Action Recognition

CS 280, Spring 2015

by Georgia Gkioxari

Examples of actions

- Movement and posture change
  e.g. run, walk, crawl, jump, hop, swim, dance, sit
- Object manipulation

  e.g. pick, carry, hold, push, pull, touch, drive, bike, play musical instrument
- Conversational gesture
  e.g. point ...
- Sign Language

Key cues for action recognition

- Morpho-kinetics of action
  shape and movement of the body
- Identity of objects
- Activity context
  scene or other people performing actions

Action recognition

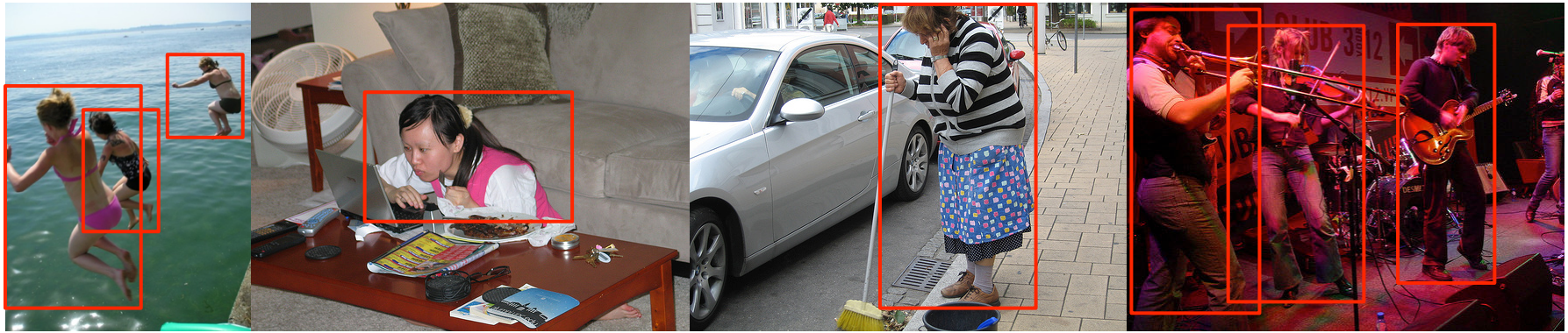
- Static action recognition
  from 2D images

- Video action recognition
  from videos

Static Action Recognition - Datasets

PASCAL VOC Action

- 10 actions [jumping, phoning, riding bike…]
- 12000 instances for train & test



Static Action Recognition - Datasets

MPII Human Dataset

- 410 actions
- 40000 instances for train & test

Static Action Recognition - Approach

CNN for action recognition:

