Asynchronous Visual-Inertial Odometry for Event Cameras

Peijing Li, Hexiong Yao, Zhiqiang Dai, Xiangwei Zhu
School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China
(lipj, yaohx5)@mail2.sysu.edu.cn, (daizhiqiang, zhuxw666)@mail.sysu.edu.cn

Keywords: Event Camera, SLAM, Event Visual-inertial Odometry, DVS, Event Feature.

Abstract

Event cameras are bio-inspired visual sensors that output changes in pixel-level brightness asynchronously instead of standard intensity frames at a fixed rate. These cameras offer reliable visual information in high-speed motion and high dynamic range (HDR) scenes, addressing the limitations of traditional cameras in such scenarios. Therefore, research of integrating event cameras into established visual algorithms holds significant value. In this study, based on traditional Visual-Inertial Odometry (VIO) frameworks, we proposed an innovative asynchronous monocular event-based inertial odometry method to fully exploit the benefits of event cameras. First, the corner features are extracted separately from the raw event stream and the time surface map, followed by uniform feature selection to accurately describe three-dimensional spatial geometry. Then, feature tracking is achieved by integrating these two event representation methods. In addition, our method obtains stable and high frequency state estimation by fusing event and IMU measurements through graph optimization. We validate the effectiveness of our proposed approach, comparing with several state-of-the-art EVIO systems and VIO systems.

1. Introduction

The task of estimating a sensors ego-motion has important applications in various fields, such as augmented/virtual reality or autonomous robot control. In recent years, significant attention has been directed towards the development of visual-inertial state estimation systems. This is primarily due to the complementary nature of information provided by visual sensors, such as cameras, and inertial sensors, specifically Inertial Measurement Units (IMUs), which enhance the robustness of conventional camera pose estimation systems. However, due to the limitations of standard cameras, visual-inertial systems still struggle to cope with some challenging situations.

Event cameras, also referred to as Dynamic Vision Sensors (DVS), hold significant promise in addressing challenges encountered by SLAM systems in real-world scenarios(Gallego et al., 2020). Unlike conventional cameras that capture intensity values of all pixels at fixed time intervals, event cameras transmit information via an asynchronous event stream, which records changes in luminance. This unique characteristic equips event cameras with superior performance in scenarios involving high-speed motion and High Dynamic Range (HDR), but also posing challenges for their integration into traditional frame-based visual SLAM algorithms.

Existing research attempts to tackle this challenge through various approaches, including reconstructing images from events(Kim et al., 2016), accumulating events over time to form event frames(Rebecq et al., 2017), or integrating event cameras with other sensors like standard cameras(Hidalgo-Carrió et al., 2022) and RGBD cameras(Zuo et al., 2022). However, these methods often lack real-time capability, limiting the potential effectiveness of event cameras in high-speed motion scenarios.

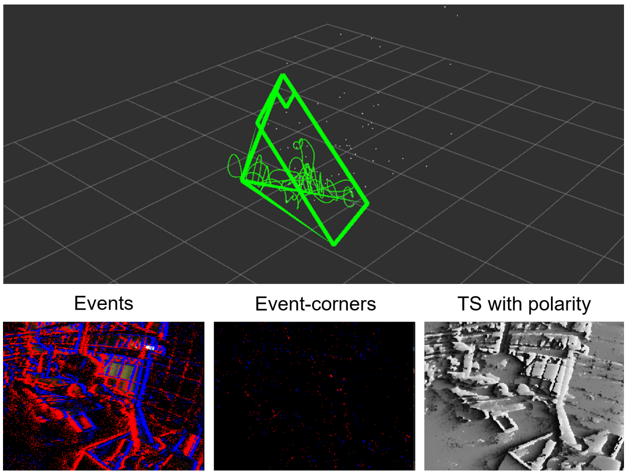


Figure 1. Our proposed method combined event-corners from raw event stream and corners from TS without polarity to provide robust state estimation in real-time. Bottom Left: Events, Bottom Middle: Event-corners, Bottom Right: Time Surface with polarity (blue: positive events, red: negative events).

To maximize the benefits of event cameras, our study proposes a monocular Visual-Inertial Odometry method that integrates event stream corner extraction and matching. This approach detects corners in the event stream through two distinct methods and utilizes them for matching and tracking to achieve asynchronous real-time pose estimation. The primary contributions of this study are outlined as follows:

- We propose a method for feature selection and tracking that integrates raw event stream and event frame data, which decrease the computational load of feature tracking and pose estimation in SLAM systems. Consequently, our system is capable of real-time pose estimation using a maximum of 60Hz feature tracking output.
- Our EVIO algorithm presents an asynchronous event-IMU tightly coupled framework. Additionally, our proposed system cam bootstrap from unknown initial states stably.
- We validate the feasibility of the algorithm across challenging scenarios on multiple datasets, including high-

speed motion and significant changes in brightness. Furthermore, we conduct various comparative experiments to comprehensively evaluate the algorithm's performance from multiple perspectives.

