Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Correspondence:

SEVAR: a stereo event camera dataset for virtual and augmented reality*

Yuda DONG[†1,3], Zetao CHEN[†‡4], Xin HE[†‡1,2], Lijun LI[4], Zichao SHU[4], Yinong CAO[1,3], Junchi FENG[1,3], Shijie LIU[1], Chunlai LI[1,2], Jianyu WANG[1,2]

[1]*Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China*

[2]*Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China*

[3]*University of Chinese Academy of Sciences, Beijing 100049, China*

[4]*Yongjiang Laboratory, Ningbo 130021, China*

[†]E-mail: dongyuda21@mails.ucas.ac.cn; zetao-chen@ylab.ac.cn; xinhe@ucas.ac.cn

Received Jan. 7, 2024; Revision accepted Feb. 29, 2024; Crosschecked Apr. 24, 2024

https://doi.org/10.1631/FITEE.2400011

Event cameras, characterized by their low latency, large dynamic range, and extremely high temporal resolution, have recently received increasing attention. These features make them particularly well-suited for virtual/augmented reality (VR/AR) applications. To facilitate the development of three-dimensional (3D) perception and navigation algorithms in VR/AR applications using event cameras, we introduce the Stereo Event camera dataset for Virtual and Augmented Reality (SEVAR), which comprises a wide variety of head-mounted indoor sequences, including scenarios with rapid motion and a large dynamic range. We present the first comprehensive set of VR/AR datasets captured with an event-based stereo camera, a regular stereo camera at 30 Hz, and an inertial measurement unit at 1000 Hz. The camera placement, field of view (FoV), and resolution match those of the head-mounted device, such as Meta Quest Pro. All sensors are time-synchronized in the hardware. Ground truth poses captured by a motion capture system are also available for trajectory evaluation. The sequences include several common scenarios, and cover the specific challenges targeted by event cameras. The dataset can be found at https://github.com/sevar-dataset/sevar.

1 Introduction

VR and AR have experienced significant growth and evolution over the past few years in a wide variety of fields, such as entertainment, education, healthcare, and the military. It is imperative that VR/AR devices operate with high reliability in the environments corresponding to these specified fields. Simultaneous localization and mapping (SLAM) represents a pivotal technology that facilitates the concurrent construction of an environment map and the tracking of device location in the constructed map. Despite the substantial progress that SLAM technology has made in recent years, its performance in practical applications remains unsatisfactory. This is particularly evident in scenarios where conditions are less than optimal, such as in rooms that are insufficiently lit or have windows with intense light sources like the Sun. Furthermore,

‡ Corresponding authors

* Project supported by the Zhejiang Provincial Natural Science Foundation of China (No. 2023C03012), the Postdoctoral Preferential Funding Project of Zhejiang Province, China (No. ZJ2022116), and the Independent Project of Hangzhou Institute for Advanced Study, China (No. B02006C019014)

ORCID: Zetao CHEN, https://orcid.org/0000-0002-5596-5008; Xin HE, https://orcid.org/0000-0001-8229-8143

aggressive movements can induce motion blur in standard cameras, increasing the difficulty in achieving precise localization and mapping in such environments. Light detection and ranging (LiDAR) technology contributes significantly to mitigating these challenges, but has not been widely adopted in VR/AR devices due to its essential requirements for a compact layout, low power consumption, and low material costs. An innovative sensor, the dynamic vision sensor (DVS), commonly known as an event camera, was designed to overcome these challenges, making it a valuable addition to VR and AR systems.

An event camera is a bio-inspired visual sensor, characterized by the unique feature that each pixel operates independently and asynchronously. When the variation in a pixel's intensity surpasses a predefined threshold, the sensor promptly initiates an event that encodes details such as the corresponding pixel coordinates, the timestamp of the event, and its polarity. Event cameras have several significant advantages over traditional cameras, such as a high dynamic range (HDR, up to 140 dB), high temporal resolution (at the microsecond level), low power consumption (at the milliwatt level), and low motion blur. Therefore, event cameras are attracting growing attention and are playing a pivotal role in advancing the field of SLAM.

