Action Recognition and Video Event Detection

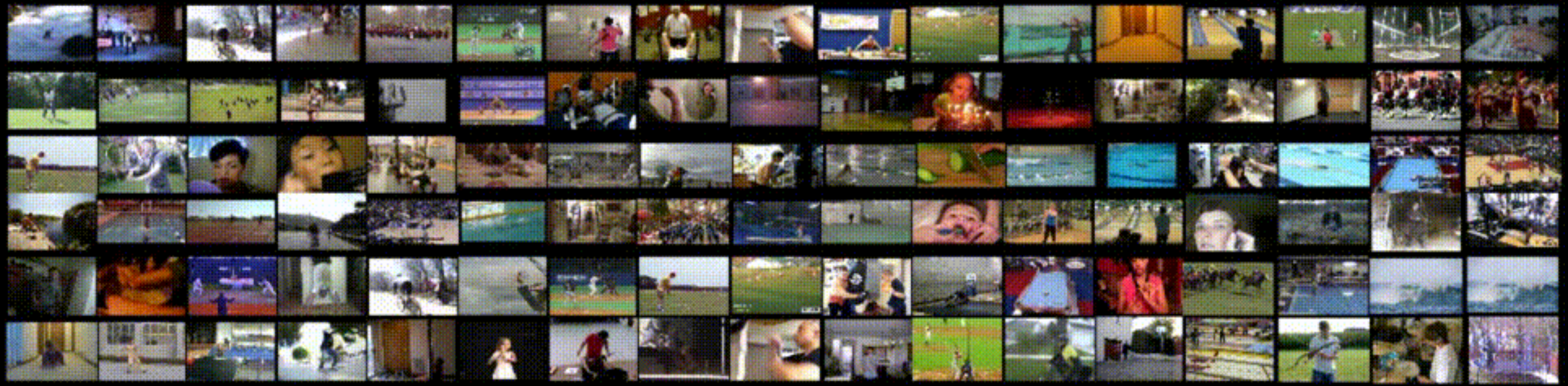
Prof. Kuan-Ting Lai

2021/5/24

Human Action Recognition

UCF-101

http://www.thumos.info/



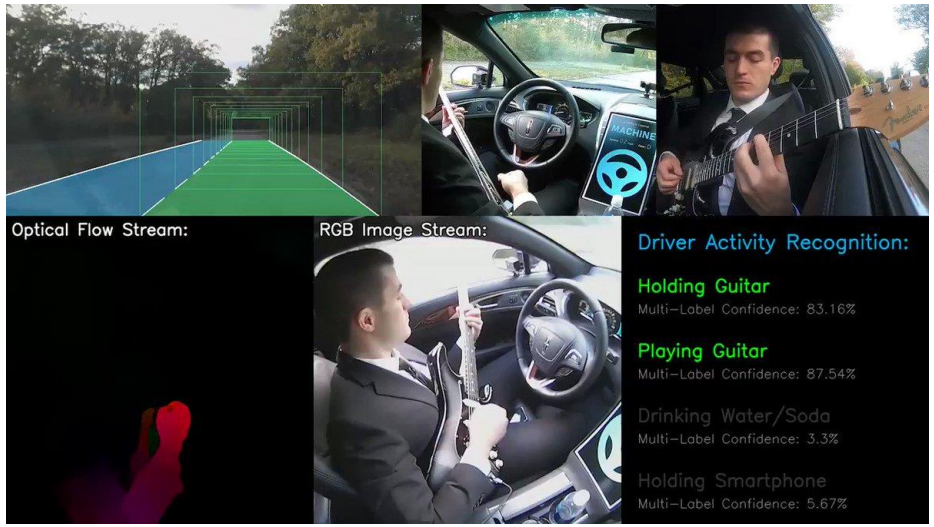
Applications of Video Event Detection

Video Search

Surveillance Video Analysis

Drone Action Recognition

Driver Activity Detection



Complex Video Events

- Basic event or action detection


push


pushup


ride bike


ride horse


run

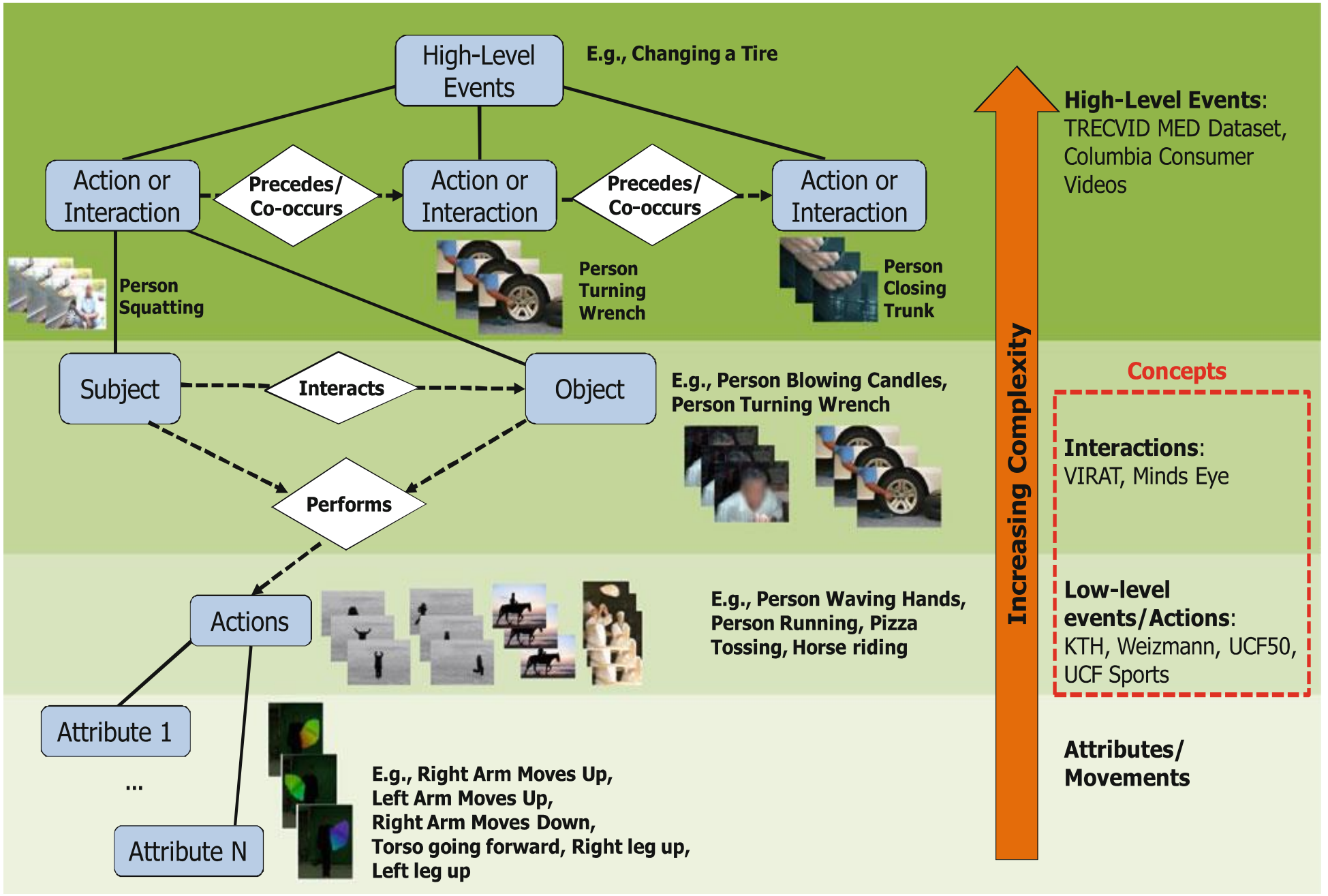

shake hands


shoot ball

- Complex or high-level event detection

Wedding Ceremony

Detect Complex (High-level) Events



- Jiang et al., high level events recognition in unconstrained videos, 2012

TRECVID Contest

- NIST TRECVID Multimedia Event Detection (MED) contest
- Detecting complex events in around 100,000 videos

MED 2011 devel. events


Attempting a board trick


Feeding an animal


Landing a fish


Wedding ceremony


Working on a woodworking project

MED 2011 testing events


Birthday party


Changing a vehicle tire


Flash mob gathering


Getting a vehicle unstuck


Grooming an animal


Making a sandwich


Parade


Parkour


Repairing an appliance


Working on a sewing project

Challenges of Video Event Detection

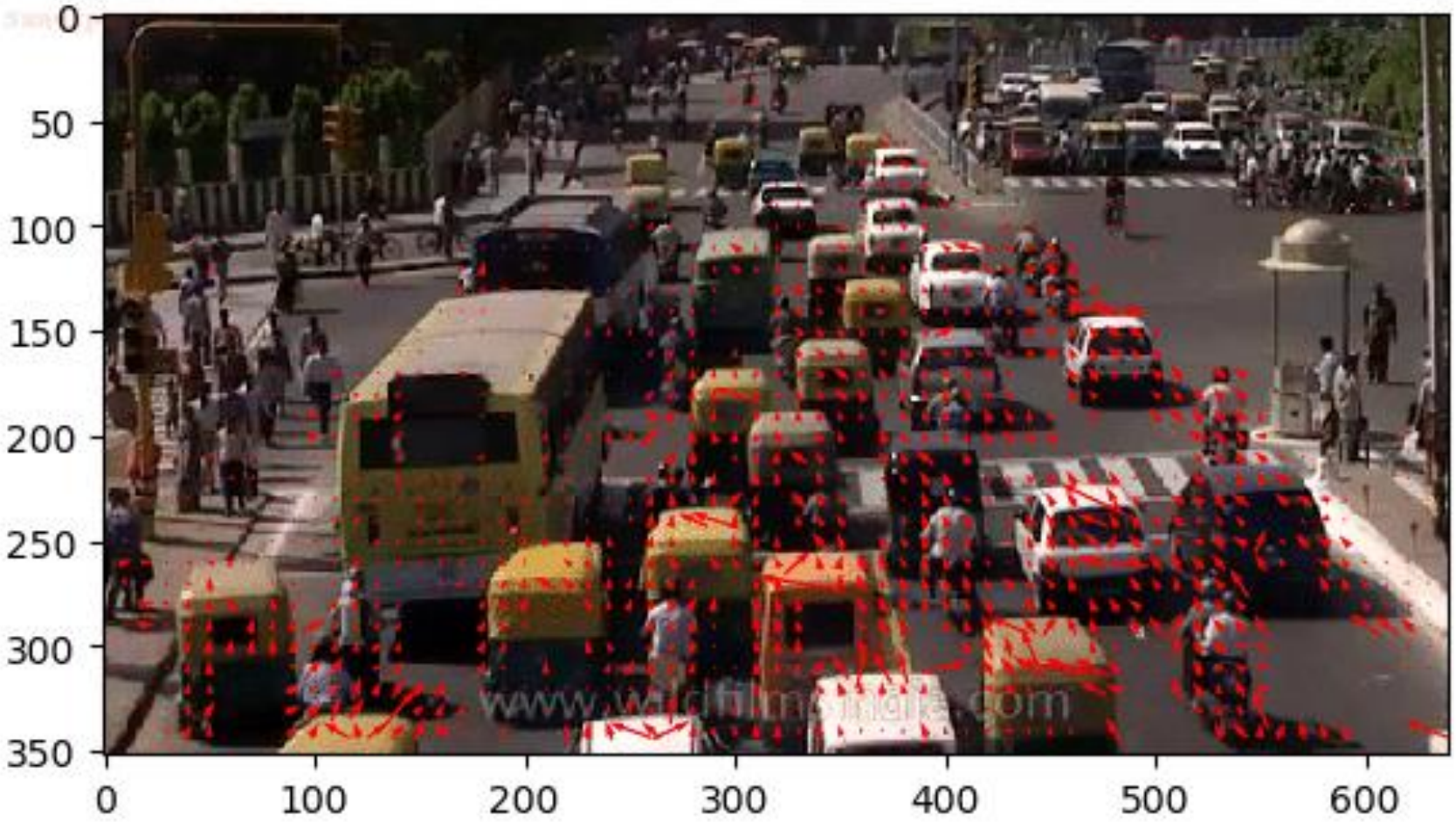
- Consist of interactions between human, objects and scenes
- Actions contain spatial data and temporal data
- Temporal data are huge and noisy
- Hard to define action classes

Optical Flow for Temporal Information

https://devblogs.nvidia.com/an-introduction-to-the-nvidia-optical-flow-sdk/

Estimate the Direction and Distance of Motion

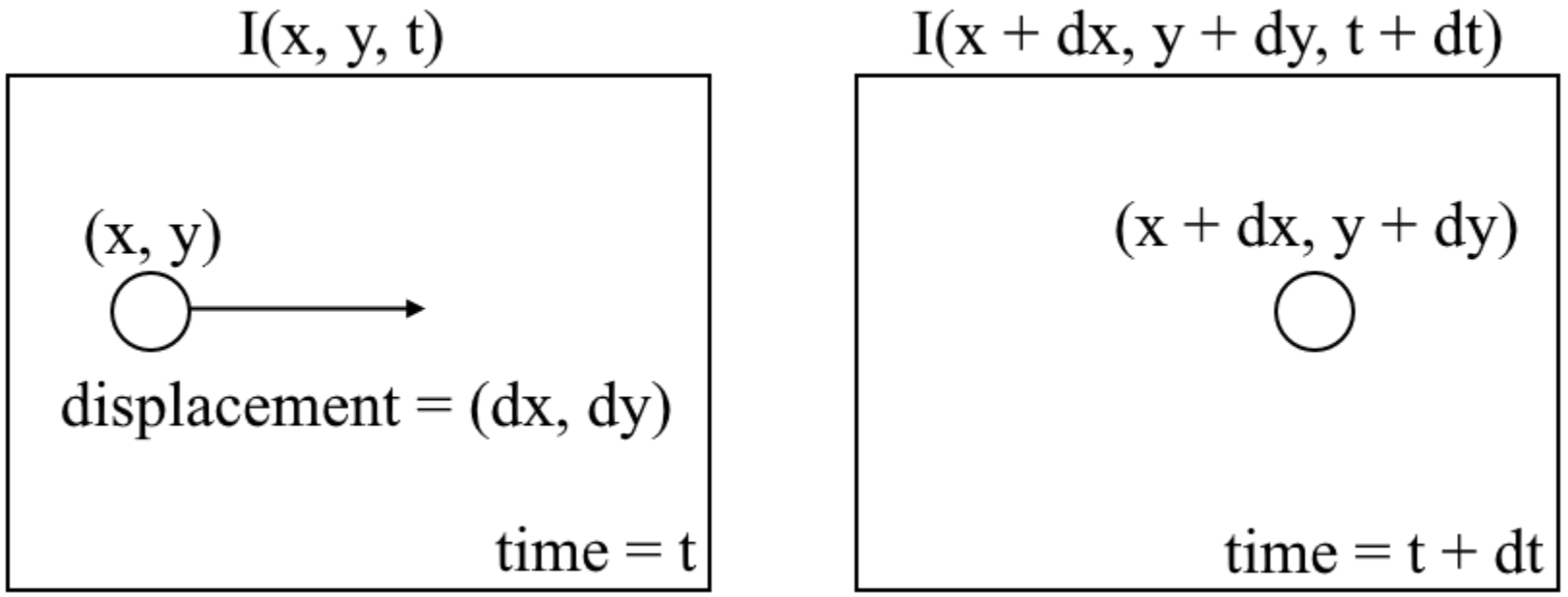
- Optical flow can be used for motion estimation



https://nanonets.com/blog/optical-flow/

What is Optical Flow?

