EgoHDM: A Real-time Egocentric-Inertial Human Motion Capture, Localization, and Dense Mapping System - Supplementary Material

BONAN LIU* and HANDI YIN*, HKUST(GZ), China
MANUEL KAUFMANN, ETH AI Center, ETH Zürich, Switzerland
JINHAO HE, HKUST(GZ), China
SAMMY CHRISTEN, Department of Computer Science, ETH Zürich, Switzerland
JIE SONG†, HKUST(GZ), China and HKUST, China
PAN HUI, HKUST(GZ), China and HKUST, China

A CALIBRATION AND ALIGNMENT

A.1 Head-Camera Calibration

When mounting the sensors and camera before a capture session, we first compute an initial estimate of $\mathbf{T}_{hc}$. To do so we fix an AprilTag marker [Olson 2011] to the back of the user's head, and then ask them to subtly move for a few seconds within the viewing range of a third camera set up opposite another AprilTag board, akin to the method described in [Pfrommer and Daniilidis 2019] as shown in Fig. 7. This must only be done once in the beginning of a capture session.

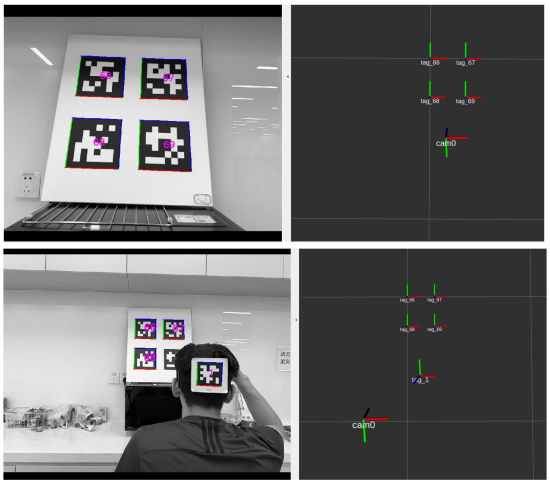

Fig. 7. Illustration of the calibration process to get an initial estimate of $\mathbf{T}_{hc}$.

A.2 Time Synchronization

To temporally align the IMU sensors with the head-mounted camera, we employ the widely used technique of cross-correlation. I.e., we compare the angular velocities of the iPhone's IMU with the sensor's values by computing an L2-based matching score and selecting the time shift with the highest score as the synchronization point (see Fig. 8). The red bar shown in the figure represents the offset we computed. For our implementation, we use camera at 30Hz and IMU at 60Hz sampling rate. Therefore the offset value is two times the peak time-axis value.

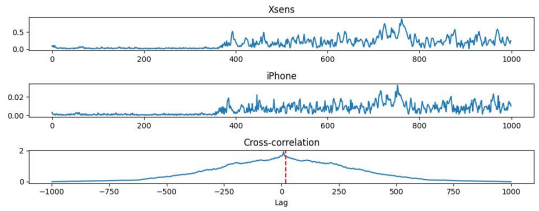

Fig. 8. Illustration of the time alignment process

B OPTIMIZATION

In Sec. 3.3 (MDBA) we perform non-linear optimization using the Gauss-Newton algorithm. Note that the visual error term E_{repr} and the inertial error term E_{iner} do not have common residuals. The parameterization of the structure leads to an extremely efficient way of solving the dense BA problem with mocap constraints, which can be decomposed into an arrow-like block-sparse matrix following [Rosinol et al. 2023], and we can divide $\mathbf{H}$ and $\mathbf{b}$ each into two independent parts.

$$\mathbf{H} = \mathbf{H}_{\text{repr}} + \mathbf{H}_{\text{inert}}, \quad \mathbf{b} = \mathbf{b}_{\text{repr}} + \mathbf{b}_{\text{inert}}, \tag{1}$$

$$\mathbf{Hx} = \mathbf{b}, \quad i.e. \quad \begin{bmatrix} \mathbf{C} & \mathbf{E} \\ \mathbf{E}^T & \mathbf{P} \end{bmatrix} \begin{bmatrix} \Delta\xi \\ \Delta\mathbf{d} \end{bmatrix} = \begin{bmatrix} \mathbf{v} \\ \mathbf{w} \end{bmatrix}, \tag{2}$$