The rest of the paper is organized as follows: Chapter 2 discusses the relevant traditional visual SLAM and event-based camera SLAM algorithms. Chapter 3 presents the overall framework of the proposed EVIO. Chapter 4 describes a series of experiments and their outcomes conducted on the publicly available Event Camera Dataset(Mueggler et al., 2017) and Indoor datasets (Guan and Lu, 2022). Chapter 5 provides the paper's conclusion.

2. Related Work

2.1 Visual Odometry with Event Camera

Event-based monocular VO has been intensively researched for challenging scenarios in recent years. The earliest purely event-based 6-DoF VO(Kim et al., 2016) recovers image intensity, performing real-time event-based SLAM through three decoupled probabilistic filters that jointly estimate the 6-DoF camera pose, 3D-map of the scene, and image intensity. EVO(Rebecq et al., 2016) proposed to solve the SLAM problem without recovering image intensity, performs a geometric approach which combines a tracking approach based on image-to-model alignment and semi-dense 3D reconstruction algorithm(Rebecq et al., 2018) in parallel. Subsequent monocular visual odometry tasks are typically enhanced by the integration of additional sensor information. EDS(Hidalgo-Carrió et al., 2022) introduces a monocular visual odometry method that integrates brightness change events from an event camera with conventional frame imageswhile Zuo et al.(Zuo et al., 2024) presents a semi-dense 6-DOF tracking method for event cameras under challenging conditions by integrating depth camera information or pre-built map data. Newest purely event-based monocular SLAM algorithms are mostly based on the contrast maximization(CM) framework(Gallego et al., 2018), which directly recover the parameters that describe the relative motion between the camera and the scene by raw events. Kim and Kim(Kim and Kim, 2021) estimates rotational motion over globally aligned events using this framework. CMax-SLAM(Guo and Gallego, 2024) the aforementioned research and proposes the first event-based SLAM system which leverages CM framework to optimize rotational motion estimation. However, these methods currently can only estimate pure rotation and require large computational resources, making them unsuitable for real-time pose estimation systems.

2.2 Visual-Inertial Odometry with Event Camera

Monocular vision-based SLAM systems encounter scale uncertainty issues. Therefore, a trend is to integrate IMUs into visual SLAM systems. IMUs can rapidly provide independent estimates of position and attitude, which is crucial for scenarios where visual information is unreliable or infrequently updated. For traditional visual SLAM, VINS(Qin et al., 2018) and ORB-SLAM3(Campos et al., 2021) are two of the most classic optimization-based VIO frameworks. For event-based SLAM, Zihao Zhu et al.(Zihao Zhu et al., 2017) proposed the first event-based VIO that tackles the incomplete estimation of scale and provides accurate 6-DoF state estimation based on Extended Kalman Filter (EKF). Rebecq et al.(Rebecq et al., 2017) obtains a discrete number of states based on a spatio-temporal window of event streams, and introduces virtual event frames to achieve nonlinear optimization that refines estimated poses. Ultimate SLAM(Vidal et al., 2018) furthered the aforementioned research by combining event streams, image frames, and IMU measurements with nonlinear optimization, which leverages the complementary advantages of event cameras and standard cameras. Mueggler et al.(Mueggler et al., 2018) adopted a continuous-time framework based on cubic spline for smooth trajectory estimation and fused both event streams and IMU together. EKLT-VIO(Mahlknecht et al., 2022) integrated an accurate state-of-the-art event-based feature tracker EKLT(Gehrig et al., 2020) with EKF backend to achieve event-based state estimation on Mars-like datasets. Xu et al.(Xu et al., 2023) proposes a tightly coupled method for direct velocity estimation using a dynamic vision sensor and an inertial measurement unit, enhancing the accuracy and robustness of velocity estimation in high-dynamic scenarios through trifocal tensor geometry and a two-layer RANSAC scheme. However, the accuracy of these loosely coupled methods still needs to be improved, and tightly coupled methods usually do not have real-time capabilities, even with low-resolution event cameras (240*180).

3. Methodology

The EVIO framework proposed in this study is illustrated in Figure 2. The algorithm primarily consists of two modules: 1) The front-end odometry module for corner extraction and feature tracking based on raw event streams and event frames. We utilize asynchronous corners directly extracted from the raw event stream and image corners extracted from event frames as features, followed by matching and tracking on the Time Surface image with polarity. Finally, asynchronous pose estimation results are obtained based on feature tracking, generating a sparse map of feature points. 2) The back-end graph optimization module tightly couples visual landmarks from the event camera and IMU pre-integration information, resulting in the final asynchronous 6-degree-of-freedom pose estimation.

