

Action Recognition and Video Event Detection



Human Action Recognition



UCF-101

<http://www.thumos.info/>



Applications of Video Event Detection

Video Search



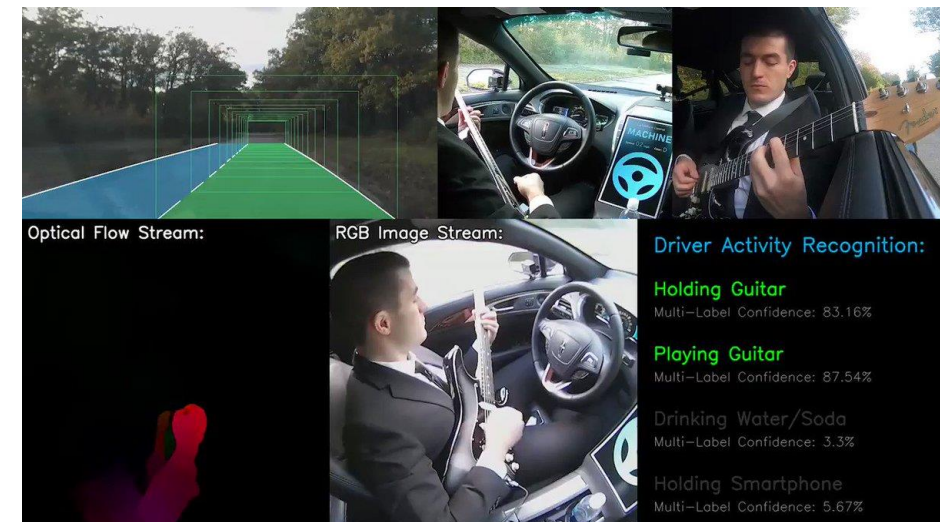
Surveillance Video Analysis



Drone Action Recognition



Driver Activity Detection



Complex Video Events

- Basic event or action detection



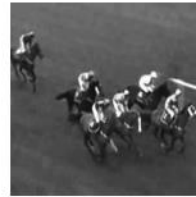
push



pushup



ride
bike



ride
horse



run



shake
hands



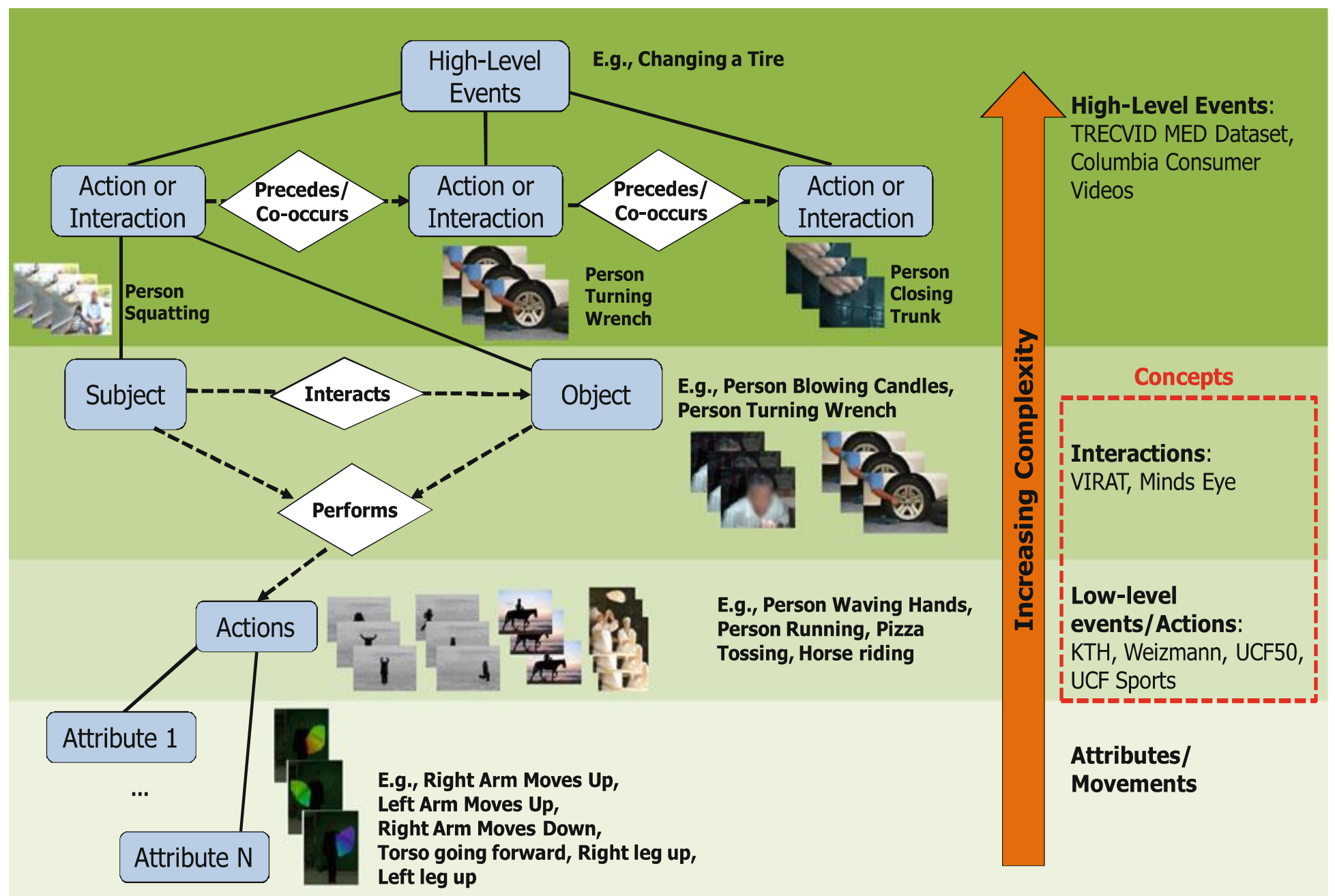
shoot
ball

- Complex or high-level event detection



Wedding Ceremony

Detect Complex (High-level) Events



- Jiang et al., high level events recognition in unconstrained videos, 2012 ⁶

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

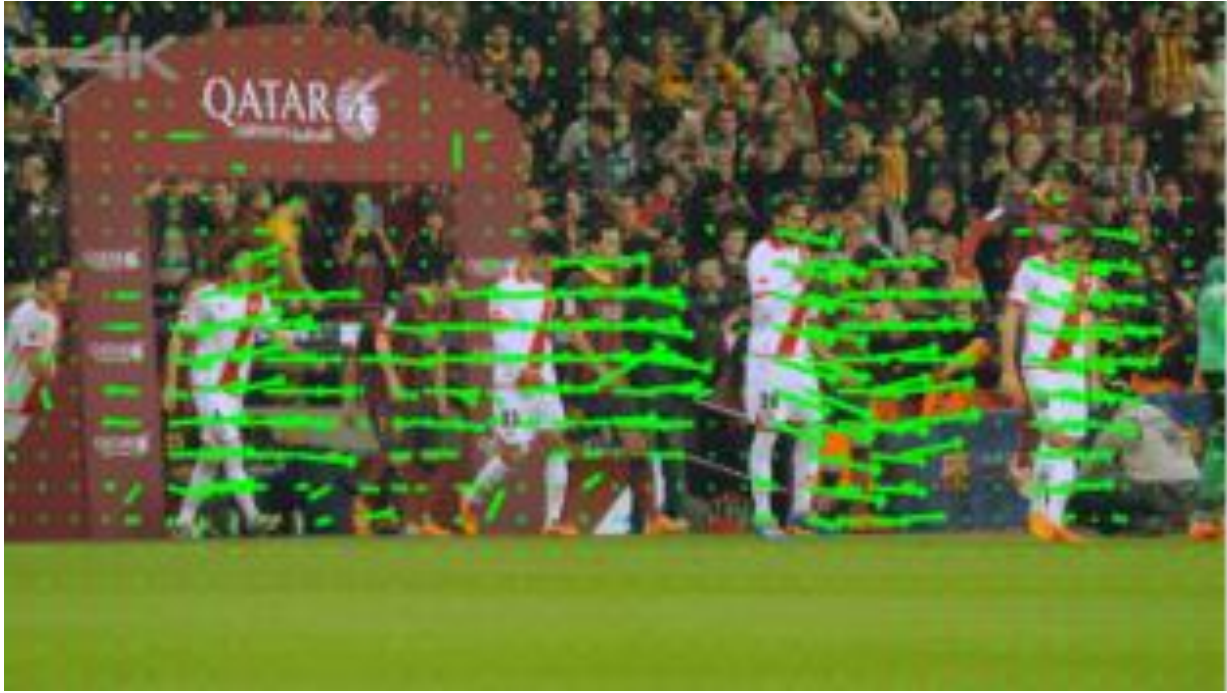


Working on a sewing project

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

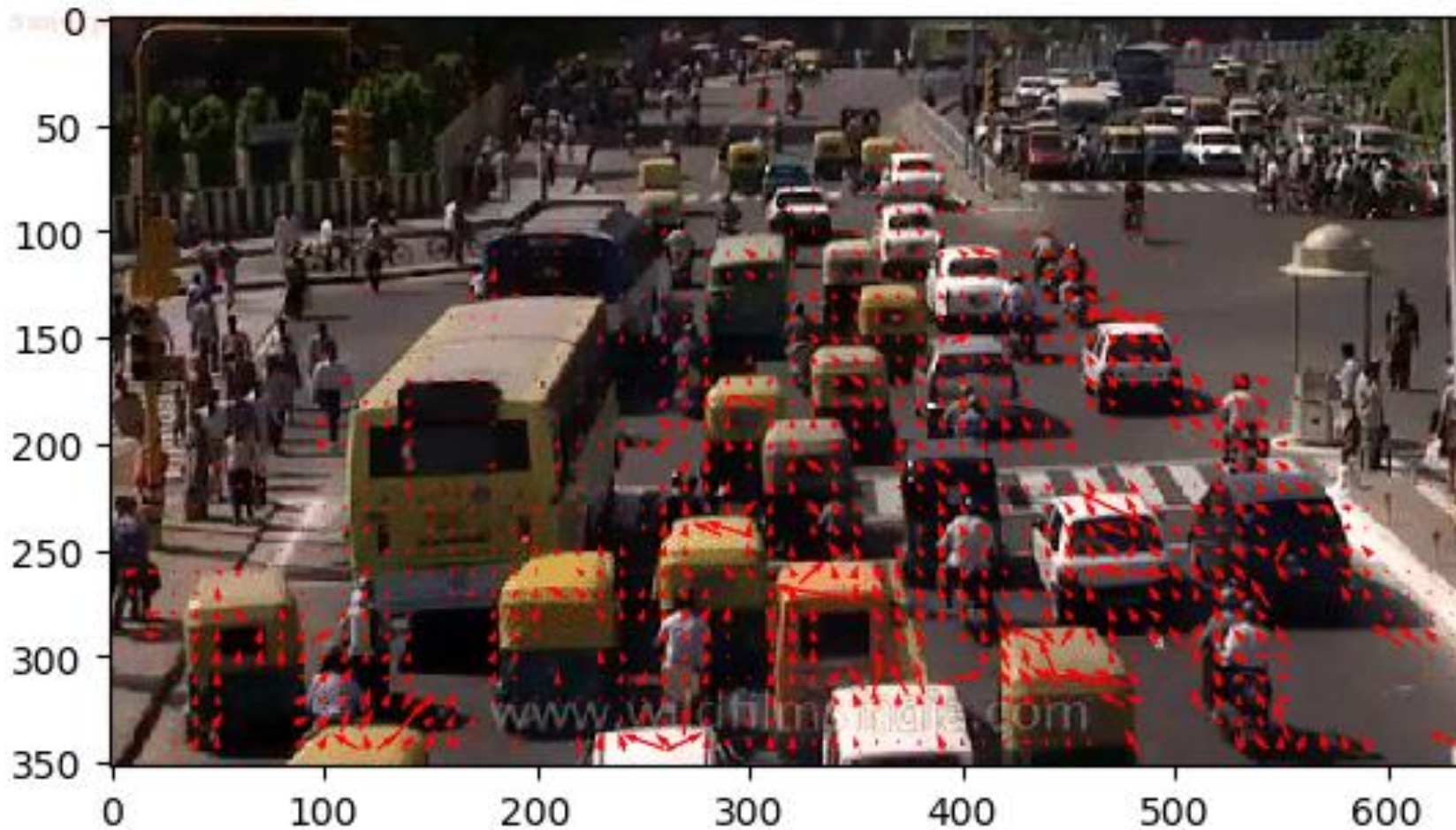
Optical Flow for Temporal Information



