Action Recognition Review & Future

胡鹤臻

2019.01.12

Outline

- Introduction
- Review on action recognition
  - Early work (hand-crafted features)
  - Deep architecture
- Spotlight & Future Work

Introduction

- ☐ What is an action?

Figure 1: Actions are "meaningful interactions" between humans and the environment.

- ☐ Why learn about action recognition?
  - ■ extends over a broad range of high-impact societal applications
    - ✓ video surveillance
    - ✓ human-computer interaction
    - ✓ retail analytics
    - ✓ user interface design
    - ✓ web-video search and retrieval
    - ✓ ……

Introduction

- Evalution
  - Classification accuracy, Inference time, GLOPS, storage
- Video benchmarks
  - Middle - scale

UCF101 (13,320 videos,101 actions)

HMDB51 (6,849 videos, 51 actions)

Introduction

- Video benchmarks
  - Large - scale

Benchmarks	Year	Team	Task
ActivityNet http://activity-net.org/index.html	2015	Universidad del Norte & KAUST	• Untrimmed Action Recognition • Temporal Action Proposals • Temporal Action Localization • Dense-Captioning Events in Videos
Youtube8M https://research.google.com/youtube8m/index.html	2016	Google	• Video Classification
Kinetics https://deepmind.com/research/open-source/open-source-datasets/kinetics/	2017	Google (DeepMind)	• Trimmed Activity Recognition
AVA https://research.google.com/ava/index.html	2017	Google	• Spatio-temporal Action Localization
Moments in Time http://moments.csail.mit.edu/	2018	MIT	• Trimmed Event Recognition

Outline

- Introduction
- Review on action recognition
  - Early work (hand-crafted features)
  - Deep architecture
- Spotlight & Future Work

Early works for action representation

- Earliest works make use of 3D models to describe actions.

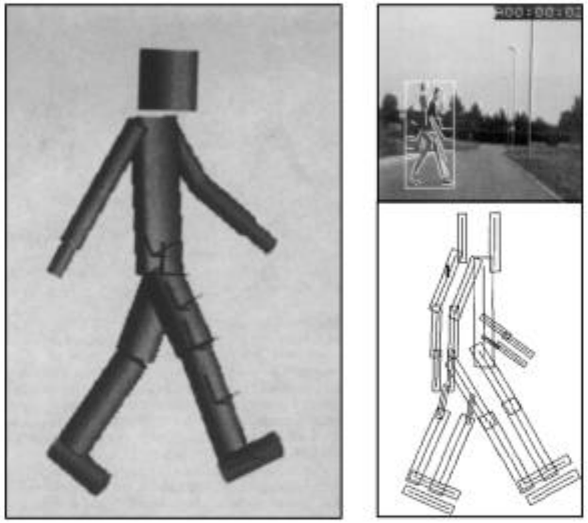


Figure 3: Early approaches represent actions by 3D models. **Left:** Hogg (1983) introduce the *WALKER* framework to represent walking action using 3D models. The walking pattern is modeled by a sequence of 3D structures. **Right:** Rohr (1994) extended the WALKER framework for pedestrian recognition. The model uses connected cylinders and their evolution to identify pedestrians.

Hogg (1983): David Hogg. Model-based vision: a program to see a walking person. Image and Vision Computing, 1:5–20, 1983.
Rohr (1994): K. Rohr. Towards model-based recognition of human movements in image sequences. CVGIP: Image Underst., 1994

Early works for action representation

- Holistic representations
  - A global representation of human body structure, shape and movements.

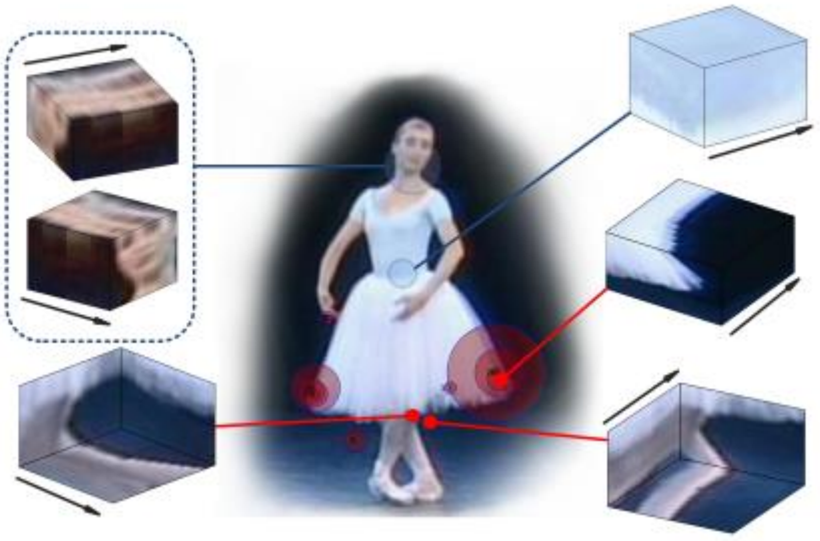
Figure 4: **Top:** A jumping sequence. **Middle:** The MEI template Bobick and Davis (2001). **Bottom:** The MHI template Bobick and Davis (2001). The MEI captures where the motion happens while the MHI template shows how the motion image is moving. The templates at the end of the action, shown in the rightmost column are used for representations.

Figure 5: **Left:** The spatiotemporal volumes used by Blank et al. (2005) to describe the evolution of an action. The 3D representation is converted to a 2D map by computing the average time taken by a point to reach the boundary. **Right:** The spatiotemporal surfaces of Yilmaz and Shah (2005) for a tennis serve and a walking sequence. The surface geometry (*e.g.*, peaks, valleys) is used to characterize the action.

Bobick and Davis (2001): A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. TPAMI, 2001
Blank et al. (2005): M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. ICCV, 2005
Yilmaz and Shah (2005): Alper Yilmaz and Mubarak Shah. Actions sketch: a novel action representation. CVPR, 2005

Early works for action representation

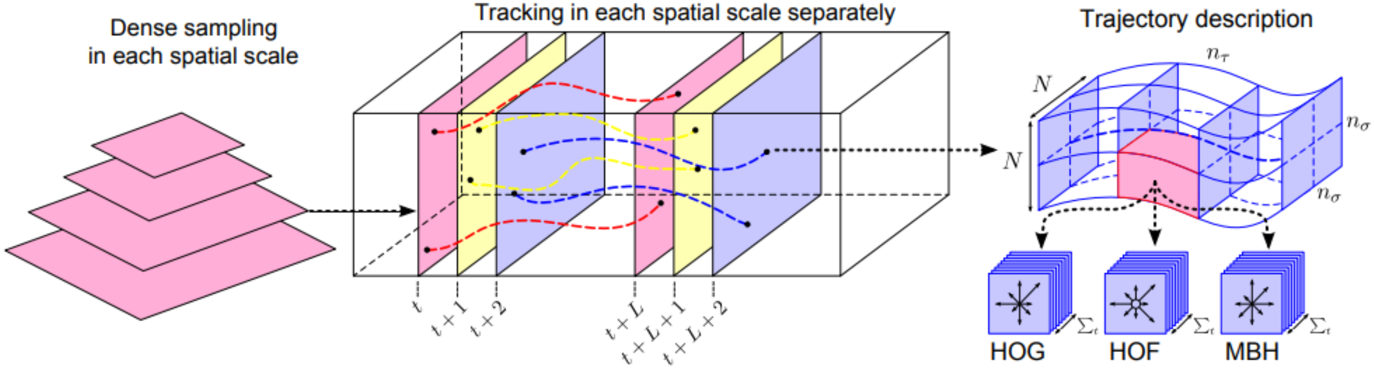
- Local representations
  - Interest point detection
    - ✓ 3D-Harris detector
    - ✓ 3D-Hessian detector
  - Local descriptor extraction
    - ✓ Edge and motion descriptors
    - ✓ Binary pattern descriptors
  - Aggregation of local descriptors
    - ✓ Bag-of-Visual Words (BoV)
    - ✓ Fisher Vector (FV)



Marked in red are the detected spatiotemporal interest points

Early works for action representation

- Dense Trajectories (DT)[1]

