Action Recognition and Detection with Deep Learning

Yue Zhao
Multimedia Lab, CUHK
https://zhaoyue-zephyrus.github.io

Why do we need to understand action?

- Various real-world applications
  - Anomaly detection in video surveillance
  - Gesture recognition for VR
  - Personalized recommendation/retrieval for video websites/apps (YouTube,Tik-Tok)

Video adapted from https://www.youtube.com/watch?v=QcCjmWwEUgg

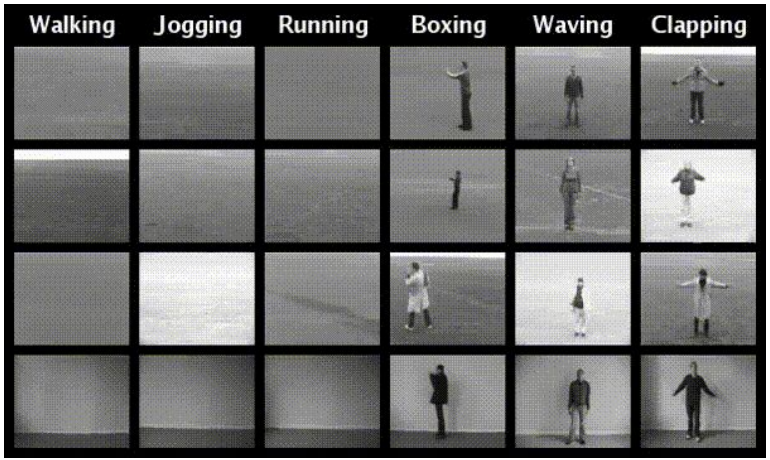
Video adapted from https://www.youtube.com/watch?v=PJqbivkm0Ms.

Overview

- Datasets for video-based action understanding
- Methods for action recognition
  - Before Deep Learning
  - After Deep Learning
- Cutting-edge action recognition
- More for action understanding
  - Temporal action detection
  - Spatial temporal action detection

Datasets (1)

- From restricted scenarios (e.g. KTH) to videos in the wild (e.g. THUMOS'14)

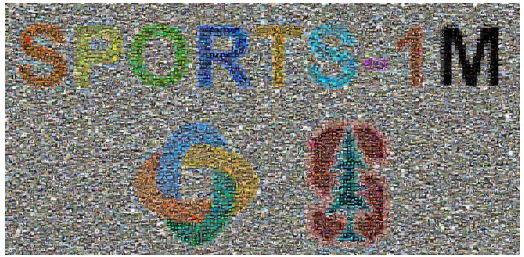


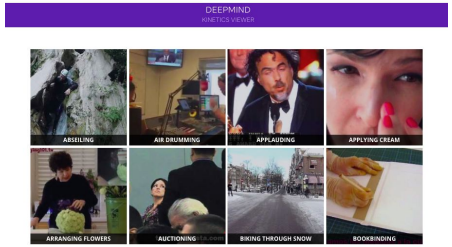
KTH Dataset
(https://www.youtube.com/watch?v=Jm69kbCC17s)

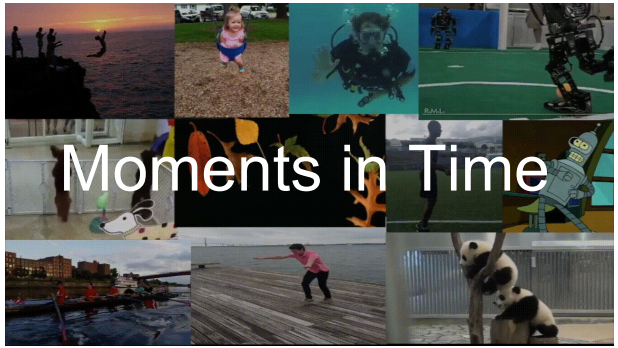
THUMOS'14 Dataset (https://www.crcv.ucf.edu/THUMOS14/)

Datasets (2)

- From small-scale (e.g. Olympic Sports) to larger-scale (Sports-1M, YouTube-8M, Moments in Time, Kinetics-400/600)
- Challenges arise:
  - Storage (It costs many TBs to save the Sports-1M videos.)
  - Computation (It takes multiple GPUs to train a network for days or even weeks.)
  - Imbalanced data (long-tail distribution)

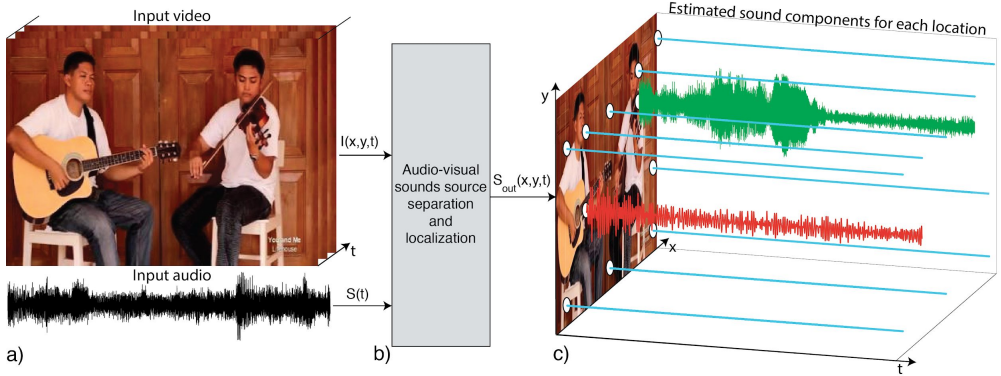


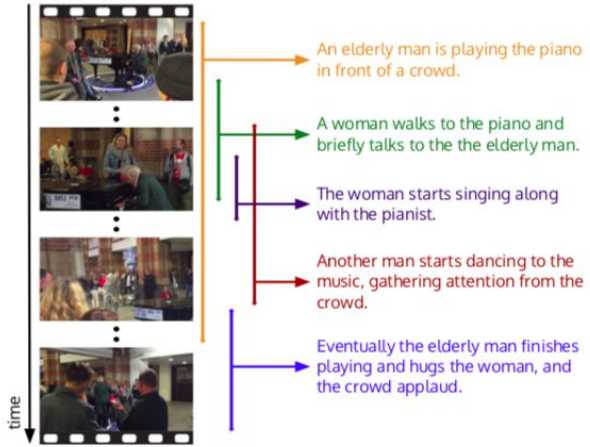




Datasets (3)

- Daily-life: Charades, VLOG
- Egocentric: Epic-Kitchens, Charades-Ego
- Multimodal: Visual + X
  - + language => ActivityNet Captions
  - + sound => The sound of pixels
  - + speech => AVA ActiveSpeaker, AVA Speech



