Robust Metric Reconstruction from Challenging Video Sequences

Guofeng Zhang[1] Xueying Qin[1*] Wei Hua[1] Tien-Tsin Wong[2] Pheng-Ann Heng[2] Hujun Bao[1*]

[1]State Key Lab of CAD&CG, Zhejiang University, P.R.China[†]
[2]The Chinese University of Hong Kong [‡]

Abstract

Although camera self-calibration and metric reconstruction have been extensively studied during the past decades, automatic metric reconstruction from long video sequences with varying focal length is still very challenging. Several critical issues in practical implementations are not adequately addressed. For example, how to select the initial frames for initializing the projective reconstruction? What criteria should be used? How to handle the large zooming problem? How to choose an appropriate moment for upgrading the projective reconstruction to a metric one? This paper gives a careful investigation of all these issues. Practical and effective approaches are proposed. In particular, we show that existing image-based distance is not an adequate measurement for selecting the initial frames. We propose a novel measurement to take into account the zoom degree, the self-calibration quality, as well as image-based distance. We then introduce a new strategy to decide when to upgrade the projective reconstruction to a metric one. Finally, to alleviate the heavy computational cost in the bundle adjustment, a local on-demand approach is proposed. Our method is also extensively compared with the state-of-the-art commercial software to evidence its robustness and stability.

1. Introduction

Structure and motion (SAM) recovery has been a long research topic for its importance in computer vision and wide applications in industrials, e.g. advertisement, film production and architecture design [1, 15]. One of the most challenging issues in SAM is to handle the long video sequences with varying focal length [10, 14, 13]. There are two main difficulties. First, in auto-calibration, zooming in/out may be confused with a forward/backward translation. Second, while processing long video sequences, the computational cost and time grow rapidly as the number of frames increases, and the accumulation error may cause the recovery to fail. However, metric reconstruction from long video sequences with varying focal length is very important, since such sequences are very common in professional movies and freelance personal videos. Most existing techniques handle the cases with unknown but constant focal length [3, 8, 9, 20].

Traditionally, SAM estimation starts with an algebraic initialization of the projective structure and motion using two- [22] or three- [2] view epipolar constraints, then upgrades the projective reconstruction to a metric one by some self-calibration techniques. These steps generally adopt the bundle adjustment (BA) method [7, 21] to refine the solution. To handle the unknown and varying intrinsic camera parameters, the selection of initial frames in projective reconstruction initialization is very crucial. In [13, 18], two initial frames are selected if they contain sufficient feature matches and their viewpoints are not too close to each other. However, it was observed in our experiments that such selection criteria is not sufficient for sequences with varying focal length, and often produce unreliable recovery results.

For long video sequences, the accumulation error in the projective reconstruction may ultimately cause the self-calibration to fail. Recently, Repko and Pollefeys [16] proposed a self-calibration approach to alleviate this problem. It still requires that the intrinsic camera parameters are constant. Moreover, the SAM estimation for a long sequence is time consuming, since the computational cost is dominated by BA, which grows rapidly with the number of frames. Although the computational cost can be reduced by taking advantage of the sparseness during optimization [21] or using key frames [6, 5, 19, 16], it is not efficient enough.

In this paper, we focus on the metric SAM recovery problem for challenging video sequences, and propose a robust approach, which can efficiently and reliably handle long sequences with varying focal length. Our main contributions are:

- A novel measurement for selecting the initial frames for projective reconstruction initialization, by com-

*Corresponding Author
[†]Email: {zhangguofeng, xyqin, huawei, bao}@cad.zju.edu.cn
[‡]Email:{ttwong, pheng}@cse.cuhk.edu.hk

1-4244-1180-7/07/$25.00 ©2007 IEEE

bining the zoom degree, self-calibration quality and image-based distance;

- A new strategy for upgrading the projective reconstruction to a metric one, which significantly improves the accuracy of metric SAM initialization;
- A local on-demand scheme in bundle adjustment, which dramatically accelerates the computation.

The rest of this paper is organized as follows. Section 2 gives an overview of our framework. Then our three main contributions are described in Section 3 (selection of initial frames), Section 4 (selection of upgrading moment) and Section 5 (local on-demand bundle adjustment). Experimental results and discussions are described in Section 6. Finally, we draw the conclusion in Section 7.

2. Framework Overview

In this section, we give an overview of our framework, as shown in Table 1. We first initialize the sequential metric structure and motion. Then we add and process the remaining frames one by one, in an order so that frames closer to the initization position are processed earlier. For each newly added frame, we use pose estimation technique to initialize its camera parameters and 3D points, and then refine existing structure and motion with BA. A key to the success of the above algorithm is to *accurately initialize the metric structure and motion*.

1	Track feature points over the whole sequence.
2	Select superior tracks and key frames.
3	*Initialize the metric structure and motion.*
3.1	Select the reference triple frames (RTFs) to initialize the projective reconstruction
3.2	Estimate the projective SAM with an incremental approach, and select an appropriate moment to upgrade it to a metric framework
4	For every additional key frame,
4.1	Initialize the newly added camera and 3D points
4.2	Refine the existing SAM by local on-demand BA
5	Solve the camera parameters of all non-key frames
6	[Optional] Refine the whole SAM through BA

Table 1. An overview of our framework.

