Supplementary Materials: Fast Monocular Visual-Inertial Initialization Leveraging Learned Single-View Depth

Nathaniel Merrill Patrick Geneva Saimouli Katragadda Chuchu Chen Guoquan Huang

I. Single-Image Depth Aided Initialization

We now consider we are given a *single* affine-invariant (up-to scale and offset) depth map, $\mathbf{D}$, in the first frame of reference at time t_0. We formulate all features as a function of this depth map and the feature bearing in the first camera frame $\{C_0\}$. The minimal state we wish to recover is:

$$\mathbf{x}' = \begin{bmatrix} a & b & {}^{I_0}\mathbf{v}_{I_0}^\top & {}^{I_0}\mathbf{g}^\top \end{bmatrix}^\top \tag{1}$$

where we have assumed that the affine-invariant depth map $\mathbf{D}$ is sufficiently accurate and can provide an estimate of the 3D structure in front of the camera up to a scale a and offset parameter b from just a single frame [1].

A. Inertial Measurement Model

The inertial measurement unit (IMU) provides angular velocities $\boldsymbol{\omega}$ and linear accelerations $\mathbf{a}$ in the inertial frame. These can be used to recover how the state evolves from one timestep to the next with the following state dynamics:

$$\prescript{I_{k+1}}{G}{\mathbf{R}} = \prescript{I_{k+1}}{I_k}{\Delta\mathbf{R}}\ \prescript{I_k}{G}{\mathbf{R}} \tag{2}$$

$${}^{G}\mathbf{p}_{I_{k+1}} = {}^{G}\mathbf{p}_{I_k} + {}^{G}\mathbf{v}_{I_k}\Delta T - \frac{1}{2}{}^{G}\mathbf{g}\Delta T^2 + \prescript{I_k}{G}{\mathbf{R}}^\top \int_{t_k}^{t_{k+1}} \int_{t_k}^{s} \prescript{I_k}{u}{\Delta\mathbf{R}} \left(\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a\right) duds \tag{3}$$

$${}^{G}\mathbf{v}_{I_{k+1}} = {}^{G}\mathbf{v}_{I_k} - {}^{G}\mathbf{g}\Delta T + \prescript{I_k}{G}{\mathbf{R}}^\top \int_{t_k}^{t_{k+1}} \prescript{I_k}{u}{\Delta\mathbf{R}} \left(\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a\right) du \tag{4}$$

$$\mathbf{b}_{\omega_{k+1}} = \mathbf{b}_{\omega_k} + \int_{t_k}^{t_{k+1}} \mathbf{n}_{\omega b}\ du \tag{5}$$

$$\mathbf{b}_{a_{k+1}} = \mathbf{b}_{a_k} + \int_{t_k}^{t_{k+1}} \mathbf{n}_{ab}\ du \tag{6}$$

From the above, we define the following integration terms:

$${}^{I_k}\boldsymbol{\alpha}_{I_{k+1}} = \int_{t_k}^{t_{k+1}} \int_{t_k}^{s} \prescript{I_k}{u}{\Delta\mathbf{R}} \left(\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a\right) duds \tag{7}$$

$${}^{I_k}\boldsymbol{\beta}_{I_{k+1}} = \int_{t_k}^{t_{k+1}} \prescript{I_k}{u}{\Delta\mathbf{R}} \left(\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a\right) du \tag{8}$$

We then remove the global frame by integrating relative to the first inertial frame. This can be derived for the position as:

$$\begin{aligned}
{}^{I_0}\mathbf{p}_{I_{k+1}} &= \prescript{I_0}{G}{\mathbf{R}}({}^{G}\mathbf{p}_{I_{k+1}} - {}^{G}\mathbf{p}_{I_0}) && (9)\\
&= \prescript{I_0}{G}{\mathbf{R}} \left({}^{G}\mathbf{p}_{I_k} + {}^{G}\mathbf{v}_{I_k}\Delta T - \frac{1}{2}{}^{G}\mathbf{g}\Delta T^2 + \prescript{I_k}{G}{\mathbf{R}}^\top {}^{I_k}\boldsymbol{\alpha}_{I_{k+1}} - {}^{G}\mathbf{p}_{I_0} \right) && (10)\\
&= \prescript{I_0}{G}{\mathbf{R}}({}^{G}\mathbf{p}_{I_k} - {}^{G}\mathbf{p}_{I_0}) + \prescript{I_0}{G}{\mathbf{R}}{}^{G}\mathbf{v}_{I_k}\Delta T - \frac{1}{2}\prescript{I_0}{G}{\mathbf{R}}{}^{G}\mathbf{g}\Delta T^2 + \prescript{I_0}{G}{\mathbf{R}}\prescript{I_k}{G}{\mathbf{R}}^\top {}^{k}\boldsymbol{\alpha}_{k+1} && (11)\\
&= {}^{I_0}\mathbf{p}_{I_k} + {}^{I_0}\mathbf{v}_{I_k}\Delta T - \frac{1}{2}{}^{I_0}\mathbf{g}\Delta T^2 + \prescript{I_k}{I_0}{\mathbf{R}}^\top {}^{I_k}\boldsymbol{\alpha}_{I_{k+1}} && (12)
\end{aligned}$$

