Covariance Propagation and Next Best View Planning for 3D Reconstruction

Sebastian Haner and Anders Heyden

Centre for Mathematical Sciences
Lund University, Sweden
{haner,heyden}@maths.lth.se
http://www.maths.lth.se

Abstract. This paper examines the potential benefits of applying next best view planning to sequential 3D reconstruction from unordered image sequences. A standard sequential structure-and-motion pipeline is extended with active selection of the order in which cameras are resectioned. To this end, approximate covariance propagation is implemented throughout the system, providing running estimates of the uncertainties of the reconstruction, while also enhancing robustness and accuracy. Experiments show that the use of expensive global bundle adjustment can be reduced throughout the process, while the additional cost of propagation is essentially linear in the problem size.

Keywords: Structure and motion, covariance propagation, next best view planning.

1 Introduction

Three-dimensional reconstruction from unordered image sequences is a well-studied problem in the computer vision literature, see e.g. [1,2,3,4,5]. Part of the challenge is that little is known about the input data at the outset in terms of scene coverage or camera calibration. Active sensor planning, on the other hand, is the problem of finding the optimal input data to a reconstruction algorithm, given full control over image acquisition (see [6] for an overview). In the photogrammetry literature this is known as the 'camera network design' problem. For example, in [7] a genetic algorithm is used to search a high-dimensional parameter space of camera placements to find the optimal measurement setup, given a limited number of cameras. In a serial acquisition process, the 'next best view' (NBV) problem asks from which viewpoint to capture the next image, given a partial reconstruction, to minimize some objective such as the reconstruction error. NBV planning is most effective when the user has full control over image acquisition, and has been applied to vision metrology using cameras mounted on robotic arms [8,9], and autonomous robot exploration [10,11].

This paper applies view planning to the unordered image reconstruction problem; although we are not free to choose any viewpoint, there is usually a choice between a subset of the images at every step of a sequential algorithm. The

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 545–556, 2012.

aim is to choose the image giving the smallest error, which we approximate as the trace of the camera covariance matrix times the reprojection error. To be able to determine the covariance, it is necessary to know the uncertainty of the observed geometry. In the following sections, it is shown how this is achieved by propagating covariances when resectioning cameras and triangulating points, and how as a side effect the algorithms gain robustness and better approximate the maximum likelihood estimate.

2 Estimation from Uncertain Geometry

The cornerstones of sequential structure-and-motion are triangulation and camera pose estimation. Usually, one attempts to find the maximum likelihood solution given noisy image measurements, but assuming that all other parameters are known exactly. This is of course rarely the case, since points and cameras are triangulated and resectioned using noisy data. Below, we derive algorithms that also take the uncertainty of the 3D structure or camera parameters into account.

2.1 Pose Estimation

Consider the problem of camera pose estimation given N 3D point coordinates X and their measured projections in one image, $\tilde{x}$. Assuming there are errors in the image measurements, the problem is to find the maximum likelihood solution, i.e. the camera parameters θ^* satisfying

$$\theta^* = \arg\max_{\theta} \ \mathcal{L}(\theta)\,, \tag{1}$$

where

$$\mathcal{L}(\theta) = \mathcal{L}(\theta \,|\, \tilde{x}, X) = p(\tilde{x} \,|\, \theta, X) \tag{2}$$

is the likelihood function. In this formulation it is assumed that the structure parameters X are precisely known. More generally, given a probability distribution of X, the problem is to maximize

$$\mathcal{L}(\theta) = \int_{\mathbb{R}^{3N}} p(\tilde{x} \,|\, \theta, X) p(X)\, dX\,. \tag{3}$$

We restrict our attention to the case of Gaussian distributions. Then we have

$$\mathcal{L}(\theta) \propto \int_{\mathbb{R}^{3N}} e^{-\|\tilde{x} - f(X,\theta)\|^2_R} \cdot e^{-\|X - \bar{X}\|^2_Q}\, dX\,, \tag{4}$$

where $f(X, \theta)$ is the projection of the points X using camera parameters θ, R the measurement error covariance, Q and $\bar{X}$ are the covariance matrix and mean of the distribution of X and $\|y\|^2_\Sigma = y^\top \Sigma^{-1} y$ the squared Mahalanobis distance. Next, we project the distribution of X onto the image plane, by integrating along

the light rays. Formally, for a given θ we parametrize each 3D point by its image projection $x = f(X, \theta)$ and depth ρ, so that

$$\mathcal{L}(\theta) \propto \int_{\mathbb{R}^{2N}} e^{-\|\tilde{x}-x\|^2_R} \left(\int_{\mathbb{R}^N} e^{-\|(x,\rho)-\bar{X}\|^2_Q} \, d\rho \right) dx \,. \tag{5}$$