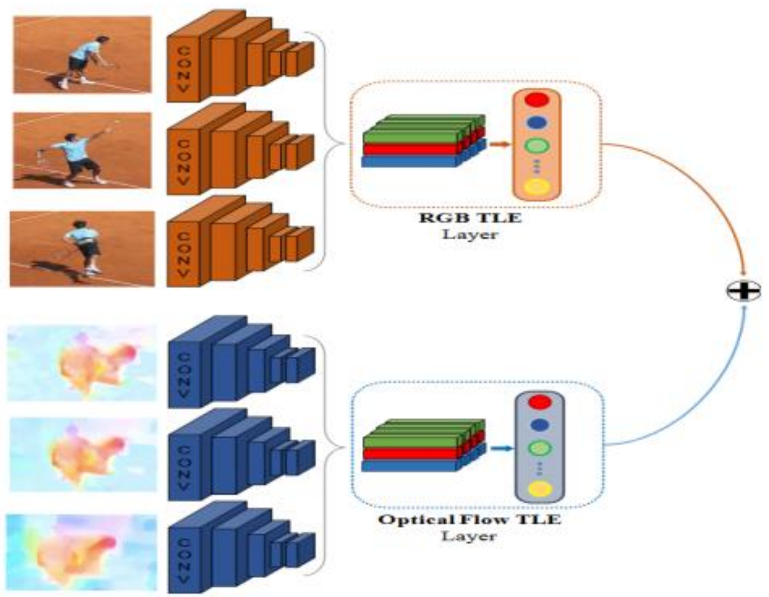


- Improved Dense Trajectories (IDT)[2]
  - Explicit camera motion estimation
  - Assumption: two consecutive frames are related by a homography.
  - Match feature points between frames using SURF descriptors and dense optical flow
  - Removing inconsistent matches due to humans: use a human detector to remove matches from human regions (computation expensive)
  - Estimate a homography with RANSAC with these matches

[1] Wang H, Kläser A, Schmid C, et al. Action recognition by dense trajectories[C]//CVPR 2011
[2] Wang H, Schmid C. Action recognition with improved trajectories[C]//ICCV 2013

Deep architecture for action recognition

- 2D CNN
  - Deep Temporal Linear Encoding (TLE) Networks
    - ✓ Aggregating K segments into a video representation
    - ✓ Bilinear encoding for feature interactions

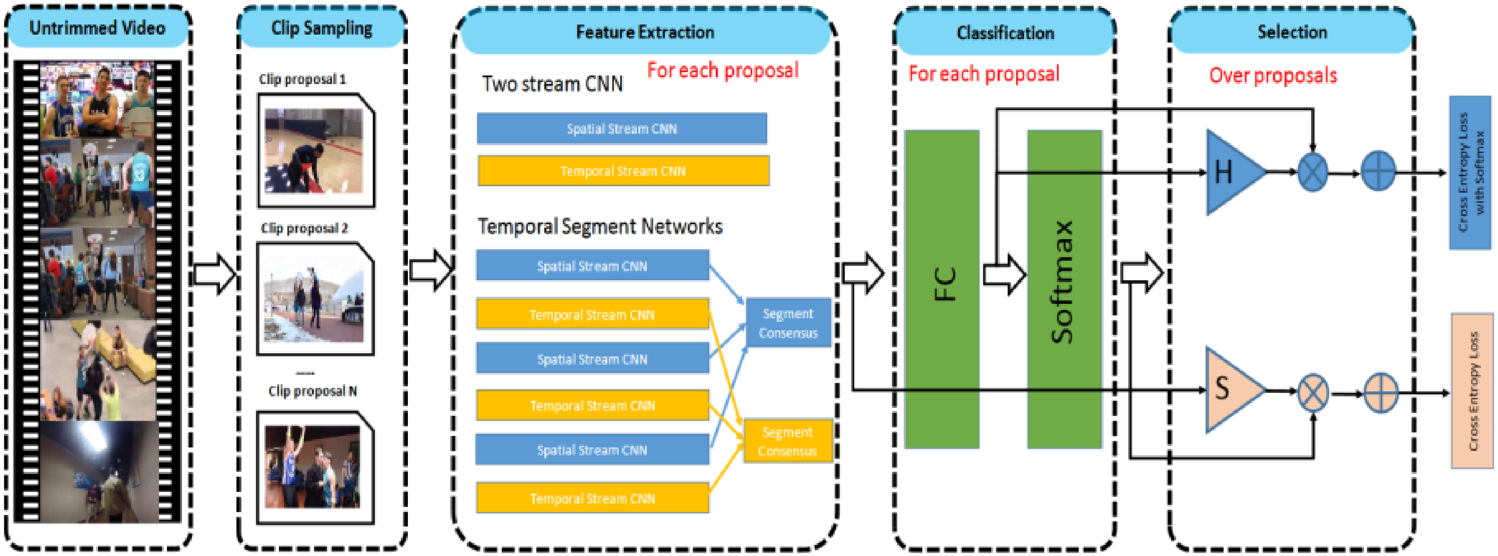


Method	UCF101	HMDB51
DT+MVSM [2]	83.5	55.9
iDT+FV [35]	85.9	57.2
Two Stream [25]	88.0	59.4
VideoDarwin [9]	–	63.7
C3D [33]	82.3	56.8
Two Stream+LSTM [41]	88.6	–
F_{ST}CV (SCI fusion) [30]	88.1	59.1
TDD+FV [37]	90.3	63.2
LTC [34]	91.7	64.8
KVMF [44]	93.1	63.3
TSN [38]	94.0	68.5
3DConv+3DPool [8]	93.5	69.2
TLE: FC-Pooling (ours)	92.2	68.8
TLE: Bilinear+TS (ours)	95.1	70.6
TLE: Bilinear (ours)	**95.6**	**71.1**

Ali Diba et al., Deep Temporal Linear Encoding Networks, CVPR 2017

Deep architecture for action recognition

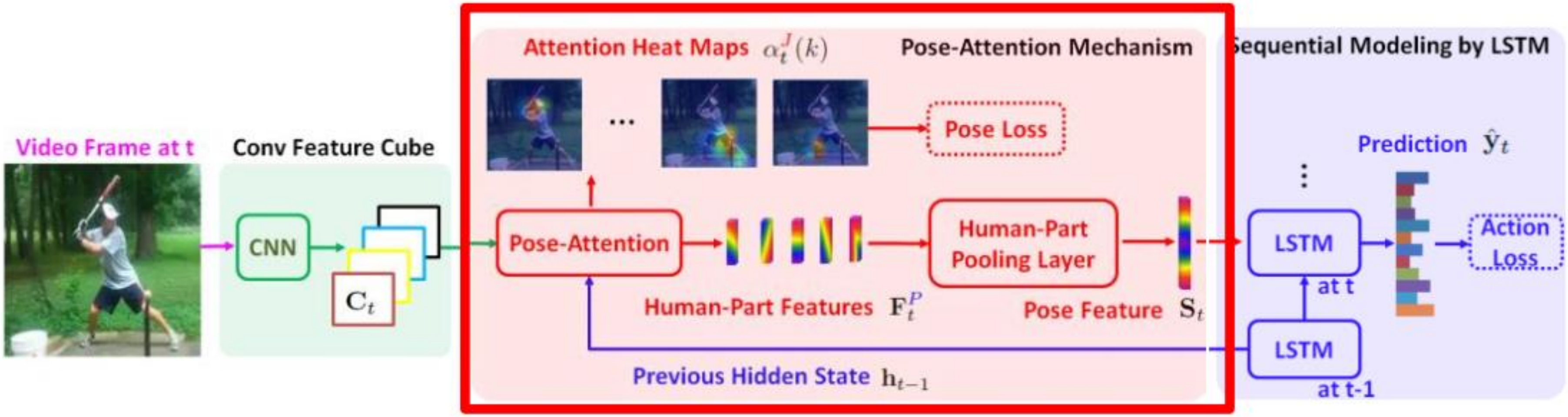
- 2D CNN
  - UntrimmedNet
    - ✓ Attention for proposal selection
    - ✓ Weakly-supervised detection



Limin Wang et al., UntrimmedNets for Weakly Supervised Action Recognition and Detection, CVPR 2017

Deep architecture for action recognition

- RNN
  - Recurrent Pose Attention Network
    - ✓ Pose attention as dynamical guidance for LSTM
    - ✓ Byproduct: pose estimation in videos

