighofer *et al.* suggest that such an initial estimate can be essentially random, meaning that the algorithm should be capable of converging globally from any initial estimate.

Rather than attempting to minimise the image space error (or pixel error) in normal bundle adjustment, the robust method attempts to minimise object space error, or reduce the error in the estimated position of an observed point in 3D space instead. This has been suggested to be a far better method of error modelling in

pose and scene structure estimation tasks [Schweighofer and Pinz, 2008; 2006; Schweighofer *et al.*, 2008].

As the robust method attempts to minimise the global object space error (for all points) and does not itself reject outliers, an outlier-rejection initialisation step, such as a RANSAC-based algorithm, should be used to increase accuracy. This initialisation is not considered in this paper.

The method presented by Schweighofer has been theoretically shown to be significantly faster than traditional bundle adjustment as it avoids expensive and sometimes unstable large, sparse matrix inversions and reaches a minimum extremely quickly, approaching the optimal solution within several iterations in certain cases.

Improvements to the theoretical algorithm have been proposed [Schweighofer *et al.*, 2008], and results presented to the effect, but as yet no significant application to robotic pose estimation or scene reconstruction has been presented.

The method presented by Schweighofer, however, contains some deficiencies. It is not as versatile as bundle adjustment in that it is not capable of estimating the parameters of the general camera model. Such parameters can include lens distortions in perspective cameras or extrinsic parameters in multi-camera rigs. The original method has also been shown to be not globally convergent in all cases, whereby local minima can still occur. This becomes apparent when many cameras are used and poor initial rotation estimates are given.

3 Theory

The optimal structure and motion method presented by Schweighofer in its original form [Schweighofer and Pinz, 2006] is reviewed here for clarity, but with a differing notation to take account of differences in implementation. We describe the object space error model as the fundamental component of the algorithm, the general camera model used in the algorithm (including the definition of raxels) and the divergence of our implementation from the previously described method.

3.1 General Camera Model

The algorithm makes use of the general camera model in an attempt to abstract away from the implementation details of the task. By using such a model, the method is capable of being applied to a wide range of scenarios, including those in robotics, scene reconstruction and object modelling.

The simplest general camera model is a single light sensor (raxel) with a ray of direction indicating the path of incoming light. It can also be extended to incorporate multiple constrained perspective cameras, mirrored cameras and compound cameras.

The model is capable, for a perspective camera, of incorporating distortion and focal length parameters in its mapping function. In our method we consider the general camera model implementation of such a perspective camera. As such, we work with calibrated cameras only.

Raxels

The general camera model uses the concept of a raxel as a fundamental atomic element. A raxel is essentially the equivalent of a pixel but with a vector indicating the direction of the light ray illuminating the pixel. The raxel can be denoted as the tuple $(\mathbf{c}, \mathbf{v})$, where $\mathbf{c}$ is the 3×1 translation portion of the raxel within the local co-ordinate system of the general camera, and $\mathbf{v}$ denotes the 3×1 vector of direction of the incoming ray. For the purposes of this paper, the magnitude of vector $\mathbf{v}$ is unimportant.

For a theoretical perspective camera, all raxels will be situated at the focal point of the camera. In a real camera, each raxel will exist in a plane at positions equivalent to pixels on the CCD. For an insect eye lens, each raxel will occur at each light sensing diode, with a ray of direction perpendicular to its surface.

For a single perspective camera, the conversion from pixel coordinates $\mathbf{p}$ to raxels is straight-forward and follows Equations 1 and 2.

$$\mathbf{c} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T \quad (1)$$

$$\mathbf{v} = \begin{bmatrix} p_x & p_y & 1 \end{bmatrix}^T \quad (2)$$

3.2 New Robust Structure and Motion Estimation Algorithm

In our implementation, we currently only consider the original implementation of the algorithm [Schweighofer and Pinz, 2006]. We assume that all features are observed in all frames, but take account that each feature may be observed more than once by a single general camera. Schweighofer *et al.* state that the algorithm is easily extendable to hidden views of features by setting certain parameters to zero, however we do not implement this functionality here.

