视频行为识别年度进展

-Toward Deep Understanding of Human Actions in the Wild-

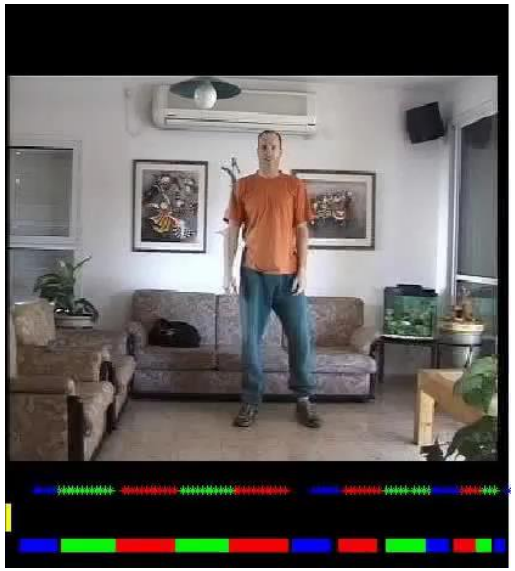
乔宇

中国科学院深圳先进技术研究院

2017-April-23

Action Understanding

- The goal of human action recognition is to automatically detect and classify ongoing activities from an input video (i.e. a sequence of images frames).
  - Human vision system is very effective in perceiving and predicting actions through visual information.
  - A basic problem in computer vision, with wide applications.



Action recognition

Punck

Kick

Duck

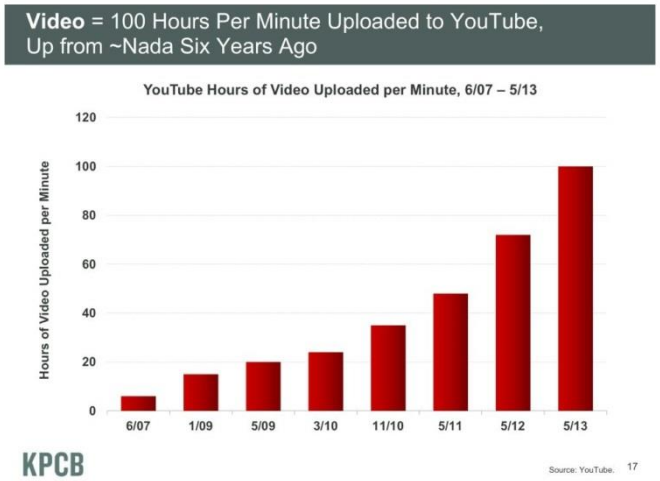
…

Application 1-Surveillance

- Ubiquitous surveillance cameras in public places.
  - More than 60M in China
  - A person is monitored average 300 times / day in London.
  - Current surveillance system mainly record without understanding human action and event in video.
- Understanding human activity is important for intelligent surveillance system.

Application 2- Online Videos

- Explosive growth of online videos



YOUKU 优酷

Users > 263million

User online time>22billion hrs/season

- Many online videos contain action or activity
- Action recognition has important applications in video tagging and content based video retrieval.

More Applications

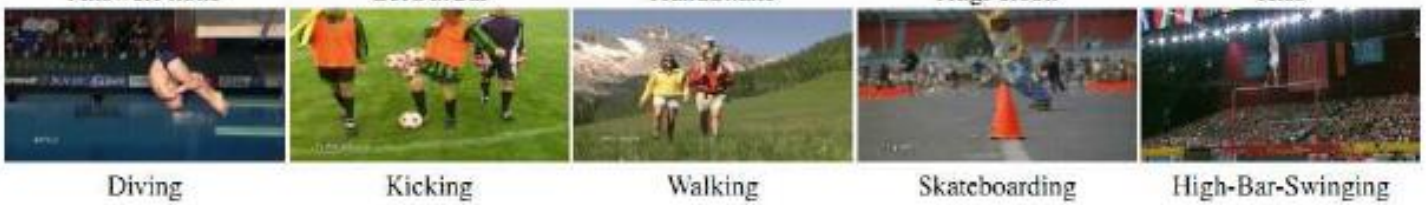
Action Datasets

KTH

UCF Sports

Hollywood

Olympic



Action recognition “in the lab”: KTH, Weizmann etc.
Action recognition “in TV, Moive”: UCF Sports, Hollywood etc.
Action recognition “in the wild”: Olympic, HMDB51, UCF101 etc.

Action Video Dataset list

Dataset	Year	Actions	Clips per Action	Settings	SoTA
KTH [82]	2004	6	400	controlled settings	≥ 0.95
Weizmann [34]	2005	9	9	controlled settings	≥ 0.95
IXMAS [114]	2006	11	33	multiview	≥ 0.95
Hollywood [54]	2008	8	30 - 140	movie	≥ 0.60
UCF Sports [76]	2009	9	14 - 36	sports broadcast	≥ 0.95
Hollywood2 [61]	2009	12	61 - 278	movie	$\simeq 0.70$
UCF YouTube [59]	2009	11	100	web video	≥ 0.90
MSR [120]	2009	3	14-25	indoor and outdoor	≥ 0.95
High Five [68]	2010	4	50	TV shows	$\simeq 0.65$
UT-interaction [78]	2010	6	20	controlled settings	≥ 0.95
Olympic Sports [63]	2010	16	21 - 67	web video	$\simeq 0.90$
UCF50 [75]	2010	50	min. 100	web video	$\simeq 0.95$
HMDB51 [51]	2011	51	min. 101	movie & web video	$\simeq 0.65$
Cooking Dataset [77]	2012	65	-	controlled settings	$\simeq 0.50$
UCF101 [89]	2012	101	min. 100	web video	$\simeq 0.90$
Sports-1M [47]	2014	487	average 2327	web video	$\simeq 0.64$