3.1 Event Representation

The event camera outputs a series of asynchronous event streams. Each event $e_k = \{x_k, y_k, t_k, p_k\}$ comprises spatio-temporal coordinates of intensity changes along with their polarity p, taking values from $\{-1, +1\}$. Our system utilizes the Time Surface (TS) with polarity generated by the Surface of Active Events (SAE) technique for front-end odometry. The time surface, serving as a two-dimensional map, assigns each pixel the timestamp of the last event recorded at that pixel, denoted as $t_{\text{last}}(x, y)$. Using an exponential decay kernel, the time surface emphasizes recent events over past events. Specifically, at any given event time t, the intensity value at pixel (x, y) when $t \geq t_{\text{last}}(x, y)$ is defined as

$$T(x, y, t) = p \cdot \exp\left(-\frac{t - t_{\text{last}}(x, y)}{\eta}\right) \tag{1}$$

where η is the decay rate parameter (30-40ms in our experiments). We use TS with polarity due to their computational and memory efficiency, along with their rich information content on edges. The polarity proves advantageous for feature tracking as it indicates the direction of event changes, which easily

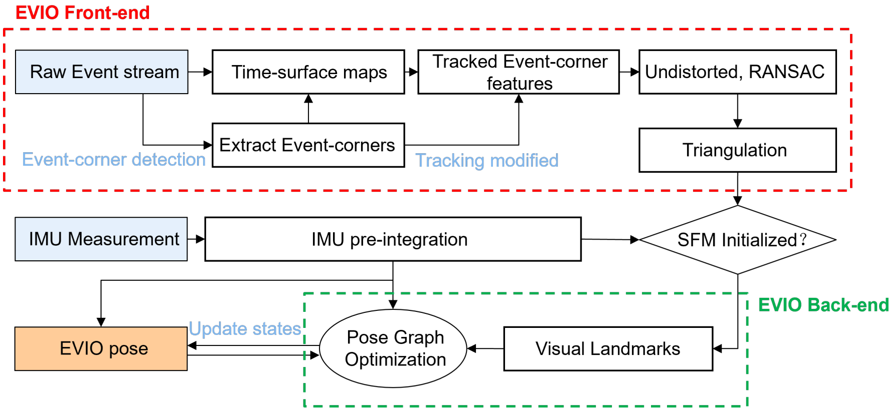


Figure 2. Overview of our proposed EVIO pipline

respond to edges in the scene in presence of different relative motions. However, in some situation, as shown in Figure 3, TS maps exhibit poor quality when there are few events, leading to unstable feature tracking. We address these scenarios by histogram equalization or other strategies, which will be introduced in Section 3.2.

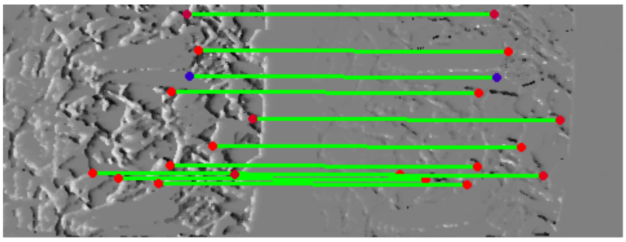


Figure 3. Tracking under reverse motion.The decreasing speed results in fewer events being generated, which consequently degrades the quality of TS map, leading to the failure of a significant number of tracking points.

3.2 Feature Extraction and Tracking based on Event Stream and Time Surface Map.

Traditional methods fail to yield an adequate number of consistently trackable feature points, while event-based corner detection struggles to distinguish noise from true corner points, thus hard to accurately representing spatial geometric relationships. To tackle this challenge, our study proposes a method that combines event-based corner detection with traditional techniques for feature extraction and tracking. This integration enhances the accuracy of feature tracking results in expressing spatial geometric relationships, reducing the required number of feature points for tracking. Specific comparative results are illustrated in Figure 4. We emphasize the distinctions between the feature extraction outcomes of the two methods using yellow boxes for clarity. By incorporating the camera’s motion direction into corner extraction, our method achieve a more accurate mapping of the 3D object edges onto the TS maps, which enhance the stability and precision of feature tracking. In contrast, the Harris method relies solely on TS maps, tends to lose track of features across successive multiple continuous TS maps. Furthermore, our method transcends the limitations imposed by varying image qualities within a single TS plot, ensuring a more consistent feature point acquisition process. This results in a superior representation of the three-dimensional spatial information, capturing the essence of the scene more effectively.

Our study enhance the publicly available Arc* algorithm to perform corner detection based on SAE for event-based corner extraction. The Arc* detector creates two sets of elements with different radii circular regions centered at the location of a new event, then calculates the lengths of contiguous circular arcs within these sets. If the length of a circular arc falls within a certain threshold, a corner point is considered to be detected. Notably, unlike traditional methods such as FAST corners, event-based corner detection methods lack the capability to assign scores to corners. The prioritization of corners relies solely on the time interval between the moment of corner detection and the current moment. This approach is susceptible to noise interference in regions with low texture, often leading to inaccuracies in selecting event corners that fully represent spatial geometric relationships. To overcome this challenge, we employ a supplementary strategy. When the number of points requiring detection exceeds a certain threshold, a limited number of corners with significant geometric information are extracted from the TS map using traditional corner detection methods. Otherwise, only event corners are utilized for spatial geometric information supplements. To enforce the uniform distribution, a minimum distance (10-20 pixels for different resolution event camera) is set between two neighboring event-corner features.

