Object Tracking by Jointly Exploiting Frame and Event Domain

Supplementary Material

Jiqing Zhang[1,*], Xin Yang[1,*], Yingkai Fu[1], Xiaopeng Wei[1], Baocai Yin[1,†], Bo Dong[2,†]
[1]Dalian University of Technology, [2] SRI International

1. Dataset

1.1. Dataset Collection

The FE108 dataset is simultaneously recorded by a DAVIS346 camera and the Vicon motion capture system [2]. The hardware setup used for collecting the FE108 is illustrated in Figure 1. The DAVIS346 equips with both dynamic vision sensor (DVS) and a frame-based active pixel sensor (APS), and it can captures both events and grayscale frames simultaneously. The detailed specifications of DAVIS346 can be found in Table 1. Since event-based cameras only use power to process changing pixels, the power consumption is significantly lower than conventional cameras (*i.e.*, $\leq$ 100 mW vs. $\geq$ 3W).

The Vicon system can provide the 3D position and trajectory of targets in a high sample rate and sub-millimeter precision by 12 Vero motion capture infrared cameras. Since Vicon system leverages active sensing to track objects. The infrared light emitted from the system becomes noise in the events domian. To deal with it, we place an infrared filter in front of the DAVIS346 to filter out the light with wavelength above 700nm. We set the sampling rate of the DAVIS346 camera's APS to 20/40Hz, and the sampling rate of the Vicon to 240Hz.

Table 1. Specifications of DAVIS346.

DVS Resolution	346 $\times$ 260 pixels
Frame Resolution	346 $\times$ 260 pixels, grayscale
DVS Dynamic Range	120 dB
APS Dynamic Range	56.7 dB
Min Latency	20 μs
Bandwidth	12 *M*Events / second
Weight	100 g
Dimensions	H$\times$W$\times$D (40$\times$60$\times$25 mm)

* Joint first authors. †Baocai Yin (ybc@dlut.edu.cn) and Bo Dong (bo.dong@sri.com) are the corresponding authors.

1.2. Dataset Annotation

The data annotation is divided into two steps: a) the coordinate system transformation between DAVIS346 and Vicon; b) transforming the 3D point on the target to a 2D point.

The coordinate system transformation between DAVIS346 and Vicon. We first use the calibration board to determine the event camera matrix *K* and distortion coefficients *d*. Then we can obtain the rotation vector *r* and the translation vector *t* of the DAVIS346 through the following equation,

$$r, t = S\left(K, d, p_i, P_i\right), \quad i = 1, 2, ..., 25, \tag{1}$$

where *S* denotes the SolvePnP [1] method, p_i is a set of 2D points on the grayscale images from APS, and P_i is a set of 3D points on the targets from the Vicon. To obtain p_i and P_i, we use a wand that can be tracked by both Vicon and APS. There are 5 markers on the wand, and we place the wand in 5 different positions. In this way, we can collect a total of 25 paired p_i and P_i.

Transforming the 3D point on the target to a 2D point. We transform the 3D points provided by the Vicon to 2D points of grayscale images by the following equation,

$$\begin{bmatrix} x_j \\ y_j \\ 1 \end{bmatrix} = \begin{bmatrix} r & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_j \\ Y_j \\ Z_j \end{bmatrix}, \tag{2}$$

where $[x_j, y_j, 1]^T$ and $[X_j, Y_j, Z_j]^T$ denote the 2D and the 3D coordinates of the *j*th marker on targets, respectively. Then the label bounding box of the target can be acquired by calculating the maximum and minimum values of all 2D points,

$$\begin{aligned} x_l &= min\left\{x_j\right\}, \qquad x_r = max\left\{x_j\right\}, \\ y_l &= min\left\{y_j\right\}, \qquad y_r = max\left\{y_j\right\}, \end{aligned} \tag{3}$$

where (x_l, y_l) is the point in the upper left corner of the label bounding box, (x_r, y_r) is the point in the lower right

