— Supplementary Material —
Finding Multiple Geometric Models by Clustering in the Consensus Space

Daniel Barath[1], Denys Rozumnyi[1,2], Ivan Eichhardt[3], Levente Hajder[3], Jiri Matas[2]
[1]Computer Vision and Geometry Group, ETH Zurich, Switzerland,
[2]VRG, Faculty of Electrical Engineering, CTU in Prague, Czech Republic,
[3]Eötvös Loránd University, Budapest, Hungary

1. Explanation of the Hyper-parameters

In this section, we describe the hyper-parameters of the proposed algorithm, their purpose and the ways to set them. Parameters of the proposed algorithm:

1. An upper-bound for the inlier-outlier threshold on the point-to-model residual used inside the MAGSAC++ scoring. This parameter is problem-dependent. It usually is defined in pixels. It is easier to set [3] than the usual inlier-outlier threshold of RANSAC.

2. Parameter q_{min} is similar to what structure-from-motion algorithms use to decide if the relative pose of an image pair is estimated successfully. For example, COLMAP [5] uses q_{min} = 15, we use 20.

3. The termination confidence is the same as in RANSAC. Its typical values are 0.95 and 0.99. We use 0.99 in our experiments.

4. The model-to-model distance threshold is from interval $\in [0, 1]$. It measures the overlap of the inlier sets of two models (0 - non-overlapping, 1 - fully overlapping). Setting it to 0.2 works on a wide range of problems and datasets.

2. SfM Results in Section 4.2

Detailed results. The results of the global SfM from [6] on each scene from the 1DSfM dataset are reported in Table 1. Note that we omitted the results on scenes Gendarmenmarkt and Union Square since [6] failed to reconstruct them with all tested pose-graph estimation techniques.

Additional visualizations are put in Figures 1 and 2, where the top rows show the results of [6] when initialized by a pose-graph estimated in the proposed way, exploiting an essential matrix and multiple homographies. The bottom rows show results when the pose-graph is estimated from essential matrices in the traditional way. Colored ellipses mutually highlight parts of the two reconstructions with noticeable differences. The traditional approach leads to reconstructions with fewer details and reduced precision compared to the proposed technique.

3. Translation from Known Rotation

In Section 4.2., we propose to estimate the relative pose from multiple homographies and the essential matrix by decomposing them and choosing the pose that leads to the most inliers when thresholding the re-projection error. We found that, while the estimated rotation matrix often is accurate, the translation can be improved by re-estimating it from the found inliers considering the known rotation.

In this section, we briefly describe the translation estimation procedure given a known rotation matrix. It is well-known [2] that the essential matrix is defined as

$$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R},$$

where $\mathbf{t} \in \mathbb{R}^3$ and $\mathbf{R} \in \text{SO}(3)$ are, respectively, the translation vector and rotation matrix, and $[\mathbf{t}]_\times$ is the cross-product matrix of $\mathbf{t}$ as follows:

$$[\mathbf{t}]_\times = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}.$$

Essential matrix $\mathbf{E}$ describes the relationship of a point correspondence in the images via the well-known epipolar constraint as follows:

