Supplementary Document to CVPR submission #0041

Je Hyeong Hong
University of Cambridge
jhh37@cantab.net

Christopher Zach
Toshiba Research Europe
christopher.m.zach@gmail.com

1. Derivation of Eq. (6) in [4]

We expand the terms in Eq. (5) in [4],

$$f_{ij} := \left\| \begin{matrix} \sqrt{1-\eta}\left(\mathtt{P}_{i,1:2}\mathbf{x}_j - (\mathbf{p}_{i,3}^\top \mathbf{x}_j)\mathbf{m}_{ij}\right) \\ \sqrt{\eta}\left(\mathtt{P}_{i,1:2}\mathbf{x}_j - \mathbf{m}_{ij}\right) \end{matrix} \right\|_2^2 . \quad (1)$$

We will drop the subscripts i and j and introduce $\mathbf{y} := \mathtt{P}\mathbf{x}$ for brevity. Thus, we obtain

$$\begin{aligned} f &= \left\| \begin{matrix} \sqrt{1-\eta}\left(\mathbf{y}_{1:2} - y_3\mathbf{m}\right) \\ \sqrt{\eta}\left(\mathbf{y}_{1:2} - \mathbf{m}\right) \end{matrix} \right\|_2^2 \\ &= (1-\eta)\left\|\mathbf{y}_{1:2} - y_3\mathbf{m}\right\|_2^2 + \eta\left\|\mathbf{y}_{1:2} - \mathbf{m}\right\|_2^2 \\ &= (1-\eta)\left(\|\mathbf{y}_{1:2}\|_2^2 - 2y_3\mathbf{m}^\top\mathbf{y}_{1:2} + \|y_3\mathbf{m}\|_2^2\right) \\ &+ \eta\left(\|\mathbf{y}_{1:2}\|_2^2 - 2\mathbf{m}^\top\mathbf{y}_{1:2} + \|\mathbf{m}\|_2^2\right) \\ &= \|\mathbf{y}_{1:2}\|_2^2 - 2\mathbf{m}^\top\big((1-\eta)y_3 + \eta\big)\mathbf{y}_{1:2} \\ &+ (1-\eta)\|y_3\mathbf{m}\|_2^2 + \eta\|\mathbf{m}\|_2^2 \\ &= \left\|\mathbf{y}_{1:2} - \big((1-\eta)y_3 + \eta\big)\mathbf{m}\right\|_2^2 \\ &- \big((1-\eta)y_3 + \eta\big)^2\|\mathbf{m}\|_2^2 + \big((1-\eta)(y_3^2+\eta)\big)\|\mathbf{m}\|_2^2 . \end{aligned} \quad (2)$$

We expand the coefficient c of $\|\mathbf{m}\|_2^2$ from the last line,

$$\begin{aligned} c &:= (1-\eta)y_3^2 + \eta - \big((1-\eta)y_3 + \eta\big)^2 \\ &= (1-\eta)y_3^2 + \eta - (1-\eta)^2y_3^2 - 2(1-\eta)\eta y_3 - \eta^2 \\ &= (1-\eta)(1-(1-\eta))y_3^2 - 2(1-\eta)\eta y_3 + \eta(1-\eta) \\ &= (1-\eta)\left(\eta y_3^2 - 2\eta y_3 + \eta\right) \\ &= \eta(1-\eta)(y_3-1)^2 . \end{aligned} \quad (3)$$

Hence, f simplifies to

$$\begin{aligned} f &= \left\|\mathbf{y}_{1:2} - \big((1-\eta)y_3 + \eta\big)\mathbf{m}\right\|_2^2 \\ &+ \eta(1-\eta)(y_3-1)^2\|\mathbf{m}\|_2^2 . \end{aligned} \quad (4)$$

Inserting the expression for $\mathbf{y} = \mathtt{P}\mathbf{x}$ and summing over all image observations yields Eq. (6) in [4].

2. Visualization of results

We show iterative visualizations of Fountain-P11 [11], Vercingetorix [8] and Alcatraz Courtyard [8] in the separately attached video. Note that:

1. reconstructions shown during pOSE optimization and projective refinement stages are not in the metric frame and therefore may look slightly distorted, and
2. these visualizations include unsuccessful steps during which the video only shows the so-far best solution.