➔ Input: Region containing the actor

➔ Output: One of A action labels

➔ Loss during CNN training: log loss of softmax probabilities

mean AP (%)	8-layer CNN	16-layer CNN
CNN	68.2	77.8

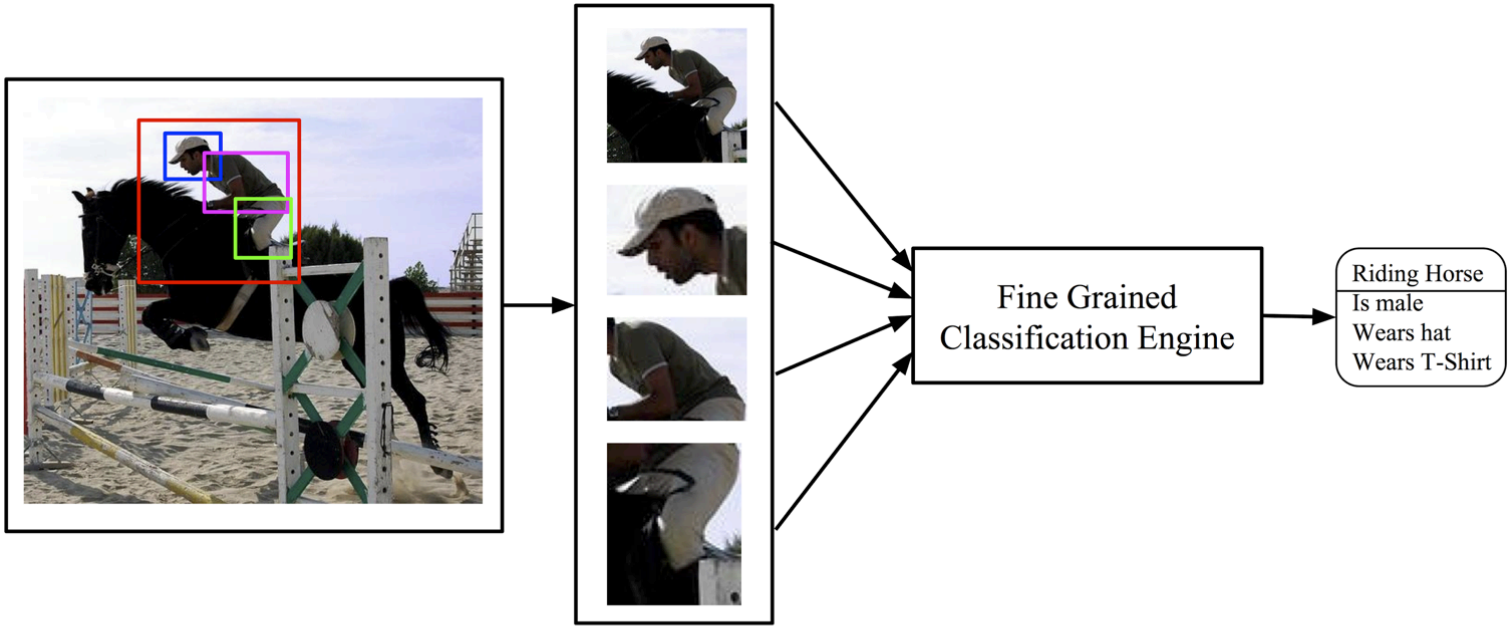
Static Action Recognition - Approach

Observations:

- Some regions in the image matter more than others

- Max pooling layers hide important cues from subsequent layers

Static Action Recognition - Approach



(a) Given an instance hypothesis, we detect parts

(b) The instance and its parts are fed into our classification engine

[1] G. Gkioxari, R. Girshick and J. Malik, Actions and Attributes from Wholes and Parts, arXiv 2014

Part detectors[1]

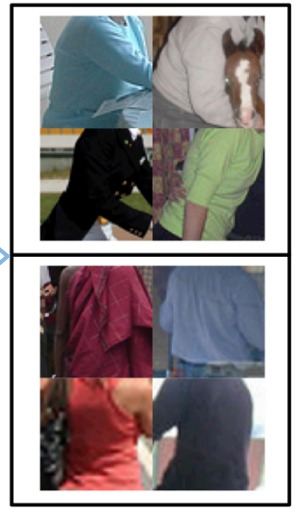
Definition: Parts should capture human body parts of distinct pose and viewpoint

Collection: Given locations of landmarks on the human body (e.g. nose, shoulder) we can obtain examples of pose clusters

Part detectors[1]

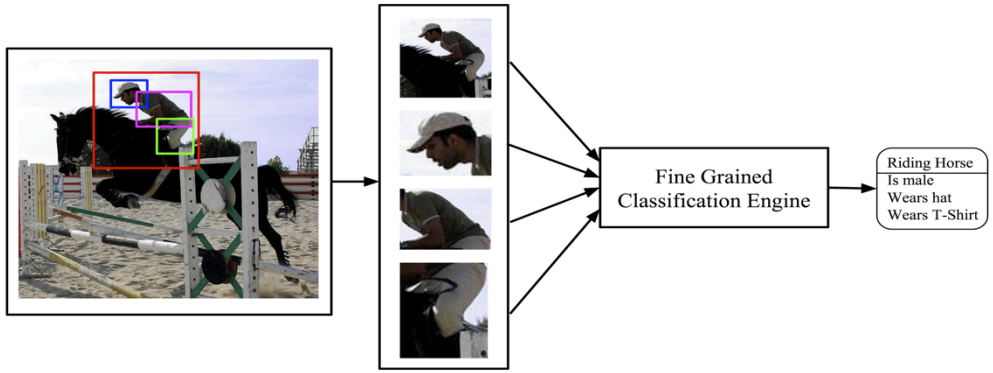
Learning: Train models (e.g. SVM) for each pose cluster on features (e.g. pool5)

