Action recognition - tasks

- Action classification: assigning an action label to a video clip

→

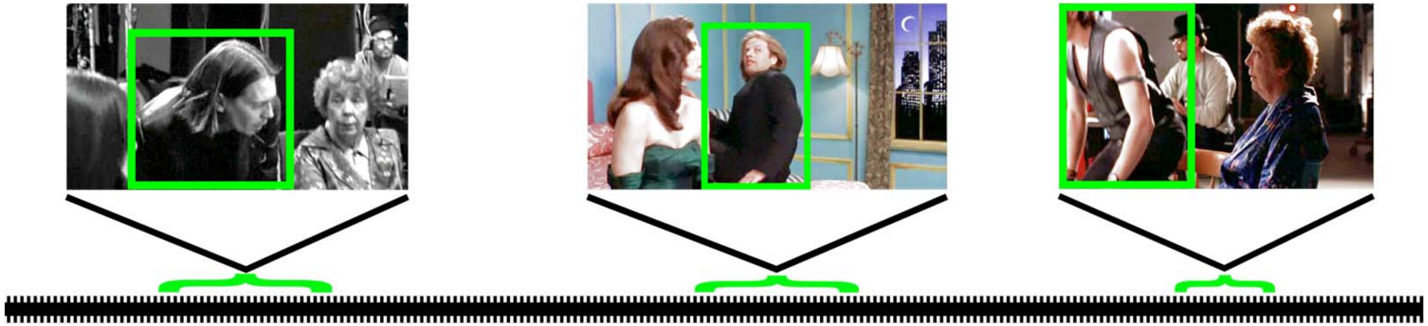
Making sandwich: present
Feeding animal: not present
...

Action recognition - tasks

- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present
...

- Action localization: search locations of an action in a video



Action classification in videos

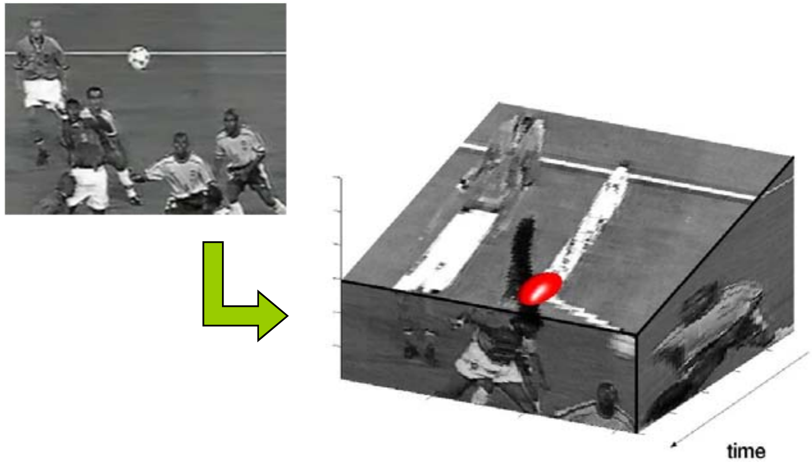
- Space-time interest points [Laptev, IJCV'05]
- Dense trajectories [Wang and Schmid, ICCV'13]
- Video-level CNN features

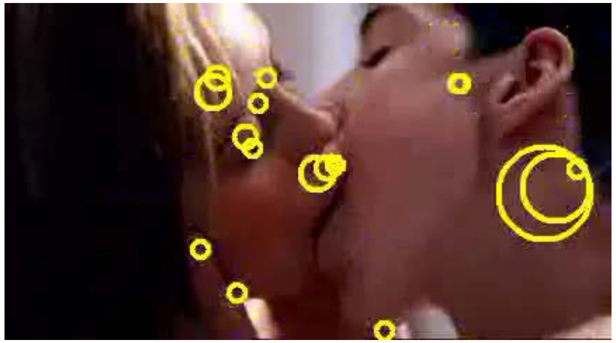
Space-time interest points (STIP) [Laptev'05]

- Space-time corner detector
  [Laptev, IJCV 2005]

$$H = \mathrm{det}(\mu) + k\,\mathrm{tr}^3(\mu)$$

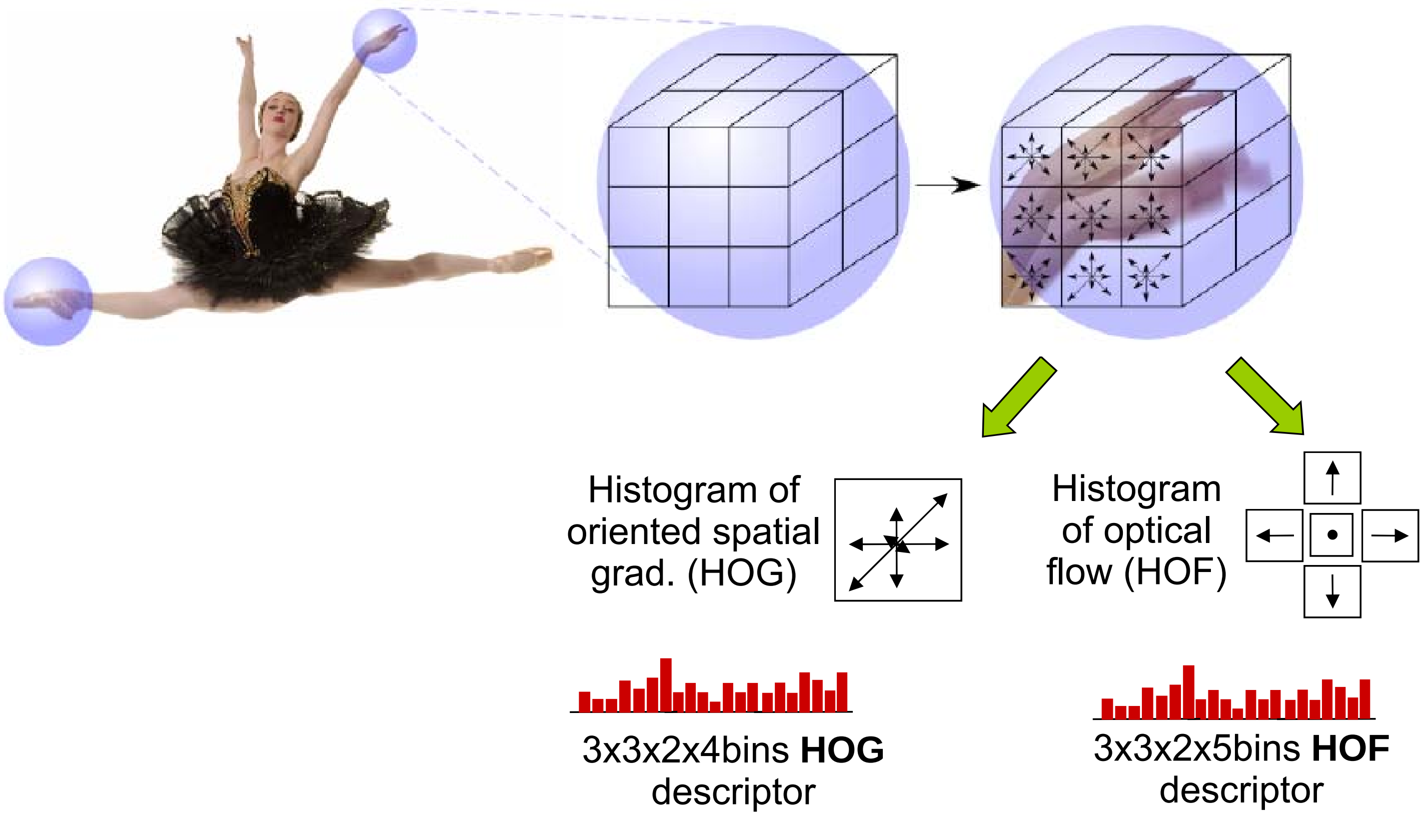
$$\mu = \begin{pmatrix} I_xI_x & I_xI_y & I_xI_t \\ I_xI_y & I_yI_y & I_yI_t \\ I_xI_t & I_yI_t & I_tI_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$





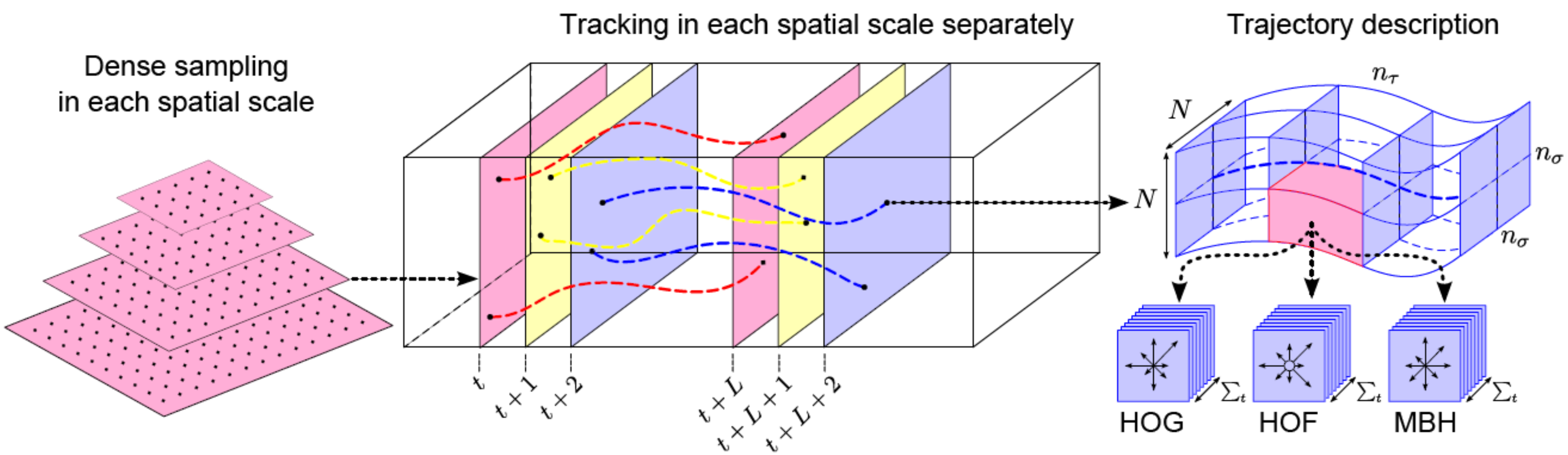
STIP descriptors

Space-time interest points



Dense trajectories [Wang et al., IJCV'13]

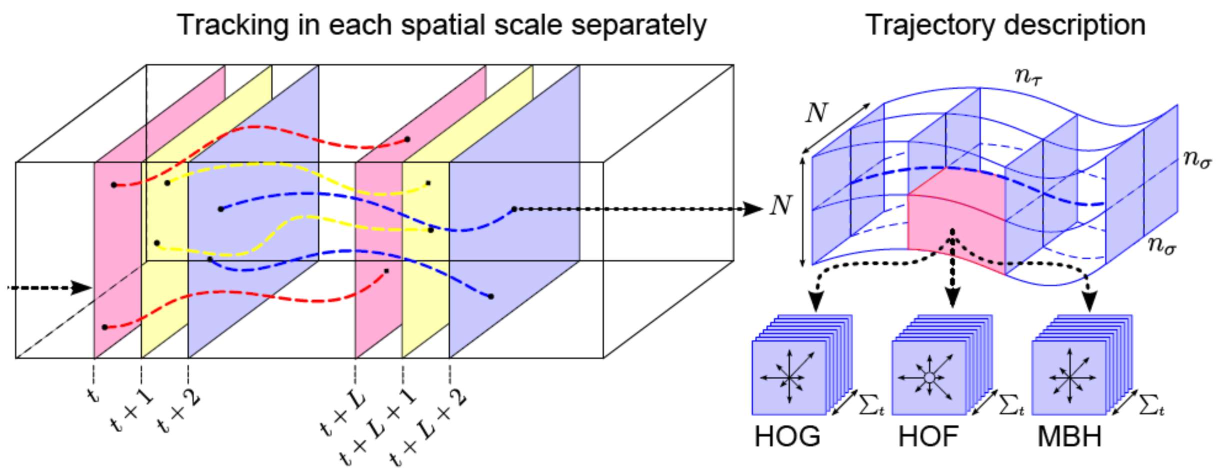
- Dense trajectories [Wang et al., IJCV'13] and Fisher vector encoding [Perronnin et al. ECCV'10]
  - Dense sampling at several scales
  - Feature tracking based on optical flow for several scales
  - Length 15 frames, to avoid drift