We can thus have the following relative preintegration equations:

$$\prescript{I_{k+1}}{I_0}{\mathbf{R}} = \prescript{I_{k+1}}{I_k}{\Delta\mathbf{R}}\ \prescript{I_k}{I_0}{\mathbf{R}} \tag{13}$$

$$
\begin{aligned}
{}^{I_0}\mathbf{p}_{I_{k+1}} &= {}^{I_0}\mathbf{p}_{I_k} + {}^{I_0}\mathbf{v}_{I_k}\Delta T - \frac{1}{2}{}^{I_0}\mathbf{g}\Delta T^2 + {}^{I_k}_{I_0}\mathbf{R}^{\top\, I_k}\boldsymbol{\alpha}_{I_{k+1}} && (14)\\
{}^{I_0}\mathbf{v}_{I_{k+1}} &= {}^{I_0}\mathbf{v}_{I_k} - {}^{I_0}\mathbf{g}\Delta T + {}^{I_k}_{I_0}\mathbf{R}^{\top\, I_k}\boldsymbol{\beta}_{I_{k+1}} && (15)
\end{aligned}
$$

We now define the integration from the first $\{I_0\}$ frame:

$$
\begin{aligned}
{}^{I_{k+1}}_{I_0}\mathbf{R} &= {}^{I_{k+1}}_{I_0}\Delta\mathbf{R}\ {}^{I_0}_{I_0}\mathbf{R} && (16)\\
&\triangleq {}^{I_{k+1}}_{I_0}\Delta\mathbf{R} && (17)\\
{}^{I_0}\mathbf{p}_{I_{k+1}} &= {}^{I_0}\mathbf{p}_{I_0} + {}^{I_0}\mathbf{v}_{I_0}\Delta T - \frac{1}{2}{}^{I_0}\mathbf{g}\Delta T^2 + {}^{I_0}_{I_0}\mathbf{R}^{\top\, I_0}\boldsymbol{\alpha}_{I_{k+1}} && (18)\\
&\triangleq {}^{I_0}\mathbf{v}_{I_0}\Delta T - \frac{1}{2}{}^{I_0}\mathbf{g}\Delta T^2 + {}^{I_0}\boldsymbol{\alpha}_{I_{k+1}} && (19)\\
{}^{I_0}\mathbf{v}_{I_{k+1}} &= {}^{I_0}\mathbf{v}_{I_0} - {}^{I_0}\mathbf{g}\Delta T + {}^{I_0}_{I_0}\mathbf{R}^{\top\, I_0}\boldsymbol{\beta}_{I_{k+1}} && (20)\\
&\triangleq {}^{I_0}\mathbf{v}_{I_0} - {}^{I_0}\mathbf{g}\Delta T + {}^{I_0}\boldsymbol{\beta}_{I_{k+1}} && (21)
\end{aligned}
$$

Note that the time offset ΔT is now from time t_0 to t_{k+1}.

B. Depth-Aided Feature Bearing Model

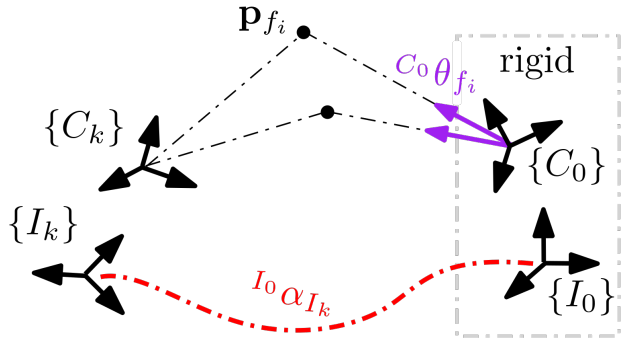

Fig. 1: Frame of references used in the problem. Two features observed from both the $\{C_k\}$ and $\{C_0\}$ frame are shown. The transformation from the $\{I_k\}$ and $\{I_0\}$ is found through IMU integration. The bearing ${}^{C_0}\boldsymbol{\theta}_{f_i}$ is used along with the scale-less depth to recover the scale a and shift b.

