Video: Ego-centric and Summarization

Presentation: Constance Clive

Computer Science Department

University of Pittsburgh

Nonchronological Video Synopsis and Indexing

Yael Pritch, Alex Rav-Acha, Shmuel Peleg

School of Computer Science and Engineering

The Hebrew University of Jerusalem

IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008

Motivation

- Effectively summarize activities from captured surveillance video
- Address queries on generated database objects

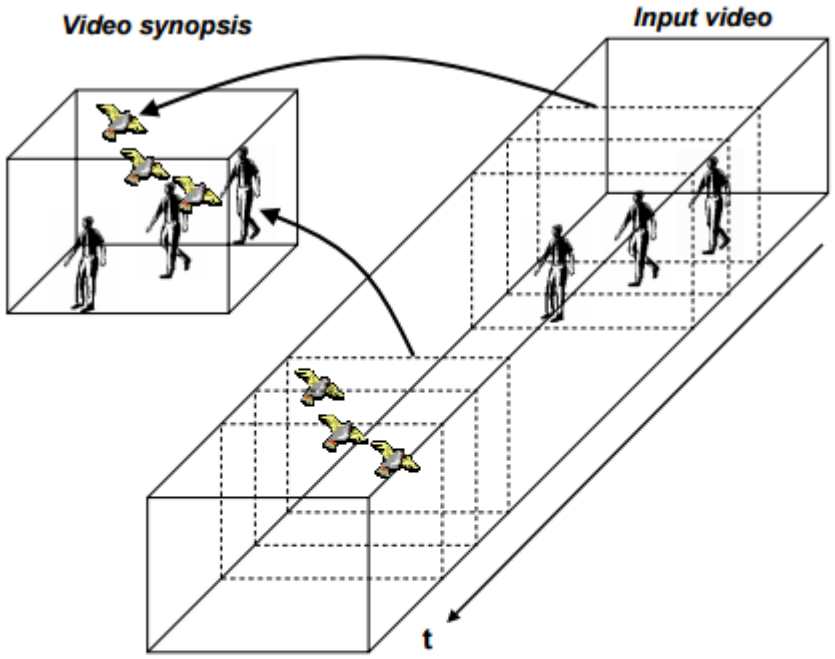
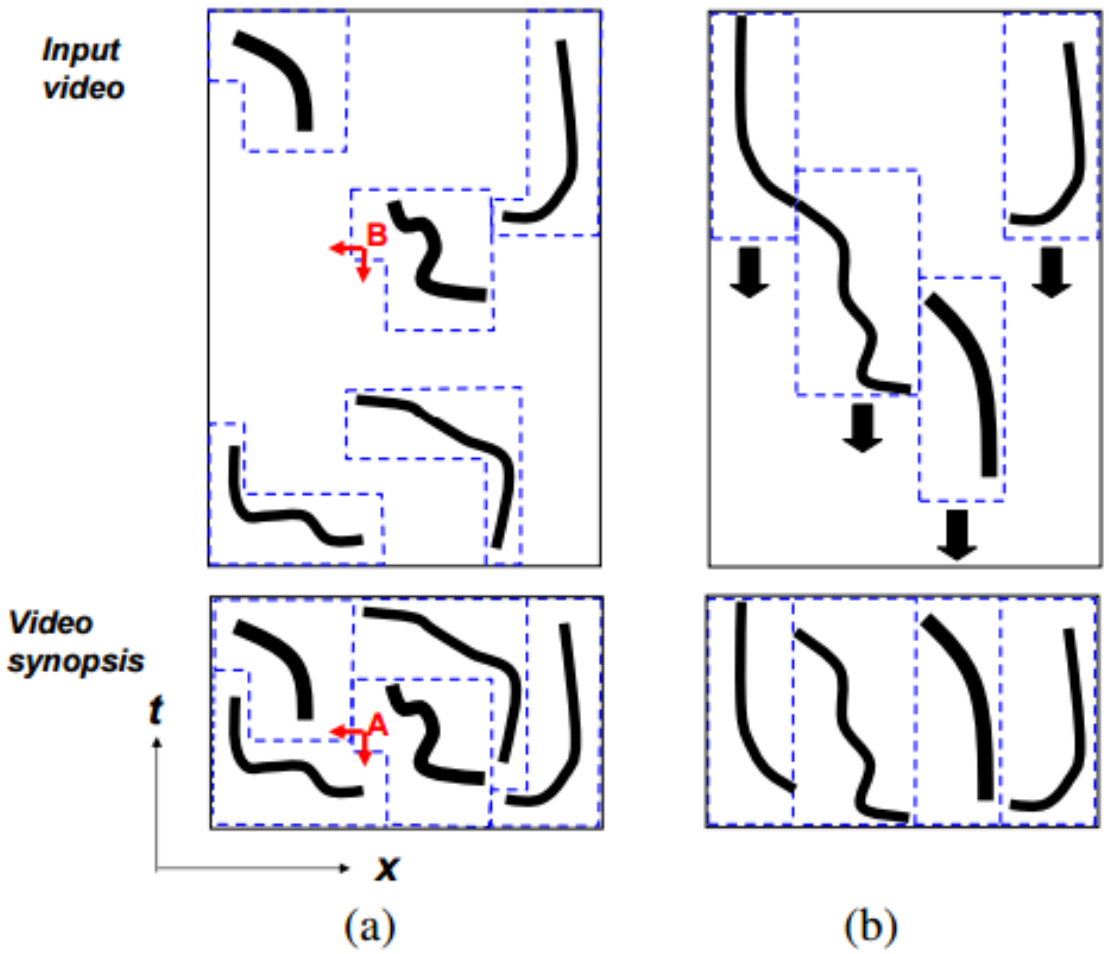