Other reconstructions are included in §5 and also separately attached as PLY files.

3. Further implementation details

In this section, we illustrate further implementation details which may help readers in understanding our strategy.

3.1. Initialization

In §6 of [4], we mentioned that each camera parameter is drawn from $\mathcal{N}(0,1)$, after which each row of every camera matrix is normalized to have unit norm to improve numerical stability. This procedure is equivalent to taking a sample from the uniform distribution on a 4D hypersphere (see [3]).

3.2. Success criterion

We first illustrate components required for understanding our success criterion used in assessing the performance of the algorithms described in [4].

Camera position error For each run on each dataset, we measure the mean deviation of camera centers from corresponding baseline values. Since our reconstructions are resolved up to metric, we have to apply optimal similarity transform [2] on the obtained solution in order to compare against baseline camera centers.

Above is a useful metric especially in detecting failure cases where a solution has a low average reprojection error but has one or two cameras positioned incorrectly. (e.g.

Property	Value
Relative function tolerance	10^{-9}
Relative gradient tolerance	10^{-6}
Relative parameter tolerance	10^{-9}
Initial damping factor	10^{-1}
Max. num. of iterations for pOSE	400
Max. num. of iterations for projective BA	400
Max. num. of iterations for metric upgrade	50
Max. num. of iterations for metric BA	200
Min. num of inliers threshold for 2-view	30
Min. length of each point track	4

Table 1. Settings used in our experiments

cameras looking behind a planar scene, which sometimes occurs in Castle-P*s.)

A downside of having metric ambiguity resolved via optimal similiarity transform is that any global translational drift present will go unnoticed. Hence, this measure is likely to output more optimistic value than in reality. Nevertheless, it is a useful metric for comparing reconstructions (also used in some global SfM pipelines [5, 12]).

Baseline For each dataset in Tables 1 and 3 of [4], the baseline data used is either a set of ground truth measurements (Strecha's sequences: Fountain-P11, Entry-P10, Herz-Jesu-P*s and Castle-P*s) or is obtained from COLMAP [10], which is a robust and reliable incremental SfM pipeline. (For datasets in Table 1 of [4], point tracks are converted to custom inlier feature matches before going through an incremental pipeline.)

Classifying success Each run is deemed successful if its mean camera position deviation from the baseline is less than two times that of the best minimum (observed in total of N runs). In terms of equations, if we denote the mean camera position error from the baseline in run k by $\bar{\varepsilon}_k$, the total number of runs made by N (≈ 100) and the *success threshold* by $\epsilon_{\text{success}}$, then

$$\bar{\varepsilon}_{\min} := \min(\bar{\varepsilon}_1, ..., \bar{\varepsilon}_N), \qquad \text{and}$$
$$\epsilon_{\text{success}} = 2\bar{\varepsilon}_{\min}.$$

We have manually checked the visual feasibility of each reconstruction classified *successful*. We believe that the choice of factor 2 is strict (see Figure 1 for a failure example) but somewhat arbitrary, and future work should introduce more rigorous statistical analysis such as median absolute deviation (MAD) of reached optima. Nevertheless, for most datasets (except for Castle-P*s), there are usually finite number of distinct observable basins which are very close in terms of camera position error such that the corresponding success rates are not very sensitive to changes in this factor.

