DeepMind

UNIVERSITY OF OXFORD

Human action recognition and the Kinetics dataset

Andrew Zisserman

Includes slides from Joao Carreira and Rohit Girdhar

Outline

1. The Kinetics human action video dataset

2. Action recognition by pre-training on Kinetics

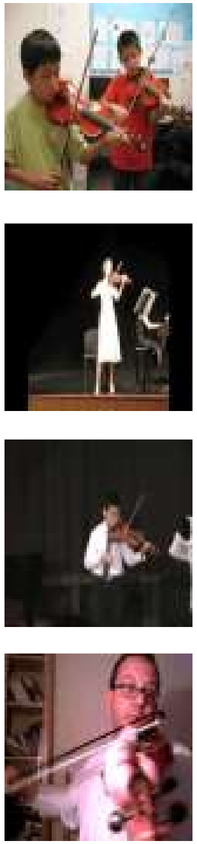
3. Where next in action recognition?

The Kinetics Human Action Video Dataset

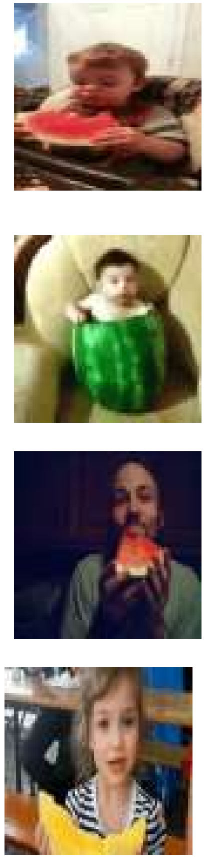
archery

country line dancing

riding or walking with horse



playing violin



eating watermelon

Motivation

Objective: A large scale human action classification video dataset

- An ImageNet for human action recognition
  - Trimmed videos
  - Actions performed by humans
  - Action classification
- Large enough to use for architecture design and comparison
- Large enough to pre-train networks for other tasks, e.g.
  - Temporal action localization in untrimmed videos

Kinetics overview

- Stats:

	Year	Actions	Clips per class	Total
Kinetics-400	2017	400	400-1000	300k
Kinetics-600	2018	600	600-1000	500k

- 10s clips
- Every clip is from a different YouTube video
  - For each action, huge variety in people, viewpoint, execution …

- *The Kinetics Human Action Video Dataset*. Kay, Carreira, Simonyan, Zhang, Hillier, Vijayanarasimhan, Viola, Green, Back, Natsev, Suleyman and Zisserman, arXiv 2017
- *A Short Note about Kinetics-600*, Carreira, Noland, Banki-Horvath, Hillier, Zisserman, arXiv 2018

Action Classes

Person Actions (Singular)

e.g. waving, blinking, running, jumping

Person-Person Actions

e.g. hugging, kissing, shaking hands

Person-Object Actions

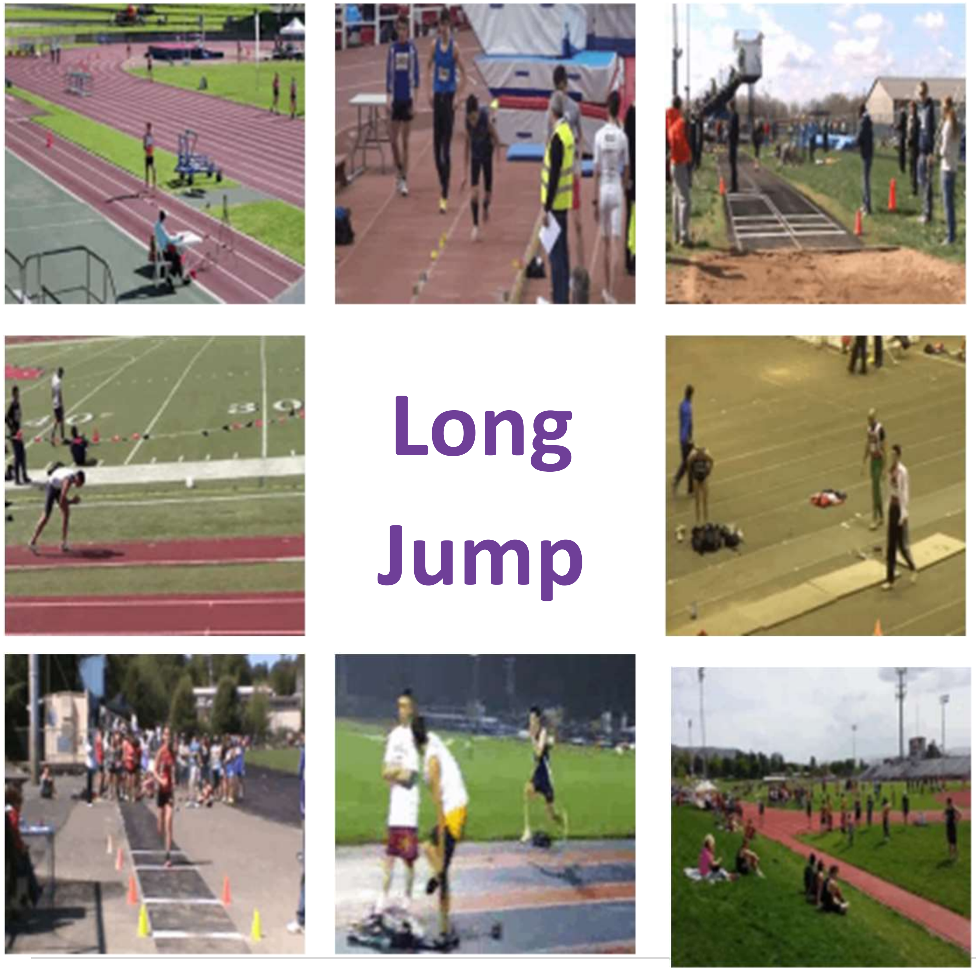
e.g. opening door, mowing lawn, washing dishes

Person Actions (Singular)

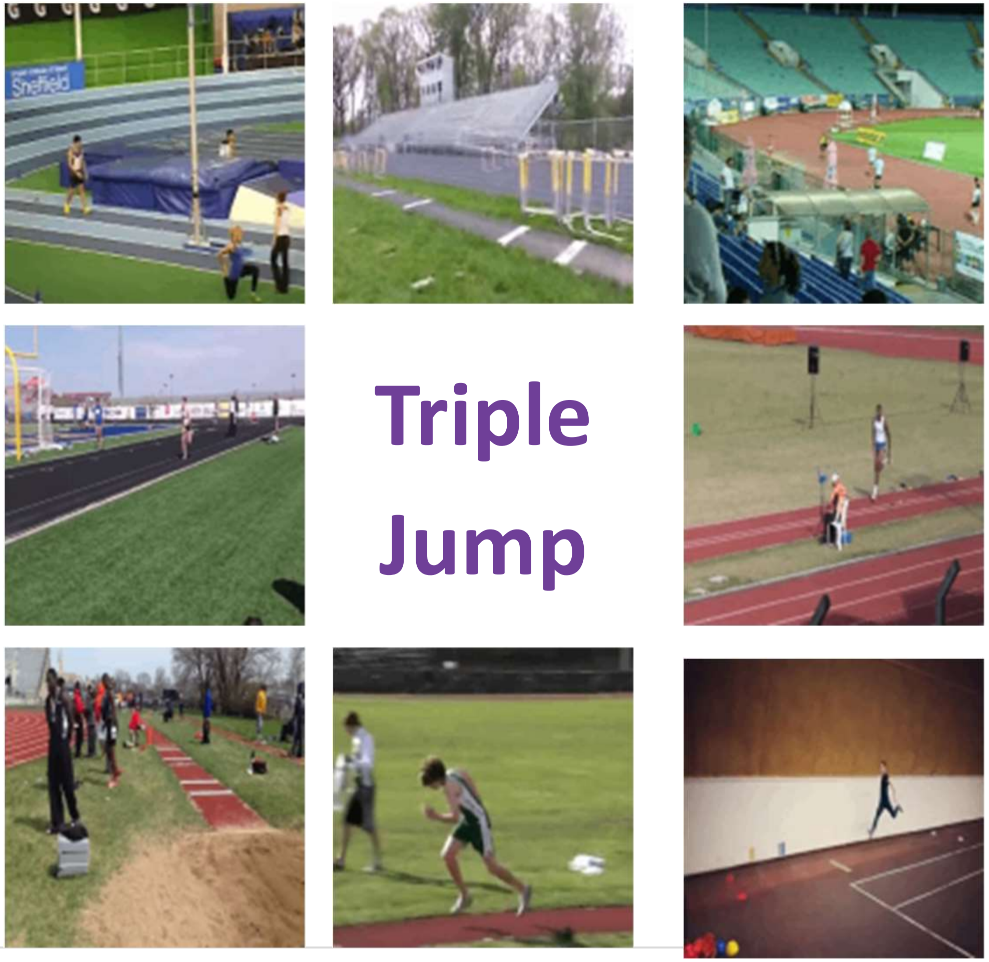
Pumping Fist

Shaking Head

Person Actions (Singular)



Long Jump



Triple Jump

Person-Person Actions

Shaking Hands

Making People Feel Welcome on University of Michigan's North Campus

Massaging Back

Person-Object Actions

Playing Violin

Playing Trumpet

Person-Object Actions

Folding Clothes

Folding Napkin

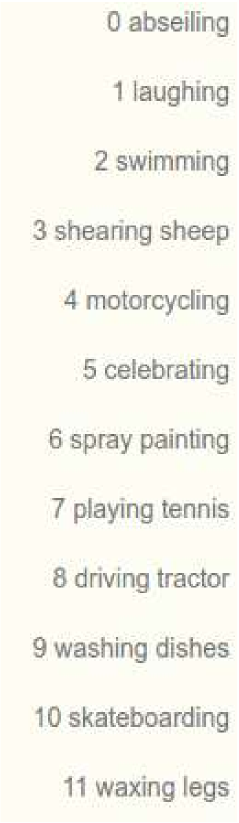
Person-Object Actions

Planting Flowers

Arranging Flowers

Dataset Collection Pipeline

Class list



0 abseiling

1 laughing

2 swimming

3 shearing sheep

4 motorcycling

5 celebrating

6 spray painting

7 playing tennis

8 driving tractor

9 washing dishes

10 skateboarding

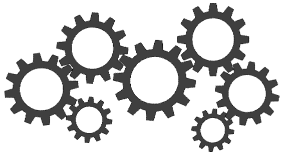
11 waxing legs

YouTube querying

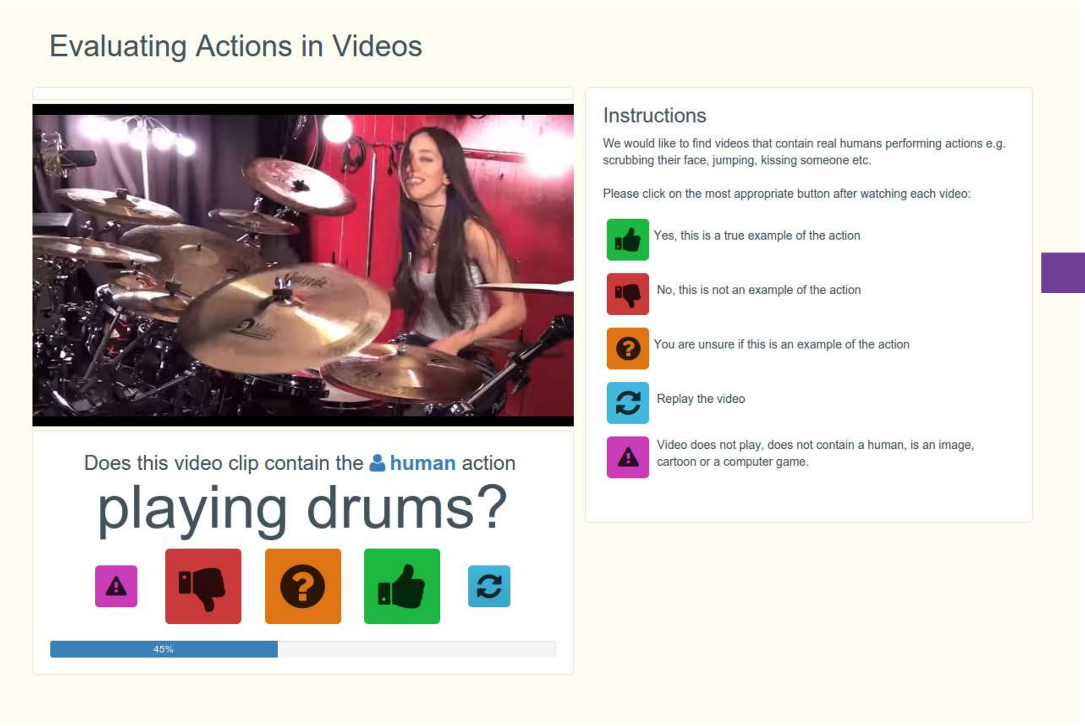
“Playing drums”

→

Image Classifiers



Human verification using **Mechanical Turk**



Evaluating Actions in Videos

Does this video clip contain the 👤 **human** action

playing drums?