Camera Setup for Mapping Point Correspondences

Our implementation of observation mappings differs slightly from the original work by Schweighofer *et al.* to account for cases where individual points imaged by multiple constrained cameras (e.g. a stereo rig) are correctly treated independently if mapped to a single general camera. In Schweighofer's original work, the implementation causes the algorithm to generate $n_c n_i$ 3D structure points given a set of n_c constrained cameras with n_i unique imaged points. We have extended the algorithm to support mapping between corresponding

points. The camera setup for the algorithm is shown in Figure 1. For each constrained camera rig (e.g. a stereo

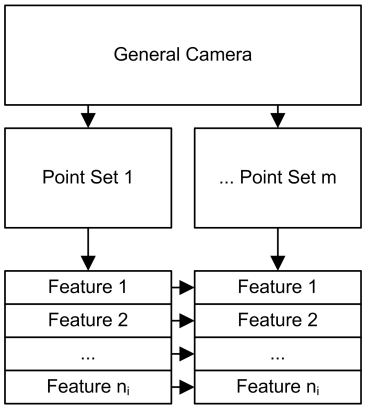

Figure 1: Camera setup for mapping point correspondences between constrained cameras.

rig consisting of two cameras) we identify one general camera, camera k, under which we identify one point set q for every constrained camera. For each unique feature imaged throughout all general cameras, we identify one feature, feature i. As an example, for 10 pairs of images taken from a stereo rig with 100 unique features over all images, there are $n_k = 10$ general cameras, $m_k = 2$ constrained cameras under each general camera and $n_i = 100$ unique features across all images.

This modification of the original algorithm effectively reduces the size of the matrix that needs to be solved in the translation step by a factor of how many cameras are in each set. In the case of stereo cameras, this leads to a speed-up factor of 8 compared to the original algorithm.

The algorithm attempts to find the pose for each general camera as a whole. The pose of the individual subcameras can easily be found as their positions are relative

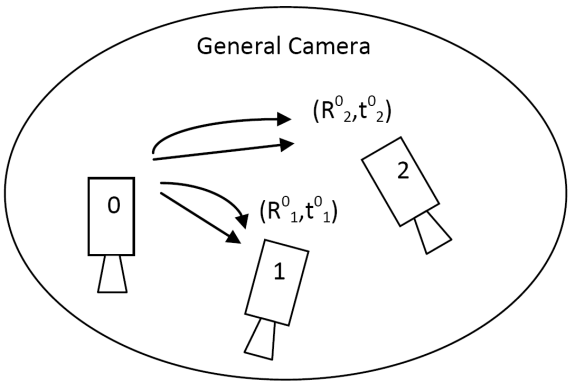

Figure 2: Mapping of cameras into a single general camera.

to the pose of the general camera.

To determine the raxel parameters for constrained cameras we use Equations 3 and 4

$$\mathbf{c}_{qi} = -\left(R_q^0\right)^T \mathbf{t}_q^0 \quad (3)$$

$$\mathbf{v}_{qi} = \left(R_q^0\right)^T \begin{bmatrix} p_x & p_y & 1 \end{bmatrix}^T \quad (4)$$

where R_q^0 and t_q^0 are the rotations and translations respectively relative to the first camera in the general camera to the qth camera (Figure 2), and $(\mathbf{c}_{qi}, \mathbf{v}_{qi})$ are the raxel parameters for the ith point under the qth camera.

Object Space Error

The essential component of the optimal structure and motion method is the Object Space Error cost function:

$$e_{kqi} = \|(I - V_{kqi})(R_k \mathbf{X}_i + \mathbf{t}_k - \mathbf{c}_{kqi})\|^2 \quad (5)$$

where

$$V_{kqi} = \frac{\mathbf{v}_{kqi}\mathbf{v}_{kqi}^T}{\mathbf{v}_{kqi}^T\mathbf{v}_{kqi}} \quad (6)$$

The algorithm attempts to minimise the cost e_{kqi} (cost for general camera k, point set q, feature i) for the 3D position X_i of point i and camera extrinsics R_k and t_k. This is in contrast to traditional bundle adjustment where the image-space error is the usual metric. Schweighofer suggests that as well as being a far more accurate error modelling metric, the object space error method also enhances the speed at which a solution is reached.

Review of the Algorithm

The general procedure to solve the structure and pose is given in Algorithm 1. Each step in the process is elaborated on in subsequent sections.

Algorithm 1 Robust Structure and Motion

Input: A set of 2D point correspondences occurring in a set of images. An initial estimate of rotation R_k for each camera pose k.