Example for dense trajectories

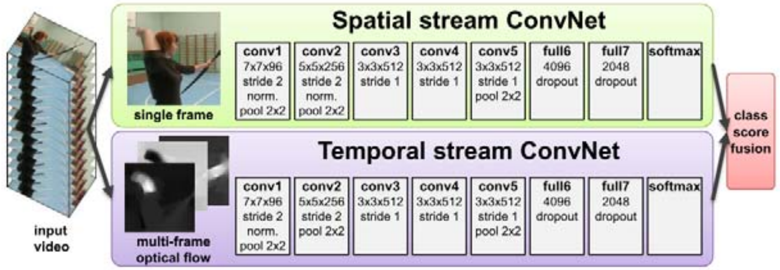
Descriptors for dense trajectory

- Histogram of gradients (HOG: 2x2x3x8)
- Histogram of optical flow (HOF: 2x2x3x9)
- Motion-boundary histogram (MBHx + MBHy: 2x2x3x8)

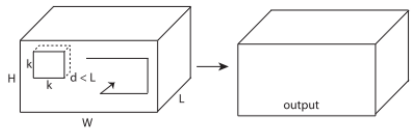


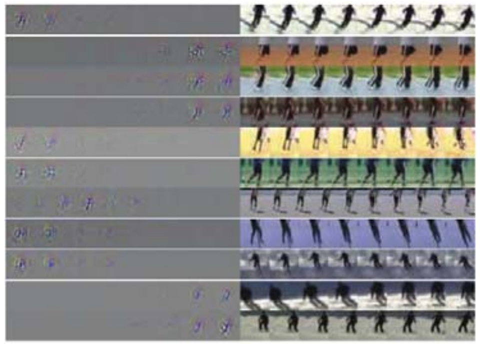
Recent CNN methods

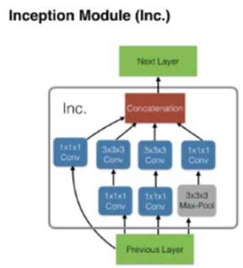
Two-Stream Convolutional Networks for Action Recognition in Videos [Simonyan and Zisserman NIPS14]



Learning Spatiotemporal Features with 3D Convolutional Networks [Tran et al. ICCV15]







Quo vadis action recognition? A new model and the Kinetics dataset [Carreira et al. CVPR17]

Recent CNN methods

Learning Spatiotemporal Features with 3D Convolutional Networks [Tran et al. ICCV15]

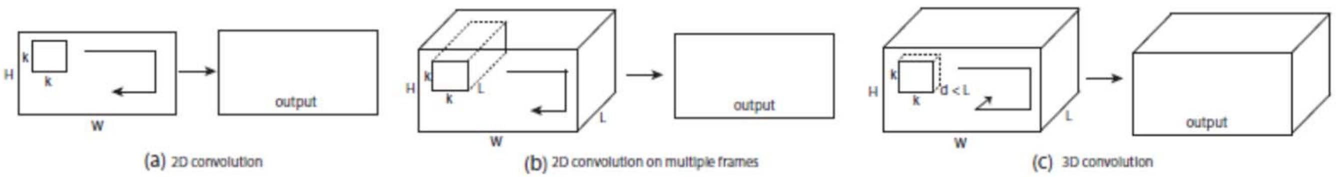
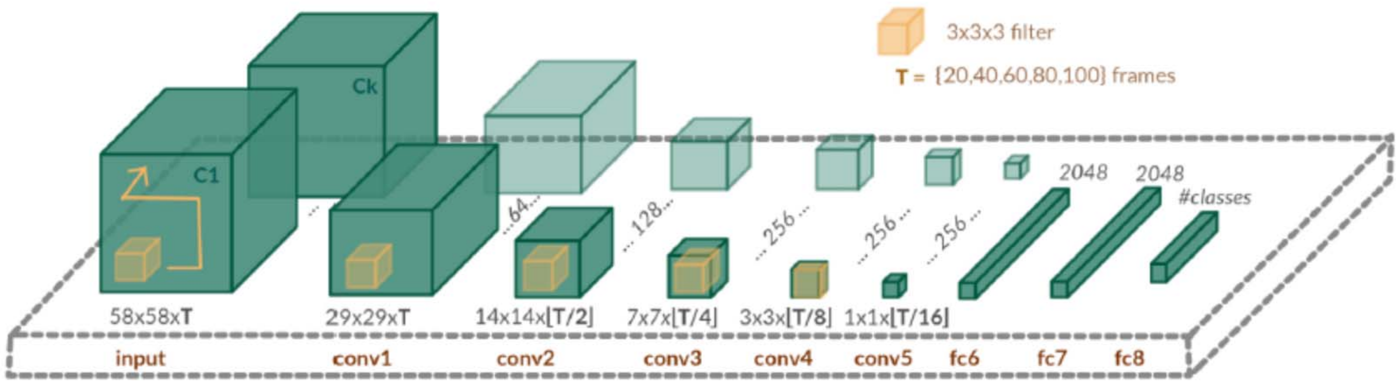
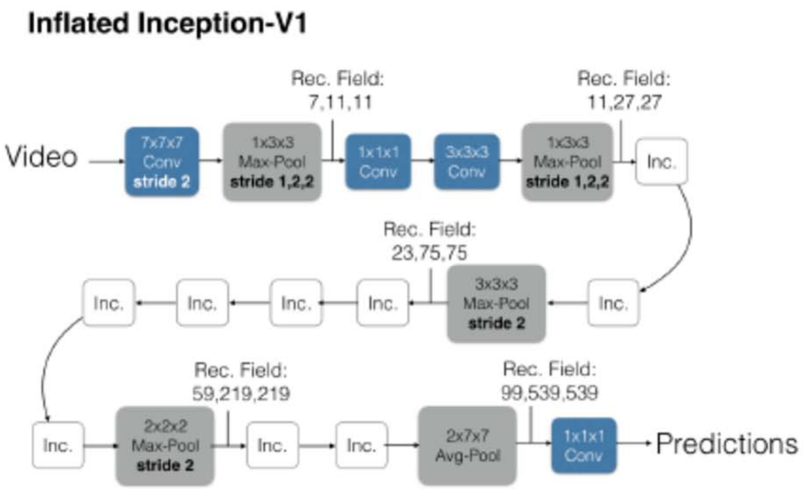


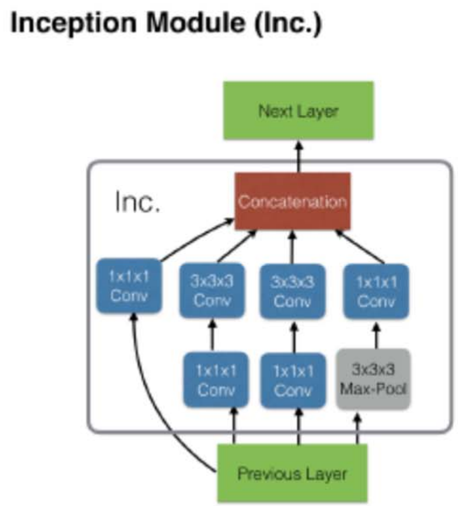
Figure 1. **2D and 3D convolution operations.** a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.



Recent CNN methods

Quo vadis, action recognition? A new model and the Kinetics dataset [Carreira et al. CVPR17]





Pre-training on the large-scale Kinetics dataset 240k training videos
→ significant performance grain

