What to Keep?:
Summarizing Long Egocentric Videos

Kristen Grauman
Department of Computer Science
University of Texas at Austin

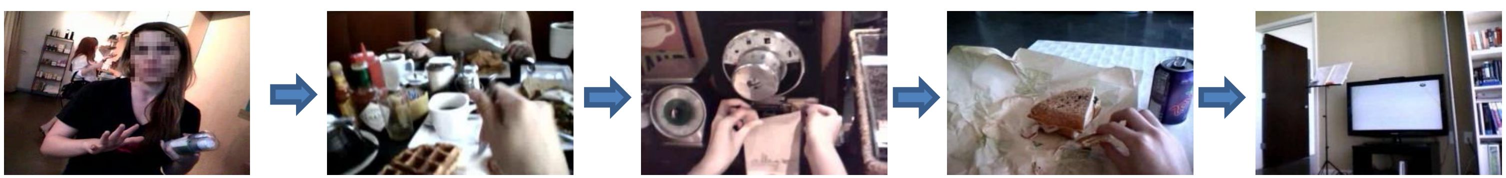
With Yong Jae Lee, Bo Xiong, Lu Zheng

THE UNIVERSITY OF
TEXAS
AT AUSTIN

Goal: Summarize egocentric video

Wearable camera

Input: Egocentric video of the camera wearer's day



9:00 am 10:00 am 11:00 am 12:00 pm 1:00 pm 2:00 pm

Output: Storyboard (or video skim) summary

Potential applications of egocentric video summarization

Memory aid

Law enforcement

Mobile robot discovery

RHex Hexapedal Robot, Penn's GRASP Laboratory

What makes egocentric data hard to summarize?

- Subtle event boundaries
- Subtle figure/ground
- Long streams of data

Prior work

- **Egocentric recognition**

  [Starner et al. 1998, Doherty et al. 2008, Spriggs et al. 2009, Jojic et al. 2010, Ren & Gu 2010, Fathi et al. 2011, Aghazadeh et al. 2011, Kitani et al. 2011, Pirsiavash & Ramanan 2012, Fathi et al. 2012,…]

