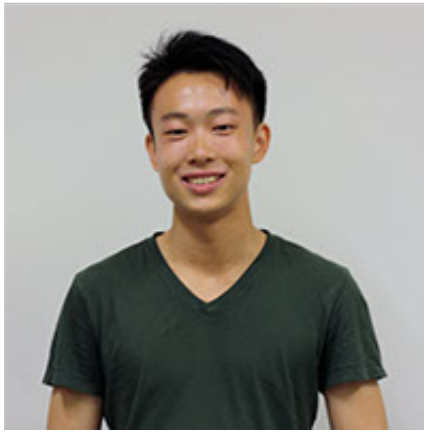
Event-based Feature Tracking and Visual Inertial Odometry



Kostas Daniilidis
with Alex Zhu and Nikolay Atanasov
University of Pennsylvania

Papers at ICRA 2017 and CVPR 2017

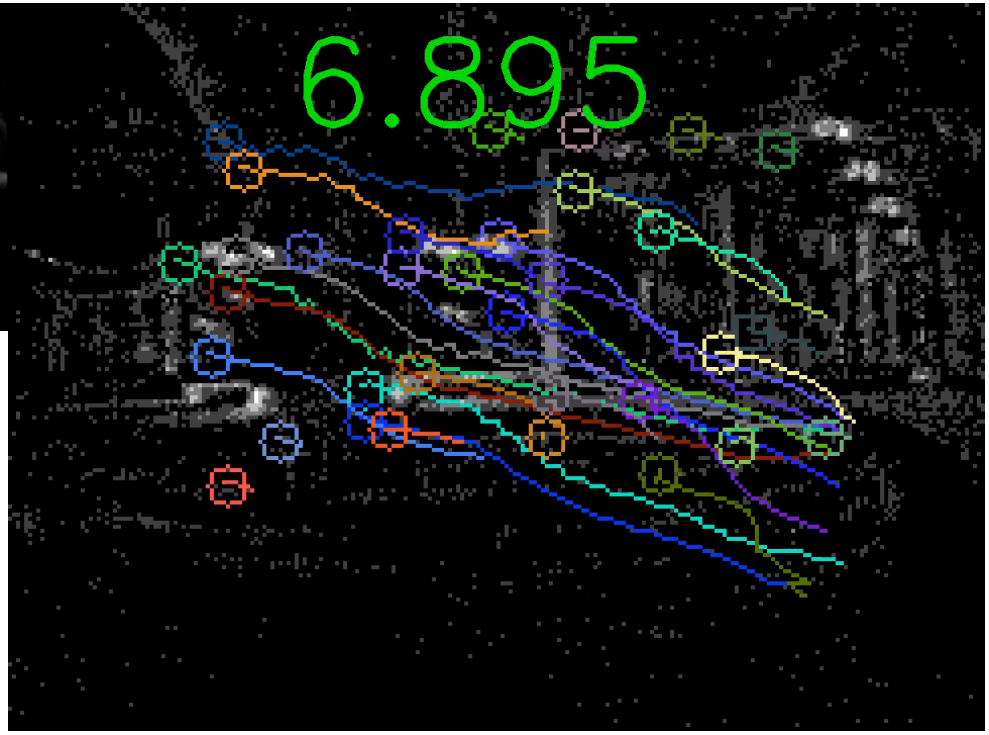
www.youtube.com/watch?v=m93XCqAS6Fc

www.youtube.com/watch?v=X3QIFj5Qc4A

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

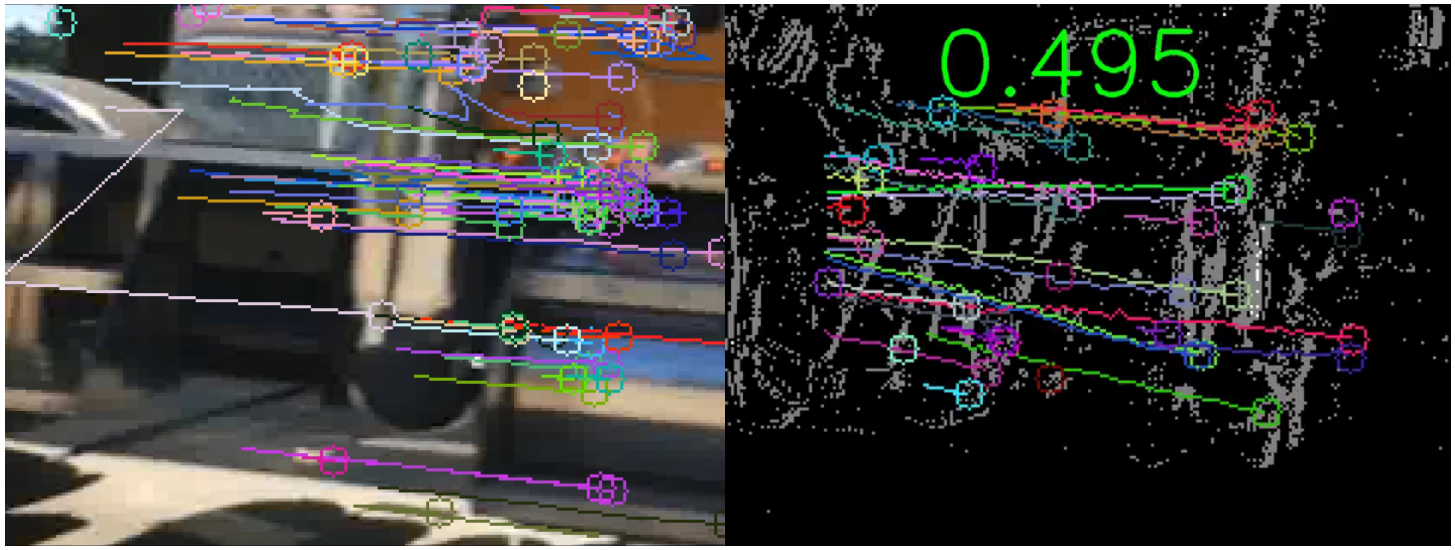
Night Scene, Very Low Lighting
0.1x Real Time

Corresponding grayscale image, 50ms exposure



Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Truck Passing 3m from the Camera at 60 miles/hr, 0.06x Realtime