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

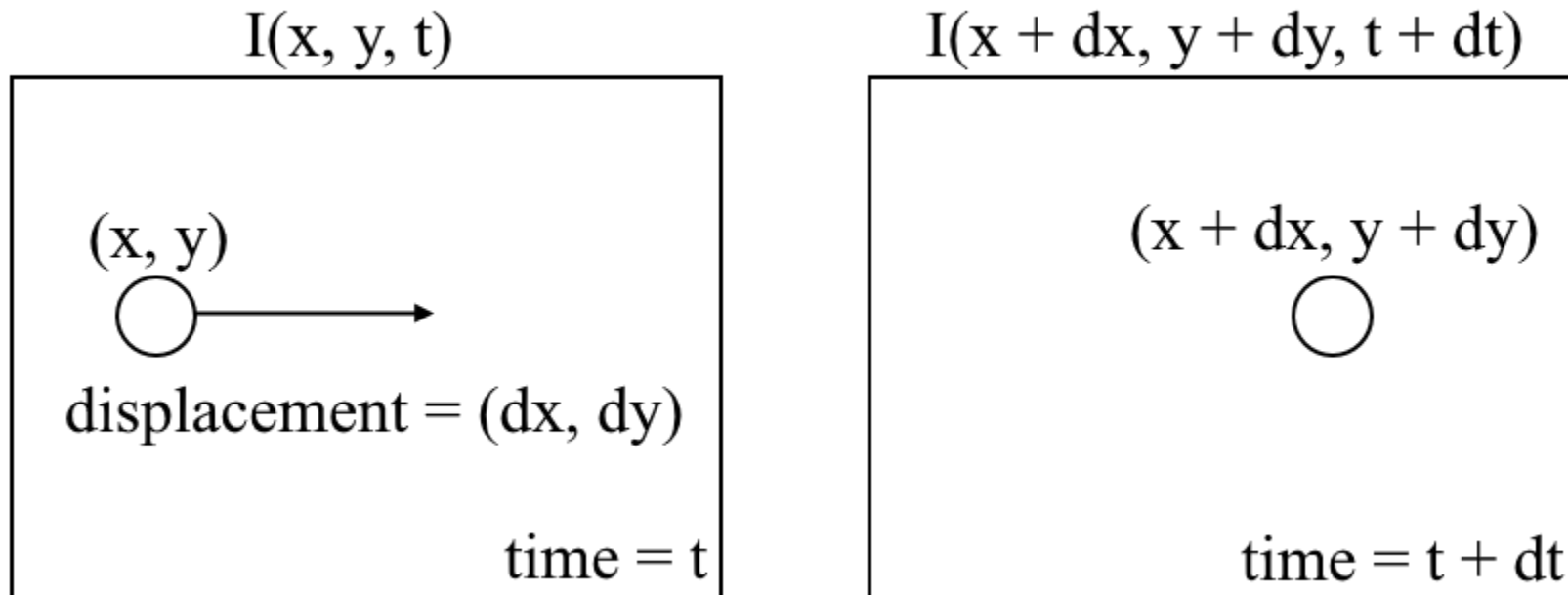
Estimate the Direction and Distance of Motion

- Optical flow can be used for motion estimation



What is Optical Flow?

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



Optical Flow

- Brightness constancy

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

- Taylor series

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

- Truncating higher order terms and dividing by Δt

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

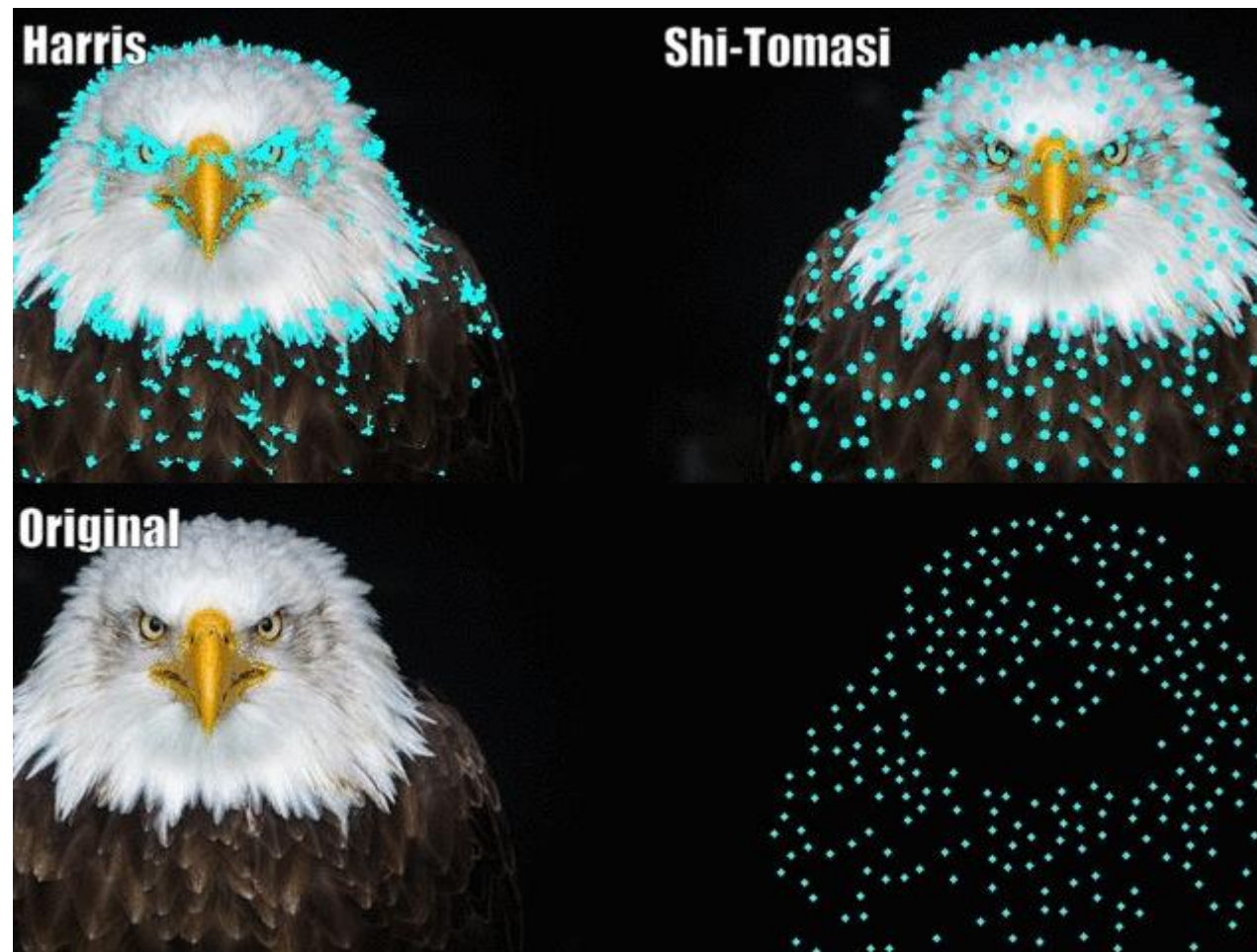
$$\Rightarrow I_x V_x + I_y V_y = -I_t \Rightarrow \nabla I \cdot \vec{V} = -I_t$$

Spare vs. Dense Optical Flow



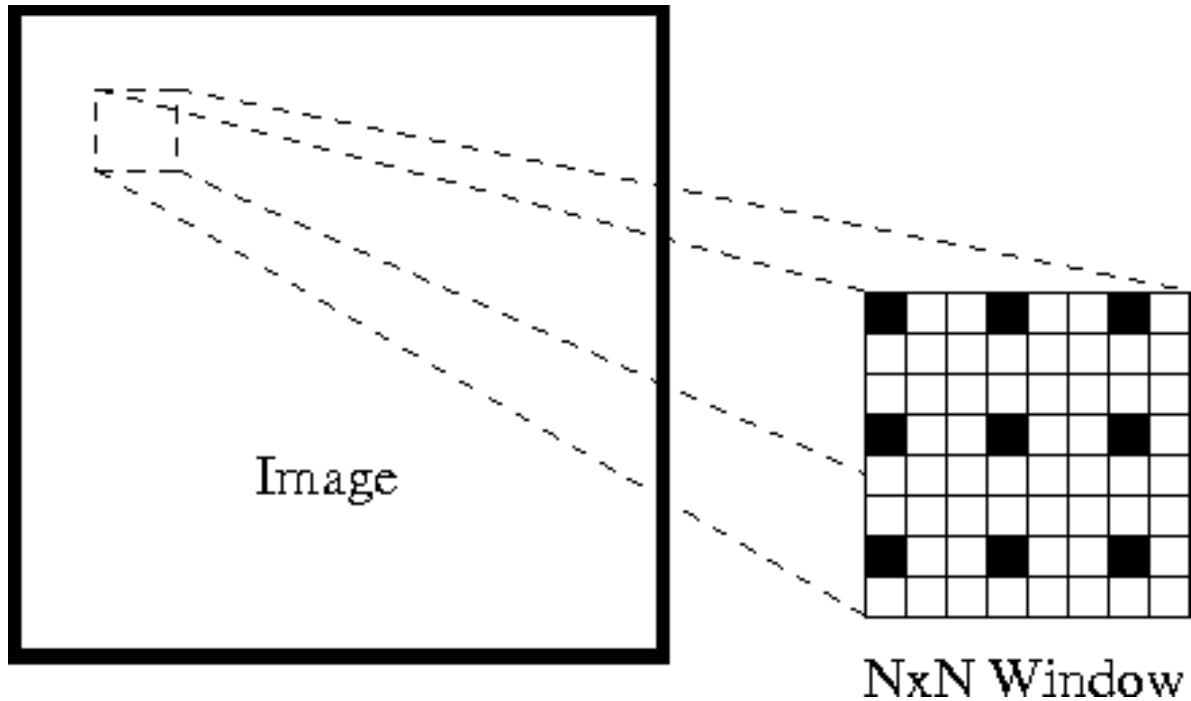
Selecting Feature Points

- 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

- $Av = b$

$$A = \begin{bmatrix} I_x(q_1) & I_y(q_1) \\ I_x(q_2) & I_y(q_2) \\ \vdots & \vdots \\ I_x(q_n) & I_y(q_n) \end{bmatrix}$$