- Motion of objects between consecutive frames of sequence
- Caused by the relative movement between the object and camera



https://nanonets.com/blog/optical-flow/

Optical Flow

- Brightness constancy

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$$

- Taylor series

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \ldots$$

- Truncating higher order terms and dividing by Δt

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0$$

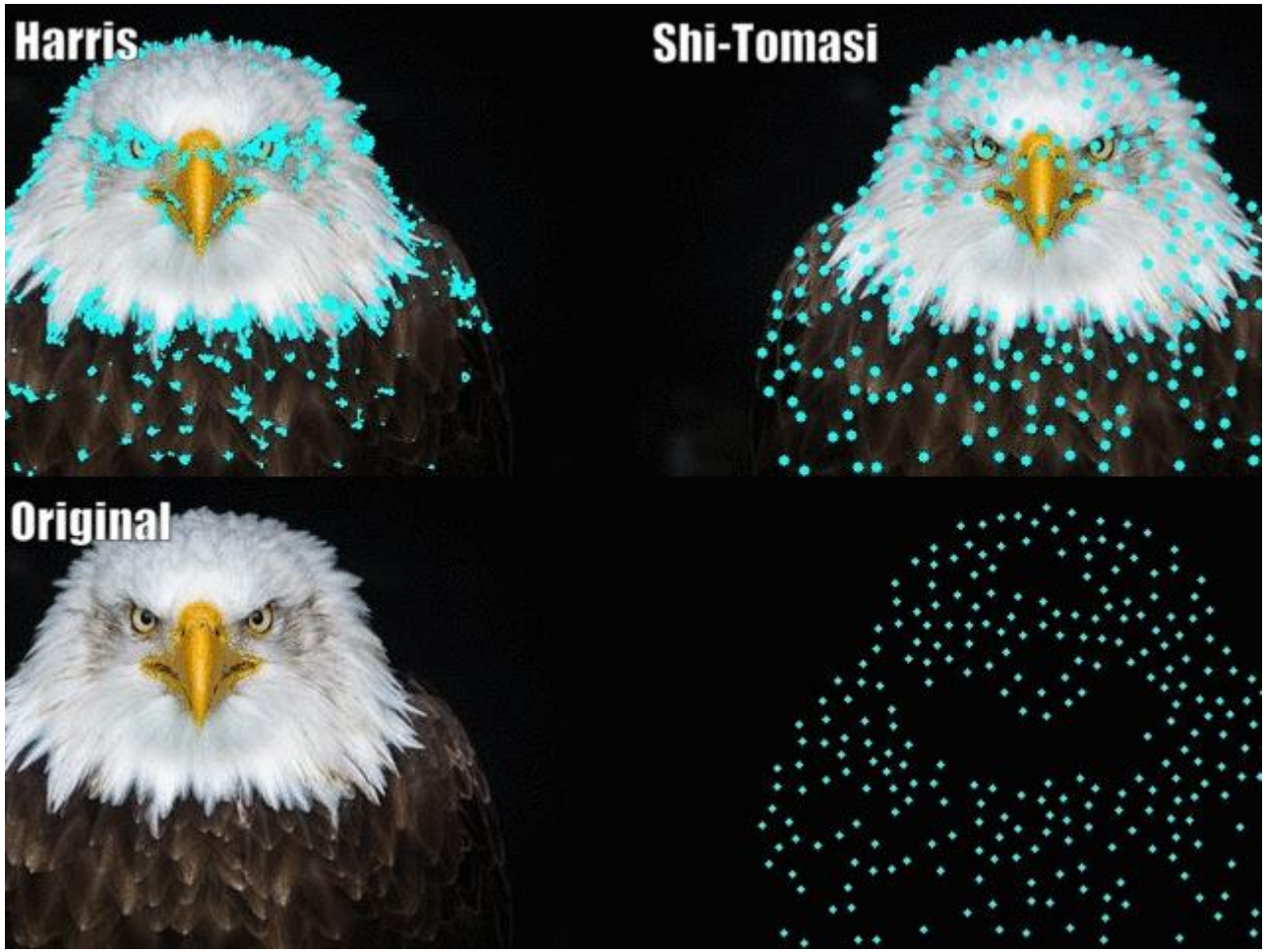
$$\Rightarrow \quad I_x V_x + I_y V_y = -I_t \quad \Rightarrow \quad \nabla I \cdot \vec{V} = -I_t$$

Spare vs. Dense Optical Flow

https://nanonets.com/blog/optical-flow/

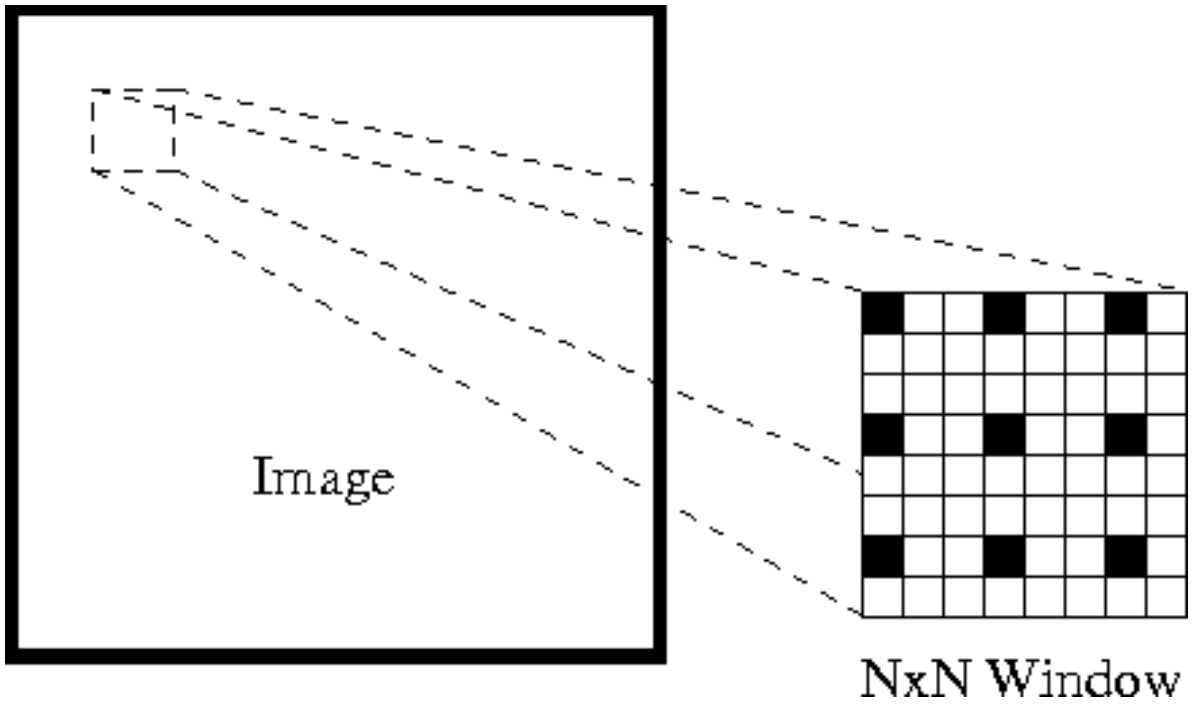
Selecting Feature Points

- Shi-Tomasi Corner Detector



Lucas-Kanade

- Take a small 3x3 window with the features detected by Shi-Tomasi



$$I_x(q_1)V_x + I_y(q_1)V_y = -I_t(q_1)$$
$$I_x(q_2)V_x + I_y(q_2)V_y = -I_t(q_2)$$
$$\vdots$$
$$I_x(q_n)V_x + I_y(q_n)V_y = -I_t(q_n)$$

Solve the 3x3 Optical Flow Equation

- $Av = b$

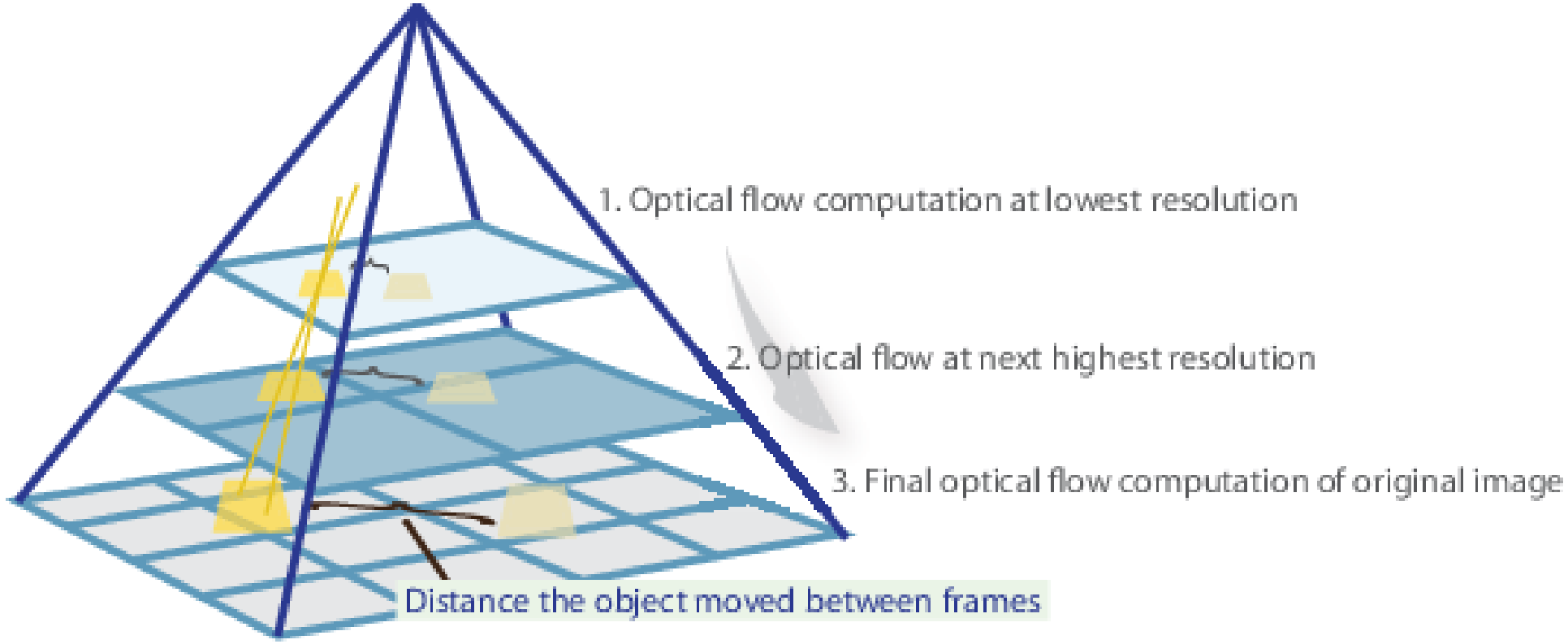
$$A = \begin{bmatrix} I_x(q_1) & I_y(q_1) \\ I_x(q_2) & I_y(q_2) \\ \vdots & \vdots \\ I_x(q_n) & I_y(q_n) \end{bmatrix} \qquad v = \begin{bmatrix} V_x \\ V_y \end{bmatrix} \qquad b = \begin{bmatrix} -I_t(q_1) \\ -I_t(q_2) \\ \vdots \\ -I_t(q_n) \end{bmatrix}$$

Least Square Fitting

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} \sum_i I_x(q_i)^2 & \sum_i I_x(q_i) I_y(q_i) \\ \sum_i I_y(q_i) I_x(q_i) & \sum_i I_y(q_i)^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i I_x(q_i) I_t(q_i) \\ -\sum_i I_y(q_i) I_t(q_i) \end{bmatrix}$$