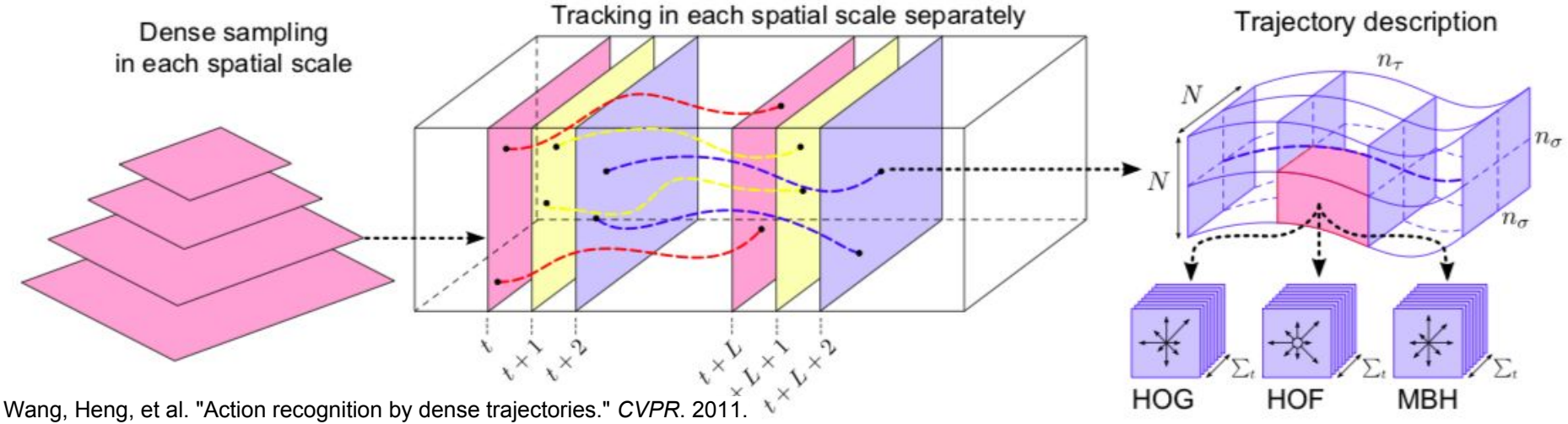


The basic problem - Action recognition

- Given a video clip, output an action prediction.
- Similar to image classification (object recognition)
- The difference is that the input is a sequence of 2D images (3D).

Pre-Deep Learning Methods

- Tracking points of interest (trajectory) and extract local descriptors (HOG, HOF, MBH) thereon.
- The trajectory can be improved by compensating the camera motion.

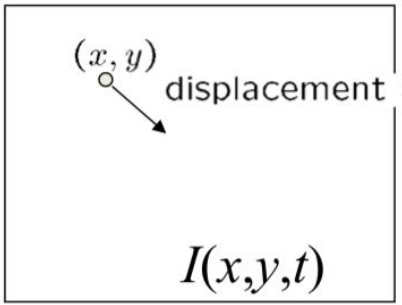


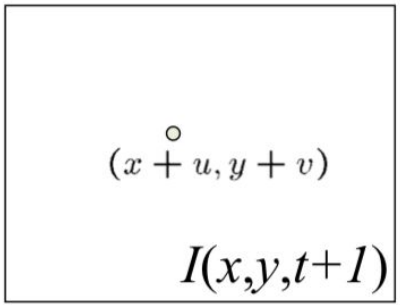
Wang, Heng, et al. "Action recognition by dense trajectories." *CVPR*. 2011.

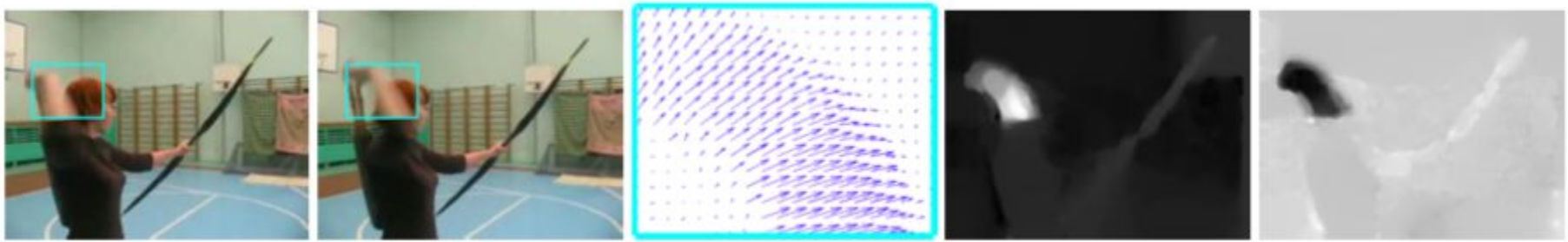
Optical Flow

- $I(x, y, t) = I(x + u, y + v, t + 1)$ (Brightness constancy equation)
- $I(x + u, y + v, t + 1) \approx I(x, y, t)$

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



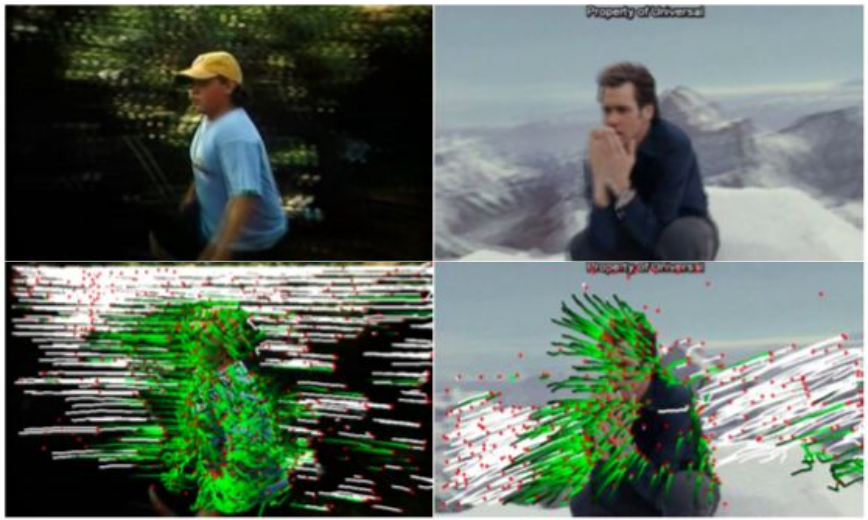


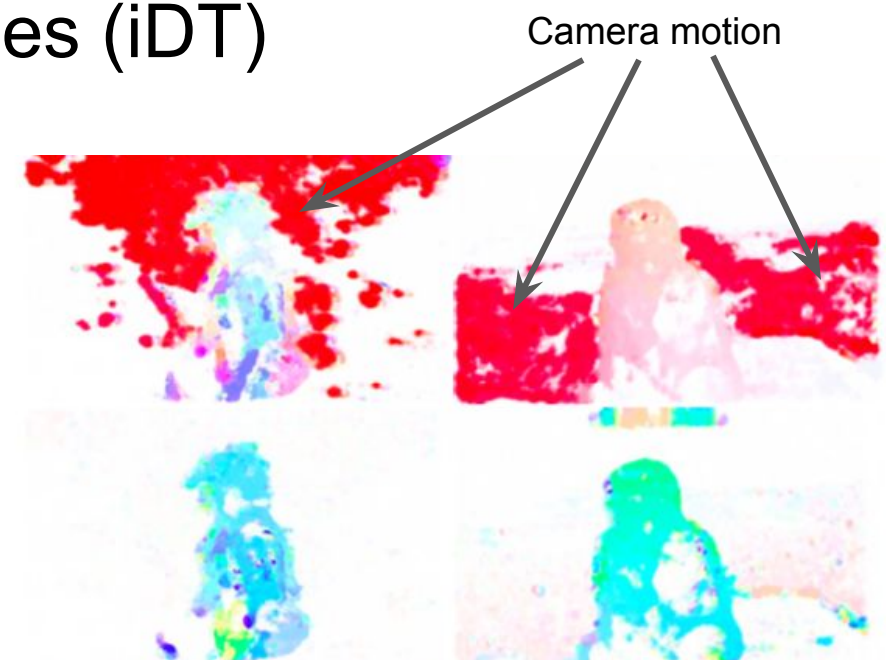


horizontal component

vertical component

Improved Dense Trajectories (iDT)