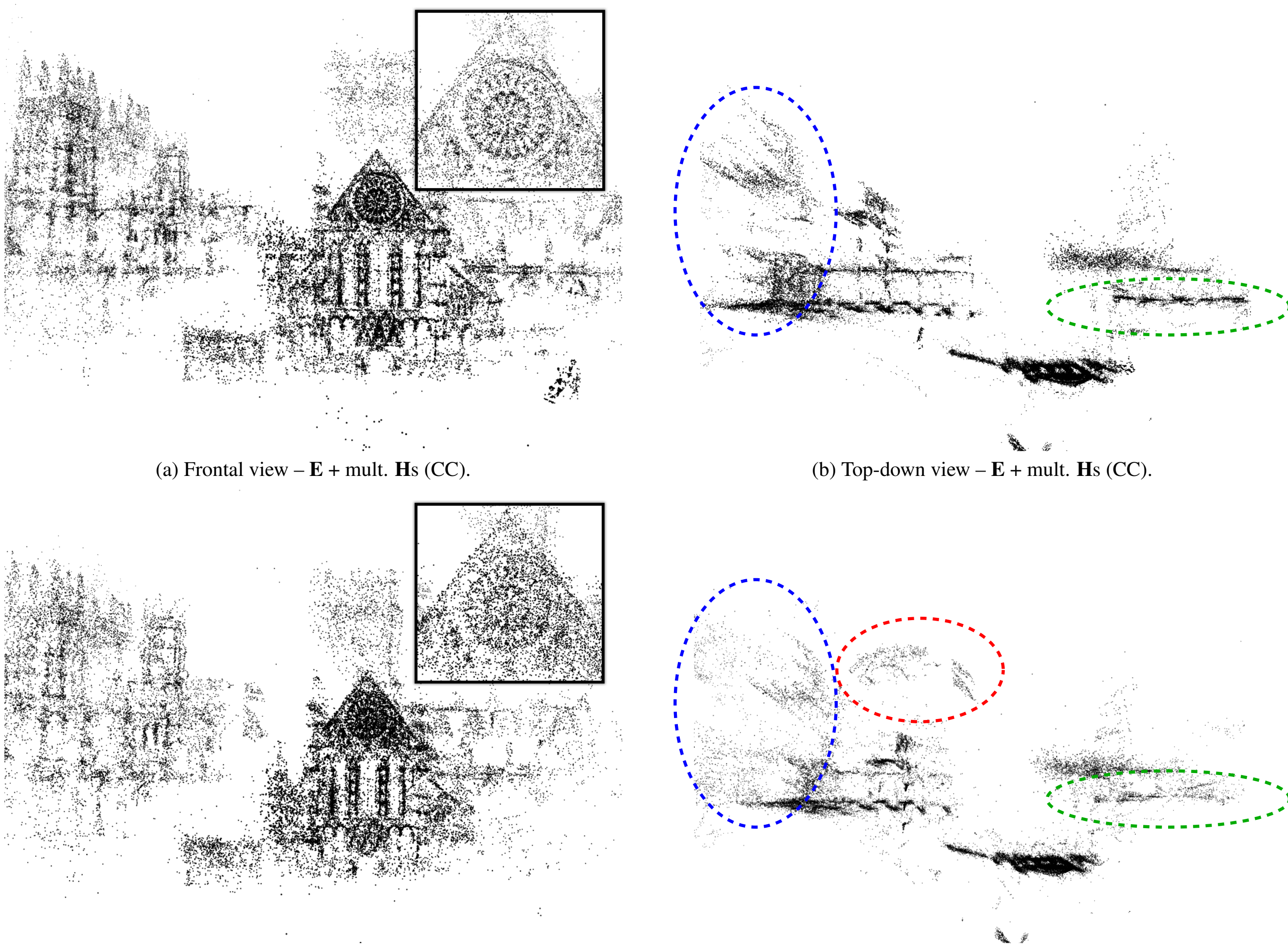
$$\mathbf{p}_2^\text{T} \mathbf{E} \mathbf{p}_1 = 0,$$

where $\mathbf{p}_1 = [u_1 \; v_1 \; w_1]^\text{T}$ and $\mathbf{p}_2 = [u_2 \; v_2 \; w_2]^\text{T}$ are homogeneous points in the normalized image plane, *i.e.*, normalized by the intrinsic camera matrices. Considering $\mathbf{R}$ to be known, we are given the following constraint

$$\mathbf{p}_2^\text{T} [\mathbf{t}]_\times \mathbf{R} \mathbf{p}_1 = 0,$$

Table 1. Results of the global SfM algorithm from [6] on the scenes from the 1DSfM dataset [7] when initialized by the pose-graph estimated from essential matrices (**E** matrix), and the proposed method combined either with Progressive NAPSAC [1] or the proposed Connected Components (CC) samplers. As ground truth, we used reconstructions from COLMAP [5]. The averages and average medians of the rotation and position errors are reported in Table 1.