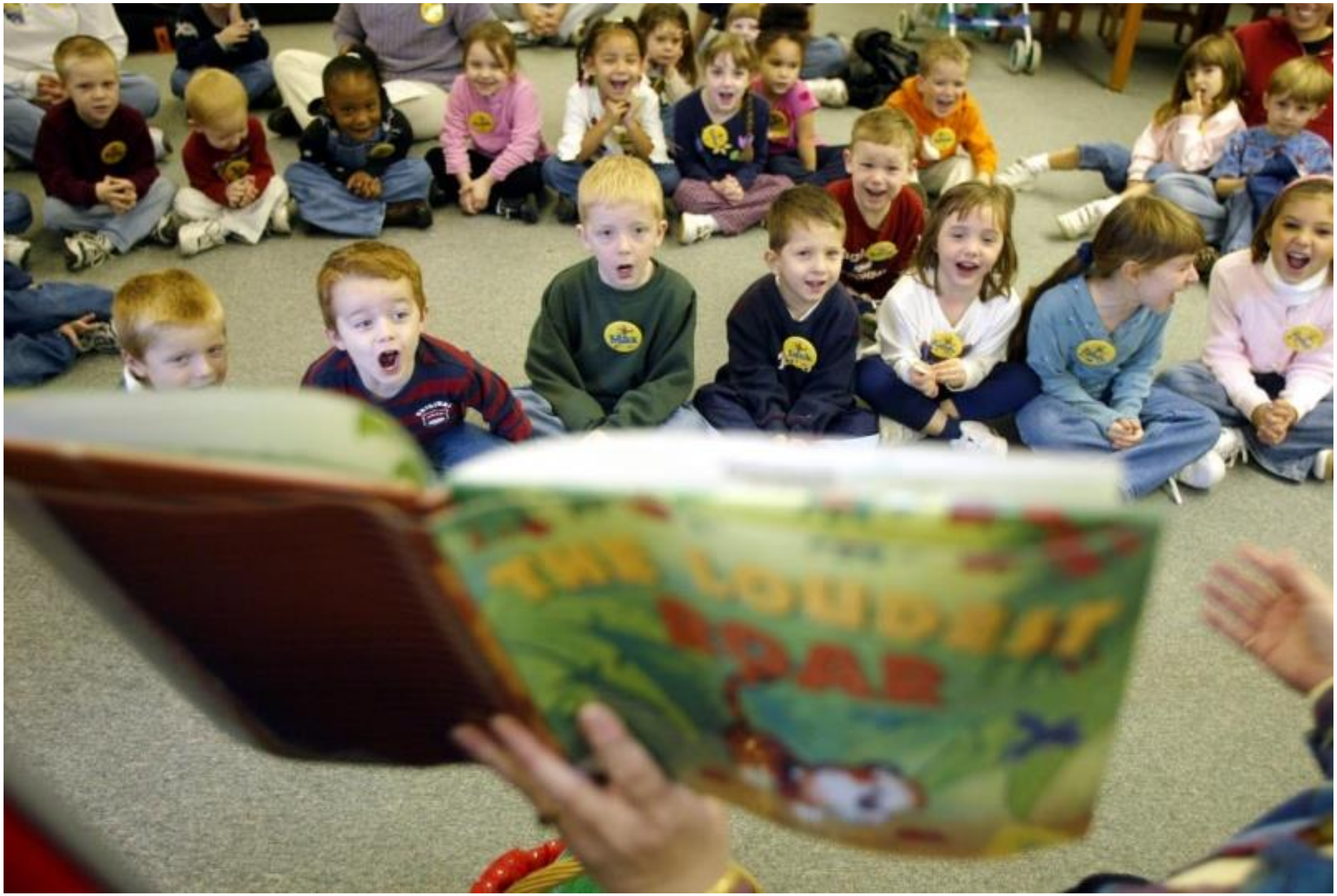
- **Video summarization**

  [Wolf 1996, Zhang et al. 1997, Ngo et al. 2003, Goldman et al. 2006, Caspi et al. 2006, Pritch et al. 2007, Laganiere et al. 2008, Liu et al. 2010, Nam & Tewfik 2002, Ellouze et al. 2010,…]

→ Low-level cues, stationary cameras

→ Consider summarization as a *sampling* problem

Our idea:
Story-driven summarization



[Lu & Grauman, CVPR 2013]

Our idea:
Story-driven summarization

Good summary captures the progress of the story

1. Segment video temporally into subshots

2. Select chain of *k* subshots that maximize both weakest link's influence and object importance

[Lu & Grauman, CVPR 2013]

Egocentric subshot detection

Define 3 generic ego-activities:

~Static

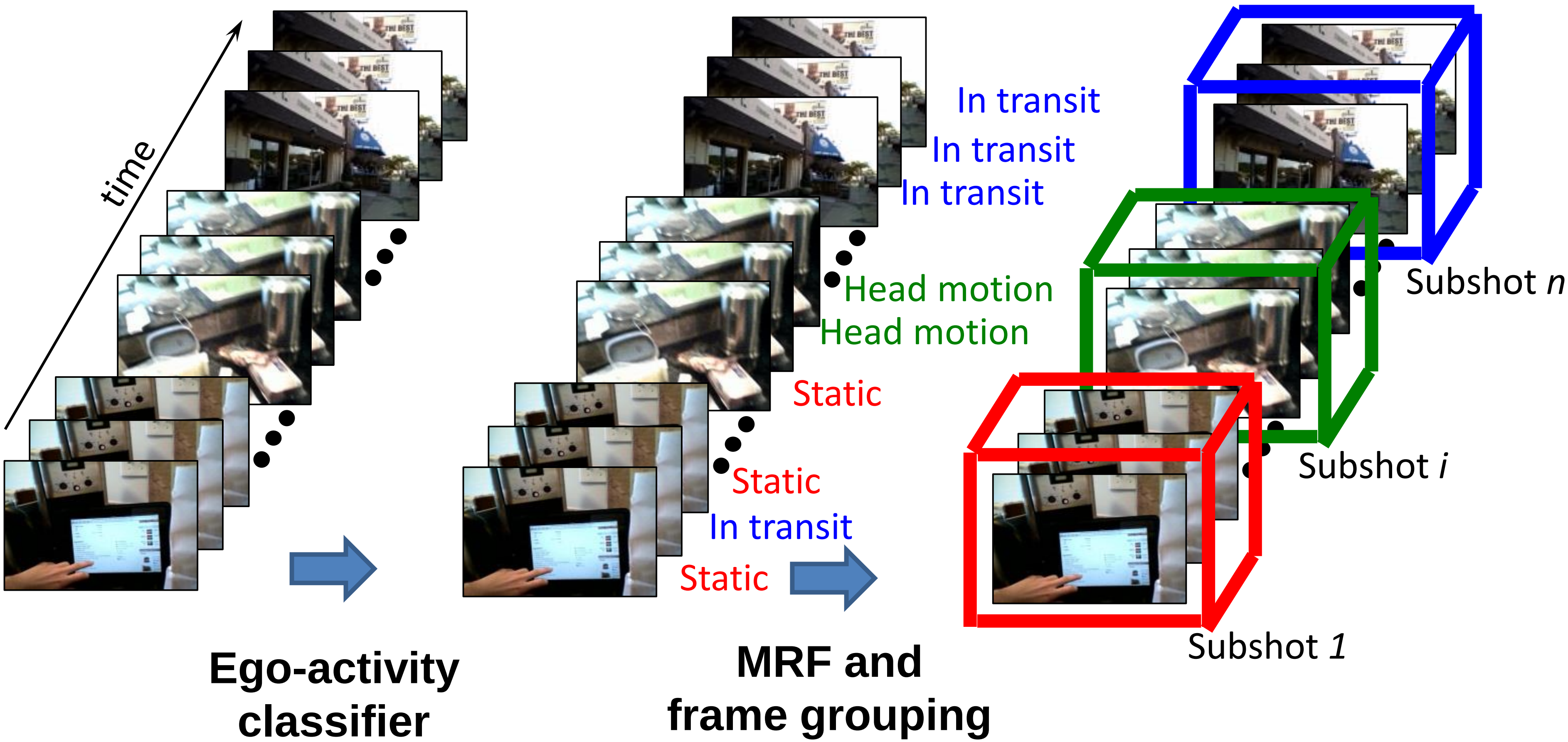
In transit

Head moving

- Train classifiers to predict these activity types
- Features based on flow and motion blur

[Lu & Grauman, CVPR 2013]

Egocentric subshot detection



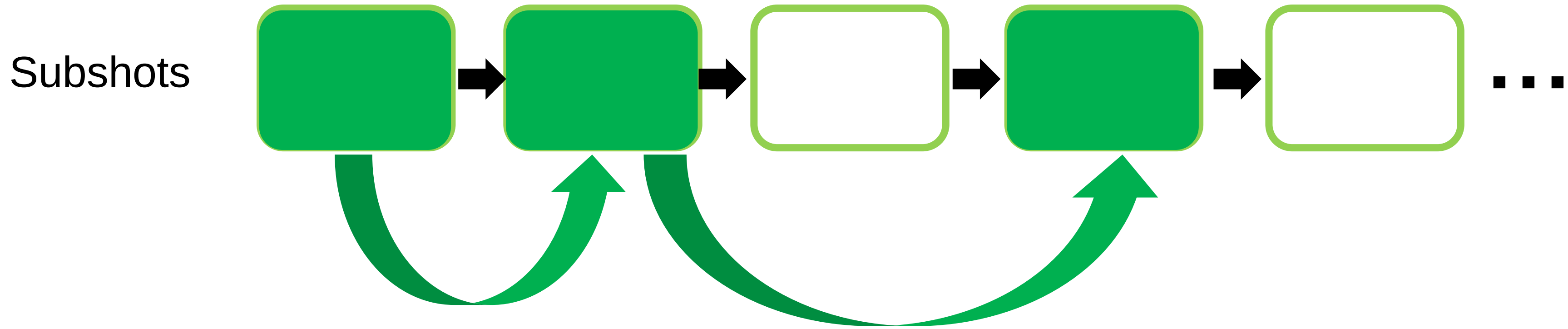
[Lu & Grauman, CVPR 2013]

Subshot selection objective

Good summary = chain of *k* selected subshots in which each influences the next via some subset of key objects

$$S^* = \arg\max_{S \subset \mathcal{V}} \; \lambda_s \, \mathcal{S}(S) + \lambda_i \, \mathcal{I}(S) + \lambda_d \, \mathcal{D}(S)$$

influence **importance** **diversity**



[Lu & Grauman, CVPR 2013]