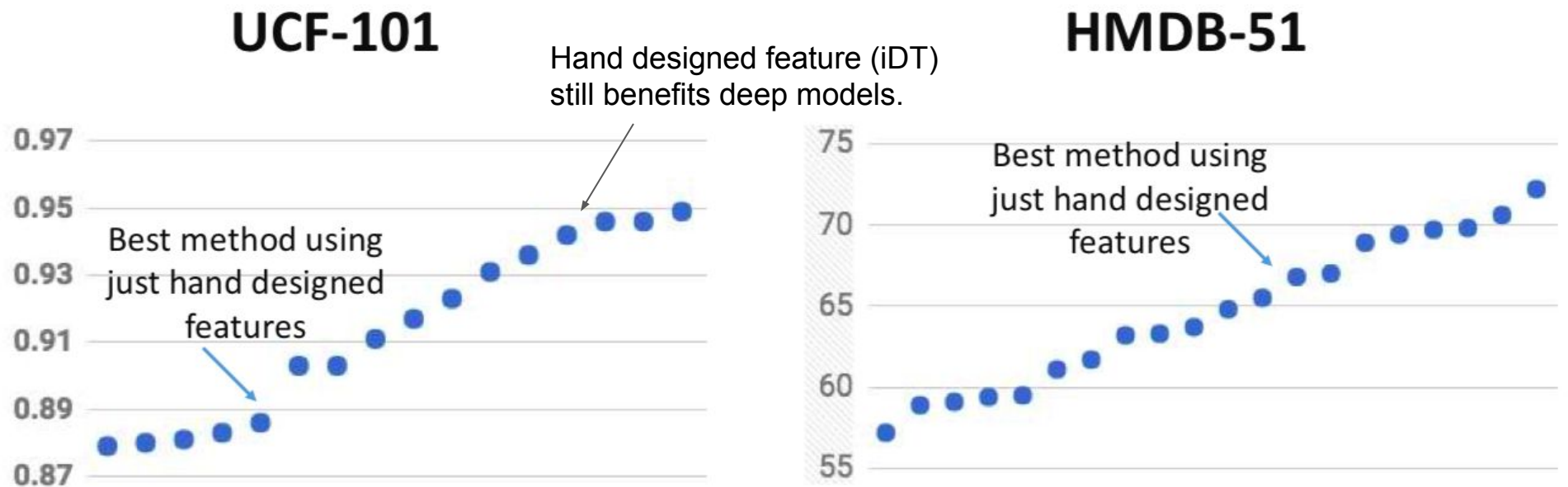




Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." *ICCV*. 2013.

Post-Deep Learning Methods

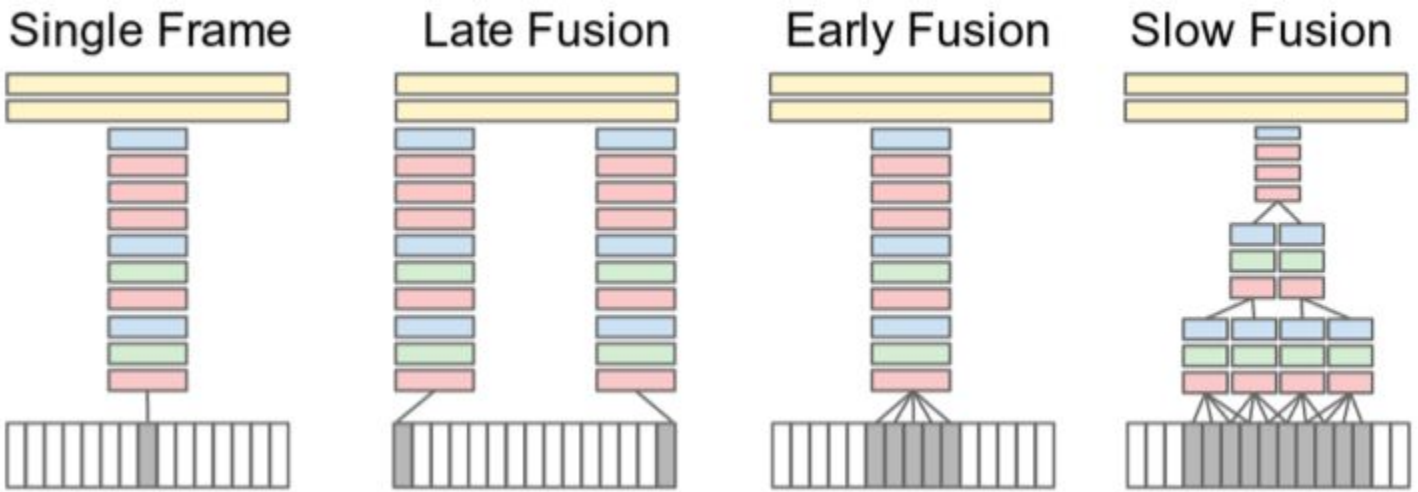
- Follow the roadmap of image classification: AlexNet, VGG, Inception, ResNet



Adapted from AZ's slides at YouTube-8M challenge workshop at ECCV 2018.
https://static.googleusercontent.com/media/research.google.com/zh-CN//youtube8m/workshop2018/p_i01.pdf

Key issue

- Extend CNN in the time domain to exploit the spatio-temporal information.



Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *CVPR*. 2014.

