Detecting activities of daily living in first person camera views

Hamed Pirsiavash, Deva Ramanan

Presented by Dinesh Jayaraman

Wearable ADL detection

It is easy to collect natural data

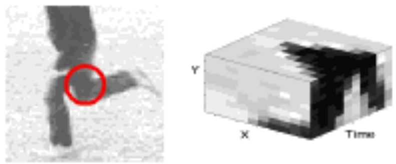
ADL actions derived from medical literature on patient rehabilitation

Slides from authors (link)

Method - Choice of features

Low level features

(Weak semantics)



Space-time interest points

Laptev, IJCV'05

High level features

(Strong semantics)

Human pose

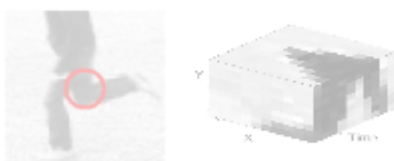
Difficulties of pose:
- Detectors are not accurate enough
- Not useful in first person camera views

Slides from authors (link)

Method - Choice of features

Low level features

(Weak semantics)



Space-time interest points

Laptev, IJCV'05

High level features

(Strong semantics)

Human pose

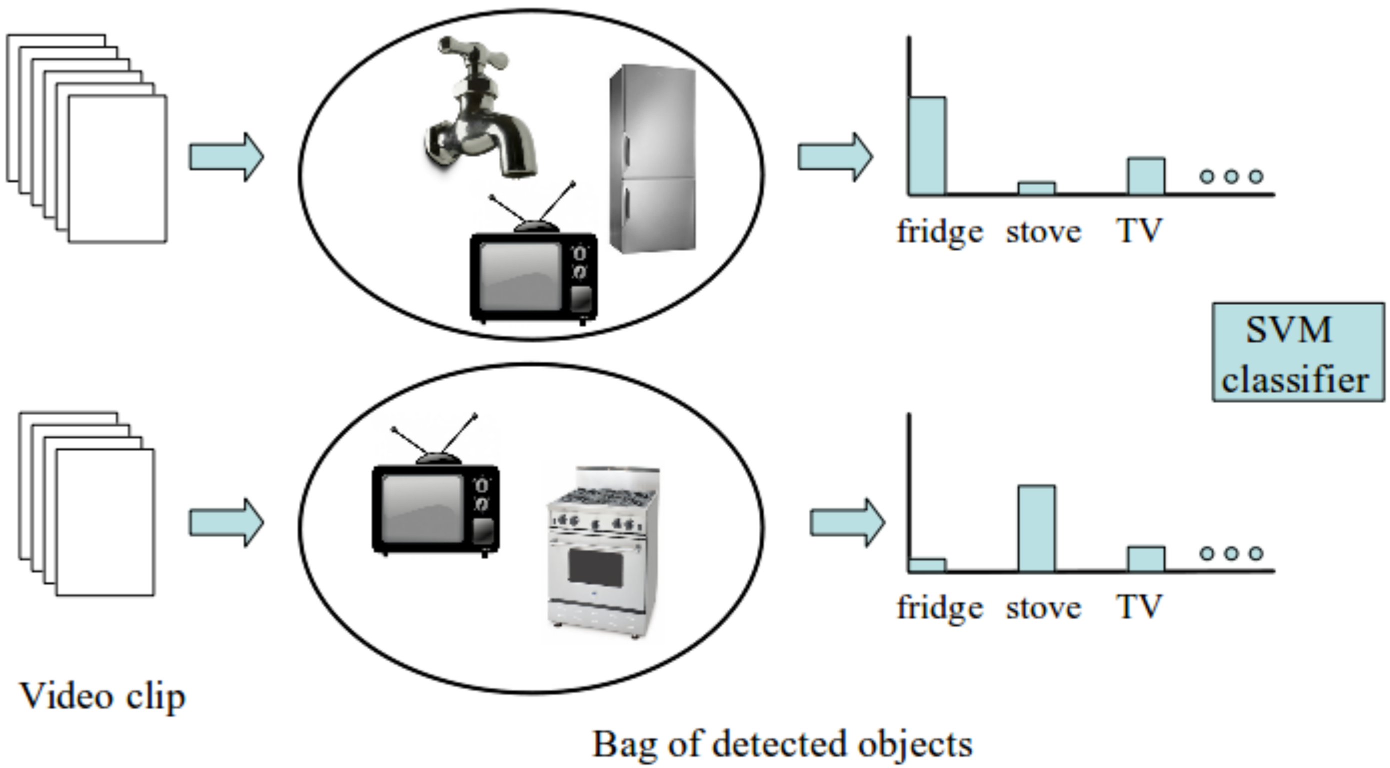
Object-centric features

Difficulties of pose:

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

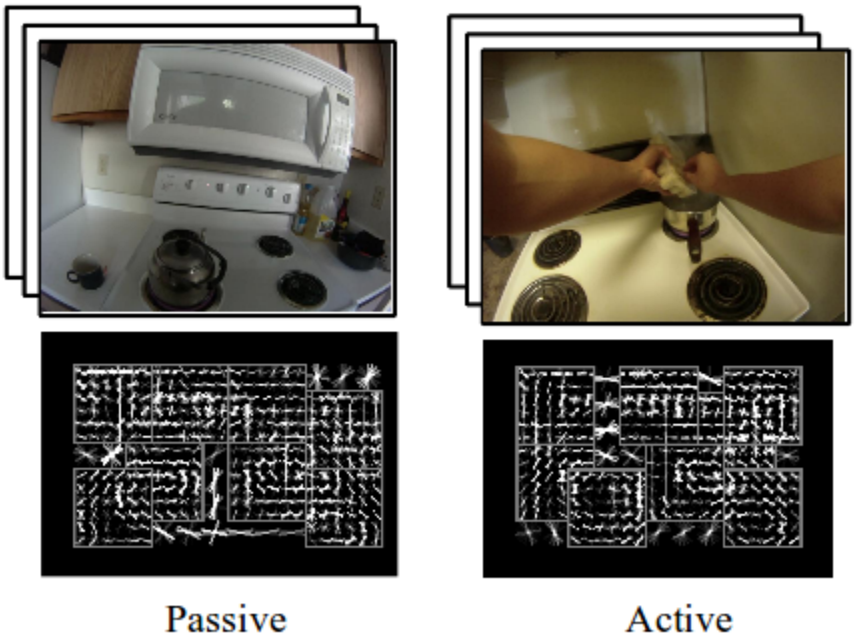
Slides from authors (link)

Bag of objects



Slides from authors (link)

Method - Active/Passive objects

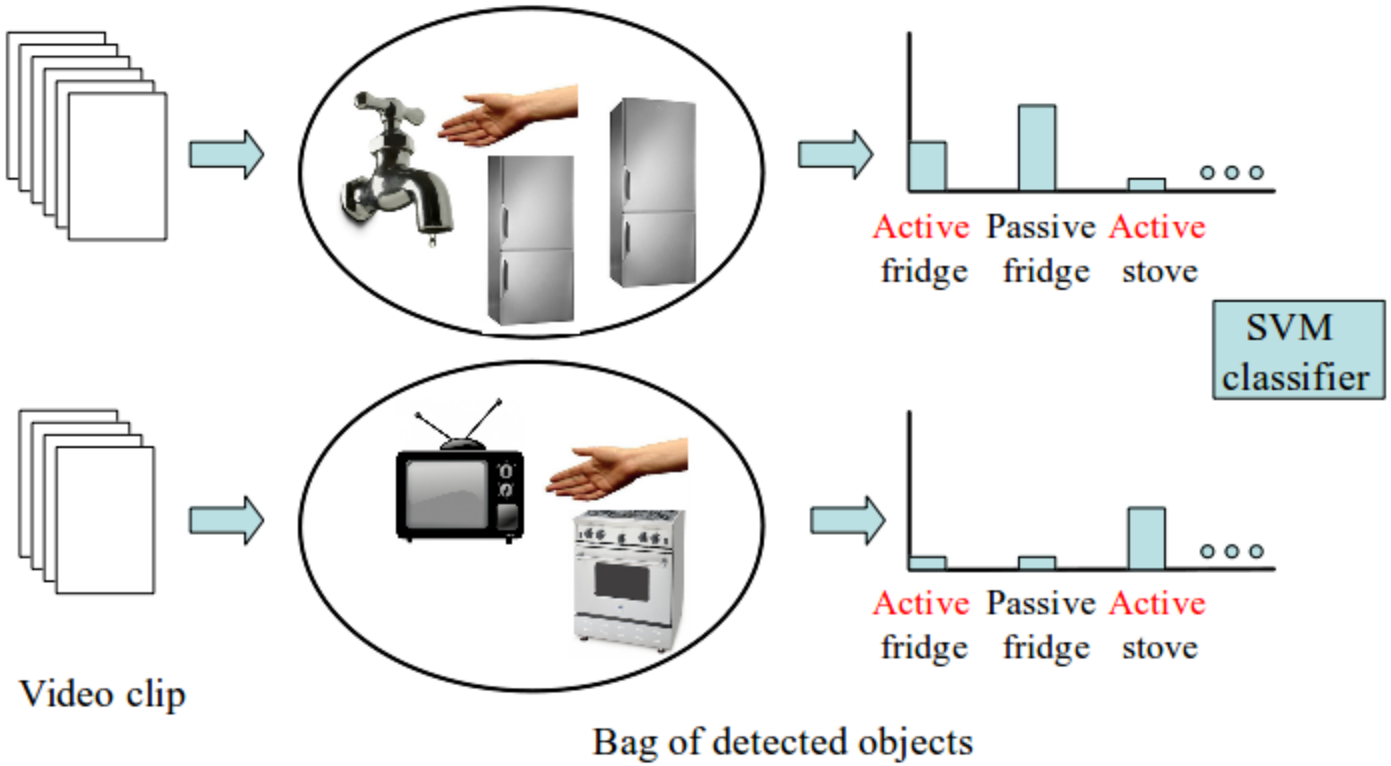


Better object detection (visual phrases CVPR'11)

Better features for action classification (active vs passive)

Slides from authors ([link]())

Method - Active/Passive objects



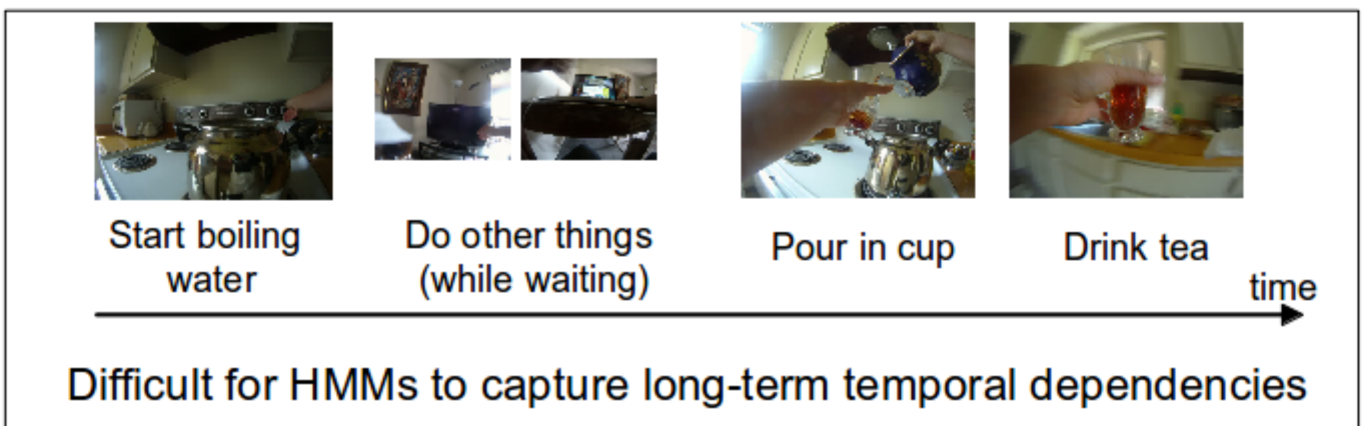
Slides from authors (link)

