Learned Monocular Depth Priors in Visual-Inertial Initialization Supplemental Material

Yunwen Zhou, Abhishek Kar, Eric Turner, Adarsh Kowdle, Chao X. Guo, Ryan C. DuToit, and Konstantine Tsotsos

Google AR
{verse,abhiskar,elturner,adarshkowdle,chaoguo,rdutoit,ktsotsos}@google.com

1 Depth Residual function

Reiterating our paper, the proposed depth residual in VI-BA for keyframe k and feature point i takes the form of the *log* of the ratio between the measured depth scaled/shifted by $\boldsymbol{S_k}$ and the feature point's estimated depth:

$$r_{\mathcal{L}_{ik}} = \log\left((a_k d_{ik} + b_k) \cdot \Omega({}^{\boldsymbol{C_j}}\boldsymbol{f_i}, \boldsymbol{q_j}, \boldsymbol{p_j}, \boldsymbol{q_k}, \boldsymbol{p_k})\right) \tag{1}$$

Cross-Keyframe Depth Transformation $\Omega(\cdot)$ in Eq. (1) is the function transforming the feature point depth w_{ij}^{-1} from its first observed j^{th} camera keyframe to the depth measurement camera k^{th} keyframe $\{C_k\}$. Since feature point ${}^{\boldsymbol{C_j}}\boldsymbol{f_i}$ is parameterized as inverse depth, to recover it into euclidean space ${}^{\boldsymbol{C_j}}\boldsymbol{l_i}$, we have

$${}^{\boldsymbol{C_j}}\boldsymbol{l_i} = w_{ij}^{-1}\begin{bmatrix} u_{ij} & v_{ij} & 1 \end{bmatrix}^T$$

where inverse parameterization is known to be unstable in linearizations when used directly. Instead, we define ${}^{\boldsymbol{C_j}}\boldsymbol{h_i} = [u_{ij}, v_{ij}, 1]^T$, so

$${}^{\boldsymbol{C_j}}\boldsymbol{h_i} = w_{ij}{}^{\boldsymbol{C_j}}\boldsymbol{l_i}$$

where we are able to deal with w_{ij} separately from the geometry transform without explicitly unfolding ${}^{\boldsymbol{C_j}}\boldsymbol{h_i}$ using w_{ij}. We define

$${}^{\boldsymbol{B}}\boldsymbol{y_i} = w_{ij}{}^{\boldsymbol{B}}\boldsymbol{l_i} \tag{2}$$

where ${}^{\boldsymbol{B}}\boldsymbol{l_i}$ is the i^{th} feature position in $\{B\}$ frame other than global $\{G\}$. If it is global $\{G\}$, Eq. (2) becomes $\boldsymbol{y_i} = w_{ij}\boldsymbol{l_i}$.

We can infer ${}^{\boldsymbol{C_k}}\boldsymbol{y_i}$ through a series of transforms only on top of ${}^{\boldsymbol{C_j}}\boldsymbol{h_i}$ using the IMU-camera extrinsics $(\boldsymbol{q_C}, \boldsymbol{p_c})$, j^{th} and k^{th} keyframe poses. Firstly, we show how to get $w_{ij}{}^{\boldsymbol{j}}\boldsymbol{l_i}$ that is ${}^{\boldsymbol{j}}\boldsymbol{y_i}$ using ${}^{\boldsymbol{C_j}}\boldsymbol{h_i}$, where ${}^{\boldsymbol{j}}\boldsymbol{l_i}$ is in j^{th} IMU keyframe coordinate system

$${}^{\boldsymbol{j}}\boldsymbol{y_i} = R(\boldsymbol{q_C}){}^{\boldsymbol{C_j}}\boldsymbol{h_i} + w_{ij}\boldsymbol{p_C}$$

Similarly, we can keep transforming all the way to the k^{th} camera frame and get ${}^{C_k}\boldsymbol{y_i}$

$$\begin{aligned} \boldsymbol{y_i} &= R(\boldsymbol{q_j})^{j}\boldsymbol{y_i} + w_{ij}\boldsymbol{p_j} \\ {}^{k}\boldsymbol{y_i} &= R(\boldsymbol{q_k^{-1}})(\boldsymbol{y_i} - w_{ij}\boldsymbol{p_k}) \\ {}^{C_k}\boldsymbol{y_i} &= R(\boldsymbol{q_C^{-1}})({}^{k}\boldsymbol{y_i} - w_{ij}\boldsymbol{p_C}) \end{aligned}$$