The right-hand factor is a distribution on the $2N$-dimensional generalized image plane, and may be seen as the projection of a random variable, i.e. $f\big(\mathcal{N}(\bar{X}, Q), \theta\big)$. By Taylor expansion about $\bar{X}$, f can be approximated by $\tilde{f}(X, \theta) = f(\bar{X}, \theta) + J(X - \bar{X})$, and for affine functions $\tilde{f}\big(\mathcal{N}(\mu, \Sigma), \theta\big) = \mathcal{N}\big(\tilde{f}(\mu, \theta), J_X \Sigma J_X^\top\big)$ with $J_X = \frac{\partial f}{\partial X}\big|_\theta$. We now have

$$\mathcal{L}(\theta) \overset{\propto}{\sim} \int_{\mathbb{R}^{2N}} e^{-\|\tilde{x}-x\|^2_R} \cdot e^{-\|f(\bar{X},\theta)-x\|^2_{JQJ^\top}} \, dx \,, \tag{6}$$

which is just the convolution $\mathcal{N}(\tilde{x}, R) * \mathcal{N}(0, JQJ^\top) = \mathcal{N}(\tilde{x}, R + JQJ^\top)$. Maximizing

$$\mathcal{L}(\theta) \overset{\propto}{\sim} e^{-\|\tilde{x}-f(\bar{X},\theta)\|^2_{R+JQJ^\top}} \tag{7}$$

is then equivalent to minimizing

$$-\log \mathcal{L}(\theta) \overset{\propto}{\sim} \|\tilde{x} - f(\bar{X}, \theta)\|^2_{R+JQJ^\top} \,, \tag{8}$$

which can be solved using an iteratively reweighted nonlinear least-squares algorithm. In fact, only a minor modification to a standard algorithm for minimizing the reprojection error is required. For example, a Levenberg-Marquardt optimization loop would be modified to

while not converged **do**
  ...
  $W \leftarrow (R + J_X Q J_X^\top)^{-1}$
  $\delta\theta \leftarrow (J_\theta^\top W J_\theta + \lambda I)^{-1} J_\theta^\top W b$
  ...
end while

where $J_\theta = \frac{\partial f}{\partial \theta}\big|_\theta$ and J_X as above. After convergence, the covariance matrix of the recovered camera parameters θ^* can be estimated by the inverse of the Hessian matrix evaluated at the minimum, $\Sigma_\theta \approx (J_{\theta^*} W^* J_{\theta^*}^\top)^{-1}$ [12,13].

Of course, a good initial guess is required to start the iterative algorithm, and can be obtained using standard minimal or linear solvers. The general effect of taking the distribution of X into account is to give more weight to well-determined 3D points than uncertain ones when finding the camera pose.

2.2 Triangulation

Handling uncertainty in camera parameters when triangulating 3D structure is completely analogous to the pose estimation case. The linearized problem formulation is to find

$$\theta^* = \arg\min_\theta \|\tilde{x} - f(\theta, \bar{P})\|^2_{R+JSJ^\top} \,, \tag{9}$$

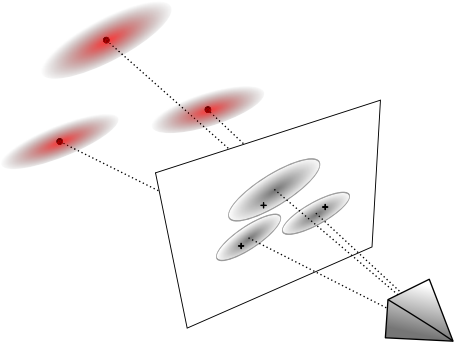


Fig. 1. Resectioning: the uncertainties of the 3D points are projected onto the image plane and convolved with the image measurement uncertainty giving the reprojection error metric. Note that the projections are not necessarily independent; however, in this work inter-point covariances are discarded for computational reasons.

where θ now represents the 3D structure, $\bar{P}$ is the mean of the distribution of the cameras with covariance S and $J = \frac{\partial f}{\partial P}\big|_{\theta}$.