Pose clusters for torso (**collection**)



Part detections on test set

Static Action Recognition - Approach



(a) Given an instance hypothesis, we detect parts

(b) The instance and its parts are fed into our classification engine

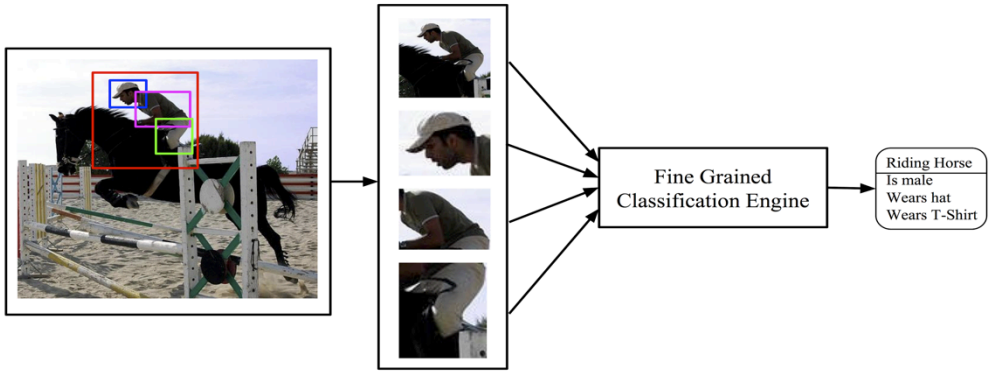
mean AP (%)	8-layer CNN	16-layer CNN
CNN	68.2	77.8
Whole & Parts CNN[1]	**71.5**	**80.4**

Static Action Recognition - Approach

- **Object Context**: The objects surrounding the actor, e.g. horse, bike
- **Action Context**: Actions other people in the image perform, e.g. running in a marathon
- **Scene**: The scene the action is taking place, e.g. swimming pool

mean AP (%)	8-layer CNN	16-layer CNN
CNN	68.2	77.8
Whole & Parts CNN[1]	71.5	80.4
Whole & Parts CNN[1] with context rescoring	**73.5**	**82.6**

Attribute Recognition



(a) Given an instance hypothesis, we detect parts

(b) The instance and its parts are fed into our classification engine

mean AP (%)	8-layer CNN	16-layer CNN
CNN	79.1	88.4
Whole & Parts CNN[1]	**86.0**	**89.5**