Method - Temporal pyramid

long-scale temporal structure

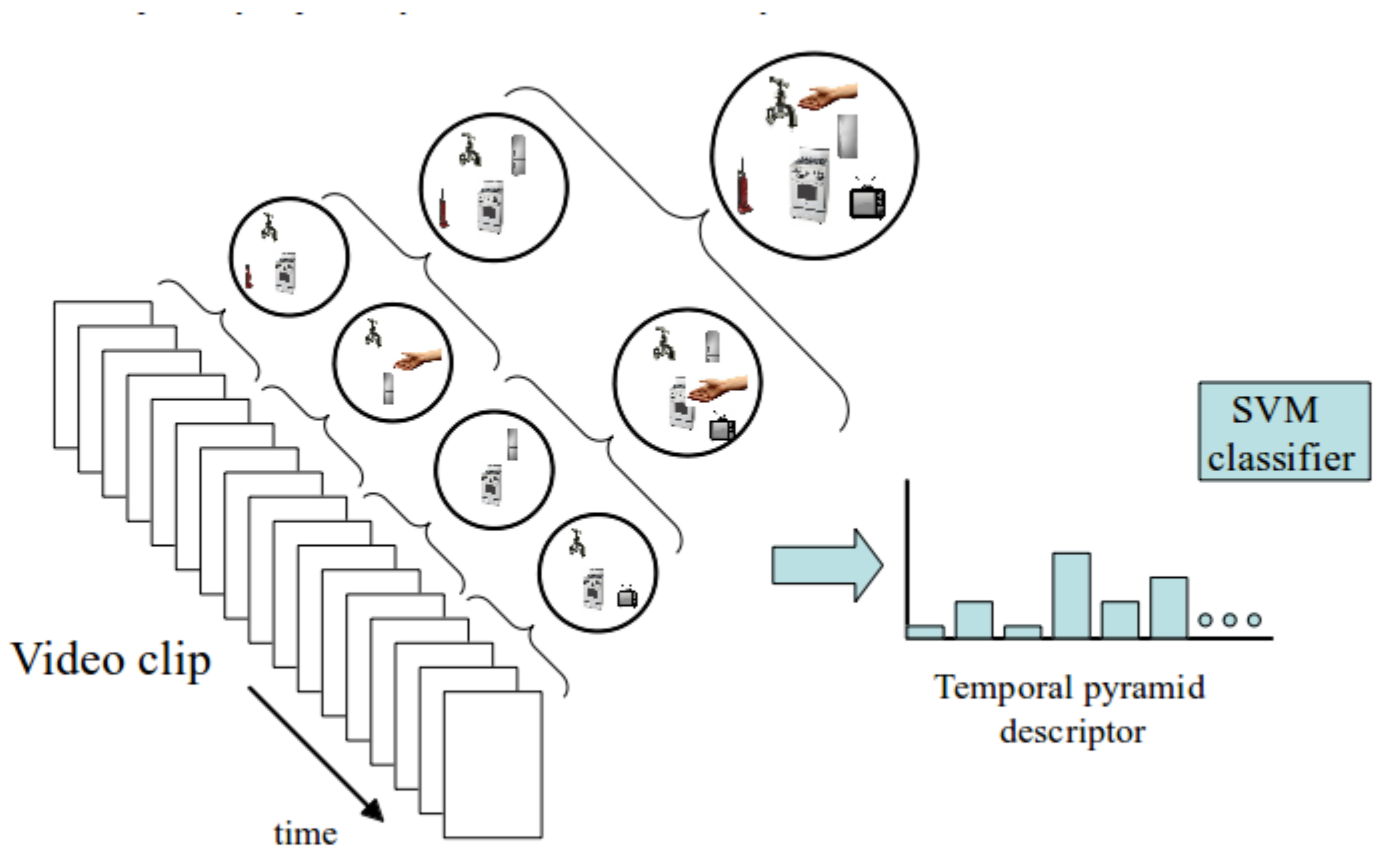
“Classic” data: boxing

Wearable data: making tea



Slides from authors (link)

Method - Temporal pyramid



Slides from authors (link)

Data

- 40 GB of video data
- Annotations
  - Object annotations
  - 30-frame intervals
  - Present/absent
  - Bounding boxes
  - Active/passive
- Action annotations
  - Start time, end time
- Pre-computed:
  - DPM object detection outputs
  - Active/passive models

Examples

Implementation differences

Temporal pyramid is not really implemented as a pyramid - linear SVM in place of kernel SVM

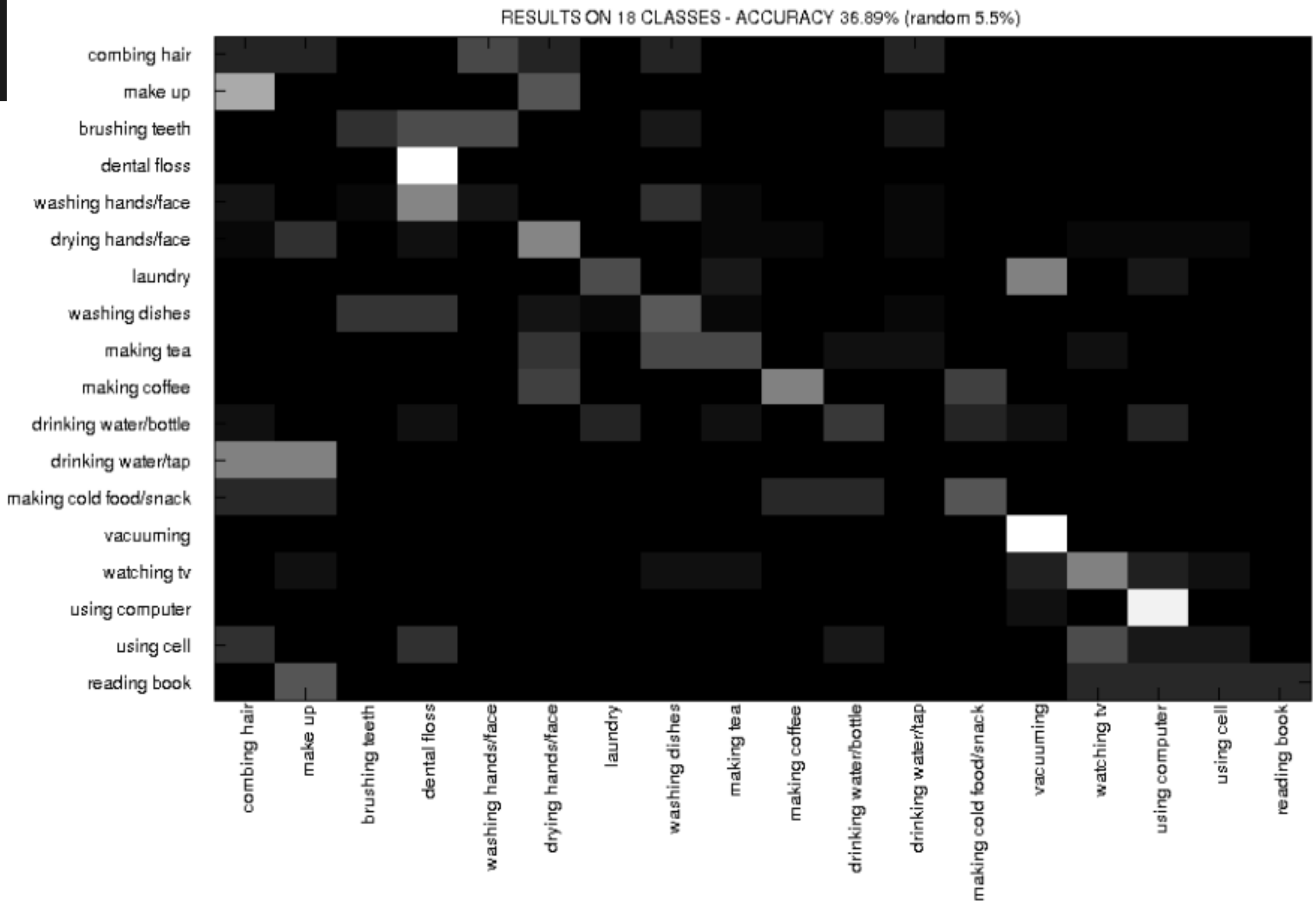
Locations are not used

Recap - Key ideas

- Bag-of-objects representation (instead of low-level STIP-type approach)
- Separate models for active/passive objects
- Temporal pyramid

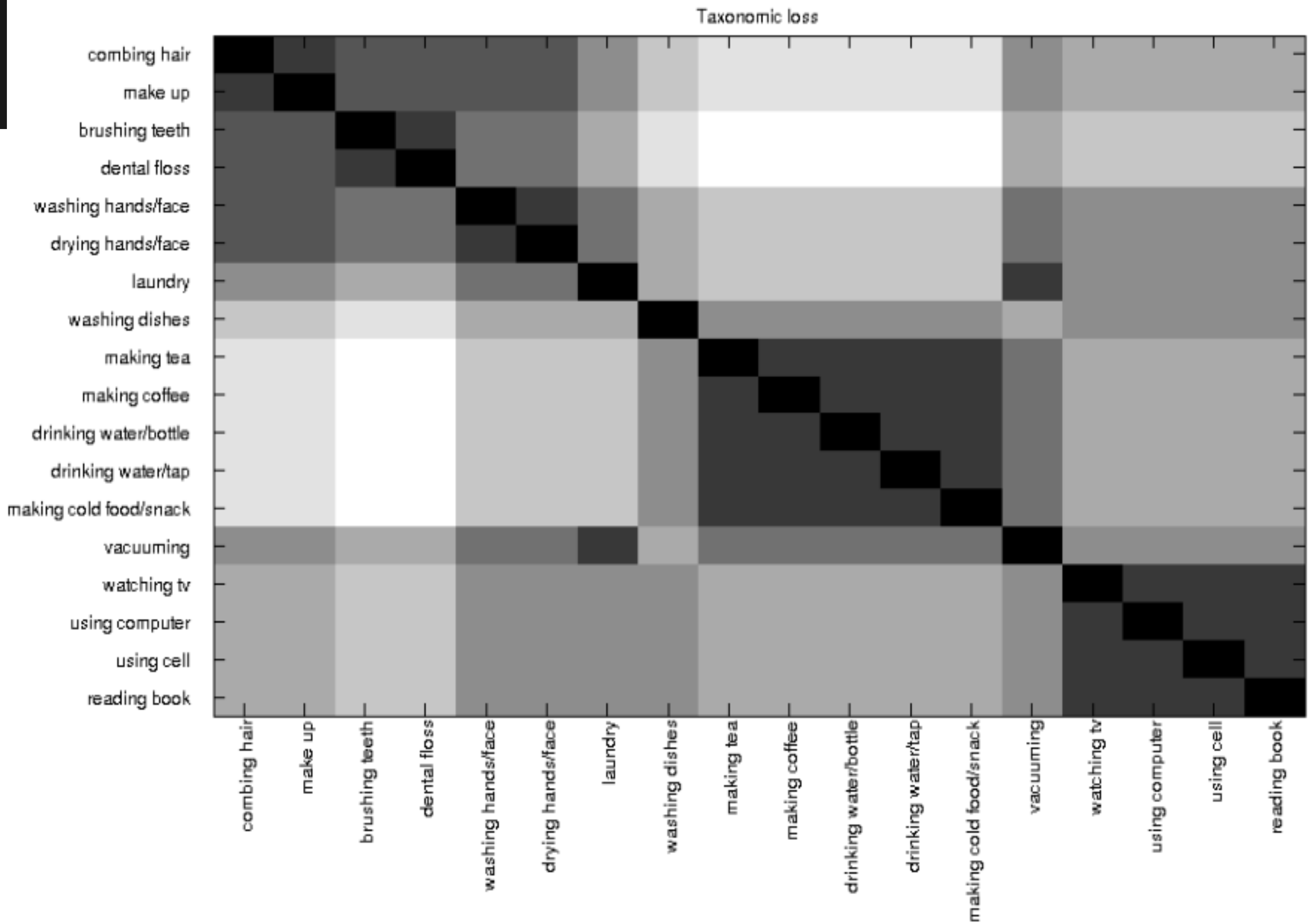
We will now study the impact of each of these

