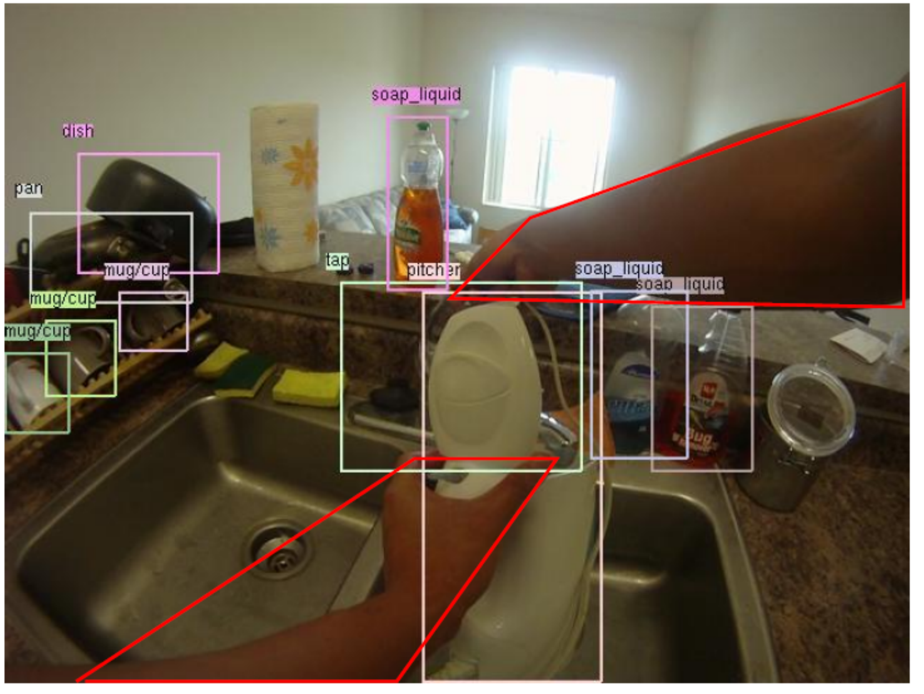
Detecting Activities of Daily Living in First-person Camera Views



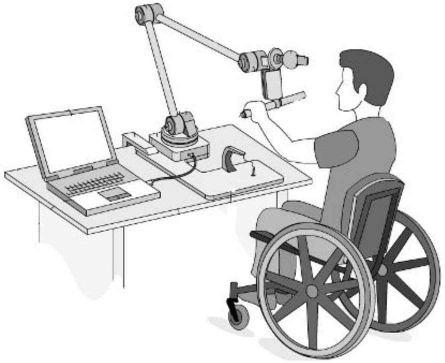
Hamed Pirsiavash, Deva Ramanan

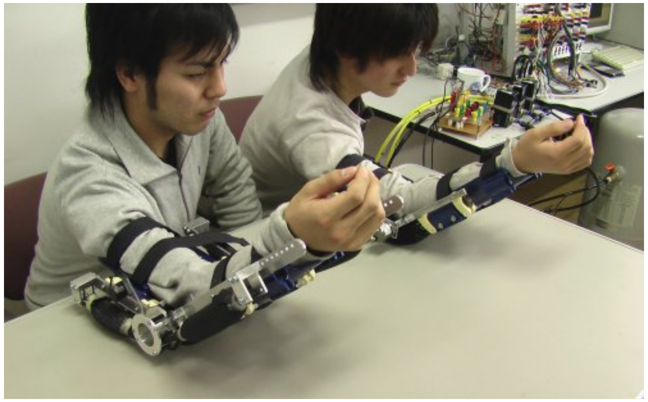
Computer Science Department, UC Irvine

Motivation

A sample video of Activities of Daily Living

Applications
Tele-rehabilitation





Long-term at-home monitoring

- Kopp et al,, Arch. of Physical Medicine and Rehabilitation. 1997.
- Catz et al, Spinal Cord 1997.

Applications
Life-logging

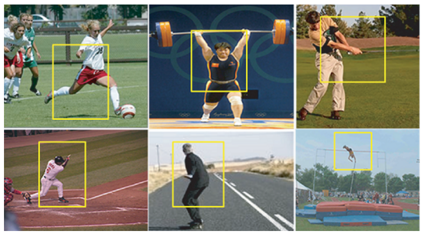
So far, mostly "write-only" memory!

This is the right time for computer vision community to get involved.

- Gemmell et al, "MyLifeBits: a personal database for everything." Communications of the ACM 2006.
- Hodges et al, "SenseCam: A retrospective memory aid", UbiComp, 2006.

Related work: action recognition

There are quite a few video benchmarks for action recognition.



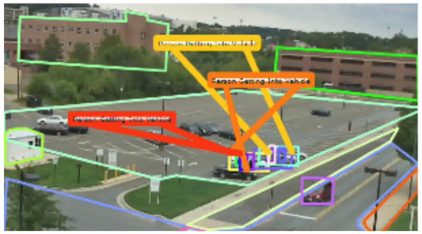
UCF sports, CVPR'08

KTH, ICPR'04

Olympics sport, BMVC'10

Hollywood, CVPR'09

UCF Youtube, CVPR'08



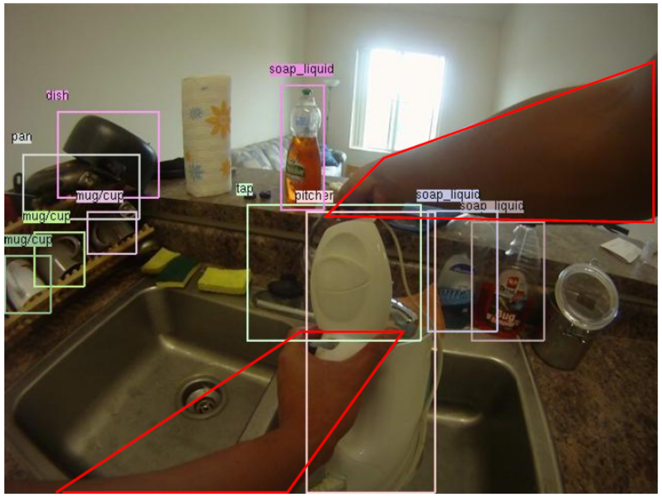
VIRAT, CVPR'11

Collecting interesting but natural video is surprisingly hard.