Reiterating Eq. (2), along with $[\cdot]_3$ that extracts the 3^{rd} element from the vector, $\Omega(\cdot)$ then becomes

$$\Omega({}^{C_j}\boldsymbol{f_i}, \boldsymbol{q_j}, \boldsymbol{p_j}, \boldsymbol{q_k}, \boldsymbol{p_k}) = \frac{[{}^{C_k}\boldsymbol{y_i}]_3}{w_{ij}} \tag{3}$$

As shown below, Eq. (1) can be simply written as followings by substitution using Eq. (3)

$$r_{\mathcal{L}_{ik}} = \log(a_k d_{ik} + b_k) + \log([{}^{C_k}\boldsymbol{y_i}]_3) - \log(w_{ij}) \tag{4}$$

Eq. (4) fits our problem space since our ML model yields inverse depth and our feature point is parameterized as inverse depth, which results in a more stable first-order approximation that is helpful for optimization purposes.

2 Formulation and Derivation of Closed-Form Solver For Initializing VI-BA

It is well-known that VI-SFM is a highly non-linear problem, therefore it is important to find an accurate initial linearization point. Our solution is based on [4]. We employ visual reprojection error to approximate the 0^{th} keyframe's pose and velocity in gravity-aligned global coordinate frame $\{G\}$. Then, the remaining keyframe poses and velocities are inferred from IMU integration.

Given N keyframes, we define our estimated states $\boldsymbol{X}$ for closed-form solver as following,

$$\boldsymbol{X} = [\boldsymbol{v}; \boldsymbol{g}; \Delta\boldsymbol{p_0}; \Delta\boldsymbol{p_1}; \ldots; \boldsymbol{\Delta p_{N-1}}] \tag{5}$$

where $\boldsymbol{v}$, $\boldsymbol{g}$ are the initial velocity and gravity vector, $\Delta\boldsymbol{p}_k$ is the k^{th} keyframe position estimation difference. All parameters are expressed with respect to the 0^{th} keyframe, and we re-express them with respect to $\{G\}$ through $\boldsymbol{g}$ after the solve. The rest of this section shows the derivation of the linear equation to solve the problem.

As in [2, 5], we marginalize feature points, yielding constraints among keyframes. Then, for feature points initial values in VI-BA, they are triangulated after the closed-form solve.

IMU Measurement Model. First, we recall the IMU measurement model:

$$\boldsymbol{\omega^m}(t) = \boldsymbol{\omega}(t) + \boldsymbol{b^w}(t) + \boldsymbol{\eta^w}(t) \tag{6}$$

$$\boldsymbol{a^m}(t) = R(t)^T({}^{G}\boldsymbol{a}(t) - {}^{G}\boldsymbol{g}) + \boldsymbol{b^a}(t) + \boldsymbol{\eta^a}(t) \tag{7}$$

where $\boldsymbol{\omega^m}(t)$ and $\boldsymbol{a^m}(t)$ are the gyro and acceleration measurement at timestamp t, ${}^{G}\boldsymbol{g} = [0; 0; -G]$ is the gravity vector in gravity-aligned global $\{G\}$. $\boldsymbol{b^a}(t)$

and $\boldsymbol{b^w}(t)$ are IMU accelerometer bias and gyro bias respectively, and their corresponding noises are $\boldsymbol{\eta^a}(t)$, $\boldsymbol{\eta^g}(t)$. $R(t)$ is the IMU body frame rotation w.r.t gravity-aligned global $\{G\}$.

Integrated Keyframe Positions. According to [4] and IMU measurement model Eq. (6), Eq. (7), we can write k^{th} keyframe position as

$$\boldsymbol{p_k} = \boldsymbol{v}\Delta t_k + \boldsymbol{g}\frac{\Delta t_k^2}{2} + \boldsymbol{\xi_k} \tag{8}$$

where $\Delta t_k = t_k - t_0$ is the elapse timestamp up to t_k since t_0. $\boldsymbol{\xi_k}$ is the k^{th} keyframe integrated position in the 0^{th} keyframe at timestamp t_k without the impact of gravity. The integrator is described in [5]

$$\boldsymbol{\xi_k} = \int_{t_0}^{t_k}\int_{t_0}^{\tau} R_0(\eta)(\boldsymbol{a^m}(\eta) - \boldsymbol{b^a}(\eta))d\eta d\tau$$

where $R_0(\eta)$ is integrated using gyro measurement from t_0 to timestamp η in 0^{th} keyframe coordinate system. To remove the non-linearity of the estimation, $\boldsymbol{b^w}(t)$ is assumed to be 0. $\boldsymbol{b^a}(t)$ is also assumed to be 0, since the small baseline closed-form solution is resilient to accelerometer bias, which is studied in [3]. The bias random walk noises $\boldsymbol{\eta^a}(t)$, $\boldsymbol{\eta^w}(t)$ are treated as zero mean so it doesn't appear in the equation.