2.3 Complexity

The introduction of the weight matrix W in the algorithms above inevitably incurs extra computational costs. In particular, if the input variables are correlated, W will be a full matrix and the natural sparsity of the problems is lost. To mitigate this, we will assume no correlation between pairs of cameras or points, so that W is block diagonal. Such simplification is also necessary since the full covariance matrix of even a moderately sized reconstruction problem would occupy hundreds of gigabytes of memory. Furthermore, it may not be necessary to recompute W every iteration, since the projection is not expected to change significantly given a good initialization.

3 Covariance Propagation

The proposed algorithms open the possibility of covariance propagation throughout the reconstruction process. Uncertainties in 3D points are transferred to uncertainty in resectioned cameras, which in turn transfer uncertainty to triangulated points, and so on. In this manner, a rough estimate of the covariances is available at any time and can be used, for example, to improve reconstruction accuracy and for next best view planning, which we exploit to reduce error accumulation.

Below we detail a system for 3D reconstruction from unordered image sequences and show the benefits that can be gained.

3.1 Selecting the Seed

In choosing the set of images on which to initialize the reconstruction, we strive for the following: the initial reconstruction should be stable, contain many structure points and it should be near the center of the camera graph (the graph with each camera a vertex and edges between cameras observing common features). The latter is motivated by the fact that error accumulation is a function of the distance from the seed; if the 'degrees of separation' from the seed is kept low, error accumulation can be minimized. We therefore wish to minimize the distance of every camera to the seed. For our purposes we define the center as any vertex of the camera connectivity graph with minimal *farness*, the sum of shortest distances from the node to all others. We define the edge weights of the graph as $1/\max(0, n_c - 4)$, where n_c is the number of observed points common to both cameras. This heuristic, while ignoring the actual two-view geometry, is based on the assumption that cameras sharing many observed points are well-determined relative to each other. The maximum imposes a 5 point overlap threshold, needed to determine relative motion between views. Now, all shortest paths in the graph can be computed and summed for each camera, the k lowest scoring yielding a set of candidate images. For each candidate, an adjacent view with a balance between many common image points and good parallax is selected as in [1], i.e. each pairing is scored according to the proportion of outliers to a homography fit. The top-scoring pair is selected, and standard two-view reconstruction is performed, followed by bundle adjustment.

In experiments, the effect of choosing a seed near the center of the graph turns out to be relatively small, so this step is not essential.

3.2 Fixing the Gauge

Reconstruction from image measurements only is subject to global translation, rotation and scale ambiguity. Unlike [14], which measured pairwise covariances in local coordinate systems, we need globally referenced covariances and so must compute these for the seed reconstruction. For the covariances to be defined we must fix the gauge, especially the scale, since the dimension of the nullspace of the Hessian matches the number of degrees of freedom of the system. From a theoretical standpoint, taking the pseudoinverse of the unconstrained Hessian is the most satisfying solution [12,13], however it can be computationally very expensive if the seed views share many points (i.e. > 1000). An alternative approach is to constrain the parameters of the system by adding penalties to the cost function, making the Hessian full rank so it can be inverted without finding an SVD. Different constraints lead to somewhat different estimates of the covariance; one way is to lock the first camera and impose a distance constraint on the mean of the structure points, as was done in [14], or one can simply fix the distance between the first and second camera. The first prior gives results closer to the pseudoinverse, but also destroys the sparsity of the Hessian matrix making inversion more expensive. In cases where the pseudoinverse is too expensive we choose the second option which preserves sparsity.

After fixing the scale, there is still a difficulty in quantifying just how large an uncertainty is, since it must be put in relation to the overall size of the reconstruction. The scale is unknown in the beginning, since there is no guarantee that the distance between the seed cameras is representative of the whole scene. This has implications for the various outlier rejection thresholds used in the reconstruction pipeline.