Output: An estimate of camera pose for each image and a 3D point cloud representing visible features

1: **while** summed structure error difference $e_i - e_{i-1} > \epsilon$ **do**
2: Estimate the parametrisation of the structure: $\tilde{Y}_i$, $\tilde{X}_{kqi}$, $\tilde{x}_i$
3: Estimate the optimal translation t_k for all cameras k
4: Estimate the optimal structure X_i
5: Find a new R_k for all cameras k using optimal absolute orientation
6: **end while**

1. Estimation of Parametrisation of Structure

The parametrisation of the structure is required for future steps in the algorithm, and can be generated from simple summations:

$$\tilde{X}_{kqi}(R) = -\tilde{Y}_i R_k^T Q_{kqi} \tag{7}$$

$$\tilde{x}_i(R) = \tilde{Y}_i \sum_{k=1}^{n_k} \sum_{q=1}^{m_k} R_k^T Q_{kqi} \mathbf{c}_{kqi} \tag{8}$$

$$\tilde{Y}_i(R) = \left(\sum_{k=1}^{n_k} \sum_{q=1}^{m_k} R_k^T Q_{kqi} R_k \right)^{-1} \tag{9}$$

$$Q_{kqi} = (I - V_{kqi})^T (I - V_{kqi}) \tag{10}$$

2. Estimating Translation

The optimum translation t_j can be determined with knowledge of R_j and the parameterisation of the structure. For each general camera j the following linear system must be solved. Note that subscript j is used to denote individual equations for each general camera j, while other subscripts used in the equations are purely to support the summation notation.

$$\sum_{k=1}^{n_k} \left(\sum_{q=1}^{m_k} \sum_{w=1}^{m_j} \sum_{i=1}^{n_i} \tilde{X}_{jwi}^T R_k^T Q_{kqi} \right) \mathbf{t}_k + \sum_{w=1}^{m_j} \sum_{i=1}^{n_i} Q_{jwi} \mathbf{t}_j + \sum_{w=1}^{m_j} \sum_{i=1}^{n_i} Q_{jwi} \left(R_q \tilde{x}_i - \mathbf{c}_{jwi} \right) = 0$$

By compiling all n_k equations into the form $A\mathbf{t} = \mathbf{b}$, the translation can be solved by letting $\mathbf{t} = A^{-1}\mathbf{b}$ provided matrix A is fully ranked. In the case where there are at least two constrained cameras in the system, matrix A has a rank of $n_k - 3$ and is rank deficient. This corresponds to the free translation of the system along each of the 3 coordinate axes. The problem can be fixed by constraining the translation of one general camera (e.g. $\mathbf{t}_0$). This brings matrix A back to full rank without overconstraining the system.

3. Estimating Structure

The optimum structure can be determined with foreknowledge of R and t for all poses using the following formula:

$$\mathbf{X}_i(R,t) = \tilde{x}_i + \sum_{k=1}^{n_k} \sum_{q=1}^{m_k} \tilde{X}_{kqi} \mathbf{t}_k \tag{11}$$

4. Estimating Rotation

Given an estimate of the structure and pose of the general cameras, optimal absolute orientation can be used to determine the optimum rotation of each general camera. Assuming that the λ^{th} estimate of R_k is $R_k^{(\lambda)}$, the next estimate can be estimated as the solution of the optimal absolute orientation problem:

$$R_k^{(\lambda+1)} = \arg\min_R \sum_{q=1}^{m_k} \sum_{i=1}^{n_i} \left\| R_k^{\lambda} \mathbf{X}_i + \mathbf{t}_k^{(\lambda)} - \left(\mathbf{c}_{kqi} - V_{kqi} \mathbf{q}_{kqi}^{(\lambda)} \right) \right\| \tag{12}$$

where

$$\mathbf{q}_{kqi} = R_k \mathbf{X}_i + \mathbf{t}_k - \mathbf{c}_{kqi}$$

This can be solved using SVD [Arun *et al.*, 1987] or quaternions [Horn, 1987]. It has been experimentally observed that on occasions $\det\left(R_k^{(\lambda+1)}\right) = -1$. This result corresponds to a reflection instead of a desired rotation. In these situations, the algorithm presented in this paper converges to a stable point which may not be the global minimum. The strategy employed to fix this situation is to reinitialise R_k to a random rotation matrix or another suitable initialisation when a reflection is detected. This has been empirically shown to, given a sufficient number of re-initialisations, set the orientation close enough to the correct orientation so as to reach the global minimum.

Speed of the algorithm

Schweighofer *et al.* have shown that the algorithm is approximately 8 times faster than sparse versions of bundle adjustment. The time complexity of the algorithm is $O_{inv}(3n_k) + O(n_i m_k n_k^2)$ where O_{inv} represents the complexity of performing a matrix inversion. Where $n_i \gg n_k$ the algorithm is linear with the number of points and when $n_k \gg n_i$ the algorithm is cubic with the number of cameras. This is an important result when considering the online application of the algorithm on robotic platforms and other pose estimation scenarios.

