Structureless VIO

Junlin Song and Miguel Olivares-Mendez

Abstract— **Visual odometry (VO) is typically considered as a chicken-and-egg problem, as the localization and mapping modules are tightly-coupled. The estimation of a visual map relies on accurate localization information. Meanwhile, localization requires precise map points to provide motion constraints. This classical design principle is naturally inherited by visual-inertial odometry (VIO). Efficient localization solutions that do not require a map have not been fully investigated. To this end, we propose a novel structureless VIO, where the visual map is removed from the odometry framework. Experimental results demonstrated that, compared to the structure-based VIO baseline, our structureless VIO not only substantially improves computational efficiency but also has advantages in accuracy.**

I. INTRODUCTION

In the last two decades, the robotics and computer vision communities have designed various VO/SLAM systems that use only a monocular camera [1], [2], [3], [4]. According to the formulation of visual constraints, these systems can be broadly classified as feature-based methods and direct methods. The common visual constraint used in feature-based methods is the reprojection error whose observed pixel location is from feature tracking or matching. And its predicted pixel location depends on the estimated 3D position of feature point through triangulation with multiple camera poses. The photometric error employed in the direct method typically relies on the depth estimation from coarse to fine. The formulation of aforementioned visual constraints tightly couples the visual map (structure) with localization (pose).

Merely using a monocular camera can not restore the true physical scale due to the scale ambiguity. To address this issue, researchers have proposed to fuse the camera with an additional sensor, like an IMU. Thanks to the built-in accelerometer, IMU can provide scale information to aid visual localization. This fusion scheme is well-known as visual-inertial odometry (VIO). The introduction of IMU significantly increases the output frequency, robustness, and accuracy of odometry; therefore, VIO is widely used in AR/VR [5], [6], [7], [8], robotics [9], [10], [11], [12], and planetary exploration [13], [14], [15].

For visual constraints, most classical VIO systems inherit the design idea from VO, either using feature-based reprojection error [16], [17], [18], [19], [20] or direct photometric error [21], [22]. The estimation of visual map (structure) and localization (pose) is interdependent. In order to remove the 3D positions or 1D depths of feature points from the state vector, the authors of Multi-State Constraint Kalman Filter (MSCKF) [16] proposed the idea of nullspace projection to modify the residual equation of the reprojection error and
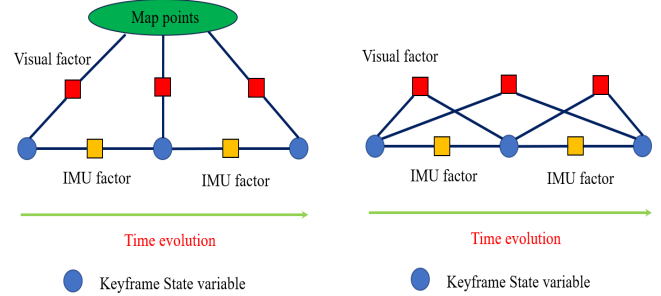
Fig. 1: Left: factor graph for structure-based VI-BA. Right: factor graph for structureless VI-BA.

marginalize feature points. However, triangulation or depth estimation for feature points remains indispensable for the entire MSCKF system [16], [19].

To fully decouple the localization (pose) from visual map (structure), we design a novel structureless visual-inertial bundle adjustment (VI-BA) that naturally eliminates the dependence on the map (structure) by leveraging the epipolar constraint. The key difference between the classic structure-based VI-BA and the proposed structureless VI-BA is presented in Fig. 1. By applying structureless VI-BA, this paper extends our previous work on monocular VIO initialization [23] to subsequent sliding-window based optimization, yielding a novel and efficient structureless VIO.

II. STRUCTURE-BASED VIO

Before introducing the structureless VIO, we first briefly describe the structure-based VIO to better illustrate the differences in the context of keyframe-based VIO. Taking VINS-Mono [18] as an example, a structure-based VIO typically optimizes the following state in the sliding window

$$x = \begin{bmatrix} x_{c_0}^T & \cdots & x_{c_N}^T & \lambda_{f_1} & \cdots & \lambda_{f_M} \end{bmatrix}^T$$
$$x_{c_k} = \begin{bmatrix} {}^G p_{I_k}^T & {}^G v_{I_k}^T & {}^G_{I_k} q^T & b_{a_k}^T & b_{g_k}^T \end{bmatrix}^T \quad (1)$$

Where state vector x includes $N+1$ keyframe states and M environmental features observed by these keyframes. x_{c_k} represents the IMU state at keyframe timestamp t_k, including the IMU's position ${}^G p_{I_k}$, velocity ${}^G v_{I_k}$, orientation ${}^G_{I_k} q$, accelerometer bias b_{a_k}, and gyroscope bias b_{g_k}. λ_{f_l} denotes the inverse depth of an environmental feature point f_l.