2.1. Feature Matching and Key Frames

Our automatic feature tracking algorithm is based upon the iterative KLT algorithm [11, 17]. The matches between consecutive frames are constrained by the epipolar geometry [22]. We use RANSAC algorithm [4] to find a set of inliers that have consistent epipolar geometry. The matched feature points constitute the feature tracks.

As we know, structure and motion estimation with longer tracks is more reliable and robust than with that shorter tracks [5]. Let N be the minimal track length we required. Then we select the tracks not shorter than N as *superior tracks* for reconstruction. We use the interval $\frac{(N-1)}{2}$ to select the key frames to ensure that all superior tracks stride over at least two key frames. We also require that three consecutive key frames have sufficient of common tracks (at least 30 in our experiments) for robust estimation. If the common tracks are less than the required, we temporarily decrease the interval to select the key frames until there are enough common tracks or the interval is 1. The following estimation is mainly based on key frames and superior tracks.

2.2. Camera Model

We model each camera using seven parameters, i.e., the rotation expressed by three Euler angles $\Theta = (\theta_x, \theta_y, \theta_z)$, the translation $\mathbf{t} = (t_x, t_y, t_z)$, and the focal length f. The intrinsic matrix is then

$$\mathbf{K} = \begin{pmatrix} f & 0 & c_x \\ 0 & \alpha f & c_y \\ 0 & 0 & 1 \end{pmatrix} \quad (1)$$

where aspect ratio α and principal point (c_x, c_y) are assumed known. In our experiments, they are set to 1 and the image center, respectively.

3. Selection of Initial Frames

Certain factors may affect the precision of SAM estimation. Firstly, sufficient features should be matched. Secondly, the structure and motion also should not be near-degenerate so that the structure is well-conditioned. The second requirement is usually verified by the median distance between points transferred through an average planar-homography and the corresponding points in the target image [13], called *image-based distance*:

$$b = \text{median}(d(H\mathbf{u}, \mathbf{u}')) \quad (2)$$

where H is the planar-homography and can be solved by minimizing b. These two criteria is usually employed to select initial pairs for initialization in previous work [13, 18]. However, from on our experience, Pollefeys et al's selection criteria, which maximizes the product of the number of matches and the image-based distance, is not always reliable in the focal-length-varying configurations. Therefore, more factors should be analyzed.

As we know, self-calibration from two views contains uncertainty and is not reliable. Triple views are much more robust and have a nice cost performance. Therefore, we choose a triplet of views as the basic building block for

structure and motion recovery. After the step 2 in Table 1, we obtain key frames indexed with 1, 2, ..., etc. Then we group them into a series of triplets, such as (1,2,3), (2,3,4), (3,4,5), ... , etc., denote them as triplets 1, 2, 3, ..., etc.

3.1. Stabilizing Self-Calibration

The linear algorithm proposed in [13, 14] can deal with varying intrinsic camera parameters. Readers are referred to Appendix for more details. However, problems still exist. In practice, the estimated f_k^2 may be very small or even negative. The reason is that the first three items of Equation 10 are not symmetric for the constraint in f_k^2, which deviates $(P_k[1]\Omega^* P_k[1]^\top)/(P_k[3]\Omega^* P_k[3]^\top)$ and $(P_k[2]\Omega^* P_k[2]^\top)/(P_k[3]\Omega^* P_k[3]^\top)$ from 1.0 to zero or even negative. A similar discussion is also given by Pollefeys in his online tutorial [12]. In order to alleviate this problem, we define the following non-linear cost function based on Equation 10,

$$E_{calib} = \frac{1}{N-1}\sum_{k=2}^{N} E_k, \tag{3}$$

where

$$\begin{aligned}
E_k &= (\tfrac{1}{9})^2((\tfrac{P_k[1]\Omega^* P_k[1]^\top}{P_k[3]\Omega^* P_k[3]^\top} - 1)^2 + (\tfrac{P_k[3]\Omega^* P_k[3]^\top}{P_k[1]\Omega^* P_k[1]^\top} - 1)^2) \\
&+(\tfrac{1}{9})^2((\tfrac{P_k[2]\Omega^* P_k[2]^\top}{P_k[3]\Omega^* P_k[3]^\top} - 1)^2 + (\tfrac{P_k[3]\Omega^* P_k[3]^\top}{P_k[2]\Omega^* P_k[2]^\top} - 1)^2) \\
&+(\tfrac{1}{0.2})^2((\tfrac{P_k[1]\Omega^* P_k[1]^\top}{P_k[2]\Omega^* P_k[2]^\top} - 1)^2 + (\tfrac{P_k[2]\Omega^* P_k[2]^\top}{P_k[1]\Omega^* P_k[1]^\top} - 1)^2) \\
&+(\tfrac{1}{0.1})^2(\tfrac{P_k[1]\Omega^* P_k[3]^\top}{P_k[3]\Omega^* P_k[3]^\top})^2 + (\tfrac{1}{0.1})^2(\tfrac{P_k[2]\Omega^* P_k[3]^\top}{P_k[3]\Omega^* P_k[3]^\top})^2 \\
&+(\tfrac{1}{0.01})^2(\tfrac{P_k[1]\Omega^* P_k[2]^\top}{P_k[3]\Omega^* P_k[3]^\top})^2
\end{aligned}$$