It is difficult to define action categories outside "sports" domain

Wearable ADL detection

It is easy to collect natural data



Wearable ADL detection

It is easy to collect natural data

ADL actions derived from medical literature on patient rehabilitation

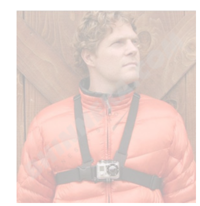
- actions
  - hygiene
    - personal hygiene
      - 5 washing hand/face
      - 6 drying hand/face
      - facial hygiene
        - 1 combing hair
        - 2 make up
        - oral hygiene
          - 3 brushing teeth
          - 4 dental floss
    - external hygiene
      - 7 laundry
      - 15 vacuuming
  - food
    - preparing food
      - 8 washing dishes
      - 9 moving dishes
    - eating food
      - liquid
        - 10 making tea
        - 11 making coffee
        - 12 drinking water/bottle
        - 13 drinking water/tap
      - solid
        - 14 making cold food/snack
  - entertainment
    - 16 watching TV
    - 17 using computer
    - 18 using cell

Outline

- Challenges
  - What features to use?
  - Appearance model
  - Temporal model

- Our model
  - “Active” vs “passive” objects
  - Temporal pyramid

- Dataset



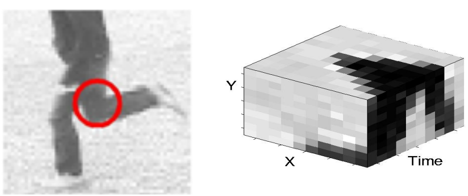
- Experiments

Challenges

What features to use?

<u>Low level features</u>

(Weak semantics)

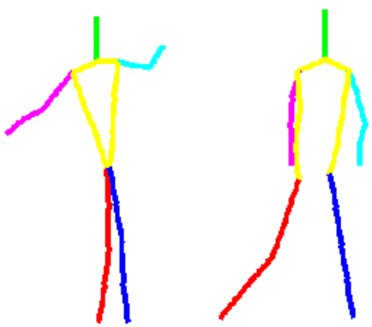


Space-time interest points

Laptev, IJCV'05

<u>High level features</u>

(Strong semantics)



Human pose

Difficulties of pose:

- Detectors are not accurate enough
- Not useful in first person camera views

Challenges

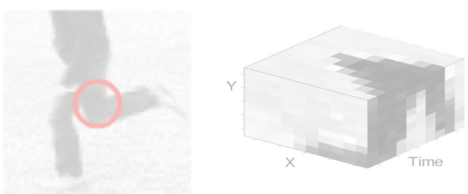
What features to use?

Low level features

(Weak semantics)

High level features

(Strong semantics)



Space-time interest points

Laptev, IJCV'05

Human pose

Object-centric features

Difficulties of pose:

- Detectors are not accurate enough
- Not useful in first person camera views

Challenges
Occlusion / Functional state

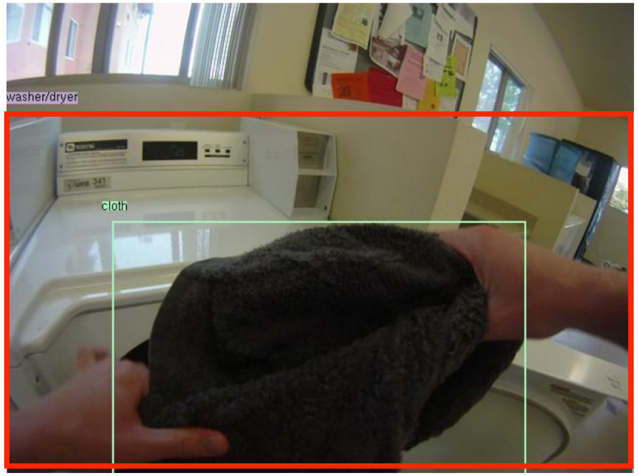
"Classic" data

Challenges

Occlusion / Functional state

“Classic” data

Wearable data



Challenges
long-scale temporal structure

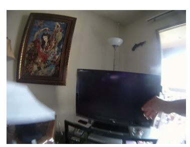
"Classic" data: boxing

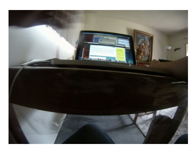
Challenges
long-scale temporal structure

"Classic" data: boxing

Wearable data: making tea

Start boiling water





Do other things (while waiting)