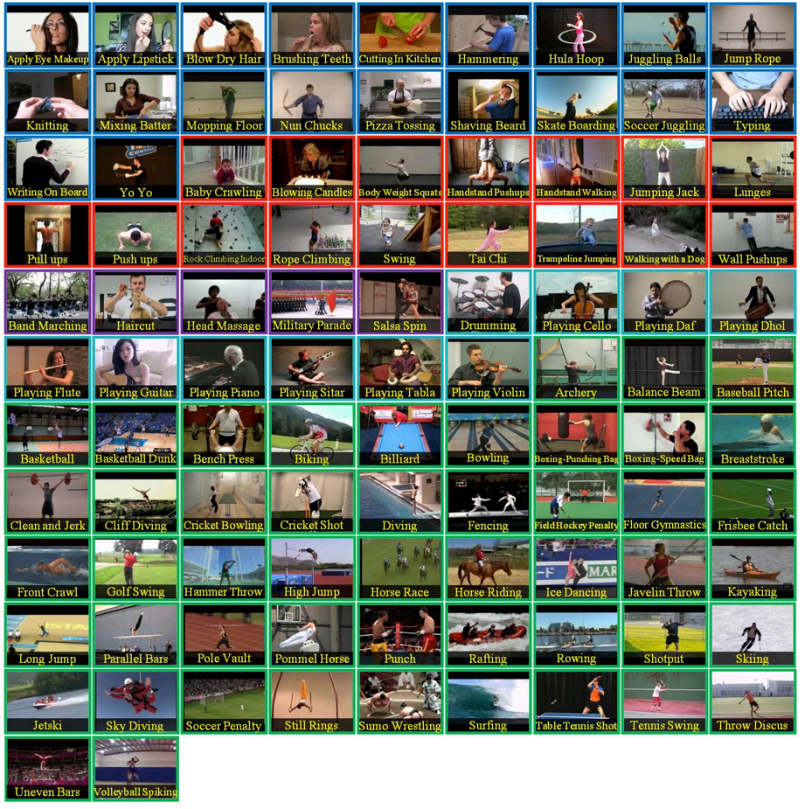
Video action recognition

Freestyle swimming

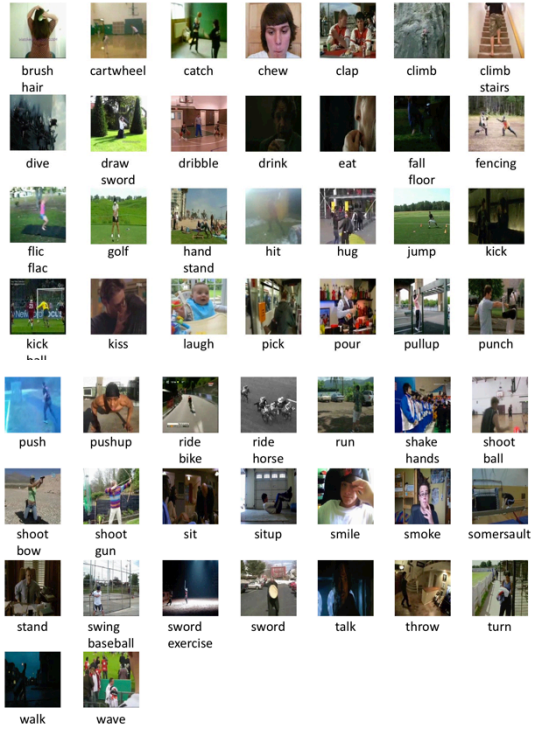
Sailing

Video action recognition - Datasets

UCF 101



HMDB

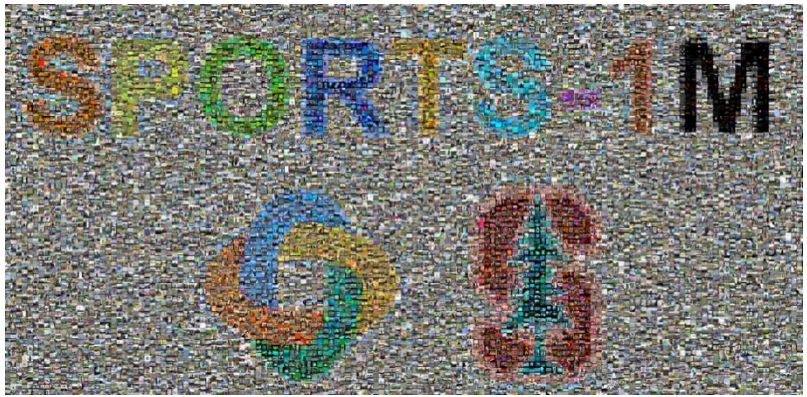


Video action recognition - Datasets

UCF 101
- 101 actions
- 13302 videos

HMDB
- 51 actions
- 6849 videos