Fig. 1. The input video shows a walking person, and after a period of inactivity displays a flying bird. A compact video synopsis can be produced by playing the bird and the person simultaneously.

Approach

Fig. 4. In this space-time representation of video, moving objects create the "activity tubes". The upper part represents the original video I, while the lower part represents the video synopsis S.
(a) The shorter video synopsis S is generated from the input video I by including most active pixels together with their spatio-temporal neighborhood. To assure smoothness, when pixel A in S corresponds to pixel B in I, their "cross border" neighbors in space as well as in time should be similar.
(b) An approximate solution can be obtained by restricting consecutive synopsis pixels to come from consecutive input pixels.



Results

- Online phase requires less than one hour to process an hour of vides (for typical surveillance video)
- Queries returned on the order of minutes depending on POI (Period of Interest)

Fig. 9. Video synopsis with the dynamic stroboscopic effect as illustrated schematically in Fig. 2.b.
The video can be seen in *http://www.vision.huji.ac.il/video-synopsis*.

Examples

- http://www.vision.huji.ac.il/video-synopsis

Detecting Activities of Daily Living in First-Person Camera Views

Hamed Pirsiavash, Deva Ramanan

Department of Computer Science, University of California, Irvine

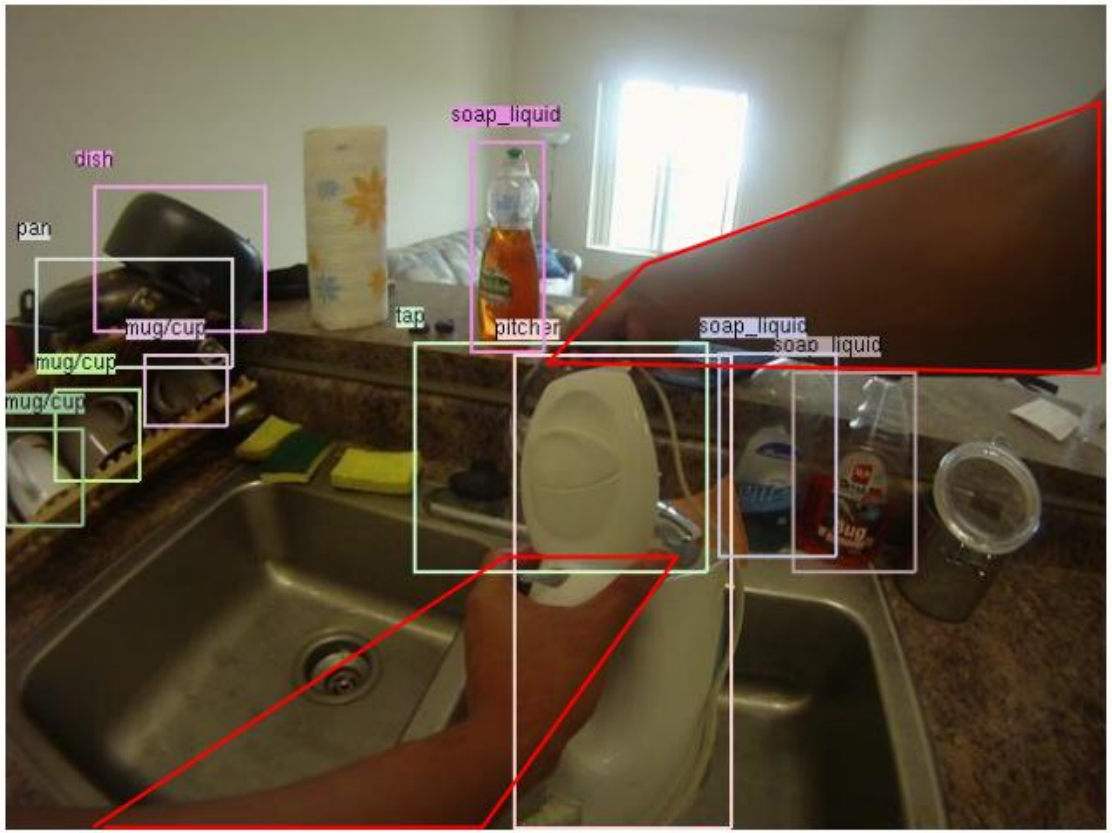


Figure 1: Activities of daily living (ADL) captured by a wearable camera.

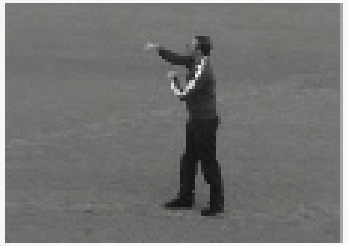
*slide courtesy of Piriavash and Ramanan

Motivation

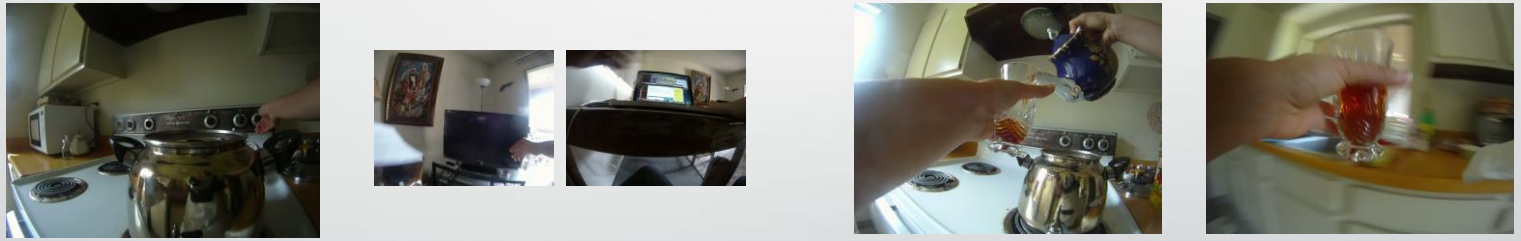
- Tele-rehabilitation
- Life-logging for patients with memory loss
- represent complex spatial-temporal relationships between objects
- Provide a large dataset of fully annotated ADLs

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

*slide courtesy of Piriavash and Ramanan

Features

t = particular frame
i = a single object
p = pixel location and scale
T = set of frames to be analyzed

