ROBOTICS & PERCEPTION GROUP

University of Zurich[UZH]

Towards Agile Flight of Vision-controlled Drones:

From Active Perception to Event-based Vision

Davide Scaramuzza
http://rpg.ifi.uzh.ch

Research Background

Computer Vision

- Visual Odometry and SLAM
- Sensor fusion
- Camera calibration

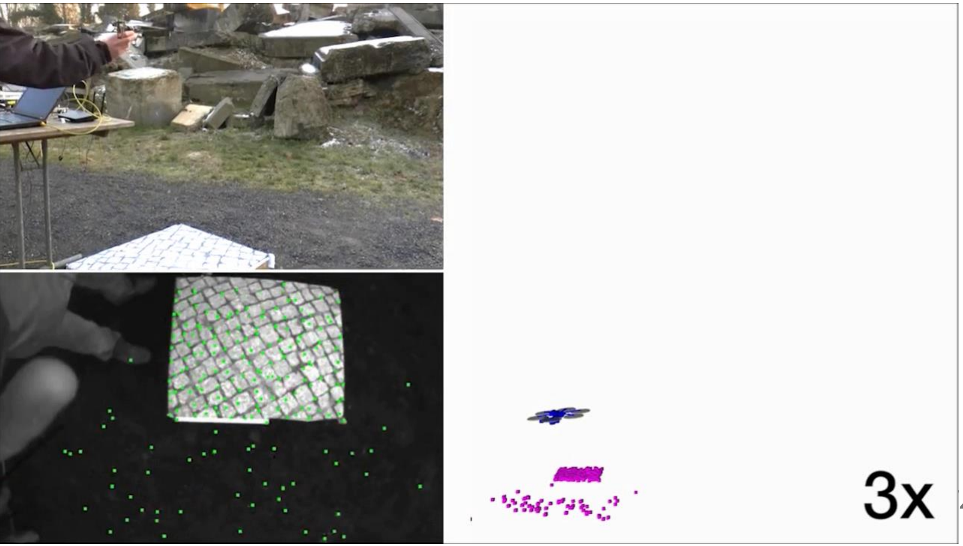
Autonomous Robot Navigation

- Self driving cars
- Micro Flying Robots

[IROS’06, PAMI’13]

[ICCV’09, CVPR’10, IJCV’11]

[JFR’10, AURO’11, RAM’14, JFR’15]



zh.ch

My Vision: Flying Robots to the Rescue!

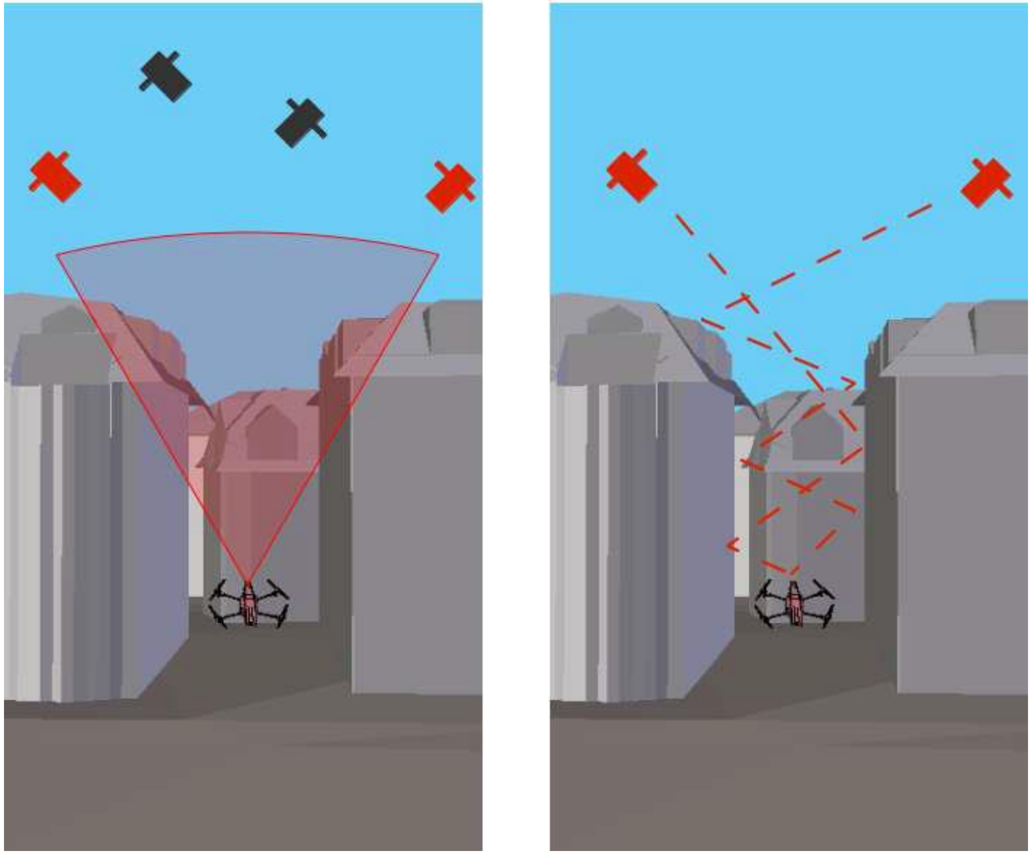
How to fly a drone

- **Remote control**
  - Requires line of sight or communication link
  - Requires skilled pilots

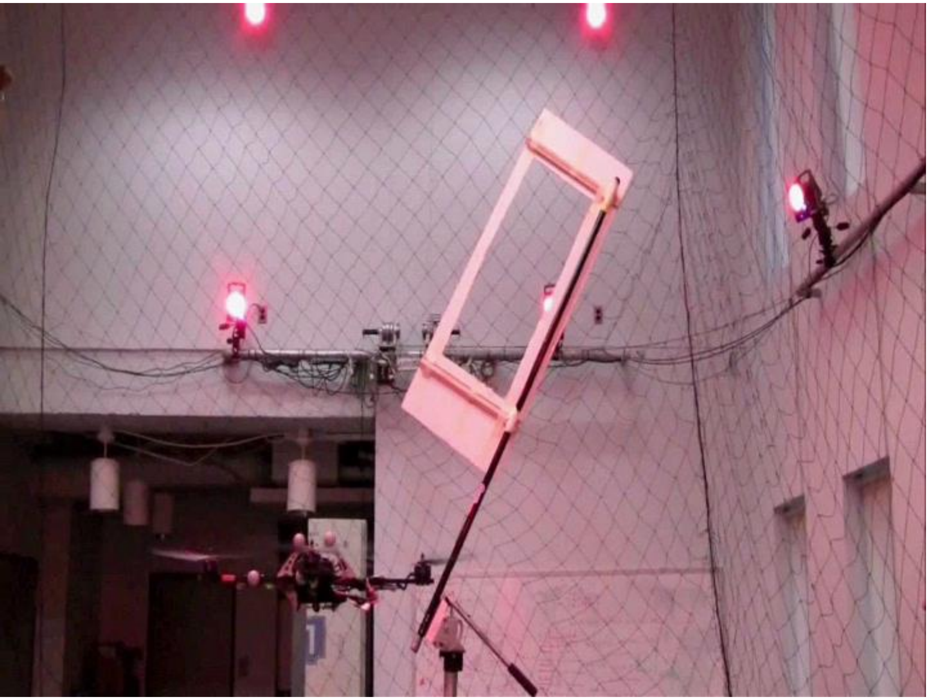
Drone crash during soccer match, Brasilia, 2013

How to fly a drone

- **GPS-based navigation**
  - Doesn't work indoors
  - Can be unreliable outdoors



How do we Localize without GPS ?



Mellinger, Michael, **Kumar**

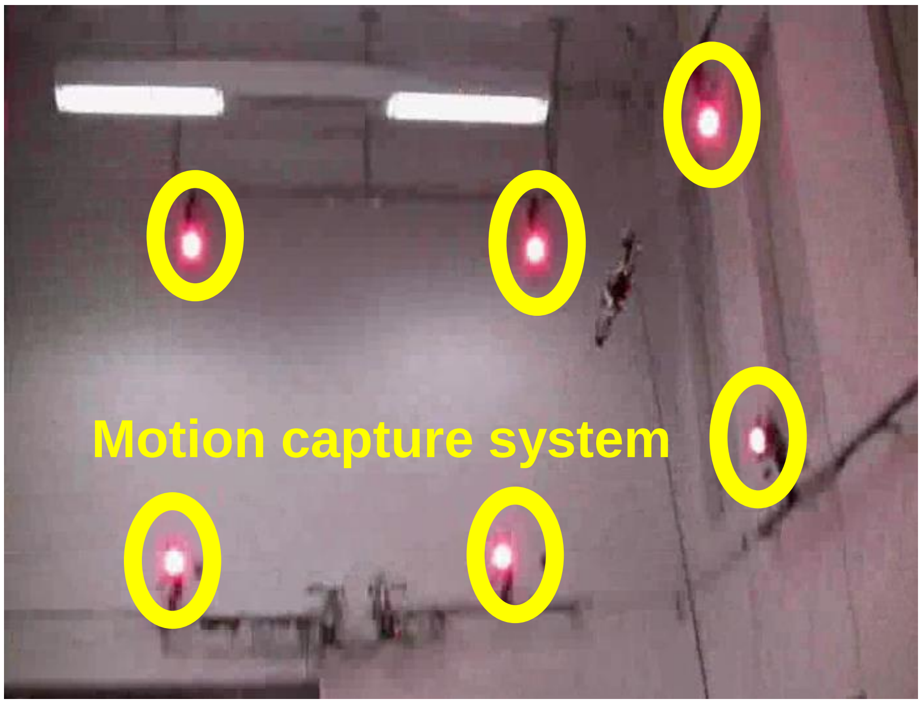
Penn University of Pennsylvania GRASP Laboratory

Fontana, Faessler, **Scaramuzza**

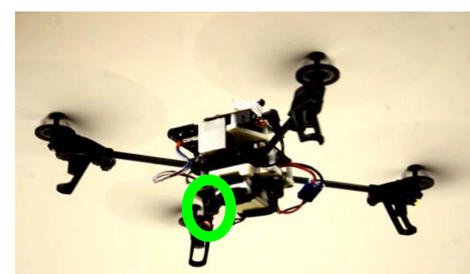
University of Zurich UZH

How do we Localize without GPS ?

This robot is «***blind***»



This robot can «***see***»



Problems with Vision-controlled Drones

Drones have the **potential to navigate quickly** through unstructured environments but

- Autonomous operation is currently **restricted to controlled environments**
- **Vision-based** maneuvers still **slow** and **inaccurate** wrt motion-capture systems

Why?

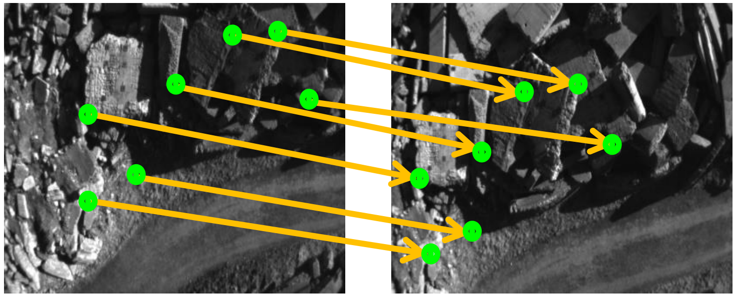
- Perception algorithms are **mature but not robust**
  - Unlike mocap systems, **localization accuracy** depends on **distance & texture**!
  - **Control & perception** have been mostly **considered separately!**
  - Algorithms and sensors have **big latencies** (50-200 ms) → need faster sensors!