$$v = \begin{bmatrix} V_x \\ V_y \end{bmatrix}$$

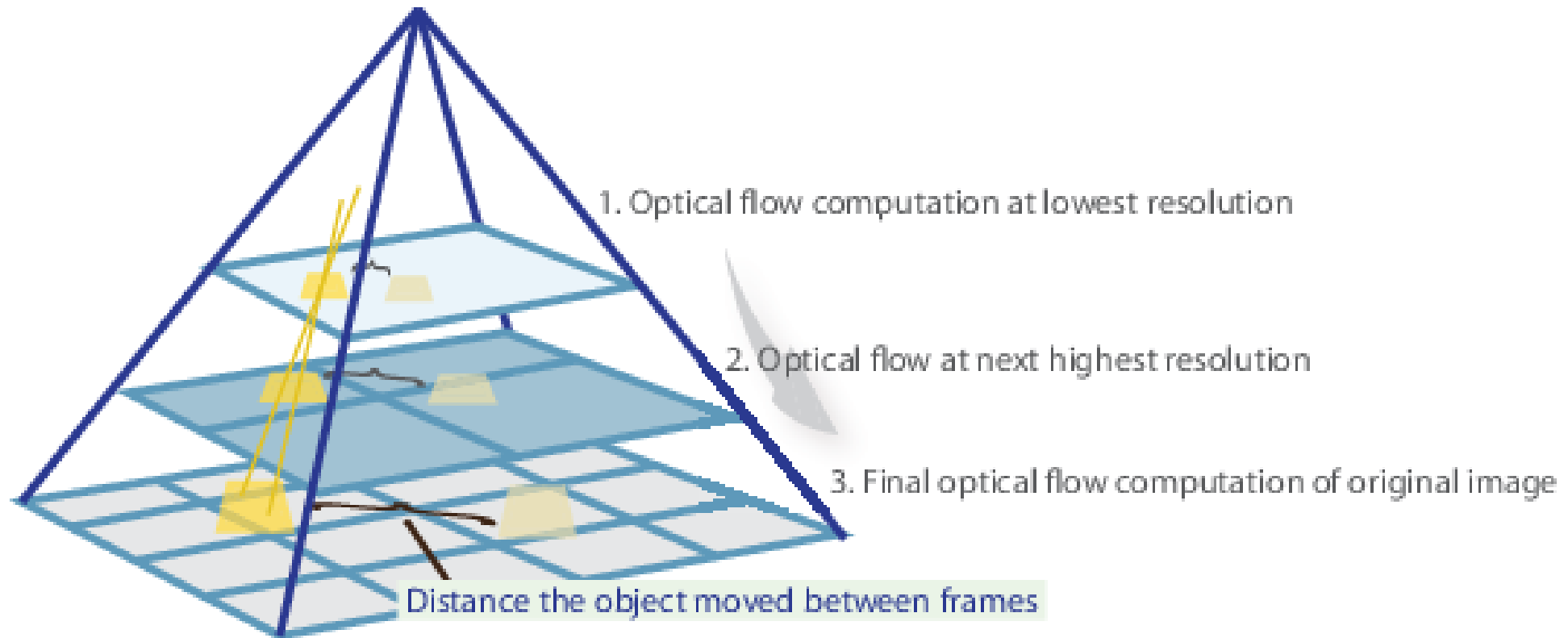
$$b = \begin{bmatrix} -I_t(q_1) \\ -I_t(q_2) \\ \vdots \\ -I_t(q_n) \end{bmatrix}$$

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)

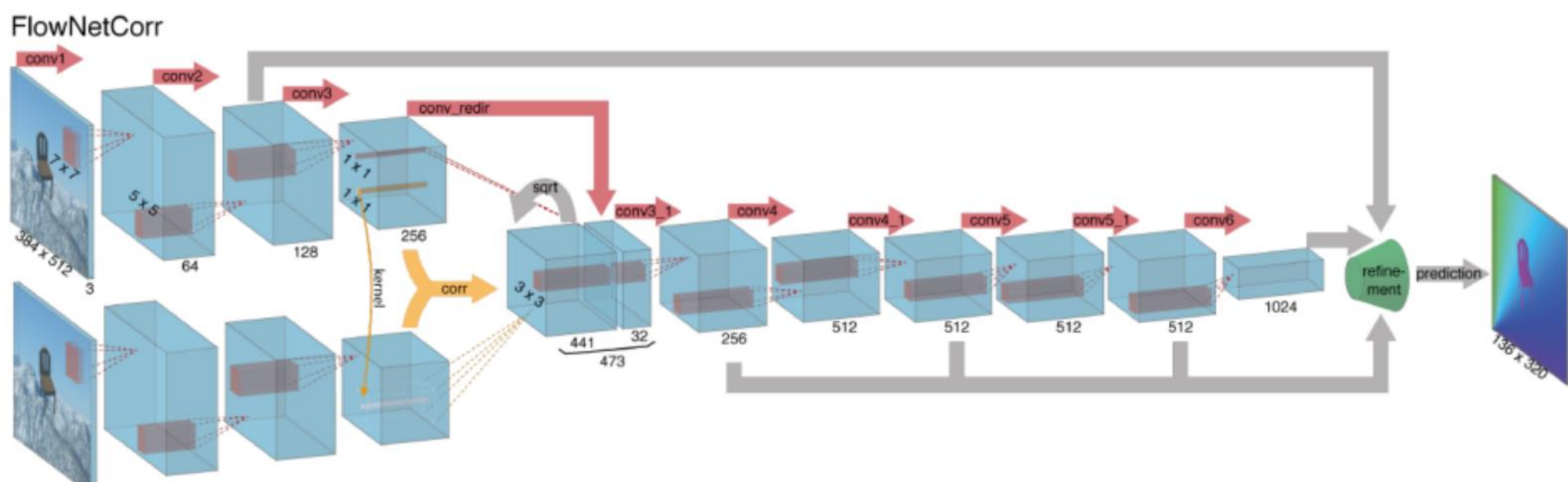
- Approximate each neighborhood of both frames by quadratic polynomials

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




FlowNet: Learning Optical Flow with Convolutional Networks

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

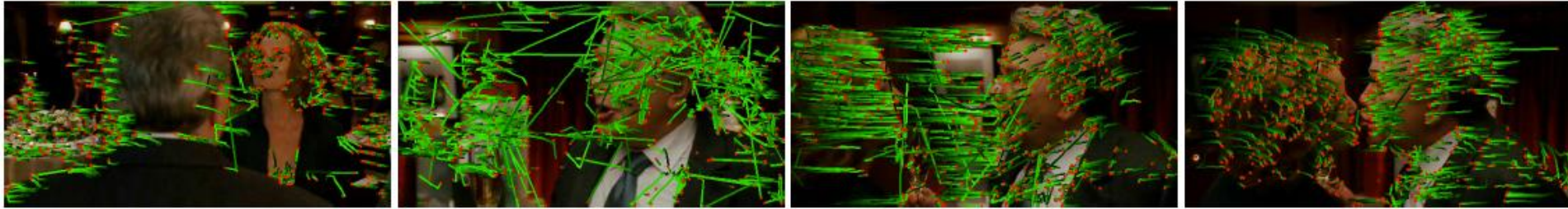


Synthetic Dataset

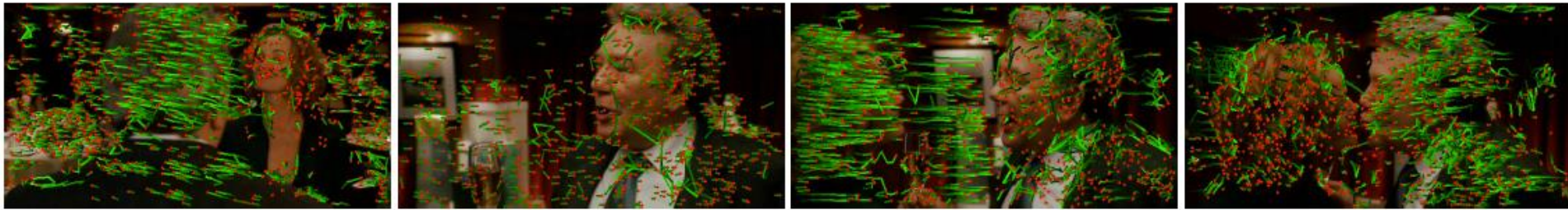
Images	Ground truth	EpicFlow	FlowNetS	FlowNetC
		 EPE: 0.27	 EPE: 1.06	 EPE: 0.91
		 EPE: 13.62	 EPE: 7.17	 EPE: 11.18
		 EPE: 32.56	 EPE: 20.82	 EPE: 26.63
		 EPE: 24.98	 EPE: 35.33	 EPE: 46.68
		 EPE: 0.33	 EPE: 0.89	 EPE: 0.71
		 EPE: 1.56	 EPE: 3.43	 EPE: 3.07
		 EPE: 5.45	 EPE: 8.11	 EPE: 7.12
		 EPE: 2.29	 EPE: 3.84	 EPE: 3.78

Dense Trajectories

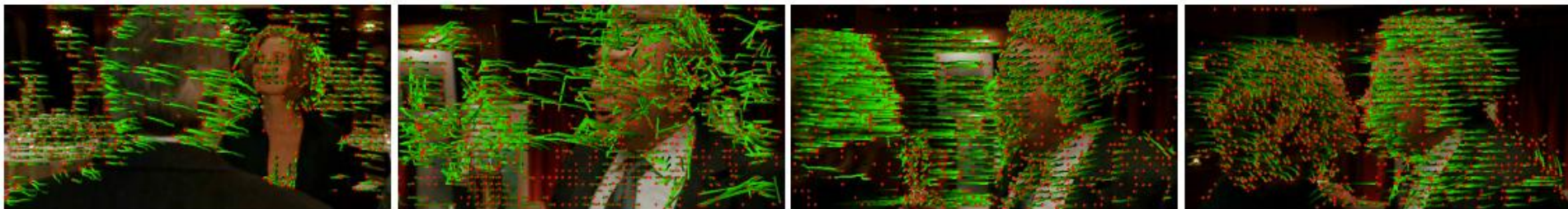
- H. Wang, et al. Dense trajectories and motion boundary descriptors for action recognition. IJCV, 2013
- Tracking dense pixels has shown better than KLT (corner) features and SIFT points