⚠ 👎 ? 👍 ↻

45%

Instructions

We would like to find videos that contain real humans performing actions e.g. scrubbing their face, jumping, kissing someone etc.

Please click on the most appropriate button after watching each video:

- 👍 Yes, this is a true example of the action
- 👎 No, this is not an example of the action
- ? You are unsure if this is an example of the action
- ↻ Replay the video
- ⚠ Video does not play, does not contain a human, is an image, cartoon or a computer game.

→

Combine, split, and filter classes

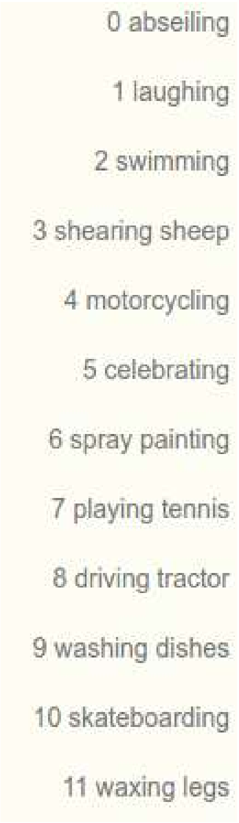
Scaling up from 400x400 to 600x600

- Finding candidate videos
  - Kinetics-400: text query for class name
  - Kinetics-600: decouple class and query text, add concept of language

- e.g: "folding paper" now matches against
  - "folding paper" (en)
  - "origami" (en)
  - "dobrar papel" (pt)

Dataset Collection Pipeline

Class list



0 abseiling

1 laughing

2 swimming

3 shearing sheep

4 motorcycling

5 celebrating

6 spray painting

7 playing tennis

8 driving tractor

9 washing dishes

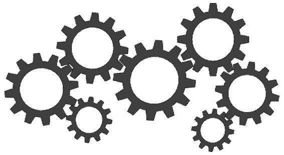
10 skateboarding

11 waxing legs

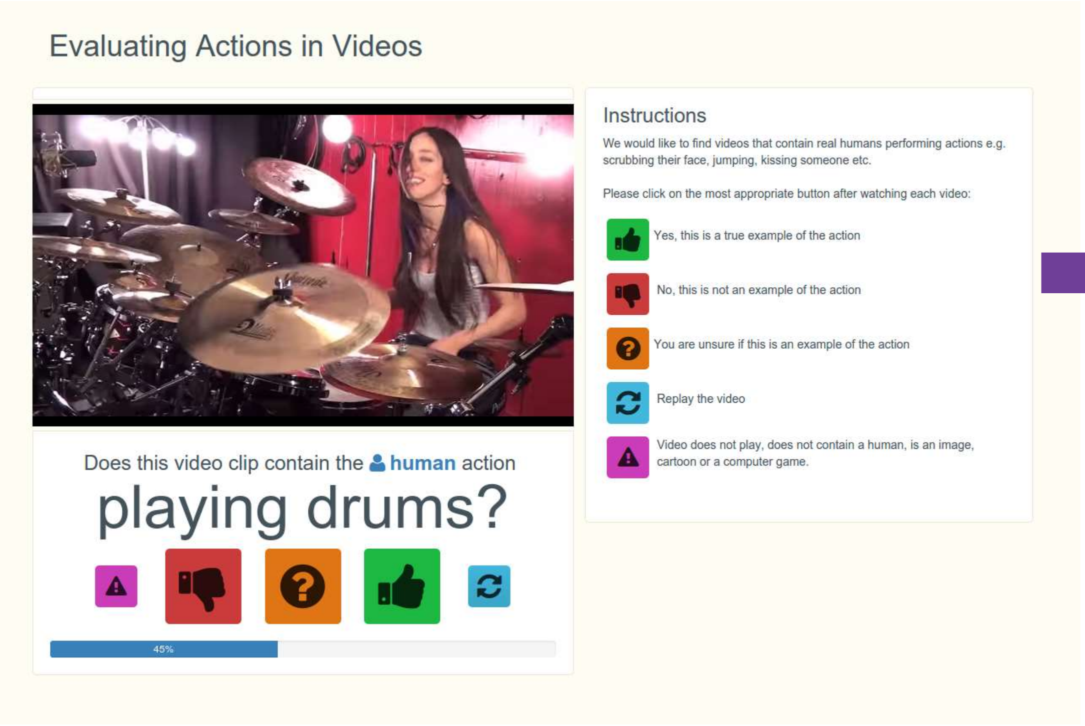
YouTube querying

“Drumming”
“Playing drums”
“Tocar bateria”

Image Classifiers



Human verification using **Mechanical Turk**



Evaluating Actions in Videos

Does this video clip contain the **human** action

playing drums?

Instructions

We would like to find videos that contain real humans performing actions e.g. scrubbing their face, jumping, kissing someone etc.

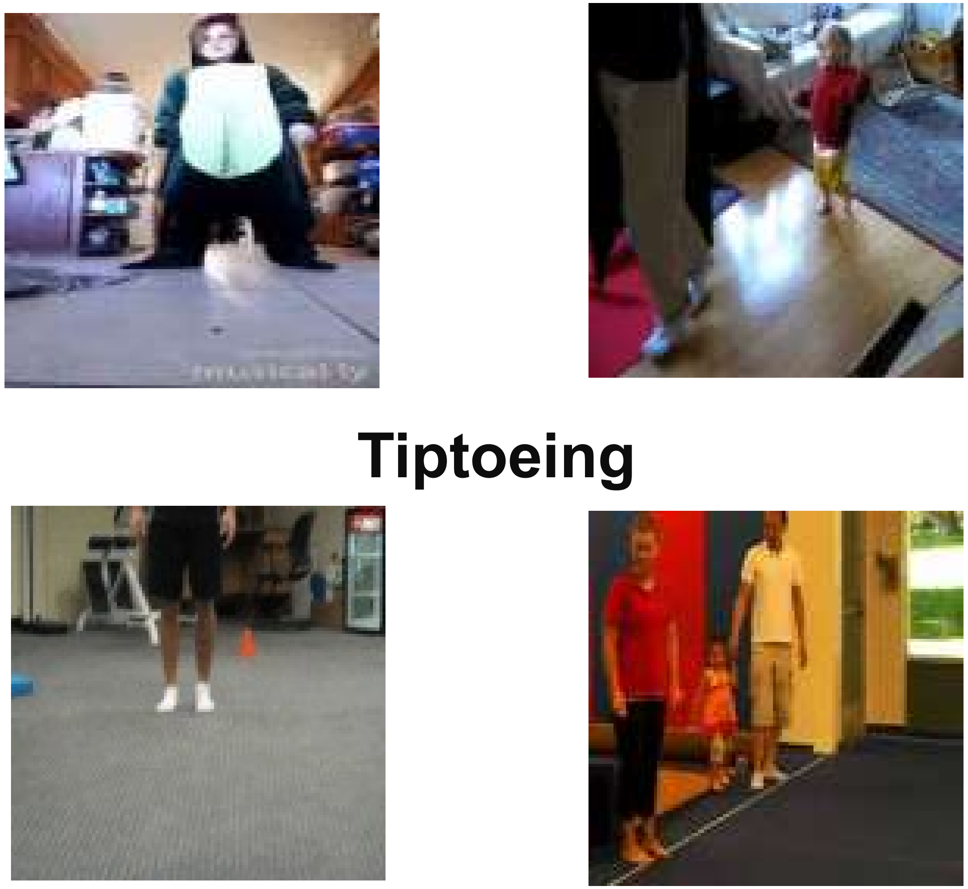
Please click on the most appropriate button after watching each video:

- 👍 Yes, this is a true example of the action
- 👎 No, this is not an example of the action
- ? You are unsure if this is an example of the action
- ↻ Replay the video
- ⚠ Video does not play, does not contain a human, is an image, cartoon or a computer game.

Combine, split, and filter classes

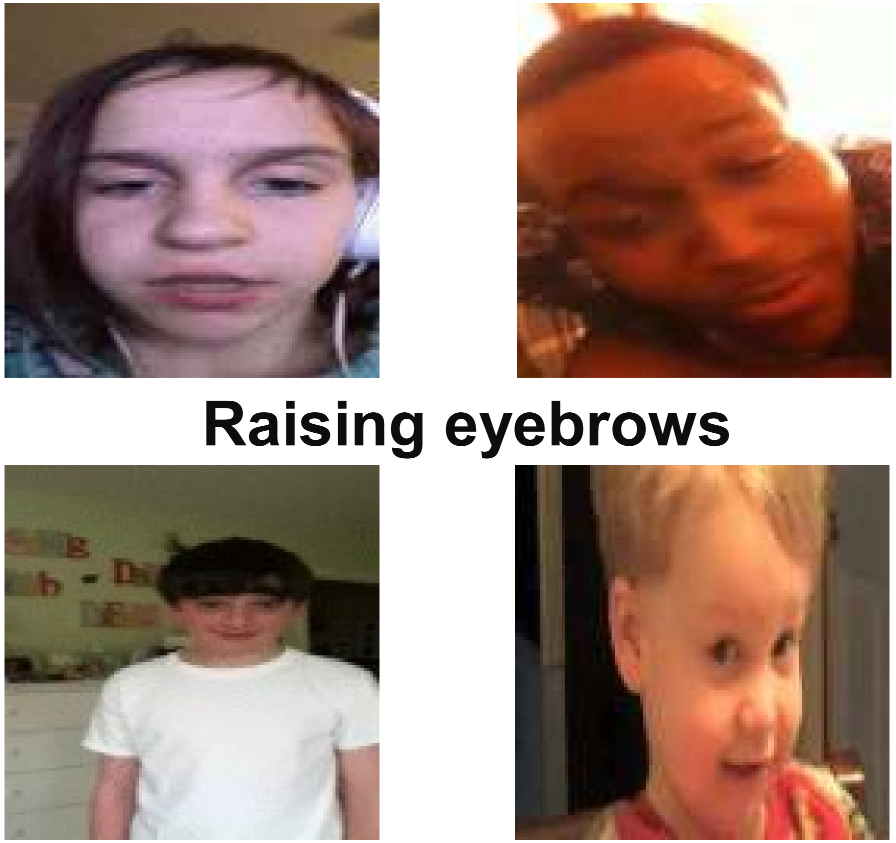
New in Kinetics-600: more body-only classes