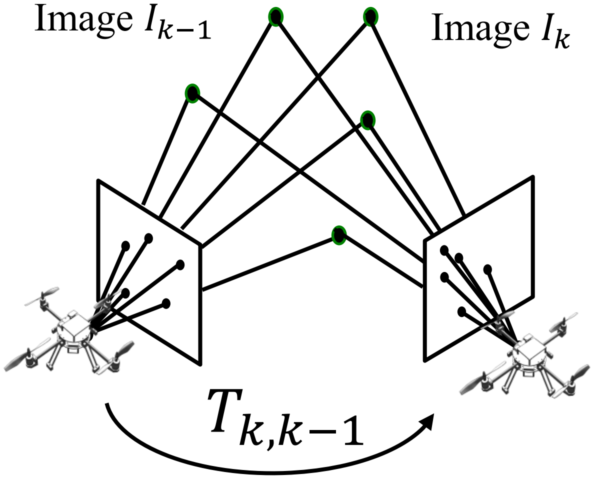
Outline

- Visual-Inertial State Estimation
- Active Vision
- Low-latency, Agile Flight

Visual-Inertial State Estimation

Working Principle: Structure from Motion





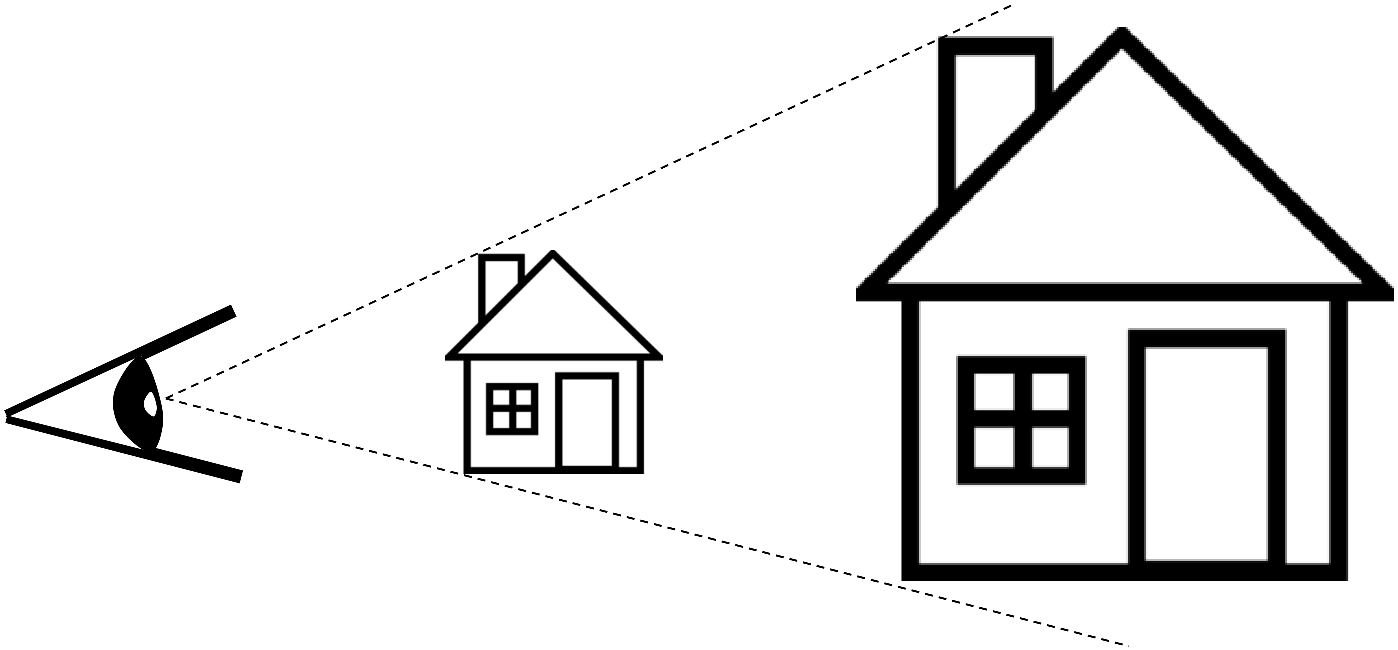
$$\mathbf{T}_{k,k-1} = \arg\min_{\mathbf{T}} \iint_{\bar{\mathcal{R}}} \rho\, I_k\Big(\pi\big(\mathbf{T}\cdot\pi^{-1}(\mathbf{u}, d_{\mathbf{u}})\big)\Big) - I_{k-1}(\mathbf{u})\, d\mathbf{u}$$

Several open source tools are available:
PTAM; OKVIS; LIBVISO; ORBSLAM; LSD-SLAM; **SVO**

Scaramuzza, Fraundorfer. Visual Odometry Tutorial, IEEE Robotics and Automation Magazine, 2011

Scale Ambiguity

- With a single camera, we only know the relative scale
- No information about the *metric scale*



Absolute Scale Determination

- The absolute pose x is known up to a scale s, thus

$$x = s\tilde{x}$$

- IMU provides accelerations, thus

$$v = v_0 + \int a(t)dt$$

- By derivating the first one and equating them

$$s\dot{\tilde{x}} = v_0 + \int a(t)dt$$

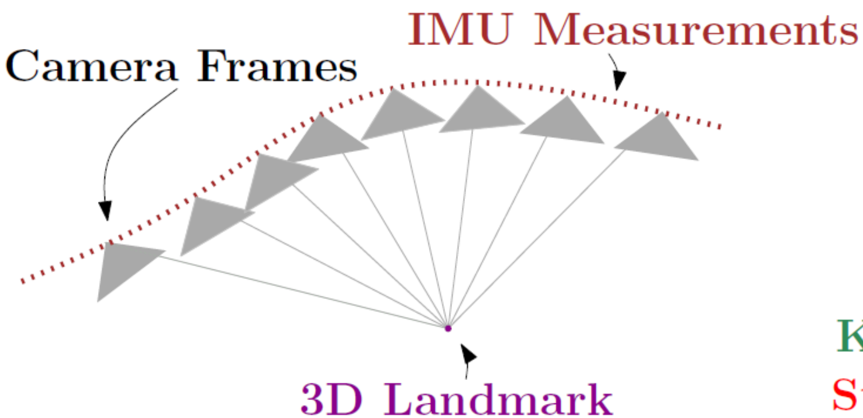
- As shown in [Martinelli, TRO'12], for 6DOF, both s and v_0 can be determined in closed form from a **single feature observation and 3 views**
- This is used to initialize the asbolute scale [Kaiser, ICRA'16]
- The scale can then be tracked with
  - EKF [Mourikis & Mourikis, IJRR'10], [Weiss, JFR'13]
  - or non-linear optimization methods [Leutenegger, RSS'13] [Forster, RSS'15]

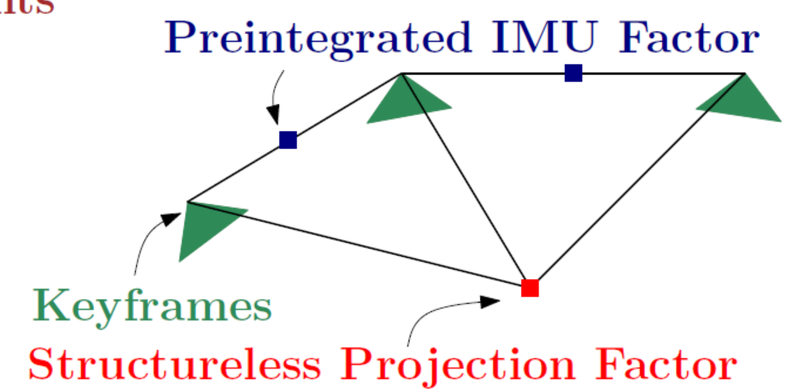
Martinelli, "Vision and IMU Data Fusion: Closed-Form Solutions for Attitude, Speed, Absolute Scale, and Bias Determination“, IEEE Transaction on Robotics, 2012

J. Kaiser, A Martinelli, F. Fontana, D. Scaramuzza, Simultaneous State Initialization and Gyroscope Bias Calibration in Visual Inertial aided Navigation, IEEE RA-L'16

Visual-Inertial Fusion [RSS'15]

- Fusion solved as a *non-linear optimization problem*
- Increased accuracy over filtering methods





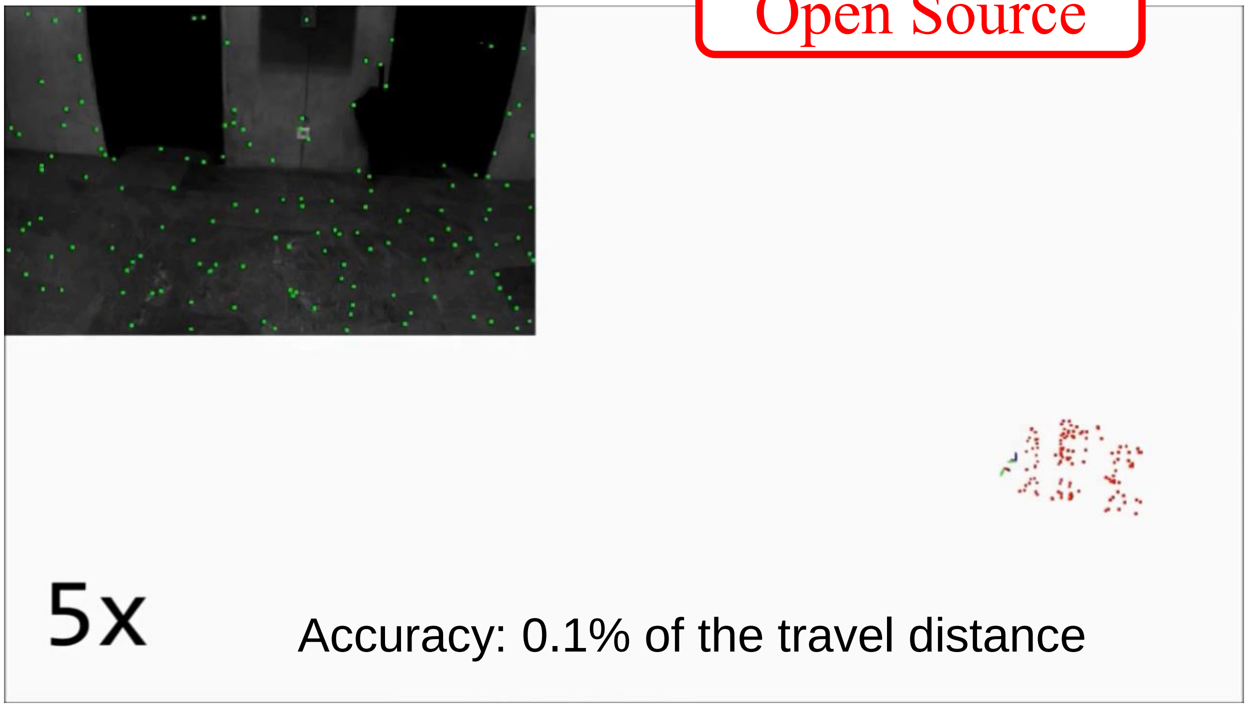
$$\sum_{(i,j)\in\mathcal{K}_k} \|\mathbf{r}_{\mathcal{I}_{ij}}\|^2_{\Sigma_{ij}} + \sum_{i\in\mathcal{K}_k} \sum_{l\in\mathcal{C}_i} \|\mathbf{r}_{\mathcal{C}_{il}}\|^2_{\Sigma_{\mathcal{C}}}$$

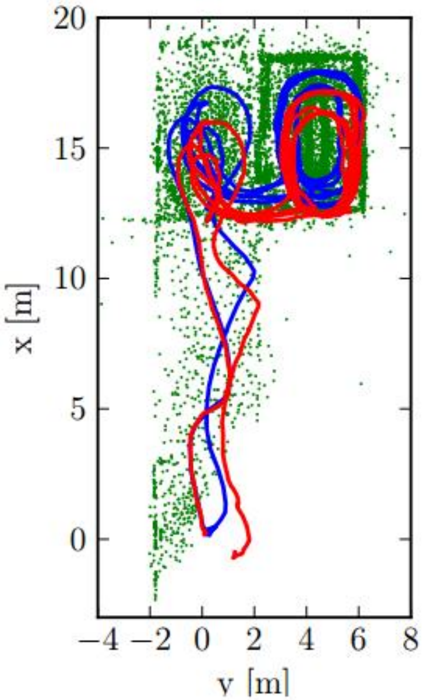
IMU residuals *Reprojection residuals*

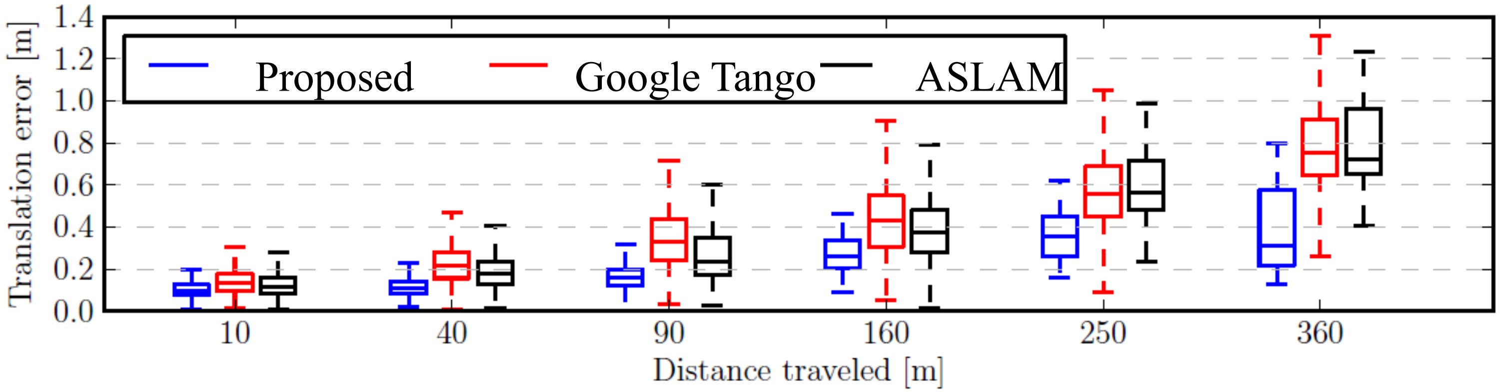
Forster, Carlone, Dellaert, Scaramuzza, IMU Preintegration on Manifold for efficient Visual-Inertial Maximum-a-Posteriori Estimation, *Robotics Science and Systens*'15, **Best Paper Award Finalist**