KLT trajectories



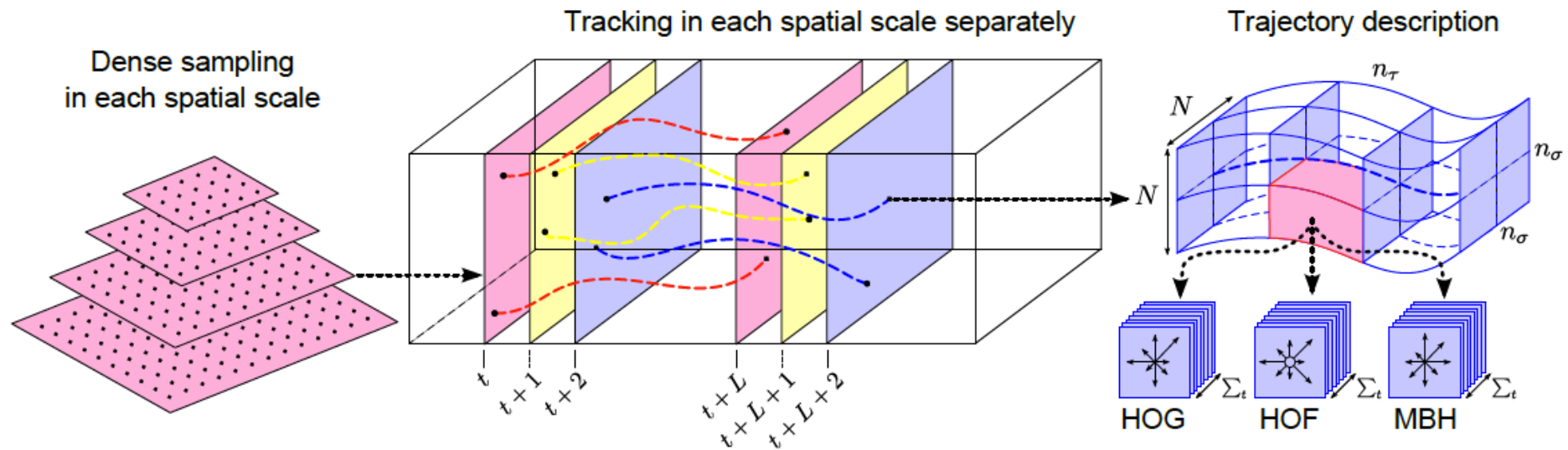
SIFT trajectories



Dense trajectories

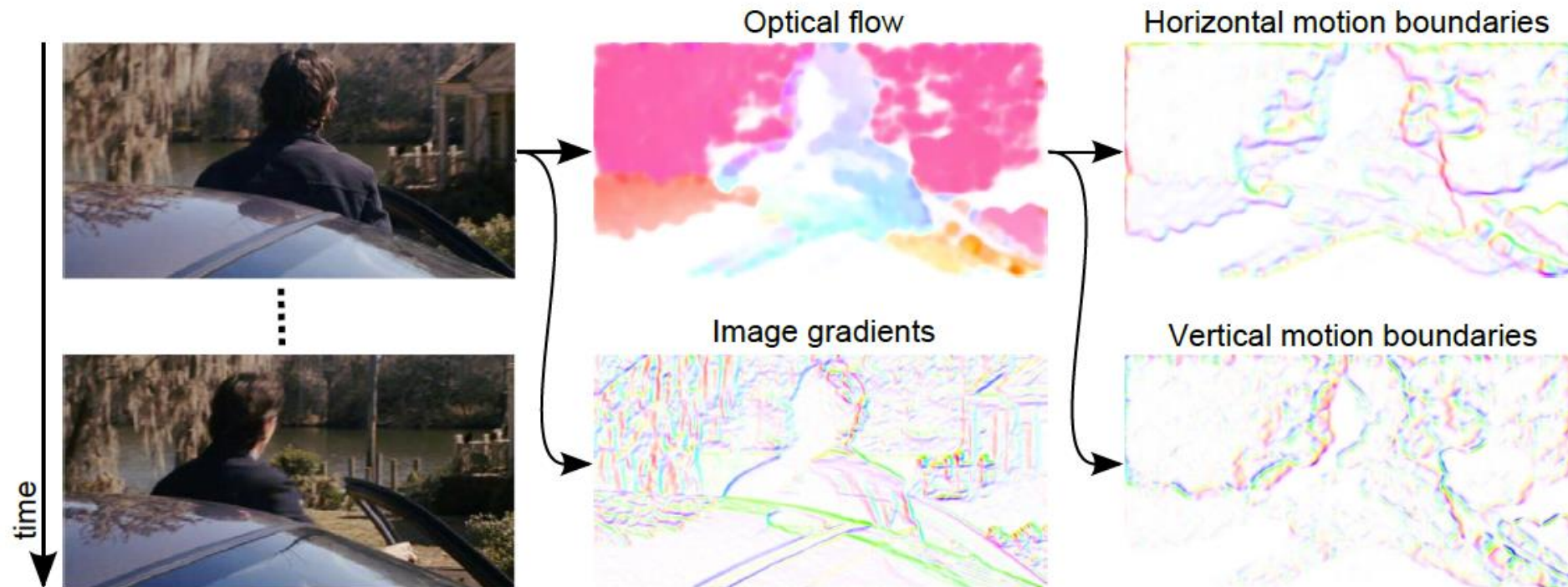
Extracting Dense Trajectories

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



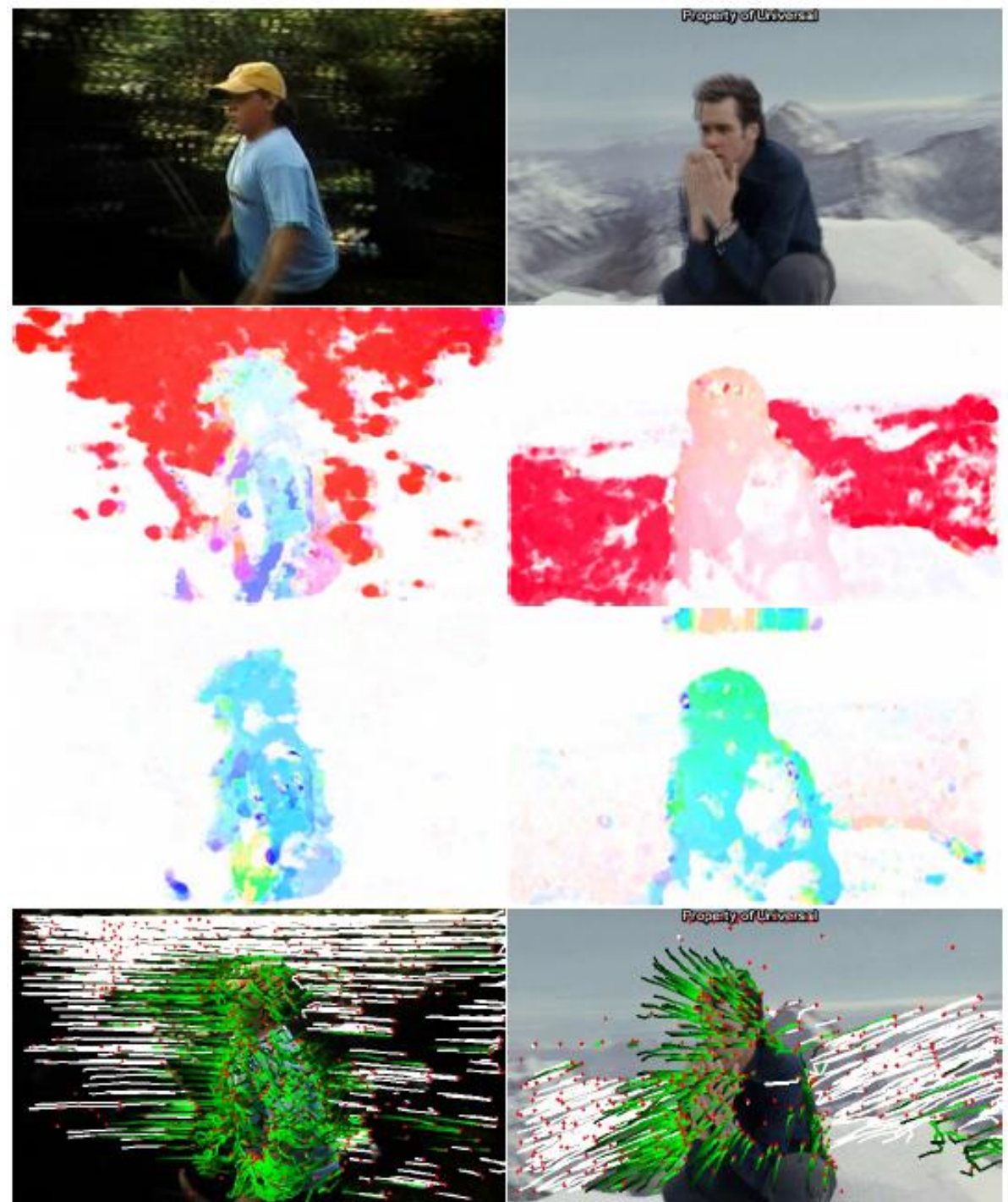
Motion Boundary Histogram (MBH)

- First proposed by Dalal & Triggs
 - “Human Detection Using Oriented Histograms 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