The notations are explained in the Appendix. Firstly, we use the linear algorithm introduced in the Appendix to estimate the focal lengths f_k. The weights of the first two terms of Equation 10 are set to values higher than the default values ($\frac{1}{\nu}$ in our experiment) to avoid that the estimated f_k^2 being too small or negative. Then these results are refined by minimizing the cost function E_{calib}. This modification can significantly improve the robustness of self-calibration, especially in the focal-length-varying configurations.

3.2. Zoom Degree

Although some algorithms for varying and unknown focal length have been proposed, the robustness is still an issue. First, in auto-calibration, a forward translating camera may be confused with zooming, especially when the scene is near planar. Second, zoom may bring up some other problems, such as motion blur may occur, and feature matching is more challenging. In general, the matching noise is larger than the cases without zoom. These problems aggravate further the confusion problem between a zoom in/out and a forward/backward translation. The following problem often occurs. The estimated SAM deviates largely from the ground truth, although the reprojection error is small. We find that a smaller zoom is more suitable for initialization. We propose a criteria to evaluate the degree of zooming between two views. The pure focal length difference between two views is normalized with the image-based distance (Equation 2):

$$\Delta f_{ij} = \frac{|f_i/f_j - 1| + |f_j/f_i - 1|}{2b_{ij}} \tag{4}$$

In our implementation, we perform projective reconstruction on each triplet of key frames *independently*, and then estimate their focal length by self-calibration.

3.3. Self-Calibration Quality

Since we begin with a projective reconstruction, and then upgrade it to a metric one through self-calibration, the quality of projective reconstruction is very important. As we know, self-calibration is sensitive to noises, and requires very accurate projective reconstruction. The reprojection error is usually used to assess the precision of projection matrices. In this paper, we argue that the reprojection error is not reliable for evaluating the precision of projective reconstruction in the case that no prior knowledge is available. In practice, although the reprojection error is quite small, projection matrices are still ill-posed, which leads to the failure of self-calibration. Since self-calibration is sensitive to the precision of projective reconstruction, the quality of self-calibration is a reasonable way to evaluate the quality of projective reconstruction. Based on the cost function in Equation 3, we define the criteria for the self-calibration quality as the following,

$$C(E_{calib}) = \frac{\varepsilon}{\varepsilon + \sqrt{E_{calib}}} e^{-\frac{E_{calib}}{2\sigma^2}} \tag{5}$$

where $\varepsilon = 0.1$, $\sigma = 0.2$ in our experiments. The value of $C(E_{calib})$ is in the range of [0,1]. No matter whether E_{calib} is small or large, $C(E_{calib})$ manifests a good discriminative ability.

3.4. Criteria for Selecting Initial Frames

Generally, a subsequence or initial views suitable for metric reconstruction should satisfy the following factors:

1. There are sufficient feature matches
2. The configurations are not near-degenerate
3. The zoom degree is small
4. The self-calibration quality is good

Generally, more feature matches produce more robust estimation. But it is difficult to determine how many matches

are sufficient. In fact, this factor can be suggested by the self-calibration quality. Therefore, we do not set a threshold value for this factor. For a triplet i, we define the following criteria to account for the image-based distance, zoom degree and self-calibration quality:

$$S_i = C(E_{calib})(B_{i,i+1} + B_{i+1,i+2} + B_{i,i+2}) \quad (6)$$

where $B_{ij} = \frac{b_{ij}}{\beta + \Delta f_{ij}}$, and $\beta = 0.04$ in our experiments.

We employ a method similar to [13] with our stabilized self-calibration to reconstruct each triplet *independently* and compute its S_i according to Equation 6.

Furthermore, the initial frames should be in a suitable subsequence for more robustness. In order to evaluate the suitability of the subsequence associated with the triplet i, we apply a Gaussian filter on S_i as the following:

$$\tilde{S}_i = \Big(\sum_{k=i-3w}^{i+3w} e^{-\frac{(k-i)^2}{2w^2}} S_k\Big) \Big/ \sum_{k=i-3w}^{i+3w} e^{-\frac{(k-i)^2}{2w^2}} \quad (7)$$

where $w = 3$ in our experiments. Finally, we use the following criteria

$$S_i^b = \sqrt{\tilde{S}_i S_i} \quad (8)$$

to select the best triplet with maximum S_i^b. Such as triplet l, i.e. $(l, l+1, l+2)$, is selected as our initial frames, which is defined as *reference triple frames* (RTFs). In the following estimation, other key frames will be processed incrementally with the ordering: $l-1, l+3, l-2, l+4, ...$, etc.

