Integrating Inertial Data to a Hybrid Direct-Indirect Visual SLAM System

Yan Song Hu[1] **John S. Zelek**[1]
[1] Vision and Image Processing Group, System Design Engineering, University of Waterloo
{y324hu,jzelek}@uwaterloo.ca

Abstract

A contemporary trend in the field of simultaneous localization and mapping (SLAM) is the application of sensor fusion to improve performance. There are many sources of additional data, including but not limited to inertial measurement units (IMU), event cameras, and depth data. This paper introduces a visual monocular SLAM system that tightly combines visual photogrammetric data, visually extracted geometric information, and inertial data. Our work improves on the energy function developed by H-SLAM [1], designed for joint optimization of photometric and geometric residuals in tracking, by allowing it to also handle inertial residuals. Furthermore, our SLAM system shares H-SLAM's [1] loop-closure mechanisms that are tightly coupled with the tracking process to ensure global consistency across large-scale maps. When tested on benchmarks, our system performs well compared to past SLAM systems that use photogrammetric, geometric, and inertial data and is competitive compared to state-of-the-art SLAM systems.

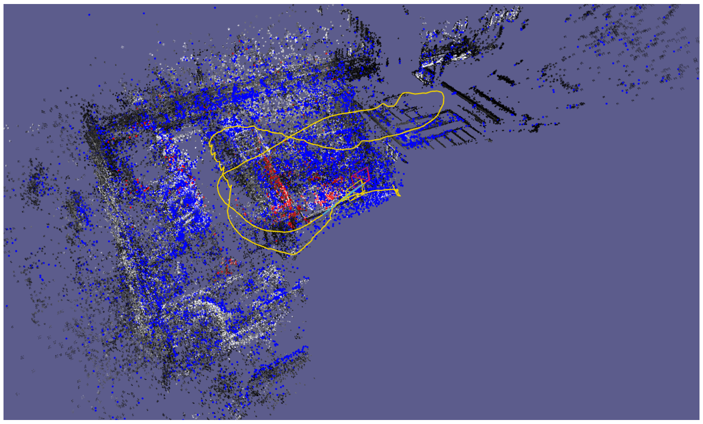

Fig. 1: The proposed system in operation on the EUROC MAV MH04 sequence. The coloured points are indirect key points and the gray points are the dense point cloud from the direct features.

1 INTRODUCTION

In situations where a mobile robot finds itself navigating within an unfamiliar environment, simultaneous localization and mapping (SLAM) is required. Because the environment is unknown, constructing a detailed map of the surrounding terrain is necessary to determine a precise location. Using a map for navigation, as enabled by SLAM, is more effective than map-less methods like odometry because it can correct accumulated drift errors. Additionally, the generated map has practical uses beyond navigation.

A specific variant of the SLAM problem is monocular visual SLAM. In visual SLAM, only 2D pixel data from a single camera is used. However, before mapping and localization can be done, visual sensor data has to be converted to a representation that is well-suited for navigation. This conversion process follows two distinct approaches: feature-based (indirect) and photogrammetric (direct). Feature-based methods transform images into a collection of distinctive and readily trackable 3D key points. In contrast, photogrammetric methods monitor pixel movements and aim to calculate the positional changes that are reflected by the movements.

Indirect and direct approaches work well together because they cover each other's limitations. Direct methods leverage a more extensive portion of the available pixel data in comparison to indirect methods, enabling them to function effectively in areas with lower textures. However, direct methods are much more susceptible to variations in lighting compared to indirect methods. Additionally, it is difficult to perform long-term loop closure with direct methods because direct pixel information is only relevant for a short time span. When used in tandem, the complementary strengths of the two methods can compensate for each other's limitations. This work builds upon the foundations laid by H-SLAM [1], which combines indirect and direct methods to become a hybrid SLAM system. We further add sensor information from an inertial measurement unit (IMU) to enhance performance.

IMUs are sensors that measure acceleration and angular rotation. When added to a SLAM system, IMUs allow visual SLAM systems to operate without visual input, calculate scale, and improve overall performance by providing additional data. Given the occasional loss of visual data in real-life robot operations, the inclusion of an IMU significantly improves robustness. This paper's novel contribution is the integration of an IMU into the hybrid direct-indirect SLAM framework of H-SLAM [1] through a novel energy function that fuses direct, indirect, and inertial residuals. H-SLAM [1] uses a unique descriptor-sharing approach to fuse the direct and indirect representations, which allows it to outperform contemporary hybrid SLAM systems. The proposed system stands out as one of the few SLAM systems that combines direct, indirect, and inertial data, achieving performance levels comparable the current state-of-the-art.