				orientation err (°)			position err (m)			focal err ($\times 10^{-2}$)		
		# views	# tracks	AVG	MED	STD	AVG	MED	STD	AVG	MED	STD
Alamo	**E** matrix	493	104 894	**2.46**	**0.59**	3.76	**1.60**	**1.36**	**3.98**	**0.02**	**0.01**	**0.05**
	E + mult. **H**s	**495**	**110 243**	2.80	0.81	3.91	1.79	1.88	4.73	**0.02**	**0.01**	**0.05**
	E + mult. **H**s (CC)	494	105 920	2.59	0.62	**3.63**	1.68	1.58	4.19	**0.02**	**0.01**	**0.05**
Ellis Isl.	**E** matrix	211	**31 200**	4.21	2.90	4.69	5.59	3.43	10.57	**0.02**	**0.01**	**0.02**
	E + mult. **H**s	210	30 610	**3.49**	**2.33**	**3.00**	**4.27**	**3.09**	**8.22**	**0.02**	**0.01**	**0.02**
	E + mult. **H**s (CC)	**215**	31 182	4.61	2.61	3.87	5.86	3.89	11.59	**0.02**	**0.01**	**0.02**
Madrid M.	**E** matrix	299	56 102	11.38	0.69	14.50	1.09	8.06	1.36	**0.06**	**0.03**	**0.10**
	E + mult. **H**s	**327**	50 438	**4.00**	**0.30**	**5.63**	**0.60**	**2.86**	**0.90**	0.07	**0.03**	0.14
	E + mult. **H**s (CC)	298	**57 457**	8.06	0.58	12.11	1.00	4.77	1.18	**0.06**	**0.03**	**0.10**
Montreal	**E** matrix	432	106 101	**1.34**	**0.41**	8.64	**0.82**	**0.38**	**1.22**	**0.02**	**0.01**	**0.03**
	E + mult. **H**s	435	**106 498**	1.52	0.46	**7.84**	0.89	0.47	1.31	**0.02**	**0.01**	**0.03**
	E + mult. **H**s (CC)	**436**	104 802	1.45	0.46	8.03	0.97	0.45	1.61	**0.02**	**0.01**	**0.03**
NYC Lib.	**E** matrix	270	**57 235**	53.59	14.08	**3.86**	14.10	52.95	7.26	**0.03**	**0.01**	**0.04**
	E + mult. **H**s	**271**	56 435	**5.20**	**2.94**	3.96	**4.97**	**4.23**	6.73	**0.03**	**0.01**	**0.04**
	E + mult. **H**s (CC)	270	55 418	6.44	3.11	4.26	5.04	5.54	**6.59**	**0.03**	**0.01**	**0.04**
Piazza d. P.	**E** matrix	**291**	42 823	7.24	3.82	3.33	4.91	7.61	4.34	**0.03**	**0.02**	**0.04**
	E + mult. **H**s	288	**44 457**	6.99	3.32	3.26	4.16	7.51	4.19	**0.03**	**0.02**	0.05
	E + mult. **H**s (CC)	**291**	43 510	**5.37**	**2.53**	**1.46**	**3.28**	**5.46**	**3.54**	**0.03**	**0.02**	**0.04**
Piccadilly	**E** matrix	**1869**	210 821	**4.71**	0.35	**13.53**	0.70	2.00	**1.05**	**0.05**	**0.03**	0.15
	E + mult. **H**s	1656	141 661	10.15	0.48	24.75	0.87	2.55	1.11	**0.05**	**0.03**	**0.14**
	E + mult. **H**s (CC)	1860	**220 045**	4.96	**0.31**	14.83	**0.66**	**1.68**	**1.05**	**0.05**	**0.03**	0.15
Roman F.	**E** matrix	989	**208 457**	4.87	**14.76**	4.68	**22.25**	3.86	82.77	**0.03**	**0.02**	**0.07**
	E + mult. **H**s	991	204 432	**4.56**	15.64	**3.37**	22.90	**3.78**	82.49	**0.03**	**0.02**	**0.07**
	E + mult. **H**s (CC)	**995**	206 641	4.85	15.78	3.59	23.61	4.01	**82.29**	**0.03**	**0.02**	**0.07**
Tower	**E** matrix	**406**	**96 481**	6.03	**9.48**	12.55	**25.04**	**2.42**	**38.79**	**0.02**	**0.01**	**0.03**
	E + mult. **H**s	397	95 394	**5.29**	10.58	**6.29**	26.47	3.39	40.70	**0.02**	**0.01**	**0.03**
	E + mult. **H**s (CC)	405	96 088	5.83	10.94	8.87	26.56	3.54	40.54	**0.02**	**0.01**	**0.03**
Trafalgar	**E** matrix	**4111**	**354 494**	18.03	16.79	32.09	23.92	10.70	**29.63**	**0.02**	**0.01**	**0.03**
	E + mult. **H**s	4097	349 621	19.10	**16.14**	41.86	**23.74**	**7.20**	30.73	**0.02**	**0.01**	**0.03**
	E + mult. **H**s (CC)	4088	349 784	**18.00**	17.09	**31.97**	24.79	10.93	30.64	**0.02**	**0.01**	**0.03**
Vienna C.	**E** matrix	705	160 363	14.47	7.49	9.86	10.96	9.40	11.55	**0.02**	**0.01**	**0.05**
	E + mult. **H**s	612	92 051	26.35	13.79	29.71	22.85	13.10	25.45	**0.02**	**0.01**	**0.05**
	E + mult. **H**s (CC)	**707**	**160 503**	**4.72**	**6.97**	**4.84**	**10.15**	**3.18**	**11.03**	**0.02**	**0.01**	**0.05**
Yorkmins.	**E** matrix	399	98 396	5.52	7.61	3.57	12.13	4.99	17.70	**0.03**	**0.01**	**0.04**
	E + mult. **H**s	**402**	100 985	5.68	7.74	3.46	12.68	5.11	20.03	**0.03**	**0.01**	**0.04**
	E + mult. **H**s (CC)	399	**109 132**	**3.49**	**6.27**	**2.90**	**11.26**	**2.91**	**17.12**	**0.03**	**0.01**	**0.04**