4. Selection of Upgrading Moment

For a long sequence, the accumulation error in the projective reconstruction may ultimately cause the failure of self-calibration. Therefore, we should manage the accumulation error, and select an appropriate moment to upgrade the projective reconstruction to a metric one before the accumulation error damages the self-calibration.

The feature tracks in the RTFs are called *reference tracks*. Their corresponding 3D points are called *reference 3D points*. Since the RTFs are suitable for initialization, its projective reconstruction can be regarded reasonable close to the ground truth. Hence, the reference 3D points are well-conditioned, and their reprojection error can be used to estimate the precision of reconstruction reliably. Therefore, we use the reference 3D points to manage the accumulation error in the projective reconstruction. For every additional key frame, we check if it satisfies that there are at least n_p projections of reference 3D points, and the average reprojection error of these reference 3D points is less than e_p pixels. In our experiments, $n_p = 15$ and $e_p = 3.0$. If no more additional key frames satisfy this condition, the projective reconstruction is stopped and upgraded to a metric one through our stabilized self-calibration with these key frames. The results are refined through metric bundle adjustment immediately. So far, we have accomplished the task for initializing the metric structure and motion recovery.

5. Local On-Demand Bundle Adjustment

In step 4.2 of Table 1, if we refine the existing SAM with traditional BA for every additional frame, the computational cost is prohibitively large for long sequences. To reduce the computational cost of BA, we propose a local on-demand scheme. Firstly, we *only* refine the additional key frame and its visible m_l 3D points (i.e., those 3D points that have the corresponding 2D feature points in this frame). Other cameras and 3D feature points are fixed, in order to reduce the computational cost. In fact, we only need to fix the cameras of the n_l key frames that also present ("see") one or more of these m_l 3D points. All other unrelated cameras and 3D feature points are not touched.

If the refined reprojection error exceeds the threshold, more key frames and the associated 3D points should be considered in the next round of refinement. The cameras of these $n_l + 1$ key frames and their associated 3D points are treated as variables and refined. Other cameras and 3D points are fixed (unrelated cameras and 3D points are not touched). This *local on-demand* approach to bundle adjustment continues until either the threshold is satisfied or all key frames and 3D points have been refined. In practice, the threshold is usually satisfied even in the first round of refinement. That means the computational time can be significantly reduced.

6. Experiments

We have examined our algorithms with different video sequences on a desktop PC with Intel Xeon 3.0 GHz CPU and 2 GB memory. Table 2 shows the statistics of three video sequences we tested. In the first 100 frames of the Synthetic Campus sequence, the camera turns right and zooms in with a slight translation, and then moves freely and rapidly. In some subsequences, the camera simultaneously zooms in and moves backward, or zooms out and moves forward. In addition, 11.7% image noise is added, i.e., each pixel is perturbed by a random number uniformly distributed in [-30,30] (pixel values are in the range [0,255]). The Building and the Garden sequence are two real outdoor video sequences captured by a hand-held video camera. They are long sequences with varying focal length.

Table 2 also summarizes the performance. In our method, the computational time mainly depends on the number of key frames, 3D points and image projections. For the Building sequence of 2,410 frames, we choose 98 key frames and reconstruct 2,908 3D points with 462,621 image projections, which take around 16 minutes. The root

sequence	Syn. Campus	Building	Garden
frames	601	2410	1608
key frames	51	98	55
3D points	2825	2908	4283
image projections	171,342	462,621	803,750
RMSE. (pixels)	0.586	1.327	1.065
matching time (min.)	7	28	21
solving time (min.)	6	16	12
performance list of boujou three			
matching time (min.)	6	21	14
solving time (min.)	55	51	34

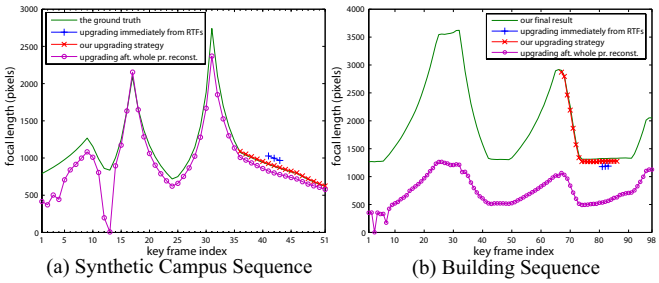
Table 2. The statistical results of our approach and the performance of boujou three.

mean squared error (RMSE) of the reprojections is 1.327 pixels. For extensive comparison, these sequences are also tested with the state-of-the-art software product, "boujou three" by 2d3 company [1], which is the best available to us. Compared with the performance of boujou three shown in the last rows of Table 2, our method is 3 to 9 times faster than that of boujou three.

In the following, we first show the efficiency of our criteria in selecting initial frames, including a comparison with the criteria of Pollefeys et al. [13] (Pollefeys criteria for short). Then we compare the results of different upgrading strategies. Finally, we verify our algorithm by generating augmented videos based on our reconstruction results, and compare them with that of boujou three.