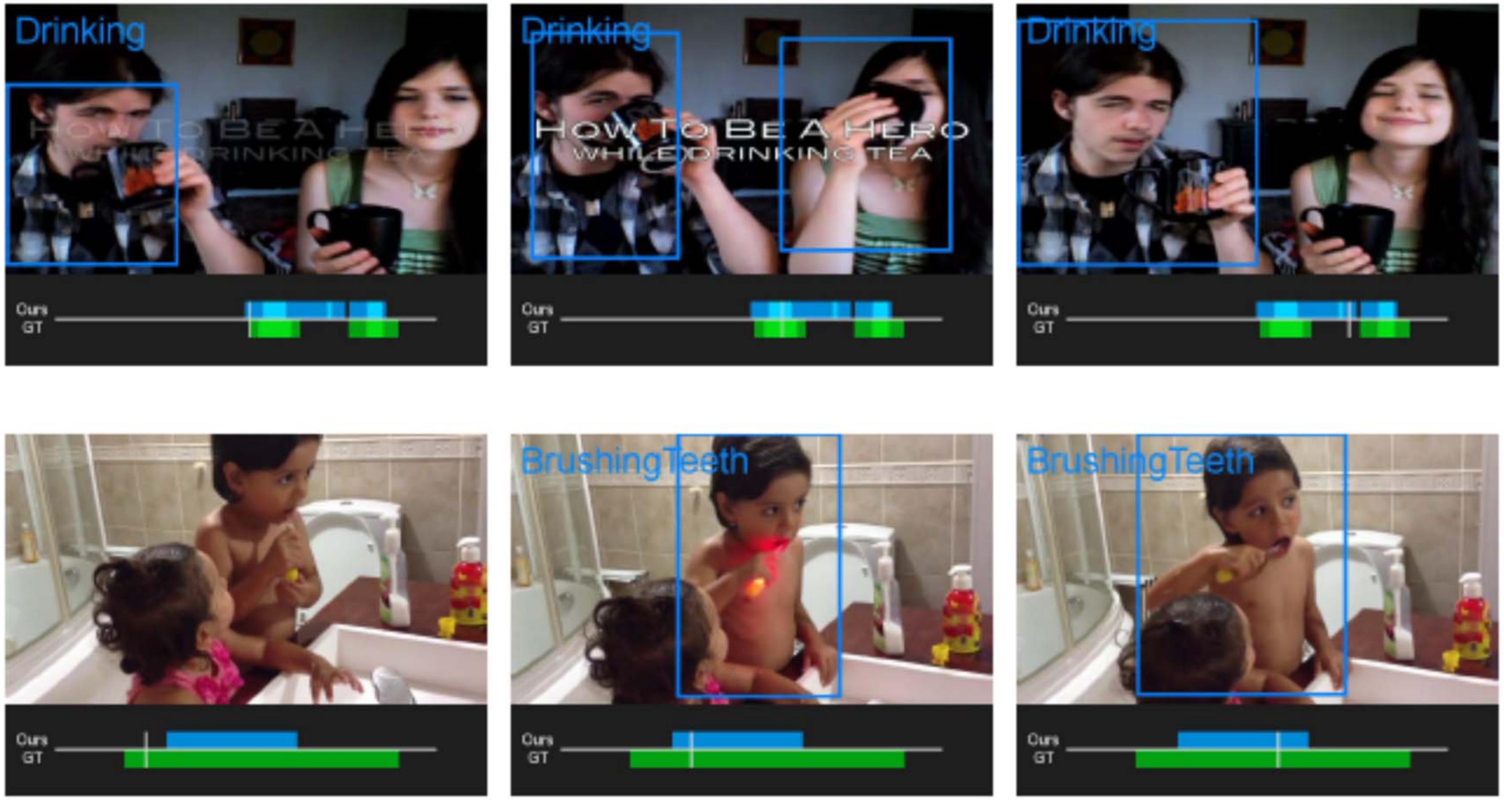
Overview

- Optical flow

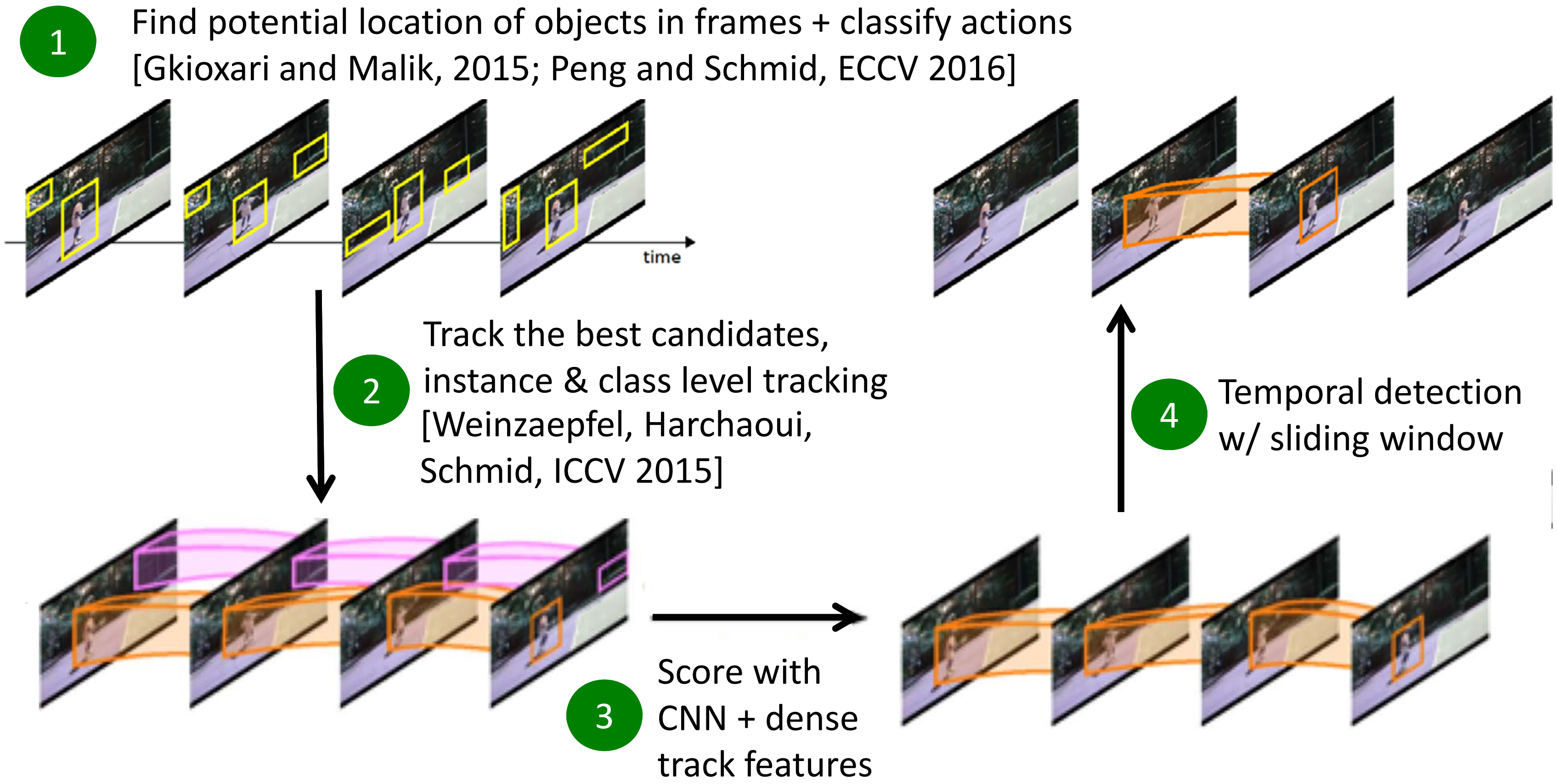
- Video classification
  - Bag of spatio-temporal features

- *Action localization*
  - *Spatio-temporal human localization*

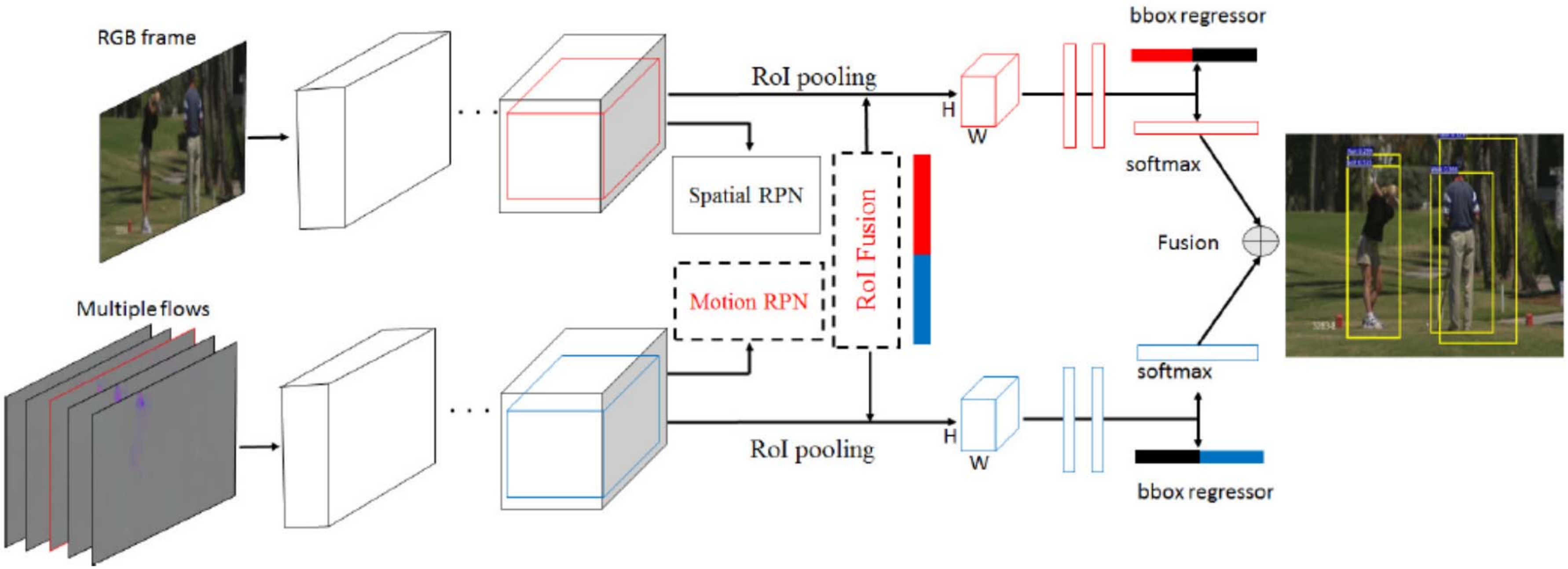
Spatio-temporal action localization



Spatio-temporal action localization



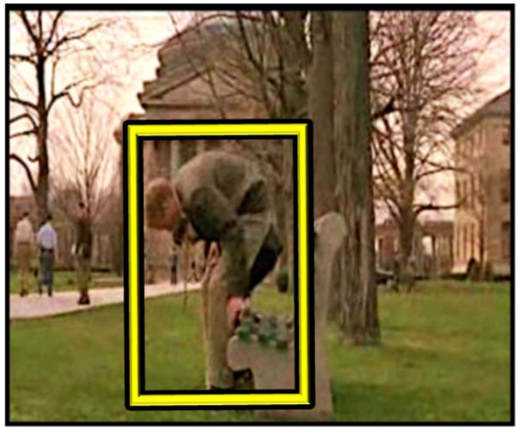
Frame-level detection: two stream Faster R-CNN [Peng & Schmid'16]



Better proposals: obtained on RGB and flow
Better features: flow from multiple frames + fusion with RGB

Action tubelets - motivation

Ambiguous action given only one frame



- Jump
- Sitting down ?

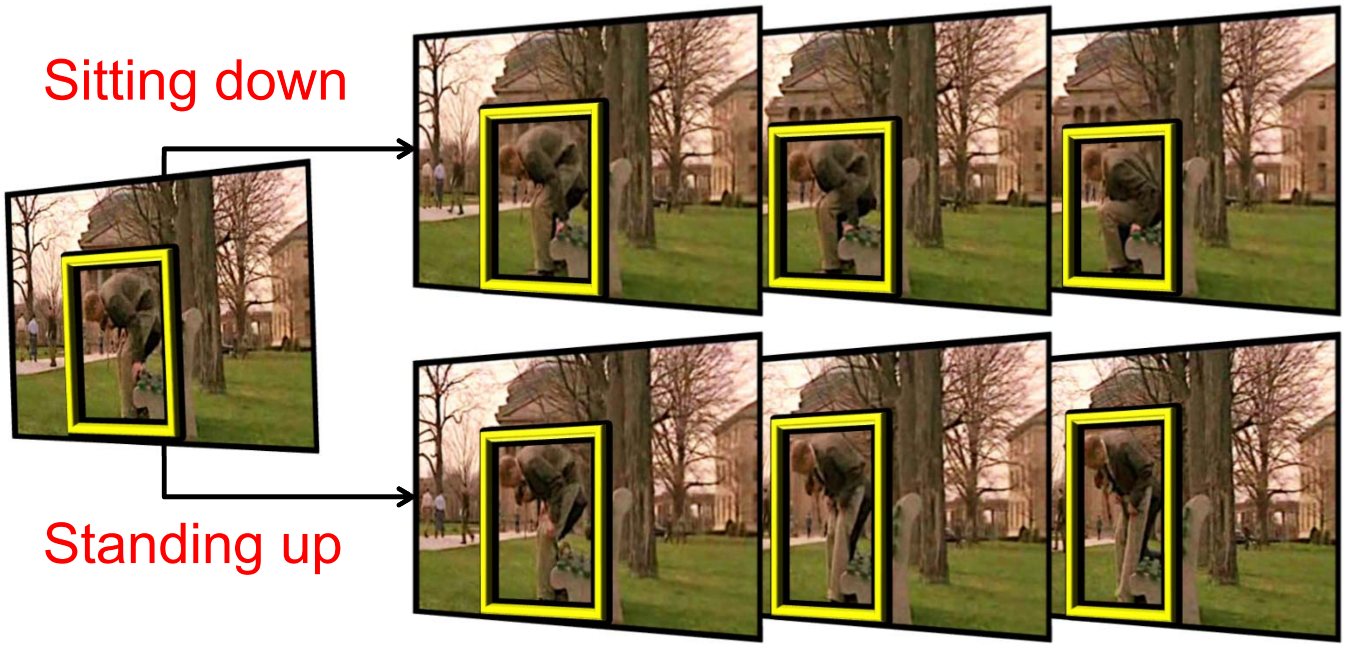
…

- Standing up ?
- Walk

[Action tublet detector for spatio-temporal action localization,
V. Kalogeiton, P. Weinzaephel, V. Ferrari, C. Schmid, ICCV'17]