Comparison with Previous Works

Open Source







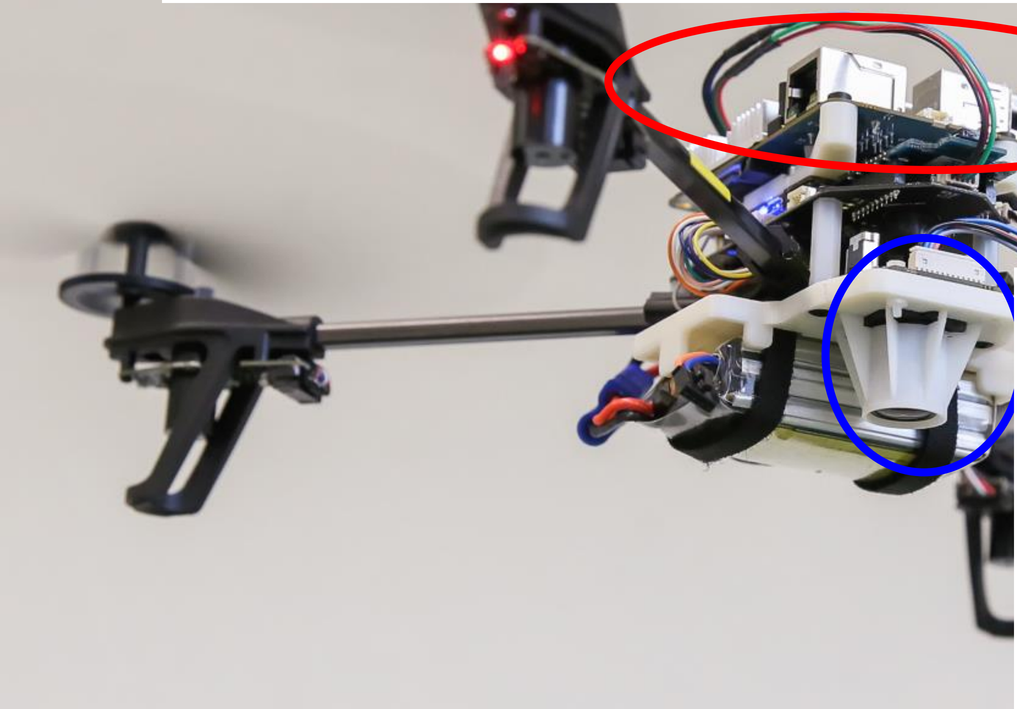
Forster, Carlone, Dellaert, Scaramuzza, IMU Preintegration on Manifold for efficient Visual-Inertial Maximum-a-Posteriori Estimation, *Robotics Science and Systens*'15, **Best Paper Award Finalist**

Integration on a Quadrotor Platform

Quadrotor System

Odroid U3 Computer

- Quad Core Odroid (ARM Cortex A-9) used in Samsung Galaxy S4 phones
- Runs Linux Ubuntu and ROS

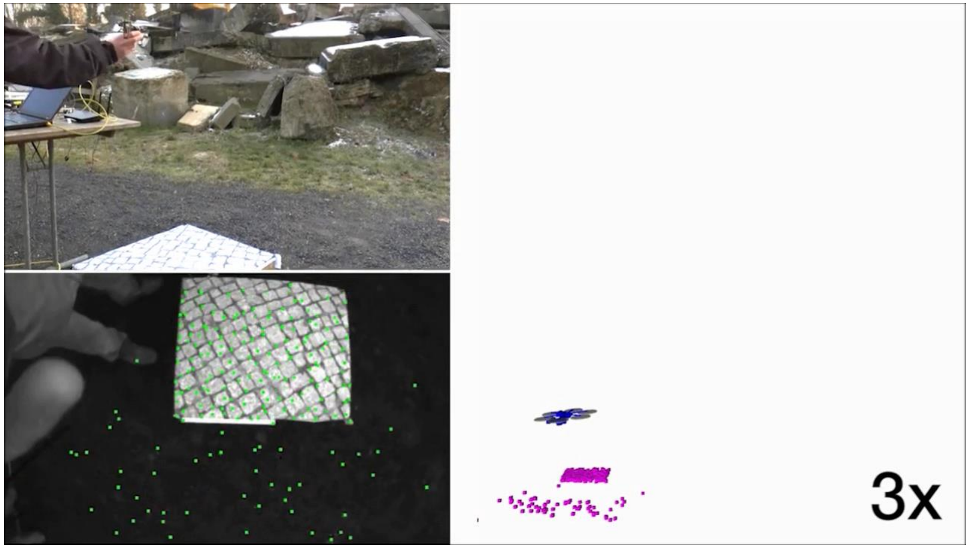


450 grams

Indoors and outdoors experiments

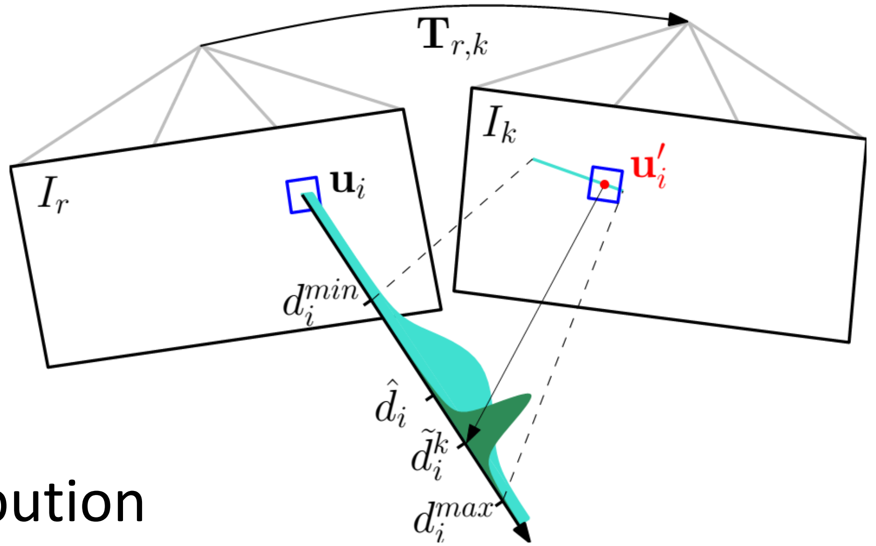
RMS error: 5 mm, height: 1.5 m – Down-looking camera

Speed: 4 m/s, height: 1.5 m – Down-looking camera



Faessler, Fontana, Forster, Mueggler, Pizzoli, Scaramuzza, Autonomous, Vision-based Flight and Live Dense 3D Mapping with a Quadrotor Micro Aerial Vehicle, **Journal of Field Robotics, 2015**.

Probabilistic Depth Estimation

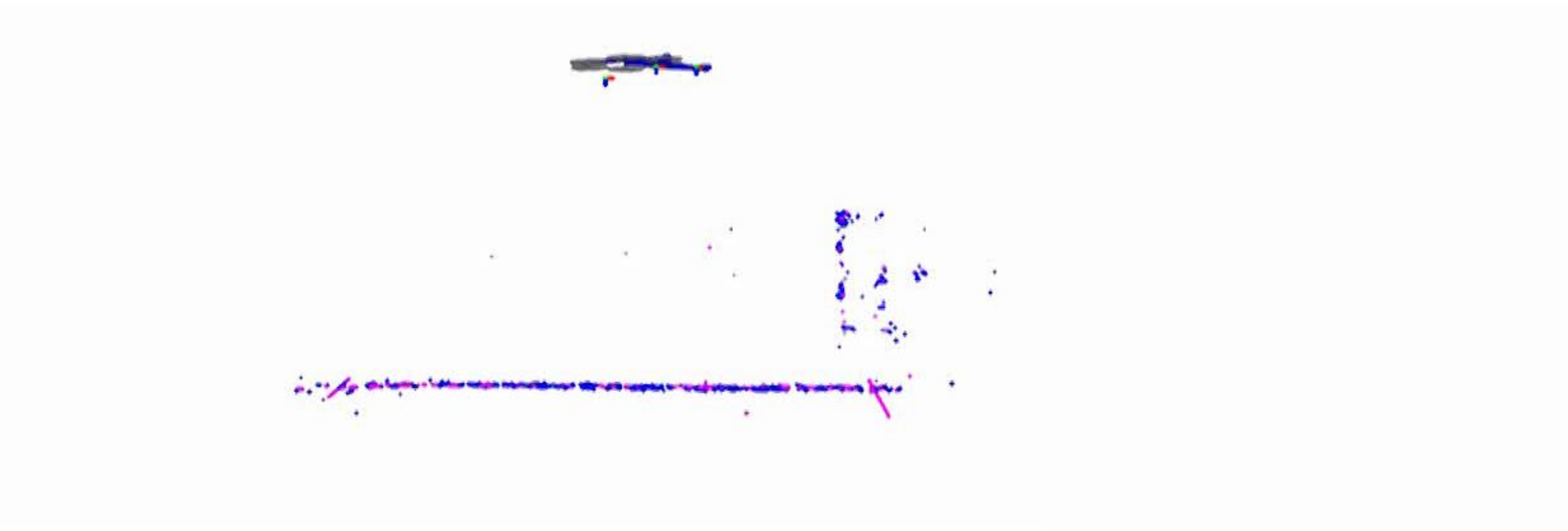


Depth-Filter:

- Depth Filter for every feature
- Recursive Bayesian depth estimation

Mixture of Gaussian + Uniform distribution

$$p(\tilde{d}_i^k | d_i, \rho_i) = \rho_i \mathcal{N}\left(\tilde{d}_i^k | d_i, \tau_i^2\right) + (1 - \rho_i)\mathcal{U}\left(\tilde{d}_i^k | d_i^{\min}, d_i^{\max}\right)$$



[Forster, Pizzoli, Scaramuzza, SVO: Semi Direct Visual Odometry, IEEE ICRA'14]

Robustness to Dynamic Objects and Occlusions

- Depth uncertainty is crucial for safety and robustness
- Outliers are caused by wrong data association (e.g., moving objects, distortions)
- Probabilistic depth estimation models outliers

Faessler, Fontana, Forster, Mueggler, Pizzoli, Scaramuzza, Autonomous, Vision-based Flight and Live Dense 3D Mapping with a Quadrotor Micro Aerial Vehicle, **Journal of Field Robotics, 2015**.

Robustness: Adaptiveness and Reconfigurability [ICRA'15]

Automatic recovery from aggressive flight; fully onboard, single camera, no GPS

Faessler, Fontana, Forster, Scaramuzza, Automatic Re-Initialization and Failure Recovery for Aggressive Flight with a Monocular Vision-Based Quadrotor, ICRA'15. **Demo at ICRA'15**, **Featured on BBC and IEEE Spectrum.**

Appearance-based

Active Perception

Active Perception [Bajcsi’88]

My Goal

- Autonomously generate and track a trajectory that satisfies a given task
  - Which trajectory minimizes the **pose uncertainty** and reduces the **control effort**?
  - Which trajectory minimize **perception ambiguities**?
  - How rapidly can it explore an area in order to find an object/person?

The Problem

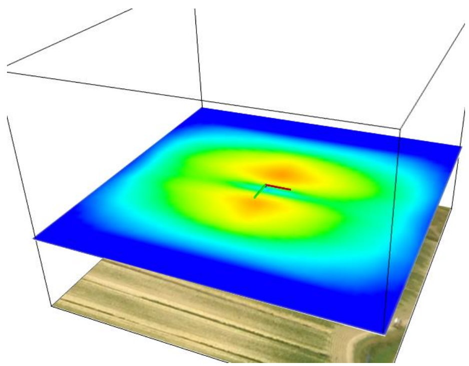
- Previous works on active perception only retained **geometric** information [Davison’02, Burgard’05, Valencia’12] while **discarding scene appearance** (i.e., texture)

Appearance-based Active Vision [RSS'14]

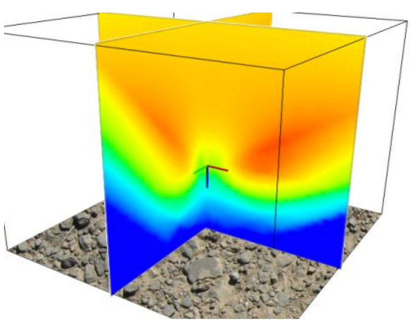
- Select movements that resolve **perception ambiguities** [RSS'14]

$$\Sigma = 2\sigma_i^2 (JJ^T) \qquad J = \sum_P \left[\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right]$$