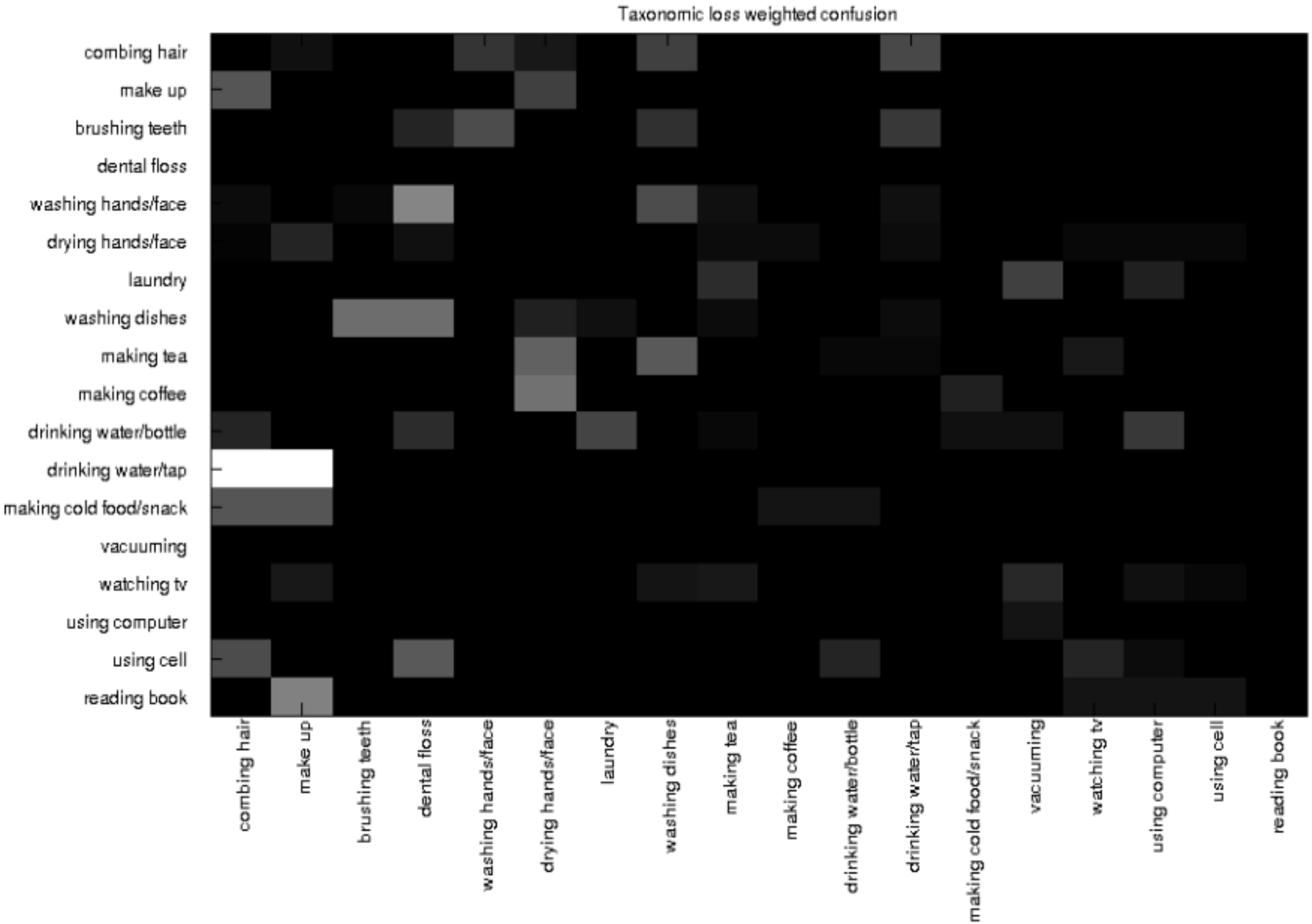
Accuracy- 37%



RESULTS ON 18 CLASSES - ACCURACY 36.89% (random 5.5%)
combing hair
make up
brushing teeth
dental floss
washing hands/face
drying hands/face
laundry
washing dishes
making tea
making coffee
drinking water/bottle
drinking water/tap
making cold food/snack
vacuuming
watching tv
using computer
using cell
reading book

Taxonomic loss function





Taxonomic loss weighted confusion
combing hair
make up
brushing teeth
dental floss
washing hands/face
drying hands/face
laundry
washing dishes
making tea
making coffee
drinking water/bottle
drinking water/tap
making cold food/snack
vacuuming
watching tv
using computer
using cell
reading book

Understanding data - 32 ADL actions, 18 selected

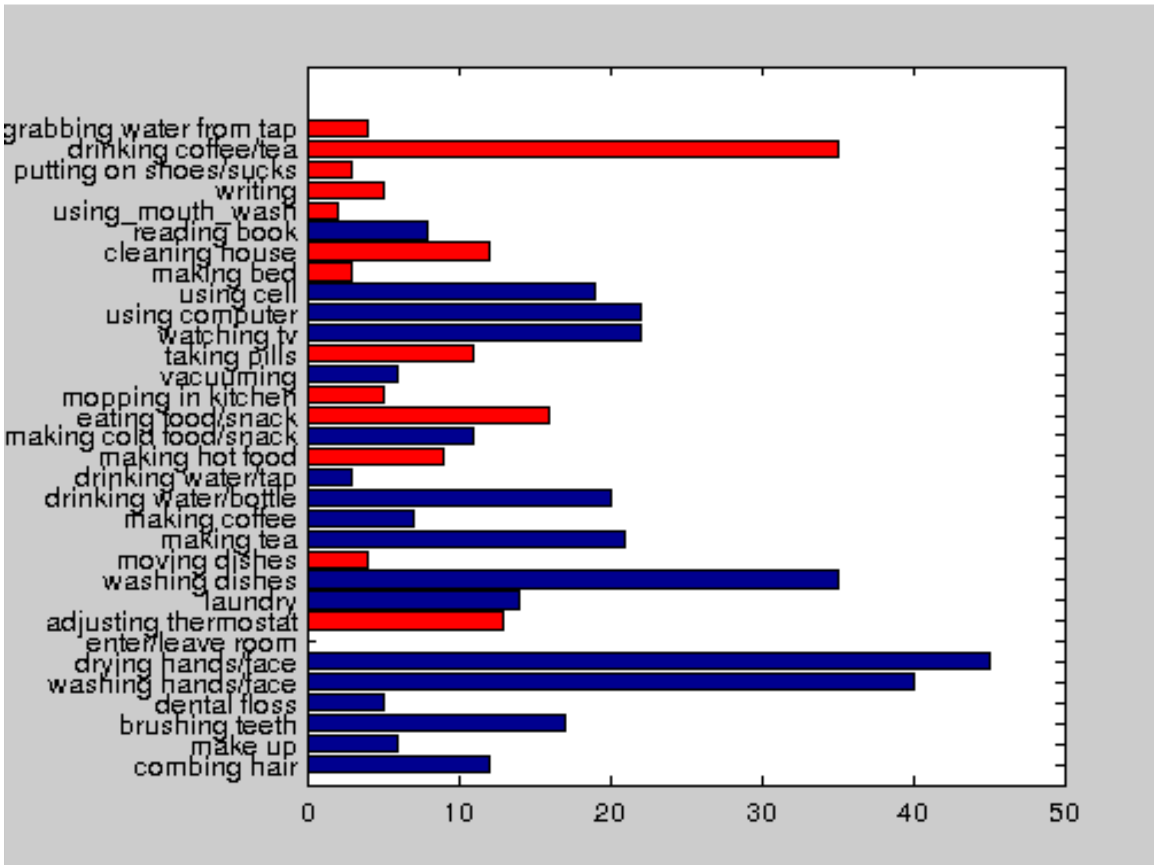
- 'combing hair'
- 'make up'
- 'brushing teeth'
- 'dental floss'
- 'washing hands/face'
- 'drying hands/face'
- 'enter/leave room'
- 'adjusting thermostat'
- 'laundry'
- 'washing dishes'
- 'moving dishes'
- 'making tea'
- 'making coffee'
- 'drinking water/bottle'
- 'drinking water/tap'
- 'making hot food'
- 'making cold food/snack'
- 'eating food/snack'
- 'mopping in kitchen'
- 'vacuuming'
- 'taking pills'
- 'watching tv'
- 'using computer'
- 'using cell'
- 'making bed'
- 'cleaning house'
- 'reading book'
- 'using_mouth_wash'
- 'writing'
- 'putting on shoes/sucks'
- 'drinking coffee/tea'
- 'grabbing water from tap'