Though research on event cameras has gained momentum in the last few years, event-based algorithms are still in the early stages of development when compared to well-established frame-based algorithms. To promote the research and development of event-based SLAM, there is a substantial need for high-quality publicly available datasets and benchmarks. There are several popular event-based datasets, such as TUM-VIE (Klenk et al., 2021), VECtor (Gao et al., 2022), and DSEC (Gehrig et al., 2021). However, note that a majority of current SLAM datasets are tailored mainly for autonomous driving or robots, as stated in Liu et al. (2021). To advance the research on event cameras for VR/AR devices, we introduce the SEVAR dataset. It is more suitable for VR and AR scenarios compared to other datasets. To the best of our knowledge, SEVAR is the first dataset involving event cameras with such a large field-of-view (FoV, more than 150°). Furthermore, the cameras are strategically positioned and oriented to closely resemble the setup found in mainstream head-mounted devices such as Meta Quest Pro.

The acquisition platform also includes regular stereo cameras and inertial sensors, similar to those found in commercial VR/AR devices. Video graphics array (VGA) cameras, with a resolution of 640×480, are equipped with 165° wide-angle lenses. The inertial sensors provide three-axis accelerometer and gyroscope data at a frequency of 1000 Hz. We provide a range of challenging sequences, including those with a high dynamic range and motion blur, as well as additional sequences optimized for real-world application scenarios. To summarize, this paper makes the following major contributions:

1. We present the first event-based dataset explicitly designed for VR and AR scenarios in terms of camera placement, FoV angle, and resolution.

2. We provide a variety of motion modes, indoor settings, and lighting conditions, offering sequences that span from easy to extremely challenging scenarios. Furthermore, we have incorporated other challenging situations commonly encountered in practical applications of head-mounted devices, such as picking up items and rapid head movements.

The datasets are captured by a precisely hardware-synchronized sensor suite that includes a stereo event camera, a stereo regular camera, and an inertial measurement unit (IMU), simultaneously recording a six-degree-of-freedom (6-DoF) ground truth pose for each sequence.

2 Related works

With the rapid progress in SLAM research over the past decade, an increasing number of datasets and benchmarks have been introduced. Typically, when considering the input modality, SLAM can be categorized into two primary types: those that are vision-based and those that use laser-based methodologies. VR and AR rely mainly on vision-based SLAM. Notably, event cameras have gained significant popularity in vision-based SLAM in recent years. This section offers a concise overview of previous event camera datasets. Table 1 presents the key features of all datasets, such as sensor setup, camera wide-angle, motion mode, ground truth acquisition method, and sensor

synchronization.

Weikersdorfer et al. (2014) presented an early event-based dataset including an event camera with 128×128 resolution (events only) and an RGB-D sensor. They provided a small number of indoor sequences with ground truth information provided by a Motion Capture (MoCap) system.

Barranco et al. (2016) provided a dataset captured by a dynamic and active pixel vision sensor (DAVIS) and a Microsoft Kinect RGB-D sensor. The DAVIS240B sensor combines event data, active pixel sensor (APS) frames (240×180 resolution), and a built-in six-axis IMU. Notably, they provided only indoor sequences with five DoFs of motion. The ground truth poses in this dataset were obtained from wheel odometry. It can be subject to drift, which may affect the accuracy of the ground truth poses.

MVSEC was presented as the first dataset to provide synchronized stereo event cameras with accurate ground truth for both depth and pose, a feature of paramount importance in advancing SLAM research. It deploys sensors including a pair of DAVIS346 sensors (346×260 resolution, events and APS frames, six-axis built-in IMU), stereo cameras, a nine-axis built-in IMU from the VI-Sensor, and a Velodyne VLP-16 LiDAR. The dataset provides rich platform usage scenarios as these sensors were flown by a hexacopter, carried on a hand-held device, driven on top of a car, and mounted on a motorcycle. However, the sensors were only partially synchronized, and temporal alignment was performed in offline mode.

TUM-VIE is the first one-mega-pixel stereo event dataset, featuring a pair of Prophesee Gen4CD event cameras with 1280×720 resolution (events only), stereo global shutter grayscale frames, and a six-axis IMU. The dataset includes sequences that encompass environments of various scales and involve different motion conditions, such as walking, running, skating, and biking. Ground truth poses were acquired through a MoCap system.