- Identify object:

$$f_i^t = \max_p \ \text{score}_i^t(p) \tag{2}$$

- Aggregate features over time:

$$x_i^0 = \frac{1}{|T|} \sum_{t \in T} f_i^t \tag{3}$$

Temporal Pyramid

$x_i^{j,k}$ = *a histogram over a video clip*

j = depth of the pyramid (level)

- Generate temporal pyramid

$$x_i^{j,k} = \frac{2^{j-1}}{|T|} \sum_{t \in T^{j,k}} f_i^t; \quad \forall k \in \{1 \ldots 2^j\} \qquad (4)$$

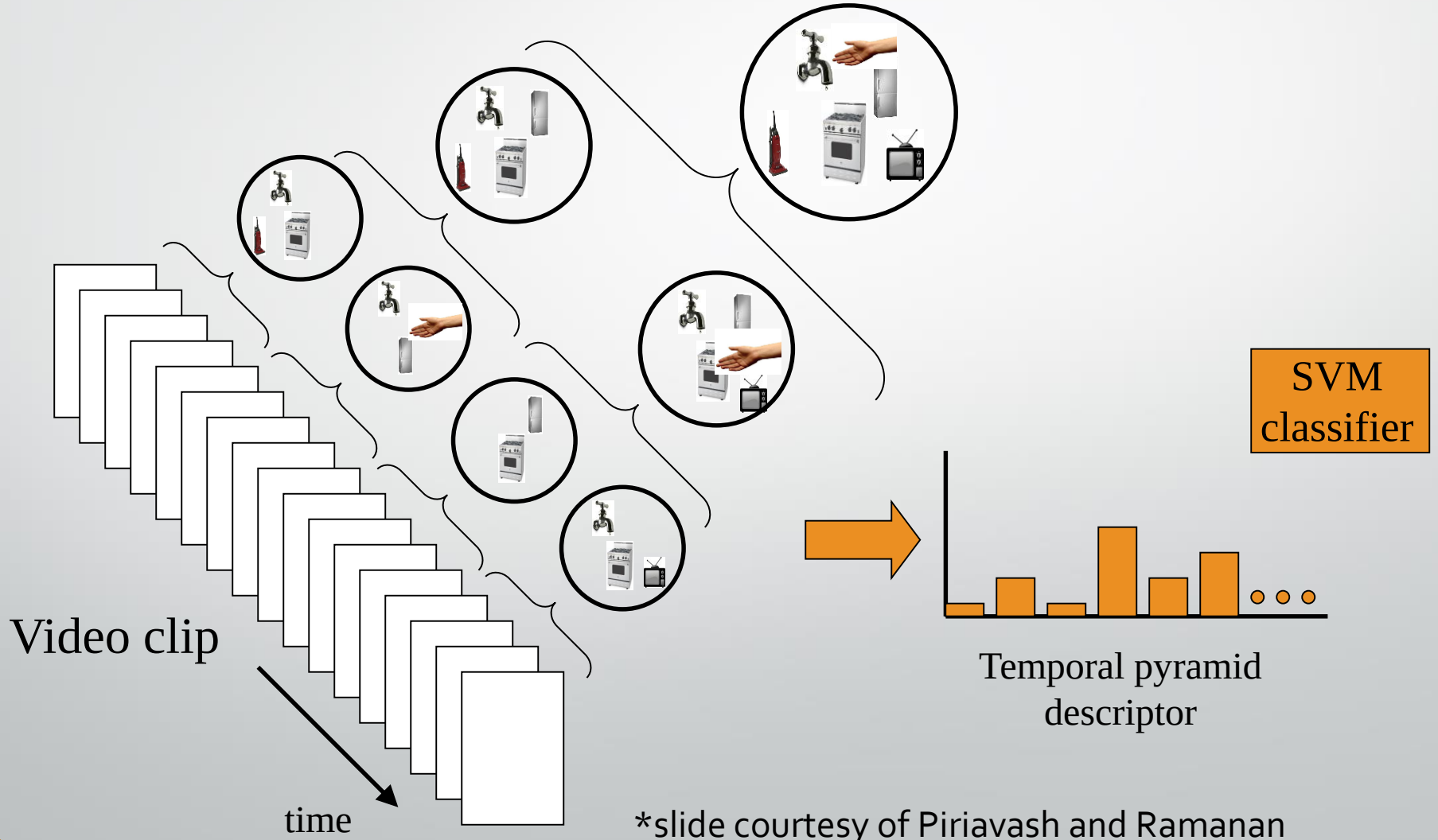
- Learn SVM classifiers on features for activity recognition:

$$x = \min\left(\left[x_1^0 \quad \ldots \quad x_i^{j,k} \ldots \quad x_K^{L,2^L} \right]^T, 0.01 \right)$$

Temporal pyramid

Coarse to fine correspondence matching with a multi-layer pyramid

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

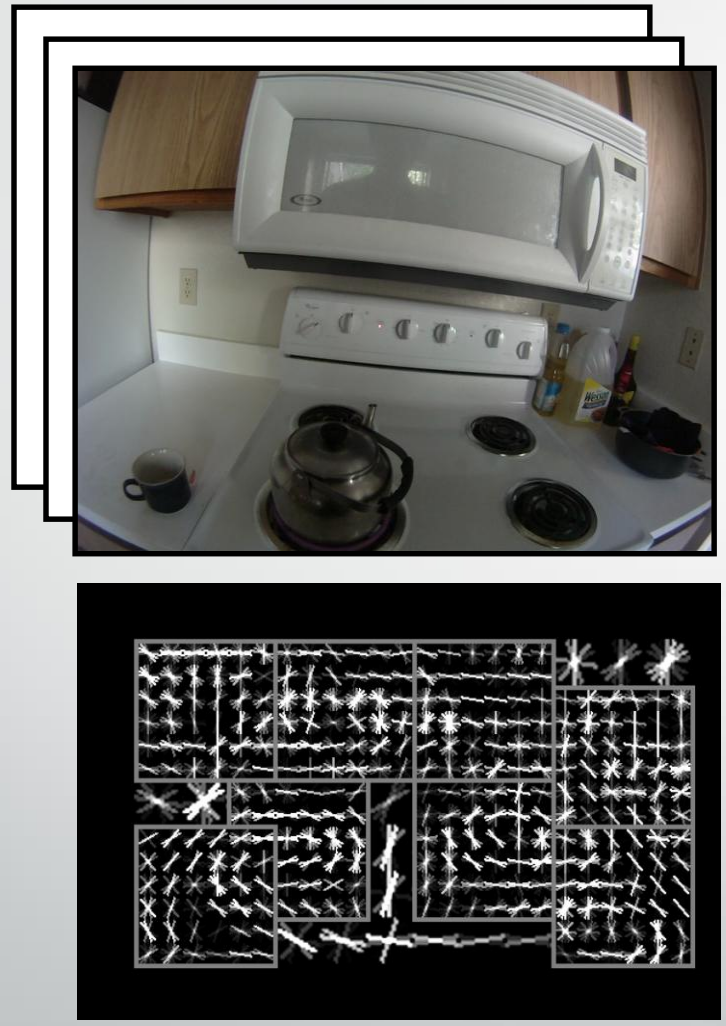


*slide courtesy of Piriavash and Ramanan

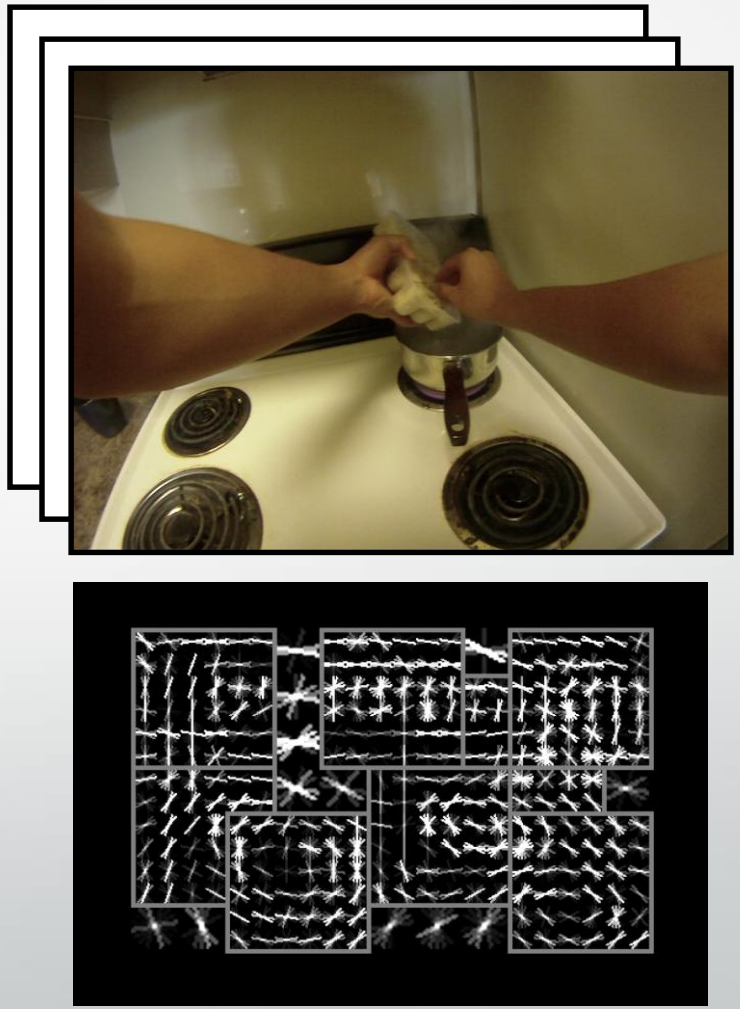
Active Object Models

- How to tell that an open fridge and a closed fridge are the same object?
- Train an additional object detector using the subset of “active” training images for a particular object

“Passive” vs “active” objects



Passive



Active

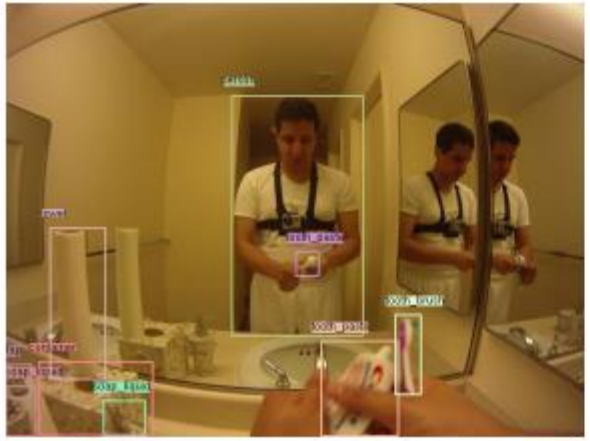
*slide courtesy of Piriavash and Ramanan

Dataset

- 20 people
- 30 minutes of footage a day
- 10 hours of footage per person
- 18 different identified ADLs

action name	mean of length (secs)	std. dev. of length
combing hair	26.50	9.00
make up	108.00	85.44
brushing teeth	128.86	45.50
dental floss	92.00	23.58
washing hands/face	76.00	36.33