Understanding data - 32 ADL actions, 18 selected

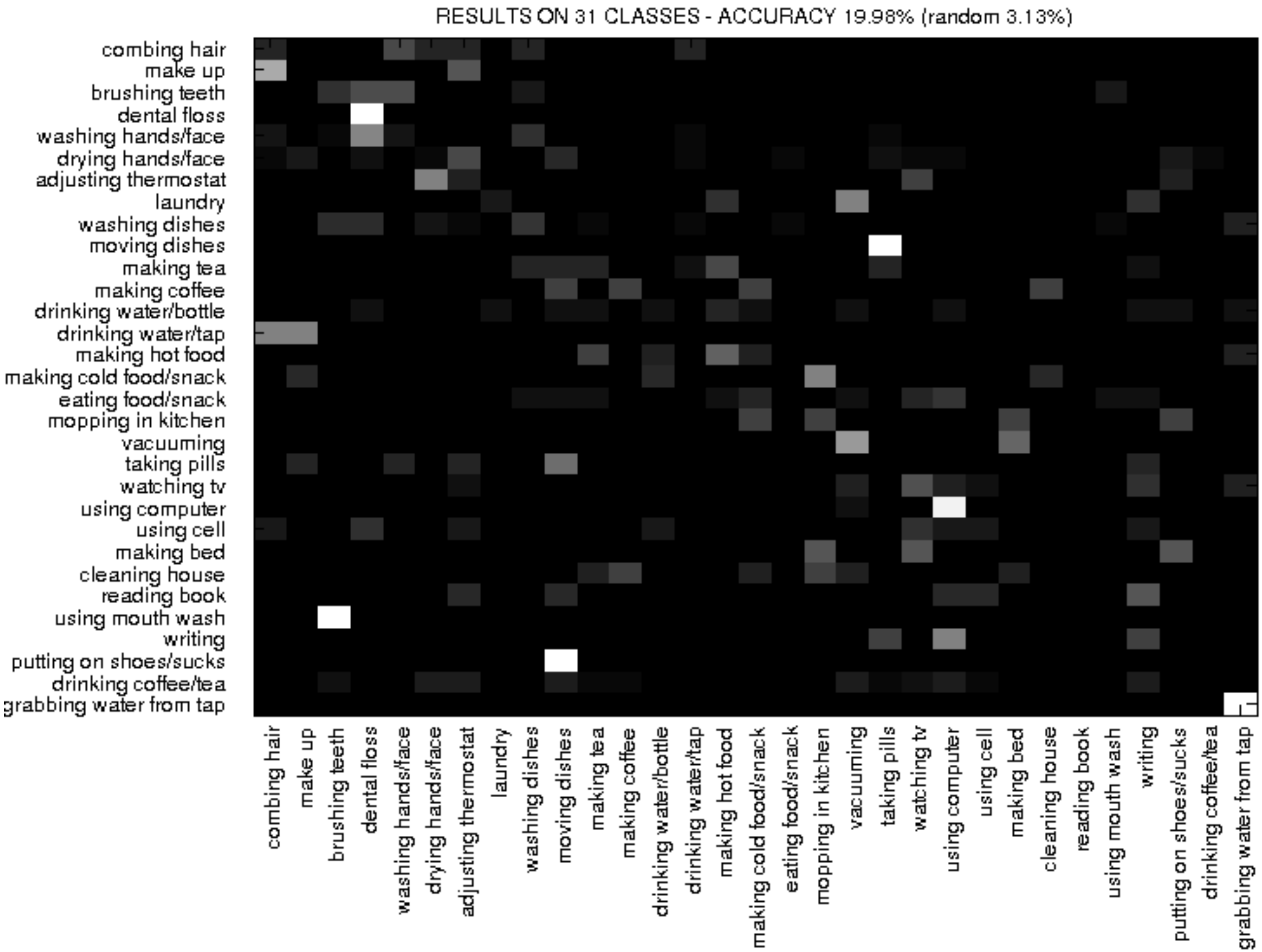
- 'combing hair'
- 'make up'
- 'brushing teeth'
- 'dental floss'
- 'washing hands/face'
- 'drying hands/face'
- 'enter/leave room'
- 'adjusting thermostat'
- 'laundry'
- 'washing dishes'
- 'moving dishes'
- 'making tea'
- 'making coffee'
- 'drinking water/bottle'
- 'drinking water/tap'
- 'making hot food'
- 'making cold food/snack'
- 'eating food/snack'
- 'mopping in kitchen'
- 'vacuuming'
- 'taking pills'
- 'watching tv'
- 'using computer'
- 'using cell'
- 'making bed'
- 'cleaning house'
- 'reading book'
- 'using_mouth_wash'
- 'writing'
- 'putting on shoes/sucks'
- 'drinking coffee/tea'
- 'grabbing water from tap'

Data available for actions

Number of instances in data



Not a data issue


RESULTS ON 31 CLASSES - ACCURACY 19.98% (random 3.13%)
combing hair
make up
brushing teeth
dental floss
washing hands/face
drying hands/face
adjusting thermostat
laundry
washing dishes
moving dishes
making tea
making coffee
drinking water/bottle
drinking water/tap
making hot food
making cold food/snack
eating food/snack
mopping in kitchen
vacuuming
taking pills
watching tv
using computer
using cell
making bed
cleaning house
reading book
using mouth wash
writing
putting on shoes/sucks
drinking coffee/tea
grabbing water from tap

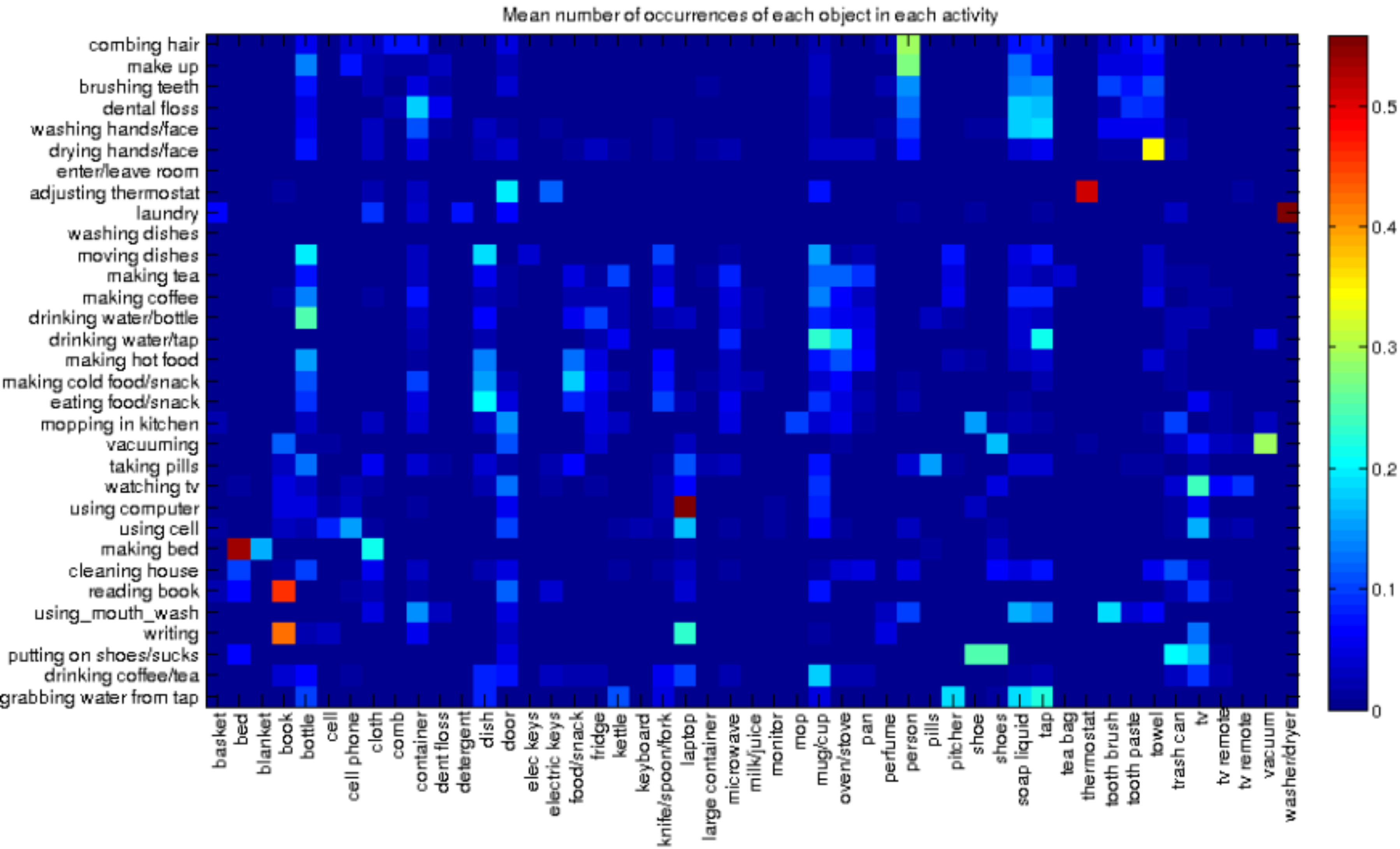
Results

Method	Accuracy
DPM | act +pass | 2 temp levels	**19.98%**

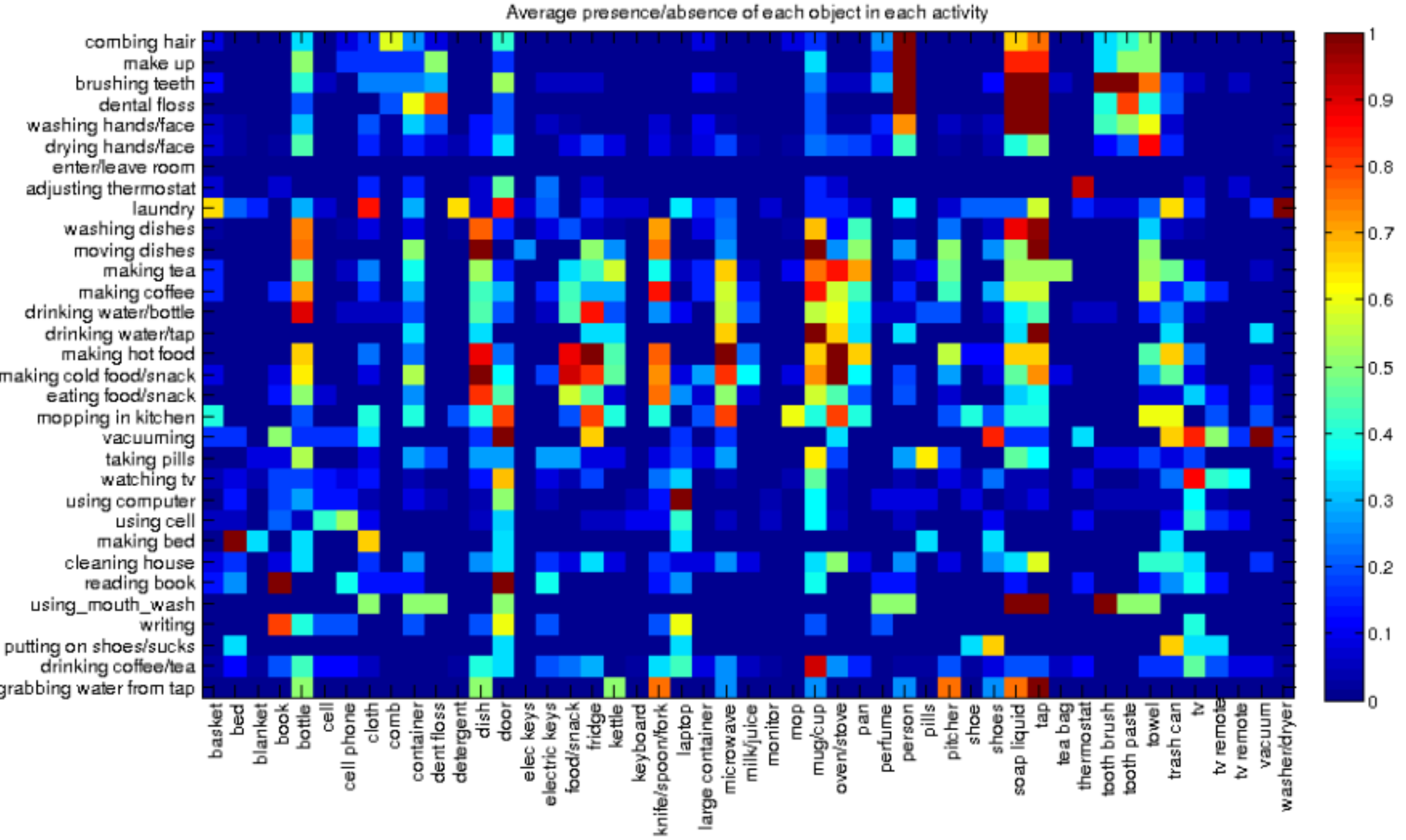
What does each stage contribute?

