ROBUST BUNDLE ADJUSTMENT FOR STRUCTURE FROM MOTION

Ji Zhang, Mireille Boutin, Daniel G. Aliaga

Dept. of Mathematics, School of ECE, Dept. of Computer Science
Purdue University, West Lafayette, IN, USA

ABSTRACT

Structure from motion (SFM) is the problem of reconstructing the geometry of a scene from a stream of images. In this problem, the geometry of the scene must be inferred from images, along with the camera pose parameters. Bundle Adjustment (BA) is a refinement method used to improve SFM solutions. It consists in simultaneously improving a set of initial estimates for all parameters (structure and camera pose) by minimizing a global cost function. It is generally considered to be highly accurate, and so is typically used as a last refinement step in most current SFM methods. Unfortunately, estimating the pose of the camera from a stream of images is an ill-conditioned problem. We thus propose a BA adjustment formulation which does not involve solving for the camera orientations. We tested this approach on several real world models. The numerical results obtained show that this approach is much less affected by noise than traditional BA.

1. INTRODUCTION

Reconstructing the geometry of a 3D scene from a set of images is an important aspect of image processing, computer vision, and the simulation of 3D environments. A popular way to obtain the geometry of a scene is to acquire and process images obtained by a camera. The problem of reconstructing the 3D geometry of a scene from a set of pictures is called *structure from motion* (SFM). An important case is when the external camera parameters are unknown (i.e. uncalibrated), since accurately measuring these requires a complex and expensive setup.

Unfortunately, SFM with an externally uncalibrated camera is a very difficult problem. Despite decades of research, a satisfying solution still has not been found. Why is SFM so difficult? One main reason is that when the external camera calibration is unknown, the problem of pose estimation is naturally embedded in SFM. The camera parameters constitute a nuisance because they negatively impact the robustness of the reconstruction. Indeed, it has been shown that estimating the pose of a camera is an ill-conditioned problem [1]. This is due to an inherent confusion between the camera position and the camera orientation which simply cannot be resolved based on the pictures alone, regardless of the solution scheme used. Consequently, numerical instabilities are observed in the reconstruction, since the geometric structure of a scene is linked to the camera pose estimates in a highly unstable manner.

The most popular solution available to improve the results of current SFM methods is called *bundle adjustment* (BA). BA is a refinement technique for SFM. It takes as input an imperfect solution for the camera pose and 3D position of features of the scene and refines this solution by minimizing a cost function based on the difference between the projection of the feature points and the tracked features on the images. But in this approach, the camera pose parameters are still being solved for, which leads to numerical instabilities.

In order to improve the numerical results of SFM, we thus propose a new formulation of BA which does not involve solving for the camera orientation. This formulation is obtained using a computational technique from invariant theory [2] which, under certain circumstances, can be used to eliminate variables from a set of equations [3]. By eliminating the camera orientation parameters from the standard SFM equations, we obtain a new set of SFM equations which does not involve any rotation matrix. This new set of equations naturally leads to the formulation of a rotation matrix free cost function to be minimized by BA. From a theoretical point of view, one naturally expects that eliminating the rotation matrices from the cost function will lead to better numerical results than standard BA. We demonstrate this experimentally using several real-world models.

2. PREVIOUS WORK

The BA method was first proposed by Brown in the context of photogrammetry [4]. It was later popularized in the computer vision community by Hartley [5] and Triggs et al. [6]. BA consists in solving for all the scene structure and the camera parameters simultaneously by minimizing a cost function equal to the sum of the squares of the distances between the reprojections of the 3D reconstructed points and the observed projections. The term *bundle* refers to the rays of light joining the camera centers and the scene points which are being adjusted to minimize the cost function. The minimization is performed numerically using a non-linear least squares method, which is typically chosen to be the Levenberg-Marquardt minimization algorithm. But since the number of variables to be optimized is enormous, this minimization is very costly. Fortunately, the problem exhibits a sparse structure which can be exploited to speed up the computations. BA methods which exploit this sparsity are called *sparse bundle adjustment methods* or SBA. Given a good initial guess, BA is highly accurate, much more so than any other SFM method currently available. So, in practice, BA is almost always applied to the results obtained with other methods, as a last refinement step. But as we mentioned previously, the numerical problems created by the need to estimate the pose are seen in this approach as well, since the camera pose parameters are an intrinsic part of the equations to be minimized.