Beyond 3x3 Window

- OpenCV adopts pyramid for LK optical flow



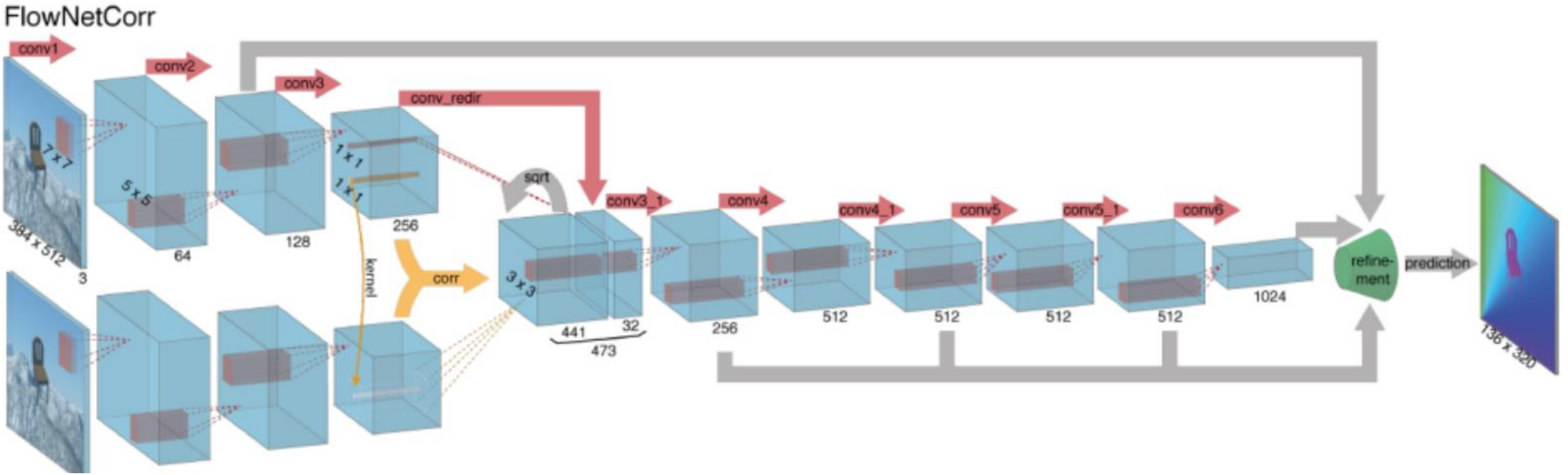
Farneback Optical Flow (Dense)

- Approximate each neighborhood of both frames by quadratic polynomials

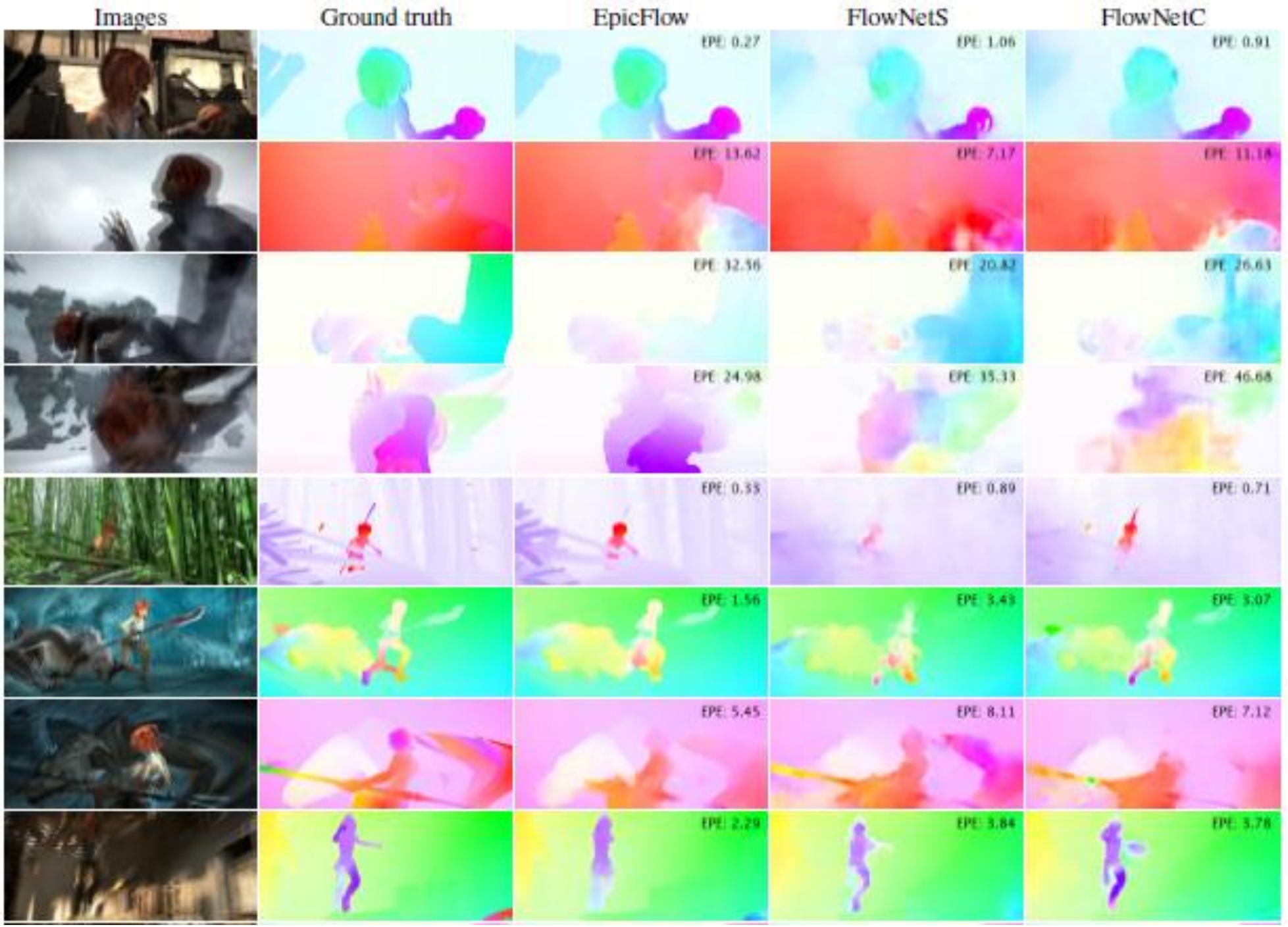
$$f(\mathbf{x}) \sim \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$$

FlowNet: Learning Optical Flow with Convolutional Networks

- Correlation layer compares each patch from ConvNet branch 1 with each path from ConvNet branch 2.

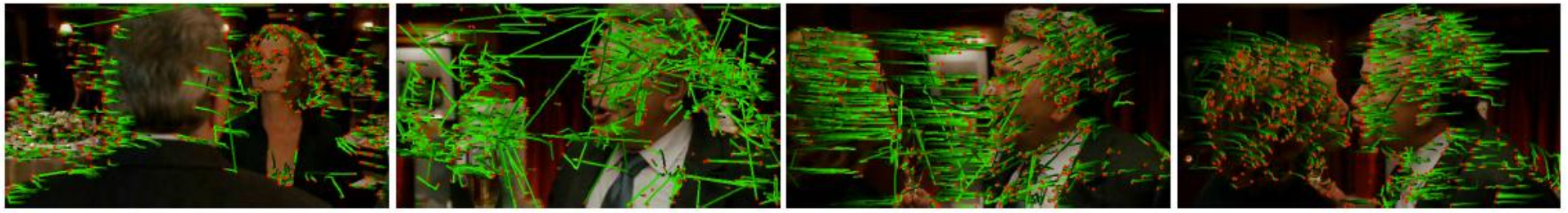


Synthetic Dataset

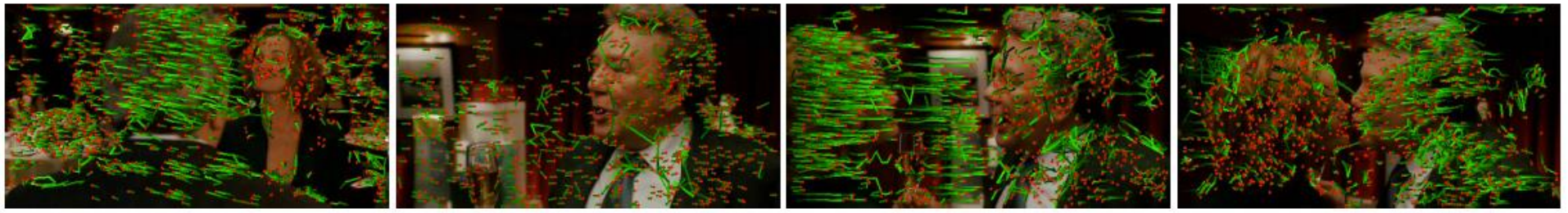


Dense Trajectories

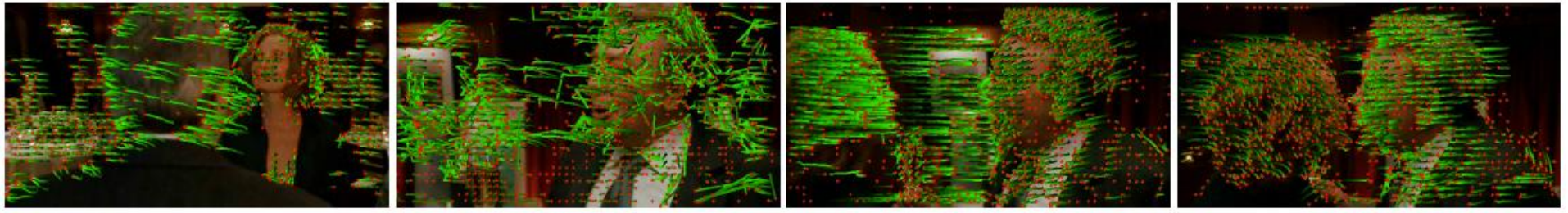
- H. Wang, et al. Dense trajectories and motion boundary descriptors for action recognition. IJCV, 2013
- Tracking dense pixels has shown better than KLT (corner) features and SIFT points



KLT trajectories



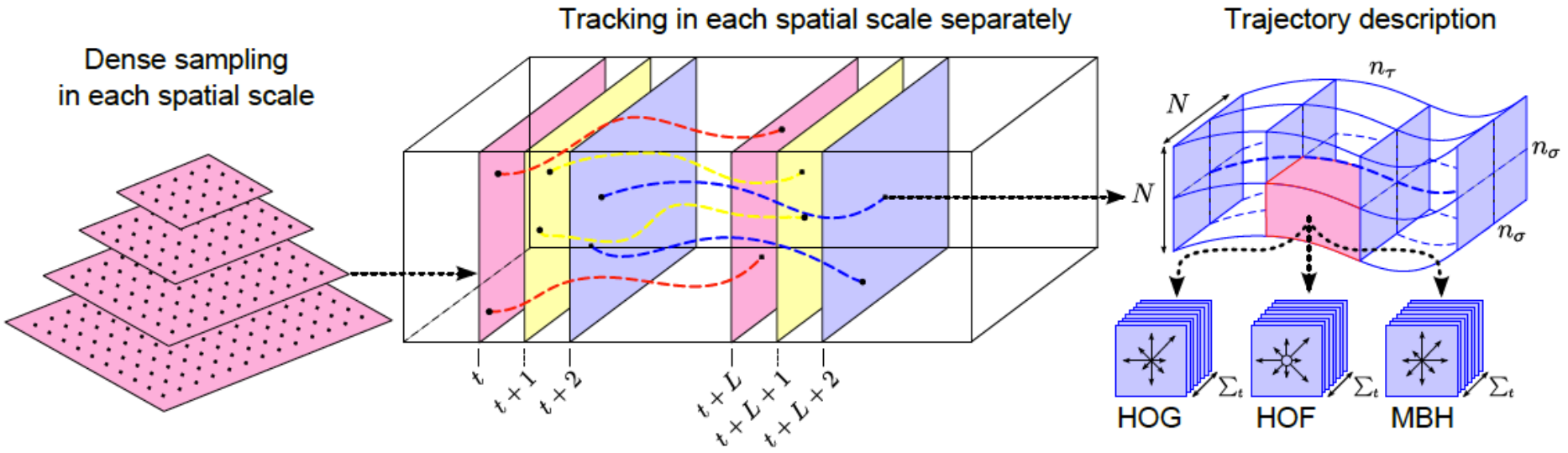
SIFT trajectories



Dense trajectories

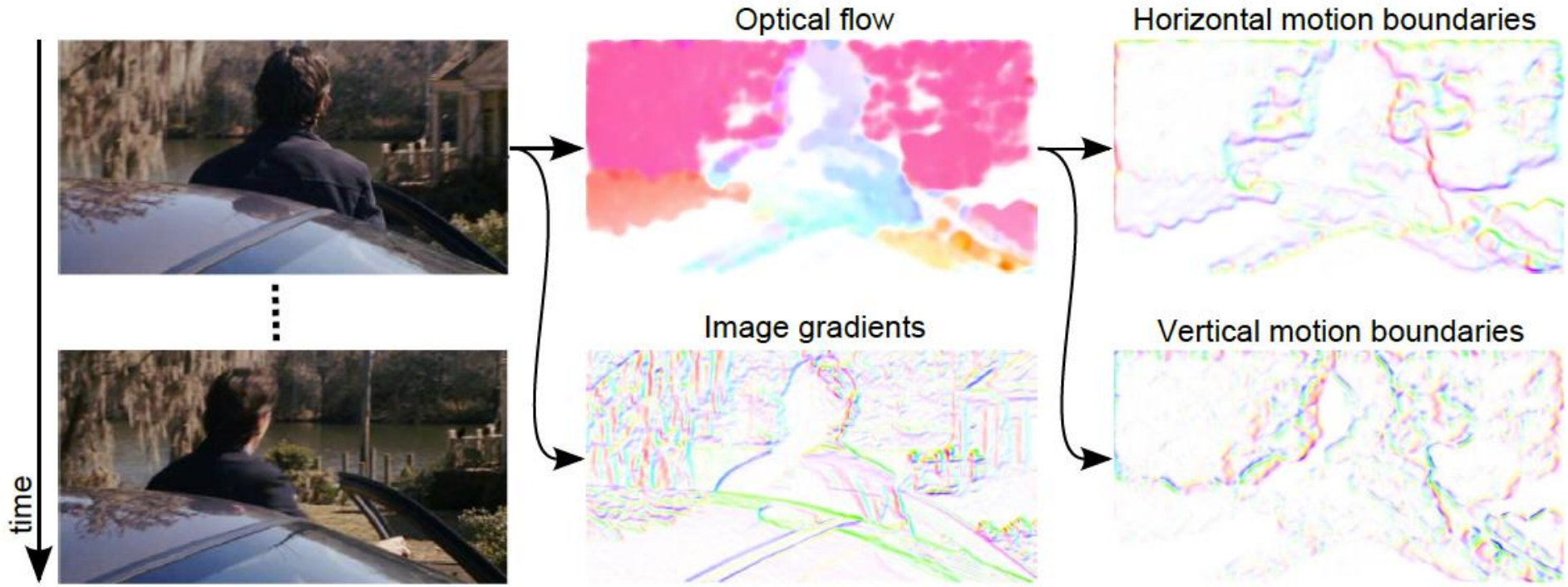
Extracting Dense Trajectories

1. Sample feature points with 5-pixel step
2. Remove features in homogeneous areas
3. Track optical flows smoothed by 3x3 median filter in 15 frames
4. Extract HOG, HOF and MBH along each trajectory