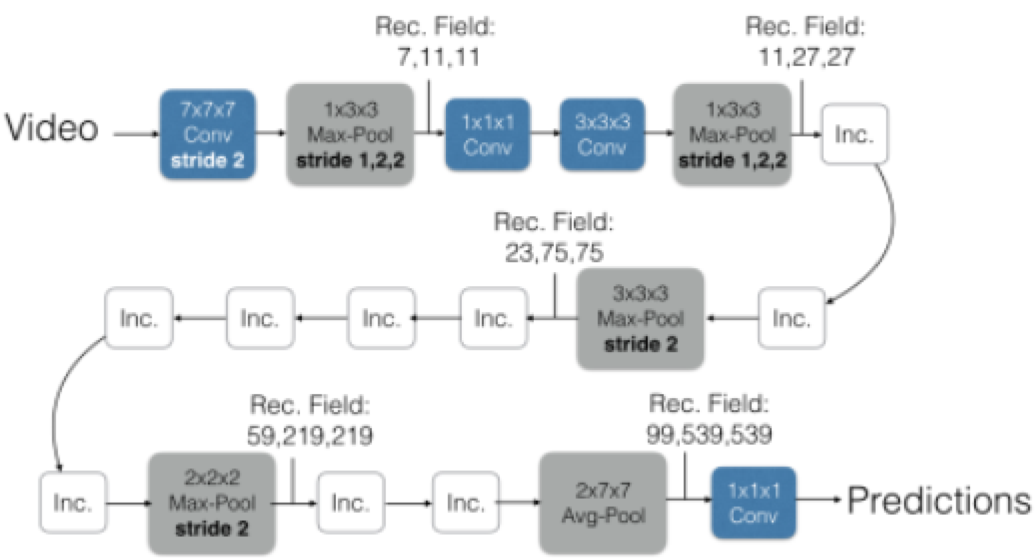


Wenbin Du et al., RPAN: An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos, ICCV2017

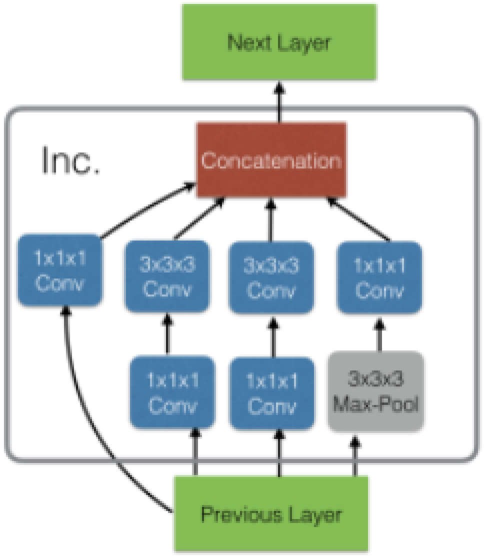
Deep architecture for action recognition

- 3D CNN
  - Inflated 3D(I3D) ConvNets
    - ✓ Inflating 2D ConvNets into 3D
    - ✓ Bootstrapping 3D filters from 2D filters
    - ✓ Propose Kenetics dataset

Inflated Inception-V1



Inception Module (Inc.)

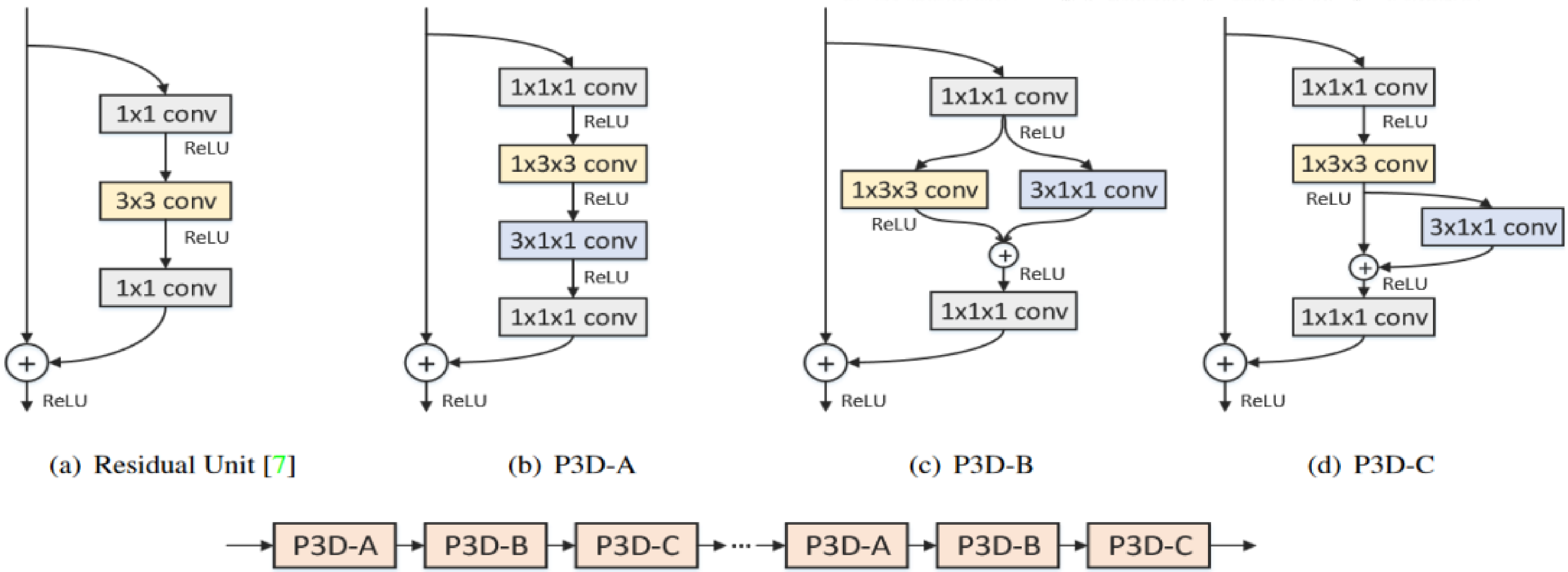


Joao Carreira et al., Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, CVPR,2017

Deep architecture for action recognition

- 3D CNN
  - Pseudo-3D Residual Networks
    - ✓ 3 types of P3D blocks
    - ✓ Interleaving design for ResNet

ActivityNet	Top-1	Top-3	MAP
IDT [34]	64.70%	77.98%	68.69%
C3D [31]	65.80%	81.16%	67.68%
VGG_19 [26]	66.59%	82.70%	70.22%
ResNet-152 [7]	71.43%	86.45%	76.56%
P3D ResNet	**75.12%**	**87.71%**	**78.86%**



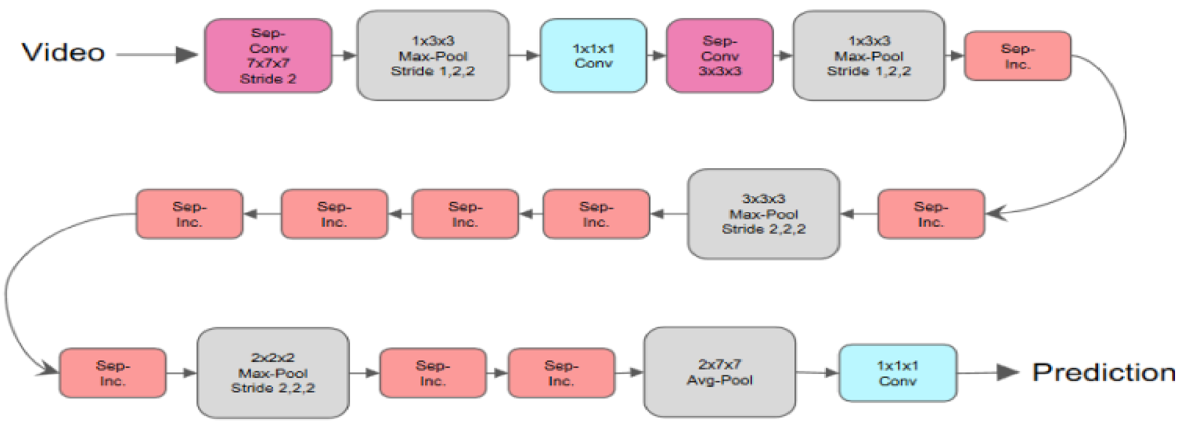
(a) Residual Unit [7] (b) P3D-A (c) P3D-B (d) P3D-C

Zhaofan Qiu et al., Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, ICCV2017