2 RELATED WORK AND BACKGROUND

This section serves a twofold purpose. Firstly, it offers context by providing an overview of direct and indirect SLAM methods whose ideas have contributed to the proposed system. Secondly, it compares the proposed system with existing hybrid inertial SLAM methods.

One of the first direct odometry approaches was Dense Tracking and Mapping (DTAM) [2]. Its use of direct features enhanced the robustness of the algorithm and facilitated a more comprehensive 3D reconstruction. Semi-direct Visual Odometry (SVO) [3] and Large Scale Direct SLAM (LSD-SLAM) [2] built upon DTAM's [2] work. SVO [3] used direct motion calculation methods on extracted key points, making it one of the first hybrid direct-indirect SLAM systems. LSD-SLAM [2] introduced the novel approach of using high-gradient regions instead of all available pixels for direct SLAM. Additionally, it incorporated loop closure by passing the data to OPENFABMAP [4], an independent appearance-based SLAM system. Direct Sparse Odometry (DSO) [5] further improved LSD-SLAM's [2] work by improving upon the pixel sampling method used. DSO would be further expanded to have loop closure as presented in the paper DSO with Loop Closure (LDSO) [6] and state-of-the-art inertial integration in Delayed-Marginalization Visual Inertial Odometry (DM-VIO) [7]. It's important to note that the principles introduced by DSO form the basis for the direct SLAM aspects of the proposed system.

The most important indirect SLAM system is ORBSLAM [8] [9]. One of its significant innovations is the usage of the same key point features for tracking, mapping, and loop closure, resulting in a tight integration of these components. In contrast, a common drawback in many direct SLAM formulations is the loose coupling between their local and global maps, as they do not employ the same features for both tracking and global loop closure. For example, LSD-SLAM [2] uses an entirely separate appearance-only SLAM system for its loop closure. Similarly, LDSO [6] extracts ORB features, which are not used for tracking, to encode a bag-of-visual-words [10] for loop closure. H-SLAM [1], which this work builds upon, attempts to solve this issue by using descriptor sharing, which associates patches of pixels to key points. Both pixel patches and key points are optimized for short-term tracking, while the ORB descriptor attached to the key points can be used for loop closure. Consequently, H-SLAM [10] uses its features for both tracking and global loop closure, resulting in a tightly integrated system.

By adopting the H-SLAM [10] framework for the fusion of direct and indirect points, the proposed system achieves a competitive edge when compared to other direct-indirect hybrid inertial SLAM systems. One example is SVO-Pro [11], a system that builds upon the foundations of SVO [3] by adding inertial information and introducing a SLAM module using through iSAM2 [12], an independent graph-based SLAM system. SD-VIS [13] is another comparable example. Similarly to SVO [3], SD-VIS [13] does tracking using inertial and direct visual information on non-keypoint frames and extracts descriptor features for keyframes, resulting in a loosely coupled approach. In contrast, the proposed method tightly integrates the loop closure process, setting it apart from other systems.

3 METHOD

3.1 Integration of the IMU into Tracking

The proposed system's architecture is fairly standard for visual SLAM and based on the framework developed by H-SLAM [10]. The novel improvement is the application of a new energy function that jointly optimizes photometric, geometric, and inertial residuals. The system performs a multi-objective optimization to minimize said energy function:

$$\mathbf{E}(\xi) = W(e_{visual}) \left[\frac{\|\mathbf{E}_p(\xi)\|}{n_p \sigma_p^2} + K \frac{\|\mathbf{E}_g(\xi)\|}{n_g \sigma_g^2} \right] + E_{imu}(\xi) + E_{prior}(\xi) \quad (1)$$

Where $E_p(\xi)$ being the photometric residuals, $E_g(\xi)$ being the geometric residuals, and $E_{imu}(\xi)$ being the inertial residuals. To balance the influence between the photometric and geometric residuals, each energy is divided by n, the count of each feature type, and σ^2, the residual variance. There is an additional K factor to reduce the influence of geometric residuals in low texture or blurry environments defined as:

$$K = \frac{5\exp(-2l)}{1+\exp(\frac{30-N_g}{4})} \quad (2)$$

Where l is the pyramid level and N_g is the number of inlier geometric matches.