ActivityNet 2016

ACTIVITYNET

200 categories， 648 hrs video， 10k for training , 5k for testing

http://activity-net.org/challenges/2016/

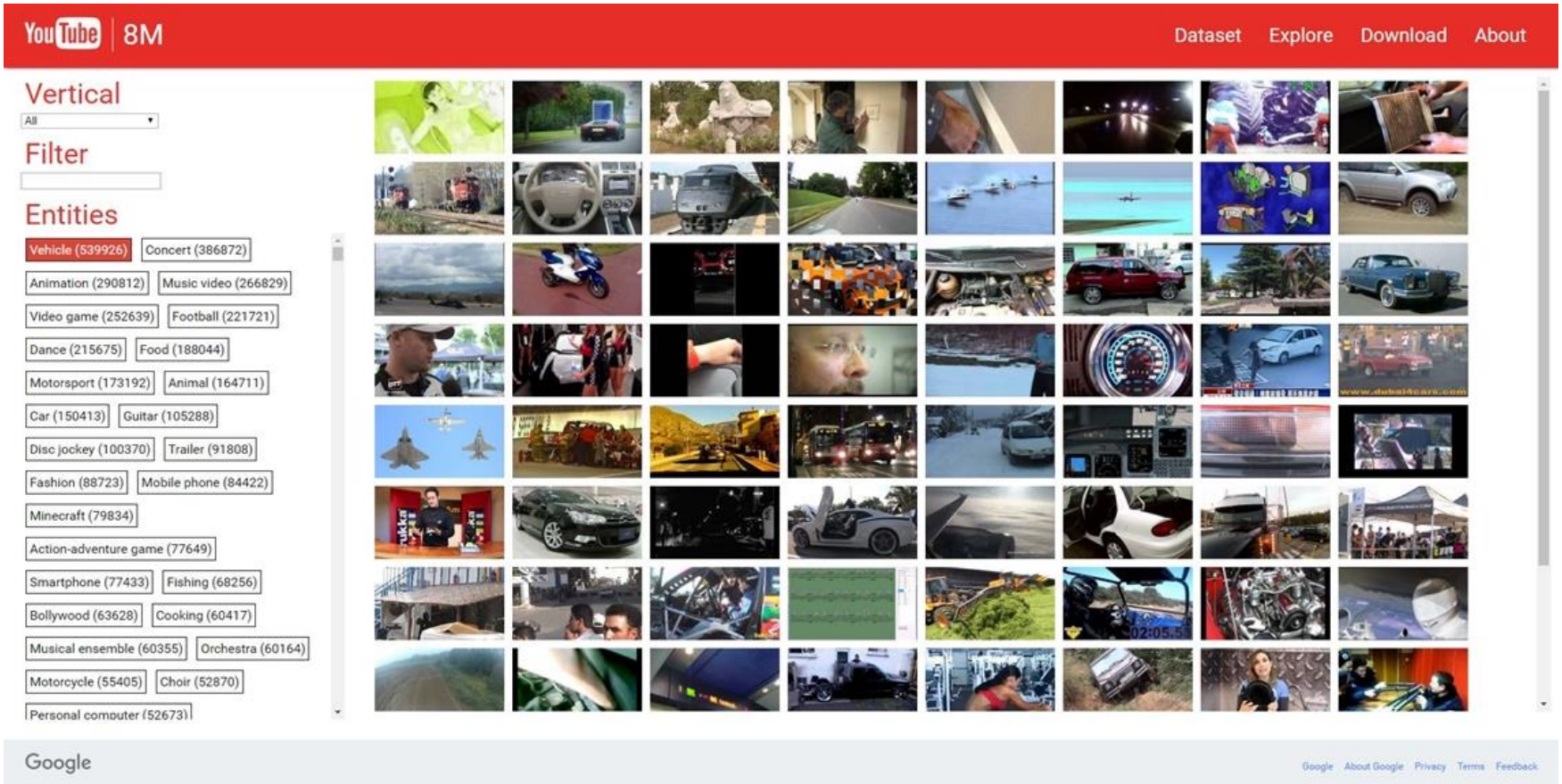
Achieve NO 1 in classification task in ActivityNet 2016 among 24 teams.

Validation Set	mAp	Top-3 Acc.
Visual	90.4%	95.2%
Audio	15.2%	29.1%
Visual + Audio	90.9%	95.6%

Testing Set	mAP	Top-3 Acc.
Visual CNN (Single)	91.2%	95.6%
Final Ensemble	93.2%	96.4%

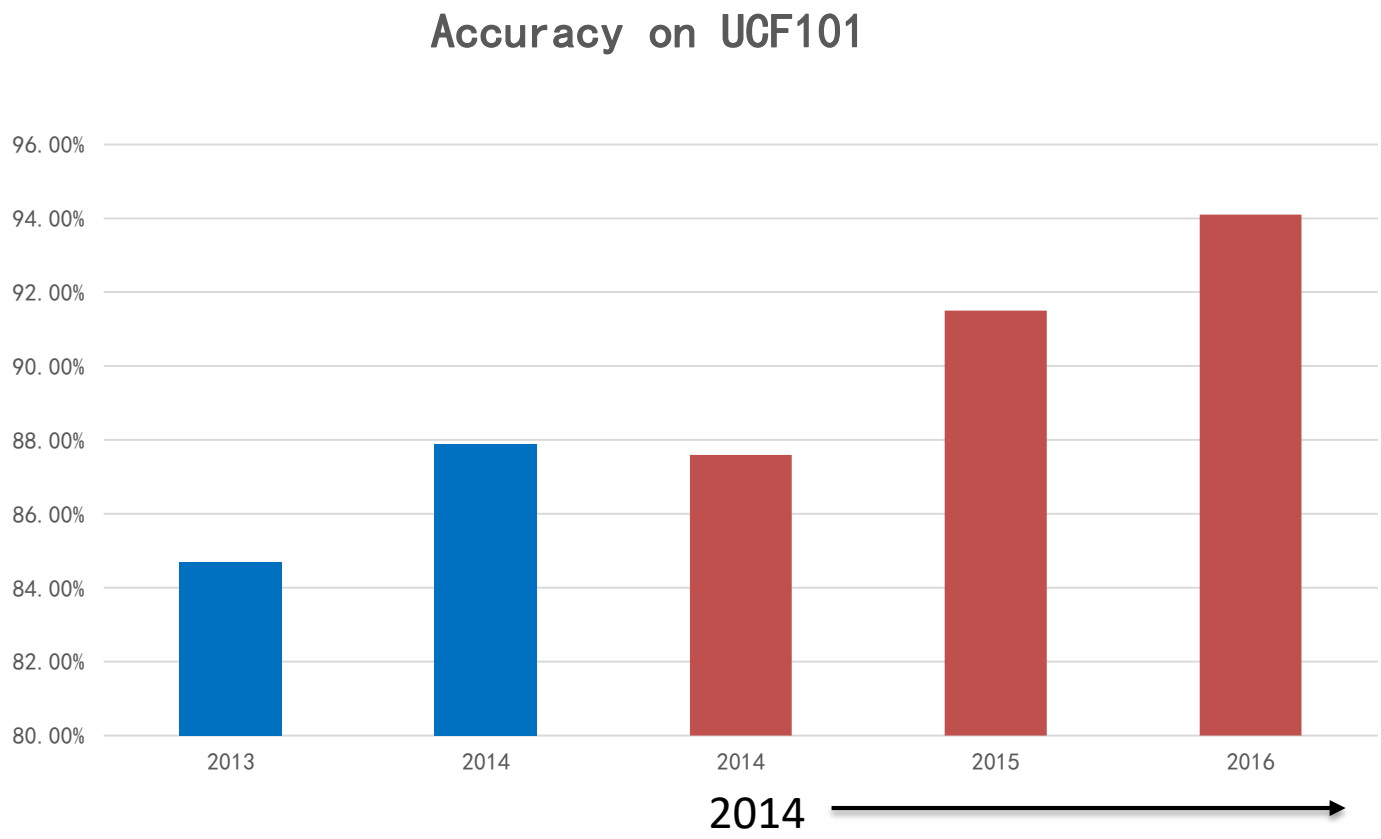
Youtube-8M

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



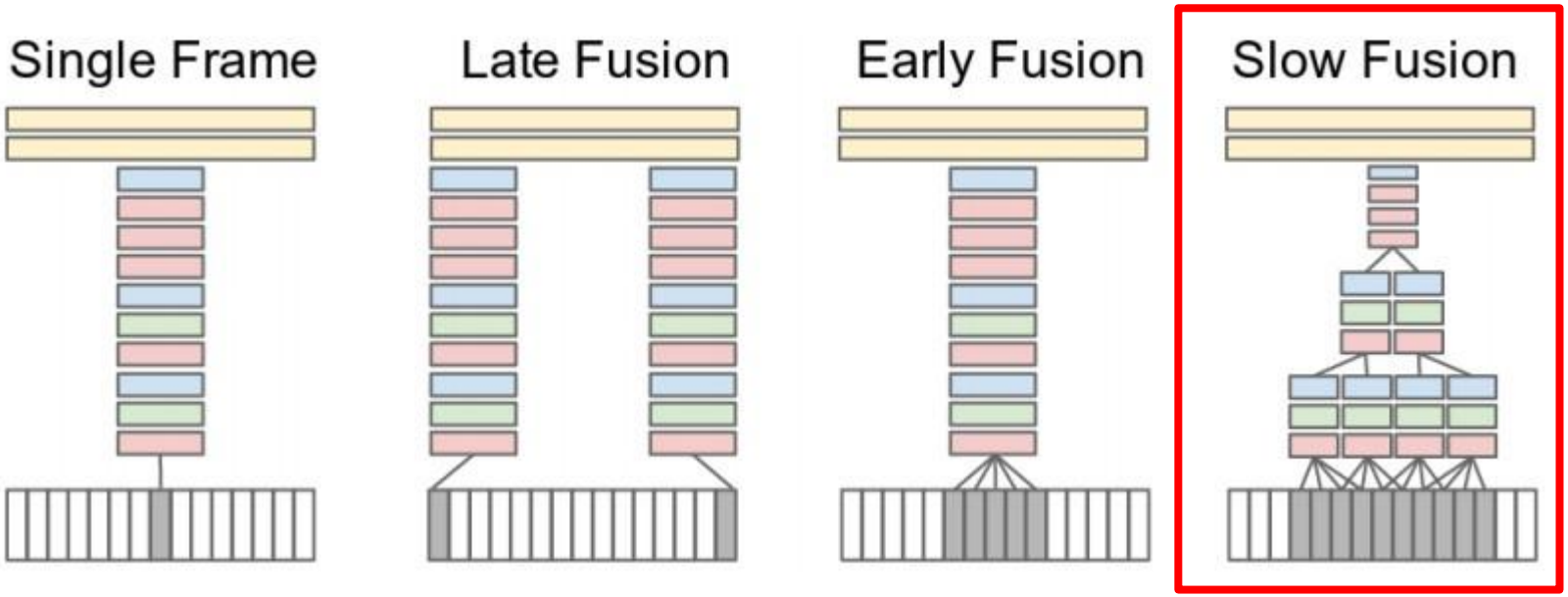
7 Million	450,000	3.2 Billion	4716	3.4
Video URLs	Hours of Video	Audio/Visual Features	Classes	Avg. Labels / Video