Initialization Criteria Analysis: We first analyze the criteria for selecting initial frames using the Synthetic Campus sequence, as shown in Figure 1(A). Figure 1(A1)-(A4) respectively show the average number of matches, average image-based distance, zoom degree, and self-calibration quality of each triplet. Figure 1(A5) shows the translation initialization error of each triplet. Since the accuracy of the recovered 3D structure is highly related to the camera translation, triple frames with smaller translation initialization error are more suitable for initialization. Therefore, the triplets around 35∼45 are good candidates for initialization.

Figure 1(A6) and (A7) respectively show the suitable scores of the triplets computed using Pollefeys criteria [13] and our criteria in Equation 8. The peak score in Figure 1(A6) is at 12, implying that Pollefeys criteria will choose triplet 12 for initialization, which is a poor candidate as indicated by the translation initialization error curve in Figure 1(A5). This is because that Pollefeys criteria, as a product of the average number of feature matches and the average image-based distance, may not adequately reflect the translation initialization error. In contrast, the peak score in Figure 1(A7) is at 41, implying that our criteria will choose triplet 41 for initialization, which is a good candidate as indicated by the translation initialization error curve in Figure 1(A5).



(a) Synthetic Campus Sequence (b) Building Sequence

Figure 2. Focal length obtained by different upgrading strategies.

Then we analyze the impact of three factors included in our selection criteria. The image-based distance criteria is useful to detect the case of near-pure rotation or near-degenerate configurations, e.g. triplets 1-7. However, it is not sensitive to reflect the instability of reconstruction, especially in the large zoom case, e.g. triplets 11-17. Fortunately, the self-calibration quality measures reconstruction quality well and remedy this flaw. Small initialization error, e.g. triplets 35-45, usually occurs in the case of small zoom degree. Therefore, it is quite necessary to include zoom degree and self-calibration quality in our selection criteria.

In addition, Figure 1(A8) shows the reprojection error of projective/metric reconstruction (upgrading to metric without being refined by BA) from each triplet. It shows the reprojection error of most triplets is very small and close to each other, and therefore it is unreliable to measure the projective reconstruction quality.

We perform the same analysis in Figure 1(B) for the real video sequence, Building. Note that the final result by our algorithm is used as the ground-truth to compute the translation initialization error, since there is no ground-truth for this real captured sequence and our result is accurate as demonstrated by our augmented video result. Figure 1(A) and (B) show that our criteria is superior for not only synthetic but also real videos. Figure 1 also shows that it is a bad choice to initialize the sequential structure and motion computation from the beginning of these two sequences, which may cause the initial 3D point structure and camera motions to be ill-posed and eventually result in bad reconstruction or even failure.

Upgrading Strategy Evaluation: We evaluate our upgrading strategy using the Synthetic Campus sequence and the real Building sequence, and compare them to different strategies. The focal length after upgrading and the ground-truth focal length are shown in Figure 2.

Figure 2(a) shows the results for the Synthetic Campus sequence. One can easily see that our upgrading scheme produces the best results on focal length, and upgrading immediately from RTFs even produces much better results than upgrading after the complete projective reconstruction.

