The Multivehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception

Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis

***Abstract*—Event-based cameras are a new passive sensing modality with a number of benefits over traditional cameras, including extremely low latency, asynchronous data acquisition, high dynamic range, and very low power consumption. There has been a lot of recent interest and development in applying algorithms to use the events to perform a variety of three-dimensional perception tasks, such as feature tracking, visual odometry, and stereo depth estimation. However, there currently lacks the wealth of labeled data that exists for traditional cameras to be used for both testing and development. In this letter, we present a large dataset with a synchronized stereo pair event based camera system, carried on a handheld rig, flown by a hexacopter, driven on top of a car, and mounted on a motorcycle, in a variety of different illumination levels and environments. From each camera, we provide the event stream, grayscale images, and inertial measurement unit (IMU) readings. In addition, we utilize a combination of IMU, a rigidly mounted lidar system, indoor and outdoor motion capture, and GPS to provide accurate pose and depth images for each camera at up to 100 Hz. For comparison, we also provide synchronized grayscale images and IMU readings from a frame-based stereo camera system.**

***Index Terms*—SLAM, visual-based navigation, event-based cameras.**

I. Introduction

EVENT based cameras sense the world by detecting changes in the log intensity of an image. By registering these changes with accuracy on the order of tens of microseconds and asynchronous, almost instant, feedback, they allow for extremely low latency responses compared to traditional cameras which typically have latencies on the order of tens of milliseconds. In addition, by tracking changes in log intensity, the cameras have very high dynamic range (>130 dB vs about 60 dB with traditional cameras), which make them very useful for scenes with dramatic changes in lighting, such as indoor-outdoor transitions, as well as scenes with a strong light source, such as the sun.

However, most modern robotics algorithms have been designed for synchronous sensors, where measurements arrive at fixed time intervals. In addition, the generated events do not carry any intensity information on their own. As a result, new algorithms must be developed to fully take advantage of the benefits provided by this sensor. Unfortunately, due to the differences in measurements, we cannot directly take advantage of the enormous amounts of labeled data captured with traditional cameras. Such data has shown to be extremely important for providing realistic and consistent evaluations of new methods, training machine learning systems, and providing opportunities for new development for researchers who do not have access to these sensors.

In this work, we aim to provide a number of different sequences that will facilitate the research and development of novel solutions to a number of different problems. One main contribution is the first dataset with a synchronized stereo event camera system. A calibrated stereo system is useful for depth estimation with metric scale, which can contribute to problems such as pose estimation, mapping, obstacle avoidance and 3D reconstruction. There have been a few works in stereo depth estimation with event based cameras, but, due to the lack of accurate ground truth depth, the evaluations have been limited to small, disparate sequences, consisting of a few objects in front of the camera. In comparison, this dataset provides event streams from two synchronized and calibrated Dynamic Vision and Active Pixel Sensors (DAVIS-m346b), with long indoor and outdoor sequences in a variety of illuminations and speeds, along with accurate depth images and pose at up to 100 Hz, generated from a lidar system rigidly mounted on top of the cameras, as in Fig. 1, along with motion capture and GPS. We hope that this dataset can help provide a common basis for event based algorithm evaluation in a number of applications.

The full dataset can be found online at https://daniilidis-group.github.io/mvsec.

The main contributions from this letter can be summarized as:

- The first dataset with synchronized stereo event cameras, with accurate ground truth depth and pose.
- Event data from a handheld rig, a flying hexacopter, a car, and a motorcycle, in conjunction with calibrated sensor data from a 3D lidar, IMUs and frame based images, from

Manuscript received September 10, 2017; accepted January 12, 2018. Date of publication February 9, 2018; date of current version March 23, 2018. This letter was recommended for publication by Associate Editor J. Nieto and Editor C. Stachniss upon evaluation of the reviewers' comments. This work was supported in part by NSF-DGE-0966142 (IGERT), in part by NSF-IIP-1439681 (I/UCRC), in part by NSF-IIS-1426840, in part by ARL MAST-CTA W911NF-08-2-0004, in part by ARL RCTA W911NF-10-2-0016, in part by ONR N00014-17-1-2093, in part by ONR STTR (Robotics Research), in part by NSERC Discovery, and in part by the DARPA FLA Program. *(Corresponding author: Alex Zihao Zhu.)*

The authors are with the GRASP Laboratory, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: alexzhu@seas.upenn.edu; tdinesh@seas.upenn.edu; ozaslan@seas.upenn.edu; pfrommer@seas.upenn.edu; kumar@seas.upenn.edu; kostas@seas.upenn.edu).

Digital Object Identifier 10.1109/LRA.2018.2800793

Fig. 1. Full sensor rig, with stereo DAVIS cameras, VI Sensor and Velodyne lidar.

a variety of different speeds, illumination levels and environments.

II. Related Work

A. Related Datasets

At present, there are a number of existing datasets that provide events from monocular event based cameras in conjunction with a variety of other sensing modalities and ground truth measurements that are suitable for testing a number of different 3D perception tasks.

Weikersdorfer *et al.* [1] combine the earlier eDVS sensor with 128×128 resolution, with a Primesense RGBD sensor, and provide a dataset of indoor sequences with ground truth pose from a motion capture system, and depth from the RGBD sensor.

Rueckauer *et al.* [2] provide data from a DAVIS 240C camera undergoing pure rotational motion, as well as ground truth optical flow based on the angular velocities reported from the gyroscope, although this is subject to noise in the reported velocities.