Two-stream Architecture

- Spatial: appearance
- Temporal: motion (optical flow)

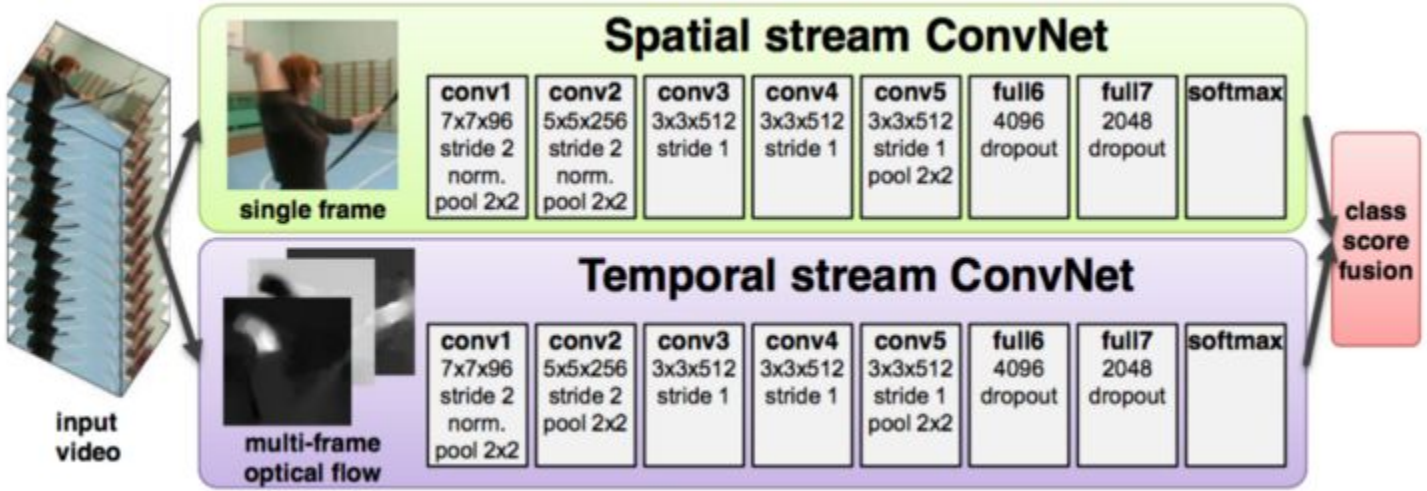
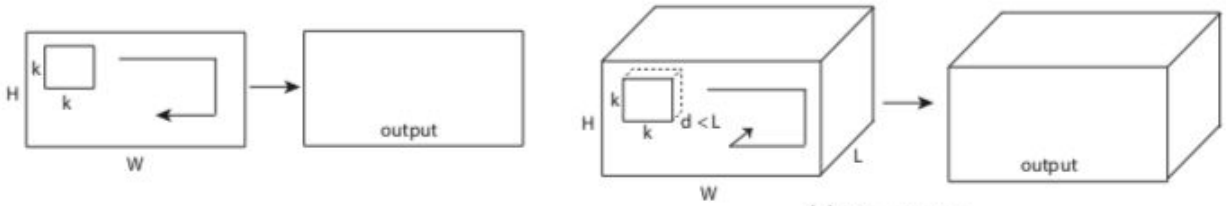


Figure 1: **Two-stream architecture for video classification.**

Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *NIPS*. 2014.

3D Networks

- Applying 3D convolution on a video volume results in another volume, preserving the temporal information of the input signal.
- Problem: model complexity increases drastically
- Tricks:
  - Leverage the good representation of 2D networks by inflating 2D conv kernels to 3D.
  - Feed it with more data! (Kinetics)



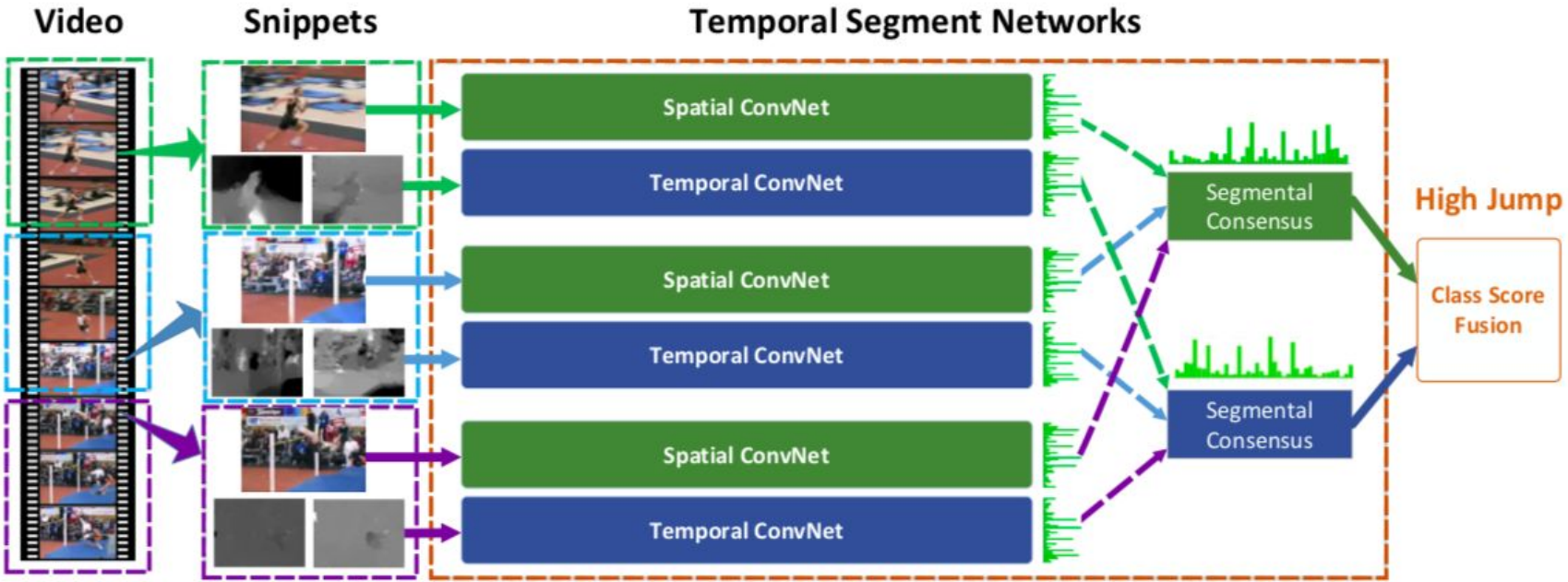
Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." *ICCV*. 2015.

Cutting-edge Action Recognition

- How can we model the long-term temporal information? (TSN, **Wang, Limin**, et al. ECCV 2016 & PAMI 2018, TRN, **Zhou, Bolei**, et al. ECCV 2018)
- How can we better model the short-term motion information?
  - Is optical flow good enough for action recognition? (Sevilla-Lara, Laura, et al. GCPR 2018)
  - Insert a CNN for motion estimation into the two-stream architecture. (Zhu, Yi, et al. ACCV 2018)
  - Use cost volume to coarsely estimate the motion. (Zhao, Yue, et al. CVPR, 2018)
- How can we take advantage of the motion information?
  - Use motion information to align appearance feature. (Zhao, Yue, et al. NeurIPS, 2018)
- How can we leverage the interaction between human (subject) and object?
  - (Wang, Xiaolong, and Gupta. ECCV 2018)
- More efficient action recognition
  - 2D convolution operation at early stage + low-cost 3D convolution operation at higher level (ECO, Zolfaghari et. al. ECCV, 2018)
  - 2D convolution operation + exchange temporal information across frames by temporal shuffle (TSM, Lin, Ji, et al. arXiv: 1811.08383)

