Technische Universität Berlin

EINSTEIN CENTER Digital Future

Event-based Robot Vision

Prof. Dr. Guillermo Gallego
Chair: Robotic Interactive Perception

guillermo.gallego@tu-berlin.de

http://www.guillermogallego.es

Event Representations

Welcome to the Zoo

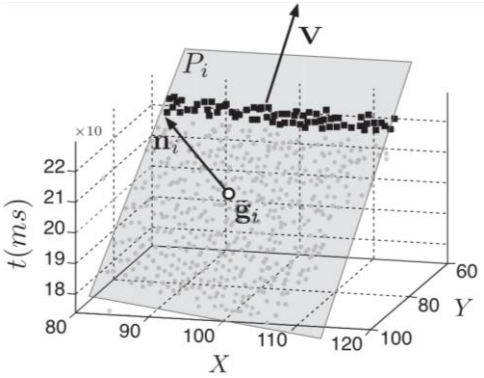
A Zoo of Representations

- Individual events (filters, SNNs…)
- Point sets on image plane
- 3D point set
- Event frame (2D grid)
  - Histogram
  - Time surface
  - Motion-compensated event frames (given motion hypothesis)
- Voxel grid (3D histograms)
- Reconstructed intensity / brightness images
- …

Point sets

- **Individual Events**
  - An event $e_k = (\boldsymbol{x}_k, t_k, p_k)$ is represented by its (four) numbers.
  - This representation is used in event-by-event processing methods, such as filters and Spiking Neural Networks (SNNs).

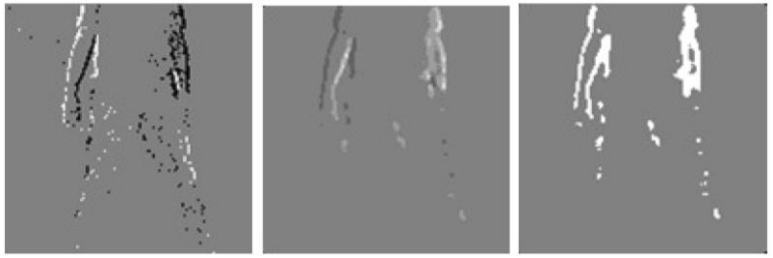
- **3D Point set**
  - Treat events as points in space-time R^3.
    Think geometrically in terms of point "clouds"
  - **Pros**: Preserves space-time information
  - **Cons**: Introduces some latency and might discard polarity



- **Evolving point set on the image plane**
  - Treat events as samples of a time-evolving "**shape**" on the image plane. Intuitive since events are caused by moving edges.

Event Frames

- Events $\mathcal{E} = \{e_k\}_{k=1}^{N_e}$ in a space-time neighborhood are **converted** into (2D) images

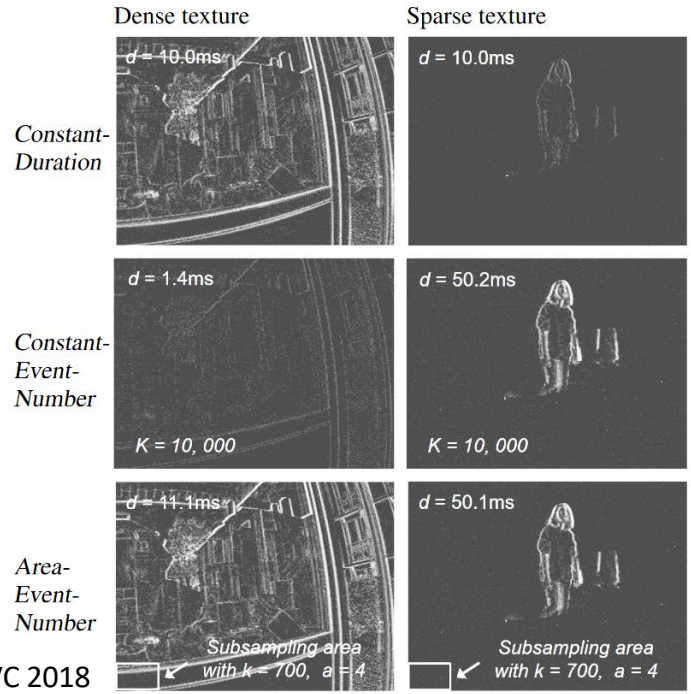


Kogler et al. ICVS 2009
"Address-event to frame *converter*"