Pour in cup

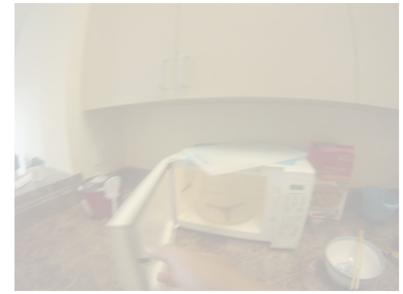
Drink tea

time

Difficult for HMMs to capture long-term temporal dependencies

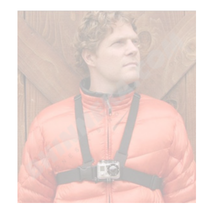
Outline

- Challenges
  - What features to use?
  - Appearance model
  - Temporal model



- Our model
  - “Active” vs “passive” objects
  - Temporal pyramid

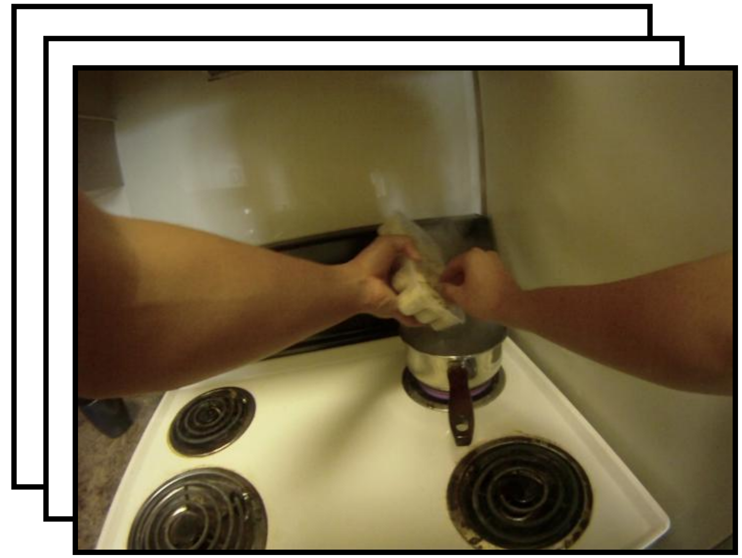
- Dataset



- Experiments