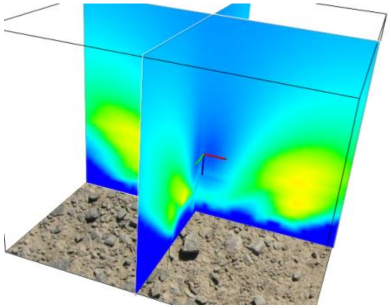
Striped texture



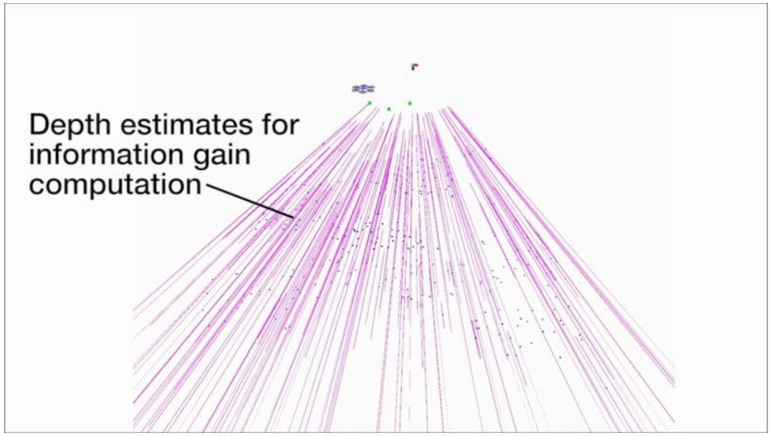
Isotropic texture



After 1 iteration



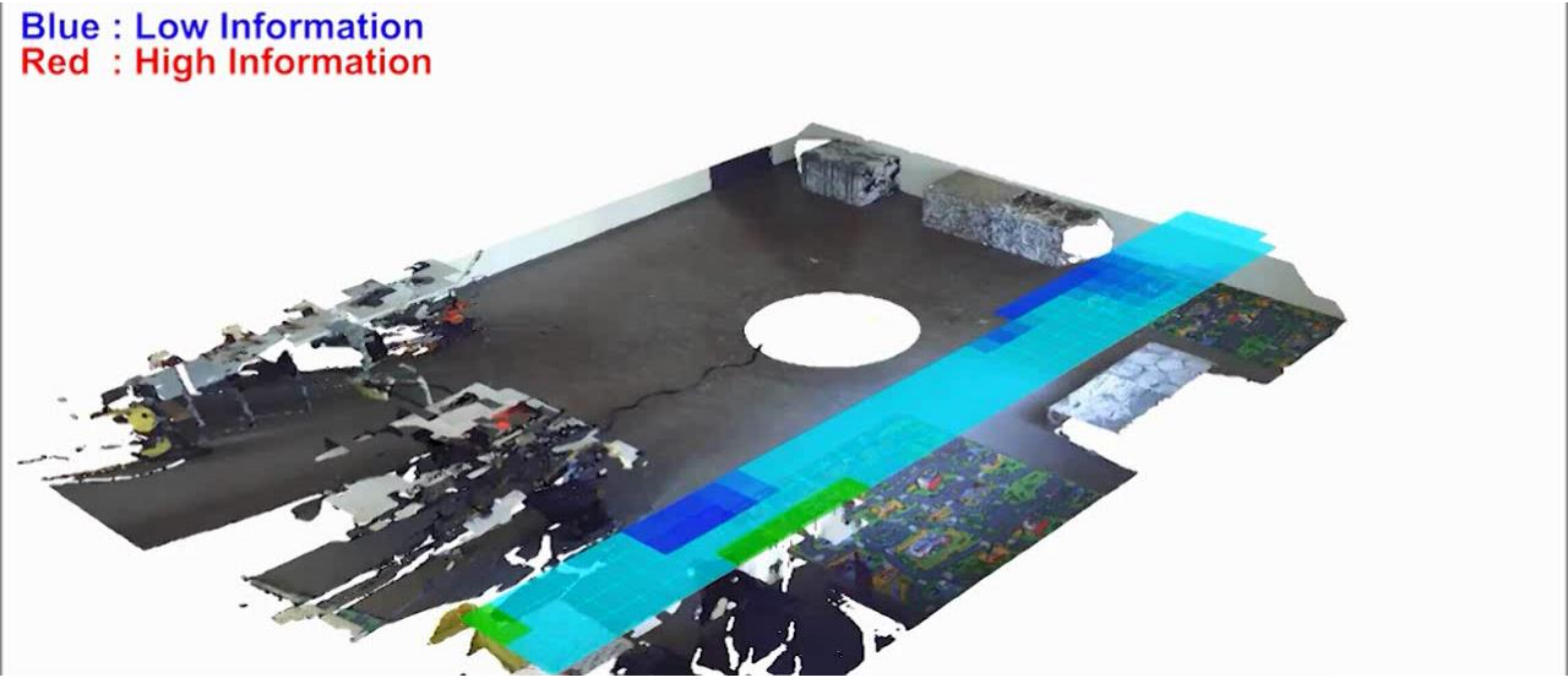
After 10 iterations



Forster, Pizzoli, Scaramuzza, Appearance-based Active, Dense Reconstruction for Micro Aerial Vehicles, RSS'14.

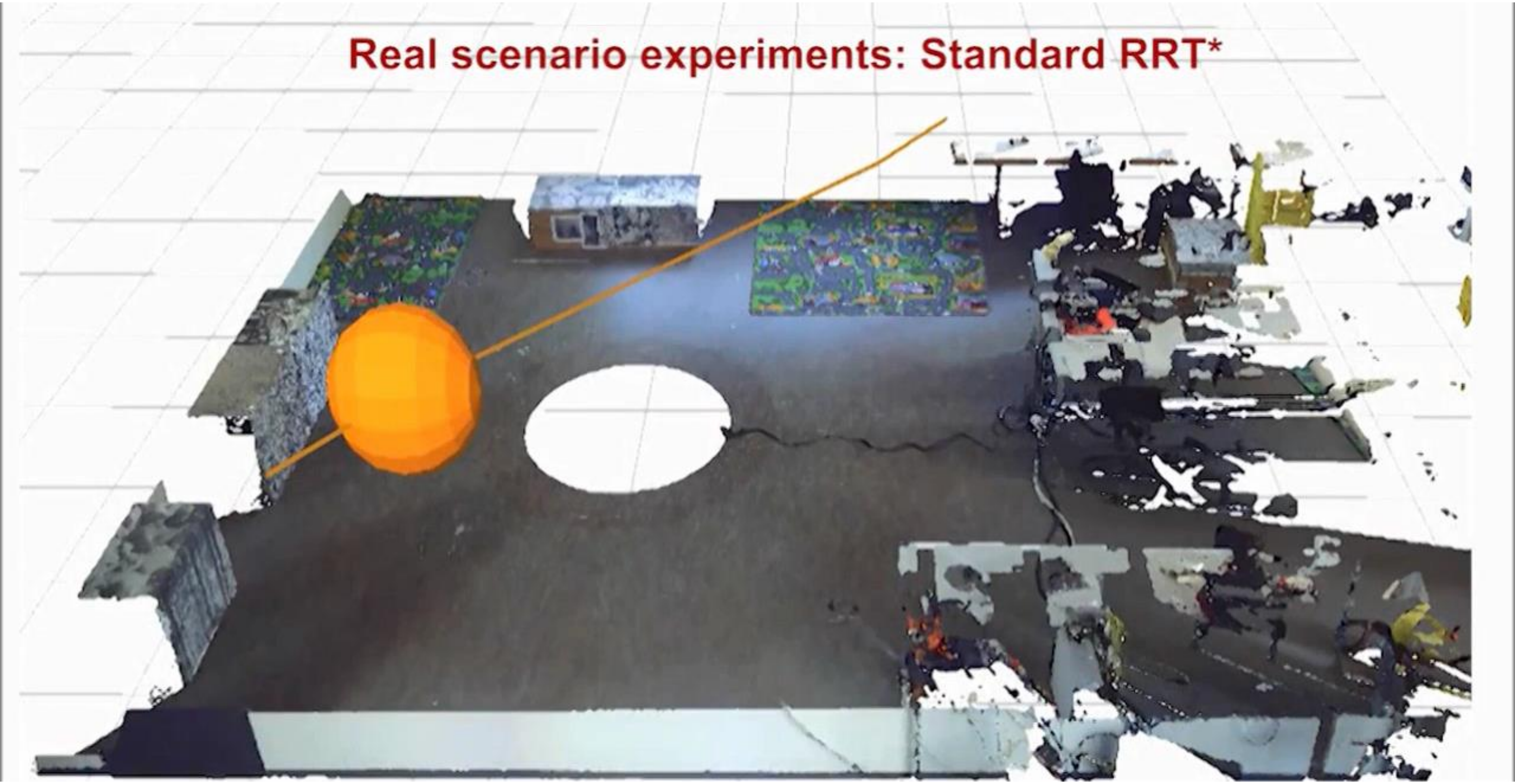
Perception Aware Path Planning [TRO'16]

Favor texture-rich environments to guarantee good tracking quality



Costante, Forster, Scaramuzza, *Perception Aware Path Planning*, IEEE Trans. on Robotics, 2016.

Perception Aware Path Planning [TRO'16]



Costante, Forster, Scaramuzza, *Perception Aware Path Planning*, IEEE Trans. on Robotics, 2016.

Low-latency, Agile Flight

Open Problems and Challenges with Micro Helicopters

Current flight maneuvers achieved with onboard cameras are still to slow compared with those attainable with Motion Capture Systems

Penn Mellinger, Kumar

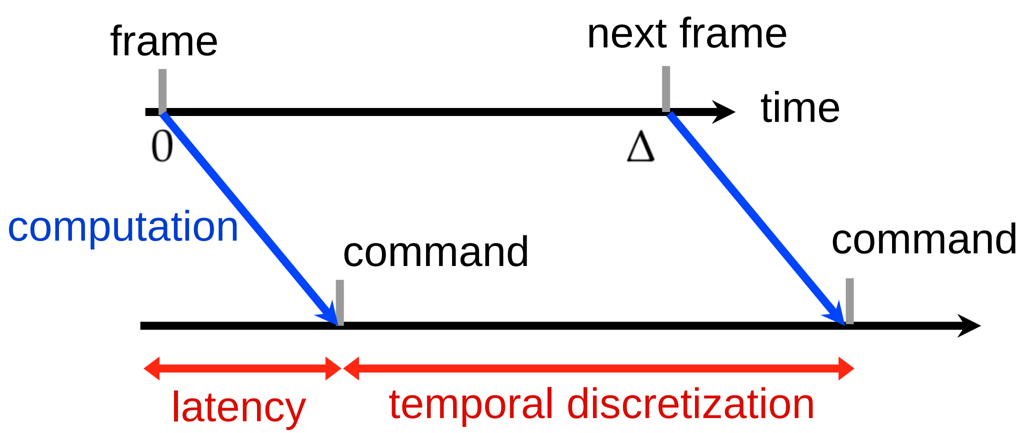
ETH Mueller, D'Andrea

The acrobatics shown in these videos were done with a motion capture system

To go faster, we need faster sensors!

- At the current state, the agility of a robot is limited by the latency and temporal discretization of its sensing pipeline.

- Currently, the average robot-vision algorithms have latencies of 50-200 ms. This puts a hard bound on the agility of the platform.



- **Can we create a low-latency, low-discretization perception pipeline?**
  - Yes, if we combine **cameras with event-based** sensors

[Censi & Scaramuzza, «Low Latency, Event-based Visual Odometry», ICRA'14]

Dynamic Vision Sensor (DVS)

- **Event-based camera** developed by Tobi Delbruck's group (ETH & UZH).
- Temporal resolution: **1 μs**
- High dynamic range: **120 dB**
- Low power: **20 mW**
- Cost: 2,500 EUR

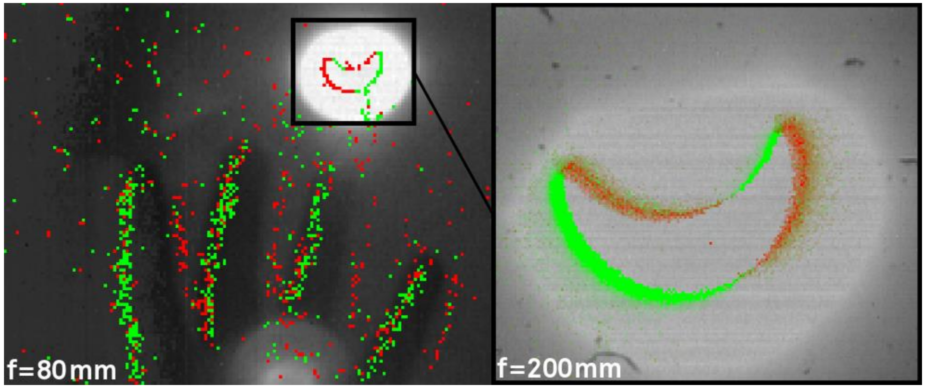
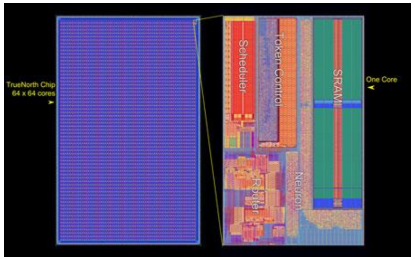


Image of the solar eclipse (March'15) captured by a DVS (courtesy of IniLabs)

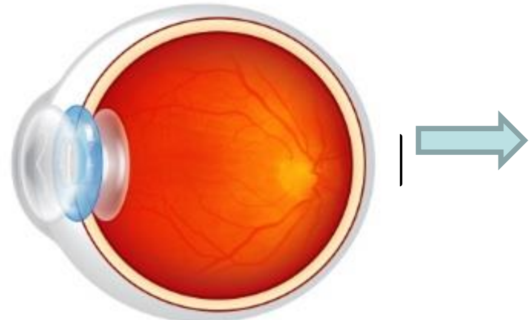


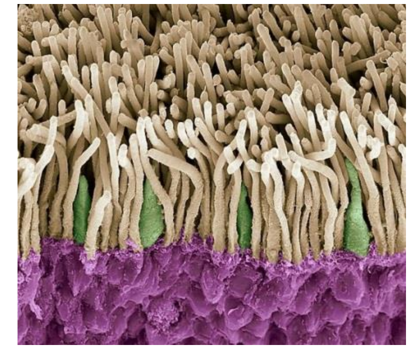
DARPA project Synapse: 1M neuron, brain-inspired processor: IBM TrueNorth