Motion Boundary Histogram (MBH)

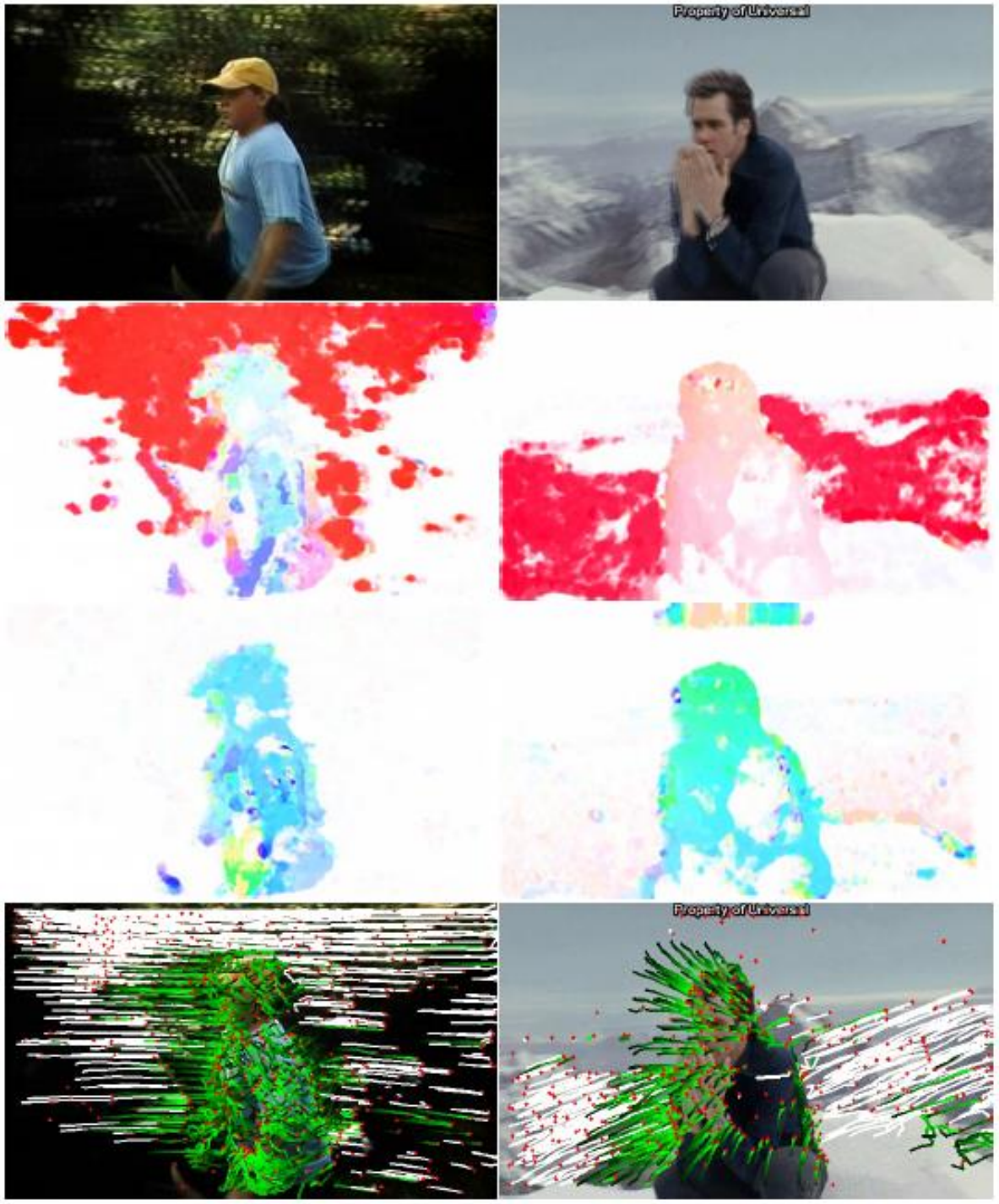
- First proposed by Dalal & Triggs
  - "Human Detection Using Oriented Historgrams of Flow and Appearance", *ECCV*, 2006
- Show best results on HMDB51 & TRECVID 2011*
- Based on oriented histogram of differential optical flows
- Can effectively cancel camera motions



*A. Tamarakar, et al., Evaluation of Low-Level Features and their Combinations for Complex Event Detection in Open Source Videos, *CVPR,* 2012

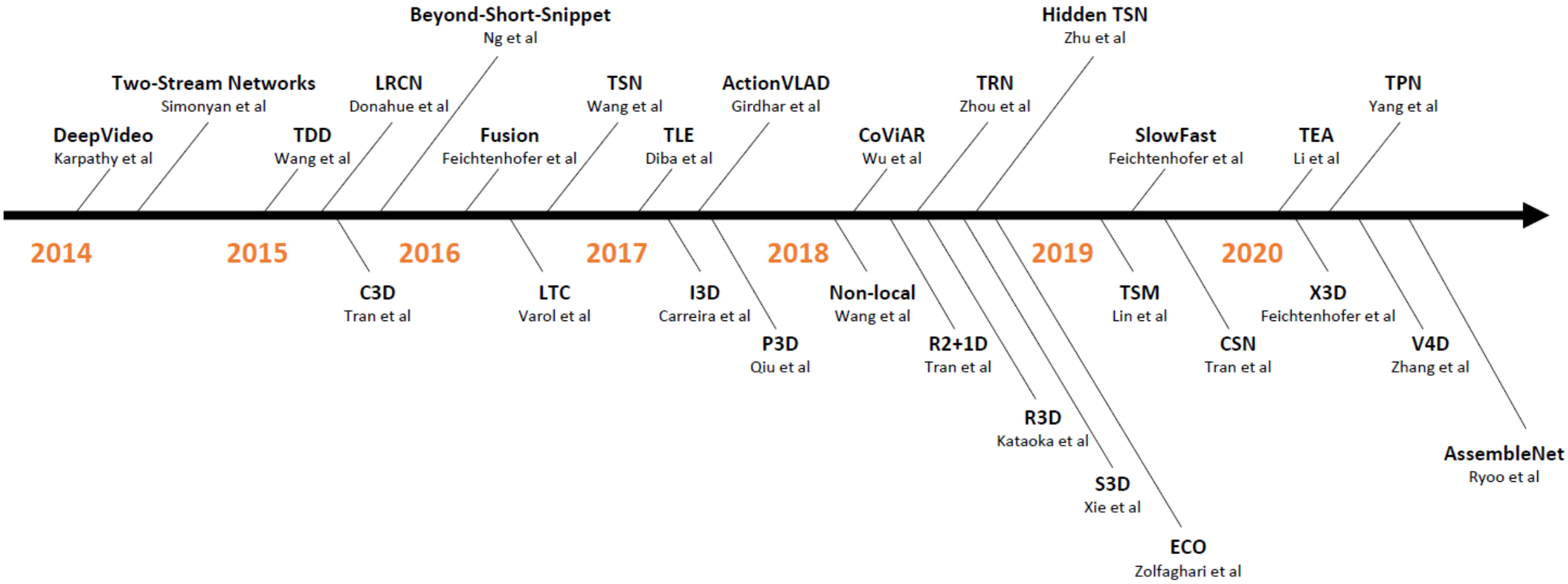
Improved Trajectories

- Cancel camera motions by matching SURF points & dense optical flows between frames
- A human detector is also used

