Introduction to Human Activity Understanding in Videos

CVPR 2020 – Tutorial on Video Modeling

Yuanjun Xiong

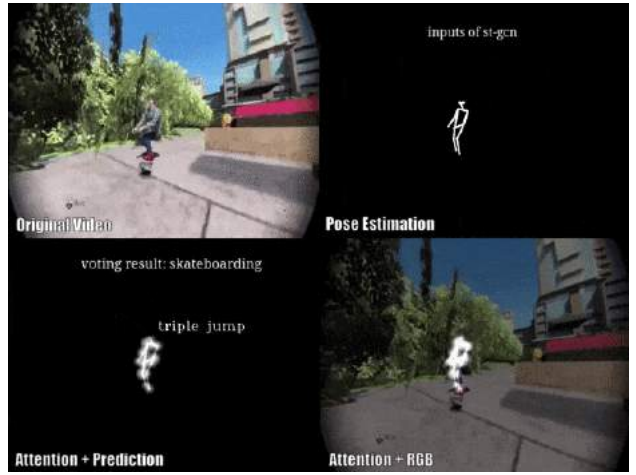
AWS AI

Outline

- **Data**
- Tasks
- Models
- Challenges & Problems

Data

- Inputs
  - **Video**
  - Images
  - Others
    - **Skeleton**
    - Sensor Data
    - …

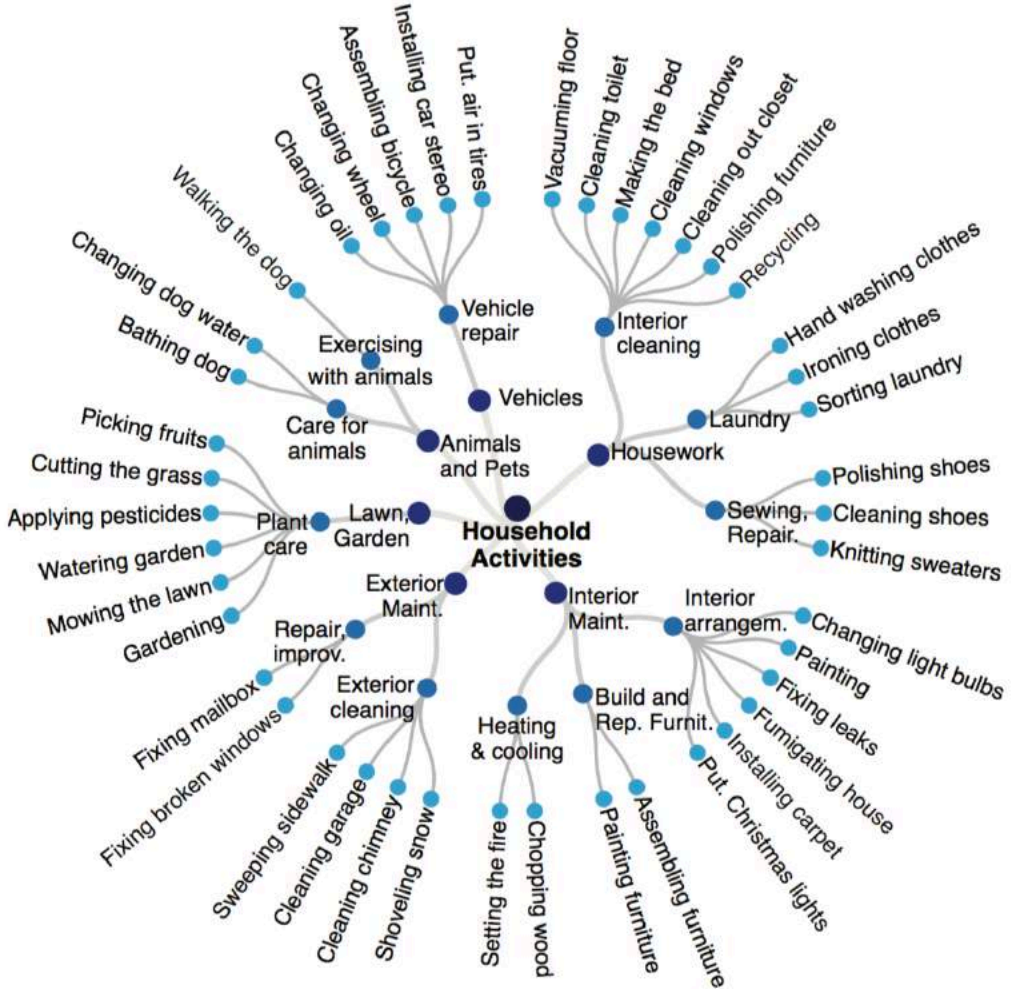


Data

- Sources of data
  - Controlled Collection
    - Weizmann Dataset[1]
    - KTH Dataset[2]
  - In the wild
    - Movies
      - HMDB51
    - Youtube videos
      - UCF101, ActivityNet, Kinetics, …
    - Surveillance footages
      - HiEve[3]