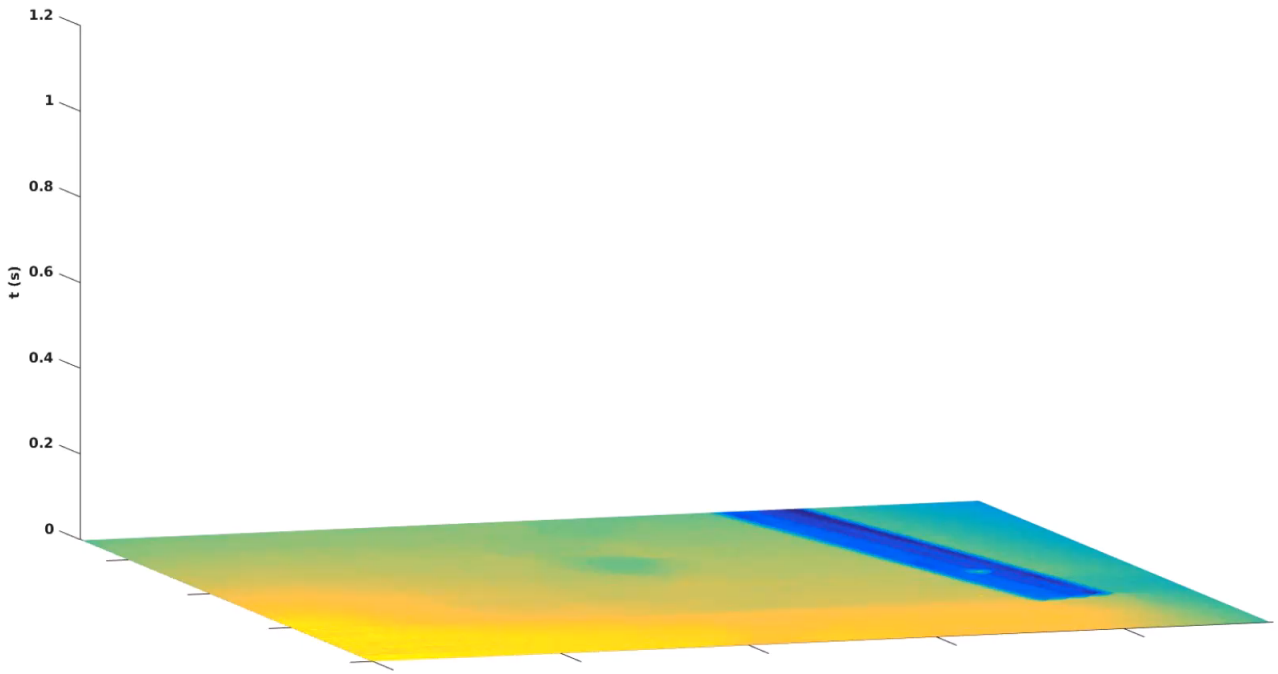
240FPS iPhone 6 image with KLT.

Event-based tracking on DAVIS 240C.

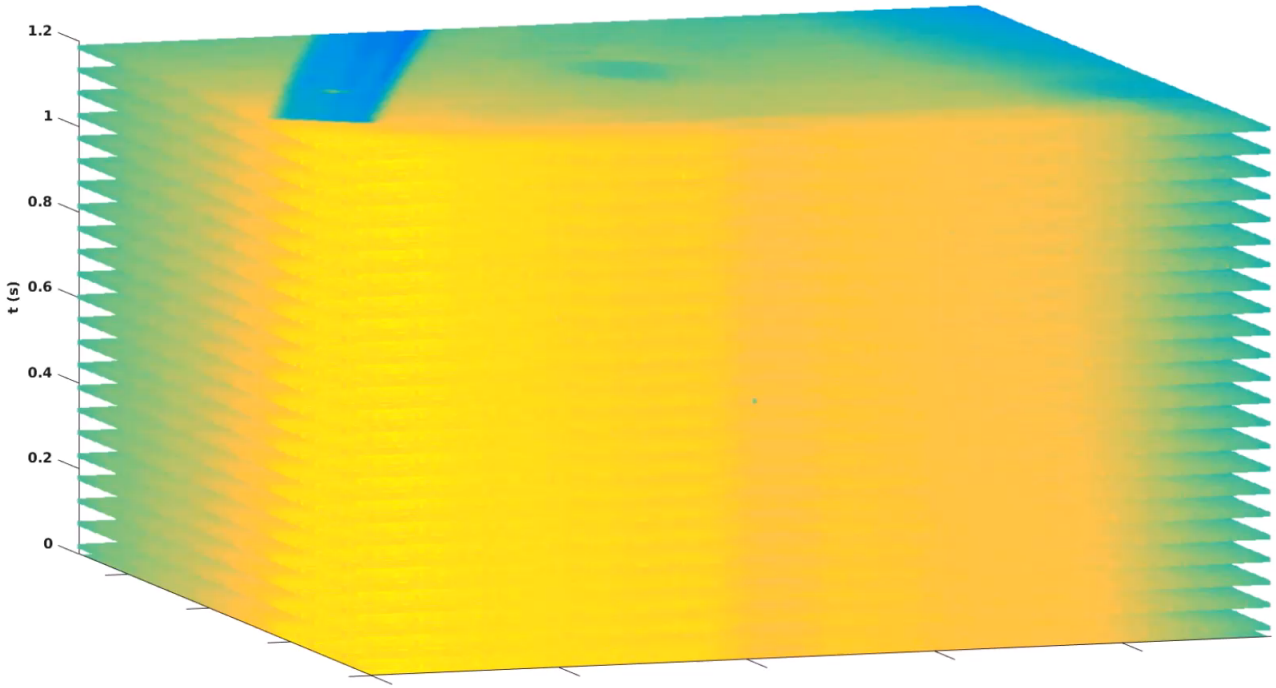
Optical flow is on the order of 5000 pixels/s.
Sequence is 600ms in realtime.

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Frame-based Cameras

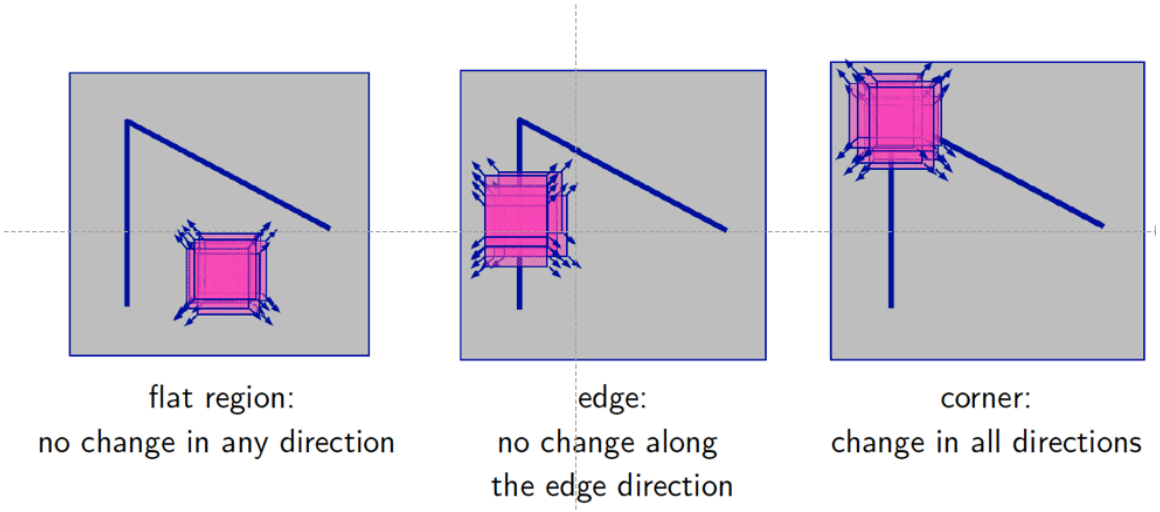


Event-based Cameras



Event-based cameras output asynchronous events (x, y, t, p) at <u>microsecond</u> resolution when $|\log(I(x,t_i)) - \log(I(x,t_{i-1})| \geq \theta$

What is a feature in classic vision?



Features are defined through motion: good flow means good features!

But they defined with a spatial neighborhood!