- **Types**:
  - Histograms of events (pixelwise)
  - Balance of event polarities
  - Saturated histogram (e.g., ternary image)
  - Time surfaces (they are 2D images after all); discuss separately
- Event **selection** (sliding window):
  - Constant time: "all events in a time interval of given size T"
  - Constant number of events N_e
  - Adaptive Area-Event-number
  - Other strategy

Event Frames

- **Advantages**: Why do they have such a high impact?
  - **Compatible** with conventional computer vision: convert the unfamiliar event output to the familiar one (images) and **re-utilize** methods from standard cameras.
  - Events are caused by moving edges, so frames have an **intuitive** and **informative** interpretation: “edge maps”. Edges convey a lot of the information of a scene.
  - They are good as a **baseline** of what is achievable.
  - Frames are HDR and may be asynchronous
- **Disadvantages**:
  - Not the event-based paradigm
    - **Quantizes** the event timestamps
    - **Power** is spent in creating frames
    - Some **latency** is also introduced
  - How many events to use? Key parameter; may be difficult to tune



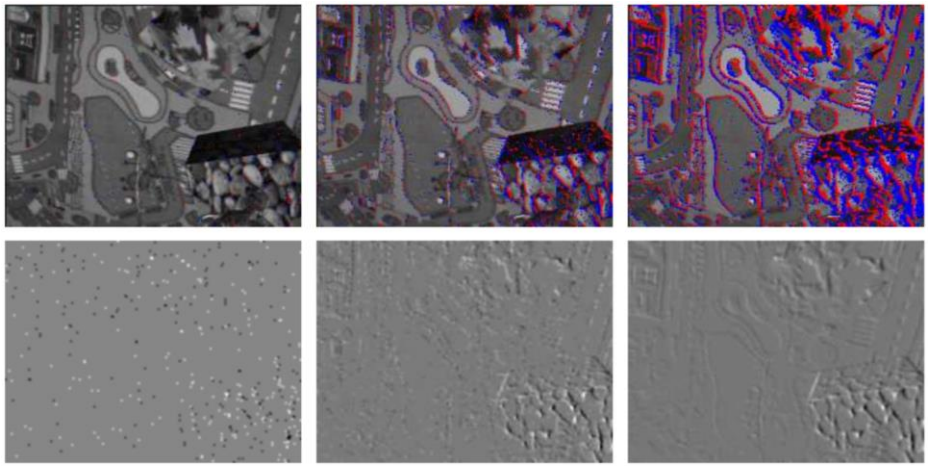
Liu et al., BMVC 2018

Brightness Increment Images

Obtained by accumulating event polarities, pixelwise.

- **Advantages**:
  - Very intuitive interpretation $\Delta L(\boldsymbol{x})$ images, like the type of images obtained by subtracting two video frames, related to “brightness constancy” equation.
  - Polarity is used

- Usage cases:
  - Stereo depth estimation
  - Camera pose estimation
  - Optical Flow estimation
  - Grayscale frame prediction
  - …