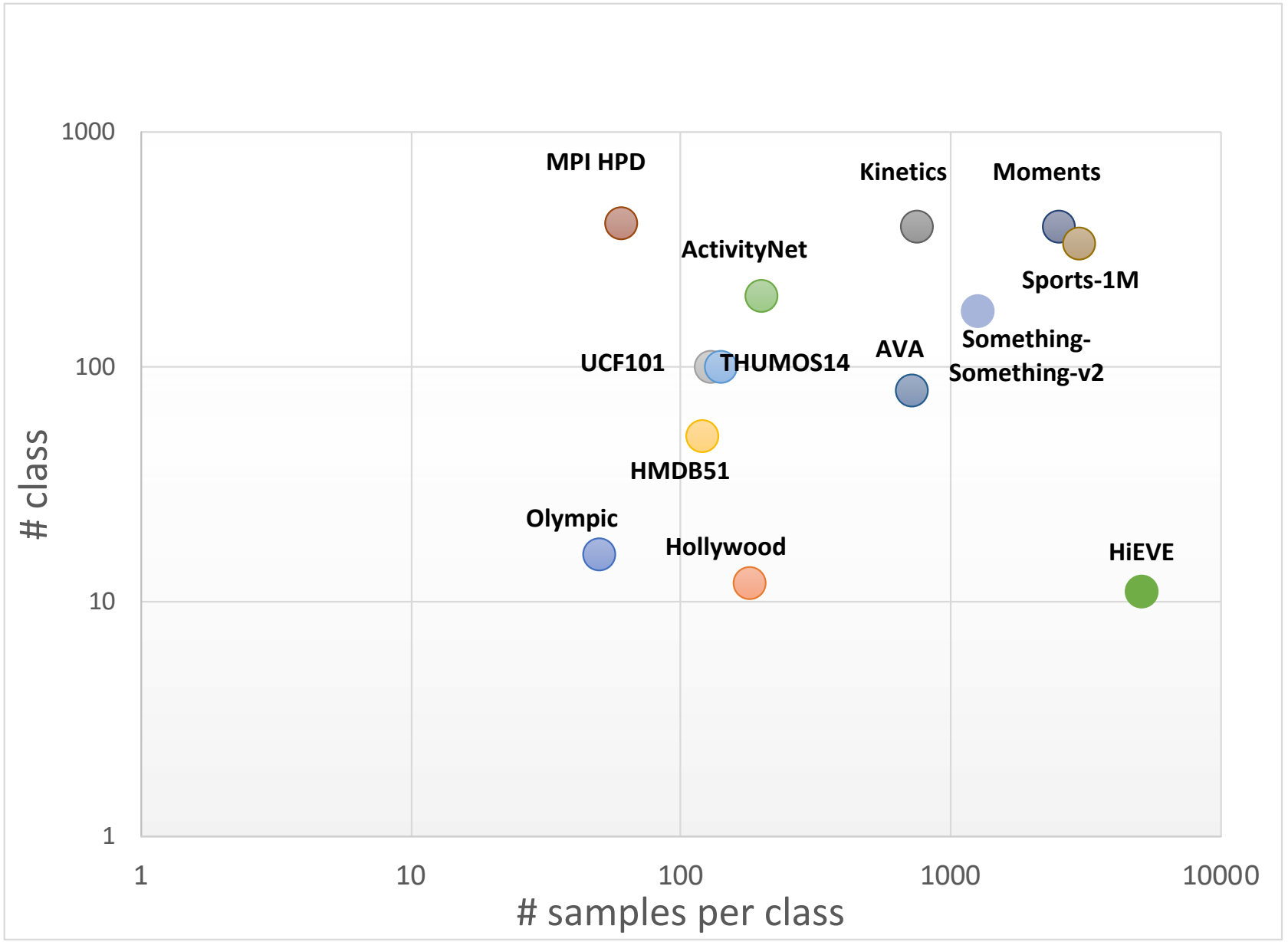
[1] “Actions as Space-Time Shapes”, Gorelick, et. Al, TPAMI
[2] “Recognizing Human Actions: A Local SVM Approach”, Schuldt, et. Al, ICPR’04
[3] “Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events”, Lin, et. Al, Arxiv 2020.

Data


Household Activities
Vehicles
Vehicle repair
Put. air in tires
Installing car stereo
Assembling bicycle
Changing wheel
Changing oil
Animals and Pets
Exercising with animals
Walking the dog
Care for animals
Changing dog water
Bathing dog
Housework
Interior cleaning
Vacuuming floor
Cleaning toilet
Making the bed
Cleaning windows
Cleaning out closet
Polishing furniture
Recycling
Laundry
Hand washing clothes
Ironing clothes
Sorting laundry
Sewing, Repair.
Polishing shoes
Cleaning shoes
Knitting sweaters
Lawn, Garden
Plant care
Picking fruits
Cutting the grass
Applying pesticides
Watering garden
Mowing the lawn
Gardening
Exterior Maint.
Repair, improv.
Fixing mailbox
Fixing broken windows
Exterior cleaning
Sweeping sidewalk
Cleaning garage
Cleaning chimney
Shoveling snow
Interior Maint.
Heating & cooling
Setting the fire
Chopping wood
Build and Rep. Furnit.
Painting furniture
Assembling furniture
Interior arrangem.
Changing light bulbs
Painting
Fixing leaks
Fumigating house
Installing carpet
Put. Christmas lights

Data

- Datasets



Outline

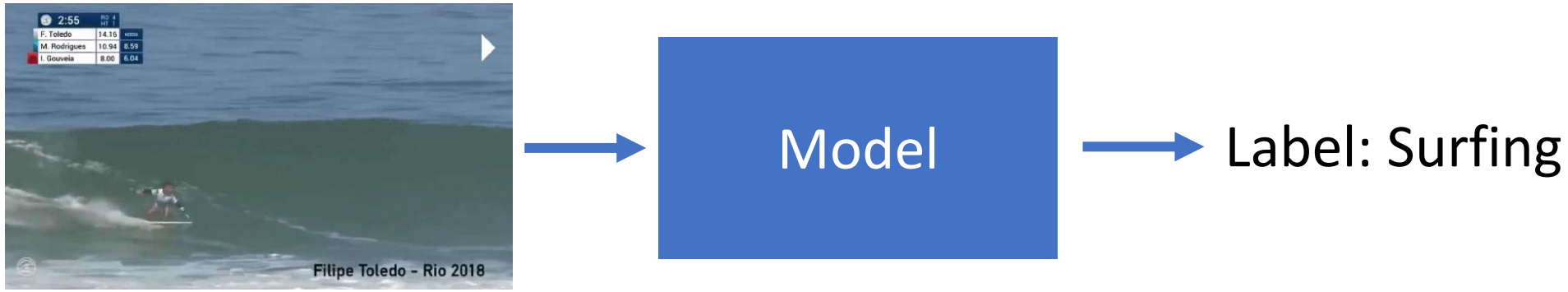
- Data
- **Tasks**
- Models
- Challenges & Problems