Barranco *et al.* [3] present a dataset with a DAVIS 240B camera mounted on top of a pan tilt unit, attached to a mobile base, along with a Microsoft Kinect sensor. The dataset provides sequences of the base moving with 5dof in a indoor environment, along with ground truth depth, and optical flow and pose from the wheel encoders on the base and the angles from the pan tilt unit. While the depth from the Kinect is accurate, the optical flow and pose are subject to drift from the position estimates of the base's wheel encoders.

Mueggler *et al.* [4] provide a number of handheld sequences intended for pose estimation in a variety of indoor and outdoor environments, generated from a DAVIS 240C. A number of the indoor scenes have provided pose ground truth, captured from a motion capture system. However, there are no outdoor sequences, or other sequences with a significant displacement, with ground truth information.

Binas *et al.* [5] provide a large dataset of a DAVIS 346B mounted behind the windshield of a car, with 12 hours of driving, intended for end to end learning of various driving related tasks. The authors provide a number of auxiliary measurements from the vehicle, such as steering angle, accelerator pedal position, vehicle speed etc., as well as longitude and latitude from a GPS unit. However, no 6dof pose is provided, as only 2D translation can be inferred from the GPS output as provided.

These datasets provide valuable data for development and evaluation of event based methods. However, they have, to date, only monocular sequences, with ground truth 6dof pose limited to small indoor environments, with few sequences with ground truth depth. In contrast, this work provides stereo sequences with ground truth pose and depth images in a variety of indoor and outdoor settings.

B. Event Based 3D Perception

Early works in [6], [7] present stereo depth estimation results with a number of spatial and temporal costs. Later works in [8], [9] and [10] have adapted cooperative methods for stereo depth to event based cameras, due to their applicability to asynchronous, point based measurements. Similarly, [11] and [12] apply a set of temporal, epipolar, ordering and polarity constraints to determine matches, while [13] compare this with matching based on the output of a bank of orientation filters. The authors in [14] show a new method to determine the epipolar line, applied to stereo matching. In [15], the authors propose a novel context descriptor to perform matching, and the authors in [16] use a stereo event camera undergoing pure rotation to perform depth estimation and panoramic stitching.

There are also a number of works on event based visual odometry and SLAM problems. The authors in [17] and [18] proposed novel methods to perform feature tracking in the event space, which they extended in [19] and [20] to perform visual and visual inertial odometry, respectively. In [1], the authors combine an event based camera with a depth sensor, to perform visual odometry and SLAM. The authors in [21] use events to estimate angular velocity of a camera, while [22] and [23] perform visual odometry by building an up to a scale map. In addition, [24] and [25] also fuse events with measurements from an IMU to perform visual inertial odometry.

While the more recent works evaluate based on public datasets such as [4], the majority are evaluated on small datasets generated solely for the paper, making comparisons of performance difficult. This is particularly the case for stereo event based cameras. In this work, we try to generate more extensive ground truth, for more meaningful evaluations of new algorithms that can provide a basis for comparisons between methods.

III. Dataset

For each sequence in this dataset, we provide the following measurements in ROS bag[1] format:

- Events, APS grayscale images and IMU measurements from the left and right DAVIS cameras.
- Images and IMU measurements from the VI Sensor.

[1]http://wiki.ros.org/Bags

TABLE I
SENSORS AND CHARACTERISTICS

Sensor	Characteristics
DAVIS m346B	346 × 260 pixel APS+DVS FOV: 67° vert., 83° horiz. IMU: MPU 6150
VI-Sensor	Skybotix integrated VI-sensor stereocamera: 2 Aptina MT9V034 gray 2 × 752 × 480 @ 20 fps, global shutter FOV: 57° vert., 2 × 80° horiz. IMU: ADIS16488 @200 Hz
Velodyne Puck LITE	VLP-16 PUCK LITE 360° Horizontal FOV, 30° Vertical FOV 16 channel 20 Hz 100 m Range
GPS	UBLOX NEO-M8N 72-channel u-blox M8 engine Position accuracy 2.0 m CEP

- Pointclouds from the Velodyne VLP-16 lidar.[2]
- Ground truth reference poses for the left DAVIS camera.
- Ground truth reference depth images for both left and right DAVIS cameras.

A. Sensors

A list of sensors and their characteristics can be found in Table I. In addition, Fig. 2(a) shows the CAD drawing of the sensor rig, with all sensor axes labeled, and Fig. 2 shows how the sensors are mounted on each vehicle. The extrinsics between all sensors are estimated through calibration, as explained in Section V.

For event generation, two experimental mDAVIS-346B cameras are mounted in a horizontal stereo setup. The cameras are similar to [26], but have a higher, 346 × 260 pixel, resolution, up to 50 fps APS (frame based images) output, and higher dynamic range. The baseline of the stereo rig is 10 cm, and the cameras are timestamp synchronized by using the trigger signal generated from the left camera (master) to deliver sync pulses to the right (slave) through an external wire. Both cameras have 4 mm lenses with approximately 87 degrees horizontal field of view, with an additional IR cut filter placed on each one to suppress the IR flashes from the motion capture systems. The APS exposures are manually set (no auto exposure) depending on lighting conditions, but are always the same between the cameras. While the timestamps of the grayscale DAVIS images are synced, there is unfortunately no way to synchronize the image acquisition itself. Therefore, there may be up to 10 ms of offset between the images.

To provide ground truth reference poses and depths (see Section IV), we have rigidly mounted a Velodyne Puck LITE above the stereo DAVIS cameras. The Velodyne lidar system provides highly accurate depth of a large number of points around the sensor. The lidar is mounted such that there is full overlap between the smaller vertical field of view of the lidar and that of the stereo DAVIS rig.