Learning region importance

Man wearing a blue shirt and watch in coffee shop

Yellow notepad on table

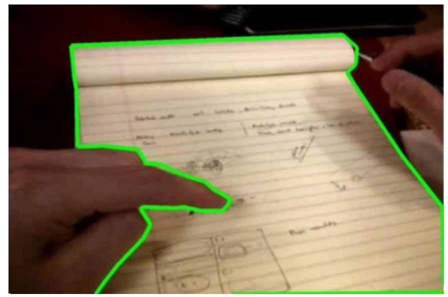
Coffee mug that cameraman drinks

- First task: watch a short clip, and *describe in text* the essential people or objects necessary to create a summary

[Lee et al. CVPR 2012]

Learning region importance

Man wearing a blue shirt and watch in coffee shop



Yellow notepad on table

Coffee mug that cameraman drinks

Iphone that the camera wearer holds

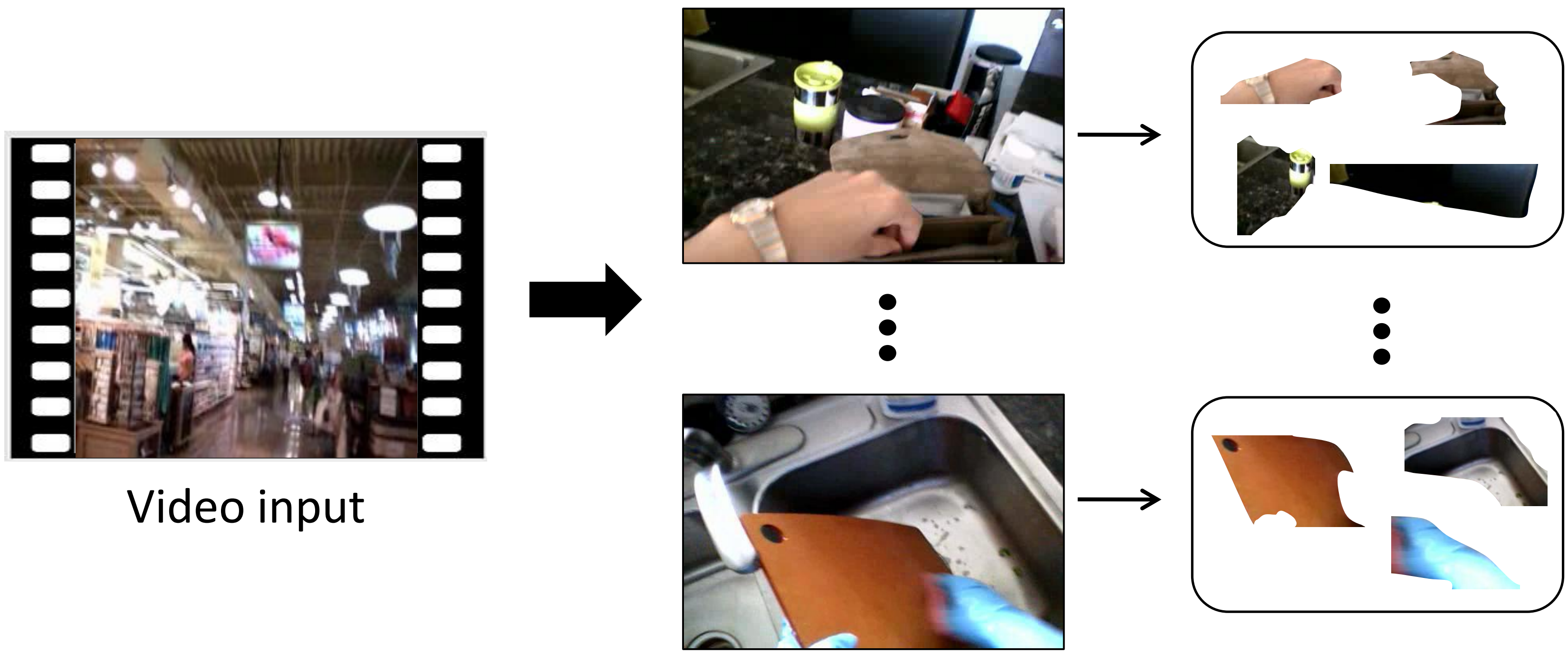
Camera wearer cleaning the plates

Soup bowl

- Second task: draw polygons around any described person or object *obtained from the first task* in sampled frames

[Lee et al. CVPR 2012]