Action tubelets - motivation

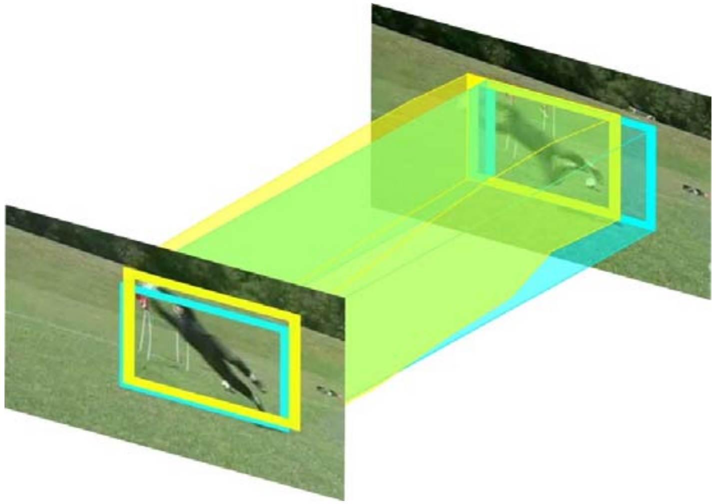
Ambiguity resolved given several frames

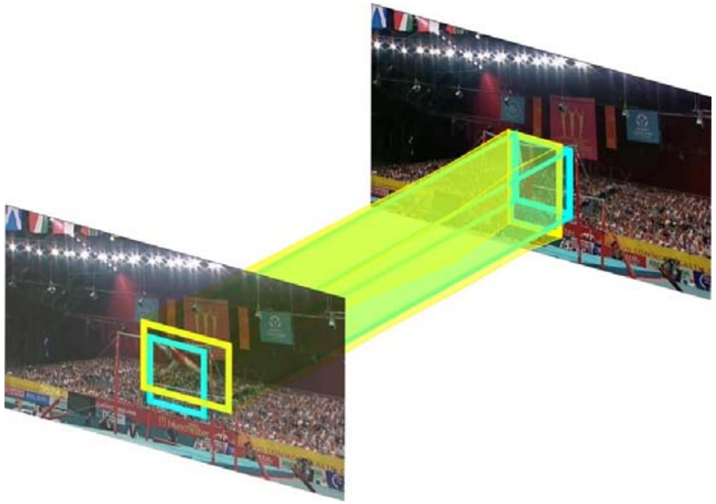


ACtion tubelet detector

Classify and regress spatio-temporal volumes

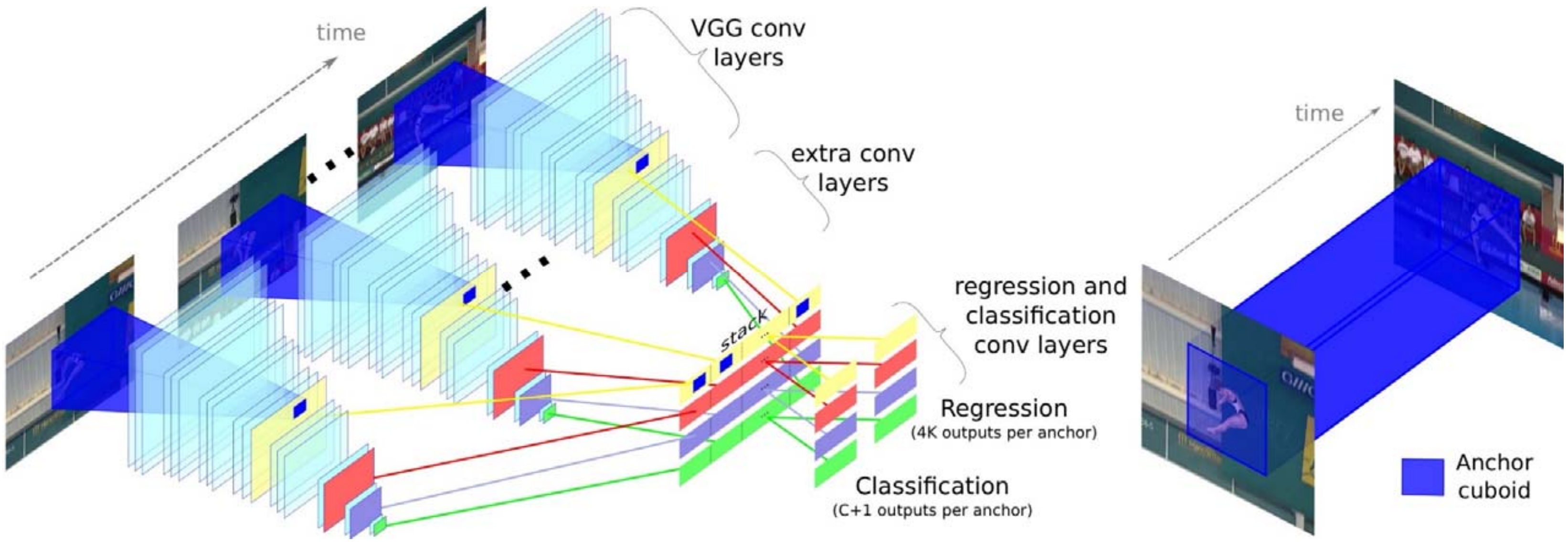
- *Anchor cuboids*: fixed spatial extent over time
- *Regressed tubelets*: score + deform the cuboid shape





ACtion tubelet detector

Use sequences of frames to detect *tubelets*: anchor cuboids

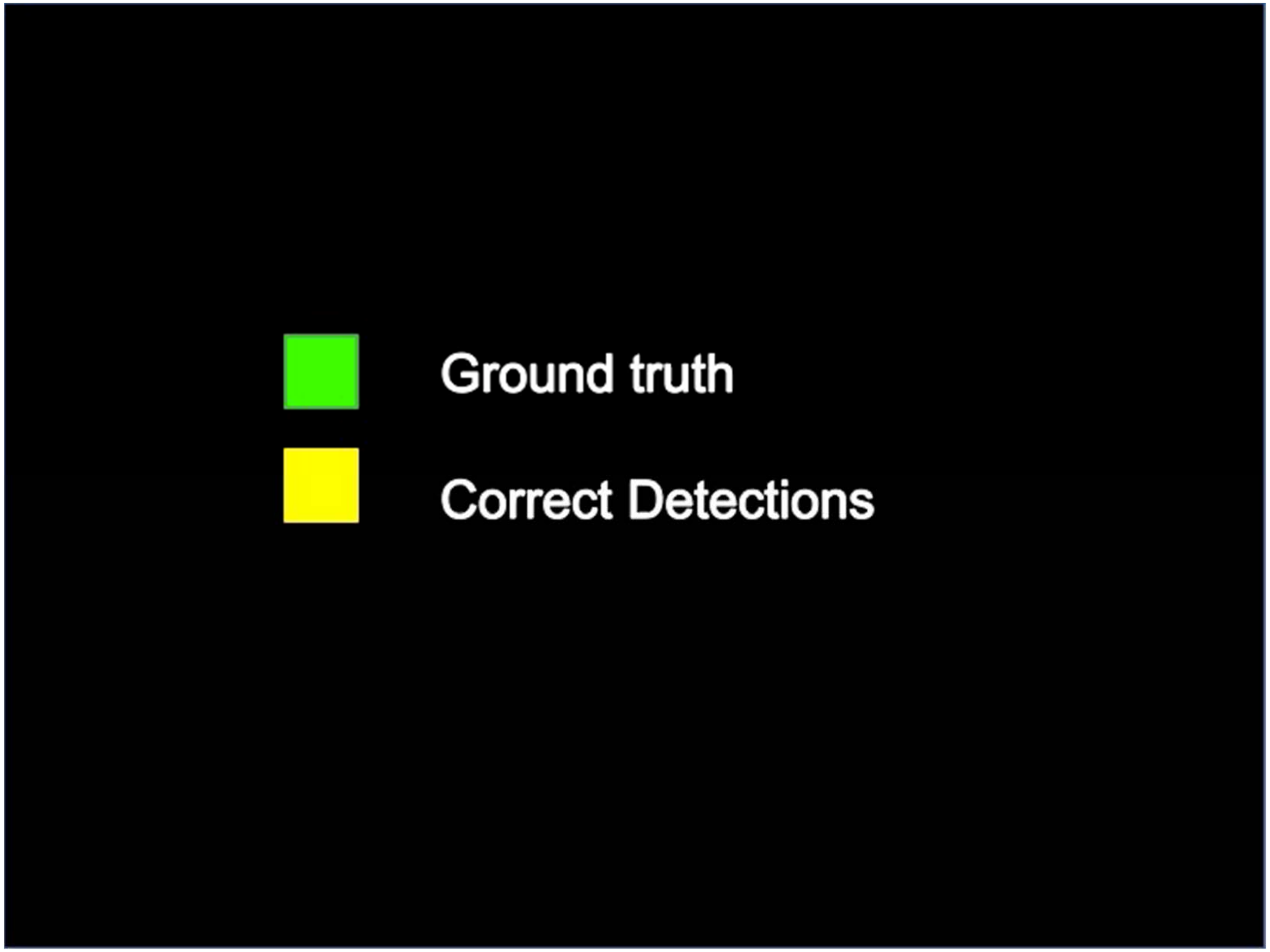


SSD detector [Liu et al., ECCV'16]

Quantitative results: Video-mAP results on **UCF-101**

detector	method	0.2	0.5	0.75	0.5:0.95
actionness	Wang et al, CVPR1	-	-	-	-
R-CNN	Gkioxari et al, CVPR15	-	-	-	-
	Weinzaepfel et al, ICCV15	51.7	-	-	-
Faster R-CNN	Peng et al, ECCV16 with MR	71.8	35.9	1.6	8.8
	Peng, et al, ECCV16 w/o MR	72.9	-	-	-
	Saha et al, BMVC16	66.7	35.9	7.9	14.4
SSD	Singh et al, arXiv17	73.5	46.3	15.0	20.4
	Ours	**75.8**	**51.5**	**22.5**	**24.8**

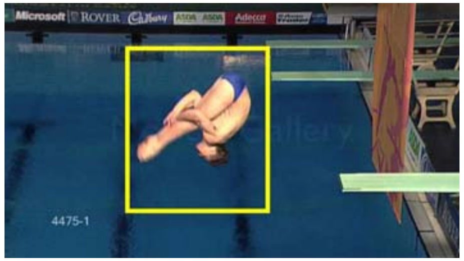
informatics mathematics



Ground truth
Correct Detections

Datasets for action localization

- Existing datasets are limited by diversity, duration, resolution
  - UCF-Sports (10 sports actions, 150 trimmed videos, similar context)



diving

lifting

running

  - J-HMDB (21 daily actions, 928 trimmed videos, avg length: 1.5s, low resolution)

climbing stairs

jumping

pushing

Datasets for action localization

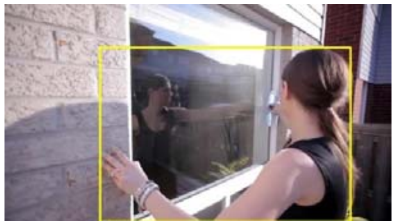
- Existing datasets are limited by diversity, duration, resolution
  - UCF-101 (24 sports actions, 3207 almost-trimmed low-res. videos)


basketball


long jump


rope climbing

  - DALY (10 actions, 3724 videos, 31 hours)


cleaning windows


playing harmonica


brushing teeth


ironing

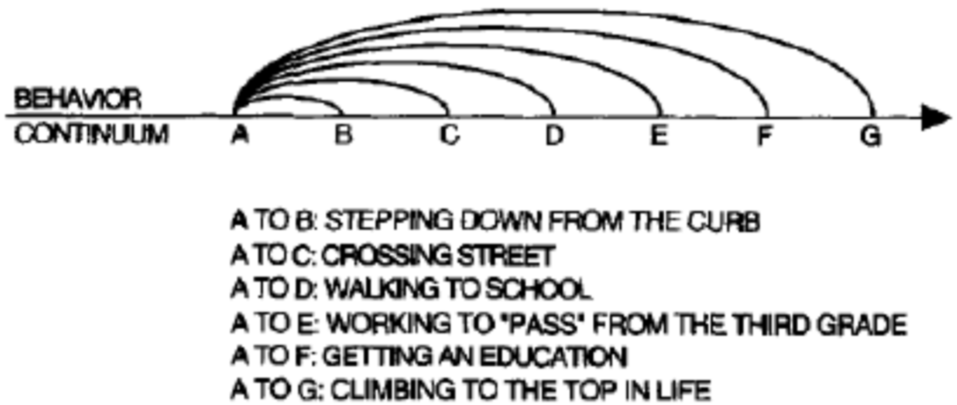
Atomic Visual Actions (AVA) dataset

- Towards a definition of atomic actions + large scale collection → Atomic Visual Actions (AVA) dataset

[AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions; Gu, Sun, Ross, Pantofaru, Li, Vijayanarasimhan, Toderici Ricco, Sukthankar, Schmid, Malik, CVPR'18]

AVA dataset – motivation

- Hierarchical nature of activities



[Barker and Wright'54]

- Basic units: atomic actions

Ava dataset – atomic actions

- Three categories of atomic actions:
  1) Pose of the person, eg., stand, sit, walk, kneel, swim
  2) Interactions with objects, eg., drive, carry, pick up
  3) Human-human interactions, eg., talk to, hug, fight

- Multiple labels per person

- Exhaustive annotation of all humans

run/jog
walk
jump
stand
sit
lie/sleep
bend/bow
crawl
swim
dance
get up
fall down
crouch/kneel
martial art

Pose (14)

talk to
watch
listen to
sing to
kiss
hug
grab
lift
kick
give/serve … to …
take … from …
play with kids
hand shake
hand clap
hand wave
fight/hit
push

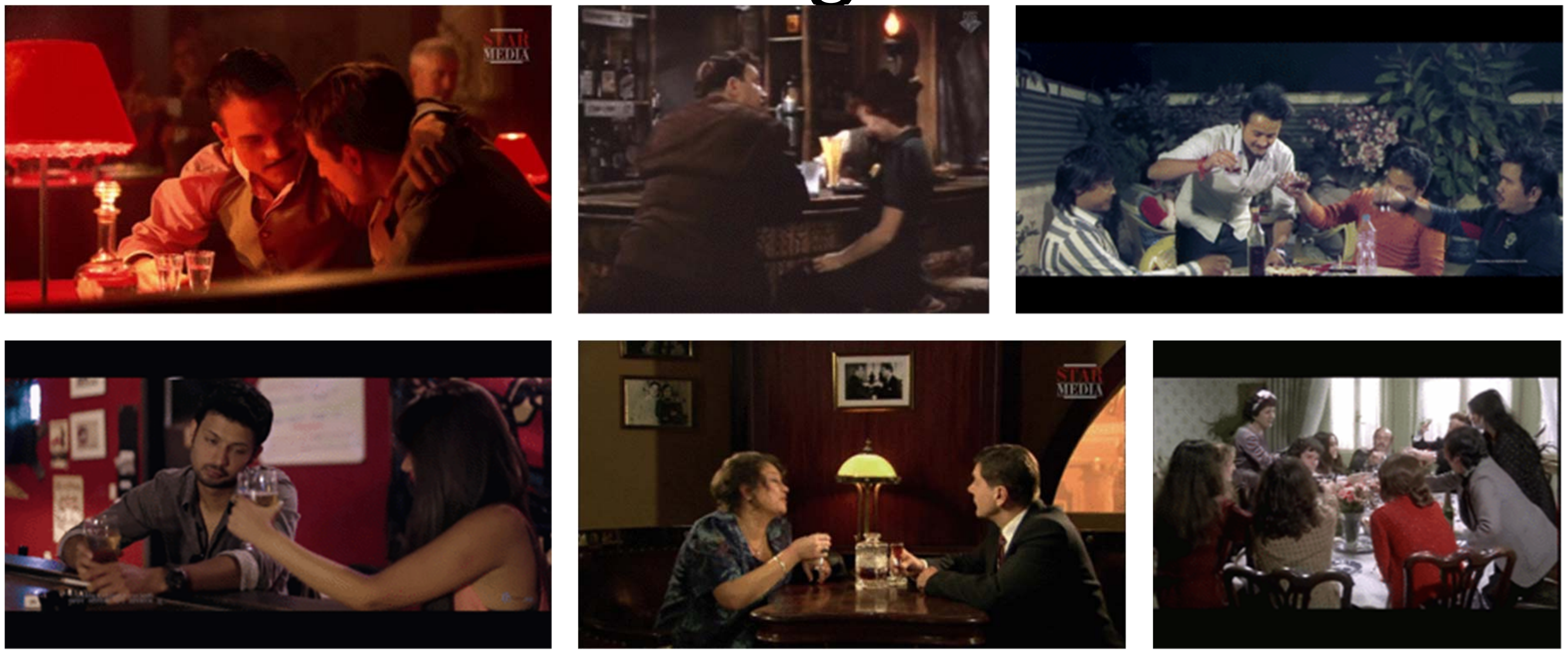
Person-person (17)

lift/pick up
put down
carry
hold
throw
catch
eat
drink
cut
hit
stir
press
extract
read
write
smoke
sail boat
row boat
fishing
touch
cook
kick
paint
dig
shovel
chop
shoot
take a photo
brush teeth
clink glass
work on a computer
answer phone
climb (e.g., mountain)
play board game
play with pets
drive (e.g., a car)
push (an object)
pull (an object)
point to (an object)
play musical instrument
text on/look at a cellphone
turn (e.g., screwdriver)
dress / put on clothing
ride (e.g., bike, car, horse)
watch (e.g., TV)
open
close
enter
exit

Person-object (49)

Answer phone

Clink glass



Hug (a person)

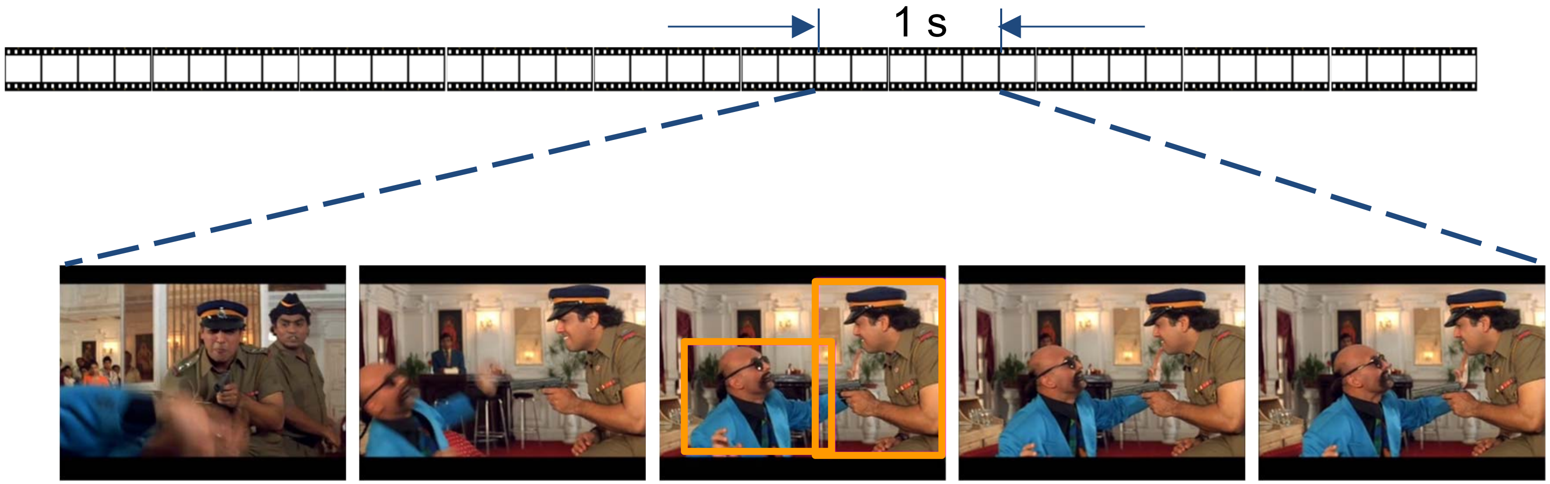




Left: Sit, Talk to, Watch; Right: Crouch/Kneel, Listen to, Watch

Left: Stand, Carry/Hold, Read; Middle: Stand, Take (object) from; Right: Stand, Give (object) to

AVA dataset - annotation

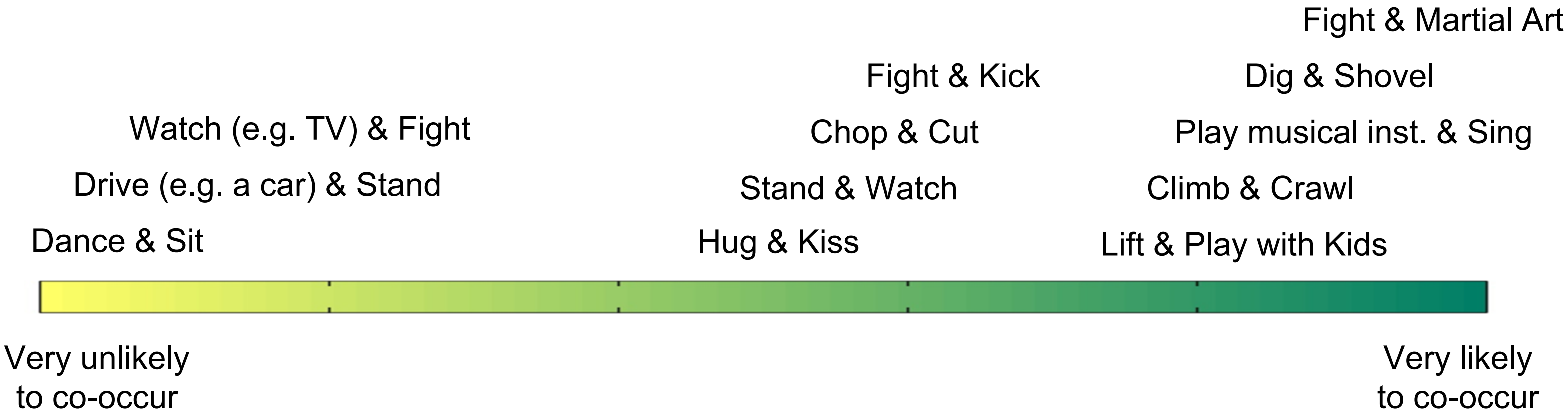


Left: Kneel, Talk to
Right: Stand, Listen, Shoot

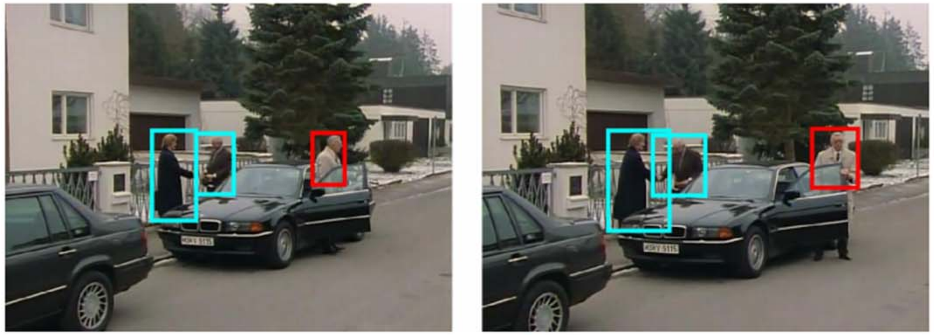
Ava dataset

- 192 videos with annotations for 15 minute intervals
- Annotation every 1 seconds
- 80 atomic actions in 107k movie segments with 740k labels with multiple labels per person
- Exhaustive annotation of all humans
  - Human are detected automatically and corrected manually