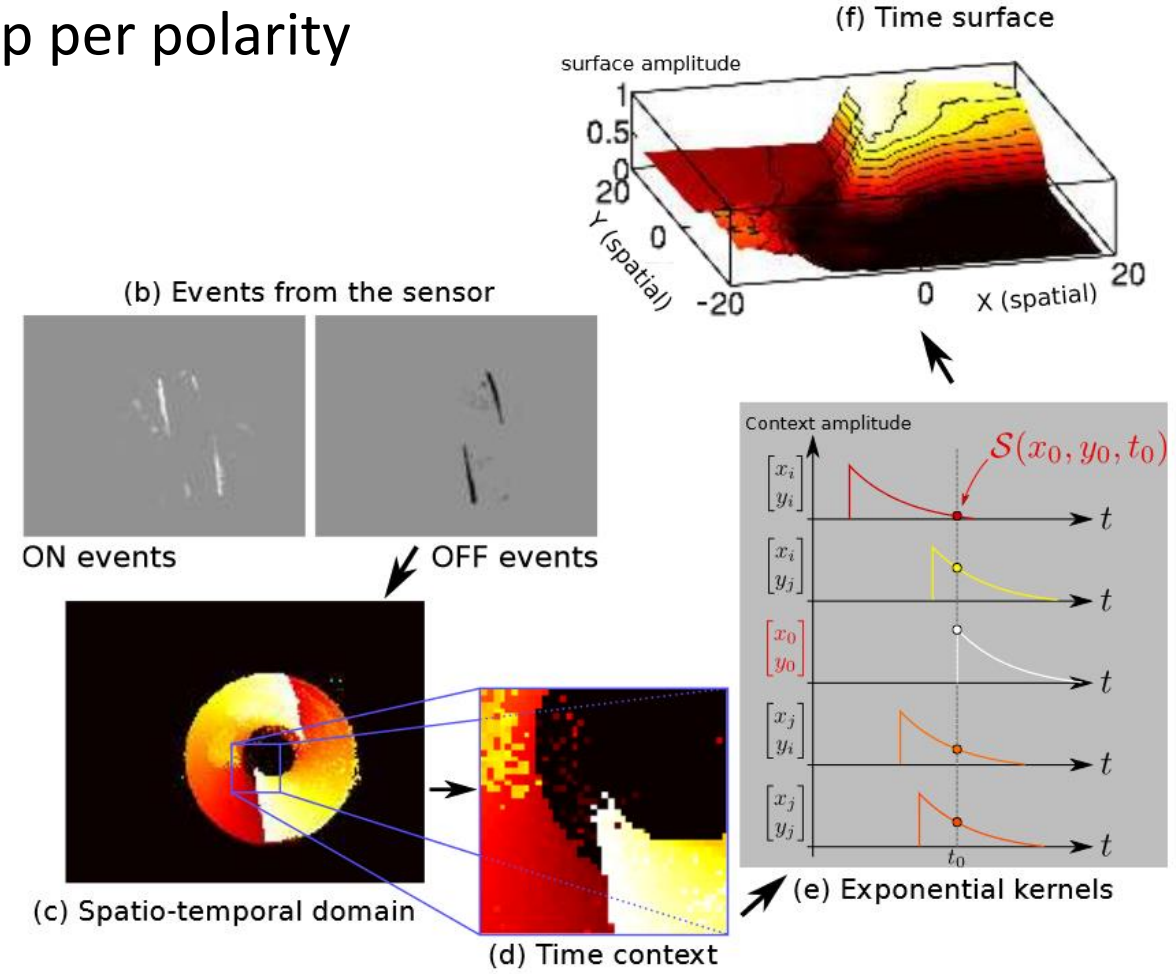


~ 500 events ~ 4300 events ~ 21000 events

Bryner et al. ICRA 2019

Time Surfaces

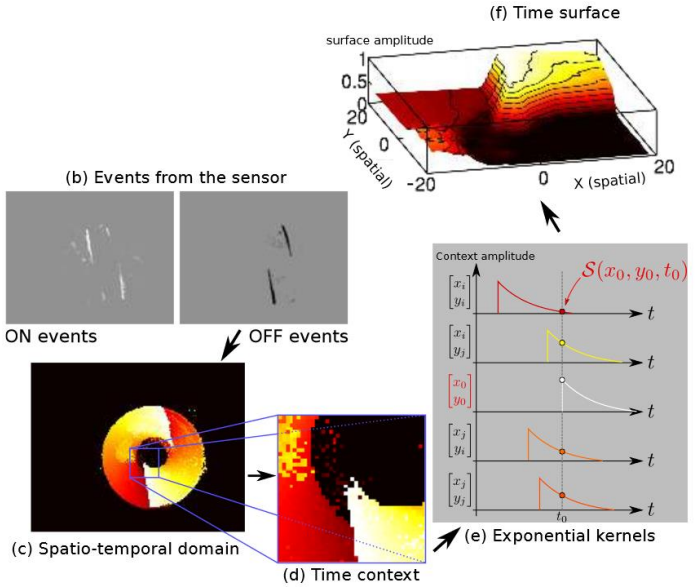
- A time map / image
  - Pixel value = *f* (time of last event)
  - One map per polarity