Learning region importance



Video input

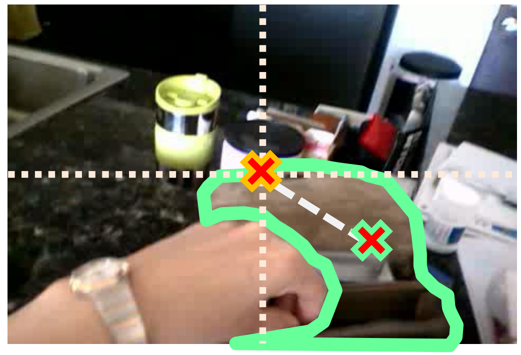
Generate candidate object regions
for uniformly sampled frames

[Lee et al. CVPR 2012]

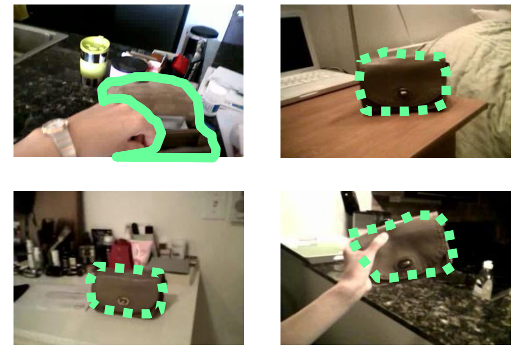
Learning region importance

Egocentric features:

distance to hand



distance to frame center



frequency

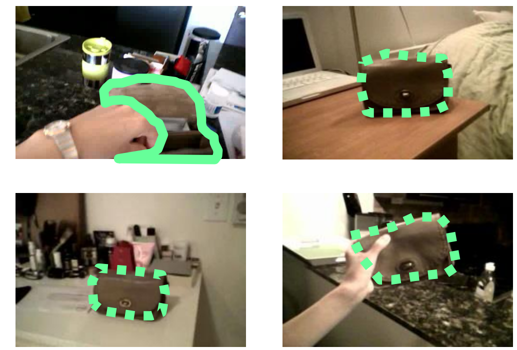
[Lee et al. CVPR 2012]

Learning region importance

Egocentric features:

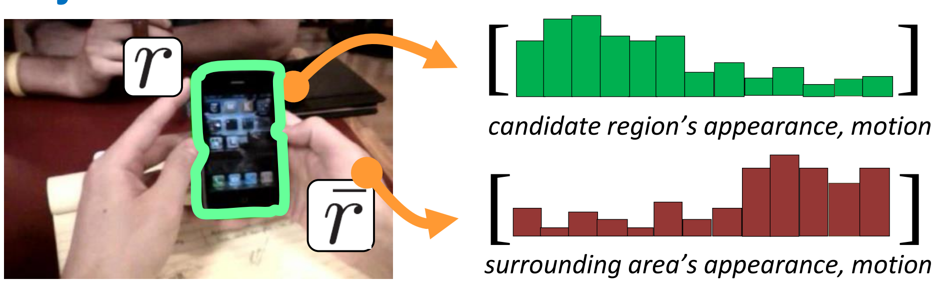
distance to hand

distance to frame center



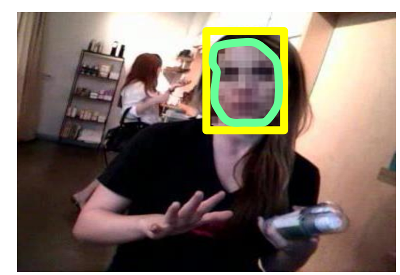
frequency

Object features:



"Object-like" appearance, motion

[Endres et al. ECCV 2010, Lee et al. ICCV 2011]