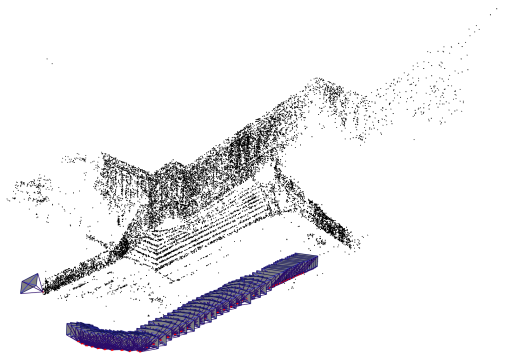


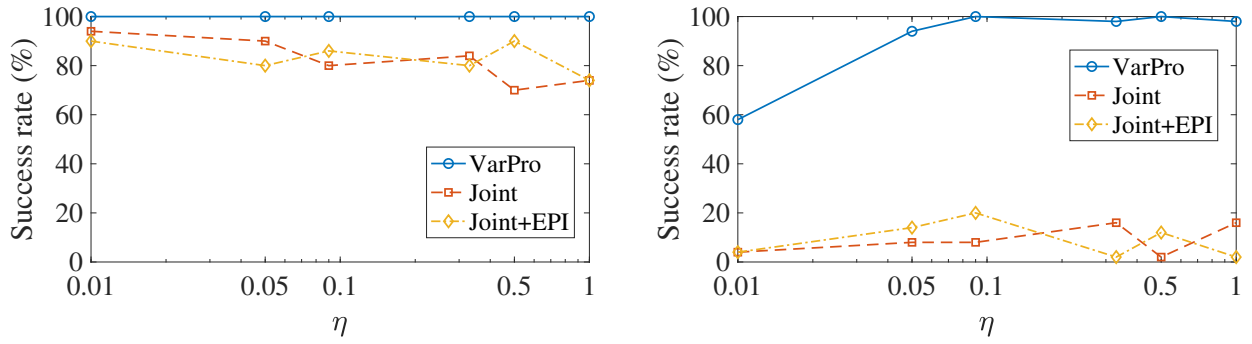
Figure 1. An almost successful reconstruction of Alcatraz Courtyard from inlier tracks. As the leftmost camera center is far off from its baseline position, this is considered a failure run (one of 6% failure cases in SRI [4]).

3.3. Determining the value of η

In order to determine an appropriate value of η, a non-uniform grid search was carried out on Strecha's benchmark datasets (listed in Table 2 of [4]), which have ground truth camera positions, the small dinosaur and vercingetorix datasets, which are circular motion sequences with more-affine views, and [8]'s strongly-perspective Lund cathedral dataset (small trimmed version). The considered values of η were 0.01, 0.05, 0.09, 0.33, 0.5 and 1. For each value of η on each dataset, the number of runs (out of 20 random initialization points) yielding accurate 3D reconstructions (through the stratified bundle adjustment strategy in § 4 of [4]) was recorded. While "easier" datasets such as Fountain-P11 and Herz-Jesu-P8 retrieved accurate 3D reconstructions across all the considered values of η, more-affine circular motion sequences had convergence issues with pOSE for $\eta = 0.01$ (e.g. Fig. 2a). On the other hand, the perspective Lund cathedral sequence did not have convergence issues during the pOSE stage but the pOSE solution with high η values led to perturbed 3D reconstructions. By considering the worst success rate performance for each η value, it was decided to set η to 0.05.

3.4. GPRT

The General Projective Reconstruction Theorem (GPRT) proposed by Nasihatkon et al. [6] is only illustrated for noise-free and fully-visible measurements. As briefly mentioned in §3 of the paper, we attempted to extend this approach to account for situations with noise and missing data. As shown in Figure 3, we generated a steplike mask which overlays on top of the visible region of data. This is from the empirical finding that constraining visible depths performs better than constraining missing depths or a mixture of visible and missing depths. However, a more thorough investigation is necessary to investigate how GPRT could be



(a) Fountain-P11

(b) Small dinosaur

Figure 2. Some demonstration of success rates of VarPro and joint optimization algorithms for different η values. The success rate counts how many times out of some fixed number of runs each algorithm yields the best observed optimum for each setting of η from arbitrary initialization. On Fountain-P11, which is a relatively "easy" dataset with dense and mostly unique correspondences, the success rates are less sensitive to the value of η across all tested algorithms. However, on the small dinosaur sequence, which is a more difficult dataset due to its banded missing data pattern, the success rates of all algorithms decrease as η decreases, and it can only be solved efficiently by VarPro above certain value of η.

further generalized for handling these practical cases with noise and missing data.

3.5. Other optimization details

Optimization settings Table 1 shows the optimization settings used in our experiments.

Number of experimental runs At the time of hong18, we obtained results for each dataset from 50 runs of pOSE-based stratified bundle adjustment (BA) and 20 runs of GPRT-based stratified BA. We verified that an increase in the number of GPRT-based BA runs does not change the overall trend observed in Tables 1 and 3 of [4].