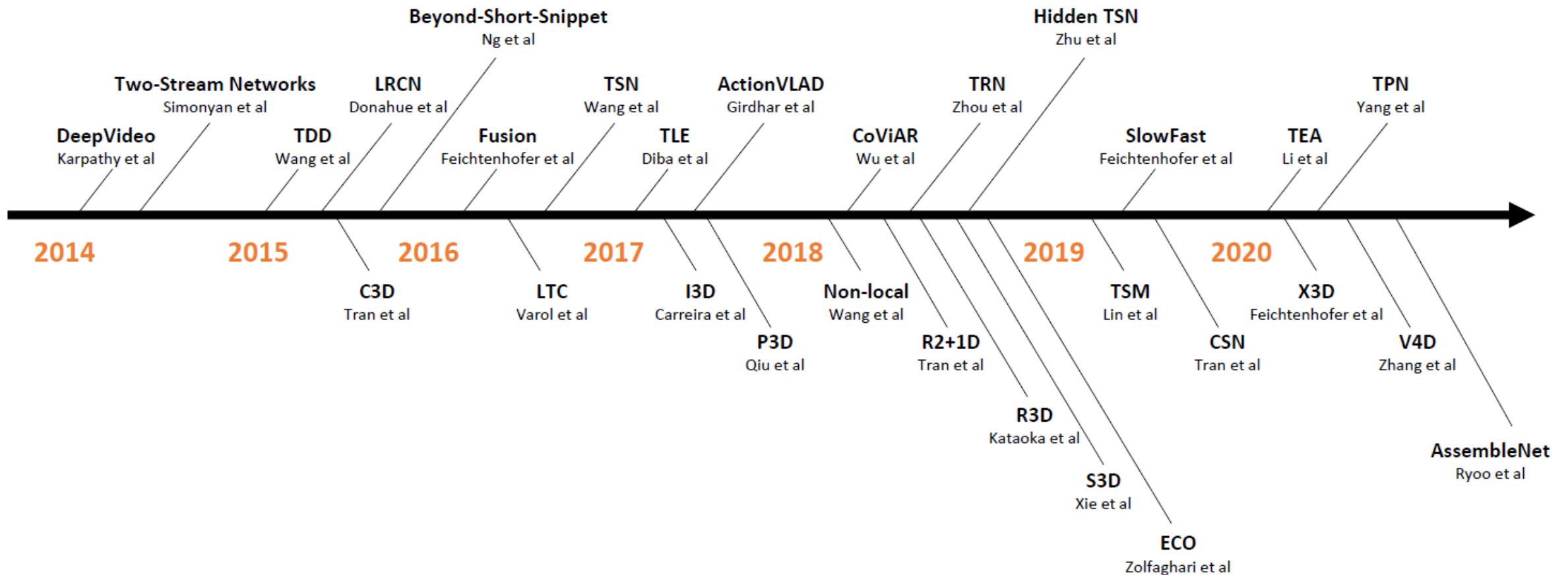
Improved Trajectories

- 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

UCF101



Cricket bowling



Skate boarding



Cutting in kitchen

Kinetics400



Riding a bike



Salsa dancing



braiding hair

Dropping something



Something-Something

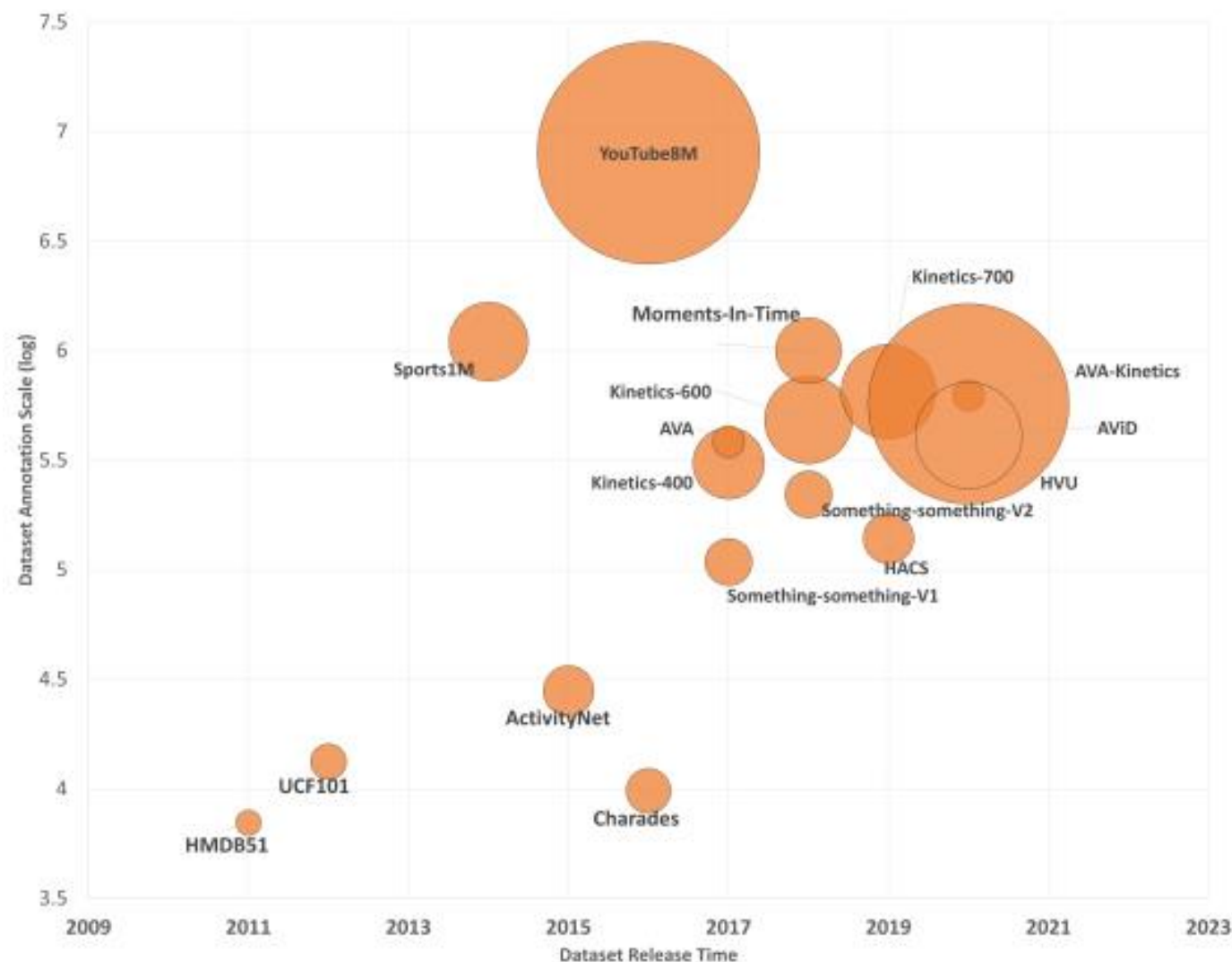


Picking something up

Moments in time



Climbing



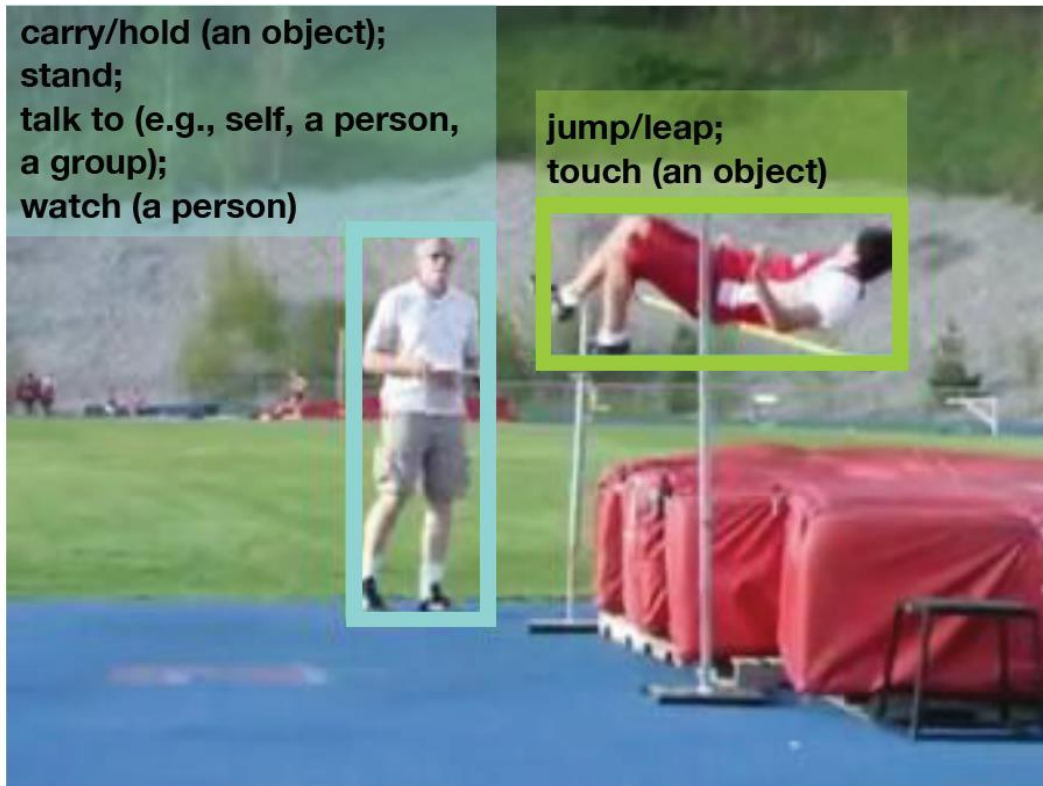
YouTube-8M

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



DeepMind Kinetics Dataset

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



DeepMind > Research > Kinetics

OPENSOURCE

22 MAY 2017

SHARE

VIEW SOURCE

VIEW PUBLICATION

FURTHER READING

Datasets

Open Source Software

Deep Learning

Vision

Kinetics

A large-scale, high-quality dataset of URL links to approximately 650,000 video clips that covers 700 human action classes, including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging. Each action class has at least 600 video clips. Each clip is human annotated with a single action class and lasts around 10s.

AVA Kinetics

[View paper](#) • [Download dataset](#)

Kinetics 700

[View paper](#) • [Download dataset](#)

Kinetics 600

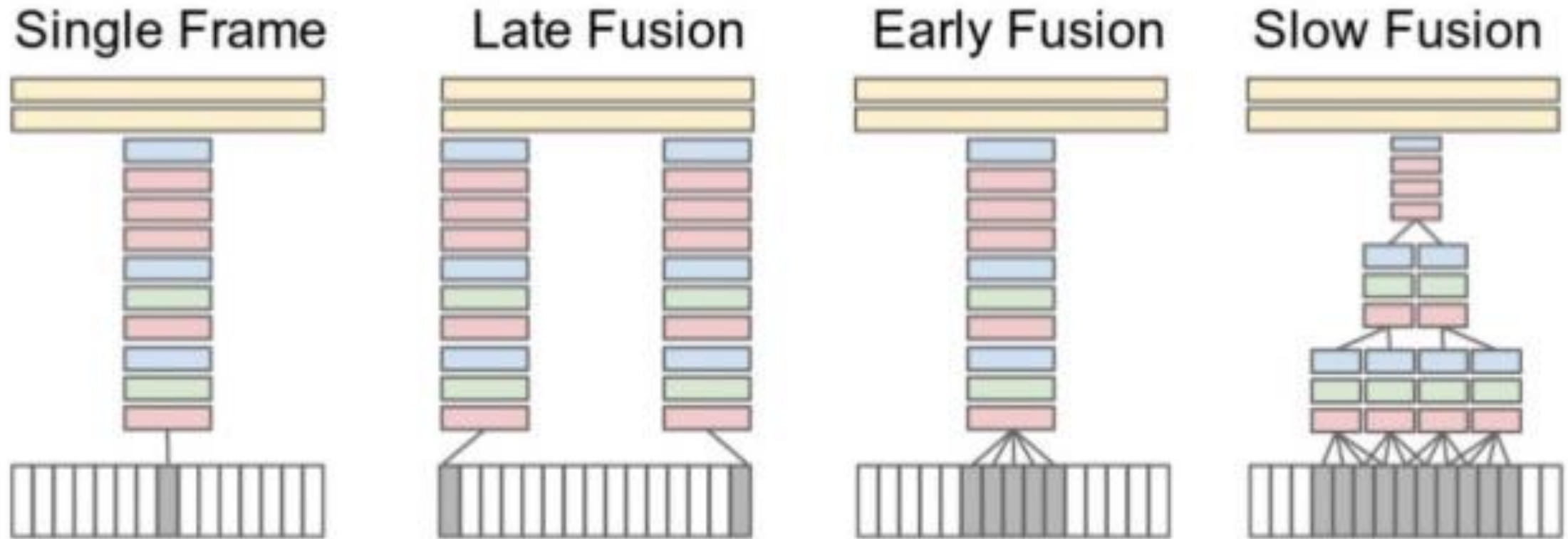
[View paper](#) • [Download dataset](#)

Kinetics 400

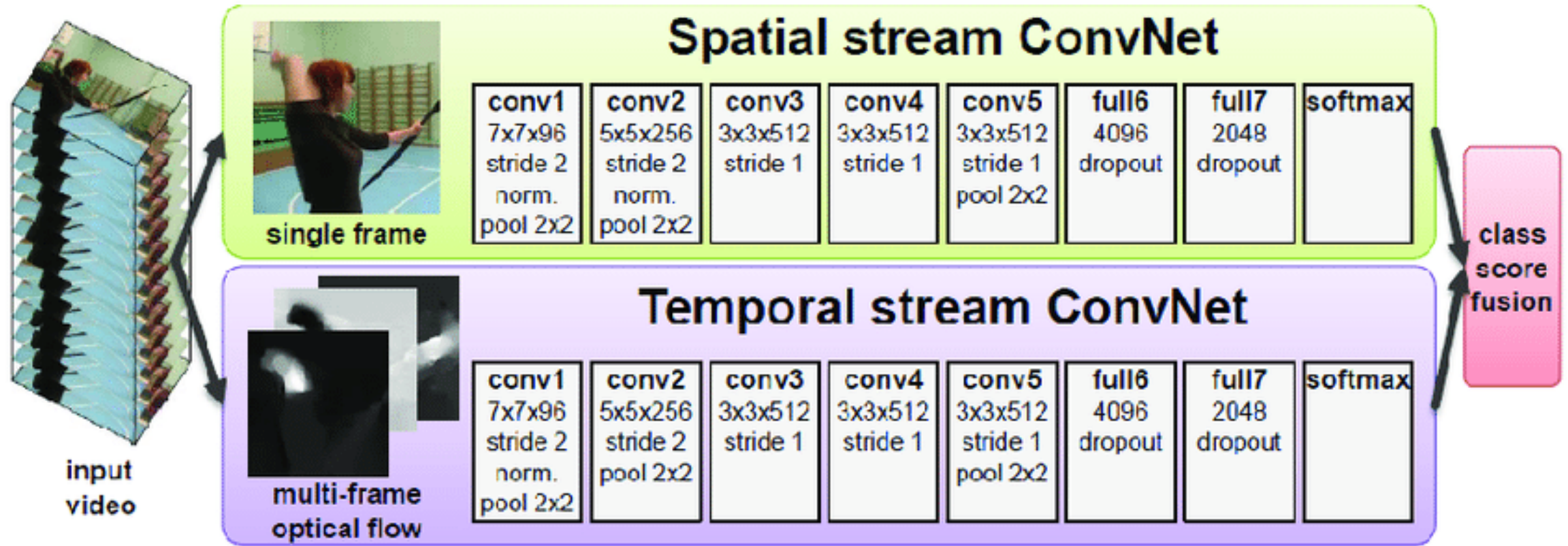
[View paper](#) • [Download dataset](#)

Learning Optical Flow with Deep Learning?