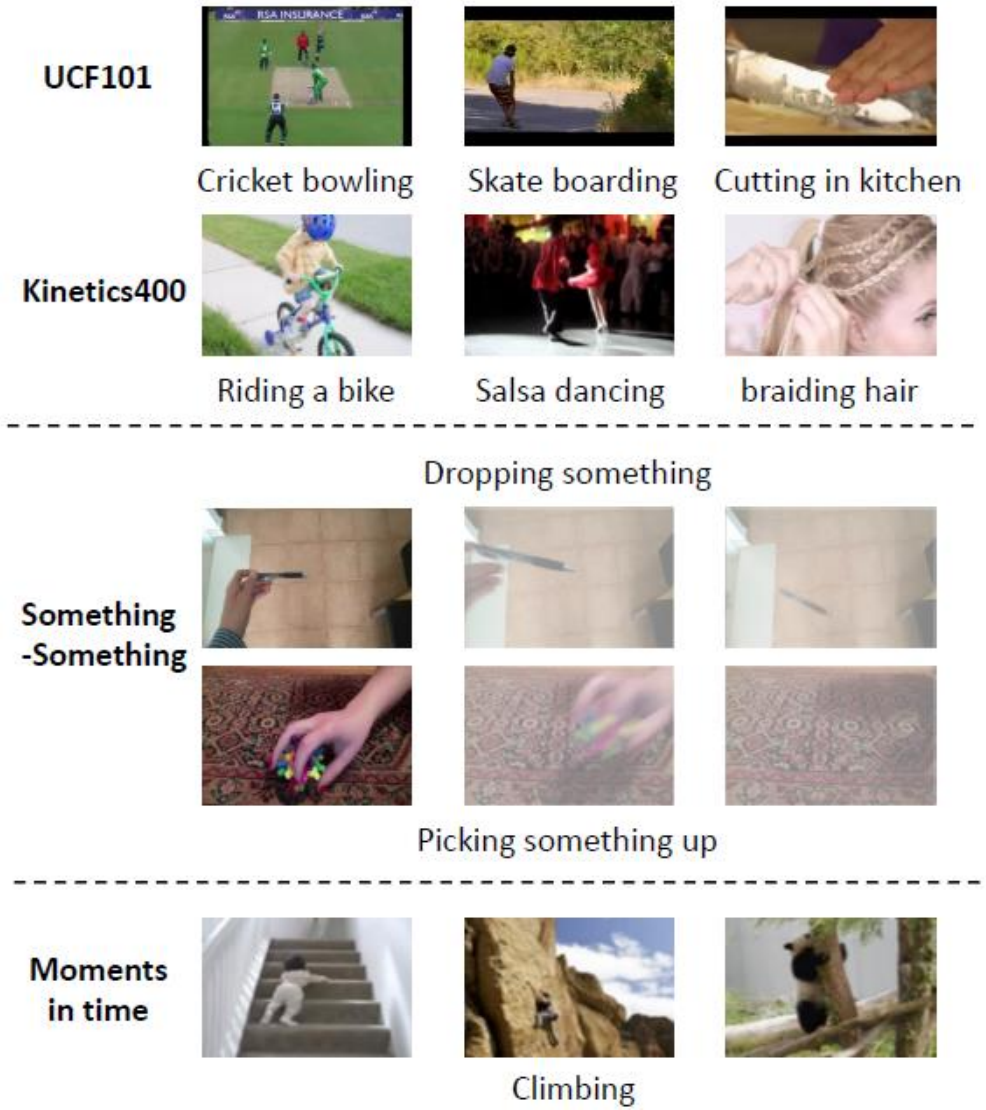


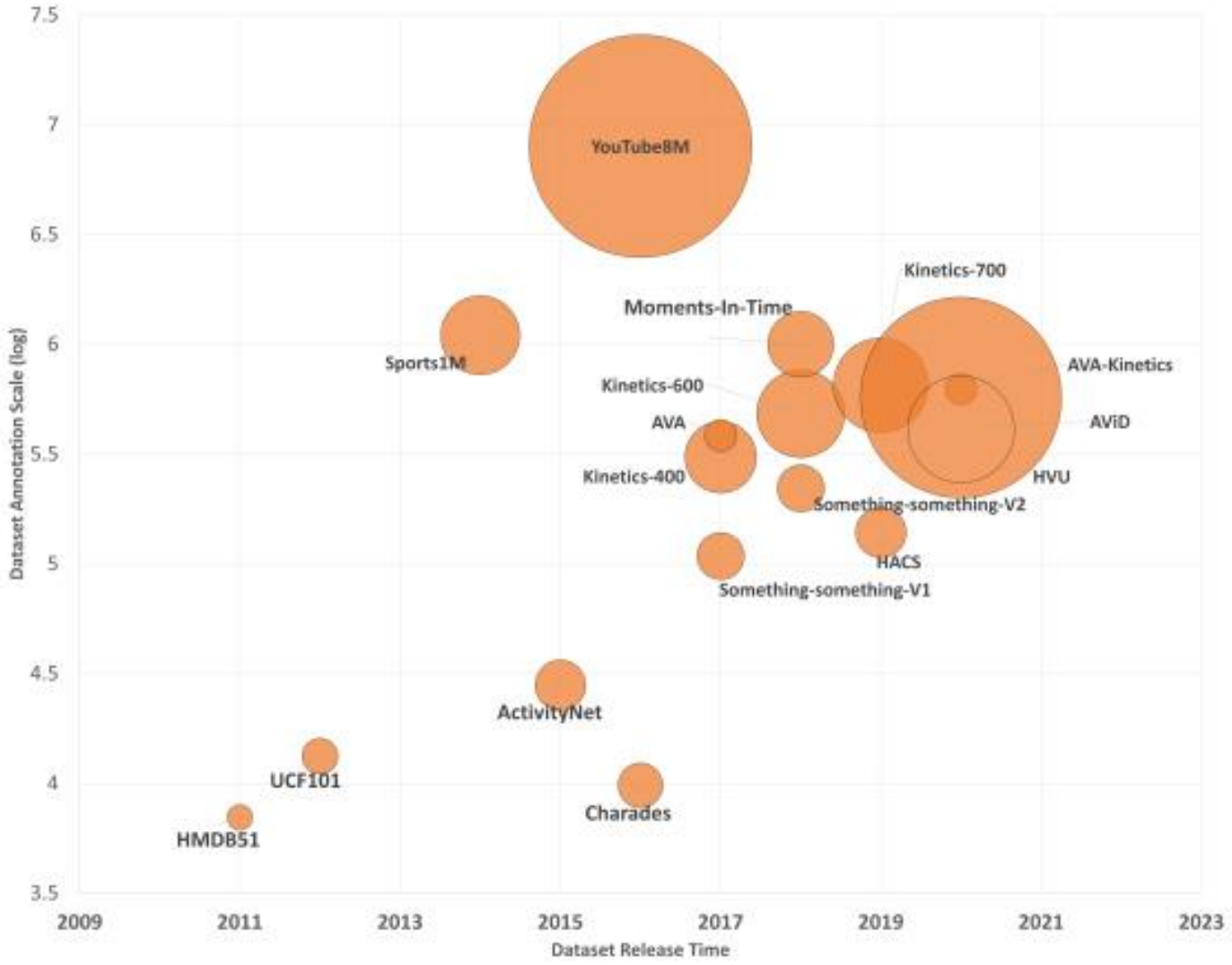
Recent DL Models for Action Recognition

- Yi Zhu et al., “A Comprehensive Study of Deep Video Action Recognition,” Amazon Web Services, 2020



Action Recognition Datasets



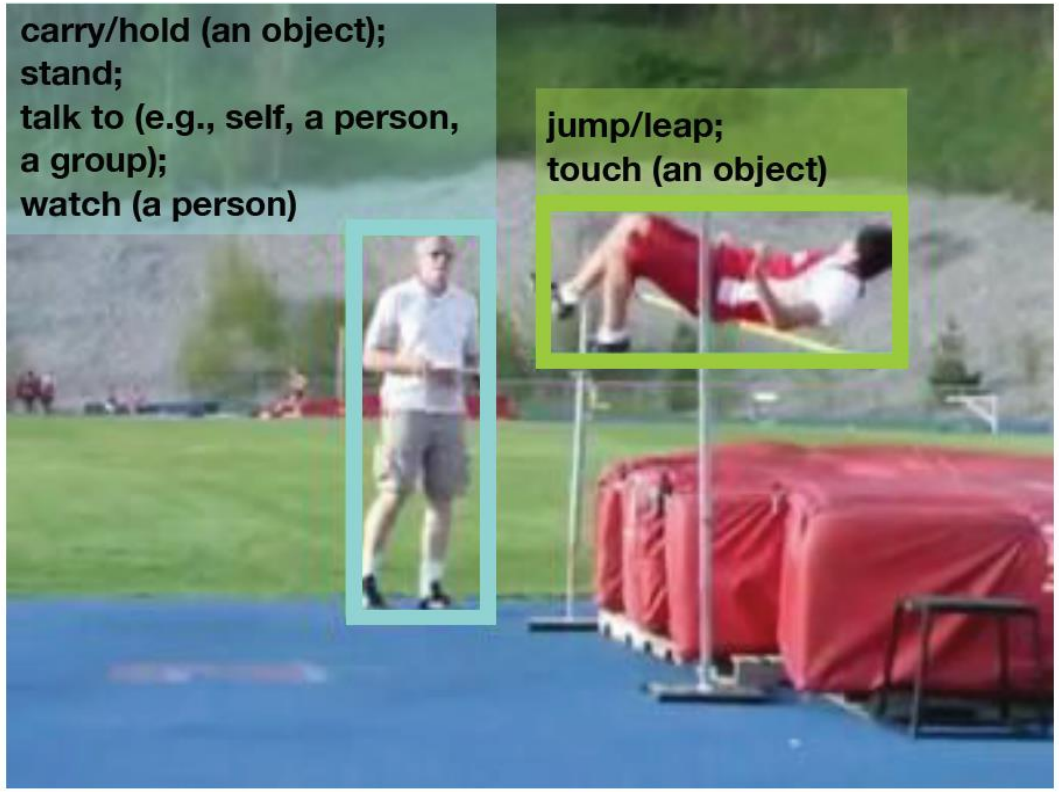


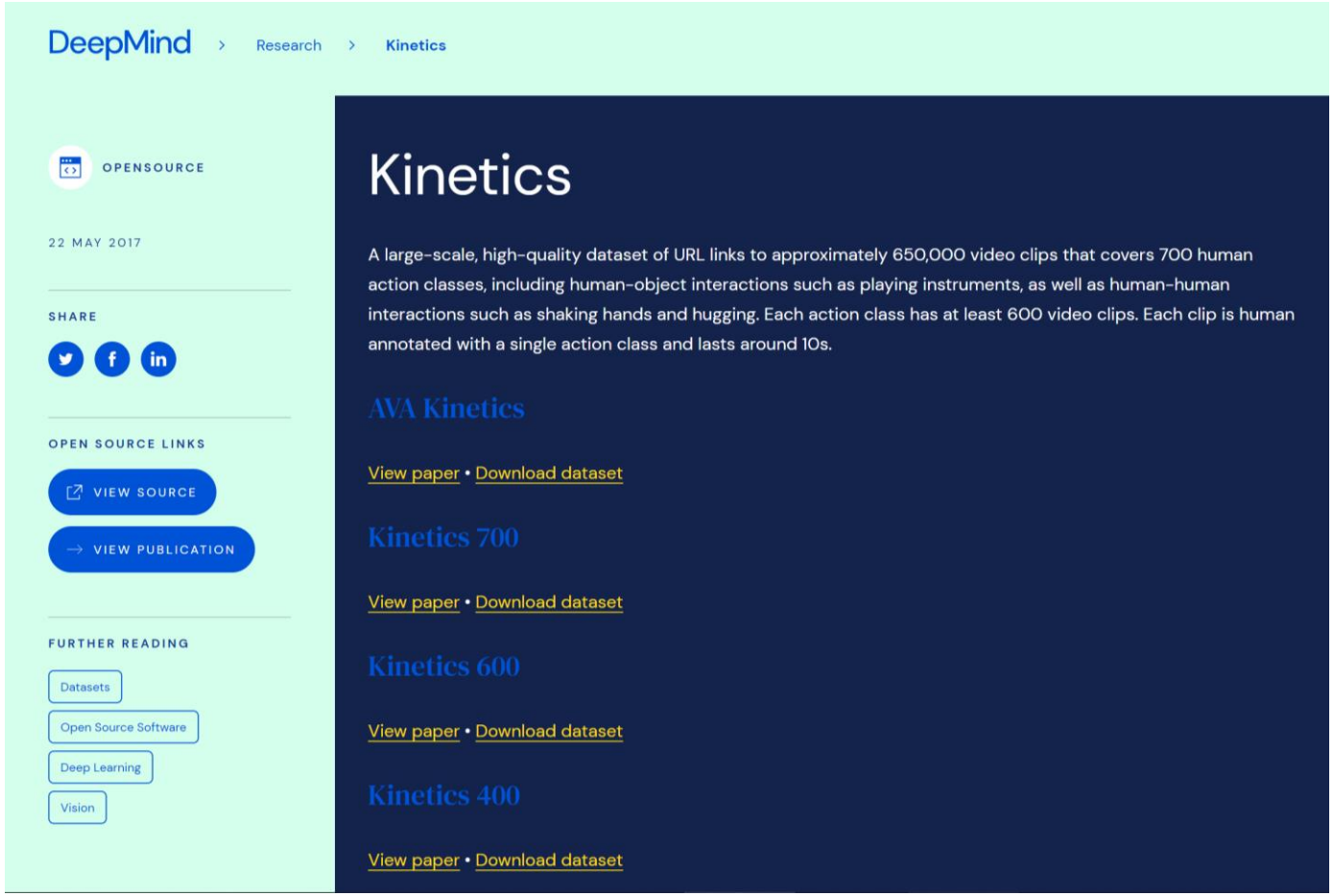
YouTube-8M

- https://research.google.com/youtube8m/

DeepMind Kinetics Dataset

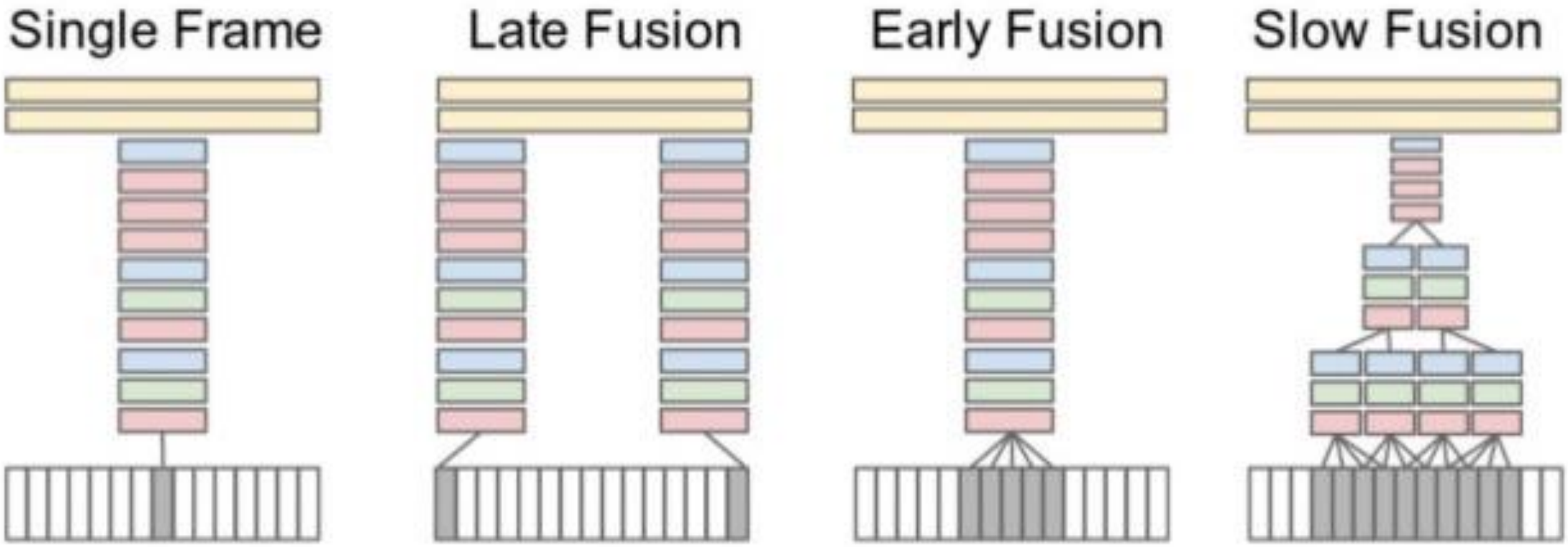
- https://deepmind.com/research/open-source/kinetics





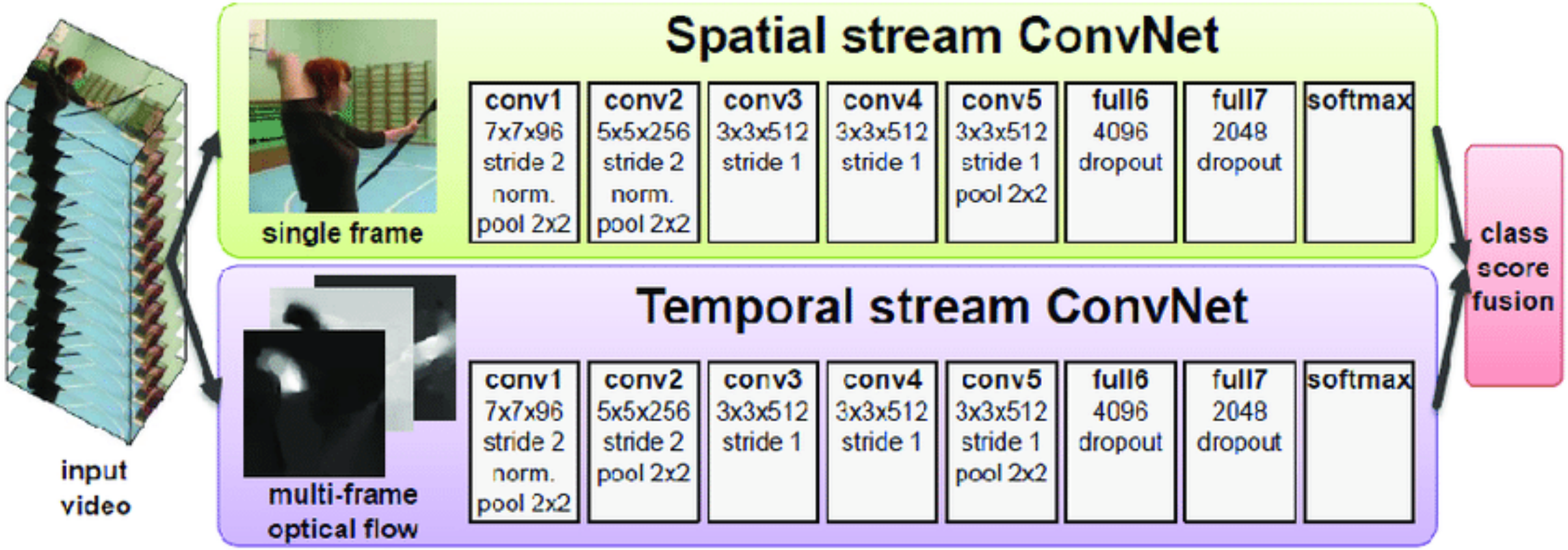
Learning Optical Flow with Deep Learning?