Robust kernel For robust projective and metric refinements, we applied the Cauchy loss function, which is defined in [1] as

$$\rho\left(s\right) := \ln(1+s), \tag{5}$$

where s is the sum of squared residuals. Above kernel is more robust to outliers than L_1 or Huber loss functions but less than Smooth truncated quadratic or Geman-McClure kernels (see [15] for a comprehensive list of loss functions).

In solving the above loss function, we use Trigg's correction [13] since this approach is already implemented in Ceres Solver [1]. Future work could incorporate the robust optimization strategy proposed in [14].

$SE(3)$ **parameterization** During the metric refinement stage, we use the axis-angle representation to formulate the rotation at each frame. We use the Jet library in Google Ceres Solver [1] to obtain auto-differentiated derivatives.

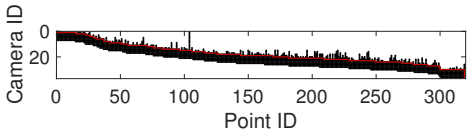
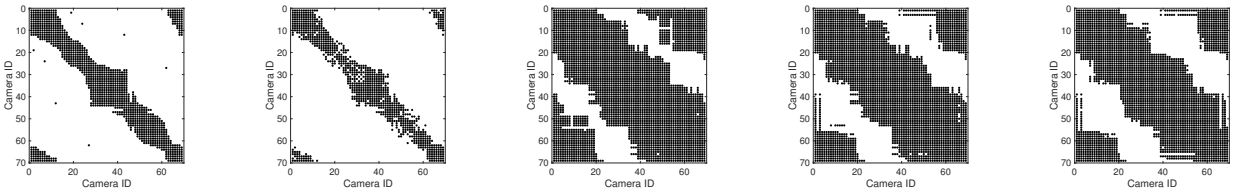


Figure 3. Visibility (black) of the Dino (S) dataset and our steplike mask (red) used to attempt GPRT [6] in presence of missing data. Note that the above steplike mask is only constraining depths which are visible (i.e. overlays on the black region).

Initializion for metric upgrade In the paper [4], solving (13) requires nonlinear least squares optimization, which can be sensitive to initialization. In [9], Pollefeys et al. proposed to first solve a linearized form of (13) in which the rank-3 constraint of $\tilde{\mathtt{H}}\tilde{\mathtt{H}}^\top \in \mathbb{R}^{4\times 4}$ is relaxed by replacing $\|\mathbf{c}\|^2$ with a new independent variable. The scale factors $\{\alpha_i\}$ are effectively eliminated by normalizing the scale of each camera prior to optimization and penalizing the differences between the diagonal entries of $\tilde{\mathtt{P}}_i\tilde{\mathtt{H}}(\mathbf{c})\tilde{\mathtt{H}}(\mathbf{c})^\top\tilde{\mathtt{P}}_i$. After the linear solution is computed, the rank-3 constraint is imposed by taking the closest rank-3 approximation via SVD. This is then used to initiate VarPro. (Empirical results in [7] show that VarPro does not have as wide basin of convergence for non-bilinear problems as it does for bilinear ones.)



(a) Two-views matched (b) Triplets matched (c) Points tracked [8] (d) Two-views revisited (e) Inlier tracks

Figure 4. Evolution of the camera adjacenty matrix for Vercingetorix [8] during the track generation stage illustrated in §5 of [4]. (d) is used for the experiment and (e) is outputted from the best solution obtained. Note that connecting pairwise matches to form point tracks usually **densifies** adjacency matrix.

4. Visualization of Camera connectivity

Figure 4 shows the evolution of the camera adjacency matrix over the track generation stage illustrated in §5 of [4]. This shows that converting triplet-verified pairwise matches to point tracks inevitably **densifies** connections as discussed in [4]. Figure 4 (d) also shows that revisiting two-view matches removes many geometrically inconsistent connections, taking one step closer to the baseline adjacency matrix shown i 4 (e).

5. Reconstructions

Reconstructions of the best obtained solutions are shown in Figures 5, 6 and 7.