Deep models boost action recognition performance



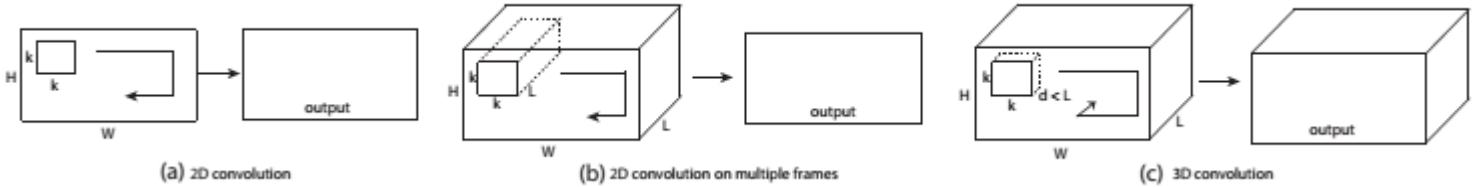
Spatio-Temporal ConvNets （CVPR14）

spatio-temporal convolutions;
worked best.



[Large-scale Video Classification with Convolutional Neural Networks, Karpathy et al., CVPR, 2014]

C3D （CVPR 15）



2D Convolution → 3D Convolution

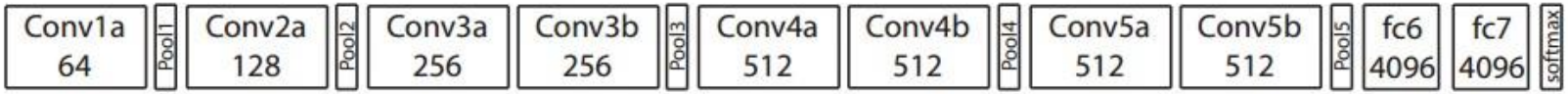


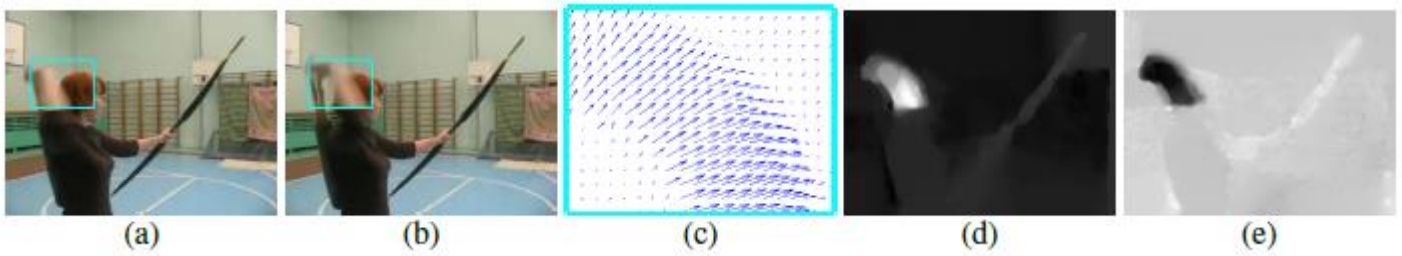
Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

C3D: 3D VGGNet

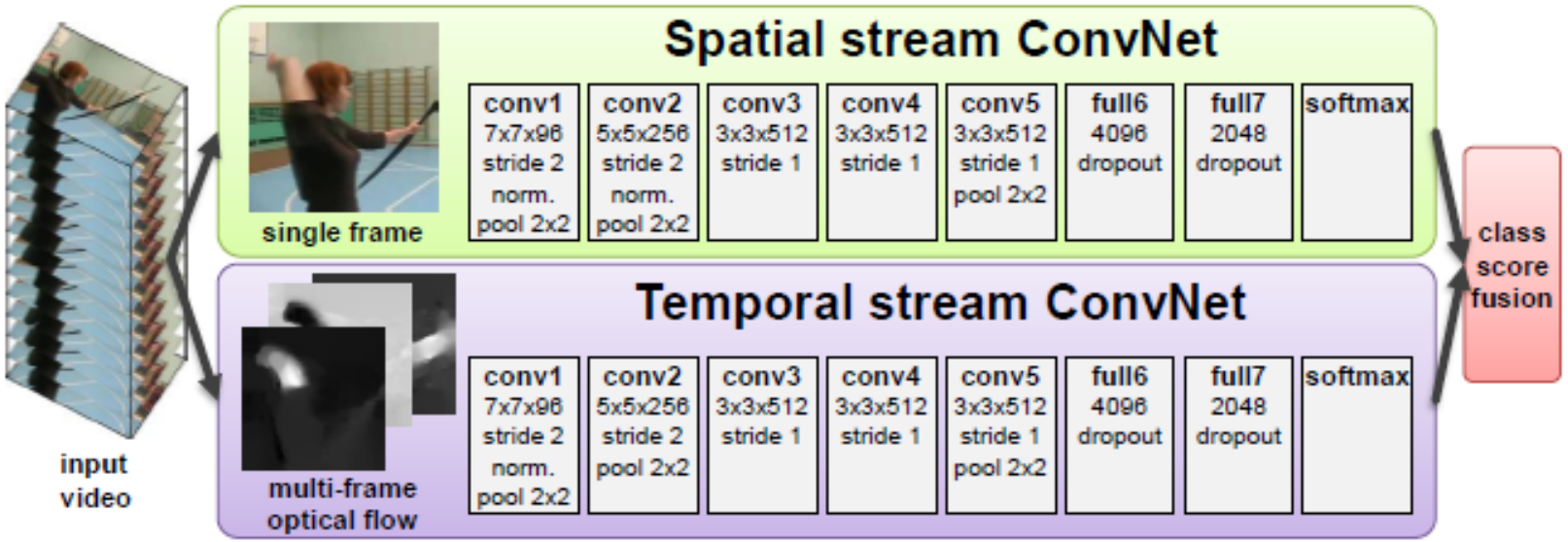
[Learning Spatiotemporal Features with 3D Convolutional Networks, Tran et al. 2015]

Two stream CNN (NIPS 2014)