VECtor introduces a fully hardware-synchronized dataset captured by a Prophesee Gen3CD event camera (640×480 resolution, events only), an RGB-D sensor, a regular stereo camera, a nine-axis IMU, and LiDAR. It covers a range of scenarios, including various motion dynamics, complex environments (e.g., rooms with ambiguous geometries, or unfurnished basements lacking texture), and diverse illumination conditions (e.g., HDR scenes, low or dynamically changing lighting situations). It provides accurate 6-DoF ground truth poses for all sequences, as well as precise intrinsic and extrinsic calibration.

There are also event-based datasets for autonomous driving. DSEC addresses the need for large-scale outdoor event-based datasets on autonomous driving. Compared to typical autonomous driving datasets (Geiger et al., 2013), DSEC stands out by incorporating a stereo event camera, which significantly enhances the dataset's application scenarios. There are also various challenging illumination conditions, such as tunnels and driving at night. Furthermore, the UZH-FPV drone racing dataset (Delmerico et al., 2019) was specifically designed for the localization of drones during high-speed motion in 6 DoFs. These aggressive flight sequences were captured by a drone equipped with a DAVIS346 event camera with a wide-angle lens and a stereo fisheye camera.

Among the relevant datasets, the one closest to our work is the TUM-VIE dataset. It provides several head-mounted sequences that are relevant for VR.

Table 1 Comparison of different event-based datasets

Dataset	Event stereo	Event FoV	Regular stereo	Regular FoV	IMU	Motion	Ground truth pose	Syn.
EDVS (Weikersdorfer et al., 2014)	×	NA	×	×	×	Hand-held	MoCap	×
EVneur (Barranco et al., 2016)	×	NA	×	×	6	Mobile robot	Odometer	×
MVSEC (Zhu et al., 2018)	√	106°D	√	98°D	6	Diverse	MoCap+Cartographer	√
TUM-VIE (Klenk et al., 2021)	√	111°D	√	126°D	6	Diverse	MoCap	√
VECtor (Gao et al., 2022)	√	106°D	√	102°D	9	Diverse	MoCap+ICP	√
DSEC (Gehrig et al., 2021)	√	NA	√	NA	6	Driving	RTK GPS	√
UZH-FPV (Delmerico et al., 2019)	×	120°D	×	90°D	6	Drone	Total station	√
Ours	√	150°D	√	165°D	6	Head-mounted	MoCap	√

FoV: field of view; IMU: inertial measurement unit; Syn.: synchronization. NA: not available

However, we have observed the smearing effect on top of the surface of active events from its high definition (HD) resolution (1280×720) cameras, and a similar phenomenon was reported by Alzugaray and Chli (2018) and Hu et al. (2021). SEVAR is the first dataset completely designed to closely resemble mainstream commercial VR/AR devices in terms of camera placement, FoV angle, and resolution. Thus, we believe that the presented dataset is valuable for integrating event cameras into head-mounted devices.

3 Dataset

3.1 Acquisition platform

We assembled a head-mounted device for data collection, as depicted in Fig. 1. In this configuration, all sensors, including the IMU, regular cameras, and event cameras, are rigidly mounted on a 3D-printed holder. The configuration and technical specifications of the regular cameras and IMU closely resemble those found in commercial VR devices, with the event cameras being positioned directly above this arrangement. To ensure stable data transmission, we employed 10 Gb fiber optics to connect regular cameras to the Ethernet port of the host computer and event cameras to the USB3.0 port of the laptop. All data were recorded using a high-speed NVMe solid state disk (SSD).

3.2 Sensor setup

We set up stereo DAVIS346 event cameras with a resolution of 346×260 and a baseline of about 13 cm. Since the MoCap system emits 850 nm infrared light to track the passive spherical markers that are attached to the acquisition platform, when the acquisition platform moves in the MoCap room, the DAVIS346 cameras may experience interference from 850 nm infrared light. We positioned an infrared filter (Calabrese et al., 2019) (PHTODE IR690, with a cutoff frequency of 400–690 nm) in front of the lenses to reduce most of the infrared noises in the event output.