4 Next Best View Planning

View planning in a sequential reconstruction process aims to actively choose the most favorable sensor configuration (i.e. camera position and orientation) to achieve a certain goal, in this case geometric accuracy. In each iteration, we can choose which camera to resection among those observing a sufficient number of triangulated points. Usually, the camera observing the largest number of triangulated points is chosen first. However, if the geometry is such that the pose is poorly determined, triangulations using the image will have larger errors, propagating to subsequently resectioned cameras, etc. It therefore makes sense to minimize the error accumulation in every step. To this end, we propose to select the camera with lowest estimated reconstruction error, by exhaustive search among candidate images. The covariance is computed by first resectioning the camera using a linear or minimal solver and taking the inverse of the Hessian, $\Sigma_{\text{cam}} \approx (J_\theta W J_\theta^\top)^{-1}$ as defined in section 3. As a scalar measure of reconstruction error we use $\text{trace}(\Sigma_{\text{cam}}) \cdot \epsilon_{\text{rp}}$, where ϵ_{rp} is the mean reprojection error. This turns out to give better results than the covariance alone; a small estimated covariance does not necessarily imply a low reprojection error, and a well-determined camera should ideally have both. Note that the score can be cached for each camera between iterations and need only be recomputed if more points in the camera's view have been triangulated. While the number of views that need to be resectioned in each iteration is dependent on the particular data set and could theoretically grow with the number of triangulated points, in practice this number is found to be approximately constant throughout the reconstruction process and typically between 10 and 50.

5 Reconstruction Pipeline

We apply NBV planning and covariance propagation to the problem of reconstruction from unordered image collections. We will assume that matching and tracking of image features has been performed and is outlier free. If not, the proposed method is easily integrated with outlier detection schemes such as RANSAC. The algorithm is mainly standard fare:

1. Find initial seed views (section 3.1).
2. Reconstruct and bundle adjust the seed geometry.
3. Compute the covariance of the seed (section 3.2).

4. Choose a camera to resect following section 4. Resect using a linear method; if it fails (i.e. large reprojection error or points behind the camera) try an L_∞ formulation [15] instead. If that also fails, choose another camera and try again. Else, refine the camera pose by minimizing (8) and store its covariance.
5. Triangulate all points seen by at least two resectioned cameras using a linear method. Compute an approximate uncertainty by evaluating the Hessian of the standard reprojection error and taking the trace of the inverse. Well-determined points, i.e. with low covariance and reprojection error, as specified by thresholds, are kept and further refined by minimizing (9). Store the covariance derived from this reprojection error.
6. If possible, goto 4) and find the next view, else terminate.

Note that global optimization is only performed on the seed. In real use, bundle adjustment cannot be avoided, but as one aim of this algorithm is to reduce the need for this relatively expensive operation, this step has been left out in the experiments that follow to illustrate the reduction in error accumulation possible with the proposed method.

6 Experiments

Figure 2 shows a simple synthetic example of the dependence on the seed of the propagated covariances. Although the *relative* uncertainties between all cameras remain the same in our linearized Gaussian propagation model, in reality the reconstruction errors depend heavily on the path taken.

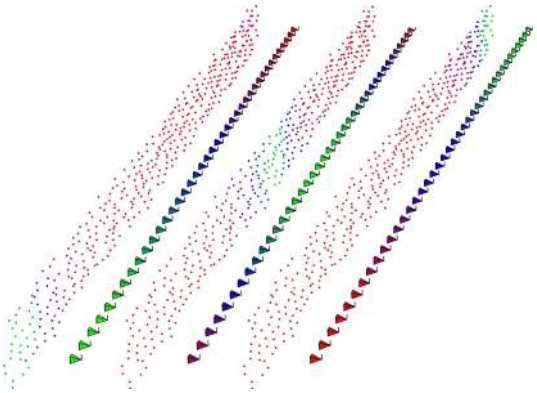


Fig. 2. Toy example of camera array observing a wall illustrating the covariance estimation results depending on which seed is chosen (from left to right, cameras 1 and 2, 19 and 20, 39 and 40). The points and cameras are color coded by the trace of their covariance, with green through blue to red for increasing uncertainty. Choosing the seed in the middle reduces the maximum camera uncertainty with respect to the seed.