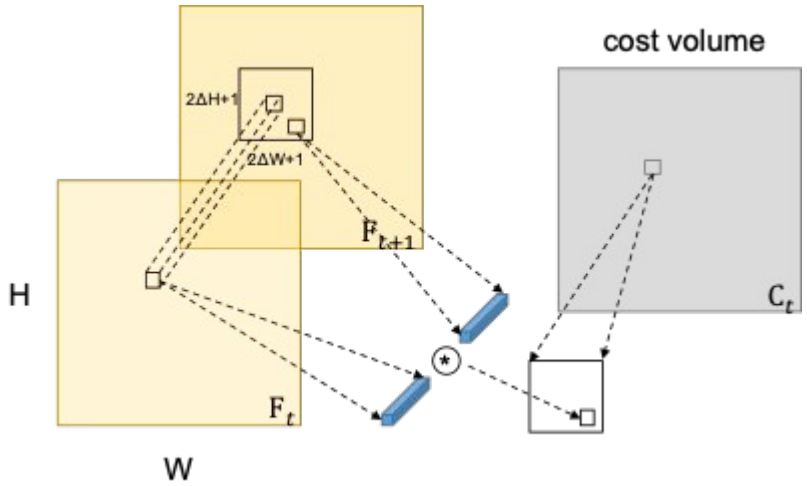
Temporal Segment Networks

- Long-term temporal modeling.



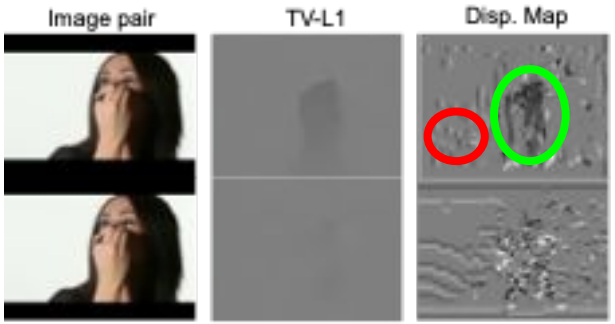
Motion estimation via cost volume

- Cost volume construction via matching similarities.
- Cost volume processing by computing expected displacement.
- Directly from RGB frames without optical flow.



$$v^y_{i,j} = \sum_{\delta i=-\Delta H}^{\Delta H} \rho^y_{(i,j)}(\delta i) \cdot \delta i,$$

$$\rho^y_{(i,j)}(\delta i) = \frac{\exp(c_{i,j}(\delta i, \delta j)/\tau)}{\sum_{\delta i'} \exp(c_{i,j}(\delta i', \delta j)/\tau)},$$



Zhao, Yue, Yuanjun Xiong, and Dahua Lin. "Recognize actions by disentangling components of dynamics." CV*PR*. 2018.

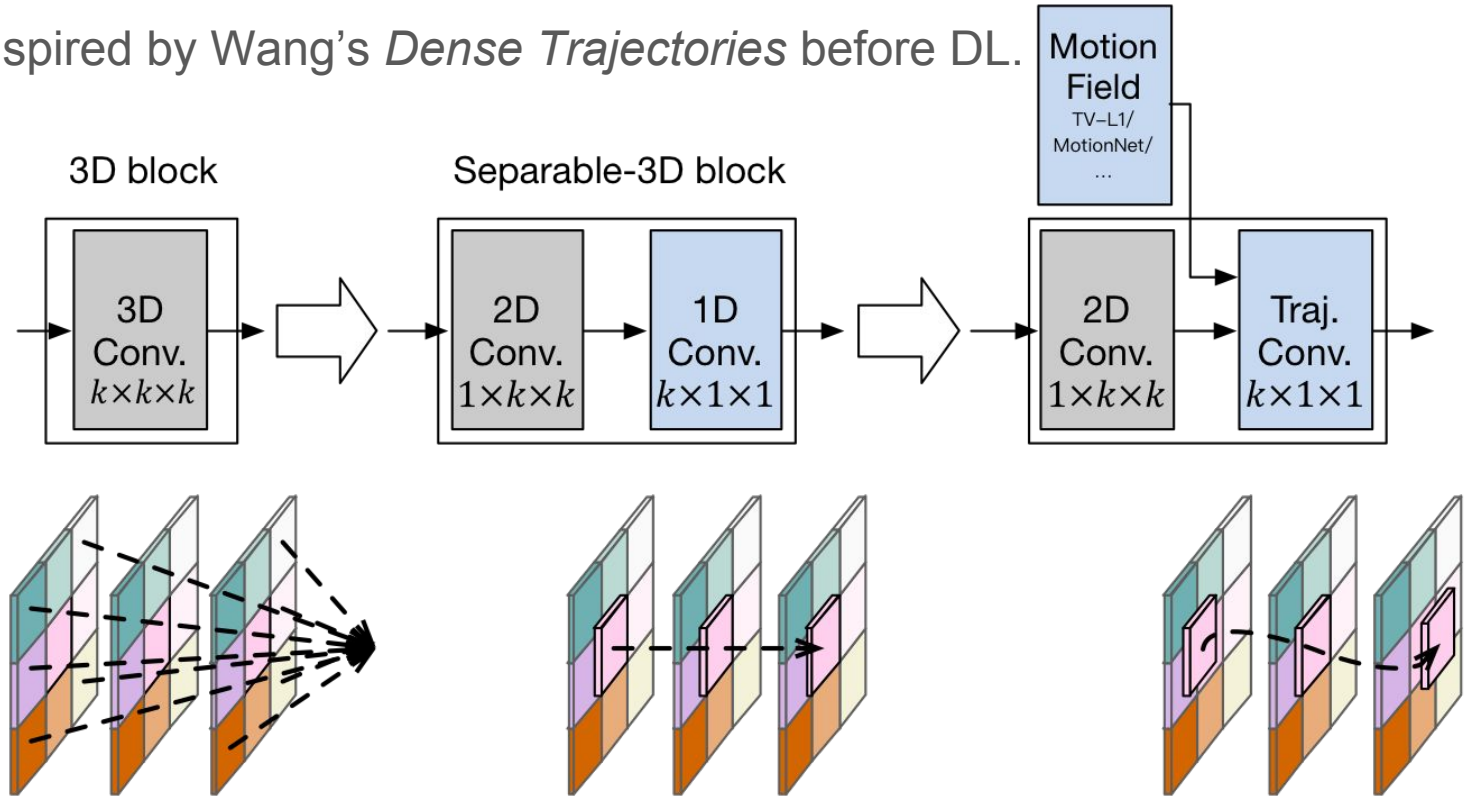
Motion estimation via cost volume

- The whole architecture outperforms other methods which only take RGB frames as input, maintaining real-time speed (>40 FPS).