Solutions of the algorithm for the monocular case

In the general camera model as described by Schweighofer *et al.* [Schweighofer and Pinz, 2006], the implementation of a monocular perspective camera has a fundamental failing. In this case the position of all raxels, c, are at a single infinitesimal point. While not a fundamentally critical issue in and of itself, the failure occurs during the optimal translation step. The solution attempts to find a system of equations $A\mathbf{t} = \mathbf{b}$, such that:

$$\mathbf{b} = -\sum_{w=1}^{m_j} \sum_{i=1}^{n_i} Q_{jwi} \left(R_q \tilde{x}_i - \mathbf{c}_{jwi} \right) \tag{13}$$

where

$$\tilde{x}_i = \tilde{Y}_i \sum_{k=1}^{n_k} \sum_{q=1}^{m_k} R_k^T Q_{kqi} \mathbf{c}_{kqi} \tag{14}$$

In this monocular case, it is seen that with all $\mathbf{c} = 0$, subsequently $\mathbf{b} = \mathbf{0}$, generating the trivial solution $\mathbf{t} = \mathbf{0}$. As such the algorithm fails in a purely monocular setup. The monocular method is not considered in this paper, although it should be noted that we have obtained preliminary results indicating the solution to monocular configurations is possible using a variant of this algorithm. This is achieved by introducing a non-linear Euclidean distance constraint between two cameras and solving accordingly, however, this makes convergence much more susceptible to errors in the initial rotation estimate. In this paper, the stereo scheme is considered without loss of impact.

4 Experiments

We have developed a number of experiments to evaluate our implementation of the robust structure and motion method, and evaluate it's effectiveness given poor initialisations and large observation noise. Our implementation runs in Python and uses the Numpy and SciPy libraries to evaluate and plot results. Three separate experiments were conducted:

4.1 Experiment 1: Random Orientation Trial

A simulated scene of 20 features in a 3m side-length cube was randomly generated. A set of 5 different geometrically constrained stereo pairs was generated that observed all scene features from 5 randomly selected viewpoints at a radius of approximately 6m from the centre of the observed scene. An example of this scenario can be seen in Figure 3. The initial estimates of the camera poses were randomly generated and the algorithm was given no knowledge of the actual 3D structure. No noise was added to observations of scene features. The experiment was repeated 100 times with different observed scenes and different camera viewpoints to demonstrate its robustness.

Results

The total error for each iteration over all 100 trials of Experiment 1 is shown in Figure 4. Our algorithm demonstrates 100% convergence rate and reaches a high precision of accuracy within 5 to 15 iterations. Cases where the error does not immediately decrease, found to be caused by the previously discussed local minima at reflections in the optimal absolute orientation step, were successfully detected and corrected by our implementation of the algorithm. In these cases the re-initialisation of the affected poses to random rotation matrices successfully corrected the convergence.

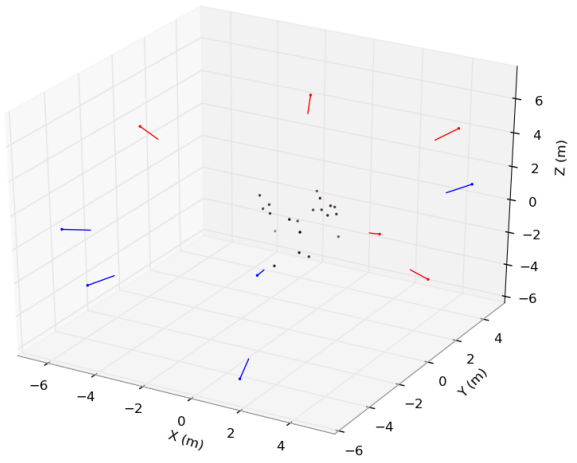

Figure 3: Example simulation setup for Experiments 1 and 2. Ten randomly placed cameras observe a scene of visible features. Every second camera (blue) is geometrically constrained to the camera generated before it (red) to form five general cameras. Each line represents the direction the respective camera is facing.

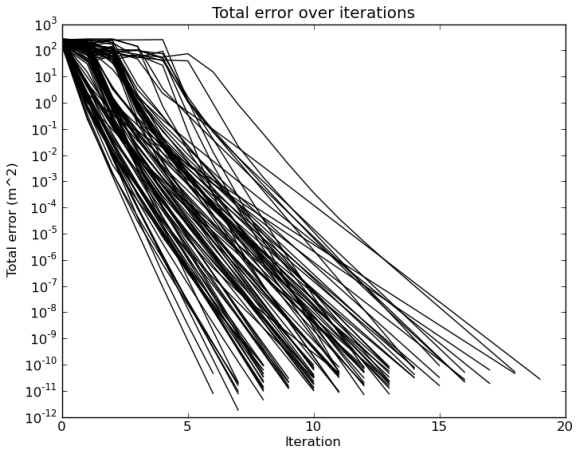

Figure 4: Total object space error for each iteration of the first experiment. Global convergence is observed in every trial.