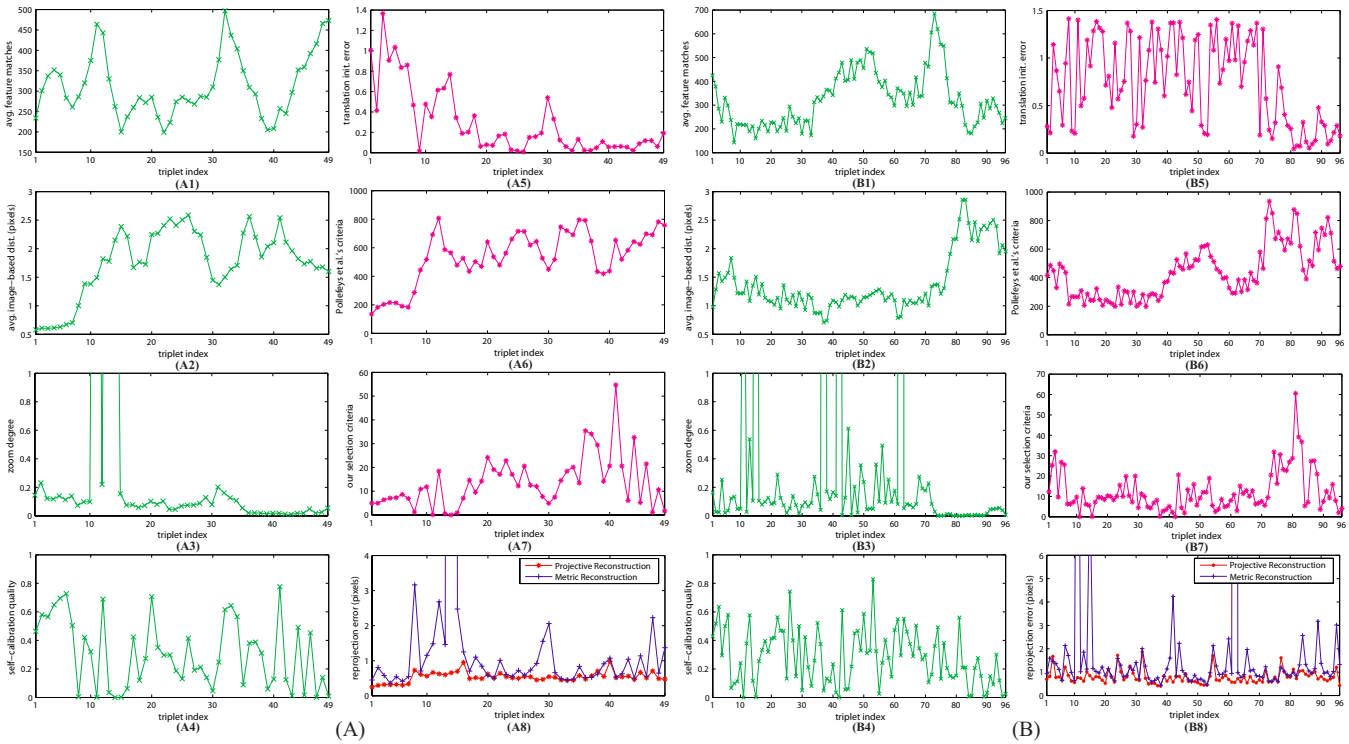


Figure 1. Selecting initial frames: (A) the Synthetic Campus sequence; (B) the Building sequence. The horizontal axis is the index of triplets, while the vertical axis is: (A1) & (B1): the average number of feature matches; (A2) & (B2): image-based distance; (A3) & (B3): zoom degree; (A4) & (B4): self-calibration quality; (A5) & (B5): translation initialization error; (A6) & (B6): Pollefeys et al.'s selection criteria; (A7) & (B7): our selection criteria; (A8) & (B8): reprojection error.

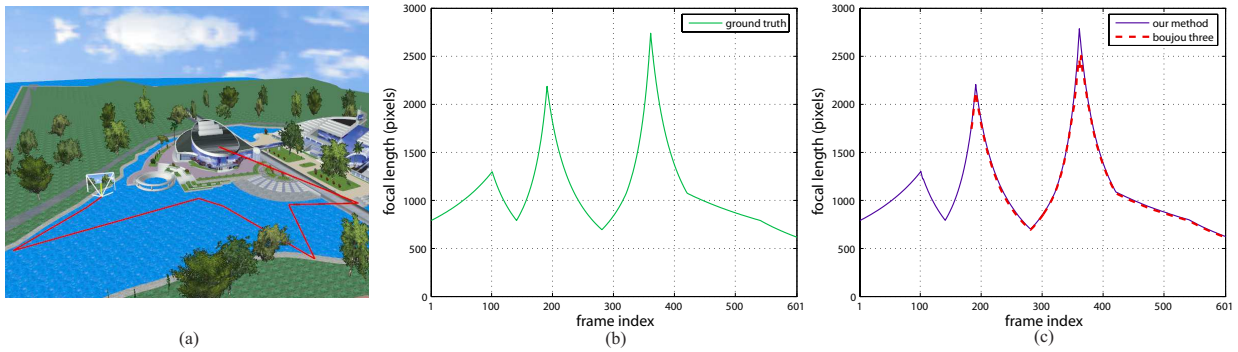


Figure 3. The Synthetic Campus sequence. (a) the synthetic scene with the camera trajectory (red line); (b) the ground truth of the focal length; (c) the focal length recovered by our method and boujou three.

Upgrading after the complete projective reconstruction produces extremely large focal length error at key frame 13. Figure 2(b) shows another set of results for the real Building sequence, and evidences that our upgrading strategy works nicely for real sequence as well.

Verification: In the following, we show more metric reconstruction results and corresponding augmented videos, including a comparision with that of boujou three. Figure 3 compares the focal length recovered by our method and boujou three. As shown by the dashed curve of Figure 3(c), most of the focal length recovered by boujou three are quite close to the ground truth, while there is significant deviations at the frames with peak local maximum focal length values, e.g. frame 361. Our method produces more accurate results at these frames than boujou three. The maximum deviation from the ground truth is less than 2%. In addition, our method successfully reconstructs the whole sequence, while boujou three does not provide a solution for the first 183 frames.