Treat optical flow as images

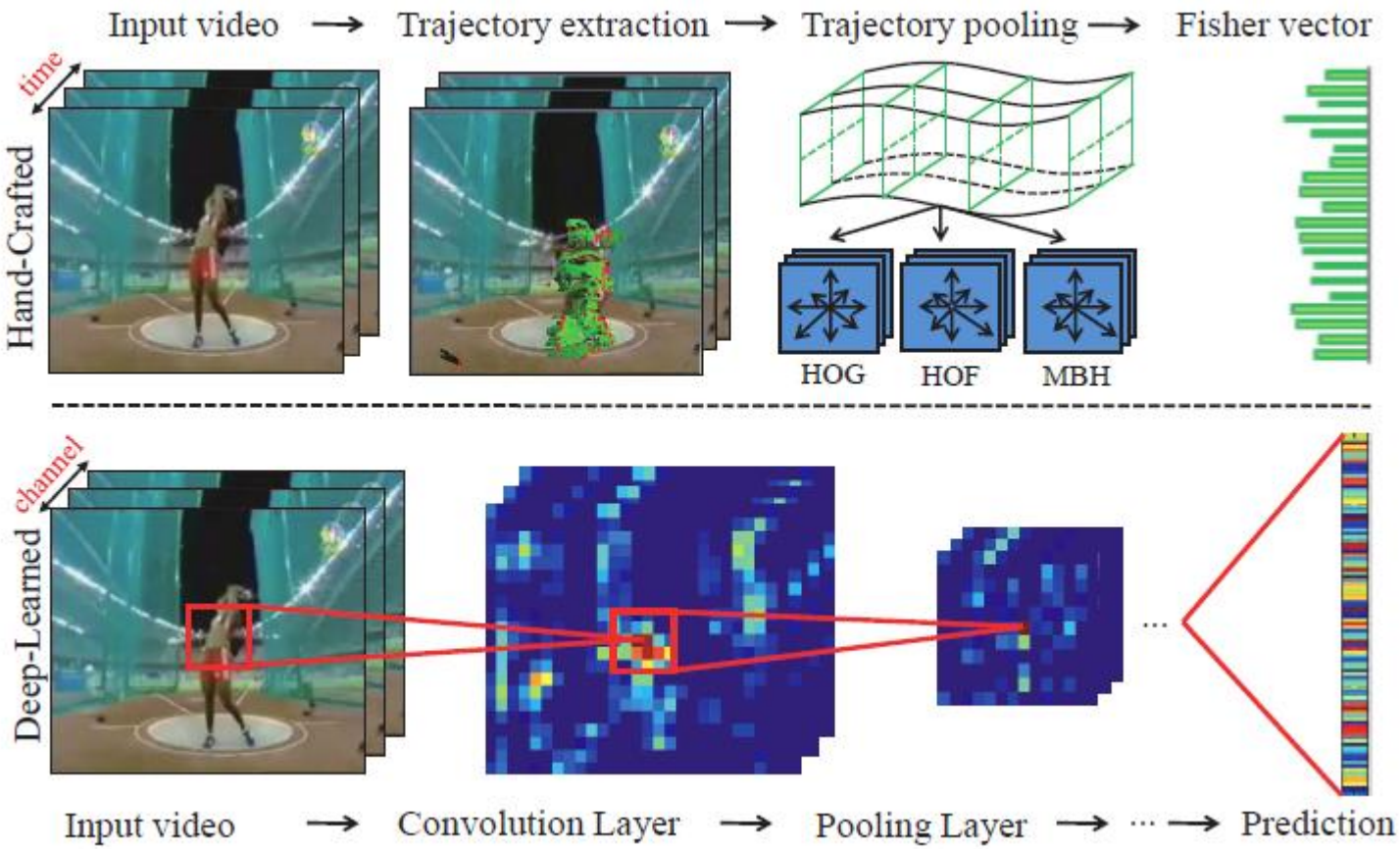


Train spatial CNN from images and temporal CNN from optical flows



Karen Simonyan Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos“, NIPS, 2014

Trajectory-Pooled Deep-Convolutional Descriptors (CVPR15)



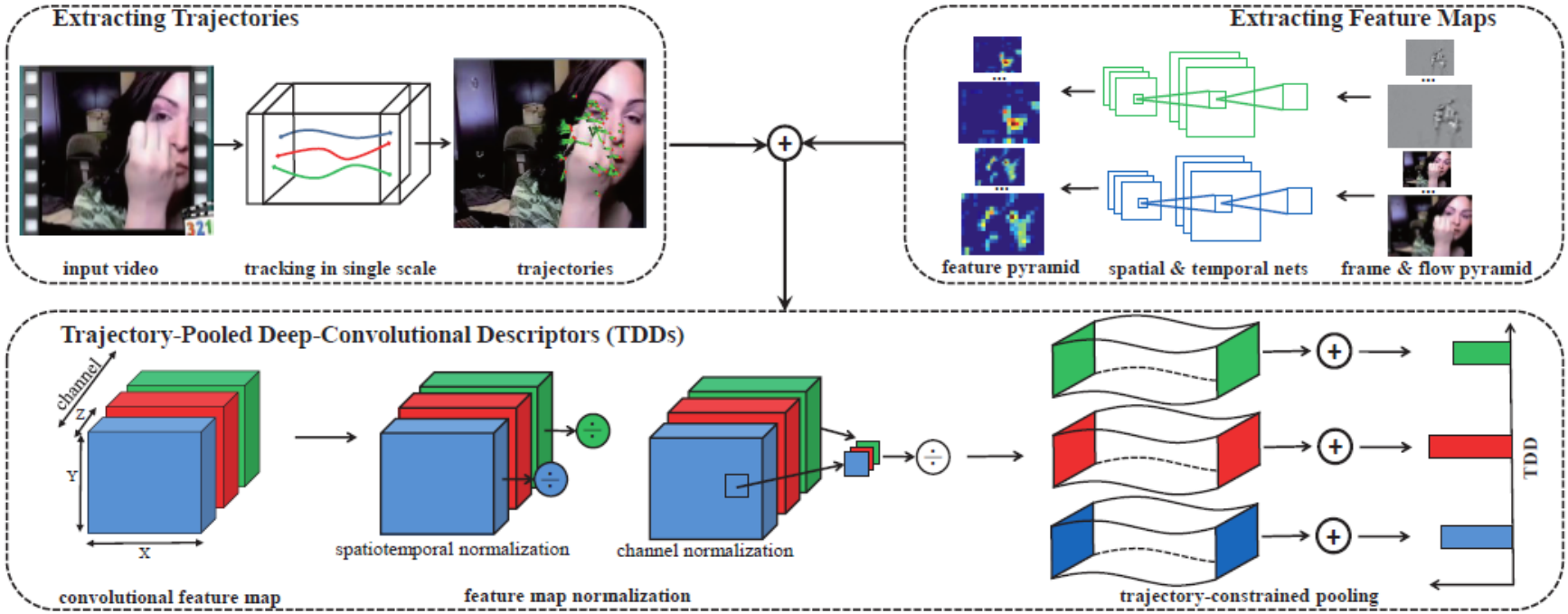
How to combine the merits of two approaches

Limin Wang, Yu Qiao, Xiaoou Tang "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors ", Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2015