Human Action Co-occurrence



Action transitions



open → close

turn → open

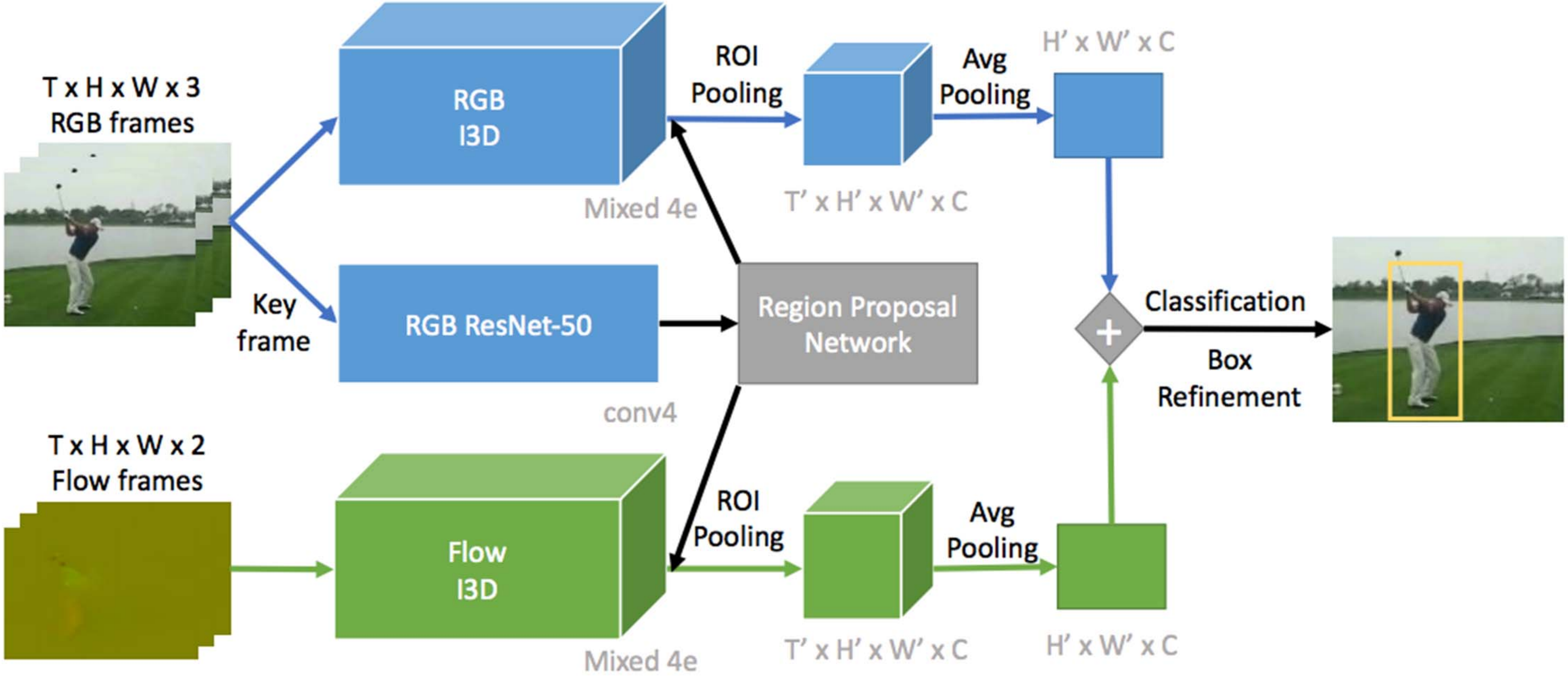
look at phone → answer phone

fall down → lie/sleep

Experimental setup

- AVA dataset (63 classes with at least 25 test instances)
- J-HMDB (928 clips, 21 classes)
- UCF101-24 (24 classes, 3k video clips)
- Evaluation metric: average precision at 50% IoU threshold

Action Detection Model



State-of-the-art performance

Frame-mAP	JHMDB	UCF101-24
Actionness [41]	39.9%	-
Peng w/o MR [29]	56.9%	64.8%
Peng w/ MR [29]	58.5%	65.7%
ACT [40]	65.7%	69.5%
Our approach	**73.3%**	**76.3%**

Video-mAP	JHMDB	UCF101-24
Peng w/ MR [29]	73.1%	35.9%
Singh *et al.* [37]	72.0%	46.3%
ACT [40]	73.7%	51.4%
TCNN [16]	76.9%	-
Our approach	**78.6%**	**59.9%**

Performance on AVA and Impact of Temporal Context

Model	Temp.+ Mode	JHMDB	UCF101-24	AVA
2D	1 RGB + 5 Flow	52.1%	60.1%	12.8%
3D	5 RGB + 5 Flow	67.9%	76.1%	13.4%
3D	10 RGB + 10 Flow	73.4%	78.0%	13.9%
3D	20 RGB + 20 Flow	76.4%	78.3%	14.9%
3D	40 RGB + 40 Flow	76.7%	76.0%	**16.2%**
3D	50 RGB + 50 Flow	-	73.2%	15.8%
3D	20 RGB	73.2%	77.0%	14.1%
3D	20 Flow	67.0%	71.3%	10.9%

Failure modes on AVA

FA for “hand shake”:
Reaching out arm

FA for “smoke”:
Hand covering mouth

FA for “write”:
Looking downwards

Failure modes on AVA

FA for "hand shake":
Reaching out arm

Other person does not reach out arm

FA for "smoke":
Hand covering mouth

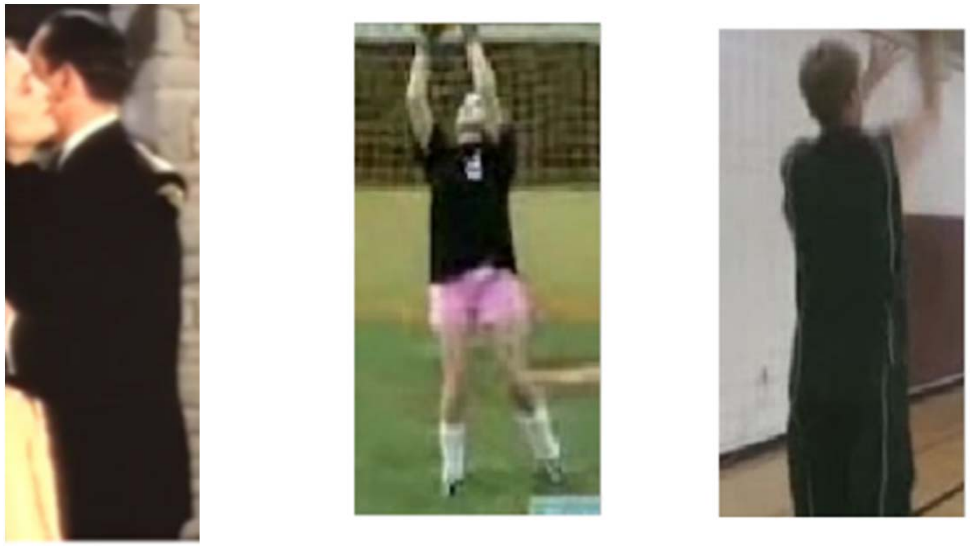
No cigarette in hand

FA for "write":
Looking downwards

Dining table with plates

Actor-centric Relation Network (ACRN)

- Faster RCNN look only at the actors (appearance, pose, etc.)
- Often we need to look at the relationship between an actor and other objects/ people to infer what they are doing



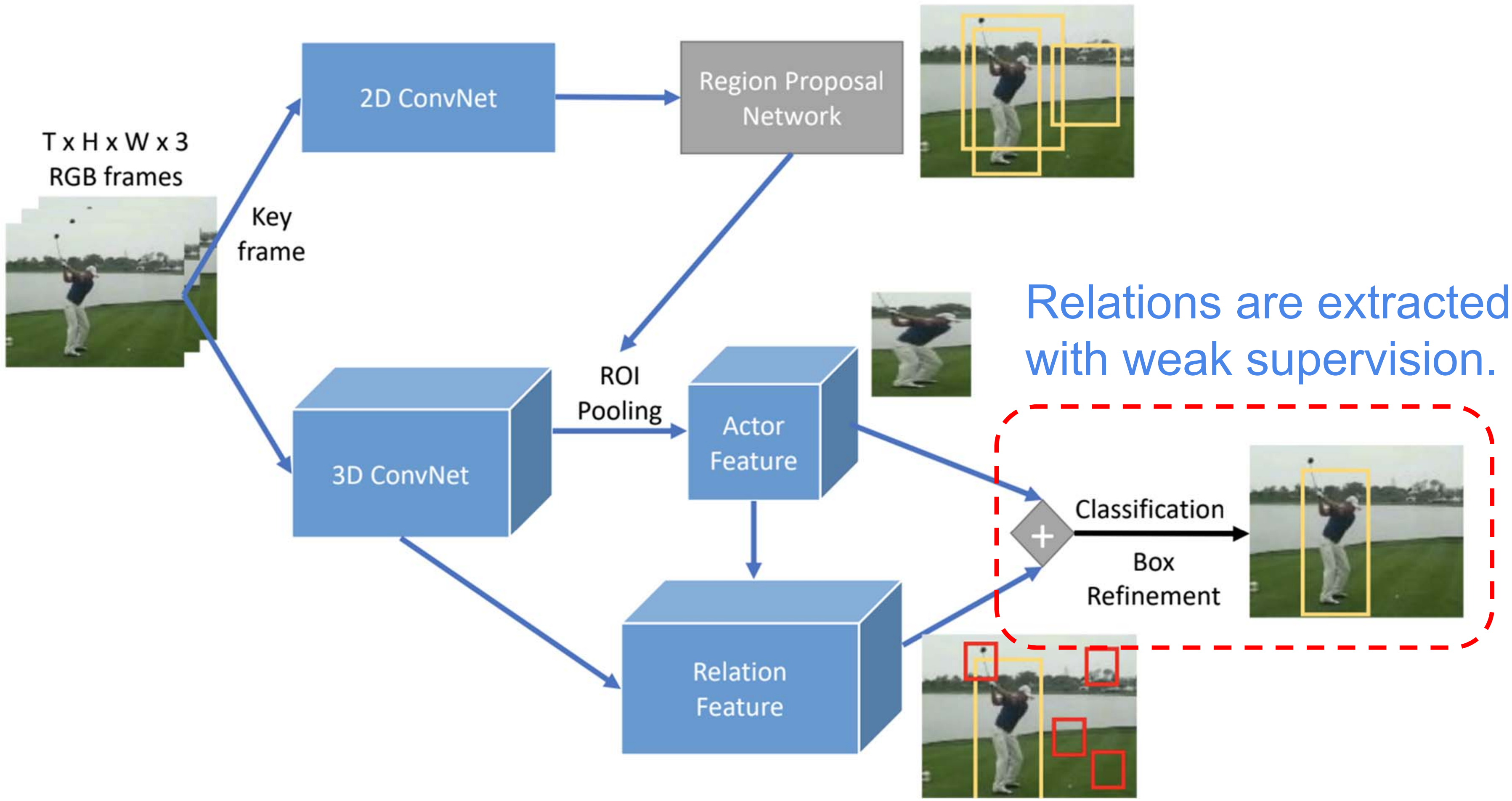
[Actor-centric relation network. C. Sun, A. Shrivastava, C. Vondrick, C. Schmid, R. Sukthankar and K. Murphy. arXiv, 2018.]