References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org, 2014. 3

[2] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 1

[3] J. Dezert and C. Musso. An efficient method for generating points uniformly distributed in hyperellipsoids. Technical report, 2001. 1

[4] J. H. Hong and C. Zach. pose: Pseudo object space error for initialization-free bundle adjustment. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3, 4, 5, 6

[5] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *2013 IEEE International Conference on Computer Vision (ICCV)*, pages 3248–3255, 2013. 2

[6] B. Nasihatkon, R. Hartley, and J. Trumpf. A generalized projective reconstruction theorem and depth constraints for projective factorization. *International Journal of Computer Vision (IJCV)*, 115(2):87–114, 2015. 2, 3

[7] D. P. O'Leary and B. W. Rust. Variable projection for nonlinear least squares problems. *Computational Optimization and Applications*, 54(3):579–593, Apr 2013. 3

[8] C. Olsson and O. Enqvist. Stable structure from motion for unordered image collections. In *Proceedings of 17th Scandinavian Conference on Image Analysis (SCIA)*, pages 524–535, 2011. 1, 2, 4, 6

[9] M. Pollefeys, R. Koch, and L. V. Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision (IJCV)*, 32(1):7–25, 1999. 3

[10] J. L. Schönberger and J. M. Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2

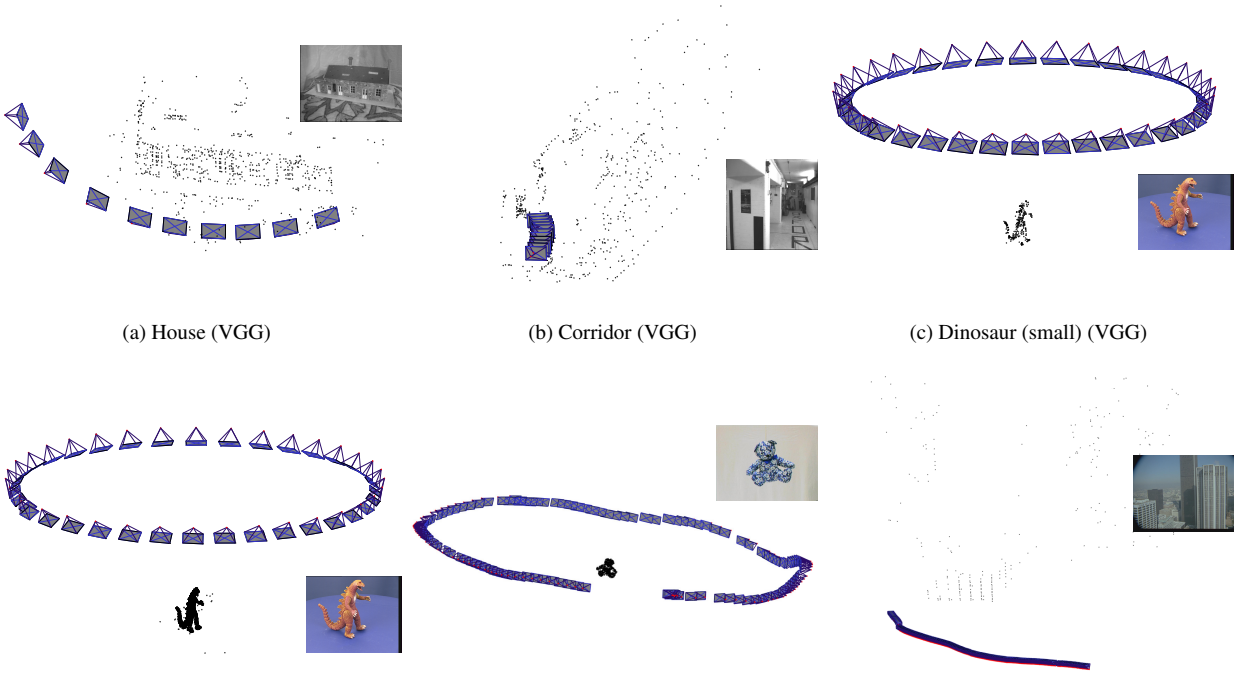
[11] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 1, 5