drying hands/face	26.67	13.06
laundry	215.50	142.81
washing dishes	159.60	154.39
moving dishes	143.00	159.81
making tea	143.00	71.81
making coffee	85.33	54.45
drinking water/bottle	70.50	30.74
drinking water/tap	8.00	5.66
making cold food/snack	117.20	96.63
vacuuming	77.00	60.81
watching tv	189.60	98.74
using computer	105.60	32.94
using cell	18.67	9.45

ADL vs. Image-Net



Object	ADL	ImageNet
tap	40.4 ± 24.3	0.1
soap liquid	32.5 ± 28.8	2.5
fridge	19.9 ± 12.6	0.4
microwave	43.1 ± 14.1	20.2
oven/stove	38.7 ± 22.3	0.1
bottle	21.0 ± 27.0	9.8
kettle	21.6 ± 24.2	0.1
mug/cup	23.5 ± 14.8	14.8
washer/dryer	47.6 ± 15.7	1.8
tv	69.0 ± 21.7	26.9

Annotation

- 10 annotators, one annotation per 30 frames (1 second
  - Action Label
  - Object bounding box
  - Object identity
  - human-object interaction
- For co-occurring actions, the shorter interrupts the longer

Annotation

Figure 6: We show different kitchen scenes in our dataset. Unlike many other manually constructed action datasets, we exhibit a large variety of scenes and objects.