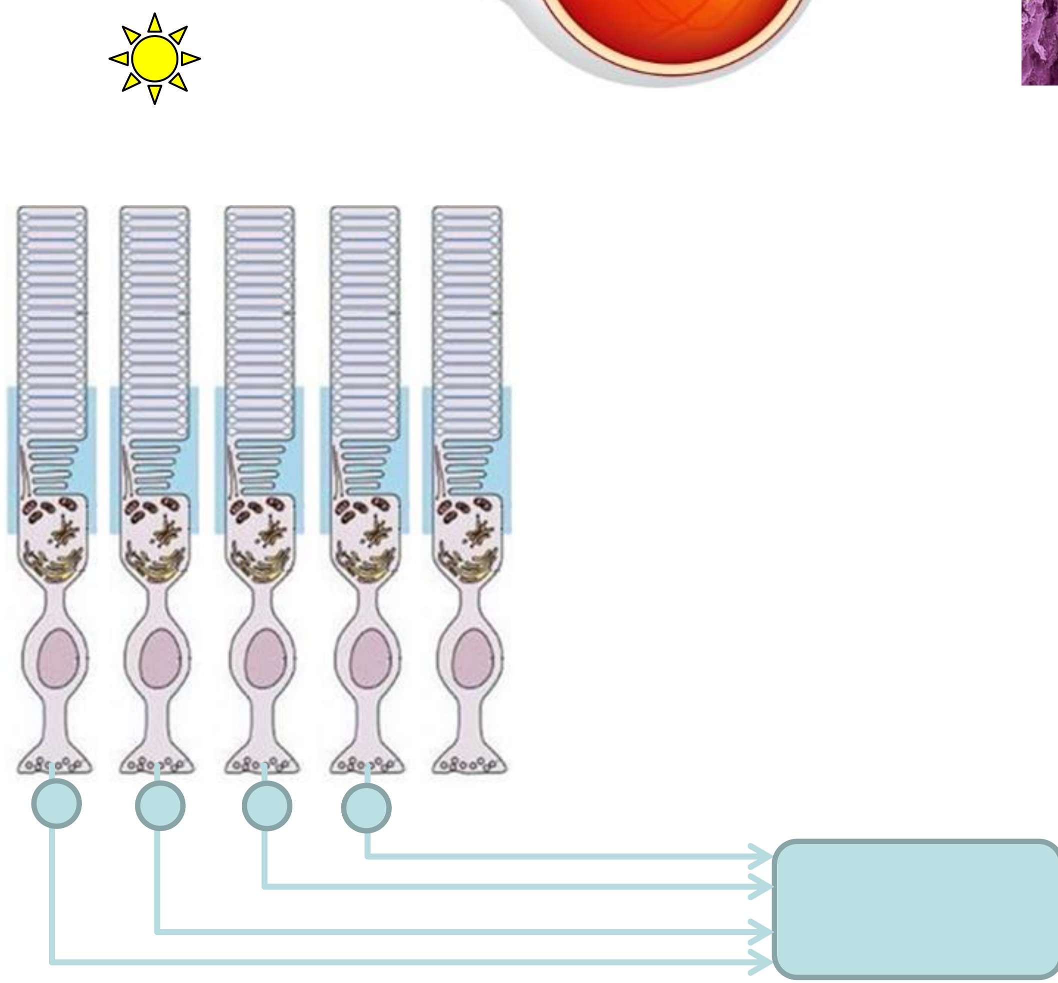
[Lichtsteiner, Posch, Delbruck. A 128x128 120 dB 15μs Latency Asynchronous Temporal Contrast Vision Sensor. 2008]

Human Vision System

- 130 million **photoreceptors**
- But only 2 million **axons**!



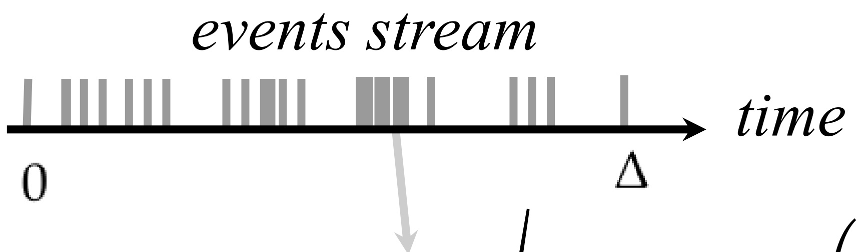




Camera vs DVS

- A **traditional camera** outputs frames at **fixed time intervals**:

- By contrast, a **DVS** outputs **asynchronous events** at ***microsecond* resolution**. An event is generated each time a single pixel detects an intensity changes value

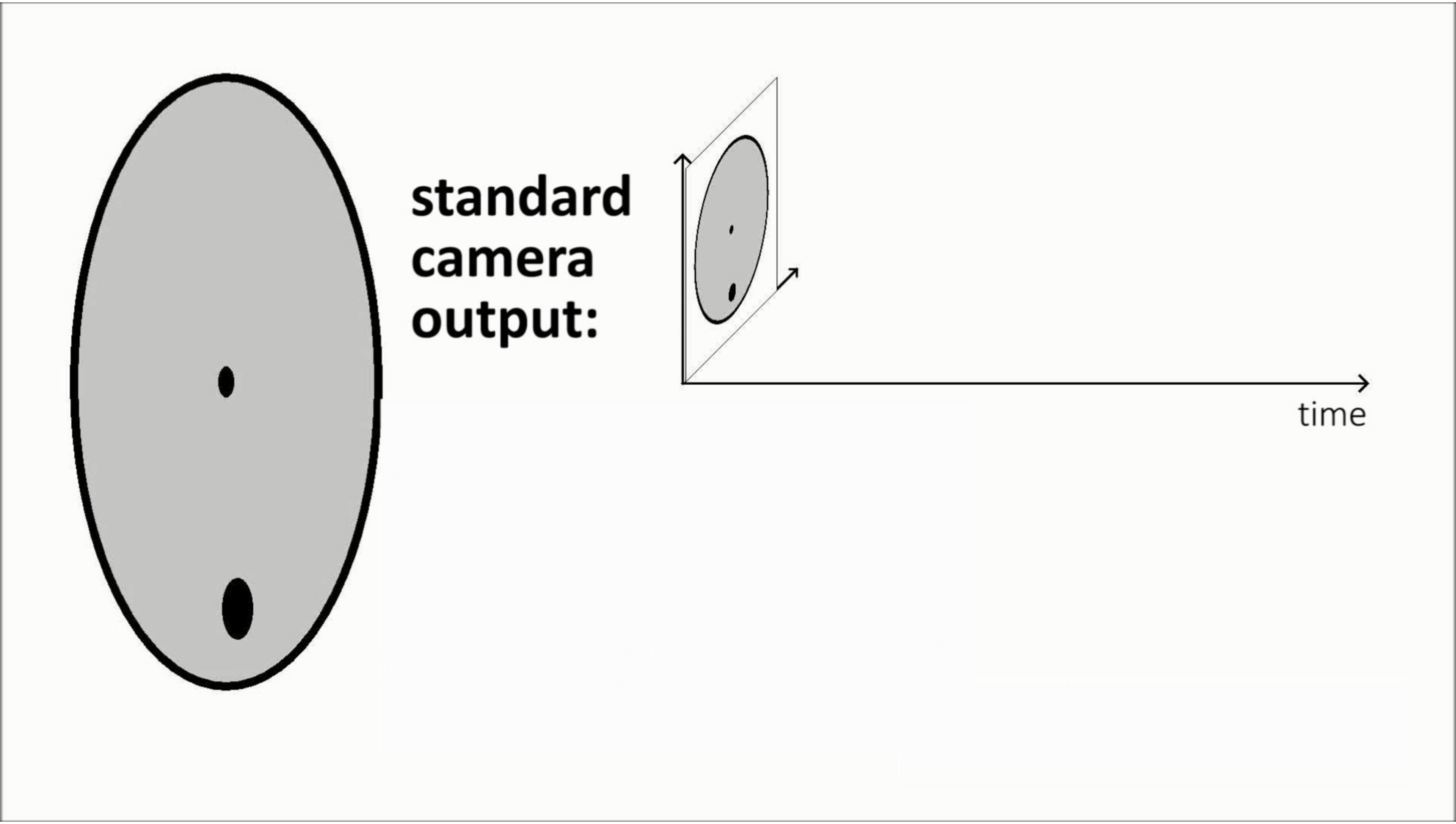


$$\textit{event:}\ \left\langle t, \langle x, y\rangle, sign\left(\frac{dI(x,y)}{dt}\right)\right\rangle$$

sign (+1 or -1)

Lichtsteiner, Posch, Delbruck. *A 128x128 120 dB 15μs Latency Asynchronous Temporal Contrast Vision Sensor.* 2008

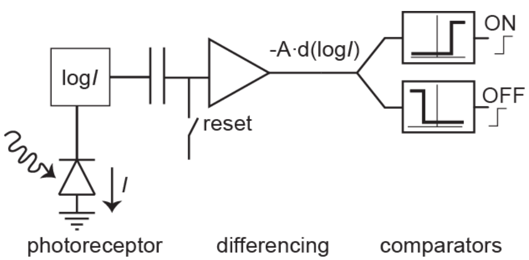
Camera vs Dynamic Vision Sensor



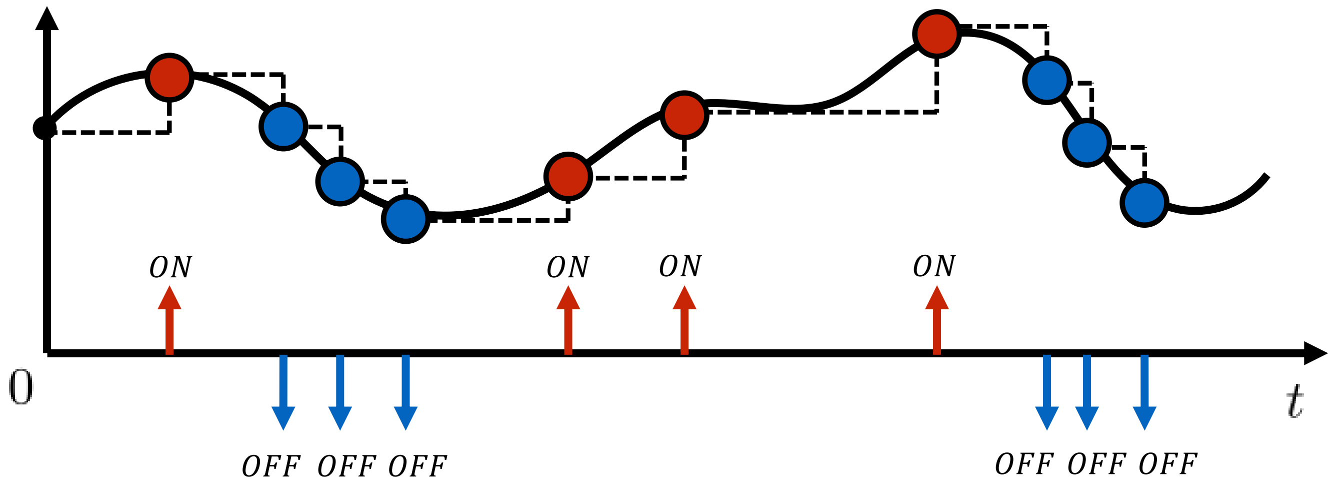
DVS Operating Principle [Lichtsteiner, ISCAS'09]

Events are generated any time a single pixel sees a change in brightness larger than C

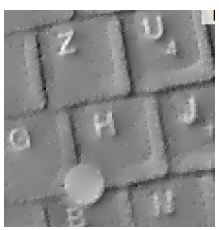
$$\Delta \log I \geq C$$



$$\mathrm{V} = \log I(t)$$



The intensity signal at the event time can be reconstructed by integration of $\pm C$