(a) Frontal view – **E** + mult. **H**s (CC).

(b) Top-down view – **E** + mult. **H**s (CC).

(c) Frontal view – **E** matrices only.

(d) Top-down view – **E** matrices only.

Figure 1. Visual comparison of the reconstructions of Yorkminster by [6] when initialized by the proposed (**E** + mult. **H**s (CC); top row) and traditional (**E** matrices; bottom) techniques. Blue and green ellipses highlight areas that the proposed algorithm reconstructs significantly more accurately than the traditional approach. The red ellipse points to an erroneous area. “CC” stands for using the proposed sampler in the proposed method for multi-homography fitting.

where the only unknowns are the three translation components $\mathbf{t} = [t_x \; t_y \; t_z]^{\mathrm{T}}$. Multiplication $\mathbf{R}\mathbf{p}_1$ can be pre-calculated as $\mathbf{p}_1' = \mathbf{R}\mathbf{p}_1$. Formula $\mathbf{p}_2^{\mathrm{T}} [\mathbf{t}]_\times \mathbf{p}_1'$ leads to:

$$-u_2 t_z v_1' + u_2 t_y w_1' + v_2 t_z u_1' - v_2 t_x w_1' - w_2 t_y u_1' + w_2 t_x v_1' = 0 \tag{1}$$