Functional Taxonomy

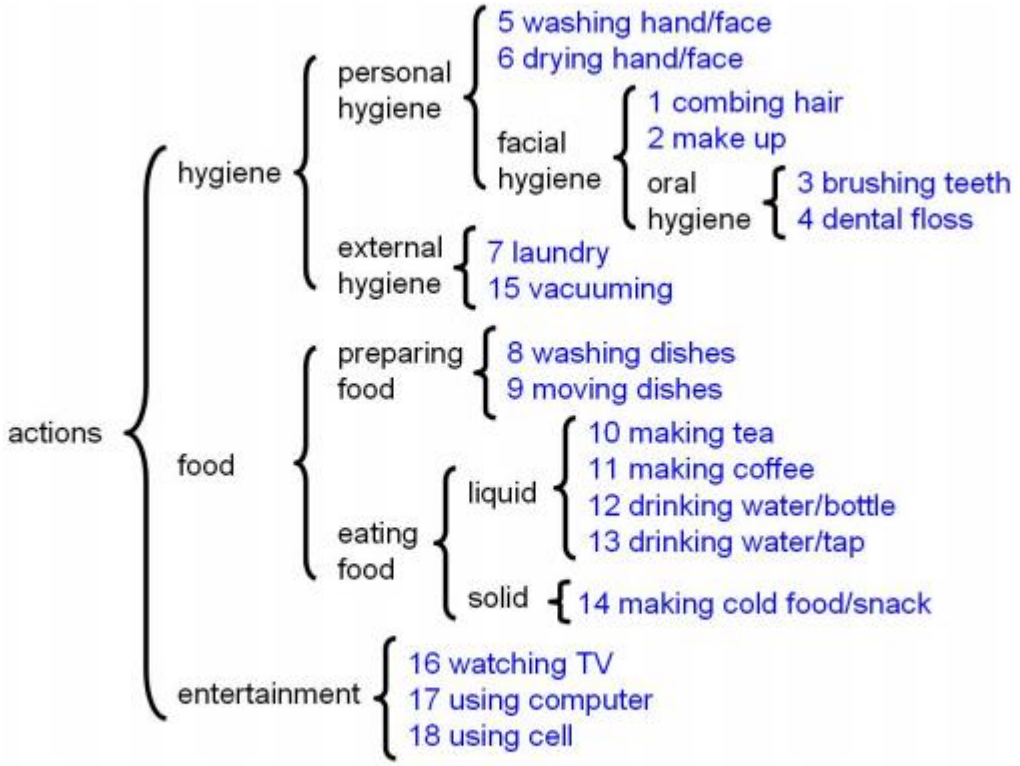


Figure 7: Our manually-designed functional ADL taxonomy.

Experiment

- Leave-one-out cross-validation
- Average precision
- Class confusion matrices for classification error and taxonomy-derived loss

Training

- Off-the-shelf parts model for object detection
- 24 object categories
- 1200 training instances
- Inherent differences between training datasets:

Action Recognition results

Space-time interest points (STIP)
Bag-of-objects model (O)
Active-object model (AO)
Idealized perfect object detectors (IO)
Augmented Idealized object detectors (IA+IO)

	pre-segmented			
	segment 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

	sliding window			
	frame class. accuracy		taxonomy loss	
	pyramid	bag	pyramid	bag
STIP	15.6	12.9	2.1957	2.1997
O	23.8	17.4	1.5975	1.8123
AO	28.8	23.9	1.5057	1.6515
IO	43.5	36.6	1.1047	1.2859
IA+IO	60.7	53.7	0.79532	0.9551