The idea of eliminating the camera pose parameters to improve numerical stability has been exploited in other contexts. For example, in some early work Tomasi and Shi [7] proposed some SFM equations where the camera orientation parameters do not appear by considering the camera rays. They used these equations to compute the direction of heading of the camera. Numerical experiments demonstrated the robustness of this approach. Similarly, Tomasi [8] described the image changes through the angles between the projection rays and showed how these can be used to reconstruct both

This work was partly supported by NSF grant 0434398

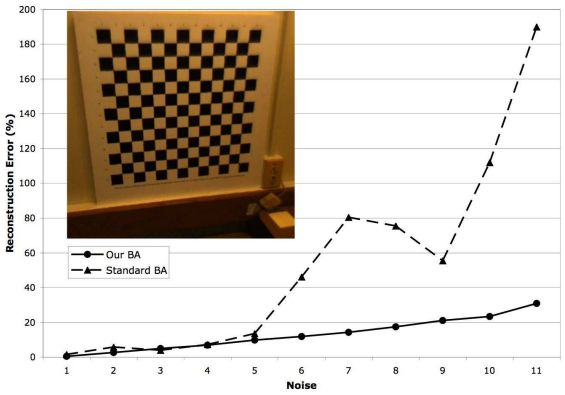

Fig. 1. **Reconstruction Comparison for Structure and Motion Optimization.** Using a chessboard equipped with a mechanically tracked arm, we obtained a ground truth reconstruction and compared it with the results of BA using cost function (4) and a standard SBA.

structure and motion in a two-dimensional world. Immunity to noise of this method was also noted in experiments, although the results were observed to be critically dependent on camera calibration.

More than ten years have passed since the publication of Tomasi and Shi's work and still no complete mathematical framework for SFM without camera parameters has been developed. Why has the idea of pose parameter elimination never been exploited to its full extent? In particular, why was a BA without camera pose parameters never proposed. This is probably due to the complexity of the problem of variable elimination. Indeed, algebraically eliminating variables from a set of equations is difficult, especially when the number of unknowns in the equations is high, as is the case here. In particular, we still have not been able to completely eliminate the camera pose parameters from the SFM equations, although this is the subject of ongoing research. However, Bazin and Boutin [3] recently suggested a simple procedure to algebraically remove the camera orientation parameters from the SFM equations. They also illustrated the use of their method in a few different settings. In the next section, we use a variation of their approach to obtain a camera orientation free formulation of SFM leading to a camera orientation free BA. Our SFM formulation is different than the ones originally proposed by Bazin and Boutin. In particular, an assumption made on the fourth coordinates used to parameterize the projective space greatly simplifies the resulting equations, which will be stated in the next section.

3. A NEW FORMULATION OF BUNDLE ADJUSTMENT

We are interested in reconstructing the geometry of a scene observed by an image stream. More precisely, we want to determine the 3D positions of the tracked features from their observed positions on the pictures. The equations relating the tracked features and their projections on the pictures can be written as follows:

$$c_{ij} F_j \begin{pmatrix} P_i \\ 1 \end{pmatrix} - \begin{pmatrix} p_{ij} \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \qquad (1)$$

where p_{ij} represents the 2D coordinates of the 3D feature point P_i observed on picture j, c_{ij} is a constant, and F_j is a 3-by-4 matrix containing the camera parameters corresponding to picture j. Let us assume that the index i takes values from 1 to N, where N is the number of features tracked on the image, and that the index j takes values from 1 to J, where J is the number of pictures taken. When the camera is internally calibrated, one can assume that the matrix F_j takes the form $F_j = \begin{pmatrix} R_j & t_j \end{pmatrix}$, where R_j is a 3D rotation matrix and t_j is a 3D translation vector.