4.2 Experiment 2: Random Orientation Trial with Observation Noise

The first experiment was repeated, but Gaussian noise of 0.1%, 0.2% and 0.4% covariance was added to all feature observations in 20 separate trials for each case. These noise values correspond to observation pixel noise of approximately 1, 2 and 4 pixels respectively in a 1024×1024 pixel image. If, for example, a constrained stereo pair

with $1m$ baseline were to observe this scene from 6 metres, this is reflective of an approximate error in depth resolution of 10, 20 and 50mm respectively. No outlier rejection was implemented in any scenario. This experiment assists in evaluating the algorithm's ability to generate an accurate solution even under high noise scenarios.

For this experiment, the initial estimates of camera pose were again randomly generated and the algorithm was given no initial knowledge of scene structure. The experiment was repeated for each of the three error covariance cases in a demonstration of robustness with high observation noise.

Results

For each trial in Experiment 2 the algorithm successfully converged close to the global minimum in all noise cases (Figure 5). As expected, the algorithm did not converge to the true solution, but converged successfully to an approximation within the assumed error.

The maximum number of iterations required to converge such that the breakout policy was satisfied was 36 iterations, however, the majority of trials converged within 15 iterations. The outlying convergence case can be attributed to a reflection that lies in a minimum close to the convergence path of this scenario. It can be seen, however, that a re-initialised pose successfully avoids the local minimum and converges globally.

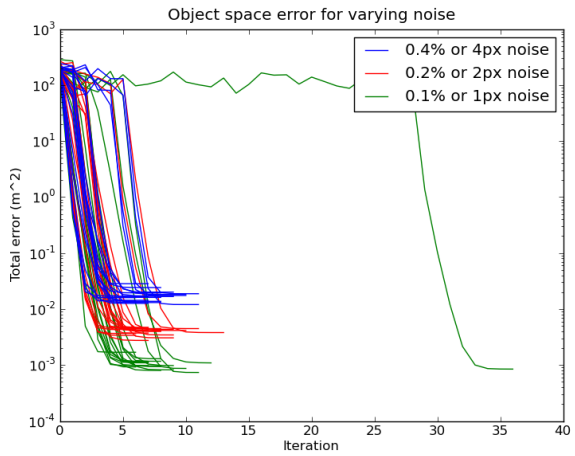

Figure 5: Total object space error for each iteration of the second experiment for (green) the 0.1% noise case, (red), the 0.2% noise case, (blue) the 0.4% noise case. Convergence to a solution that approximates the truth is seen in all trials.

The maximum final errors in rotation for all trials in the 0.1%, 0.2% and 0.4% noise cases were 0.06°, 0.15° and 0.18° respectively, and the maximum errors in translation were 0.7, 1.9 and 2.7 cm respectively. Averaged total object space errors were 0.001, 0.004 and 0.018 m^2 respectively.

4.3 Experiment 3: Simulated UAV Camera Pose Trial

In the final experiment, to emphasise the relevance of the pose estimation method to robotic applications, a simulation was developed that is reflective of a robotic UAV using vision as the input for pose estimation. A stereo pair of cameras was placed in a curved trajectory reflective of the path of an aircraft at a distance of 50m above a ground surface plane (Figure 6). The stereo pair was given a 1 metre baseline and the rig was initially oriented towards the ground plane. A roll around the aircraft's flight direction vector was applied to the aircraft pose over the trajectory in a reflection of the conditions experienced in a real scenario.

Scene features were randomly placed in a $10 \times 100m$ grid on the ground plane with random Gaussian noise applied to their height from the ground plane of 2m to simulate the ground surface.

Three separate trials were run, where features were observed with 0.1%, 0.2% and 0.4% Gaussian noise, again corresponding to approximately 1, 2 and 4 pixels noise respectively. Such observation noise is especially detrimental to accurate feature observation in these scenarios due to the distance of the observed features from the rig. In each of the scenarios, pixel noise of 1, 2 and 4 pixels resolves to approximately 0.8, 1.7 and 3.6 m variance in depth resolution for our setup, indicating how even small observation noise can be significant when use of cameras is applied to highly distant observations. Again, no outlier rejection was implemented. These trials are deemed to be a stringent test of the ability of the algorithm to handle noise in a real robotic scenario.