Head stand

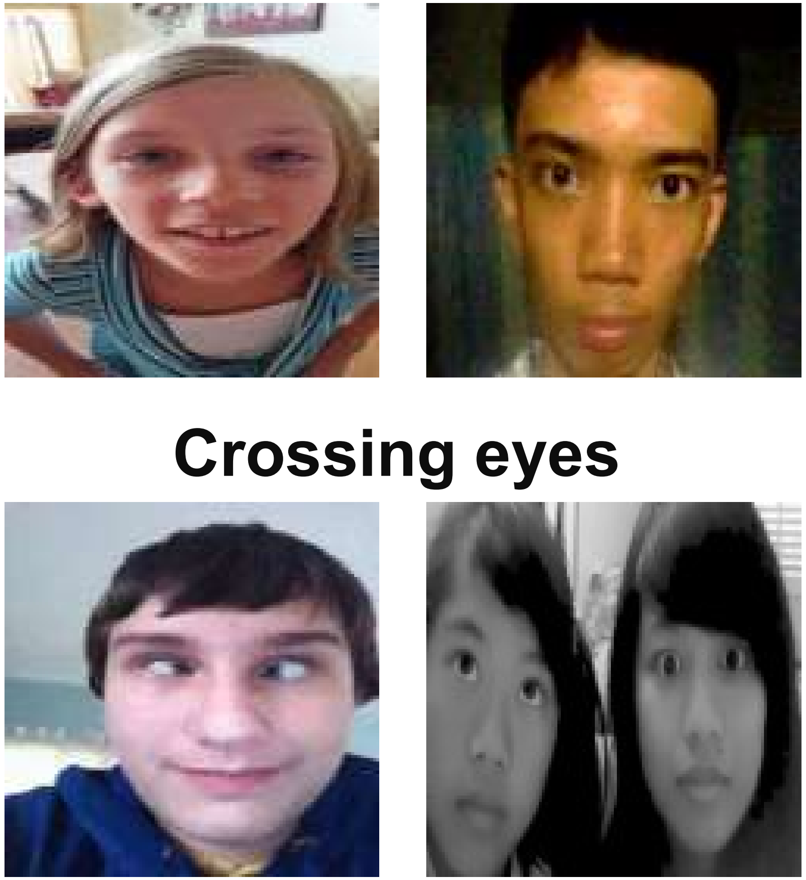


Tiptoeing

More face classes

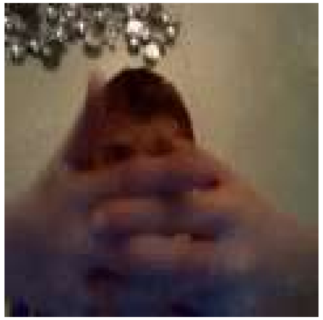
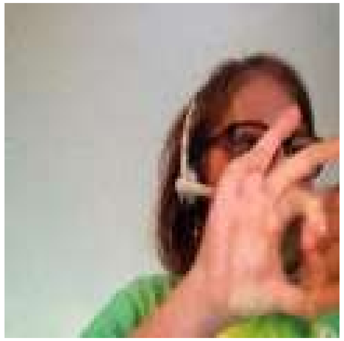


Raising eyebrows

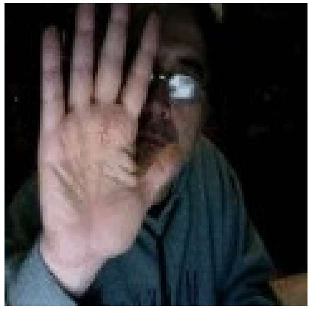
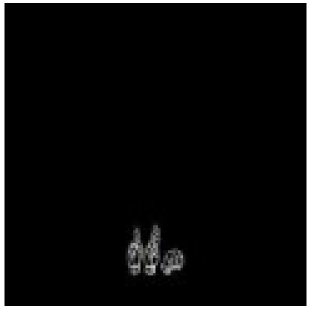


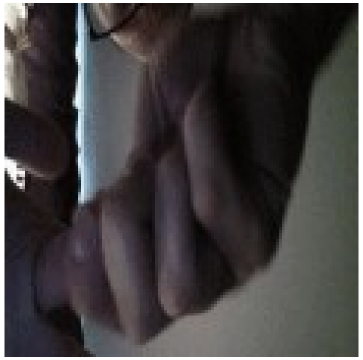
Crossing eyes

More hand classes

Twiddling fingers

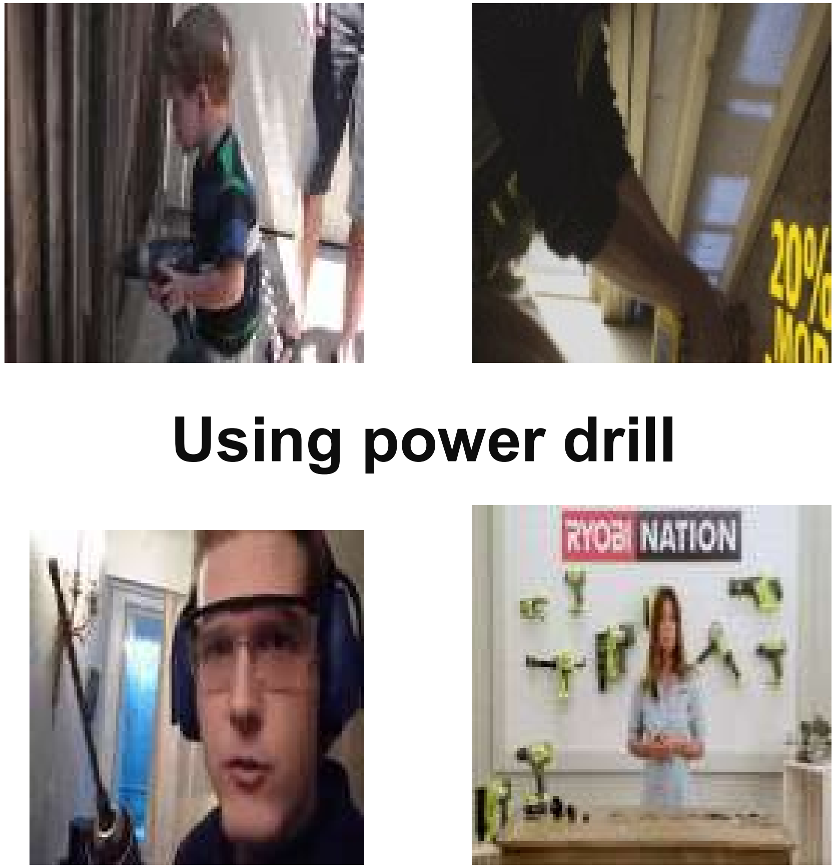





Cracking knuckles

More basic tool use

Using sledgehammer

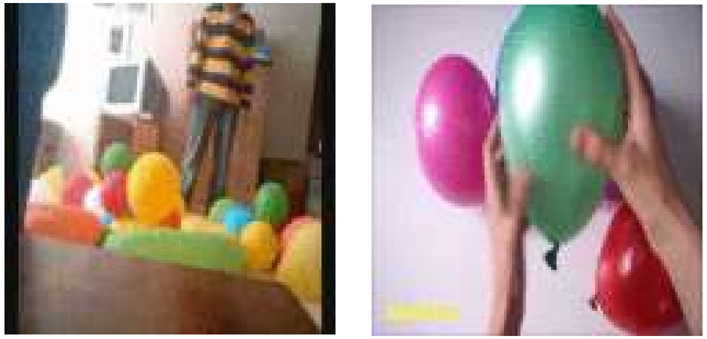


Using power drill

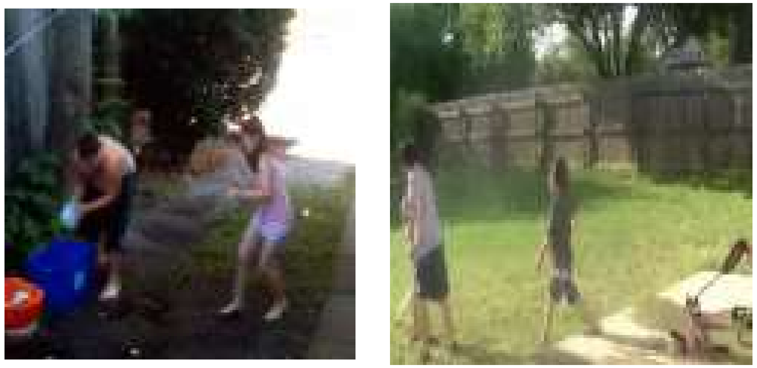
Also using paint roller, circular saw, wrench, others

More actions around similar objects

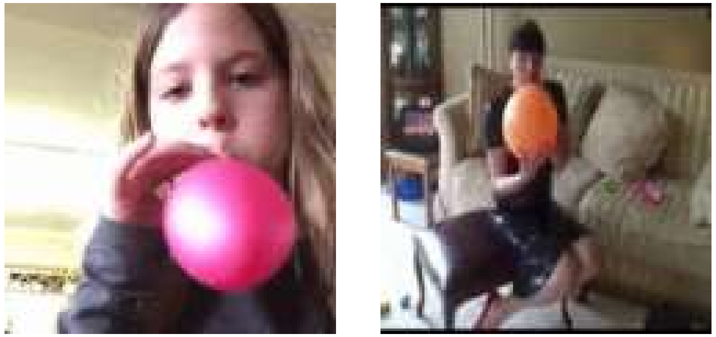
Popping balloons



Throwing water balloons



Inflating balloons



Making balloon shapes

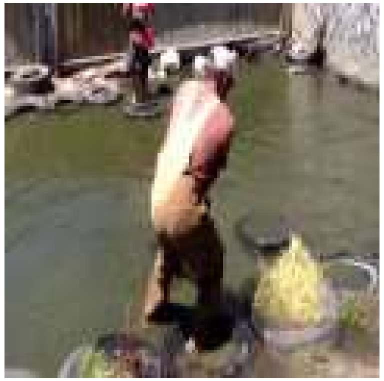
More dances

Mosh pit dancing

Square dancing

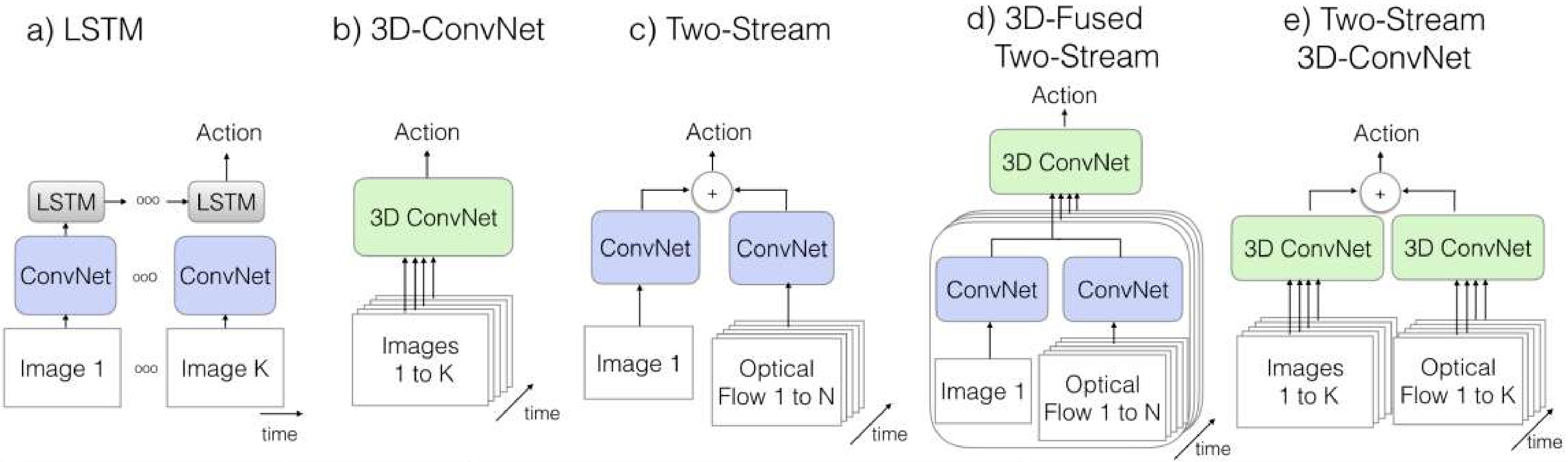
More random stuff many people do

Contact juggling

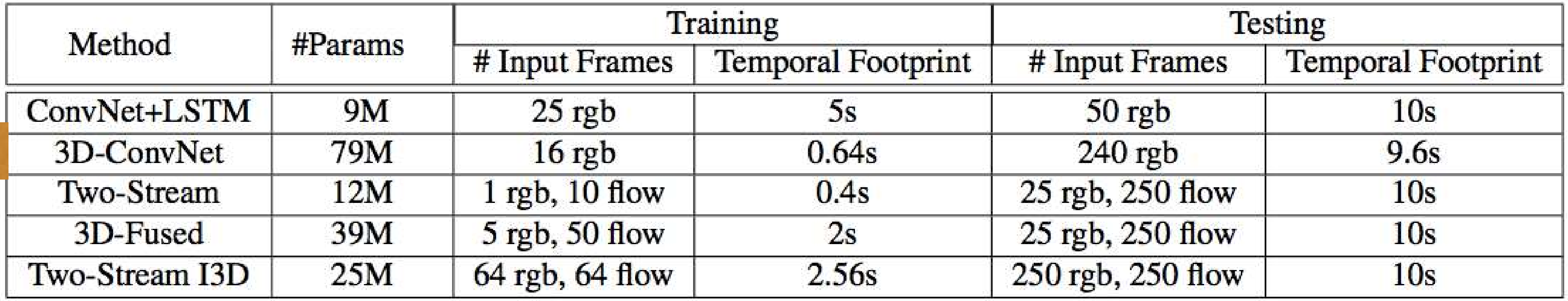


Alligator wrestling

Comparison of networks on Kinetics



(C3D)