overlap w/ face detection

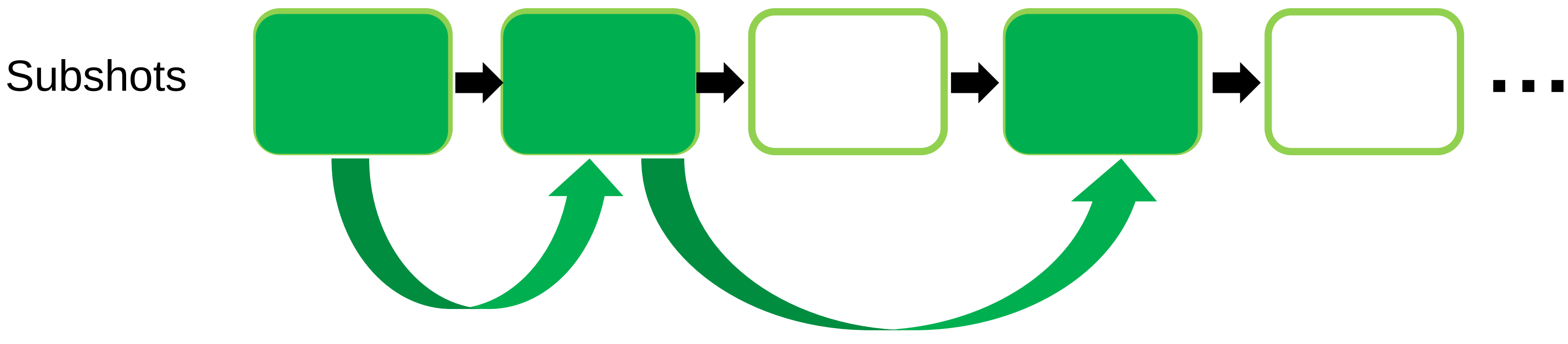
Region features: *size, width, height, centroid*

[Lee et al. CVPR 2012]

Influence criterion

- Want the *k* subshots that maximize the weakest link's influence, subject to coherency constraints

$$\mathcal{S}(S) = \max_{a} \min_{j=1,\ldots,K-1} \sum_{o_i \in O} a_{i,j} \textsc{Influence}(s_j, s_{j+1} | o_i)$$

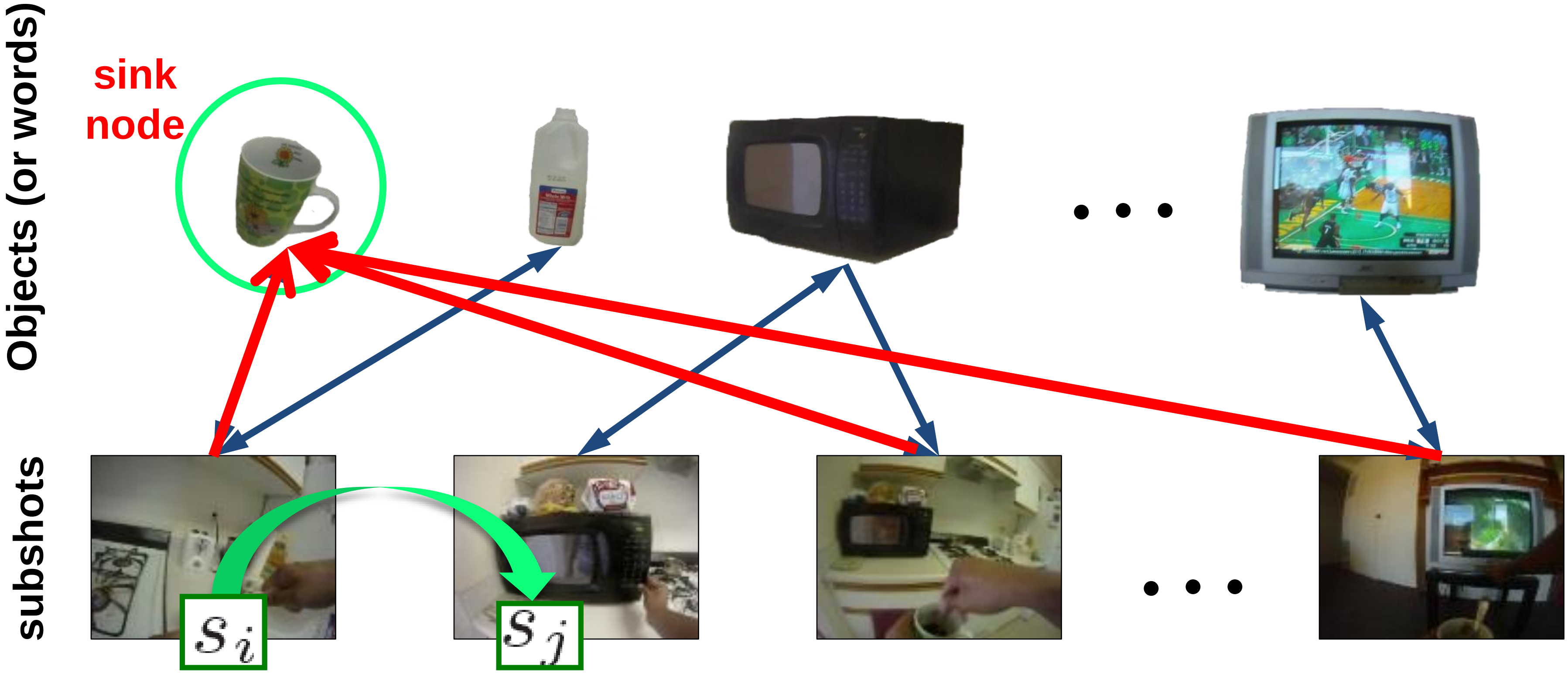


[Lu & Grauman, CVPR 2013]