Lagorce et al., *HOTS: A Hierarchy Of event-based Time-Surfaces for pattern recognition*, PAMI 2017

Time Surfaces

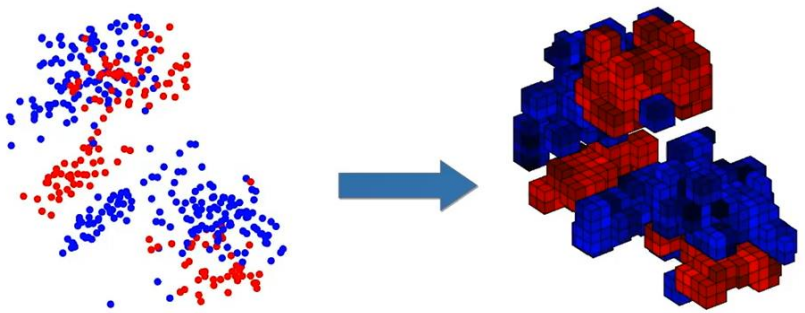
- A time map / image
  - Pixel value = *f* (time of last event)
  - Separate by polarity
  - Kernels may emphasize recent events



Lagorce et al. PAMI 2015

- **Advantages**:
  - Expose rich temporal information of events
  - Intuitive: intensity is a function of motion history
  - Can be robust to noise by local filtering (Sironi et al. CVPR 2018)
  - Asynchronous update with every event
  - Compatible with conventional computer vision
- **Disadvantages**:
  - Only one value per pixel even if multiple events on same pixel
  - Not good for textured scenes (pixel overwrite frequently)