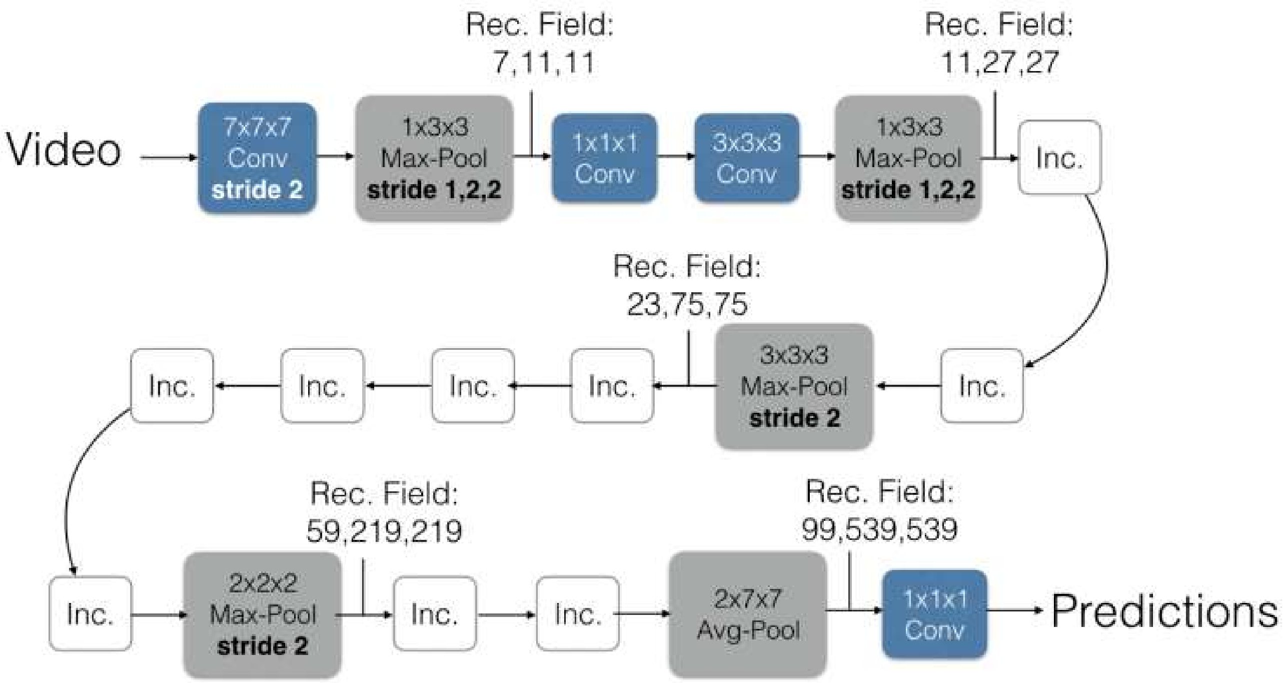
Method	#Params	Training		Testing	
		# Input Frames	Temporal Footprint	# Input Frames	Temporal Footprint
ConvNet+LSTM	9M	25 rgb	5s	50 rgb	10s
3D-ConvNet	79M	16 rgb	0.64s	240 rgb	9.6s
Two-Stream	12M	1 rgb, 10 flow	0.4s	25 rgb, 250 flow	10s
3D-Fused	39M	5 rgb, 50 flow	2s	25 rgb, 250 flow	10s
Two-Stream I3D	25M	64 rgb, 64 flow	2.56s	250 rgb, 250 flow	10s

Table 1. Number of parameters and temporal input sizes of the models.

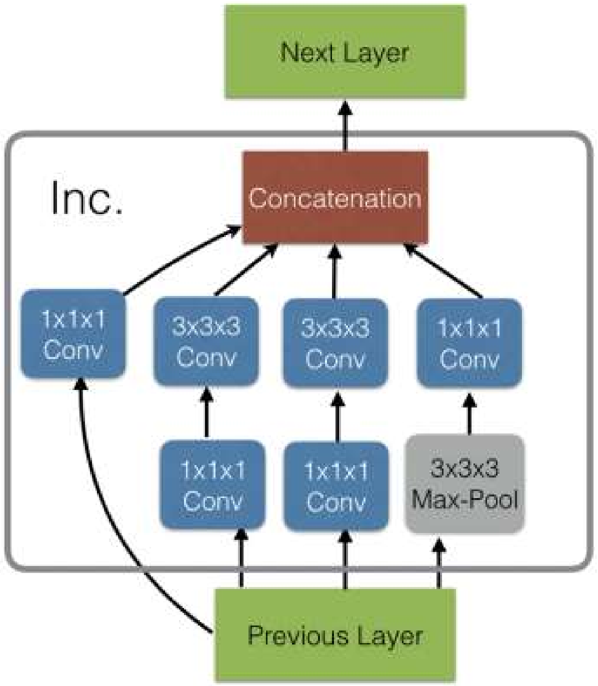
Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset
Joao Carreira, Andrew Zisserman, CVPR 17

Inflated 3D Inception (I3D)

Inflated Inception-V1

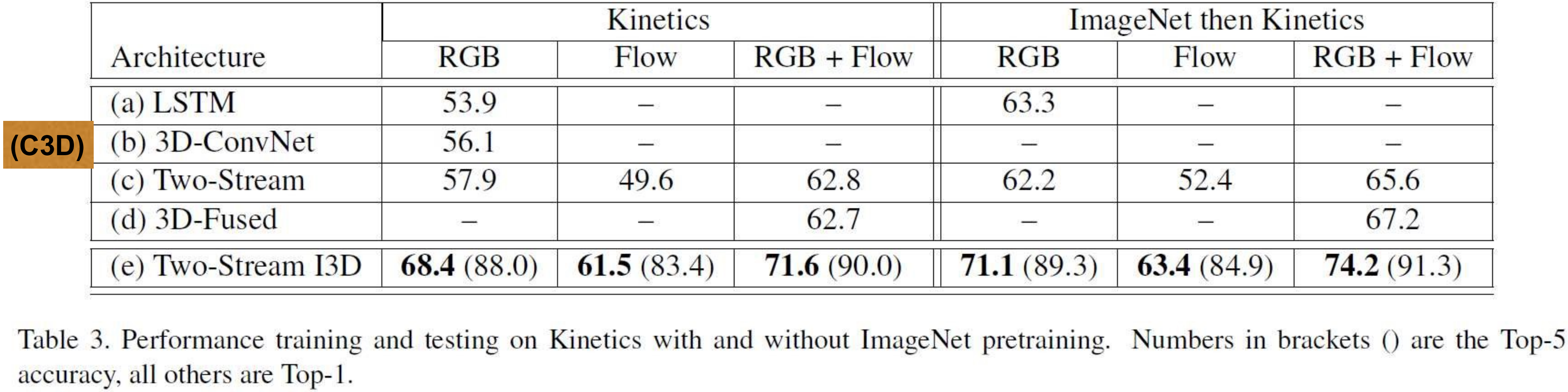


Inception Module (Inc.)



Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset
Joao Carreira, Andrew Zisserman, CVPR 17

Network comparison on Kinetics-400



(C3D)

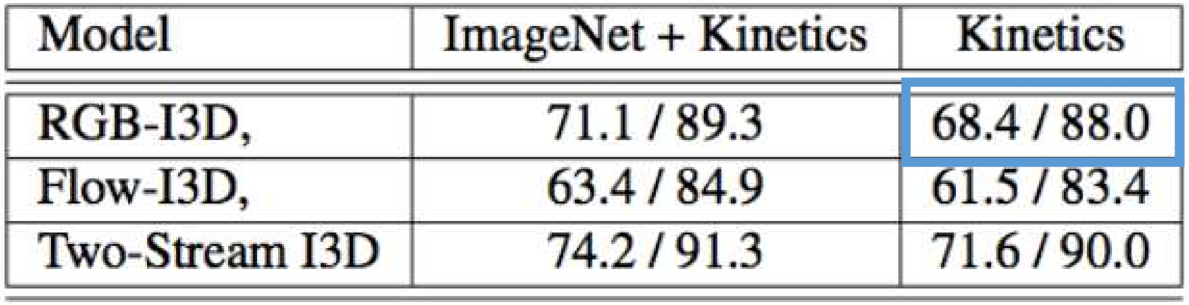
Architecture	Kinetics			ImageNet then Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	53.9	–	–	63.3	–	–
(b) 3D-ConvNet	56.1	–	–	–	–	–
(c) Two-Stream	57.9	49.6	62.8	62.2	52.4	65.6
(d) 3D-Fused	–	–	62.7	–	–	67.2
(e) Two-Stream I3D	**68.4** (88.0)	**61.5** (83.4)	**71.6** (90.0)	**71.1** (89.3)	**63.4** (84.9)	**74.2** (91.3)

Table 3. Performance training and testing on Kinetics with and without ImageNet pretraining. Numbers in brackets () are the Top-5 accuracy, all others are Top-1.

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset
Joao Carreira, Andrew Zisserman, CVPR 17

I3D comparison from Kinetics-400 to Kinetics-600

Kinetics-400



Model	ImageNet + Kinetics	Kinetics
RGB-I3D,	71.1 / 89.3	68.4 / 88.0
Flow-I3D,	63.4 / 84.9	61.5 / 83.4
Two-Stream I3D	74.2 / 91.3	71.6 / 90.0

Kinetics-600, RGB-I3D, training/testing on Kinetics-600 72.0 / 91.0

A Short Note about Kinetics-600
Authors: Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, Andrew Zisserman, arXiv 2018

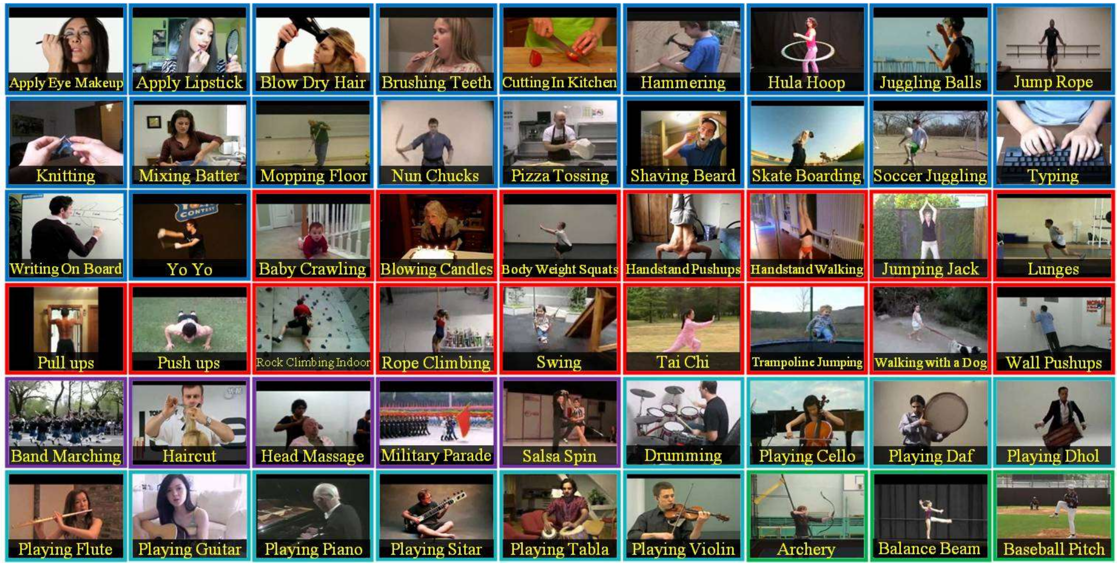
Part II

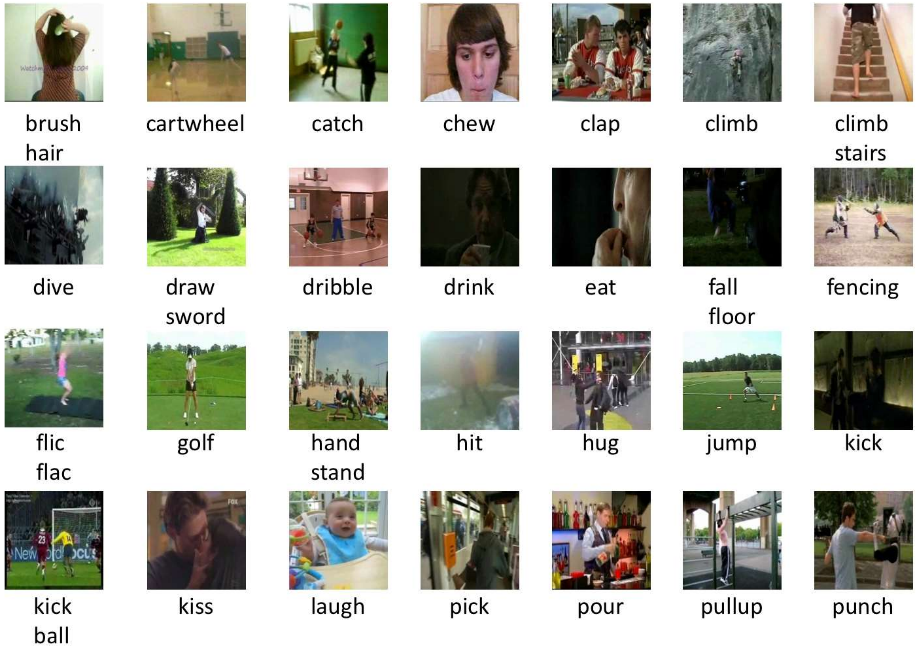
Action recognition by pre-training on Kinetics