1M sports dataset
- 487 actions
- 1,133,158 videos

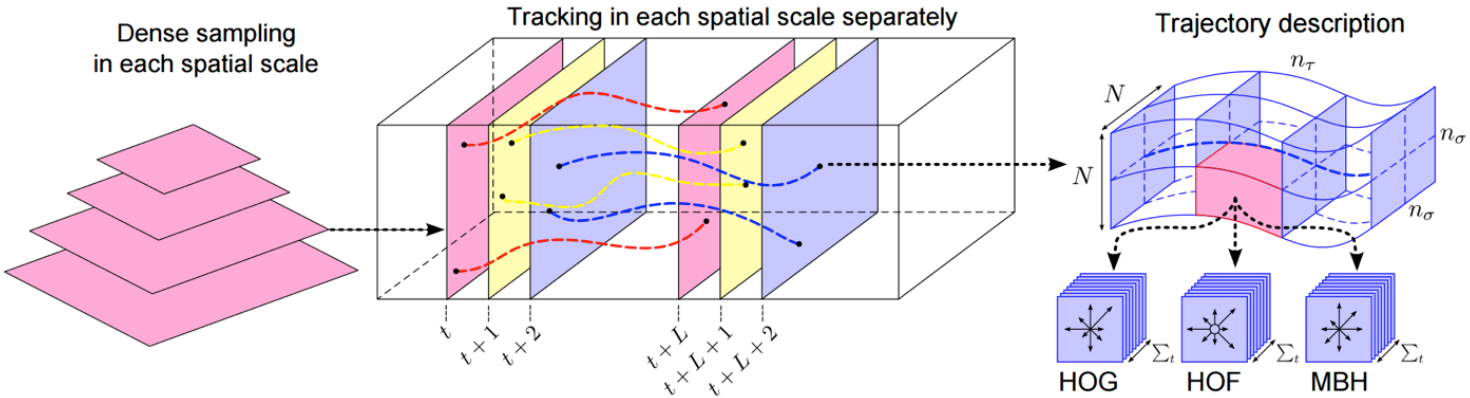


Video action recognition

- Motion cues play important role in videos
- We capture motion with optical flow

Optical flow between frames no.8 - no.9

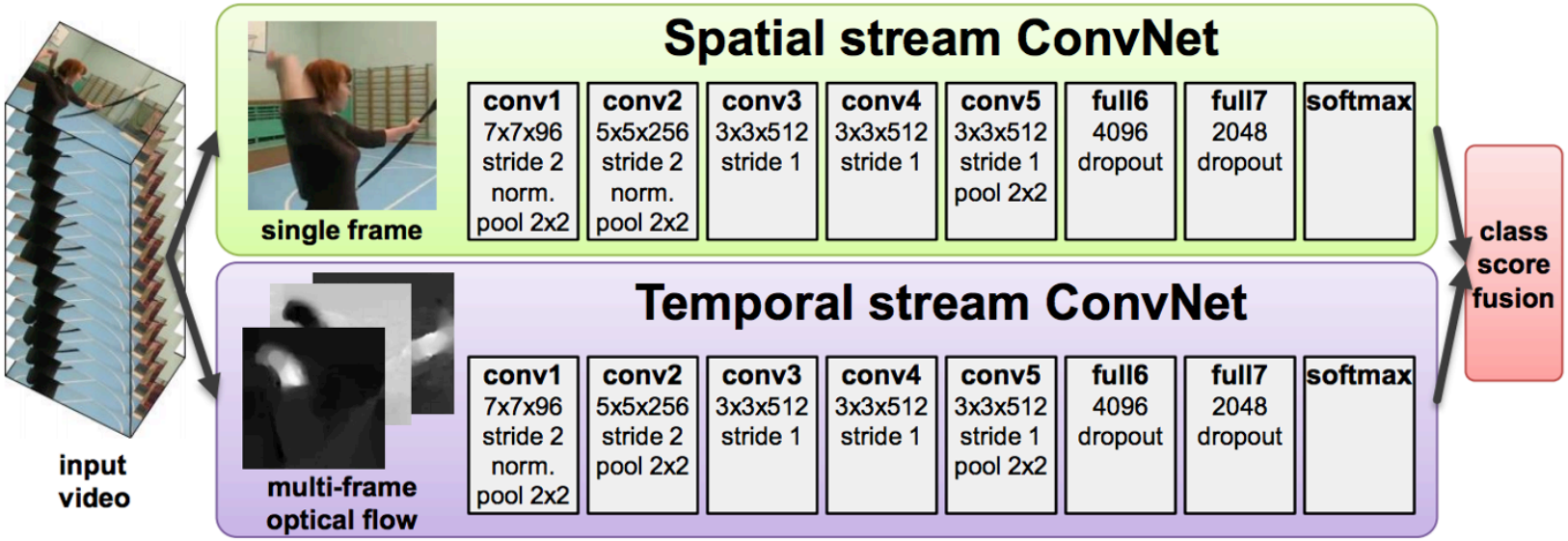
Video action recognition - approach[1]