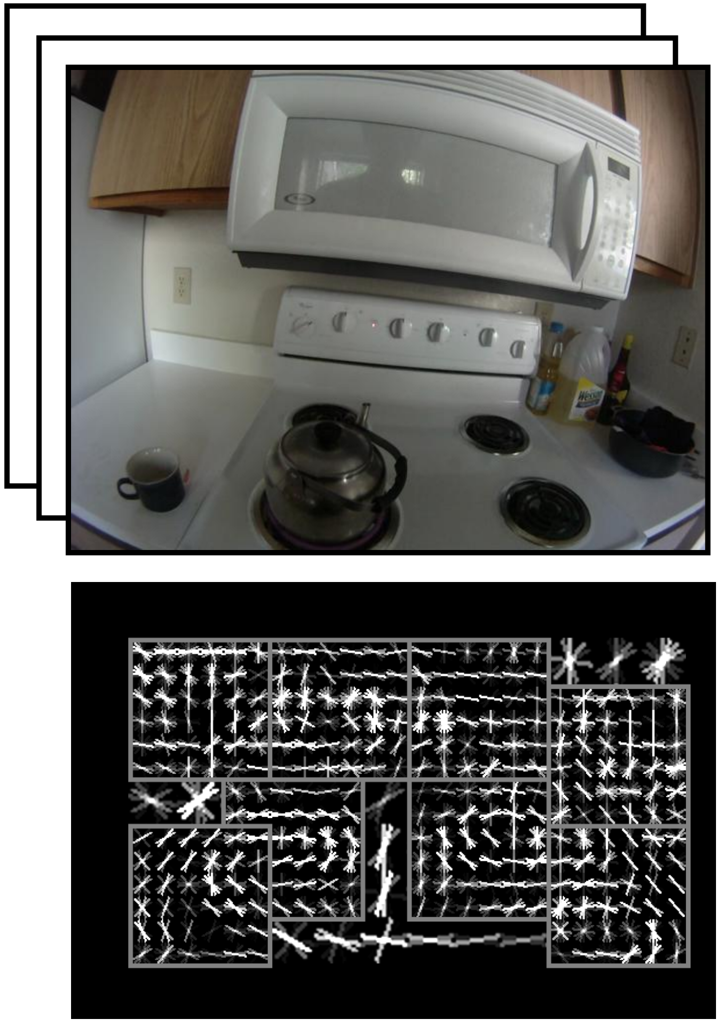
“Passive” vs “active” objects

Passive

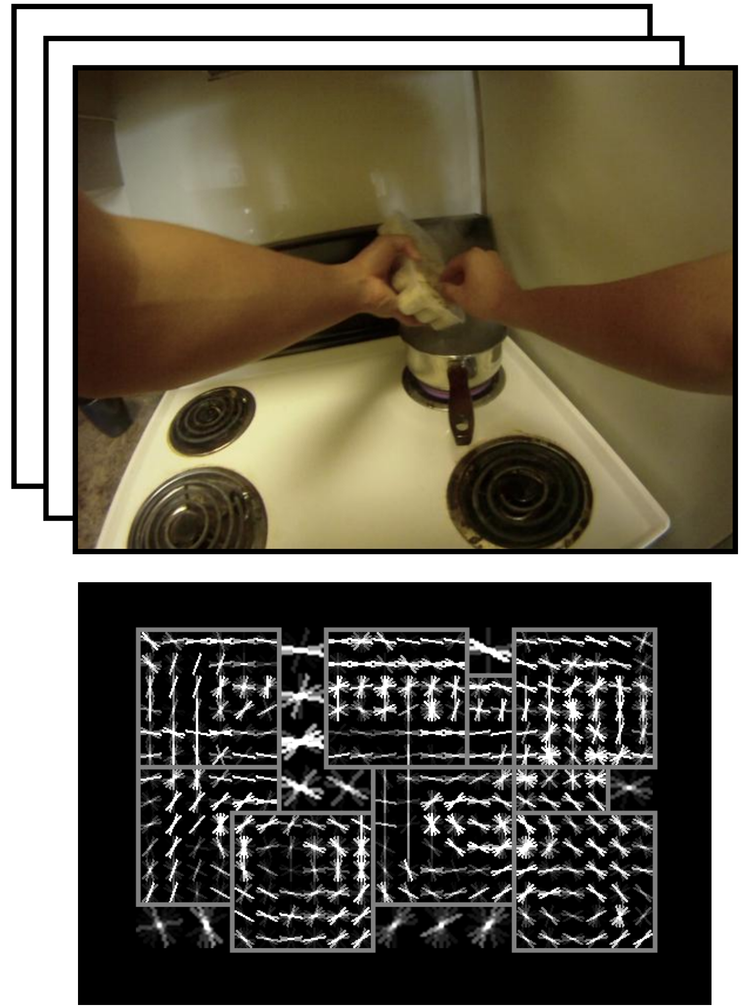


Active

“Passive” vs “active” objects

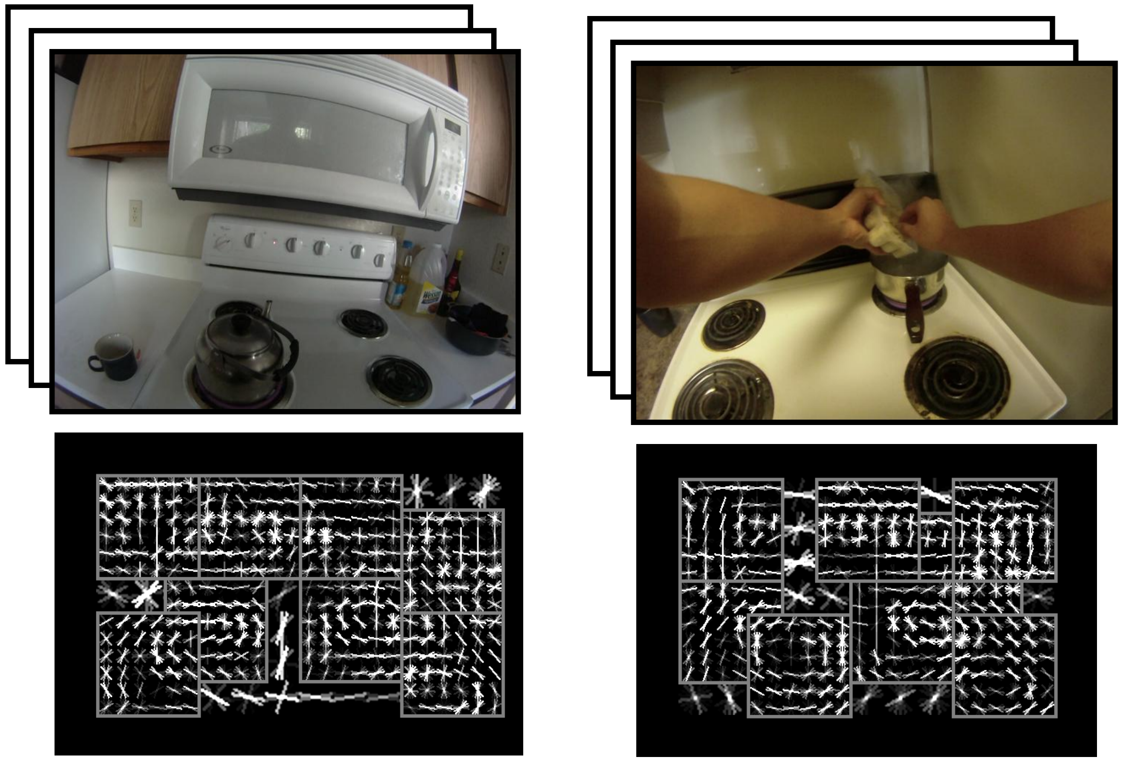


Passive



Active

“Passive” vs “active” objects

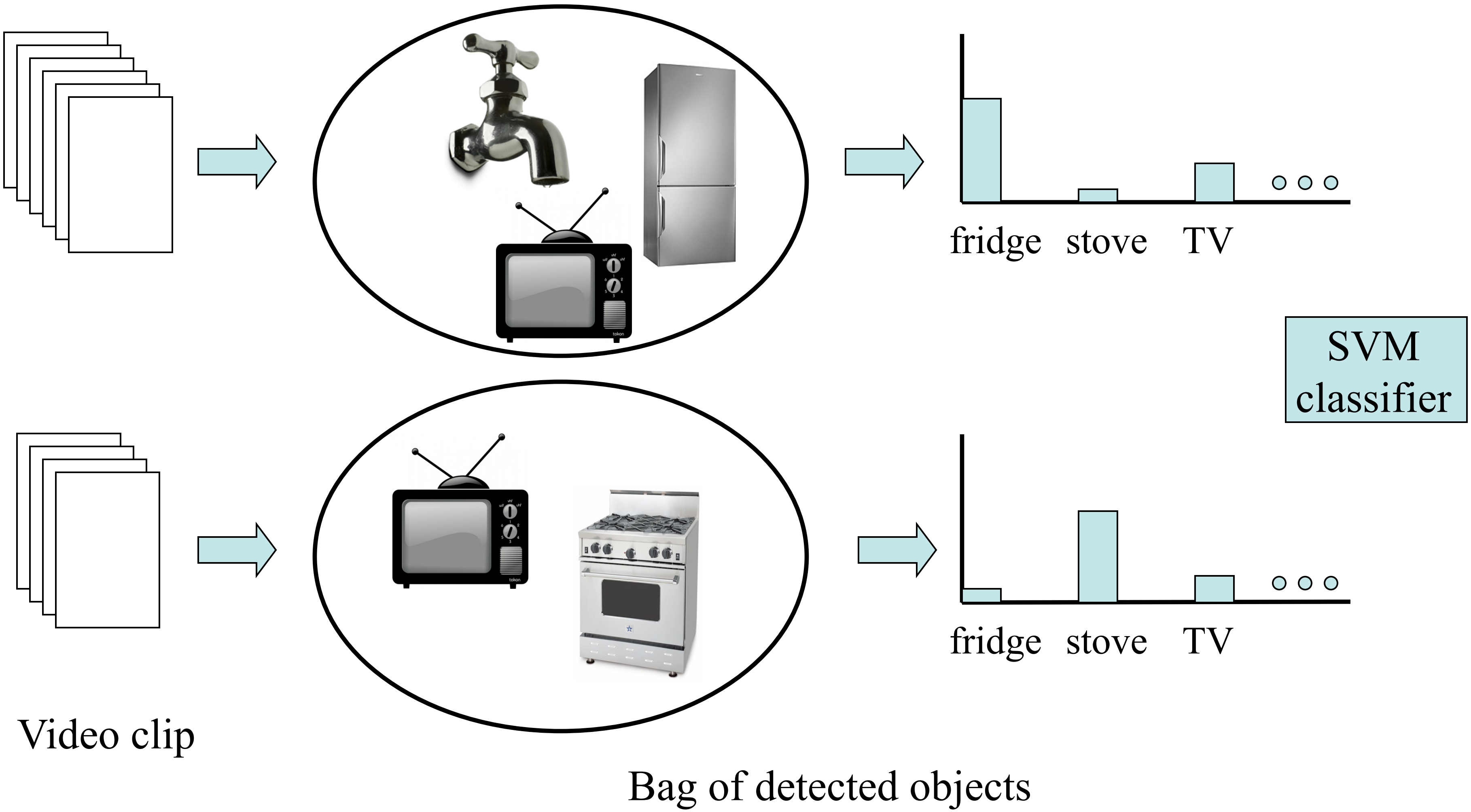


Passive   Active

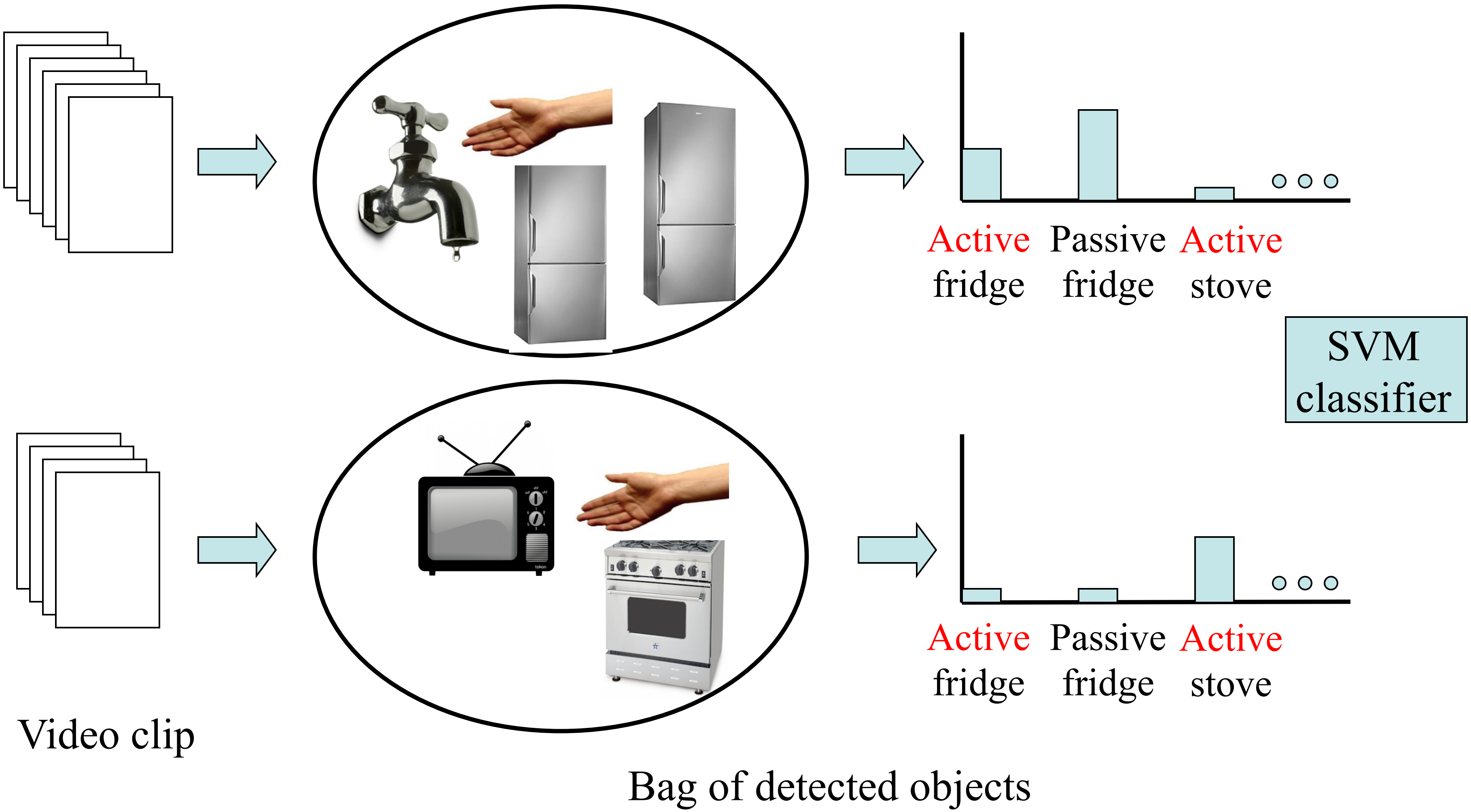
Better object detection (visual phrases CVPR’11)