Defining integrated keyframe positions in 0^{th} keyframe coordinate system as $\boldsymbol{P} = [\boldsymbol{p_0}; \boldsymbol{p_1}; \ldots; \boldsymbol{p_{N-1}}]$, we form a linear equation with states in $\boldsymbol{X}$ and $\boldsymbol{\Xi} = [\boldsymbol{0_{3\times1}}; \boldsymbol{\xi_0}; \ldots; \boldsymbol{\xi_{N-1}}]$

$$\boldsymbol{P} = F\boldsymbol{g} + W\begin{bmatrix}\boldsymbol{v}\\ \Delta\boldsymbol{p_0}\\ \vdots\\ \Delta\boldsymbol{p_{N-1}}\end{bmatrix} + \boldsymbol{\Xi} \tag{9}$$

where

$$F = \begin{bmatrix}0_{3\times3}\\ \vdots\\ \frac{\Delta t_{N-1}^2}{2}I_3\end{bmatrix},\ W = \begin{bmatrix}0_{3\times3} & I_3 & \\ \vdots & & \ddots\\ \Delta t_{N-1}I_3 & & I_3\end{bmatrix}$$

Visual Constraint. We define the estimated i^{th} feature point position at k^{th} camera frame as ${}^{\boldsymbol{C_k}}\boldsymbol{l_i} = [x, y, z]^T$. With undistorted 2D perspective projection measurement $[u_{ik}, v_{ik}]^T$, we can formulate the visual constraint for a single feature point in a linear equation:

$$\begin{bmatrix}1 & 0 & -u_{ik}\\ 0 & 1 & -v_{ik}\end{bmatrix}{}^{\boldsymbol{C_k}}\boldsymbol{l_i} = K_{ik}^{\boldsymbol{C_k}}\boldsymbol{l_i} = 0 \tag{10}$$

Then we transform ${}^{\boldsymbol{C_k}}\boldsymbol{l_i}$ to $\boldsymbol{l_i}$ in 0^{th} keyframe coordinate system though pre-calibrated IMU-camera extrinsics $[R_C, \boldsymbol{p_C}]$ and k^{th} IMU keyframe pose $[R_k, \boldsymbol{p_k}]$, in which R_k is computed by zero bias gyro measurements integration.

$${}^{\boldsymbol{C_k}}\boldsymbol{l_i} = R_C^T R_k^T \boldsymbol{l_i} - R_C^T R_k^T \boldsymbol{p_k} - R_C^T \boldsymbol{p_C} \tag{11}$$

Substitute ${}^{\boldsymbol{C_k}}\boldsymbol{l_i}$ in Eq. (10) with Eq. (11), we can write Eq. (10) as

$$A_{ik}\boldsymbol{p_k} + H_{ik}\boldsymbol{l_i} = \boldsymbol{b_{ik}} \tag{12}$$

where

$$\begin{aligned} &A_{ik} = -K_{ik}R_C^T R_k^T,\ H_{ik} = K_{ik}R_C^T R_k^T, \\ &\boldsymbol{b_{ik}} = K_{ik}R_C^T \boldsymbol{p_C} \end{aligned}$$

The i^{th} feature point should be observed by at least by 2 keyframes among N keyframes, and we can stack Eq. (12) together for all visual constraints w.r.t one feature point as

$$A_i\boldsymbol{P} + H_i\boldsymbol{l_i} = \boldsymbol{b_i} \tag{13}$$

where

$$\begin{aligned} &A_i = Diag(A_{ik}),\ H_i = [H_{i0}; \ldots; H_{iN-1}] \\ &\boldsymbol{b_i} = [\boldsymbol{b_{i0}}; \ldots; \boldsymbol{b_{iN-1}}] \end{aligned}$$