- Karpathy et al., “Large-scale Video Classification with Convolutional Neural Networks”, 2014



http://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review

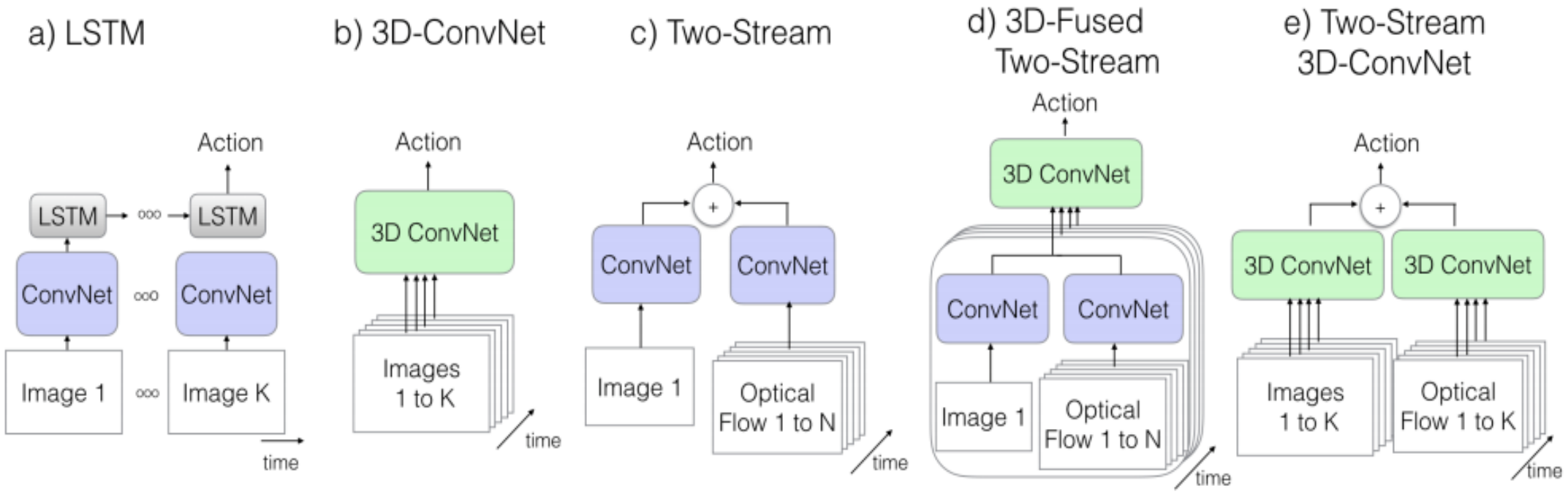
Two-stream Convolutional Neural Networks



Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *NIPS*

State-of-the-Arts

- LRCN [15], C3D [16], Conv3D & Attention [17], TwoStreamFusion [18], TSN [19], ActionVlad [20], HiddenTwoStream [1] I3D [21] and T3D [22]

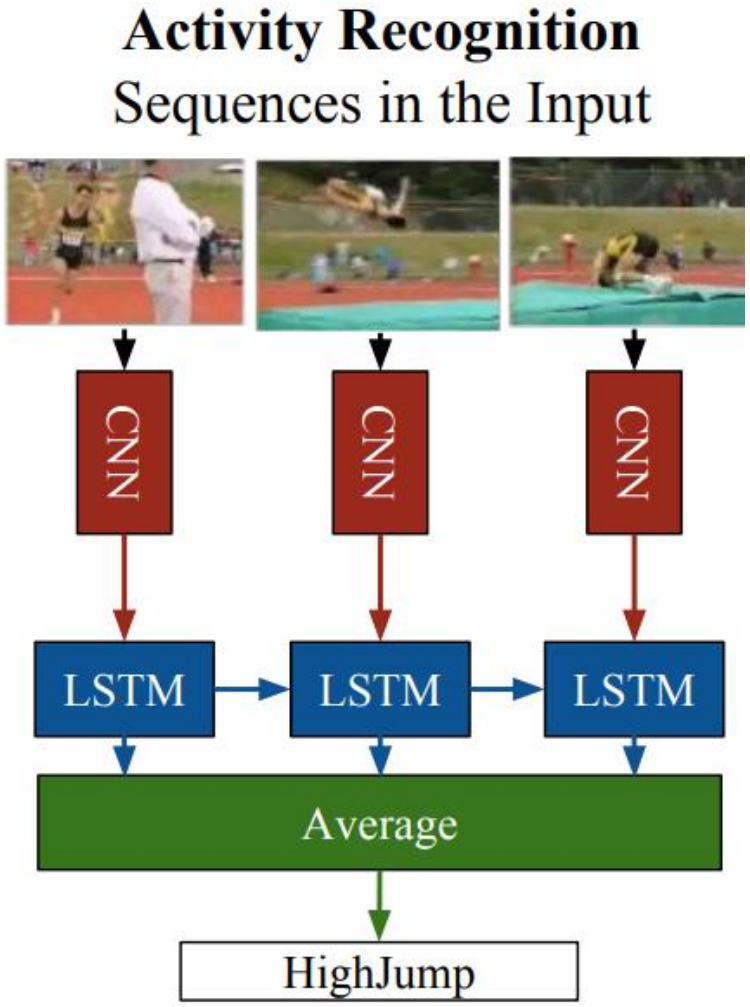


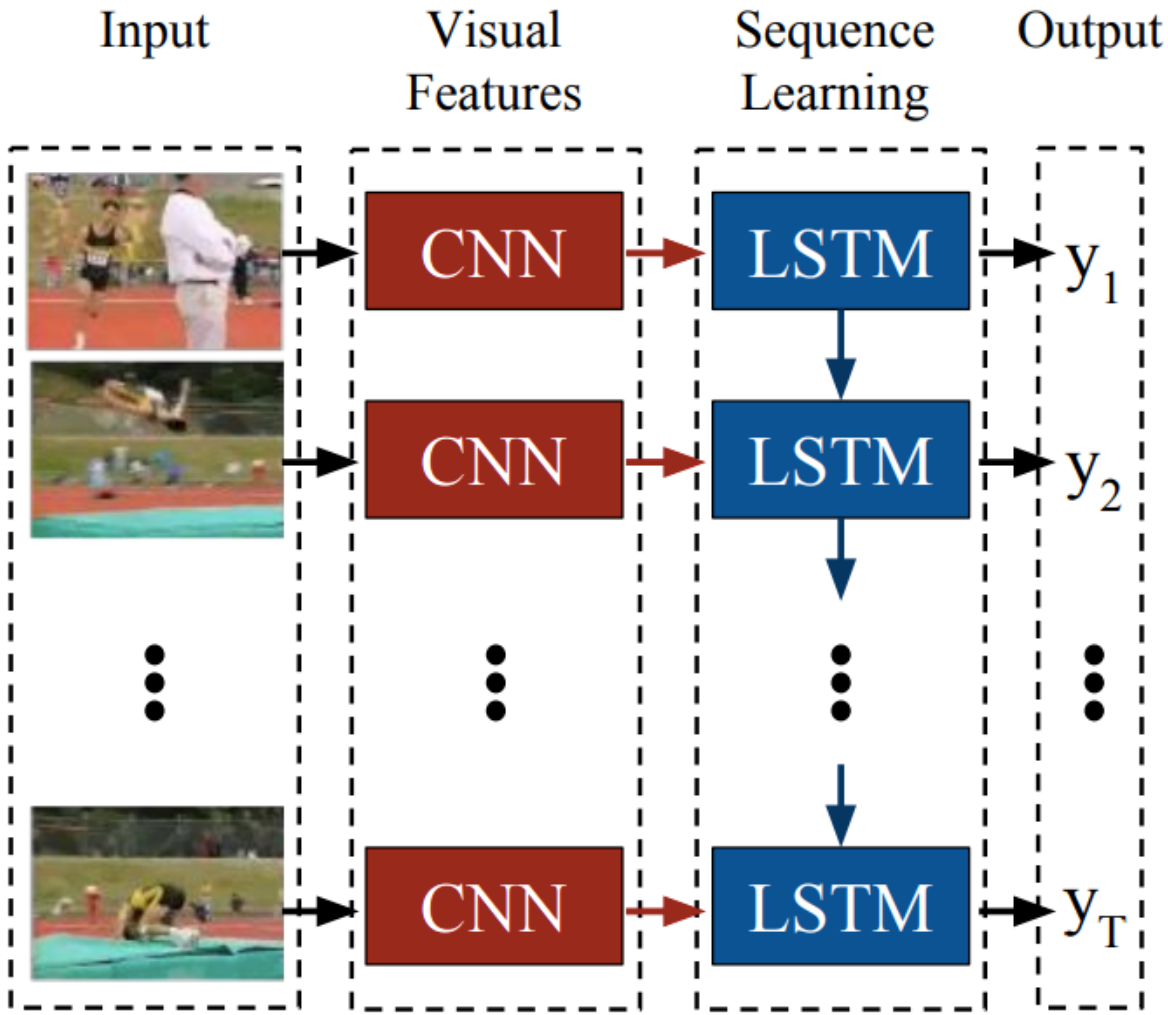
http://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review

Long-term Recurrent Convolutional Networks

- Donahue et al., “Long-term Recurrent Convolutional Networks for Visual Recognition and Description,” 2014 (Arxiv Link)
- Key Contributions:
  - Building on previous work by using RNN as opposed to stream based designs
  - Extension of encoder-decoder architecture for video representations
  - End-to-end trainable architecture proposed for action recognition

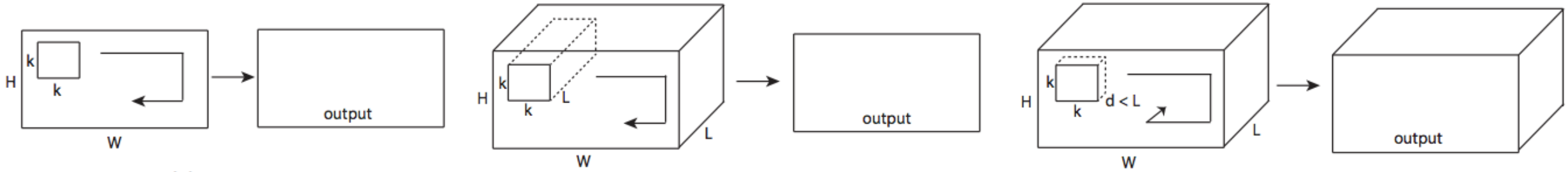
Long-term Recurrent Convolutional Networks





3D Convolutional Networks (C3D)

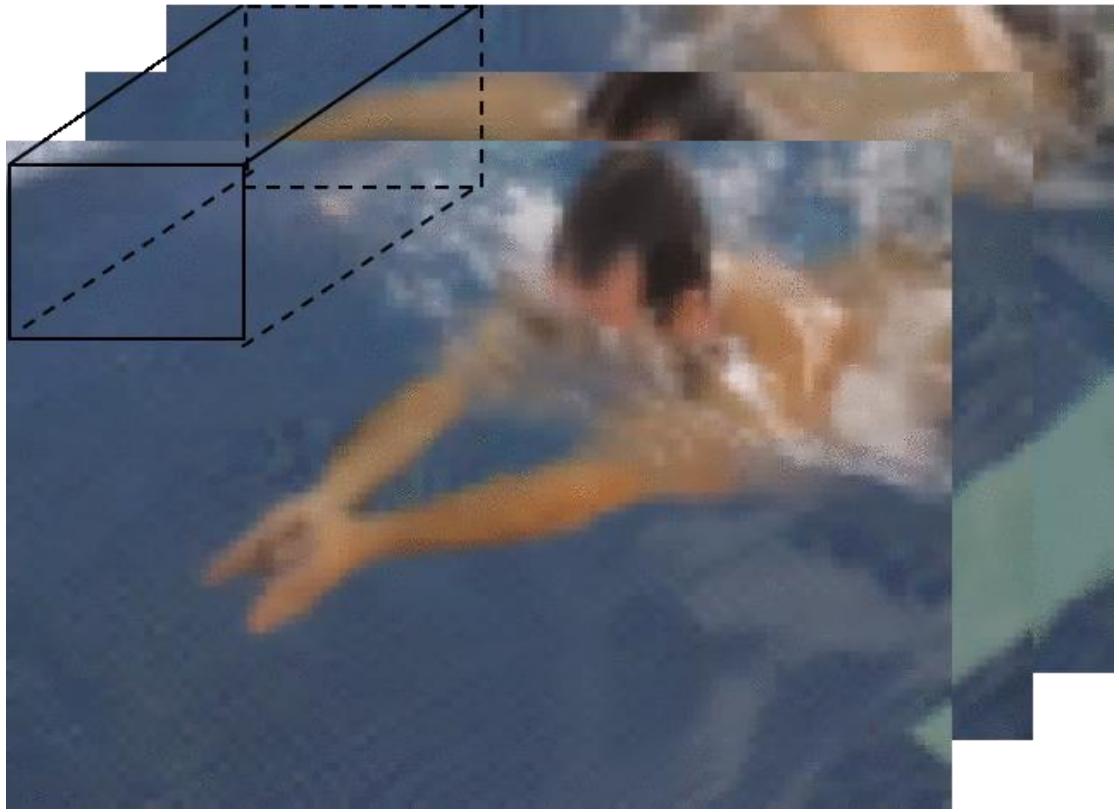
- Du Tran et al., “Learning Spatiotemporal Features with 3D Convolutional Networks,” 2014 (Arxiv Link)
- Key Contributions
  - Repurposing 3D convolutional networks as feature extractors
  - Extensive search for best 3D convolutional kernel and architecture
  - Using deconvolutional layers to interpret model decision