Actor-centric Relation Network (ACRN)

Catch ball

Throw ball

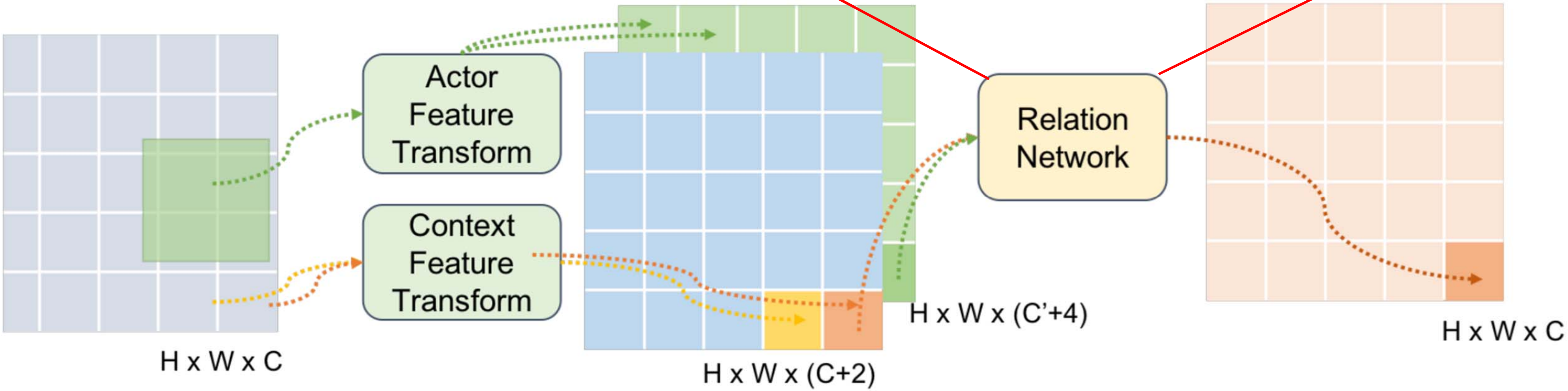
ACRN Architecture



Relations are extracted with weak supervision.

ACRN Arc

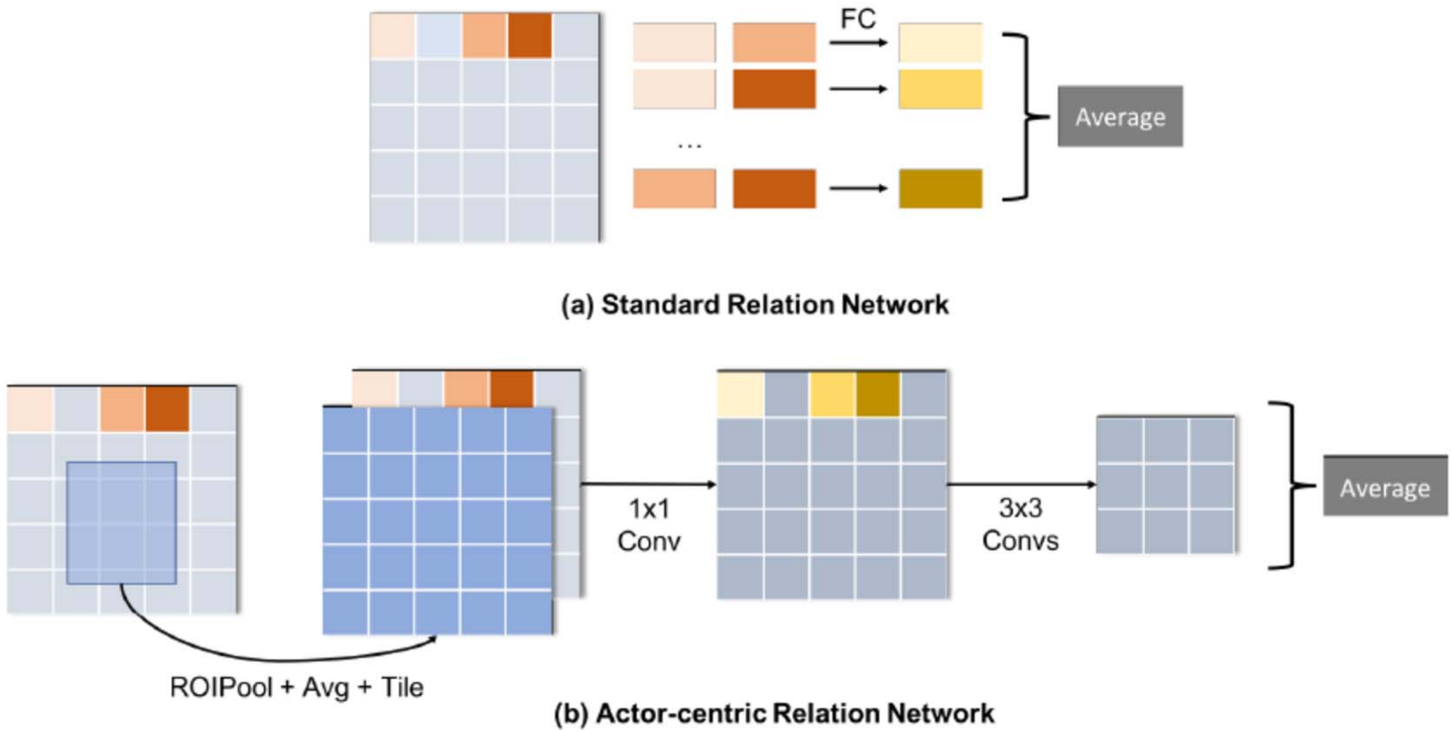
$$\mathrm{ACRN}(\mathcal{A}_i, I) = f_\phi\left(\sum_{j,k} g_\theta\left(a_i, o_{j,k}\right)\right)$$



- Pairwise relation between actor and "objects"
- No explicit objectness proposals, use feature cells
- Implemented as 1x1 convolutions

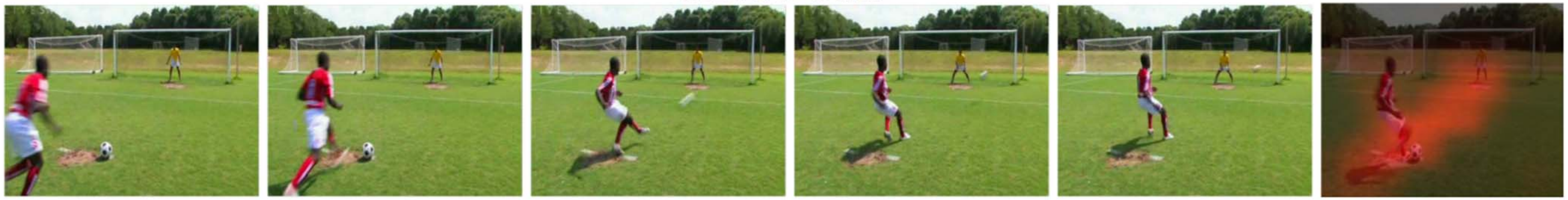
Related work: Santoro et al., A simple neural network module for relational reasoning. NIPS 2017.

ARCN architecture



Visualizations

shoot ball



kick ball

pour

Visualizations

Comparison with SOTA

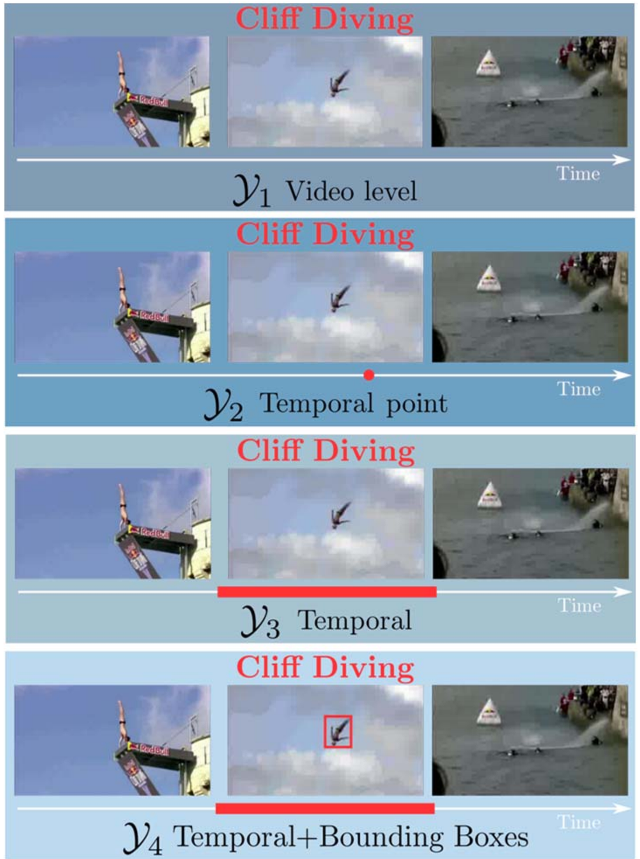
Model	frame-AP	video-AP
Peng et al. [9]	58.5	73.1
Singh et al. [36]	-	72.0
ACT [21]	65.7	73.7
I3D [5]	73.3	78.6
Base Model	75.2	78.8
ACRN	**77.9**	**80.1**

(a) JHMDB (3 splits)

Model	frame-AP
Two Stream [5]	14.2
I3D [5]	15.1
Base Model	15.5
ACRN	**17.4**

(b) AVA

Action localization with varying levels of supervision



Annotation: actions performed in the videos

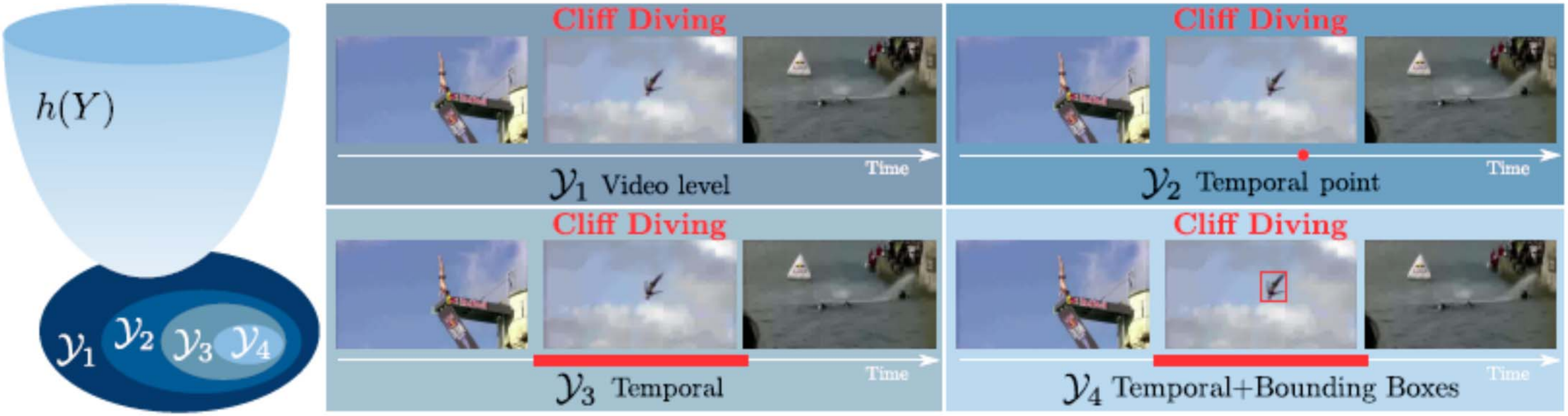
Annotation: one frame inside the action (temporal point)