Better features for action classification (active vs passive)

Appearance feature: bag of objects



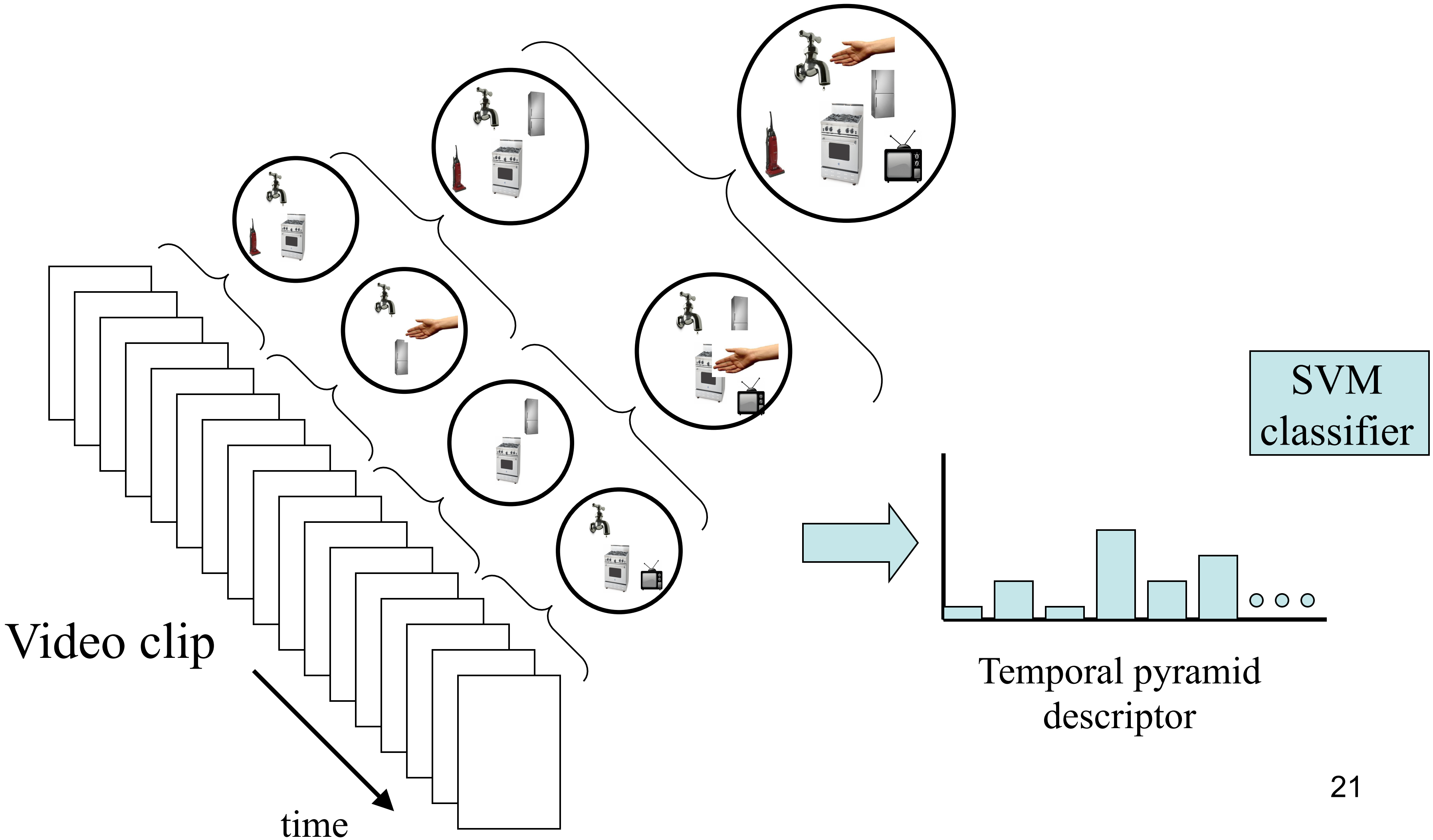
Appearance feature: bag of objects



Temporal pyramid

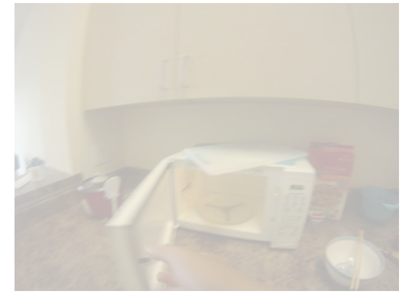
Coarse to fine correspondence matching with a multi-layer pyramid