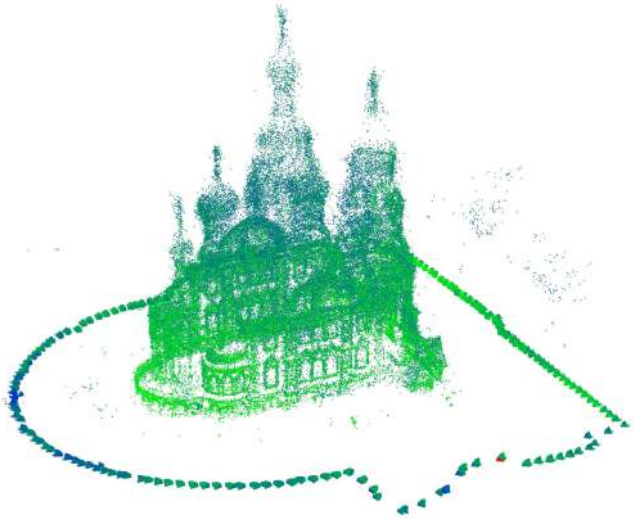


Fig. 3. Resulting point cloud reconstruction of the 'Spilled blood' dataset using the proposed algorithm, color coded by estimated covariance. No bundle adjustment has been performed. The dataset has 781 images, 162,521 points and 2,541,736 feature measurements.

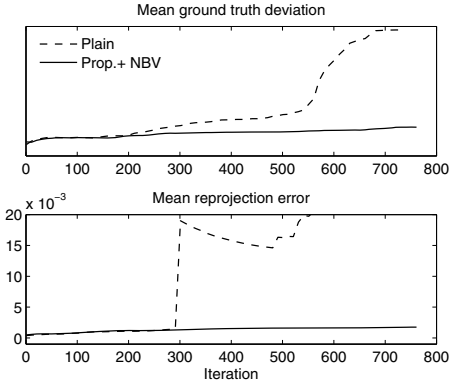


Fig. 4. Comparison between the proposed algorithm and a 'plain' method on the 'Spilled blood' dataset. In the plain method the standard reprojection error is minimized instead, and the next camera is chosen by the maximum overlap principle. Running times were 22 and 13 min respectively. There is no absolute scale on the top graph since it depends on the overall scale of the reconstruction, which is arbitrary.

Next, the algorithm is applied to a dataset extracted from photos of the 'Spilled blood' church in St. Petersburg. The reconstruction and a comparison with a standard method is shown in figures 3 and 4. The comparison shows the mean standard reprojection error and the ground truth deviation, defined

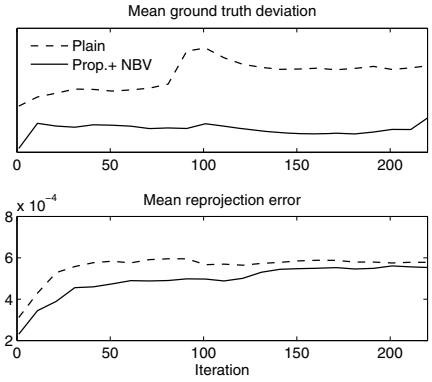


Fig. 5. Results for the Trafalgar dataset (256 images). Running times were 83 and 138 s respectively.