Annotation: temporal interval

Annotation: temporal interval + one spatial human box

[A flexible model for training action localization with varying levels of supervision; G. Cheron, JB Alayrac, I. Laptev and C. Schmid; arXiv 2018]

Action localization with varying levels of supervision



Y: assignment of human tracklets to action labels
h(Y): objective function
$\mathcal{Y}_1 \supset \mathcal{Y}_2 \supset \mathcal{Y}_3 \supset \mathcal{Y}_4$ increasingly stricter constraints

Approach

- Person tracks are obtained by automatic detection + linking
- Tracks are subdivided into short elementary segments called "*tracklets*"
- Given M tracklets and K possible action classes (including background), assign "correct" action class to each tracklet

$$Y \in \{0,1\}^{M\times K}$$

- Discriminative clustering

$$\min_{Y\in\mathcal{Y}_s} h(Y) \qquad h(Y) = \min_{W\in\mathbb{R}^{d\times K}} \frac{1}{2M}\|XW - Y\|_F^2 + \frac{\lambda}{2}\|W\|_F^2$$

$\mathcal{Y}_s$ Set of constraints W is the classifier

- Optimization with block coordinate Frank-Wolfe algorithm

Experimental setup

- Person tracks:
  - If BB are available fine-tuned Faster R-CNN + on-line linking (score + overlap)
  - Otherwise off-self Faster R-CNN detector trained on COCO + KLT
- Tracklet feature representation
  - Average I3D features (RGB + OF) pooled from tracklet bounding boxes
- Datasets
  - UCF101-24
  - DALY

Experimental results

Supervision		Video level	Shot level	Temporal point	Temporal	Temporal + spatial points			1 BB	Temp. + 1 BB		Temp + 3 BBs		Fully Suppervised	
Method		Our	Our	Our	Our	Our	[46]	[27]	Our	Our	[46]	Our	[46]	Our	[46]
UCF101-24	@0.2	43.9	-	45.5	47.3 (69.5)	49.1 (69.8)	57.5	34.8	66.8	70.6	57.4	74.5	57.3	76.0	58.9
(mAP)	@0.5	17.7	-	18.7	20.1 (38.0)	19.5 (39.5)	-	-	36.9	38.6	-	43.2	-	50.1	-
DALY	@0.2	7.6	12.3	26.7	31.5 (33.4)	No continuous spatial GT			28.1	32.5	14.5	32.5	13.9	No full GT available	
(mAP)	@0.5	2.3	3.9	8.1	9.8 (14.3)				12.2	13.3	-	15.0	-		

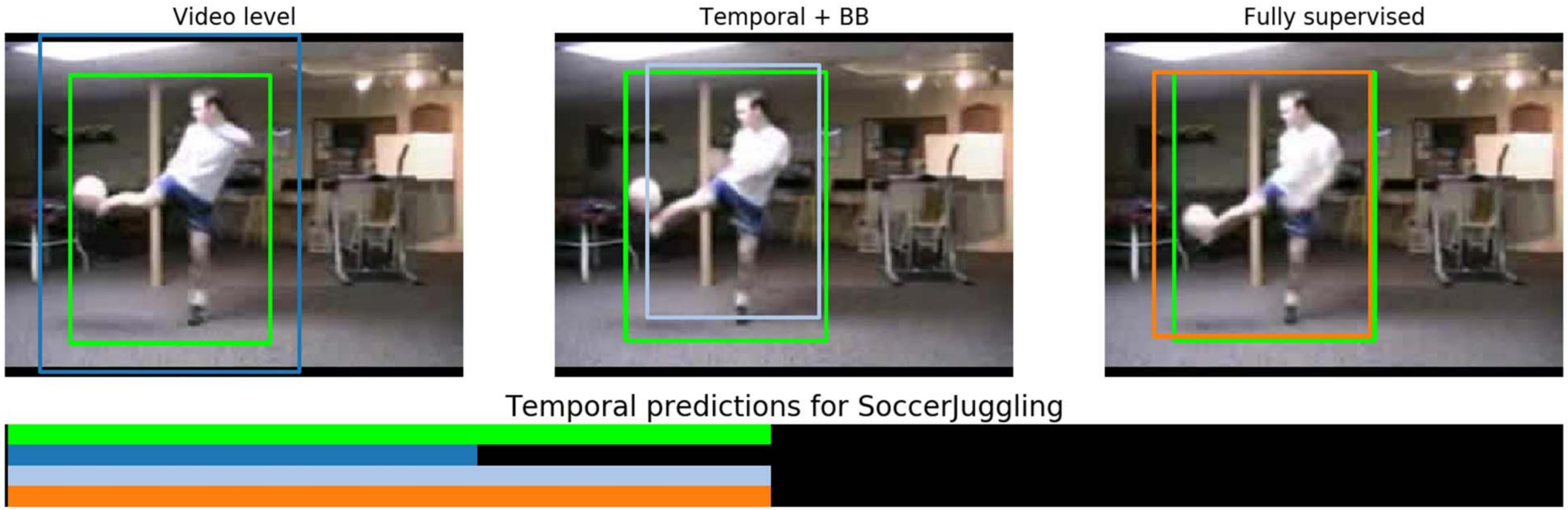
[27] Pascal Mettes, Jan C. van Gemert, and Cees G. M. Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV*, 2016. 1, 3, 7

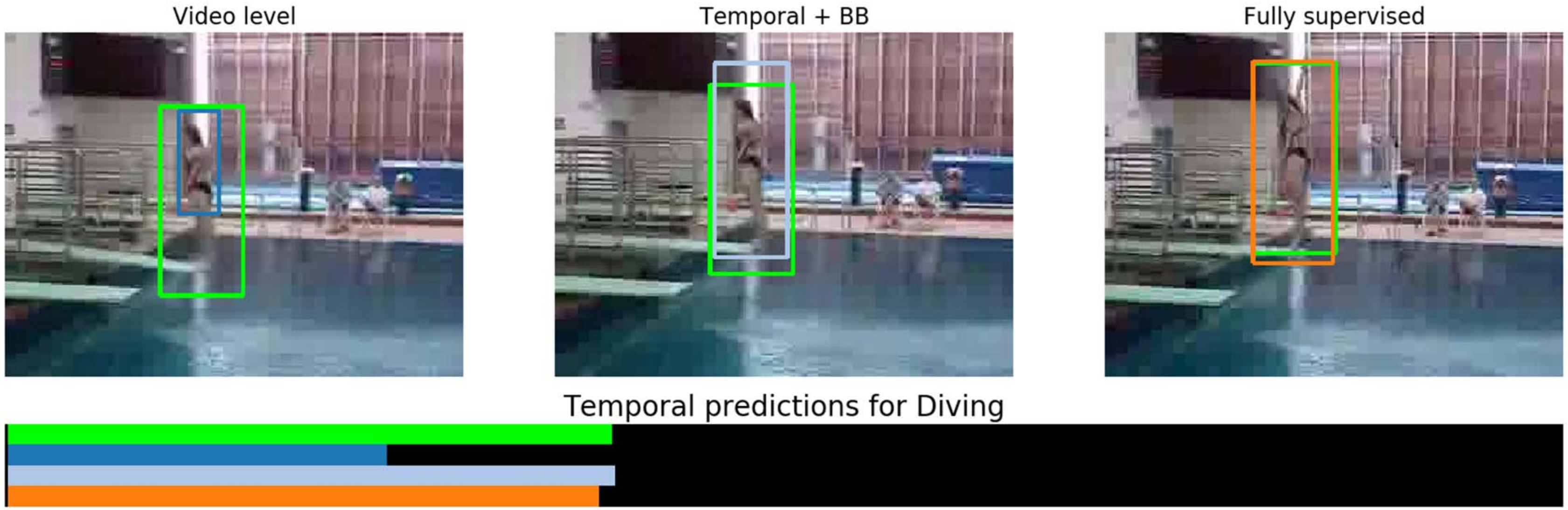
[46] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. In *CoRR*, 2016. 1, 3, 5, 6, 7, 8

UCF101-24: good results for temporal annotation+ 1 BB
without BB annotation decrease in performance due to human detections

DALY: excellent results for temporal annotation only, video level difficult due ratio action length / video length

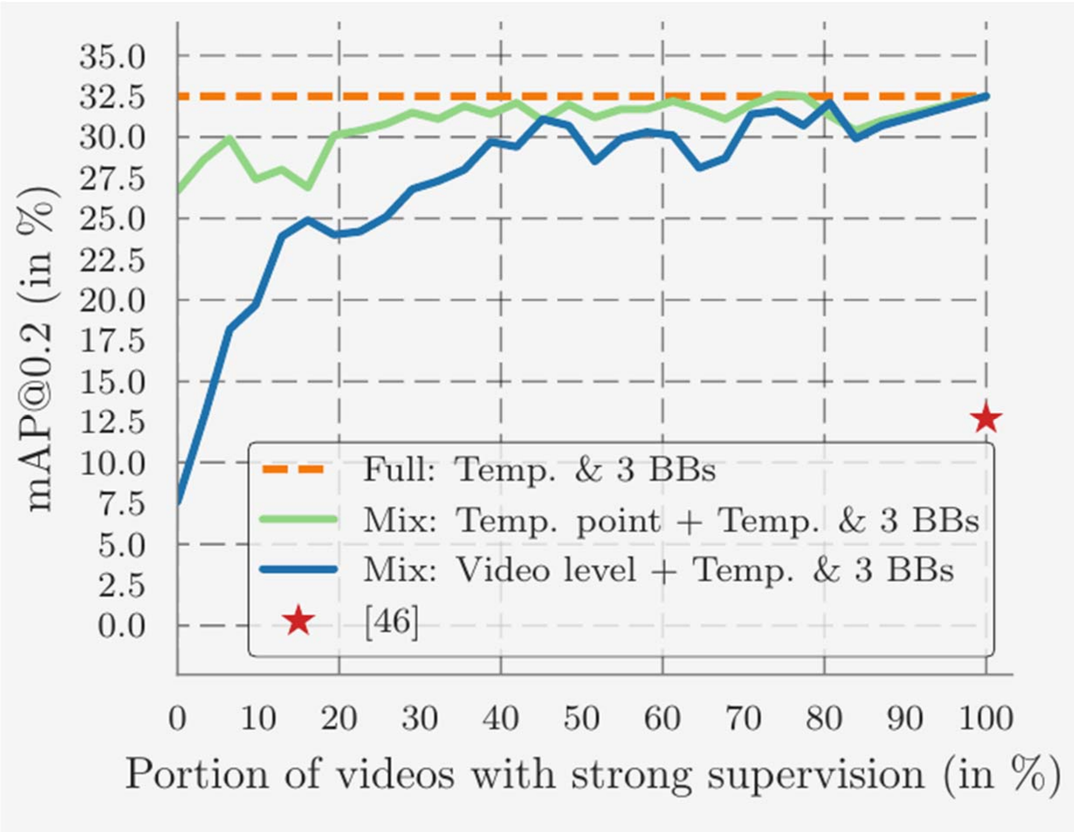
Results on DALY





Results on UCF101-24

Mixing different levels of supervision (DALY)



Video level:
Significant improvement with small number of full annotations

Temp. point:
Excellent results
Improvement by adding the bounding box annotations

Conclusion

- Importance of dataset for evaluation
- Design of a new model to take into account spatial relations
- Excellent results for weakly supervised training + mixed trainig
- Human detection could be still improved
- Cross-model integration with text and sound

JOB OPENINGS

Inria postdocs
Google research scientists

cordelia.schmid@inria.fr