Method	Dataset	Accuracy
C3D (1 nets) [23]	Sports-1M	82.3%
C3D (3 nets) [23]	Sports-1M	85.2%
Pseudo-3D ResNet [18]	ImageNet+Sports-1M	88.6%
RGB-I3D [3]	ImageNet	84.5%
RGB+EMV [33]	ImageNet	86.4%
TSN (RGB) [29]	ImageNet	85.7%
TSN (RGB+RGB-Diff) [29]	ImageNet	91.0%
Ours	ImageNet	91.8%
Ours	ImageNet+Kinetics	95.9%

TrajectoryNet

- Inspired by Wang's *Dense Trajectories* before DL.

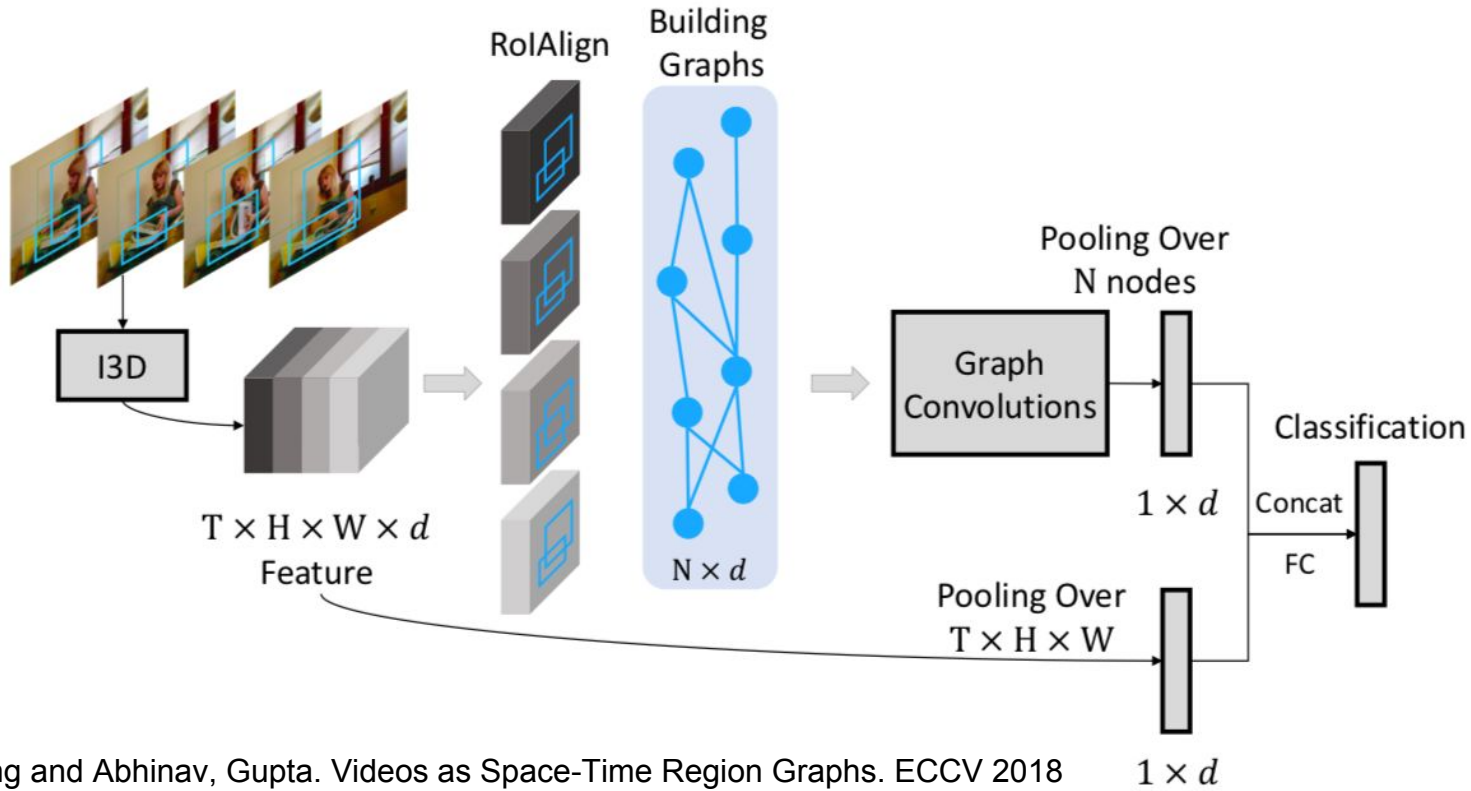


Method	Use deep feature?	Feature tracking?	End-to-end?
STIP	✗	✗	✗
DT, iDT	✗	✓	✗
TSN, I3D	✓	✗	✓
TDD	✓	✓	✗
TrajectoryNet (Ours)	✓	✓	✓

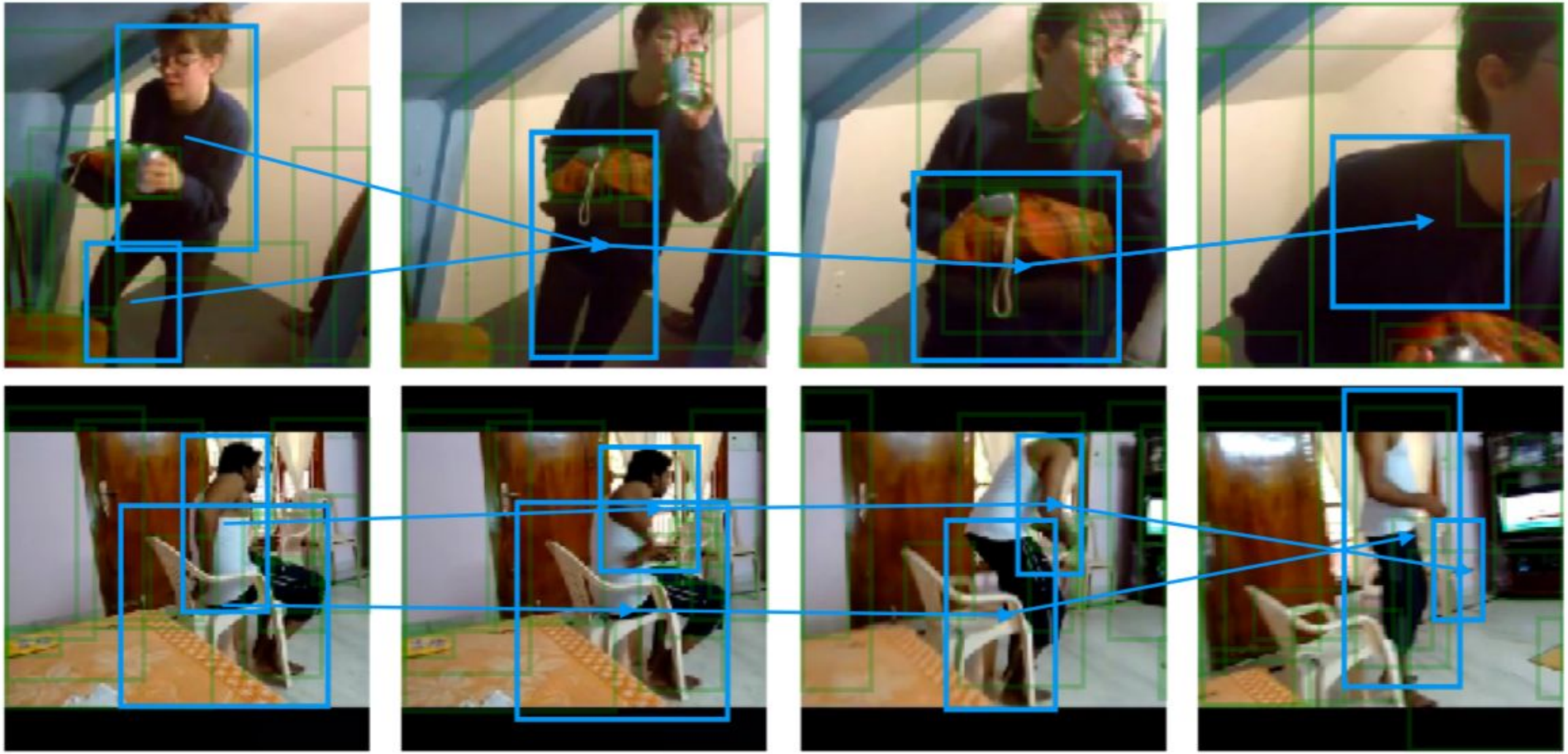
- Achieve competitive results with a relatively small model.