Tasks

- Basics
  - Action Classification: what
  - Temporal Action Detection/Localization: when
  - Spatial Temporal Action Detection: where
- Extensions
  - Online vs. Offline
  - Supervised vs. Unsupervised
  - Single Person vs. Multi-Person
  - Labels vs. natural language

Video Action Classification

- One video to one label



Trimmed vs. Untrimmed

Untrimmed Video Classification

Trimmed Video Classification

- Trimmed
  - Datasets: UCF101, HMDB51, Kinetics, Something-Something, …
  - Methods: two-stream CNN, 3D CNN, …
- Untrimmed
  - Datasets: ActivityNet, THUMOS15, …
  - Methods: UntrimmedNet, …

Single Person vs. Group

Single Person Action Classification

No inter-person interaction

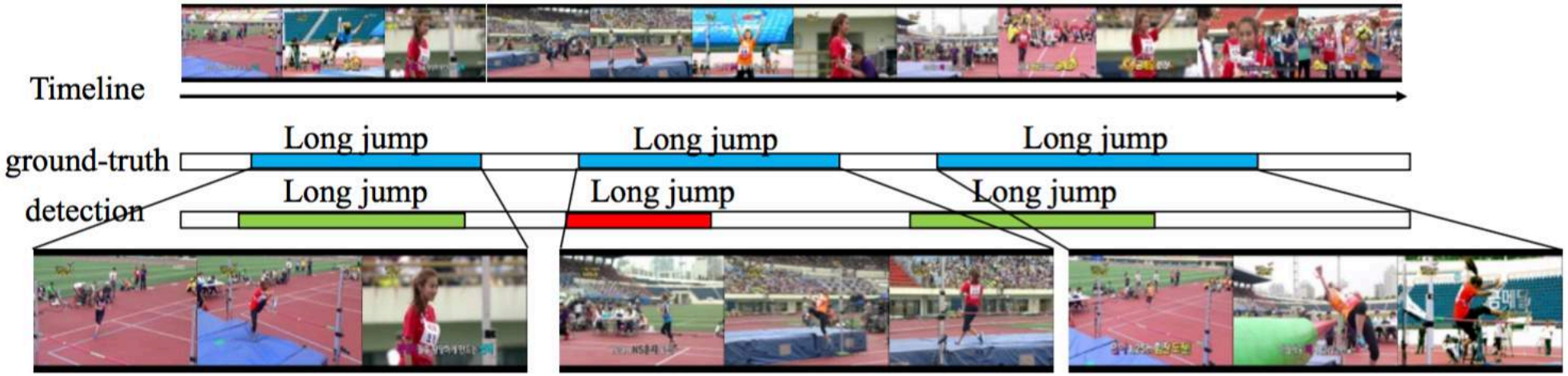
Multi-Person/Group Action Classification

Inter-person interaction

Temporal Action Detection/Localization

Input: An Untrimmed Video

Output: 1) action labels; 2) start & end timestamps



Supervision Matters

Strongly Supervised Temporal Action Detection

Weakly Supervised Temporal Action Detection

Labels: Jumping, walking,…

Spatial Temporal Action Detection

Input: Video

Output:

- 1) action labels; 2) person tracks; 3) start-end time
- 1) action labels + bounding boxes at every frame

Action Segmentation

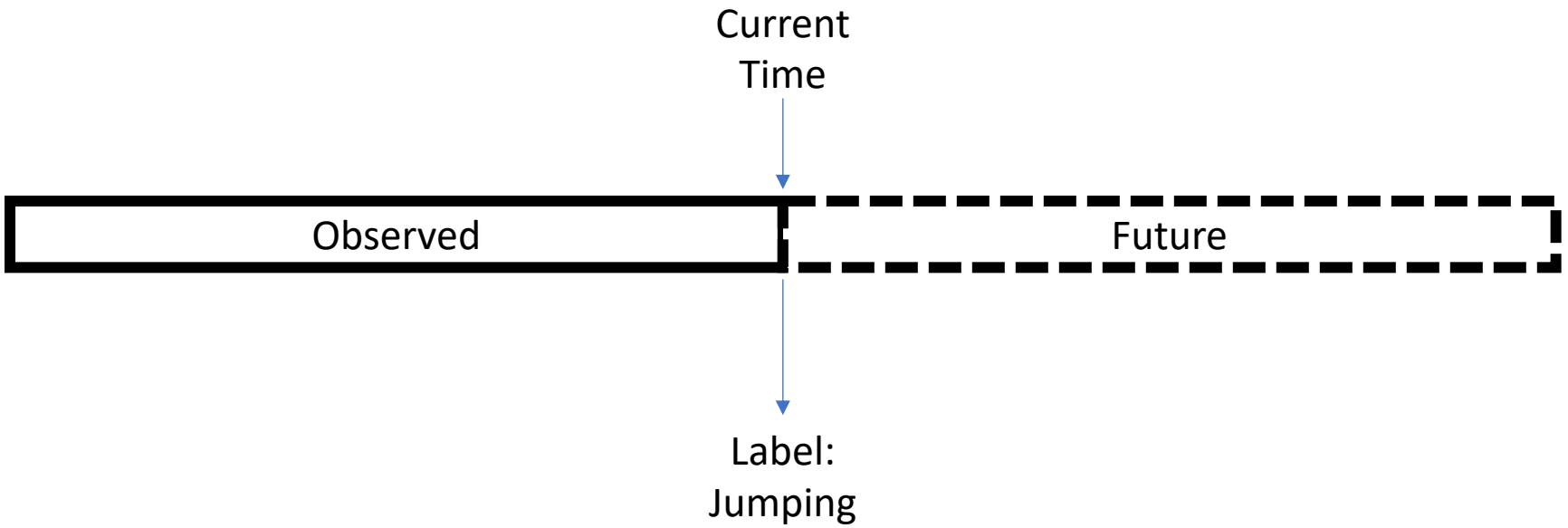
- Temporal Segmentation
  - label + confidence at every frame

