Detecting Activities of Daily Living in First-person Camera Views

Hamed Pirsiavash, Deva Ramanan Univesity of California, Irvine

Activities of Daily Living (ADL)

- In healthcare
 - daily self-care activities within an individual's place of residence, in outdoor environments, or both
 - Feeding oneself
 - Maintaining hygiene
 - Dressing and undressing

Visual Activity Recognition

- Automated analysis of ongoing events and their context in videos or still images.
 - Human activity recognition
 - Single human actions
 - Human human interaction
 - Human-Object interaction
 - Group activities
 - General events











Detecting ADL

- Mostly human interacting with objects
 - tv, sofa, tooth brush, food, car, stove, ...
- Applications
 - Tele-rehabilitation
 - Evaluate everyday functional activities
 - Long term, efficient monitoring
 - Life logging
 - Visual history or memory
 - Large scale
- "It is all about objects being interacted with"

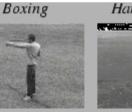
Other Datasets

- Actor-scripted video footage
- Movies
- Sports
- Videos in the wild













hooting













Skateboarding

Swing-Bench

Swing-Side

This dataset

- Hard to define canonical activities
 - ADL from medical literature and rehabilitation
- Hard to capture intra class variation
 - Wearable camera on different persons
- Wearable camera (GoPro)
 - HD (1280x960)
 - 170 fov
 - 30 hz



This dataset

- 20 people in their own apartments
- 18 actions
- Unscripted
- 10 hours
- Annotation
 - Action label
 - Object bounding box
 - Object identity
 - tracking
 - Interaction
 - active/passive

action name	mean of	std. dev. of
	length (secs)	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

This dataset - characteristics

Large variation in scenes and objects

























This dataset - charactristcs

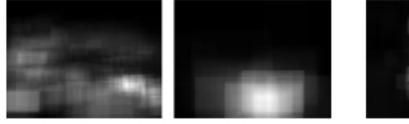
Various object view point and occlusion level



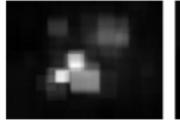
This dataset - characteristics

- Biases
 - active/passive objects
 - Location
 - pose





pan









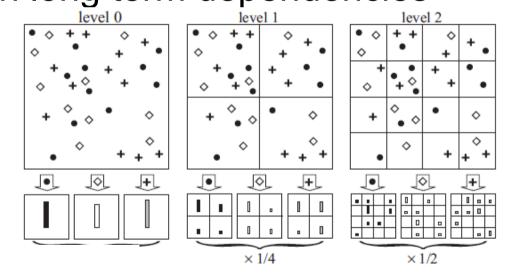
mug/cup

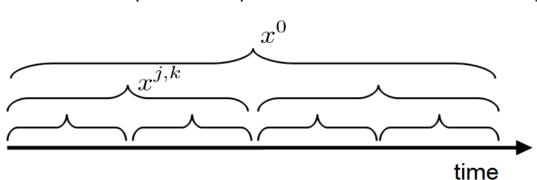
dish

This dataset - characteristics Inherent functional taxonomy 5 washing hand/face 6 drying hand/face personal 1 combing hair hygiene combing hair 2 make up facial hygiene make up hygiene oral 3 brushing 1 hygiene 4 dental flos brushing teeth external 7 laundry hygiene 15 vacuuming dental floss washing hands/face 8 washing dishes preparing 9 moving dishes food drying hands/face actions 10 making tea laundry food 11 making coffee liquid washing dishes 12 drinking water/bottle eating 13 drinking water/tap moving dishes food solid 14 making cold food/snack making tea 16 watching TV making coffee entertainment -17 using computer drinking water/bottle 18 using cell drinking water/tap making cold food/snack vacuuming watching tv using computer using cell doing nothing using colling ashing lands laco dying rands lace bushing toth washing disters noinguistes TINING WARE TOTHO dinking watertag Ing cold lood spect using computer makingles Vacuming combing hair Watchingh

Learning - Features

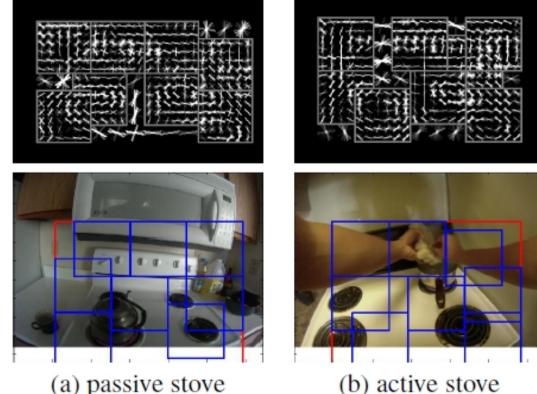
- Temporal pyramids
 - Spatial pyramids
 - Visual words \rightarrow Object models
 - Modelling actions with long term dependencies
 - Felzenswalb's DPM
 - Location bias





Learning - Features

- Active Object Models
 - Objects look different when being interacted with
 - Detection of visual phrases, Farhad et. al. (cvpr 11)

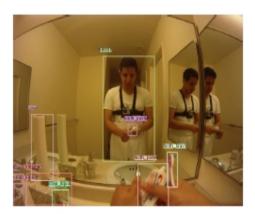


Experiments

- Same individual does not appear in train and test
- Leave one out cross validation testing
- AP for object detection
- Action recognition
 - Classification error
 - Weighted taxonomy derived loss

Experiments object detection

- 24 objects
- 1200 bbox/object
- low number of instances
- High variation in viewpoint and occlusion state





ImageNet

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

Experiments – action recognition

- Spatio-temporal interest points(stip)
- Object bag(O)
- Active object bag(AO)
- Ideal object detector(IO)
- Ideal active/passive object (IA + IO)
- Pre-segmented or sliding window
- Accuracy or taxonomy loss

	pre-segmented			
	segment class. accuracy		taxonomy loss	
	pyramid	bag	pyramid	bag
STIP	22.8	16.5	1.8792	2.1092
0	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	
0	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	

• "It is all about objects being interacted with"

Experiments – action recognition

www.vacuuming.comping.com using.comping.com

making cold tood shack

auro

Washingdist

moving

drinking

drying hands te

washing hands

brushint

- Small objects
- Scene based features

combing hair make up brushing teeth dental floss washing hands/face drving hands/face laundry washing dishes moving dishes making tea making coffee drinking water/bottle drinking water/tap making cold food/snack vacuuming watching ty using computer using cell

summary

- Most human activities involve objects
- Good detection of object and its state (active/passive) help activity recognition a lot
- Naturally captured datasets is more realistic...
- Objects appear visually different in various scenarios due to occlusion and interactions