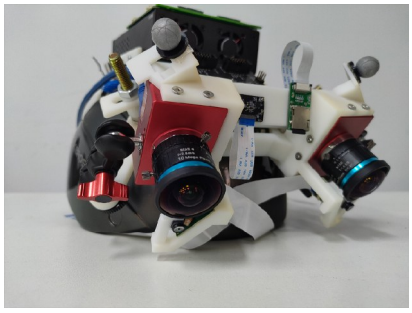

Fig. 1 Acquisition platform overview
IMU (middle), stereo regular cameras (bottom), and event cameras (top) are rigidly mounted on a 3D-printed holder

We selected OV7251 as our regular stereo camera. It is a compact, low-power sensor with a global shutter and excellent low-light sensitivity, making it popular for space-constrained applications like head-mounted devices, smartphones, tablets, and laptops. The regular stereo cameras are positioned under the event cameras with a vertical baseline of about 3 cm. The specific characteristics of each sensor are shown in Table 2.

Table 2 Hardware specification of the acquisition platform

Sensor	Rate	Specification
2× DAVIS346, baseline: about 13 cm	N/A	346×260; *f*/2.8 mm; FoV: 150°D; up to 120 dB
2× OV7251, baseline: about 13 cm	30 Hz	640×480; *f*/0.825 mm; FoV: 165°D; global shutter
IMU ICM42688P	1000 Hz	3-axis accelerometer; 3-axis gyroscope
Vicon MoCap	300 Hz	6-DoF ground truth; 850 nm infrared light

3.3 Synchronization and calibration

An overview of time synchronization and calibration is shown in Fig. 2. The time synchronization of all sensors is achieved through the use of hardware triggers provided by a micro-controller unit (MCU). The MCU outputs 1000 Hz pulse signals to the IMU after it receives the start signal sent by the user. Meanwhile, it sends a separate 30 Hz pulse signal to the stereo regular camera to control the start of exposure. The same signal is forwarded to the left event camera. Whenever the event camera receives a pulse signal, it is recorded as “external input triggering” and is precisely timestamped by the event camera’s clock. Internal synchronization between event cameras is achieved through internal firmware and a daisy chain connection.

We used a pattern of AprilTags (Olson, 2011)

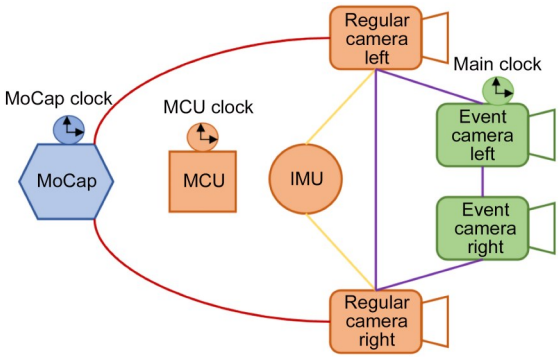

Fig. 2 Overview of time synchronization and calibration
The same color represents using the same clock domain before time synchronization. Red, yellow, and orange connecting edges represent camera–MoCap hand–eye calibration, camera–IMU extrinsic calibration, and camera–camera extrinsic calibration, respectively. References to color refer to the online version of this figure

and the Kalibr toolbox (Furgale et al., 2013) to estimate the intrinsic and extrinsic calibration. As DAVIS event cameras can produce frame images homologous to the imaging plane of the event output, we directly employed these frame images for calibration. The acquisition platform is directed towards a static AprilTag grid board with known size and performs complete 6-DoF motion under good illumination conditions to fully excite the six axes of IMU.

Before calibrating the extrinsic calibration between the camera and MoCap, time synchronization must be performed, as they operate with different clocks. Given that the camera and MoCap tracking marker are rigidly connected, synchronization can be achieved by leveraging the constraint of angular velocities derived from the screw congruence theorem (Furrer et al., 2018). With the help of AprilTags, we can obtain a reliable estimation of camera poses. Angular velocities are calculated from the camera poses and aligned with the MoCap measurements. The time offset is determined by computing the temporal correlation function of the angular velocities and mapping the peak index.