Because of the presence of noise in the tracked feature measurements, it is impossible to solve these equations simultaneously for all i's and j's. All we can hope is to try to make the left-hand side of all these equations *close* to a vector of zeros. In BA, this is formulated as a least squares problem: one demands that the sum of the squares of all the left-hand side expressions (the cost function) be as close as possible to zero. The camera pose parameters are part of the cost function through the matrix F_j. In order to formulate a better cost function, we propose to eliminate at least part of the F_j matrix parameters from Equations 1. Since the most troublesome parameters are the rotation matrices (e.g. camera rotation errors have significant effect on the 3D position of distant points), we concentrate on eliminating these.

Eliminating parameters from the SFM equations is not straightforward, even though they are polynomial. Indeed, one could think that the symbolic elimination tools developed for the case of polynomial equations (e.g., Singular [9]) would be well suited for eliminating the nuisance parameters in this case. Unfortunately, the set of equations we are dealing with is so big and involves so many variables that these programs cannot handle the size of the problem. Also, by restricting ourselves to polynomial functions, we are likely to end up with equations of a higher polynomial degree than we began with. Indeed, since division by a variable is not allowed, the more variables are eliminated, the more the degrees of the polynomials in the basis tend to increase. This approach to variable elimination thus has the undesired likely potential of increasing the complexity and the numerical instability of the problem.

In contrast with commutative algebraic approaches, the computational approaches developed in the context of differential geometry are not restricted to polynomial equations. The one we used to eliminate the camera orientation follows a systematic approach suggested in [3]. The idea is to view the problem of camera orientation elimination as a problem in invariant theory: the camera orientation parameters are seen as Lie group parameters acting on the other unknowns of the problems. To obtain a basis for all SFM equations which do not involve the camera orientation parameters, one simply needs to compute a basis for the invariants of this group action. This is accomplished by using a modern version of the moving frame method developed by Fels and Olver [2]. This method consists in finding an analytic expression for the group transformation $g = \rho(x)$ which will map any given point x back into a pre-determined canonical position. The basis of invariants naturally appear in the coordinates of the point x transformed by $\rho(x)$. Readers should refer to [10] for a more detailed, yet accessible introduction. Using this method, we obtained a set of equations equivalent to Equations 1, but which does not involve any rotation matrices. This set of equations forms a *basis* for all camera orientation free SFM equations, in the sense that any other SFM equation can be written (locally) as a function of these equations. This basis of equations can be expressed as

$$\begin{aligned} (P_i - C_j) \cdot (P_1 - C_j) - \gamma_{ij}\gamma_{1j}k_{1ij} &= 0, \\ (P_1 - C_j) \times (P_i - C_j) \cdot (P_1 - C_j) \times (P_2 - C_j) - \gamma_{ij}\gamma_{1j}^2\gamma_{2j}k_{2ij} &= 0, \\ (P_i - C_j) \cdot (P_1 - C_j) \times (P_2 - C_j) - \gamma_{ij}\gamma_{1j}\gamma_{2j}k_{3ij} &= 0, \end{aligned} \qquad (2)$$

where the values of the constants k's are given by the coordinates of the tracked features on the pictures as

$$\begin{aligned}\left(p_{ij}^T,1\right)\cdot\left(p_{1j}^T,1\right) &= k_{1ij},\\ \left(p_{1j}^T,1\right)\times\left(p_{ij}^T,1\right)\cdot\left(p_{1j}^T,1\right)\times\left(p_{2j}^T,1\right) &= k_{2ij},\\ \left(p_{ij}^T,1\right)\cdot\left(p_{1j}^T,1\right)\times\left(p_{2j}^T,1\right) &= k_{3ij},\end{aligned}\quad(3)$$

for $i = 1,\ldots,N$ and $j = 1,\ldots,J$. (Note that the first equation is always trivial for $i = 1,2$ and the second equation is also trivial for $i = 1$.) By construction, all the equations of 2 are invariant under a simultaneous rigid motion of all the camera centers and 3D feature points. They are also independent of the coordinate system used for the images. However, they are not invariant under a relabeling of the feature points, but relabeling invariant equations can be obtained by taking simple linear combinations of functions of the above equations.