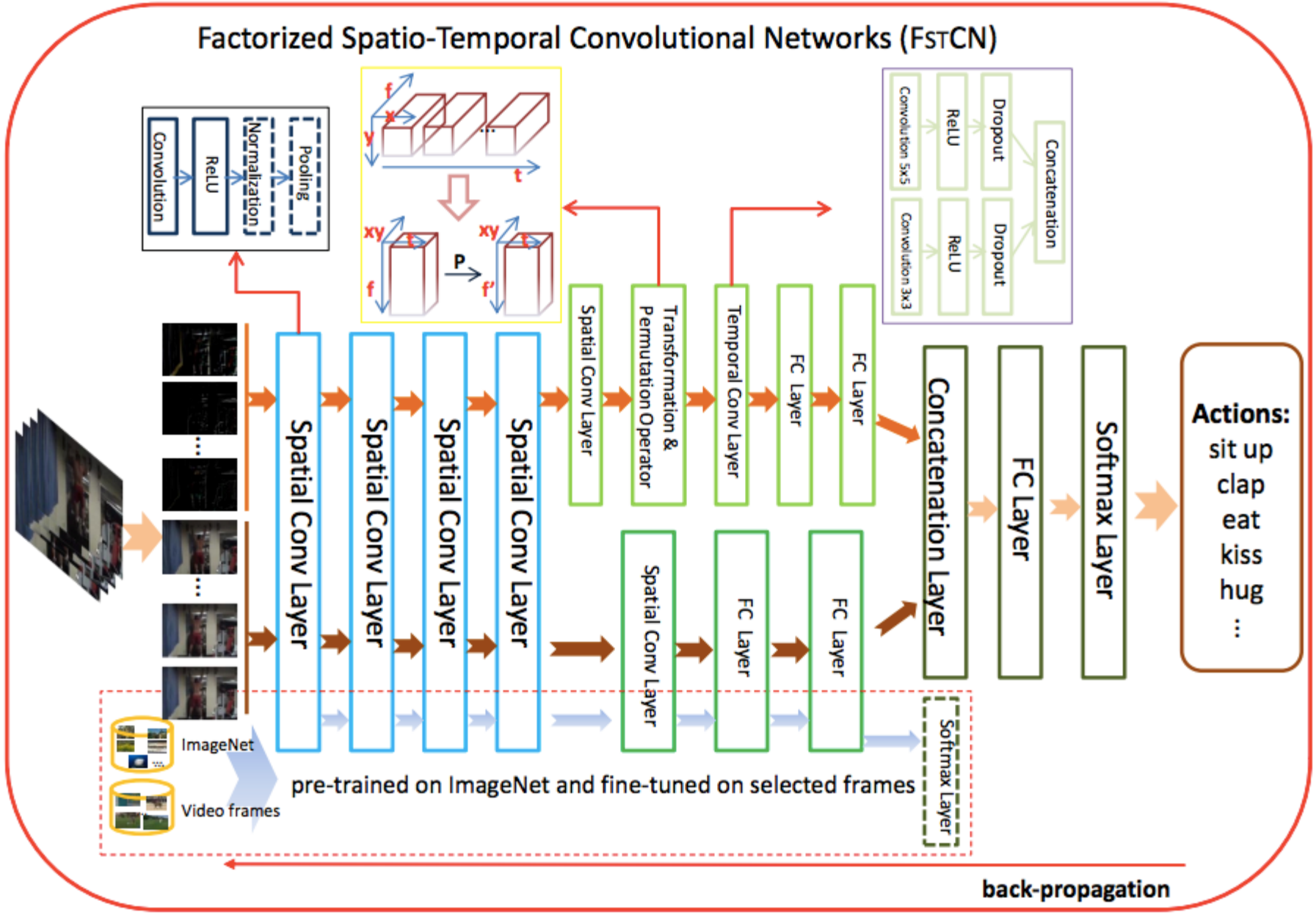
(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

3D Convolutional Networks (C3D)

- Extract features on 2-second clip
- C3D tends to focus on spatial appearance in first few frames and tracked the motion in the subsequent frames



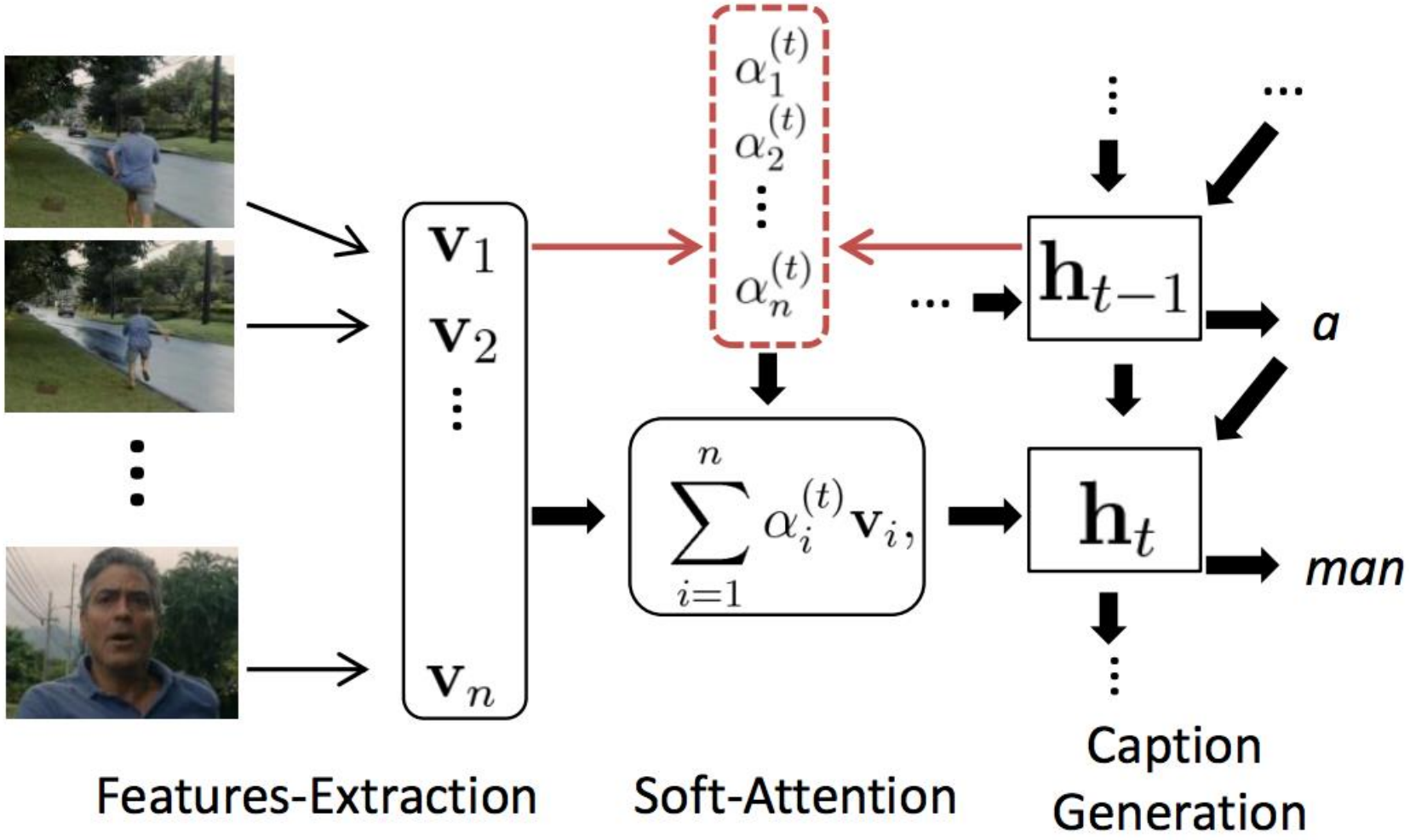
F
N



[Human Action Recognition using Factorized Spatio-Temporal Convolutional Networks]

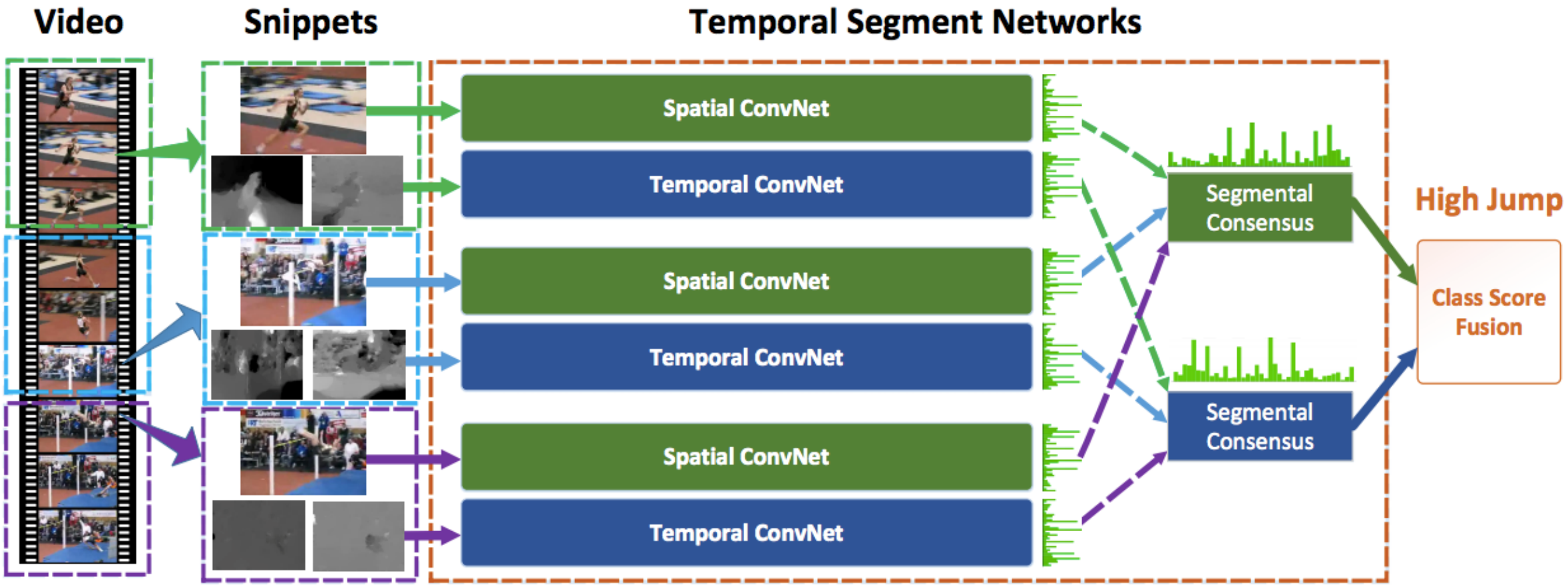
Conv3D + Attention

- Yao et al., Describing Videos by Exploiting Temporal Structure, 2015



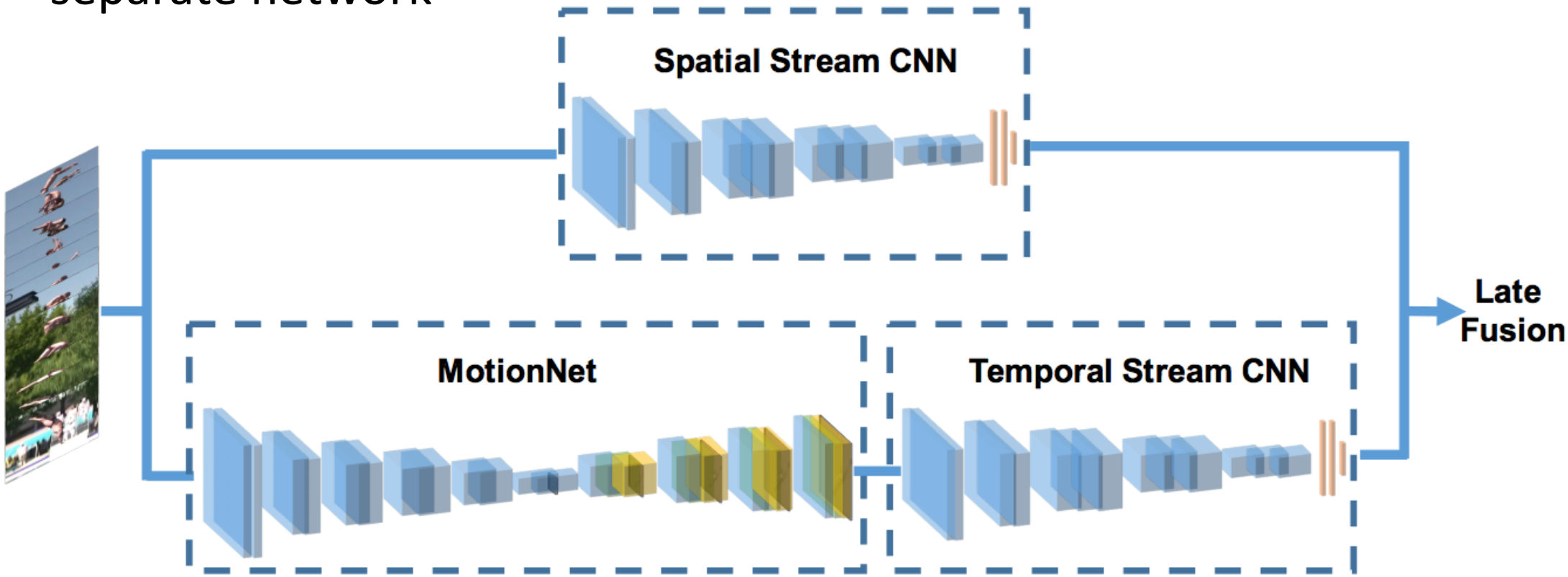
Temporal Segment Networks (TSN)