Accuracy (%)	UCF 101	HMDB
Dense Trajectories	85.9	57.2

[1] Wang et al, Action Recognition by Dense Trajectories, CVPR 2011

Video action recognition - approach[2]



[2] Simonyan and Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

Video action recognition - approach[2]

Accuracy (%)	UCF 101	HMDB
Dense Trajectories	85.9	57.2
Spatial stream CNN	73.0	40.5
Temporal stream CNN	83.7	54.6
Two-stream CNN	**88.0**	**59.4**

[2] Simonyan and Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

Video action recognition - problems

- Scene bias

  most videos can be classified correctly solely based on the scene

- Multiple actions

  the tasks assigns one label to the whole video, what if more actions are being performed?

- Localization

  the location of the predicted action is not specified

Action detection in video[3]

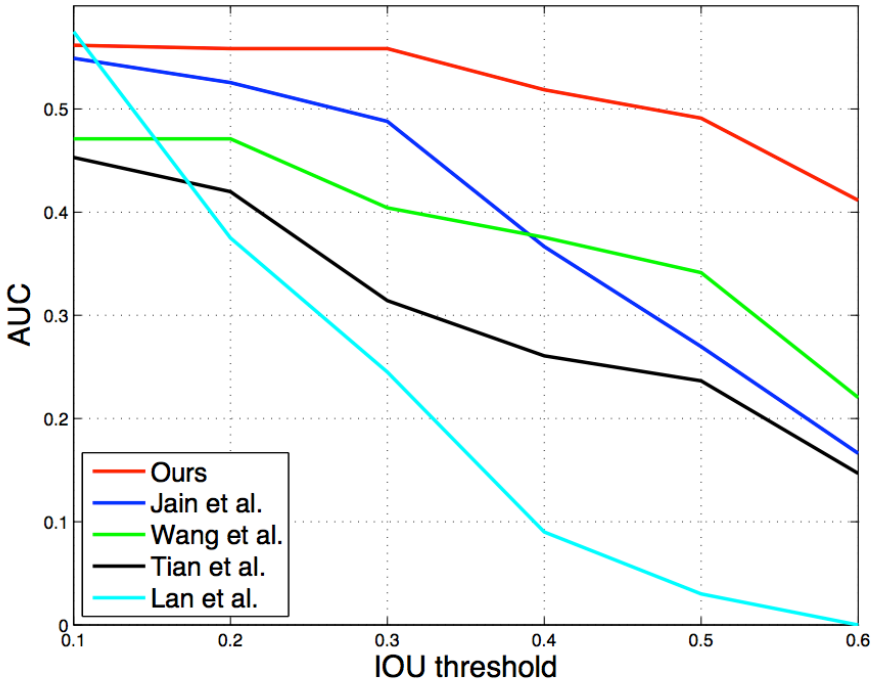
- Task: Given a video, localize the action(s) being performed in the video
- Method
  - Start from regions (prune based on motion saliency)
  - Classify each region based shape and motion cues (spatial- & motion- CNNs and fusion)
  - Link detections across frames (dynamic programming)

[3] Gkioxari and Malik, Finding Action Tubes, CVPR 2015

Action detection in video[3]

[3] Gkioxari and Malik, Finding Action Tubes, CVPR 2015

Action detection in video[3]