The photometric residuals are calculated using the same method as DSO [5], and the geometric residuals are the difference between the predicted and perceived key-point positions. To solve the energy minimization problem, the Gauss-Newton method first proposed by Leutenegger et al. [14] and then applied to direct Systems by DSO [5] is used. The optimization algorithm is a sliding window Gauss-Newton approach that uses First Estimate Jacobians [15].

Because the inertial data arrives at a faster rate than the visual data, the IMU sensor inputs are first pre-integrated using a well-known method [16] for synchronization. The energy is then calculated using a method first proposed by Visual Inertial DSO [17]:

$$E_{imu}(s_i, s_j) := (s_j \boxminus \hat{s}_j)^T \hat{\Sigma}_{s,j}^{-1} (s_j \boxminus \hat{s}_j) \quad (3)$$

Where s_i and s_j are the two states, $\hat{s}$ is a predicted state, and $\hat{\Sigma}$ is the associated covariance matrix. The operator $\boxminus$ is an increment in the opposing direction when dealing with poses and normal subtraction for other components.

Furthermore, to deal with situations where there is bad image data, the $W(e_{visual})$ term is added to lower the influence of the visual direct and indirect residuals.

$$e_{visual} = \sqrt{\frac{E_p}{n_p} + \frac{E_g}{n_g}} \quad (4)$$

$$W(e_{visual}) = \lambda \begin{cases} \frac{\theta}{e_{visual}}, & \text{if } e_{visual} \geq \theta \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

Since the inertial data is now available, the gravity and scale variables are optimized as explicit variables. To handle the marginalization of terms with IMU factors, we use the delayed marginalization method proposed by DM-VIO [7]. Finally, the IMU initialization is done using the same method as DM-VIO [7], which leverages the delayed marginalization method to improve performance and robustness.

3.2 Integration of the IMU into Loop Closure

Because IMU measurements are based on odometry, they do not provide any information for the loop closure module. However, the IMU data is modified by the loop closure once a loop is detected. This is achieved by rotating the IMU data by the corrected pose rotation. The loop closure module is based on the framework presented by H-SLAM [10]. The loop closure is done using hybrid graphs that use both the co-visibility information from the indirect key points and pose-pose constraints provided by the temporally connected frames. In order to not break the optimization, the frames in the moving optimization window are frozen from the loop closure process.

4 RESULTS AND DISCUSSION

To test the performance of the proposed system, the system was tested on all the sequences of the commonly used EUROC MAV [18] dataset. This dataset is comprised of eleven visual-inertial sequences from captured drone footage within various indoor environments. Given the dataset's challenging attributes, including rapid drone motion and adverse lighting conditions in certain sequences, it serves as an effective test for a system's robustness. Because SLAM systems are not deterministic in their tracking point selection, the results are averaged over eight runs. The system is compared to both similar and state-of-the-art methods such as ORBSLAM3 [19]. The results are listed in table 1.

Set	Ours	HSLAM	LDSO	SD-VIS	DM-VIO	ORB3
MH01	0.076	0.035	0.053	0.261	0.065	0.062
MH02	0.04	0.034	0.062	0.290	0.044	0.037
MH03	0.101	0.140	0.114	0.577	0.097	0.046
MH04	0.119	0.334	0.152	0.497	0.102	0.075
MH05	0.107	0.141	0.085	0.512	0.096	0.057
V101	0.062	0.136	0.099	0.245	0.048	0.049
V102	0.068	0.193	0.087	0.502	0.045	0.015
V103	0.07	0.823	0.536	0.389	0.069	0.037
V201	0.032	0.051	0.066	0.202	0.029	0.042
V202	0.056	0.077	0.078	0.455	0.05	0.021
V203	0.115	1.257	X	0.445	0.03	0.027
Avg	0.077	0.3	0.13*	0.400	0.069	0.043

Table 1: Results on the EUROC MAV dataset. Error is in root mean squared average trajectory error in meters. Please note the benchmarked systems have different sensors suites and features. HSLAM and LDSO are monocular SLAM systems. DM-VIO is a monocular inertial visual odometry system with no loop closure. ORBSLAM3 (shown as ORB3 in the table), SD-VIS, and the proposed system are operated as monocular inertial SLAM systems. An average with a * means at least one of the sequences did not finish. All results except the proposed system use values reported by the authors of each system