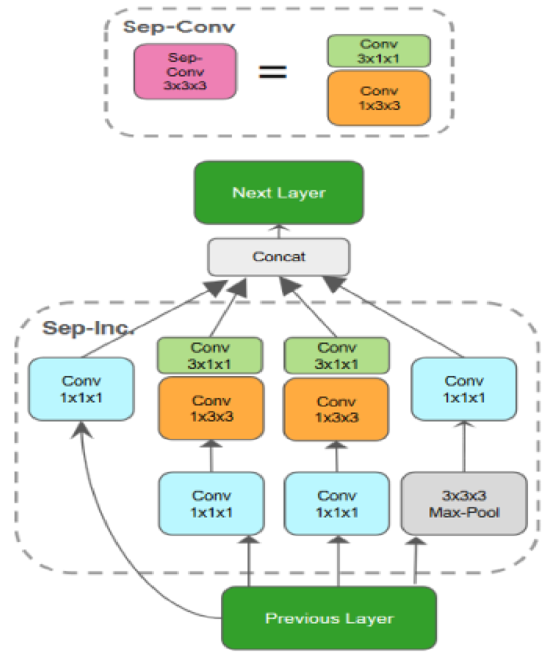
Deep architecture for action recognition

- 3D CNN
  - Spatiotemporal Separable 3D

3D Conv (Kt x K x K) → Spatial Conv (1 x K x K) + Temporal Conv (Kt x 1 x 1)



(a) S3D



(b) Separable Inception block

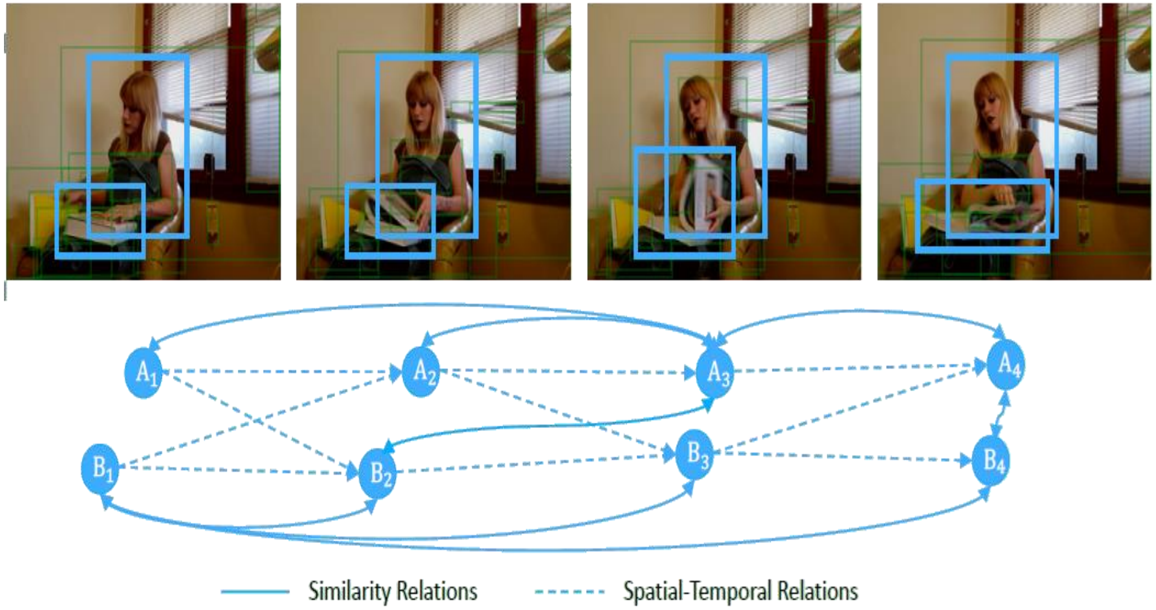
Kinetics	Inputs	Backbone	Top-1 (%)	Top-5 (%)
Shifting Attention Net [1]	RGB+Flow+Audio	Inception-ResNet-v2	77.7	93.2
Temporal Segment Net [53]	RGB+Flow	Inception	73.9	91.1
ARTNet w/ TSN [50]	RGB+Flow	ResNet-18	72.4	90.4
I3D [3]	RGB+Flow	Inception	74.1	91.6
S3D-G	RGB+Flow	Inception	**77.2**	**93.0**

Saining Xie et al., Rethinking Spatiotemporal Feature Learning For Video Understanding, CVPR, 2018

Outline

- Introduction
- Review on action recognition
  - Early work (hand-crafted features)
  - Deep architecture
- Spotlight & Future Work

Region Graphs



1. Video: space-time region graph
2. Nodes: region of interest (proposed by Faster R-CNN)
3. Edges: Similarity relations, Spatial-temporal relations
4. Reasoning: GCNs

Wang X, Gupta A. Videos as Space-Time Region Graphs. ECCV, 2018.

Region Graphs

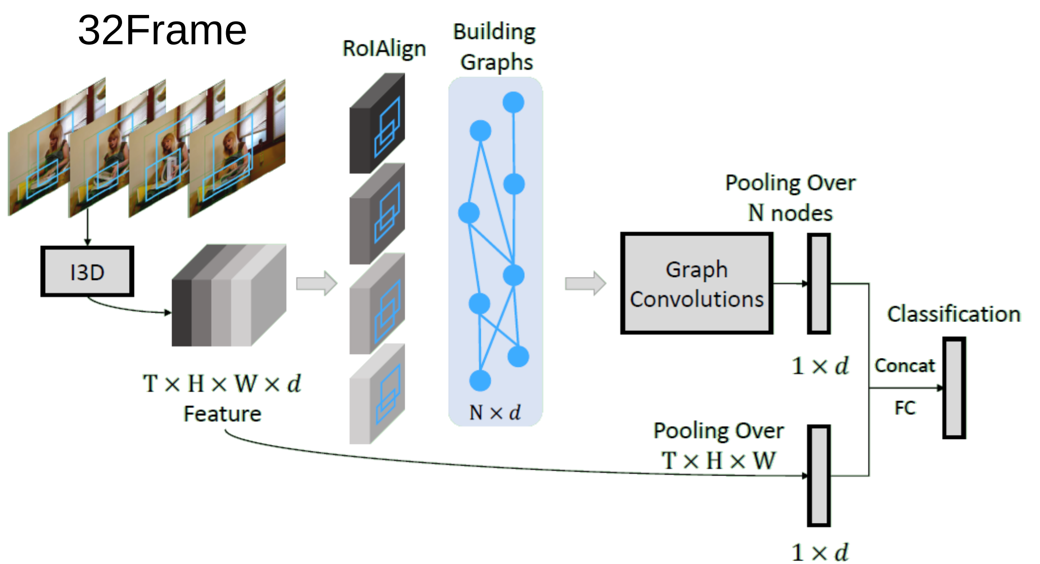


Figure 2. Model Overview. Our model uses 3D convolutions to extract visual features followed by RoIAlign extracting d-dimension feature for each object proposal. These features are provided as inputs to the Graph Convolutional Network which performs information propagation based on spatiotemporal edges. Finally, a d-dimension feature is extracted and appended to another d-dimension video feature to perform classification.

Overviews

1. Graphs: Objects by RPN→ RoIAlign → N*d dimensions
2. Relations: Similarity graph & spatial-temporal relations
3. ReasoNing: GCNs

Wang X, Gupta A. Videos as Space-Time Region Graphs. ECCV, 2018.

Region Graphs

- Similarity Graph
  - Target: correlations between different states of the same object instance across frame, but also the relations between different objects

$$F(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi'(\mathbf{x}_j), \;\; \phi(\mathbf{x}) = \mathbf{w}\mathbf{x} \text{ and } \phi'(\mathbf{x}) = \mathbf{w}'\mathbf{x}$$

$$\mathbf{G}_{ij}^{sim} = \frac{\exp F(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j=1}^{\mathrm{N}} \exp F(\mathbf{x}_i, \mathbf{x}_j)}$$

← Normalization (Softmax)

- Spatial-Temporal Graph
  - Target: encode these spatial and temporal relations between objects

$$\mathbf{G}_{ij}^{front} = \frac{\sigma_{ij}}{\sum_{j=1}^{\mathrm{N}} \sigma_{ij}}$$

Tips: also construct a backward graph (from frame t+1 to t)

→ **for richer structure info & enlarge the number of propagation neighbourhoods**

Wang X, Gupta A. Videos as Space-Time Region Graphs. ECCV, 2018.