Speed is dealt with multiple scales

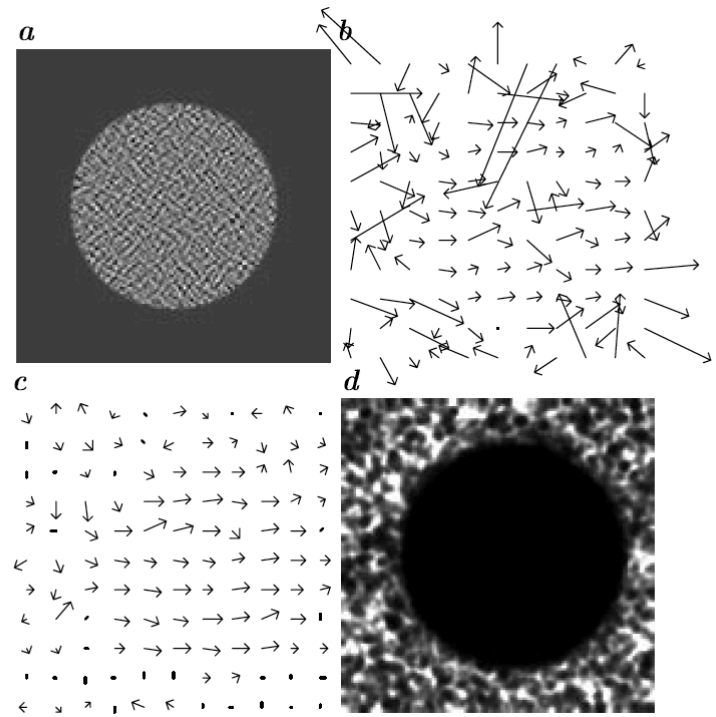
Published in: Handbook of Computer Vision and Applications, eds. B Jähne, H Haussecker, and P Geissler, volume 2, chapter 14, pages 297-422, Academic Press, Spring 1999.

Last modified: 28 Dec 1998.

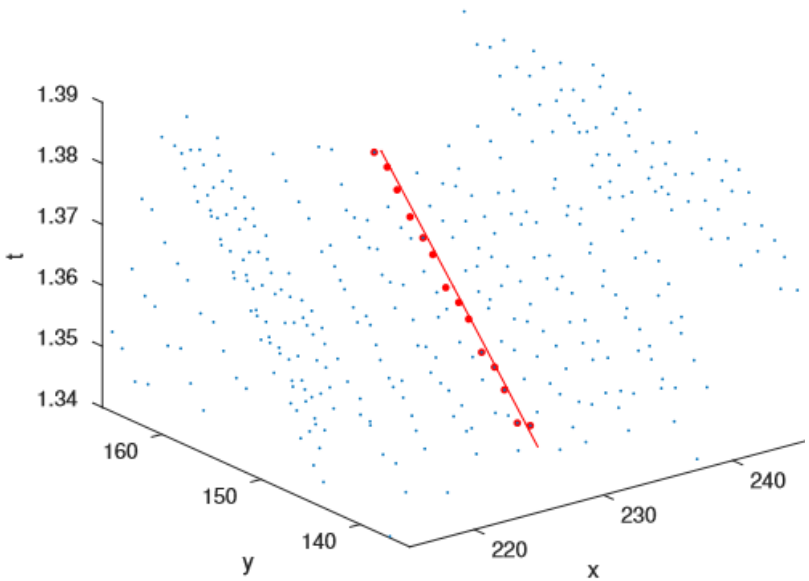
14 Bayesian Multi-Scale Differential Optical Flow

Eero P. Simoncelli

Center for Neural Science, and
Courant Institute of Mathematical Sciences
New York University

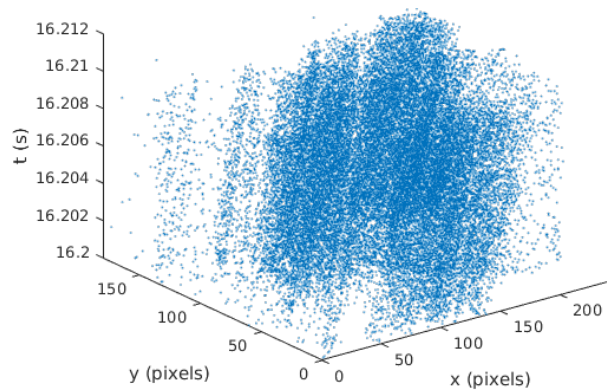


A feature is a set of 2D events induced by the same point in 3D.



$$\begin{pmatrix} f(t) \\ 1 \end{pmatrix} \sim K \begin{bmatrix} R(t) & T(t) \end{bmatrix} \begin{pmatrix} F \\ 1 \end{pmatrix}$$

A feature is a set of 2D noisy events induced by the same point in 3D.

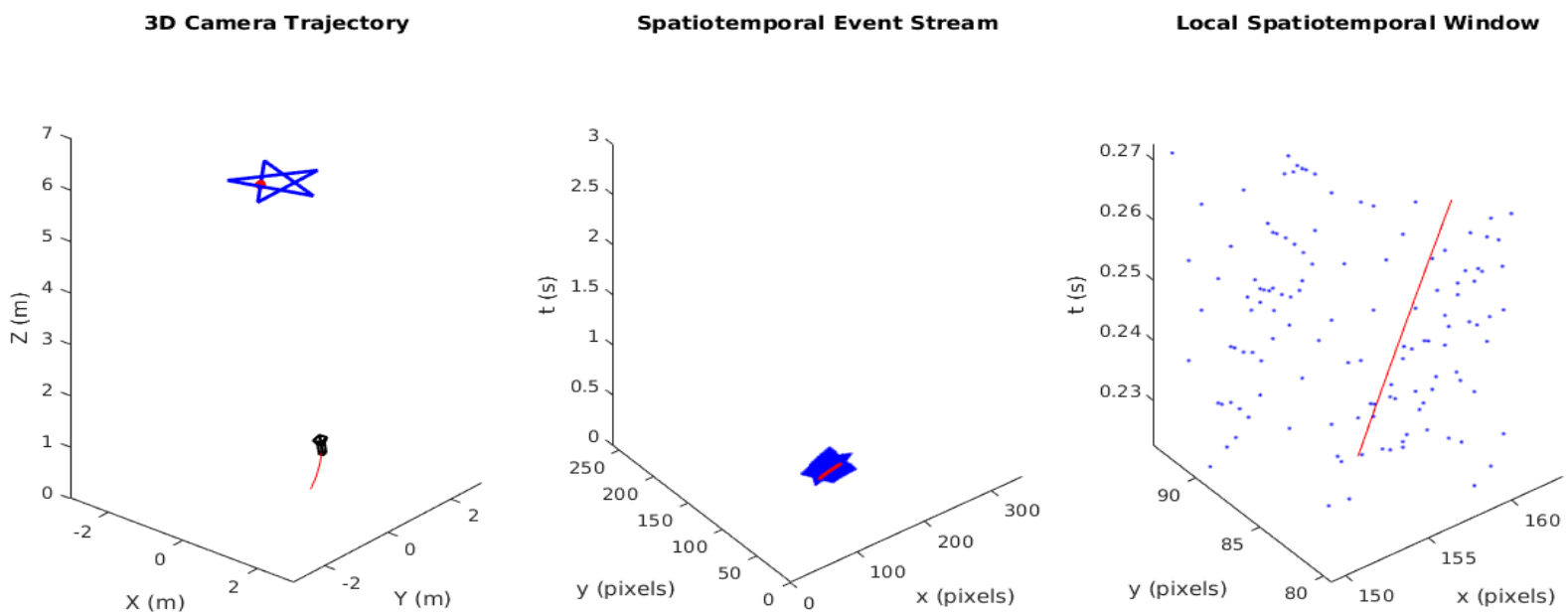


Our measurements are events $\{e_i := (x_i, t_i)\}_{i=1}^{n}$, where

$$x_i := p_{\pi(i)}(t_i) + \eta(t_i), \quad \eta(t_i) \sim \mathcal{N}(0, \Sigma), \quad \forall i$$

$\pi : \{1, \ldots, n\} \to \{1, \ldots, m\}$ is an unknown many-to-one function representing the *data association* between the events $\{e_i\}$ and projections $\{p_j\}$ that generate them.