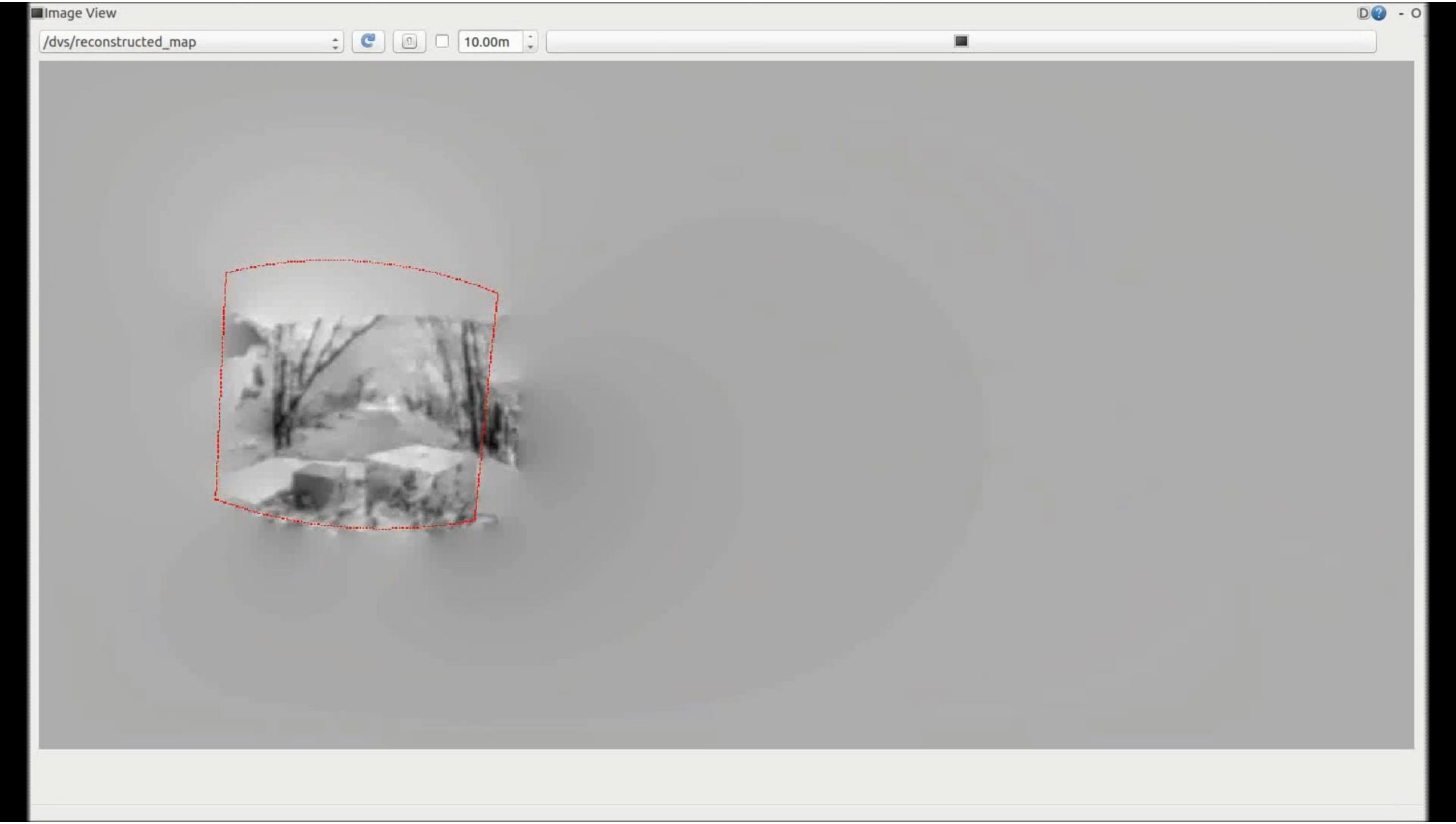


[Cook et al., IJCNN'11]

[Kim et al., BMVC'15]

[Lichtsteiner, Posch, Delbruck. A 128x128 120 dB 15µs Latency Asynchronous Temporal Contrast Vision Sensor. 2008]

Pose Tracking and Intensity Reconstruction from a DVS



Dynamic Vision Sensor (DVS)

Advantages

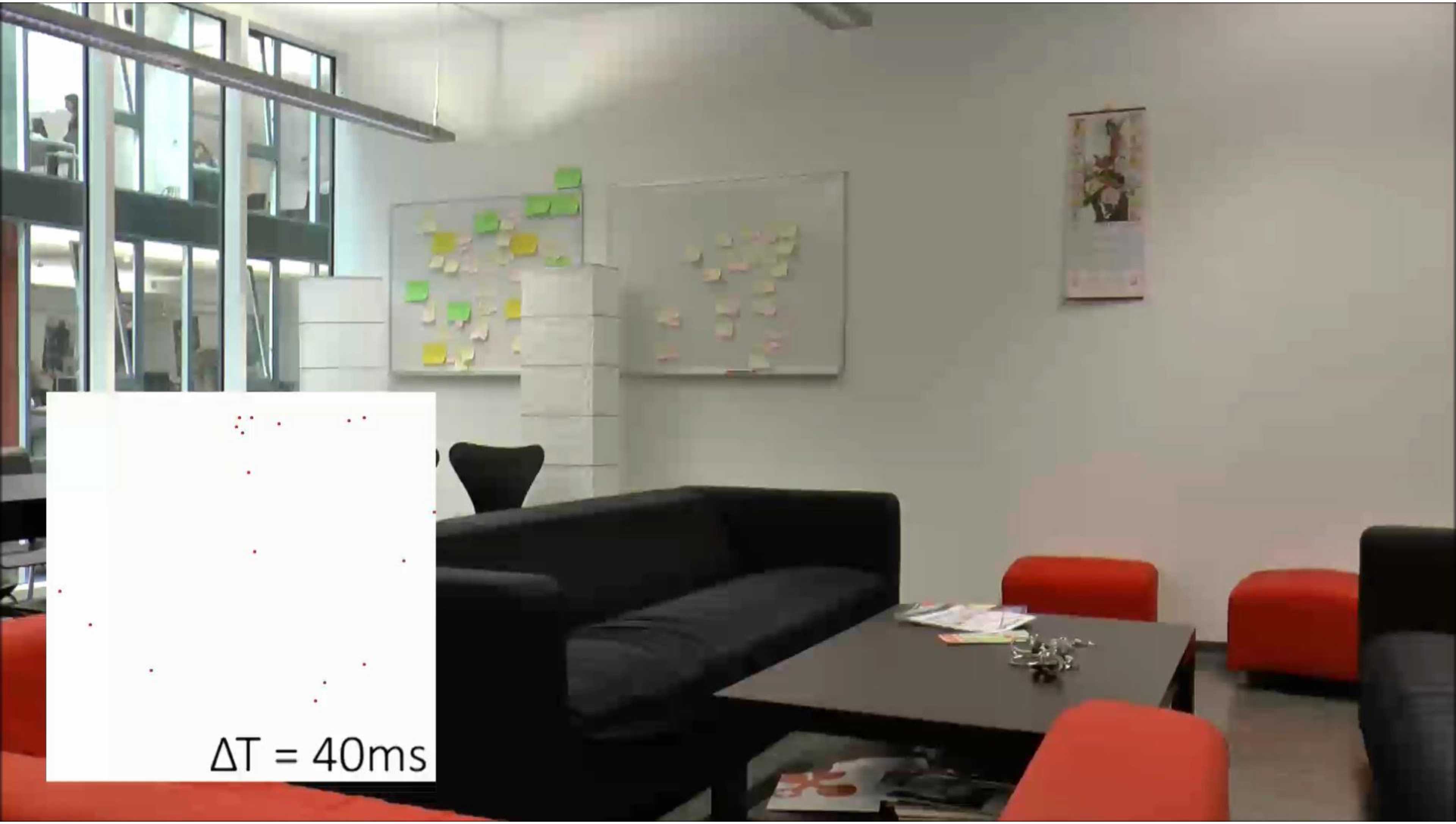
- **low-latency** (~1 micro-second)
- **high-dynamic range** (120 dB instead 60 dB)
- Very **low bandwidth** (only intensity changes are transmitted): ~200Kb/s
- **Low storage capacity, processing time, and power**

Disadvantages

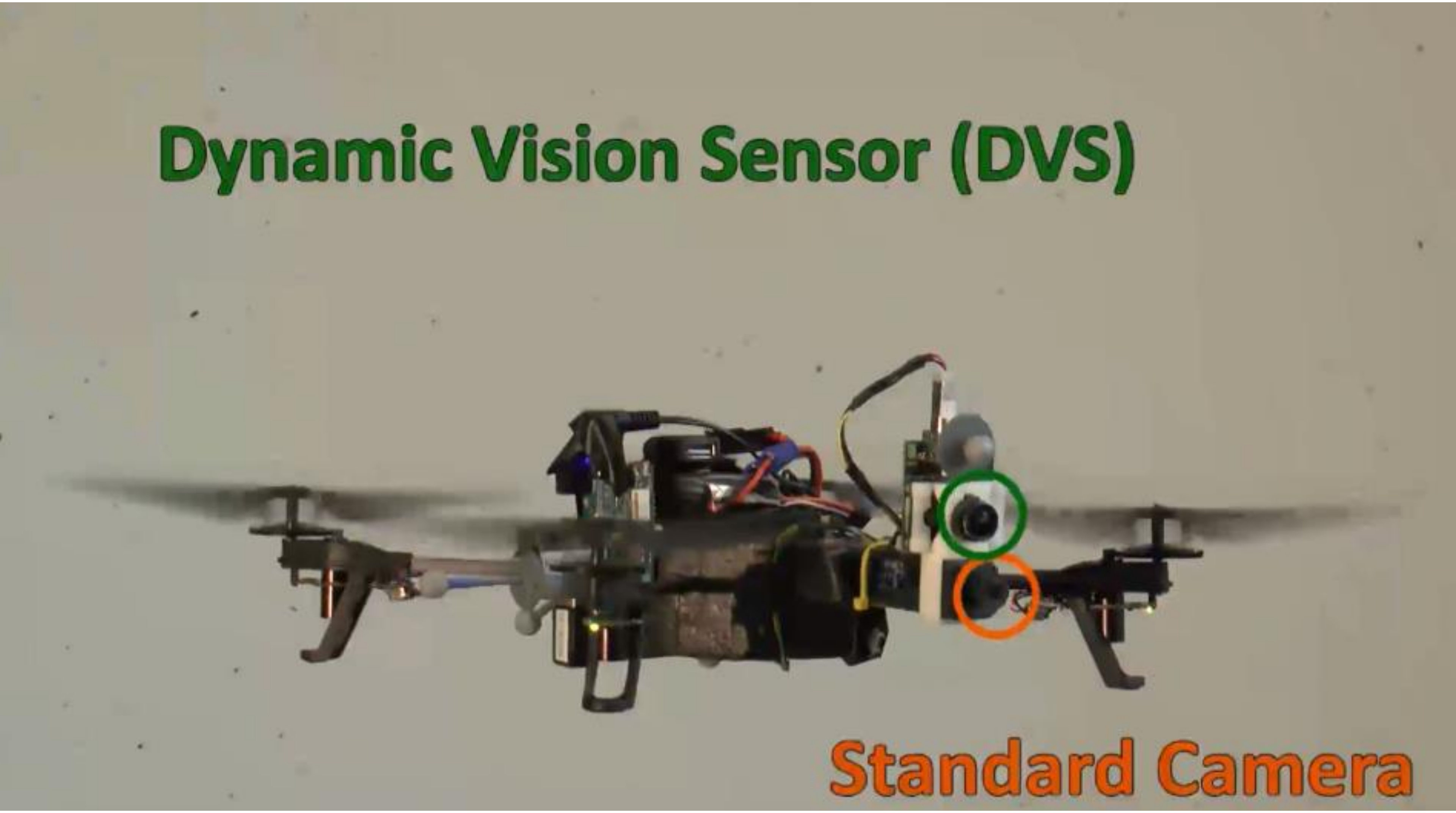
- Require totally **new vision algorithms**
- **No intensity information** (only binary intensity changes)
- **Very low image resolution**: 128x128 pixels

Lichtsteiner, Posch, Delbruck. *A 128x128 120 dB 15μs Latency Asynchronous Temporal Contrast Vision Sensor.* 2008

Camera vs Dynamic Vision Sensor



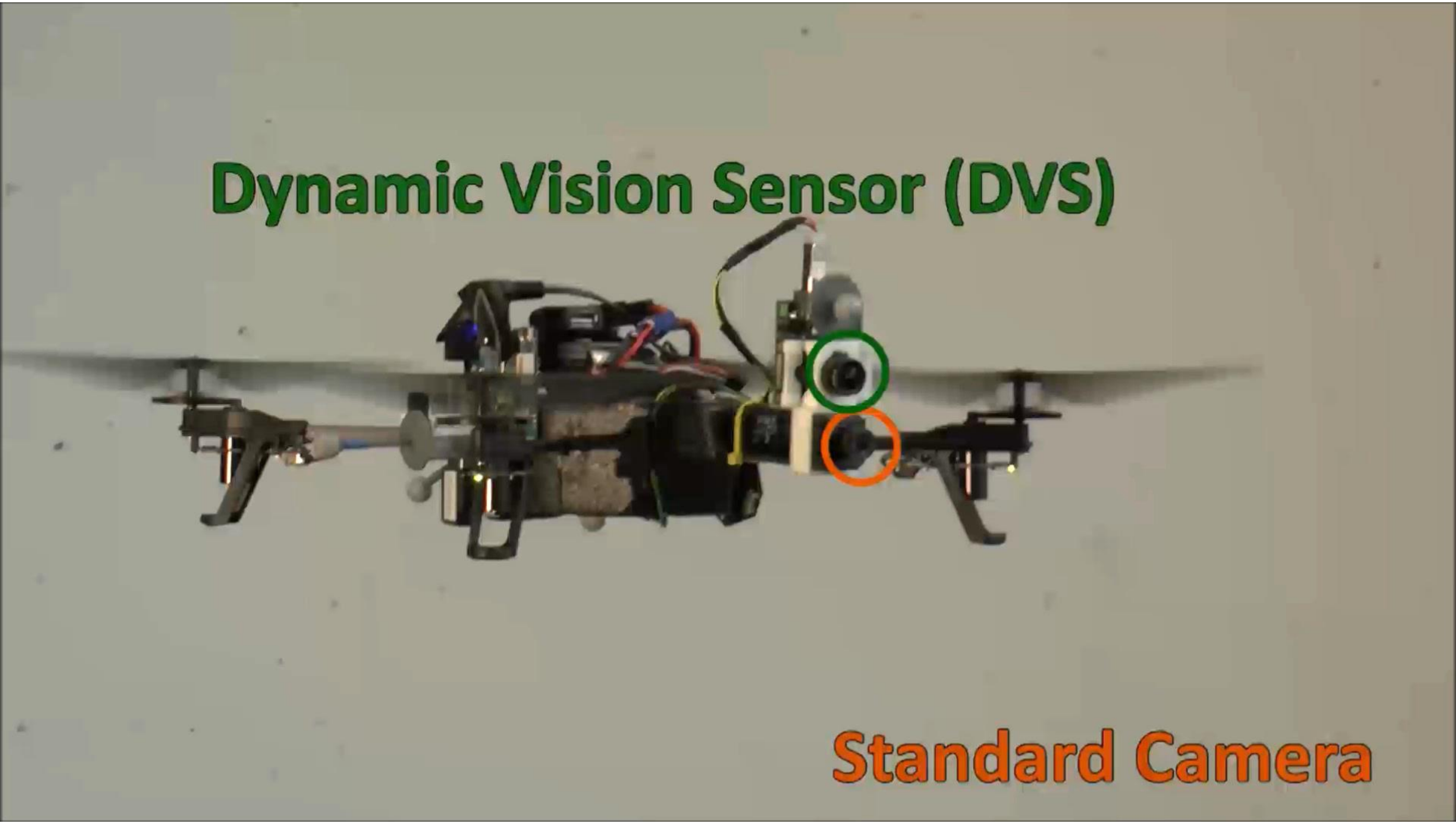
DVS mounted on a quadrotor AR Drone [IROS, RSS]