Given the sparsity pattern of the Hessian, we can extract the required marginal covariances for the per-pixel depth variables efficiently. Follwowing [Rosinol et al. 2023], the marginal covariances of the inverse depth-maps Σ_d are given by:

$$\begin{aligned} \Sigma_d &= \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{E}^T\Sigma_T\mathbf{E}\mathbf{P}^{-1} \\ \Sigma_T &= (\mathbf{H}/\mathbf{P})^{(-1)}, \end{aligned} \tag{3}$$

*Both authors contributed equally to this research.
†Corresponding author.

Authors' addresses: Bonan Liu, bliu404@connect.hkust-gz.edu.cn; Handi Yin, hyin335@connect.hkust-gz.edu.cn, HKUST(GZ), Guangzhou, Guangdong, China; Manuel Kaufmann, ETH AI Center, ETH Zürich, Zurich, Switzerland; Jinhao He, HKUST(GZ), Guangzhou, Guangdong, China; Sammy Christen, Department of Computer Science, ETH Zürich, Zurich, Switzerland; Jie Song, HKUST(GZ), Guangzhou, Guangdong, China and HKUST, Hong Kong, Hong Kong, China; Pan Hui, HKUST(GZ), Guangzhou, Guangdong, China and HKUST, Hong Kong, Hong Kong, China.

Where Σ_T is the marginal covariance of the poses.

C KALMAN FILTER

In our MDBA (Sec. 3.3), the camera tracking module utilizes both visual and inertial data, enabling the keyframe-based camera localization to refine human motion at 60 FPS captured by the inertial motion capture system. For complete constraint of the refinement process, only human translation updates are performed, and the system is implemented with a prediction-correction algorithm via Kalman Filter as per Egolocate [Yi et al. 2023]. With the refined camera pose, we consistently update the human translation, incorporating the marginal covariance Σ_T from our MDBA to reduce the impact of camera pose noise on the human.

D MAPPING ACCURACY ON OUR DATA

In our newly captured in-the-wild data, we assess the impact of each module on our system qualitatively (see Sec. 4.4, Fig. 6). Here we present the mapping quality results to evaluate our online dense map reconstruction against the ground-truth map obtained from Station Scanner BLK360, as shown in Fig. 9. For alignment of our map with the ground-truth scanning map, CloudCompare [Girardeau-Montaut 2016] is utilized for registration purposes, employing ICP [Segal et al. 2009] algorithms. Then we calculate the mean point-to-point distance error for evaluation, the correlation error shown in Fig. 9 is 0.01, 0.07, 0.05 meters.

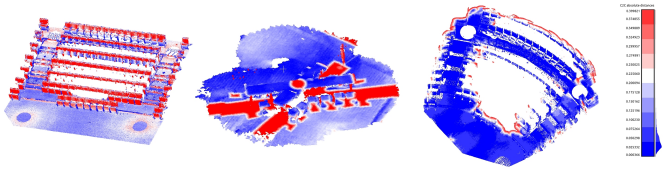

Fig. 9. The point-to-point distance error map on our newly captured in-the-wild scene with non-flat terrain.

E DETAILED PER-SCENE EXPERIMENT RESULTS.

In Sec. 4 we only report the average metrics on HPS for brevity. Here, we provide a more detailed version of our evaluation tables for the per-scene absolute root position errors in meters (Tab. 5), the per-scene camera localization errors in meters (Tab. 6) and the per-scene mapping accuracy in meters on synthetic TotalCapture (Tab. 7).

REFERENCES

Daniel Girardeau-Montaut. 2016. CloudCompare. *France: EDF R&D Telecom ParisTech* 11, 5 (2016).

Edwin Olson. 2011. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3400–3407.

Bernd Pfrommer and Kostas Daniilidis. 2019. Tagslam: Robust slam with fiducial markers. *arXiv preprint arXiv:1910.00679* (2019).

Antoni Rosinol, John J Leonard, and Luca Carlone. 2023. Probabilistic volumetric fusion for dense monocular slam. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3097–3105.

Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. 2009. Generalized-icp.. In *Robotics: science and systems*, Vol. 2. Seattle, WA, 435.

Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. 2023. EgoLocate: Real-time Motion Capture, Localization, and Mapping with Sparse Body-mounted Sensors. *ACM Transactions on Graphics (TOG)* 42, 4, Article 76 (2023), 17 pages.

Table 5. Absolute root position error in meters per scene or motion type.

Method	TotalCapture					HPS								
	acting	freestyle	rom	walking	average	BIB_AB	BIB_EG	EG	Etage6	GEB	BIB_UG	KINO	BIB_OG	average
TIP	0.43	0.87	0.21	0.49	0.45	2.23	3.41	**1.43**	3.87	1.38	2.92	0.89	3.89	3.00
PIP	0.61	0.51	**0.07**	0.49	0.37	1.26	2.59	1.89	1.78	1.35	2.49	1.50	4.81	2.75
EgoLocate	0.28	0.33	0.10	0.25	0.22	1.23	**1.54**	1.83	**1.35**	1.17	2.40	**0.87**	1.90	1.70
	±0.06	±0.06	±0.02	±0.03	±0.04	±0.37	±0.28	±0.37	±0.18	±0.29	±0.49	±0.16	±0.41	±0.34
Ours	**0.16**	**0.18**	0.09	**0.15**	**0.13**	**1.01**	1.58	1.66	1.98	**1.14**	**1.97**	1.10	**1.54**	**1.50**

Table 6. Camera localization error in meters per scene or motion type.

Method	TotalCapture					HPS								
	acting	freestyle	rom	walking	average	BIB_AB	BIB_EG	EG	Etage6	GEB	BIB_UG	KINO	BIB_OG	average
ORB-SLAM3	0.82	0.89	0.25	0.42	0.54	8.58	12.57	8.87	5.62	1.62	9.29	4.79	7.61	8.18
	±0.44	±0.17	±0.16	±0.46	±0.29	±0.92	±2.30	±2.23	±1.62	±0.15	±1.52	±0.52	±2.06	±1.71
ORB-SLAM3-I	10.54	4.75	-	1.08	4.87	-	-	-	-	-	7.31	-	-	-
	±5.48	±2.62	-	±1.88	±3.24	-	-	-	-	-	±5.98	-	-	-
Droid-SLAM (on)	0.23	0.19	0.07	0.27	0.20	-	-	-	-	-	-	-	-	-
Droid-SLAM (off)	0.14	0.10	0.07	0.24	0.14	-	-	-	-	-	-	-	-	-
EgoLocate	0.29	0.35	0.13	0.25	0.24	1.25	**1.53**	1.81	**1.34**	1.18	2.39	0.86	1.90	1.69
	±0.06	±0.06	±0.02	±0.04	±0.04	±0.37	±0.28	±0.36	±0.18	±0.29	±0.49	±0.16	±0.41	±0.33
Ours	**0.07**	**0.09**	**0.05**	**0.08**	**0.07**	**1.00**	1.57	**1.65**	1.97	**1.14**	**1.95**	1.08	**1.53**	**1.49**

Table 7. Mapping accuracy in terms of point-to-point errors in meters per scene and action type in synthetic TotalCapture.

Method	Japan Office					Flooded Grounds					SciFi Warehouse				
	acting	freestyle	rom	walking	average	acting	freestyle	rom	walking	average	acting	freestyle	rom	walking	average
Droid-SLAM (off)	0.38	0.38	0.23	0.40	0.35	1.35	1.36	1.02	1.73	1.40	0.46	0.42	0.28	0.39	0.40
EgoLocate	0.32	0.49	0.72	**0.24**	0.45	0.94	1.49	1.75	**0.80**	1.23	0.24	0.37	0.44	**0.18**	0.31
	±0.09	±0.09	±0.23	±0.06	±0.12	±0.31	±0.70	±1.13	±0.19	±0.56	±0.03	±0.10	±0.17	±0.02	±0.08
Ours	**0.07**	**0.13**	**0.20**	0.26	**0.17**	**0.72**	**0.81**	**0.90**	0.81	**0.82**	**0.06**	**0.11**	**0.30**	0.23	**0.18**