[2]http://velodynelidar.com/vlp-16-lite.html

In the outdoor scenes, we have also mounted a GPS device for a second ground truth reference for latitude and longitude. Typically, the GPS is placed away from the sensor rig to avoid interference from the USB 3.0 data cables.

In addition, we have mounted a VI Sensor [27], originally developed by Skybotix for comparison with frame based methods. The sensor has a stereo pair with IMU, all synchronized. Unfortunately, the only mounting option was to mount the cameras upside down, but we provide the transform between them and the DAVIS cameras.

B. Sequences

The full list of sequences with summary statistics can be found in Table II, and sample APS images with overlaid events can be found in Fig. 3.

1) Hexacopter With Motion Capture: The sensor setup was mounted below the compute stack of the hexacopter, with a 25 degree downwards pitch, as in Fig. 2(b). Two motion capture systems are used to generate sequences for this dataset, one indoors and one outdoors (see Fig. 4). The 26.8 m × 6.7 m × 4.6 m indoor area is instrumented with 20 Vicon Vantage VP-16 cameras. The outdoor netted area of 30.5 m × 15.3 m × 15.3 m is instrumented with an all-weather motion capture system comprised of 34 high resolution Qualisys Oqus 700 cameras. Both systems provide millimeter accuracy pose at 100 Hz by emitting infrared strobes and tracking IR reflecting markers placed on the hexacopter. We provide sequences in each area, with flights of different length and speed.

2) Handheld: In order to test performance in high dynamic range scenarios, the full sensor rig is carried in a loop through both outdoor and indoor environments, as well as indoor environments with and without external lighting. Ground truth pose and depth is provided by lidar SLAM.

3) Outdoor Driving: For slow to medium speed sequences, the sensor rig is mounted on the sun roof of a sedan as in Fig. 2(c), and driven around several West Philadelphia neighborhoods at speeds up to 12 m/s. Sequences are provided in both day and evening situations, including sequences with the sun directly in the cameras' field of view. Ground truth is provided as depth images from a lidar map, as well as pose from loop closed lidar odometry and GPS.

For high speed sequences, the DAVIS stereo rig and VI Sensor are mounted on the handlebar of a motorcycle [see Fig. 2(d)], along with the GPS device. The sequences involve driving at up to 38 m/s. Longitude and latitude, as well as relative velocity, are provided from the GPS.

IV. GROUND TRUTH GENERATION

To provide ground truth poses, motion capture poses are used when available. Otherwise, if lidar is available, Cartographer [28] is used for the driving sequences to fuse the lidar sweeps and IMU data into a loop-closed 2D pose of the lidar, which is transformed into the left DAVIS frame using the calibration in Section V-D. For outdoor scenes, we also provide raw GPS readings.

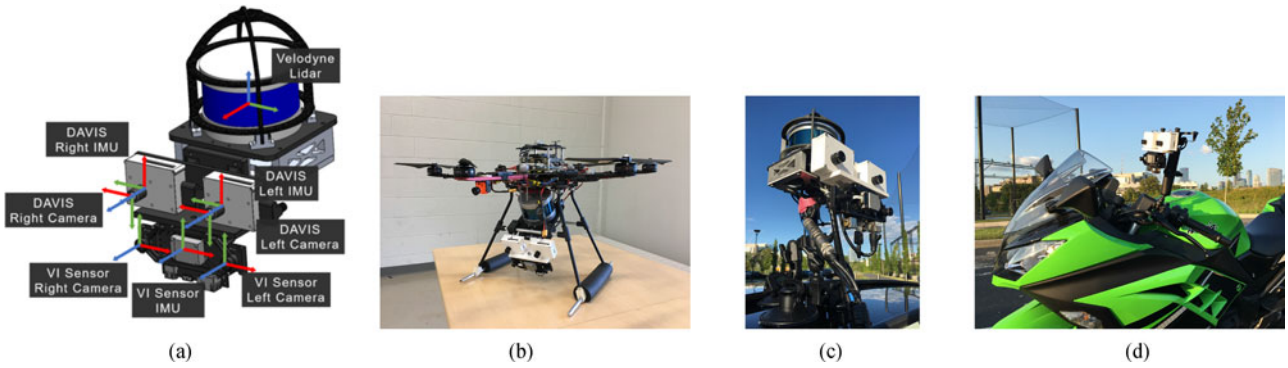

Fig. 2. Left to right: (a): CAD model of the sensor rig. All sensor axes are labeled and colored R:X, G:Y, B:Z, with only combinations of approximately 90 degree rotations between each pair of axes. (b): Sensor package mounted on hexacopter. (c): Sensor package mounted using a glass suction tripod mount on the sunroof of a car. (d): DAVIS cameras and VI Sensor mounted on motorcycle. Note that the VI-Sensor is mounted upside down in all configurations. Best viewed in color.