The only unknowns in these equations are the camera center positions for each picture C_j, $j = 1,\ldots,J$, the 3D features positions P_i, for $i = 1,\ldots,N$, and the parameters γ_{ij}, for all i,j, (which are related to the depth disparity values.) There are thus $3N + 3J + NJ$ unknowns. The number of non-trivial equations is $3J(N-1)$. But the number of functionally independent equations is actually $3J(N-1)-6$ (because of the invariance under rigid motion.) So for N and J big enough, we can attempt to solve these equations, or a subset of these equations, numerically.

Starting from these equations, there are several ways to formulate a rotation free cost function to be minimized for BA. For example, since the right-hand side of each equation is equal to zero, we could sum up the square of all the left-hand sides. But this would lead to a computationally intensive BA, as very many variables would need to be optimized at the same time. However, the sparse structure of the relationships between the variables could be used to obtain an SBA method with decreased complexity. This is the subject of ongoing research. In our experiments, we used potentially less accurate but less computationally intensive methods. For example, we solved for the position of the 3D features by dividing them into sextuplets of points and considering the projection of these six points onto three pictures. Assuming that the index of the six points considered are $i = 1,2,3,4,5,6$ and that the pictures considered are $j = 1,2,3$, the cost function takes the form

$$\begin{aligned}&\sum_{j=1}^{3}\sum_{i=1}^{6}\left[(P_i-C_j)\cdot(P_1-C_j)-\gamma_{ij}\gamma_{1j}k_{1ij}\right]^2+\\ &\left[(P_1-C_j)\times(P_i-C_j)\cdot(P_1-C_j)\times(P_2-C_j)-\gamma_{ij}\gamma_{1j}^2\gamma_{2j}k_{2ij}\right]^2\\ &\quad+\left[(P_i-C_j)\cdot(P_1-C_j)\times(P_2-C_j)-\gamma_{ij}\gamma_{1j}\gamma_{2j}k_{3ij}\right]^2.\end{aligned}\quad(4)$$

Using a Levenberg-Marquadt algorithm, we can minimize this cost function by varying the values of the unknowns 3D points P_i, for $i = 1,\ldots,6$, the camera centers C_j, for $j = 1,2,3$, and the parameters γ_{ij}, for $i = 1,\ldots,6$ and $j = 1,2,3$. But using this formulation for the cost function requires that the sextuplets of points be chosen so that they can all be seen on all three pictures. Intuitively, the more different the projections of these six points looks on the three picture, the better. This requires a pre-processing step where the sextuplets are formed based on their different observations on common pictures. This pre-processing step is significantly less work than evaluating a global cost function using all the points and all the pictures.

To speed up the computations, we formulated an alternative cost function requiring only pairs of points observed in six images, namely

$$\begin{aligned}&\sum_{j=1}^{3}\left[|P_1-C_j|^2-\gamma_{1j}^2k_{11j}\right]^2+\left[|P_2-C_j|^2-\gamma_{2j}^2k_{22j}\right]^2+\\ &\qquad\left[(P_2-C_j)\cdot(P_1-C_j)-\gamma_{2j}\gamma_{1j}k_{12j}\right]^2.\end{aligned}\quad(5)$$

This was done by considering the first equation of 2 with $i = 1$ and 2. The two resulting equations were complemented with a third equation which is the analogue of the first equation with P_1 replaced by P_2 (which holds by symmetry.) For simplicity, we assume that the camera center position is equal to the initial guess, and only optimize the remaining parameters (P_i's, γ_{ij}'s) based on the coordinates of their projections on three pictures.

4. NUMERICAL EXPERIMENTS

The following experiments demonstrate that removing the camera orientation from the cost function leads to a BA which is less sensitive to variations in the initial guess than standard BA. Our numerical experiments were done on streams of images with a set of features tracked using the Kanade-Lucas-Tomasi feature tracking package [11]. Overall, we see a clear improvement of our method over traditional BA as the error in the initial guess increases.