- Spatial temporal segmentation

Image courtesy of: *Temporal Convolutional Networks for Action Segmentation and Detection*, Lea, et. Al, in CVPR2017

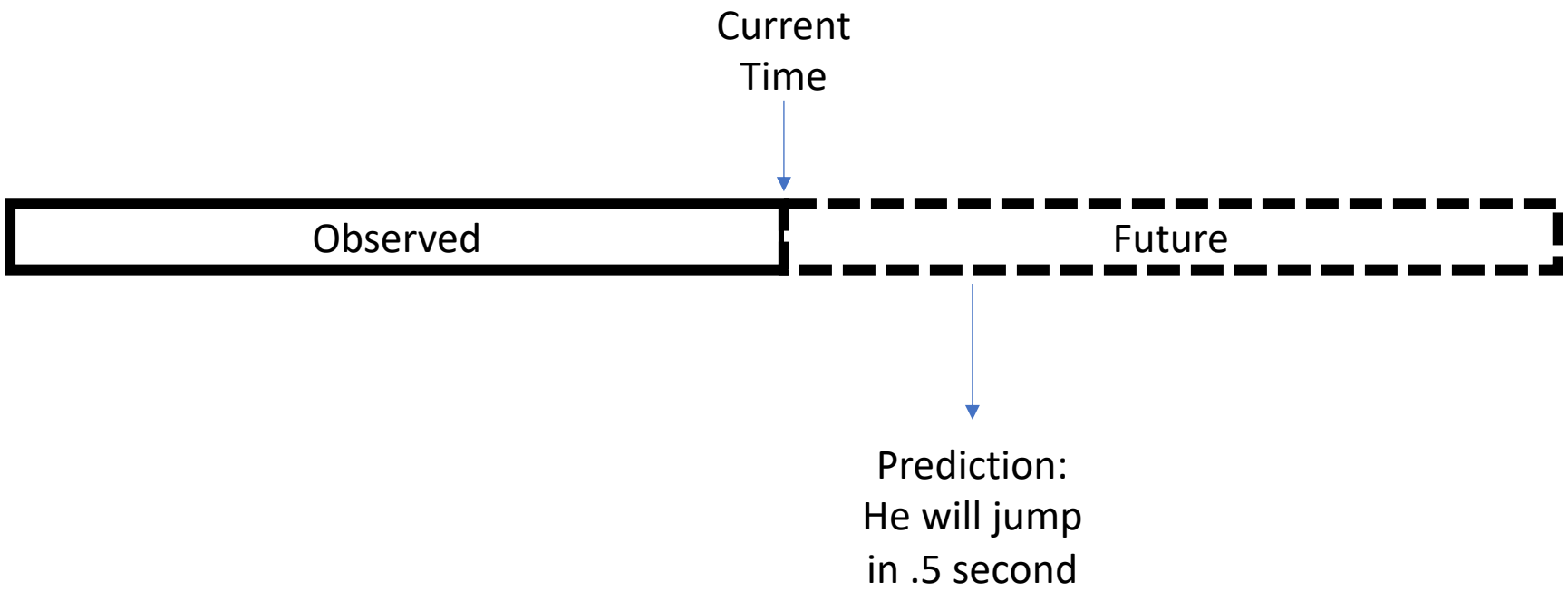
Offline vs. Online

- Online/Real-Time Action Recognition



Look into Future

- Action Prediction/Anticipation



Outline

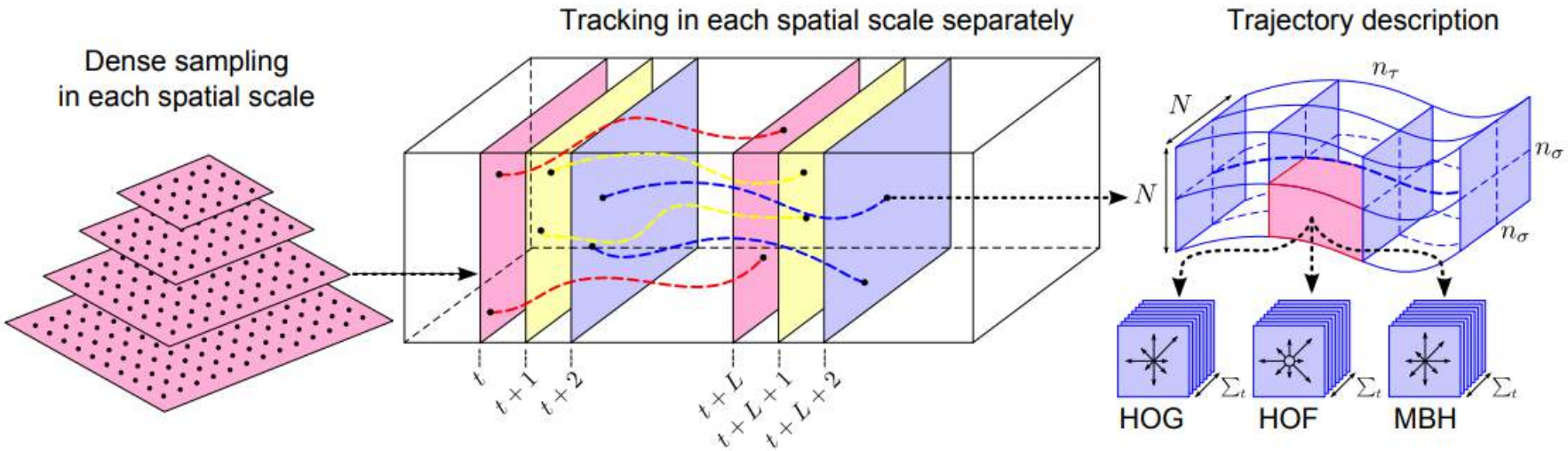
- Data
- Tasks
- **Models**
- Challenges & Problems