Once the camera and MoCap poses are synchronized, we can proceed with extrinsic calibration to conduct the homogeneous transformation. Solving the hand–eye calibration problem ultimately comes down to solving

$$AX = XB, \tag{1}$$

where A, B, and X represent rigid body motions. The transformations relevant to hand–eye calibration are depicted in Fig. 3. Using the relative transformation between two motions as

$$T_{\mathrm{BH}_1} T_{\mathrm{HE}} T_{\mathrm{WE}_1}^{-1} = T_{\mathrm{BH}_2} T_{\mathrm{HE}} T_{\mathrm{WE}_2}^{-1}, \tag{2}$$

where B represents the base, H the hand, W the world, and E the eye, at least two pairs of poses are required to solve Eq. (2). Subscripts 1 and 2 represent two poses at different positions. Based on the relationship between the two poses and the hand–eye calibration, it can be concluded that

$$T_{\mathrm{H}_1\mathrm{H}_2} = T_{\mathrm{BH}_1}^{-1} T_{\mathrm{BH}_2}, \tag{3}$$

$$T_{\mathrm{E}_1\mathrm{E}_2} = T_{\mathrm{WE}_1}^{-1} T_{\mathrm{WE}_2}. \tag{4}$$

Eq. (2) can be transformed into the form of $AX = XB$:

$$T_{\mathrm{H}_1\mathrm{H}_2} T_{\mathrm{HE}} = T_{\mathrm{HE}} T_{\mathrm{E}_1\mathrm{E}_2}. \tag{5}$$

In this study, we employ a dual quaternion based hand–eye calibration method (Eq. (6)), which provides a global linear solution. With the global initial guess, we use a continuous-time batch estimator to perform a refinement, and obtain an accurate and robust calibration result.

$$q_{\mathrm{H}_1\mathrm{H}_2} = q_{\mathrm{HE}} q_{\mathrm{E}_1\mathrm{E}_2} q_{\mathrm{HE}}^{-1}. \tag{6}$$

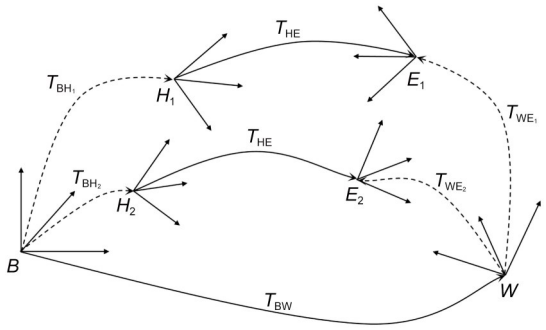

Fig. 3 Coordinate transformation relationship for hand–eye calibration, where subscripts 1 and 2 represent two poses, solid lines indicate static transformations, and dashed lines indicate transformations that change over time

3.4 Sequence overview

Data sequences were recorded in a 10 m×8 m×4 m MoCap room, and each sequence contained ground

truth poses of the complete trajectory. The MoCap system was calibrated immediately before the recording. Each sequence was individually recorded in an robot operating system (ROS) bag file, with the corresponding ground truth pose provided in a TXT file. An overview of the data sequences is presented in Table 3 and Fig. 4. The name of each sequence represents the corresponding scene, and the suffix indicates motion speed or light intensity. The sequence with the suffix "slow" contains a simple 3D translational motion, the suffix "normal" contains a full six-DoF motion, and the suffix "fast" contains a faster six-DoF motion. All sequences with the suffix "hdr" were recorded in low-light conditions. MER represents the average event rate of the left event camera for each sequence. Each sequence contains multiple closed loops.

The first four sequences prefixed with "board" in Table 3 are handheld sequences. The board sequences show a board filled with common items in daily life, such as pictures, paint brush, fruit, flowers, and wine bottles. The desk sequences display a work scenario consisting of a set of tables and chairs with computer screen, books, water cup, mobile phone, pen, desk lamp, fan, keyboard, and mouse. To distinguish it from the curtains in the background, we placed a whiteboard behind the table. The sofa sequences showed a common living room scenario consisting of a large sofa and tea table with laptops, potting, backpack, penguin doll, and teacup. The walk sequences were captured in the entire MoCap room with the aim of testing various typical behaviors such as shaking and stillness during walking.