Our first test is to optimize the structure and motion of a captured model. To be able to accurate quantify the methods, the chessboard dataset was captured with an in-house acquisition system using a mechanically tracked arm (Microscribe Arm G2LX manufactured by Immersion Corp.) to obtain very precise measurements of the camera pose (fraction of a degree precision) and chess board position (millimeter precision) to be used as ground truth in this experiment. A total of 48 images and 96 features were used in this example. These were divided into 16 groups of sextuplets seen on three images. The cost function was minimized group by group. We compared the 3D reconstruction error (using Euclidean norm) of our approach with that of a publicly available sparse bundle adjustment method [12], for various amounts of error in the initial guess. The results are plotted in Figure 1. The figure illustrates reconstruction error, a percentage of the model radius, as a function of randomly-added noise. The magnitude of the noise for camera center and point estimates varies from 0 to 40% of the model diagonal and the magnitude of the noise for camera rotation ranges from 0 to 12 degrees. These results show that the cost function (4) leads to a better refinement than standard BA, as the error in the initial guess increases.

Figure 2 represents the performance of the cost function (5) compared to standard BA, for increasing amounts of error in the initial guess. The axis are similar to Figure 1. The noise for point estimates varies from 0 to 25% and the camera rotation ranges from 0 to 12 degrees. Again we used [12] to obtain the result of SBA but optimizing only for structure. Even though our version of BA does not optimize the camera center positions and only optimizes small groups of features at the time, it yields a significant improvement. In fact, for comparison purposes, we show the performance of full optimization SBA and observe that we are able to slightly outperform the full optimization.

A more visual illustration of the results of the our experiments is given in Figure 3. In this experiment, we show reconstructions using various amounts of randomly-added noise. The original model is shown in (a). A reconstruction obtained by our method using the tracked features and initial guesses obtained via simple triangulation of the 563 feature points is shown in (b). The middle row represents the giraffe reconstruction obtained with standard SBA using [12], with increasing amounts of error in the initial guess. The bottom row represents the results obtained with cost function (5), for the same initial guesses. In each picture, we draw lines from the points of the best reconstruction to the current reconstructed set. Thus, larger reconstruction errors are clearly visible as longer and much abundant lines appear in the picture. Visually, our method is able to tolerate larger errors in the initial estimates. While the visual aspect of the giraffes in bottom row are better than the middle one,

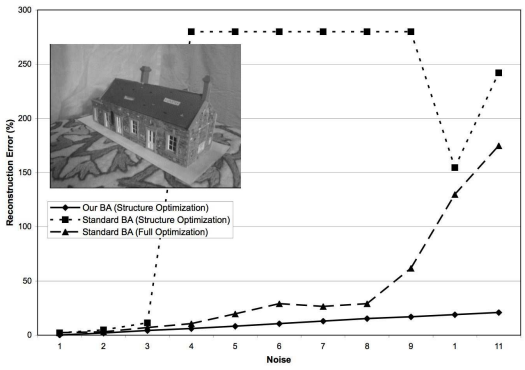

Fig. 2. **Reconstruction Comparison for Structure Optimization.** Using the house data, we optimize the structure using both SBA and our method. Overall, we achieve a clear improvement (even compared to SBA optimizing both structure and motion).

the number of outliers is also a clear indication of the better quality of our reconstruction over standard BA. Here, outliers are defined as the reconstructed points which differ from the ground truth by more than 30% of the radius of the world space.

5. CONCLUSION

We proposed a basis of equations for formulating an improved BA cost function. These involve less unknowns than the ones used in the standard formulation of BA. More precisely, the camera orientation parameters, which are the most problematic parameters in SFM, have been eliminated by algebraic manipulation. Using these equations, we formulated a simple, rotation matrix free cost function for BA. We presented numerical experiments demonstrating that this cost function leads to a BA which is more resilient to errors in the initial guess than standard BA. The improvement in the result is clear, even when the optimization is performed locally on sextuplets of points and only considering three pictures at a time. Structure-only computations can be significantly accelerated by considering a modified cost function involving only pairs of points and six pictures at a time. By assuming the camera center estimate is correct (and ignoring any camera rotation estimates), the numerical results can be obtained very quickly. Even with this simplified approach, a better BA is obtained compared with the standard approach. We thus conclude that removing the camera orientation parameters from the cost function leads to a BA which is more robustness to errors in the initial guess. In future work, we will attempt to improve the numerical results obtained by performing a global optimization of all the points and all the pictures simultaneously, while taking advantage of the sparse structure of the problem to speed up the computations.