The structure-based VIO can be formulated as the following structure-based VI-BA problem

$$\min_x \left\{ \|r_p\|^2 + \sum_{k=1}^N \|r_{I_{k-1,k}}\|_{\Sigma_{I_{k-1,k}}}^2 + \sum_{l=1}^M \sum_{i \in K^l} \rho_{Hub} \left(\|r_{il}\|_{\Sigma_C}^2 \right) \right\} \quad (2)$$

Where r_p, $r_{I_{k-1,k}}\left(x_{c_{k-1}}, x_{c_k}\right)$ and r_{il} are the prior residual from marginalization, IMU preintegration residual and visual reprojection residual, respectively. Their detailed definitions are provided in Section VI of [18]. K^l is the set of keyframes observing feature point f_l. A robust Huber kernel function $\rho_{Hub}\left(\bullet\right)$ is used to mitigate the impact of pixel observation outliers.

III. PROPOSED STRUCTURELESS VIO

We develop a novel structureless VIO based on VINS-Mono [18]. For feature tracking, IMU preintegration and sliding-window marginalization, we reuse the modules from VINS-Mono. And the key modification is the formulation of visual residual. The state variables of structureless VIO is obtained by deleting the environmental feature points

$$
\begin{aligned}
x &= \begin{bmatrix} x_{c_0}^T & \cdots & x_{c_N}^T \end{bmatrix}^T \\
x_{c_k} &= \begin{bmatrix} {}^G p_{I_k}^T & {}^G v_{I_k}^T & {}^G_{I_k} q^T & b_{a_k}^T & b_{g_k}^T \end{bmatrix}^T
\end{aligned}
\tag{3}
$$

Unlike structure-based VIO, visual measurements are formulated by epipolar constraints, completely eliminating the dependence on 3D structure. Structureless VIO can be expressed as the following structureless VI-BA problem

$$
\min_x \left\{ \|r_p\|^2 + \sum_{k=1}^N \|r_{I_{k-1,k}}\|_{\Sigma_{I_{k-1,k}}}^2 + \sum_{l=1}^M \sum_{(i,j)\in K^l} \rho_{Hub}\left(\|r_{ij}^n\|_{\Sigma_C}^2\right) \right\}
\tag{4}
$$

Where r_p and $r_{I_{k-1,k}}\left(x_{c_{k-1}}, x_{c_k}\right)$ are described in Equation (2). While r_{ij}^n is the visual measurement residual generated by epipolar geometry. If a keyframe pair (i, j) observes the same feature point f_l, this co-view relationship can be expressed using the epipolar constraint

$$
\begin{aligned}
r_{ij}^n\left(x_{c_i}, x_{c_j}\right) &= \left({}^G_{I_j} R {}^I_C R z_j^n\right)^T \left[\frac{t}{\|t\|}\right]_\times \left({}^G_{I_i} R {}^I_C R z_i^n\right) \\
t &= {}^G p_{C_i} - {}^G p_{C_j} \\
&= {}^G p_{I_i} + {}^G_{I_i} R {}^I p_C - {}^G p_{I_j} - {}^G_{I_j} R {}^I p_C
\end{aligned}
\tag{5}
$$

Where $\left\{{}^I_C R, {}^I p_C\right\}$ are the extrinsic parameters between IMU and camera. z_i^n and z_j^n represent the normalized coordinates observations of the same feature point f_l from the keyframe pair (i, j). r_{ij}^n has intuitive geometric interpretation, i.e., two feature bearing vectors should be co-planar with the frame-to-frame translation vector t, as depicted in the Fig. 2. All direction vectors are expressed in the global frame. t is normalized to prevent it from converging to 0. Detailed Jacobians are provided in Section V-B of [23].
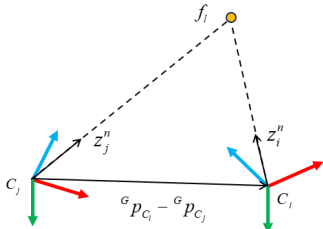

Fig. 2: Co-planar geometric relationships for feature bearing vectors with the frame-to-frame translation vector.

TABLE I: Performance comparison on EuRoC Dataset.

Sequence	ATE (m)		Avg solve time (ms)	
	VINS-Mono	Ours	VINS-Mono	Ours
MH_01_easy	**0.157**	0.172	39.27	**18.30**
MH_02_easy	**0.181**	0.223	38.96	**17.09**
MH_03_medium	**0.196**	0.233	39.90	**16.56**
MH_04_difficult	0.378	**0.257**	37.76	**15.57**
MH_05_difficult	0.303	**0.282**	38.56	**16.00**
V1_01_easy	0.082	**0.050**	39.22	**16.44**
V1_02_medium	0.112	**0.084**	33.54	**13.27**
V1_03_difficult	0.188	**0.103**	26.73	**10.84**
V2_01_easy	0.097	**0.078**	38.41	**16.66**
V2_02_medium	0.153	**0.116**	31.88	**13.50**
V2_03_difficult	0.298	**0.245**	20.87	**10.40**
Avg	0.195	**0.168**	35.01	**14.97**

TABLE II: Performance comparison on TUM-VI Dataset.