After feature extraction, we begin by applying the Lucas-Kanade (LK) optical flow method to match feature points extracted from TS maps. However, as we have previously noted, the TS plot may occasionally suffer from a dearth of events, leading to suboptimal quality. Such a shortfall substantially influences the efficacy of the Lucas-Kanade optical flow tracking, underscoring the need for robust methods to address these limitations. To address this challenge, when the tracking fails and the distribution of the surrounding feature points is relatively sparse, we search for the matching event corner by referring to the optical flow of succeeded tracked features. Especially, when the number of successfully tracked feature points falls below a certain threshold, we assume the camera to be in a stationary state, retaining the features extracted from the previous frame and the TS plot for use in the subsequent tracking process.

Feature pairs are identified after tracking. All the event-corner features are first undistorted based on the camera distortion mode. These features are then projected onto the normalized camera coordinate system. To remove outliers, we use the Random Sample Consensus (RANSAC) for further filtering. Subsequently, we recover the inverse depth of the features that are successfully tracked between two consecutive timestamps through triangulation. The landmark whose 3D position has

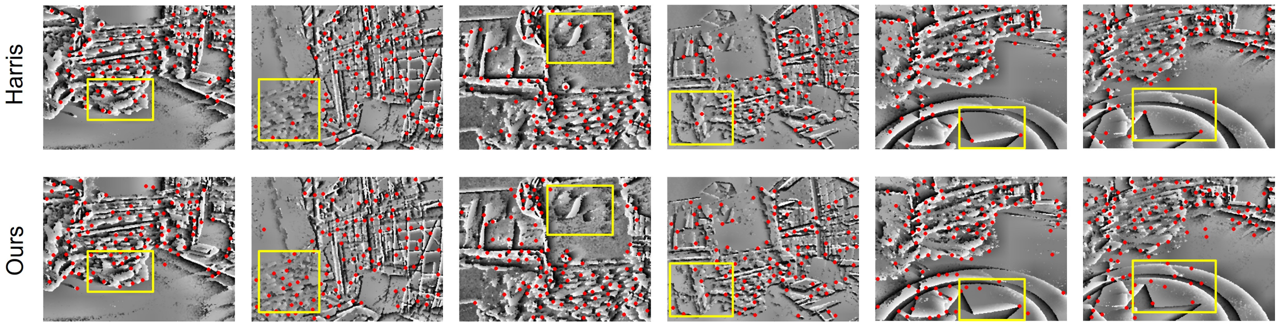
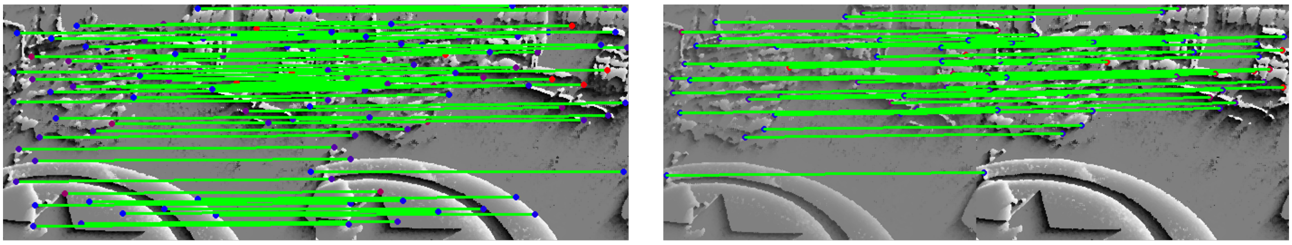


Figure 4. Comparison of our feature extraction method and Harris under several scenarios. Feature extraction from SAE takes into account the direction of motion and can uniformly extract points even in areas where the TS map quality is poor.



(a)

(b)

Figure 5. Tracking results using different method: (a) Our proposed method; (b) Harris.

been successfully calculated would be fed to the sliding window for pose graph optimization.

3.3 Pose Graph Optimization

In our proposed system, the full state vector χ in the sliding window is defined as:

$$\begin{aligned} \chi &= \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \lambda_1, \lambda_2, \ldots, \lambda_m, \mathbf{T}_c^b\right] \\ \mathbf{x}_i &= \left[\mathbf{p}_{b_i}^w, \mathbf{q}_{b_i}^w, \mathbf{v}_{b_i}^w, \mathbf{b}_{a_i}, \mathbf{b}_{g_i}\right], i \in [0, n] \end{aligned} \tag{2}$$

where $\mathbf{T}_c^b = \left[\mathbf{R}_c^b, t_c^b\right]$ is the extrinsic transformation from the camera frame c to the body (IMU) frame b; λ_j is the inverse depth of the j-th point feature from its first observation; $\mathbf{x}_i$ represents the body state in the i-th sliding window, which is made up of the following parameters: position $\mathbf{p}_{b_i}^w$, orientation quaternion $\mathbf{q}_{b_i}^w$, velocity $\mathbf{v}_{b_i}^w$ of the IMU in the world frame, acceleration bias $\mathbf{b}_{a_i}$, and gyroscope bias $\mathbf{b}_{g_i}$.