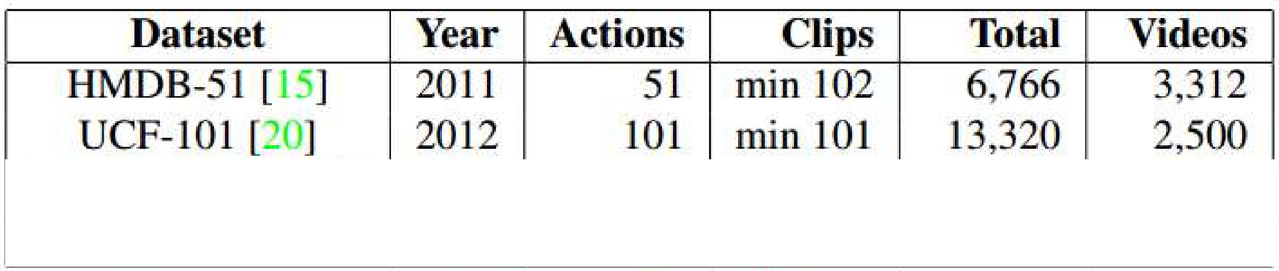
Performance on four datasets:

1. UCF-101 – classification
2. HMD-51 – classification
3. Charades – temporal localization
4. AVA – spatio-temporal localization

UCF-101 and HMDB-51



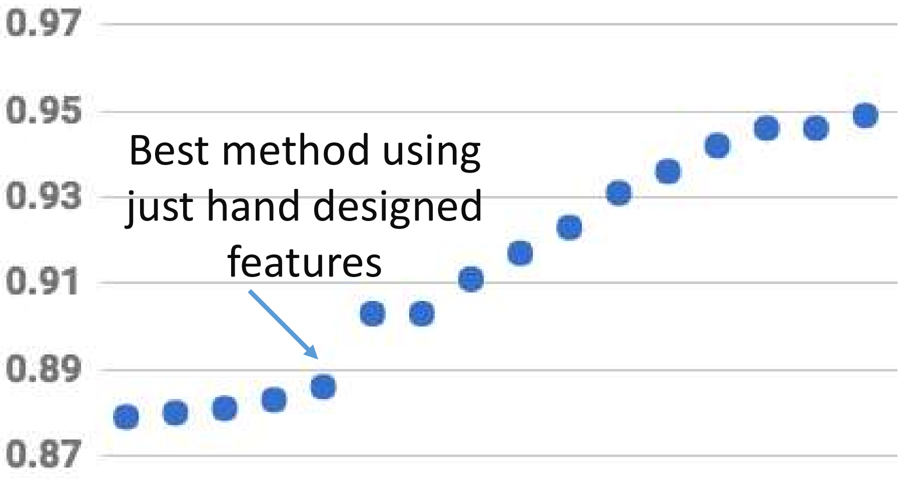




Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500

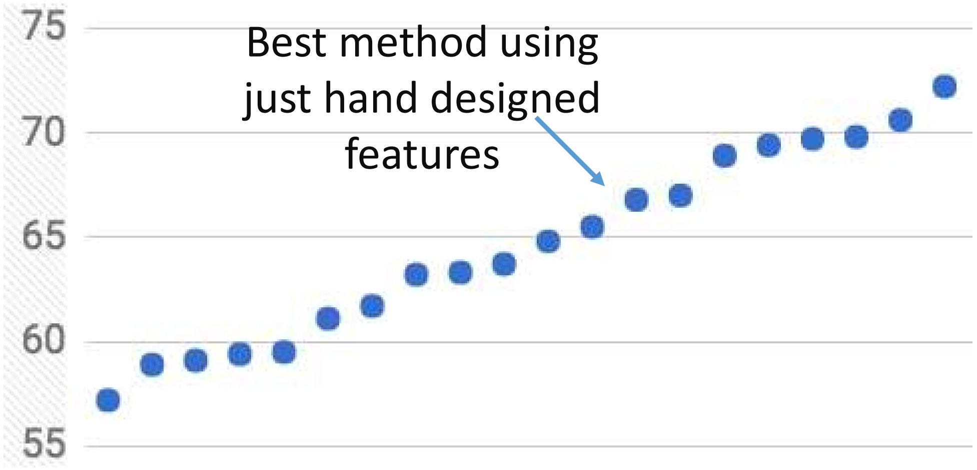
Transferring from ImageNet to Video

UCF-101



Best method using just hand designed features

HMDB-51

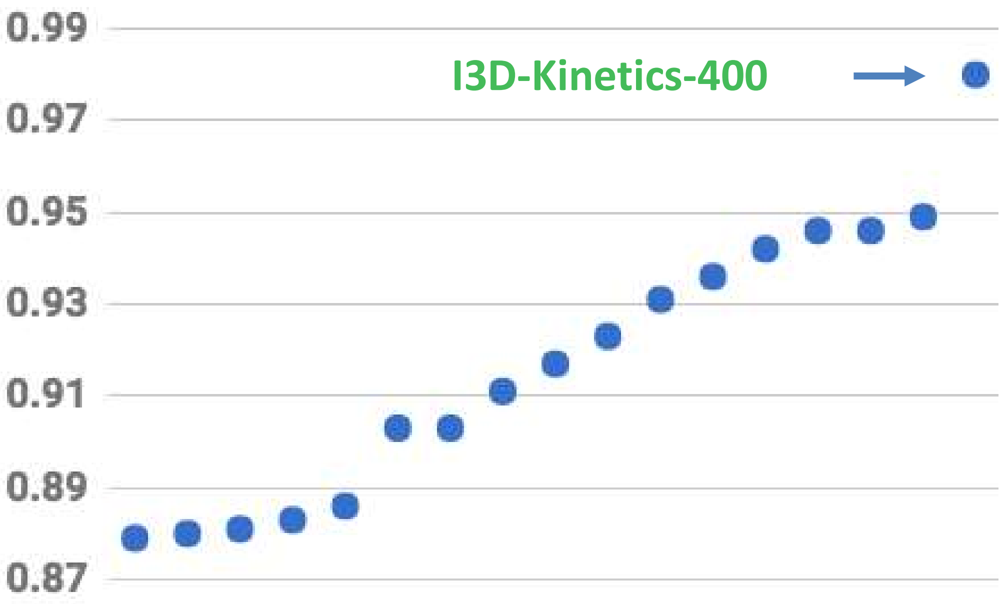


Best method using just hand designed features

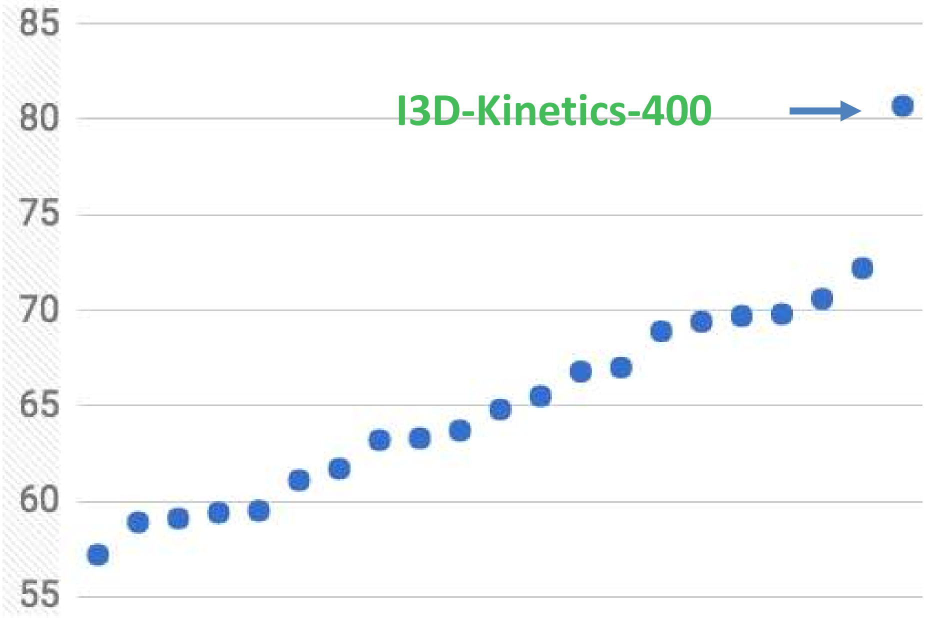
Compilation of results from actionrecognition.net

I3D-Kinetics-400 transfer performance (two stream, flow+RGB)