Results

The convergence of each trial in Experiment 3 is shown in Figure 7. For comparison, the horizontal lines in the figure indicate the simulated noise in structure given theoretically perfect pose estimates. In this experiment, in all 3 noise cases, all 20 sets of cameras converged (within error) to their correct pose along the UAV flight path and all structure points were determined to a high accuracy. Minimal reflections were encountered in this experiment due to the more accurate initial estimates given to the algorithm. However, when detected, the affected rotation matrices were successfully reinitialised to point vertically down-wards.

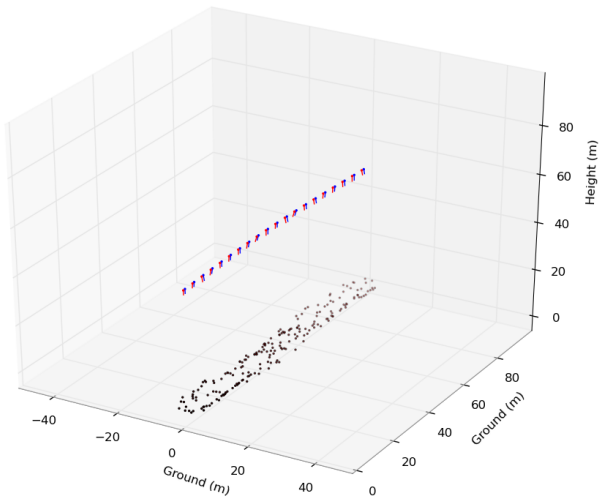

Figure 6: A view of a trial of the simulated UAV dataset showing camera trajectory, layout of scene and observed features.

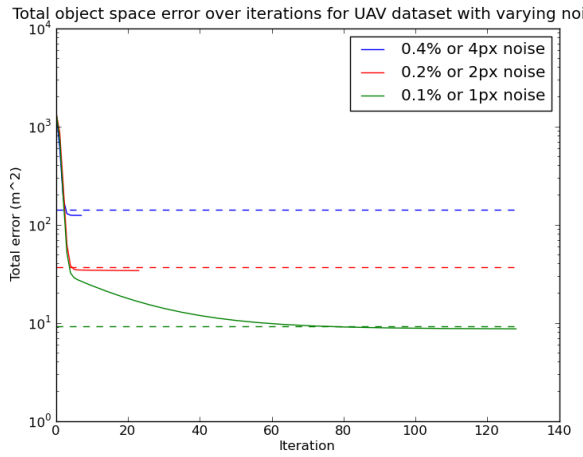

Figure 7: Accumulated total error in structure for each iteration in each trial of experiment 3. Each horizontal line represents the calculated error in structure given perfect pose estimates.

To emphasise the ability of the algorithm to converge to a solution given large observation noise, histograms reflecting the error in final pose and scene structure are shown in Figures 8, 9 and 10. As can be seen, even in high observation noise situations, the algorithm is capable of producing a very high quality rotation estimation and good estimations of translation and scene structure given the inclusion of all features and no outlier rejection.

5 Conclusion

We have shown that the algorithm presented in this paper is applicable to a variety of structure from motion problems, and that initial pose estimations are not necessary for global convergence. In all experiments the algorithm converged to an accurate solution indicating the robustness of the algorithm to local minima. It has also been demonstrated that the algorithm is capable of converging robustly even with noisy measurements, poor observation scenarios and data that is close to planar in resemblance of real-world cases.

5.1 Future Work

Following improvements in the theory behind the optimal structure and pose estimation, such as dealing with the single camera case and real-time implementation of the algorithm, we intend to extend our work by:

1. Implementing and evaluating situations where features may not be visible throughout all frames.
2. Implementing a RANSAC initialisation step or a similar algorithm to deal with and remove erroneous feature points and correspondences, and evaluate the effectiveness of the algorithm to handle erroneous feature matches.
3. Implementing a 'sliding windows' approach to make the algorithm applicable to online robotic scenarios.
4. Applying the algorithm on outdoor robotic datasets and comparing the results to traditional RANSAC-based visual pose estimation optimised with bundle adjustment.
5. Developing the algorithm on the GPU to further improve speed and applying the algorithm in an online robotic pose estimation scenario.

References

[Arun *et al.*, 1987] K S Arun, T S Huang, and S D Blostein. Least-Squares Fitting of Two 3-D Point Sets. *IEEE Trans. Pattern Anal. Mach. Intellig*, 9(5):698–700, 1987.

[Bay *et al.*, 2006] Herbert Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer VisionECCV 2006*, pages 404–417, 2006.

[Braid *et al.*, 2006] D Braid, A Broggi, and Gary Schmiedel. The TerraMax autonomous vehicle. *Journal of Field Robotics*, 23(October 2005):693–708, 2006.