Methods for Video Representations

- Hand-Crafted Spatial Temporal Features
  - Space time bag of features[1]
  - Dense Trajectories[2]
  - iDT[3]
- Deep Features
  - CNN
  - RNN
  - Graph-NN

Dense Trajectories

- Dense Trajectories



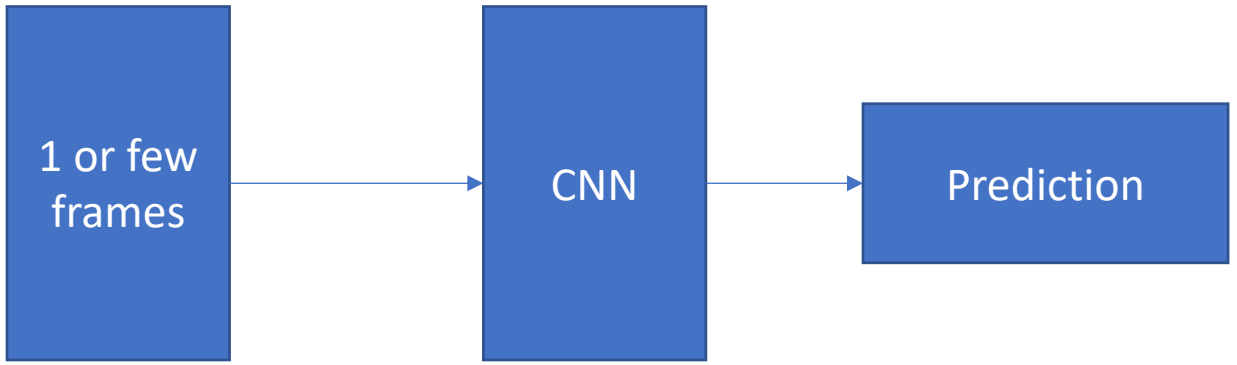
- Improved dense trajectories
  - Camera motion
  - Human mask

[1] Wang, Heng, et al. "Dense trajectories and motion boundary descriptors for action recognition." *International journal of computer vision* 103.1 (2013): 60-79.

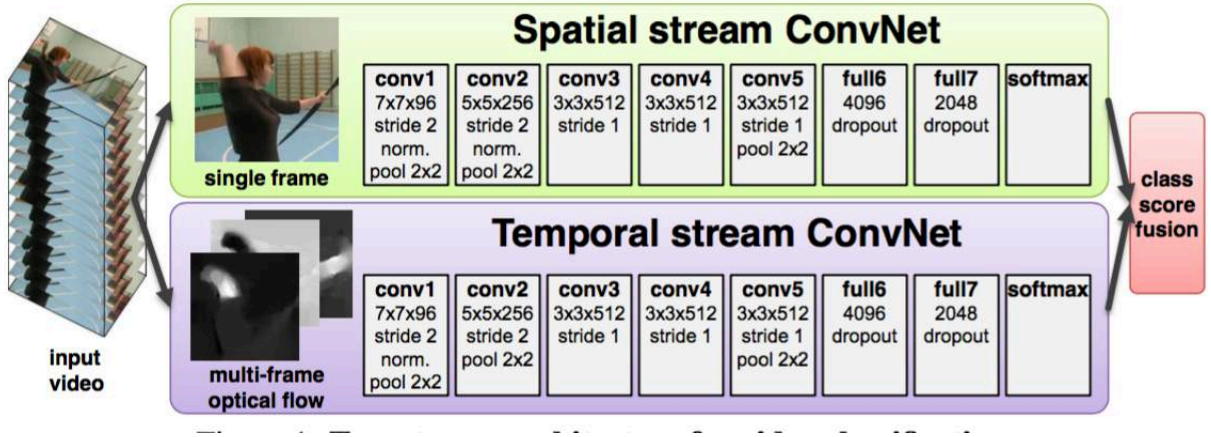
[2] Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." *Proceedings of the IEEE international conference on computer vision*. 2013.