Inspired by “Spatial Pyramid” CVPR’06 and “Pyramid Match Kernels” ICCV’05



Outline

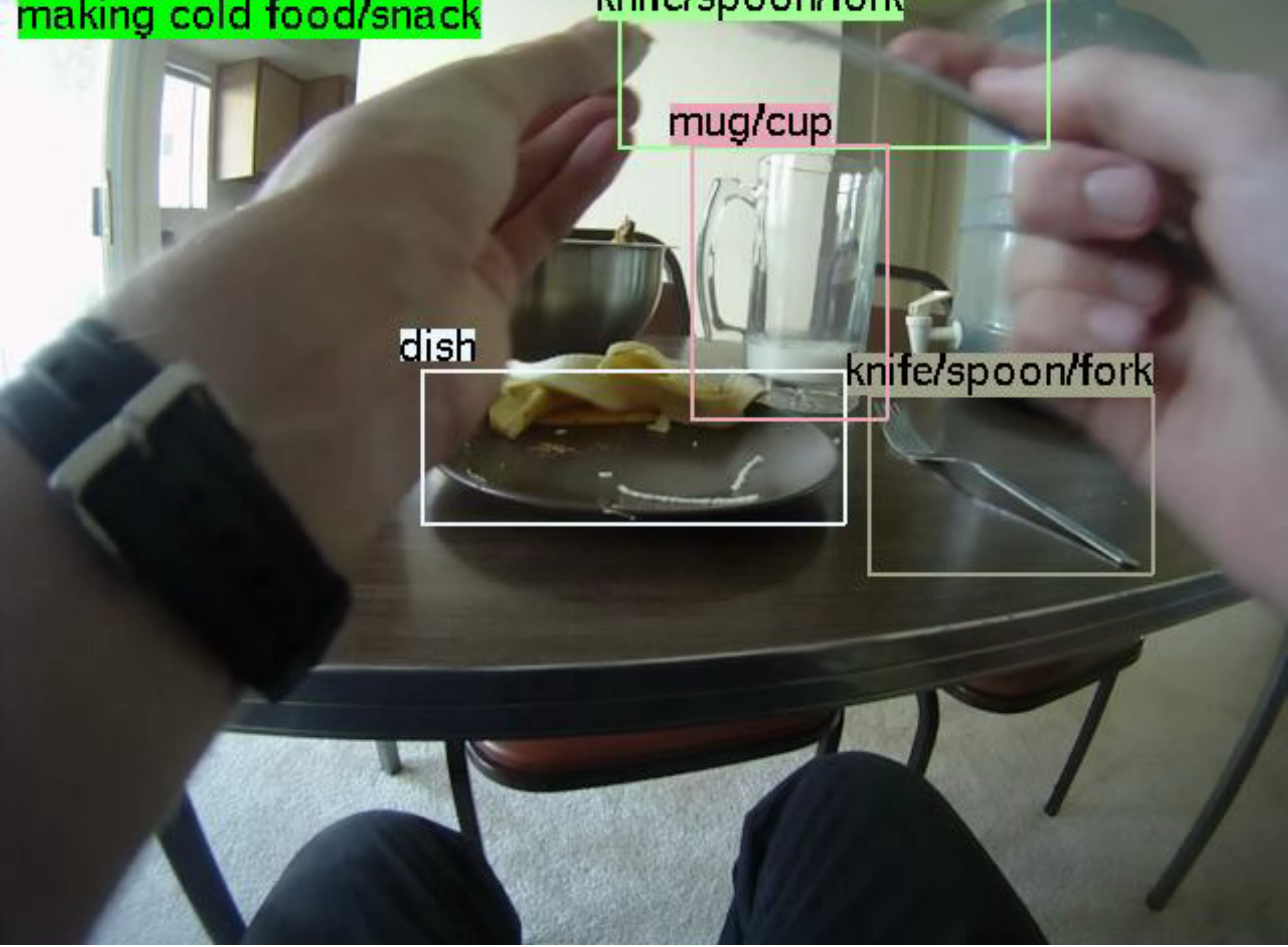
- Challenges
  - What features to use?
  - Appearance model
  - Temporal model



- Our model
  - “Active” vs “passive” objects
  - Temporal pyramid

- Dataset

- Experiments

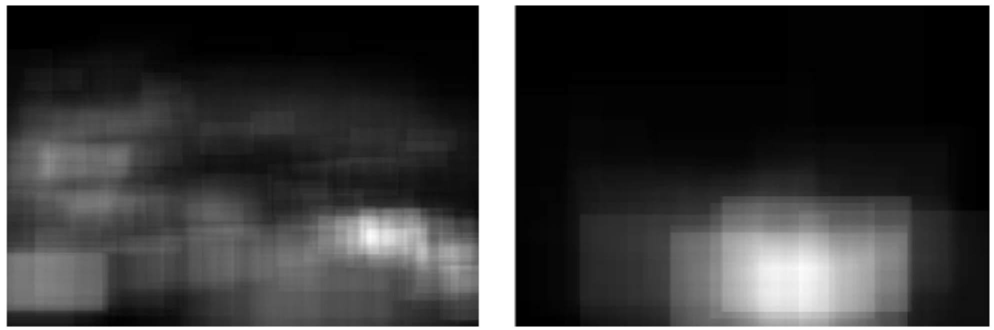

making cold food/snack
mug/cup
dish
knife/spoon/fork

Wearable ADL data collection

- 20 persons
- 20 different apartments
- 10 hours of HD video
- 170 degrees of viewing angle
- Annotated
  - Actions
  - Object bounding boxes
  - Active-passive objects
  - Object IDs

Prior work:
- Lee et al, CVPR’12
- Fathi et al, CVPR’11, CVPR’12
- Kitani et al, CVPR’11
- Ren et al, CVPR’10