Eq. 1 is linear in the elements of the translation vector. Therefore, the equation can be reformulated as

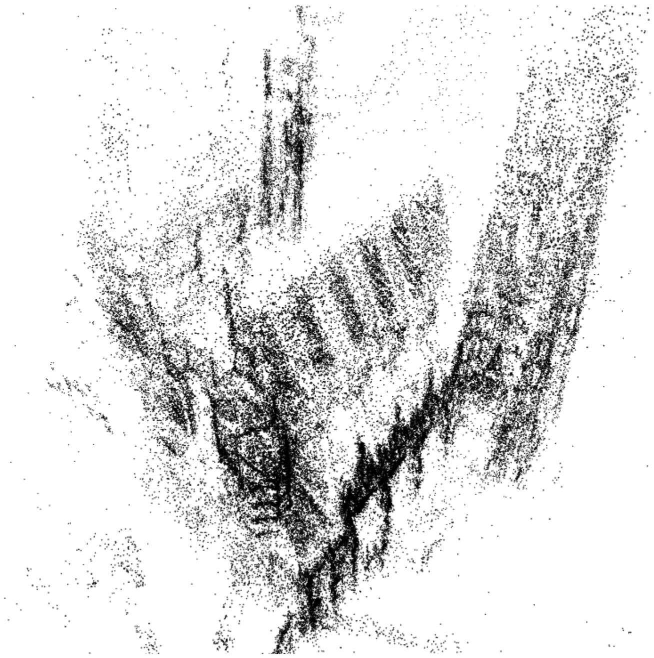
$$\begin{bmatrix} v_1' w_2 - w_1' v_2 \\ u_2 w_1' - w_2 u_1' \\ v_2 u_1' - u_2 v_1' \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} = 0.$$

If at least two point correspondences are given, a homogeneous linear system of equations is obtained. The optimal solution, in the LSQ sense, is given via calculating the null-vector of the coefficient matrix.

4. Trajectories of Fast-moving Objects

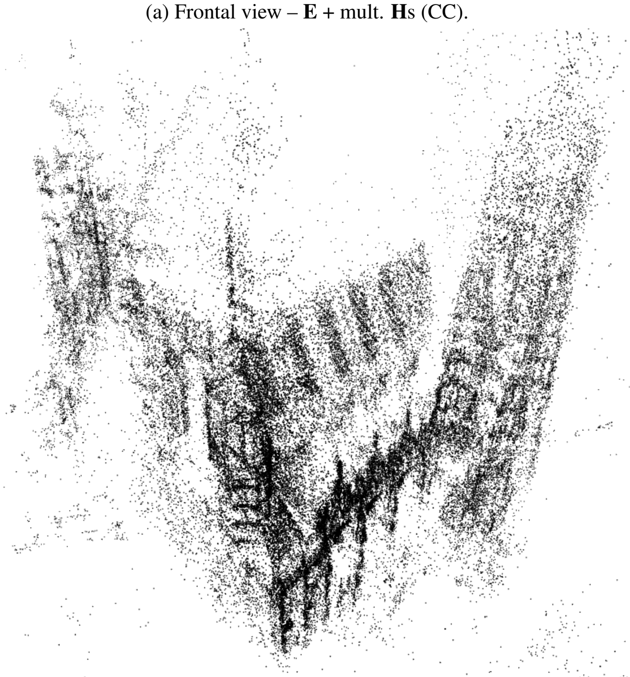
We show example visualizations of trajectory estimation of fast-moving objects in Figure 3. After extracting blur kernels that encode the object motion, we apply a multi model fitting algorithm recovering line segments. The estimated line segments are colored in red. The ground truth line segments are generated by applying a classical state-of-the-art object tracking algorithm on high-speed camera footage with manual annotations, which is shown in green. We show the results of sequential RANSAC as originally proposed in [4]. Additionally, we show final trajectories after filtering and refinement by [4]. Quantitative results are reported in the paper.