Assuming a calibrated perspective camera, the bearing measurement of the i'th feature at timestep t_k can be related to the state by the following:

$$
\begin{aligned}
\mathbf{z}_{i,k} &:= \boldsymbol{\Lambda}({}^{C_k}\mathbf{p}_{f_i}) + \mathbf{n}_k && (22)\\
{}^{C_k}\mathbf{p}_{f_i} &= {}^{C}_{I}\mathbf{R}\,{}^{I_k}_{I_0}\mathbf{R}({}^{I_0}\mathbf{p}_{f_i} - {}^{I_0}\mathbf{p}_{I_k}) + {}^{C}\mathbf{p}_{I} && (23)
\end{aligned}
$$

where $\boldsymbol{\Lambda}([x\ y\ z]^\top) = [x/z\ y/z]^\top$ is the camera perspective projection model, $\mathbf{z}_{i,k} = [u_{i,k}, v_{i,k}]^\top$ is the normalized feature bearing measurement with white Gaussian noise $\mathbf{n}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$, and $\{{}^{C}_{I}\mathbf{R}, {}^{C}\mathbf{p}_{I}\}$ are the known camera-IMU transformation.

We assume that for a single image the scale a and shift b are constant for the whole depth map. Specifically, for feature ${}^{I_0}\mathbf{p}_{f_i}$ we can express the metric depth scalar $z_i = \mathrm{Z}(u_{i,0}, v_{i,0})$ as a function of a, b, and $d_i = \mathrm{D}(u_{i,0}, v_{i,0})$:

$$
\begin{aligned}
{}^{I_0}\mathbf{p}_{f_i} &= {}^{I}_{C}\mathbf{R}\ {}^{C_0}\mathbf{p}_{f_i} + {}^{I}\mathbf{p}_{C}\\
&= z_i\ {}^{I_0}\boldsymbol{\theta}_{C_0\to f_i} + {}^{I}\mathbf{p}_{C}\\
&= (ad_i + b)\ {}^{I_0}\boldsymbol{\theta}_{C_0\to f_i} + {}^{I}\mathbf{p}_{C} && (24)
\end{aligned}
$$

where ${}^{I_0}\boldsymbol{\theta}_{C_0\to f_i} = {}^{I}_{C}\mathbf{R}[u_{i,0}\ v_{i,0}\ 1]^\top$ is the bearing vector of the feature rotated (but not translated) into the IMU frame, see Fig. 1 for example frame of references. This treats the normalized 2D coordinates of the feature in the first camera frame $u_{i,0}$ and $v_{i,0}$ as a known quantity.

We can define the following linear measurement observation, which removes the need for the division in $\boldsymbol{\Lambda}(\cdot)$. As presented in [2, 3], we consider the following:

$$
\begin{bmatrix} 1 & 0 & -u_{i,k} \\ 0 & 1 & -v_{i,k} \end{bmatrix} {}^{C_k}\mathbf{p}_{f_i} \triangleq \boldsymbol{\Gamma}_{i,k}\,{}^{C_k}\mathbf{p}_{f_i} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (25)
$$

One can check that the left side of the above equation when multiplied out, should equate to zero. This shows that the difference between the normalized feature observation and the projected feature should be zero.

Substituting Eq. (24) and Eq. (19) into Eq. (23) we can recover the following linear system:

$$\mathbf{0}_2 \overset{(23)}{=} \mathbf{\Gamma}_{i,k} \left({}^{C}_{I}\mathbf{R}^{I_k}_{I_0}\mathbf{R} \left({}^{I_0}\mathbf{p}_{f,i} - {}^{I_0}\mathbf{p}_{I_k} \right) + {}^{C}\mathbf{p}_I \right) \tag{26}$$

$$\mathbf{0}_2 \overset{(24)}{=} \mathbf{\Gamma}_{i,k} \left({}^{C}_{I}\mathbf{R}^{I_k}_{I_0}\mathbf{R} \left(\left[(ad_i + b)\ {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_i} + {}^{I}\mathbf{p}_C \right] - {}^{I_0}\mathbf{p}_{I_k} \right) + {}^{C}\mathbf{p}_I \right) \tag{27}$$