UCF-101



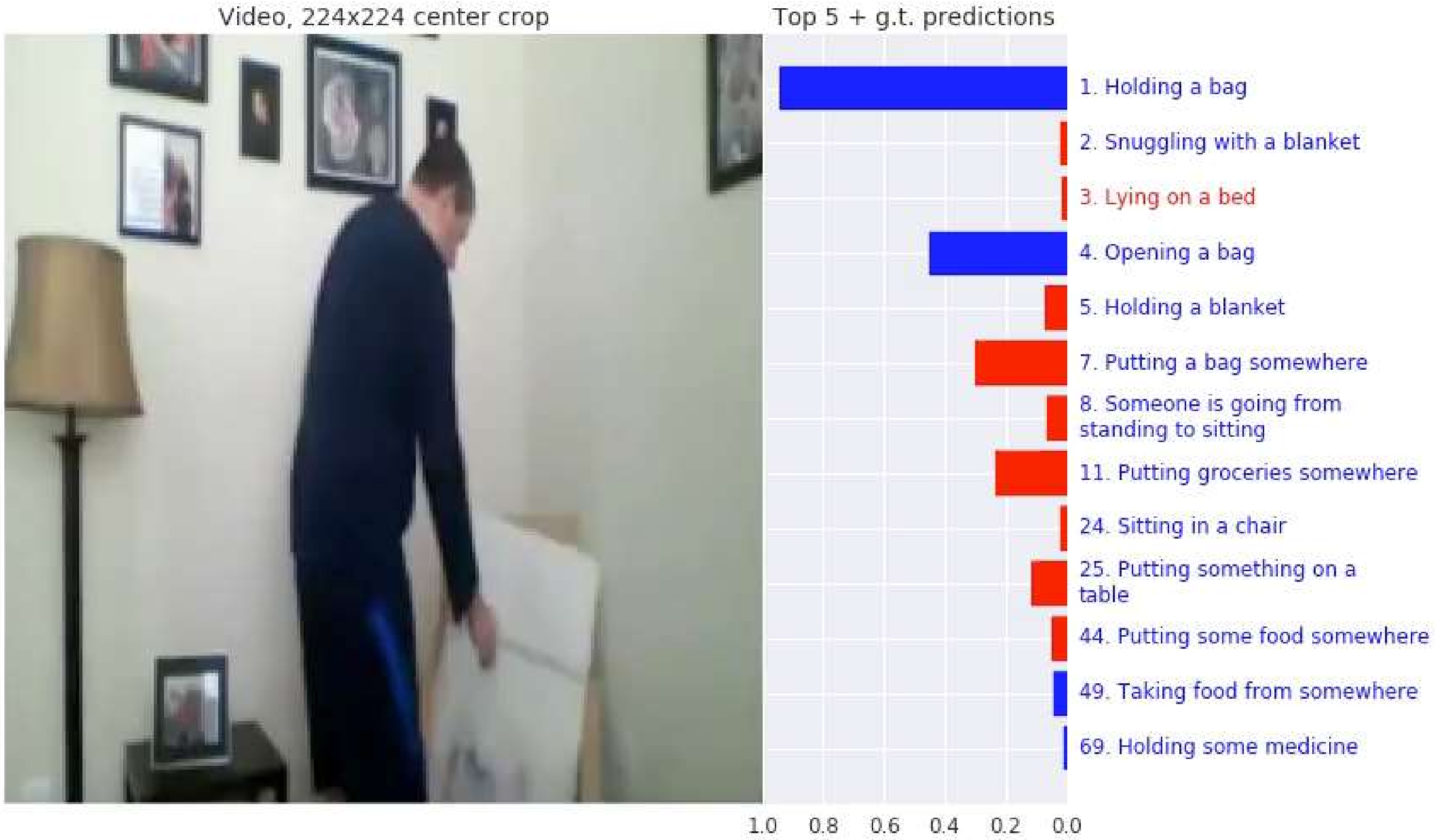
HMDB-51



Compilation of results from actionrecognition.net

Charades dataset - action localization

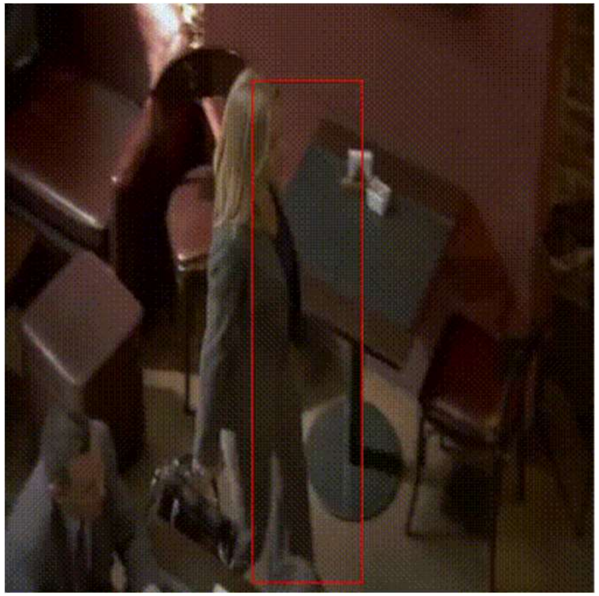
- I3D model with Kinetics-400 pre-training defined the state of the art
- Winner of the CVPR 2017 Charades challenge



Atomic Visual Actions (AVA) Dataset

- Person-centric actions
- Multiple people, multiple action labels
- Atomic actions
- Exhaustivity
- Action transitions over time
- Realistic scenes and diverse environment

Carry/Hold (an object);
Walk



AVA: A video dataset of spatio-temporally localized atomic visual actions, C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, CVPR 2018.

80 Atomic Actions in AVA

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

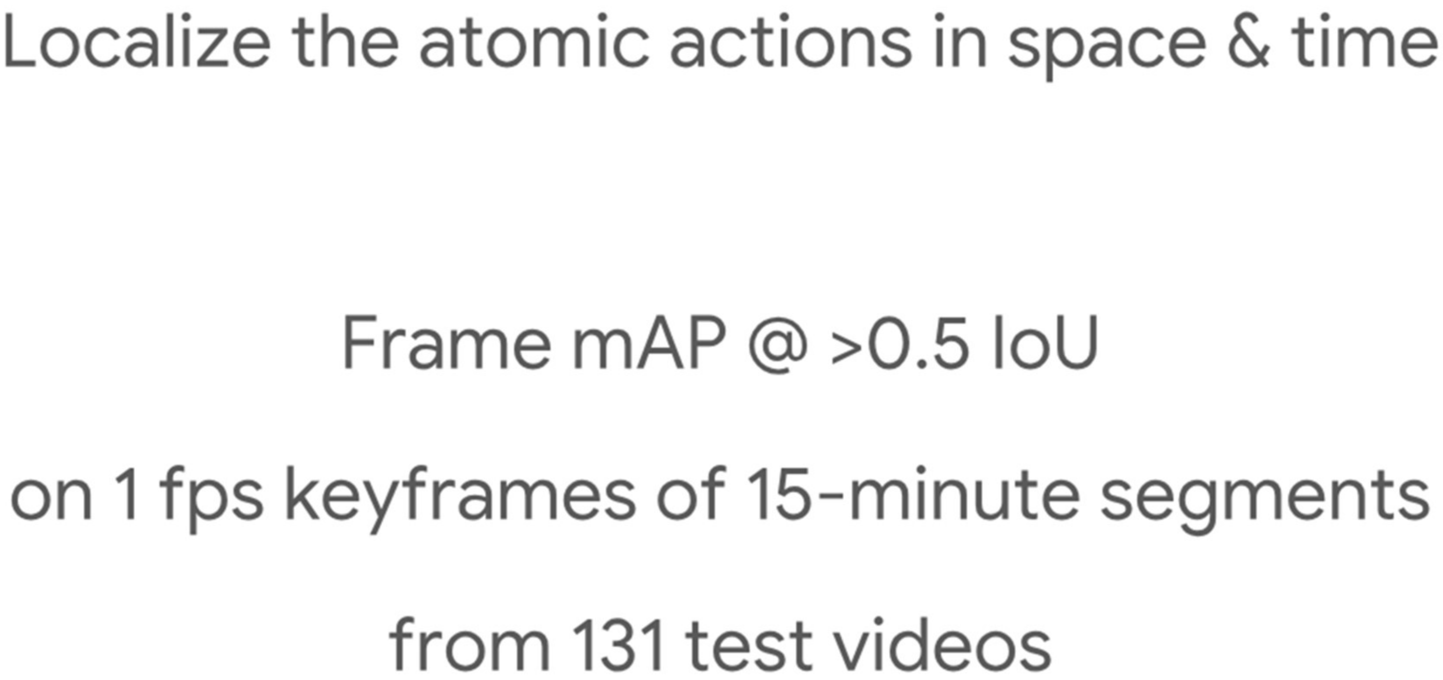
smoke
sail boat
row boat
fishing
touch
cook
kick
paint
dig
shovel
chop
shoot
take a photo
brush teeth
clink glass

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)

AVA Challenge 2018



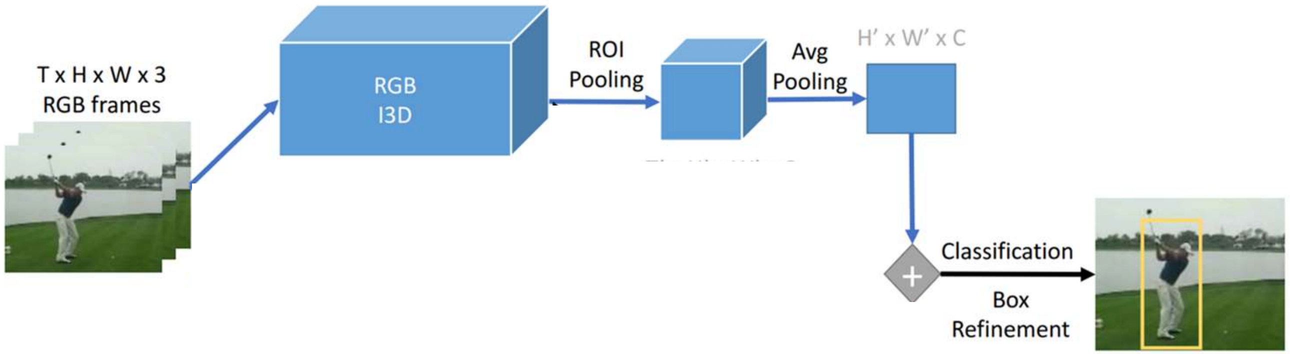
Localize the atomic actions in space & time

Frame mAP @ >0.5 IoU

on 1 fps keyframes of 15-minute segments

from 131 test videos

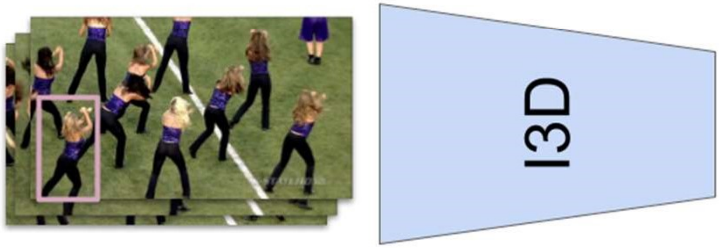
Model overview



A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018

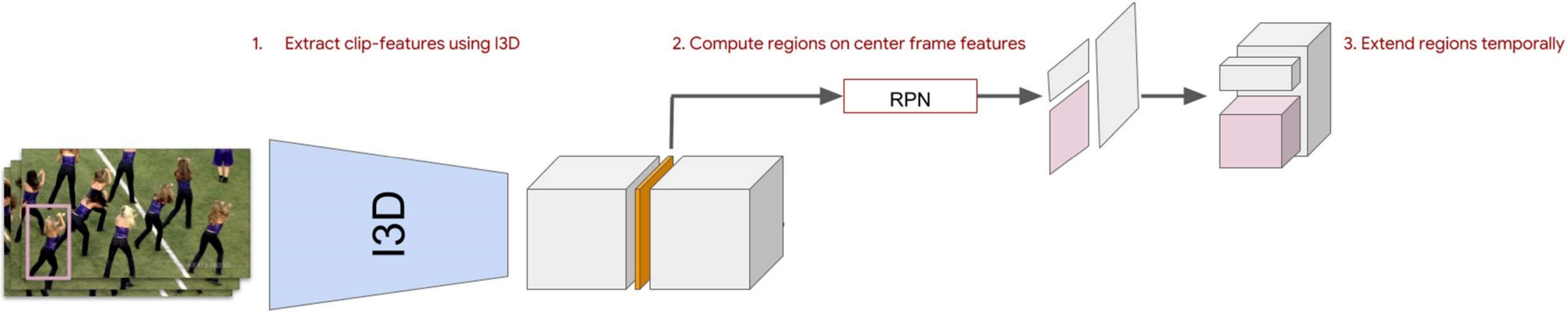
Network architecture

1. Extract clip-features using I3D



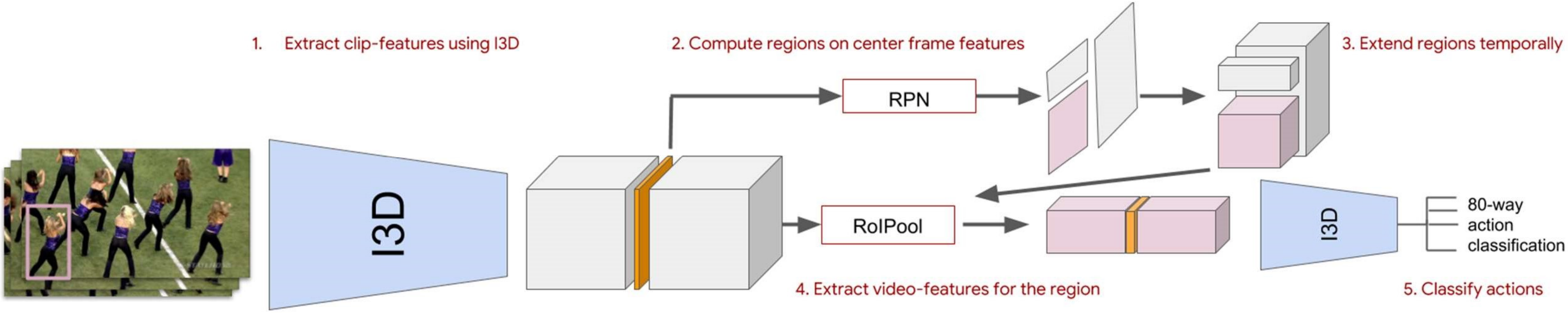
A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018