A feature is a set of events with same flow



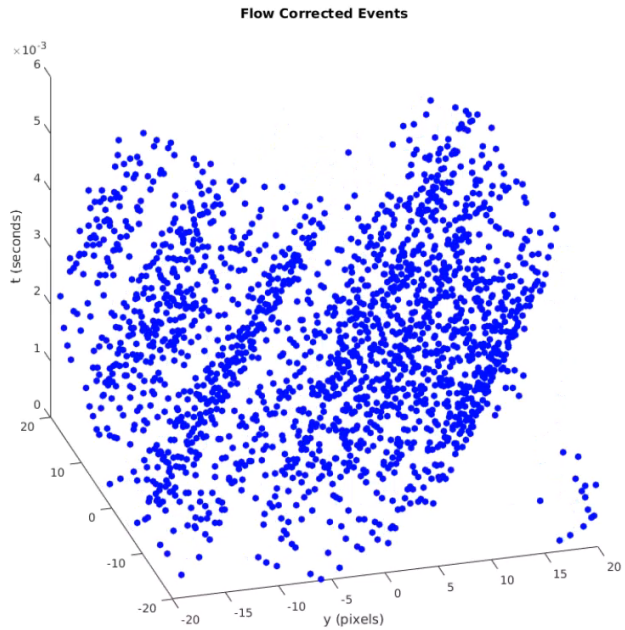
$$\|(x_i - t_i v) - (x_k - t_k v)\|^2 \mathbb{1}_{\{\pi(i)=\pi(k)=j\}} = 0, \quad \forall\, i, k \in [n]$$

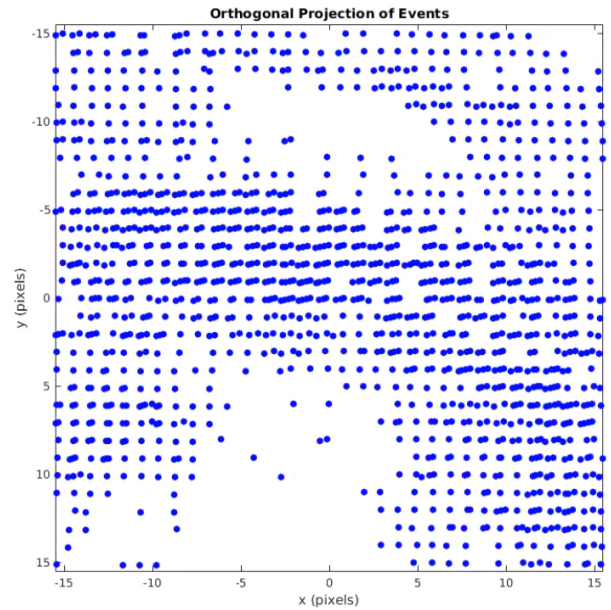
Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

But we do not know the association so we will take the expectation

$$\mathbb{E}_{\pi(i),\pi(k)}\|(x_i - t_i v) - (x_k - t_k v)\|^2 \mathbb{1}_{\{\pi(i)=\pi(k)=j\}}$$

Optical Flow Estimation





Data association probability

Propagated events through time

$$\min_{r,v} \sum_{i=1}^{n} \sum_{k=1}^{n} \left[\sum_{j=1}^{m} r_{ij} r_{kj} \right] ||(x_i - t_i v) - (x_k - t_k v)||^2$$

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

E-Step

$$r_{ij}(\{p_j\}) := \frac{\phi((x_i - t_i v); p_j, \Sigma)}{\sum_{l=1}^{m} \phi((x_i - t_i v); p_l, \Sigma)}$$

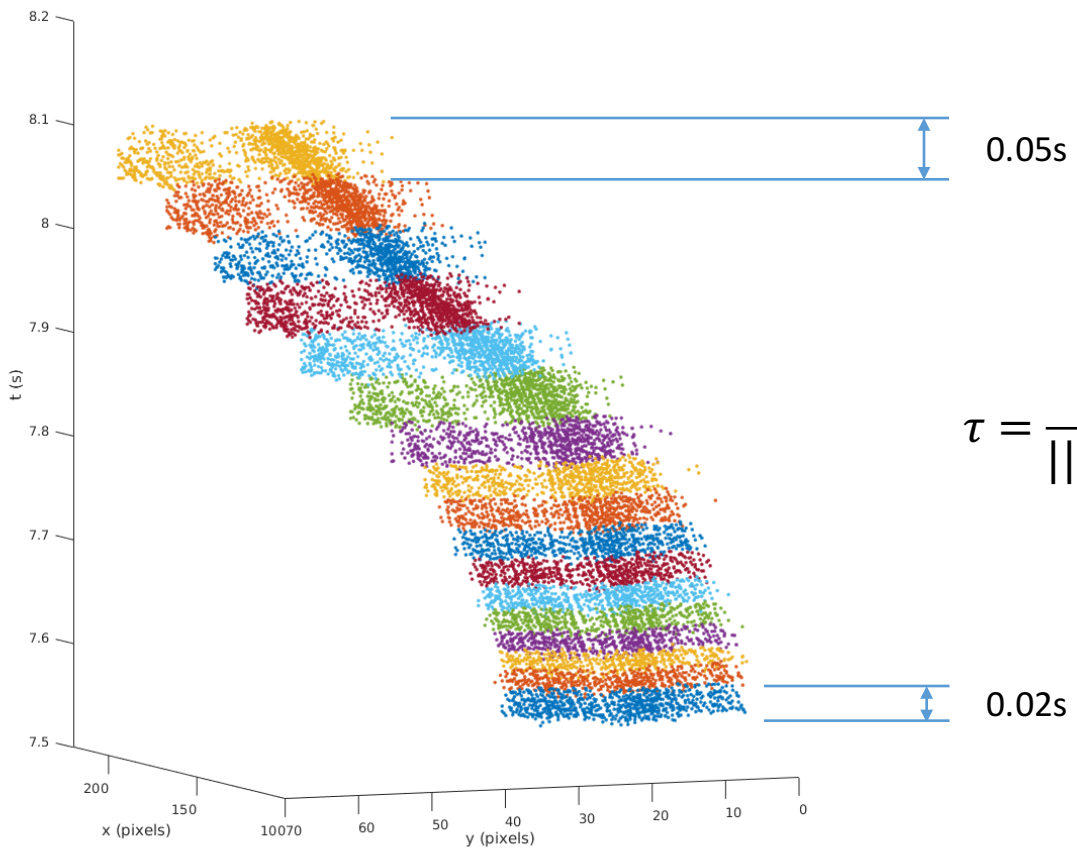
M-Step (linear least squares)

$$\min_{v} \sum_{i=1}^{n} \sum_{k=1}^{n} \left[\sum_{j=1}^{m} r_{ij} r_{kj} \right] \|(x_i - t_i v) - (x_k - t_k v)\|^2$$

How long temporal window?