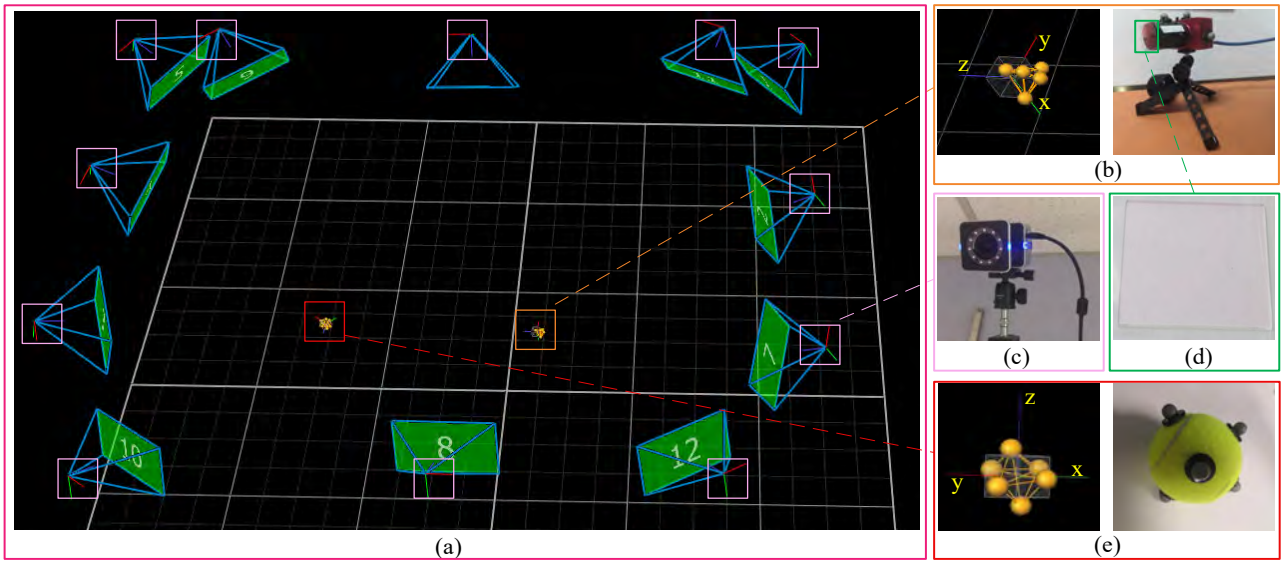
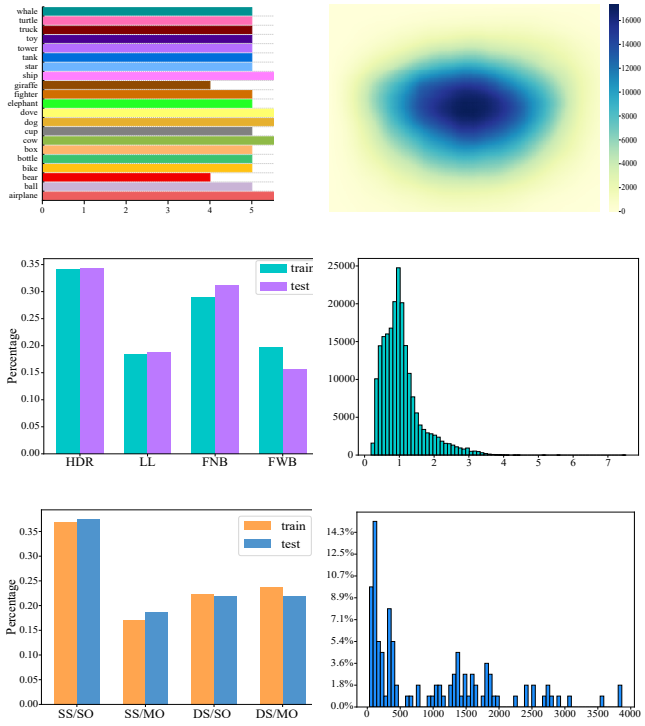


Figure 1. The FE108 dataset recording setup. (a) the targets and DAVIS346 under Vicon system. (b) the DAVIS346 and it's 3D model in Vicon (c) the Vero motion camera. (d) the filter lens. (e) the targets and it's 3D model in Vicon.