Document-document influence
[Shahaf & Guestrin, KDD 2010]

Connecting the dots between news articles. D. Shahaf and C. Guestrin. In KDD, 2010.

Estimating visual influence



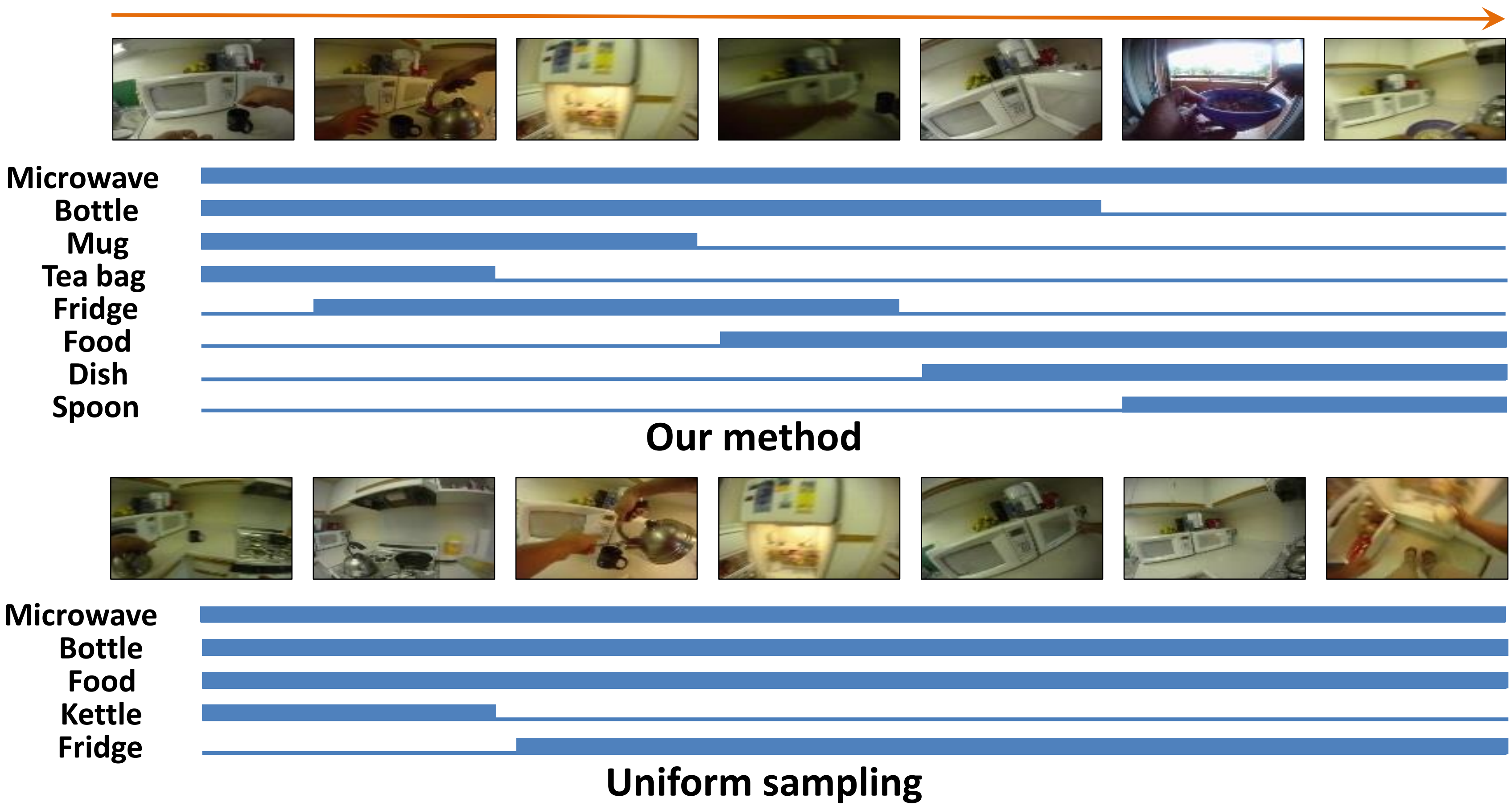
$$\textsc{Influence}(s_i, s_j | o) = \prod_i (s_j) - \prod_i^o (s_j)$$