Average object locations



pan

Passive Active

Active objects tend to appear on the right hand side and closer

- Right-handed people are dominant
- We cannot mirror-flip images in training

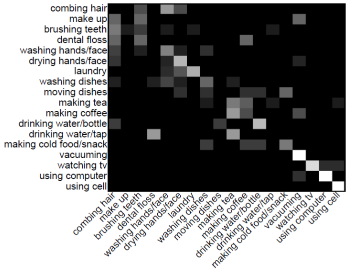
Outline

- Challenges
  - What features to use?
  - Appearance model
  - Temporal model

- Our model
  - “Active” vs “passive” objects
  - Temporal pyramid

- Dataset

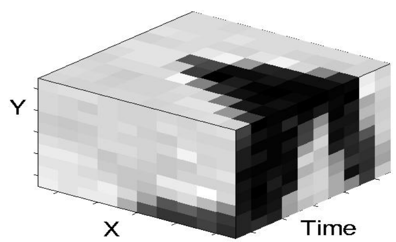
- Experiments



Experiments

Baseline

Space-time interest points
(STIP) Laptev et al, BMVC'09



Low level features

Our model

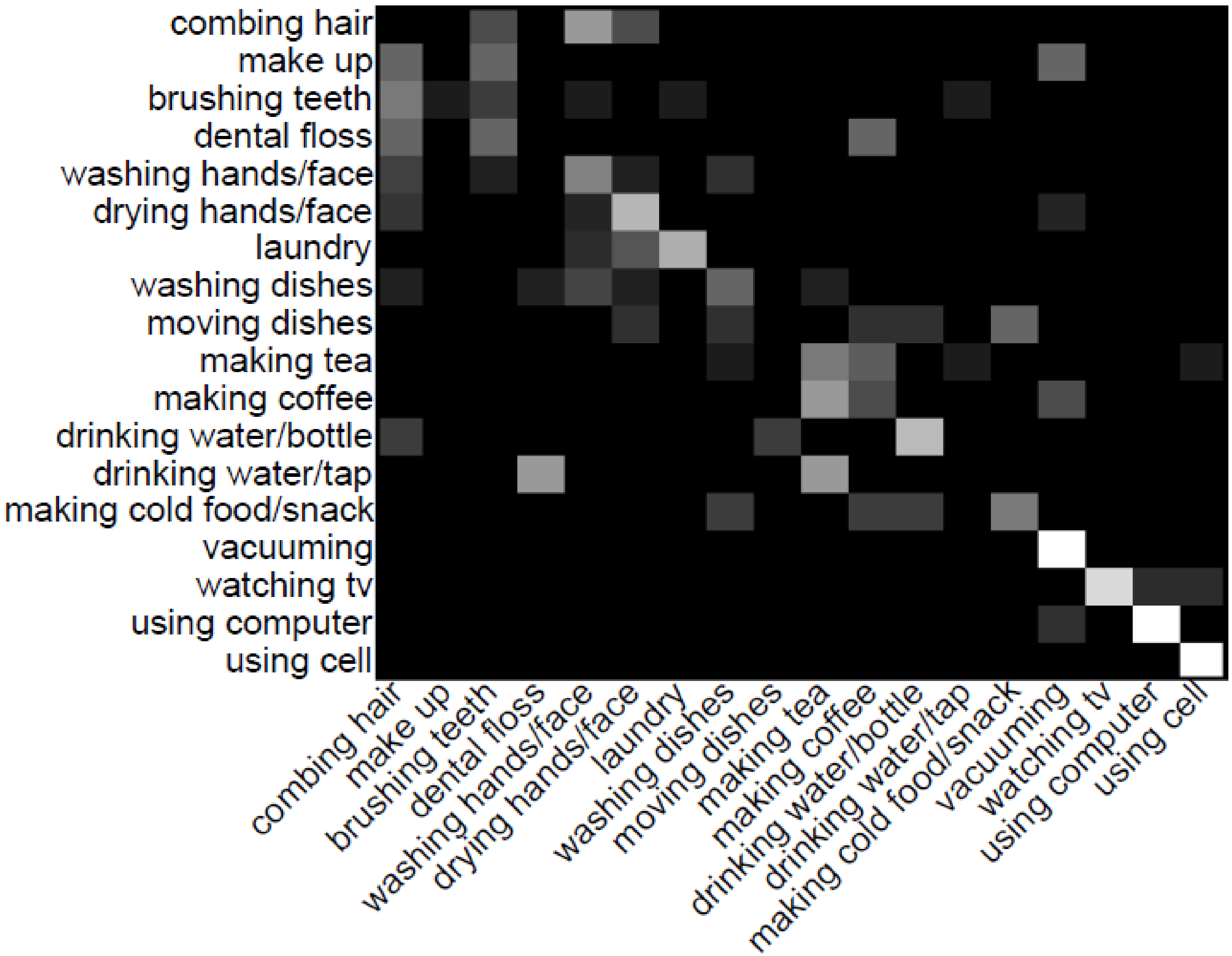
Object-centric features

24 object categories

High level features

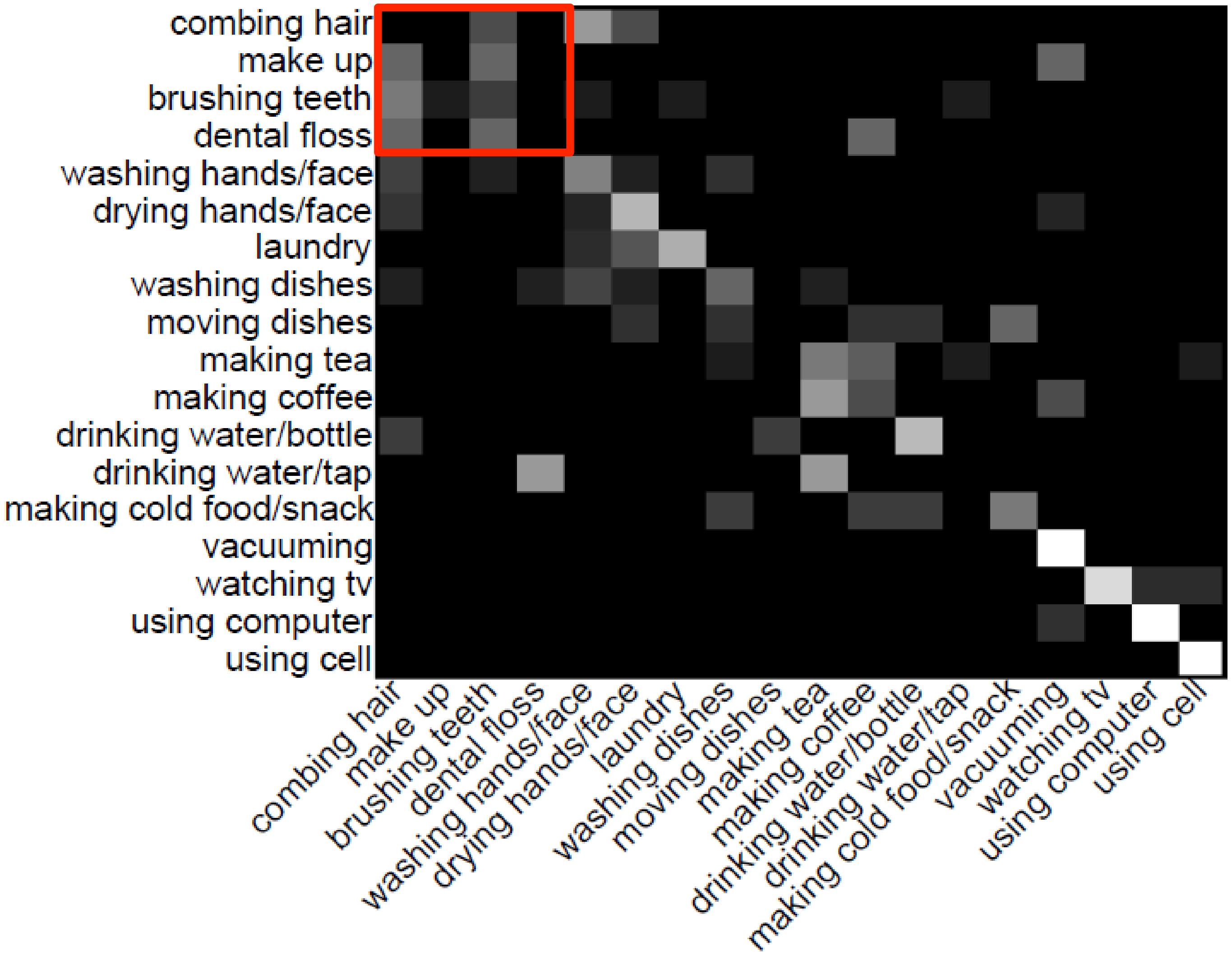
Accuracy on 18 action categories

- Our model: 40.6%
- STIP baseline: 22.8%



Accuracy on 18 action categories

- Our model: 40.6%
- STIP baseline: 22.8%



Classification accuracy

	Classification accuracy
Bag of STIP	16.5
Temporal pyramid of STIP	22.8

- Temporal model helps

Classification accuracy

	Classification accuracy
Bag of STIP	16.5
Temporal pyramid of STIP	22.8
Object detectors	32.7