(e) Scenes distribution (f) Avg event rate (Ev/ms)

Figure 2. Statistics of FE108 dataset in terms of (a) classes, (b) bounding box center position, (c) attributes, (d) aspect ratios (H/W), (e) scenes, and (f) event rate.

corner of the label bounding box. We can get the width w and height h of the bounding box by the following equation,

$$w = x_r - x_l, \qquad h = y_r - y_l. \tag{4}$$

1.3. Dataset Format

We collect a total of 108 sequences, with an average of 1932 grayscale frames and 51.9M events per sequence. We split the recordings into 76 training sets and 32 testing sets. Each sequence folder contains 5 files/folders, *i.e. img*, *img_groundtruth.txt*, *event.txt*, *event_groundtruth.txt* and *event.aedat4*. Grayscale frames (20/40Hz) are stored in the *img* folder. *img_groundtruth.txt* records the label bounding box corresponding to the grayscale frame in the form of (x, y, h, w). Each event in *event.txt* is represented by a tuple $e = (x, y, t, p)$. *event_groundtruth.txt* records the corresponding event bounding box at 240Hz. *event.aedat4* is the DAVIS346 raw output which provides IMU and Trigger data in addition to raw events and frames.

1.4. Dataset Facts

Categorical Analysis: The FE108 dataset can be categorized differently from different perspectives. The first perspective is the number of object classes. There are 21 different object class, which can be divided into three categories: animal, vehicle, and daily goods (e.g., bottle, box). More details of all object classes are illustrated in Figure 2 (a). Second, as shown in Figure 2 (c), the FE108 contains four types of challenging scenes: low-light (LL), high dynamic range (HDR), fast motion with and without motion blur on APS frame (FWB and FNB). Third, based on the camera movement and number of target objects, as shown in Fig-

ure 2 (e), FE108 has four scenes: static shots with a single object and multiple objects; dynamic shot with a single object and various objects.

Ground Truth Bounding Box statistics: In Figure 2 (b), we plot out the distribution of all annotated bounding box locations, which shows most annotations are close to frames' centers. In Figure 2 (d), we also show the distribution of the bounding box aspect ratios (*i.e.*, H/W). These two distributions evidence the annotated bounding boxes' diversity is pretty wide.

Event Rate: The FE108 dataset is collected in a constant lighting condition. It means all events are triggered by motions (*e.g.*, moving objects, camera motion). Therefore, the distribution of the event rate can represent the motion distribution of FE108. We compute the event stream rate of FE108 as follows: Firstly, for the event stream of each captured scene, we discretize its time dimension into 1ms intervals. Secondly, For each interval, we count the number of events. Finally, the average number of events of all intervals is used as the event rate. As shown in Figure 2 (f), the distribution of the event rate is pretty diverse. It indicates the captured 108 scenes offer wide motion diversity.

2. Evaluation on Standard RGB Benchmarks

The standard RGB benchmarks (such as, OTB2013[9], VOT2015[7], and GOT-10k [6]) do not have associated event data. An open event camera simulator, ESIM[8], can generate events data based on frames from the frame-domain. However, the generated events data between adjacent frames are not faithful. If frames experience over/underexposure or motion blur, the generated events are not faithful either. Therefore, without accurate event data, validation based on standard benchmarks is not convincing. Even so, we adopt ESIM[8] to generate events and compare our method with three state-of-the-art approaches (*i.e.*, ATOM [4], DiMP [3] and PrDiMP [5]) on OTB2013[9]. All methods are trained on our FE108. As shown in Table 2, we can see that our method still slightly outperforms other approaches in terms of RSR, $OP_{0.50}$ and $OP_{0.75}$ even though no real events are available, which shows the effectiveness of our approach. Since using the event simulator on the standard RGB dataset does not give full play to the advantages of events, proposing non-event datasets/challenges is outside the scope of our work. The results also demonstrate that the proposed FE108 dataset is important to stimulate more future research on multi-modal learning with real asynchronous events.

3. More Results

We provide more results (in *supp_video*) to show the effectiveness of our proposed method.

Table 2. Quantitive results on OTB2013[9] dataset

Benchmarks	Methods	RSR ↑	$OP_{0.50}$ ↑	$OP_{0.75}$ ↑	RPR ↑
OTB2013[9]	ATOM [4]	16.4	9.4	1.7	21.2
	DiMP [3]	16.0	11.5	2.3	22.3
	PrDiMP [5]	16.7	12.5	2.0	21.0
	Ours	18.0	13.8	3.4	21.4

References

[1] Solvepnp. https://docs.opencv.org/. 1

[2] Vicon motion capture. https://www.vicon.com/. 1

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 3

[4] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, 2019. 3

[5] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020. 3

[6] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[7] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *ICCVW*, 2015. 3

[8] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *CoRL*, 2018. 3

[9] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 3