0.5ms window | 2ms window | 5ms window

How do we choose the right temporal window?



$$\tau = \frac{k}{||v||_2}$$

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Over longer time: Monitor quality of feature with an affine motion model!

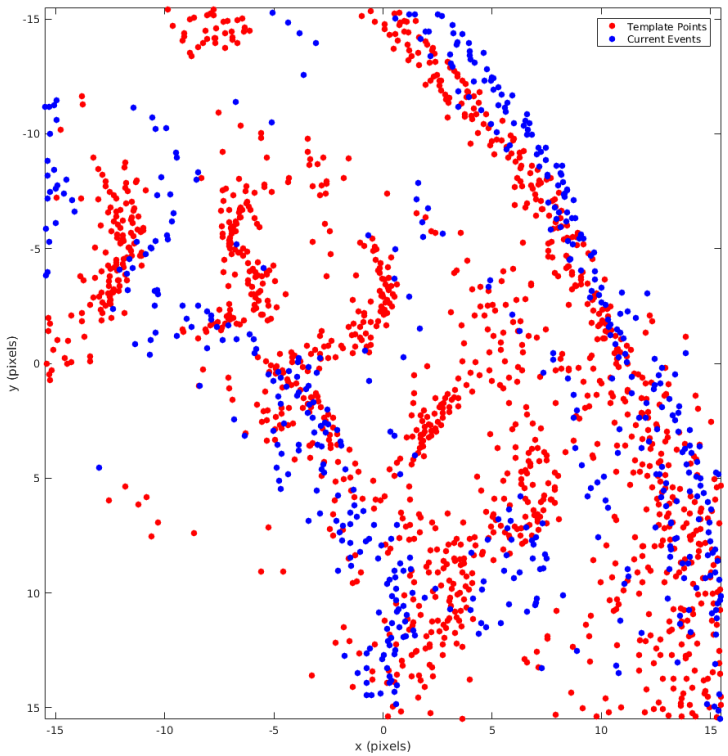
IEEE Conference on Computer
Vision and Pattern Recognition
(CVPR94) Seattle, June 1994

Good Features to Track

Jianbo Shi
Computer Science Department
Cornell University
Ithaca, NY 14853

Carlo Tomasi
Computer Science Department
Stanford University
Stanford, CA 94305

Drift Correction - Stabilization



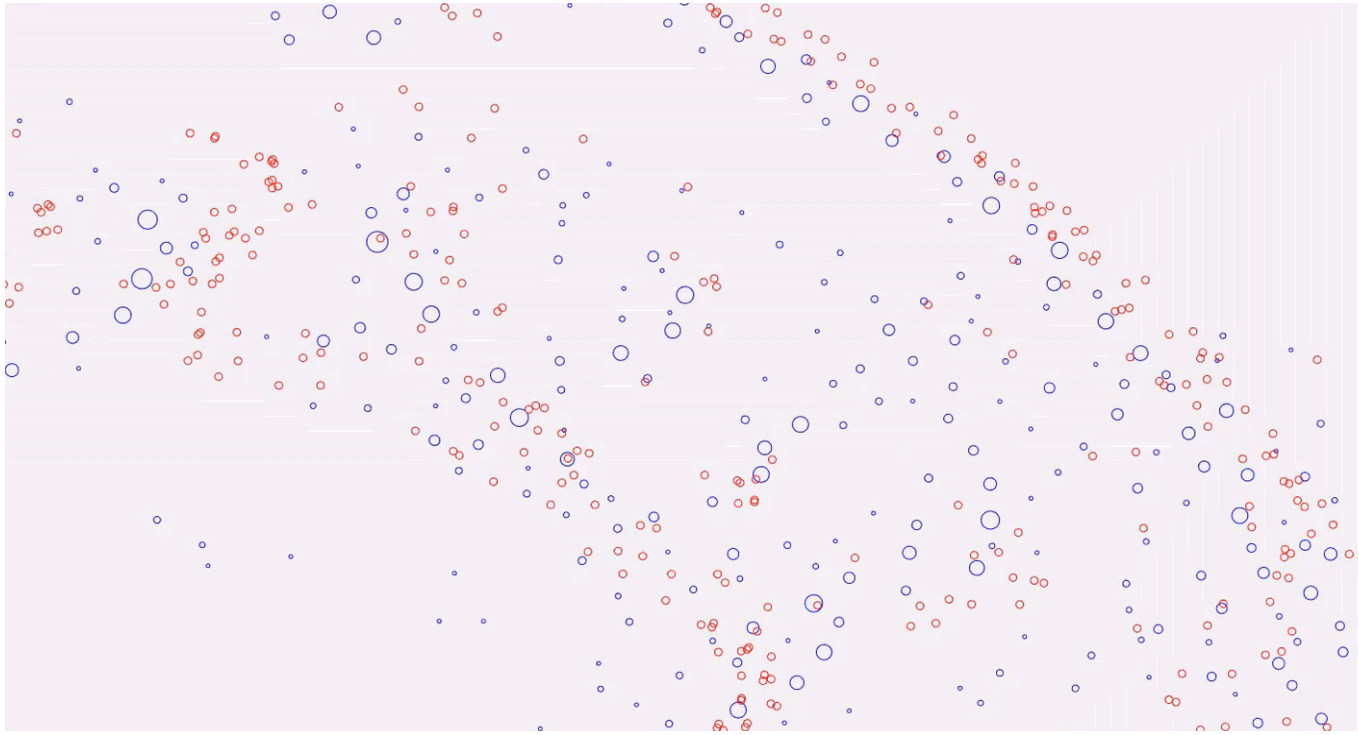
Warped propagated events

Data association

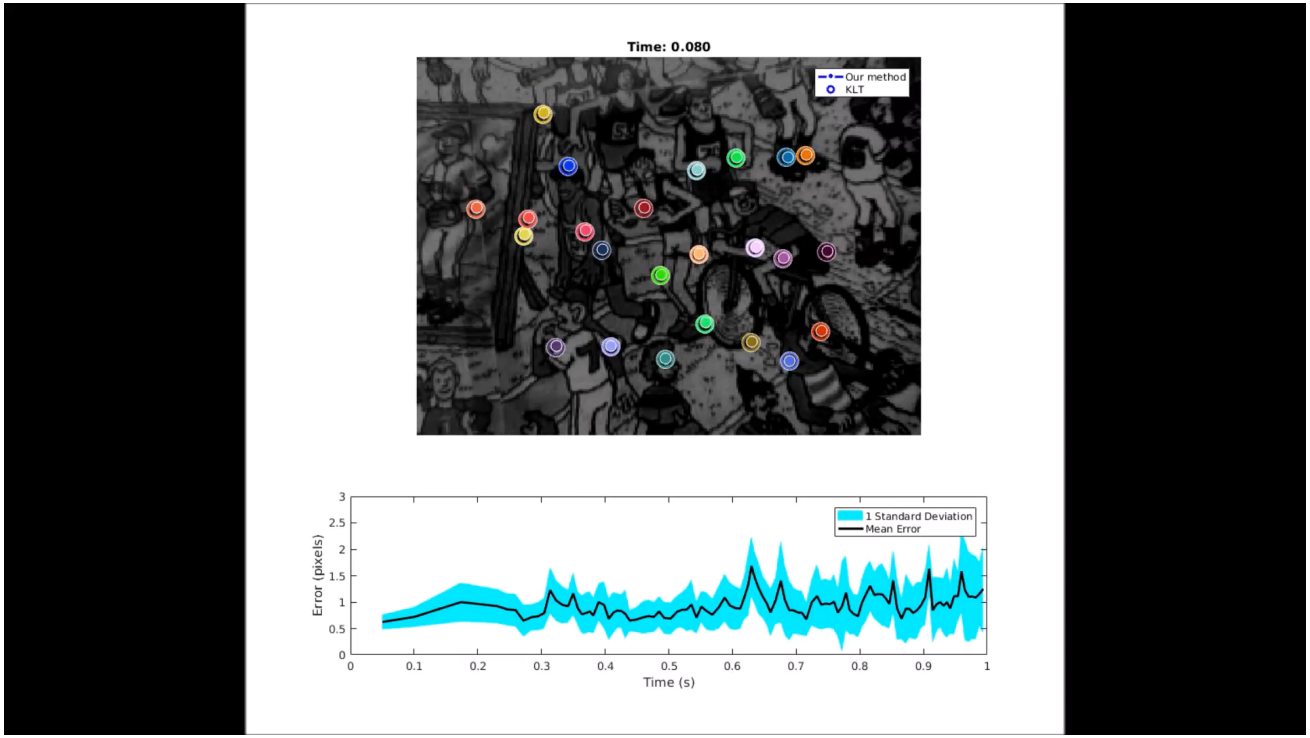
$$\min_{A,b,r} \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \left| A(x_i - t_i v) + b - p_j \right|^2$$