TABLE II
SEQUENCES FOR EACH VEHICLE

Vehicle	Sequence	T(s)	D(m)	$\|v\|_{\max}$(m/s)	$\|\omega\|_{\max}$(°/s)	MER(events/s)	Pose GT	Depth Available
Hexacopter	Indoor 1*	70	26.7	1.4	28.3	185488	Vicon	Yes
	Indoor 2*	84	36.8	1.5	29.8	273567	Vicon	Yes
	Indoor 3*	94	52.3	1.7	31.0	243953	Vicon	Yes
	Indoor 4*	20	9.8	2.0	64.3	361579	Vicon	Yes
	Outdoor 1	54	33.2	1.5	129.8	261589	Qualisys	Yes
	Outdoor 2	41	29.9	1.5	109.4	256539	Qualisys	Yes
Handheld	Indoor-outdoor	144	80.4	1.6	93.6	468675	LOAM	Yes
	Indoor corridor	249	105.2	2.7	37.4	590620	LOAM	Yes
Car	Day 1†	262	1207	7.6	30.6	386178	Cart., GPS	Yes
	Day 2†	653	3467	12.0	35.5	649081	Cart., GPS	Yes
	Evening 1	262	1217	10.4	20.6	334614	Cart., GPS	Yes
	Evening 2	374	2109	11.2	33.6	404105	Cart., GPS	Yes
	Evening 3	276	1613	10.0	25.1	371498	Cart., GPS	Yes
Motorcycle	Highway 1	1500	18293	38.4	203.4	511024	GPS	No

T: Total time, D: Total distance traveled, $\|v\|_{\max}$: Maximum linear velocity, $\|\omega\|_{\max}$: Maximum angular velocity, MER: Mean event rate. * No VI-Sensor data is available for these sequences. †A hardware failure caused the right DAVIS grayscale images to fail for these sequences.

For each sequence with lidar measurements, we run the Lidar Odometry and Mapping (LOAM) algorithm [29] to generate dense 3D local maps, which are projected into each DAVIS camera to generate dense depth images at 20 Hz, and to provide 3D pose for the handheld sequences.

Two separate lidar odometry algorithms are used as we noted that LOAM produces better, more well aligned, local maps, while Cartographer's loop closure results in more accurate global poses with less drift for longer trajectories. While Cartographer only estimates a 2D pose, we believe that this is a valid assumption as the roads driven have, for the most part, a single consistent grade.

A. *Ground Truth Pose*

For the sequences in the indoor and outdoor motion capture arenas, the pose of the body frame of the sensor rig ${}^{\text{world}}\mathbf{H}_{\text{body}(t)}$ at each time t is measured at 100 Hz with millimeter level accuracy.

For outdoor sequences we rely on Cartographer to perform loop closure and fuse lidar sweeps and IMU data into a single loop-closed 2D pose of the body (lidar in this case) with minimal drift.

In order to provide a quantitative measure of the quality of the final pose, we align the positions with the GPS measurements, and provide both overlaid on top of satellite imagery for each outdoor sequence in the dataset, as well as the difference in position between the provided ground truth and GPS. Fig. 7 provides a sample overlay for Car Day 2, where the average error between Cartographer and the GPS is consistently around 5 m without drift. This error is consistent amongst all of the outdoor driving sequences, where the overall average error is 4.7 m, and is on a similar magnitude to the error expected from GPS. Note that the spike in error around 440 seconds is due to significant GPS error, and corresponds to the section in bold on the top right of the overlay.

In both cases, the extrinsic transform, represented as a 4×4 homogenous transform matrix ${}^{\text{body}}\mathbf{H}_{\text{DAVIS}}$, for each sequence that takes a point from the left DAVIS frame to the body frame is then used to estimate the pose of the left DAVIS at time t with respect to the first left DAVIS pose at time t_0:

$$
\begin{aligned}
&{}^{\text{DAVIS}(t_0)}\mathbf{H}_{\text{DAVIS}(t)} \\
&\quad = {}^{\text{body}}\mathbf{H}^{-1}_{\text{DAVIS}}\, {}^{\text{world}}\mathbf{H}^{-1}_{\text{body}(t_0)}\, {}^{\text{world}}\mathbf{H}_{\text{body}(t)}\, {}^{\text{body}}\mathbf{H}_{\text{DAVIS}}. \quad (1)
\end{aligned}
$$