Sequence	ATE (m)		Avg solve time (ms)	
	VINS-Mono	Ours	VINS-Mono	Ours
room1	0.067	**0.051**	20.19	**11.56**
room2	**0.068**	0.104	22.67	**12.58**
room3	0.121	**0.114**	19.73	**11.89**
room4	**0.048**	0.063	21.03	**13.14**
room5	0.217	**0.127**	18.17	**11.27**
room6	**0.076**	0.085	26.24	**15.04**
Avg	0.100	**0.091**	21.34	**12.58**
corridor1	0.629	**0.398**	18.62	**11.46**
corridor2	**0.933**	0.956	19.06	**12.13**
corridor3	1.978	**0.893**	16.75	**10.98**
corridor4	0.315	**0.224**	19.92	**12.69**
corridor5	0.689	**0.456**	19.56	**12.36**
Avg	0.909	**0.585**	18.78	**11.92**

IV. RESULTS

Our structureless VIO is compared to the structure-based VINS-Mono [18]. The loop closure of VINS-Mono is disabled for fair comparison. For the setting of other parameters, we refer to VINS-Mono[1]. To verify the performance of different VIO algorithms, we employ two popular VIO datasets, EuRoC [24] and TUM-VI [25]. All the experiments are conducted on a laptop computer with an Intel(R) Xeon(R) W-10855M CPU @ 2.80GHz, and 16 GB of RAM.

Absolute trajectory error (ATE) [26] and the average solve time are recorded in TABLE I and TABLE II. Results demonstrate the structureless VIO scheme brings remarkable computational efficiency advantage over structure-based VIO. Moreover, structureless VIO can further improve accuracy and we attribute it to the independence of epipolar constraint from depth uncertainty, especially critical for corridor scenes.

V. CONCLUSION

In this paper, we introduce a novel structureless VIO that adopts epipolar constraint instead of reprojection error or photometric error to formulate visual measurement. Experimental results on two publicly benchmark datasets demonstrate that our method reduces the optimization time by a large margin and offers improvement in accuracy.

[1] https://github.com/HKUST-Aerial-Robotics/VINS-Mono/tree/master/config

REFERENCES

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[2] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *2014 IEEE international conference on robotics and automation (ICRA)*, pp. 15–22, IEEE, 2014.

[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: A versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[4] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[5] Google, "ARCore." https://developers.google.com/ar.

[6] Apple, "ARKit." https://developer.apple.com/augmented-reality.

[7] Meta, "Oculus." https://store.facebook.com/quest.

[8] Y. Fan, T. Zhao, and G. Wang, "Schurvins: Schur complement-based lightweight visual inertial navigation system," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17964–17973, 2024.

[9] K. J. Wu, C. X. Guo, G. Georgiou, and S. I. Roumeliotis, "Vins on wheels," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5155–5162, IEEE, 2017.

[10] J. Delmerico and D. Scaramuzza, "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots," in *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 2502–2509, IEEE, 2018.

[11] J. Kang, H. Kim, and K.-S. Kim, "View: Visual-inertial external wrench estimator for legged robot," *IEEE Robotics and Automation Letters*, 2023.

[12] J. Song, P. J. Sanchez-Cuevas, A. Richard, R. T. Rajan, and M. Olivares-Mendez, "Gps-vio fusion with online rotational calibration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11906–11912, IEEE, 2024.

[13] A. I. Mourikis, N. Trawny, S. I. Roumeliotis, A. E. Johnson, A. Ansar, and L. Matthies, "Vision-aided inertial navigation for spacecraft entry, descent, and landing," *IEEE Transactions on Robotics*, vol. 25, no. 2, pp. 264–280, 2009.

[14] D. S. Bayard, D. T. Conway, R. Brockers, J. H. Delaune, L. H. Matthies, H. F. Grip, G. B. Merewether, T. L. Brown, and A. M. San Martin, "Vision-based navigation for the nasa mars helicopter," in *AIAA Scitech 2019 Forum*, p. 1411, 2019.

[15] J. Delaune, D. S. Bayard, and R. Brockers, "Range-visual-inertial odometry: Scale observability without excitation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2421–2428, 2021.

[16] A. I. Mourikis and S. I. Roumeliotis, "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.

[17] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[18] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE transactions on robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[19] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4666–4672, IEEE, 2020.

[20] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[21] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, "Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback," *The International Journal of Robotics Research*, vol. 36, no. 10, pp. 1053–1072, 2017.

[22] L. Von Stumberg, V. Usenko, and D. Cremers, "Direct sparse visual-inertial odometry using dynamic marginalization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2510–2517, IEEE, 2018.

[23] J. Song, A. Richard, and M. Olivares-Mendez, "Improving monocular visual-inertial initialization with structureless visual-inertial bundle adjustment," *arXiv preprint arXiv:2502.16598*, 2025.

[24] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[25] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler, and D. Cremers, "The tum vi benchmark for evaluating visual-inertial odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1680–1687, IEEE, 2018.

[26] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7244–7251, IEEE, 2018.