Framework of the proposed methods

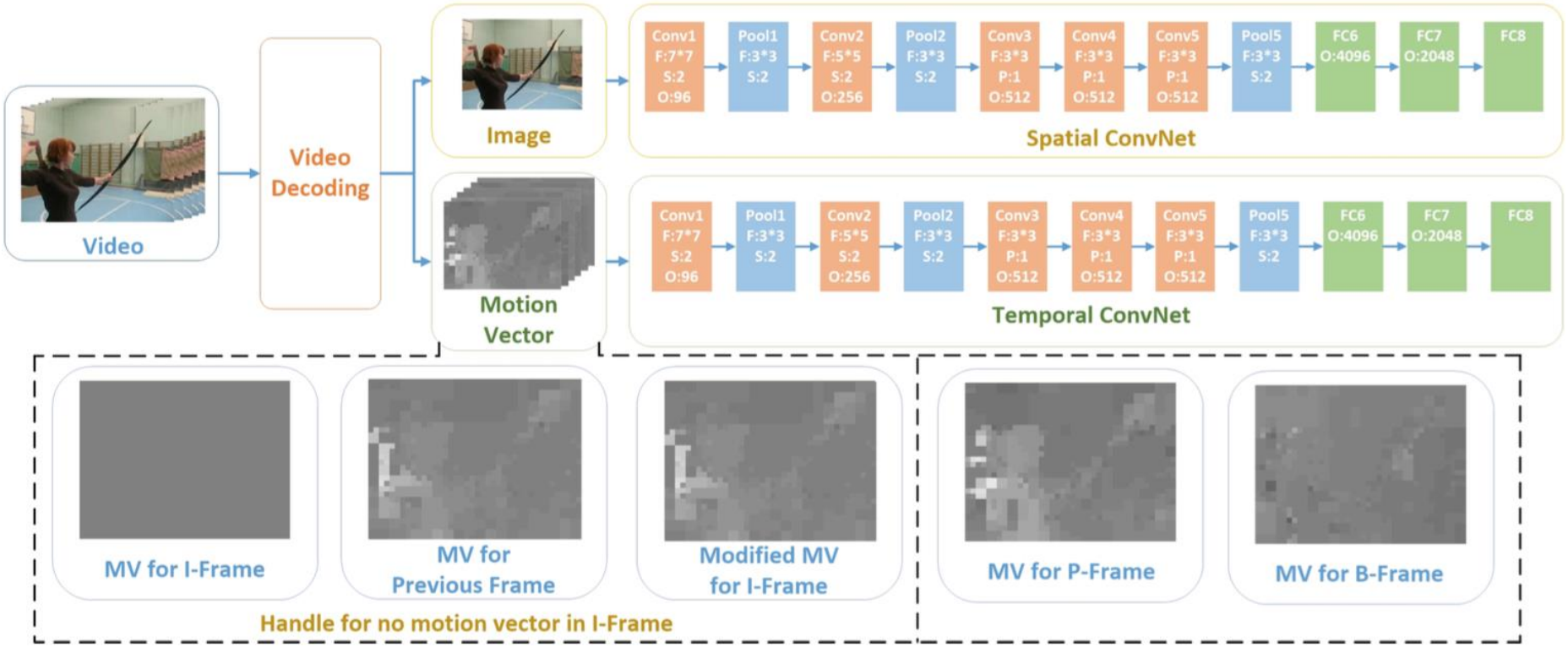
Propose *trajectory-pooled deep convolutional descriptor* (TDD) to integrate the key factors from handcrafted and deep approaches.

- Utlize two-stream ConvNets to obtain multi-scale deep convolutional features.
- Pool the local ConvNet responses over the spatiotemporal tubes centered at the trajectories.



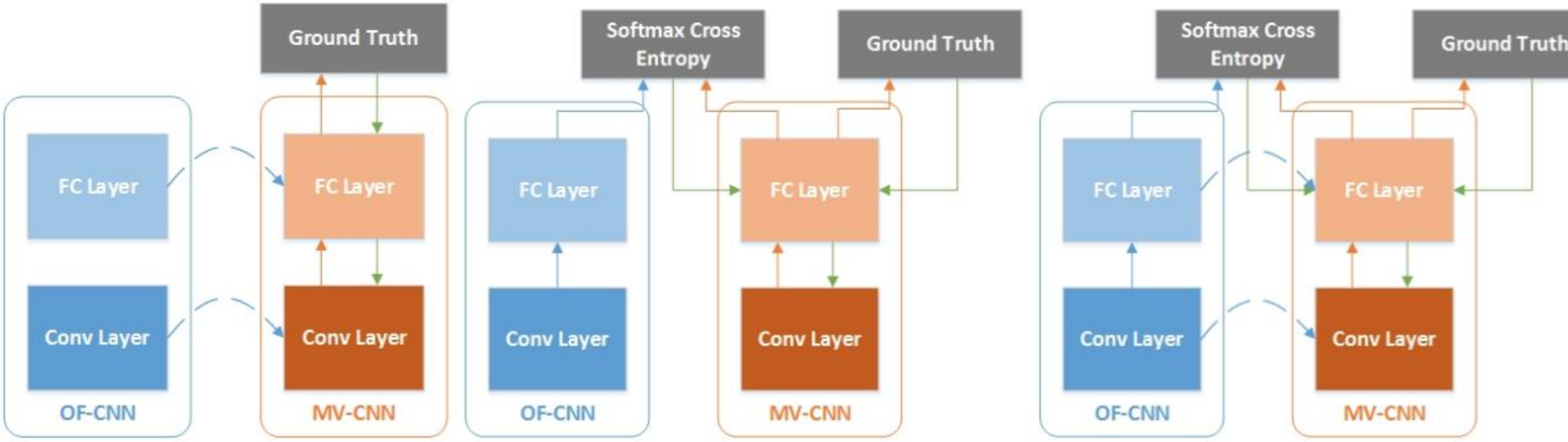
Motion Vector CNN (CVPR15)

- Many deep learning approaches for video based action recognition are computationally expensive, due to the calculation of optical flows
- Motion vector also includes motion information of local regions



B. Zhang, L. Wang, Z. Wang, Yu Qiao, and H. Wang " Real-time Action Recognition with Enhanced Motion Vector CNNs ", Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2016

