Live Demonstration: Integrating Event Based Hand Tracking Into TouchFree Interactions

Ryan Page
Ultraleap, Glass Wharf, Bristol, UK
ryan.page@ultraleap.com

Abstract

Hand tracking is becoming ever more prominent as an intuitive way to interact with digital content. There are however technical challenges at the sensing level, these include environmental robustness, power and system latency to name three. Event cameras have the potential to solve these, with their asynchronous readout, low power and high dynamic range. To explore the potential of event cameras, Ultraleap have developed a prototype stereo camera using two Prophesee IMX636ES sensors. To go from event data to hand positions the event data is aggregated into event frames. This is then consumed by a hand tracking model which outputs 28 joint positions for each hand with respect to the camera. This data is used by Ultraleap's TouchFree application to determine a users interaction with a 2D display. This is brought together in the Ball Pit demo shown in the supplementary material, where the user is using their hands to remove balls from a box.

Supplementary Material

A video showing the TouchFree Ball Pit experience, driven by real-time event based hand tracking is included as part of the submission.

1. Introduction

Hand tracking is an intuitive way to interact with the digital world. This can take the form of bringing your physical hands into virtual/augmented reality (AR/VR) or controlling machines via Human Machine Interfaces (HMIs). The demonstration here addresses the latter, with work ongoing into AR/VR platforms. HMIs are used in a number of applications including kiosks, digital signage and location based entertainment. One of the challenges often faced in these applications is they are subject to adversarial environmental conditions, either bright sunlight or overhead lighting and both daytime and nighttime use. The high dynamic range and the fact that events are driven by scene dynamics makes them an ideal modality for use in these conditions. To explore event based vision a stereo event camera rig was developed, the rig included a Ultraleap SIR170 stereo camera [4] as a reference device. This was used to create a labelled event frame dataset, where event frames were generated using OpenEB's timesurfaces implementation [3]. This dataset was used to train the state-of-art Ultraleap Gemini model [6]. This model is used in the Ball Pit demo. The demo utilises the Ultraleap TouchFree [5] toolkit which aggregates the 28 3D hand joint positions and determines the users interaction with a screen. The demo challenges a user to remove all the balls from a box by gesturing left and right, up and down in a swiping motion, this makes for an entertaining experience that introduces some key touchless ideas. The rest of this paper is dedicated to introducing the camera rig, calibration, data labelling and inference and the demo.

2. Camera Rig

The event camera rig consists of two Prophesee IMX636ES sensors with MetaVision EVK3 [2] readout boards. The cameras have ultra-wide angle lenses with visible bandpass filters. A SMA coaxial cable is used to synchronise the cameras ensuring both have a common time frame. The cameras are in a stereo arrangement with a baseline of 81 mm. A SIR170 is fixed above to enable synchronised dual recordings. All cameras are connected to the host over USB, synchronised recordings are taken with a custom Data Acquisition System (DAQ). The rig is shown in figure 1.

3. Calibration

Both cameras use ultra-wide angle lenses and were calibrated using the OpenCV fisheye model, which is based on the Kanala-Brandt model [7]. In both cases a checkerboard target was used, with calibration parameters determined using the Calib.io [1] toolkit. To generate calibration images for the event camera a monitor was placed on an articulated arm with a flickering checkerboard pattern. The events were

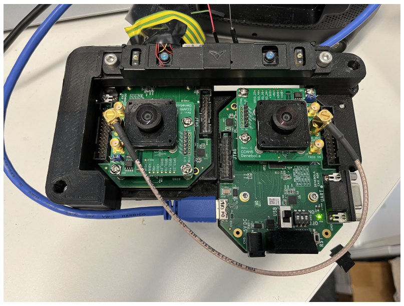

Figure 1. Figure showing stereo event camera rig consisting of two Metavision EVK3s, along with SIR170 for data labelling.

aggregated into event frames using a positive polarity histogram with an accumulation time approximately ten times the flicker rate. To enable event frames to be labelled with hand positions, the extrinsics between the two stereo cameras needed to be determined. This was done using a custom LED wand fixed to an articulated arm, using triangulation the LED positions could be determined for each stereo camera system, then the 6 degrees of freedom making up the extrinsics could be solved by minimising the root squared error of the point-to-point distance.

4. Data Labelling and Inference

The datasets used to train a model were generated by recording a sequence of hand gestures with both stereo cameras. The event stream was then preprocessed to event frames using linear time-surfaces, similar to the process used by [8]. To generate labels the recordings from the SIR170 were passed through an existing Gemini hand tracking model to determine hand positions. The calibration between the stereo cameras was then used to rebase the hand positions to the event camera frame of reference. The 3D points were then projected to pixel coordinates using the event camera calibration to check alignment. The hand points projected onto the event frame are shown in figure 2.

From figure 2 the hand positions can be seen to overlay with the most recent events in the time-surface. Model execution is on a development platform on a host PC, which includes the event signal processing to enable event frames to be generated in real-time. The 3D hand positions could then be consumed by the TouchFree application.

5. Visitor Experience

In the demonstration an interactive ball bit is shown on a monitor, with a stereo event camera positioned below to track the hands. The users hands are then used to control an invisible spherical collider, which moves through the balls. The spherical collider is moved up, down, left and right with a spotlight tracking the hand over the ball pit. If the user moves their hand towards the screen this increases the size of the collider. The aim is clear the balls as fast as possible.

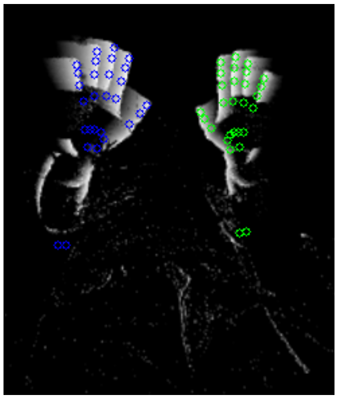

Figure 2. Figure showing hand labels overlaid on a linear time-surface generated from the event stream.

References

[1] calib.io Camera Calibration. https://calib.io/products/calib. 1

[2] Prophesee Metavision EVK3. https://www.prophesee.ai/event-based-evaluation-kits/. 1

[3] Prophesee OpenEB. https://github.com/prophesee-ai/openeb. 1

[4] Stereo IR 170. https://www.ultraleap.com/product/stereo-ir-170/. 1

[5] TouchFree. https://www.ultraleap.com/enterprise/touchless-experiences/touchfree-solution/. 1

[6] Ultraleap Gemini. https://www.ultraleap.com/tracking/gemini-hand-tracking-platform/. 1

[7] J. Kannala and S.S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1335–1340, 2006. 1

[8] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *International Conference on Computer Vision (ICCV)*, 2021. 2