Notice that the line segments found by seq. RANSAC are not continuous, *i.e.*, there is a clear gap between all of them.

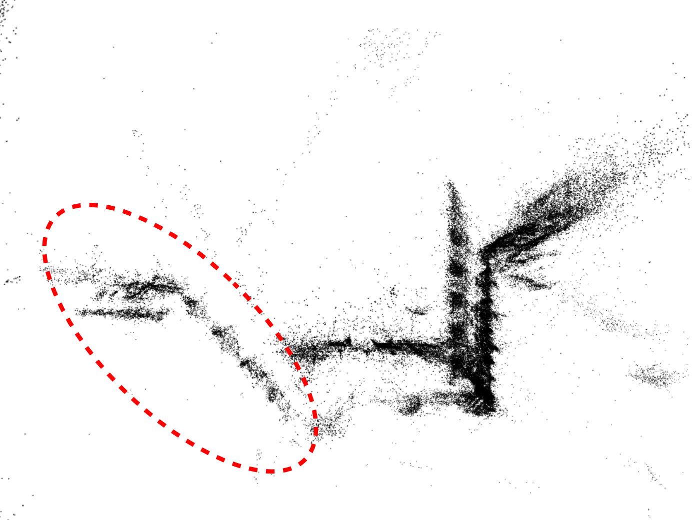


(a) Frontal view – **E** + mult. **H**s (CC).

(b) Top-down view – **E** + mult. **H**s (CC).



(c) Frontal view – **E** matrices only.



(d) Top-down view – **E** matrices only.

Figure 2. Visual comparison of the reconstructions of Vienna Cathedral by [6] when initialized by the proposed (**E** + mult. **H**s (CC); top) and traditional (**E** matrices; bottom) techniques. The proposed approach preserves the parallelism of the walls of the cathedral (red ellipse). "CC" stands for using the proposed sampler in the proposed method for multi-homography fitting.

This is caused by the hard point-to-line assignment used in seq. RANSAC and in the state-of-the-art multi-model fitting algorithms. Using the proposed method allows finding continuous chains that lead to better trajectories as shown in the last column and, also, in Table 3 in the main paper.

References

[1] Daniel Barath, Maksym Ivashechkin, and Jiri Matas. Progressive NAPSAC: sampling from gradually growing neighborhoods. *arXiv preprint arXiv:1906.02295*, 2019. 2