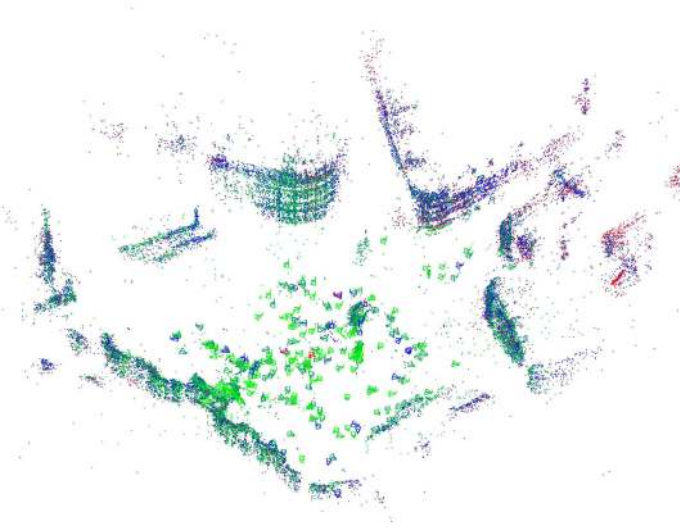
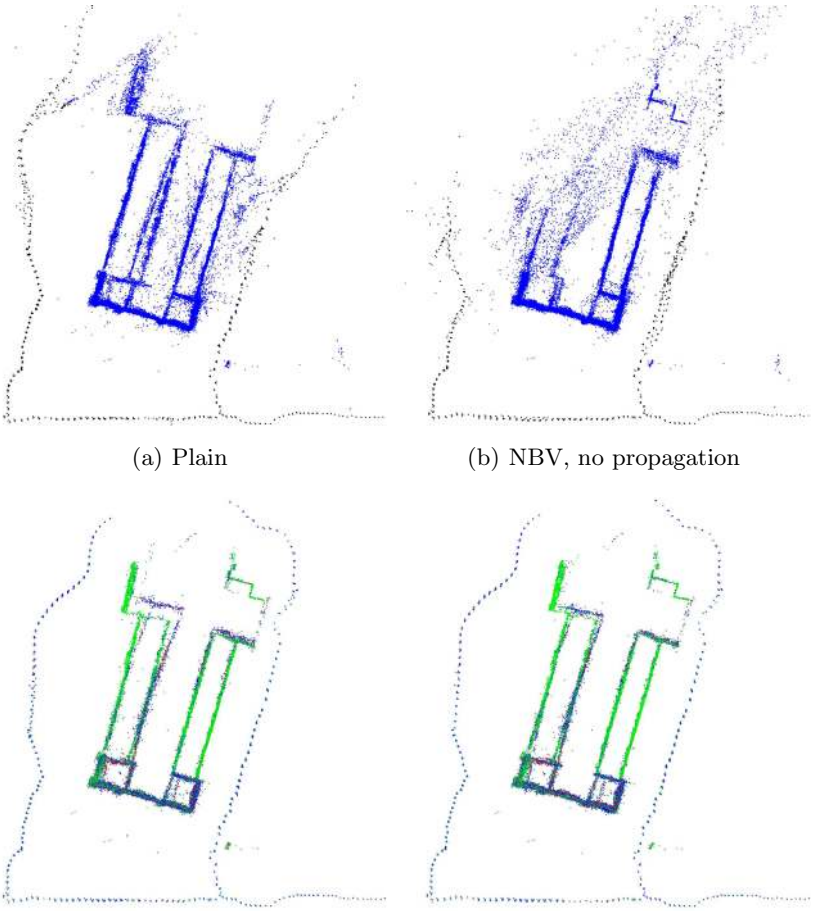


Fig. 6. Resulting point cloud reconstruction of the Trafalgar dataset

as the mean distance of each triangulated point from its ground truth position, after the two point clouds have been aligned using a Procrustes transformation. The 'ground truth' in this case has been obtained from the system described in [4]. The plain method, without covariance propagation or NBV planning, runs into trouble around iteration 300 and does not manage to resection all cameras, whereas the proposed algorithm does and is generally more robust and accurate. A similar comparison is made for the 'Trafalgar' dataset of [16] in figure 5.

Finally, we compare three variants of the proposed algorithm on the Lund Cathedral dataset. The covariance propagation and next best view-planning can be used independently, i.e. the next image can be chosen by maximum overlap

while propagating covariances, or the next view can be chosen based on camera uncertainty calculated using zero point covariances, with no propagation. As figure 7 shows, using NBV planning alone doesn't work well at all and the process breaks down, like the plain method. Propagation without planning works almost as well as both combined, and is probably the greatest contributing factor.



(a) Plain (b) NBV, no propagation

(c) Propagation only (d) Propagation and NBV

Fig. 7. Lund Cathedral dataset (1060 images, 45770 points, 408625 projections) reconstructed using the baseline algorithm, next best view-planning only without propagating covariances, propagating covariances but using the maximum overlap principle, and the proposed algorithm, using both NBV planning and propagation