Template points

Scope of each feature



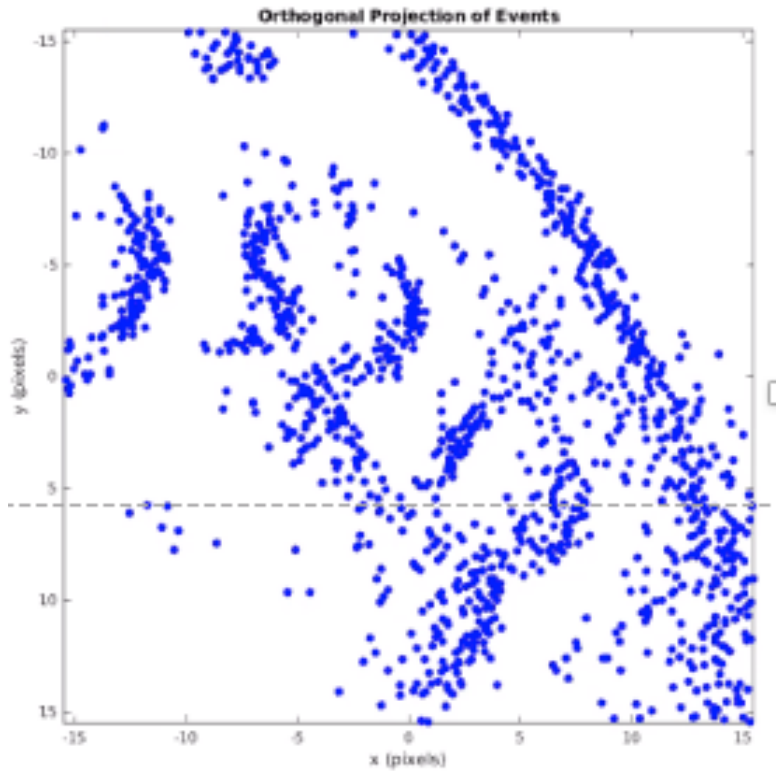
Results: KLT Comparison



20 windows initialized, no reacquisition if tracks are discarded. Frame based images from DAVIS.

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Sparsify and deal with aperture problem: FAST corner selection in the aggregation of warped images



Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Visual Inertial Odometry

- Given the event-based feature tracks and a set of IMU observations, how do we obtain an accurate estimate of the camera pose?
- MSCKF (Roumeliotis' group)
- Enforce 3D rotation in 2D tracking

Instead of affine warp…

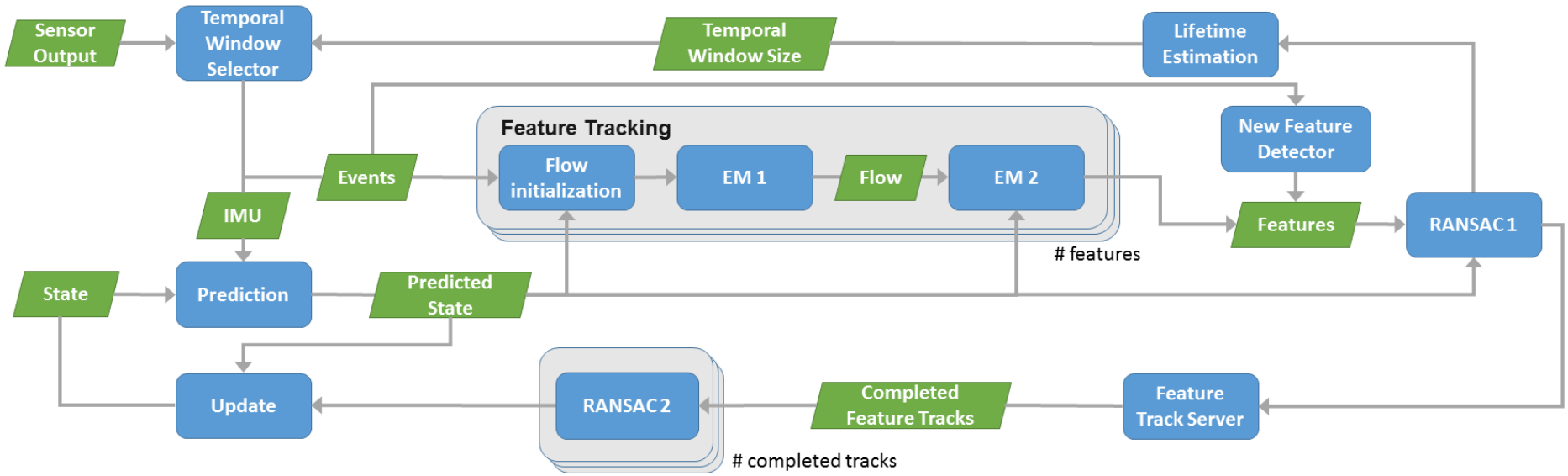
We use the current estimate of rotation and we estimate only scale and local translation

$$y_k^i = \pi\left({}^{i*}R_i \begin{pmatrix} l_k^i \\ 1 \end{pmatrix}\right) - \pi\left({}^{i*}R_i \begin{pmatrix} f(T_i) + u_i dt_i \\ 1 \end{pmatrix}\right)$$

Outlier Rejection

- The EKF uses the L2 loss, and so is very susceptible to outliers in the measurements. To remove these outliers, we apply two RANSAC steps during the tracking.
- RANSAC 1: Pure Translation
  - After each temporal window, two point RANSAC is applied given the rotation estimated from the IMU to reject failed trackers.
- RANSAC 2: Triangulation over frames
  - As each feature track is residualized, a second RANSAC step is applied to find the largest inlier set that agrees on a 3D pose of the feature, given the observations and their corresponding camera poses.