[Comport *et al.*, 2010] A.I. Comport, E. Malis, and P. Rives. Real-time Quadrifocal Visual Odometry. *The International Journal of Robotics Research*, 29(2-3):245, January 2010.

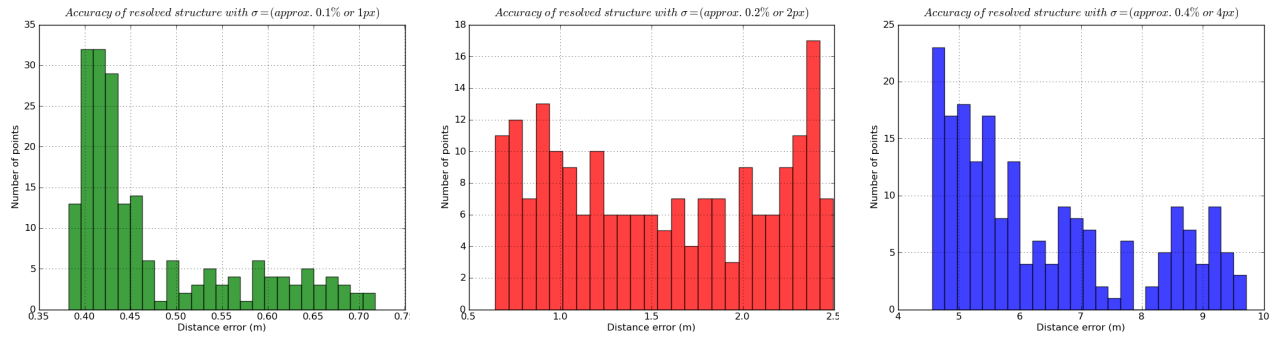

Figure 8: Histogram of error in reconstructed scene points during third experiment for a) 0.1% noise, b) 0.2% noise and c) 0.4% noise.

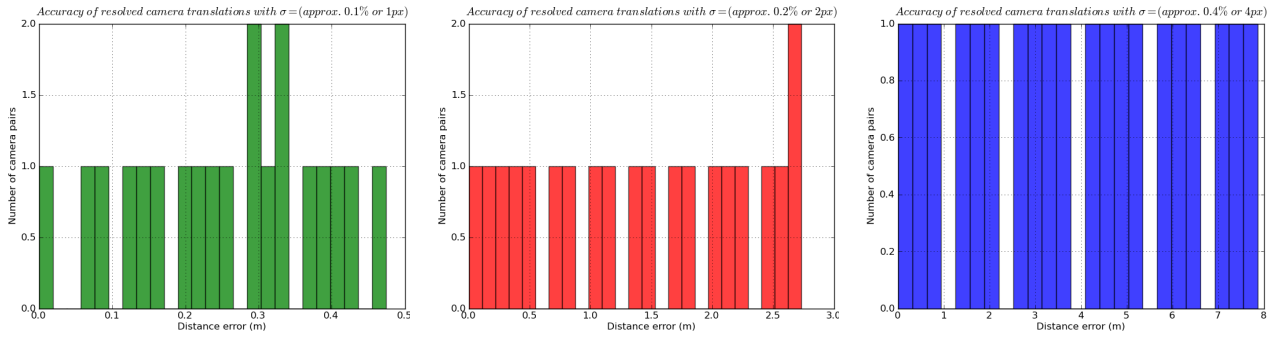

Figure 9: Histogram of error in camera translation during third experiment for a) 0.1% noise, b) 0.2% noise and c) 0.4% noise.

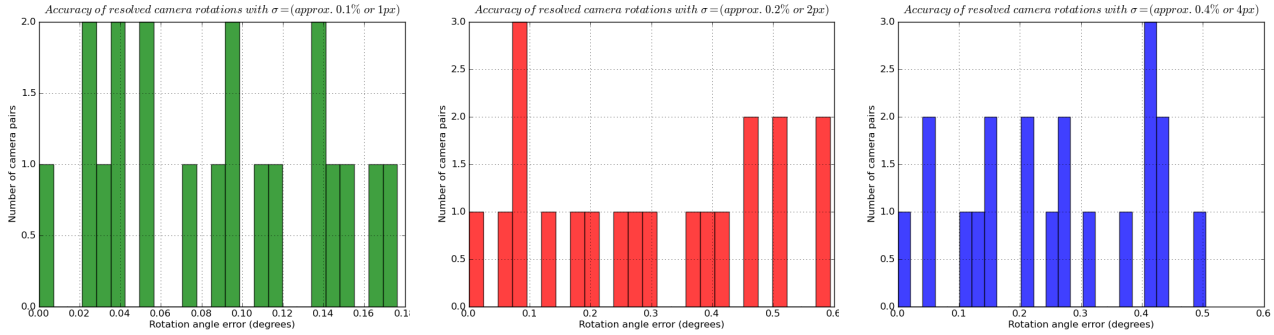

Figure 10: Histogram of error in camera rotation during third experiment for a) 0.1% noise, b) 0.2% noise and c) 0.4% noise.