$$\mathbf{0}_2 \overset{(19)}{=} \mathbf{\Gamma}_{i,k} \left({}^{C}_{I}\mathbf{R}^{I_k}_{I_0}\Delta\mathbf{R} \left(\left[(ad_i + b)\ {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_i} + {}^{I}\mathbf{p}_C \right] - {}^{I_0}\mathbf{v}_{I_0}\Delta T + \frac{1}{2}{}^{I_0}\mathbf{g}\Delta T^2 - {}^{I_0}\boldsymbol{\alpha}_{I_k} \right) + {}^{C}\mathbf{p}_I \right) \tag{28}$$

Rearrange Eq. (28) we can get:

$$\underbrace{\mathbf{\Gamma}_{i,k}\ {}^{C}_{I}\mathbf{R}^{I_k}_{I_0}\Delta\mathbf{R}}_{\mathbf{\Upsilon}_{i,k}} \left((ad_i + b)\ {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_i} - {}^{I_0}\mathbf{v}_{I_0}\Delta T_k + \frac{1}{2}{}^{I_0}\mathbf{g}\Delta T_k^2 \right) = \underbrace{\mathbf{\Gamma}_{i,k}\ {}^{C}_{I}\mathbf{R}^{I_k}_{I_0}\Delta\mathbf{R}}_{\mathbf{\Upsilon}_{i,k}} \left({}^{I_0}\boldsymbol{\alpha}_{I_k} + {}^{C}_{I}\mathbf{R}^{\top}{}^{C}\mathbf{p}_I \right) - \mathbf{\Gamma}_{i,k}\ {}^{C}\mathbf{p}_I$$

$$\begin{aligned} \Rightarrow \quad \mathbf{\Upsilon}_{i,k} \begin{bmatrix} d_i{}^{I_0}\boldsymbol{\theta}_{C_0 \to f_i} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_i} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \end{bmatrix} \mathbf{x}' &= \mathbf{\Upsilon}_{i,k} \left({}^{I_0}\boldsymbol{\alpha}_{I_k} + {}^{C}_{I}\mathbf{R}^{\top}{}^{C}\mathbf{p}_I \right) - \mathbf{\Gamma}_{i,k}\ {}^{C}\mathbf{p}_I \\ \Rightarrow \quad \mathbf{\Upsilon}_{i,k} \begin{bmatrix} \mathbf{B}_i & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \end{bmatrix} \mathbf{x}' &= \mathbf{b}'_{i,k} \\ \Rightarrow \quad \mathbf{A}'_{i,k}\ \mathbf{x}' &= \mathbf{b}'_{i,k} \end{aligned}$$

where $\boldsymbol{\Delta}\mathbf{T}_k = \Delta T_k \mathbf{I}_3$. Given M features from N images, $\mathbf{A}' \in \mathbb{R}^{2MN \times (2+6)}$ and $\mathbf{b}' \in \mathbb{R}^{2MN}$. One can see that the state size remains constant, no matter how many features are included in the problem. The general matrix form for a given k'th image is:

$$\underbrace{\begin{bmatrix} \mathbf{\Upsilon}_{1,k} & & & \\ & \mathbf{\Upsilon}_{2,k} & & \\ & & \ddots & \\ & & & \mathbf{\Upsilon}_{M,k} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} {}^{I_0}d_1{}^{I_0}\boldsymbol{\theta}_{C_0 \to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_1} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \\ {}^{I_0}d_2{}^{I_0}\boldsymbol{\theta}_{C_0 \to f_2} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_2} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M{}^{I_0}\boldsymbol{\theta}_{C_0 \to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_M} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \end{bmatrix}}_{\mathbf{K}} \begin{bmatrix} a \\ b \\ {}^{I_0}\mathbf{v}_{I_0} \\ {}^{I_0}\mathbf{g} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{b}'_{1,k} \\ \mathbf{b}'_{2,k} \\ \vdots \\ \mathbf{b}'_{M,k} \end{bmatrix}}_{\mathbf{b}} \tag{29}$$

where we have defined the matrix $\mathbf{D}$ block diagonal, and dense $\mathbf{K}$ matrices which factorise the matrix $\mathbf{A}'_{i,k}$.