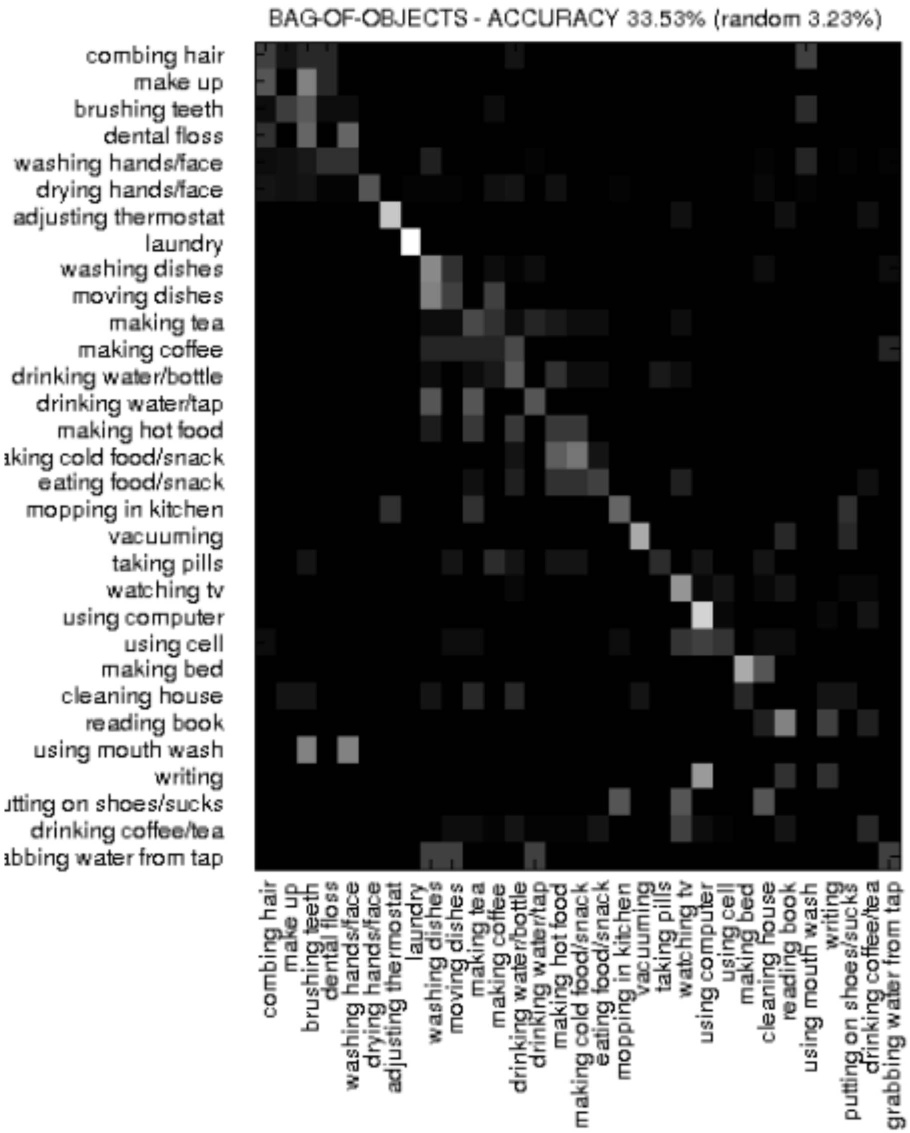
- Bag-of-objects
- Bag-of-active/passive objects
- Bag-of-active/passive objects with temporal ordering

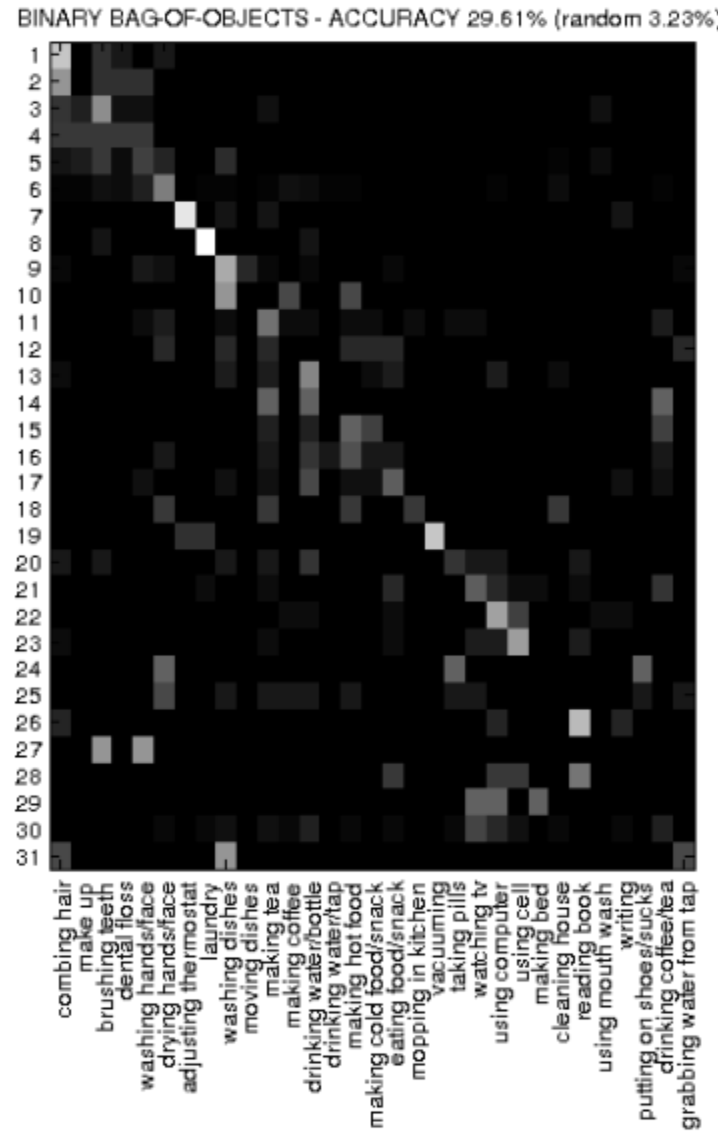
Object occurence



Object presence




BAG-OF-OBJECTS - ACCURACY 33.53% (random 3.23%)
combing hair
make up
brushing teeth
dental floss
washing hands/face
drying hands/face
adjusting thermostat
laundry
washing dishes
moving dishes
making tea
making coffee
drinking water/bottle
drinking water/tap
making hot food
aking cold food/snack
eating food/snack
mopping in kitchen
vacuuming
taking pills
watching tv
using computer
using cell
making bed
cleaning house
reading book
using mouth wash
writing
utting on shoes/sucks
drinking coffee/tea
abbing water from tap
grabbing water from tap
putting on shoes/sucks
making cold food/snack


BINARY BAG-OF-OBJECTS - ACCURACY 29.61% (random 3.23%)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
combing hair
make up
brushing teeth
dental floss
washing hands/face
drying hands/face
adjusting thermostat
laundry
washing dishes
moving dishes
making tea
making coffee
drinking water/bottle
drinking water/tap
making hot food
making cold food/snack
eating food/snack
mopping in kitchen
vacuuming
taking pills
watching tv
using computer
using cell
making bed
cleaning house
reading book
using mouth wash
writing
putting on shoes/sucks
drinking coffee/tea
grabbing water from tap

Results

Method	Accuracy
DPM | act.+pas.| 2 temp levels	19.98%
Ideal | no activity info | no ord.	**29.61%**

Thresholded bag-of-objects

- Object presence duration is an important cue, but
  - has large variance
  - assumes objects with large presence duration are also important for discrimination
- Binary approach counters these shortcomings but
  - loses object presence duration cues
  - susceptible to noise without ground truth data. Even one false positive will have large impact.

Thresholded bag-of-objects

- Thresholded bag-of-objects features compromise
  - less noisy
  - retains information about which objects are more and less important

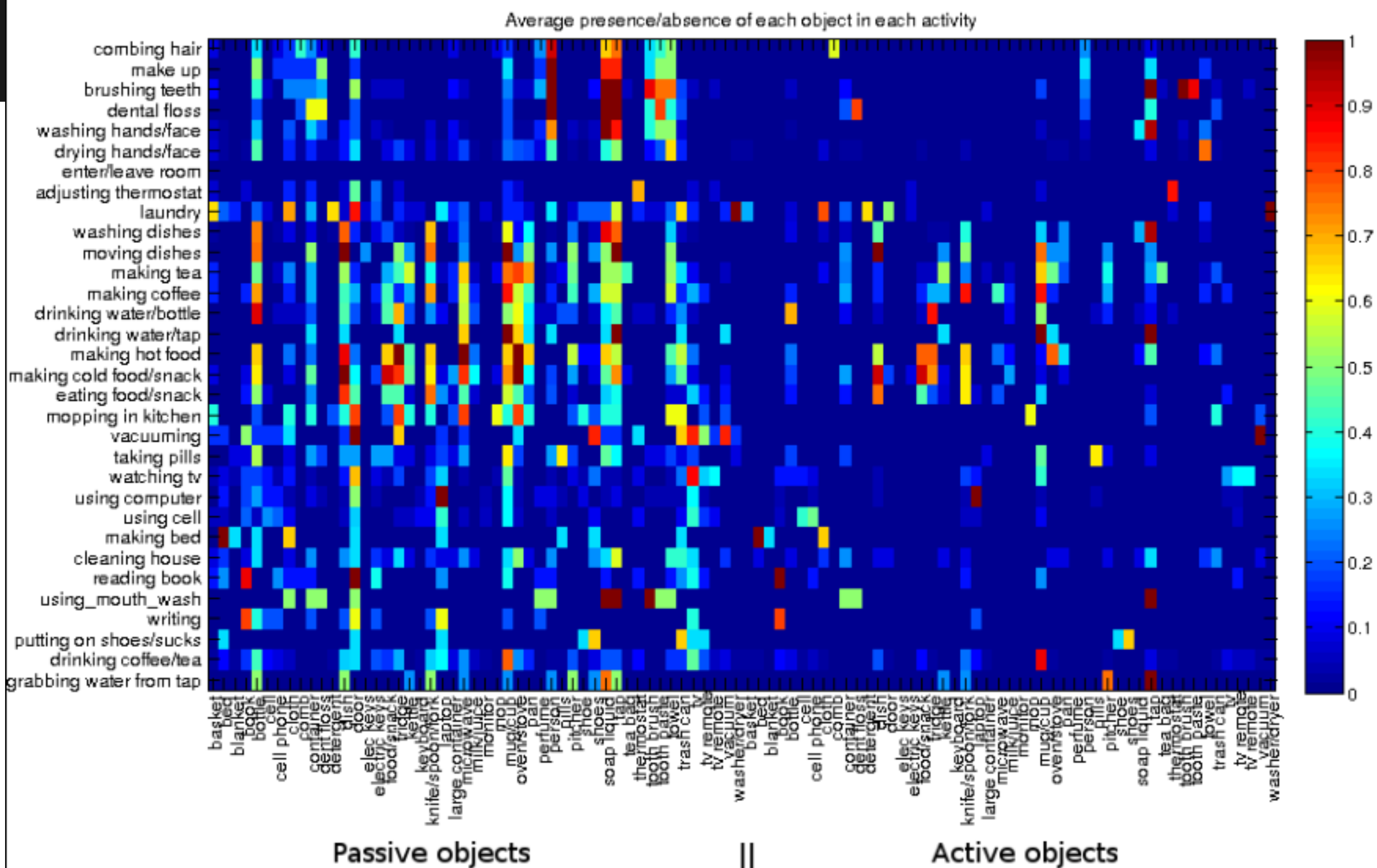
Bag-of-objects

Captures some notion of the scene.

