Supplementary Material: Relative Pose from a Calibrated and an Uncalibrated Smartphone Image

Yaqing Ding[1,4], Daniel Barath[2], Jian Yang[1], Zuzana Kukelova[3]
[1] School of Computer Science and Engineering, Nanjing University of Science and Technology
[2] Computer Vision and Geometry Group, Department of Computer Science, ETH Zürich
[3] Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague
[4] Centre for Mathematical Sciences, Lund University
dingyaqing@njust.edu.cn

1. Introduction

This supplementary material contains the following.

1. Section 2 proves Property 2 in the main paper.
2. Section 3 provides the translation errors of the compared algorithms.
3. Section 4 provides example images from the captured PHONE dataset as well as some additional results.

2. Proof of Property 2

Suppose that we are given an $N \times 3$ polynomial matrix $\mathbf{A}$ with its i^{th} row of the form

$$\mathbf{A}_{(i,:)} = (\mathbf{R}'_{align}[u'_i, v'_i, f]^\top) \times (\mathbf{R}_y \mathbf{p}_i). \tag{1}$$

Let I be an index set and $\mathbf{A}_I$ a 3×3 submatrix of the matrix $\mathbf{A}$ that contains the rows complementary to I. Then the following property holds:.

Property 2: *Determinants of* 2×2 *submatrices* $\mathcal{A}_{12}, \mathcal{A}_{22}, \mathcal{A}_{32}$ *of the matrix* $\mathcal{A} = \mathbf{A}_I$ *have* $1 + \sigma^2$ *as a common factor.*

Proof. The homogeneous coordinates of points $\mathbf{p}_i$ and $\mathbf{R}'_{align}[u'_i, v'_i, f]^\top$ do not contain σ. Therefore, multiplying them by some scalars will not affect Property 2. Hence, we can assume that $\mathbf{p}_i \sim [a_i, b_i, 1]^\top$, and $\mathbf{R}'_{align}[u'_i, v'_i, f]^\top \sim [a'_i, b'_i, 1]^\top$. Let the 3×3 matrix $\mathcal{A} = \mathbf{A}_I$ contain rows of the form (1) that correspond to point correspondences $\mathbf{p}_i \sim [a_i, b_i, 1]^\top$, and $\mathbf{R}'_{align}[u'_i, v'_i, f]^\top \sim [a'_i, b'_i, 1]^\top$ for $i = 1, \ldots, 3$. Then the submatrix $\mathcal{A}_{32}$ of the matrix $\mathcal{A}$ has the form:

$$\begin{bmatrix} (1-\sigma^2)b'_1 - (1+\sigma^2)b_1 - 2\sigma a_1 b'_1 & (1+\sigma^2)a'_1 b_1 - 2\sigma b'_1 - (1-\sigma^2)a_1 b'_1 \\ (1-\sigma^2)b'_2 - (1+\sigma^2)b_2 - 2\sigma a_2 b'_2 & (1+\sigma^2)a'_2 b_2 - 2\sigma b'_2 - (1-\sigma^2)a_2 b'_2 \end{bmatrix}. \tag{2}$$

Using the identity $(1-\sigma^2)^2 + (2\sigma)^2 = (1+\sigma^2)^2$ and by expanding $\det(\mathcal{A}_{32})$ we can find that it has the following form:

$$\begin{aligned} \det(\mathcal{A}_{32}) =& (b_1 b_2 + b'_1 b'_2)(a'_1 - a'_2)(1+\sigma^2)^2 + \ldots \\ & (1+\sigma^2)(b_1 b'_2(2\sigma - a_2(\sigma^2-1)) - b_2 a'_2 b'_1(\sigma^2 + 2a_1\sigma - 1) + b_1 a'_1 b'_2(\sigma^2 + 2a_2\sigma - 1) - b_2 b'_1(2\sigma - a_1(\sigma^2-1))), \end{aligned} \tag{3}$$

which is divisible by $1 + \sigma^2$. Similarly, we can prove that $\det(\mathcal{A}_{12})$ and $\det(\mathcal{A}_{22})$ are divisible by $1 + \sigma^2$. □

3. Translation Errors

Fig. 1 shows the translation errors for general camera motion as a function of the image noise (in pixels), field-of-view (in degrees), baseline, and the gravity vector noise (in degrees) for the synthetic experiment described in Section 6 of the main paper. Fig. 2 shows translation errors for pure translation as a function of the image noise, field-of-view, baseline, and the gravity vector noise. Due to the ambiguity between focal length and translation for this type of motion, the state-of-the-art solvers E5f_v and F7 fail to recover the correct translation and focal length. Finally, Fig. 3 shows translation errors for a planar scene as a function of the image noise, field-of-view, baseline, and the gravity vector noise. We can see that our solver E4f_s provides the most accurate results.