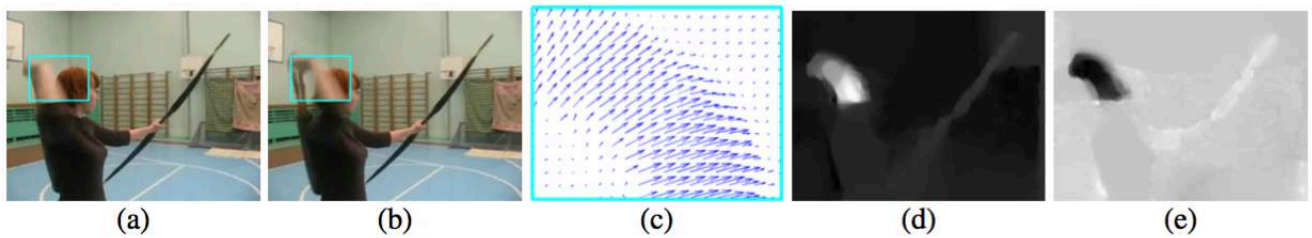
Snippet-Level Deep Representation

- Snippet – One or Few Frames
  - Learning representation with video classification
  - Methods:
    - DeepVideo [1]
    - Two-Stream CNN
    - 3D-CNN



Two-Stream CNN

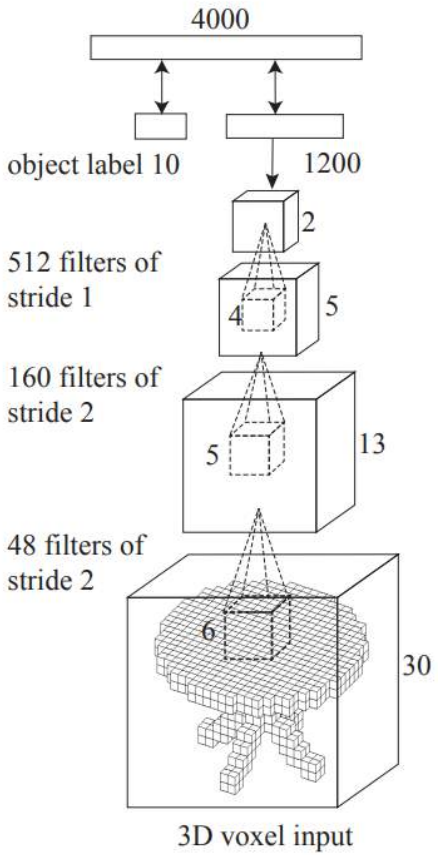




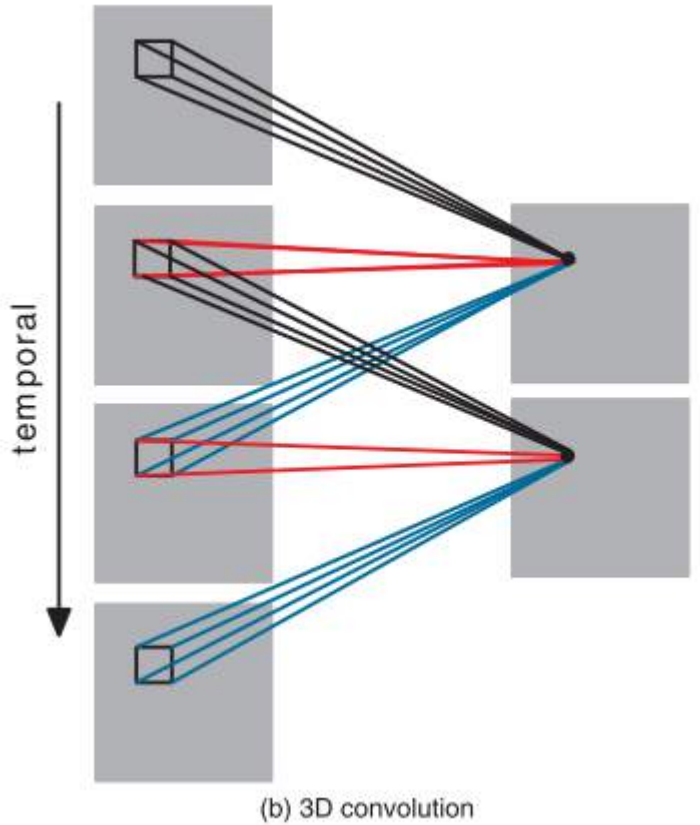
[1] Karen Simonyan and Andrew Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, in NIPS, 2014.