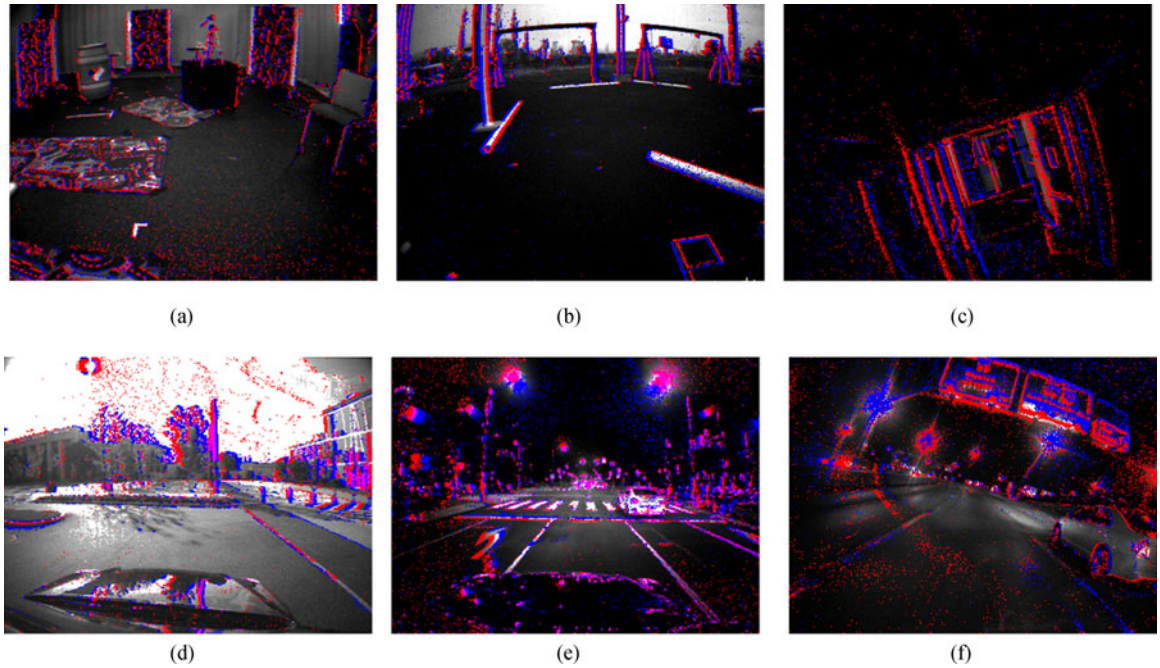

Fig. 3. Sample images with overlaid events (blue and red) from indoor and outdoor sequences, during day and evening. Best viewed in color. (a) Hexacopter indoor flight with Vicon motion capture. (b) Hexacopter outdoor flight with qualisys motion capture. (c) Handheld with difficult lighting conditions. (d) Car day 1. (e) Outdoor car evening. (f) Motorcycle highway 1.

Fig. 4. Motion capture arenas. Left: Indoor Vicon arena, right: Outdoor Qualisys arena.

B. Depth Map Generation

In each sequence where lidar is available, depth images for each DAVIS camera are generated for every lidar measurement. We first generate a local map by transforming each lidar pointcloud in a local window around the current measurement into the frame of the current measurement using the poses from LOAM. At each measurement, the window size is determined such that the distances between the current, and the first and last LOAM poses in the window are at least d meters, and that there are at least s seconds between the current, and first and last LOAM poses, where d and s are parameters tuned for each sequence. Examples of these maps can be found in Fig. 5.

We then project each point, $\mathbf{p}$, in the resulting pointcloud into the image in each DAVIS camera, using the standard pinhole projection equation:

$$\begin{pmatrix} u & v & 1 \end{pmatrix}^T = \mathbf{K}\Pi \left({}^{\text{body}}\mathbf{H}_{\text{DAVIS}} \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix} \right) \tag{2}$$

where Π is the projection function:

$$\Pi\left(\begin{pmatrix} X & Y & Z & 1 \end{pmatrix}^T\right) = \begin{pmatrix} \frac{X}{Z} & \frac{Y}{Z} & 1 \end{pmatrix}^T \tag{3}$$

and $\mathbf{K}$ is the camera intrinsics matrix for the rectified image (i.e. the top left 3×3 of the projection matrix).

Any points falling outside the image bounds are discarded, and the closest point at each pixel location in the image is used to generate the final depth map, examples of which can be found in Fig. 6.

In addition, we also provide raw depth images without any undistortion by unrectifying and distorting the rectified depth images using the camera intrinsics and OpenCV.

V. Calibration

In this section, we describe the various steps performed to calibrate the intrinsic parameters of each DAVIS and VI-Sensor camera, as well as the extrinsic transformations between each

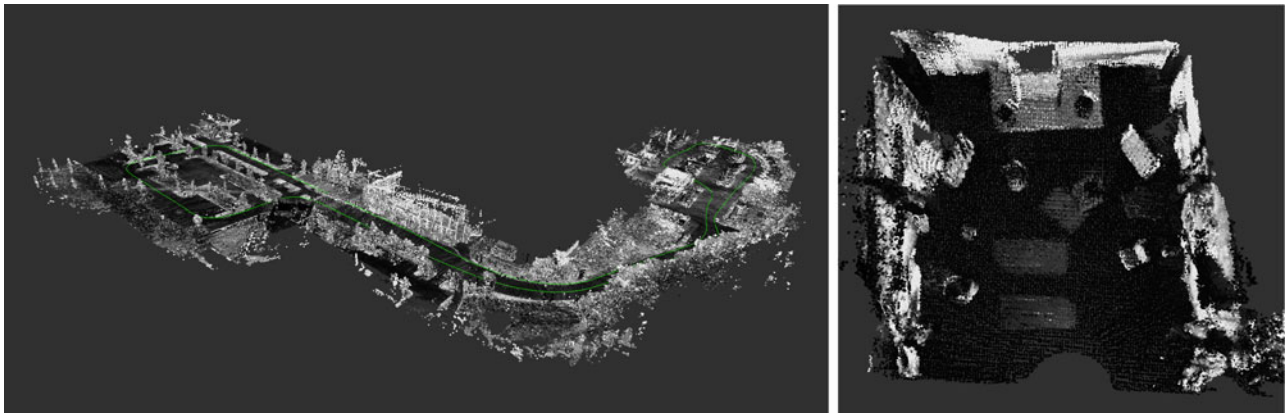

Fig. 5. Sample maps generated for ground truth. **Left**: Full map from Car Day 1 sequence, trajectory in green. **Right**: Local map from the Hexacopter Indoor 3 sequence.