Table 3 Overview of the data sequences

Sequence	Duration (s)	Length (m)	MER ($\times 10^6$ events/s)
Board-slow	21	13.7	2.01
Board-normal	25	17.5	2.73
Board-fast	23	18.9	5.13
Board-hdr	25	17.6	3.46
Desk-normal	57	21.6	2.11
Desk-fast	53	24.3	5.03
Desk-hdr	50	23.5	3.28
Sofa-normal	30	15.7	1.98
Sofa-fast	35	19.5	4.87
Sofa-hdr	33	18.6	2.77
Walk-slow	51	26.5	1.76
Walk-normal	58	30.3	2.24
VR-normal	80	28.6	3.05
VR-hdr	86	29.8	3.66
AR-normal	92	43.8	3.48
AR-hdr	98	46.9	4.12

Compared to other datasets, the VR and AR sequences are the two most unique sets of sequences in our dataset. The VR sequences simulate shooting games, while the AR sequences imitate sword games, incorporating a significant number of hand and head movements. We incorporated a diverse range of hand and head movements in the dataset, including challenging actions such as jumping, high-speed motion, picking up objects, and quickly swinging the head. Lack of texture and motion blur can make camera tracking difficult. We believe that better event-based algorithms will solve these problems. To enhance realism, we refrain from artificially adding textures (such as calibration patterns), which can pose greater challenges for algorithm evaluation.

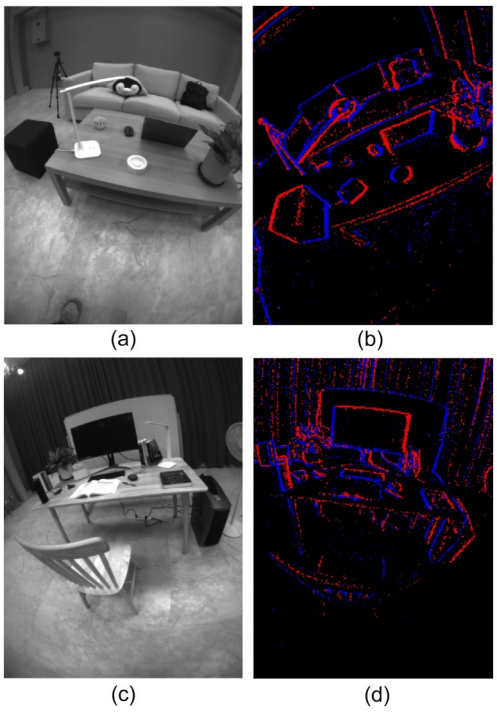

Fig. 4 Examples of sofa-fast and desk-normal sequences: (a) sofa-frame data; (b) sofa-event data; (c) desk-frame data; (d) desk-event data

4 Evaluation

To verify the quality of our data and calibration, we tested the sequences using state-of-the-art open-source algorithms, including ORB-SLAM3 (Campos

et al., 2021) with a regular stereo setting (VO), ORB-SLAM3 with a stereo-inertial setting (VIO), and Open-VINS (Geneva et al., 2020) with a VIO. The results are shown in Table 4 and Fig. 5. We used the absolute pose error (APE) to measure the global consistency of trajectories (Sturm et al., 2012). APE is based on the absolute relative pose between two poses $P_{\mathrm{ref},i}$ and $P_{\mathrm{est},i}$ at timestamp i:

$$E_i = P_{\mathrm{ref},i}^{-1} P_{\mathrm{est},i}, \tag{7}$$

where $P_{\mathrm{ref},i}$ is the ground truth pose and $P_{\mathrm{est},i}$ is the estimated pose. After obtaining the transformation E_i, each part can be extracted for analysis according to the requirements. In this study, we use the full transformation part of E_i to calculate APE:

$$\mathrm{APE}_i = \left\| E_i - I_{4\times 4} \right\|_{\mathrm{F}}. \tag{8}$$

Then, statistics can be calculated on the root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathrm{APE}_i^2}, \tag{9}$$

where N represents the number of frames in the trajectory.

