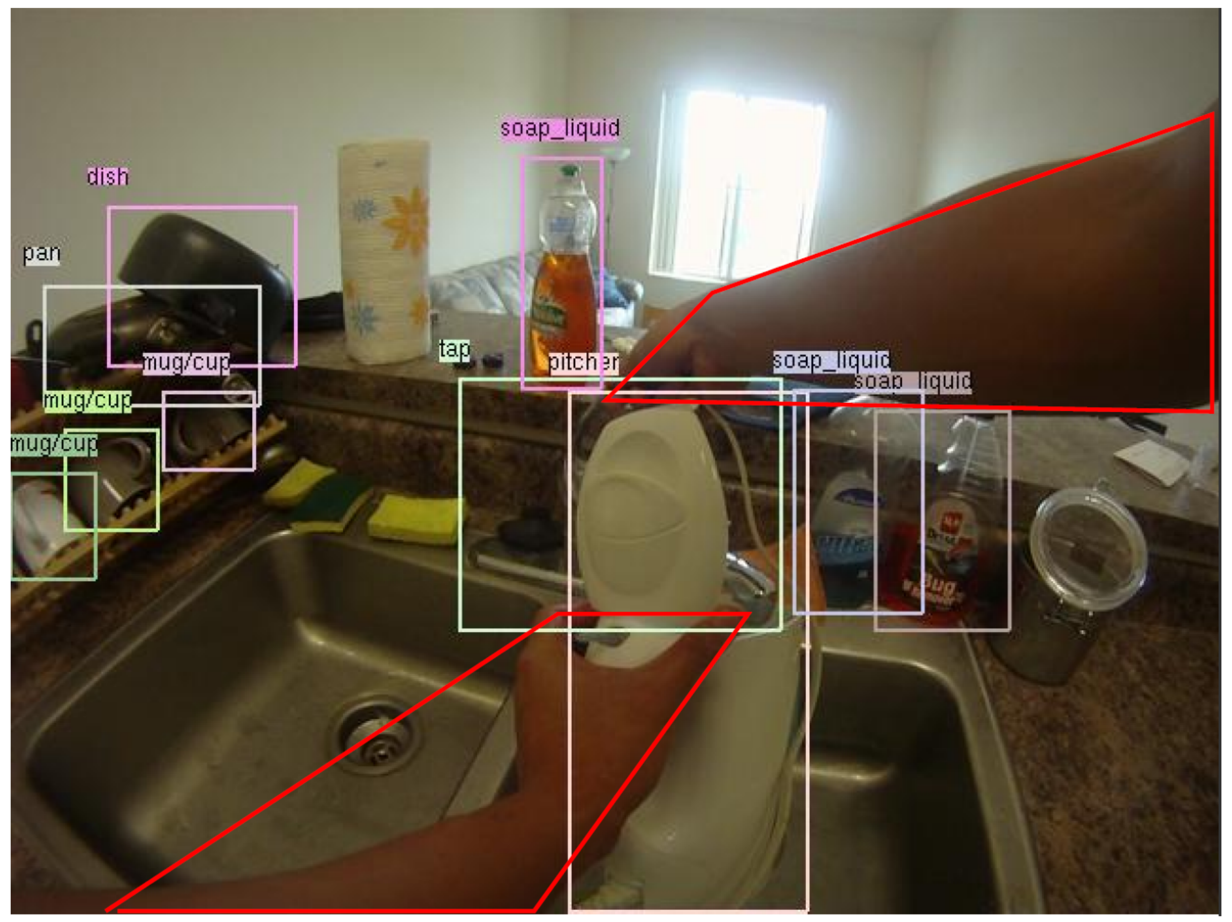
Detecting Activities of Daily Living in First-person Camera Views

Hamed Pirsiavash, Deva Ramanan   Computational Vision Lab, University of California Irvine, CA, USA   {hpirsiav, dramanan} @ ics.uci.edu



Motivation:

- Activity recognition is less well defined compared to object detection since it is difficult to:
  - Define domain independent categories
  - Collect natural footage with large intra-class variation

- We detect activities of daily living (ADLs) from first person wearable cameras
  - ADL categories derived from medical literature on rehabilitation
  - Capturing data is easy.