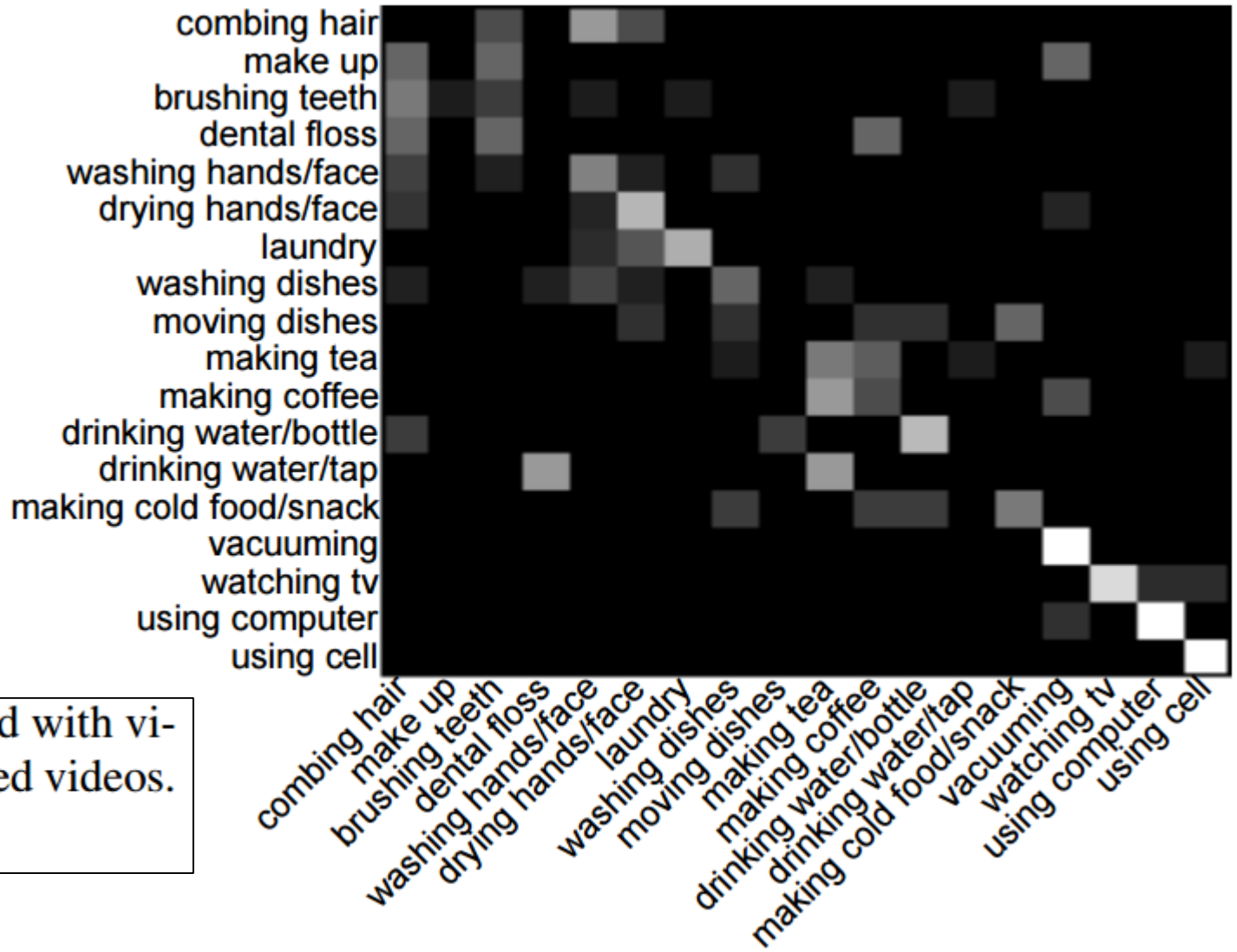


Figure 10: Confusion matrix for temporal pyramid with vision based active object detectors on pre-segmented videos. Segment classification accuracy = 40.6%.

Discussion

- Limitations?
- Future Work?