Table 4 The absolute pose errors (APEs) of state-of-the-art open-source algorithms

Sequence	APE (m)		
	ORB-SLAM3-VO	ORB-SLAM3-VIO	Open-VINS
Board-slow	0.0173	0.0047	0.0133
Board-normal	0.0430	0.0139	0.0583
Board-fast	0.0862	0.0352	0.1401
Board-hdr	–	–	–
Desk-normal	0.0216	0.0084	0.0987
Desk-fast	0.0965	0.0464	0.2678
Desk-hdr	–	–	–
Sofa-normal	0.0408	0.0217	0.0663
Sofa-fast	0.1065	0.0464	0.1571
Sofa-hdr	–	–	–
Walk-slow	0.0915	0.0542	0.0676
Walk-normal	0.0804	0.0369	0.0785
VR-normal	0.1195	0.0368	0.2945
VR-hdr	–	–	–
AR-normal	0.1411	0.0488	0.1974
AR-hdr	–	–	–

"–" indicates that the algorithm failed

According to our evaluation, ORB-SLAM3 and Open-VINS performed well in the simple sequences, such as board-slow, board-normal, desk-normal,

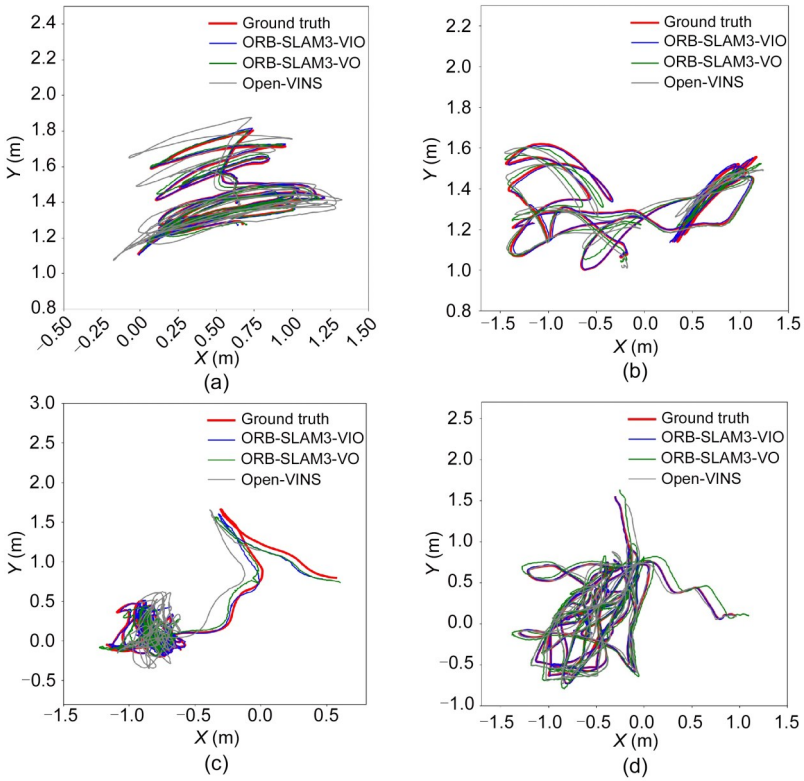

Fig. 5 Visual trajectory maps for different sequences: (a) desk-normal; (b) sofa-fast; (c) VR-normal; (d) AR-normal
References to color refer to the online version of this figure

sofa-normal, and walk-slow. The above results were in line with expectations, verifying the calibration quality of our data. However, as the sequences became more challenging, such as fast motion and low light scenes, ORB-SLAM3 and Open-VINS had a large drift or even failed. These results showed the deficiencies of the frame-based visual system in challenging scenarios. They also suggested that our datasets can be used for further research in the event-based visual system.

5 Conclusions

In this paper, we present a precisely synchronized event-based dataset, designed especially for multi-sensor fusion in SLAM applications, with a particular emphasis on VR and AR scenarios. Alongside setting up commonly used stereo regular cameras and an IMU, we have integrated stereo event cameras. We specialize in recording sequences to imitate real-life scenarios, while adding challenging sequences such as low light and fast motion. Consequently, it is our aspiration that this dataset will serve as a valuable resource for the advancement of research in the domain of event-based multi-sensor fusion algorithms.

Contributors

Yuda DONG designed the research. Yuda DONG, Junchi FENG, and Yinong CAO processed the data. Zichao SHU contributed to hand–eye calibration. Yuda DONG and Zetao CHEN drafted the paper. Xin HE, Jianyu WANG, and Lijun LI helped organize the paper. Yuda DONG, Chunlai LI, and Shijie LIU revised and finalized the paper. Xin HE provided research funding.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The dataset can be found at https://github.com/sevar-dataset/sevar.