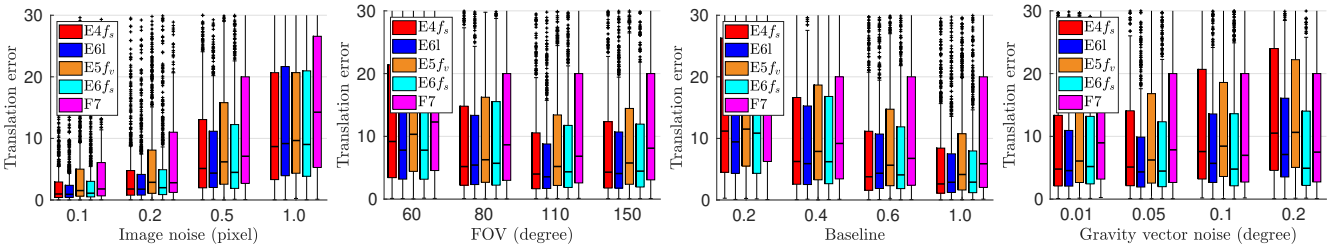

Figure 1. From **Left** to **Right**: the translation error (in degrees) of the proposed (E4f_s, E6l) and state-of-the-art solvers w.r.t. increasing image noise, field-of-view, baseline, and gravity vector noise under general motion.

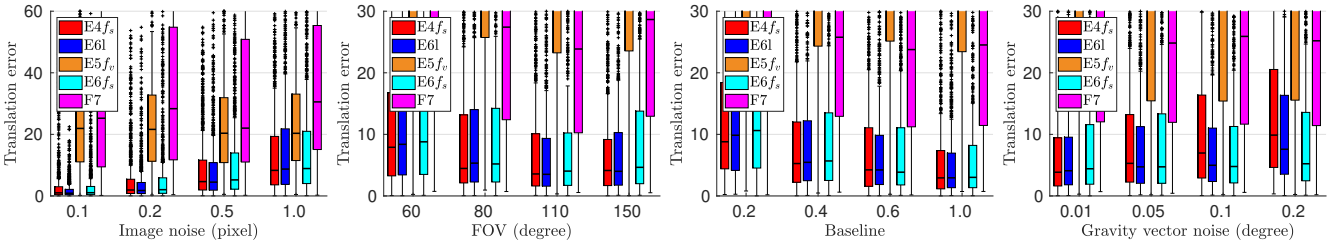

Figure 2. From **Left** to **Right**: the translation error (in degrees) of the proposed (E4f_s, E6l) and state-of-the-art solvers w.r.t. increasing image noise, field-of-view, baseline, and gravity vector noise under pure translation.

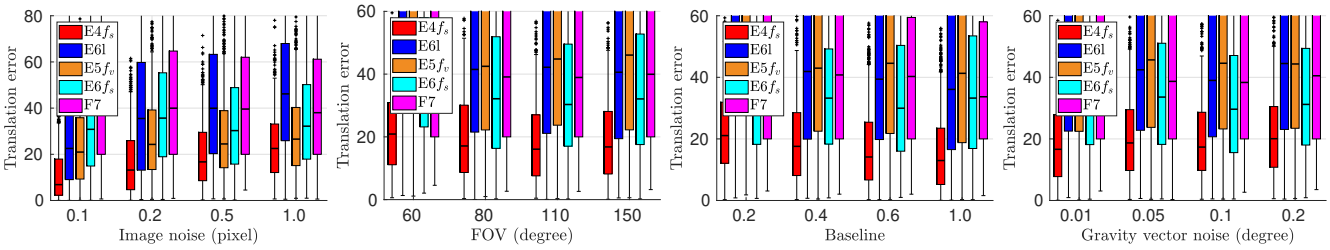

Figure 3. From **Left** to **Right**: the translation error (in degrees) of the proposed (E4f_s, E6l) and state-of-the-art solvers w.r.t. increasing image noise, field-of-view, baseline, and gravity vector noise under planar structures.