In Figure 4, we show the results of recovered focal length, 3D points, camera trajectory from the real Building

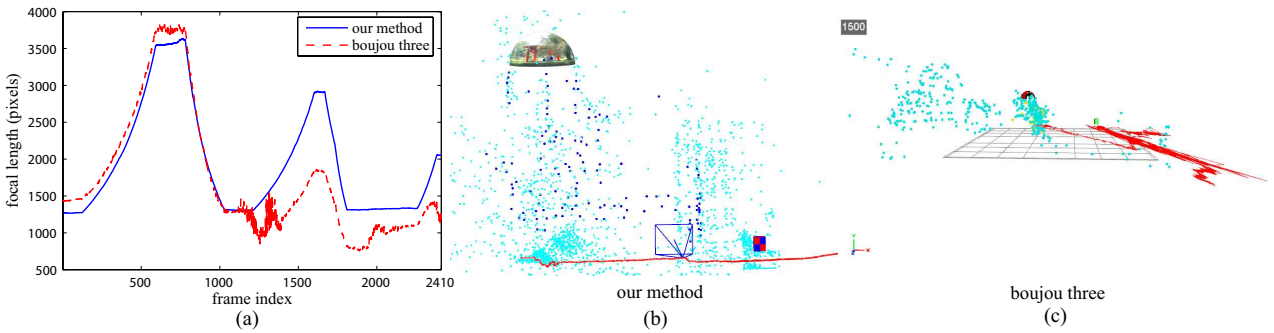


Figure 4. The result of the Building sequence. (a) shows the focal length recovered by our method and boujou three; (b) and (c) show the 3D points and the camera trajectory, captured from our system and boujou three respectively, overlaid with the synthetic objects.

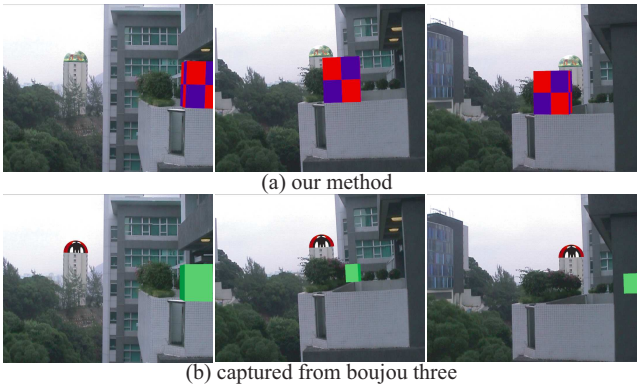


Figure 5. Some augmented frames of the Building sequence.

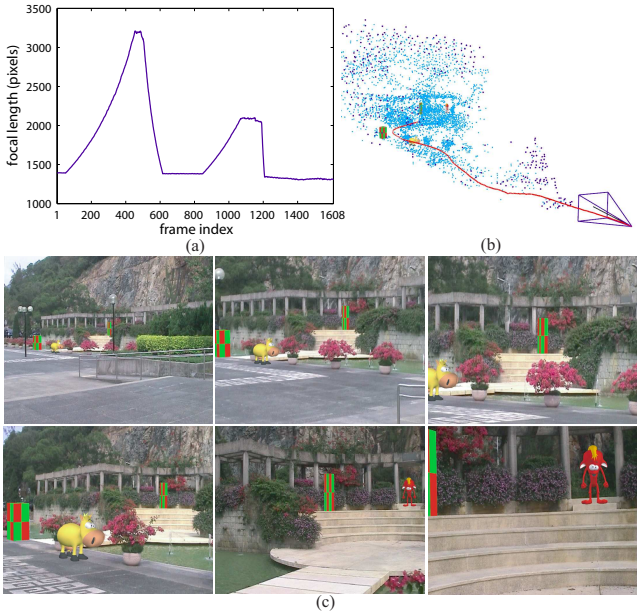


Figure 6. The result of the Garden sequence. (a) shows the focal length recovered by our method. (b) shows our recovered 3D points and the camera trajectory overlaid with the synthetic objects. (c) some augmented frames by our method.

sequence. Since there is no ground truth, we verify the recovered results by augmenting the video with two synthetic objects. The dashed curve in Figure 4(a) and the camera trajectory in Figure 4(c) indicate that the recovered camera motion by boujou three is very jittering, which is however not true. Several augmented frames captured from boujou three are shown in Figure 5(b), to examine its reconstruction quality. While the overlaid dome far away from the camera in boujou three's augmented video looks steady, the overlaid box closer to the camera is jitter and drifts, which implies the recovered SAM is not accurate. The recovered results by our method are shown in Figure 4(a) (solid curve) and Figure 4(b), the quality is also verified by the augmented video result as shown in Figure 5(a), in which both the farther dome and nearby box remain firmly registered to the scene throughout the sequence. Figure 6 shows another set of reconstruction result and augmented video result by our method with the real sequence, Garden. High quality reconstruction is resulted.

7. Conclusions

We have presented a robust approach for automatic metric reconstruction. The major advantage of our approach is that it can reliably handle long video sequences with varying focal length, which may cause problems to previous methods. In addition, our approach is very efficient. These advantages are achieved by our innovations in the selection criteria of initial frames, the upgrading strategy and the local on-demand optimization scheme. The robustness and efficiency have been demonstrated by our extensive experiments on both synthetic and real video sequences.

The limitation of the proposed work is that we do not consider the non-linear lens distortion in our current SAM estimation, thus our method may be unsuitable for the sequences that contain significant lens distortion. Addressing the lens distortion is left for future development. Another future direction is to improve the feature tracking results for sequences with motion blur, so that we can have enough long tracks for robust metric reconstruction.