Method	Backbone network	Pre-train	Val Top-1
3D-CNN [8]	C3D	Sports-1M	11.5
MultiScale TRN [50]	BN-Inception	ImageNet	34.4
ECO lite [52]	BN-Inception + 3D-ResNet18	Kinetics	46.4
Non-local I3D + GCN [44]	ResNet-50	Kinetics	46.1
TrajectoryNet-MotionNet-(17) w/o. motion	ResNet-18	ImageNet	43.3
TrajectoryNet-MotionNet-(17) w/. motion	ResNet-18	ImageNet	44.0
TrajectoryNet-MotionNet-(17) w/o. motion	ResNet-18	Kinetics	47.8

Videos as Space-Time Region Graphs



Xiaolong, Wang and Abhinav, Gupta. Videos as Space-Time Region Graphs. ECCV 2018

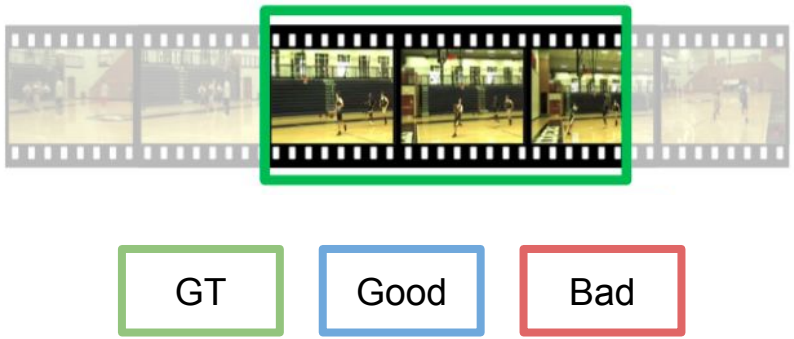


More for Action Understanding

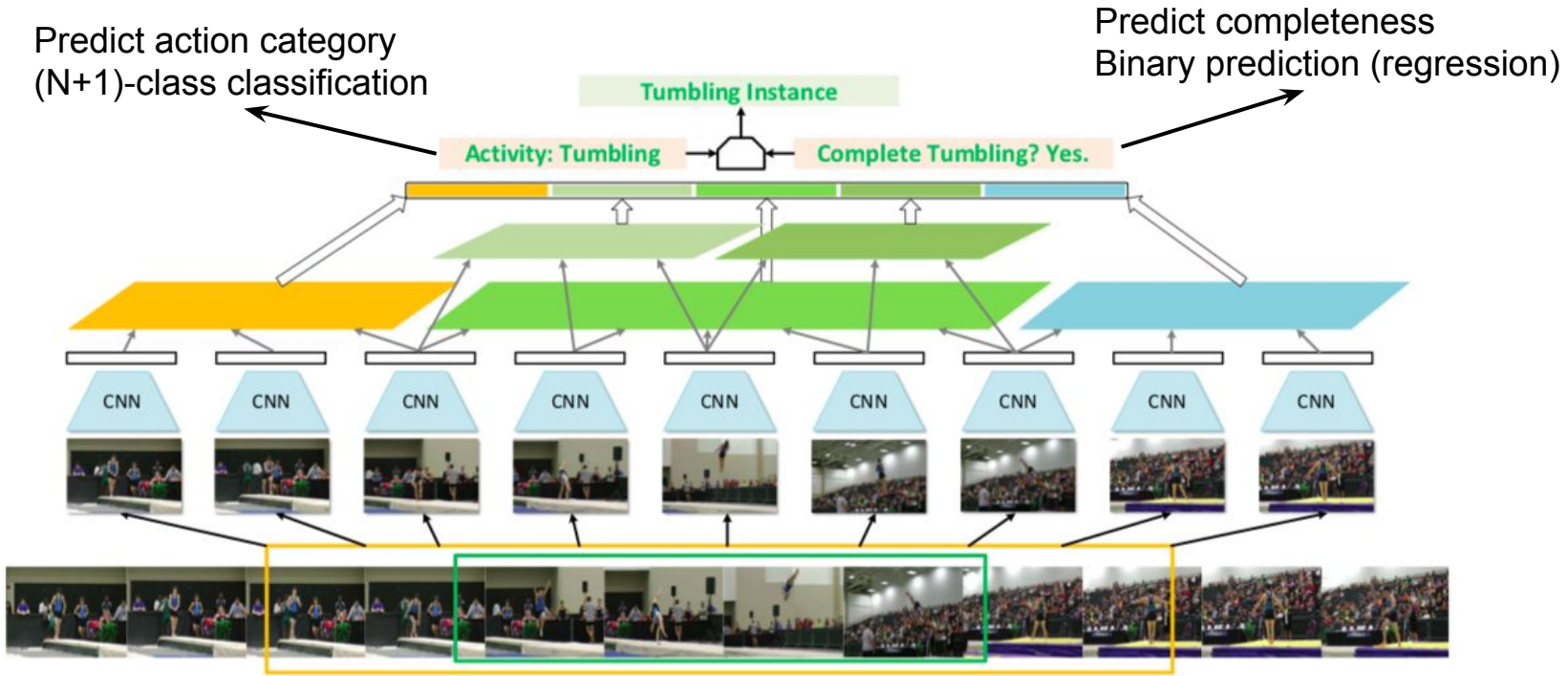
- Temporal action detection
- Spatial temporal action detection

Temporal Action Detection

- Action recognition in trimmed videos (3~10-sec clips) can be done fairly well.
  - Over 90% top-1 accuracy on ActivityNet (200 classes).
  - Nearly 80% top-1 accuracy on Kinetics-400/600.
- Precise temporal localization from untrimmed videos is unsatisfactory.
- Automatic video editing/highlighting; anomaly detection