The maximum a posteriori(MAP) estimation of χ is solved by a joint nonlinear optimization, which cost function can be written as:

$$J(\chi) = \|e_p\|^2 + \sum_{k=1}^{n} \left\|e_i^k\right\|^2 + \sum_{k=1}^{n} \sum_{l \in \xi} \left\|e_c^{k,l}\right\|^2 \tag{3}$$

where e_p, e_i^k, $e_c^{k,l}$ represent the marginalization residual, the IMU pre-integration residuals and the event measurement residuals from the re-projection function, respectively. The set χ contains the event features that have been tracked at least twice in the current sliding window, and the re-projection function is defined as:

$$e_c^{k,l} = \mathbf{z}_l^k - \pi_c \left(\mathbf{T}_c^b\right)^{-1} \mathbf{T}_{b_i}^{b_k} \mathbf{T}_c^b \pi_c^{-1} \lambda_l^{-1} \mathbf{z}_l^i \tag{4}$$

where z_l^k and z_l^i is the measured image coordinate of the l-th feature in the k-th and the i-th keyframe, respectively. π_c is the event camera projection model, obtained from prior intrinsic calibration, and $\mathbf{T}_{b_i}^{b_k}$ is the incremental transformation between the camera poses at the k-th and the i-th keyframe. In addition, we employ the Ceres solver to carry out sliding window optimization. For marginalization, we adopt a two-way marginalization strategy to eliminate states from the sliding window, implementing marginalization via the Schur complement method. This approach ensures real-time performance of the system while optimizing computational efficiency, enabling it to handle large amounts of data while maintaining high accuracy.