II. Minimal Case Analysis

In order to simplify the analysis of minimal cases for the linear problem introduced in the previous section, we consider the case that features can be observed in all images. Therefore, the number of measurements is $2MN$, where M is the number of features and N denotes the number of frames. The state size is $1 + 1 + 3 + 3 = 8$, where include scalar a and b, 3 degree of freedom (DoF) velocity, ${}^{I_0}\mathbf{v}_{I_0}$, and 3DoF gravity, ${}^{I_0}\mathbf{g}$. Thus, the necessary condition is $2MN \geq 8$. We note that if one uses the quadratically-constrained least-squares this would remove 1DoF.

We can now identify the following cases for the number of available images, where we consider the base frame I_0 as the first one:

- $N = 1$: The necessary condition is not met, regardless of the number of features.
- $N = 2$: The necessary condition will never be met regardless of the number of features.
- $N = 3$: The necessary condition is met when $M \geq 2$.

The below analysis follows that by Dong-Si and Mourikis [3, Appendix B] where we focus on the rank of the $\mathbf{K}$ sub-matrix of the linear problem. For each case we show that there exists Gaussian eliminations which can simplify the structure of the matrices, allowing for introspection.

A. Two Images ($N = 2$)

We first consider that there is one image at timestamp 0 with M features in total. Focusing on the $\mathbf{K}$ sub-matrix, we can perform a column-wise Gaussian elimination:

$$\mathbf{K} = \begin{bmatrix} {}^{I_0}d_1{}^{I_0}\boldsymbol{\theta}_{C_0 \to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_1} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \\ {}^{I_0}d_2{}^{I_0}\boldsymbol{\theta}_{C_0 \to f_2} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_2} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M{}^{I_0}\boldsymbol{\theta}_{C_0 \to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_M} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \end{bmatrix} \tag{30}$$

$$\underset{\sim}{\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k * C_3} \begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \\ {}^{I_0}d_2\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & -\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \end{bmatrix} \tag{31}$$

$$\underset{\sim}{C_3 + C_4} \begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 & \mathbf{0}_3 \\ {}^{I_0}d_2\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & -\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 & \mathbf{0}_3 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 & \mathbf{0}_3 \end{bmatrix} \tag{32}$$

We can conclude through inspection of the row rank that:

$$\texttt{rank}(\mathbf{K}) \leq 8 - 3 \tag{33}$$

Thus this matrix is not full rank and the necessary condition will never meet regardless of the number of features.

B. Three Image ($N = 3$)

The $\mathbf{K}$ matrix for the case of a base image at time t_0, and two extra images at t_1 and t_2 can be written as:

$$\mathbf{K} = \begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ {}^{I_0}d_2\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_2^2 \\ {}^{I_0}d_2\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & -\boldsymbol{\Delta}\mathbf{T}_2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_2^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_2^2 \end{bmatrix} \tag{34}$$

$$\underset{\sim}{R_{3M+i} - R_i \ \forall i \in \{1,...,3M\}} \begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ {}^{I_0}d_2\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_2 + \boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_2^2 - \boldsymbol{\Delta}\mathbf{T}_1^2) \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_2 + \boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_2^2 - \boldsymbol{\Delta}\mathbf{T}_1^2) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_2 + \boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_2^2 - \boldsymbol{\Delta}\mathbf{T}_1^2) \end{bmatrix} \tag{35}$$

$$\underset{\sim}{R_i - R_{i+1} \ \forall i \in \{3M+1,...,6M\}} \begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ {}^{I_0}d_2\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_2} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_2 + \boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_2^2 - \boldsymbol{\Delta}\mathbf{T}_1^2) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{36}$$

We can conclude through inspection of the row rank that:

$$\texttt{rank}(\mathbf{K}) = \min(3M + 3, 8) \tag{37}$$

The necessary condition will be satisfied if $3N + 3 \geq 8 \Rightarrow M \geq 2$. The minimal number of features is 2.

REFERENCES

[1] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, 2022.

[2] T.-C. Dong-Si and A. I. Mourikis, "Estimator initialization in vision-aided inertial navigation with unknown camera-imu calibration," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1064–1071, IEEE, 2012.

[3] T.-C. Dong-Si and A. I. Mourikis, "Closed-form solutions for vision-aided inertial navigation," tech. rep., Dept. of Electrical Engineering, University of California, Riverside, 2011. Available: http://tdongsi.github.io/download/pubs/2011_VIO_Init_TR.pdf.