Analyzing the results, it becomes clear that the proposed system consistently outperforms H-SLAM [10], upon which it is built, as well as older hybrid inertial SLAM systems such as SD-VIS [13]. Notably, The proposed system outperformed HSLAM [10] across the MH04, MH05, V103, and V203 datasets, which have more challenging motions and lighting conditions. This shows that adding an IMU significantly improved robustness. The proposed system also remains competitive when compared to state-of-the-art systems like ORBSLAM3 [19] and DM-VIO [7]. However, it's worth noting that the incorporation of loop closure does not yield a substantial improvement in the

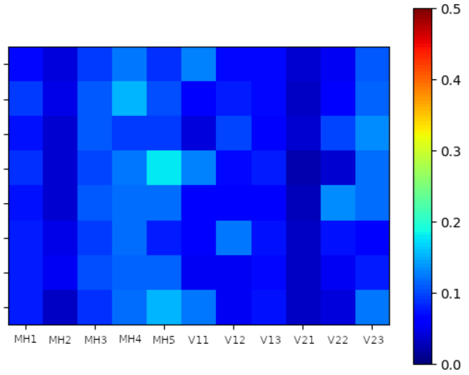

Fig. 2: Chart of the proposed system's performance in the EUROC MAV dataset. Eight runs were done for each of the sequences and then averaged.

proposed system when compared to DM-VIO [7]. This outcome may be attributed to the characteristics of the EUROC MAV dataset [18] because it lacks sequences with large loops, which are where loop closure mechanisms exhibit significant advantages. To further validate the effectiveness of the loop closure, additional testing should be done with datasets featuring large loops. Another reason may be the need for fine-tuning the loop closure process for the delayed marginalization system used. The delayed marginalization process was not extensively considered during the development of the system and should be a focus of future research.

4.1 Hardware and Implementation

The experiments were conducted on a desktop computer with an Intel Core i7-8700K CPU and NVIDIA 1080 TI. It's important to note that the GPU was not used. The execution involves three concurrent threads: tracking, mapping, and loop closure. The tracking thread processes each image as they are sequentially inputted. In comparison, mapping operations are done during keyframe insertions and loop closure is exclusively executed on loop detections. It was found that the runtime of the system was dependent on the size of the map, slowing as the map size increased. However, the proposed system consistently maintained real-time operation speeds during tests using the EUROC MAV [18] dataset.

5 CONCLUSIONS

In conclusion, this paper develops a SLAM system that utilizes a broader array of data sources, including visual photogrammetric data, visually derived geometric information, and inertial measurements. The system uniquely combines tracking that uses indirect key points, direct pixel patches, and inertial measurements with a tightly coupled loop closure system. The combination of systems allows for favorable performance when compared to similar prior SLAM systems and competitiveness compared to state-of-the-art SLAM systems. This paper underscores the potential of utilizing a diverse set of data sources in SLAM to improve the overall robustness and performance of the system.

References

[1] G. Younes, D. Khalil, J. Zelek, and D. Asmar, "H-slam: Hybrid direct-indirect visual slam," 2023.

[2] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 International Conference on Computer Vision*, 2011, pp. 2320–2327.

[3] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2017.

[4] A. Glover, W. Maddern, M. Warren, S. Reid, M. Milford, and G. Wyeth, "Openfabmap: An open source toolbox for appearance-based loop closure detection," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 4730–4735.

[5] 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, 2018.

[6] X. Gao, R. Wang, N. Demmel, and D. Cremers, "Ldso: Direct sparse odometry with loop closure," 2018.

[7] L. v. Stumberg and D. Cremers, "Dm-vio: Delayed marginalization visual-inertial odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1408–1415, 2022.

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

[9] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[10] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[11] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, 2017.

[12] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "isam2: Incremental smoothing and mapping with fluid relinearization and incremental variable reordering," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 3281–3288.

[13] Q. Liu, Z. Wang, and H. Wang, "Sd-vis: A fast and accurate semi-direct monocular visual-inertial simultaneous localization and mapping (slam)," *Sensors (Basel, Switzerland)*, vol. 20, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:212749808

[14] 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. [Online]. Available: https://doi.org/10.1177/0278364914554813

[15] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "A first-estimates jacobian ekf for improving slam consistency," in *Experimental Robotics*, O. Khatib, V. Kumar, and G. J. Pappas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 373–382.

[16] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," 07 2015.

[17] L. V. Stumberg, V. Usenko, and D. Cremers, "Direct sparse visual-inertial odometry using dynamic marginalization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2018. [Online]. Available: https://doi.org/10.1109%2Ficra.2018.8462905

[18] 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*, 2016. [Online]. Available: http://ijr.sagepub.com/content/early/2016/01/21/0278364915620033.abstract

[19] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. 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.