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



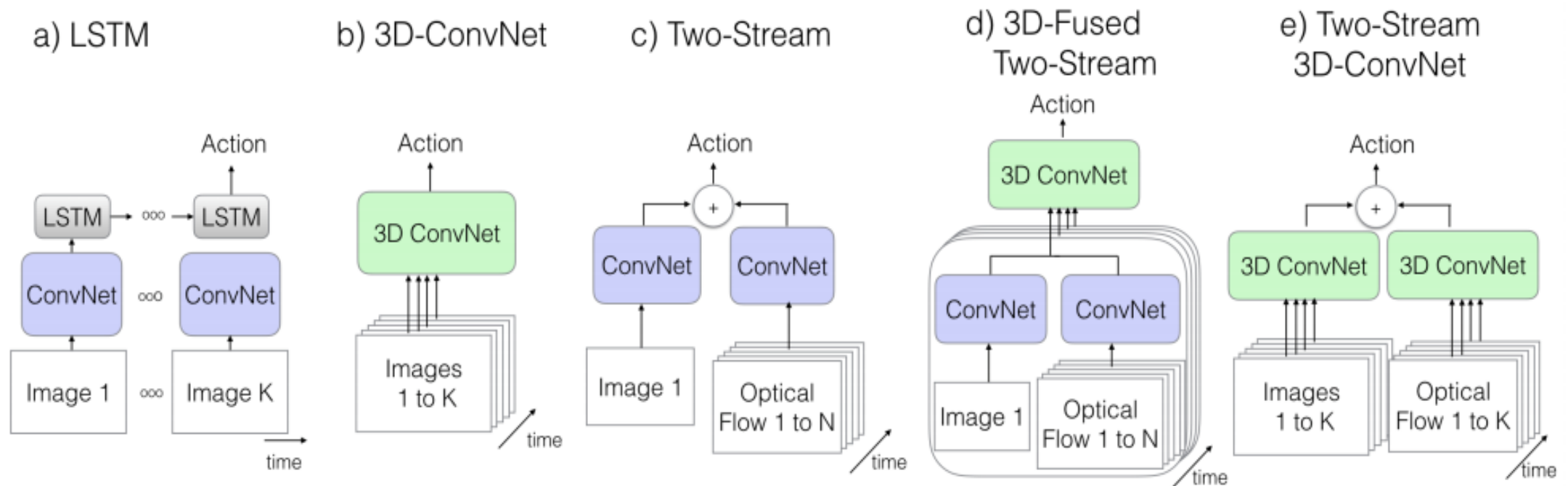
Two-stream Convolutional Neural Networks



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]

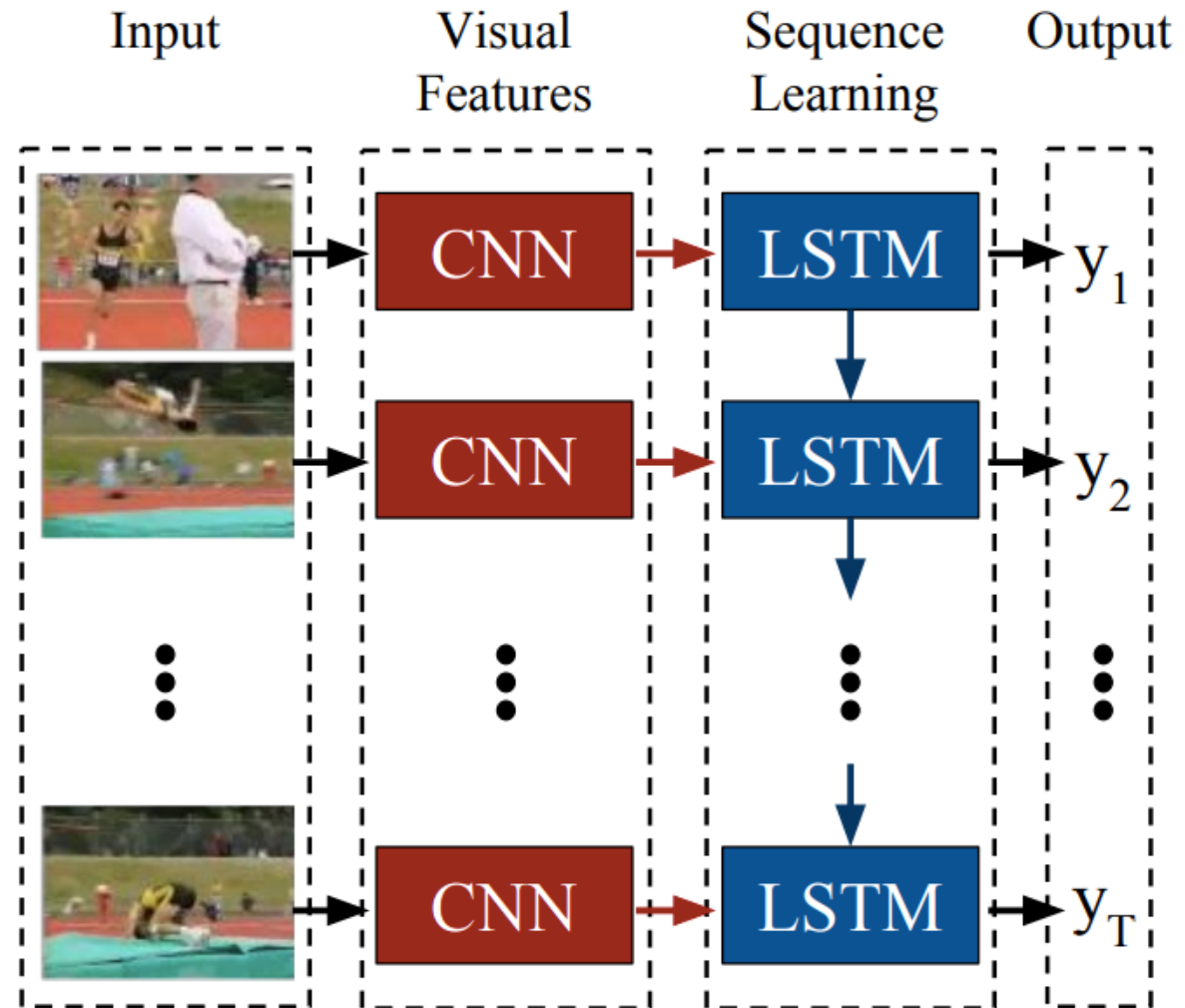
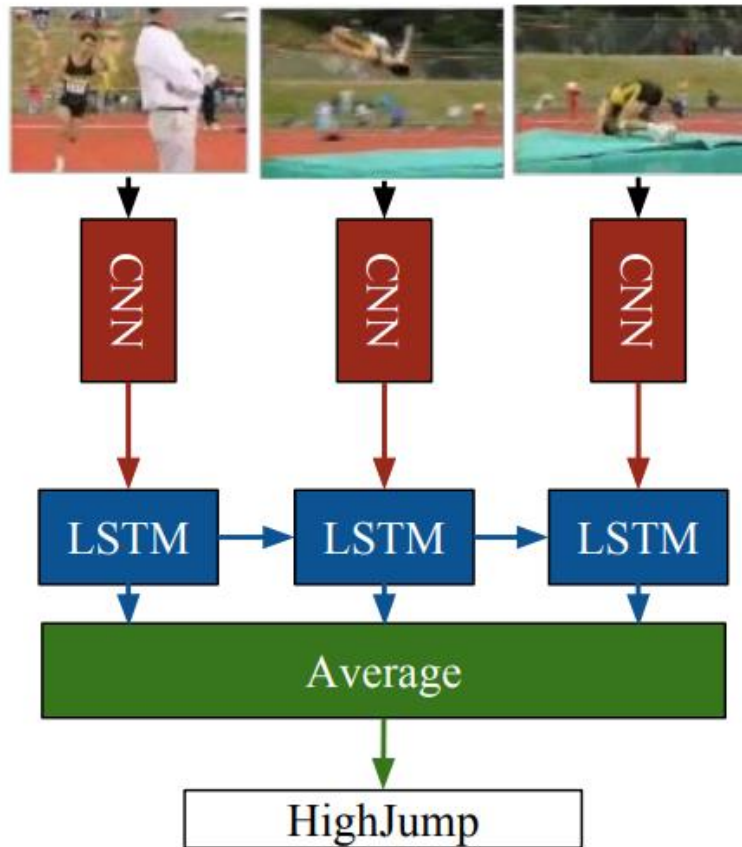


Long-term Recurrent Convolutional Networks

- 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

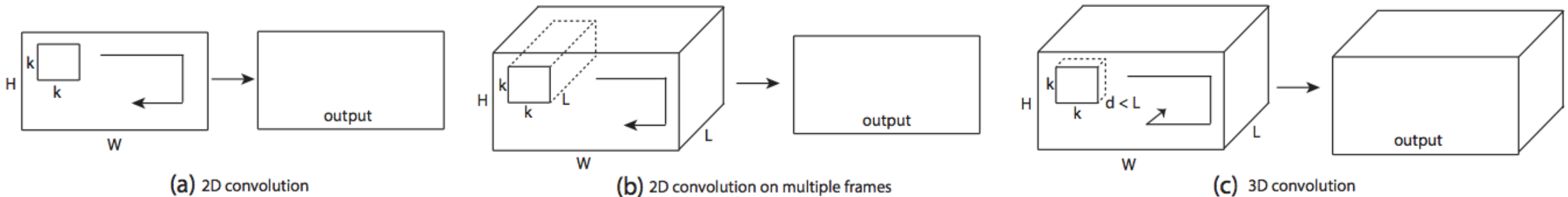
Long-term Recurrent Convolutional Networks

Activity Recognition Sequences in the Input



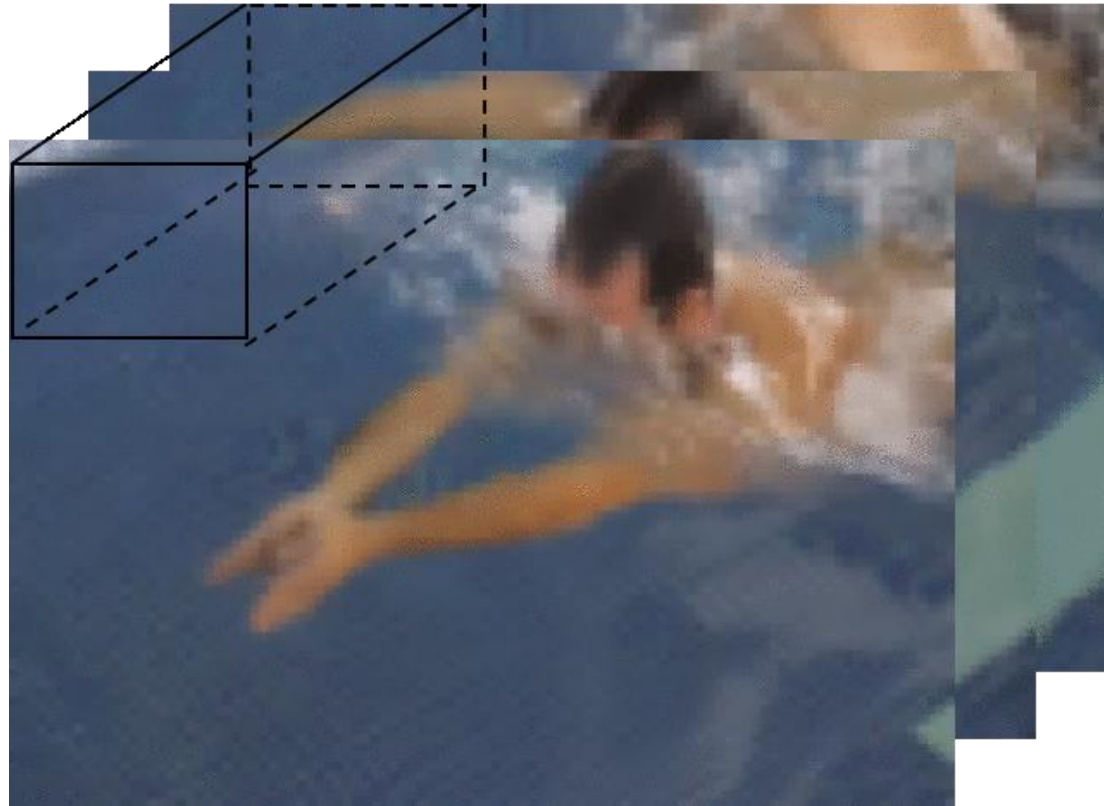
3D Convolutional Networks (C3D)