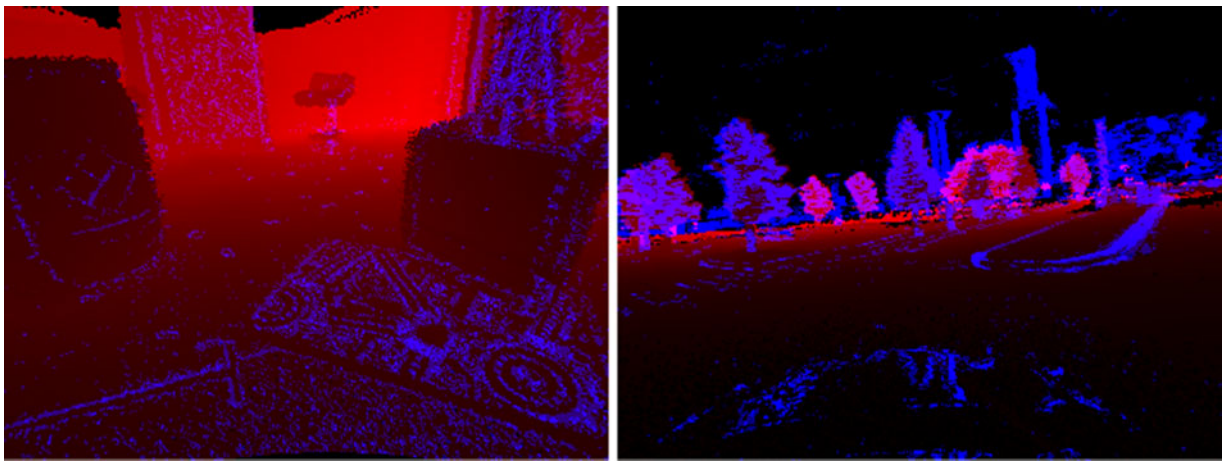

Fig. 6. Depth images (red) with events overlaid (blue) from the Hexacopter Indoor 2 and Car Day 1 sequences. Note that parts of the image (black areas, particularly the top) have no depth due to the limited vertical field of view and range of the lidar. These parts are labeled as NaNs in the data. Best viewed in color.

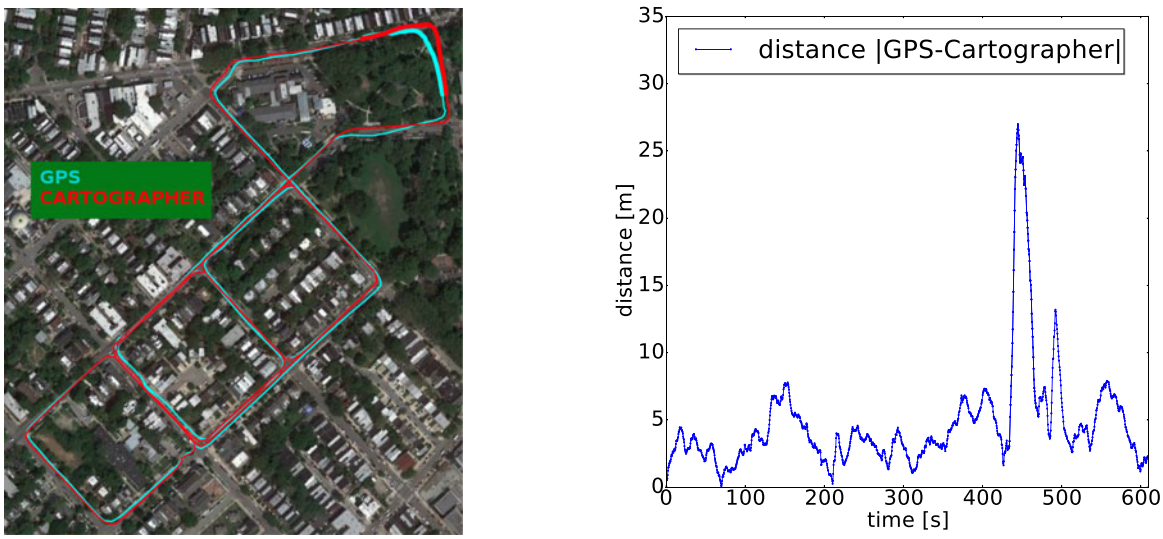

Fig. 7. Comparison between GPS and Cartographer trajectories for Car Day 2 overlaid on top of satellite imagery. Note that the spike in error between Cartographer and GPS corresponds to the bolded section in the top right of the overlay on the left, and is largely due to GPS error. Best viewed in color.

of the cameras, IMUs and the lidar. All of the calibration results are provided in yaml form.

The camera intrinsics, stereo extrinsics, and camera-IMU extrinsics are calibrated using the Kalibr toolbox[3] [30], [31], [32], the extrinsics between the left DAVIS camera and Velodyne lidar are calibrated using the Camera and Range Calibration Toolbox[4] [33], and fine tuned manually, and the hand eye calibration between the mocap model pose in the motion capture world frame and the left DAVIS camera pose is performed using CamOdoCal[5] [34]. To compensate for changes in the mounted rig, each calibration is repeated each day data was collected, and every time the sensing payload was modified. In addition to the calibration parameters, the raw calibration data for each day is also available on demand for users to perform their own calibration, if desired.

A. Camera Intrinsic, Extrinsic and Temporal Calibration

The camera intrinsics and extrinsics are estimated using a grid of AprilTags [35] that is moved in front of the sensor rig and calibrated using Kalibr. Each calibration provides the focal length and principal point of each camera as well as the distortion parameters and the extrinsics between the cameras.

In addition, we calibrate the temporal offset between the DAVIS stereo pair and the VI Sensor by finding the temporal offset that maximizes the cross correlation between the magnitude of the gyroscope angular velocities from the IMUs of the left DAVIS and the VI Sensor. The timestamps for the VI Sensor messages in the dataset are then modified to compensate for this offset.

B. Camera to IMU Extrinsic Calibration

To calibrate the transformation between the camera and IMU, a sequence is recorded with the sensor rig moving in front of the AprilTag grid. The two calibration procedures are separated to optimize the quality of each individual calibration. The calibration sequences are once again run through Kalibr using the camera-IMU calibration to estimate the transformations between each camera and each IMU, given the prior intrinsic and camera-camera extrinsic calibrations.