Network architecture


1. Extract clip-features using I3D
2. Compute regions on center frame features
3. Extend regions temporally
I3D
RPN

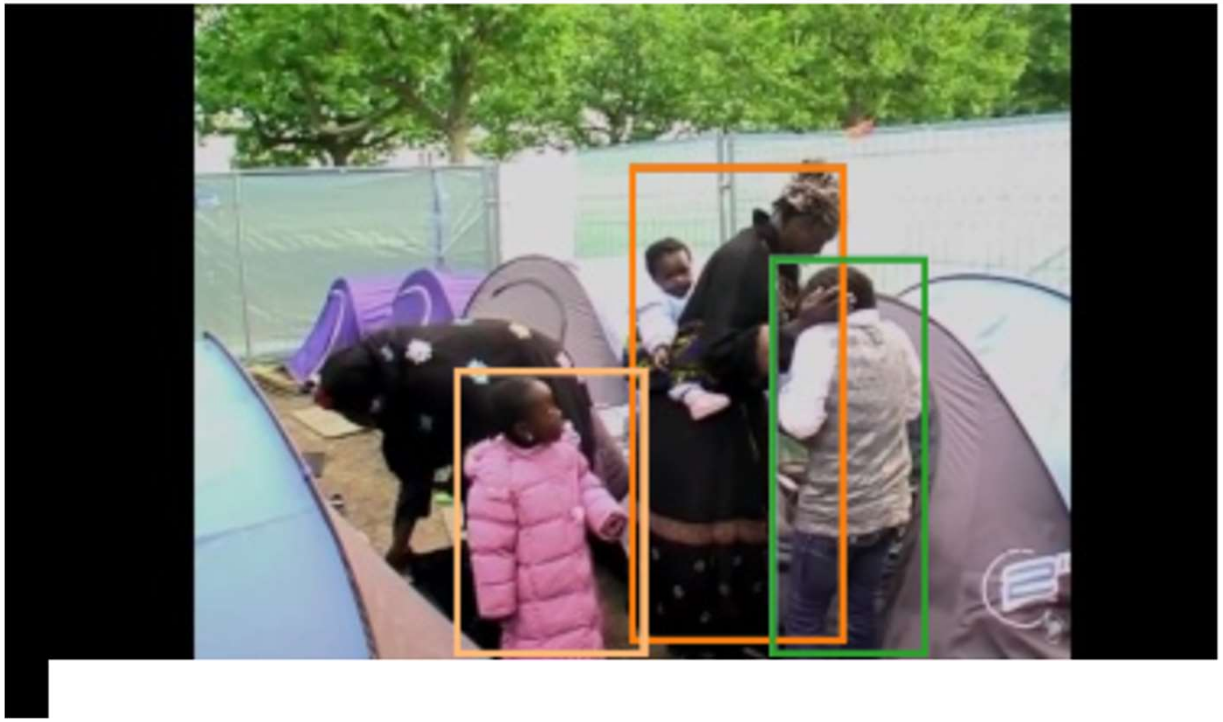
A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018

Network architecture


1. Extract clip-features using I3D
2. Compute regions on center frame features
3. Extend regions temporally
I3D
RPN
RoIPool
4. Extract video-features for the region
I3D
80-way action classification
5. Classify actions

A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018

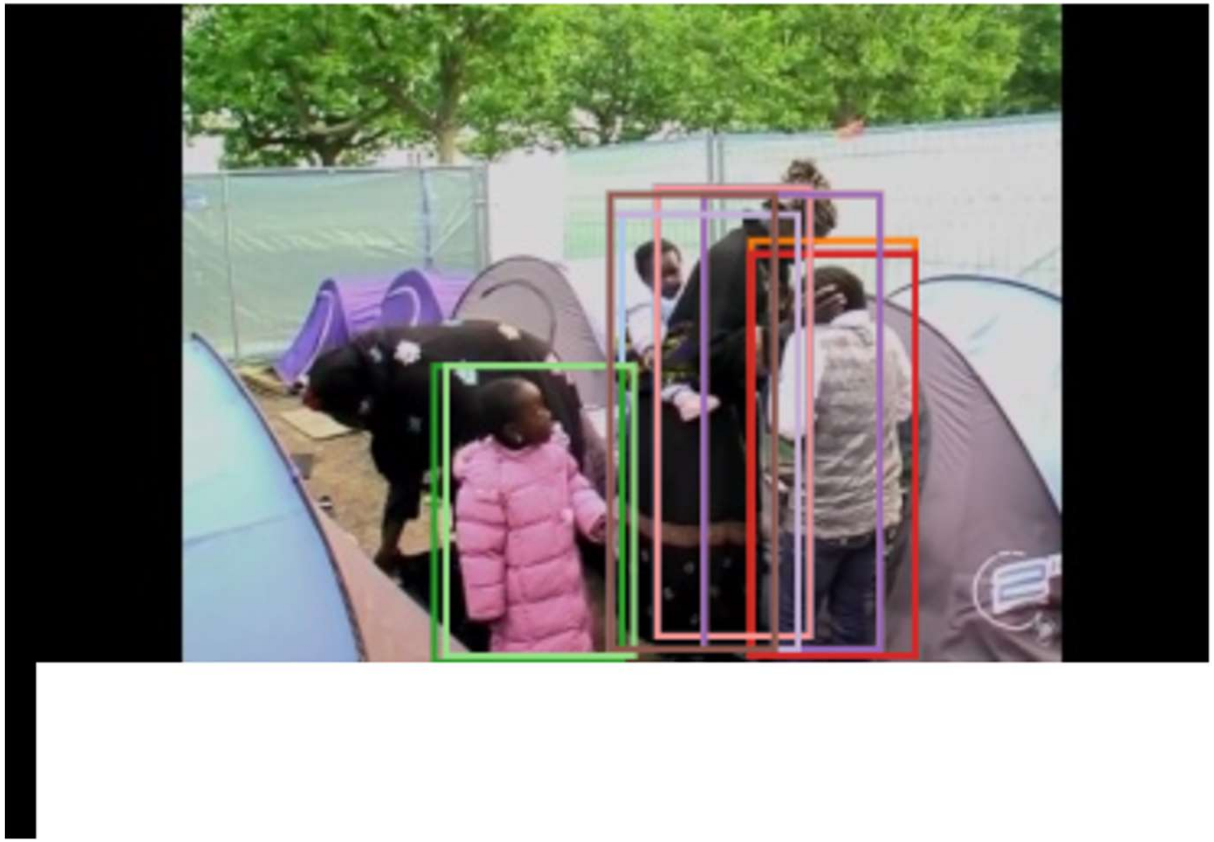
Groundtruth



- watch (a person)(50,68,25,97)
- listen to (a person)(62,75,39,99)
- watch (a person)(62,75,39,99)
- grab (a person)(50,68,25,97)
- bend/bow (at the waist)(50,68,25,97)
- watch (a person)(35,51,56,99)
- listen to (a person)(35,51,56,99)
- stand(35,51,56,99)
- stand(62,75,39,99)

A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018

Predictions

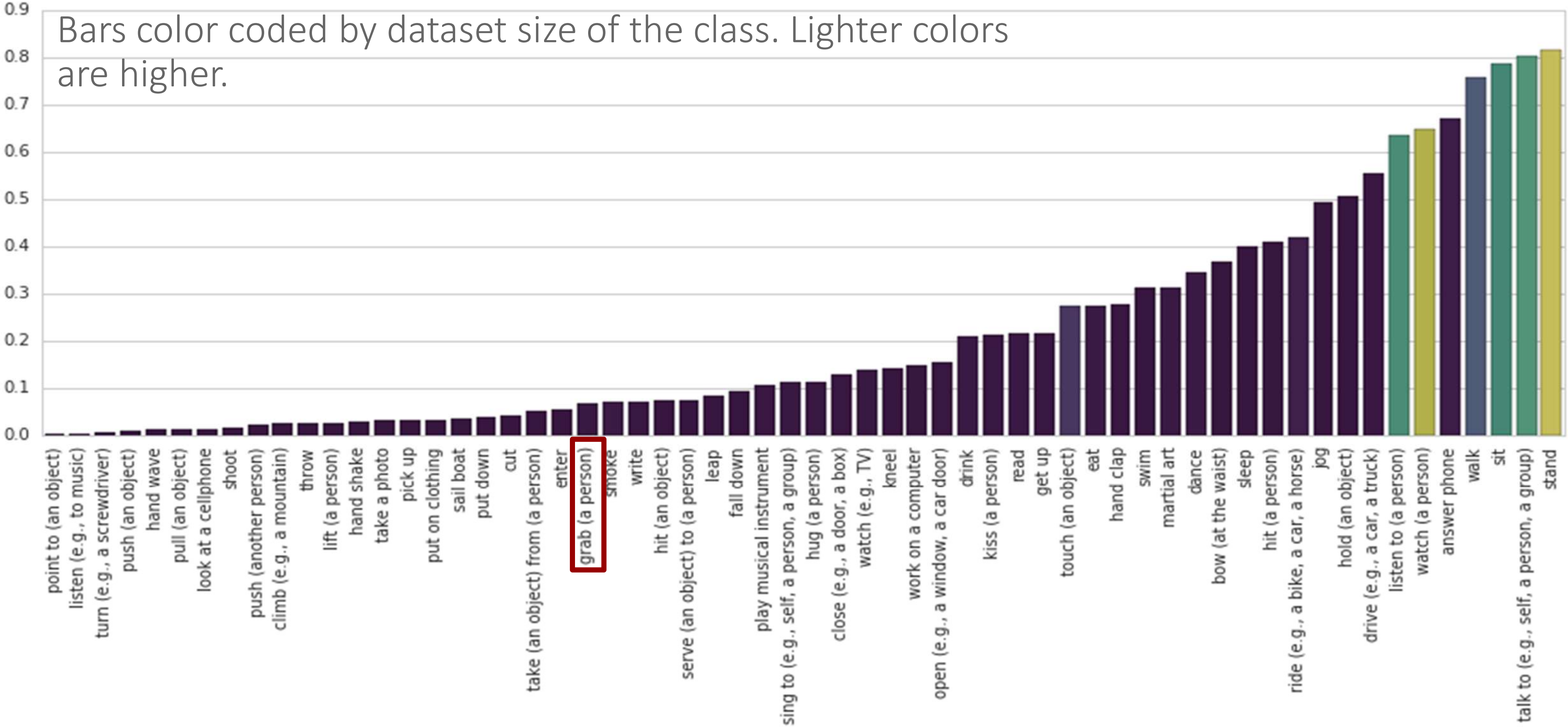


- watch (a person)(53,66,28,96)
- watch (a person)(50,63,32,98)
- listen to (a person)(61,75,36,99)
- watch (a person)(57,72,29,98)
- sit(34,50,55,100)
- stand(35,51,55,99)
- watch (a person)(61,75,38,99)
- stand(61,75,38,99)
- stand(53,66,28,96)
- stand(57,72,29,98)
- carry/hold (an object)(49,65,32,98)
- stand(49,63,29,98)

Test set mAP = 21%

A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018

Easiest and Hardest Classes



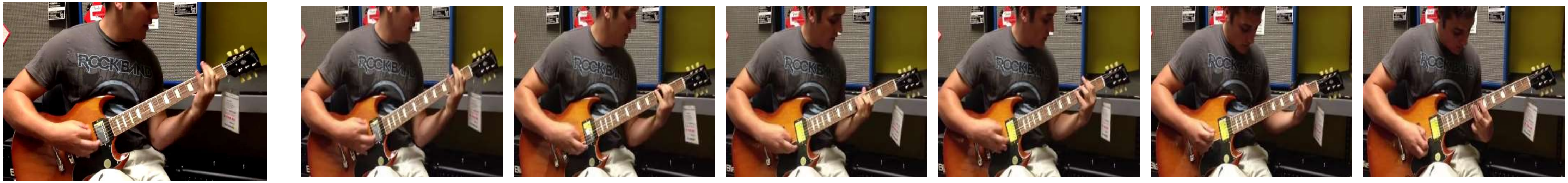
Bars color coded by dataset size of the class. Lighter colors are higher.

Part III

Where next in action recognition?

Video

A temporal sequence of frames

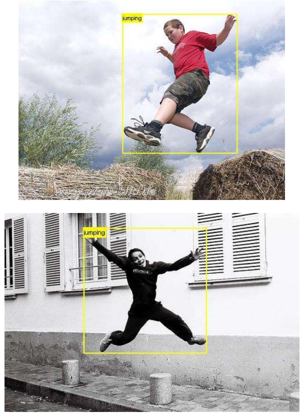


What is required to recognize the action?