Applications:

- Tele-rehabilitation
  - Long-term, at-home monitoring instead of short-term inpatient care
- Life-logging
  - Process visual personal histories, avoiding "write-only" memories

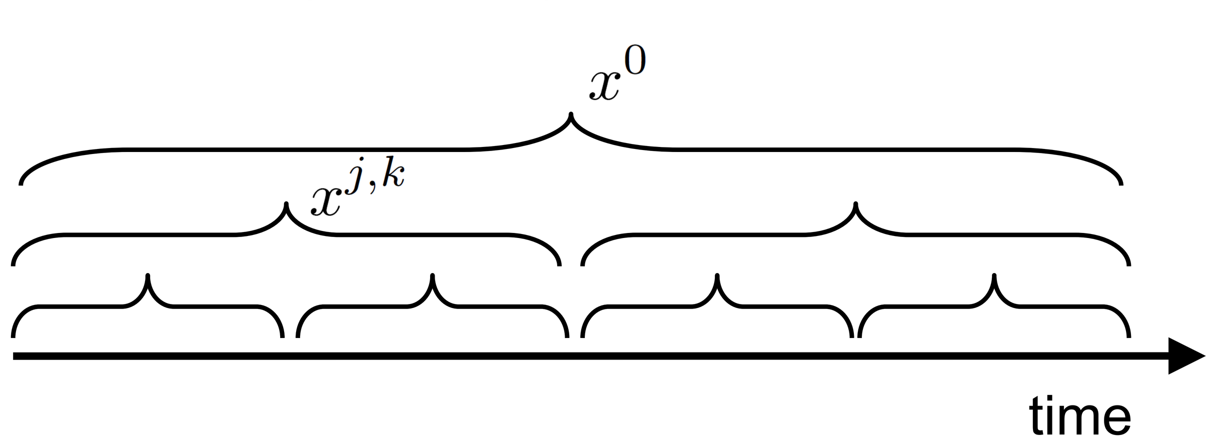
Our contributions:

- Novel representations
- Dataset

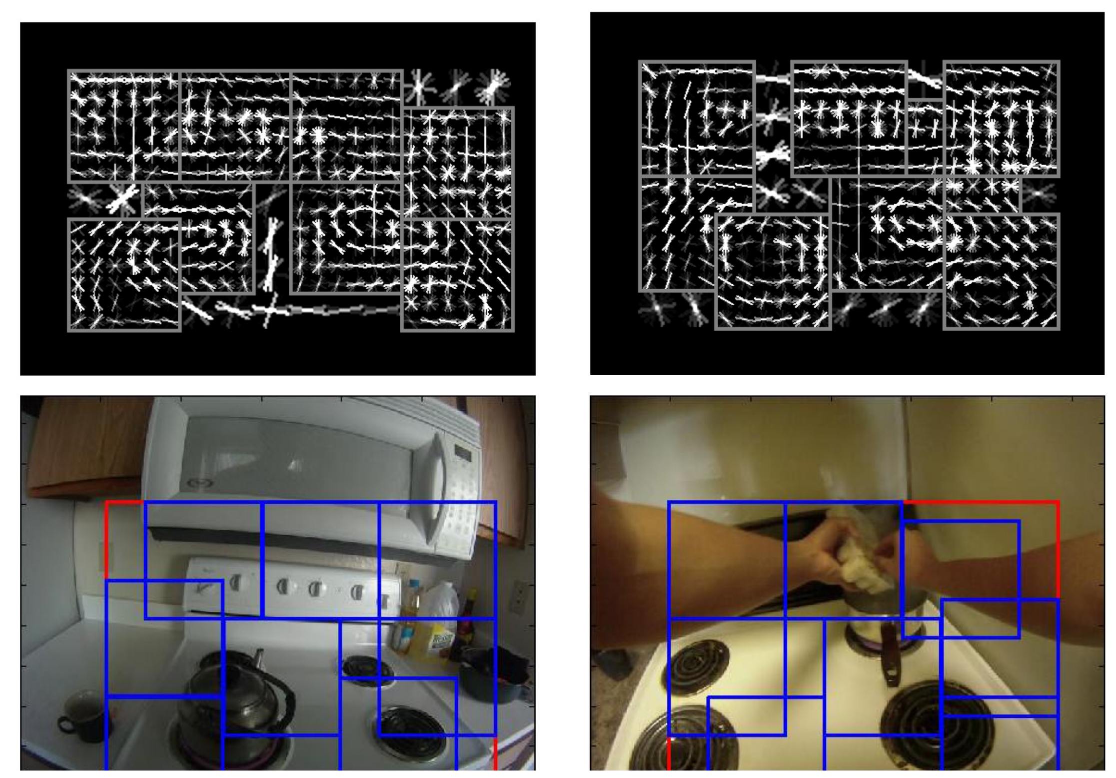
Novel representations:

- **Object-centric activity models**
  - Bag of objects instead of bag of xyt interest points (STIP)

- **Temporal pyramid**
  - Similar to spatial pyramid
  - Models actions with long-term dependencies, eg. making tea
  - Encodes temporal correspondence between model and data



- **Composite object models**
  - Objects look different when being interacted with
  - Learn separate detectors for active and passive objects



- **Activity descriptor**

  - Score of the i'th object detector at point $p = (x, y, s)$

$$\text{score}_i^t(p) \in [0, 1]$$

  - The best detection: $f_i^t = \max_p \ \text{score}_i^t(p)$

  - Bag of objects (level 0): $x_i^0 = \frac{1}{|T|} \sum_{t \in T} f_i^t$

  - j'th level of the pyramid:

$$x_i^{j,k} = \frac{2^j}{|T^{j,k}|} \sum_{t \in T^{j,k}} f_i^t \quad ; \quad \forall k \in \{1 \ldots 2^j\}$$

  - Pyramid descriptor:

$$x = \left[x_1^0 \quad \ldots \quad x_i^{j,k} \ldots \quad x_K^{L,2^L} \right]^T$$

  - Learn activity-specific classifiers (SVMs)

Dataset:

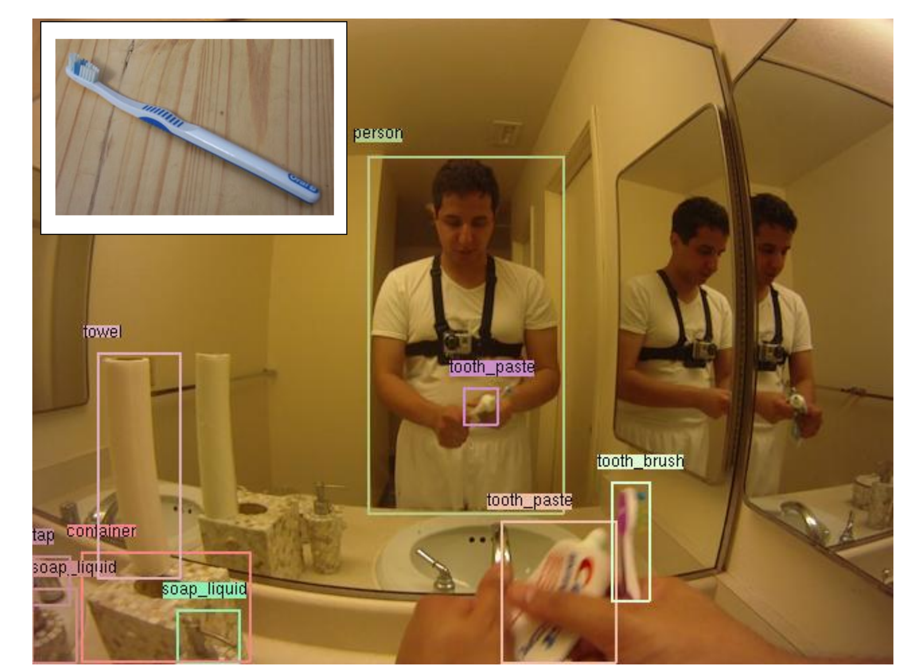
- **Size**
  - One million frames, 10 hours of near-continuous video
  - 20 people, 20 homes

- **Annotation**
  - Activity label, object bounding box, object identity, and human-object interaction

- **Characteristics:** Large variation in…
  - Scenes (cf. existing wearable datasets)
  - Object viewpoints/occlusions (cf. existing image datasets)
  - Un-segmented variable-length activities (cf. action datasets)