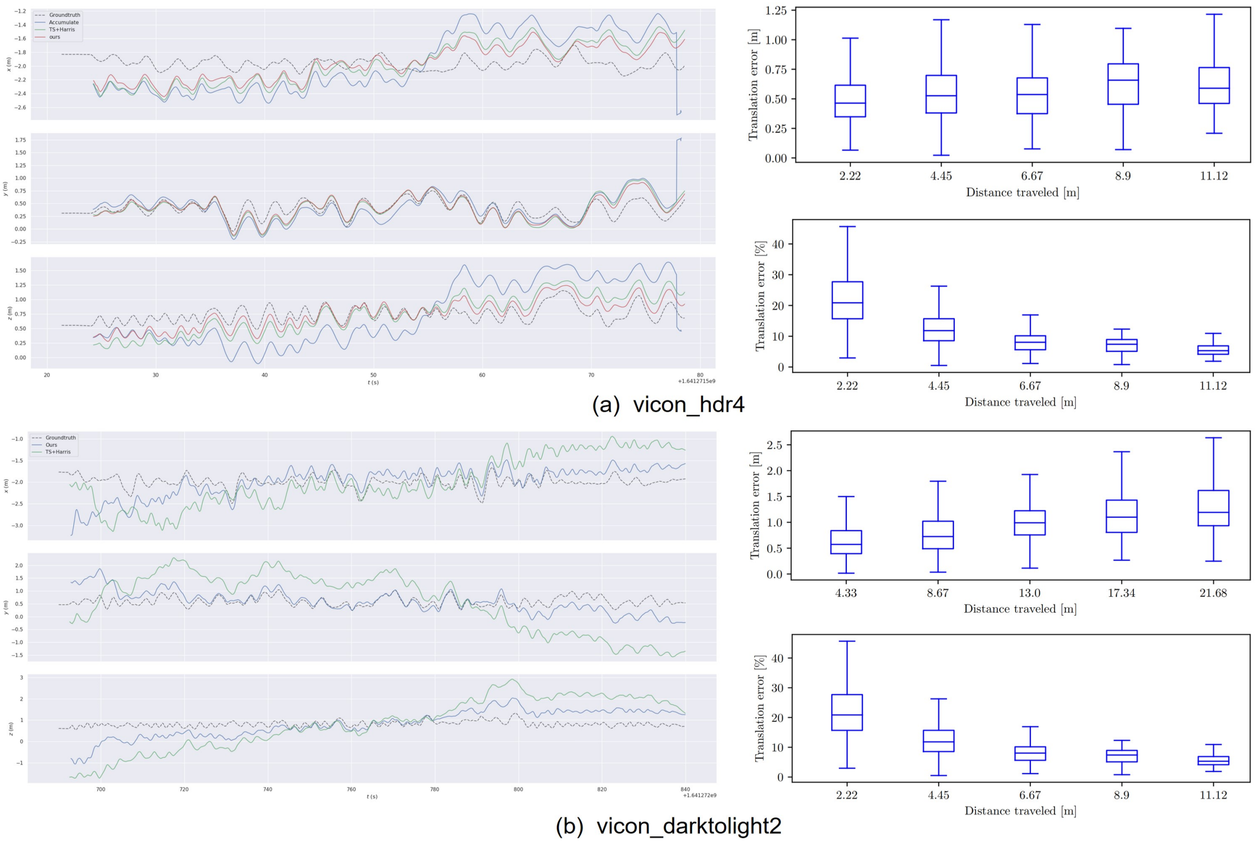
3.4 Additional Implementation Details

(a) Initialization: Adopted from (Qin et al., 2018), the initialization of our EVIO starts with a vision-only structure from motion (SfM) to build the up-to-scale structure of camera pose and event-corner feature positions, then loosely aligning the SfM with the pre-integrated IMU Measurements. During initialization, we chose to generate time-surface maps using a fixed number of events, and extract FAST corners. This strategy can reduce the effect of stationary motion and event corner from noise, improving the accuracy and stability of initialization.

(b) Still State: Since the event cameras output very little events (only noise) when the sensor is still, this will always lead to a failure feature detecting and tracking. To tackle this problem, we set a threshold, and when the number of events at the interval is less than the threshold, the old time surface map and event corners will copy to the new event frame.

4. Experiments

To validate the effectiveness of our system, we performed experiments on different sequences from two public datasets(Mueggler et al., 2017, Guan and Lu, 2022). To demonstrate



(a) vicon_hdr4

(b) vicon_darktolight2

Figure 6. Comparison of translation estimates of our proposed EVIO against ground truth, TS+Harris method and accumulated event frame method; The relative errors of the translation of our method.

the advantages of our algorithm, we first we compare it with different state-of-art EVIO methods. Then, we conduct comparative experiments VINS-Mono and ORB-SLAM3 without loop detection to highlight event cameras’ superiority in challenging scenarios. All experiments run in real-time on a computer equipped with an Intel Core i7-14700k processor. Our algorithm is implemented in C++ on Ubuntu 20.04 and ROS Noetic. Pose estimation results are evaluated using open-source tools(Grupp, 2017, Zhang and Scaramuzza, 2018).

4.1 Comparison with EVIO Works

To compare the accuracy of our EVIO framework with other state-of- art EVIO, we conducted experiments on the Event Camera Dataset(Mueggler et al., 2017). This dataset is the most commonly used publicly available datasets among EVIO research, captured using a DAVIS240C camera with a resolution of 240*180 pixels, featuring rapid six degrees of freedom motion and scenes with High Dynamic Range (HDR). The extrinsic and intrinsic parameters of the camera and IMU were calibrated using the calibration sequence of this dataset with Kalibr. Due to the lack of open-source EVIO code and difficulty of fine-tuning the parameters, we use the raw result from two established algorithms for comparison: Ultimate SLAM(Vidal et al., 2018), an optimization-based EVIO algorithm, and EKIT-VIO(Mahlknecht et al., 2022), a filter-based EVIO algorithm. The estimated and ground-truth trajectories were aligned with a 6-DOF transformation in SE3) using 5 seconds of the resulting trajectory as Ultimate SLAM(USLAM) and EKIT-VIO, calculated by open-source tool(Zhang and Scaramuzza, 2018). Superior performance highlighted in bold black. As seen in Table 1, we find that our system performs comparably to state-of-art EVIO systems.

Sequence	Ours	TS+Harris	USLAM	EKLT-VIO
poster_translation	0.29	0.45	**0.15**	0.35
poster_6dof	0.37	**0.29**	0.30	0.35
dynamic_6dof	0.40	0.60	**0.38**	0.79
hdr_boxes	**0.44**	0.57	0.67	0.46
hdr_poster	**0.29**	0.30	0.49	0.65
Average	**0.36**	0.64	0.40	0.52

Table 1. Performance Comparasion on Event Camera Dataset.

We conducted further experiments on an indoor dataset(Guan and Lu, 2022). This dataset comprise indoor scenes captured using the high-resolution event camera DAVIS346 (346*260) under HDR conditions, characterized by very low light or strong light variations. Ground truth is provided by Vicon, and some sequence spans over 140 seconds. Our experiments utilize only event streams and IMU data, with extrinsic and intrinsic parameters provided by the dataset. The frontend output frame rate souranges between 40 and 60 Hz, with 80 - 100 features extracted. We evaluate the localization accuracy using the root mean square errors (RMSE) of the absolute trajectory error (ATE) and relative pose error (RPE), calculated using the open-source evo tool(Grupp, 2017).