Region Graphs

Figure 3. Similarity Graph $\mathbf{G}^{sim}$. Above figure shows our similarity graph not only captures similarity in visual space but also correlations (similarity in functional space). The query box is shown in orange, the nearest neighbors are shown in blue. The transparent green boxes are the other unselected object proposals.

Wang X, Gupta A. Videos as Space-Time Region Graphs. ECCV, 2018.

Region Graphs

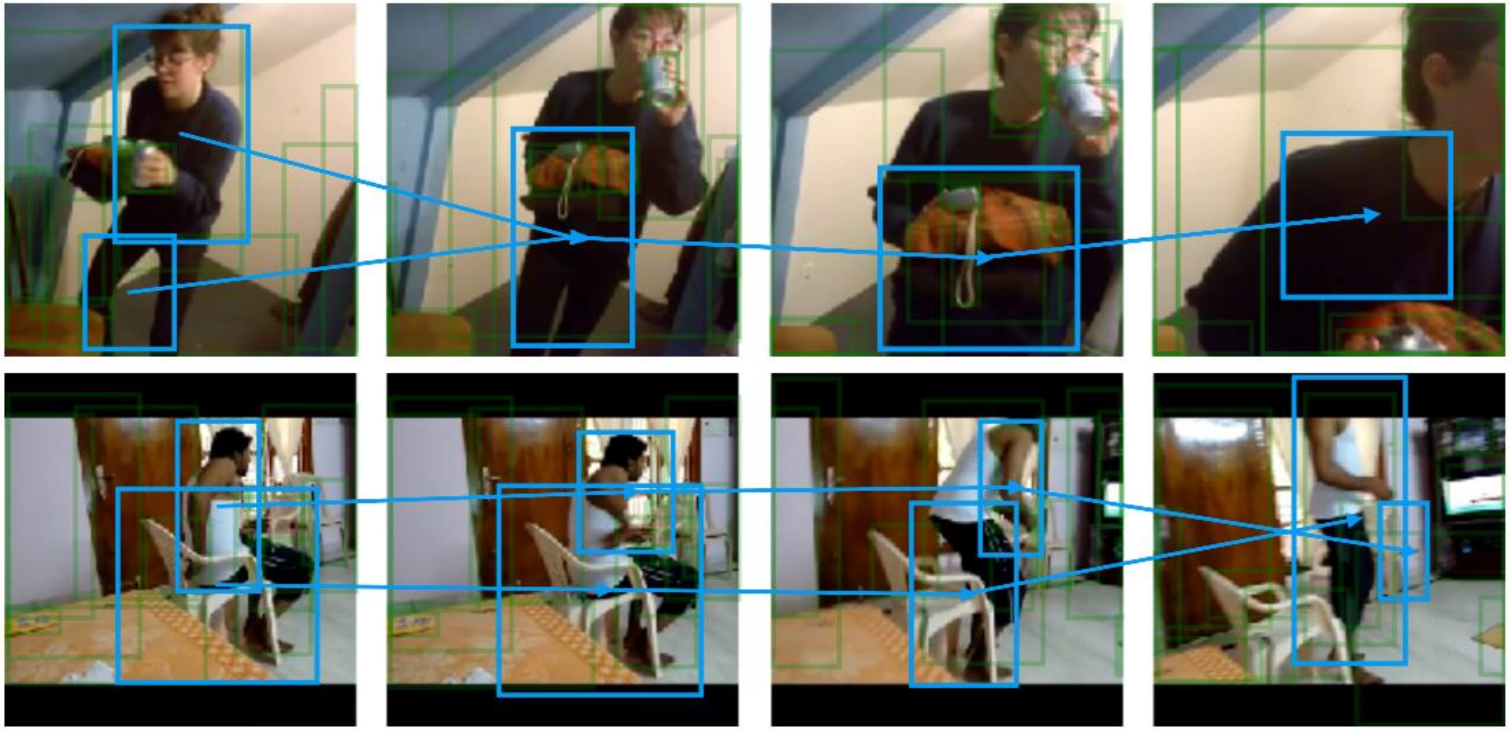


Figure 4. Spatial-Temporal Graph $\mathbf{G}^{front}$. Highly overlapping object proposals across neighboring frames are linked by directed edge. We plot some example trajectories with blue boxes and the direction shows the arrow of time.

Wang X, Gupta A. Videos as Space-Time Region Graphs. ECCV, 2018.

Region Graphs

- Graph Convolutional Networks
  - One layer of graph convolutions:

    $$\mathbf{Z} = \mathbf{G}\mathbf{X}\mathbf{W}$$ Gsim

  - Combine multiple graphs

    $$\mathbf{Z} = \sum_i \mathbf{G}_i \mathbf{X} \mathbf{W}_i$$ Gfront & Gback

  - Fuse the results from two GCNs in the end (summed together)

Wang X, Gupta A. Videos as Space-Time Region Graphs. ECCV, 2018.

Region Graphs

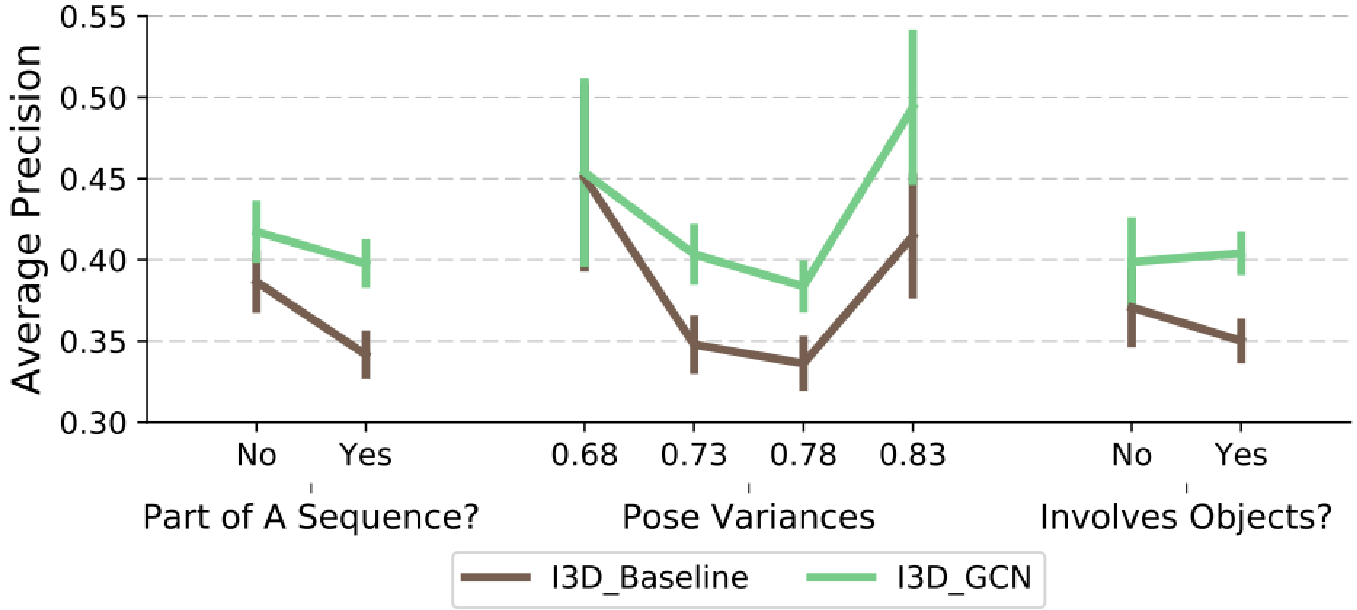


Figure 5. Error Analysis. We compare our approach against baseline I3D approach across three different attributes. Our approach improves significantly when action is part of sequence, involves interaction with objects and has high pose variance.

Wang X, Gupta A. Videos as Space-Time Region Graphs. ECCV, 2018.