Enhanced Motion Vector CNN



(a) Strategy 1: Teacher Initialization

Fine tune Optical Flow CNN

(b) Strategy 2: Supervision Transfer

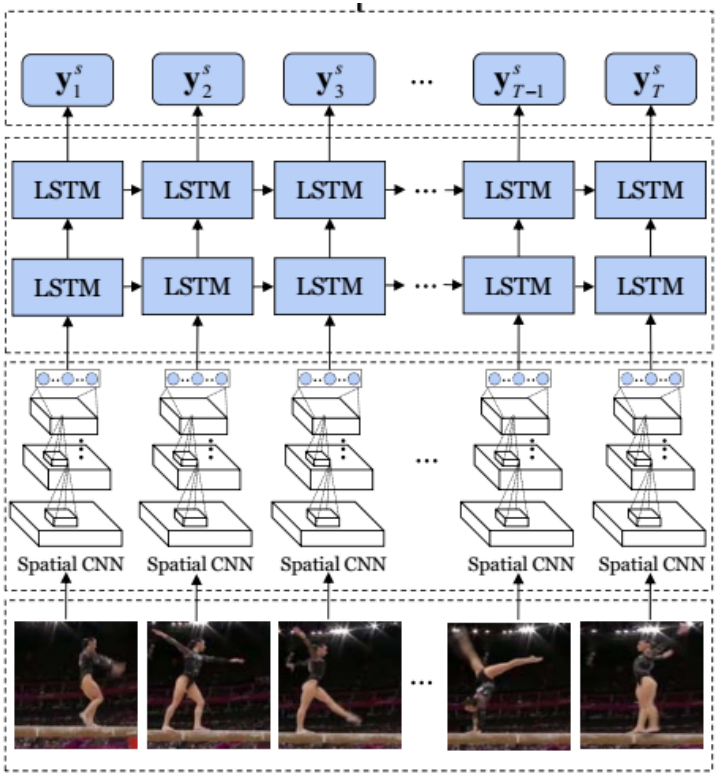
OF-CNN output as supervision

(c) Strategy 3: Combination

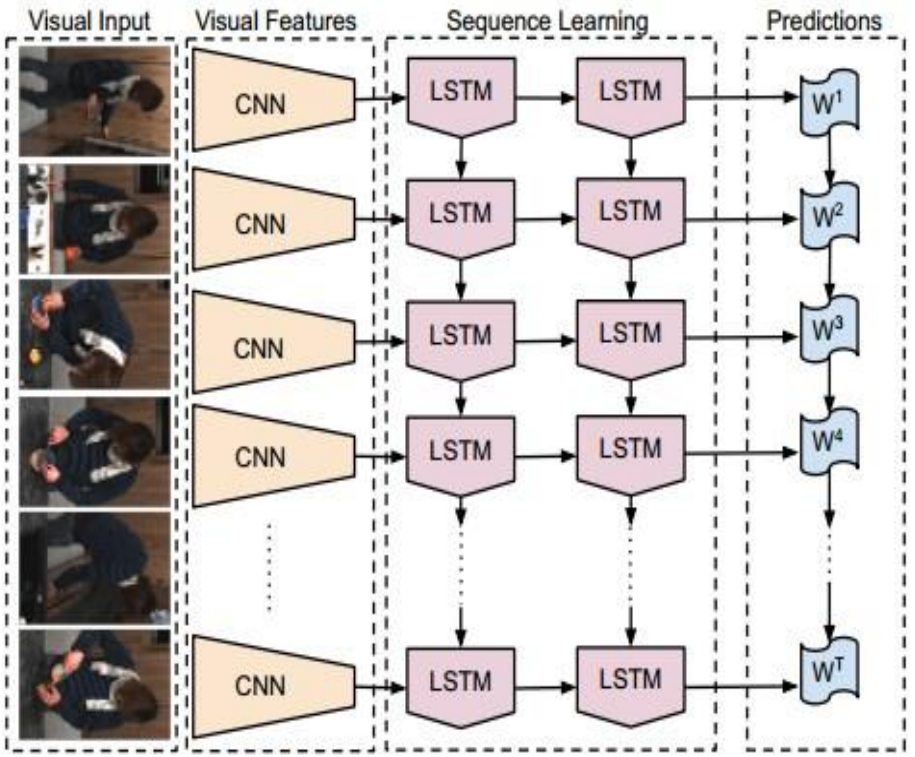
Strategy 1+2

Samples of filters for Conv1 layer. Left to right: MV- CNN, EMV-CNN and OF-CNN.

Recurrent Neural Network/LSTM for Action Recognition

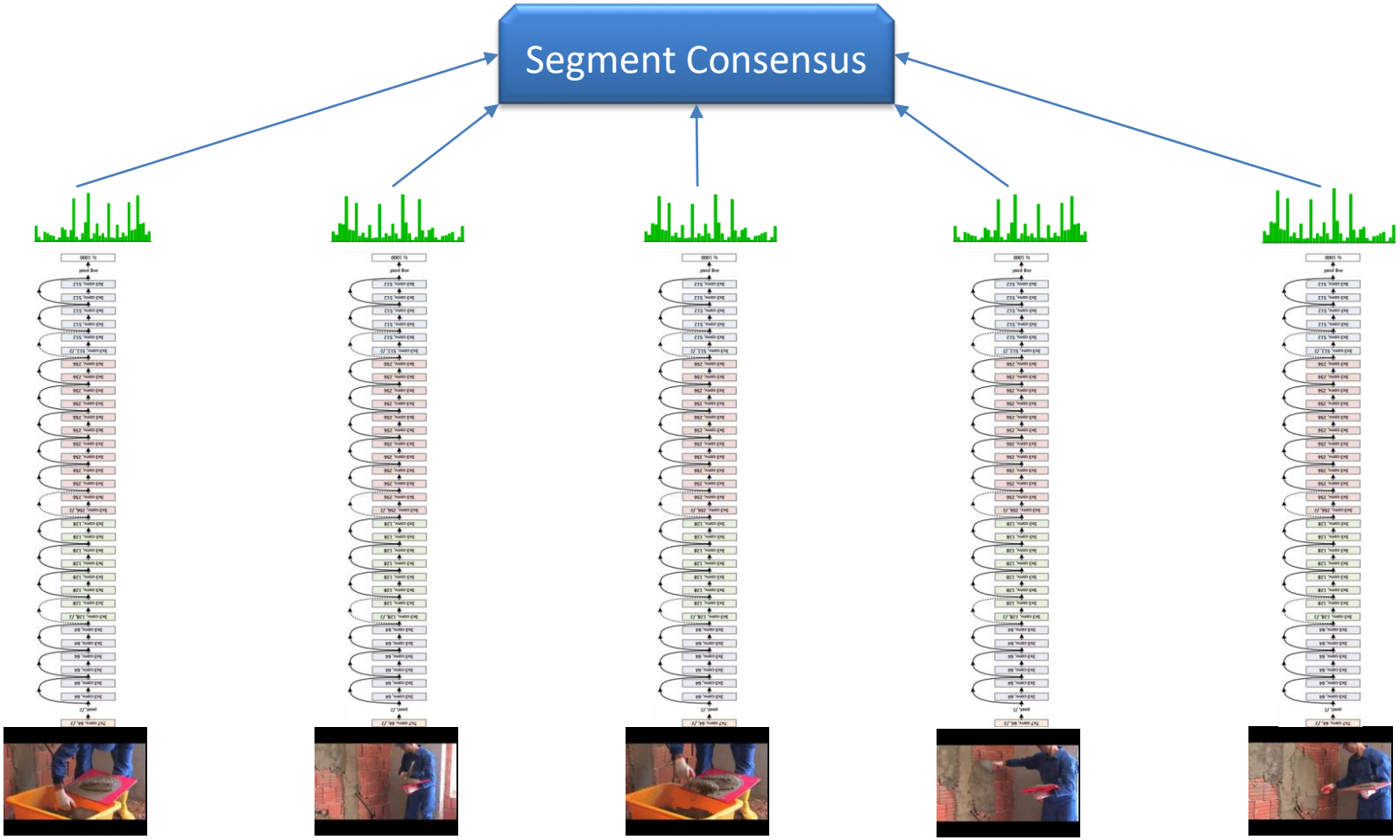


Wu Z, Wang X, Jiang Y, et al. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification[C]. acm multimedia, 2015: 461-470.



[Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al., 2015]

Deep Segmental Network (ECCV 16)



1.Limin Wang, Yuanjun Xiong, Zhe Wang, *Yu Qiao*, Dahua Lin, Xiaoou Tang,Luc Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," Proc. European Conference Computer Vision, 2016