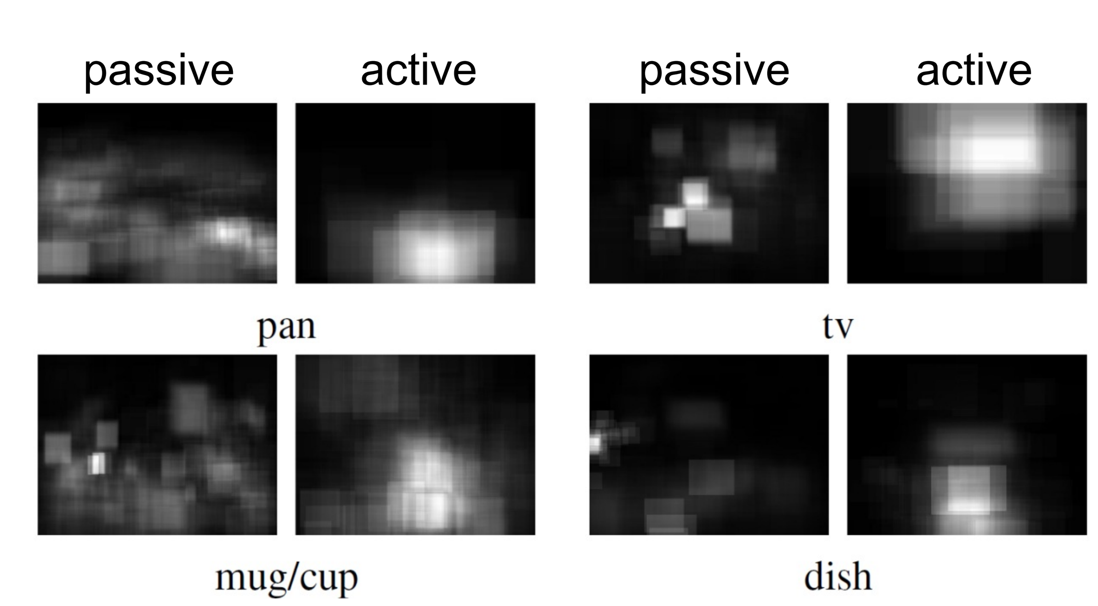
- **Functional Taxonomy**
  - Non scripted ADL's