$A_{ik} = 0_{3\times3},\ H_{ik} = 0_{3\times2},\ \boldsymbol{b_{ik}} = \boldsymbol{0_{2\times1}}$ if k^{th} keyframe doesn't observe the feature. As described in [2, 5], H_i in Eq. (13) can be projected on its left-nullspace to marginalize out $\boldsymbol{l_i}$. Then we have

$$A_i^*\boldsymbol{P} = \boldsymbol{b_i^*} \tag{14}$$

Reiterating Eq. (9), for i^{th} feature point observed by N keyframes, we can form the linear equation for the visual constraint

$$F_i^*\boldsymbol{g} + W_i^* \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\Delta p_0} \\ \vdots \\ \boldsymbol{\Delta p_{N-1}} \end{bmatrix} = \boldsymbol{r_i} \tag{15}$$

where

$$\begin{aligned} &F_i^* = A_i^* F,\ W_i^* = A_i^* W \\ &\boldsymbol{r_i} = \boldsymbol{b_i^*} - A_i^*\boldsymbol{\Xi} \end{aligned}$$

Closed-form VI-SFM Least Square Problem. Suppose we have $\mathcal{F}$ feature points in total, the closed-form VI-SFM solver is essentially to solve a least square problem with a number of visual constraints as Eq. (15)

$$\begin{aligned} &\text{minimize} \sum_{i\in\mathcal{F}} \|F_i^*\boldsymbol{g} + W_i^* \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\Delta p_0} \\ \vdots \\ \boldsymbol{\Delta p_{N-1}} \end{bmatrix} - \boldsymbol{r_i}\|^2, \\ &\text{subject to } \|\boldsymbol{g}\|^2 = G \end{aligned} \tag{16}$$

3 Experimental Results for 10KFs under 10/4Hz settings

In Tab. 1, we present results of 10KFs under $10Hz/4Hz$ settings as specified in [1]. We partition the datasets into 1.6 second trajectories for $10Hz$ and 3 seconds for $4Hz$ to run the same exhaustive initialization benchmark. Ours performs best in the $10Hz$ setting, while results are mixed at $4Hz$ where ours performs similarly to the baseline. This is expected, as lower framerates with the same number of keyframes results in overall larger baselines (i.e., more motion). Note that by construction, the $4Hz$ sequences result in slower initialization time (at least 2.25s vs. 0.897s and 0.399s for $10KFs/10Hz$ and $5KFs/10Hz$). For practical applications, faster initialization is preferred.

Table 1: Aggregated initialization benchmark for Inertial-Only, Baseline and our proposed method using various framerates on all EuRoC datasets. For each metric, lower is better.

Metrics	10KFs $10Hz$			10KFs $4Hz$		
	Intertial-Only	**Baseline**	**Ours**	**Intertial-Only**	**Baseline**	**Ours**
Scale Error (%) $\|\|\bar{a}\|\| > 0.005G$	35.67	16.36	**13.78**	14.65	8.76	**8.41**
Position RMSE (meters)	0.123	0.038	**0.034**	0.066	**0.061**	0.064
Gravity RMSE (degrees)	1.89	1.42	**1.38**	1.36	**1.32**	1.36
log**(Condition Num)** $\|\|\bar{a}\|\| < 0.005G$	n/a	12.12	**11.23**	n/a	**10.25**	10.54

References

1. Campos, C., Montiel, J.M., Tardós, J.D.: Inertial-only optimization for visual-inertial initialization. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 51–57. IEEE (2020)
2. Forster, C., Carlone, L., Dellaert, F., Scaramuzza, D.: On-manifold preintegration theory for fast and accurate visual-inertial navigation. CoRR **abs/1512.02363** (2015), http://arxiv.org/abs/1512.02363
3. Kaiser, J., Martinelli, A., Fontana, F., Scaramuzza, D.: Simultaneous state initialization and gyroscope bias calibration in visual inertial aided navigation. IEEE Robotics and Automation Letters **2**(1), 18–25 (2017). https://doi.org/10.1109/LRA.2016.2521413
4. Li, M., Mourikis, A.I.: A convex formulation for motion estimation using visual and inertial sensors. In: Proceedings of the Workshop on Multi-View Geometry, held in conjunction with RSS. Berkeley, CA (July 2014)
5. Mourikis, A.I., Roumeliotis, S.I.: A multi-state constraint kalman filter for vision-aided inertial navigation. In: Proceedings 2007 IEEE International Conference on Robotics and Automation. pp. 3565–3572. IEEE (2007)