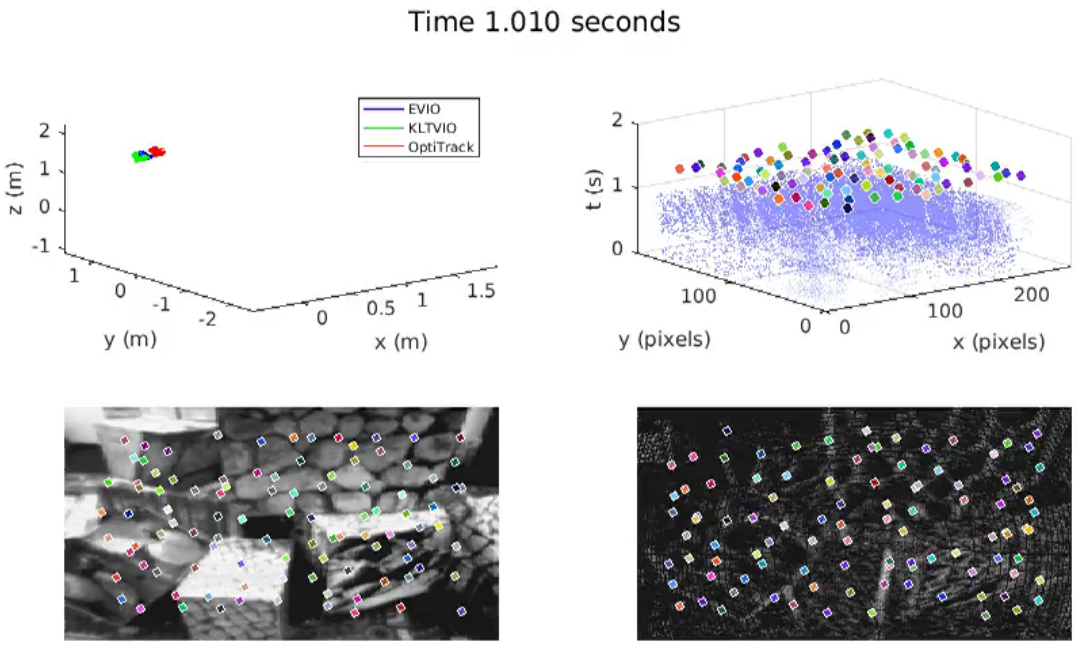
EVIO Summary



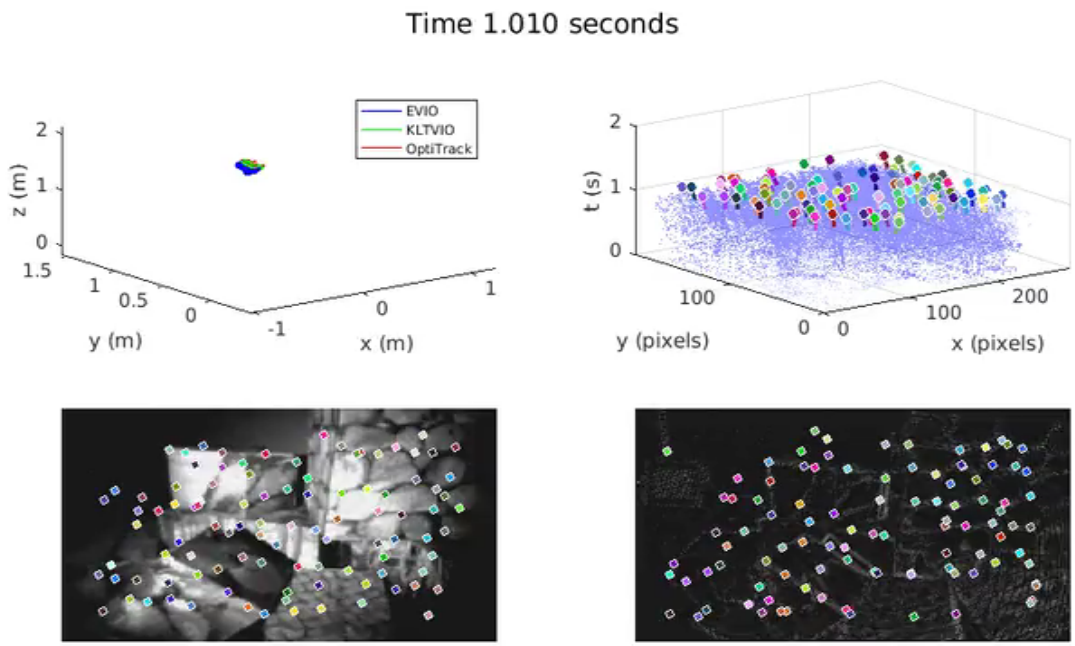
Zhu A., Atanasov N., and Daniilidis D. "Event-based Visual Inertial Odometry", *Submitted to Computer Vision and Pattern Recognition (CVPR) 2017.*