We tried our best to fine-tune the parameters of Ultimate SLAM. However, the system fail to provide comparable pose estimation results because of poor initialization and insufficient number of tracked features. Therefore, we choose to compare our system with two different methods: algorithm using Harris instead of our proposed feature extraction method and

Sequence	Ours	TS+Harris	Accumulate
hdr1	**0.61**	0.71	failed
hdr2	0.66	**0.63**	failed
hdr3	**0.40**	1.33	0.97
hdr4	**0.32**	0.43	0.68
darktolight1	**0.62**	0.84	0.71
darktolight2	**0.47**	1.39	failed
lighttodark1	**0.66**	0.80	failed
lighttodark2	**0.28**	0.37	0.31
dark1	**1.08**	1.44	failed
Average	**0.57**	0.88	0.67

Table 2. Performance Comparasion on Indoor Dataset.

Vins-fusion with event accumulated image as input(Xiao et al., 2022). Table 2 presents the computed and compared average RMSE for each VIO system. We define a situation in which pose estimation interruptions or errors within a sequence exceed the average level by an order of magnitude as "failed." From Table 2, it’s evident that the method relying on event accumulated images struggles to maintain stable state estimation output over these sequences. This is because the fixed number of accumulated events is unable to adapt fluctuations in event data generation caused by varying camera motion speeds, especially with higher-resolution cameras (346*260). This underscores the significance of using TS maps as event representation for tracking, which reduce parameter tuning complexity and improve feature extraction and tracking robustness. Our proposed approach outperforms both the TS+Harris and event accumulation methods on 9 out of 10 sequences.

We also calculate the runtime of each module, which is the average runtime across all sequences in these two publicly datasets, as shown in Table 3. It is evident that the corner detection algorithm we utilized significantly improves the feature extraction efficiency.

Models	Ours Davis240c	Harris Davis240c	Ours Davis346	Harris Davis346
TS Creation	2.506	2.526	5.251	5.250
Feature Extraction	0.017	0.613	0.018	1.174
Front-end Process	1.610	2.292	1.795	2.981
Back-end Process	8.222	8.332	6.753	7.340

Table 3. Running Time of different model(ms)

4.2 Comparison with VIO Works

We also compared our proposed EVIO method with the most common VIO methods. In the Indoor dataset, standard images were captured using Davis346, and their intrinsic and extrinsic parameters match those used for capturing the event stream. We employed these standard images and IMU data to run Vins-mono and VO version of ORB-SLAM3 with loop detection. This is because the VIO version of ORB-SLAM3 failed or cannot initialize in all the sequences, and it cannot estimate pose through the whole sequence without loop detection. The comparisons are presented in Table 4, with ORB-SLAM3 results shown in the gray columns.

Our system achieves robust and accurate pose estimation even in scenarios with high-speed motion and intense HDR conditions. The main reason is better feature extraction shown as Figure 7. Vins-mono uses histogram equalization to address dark scenarios, while both Vins-mono and ORB-SLAM3 can’t extract features on the opening light due to the principle of camera imaging. The VO version of ORB-SLAM3 seems to be outperforms our system, however, its efficacy largely depends on its relocalization strategies and loop detection. Notably, ORB-SLAM3 would track failures and lose tracking frames during the aggressive motion or too dark scenarios, affecting descriptor generation severely and causing interruptions in pose estimation as Figure 8 shown. In contrast, our system ensures continuous and smooth state estimation, rendering it better suited for real-time navigation applications.

Sequence	Ours	Vins-mono w/o loop	ORB-SLAM3 w/ loop
hdr1	**0.61**	1.07	0.24
hdr2	**0.66**	0.93	0.20
hdr3	**0.40**	0.42	0.20
hdr4	**0.32**	0.33	0.36
darktolight1	**0.62**	2.54	0.28
darktolight2	**0.47**	1.19	0.14
lighttodark1	0.66	**0.50**	failed
lighttodark2	0.28	**0.24**	0.25
dark1	1.08	**0.48**	failed
Average	**0.57**	0.86	0.24

Table 4. Performance comparasion with VIO.

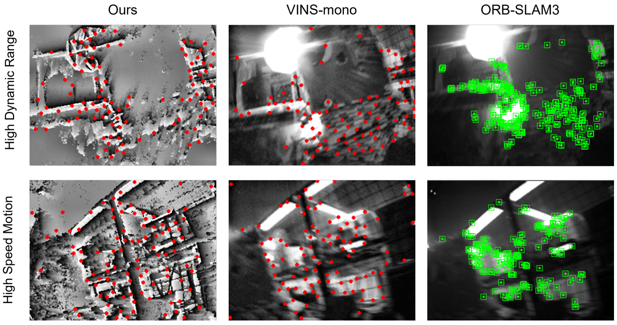


Figure 7. Comparison of feature extraction in our EVIO method, VINS-mono and ORB-SLAM3 under high dynamic range and high speed motion scenarios.