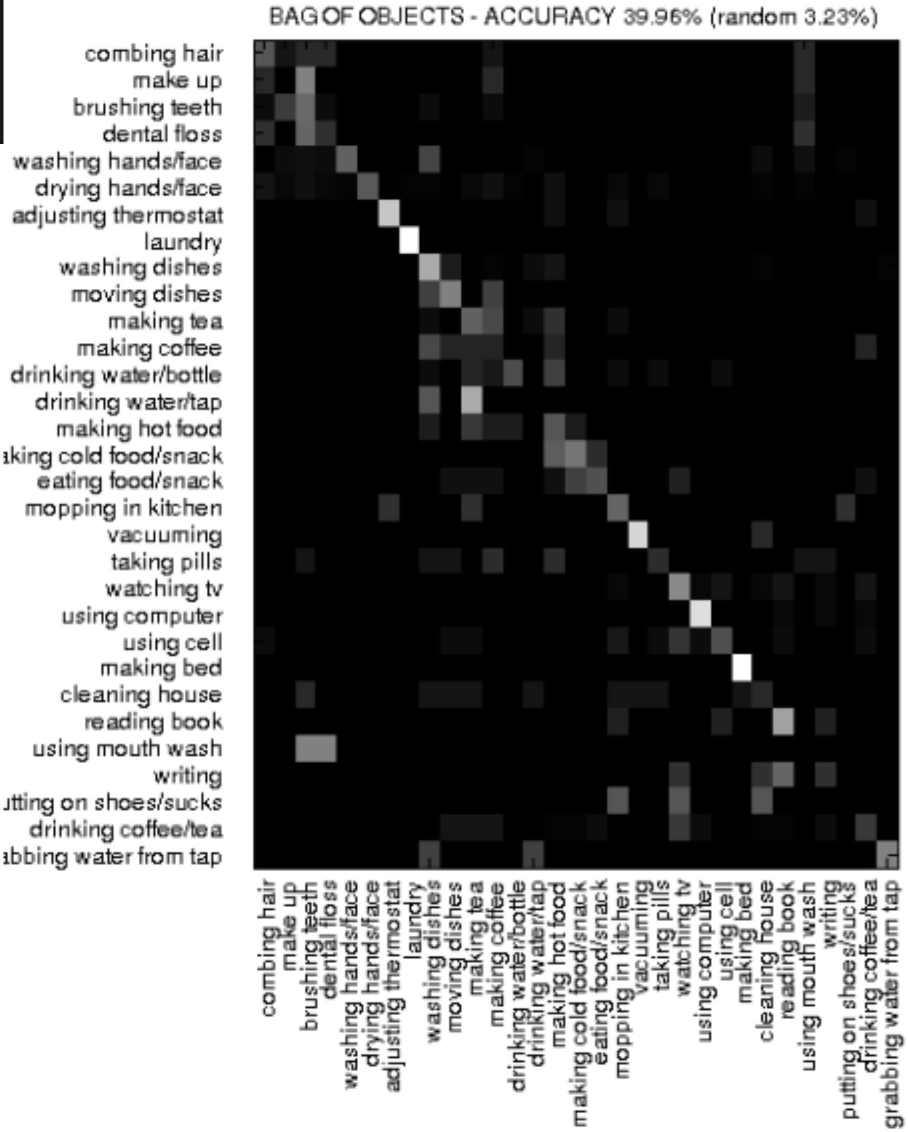
Action classes that are typically performed in similar settings tend to get confused.

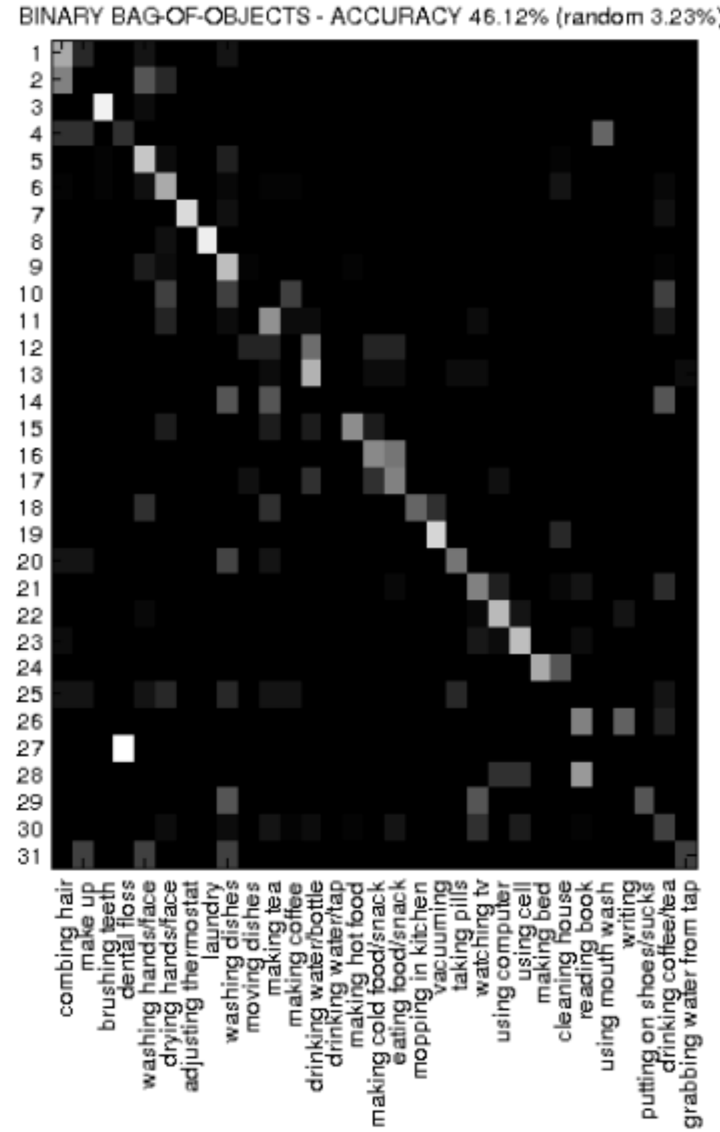
Can action recognition really just be reduced to object detection?

Active and passive objects



Active and passive objects





Results

Method	Accuracy
DPM | act.+pas.| 2 temp levels	19.98%
Ideal | no activity info | no ord.	29.61%
Ideal | act. + pas. | no ord	**46.12%**

Data ambiguity

Again, a large quantity of the data actually collected is not used in the paper, or in the implementation.

Only 21 of 49 passive objects and 5 of 49 active objects are used in the implementation.

This might be a constraint forced by object detection performance.

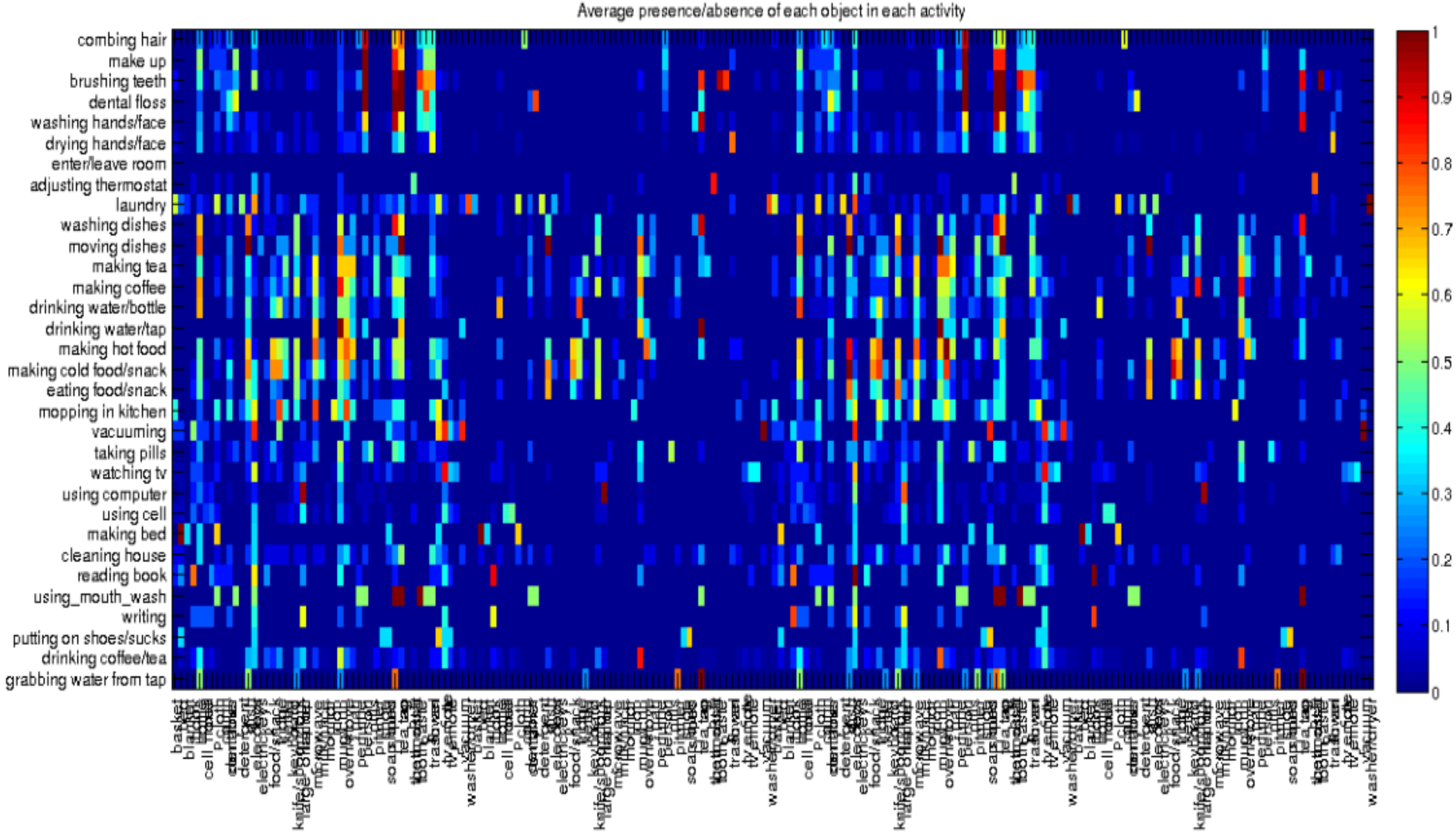
Active and passive objects

Information about which objects are being *used* - crucial cue for *action* recognition.

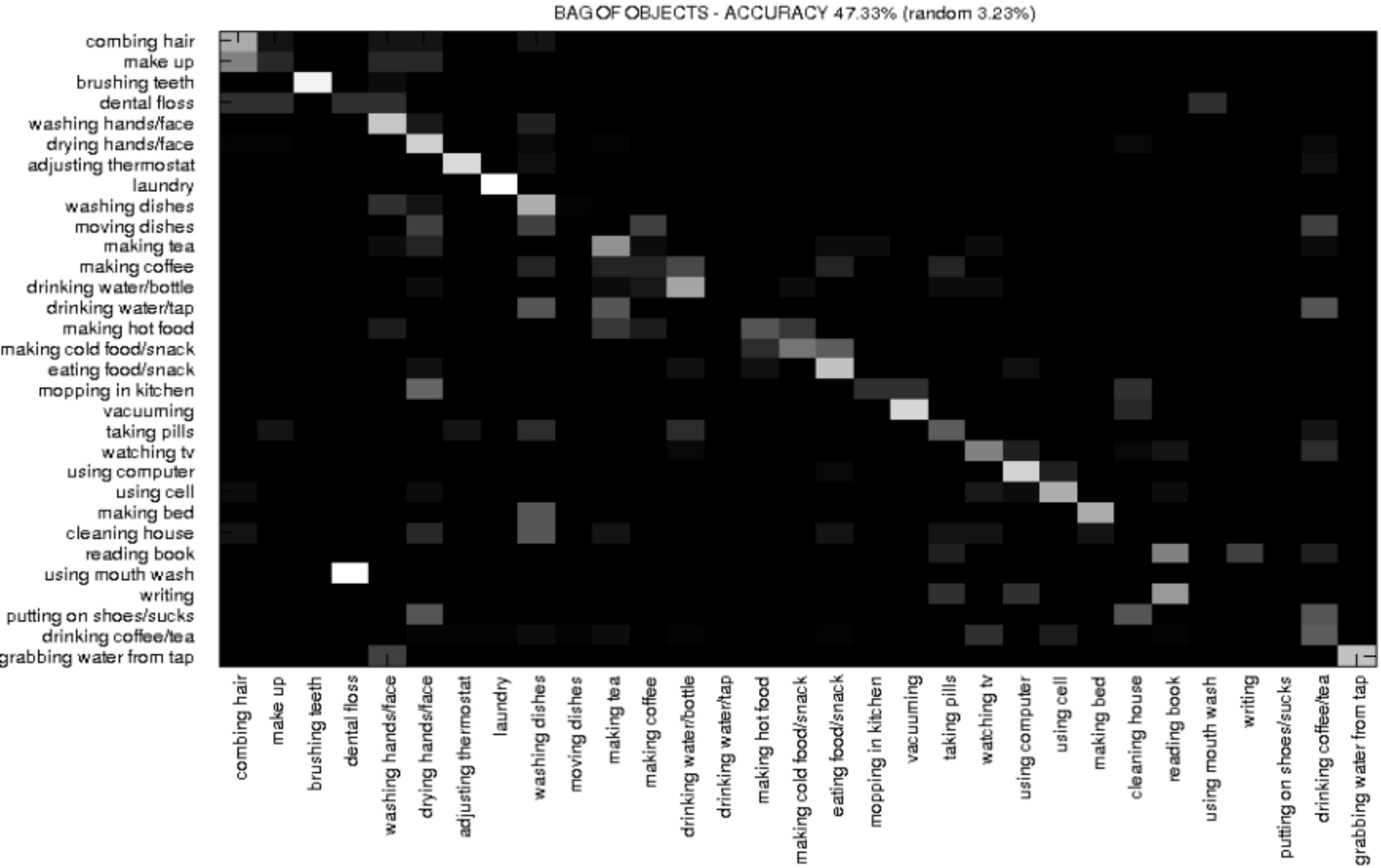
Captures important information about person's interaction with objects, rather than just looking at objects.