[12] C. Sweeney, T. Sattler, T. Höllerer, M. Turk, and M. Pollefeys. Optimizing the viewing graph for structure-from-motion. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 801–809, 2015. 2

[13] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - A modern synthesis. In *International Workshop on Vision Algorithms: Theory and Practice*, 1999 IEEE International Conference on Computer Vision (ICCVW), pages 298–372, 2000. 3

[14] C. Zach. Robust bundle adjustment revisited. In *13th European Conference on Computer Vision (ECCV)*, pages 772–787, 2014. 3

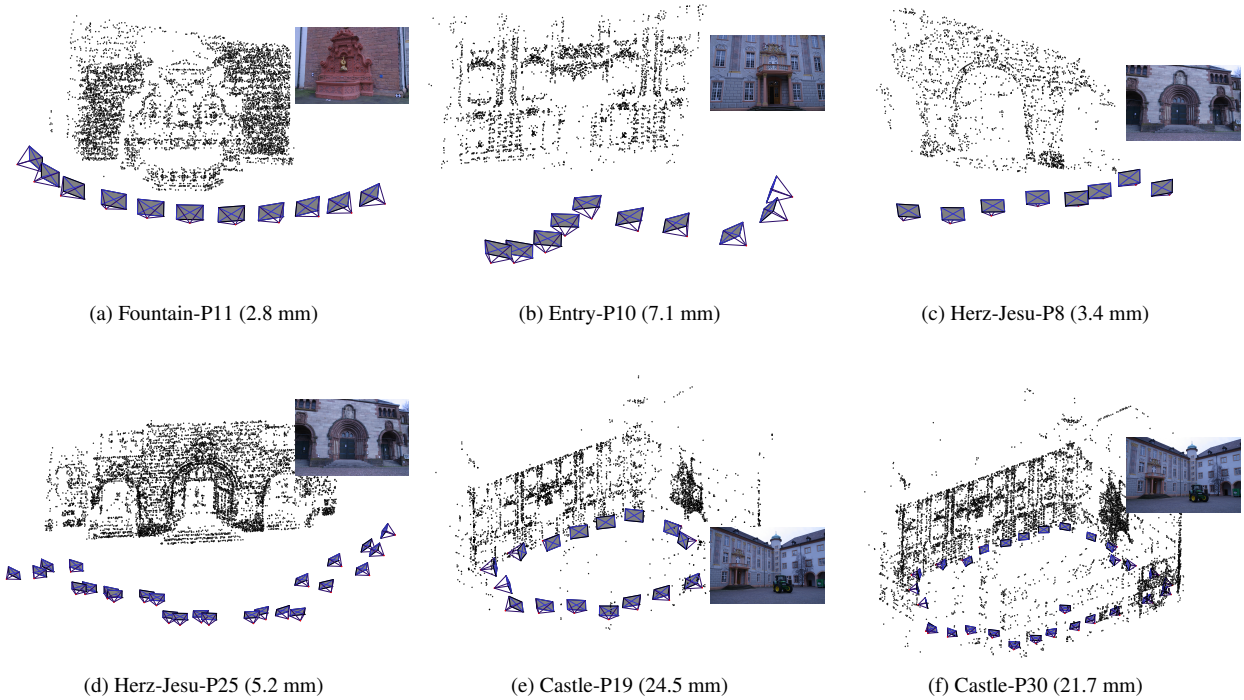
[15] C. Zach and G. Bourmaud. Iterated lifting for robust cost optimization. In *2017 British Machine Vision Conference (BMVC)*, 2017. 3



(a) House (VGG) (b) Corridor (VGG) (c) Dinosaur (small) (VGG)

(d) Dinosaur (large) (VGG) (e) Blue bear (Ponce) (f) Wilshire (Ponce)