Variation in object state



Variation in object viewpoint

Variation in scenes

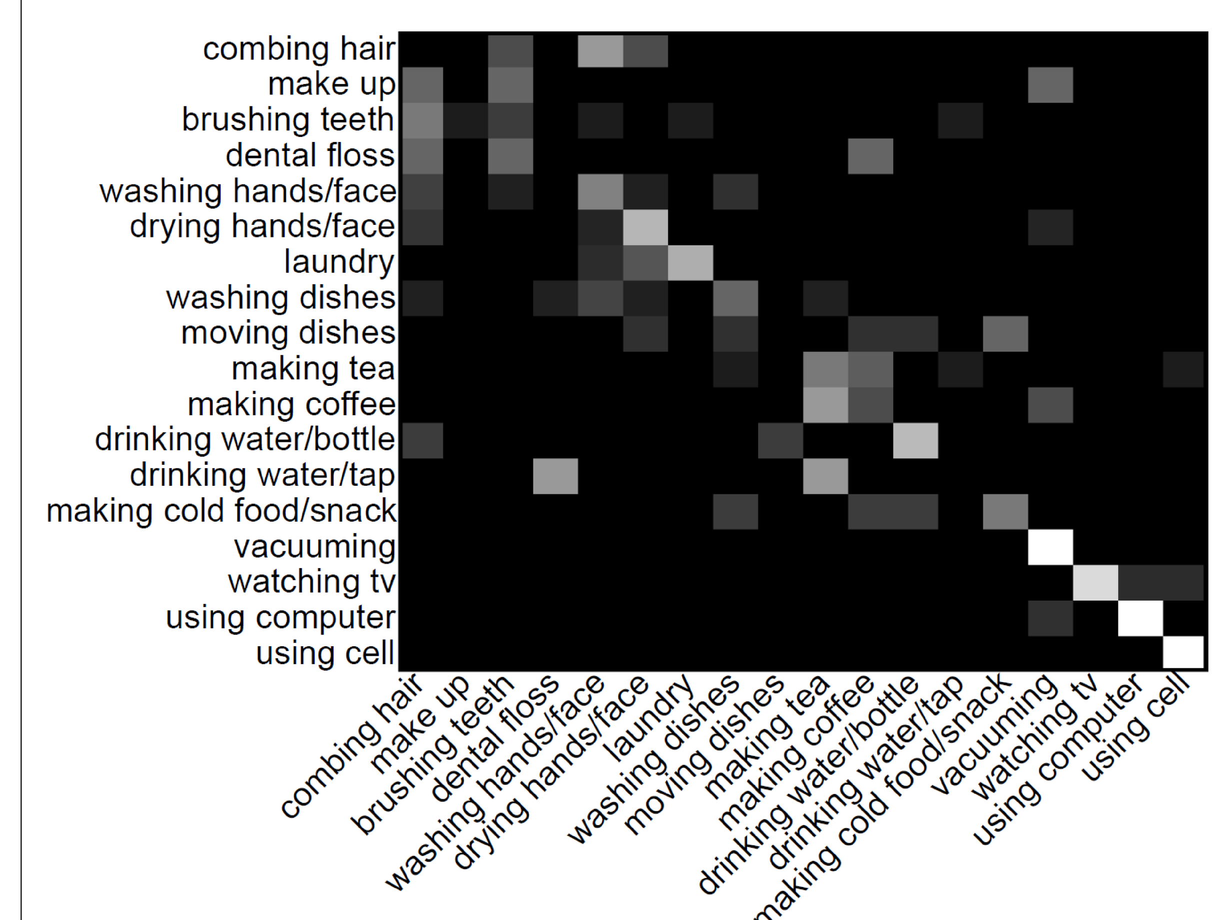


Location and scale of active versus passive objects

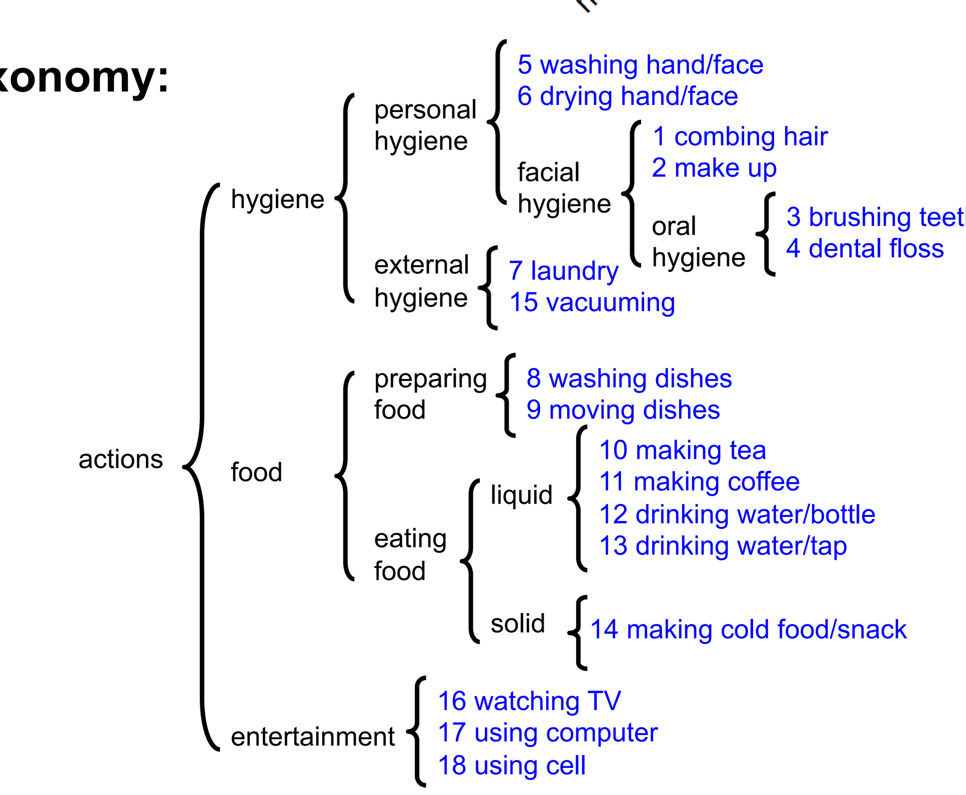
Results:

- **Models**
  - **STIP:** space-time interest points (baseline)
  - **O:** Bag/pyramid of objects
  - **AO:** Bag/pyramid of active objects
  - **IO:** Bag/pyramid of ideal objects
  - **IA+IO:** Bag/pyramid of ideal objects and ideal active objects

	pre-segmented			
	class. accuracy		taxonomy loss	
	pyramid	bag	pyramid	bag
STIP	22.8	16.5	1.8792	2.1092
O	32.7	24.7	1.4017	1.7129
AO	40.6	36.0	1.2501	1.4256
IO	55.8	49.3	0.9267	0.9947
IA+IO	77.0	76.8	0.4664	0.4851



Taxonomy:



Conclusions:

- Real-world ADL recognition is **"all about objects"** and importantly, **"objects being interacted with"**
- Functional loss correlates with scene context
- Looking ahead: better models of object viewpoint, occlusion, and functional interactions