[Mueggler, Huber, Scaramuzza, *Event-based, 6-DOF Pose Tracking for High-Speed Maneuvers*, IROS'14]

[Mueggler, G. Gallego, D. Scaramuzza, *Continuous-Time Trajectory Estimation for Event-based Vision Sensors*, Robotics: Science and Systems (RSS), Rome, 2015]

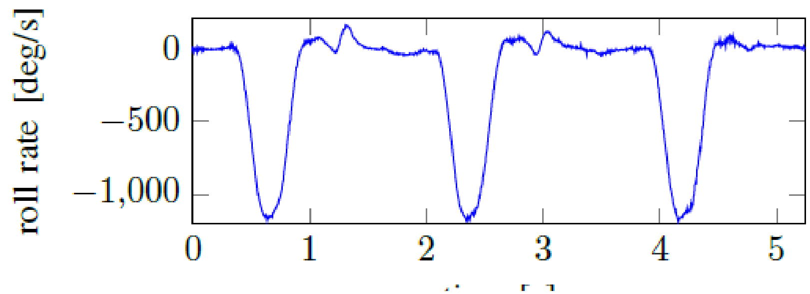
Application Experiment: Quadrotor Flip (1,200 deg/s)



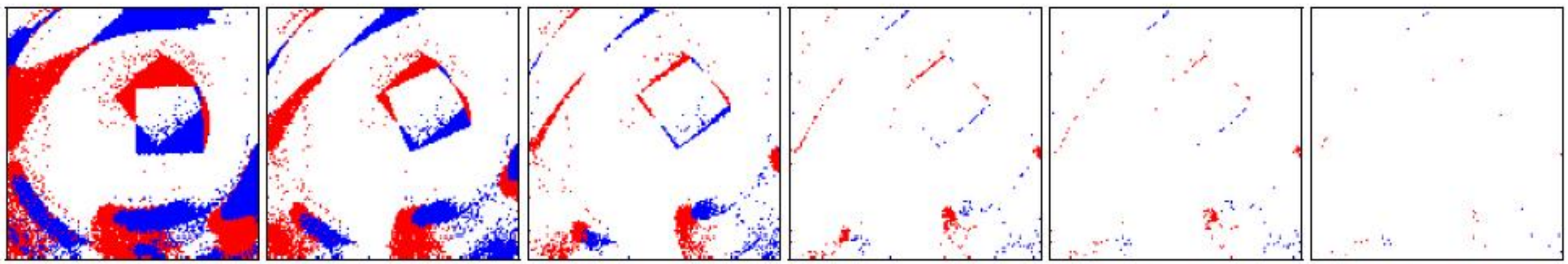
[Mueggler, Huber, Scaramuzza, *Event-based, 6-DOF Pose Tracking for High-Speed Maneuvers*, IROS'14]

[Mueggler, G. Gallego, D. Scaramuzza, *Continuous-Time Trajectory Estimation for Event-based Vision Sensors*, Robotics: Science and Systems (RSS), Rome, 2015]

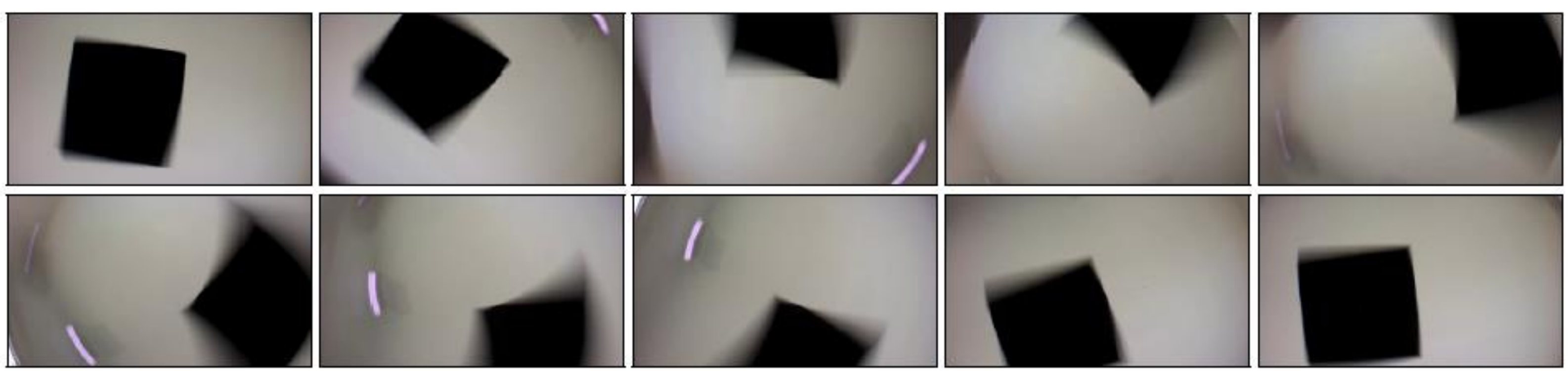
Camera and DVS renderings



Peak Angular Speed: 1,200 deg/s



(a) $\Delta T = 33\,\mathrm{ms}$ (b) $\Delta T = 15\,\mathrm{ms}$ (c) $\Delta T = 5\,\mathrm{ms}$ (d) $\Delta T = 1\,\mathrm{ms}$ (e) $\Delta T = 0.5\,\mathrm{ms}$ (f) $\Delta T = 0.1\,\mathrm{ms}$



IROS'14, RSS'15

Frame-based vs Event-based Vision

- Naive solution: **accumulate events** occurred over a certaint time interval and adapt «standard» CV algorithms.
  - Drawback: it **increases latency**

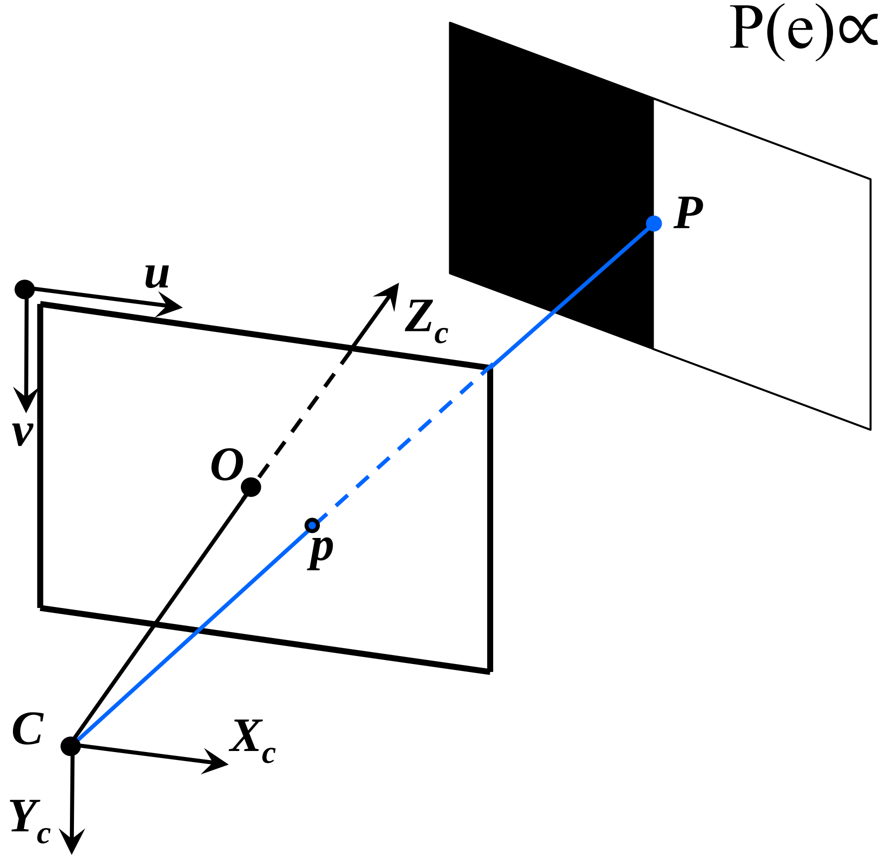
- Instead, we want **each single event** to be used **as it comes!**

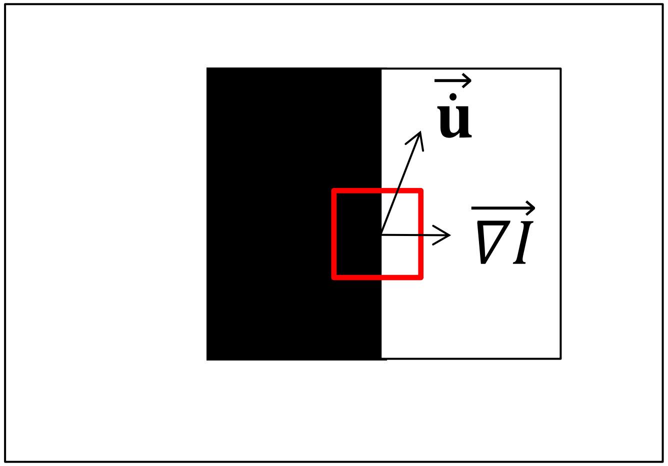
- Problems
  - DVS output is a sequence of **asynchrnous events** rather than a standard image
  - Thus, a **paradigm shift** is needed to deal with its data

Generative Model [Censi & Scaramuzza, ICRA'14]

The generative model tells us that the **probability** that an event is generated depends on the **scalar product** between the gradient ∇I and the apparent motion $\dot{\mathbf{u}}\Delta t$

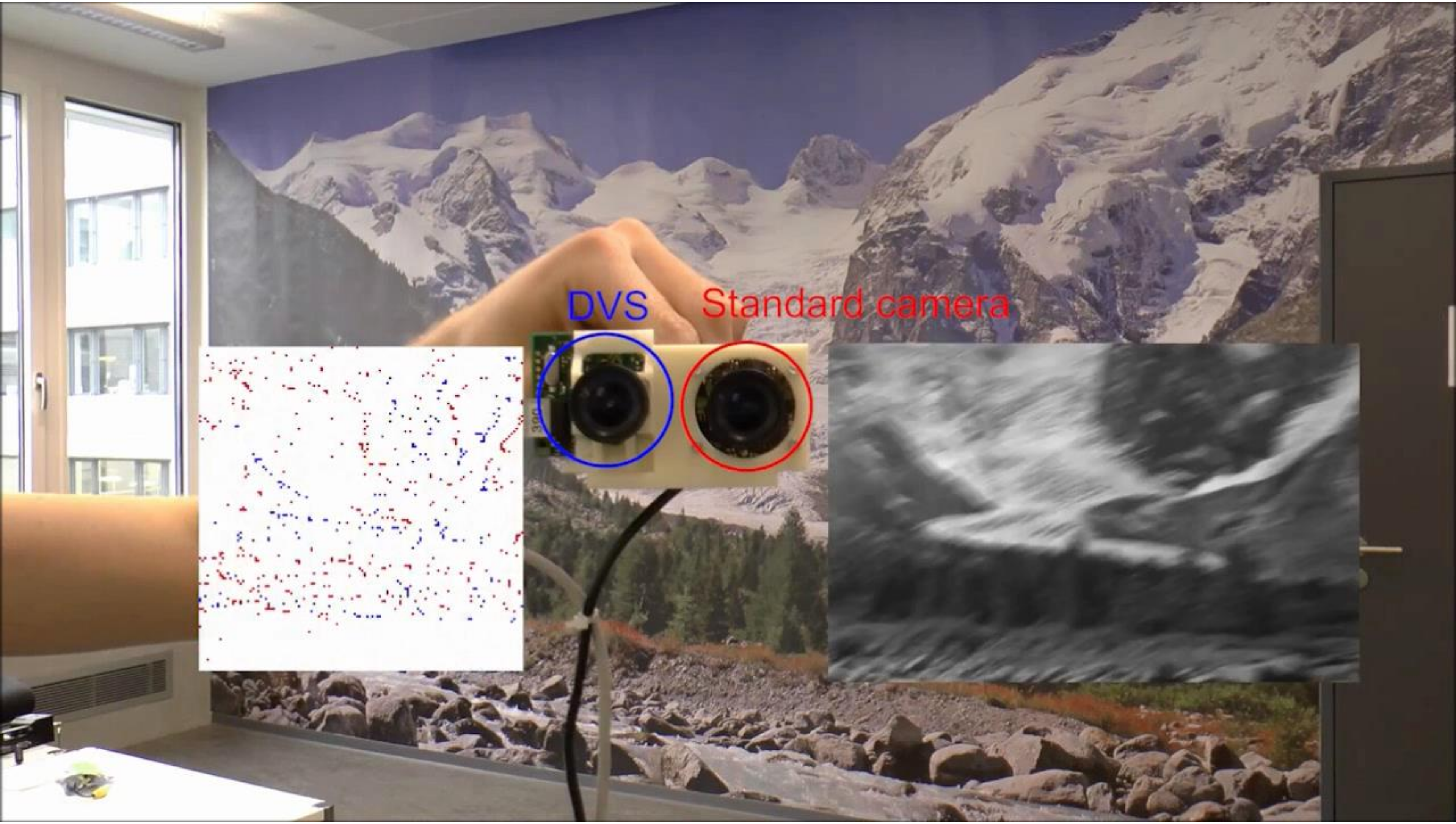
$$\mathrm{P(e)} \propto |\langle \nabla I, \dot{\mathbf{u}}\Delta t \rangle|$$