Helps disambiguate previously confused action classes performed in similar settings.Large performance boost (from 33.5% to 40% and 29.5% to 46% respectively)

Temporal ordering



Temporal ordering



Results

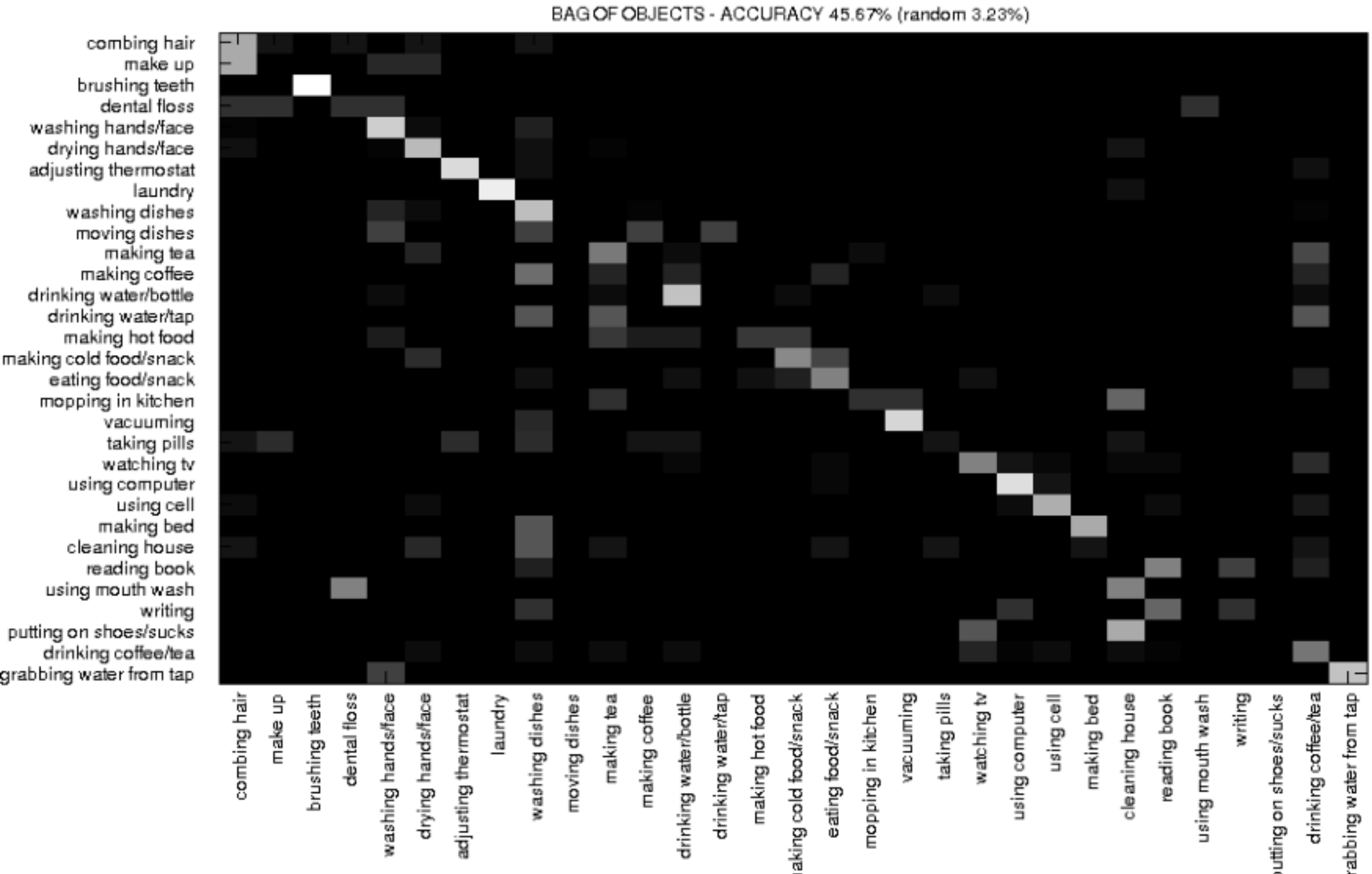
Method	Accuracy
DPM | act.+pas.| 2 temp levels	19.98%
Ideal | no activity info | no ord.	29.61%
Ideal | act. + pas. | no ord	46.12%
Ideal | act. + pas. | 2 temp levels	**47.33%**

Temporal ordering

Marginal improvement in performance

Does more temporal ordering improve performance?

Three temporal levels



Temporal ordering

Contributes little to classification when ground truth annotations for active and passive objects are known for this dataset

Without active/passive objects, temporal ordering (2 levels) boosts performance from 29.6 to 36.2%

	segment class. accuracy	
	pyramid	bag
STIP	22.8	16.5
O	32.7	24.7
AO	40.6	36.0
IO	55.8	49.3
IA+IO	77.0	76.8

Results

Method	Accuracy
DPM | act.+pas.| 2 temp levels	19.98%
Ideal | no activity info | no ord.	29.61%
Ideal | no activity inf| 2 temp lev	**36.20%**
Ideal | act. + pas. | no ord	46.12%
Ideal | act. + pas. | 2 temp levels	47.33%
Ideal | act. + pas. | 3 temp levels	45.67%

Temporal ordering

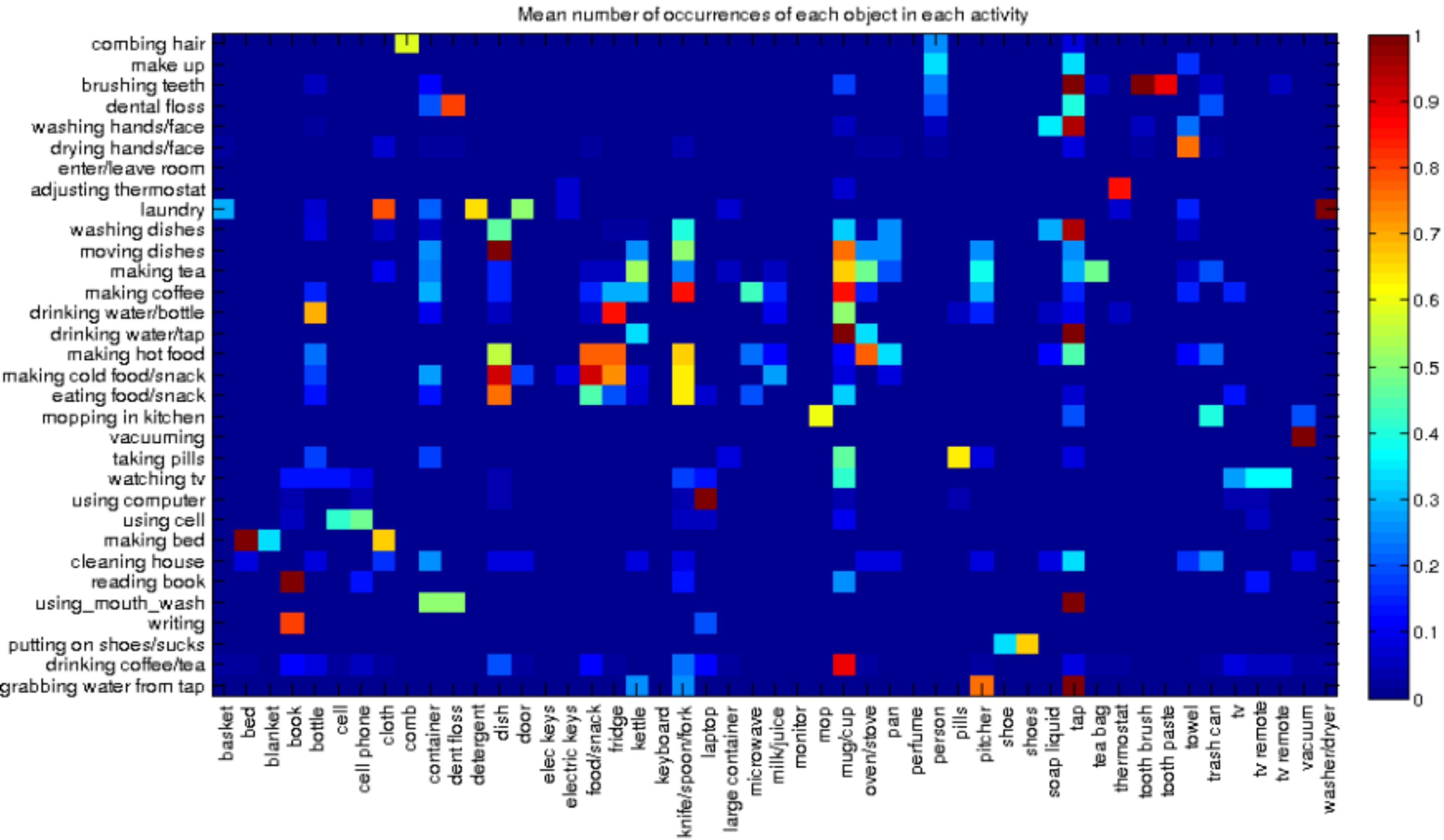
Why is temporal ordering more important when not using less data or "non-ideal detectors"?

Can we do better?