[Engels *et al.*, 2006] C Engels, H. Stewénius, and D. Nistér. Bundle adjustment rules. *Photogrammetric Computer Vision*, 2, 2006.

[Fischler and Bolles, 1981] MA Fischler and RC Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[Grossberg and Nayar, 2001] MD Grossberg and SK Nayar. A general imaging model and a method for finding its parameters. *ICCV*, pages:1100–1105, 2001.

[Haralick *et al.*, 1994] B.M. Haralick, C.N. Lee, K. Ottenberg, and M. Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *International Journal of Computer Vision*, 13(3):331–356, 1994.

[Hartley and Zisserman, 2004] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[Horn, 1987] Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629, April 1987.

[Kalantari *et al.*, 2009] Mahzad Kalantari, Franck Jung, Jean-Pierre Guedon, and Nicolas Paparoditis. *The Five Points Pose Problem : A New and Accurate Solution Adapted to any Geometric Configuration*, pages 215–226. Springer Berlin / Heidelberg, 2009.

[Konolige and Agrawal, 2008] Kurt Konolige and Motilal Agrawal. FrameSLAM: from Bundle Adjustment to Realtime Visual Mappping. *Robotics, IEEE Transactions on*, 24(5):1066–1077, 2008.

[Konolige *et al.*, 2007] K. Konolige, M. Agrawal, and J. Sola. Large scale visual odometry for rough terrain. In *Proc. International Symposium on Robotics Research*. Citeseer, 2007.

[Konolige *et al.*, 2010] Kurt Konolige, James Bowman, JD Chen, Patrick Mihelich, Michael Calonder, Vincent Lepetit, and Pascal Fua. View-based maps. *The International Journal of Robotics Research*, 29(8):941–957, 2010.

[Lowe, 1999] D.G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, IEEE International Conference on*, volume pages, page 1150. Published by the IEEE Computer Society, 1999.

[Lu *et al.*, 2000] C.-P. Lu, G.D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):610–622, June 2000.

[Luong and Faugeras, 1996] Q.T. Luong and O.D. Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17(1):43–75, 1996.

[Nistér and Stewénius, 2007] D. Nistér and H. Stewénius. A minimal solution to the generalised 3-point pose problem. *Journal of Mathematical Imaging and Vision*, 27(1):67–79, 2007.

[Nistér *et al.*, 2006] David Nistér, Oleg Naroditsky, and James Bergen. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23(1):3–20, January 2006.

[Nistér, 2004] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 756–777, 2004.

[Nuske *et al.*, 2009] Stephen Nuske, Jonathan Roberts, and Gordon Wyeth. Localization Using a Three-Dimensional-Edge Map. *Journal of Field Robotics*, 15213(9):728–756, 2009.

[Schweighofer and Pinz, 2006] Gerald Schweighofer and Axel Pinz. Fast and globally convergent Structure and Motion estimation for General Camera Models. In *Proc. of BMVC*, pages 147–157, 2006.

[Schweighofer and Pinz, 2008] Gerald Schweighofer and Axel Pinz. Globally Optimal O(n) Solution to the PnP Problem for General Camera Models. *Proceedings of the 19th British Machine Vision Conference*, pages 1–10, 2008.

[Schweighofer *et al.*, 2008] Gerald Schweighofer, Sinisa Segvic, and Axel Pinz. Online/Realtime Structure and Motion for General Camera Models. *2008 IEEE Workshop on Applications of Computer Vision*, pages 1–6, January 2008.

[Sibley *et al.*, 2009] Gabe Sibley, Christopher Mei, Ian Reid, and Paul Newman. Adaptive relative bundle adjustment. In *Robotics Science and Systems Conference*, pages 1–8. Citeseer, 2009.

[Sturm and Ramalingam, 2004] P. Sturm and S. Ramalingam. A generic concept for camera calibration. *Computer Vision-ECCV 2004*, 54:1–13, 2004.

[Sturm, 2005] P. Sturm. Multi-View Geometry for General Camera Models. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1(1063-6919):206–212, 2005.

[Triggs *et al.*, 2000] B Triggs, P McLauchlan, and R Hartley. Bundle adjustment a modern synthesis. *Vision algorithms: theory*, pages:298–375, 2000.