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

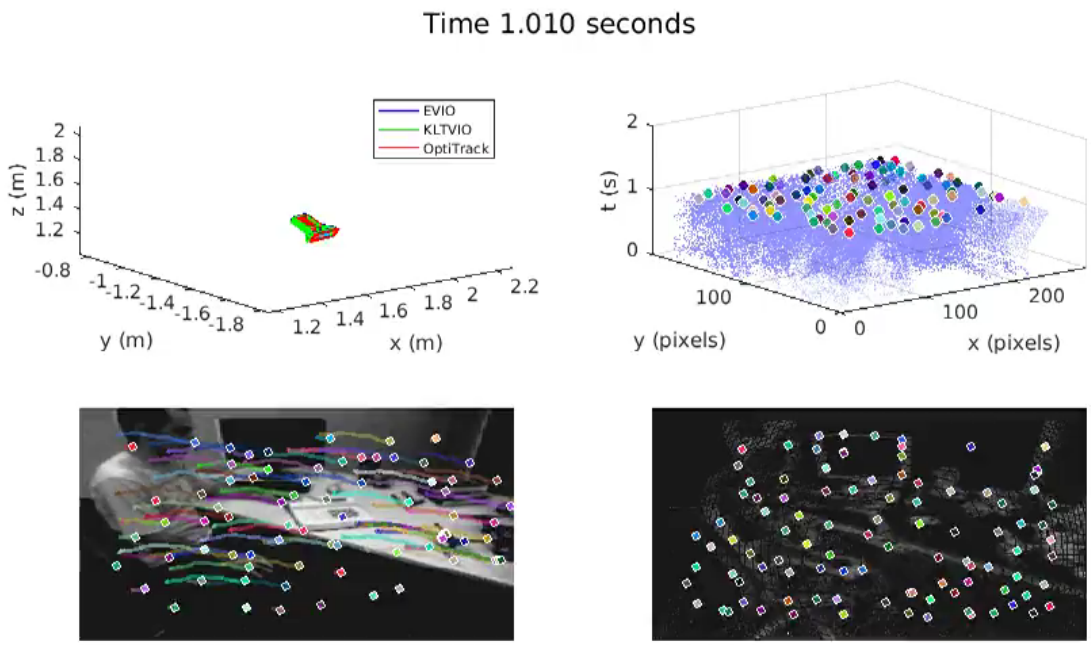
Results: General Scene

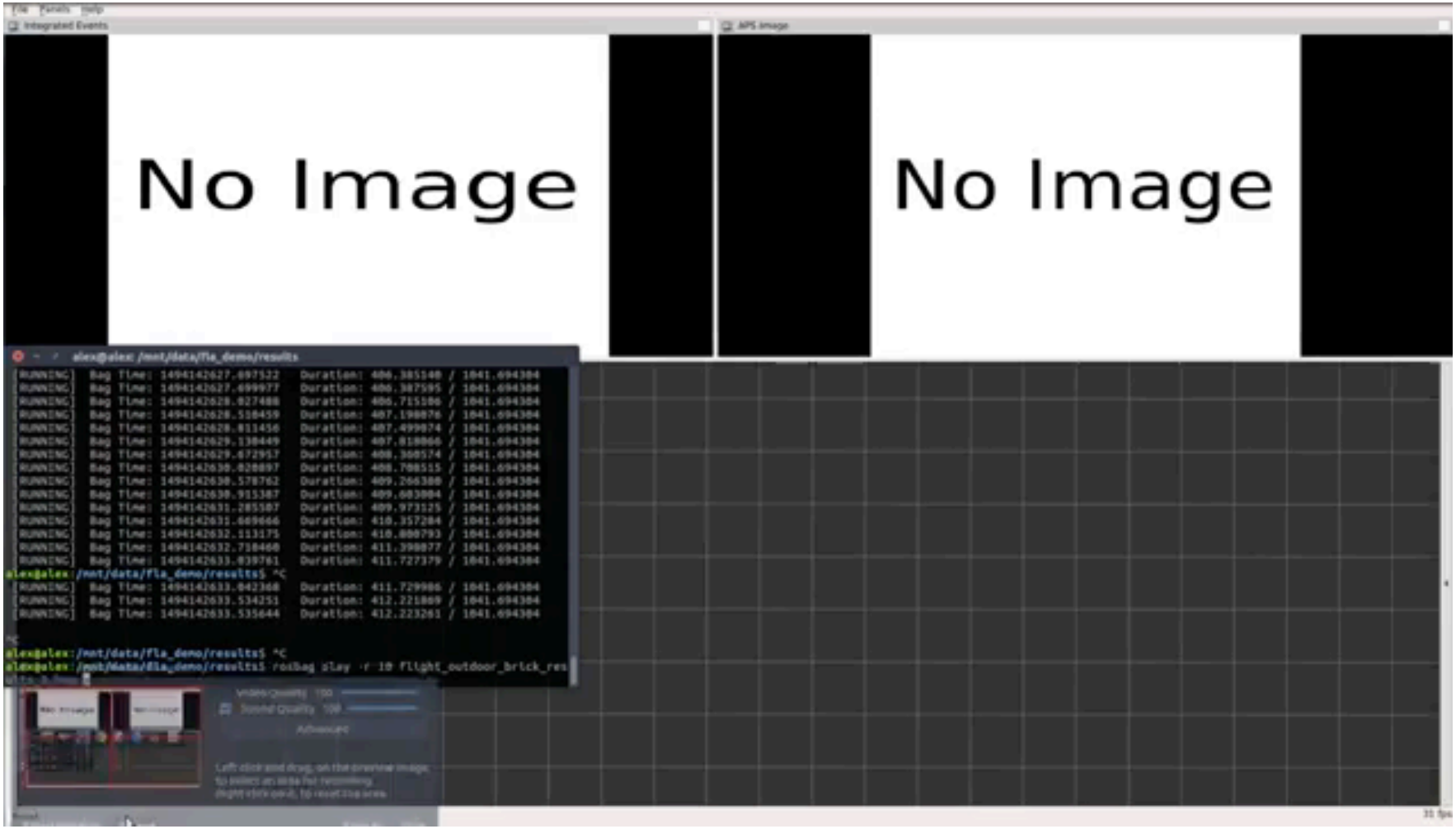


Results: HDR Scene

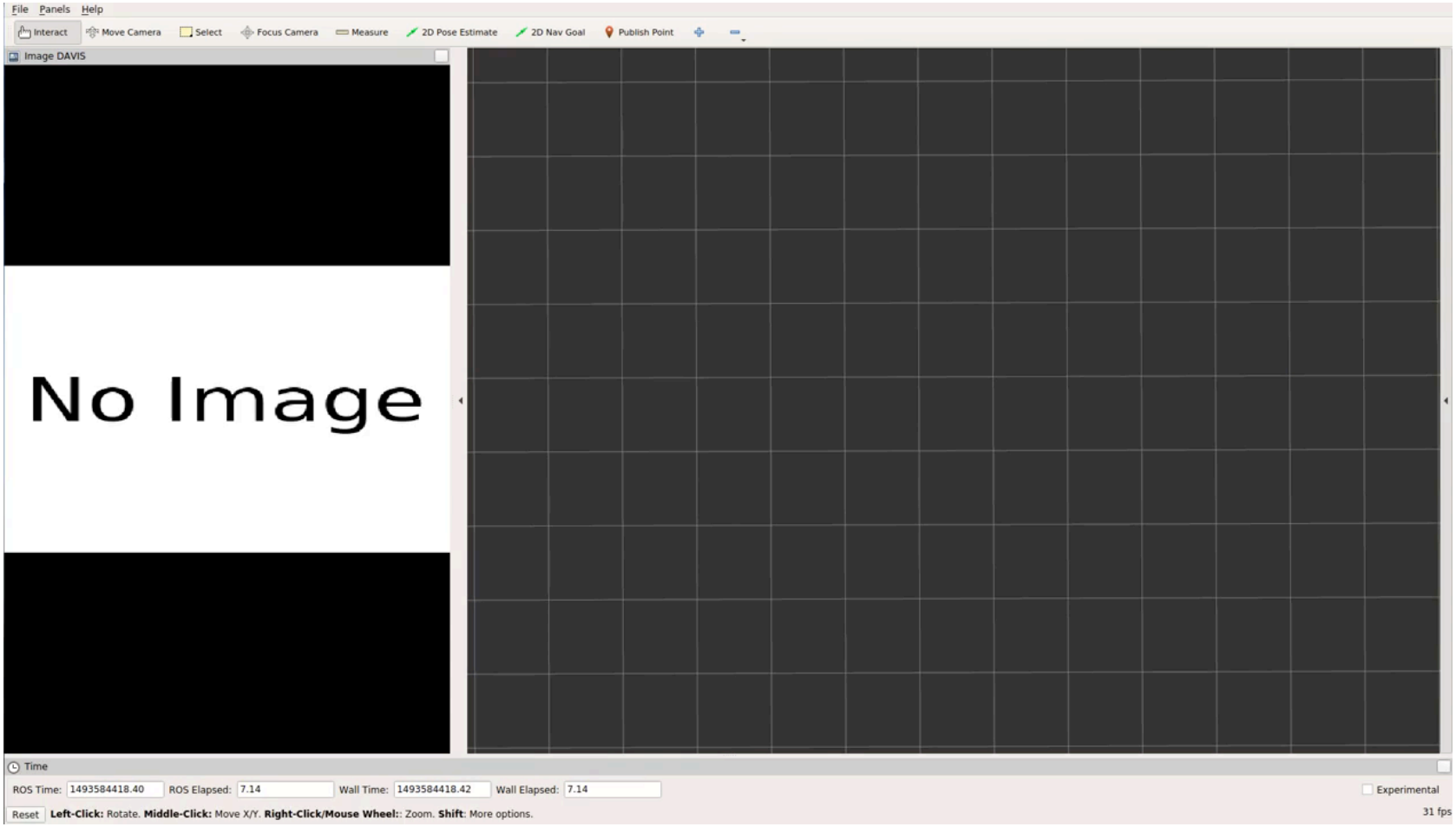


Results: Motion Independent of Camera





Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab



Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

The future of robot vision is event-based!