Region Graphs

Charades dataset

model	backbone	modality	mAP
2-Stream [93]	VGG16	RGB + flow	18.6
2-Stream +LSTM [93]	VGG16	RGB + flow	17.8
Asyn-TF [93]	VGG16	RGB + flow	22.4
MultiScale TRN [36]	Inception	RGB	25.2
I3D [8]	Inception	RGB	32.9
I3D [58]	ResNet-101	RGB	35.5
NL I3D [58]	ResNet-101	RGB	37.5
NL I3D + GCN	ResNet-50	RGB	37.5
I3D + GCN	ResNet-101	RGB	39.1
NL I3D + GCN	ResNet-101	RGB	**39.7**

Something-Something dataset

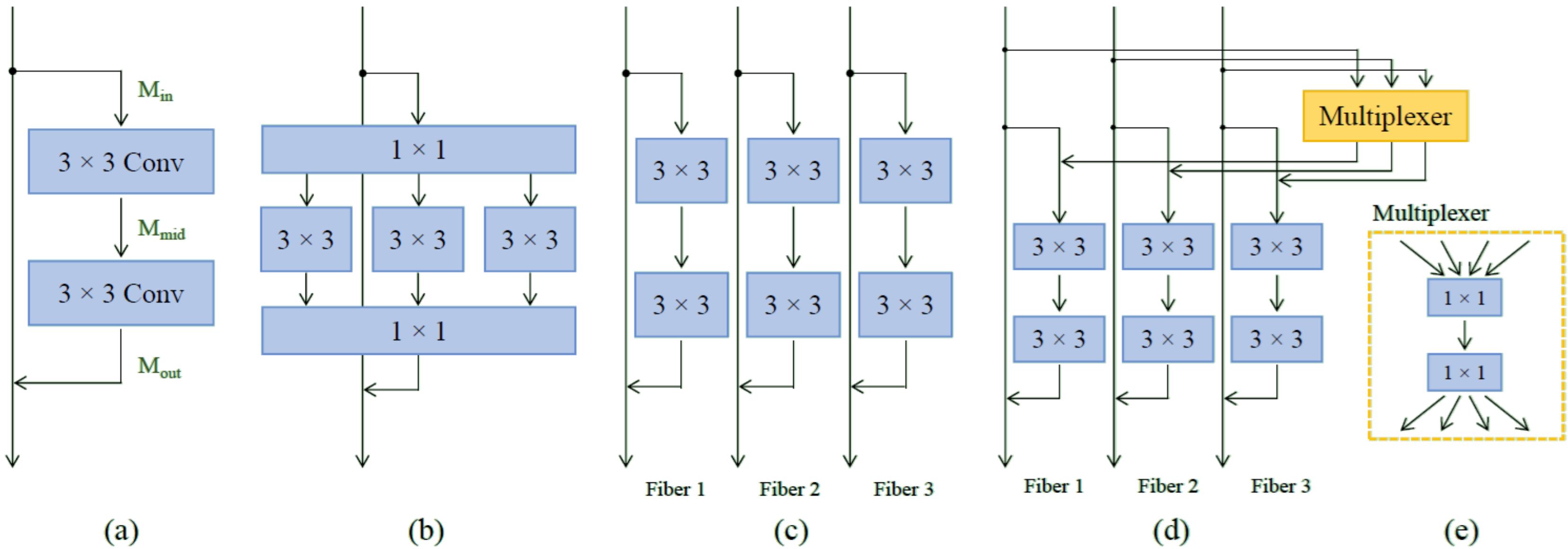
		val		*test*
model	backbone	top-1	top-5	top-1
C3D [21]	C3D[7]	-	-	27.2
MultiScale TRN [36]	Inception	34.4	63.2	33.6
I3D	ResNet-50	41.6	72.2	-
I3D + GCN	ResNet-50	43.3	75.1	-
NL I3D	ResNet-50	44.4	76.0	-
NL I3D + GCN	ResNet-50	46.1	76.8	45.0

Results

better in modeling
a long term sequence of actions &
actions that require object interactions.

Wang X, Gupta A. Videos as Space-Time Region Graphs. ECCV, 2018.

Multi-Fiber Networks

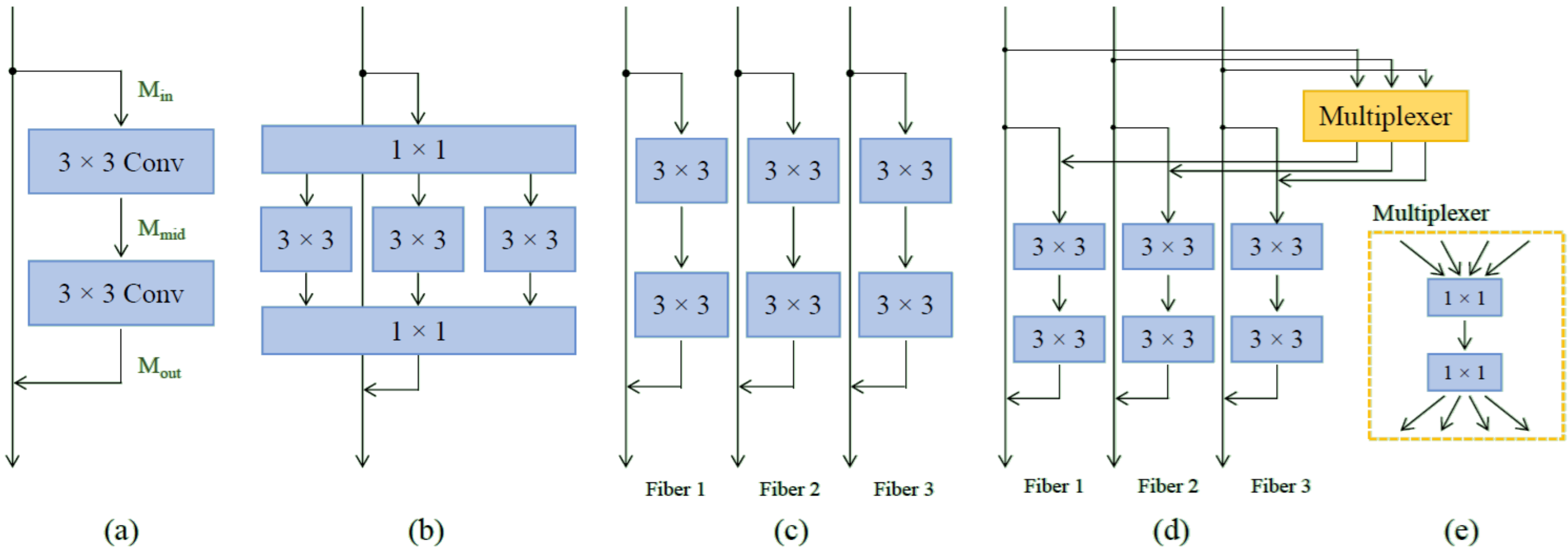


Sparse connections: Reduce computation cost

Multiplexer: Compensate the information loss

Chen Y, Kalantidis Y, Li J, et al. Multi-fiber networks for video recognition. ECCV, 2018.

Multi-Fiber Networks



Slicing Strategy

$$\# \text{ Connections} = M_{in} \times M_{mid} + M_{mid} \times M_{out}.$$

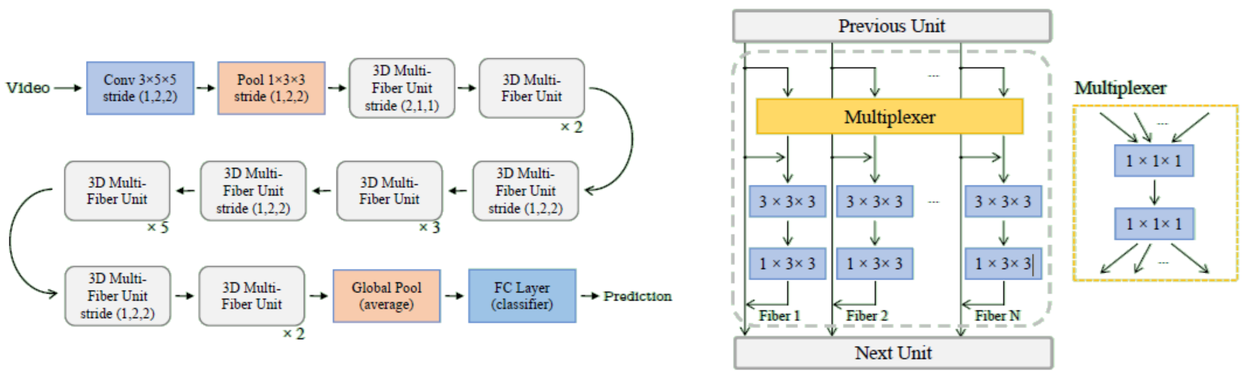
$$\begin{aligned}\# \text{ Connections} &= N \times (M_{in}/N \times M_{mid}/N + M_{mid}/N \times M_{out}/N) \\ &= (M_{in} \times M_{mid} + M_{mid} \times M_{out})/N.\end{aligned}$$