Acknowledgements

The authors are greatly indebted to Prof. Zhanyi Hu and Prof. Xinguo Liu for their insightful suggestions. This paper is supported by 973 program of China (No.2002CB312104), NSF of China (No.60633070), and affiliated with the Virtual Reality, Visualization and Imaging Research Center at The Chinese University of Hong Kong as well as the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies.

Appendix

According to computer vision theory, the projection of the absolute quadric in the image yields the dual image absolute conic:

$$w^* \simeq \mathbf{K}_k {\mathbf{K}_k}^\top \simeq \mathbf{P}_k \Omega^* {\mathbf{P}_k}^\top, \tag{9}$$

where $\mathbf{K}_k$ and $\mathbf{P}_k$ is the intrinsic matrix and projection matrix of frame k, which are normalized according to [13]. If the skew is assumed to zero, both principal point and aspect ratio are known, according to the linear self-calibration algorithm proposed in [13, 14], Equation 9 can be rewritten as follows:

$$\lambda_k \begin{bmatrix} f_k^2 & 0 & 0 \\ 0 & f_k^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{P}_k \begin{bmatrix} f_1^2 & 0 & 0 & a_1 \\ 0 & f_1^2 & 0 & a_2 \\ 0 & 0 & 1 & a_3 \\ a_1 & a_2 & a_3 & ||a||^2 \end{bmatrix} \mathbf{P}_k^\top.$$

The uncertainty is taken into account by weighting the equations accordingly.

$$\begin{array}{l} \frac{1}{9\nu}(P_k[1]\Omega^* P_k[1]^\top - P_k[3]\Omega^* P_k[3]^\top) = 0 \\ \frac{1}{9\nu}(P_k[2]\Omega^* P_k[2]^\top - P_k[3]\Omega^* P_k[3]^\top) = 0 \\ \frac{1}{0.2\nu}(P_k[1]\Omega^* P_k[1]^\top - P_k[2]\Omega^* P_k[2]^\top) = 0 \\ \frac{1}{0.1\nu}(P_k[1]\Omega^* P_k[3]^\top) = 0 \\ \frac{1}{0.1\nu}(P_k[2]\Omega^* P_k[3]^\top) = 0 \\ \frac{1}{0.01\nu}(P_k[1]\Omega^* P_k[2]^\top) = 0 \end{array} \tag{10}$$

where $P_k[i]$ is the ith row of $\mathbf{P}_k$ and ν a scale factor that is initialized to 1 and then set to $P_k[3]\tilde{\Omega}^* P_k[3]^\top$ with $\tilde{\Omega}^*$ which is the result of the previous iteration. An estimate of the dual absolute quadric Ω^* can be obtained by solving the above set of equations for all views through linear least-squares. Please refer to [13, 14] for more details.

References

[1] 2d3. http://www.2d3.com. 1, 5

[2] S. Avidan and A. Shashua. Threading fundamental matrices. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(1):73–77, 2001. 1

[3] O. D. Faugeras, Q.-T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In *ECCV*, pages 321–334, 1992. 1

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 2

[5] A. Fitzgibbon and A. Zisserman. Automatic camera tracking. In M. Shah and R. Kumar, editors, *Video Registration*, chapter 2, pages 18–35. Kluwer, 2003. 1, 2

[6] S. Gibson, J. Cook, T. Howard, R. J. Hubbold, and D. Oram. Accurate camera calibration for off-line, video-based augmented reality. In *ISMAR*, pages 37–46, 2002. 1

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000. 1

[8] R. I. Hartley. Euclidean reconstruction from uncalibrated views. In *Applications of Invariance in Computer Vision*, pages 237–256, 1993. 1

[9] A. Heyden and K. Astrom. Euclidean reconstruction from constant intrinsic parameters. In *ICPR'96*, volume 1, pages 339–343, 1996. 1

[10] A. Heyden and K. Åström. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In *CVPR*, pages 438–443, 1997. 1

[11] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981. 2

[12] M. Pollefeys. Visual 3d modeling from images. Online tutorial: http://www.cs.unc.edu/∼marc/tutorial/index.html. 3

[13] M. Pollefeys, L. J. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004. 1, 2, 3, 4, 5, 8

[14] M. Pollefeys, R. Koch, and L. J. V. Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *ICCV*, pages 90–95, 1998. 1, 3, 8

[15] REALVIZ. http://www.realviz.com. 1

[16] J. Repko and M. Pollefeys. 3d models from extended uncalibrated video sequences: Addressing key-frame selection and projective drift. In *3DIM*, pages 150–157, 2005. 1

[17] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600, 1994. 2

[18] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006. 1, 2

[19] T. Thormählen, H. Broszio, and A. Weissenfeld. Keyframe selection for camera motion and structure estimation from multiple views. In *ECCV*, pages 523–535, 2004. 1

[20] B. Triggs. Autocalibration and the absolute quadric. In *CVPR*, pages 609–614, 1997. 1

[21] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Workshop on Vision Algorithms*, pages 298–372, 1999. 1

[22] Z. Zhang, R. Deriche, O. D. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artif. Intell.*, 78(1-2):87–119, 1995. 1, 2