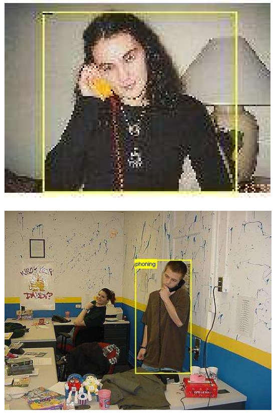
- a single frame?
- a bag of frames (unordered)?
- an ordered sequence of frames?
- …

Action Classification on Static Frames

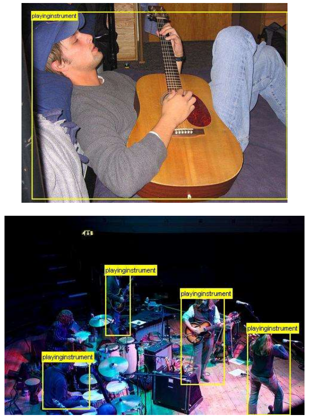
Jumping



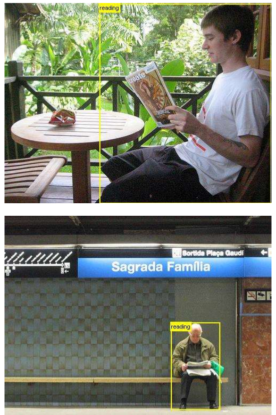
Phoning



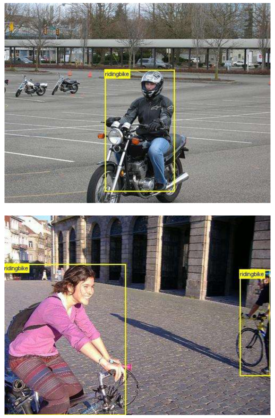
Playing Instrument



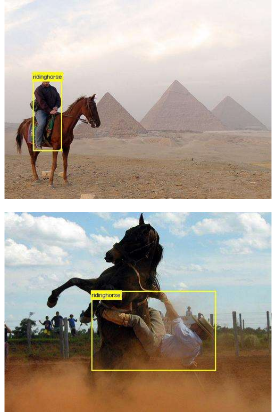
Reading



Riding Bike

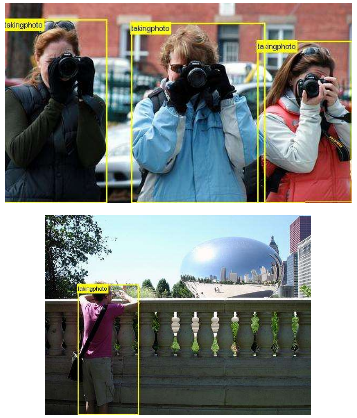


Riding Horse

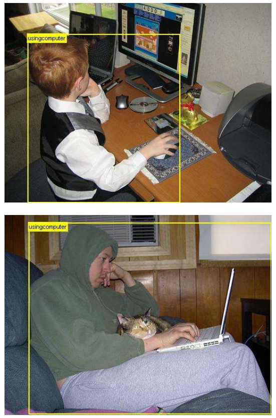


Running

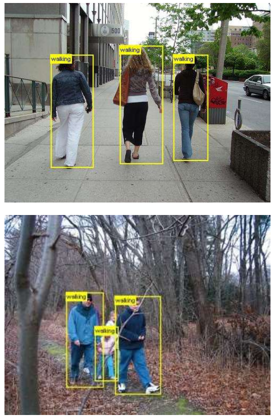
Taking Photo



Using Computer



Walking



PASCAL VOC Action Classification Challenge

Some actions require motion for classification

- Sitting down/standing up; closing/opening something
- Different dance styles ….

Some actions require motion for classification

- Sitting down/standing up; closing/opening something
- Different dance styles ….

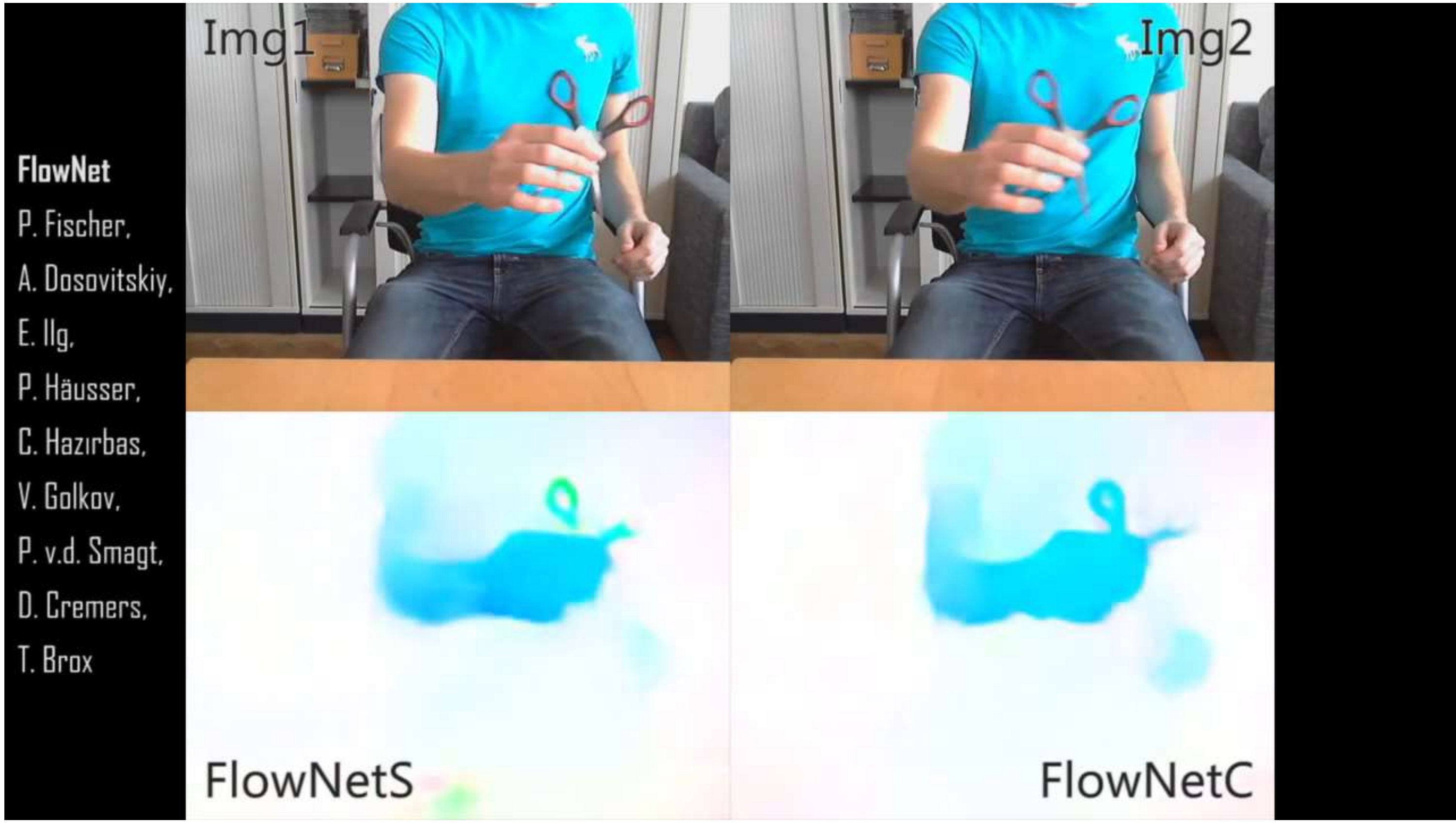
Dancing Macarena

Dancing Charleston

Zumba

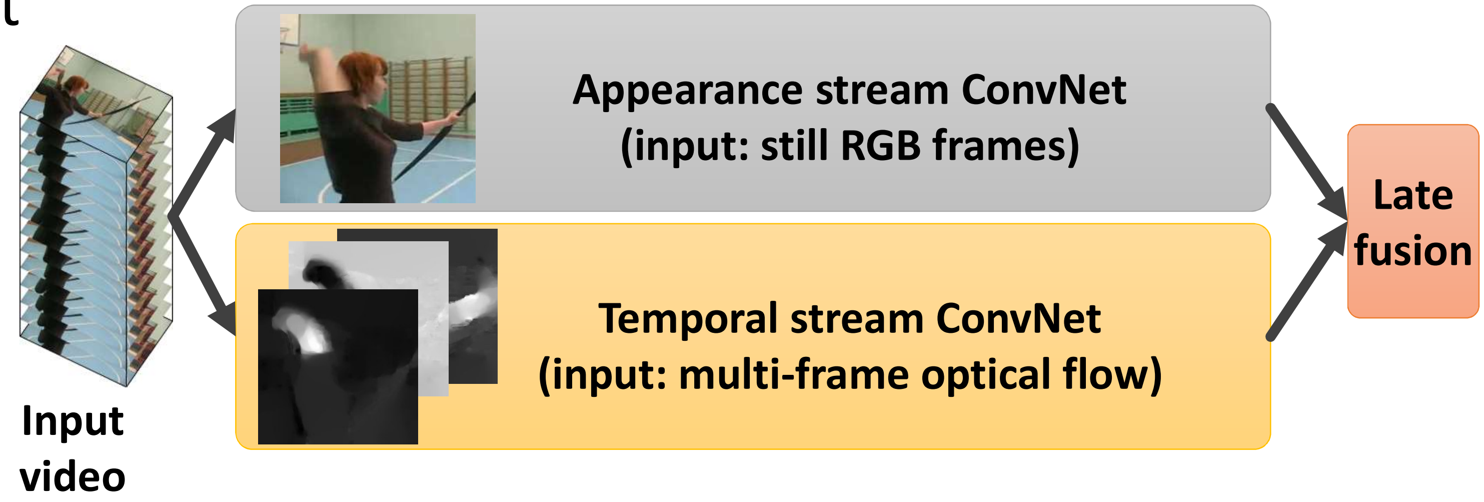
Representing motion using optical flow

- Throws away “nuisance factors” like appearance of clothes and skin
- Helps with foreground/background segmentation



The benefits of optical flow

- Two-Stream ConvNet Architecture


Appearance stream ConvNet
(input: still RGB frames)
Temporal stream ConvNet
(input: multi-frame optical flow)
Late fusion
Input video

- UCF-101 Mean Accuracy (across all splits)

Model	UCF-101
Spatial Stream ConvNet	72.6
Temporal Stream ConvNet (multi-task)	83.6
Two-stream fusion (by averaging)	86.9
Two-stream fusion (weighted averaging)	87.6

K. Simonyan, A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", NIPS 2014

State of the art on Kinetics-400

Top-1 % accuracy on Action classification performance on Kinetics-400 val

Model	RGB only	RGB + flow
S3D-G	74.7	77.2
TSN Inception V3	72.5	76.6
Non-local Neural Networks	77.7	
I3D	71.1	74.2

- Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification, Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, Kevin Murphy, ECCV 2018
- Temporal segment networks: Towards good practices for deep action recognition, Wang, L., Xiong, Y.,Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., ECCV 2016
- Non-local Neural Networks, Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, CVPR 2018
- Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Joao Carreira, Andrew Zisserman, CVPR 17

State of the art on Kinetics-400

Top-1 % accuracy on Action classification performance on Kinetics-400 val

Model	RGB only	RGB + flow
S3D-G	74.7	77.2
TSN Inception V3	72.5	76.6
Non-local Neural Networks	77.7	
I3D	71.1	74.2

- Ceiling on performance is currently less than 80%
- Adding flow boosts performance by around 3%
- Conclusion: RGB models are not able to fully learn from the motion information yet

Relevant Paper

What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets

De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles, CVPR 2018

- Conclusion: the C3D model (using 16 frames) does not use motion to classify 35% of the classes in Kinetics-400

- Consequently: either the model can not learn from the motion of those classes, or the classes do not require motion to classify them

Summary

Current generation of neural network architectures for action classification

- Have not saturated performance on Kinetics yet
- Are probably not learning motion information to its full potential

- Need for more innovation … research questions:
  - How to develop architectures that can efficiently learn motion information?
  - How to develop lighter architectures for action classification?

Notes for the future:

- Kinetics-800 will be released next year
- ActivityNet workshop for Kinetics and AVA challenges