Multiplexer

one for dimension reduction and

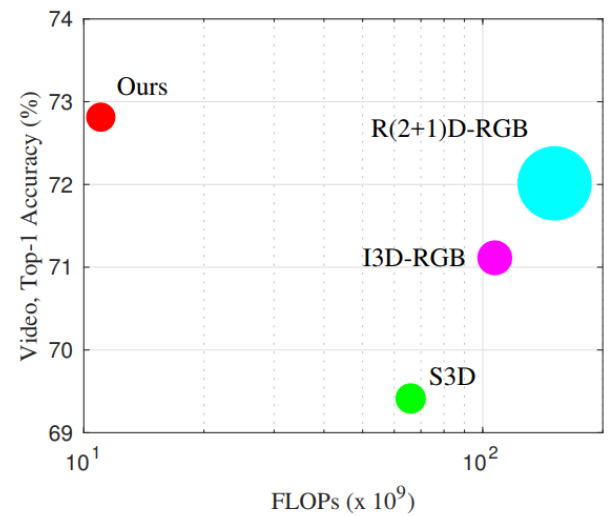
the other for dimension expansion.

Multi-Fiber Networks



(a) 3D Multi-fiber Network (b) 3D Multi-fiber Unit

Fig. 3. Architecture of 3D multi-fiber network. (a) The overall architecture of 3D Multi-fiber Network. (b) The internal structure of each Multi-fiber Unit. Note that only the first 3 × 3 convolution layer has expanded on the 3rd temporal dimension for lower computational cost.



Multi-Fiber Networks

layer	Repeat	#Channel	2D MF-Net		3D MF-Net	
			Output Size	Stride	Output Size	Stride
Input		3	224×224		$16 \times 224 \times 224$	
Conv1 MaxPool	1	16	112×112 56×56	(2,2) (2,2)	$16 \times 112 \times 112$ $16 \times 56 \times 56$	(1,2,2) (1,2,2)
Conv2	1 2	96	56×56	(1,1) (1,1)	$8 \times 56 \times 56$	(2,1,1) (1,1,1)
Conv3	1 3	192	28×28	(2,2) (1,1)	$8 \times 28 \times 28$	(1,2,2) (1,1,1)
Conv4	1 5	384	14×14	(2,2) (1,1)	$8 \times 14 \times 14$	(1,2,2) (1,1,1)
Conv5	1 2	768	7×7	(2,2) (1,1)	$8 \times 7 \times 7$	(1,2,2) (1,1,1)
AvgPooling			1×1		$1 \times 1 \times 1$	
FC			1000		400	
#Params			5.8 M		8.0 M	
FLOPs			861 M		11.1 G	

Multi-Fiber Networks

Table 3. Comparison on action recognition accuracy with state-of-the-arts on Kinetics. The complexity is measured using FLOPs, *i.e.* floating-point multiplication-adds. All results are only using RGB information, *i.e.* no optical flow. Results with citation numbers are copied from the respective papers.

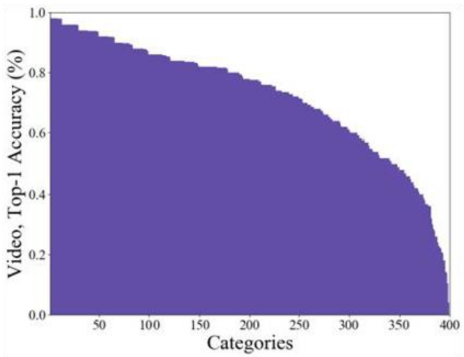
Method	#Params	FLOPs	Top-1	Top-5
Two-Stream [1]	12 M	–	62.2 %	–
ConvNet+LSTM [1]	9 M	–	63.3 %	–
S3D [8]	8.8 M	66.4 G	69.4 %	89.1 %
I3D-RGB [1]	12.1 M	107.9 G	71.1 %	89.3 %
R(2+1)D-RGB [2]	63.6 M	152.4 G	72.0 %	90.0 %
MF-Net (Ours)	**8.0 M**	**11.1 G**	**72.8 %**	**90.4 %**

Table 4. Action recognition accuracy on UCF-101 and HMDB51. The complexity is evaluated with FLOPs, *i.e.* floating-point multiplication-adds. The top part of the table refers to related methods based on 2D convolutions, while the lower part to methods utilizing spatio-temporal convolutions. Column "+OF" denotes the use of Optical Flow. FLOPs for computing optical flow are not considered.

Method	FLOPs	+OF	UCF-101	HMDB51
ResNet-50 [37]	3.8 G		82.3 %	48.9 %
ResNet-152 [37]	11.3 G		83.4 %	46.7 %
CoViAR [18]	4.2 G		90.4 %	59.1 %
Two-Stream [13]	3.3 G	✓	88.0 %	59.4 %
TSN [38]	3.8 G	✓	94.2 %	69.4 %
C3D [7]	38.5 G		82.3 %	51.6 %
Res3D [23]	19.3 G		85.8 %	54.9 %
ARTNet [16]	25.7 G		94.3 %	70.9 %
I3D-RGB [1]	107.9 G		95.6 %	74.8 %
R(2+1)D-RGB [2]	152.4 G		96.8 %	74.5 %
MF-Net (Ours)	**11.1 G**		96.0 %	74.6 %

Multi-Fiber Networks

☐ Drawbacks



assembling computer	100%	clapping	50%	drinking shots	21%
surfing crowd	100%	digging	50%	fixing hair	20%
paragliding	98%	kicking soccer ball	50%	recording music	18%
playing chess	98%	laughing	50%	sneezing	18%
playing squash or racquetball	98%	moving furniture	50%	faceplanting	14%
presenting weather forecast	98%	singing	50%	headbutting	14%
sled dog racing	98%	exercising arm	49%	sniffing	10%
snowkiting	98%	celebrating	48%	slapping	4%

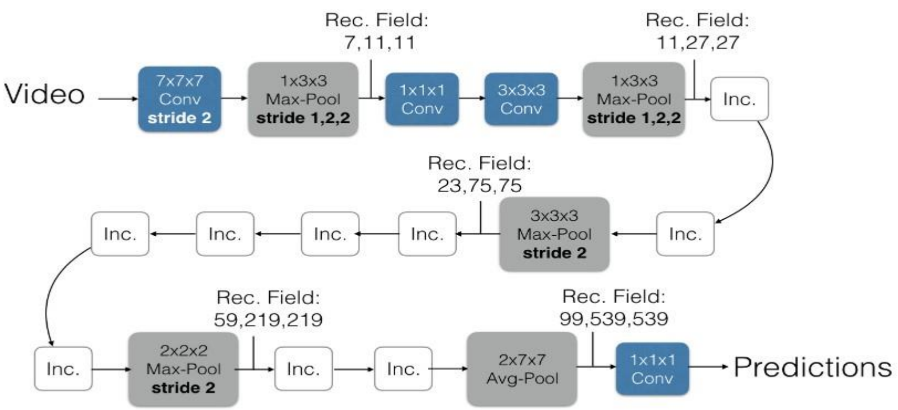
Fig. 7. Statistical results on Kinetics validation dataset. Left: Accuracy distribution of the proposed model on the validation set of Kinetics. The category is sorted by accuracy in a descending order. Right: Selected categories and their accuracy.

Highest accuracy: objects/backgrounds clearly distinguishable from other categories & actions spanning long duration.
Low accuracy: do not display any distinguishing object & the target action lasts for a very short time within a long video.

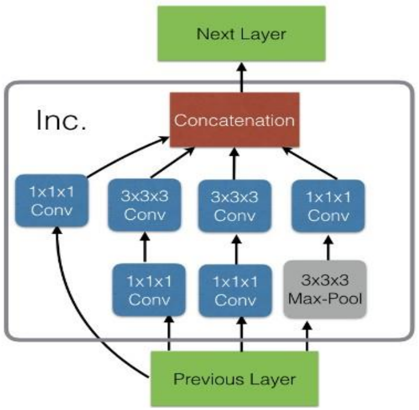
Future Work

- Designing effective modules in 3D CNNs can be crucial for lager-scale video classification

Inflated Inception-V1



Inception Module (Inc.)