C. Motion Capture to Camera Extrinsic Calibration

Each motion capture system provides the pose of the mocap model in the motion capture frame at 100 Hz. However, the mocap model frame is not aligned with any camera frame, and so a further calibration is needed to obtain the pose of the cameras from the motion capture system.

The sensor rig was statically held in front of an Aprilgrid at a variety of different poses. At each pose at time t_i, the pose of the left DAVIS camera frame in the grid frame ${}^{\text{aprilgrid}}\mathbf{H}_{\text{DAVIS}(t_i)}$, as well as the pose of the mocap model (denoted body) in the mocap frame ${}^{\text{mocap}}\mathbf{H}_{\text{body}(t_i)}$, were measured. These poses were then used to solve the handeye calibration problem for the transform that transforms a point in the left DAVIS frame into the model frame ${}^{\text{body}}\mathbf{H}_{\text{DAVIS}}$:

$$ {}^{\text{body}(t_0)}\mathbf{H}_{\text{body}(t_i)}\,{}^{\text{body}}\mathbf{H}_{\text{DAVIS}} = {}^{\text{body}}\mathbf{H}_{\text{DAVIS}}\,{}^{\text{DAVIS}(t_0)}\mathbf{H}_{\text{DAVIS}(t_i)} \quad i = 1, \ldots, n \tag{4} $$

where:

$$ {}^{\text{body}(t_0)}\mathbf{H}_{\text{body}(t_i)} = {}^{\text{mocap}}\mathbf{H}^{-1}_{\text{body}(t_0)}\,{}^{\text{mocap}}\mathbf{H}_{\text{body}(t_i)} \tag{5} $$

$$ {}^{\text{DAVIS}(t_0)}\mathbf{H}_{\text{DAVIS}(t_i)} = {}^{\text{aprilgrid}}\mathbf{H}^{-1}_{\text{DAVIS}(t_0)}\,{}^{\text{aprilgrid}}\mathbf{H}_{\text{DAVIS}(t_i)}. \tag{6} $$

The optimization is performed using CamOdoCal, using the linear method in [36], and refining using a nonlinear optimization as described in [34].

D. Lidar to Camera Extrinsic Calibration

The transformation that takes a point from the lidar frame to the left DAVIS frame was initially calibrated using the Camera and Range Calibration Toolbox [33]. Four large checkerboard patterns are placed to fill the field of view of the DAVIS cameras, and a single pair of images from each camera is recorded, along with a full lidar scan. The calibrator then estimates the translation and rotation that aligns the camera and lidar observations of the checkerboards.

However, we found that the reported transform had up to five pixels of error when viewing the projected depth images (see Fig. 6). In addition, as the lidar and cameras are not hardware time synchronized, there was occasionally a noticeable and constant time delay between the two sensors. To improve the calibration, we fixed the translation based on the values from the CAD models, and manually fine tuned the rotation and time offset in order to maximize the overlap between the depth and event images. For visual confirmation, we provide the depth images with events overlaid for each camera. The timestamps of the lidar messages provided in the dataset are compensated for the time offset.

VI. Known Issues

A. Moving Objects

The mapping used to generate the depth maps assumes that scenes are static, and typically does not filter out points on moving objects. As a result, the reported depth maps may have errors of up to two meters when tracking points on other cars, etc. However, these objects are typically quite rare compared to the total amount of data available. If desired, future work could involve classifying vehicles in the images and omitting these points from the depth maps.

B. Clock Synchronization

The motion capture and GPS are only synchronized to the rest of the system using the host computer's time. This may incur an offset between the reported timestamps and the actual measurement time. We record all measurements on one computer to minimize this effect. In addition, there may be some delay between a lidar point's measurement and the timestamp of the message due to the spin rate of the lidar.

[3] https://github.com/ethz-asl/kalibr
[4] http://www.cvlibs.net/software/calibration/
[5] https://github.com/hengli/camodocal

C. DVS Biasing

Default biases for each camera were used when generating each sequence. However, it has been noted that, for the indoor flying sequences, the ratio of positive to negative events is higher than usual ($\sim$ 2.5–5x). At this point, we are unaware of what may have caused this imbalance, or whether tuning the biases would have balanced it. We note that the imbalance is particularly skewed over the speckled floor. We advise researchers using the polarities of the events to be aware of this imbalance when working with these sequences.

VII. Conclusion

We present a novel dataset for stereo event cameras, on a number of different vehicles and in a number of different environments, with ground truth 6dof pose and depth images. We hope that this data can provide one standard on which new event based methods can be evaluated and compared.

References

[1] D. Weikersdorfer, D. B. Adrian, D. Cremers, and J. Conradt, "Event-based 3D SLAM with a depth-augmented dynamic vision sensor," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 359–364.

[2] B. Rueckauer and T. Delbruck, "Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor," *Frontiers Neurosci.*, vol. 10, p. 176, 2016.

[3] F. Barranco, C. Fermuller, Y. Aloimonos, and T. Delbruck, "A dataset for visual navigation with neuromorphic methods," *Frontiers Neurosci.*, vol. 10, p. 49, 2016.

[4] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM," *Int. J. Robot. Res.*, vol. 36, no. 2, pp. 142–149, 2017. [Online]. Available: http://dx.doi.org/10.1177/0278364917691115

[5] J. Binas, D. Niel, S.-C. Liu, and T. Delbruck, "DDD17: End-to-End DAVIS Driving Dataset.," *ICML'17 Workshop Mach. Lear. Auton. Vehicles*, Sydney, Australia, 2017.