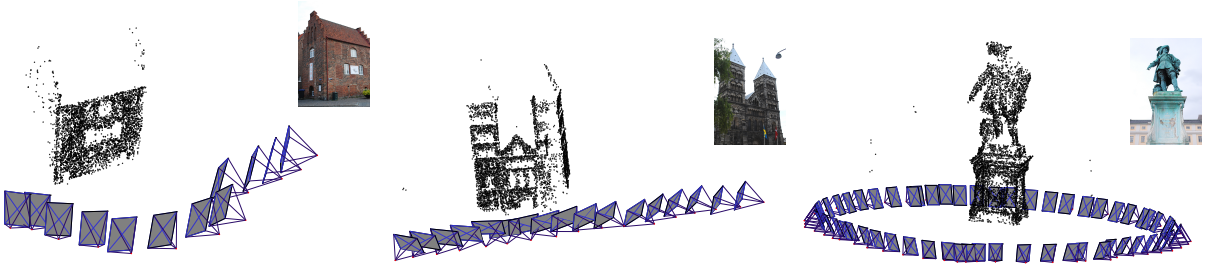
Figure 5. Successful reconstructions of point track sequences mentioned in Table 1 of [4]



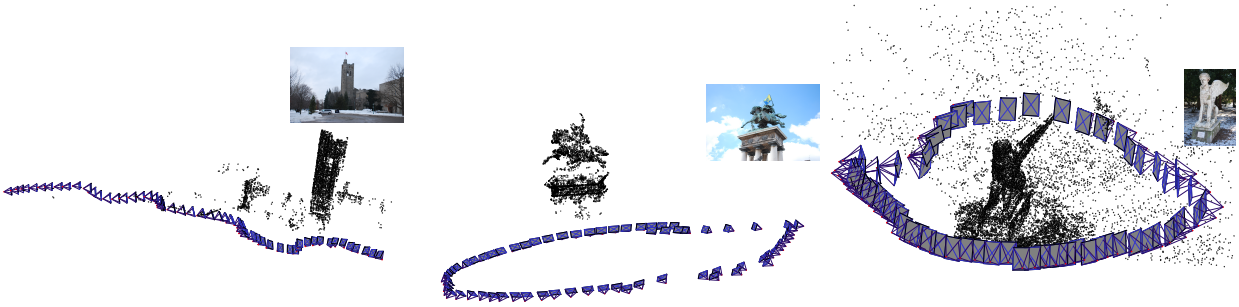
(a) Fountain-P11 (2.8 mm) (b) Entry-P10 (7.1 mm) (c) Herz-Jesu-P8 (3.4 mm)

(d) Herz-Jesu-P25 (5.2 mm) (e) Castle-P19 (24.5 mm) (f) Castle-P30 (21.7 mm)

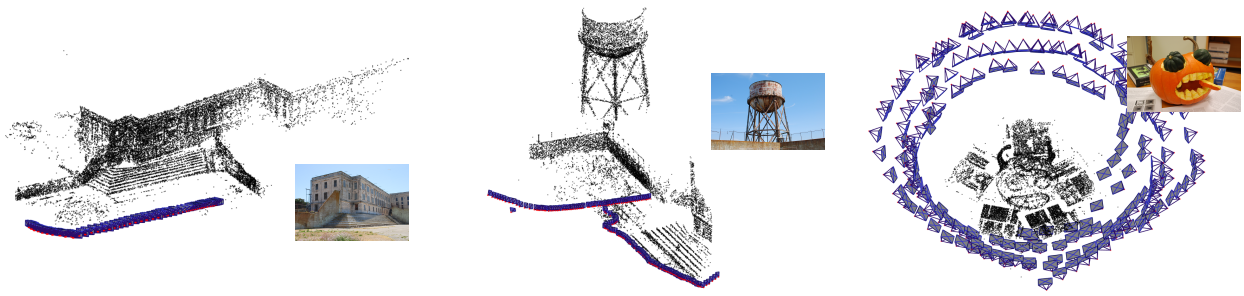
Figure 6. Successful reconstructions of Strecha's datasets [11] mentioned in Table 2 of [4]



(a) House Martenstorget (b) Lund Cathedral (small) (c) Gustav II Adolf



(d) Univ. of Western Ontario (small) (e) Vercingetorix (f) Lund Univ. Sphinx



(g) Alcatraz Courtyard (h) Water Tower (i) Pumpkin

Figure 7. Successful reconstructions of Olsson's datasets [8] mentioned in Table 2 of [4]