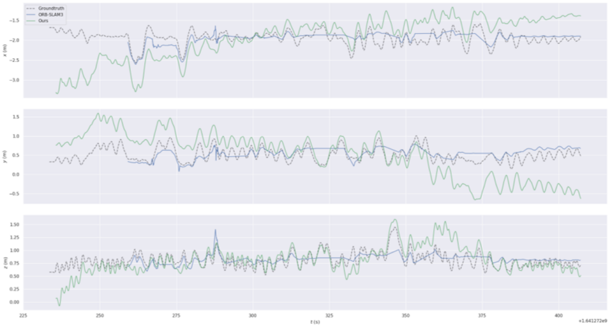


Figure 8. Comparison of translation estimates of our proposed EVIO method and ORB-SLAM3 in darktolight1 sequence.

5. Conclusion

In this paper, we introduce an asynchronous monocular event-based inertial odometry system, which provides real-time estimation of 6-dof pose up to 60Hz. To enhance the system’s real-time performance and accuracy, we propose a method that combines event stream and event frame for corner extraction and tracking. Additionally, we present an initialization strategy aimed at improving the accuracy and stability of dynamic initialization for event cameras under unknown initial conditions. Our experimental evaluations on publicly available datasets demonstrate that our approach exhibits commendable performance compared to state-of-the-art EVIO and VIO systems. In future research, we will explore the fusion of traditional imagery with EVIO systems by establishing event generate model. Moreover, leveraging deep learning techniques for event stream feature extraction and tracking may offer valuable insights for enhancing the overall system performance.

References

Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M., Tardós, J. D., 2021. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6), 1874–1890.

Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K. et al., 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1), 154–180.

Gallego, G., Rebecq, H., Scaramuzza, D., 2018. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3867–3876.

Gehrig, D., Rebecq, H., Gallego, G., Scaramuzza, D., 2020. EKLT: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3), 601–618.

Grupp, M., 2017. evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo.

Guan, W., Lu, P., 2022. Monocular event visual inertial odometry based on event-corner using sliding windows graph-based optimization. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2438–2445.

Guo, S., Gallego, G., 2024. CMax-SLAM: Event-based Rotational-Motion Bundle Adjustment and SLAM System using Contrast Maximization. *IEEE Transactions on Robotics*.

Hidalgo-Carrió, J., Gallego, G., Scaramuzza, D., 2022. Event-aided direct sparse odometry. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5781–5790.

Kim, H., Kim, H. J., 2021. Real-time rotational motion estimation with contrast maximization over globally aligned events. *IEEE Robotics and Automation Letters*, 6(3), 6016–6023.

Kim, H., Leutenegger, S., Davison, A. J., 2016. Real-time 3d reconstruction and 6-dof tracking with an event camera. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, Springer, 349–364.

Mahlknecht, F., Gehrig, D., Nash, J., Rockenbauer, F. M., Morrell, B., Delaune, J., Scaramuzza, D., 2022. Exploring event camera-based odometry for planetary robots. *IEEE Robotics and Automation Letters*, 7(4), 8651–8658.

Mueggler, E., Gallego, G., Rebecq, H., Scaramuzza, D., 2018. Continuous-time visual-inertial odometry for event cameras. *IEEE Transactions on Robotics*, 34(6), 1425–1440.

Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D., 2017. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. *The International Journal of Robotics Research*, 36(2), 142–149.

Qin, T., Li, P., Shen, S., 2018. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4), 1004–1020.

Rebecq, H., Gallego, G., Mueggler, E., Scaramuzza, D., 2018. EMVS: Event-based multi-view stereo3D reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12), 1394–1414.

Rebecq, H., Horstschaefer, T., Scaramuzza, D., 2017. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization.

Rebecq, H., Horstschäfer, T., Gallego, G., Scaramuzza, D., 2016. Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. *IEEE Robotics and Automation Letters*, 2(2), 593–600.

Vidal, A. R., Rebecq, H., Horstschaefer, T., Scaramuzza, D., 2018. Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2), 994–1001.

Xiao, K., Wang, G., Chen, Y., Xie, Y., Li, H., Li, S., 2022. Research on event accumulator settings for event-based slam. *2022 6th International Conference on Robotics, Control and Automation (ICRCA)*, IEEE, 50–56.

Xu, W., Peng, X., Kneip, L., 2023. Tight fusion of events and inertial measurements for direct velocity estimation. *IEEE Transactions on Robotics*.

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

Zihao Zhu, A., Atanasov, N., Daniilidis, K., 2017. Event-based visual inertial odometry. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5391–5399.

Zuo, Y.-F., Xu, W., Wang, X., Wang, Y., Kneip, L., 2024. Cross-modal semi-dense 6-dof tracking of an event camera in challenging conditions. *IEEE Transactions on Robotics*.

Zuo, Y.-F., Yang, J., Chen, J., Wang, X., Wang, Y., Kneip, L., 2022. Devo: Depth-event camera visual odometry in challenging conditions. *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2179–2185.