- Temporal model helps
- Our object-centric features outperform STIP

Classification accuracy

	Classification accuracy
Bag of STIP	16.5
Temporal pyramid of STIP	22.8
Object detectors	32.7
Active/passive object detectors	40.6

- Temporal model helps
- Our object-centric features outperform STIP
- Visual phrases improves accuracy

Classification accuracy

	Classification accuracy
Bag of STIP	16.5
Temporal pyramid of STIP	22.8
Object detectors	32.7
Active/passive object detectors	40.6
Ideal active/passive object detectors	77.0

- Temporal model helps
- Our object-centric features outperform STIP
- Visual phrases improves accuracy
- Ideal object detectors double the performance

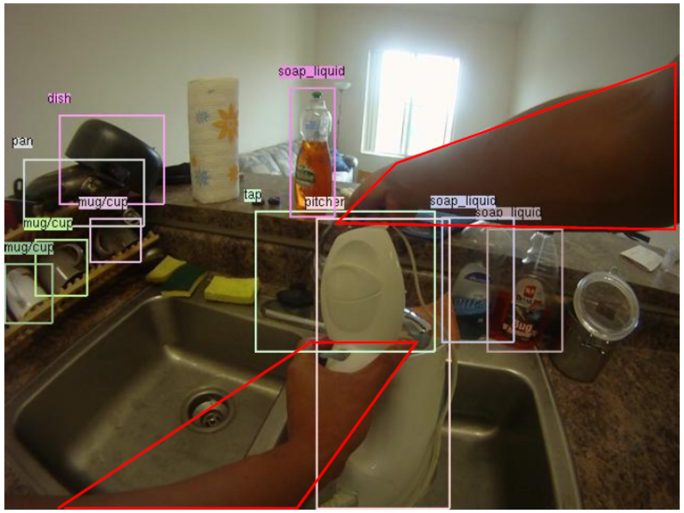
Classification accuracy

	Classification accuracy
Bag of STIP	16.5
Temporal pyramid of STIP	22.8
Object detectors	32.7
Active/passive object detectors	40.6
Ideal active/passive object detectors	77.0

- Temporal model helps
- Our object-centric features outperform STIP
- Visual phrases improves accuracy
- Ideal object detectors double the performance

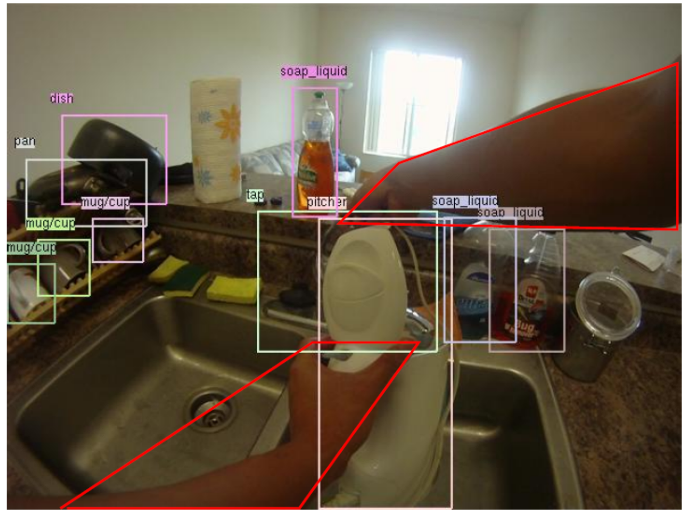
Results on temporally continuous video and taxonomy loss are included in the paper

Summary



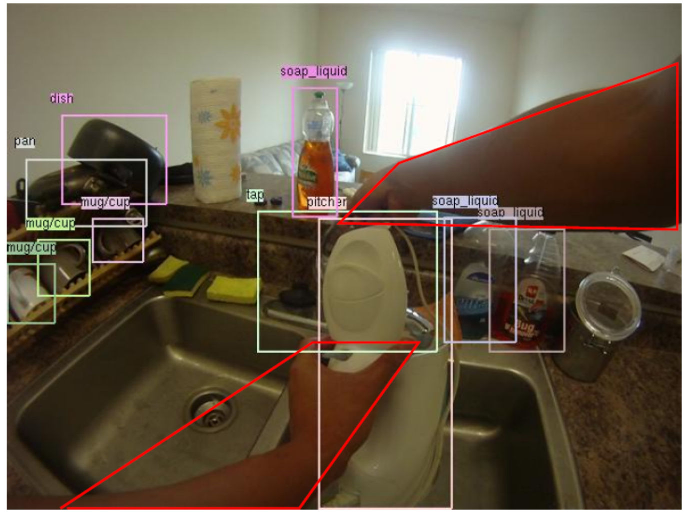
Data and code will be available soon!

Summary

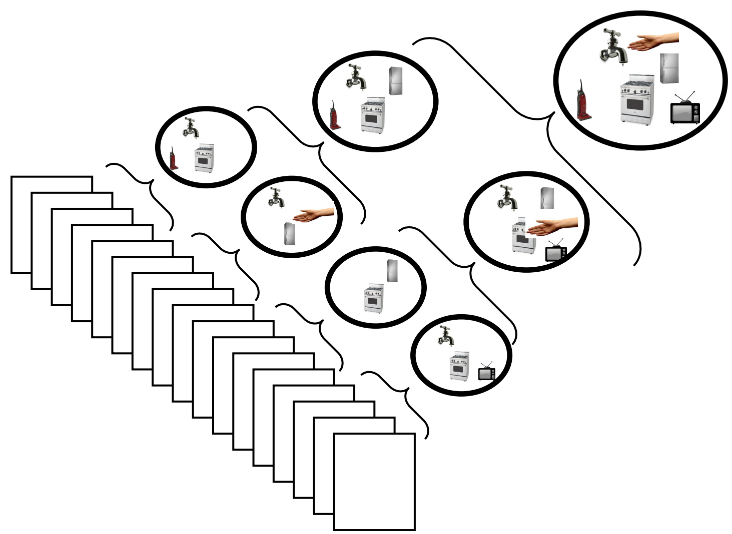


Data and code will be available soon!

Summary



Data and code will be available soon!



Thanks!