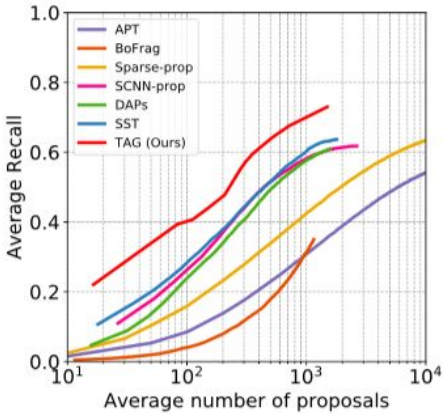
Structured Segment Networks

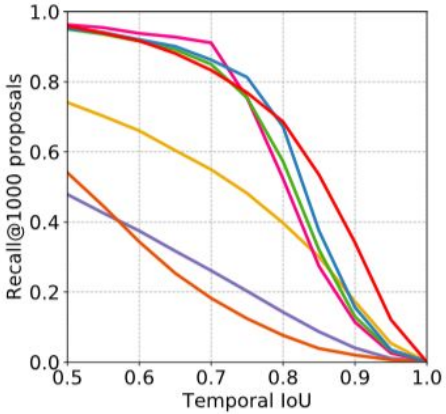


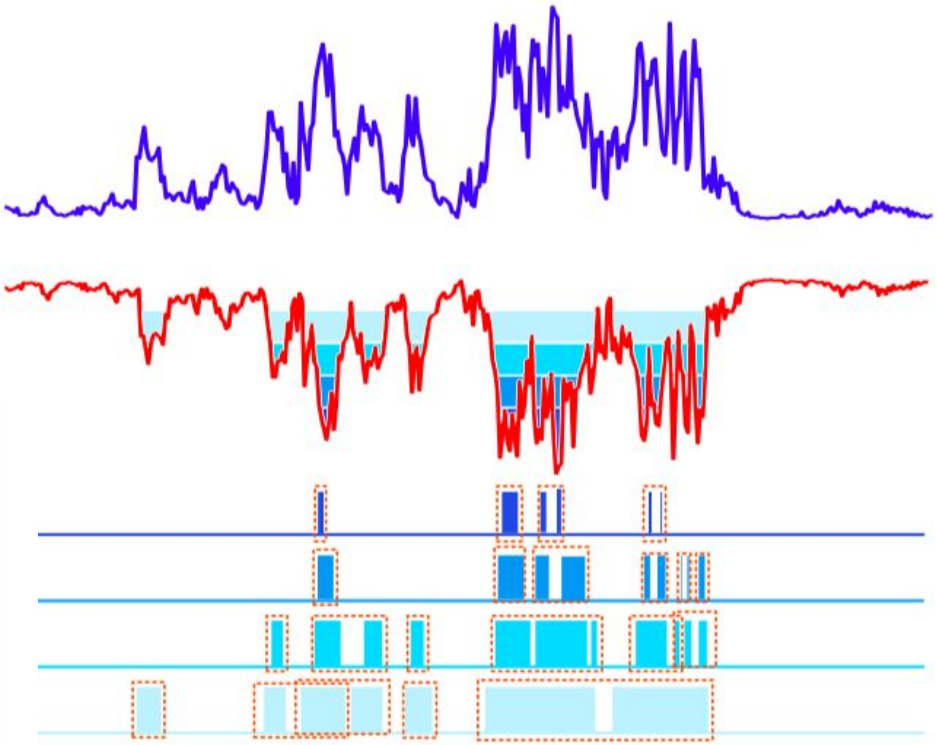
Zhao, Yue, et al. "Temporal action detection with structured segment networks." *ICCV*. 2017.

Action Proposal Generation via Actionness Grouping

- Sliding windows are redundant and inprecise.
- To alleviate this, temporal actionness group is proposed to generate proposals that are sparse and precise at boundaries.







- State-of-the-Art results on ActivityNet v1.3 and THUMOS14.
- Solid baselines for recently proposed datasets (HACS and COIN).

Hang Zhao, et al. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization, arXiv: 1712.09374.
Yansong Tang, et al. COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis. CVPR 2019

ActivityNet v1.3 (testing), **mAP@α**				
Method	0.5	0.75	0.95	Average
Wang *et. al.* (Wang and Tao, 2016)	42.48	2.88	0.06	14.62
Montes *et. al.* (Montes et al, 2016)	22.37	14.88	4.45	14.81
Singh *et. al.* (Singh et al, 2016)	28.67	17.78	2.88	17.68
Singh *et. al.* (Singh and Cuzzolin, 2016)	36.40	11.05	0.14	17.83
R-C3D (Xu et al, 2017)	28.4	-	-	-
TCN (Dai et al, 2017)	37.49	23.47	4.47	23.58
CDC (Shou et al, 2017)	43.0	25.7	0.2	22.9
SSN (ImageNet) (Zhao et al, 2017)	**43.26**	**28.70**	**5.63**	**28.28**
SSN (ImageNet+Kinetics), single model	-	-	-	**29.34**
SSN + U-SSN (ImageNet+Kinetics), ensemble	-	-	-	**31.86**

THUMOS14, **mAP@α**					
Method	0.1	0.2	0.3	0.4	0.5
Wang *et. al.* (Wang et al, 2014a)	18.2	17.0	14.0	11.7	8.3
Oneata *et. al.* (Oneata et al, 2014)	36.6	33.6	27.0	20.8	14.4
Richard *et. al.* (Richard and Gall, 2016)	39.7	35.7	30.0	23.2	15.2
S-CNN (Shou et al, 2016)	47.7	43.5	36.3	28.7	19.0
Yeung *et. al.* (Yeung et al, 2016)	48.9	44.0	36.0	26.4	17.1
Yuan *et. al.* (Yuan et al, 2016)	51.4	42.6	33.6	26.1	18.8
CDC (Shou et al, 2017)	-	-	40.1	29.4	23.3
R-C3D (Xu et al, 2017)	54.5	51.5	44.8	35.6	28.9
SSN† (ImageNet) (Zhao et al, 2017)	**60.3**	**56.2**	**50.6**	**40.8**	**29.1**
SSN* (ImageNet) (Zhao et al, 2017)	**66.0**	**59.4**	**51.9**	**41.0**	**29.8**
SSN† (ImageNet+Kinetics)	**62.2**	**58.7**	**53.3**	**44.5**	**33.3**
SSN* (ImageNet+Kinetics)	**69.3**	**63.6**	**55.2**	**44.8**	**34.3**

Spatial-temporal Action Detection

- Localize the person and determine the action he/she is performing.
- Challenges:
  - Multiple persons in one scene.
  - Diversity of action.
  - Intrinsically imbalanced data.
- Person tracking; patient monitoring

Girdhar, Rohit, et al. "Video Action Transformer Network." *CVPR*. 2019.

Conclusion

- Action recognition is important for many applications.
- Action understanding is far from being solved.
  - The good: recognition accuracy keeps improving.
  - The bad: more structured analysis is missing - temporal localization (detection), spatial-temporal detection, …
  - The ugly: open problem - how do we human perceive and understand action and how can we use such knowledge to help computer do so?

Q&A