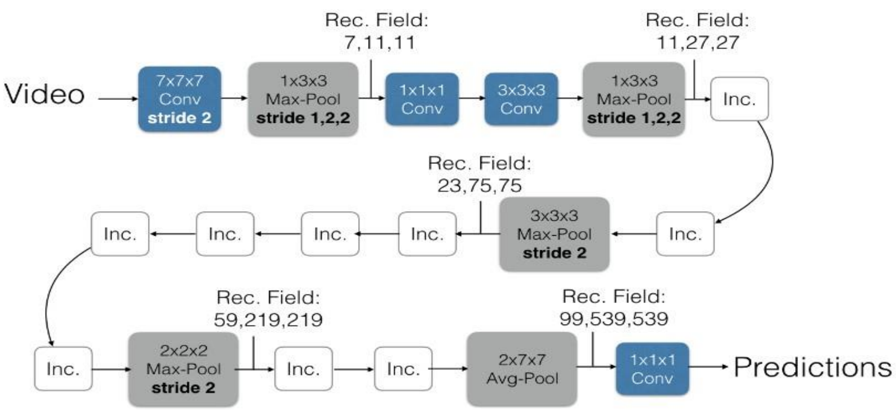
To name a few:

- Joao Carreira et al., Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, CVPR2017
- Zhaofan Qiu et al., Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, ICCV2017
- Du Tran et al., A Closer Look at Spatiotemporal Convolutions for Action Recognition, CVPR2018
- Limin Wang et al., Appearance-and-Relation Networks for Video Classification, CVPR2018
- Xiaolong Wang et al., Non-local Neural Networks, CVPR2018

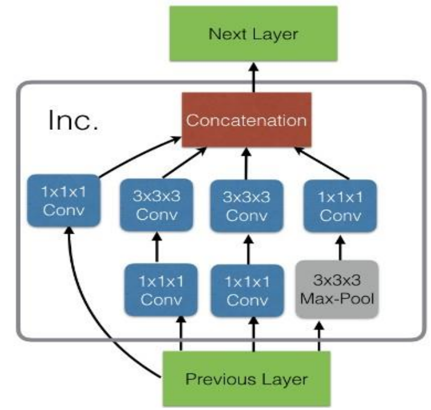
Future Work

- Designing effective modules in 3D CNNs can be crucial for lager-scale video classification

Inflated Inception-V1



Inception Module (Inc.)

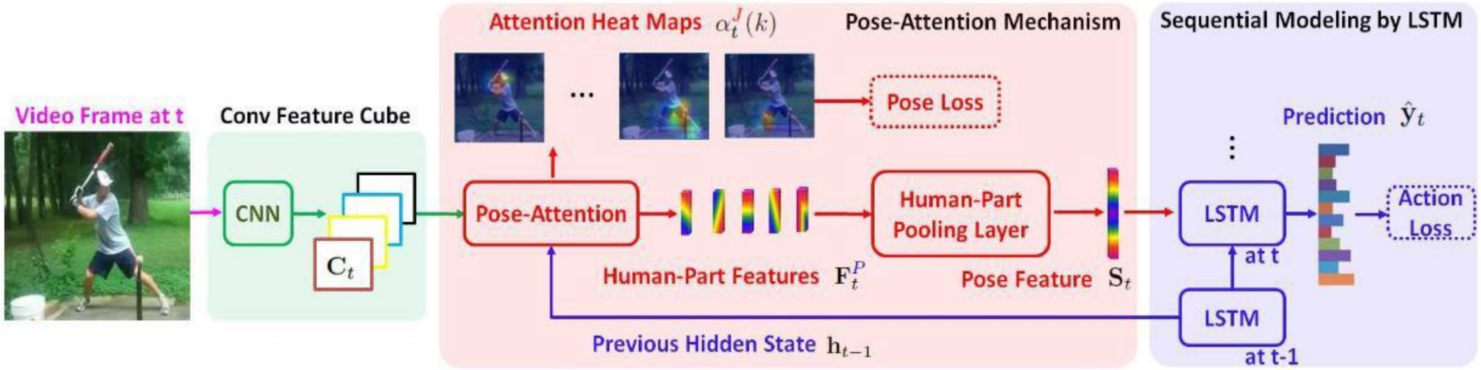


To name a few:

- Joao Carreira et al., Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, CVPR2017
- Zhaofan Qiu et al., Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, ICCV2017
- Du Tran et al., A Closer Look at Spatiotemporal Convolutions for Action Recognition, CVPR2018
- Limin Wang et al., Appearance-and-Relation Networks for Video Classification, CVPR2018
- Xiaolong Wang et al., Non-local Neural Networks, CVPR2018

Future Work

- Pose is a discriminative guidance for human actions in videos



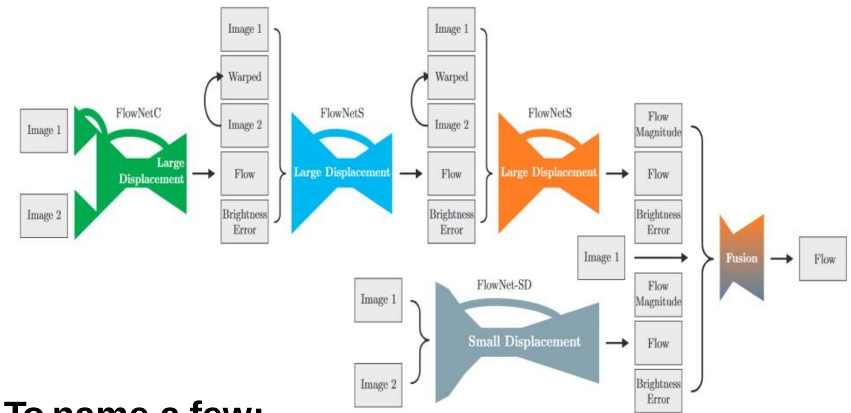
To name a few:

- Wenbin Du et al., RPAN: An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos, ICCV2017, oral (**ours**)
- Mohammadreza Zolfaghari et al., Chained Multi-stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection, ICCV2017
- Sijie Yan et al., Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, AAAI2018
- Mengyuan Liu et al., Recognizing Human Actions as Evolution of Pose Estimation Maps, CVPR2018
- Diogo Luvizon et al., 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning, CVPR2018
- Vasileios Choutas et al., PoTion: Pose MoTion Representation for Action Recognition, CVPR2018

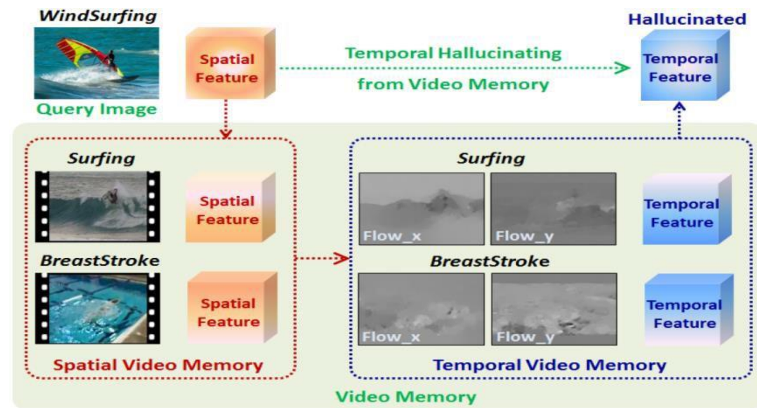
Future Work

☐ Motion prediction & flow-like features

Learning flow in the videos



Learning flow in the images?!



To name a few:

- Eddy Ilg et al., FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks, CVPR2017
- Zelun Luo,et al., Unsupervised Learning of Long-Term Motion Dynamics for Videos, CVPR2017
- Xiaodan Liang et al., Dual Motion GAN for Future-Flow Embedded Video Prediction , ICCV2017
- Shuyang Sun et al., Optical Flow Guided Feature: A Motion Representation for Video Action Recognition, CVPR2018
- Lijie Fan et al., End-to-End Learning of Motion Representation for Video Understanding, CVPR2018
- Ruohan Gao et al., Im2Flow: Motion Hallucination from Static Images for Action Recognition, CVPR2018
- Lei Zhou et al., Temporal Hallucinating for Action Recognition with Few Still Images, CVPR2018 (**ours**)

Thanks !