Voxel Grids

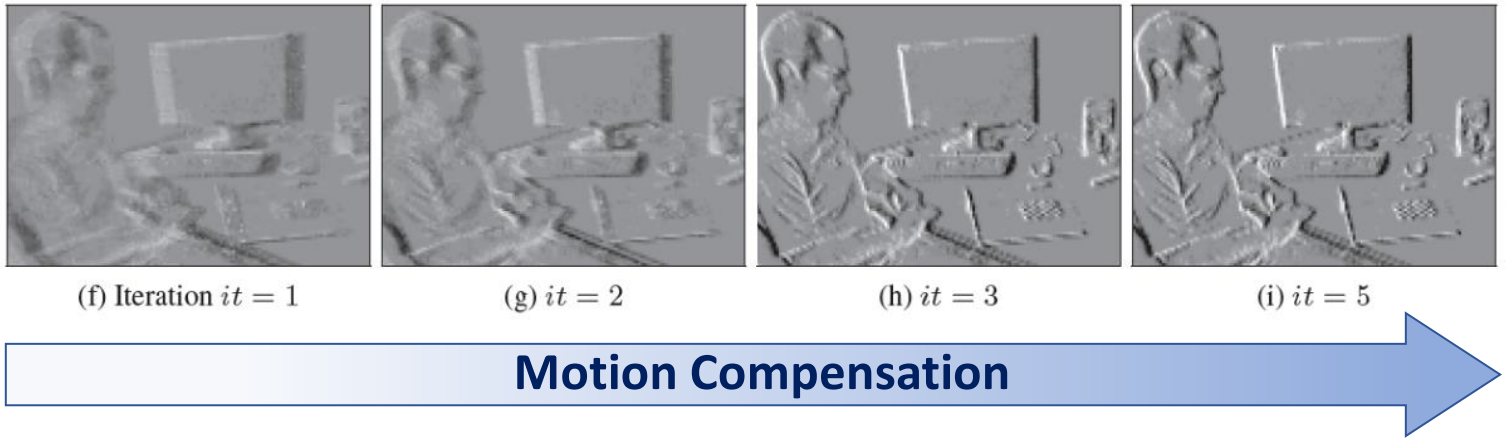


Zhu et al., CVPR 2019

- **3D histograms** of events.
  voxel = discretized space-time
- **Voting** ("insertion") **schemes**:
  - Nearest neighbor: each event votes for one cell only
  - Linear: each event splits its vote according to distance to neighboring voxels. Produces a smoother histogram.
- **Advantages**:
  - Preserves better the space-time structure of event data than 2D grid representations (such as event frames, time surfaces)
  - Compatible with conventional computer vision (CNNs, etc.)
- **Disadvantages**:
  - Memory (3D grid). Sparsity is lost: many grid values are zero.
  - Time is still quantized

Motion-Compensated Event Frames

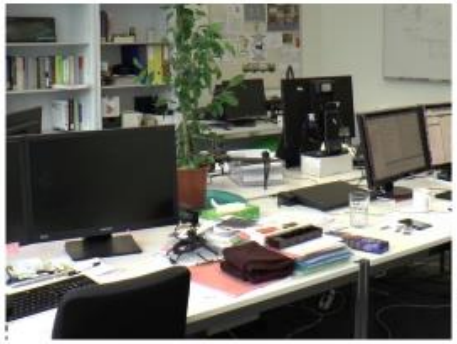
- It is a function of **events and** a candidate **motion field**



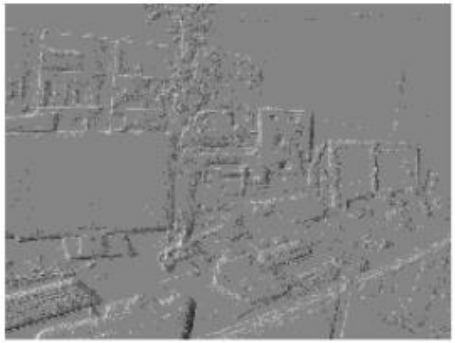
- **Advantages**:
  - Intuitive meaning: a sharp map of the edges causing the events
  - Can be used to estimate the motion field that best fit the events
  - Sharp images can be useful for later processing stages
- **Disadvantages**:
  - If motion is not given, an estimation algorithm must provide it
  - It is not fully motion invariant

Reconstructed Intensity Images

- Brightness images **reconstructed from events** can be interpreted as a more **motion-invariant** representation of the visual information contained in the events



Office scene



Events
(2D visualization)

Reconstructed
intensity image

Frame from DAVIS

- **Pros**:
  - **Compatible** with conventional computer vision
  - HDR, High-speed video recovered
- **Cons**:
  - Expensive to compute, latency, and may contain artefacts

Why so many different representations?

- Event data is **unconventional**. Cannot directly use the methods we have for standard cameras
- This is research. People **try new things** and see if they work (for their particular tasks and problem constraints)
- Representation may be **suggested by** constraints on method or platform utilization

Visual code

- Events **encode** visual information: they are a “code”
- There are other codes (e.g., provided by other sensors)

- Most representations shown are **data pre-processing**. Representations for higher levels of abstraction in visual processing can be built from events, for example, using hierarchical NN. The output of such networks is another “visual code”.

- The boundary between “representation” and **feature extraction** is fuzzy.

References

Reading:

- Section 3 of Gallego et al., Event-based Vision: A Survey, TPAMI 2020

- Gehrig et al., *End-to-End Learning of Representations for Asynchronous Event-Based Data*, ICCV 2019

- Rebecq et al., *High Speed and High Dynamic Range Video with an Event Camera*, TPAMI 2020