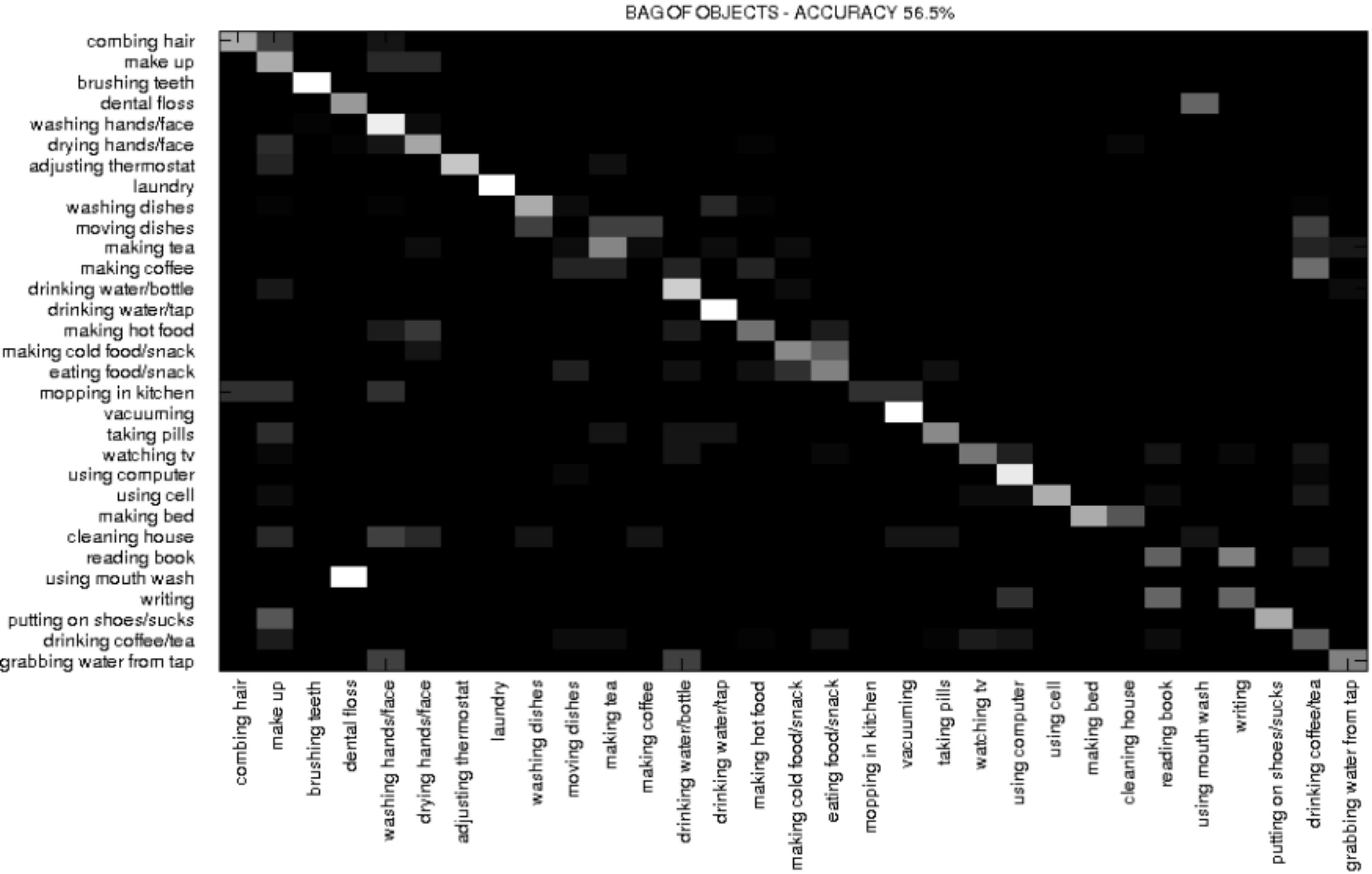
What we have learnt:

- Activity information contributes most
- Temporal ordering makes insignificant difference when activity information is available
- Training data is limited => smaller feature space is preferable

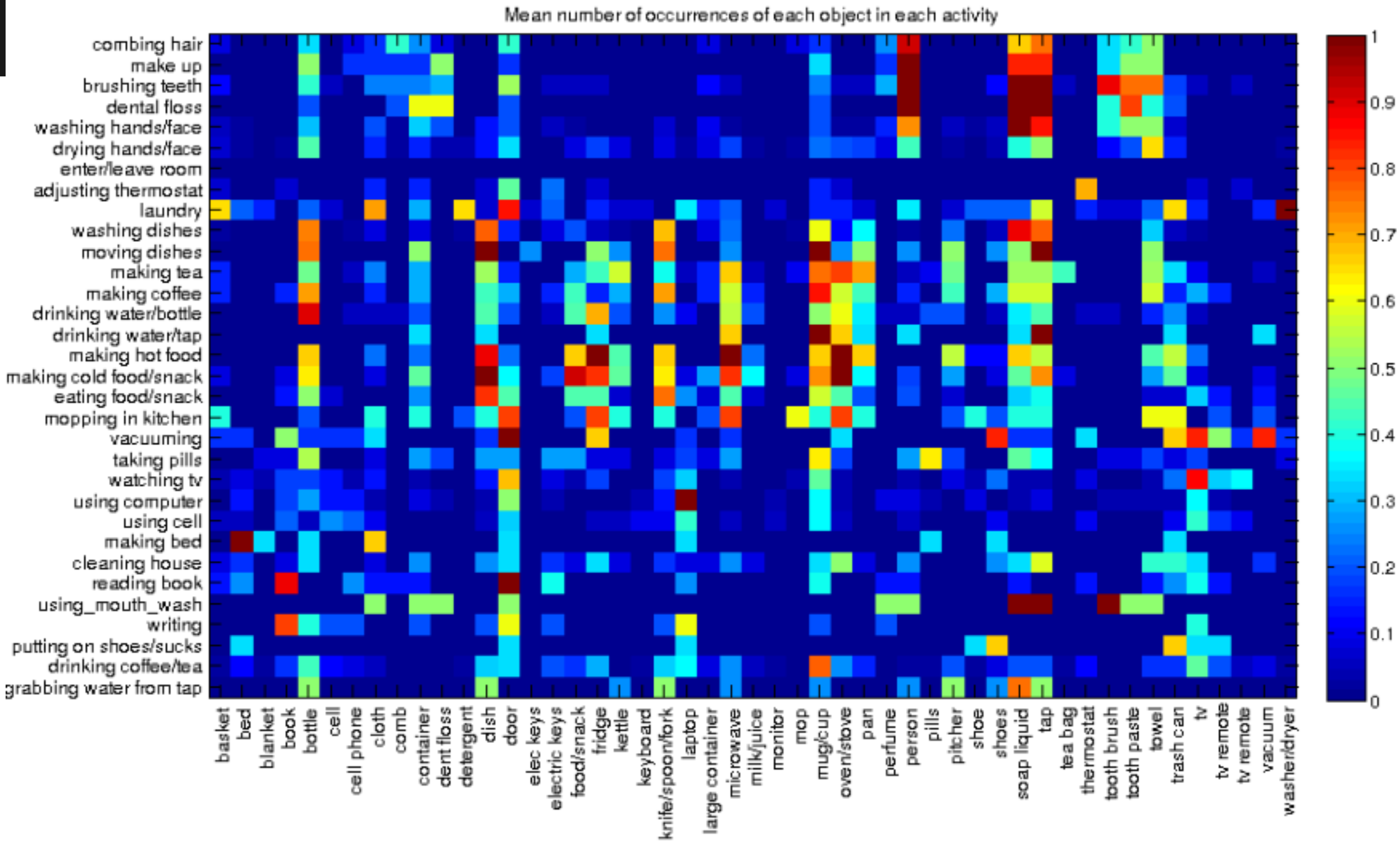
ONLY active objects



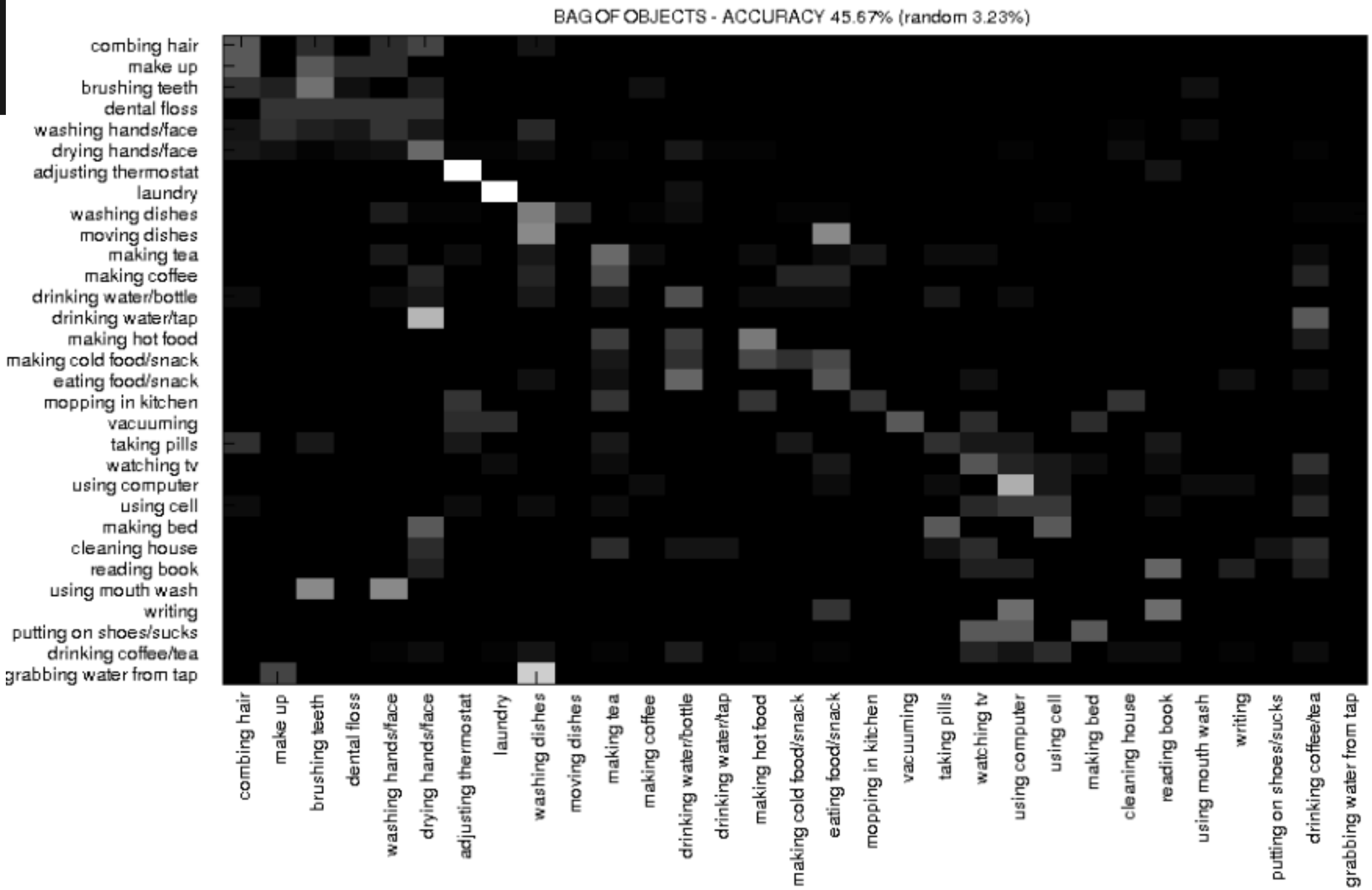
ONLY active objects



ONLY Passive objects



ONLY passive objects



Active objects

- Deteriorates to 51.63% with two temporal levels - insufficient training data
- We have side-stepped object detection by using ground truth annotations
- Near-ideal active object detection performance may be very hard to achieve - occlusions etc., so other cues are important for robust performance.

Results

Method	Accuracy
DPM | act.+pas.| 2 temp levels	19.98%
Ideal | no activity info | no ord.	29.61%
Ideal | no activity inf | 2 temp lev	36.20%
Ideal | pas. | 2 temp levels	**25.04%**
Ideal | act. | no ord	**56.50%**
Ideal | act. | 2 temp levels	**51.63%**
Ideal | act. + pas. | no ord	46.12%
Ideal | act. + pas. | 2 temp levels	47.33%
Ideal | act. + pas. | 3 temp levels	45.67%

- Hamed Pirsiavash and Deva Ramanan, "*Detecting activities of daily living in first-person camera views*", CVPR 2012

- Examples, dataset and code at http://deepthought.ics.uci.edu/ADLdataset/adl.html