[6] J. Kogler, C. Sulzbachner, F. Eibensteiner, and M. Humenberger, "Address-event matching for a silicon retina based stereo vision system," in *Proc. 4th Int. Conf. Sci. Comput. Comput. Eng.*, 2010, pp. 17–24.

[7] J. Kogler, M. Humenberger, and C. Sulzbachner, "Event-based stereo matching approaches for frameless address event stereo data," *Proc. 7th Int. Conf. Adv. Visual Comput.-Volume Part I*, pp. 674–685, 2011.

[8] E. Piatkowska, A. N. Belbachir, and M. Gelautz, "Cooperative and asynchronous stereo vision for dynamic vision sensors," *Meas. Sci. Technol.*, vol. 25, no. 5, 2014, Art. no. 055108.

[9] M. Firouzi and J. Conradt, "Asynchronous event-based cooperative stereo matching using neuromorphic silicon retinas," *Neural Process. Lett.*, vol. 43, no. 2, pp. 311–326, Apr. 2016. [Online]. Available: https://doi.org/10.1007/s11063-015-9434-5

[10] E. Piatkowska, J. Kogler, N. Belbachir, and M. Gelautz, "Improved cooperative stereo matching for dynamic vision sensors with ground truth evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 370–377.

[11] P. Rogister, R. Benosman, S.-H. Ieng, P. Lichtsteiner, and T. Delbruck, "Asynchronous event-based binocular stereo matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 2, pp. 347–353, Feb. 2012.

[12] J. Carneiro, S.-H. Ieng, C. Posch, and R. Benosman, "Event-based 3D reconstruction from neuromorphic retinas," *Neural Netw.*, vol. 45, pp. 27–38, 2013.

[13] L. A. Camuñas-Mesa, T. Serrano-Gotarredona, S. H. Ieng, R. B. Benosman, and B. Linares-Barranco, "On the use of orientation filters for 3D reconstruction in event-driven stereo vision," *Frontiers Neurosci.*, vol. 8, p. 48, 2014.

[14] R. Benosman, S.-H. Ieng, P. Rogister, and C. Posch, "Asynchronous event-based Hebbian epipolar geometry," *IEEE Trans. Neural Netw.*, vol. 22, no. 11, pp. 1723–1734, Nov. 2011.

[15] D. Zou *et al.*, "Context-aware event-driven stereo matching," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 1076–1080.

[16] S. Schraml, A. Nabil Belbachir, and H. Bischof, "Event-driven stereo matching for real-time 3D panoramic vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 466–474.

[17] A. Z. Zhu, N. Atanasov, and K. Daniilidis, "Event-based feature tracking with probabilistic data association," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 4465–4470.

[18] D. Tedaldi, G. Gallego, E. Mueggler, and D. Scaramuzza, "Feature detection and tracking with the dynamic and active-pixel vision sensor (DAVIS)," in *Proc. 2nd Int. Conf. Event-Based Control, Commun., Signal Process.*, 2016, pp. 1–7.

[19] B. Kueng, E. Mueggler, G. Gallego, and D. Scaramuzza, "Low-latency visual odometry using event-based feature tracks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 16–23.

[20] A. Z. Zhu, N. Atanasov, and K. Daniilidis, "Event-based visual inertial odometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5391–5399.

[21] G. Gallego and D. Scaramuzza, "Accurate angular velocity estimation with an event camera," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 632–639, Apr. 2017.

[22] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3D reconstruction and 6-DOF tracking with an event camera," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 349–364.

[23] H. Rebecq, T. Horstschaefer, G. Gallego, and D. Scaramuzza, "EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real time," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 593–600, Apr. 2017.

[24] H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization," in *Proc. Brit. Mach. Vis. Conf.*, vol. 3, 2017.

[25] E. Mueggler, G. Gallego, H. Rebecq, and D. Scaramuzza, "Continuous-time visual-inertial trajectory estimation with event cameras," arXiv:1702.07389, 2017.

[26] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240$\times$ 180 130 dB 3 μs latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, Oct. 2014.

[27] J. Nikolic *et al.*, "A synchronized visual-inertial sensor system with FPGA pre-processing for accurate real-time SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 431–437.

[28] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2D LIDAR SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 1271–1278.

[29] J. Zhang and S. Singh, "LOAM: Lidar odometry and mapping in real-time," *Robot., Sci. Syst.*, vol. 2, 2014.

[30] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 1280–1286.

[31] P. Furgale, T. D. Barfoot, and G. Sibley, "Continuous-time batch estimation using temporal basis functions," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 2088–2095.

[32] J. Maye, P. Furgale, and R. Siegwart, "Self-supervised calibration for robotic systems," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2013, pp. 473–480.

[33] A. Geiger, F. Moosmann, Ö. Car, and B. Schuster, "Automatic camera and range sensor calibration using a single shot," in *Proc. Int. Conf. Robot. Autom.*, St. Paul, MN, USA, May 2012.

[34] L. Heng, B. Li, and M. Pollefeys, "Camodocal: Automatic intrinsic and extrinsic calibration of a rig with multiple generic cameras and odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 1793–1800.

[35] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 3400–3407.

[36] K. Daniilidis, "Hand-eye calibration using dual quaternions," *Int. J. Robot. Res.*, vol. 18, no. 3, pp. 286–298, 1999.