[2] Richard Hartley and Andrew Zisserman. *Multiple view geom-*

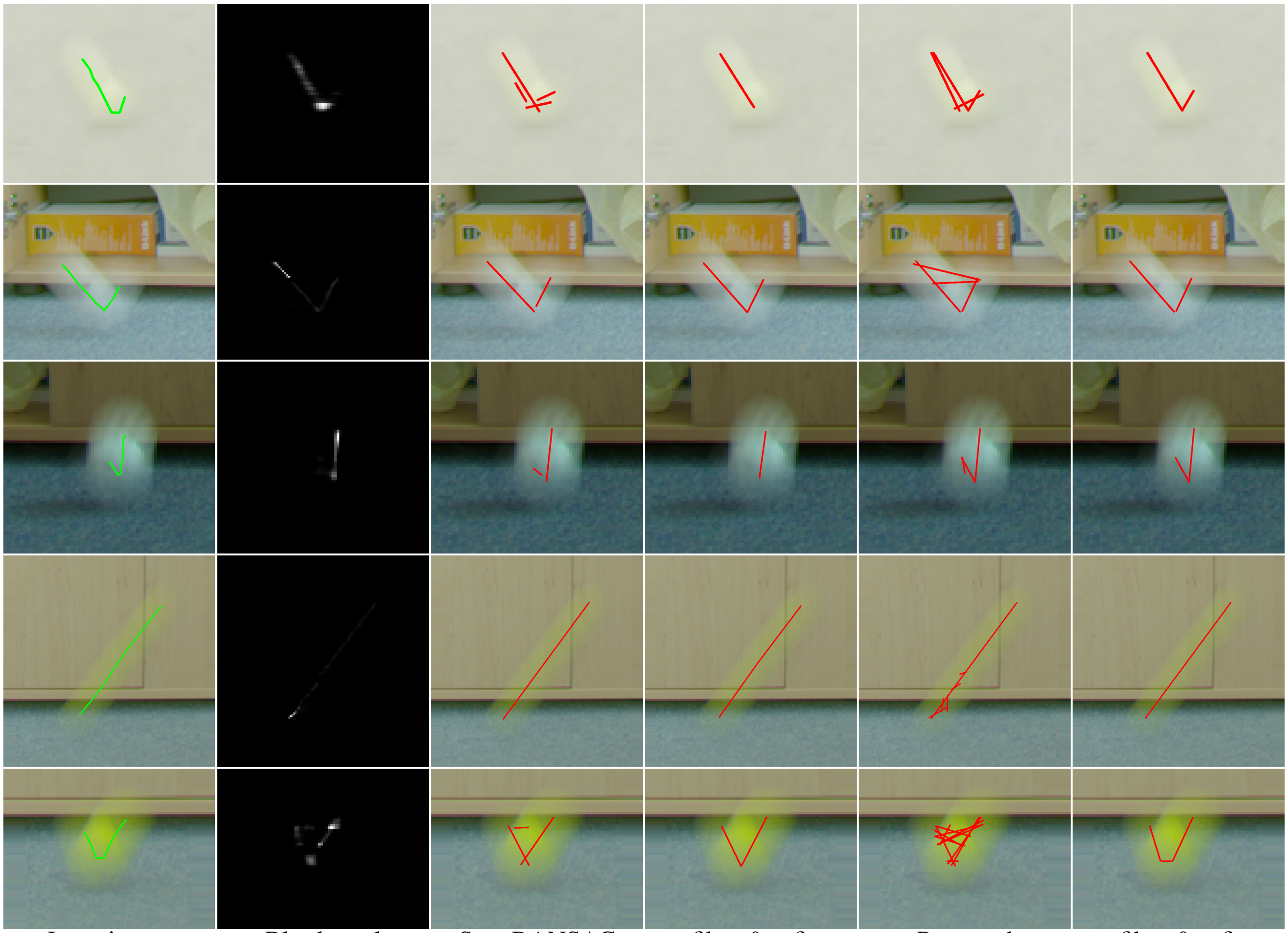


Figure 3. Fitting multiple line segments for trajectory estimation of fast-moving objects. The estimated and ground truth segments are colored by red and green, respectively. The original Tracking by Deblatting [4] method for trajectory estimation of fast-moving objects uses the sequential RANSAC algorithm. Therefore, we report results using their implementation. The filtering and refinement are done by the method proposed in [4]. After post-processing by filtering and refinement, the results from the proposed algorithm more often cover the sought trajectory than by the other methods. The results of seq. RANSAC, besides being qualitatively worse, *i.e.* missing a segment in rows 1, 3, and 5, suffer from the single-model assignment of inliers which shows as a gap between consecutive segments. The width of the gap equals to the inlier threshold of seq. RANSAC.

etry in computer vision. Cambridge University Press, 2003. 1

[3] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 2021. 1

[4] Jan Kotera, Denys Rozumnyi, Filip Šroubek, and Jiri Matas. Intra-frame object tracking by deblatting. In *International Conference on Computer Vision Workshops*, Oct 2019. 3, 5

[5] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1, 2

[6] Christopher Sweeney, Tobias Hollerer, and Matthew Turk. Theia: A fast and scalable structure-from-motion library. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 693–696, 2015. 1, 2, 3, 4

[7] Kyle Wilson and Noah Snavely. Robust global translations with 1DSfM. In *Proc. European Conf. on Computer Vision*, 2014. 2