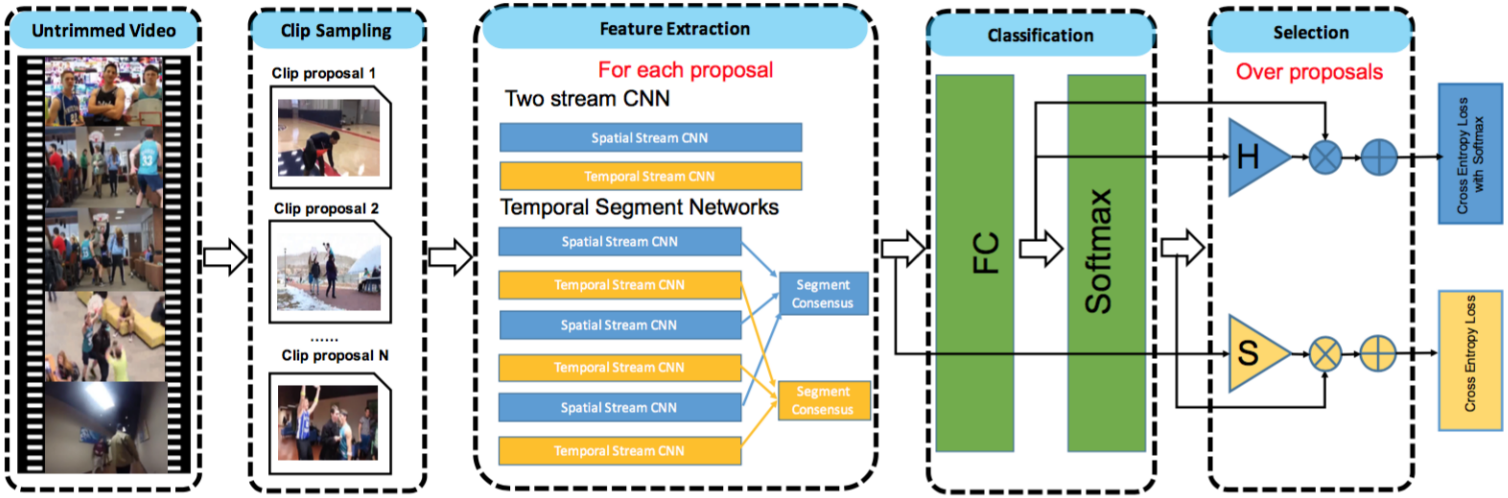
Performance of TSN

HMDB51		UCF101		THUMOS14		ActivityNet	
iDT+FV [2]	57.2%	iDT+FV [70]	85.9%	iDT+FV [70]	63.1%	iDT+FV [70]	66.5%
DT+MVSV [46]	55.9%	DT+MVSV [46]	83.5%	object+motion [71]	71.6%	Depth2Action [72]	78.1%
iDT+HSV [73]	61.1%	iDT+HSV [73]	87.9%				
MoFAP [51]	61.7%	MoFAP [51]	88.3%				
Two Stream [1]	59.4%	Two Stream [1]	88.0%	Two Stream [1]	66.1%	Two Stream [1]	71.9%
VideoDarwin [23]	63.7%	C3D (3 nets) [17]	85.2%	EMV+RGB [18]	61.5%	C3D [17]	74.1%
MPR [74]	65.5%	Two stream +LSTM [4]	88.6%				
$F_{ST}CN$ [55]	59.1%	$F_{ST}CN$ [55]	88.1%				
TDD+FV [5]	63.2%	TDD+FV [5]	90.3%				
LTC [24]	64.8%	LTC [24]	91.7%				
KVMF [75]	63.3%	KVMF [75]	93.1%				
TSN (3 seg)	70.7%	TSN (3 seg)	94.2%	TSN (3 seg)	78.8%	TSN (3 seg)	89.0%
TSN (7 seg)	**71.0%**	TSN (7 seg)	**94.9%**	TSN (7 seg)	**80.1%**	TSN (7 seg)	**89.6%**

L. Wang et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, in ECCV 2016.

L. Wang et al. Temporal Segment Networks for Action Recognition in Videos, to appear.

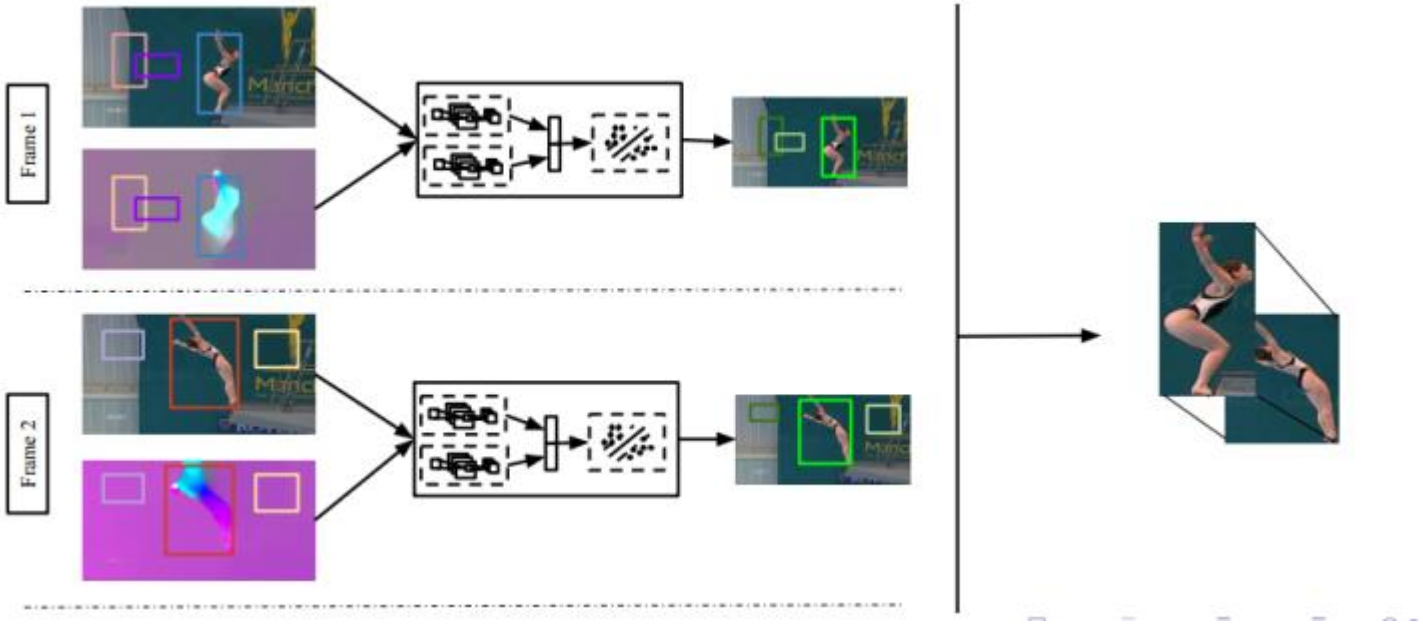
UntrimmedNet: Directly Learning from Untrimmed Video (CVPR17)



Leverage **attention modeling** in TSN for **weakly supervised action recognition and detection.**

L. Wang et al. UntrimmedNet for Weakly Supervised Action Recognition and Detection, in CVPR 2017.

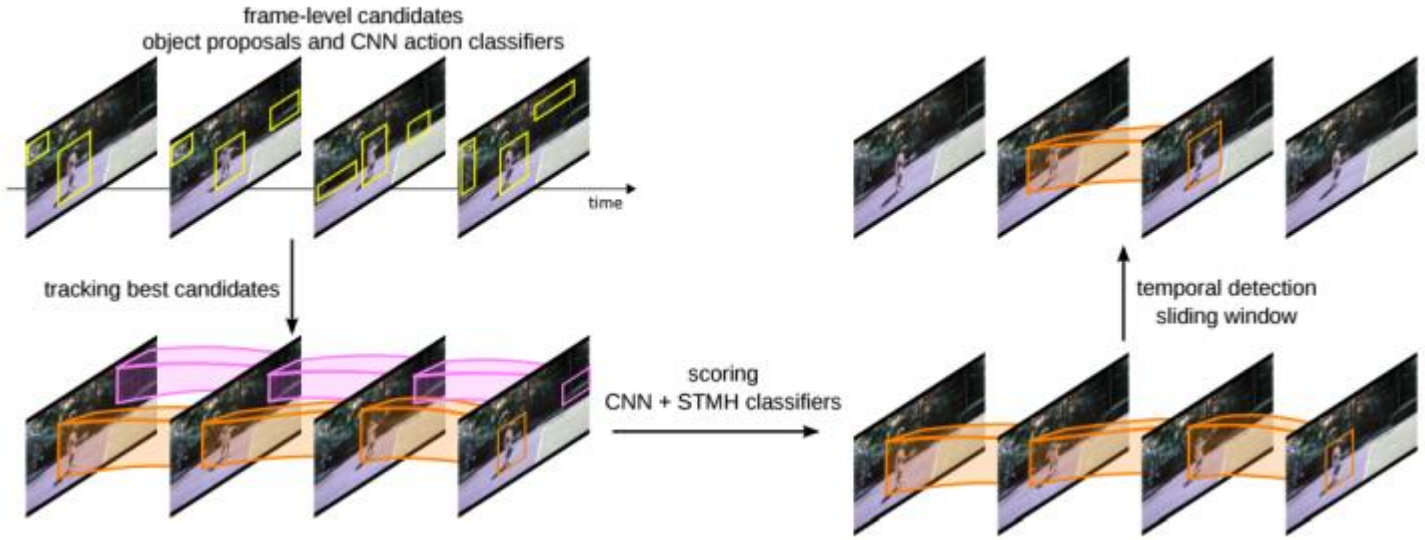
R-CNNs for Pose Estimation and Action Detection (CVPR'14)