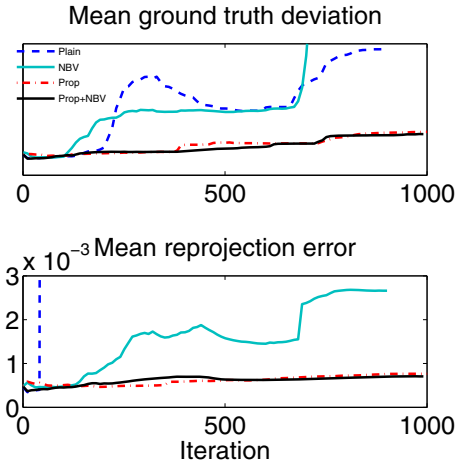


Fig. 8. Error plots for the Lund Cathedral dataset

6.1 Bundle Adjustment

In real use, bundle adjustment would be performed locally and/or globally at regular intervals during the reconstruction process. Unfortunately, after bundling the estimated covariances are no longer valid and need to be updated. Inverting the whole BA Hessian matrix is infeasible, but it is possible to compute only the block diagonal of the inverse quite efficiently, given the full covariance of only the camera parameters. The dominating cost is computing this as the inverse of the Schur complement of the camera block of the Hessian, a cost cubic in the number of cameras. As more cameras are resectioned, updating the covariances this way eventually becomes time-consuming and thus pointless.

Preliminary results indicate that the proposed algorithm still outperforms the standard method on datasets with outliers, where regular local bundle adjustment with a robust cost function and outlier removal is applied, even without updating the covariances at all. Nevertheless, an efficient mechanism for computing the new covariances is still needed and is the subject of future research.

7 Conclusion

The proposed method increases robustness to errors such as poorly resectioned cameras and poorly triangulated points, reduces error accumulation and also provides estimates of reconstruction accuracy which could be further processed for outlier detection etc. This comes at a cost of up to a twofold increase in running time. However, this cost is practically linear in the problem size, whereas iterated bundle adjustment costs between $\mathcal{O}(n^3)$ and $\mathcal{O}(n^4)$, depending on problem structure. Thus, trading less frequent bundling for covariance propagation and next best view planning should pay off for large problems.

References

1. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the World from Internet Photo Collections. International Journal of Computer Vision 80, 189–210 (2007)
2. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building rome in a day. In: ICCV, pp. 70–79 (2009)
3. Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building Rome on a Cloudless Day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 368–381. Springer, Heidelberg (2010)
4. Olsson, C., Enqvist, O.: Stable Structure from Motion for Unordered Image Collections. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 524–535. Springer, Heidelberg (2011)
5. Crandall, D., Owens, A., Snavely, N., Huttenlocher, D.P.: Discrete-Continuous Optimization for Large-Scale Structure from Motion. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3001–3008. IEEE (2011)
6. Chen, S., Li, Y.F., Zhang, J., Wang, W.: Active Sensor Planning for Multiview Vision Tasks, 1st edn. Springer Publishing Company, Incorporated (2008)
7. Dunn, E., Olague, G., Lutton, E.: Parisian camera placement for vision metrology. Pattern Recognition Letters 27, 1209–1219 (2006)
8. Wenhardt, S., Deutsch, B., Hornegger, J., Niemann, H., Denzler, J.: An Information Theoretic Approach for Next Best View Planning in 3-D Reconstruction. In: Proc. International Conference on Pattern Recognition (ICPR 2006), vol. 1, pp. 103–106. IEEE Computer Society (2006)
9. Trummer, M., Munkelt, C., Denzler, J.: Online Next-Best-View Planning for Accuracy Optimization Using an Extended E-Criterion. In: Proc. International Conference on Pattern Recognition (ICPR 2010), pp. 1642–1645. IEEE Computer Society (2010)
10. Dunn, E., van den Berg, J., Frahm, J.M.: Developing visual sensing strategies through next best view planning. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, pp. 4001–4008 (2009)
11. Haner, S., Heyden, A.: Optimal View Path Planning for Visual SLAM. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 370–380. Springer, Heidelberg (2011)
12. Hartley, R., Zisserman, A.: Multiple View Geometry. Cambridge University Press (2003)
13. Morris, D.D.: Gauge Freedoms and Uncertainty Modeling for 3D Computer Vision. PhD thesis, Carnegie Mellon University (2001)
14. Snavely, N., Seitz, S.S.M., Szeliski, R.: Skeletal graphs for efficient structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
15. Kahl, F., Hartley, R.: Multiple View Geometry Under the L_∞ Norm. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 1603–1617 (2008)
16. Agarwal, S., Snavely, N., Seitz, S.M., Szeliski, R.: Bundle Adjustment in the Large. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 29–42. Springer, Heidelberg (2010)