- Wang et al., “Temporal Segment Networks: Towards Good Practices for Deep Action Recognition”, 2016
- Sampling clips sparsely across the video to better model long range

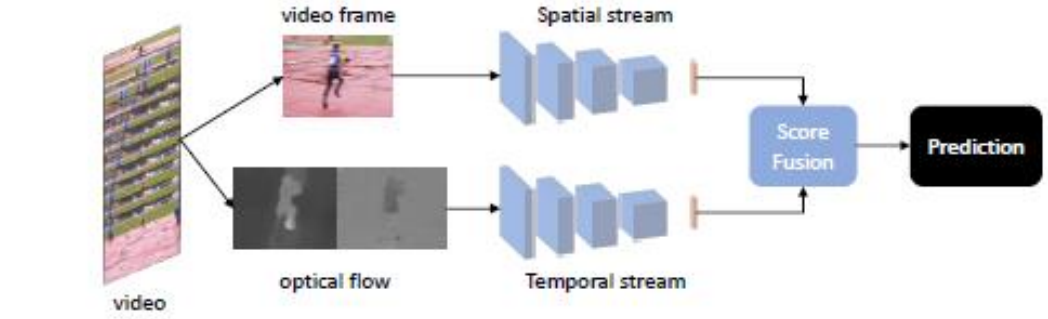


Hidden Two Stream

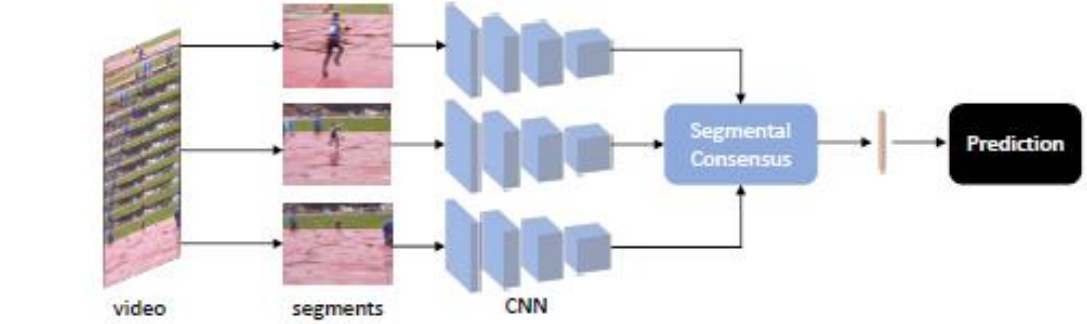
- Zhu et al., “Hidden Two-Stream Convolutional Networks for Action Recognition,” 2017
- Novel architecture for generating optical flow input on-the-fly using a separate network



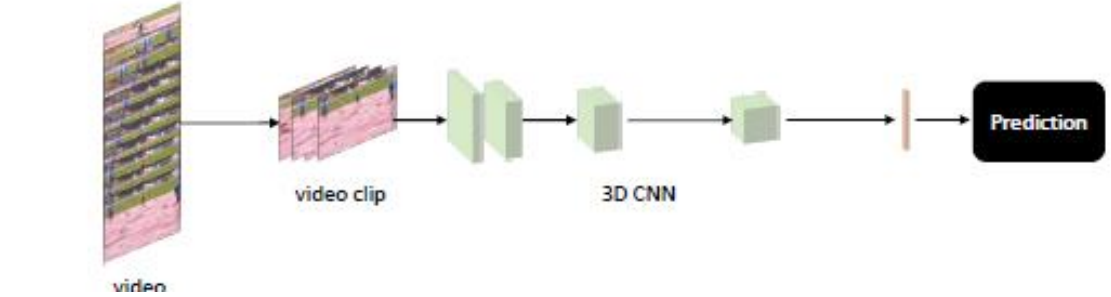
5 Important Action CNN Models



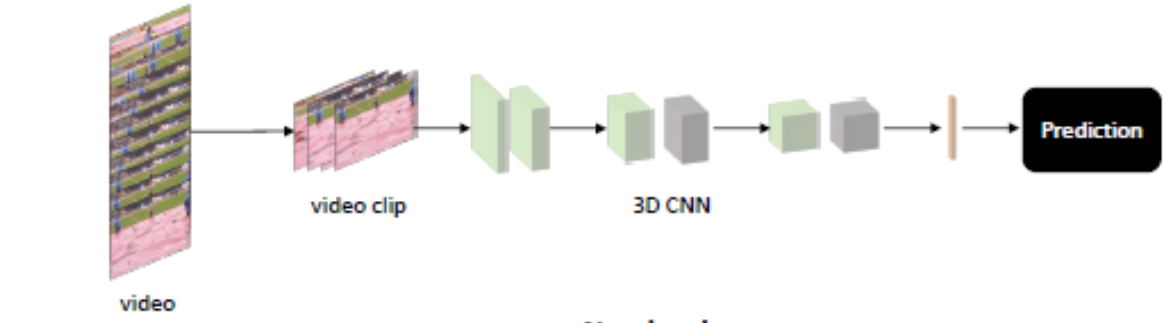
Two-stream Networks



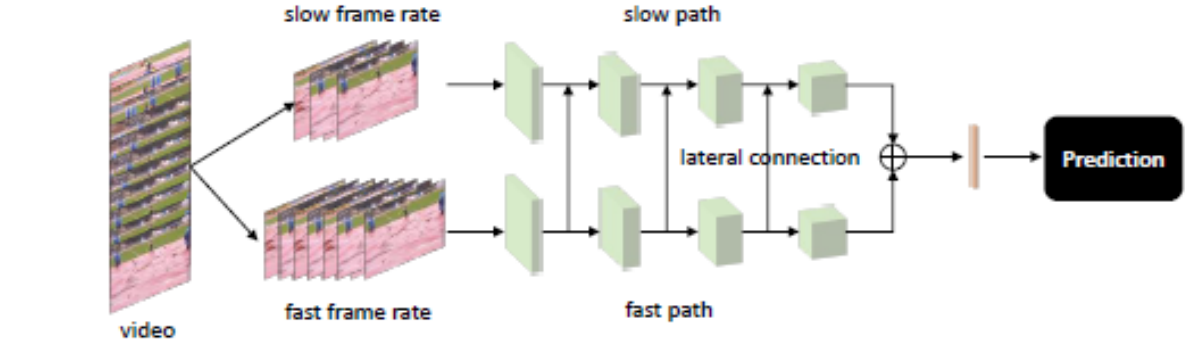
Temporal Segment Networks (TSN)



I3D



Non-local

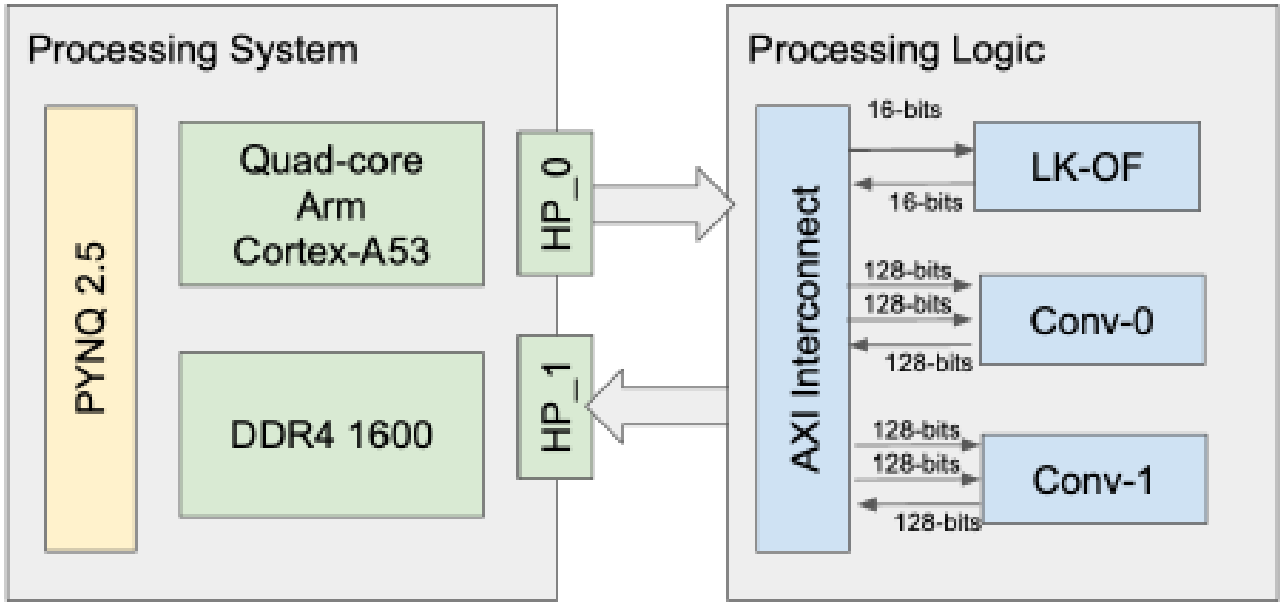


SlowFast

Efficient Two-stream Action Recognition on FPGA

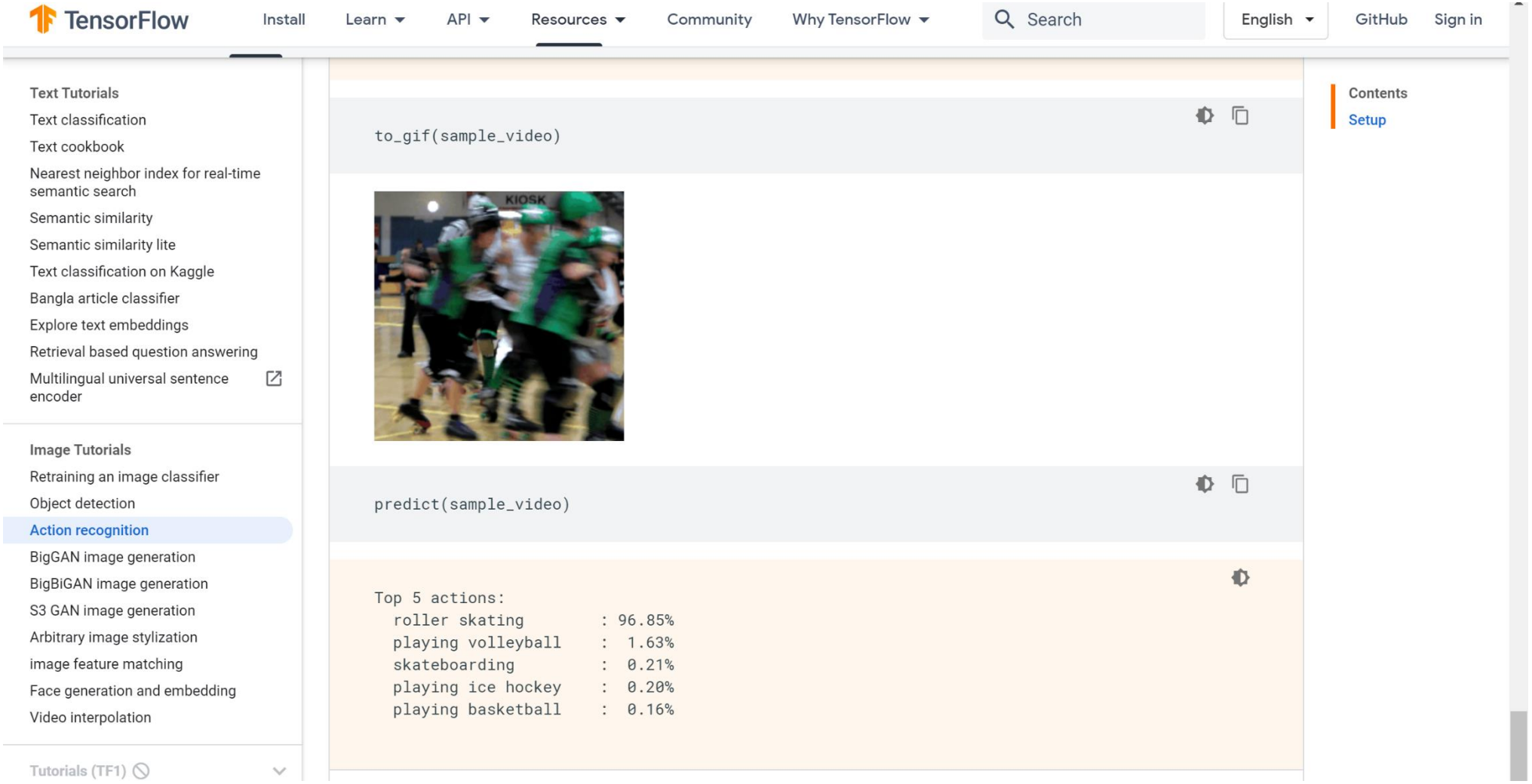
Jia-Ming Lin[1], Kuan-Ting Lai[2], Bin-Ray Wu[1] and Ming-Syan Chen[1]
[1]National Taiwan University
[2]National Taipei University of Technology
ktlai@ntut.edu.tw, {jmlin, brwu, mschen}@arbor.ee.ntu.edu.tw

- CVPR 2021 ECV Workshop
- Contributions:
  - Has 10x less operations than other models with little accuracy drop (<2%)
  - Use only 2D CNN operations
  - Processing both spatial and temporal streams parallelly.



TF-Hub Action Recognition Model

- https://www.tensorflow.org/hub/tutorials/action_recognition_with_tf_hub



References

- https://nanonets.com/blog/optical-flow/
- https://neurohive.io/en/datasets/new-datasets-for-action-recognition/
- http://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review
- Yi Zhu et al., “A Comprehensive Study of Deep Video Action Recognition,” Amazon Web Services, 2020