3D CNN

3D spatial convolution



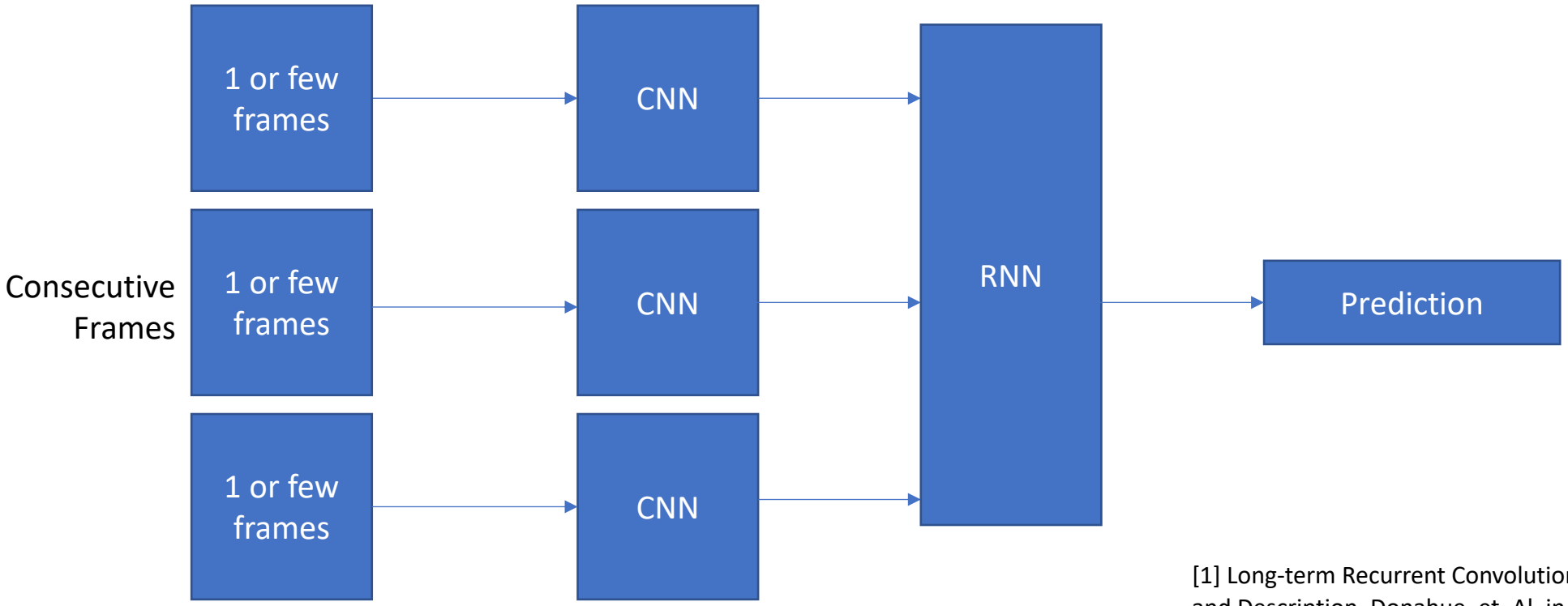
Space Time Convolution



[1] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence*35.1 (2013): 221-231.
[2] Du Tran et al. Learning Spatiotemporal Features with 3D Convolutional Networks, in ICCV, 2015.

Modeling Longer Timespan

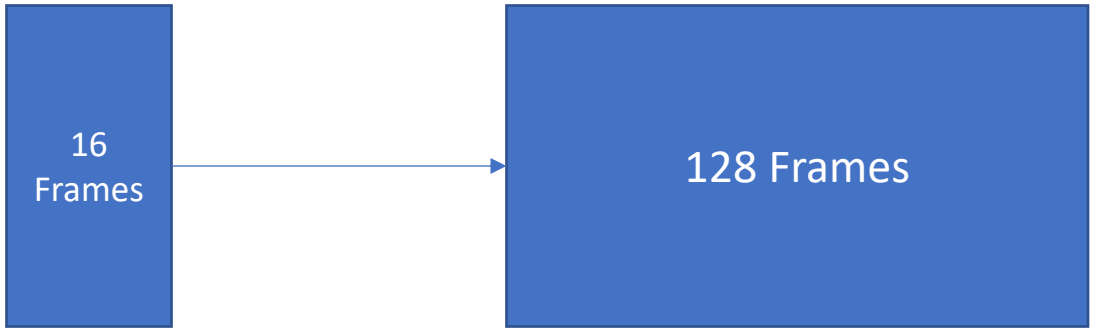
- How to model longer terms?
  - Recurrent Neural Network



[1] Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue, et. Al, in TPAMI

Long-Term 3D CNN

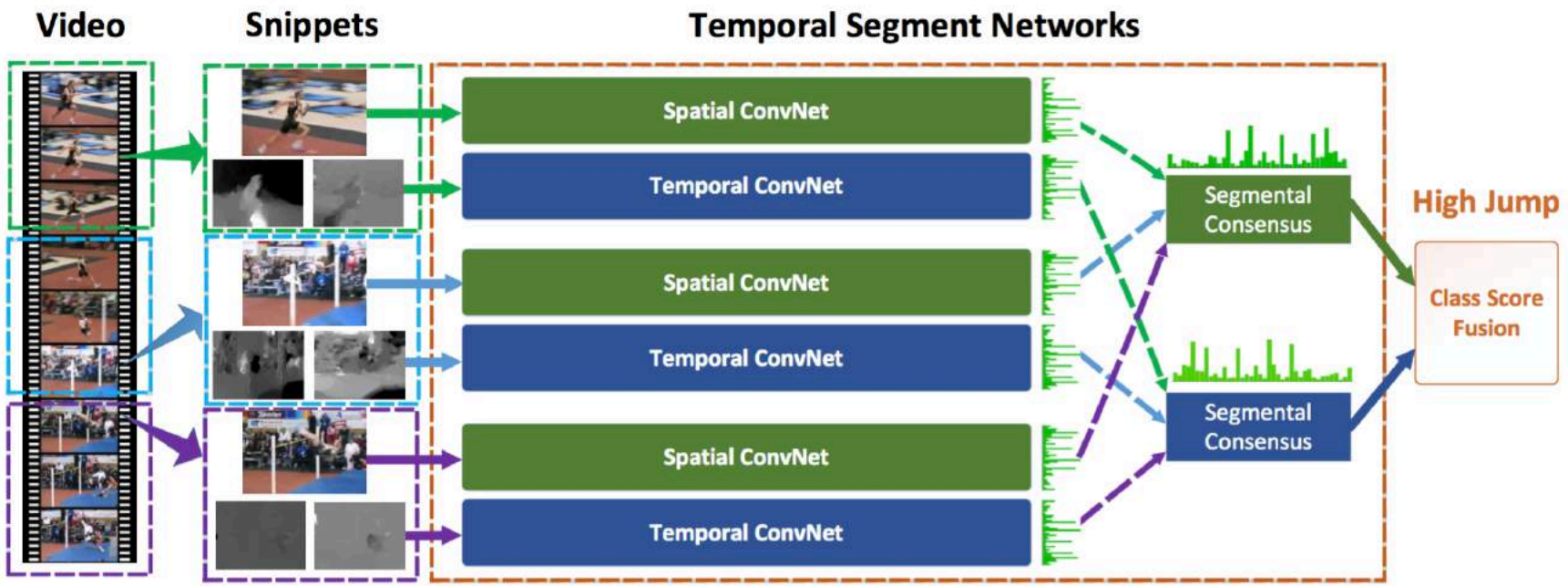
- Let’s put more frames to 3D-CNNs
  - I3D [1]
  - LTC [2]



[1] Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Carreira, et. al., in CVPR 2017.
[2] Long-term Temporal Convolutions for Action Recognition, Varol, et. al. in TPAMI.