References

Alzugaray I, Chli M, 2018. Asynchronous corner detection and tracking for event cameras in real time. *IEEE Rob Autom Lett*, 3(4):3177-3184. https://doi.org/10.1109/LRA.2018.2849882

Barranco F, Fermuller C, Aloimonos Y, et al., 2016. A dataset for visual navigation with neuromorphic methods. *Front Neurosci*, 10:49. https://doi.org/10.3389/fnins.2016.00049

Calabrese E, Taverni G, Easthope CA, et al., 2019. DHP19: dynamic vision sensor 3D human pose dataset. IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.1695-1704. https://doi.org/10.1109/CVPRW.2019.00217

Campos C, Elvira R, Rodríguez JJG, et al., 2021. ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Trans Rob*, 37(6):1874-1890. https://doi.org/10.1109/TRO.2021.3075644

Delmerico J, Cieslewski T, Rebecq H, et al., 2019. Are we ready for autonomous drone racing? The UZH-FPV drone racing dataset. Int Conf on Robotics and Automation, p.6713-6719. https://doi.org/10.1109/ICRA.2019.8793887

Furgale P, Rehder J, Siegwart R, 2013. Unified temporal and spatial calibration for multi-sensor systems. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.1280-1286. https://doi.org/10.1109/IROS.2013.6696514

Furrer F, Fehr M, Novkovic T, et al., 2018. Evaluation of combined time-offset estimation and hand-eye calibration on robotic datasets. In: Hutter M, Siegwart R (Eds.), Field and Service Robotics. Springer Proceedings in Advanced Robotics, Vol. 5. Springer, Cham, p.145-159. https://doi.org/10.1007/978-3-319-67361-5_10

Gao L, Liang YX, Yang JQ, et al., 2022. VECtor: a versatile event-centric benchmark for multi-sensor SLAM. *IEEE Rob Autom Lett*, 7(3):8217-8224. https://doi.org/10.1109/LRA.2022.3186770

Gehrig M, Aarents W, Gehrig D, et al., 2021. DSEC: a stereo event camera dataset for driving scenarios. *IEEE Rob Autom Lett*, 6(3):4947-4954. https://doi.org/10.1109/LRA.2021.3068942

Geiger A, Lenz P, Stiller C, et al., 2013. Vision meets robotics: the KITTI dataset. *Int J Rob Res*, 32(11):1231-1237. https://doi.org/10.1177/0278364913491297

Geneva P, Eckenhoff K, Lee W, et al., 2020. OpenVINS: a research platform for visual-inertial estimation. IEEE Int Conf on Robotics and Automation, p.4666-4672. https://doi.org/10.1109/ICRA40945.2020.9196524

Hu YH, Liu SC, Delbruck T, 2021. v2e: from video frames to realistic DVS events. IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.1312-1321. https://doi.org/10.1109/CVPRW53098.2021.00144

Klenk S, Chui J, Demmel N, et al., 2021. TUM-VIE: the TUM stereo visual-inertial event dataset. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.8601-8608. https://doi.org/10.1109/IROS51168.2021.9636728

Liu YZ, Fu YJ, Chen FD, et al., 2021. Simultaneous localization and mapping related datasets: a comprehensive survey. https://arxiv.org/abs/2102.04036

Olson E, 2011. AprilTag: a robust and flexible visual fiducial system. IEEE Int Conf on Robotics and Automation, p.3400-3407. https://doi.org/10.1109/ICRA.2011.5979561

Sturm J, Engelhard N, Endres F, et al., 2012. A benchmark for the evaluation of RGB-D SLAM systems. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.573-580. https://doi.org/10.1109/IROS.2012.6385773

Weikersdorfer D, Adrian DB, Cremers D, et al., 2014. Event-based 3D SLAM with a depth-augmented dynamic vision sensor. IEEE Int Conf on Robotics and Automation, p.359-364. https://doi.org/10.1109/ICRA.2014.6906882

Zhu AZ, Thakur D, Ozaslan T, et al., 2018. The multivehicle stereo event camera dataset: an event camera dataset for 3D perception. *IEEE Rob Autom Lett*, 3(3):2032-2039. https://doi.org/10.1109/LRA.2018.2800793