6. REFERENCES

[1] Cornelia Fermüller and Yiannis Aloimonos, "Observability of 3D motion," *Int. J. Comput. Vision*, vol. 37, no. 1, pp. 43–63, 2000.

[2] Mark Fels and Peter J. Olver, "Moving coframes. I. a practical algorithm," *Acta Appl. Math.*, vol. 51, pp. 161–213, 1998.

[3] Pierre-Louis Bazin and Mireille Boutin, "Structure from motion: a new look from the point of view of invariant theory," *SIAM J. Appl. Math.*, vol. 64, no. 4, pp. 1156–1174, 2004.

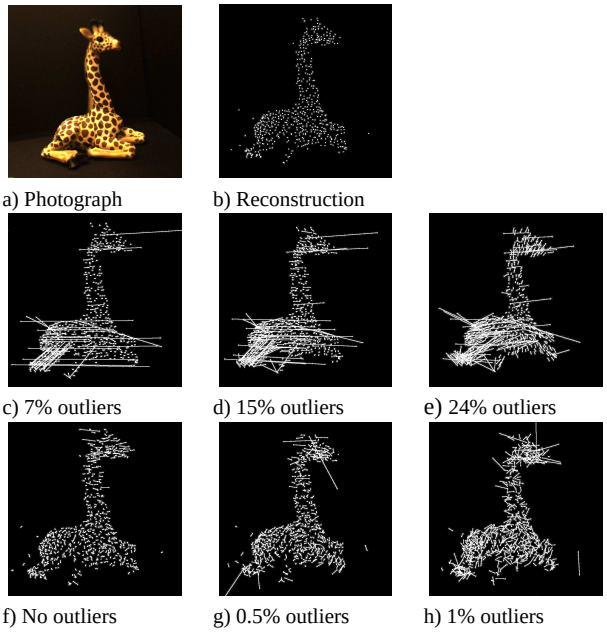

Fig. 3. **Visual Comparison of Initial Guess Sensitivity.** (a) Photograph of object; (b) Reconstruction of object; Reconstructions using small, medium, and large amounts of error in initial estimates for SBA (c,d,e) and our method (f,g,h).

[4] D.C. Brown, "A solution to the general problem of multiple station analytical stereotriangulation," Tech. Rep. 43, Patrick Airforce Base, Florida.

[5] Richard I. Hartley, "Euclidean reconstruction from uncalibrated views," in *Proc. of the Second Joint European - US Workshop on Applications of Invariance in Computer Vision*, London, UK, 1994, pp. 237–256, Springer-Verlag.

[6] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon, "Bundle adjustment - a modern synthesis," in *ICCV '99: Proc. of the International Workshop on Vision Algorithms*, London, UK, 2000, pp. 298–372, Springer-Verlag.

[7] C. Tomasi and J. Shi, "Direction of heading from image deformations," in *CVPR93*, 1993, pp. 422–427.

[8] C. Tomasi, "Pictures and trails: A new framework for the computation of shape and motion from perspective image sequences," Los Alamitos, CA, USA, June 1994, pp. 913–918, IEEE Computer Society Press.

[9] V. Levandovskyy G.-M. Greuel and H. Schönemann, "SINGULAR::PLURAL 2.1," A Computer Algebra System for Noncommutative Polynomial Algebras.

[10] Peter J. Olver, *Classical invariant theory*, vol. 44 of *London Mathematical Society Student Texts*, Cambridge University Press, Cambridge, 1999.

[11] Jianbo Shi and Carlo Tomasi, "Good features to track," Tech. Rep., Ithaca, NY, USA, 1993.

[12] M.I.A. Lourakis and A.A. Argyros, "The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm," Tech. Rep. 340, Institute of Computer Science - FORTH, Heraklion, Crete, Greece, Aug. 2004.