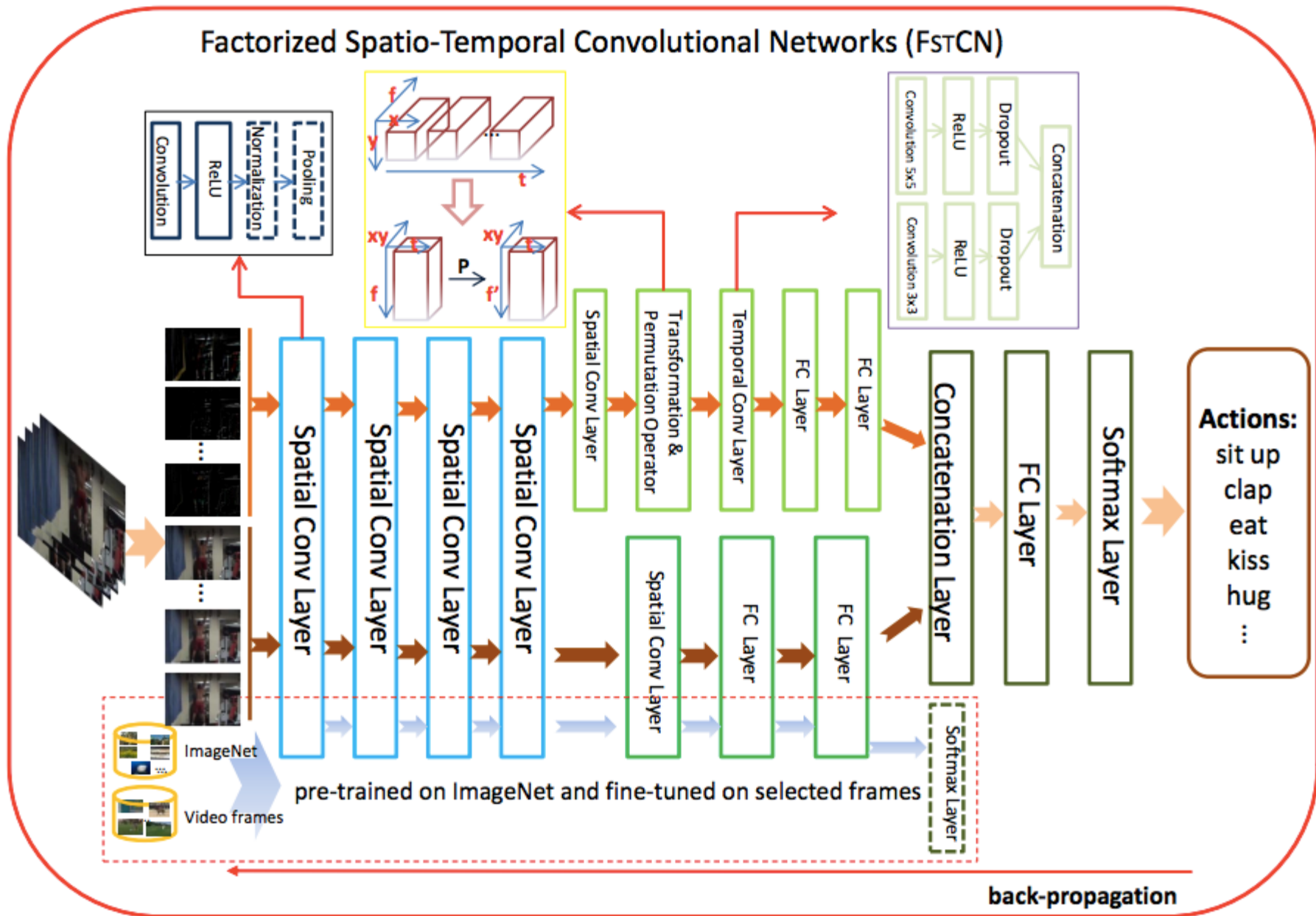
- Du Tran et al., “Learning Spatiotemporal Features with 3D Convolutional Networks,” 2014 ([Arxiv Link](#))
- Key Contributions
 - Repurposing 3D convolutional networks as feature extractors
 - Extensive search for best 3D convolutional kernel and architecture
 - Using deconvolutional layers to interpret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