[3] Gkioxari and Malik, Finding Action Tubes, CVPR 2015

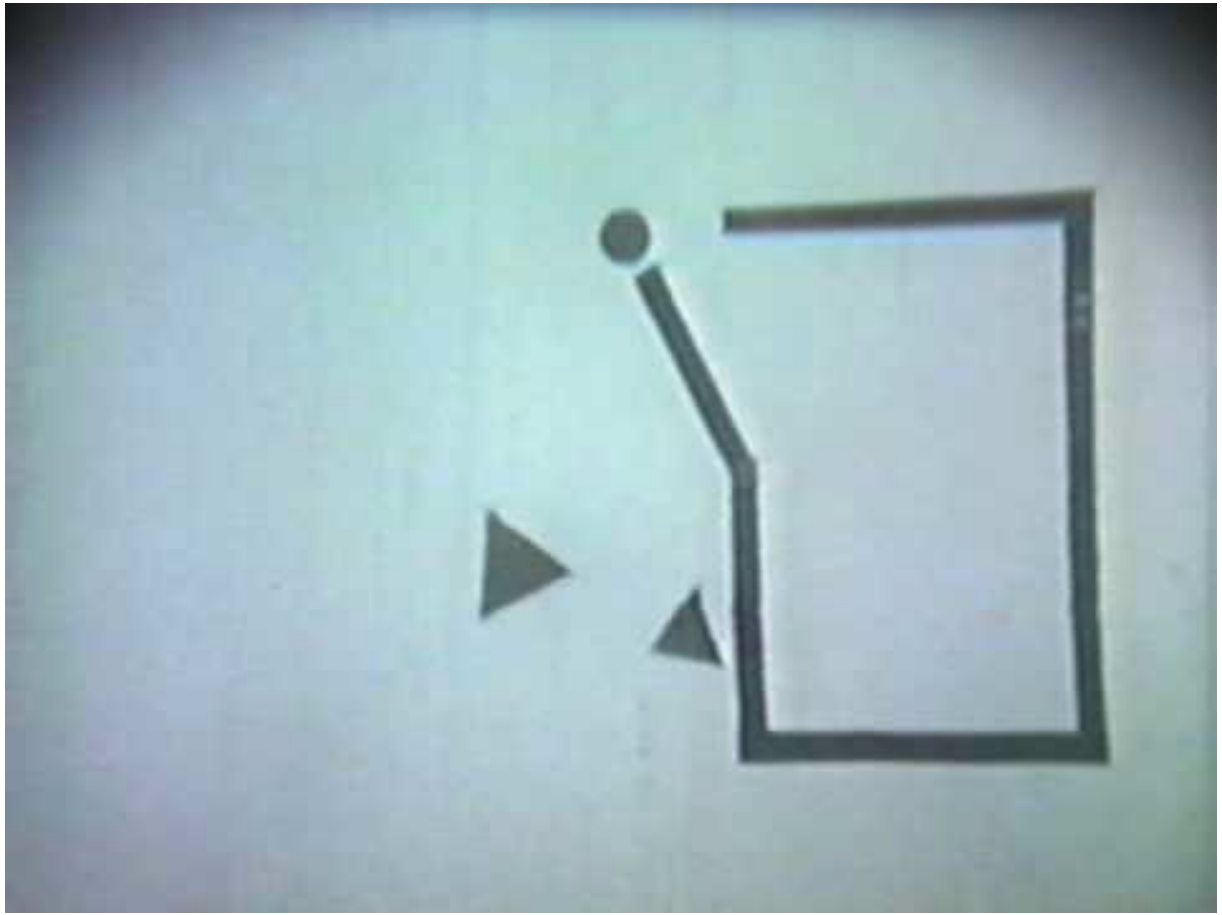
Action detection in video[3]

Action classification can be benefited from analyzing an action wrt the actor

Accuracy (%)	Wang et al.[1]	Two-stream CNN[2]	Action Tubes[3]
J-HMDB	56.6	56.5	**62.5**

[3] Gkioxari and Malik, Finding Action Tubes, CVPR 2015

Questions?



Heider & Simmel 1944