[Censi & Scaramuzza, *Low Latency, Event-based Visual Odometry*, ICRA'14]

Event-based 6DoF Pose Estimation Results



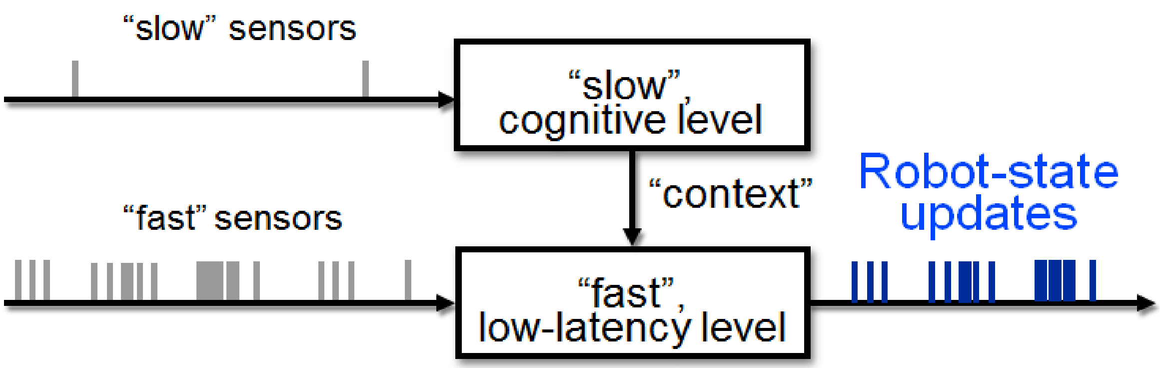
[Event-based, 6-DOF Camera Tracking for High-Speed Applications, Submitted to PAMI]

[Censi & Scaramuzza, *Low Latency, Event-based Visual Odometry*, ICRA'14]

Recap

- DVS: **revolutionary sensor** for robotics:
    1. **low-latency** (~1 micro-second)
    2. **high-dynamic range** (120 dB instead 60 dB)
    3. Very **low bandwidth** (only intensity changes are transmitted)

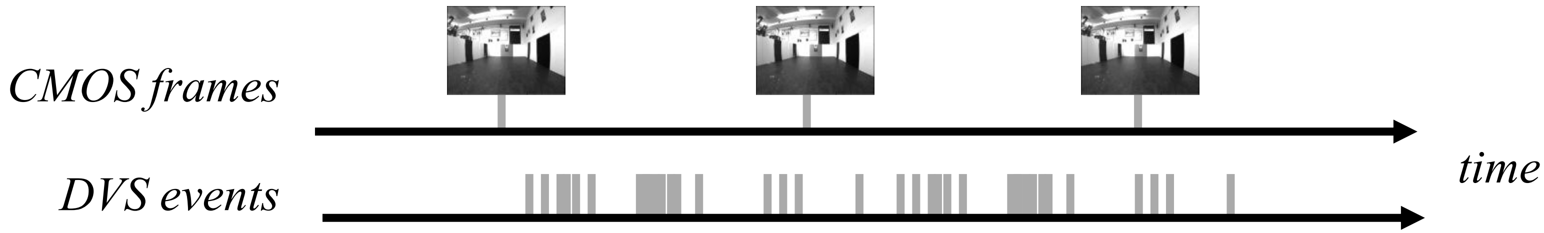
- Possible future sensing architecture:



[Censi & Scaramuzza, *Low Latency, Event-based Visual Odometry*, ICRA'14]

DAVIS: Dynamic and Active-pixel Vision Sensor [Brandli'14]

Combines the event-driven activity output of the DVS with conventional static frame output of CMOS active-pixel sensors.

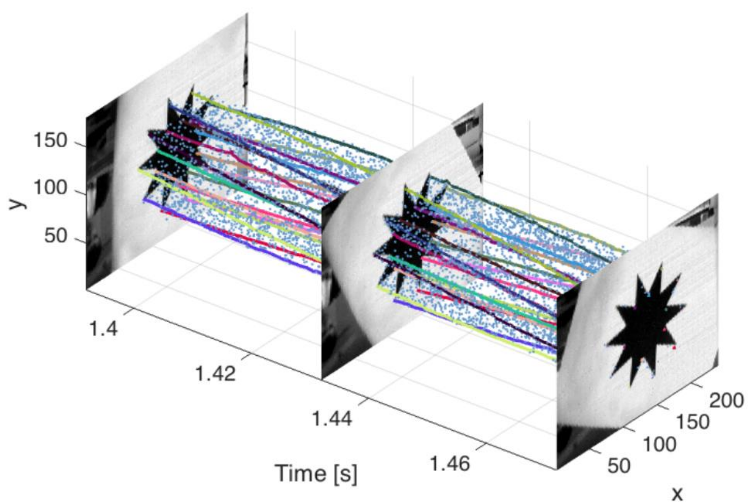


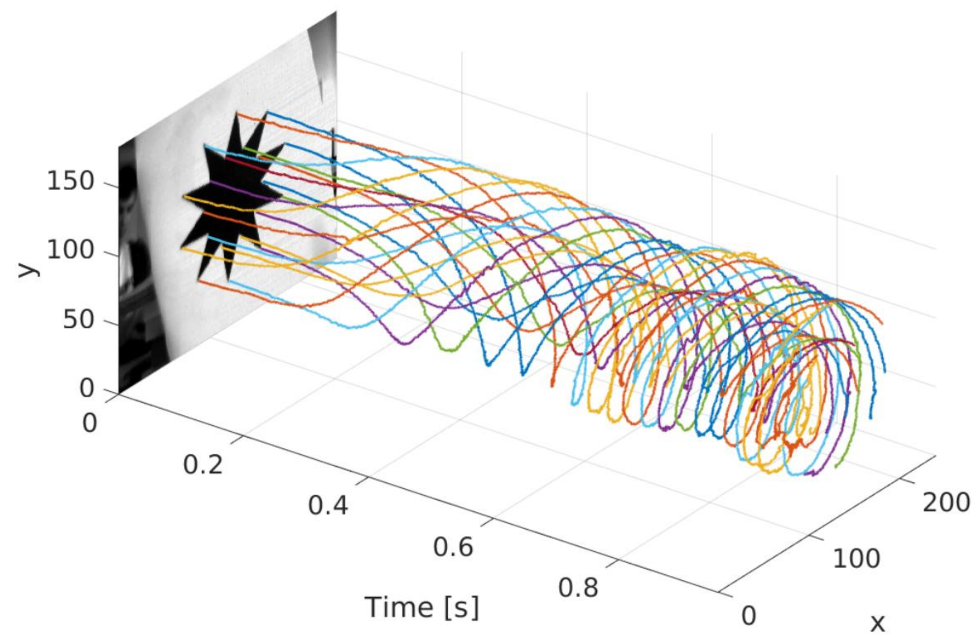
Brandli, Berner, Yang, Liu, Delbruck, "*A 240× 180 130 dB 3 µs Latency Global Shutter Spatiotemporal Vision Sensor*." IEEE Journal of Solid-State Circuits, 2014.

Event-based Feature Tracking [EBCCSP'16]

- Extract Harris corners on images
- Track corners using event-based Iterative Closest Points (ICP)

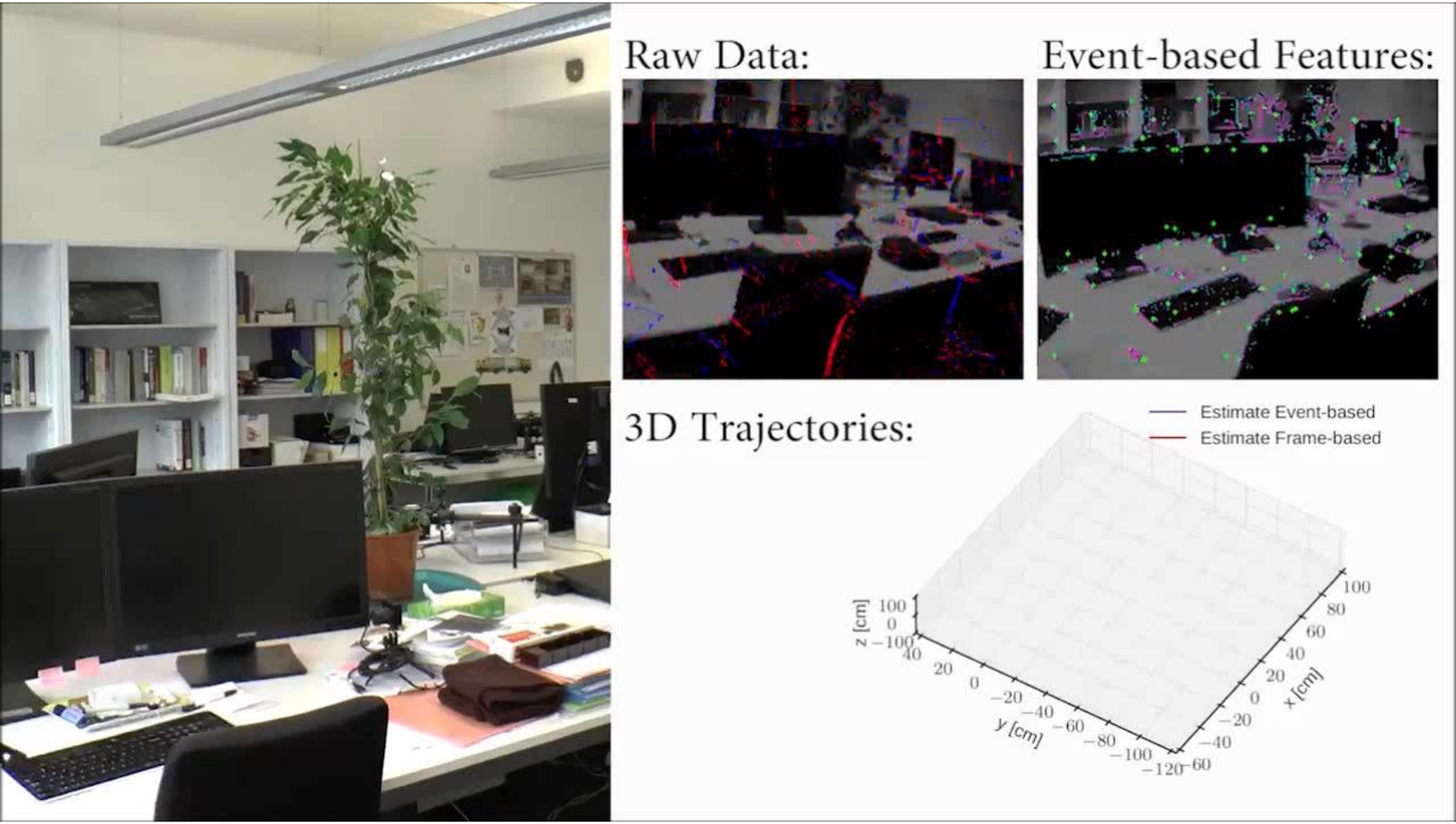
$$\arg\min_{\mathtt{A}} \sum_{(\mathbf{p}_i,\mathbf{m}_i)\in\text{Matches}} \|\mathtt{A}(\mathbf{p}_i) - \mathbf{m}_i\|^2$$





Tedaldi, Gallego, Mueggler, Scaramuzza, "Feature Detection and Tracking with the Dynamic and Active-Pixel Vision Sensor (DAVIS", IEEE Int. Conference on Event-based Control, Communication, and Signal Processing, EBCCSP'16.

Event-based, Sparse Visual Odometry [IROS'16]



Conclusions

- Agile flight (**like birds**) is still far (10 years?)
- Agile flight requires success at different levels
  - **perception, planning, and control**
- **Perception and control** need to be considered **jointly!**
- **Event cameras** open enormous possibilities! Standard cameras have been studied for 50 years!

Thanks!