R-CNN based frame-level detection + linking with dynamic programming

Gkioxari, G , Hariharan, B , Girshick, R, J. Malik, R-CNNs for Pose Estimation and Action Detection, IEEE Conference on Computer Vision & Pattern Recognition, 2014

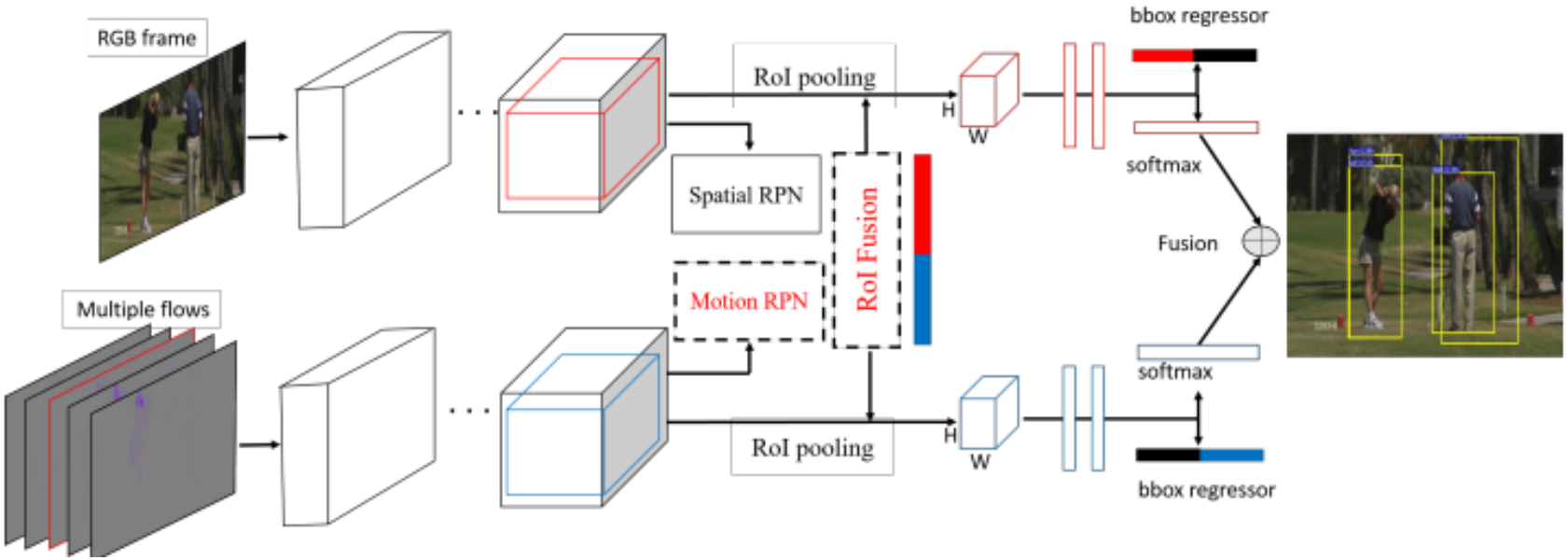
Learning to Track for Spatio-Temporal Action Localization (CVPR15)



R-CNN based frame-level detection (replacing SS by a better one{EdgeBoxes) + best candidates tracking

P Weinzaepfel，Z Harchaoui，C Schmid, "Learning to Track for Spatio-Temporal Action Localization", CVPR, 2015

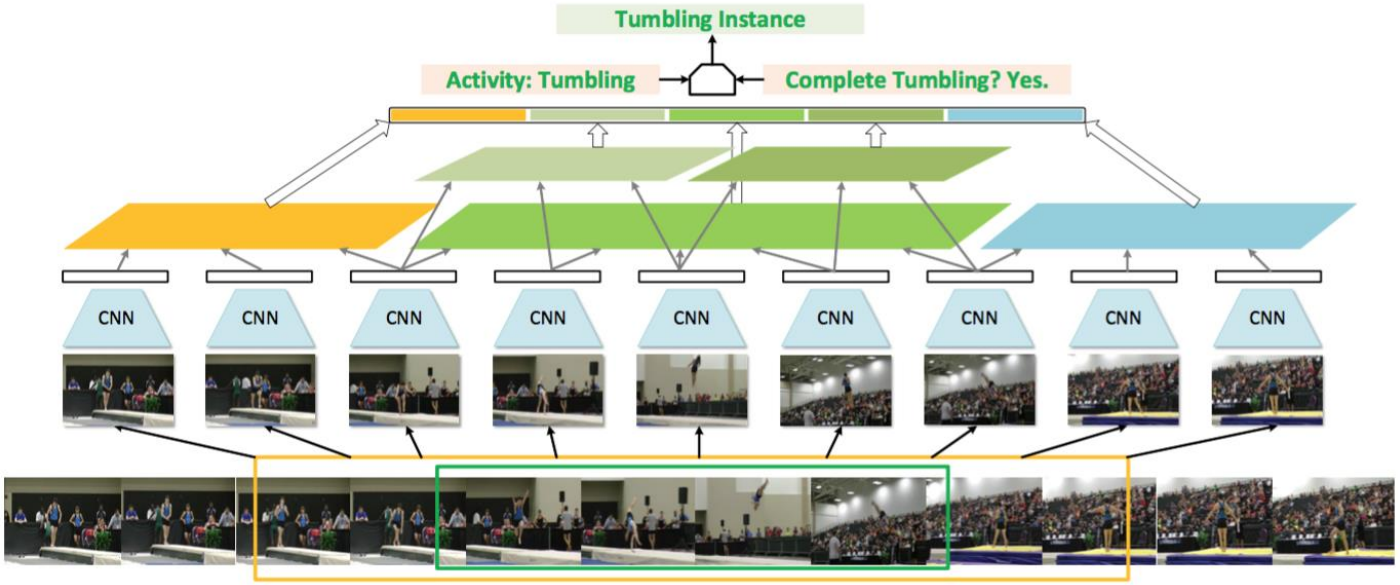
Multi-region two-stream R-CNN for action detection (ECCV 16)



RPN proposals with multiple RGB frames & optical flows + multi-region scheme for action detection

Xiaojiang Peng, Cordelia Schmid, Multi-region two-stream R-CNN for action detection European Conference on Computer Vision (ECCV), 2016.

Structured Segment Network for Action Detection (arxiv 17)



Incorporating **context and structure** modeling into TSN for action detection

Y. Zhao et al. Temporal Action Detection with Structured Segment Networks, in arXiv 1704.06228.

谢谢！

模型和代码公开

场景理解与分类

- MR-CNNs (2nd in scene classification task ImageNet 2016, 1st in LSUN 2016)
- Weakly Supervised PatchNets (Top performance in MIT Indoor67 and SUN397)

行为识别和检测

- Temporal Segment Networks (NO1 in ActivityNet 2016)
- MV-CNNS(Speed:300帧/s)
- Trajectory-Pooled Deep-Convolutional Descriptors （Top performance in UCF101 and HMDB51)

人脸检测与识别

- MJ-CNN face detection (top performance in FDDB & WIDE)
- HFA-CNN face recognition (single model 99% in LFW)

场景文字检测与识别

- Connectionist Text Proposal Network for Scene Text Detection (Top performance in ICDAR)

下载地址

http://mmlab.siat.ac.cn/yuqiao/Codes.html

Yu QIAO

2017.4

Thank you!

Q&A