Modeling Longer Timespan

- Temporal Segment Networks
  - Move from dense sampling to sparse sampling

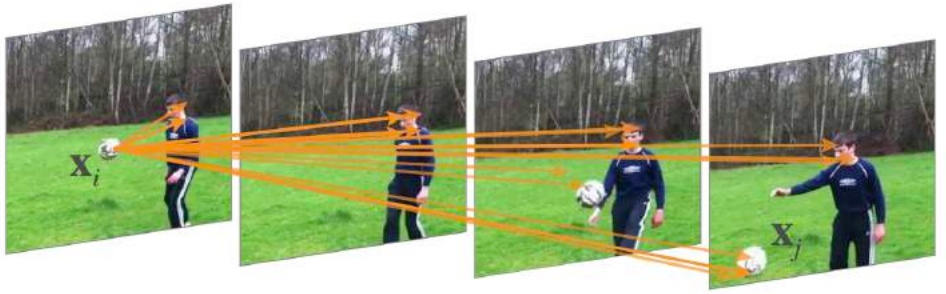


Spatial Temporal Modeling

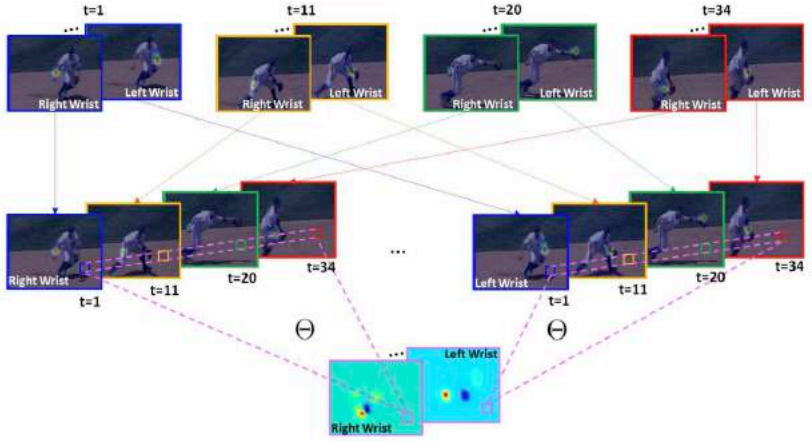
- Regular Convolution
- Attention
  - Self-Attention
  - Pose-aided attention
- Graph Convolution
  - ST-GCN

Attention

- Non-local CNN



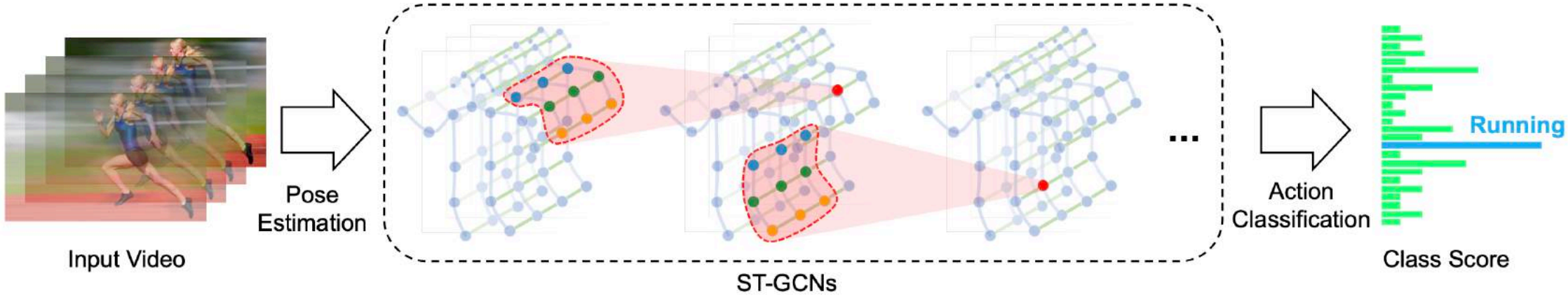
- Pose-Guided Attention



[1] Non-local Neural Networks, Wang et.al., in CVPR 2018
[2] PA3D: Pose-Action 3D Machine for Video Recognition, Yan et.al., in CVPR2019

Graph-NN

- ST-GCN



[1] Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, Yan, et. al. in AAAI 2018
[2] Videos as Space-Time Region Graphs, Wang, et. al., in ECCV 2018.

Outline

- Data
- Tasks
- Models
- **Challenges & Problems**

Challenges

- Accuracy
- Efficiency
  - Can we recognize action from compressed videos?
- Generalization
  - Can our models work on new domains?

Efficiency

- How to make video models run faster?
  - Input
    - Motion vectors
    - RGB Diffs
    - Integrated Optical Flows

Efficiency

- Model efficiency
  - 3D CNN
    - Pseudo-3D Conv
    - Decoupled 3D CNN
  - 2D-3D hybrids
    - ECO-lite

Generalization

- Activities are compositional & fine-grained concepts
  - Playing instrument
    - Playing violin
    - Playing piano….
  - Fine-grained action recognition

[1] FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding, Shao, et. al., in CVPR 2020.
[2] Mimetics: Towards Understanding Human Actions Out of Context, Weinzaepfel, et. al. in Arxiv.

Future Directions

- Data
  - Can we collect larger datasets?
- Model
  - More efficient and more accurate
  - Better spatial-temporal modeling
  - Better motion representation
  - Architecture search
- Applications
  - Entertainment
  - Human-robot interaction
  - More to come…