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

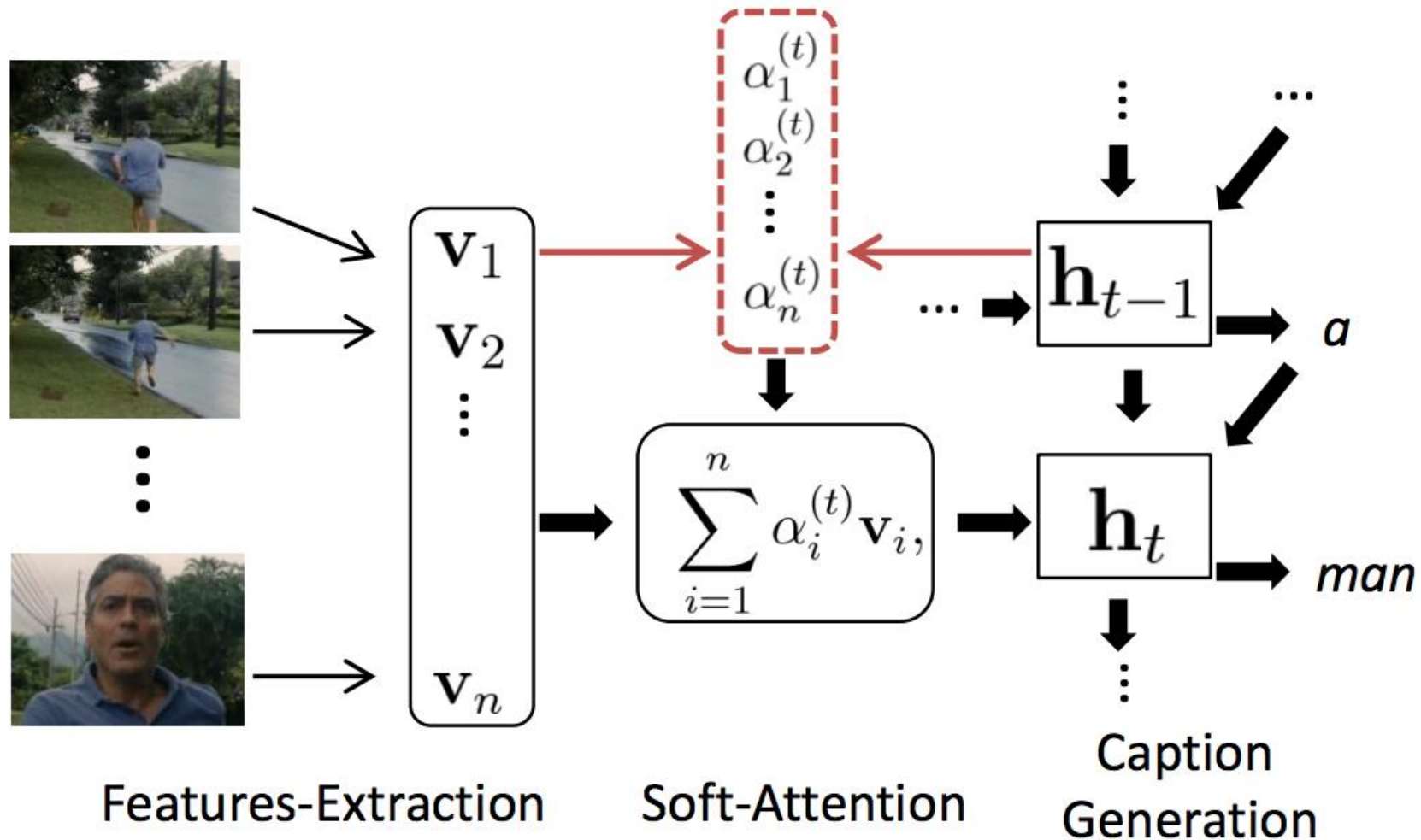




Human Action Recognition using Factorized Spatio-Temporal Convolutional Networks

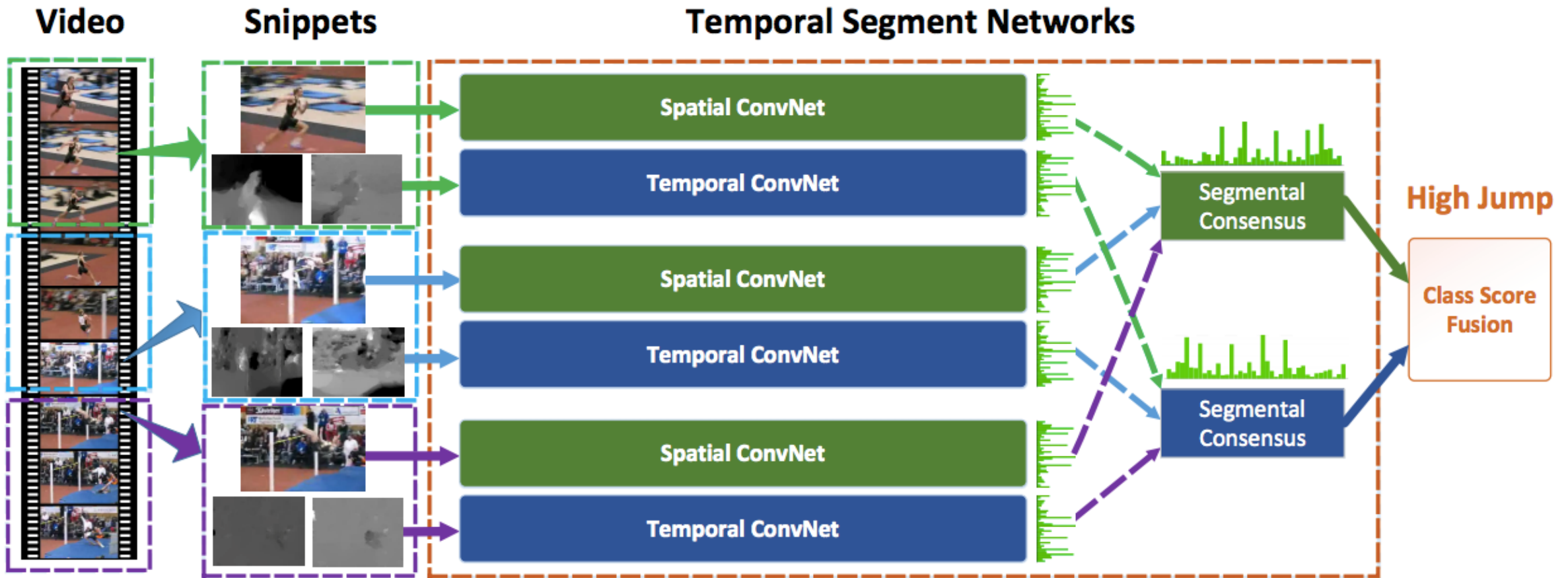
Conv3D + Attention

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



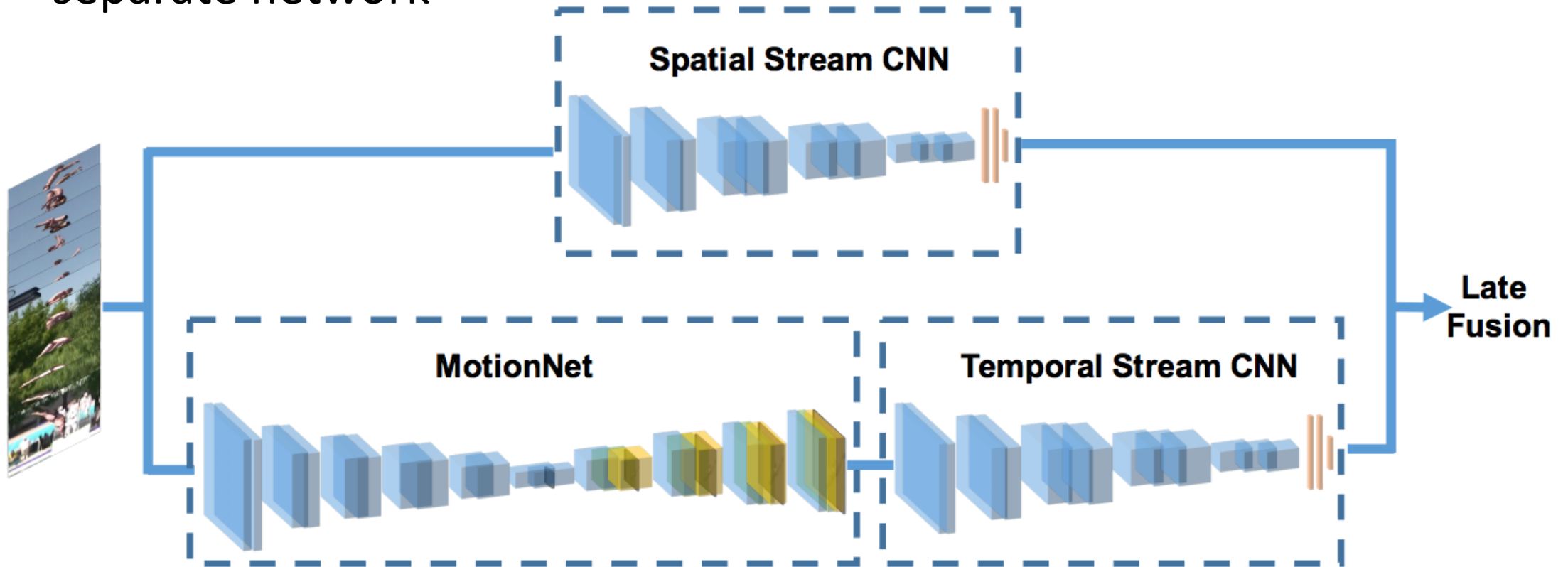
Temporal Segment Networks (TSN)

- 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

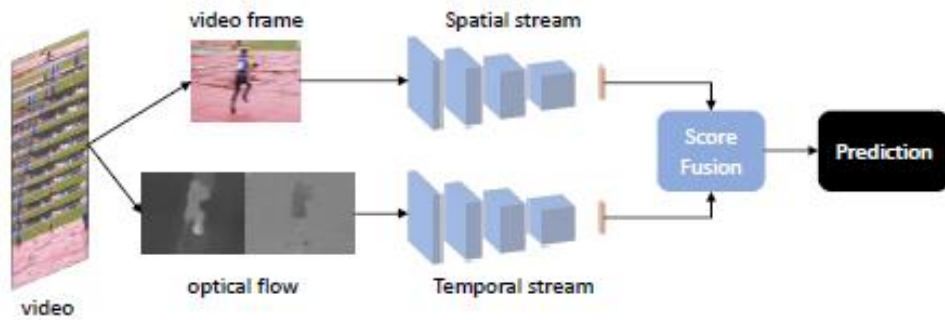


Hidden Two Stream

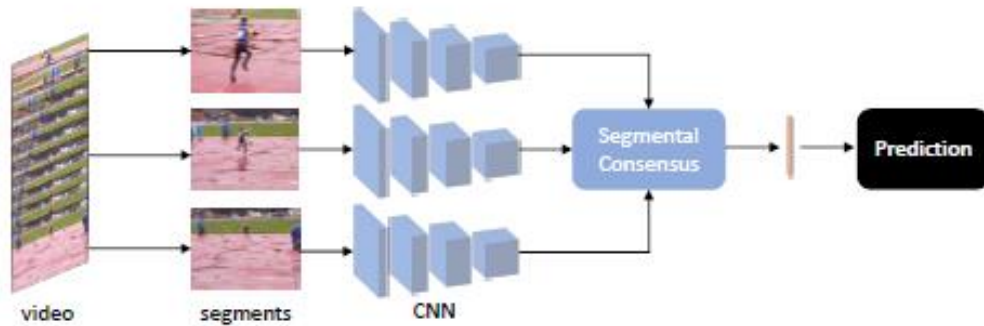
- Zhu et al., “Hidden Two-Stream Convolutional Networks for Action Recognition,” 2017
- Novel architecture for generating optical flow input on-the-fly using a separate network



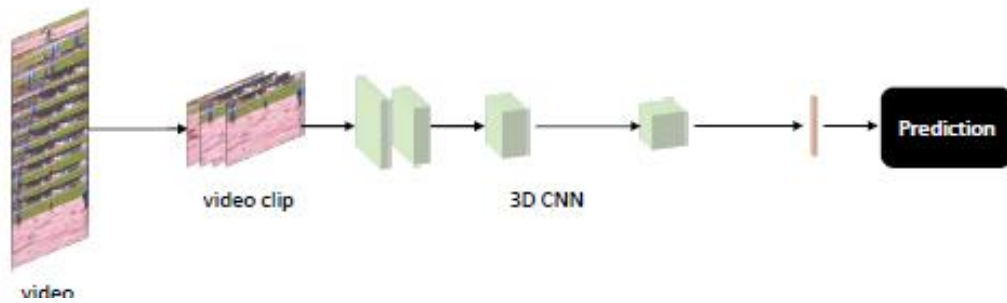
5 Important Action CNN Models



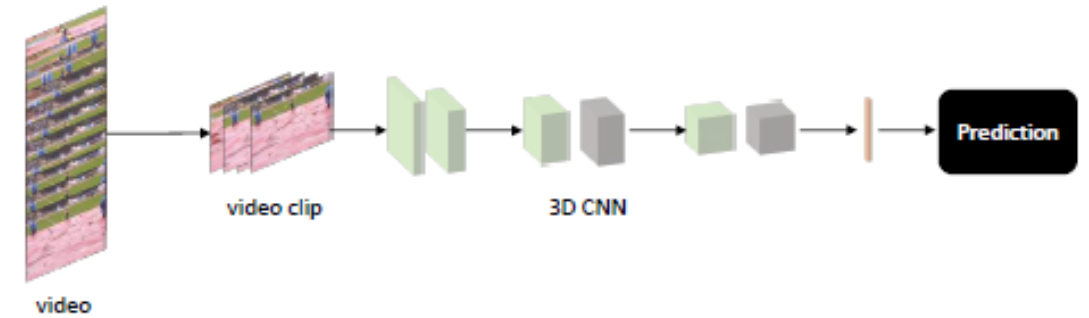
Two-stream Networks



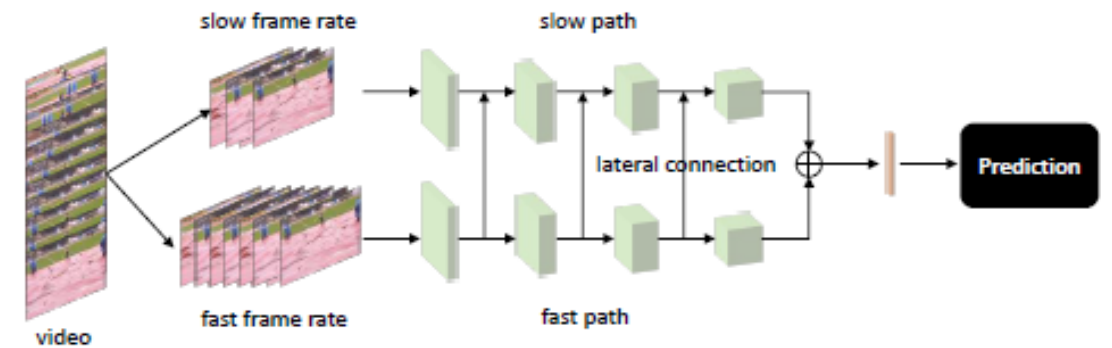
Temporal Segment Networks (TSN)



I3D



Non-local



SlowFast



Efficient Two-stream Action Recognition on FPGA

- CVPR 2021 ECV Workshop

- Contributions:

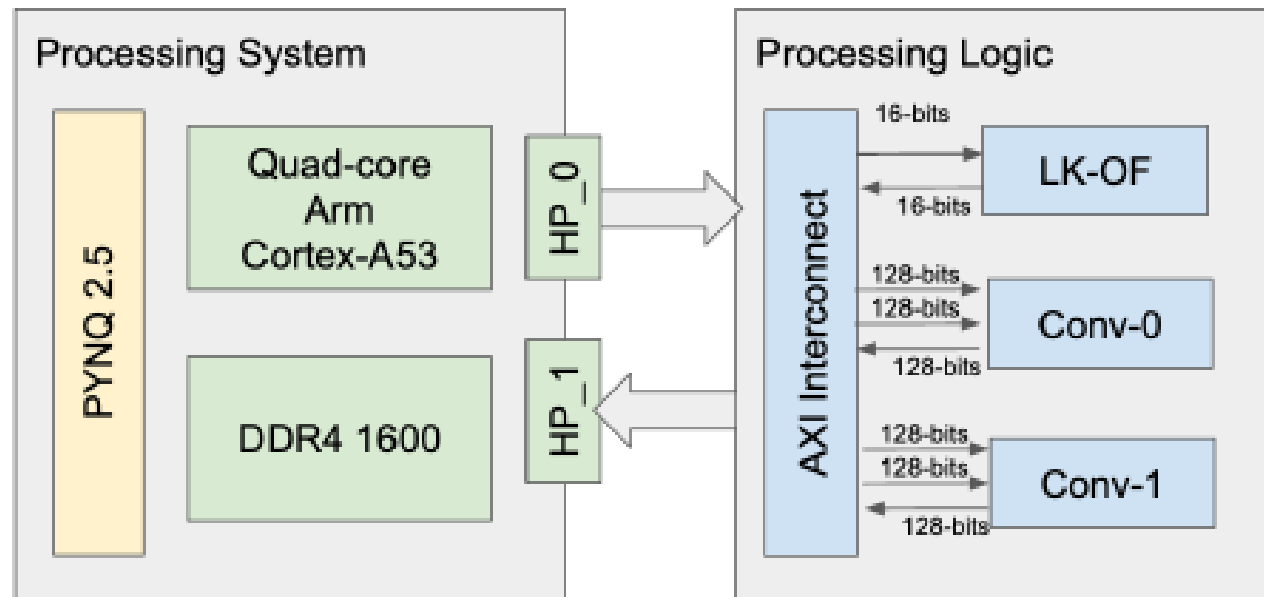
- Has 10x less operations than other models with little accuracy drop (<2%)
- Use only 2D CNN operations
- Processing both spatial and temporal streams parallelly.

Jia-Ming Lin¹, Kuan-Ting Lai², Bin-Ray Wu¹ and Ming-Syan Chen¹

¹National Taiwan University


²National Taipei University of Technology

ktlai@ntut.edu.tw, {jmlin, brwu, mschen}@arbor.ee.ntu.edu.tw



TF-Hub Action Recognition Model

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

 TensorFlow

Install

Learn ▾

API ▾

Resources ▾

Community

Why TensorFlow ▾

🔍 Search

English ▾

GitHub

Sign in

Text Tutorials

Text classification

Text cookbook

Nearest neighbor index for real-time semantic search

Semantic similarity

Semantic similarity lite

Text classification on Kaggle

Bangla article classifier

Explore text embeddings

Retrieval based question answering

Multilingual universal sentence encoder

Image Tutorials

Retraining an image classifier

Object detection

Action recognition

BigGAN image generation

BigBiGAN image generation

S3 GAN image generation


Arbitrary image stylization

image feature matching

Face generation and embedding

Video interpolation

to_gif(sample_video)



predict(sample_video)

Top 5 actions:

roller skating	: 96.85%
playing volleyball	: 1.63%
skateboarding	: 0.21%
playing ice hockey	: 0.20%
playing basketball	: 0.16%

Tutorials (TF1) ▾

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