Event-based dataset with video augmentation for scene understanding

James P. Turner[1], Neelay Shah[2], Jens E. Pedersen[3], Jörg Conradt[3], Thomas Nowotny[1]

[1]Sussex Neuroscience, School of Engineering and Informatics, University of Sussex, UK, [2]Max Planck Institute for Intelligent Systems, Germany, [3]KTH, Royal Institute of Technology, Sweden

Abstract

We have created an event-based dataset generation pipeline for visual scene understanding, with video augmentation:

- Datasets contain streams of events from one or more event cameras (x, y, time, polarity)
- Per-event semantic segmentation label for all cameras
- Translation and Rotation (pose) labels, relative to all cameras
- Event streams are augmented with spike-encoded 3D Fractal noise or video clips (from YouTube)
- Dense 3D noise or videos are converted into sequences of sparse event frames using a technique which 'assigns' events by thresholding the differences in pixel intensities between consecutive frames
- Events from these frames - identified by their location and timestamp - are 'injected' into the recorded event streams

Hardware Setup

- 8x 3D tracking (Vicon Vero 2.2) cameras
- 2x event/RGB (DVS DAVIS 346) cameras with NIR cut filters

Begin with prop STL mesh (used for 3D print) and known 3D tracking LED marker locations

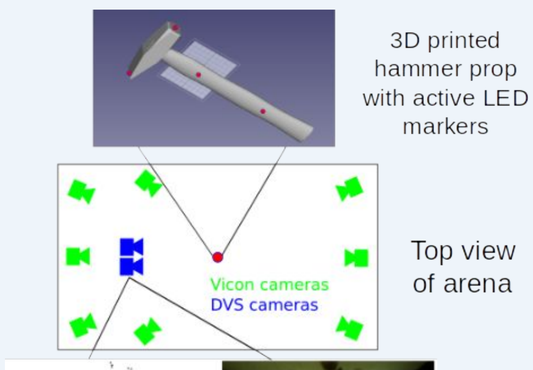
Learn optimal transformation $y = Ax + b$ from 3D Vicon space x to 3D camera space y, from known marker locations

$$\vec{y}^i = A\vec{x} + \vec{b}$$

$$\vec{z}^i = f \cdot \begin{pmatrix} y_1^i / y_3^i \cdot k \\ y_2^i / y_3^i \end{pmatrix} + \begin{pmatrix} s_x/2 \\ s_y/2 \end{pmatrix}$$

Project prop STL mesh to 2D camera image plane at coordinates z using standard pinhole camera model

A is a learnt rotation matrix, b is a learnt translation vector, f is focal length times pixel density, s_x and s_y are the horizontal and vertical camera resolution



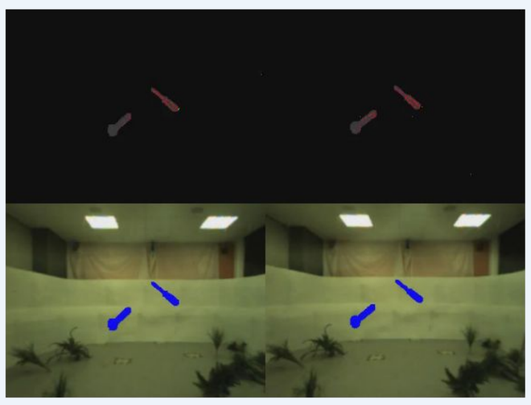
3D Tracker Props

Issue: 3D tracker system uses NIR light strobing to detect location of passive reflective markers, which introduces heavy noise to DVS cameras (even with filters)

Solution: disable NIR strobing on 3D tracking system, and use 'active' markers instead

- Custom 3D printed hollow props
- Use omnidirectional 780 nm NIR LED markers, rather than the standard passive reflectors
- Markers are wired into props at predetermined points (from STL mesh file)
- Any remaining noise from these lower-power non-strobing lights is easily filtered using 780 nm NIR cut filters

Raw Sample Dataset



Small sample dataset with 9 separate 30 second recordings of suspended moving props [1]

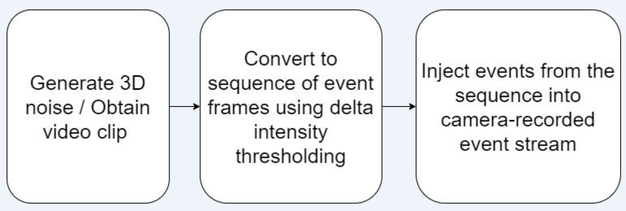
Camera-relative prop pose and per-event class labels for each of two cameras

Stored in HDF5 format online
https://doi.org/10.25377/sussex.17112080.v1

Processing code is available online
https://github.com/jamesturner246/vicon-dvs-projection

Video Augmentation

- 3D Fractal noise or custom videos can be used for augmentation
- The frames of the noise/video are converted to a sequence of events using a thresholding technique
- Between frames, If a pixel's intensity changes by an amount greater than a given threshold, an event is generated for that pixel



- Different event polarity, depending on direction of change of pixel value
- The raw dataset is then augmented by injecting events generated from the video/noise
- Code for augmenting event streams end-to-end is available at: https://github.com/NeelayS/event_aug

Augmented Dataset

- We augmented the camera-recorded event streams using both video clips downloaded from YouTube and generated 3D Fractal noise.
- The videos were selected such that they had object(s) performing relative motion with respect to their surroundings in them. This guaranteed the presence of a coherent stream of events after spike-encoding.
- The pictures below show an example of augmentation. On the left is one frame of a video containing dashcam footage of a car. The picture on the right depicts a still of an event recording of a screwdriver augmented with events from the video frame.

Summary and Discussion

- We have constructed a full pipeline for generating event-based datasets for semantic segmentation and pose estimation using custom 3D printable tracked props [1] [2]
- We have further implemented an event-based visual data augmenting process, which projects prop events from a raw dataset onto event-based fractal noise or video sources
- The prop-based data generation pipeline requires a 3D tracking system, such as Vicon, while the augmentation only requires standard video files or Youtube links
- We hope to contribute to further efforts for spiking neural network development, and ultimately neuromorphic computing in general by making training data more readily available

Acknowledgements

We thank Poppy Collis for her contributions in an early stage of this work

References

[1] James Turner, Jens Pedersen, Jörg Conradt, and Thomas Nowotny. 2022. Event-based dataset for classification and pose estimation. In *Neuro-Inspired Computational Elements Conference (NICE 2022), March 28–April 1, 2022, Virtual Event, USA.* ACM, New York, NY, USA 3 Pages. https://doi.org/10.1145/3517343.3517378

[2] James Turner, Jens Pedersen, Jörg Conradt, Thomas Nowotny. 2022. Stereo event- and frame-based benchmark dataset for scene understanding. University of Sussex. Dataset. https://doi.org/10.25377/sussex.17112080.v1

Project support
EPSRC (EP/P006094/1, EP/S030964/1),
H2020 945539 (HBP SGA3)