4. PHONE Dataset

Fig. 4 shows example images from the new PHONE dataset. The dataset is recorded by using different smartphones (iPhone 6s and iPhone 11). The sequences were captured at 30Hz with the rear camera, and the corresponding IMU data were captured at 100Hz with the built-in sensor. In addition, the sequences cover all the camera configurations we discussed in the synthetic evaluation: general motion, pure translation and rotation, and planar scenes. To obtain ground truth, we calibrated the phones and used the RealityCapture [1] software to obtain camera poses and 3D reconstructions. In total, $12,464$ image pairs with synchronized gravity directions, ground truth poses, calibrations and 3D reconstructions were generated.

The cumulative distribution functions (CDF) of the rotation, translation and focal length errors, run-time, iteration number, and inlier number for the experiment on the PHONE dataset (described in Section 7 in the main paper) are shown in Fig. 5. Being accurate is interpreted as a curve close to the top-left corner. Both of the proposed solvers, E4f_s and E6f_s, lead to more accurate rotation, translation, and focal length estimates than the tested SOTA ones.

Figure 4. Example images from the PHONE dataset.

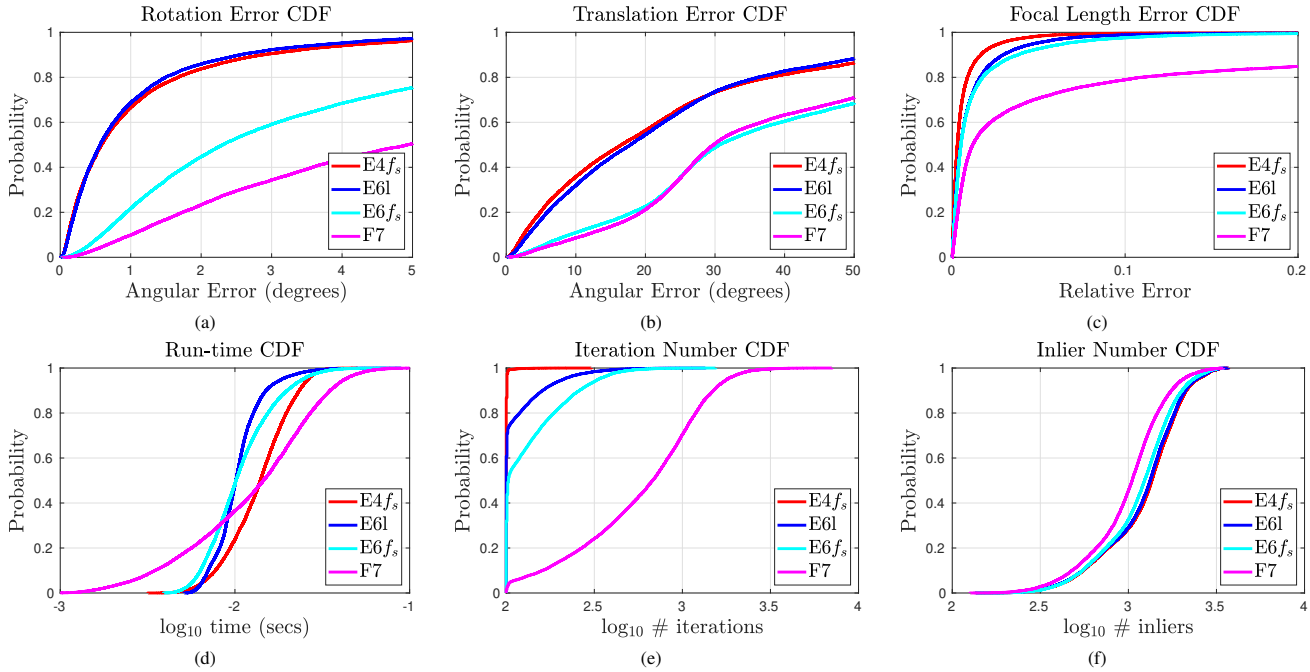

Figure 5. The cumulative distribution functions of the (a) rotation, (b) translation (both in degrees), (c) focal length errors, (d) run-time (in seconds), (e) iteration, and (f) inlier numbers of GC-RANSAC [2] on the PHONE datasets (12, 464 image pairs). Being accurate is interpreted as a curve close to the top-left corner.

References

[1] Realitycapture. http://www.capturingreality.com. 2

[2] Daniel Barath and Jiří Matas. Graph-cut RANSAC. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3