Captures how reachable subshot *j* is from subshot *i*, via any object *o*

[Lu & Grauman, CVPR 2013]

Estimating visual influence

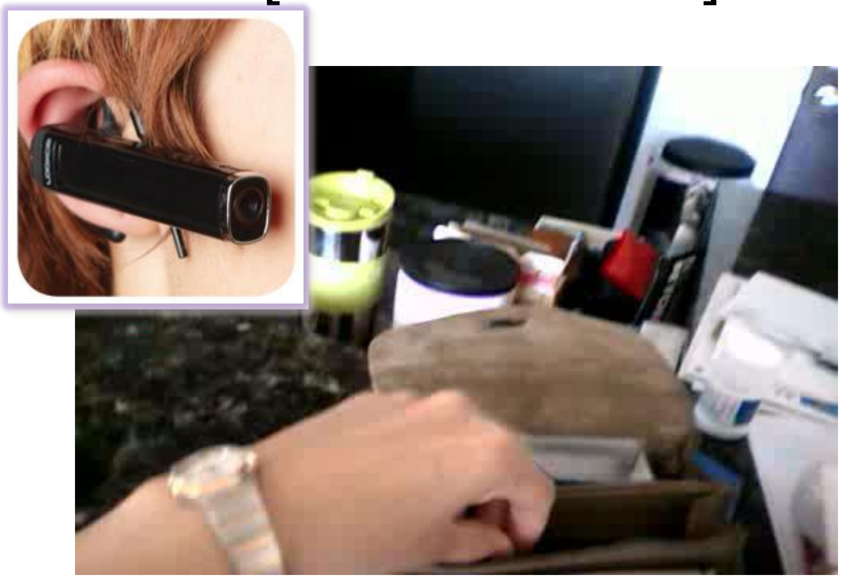
- Prefer small number of objects at once, and coherent (smooth) entrance/exit patterns



[Lu & Grauman, CVPR 2013]

Datasets

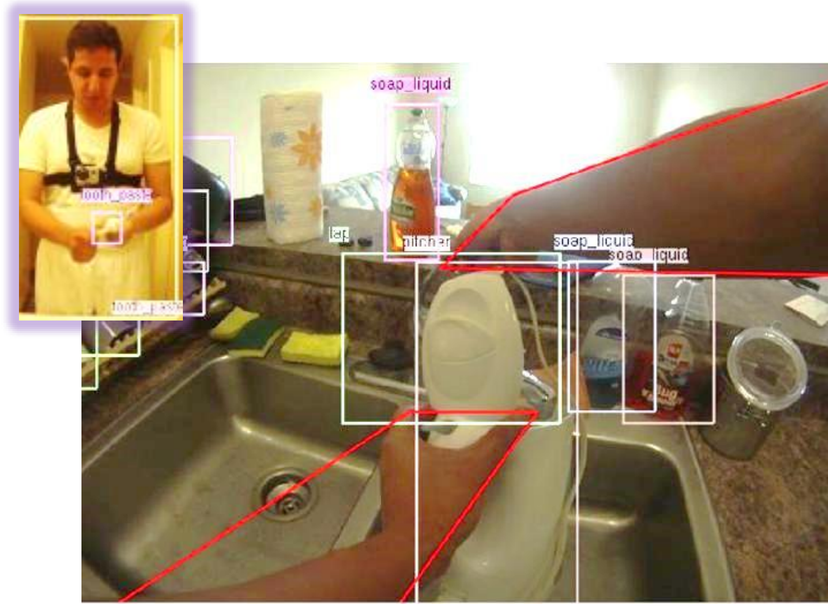
UT Egocentric (UT Ego)
[Lee et al. 2012]



4 videos, each 3-5 hours long, uncontrolled setting.

We use visual words and subshots.

Activities of Daily Living (ADL)
[Pirsiavash & Ramanan 2009]

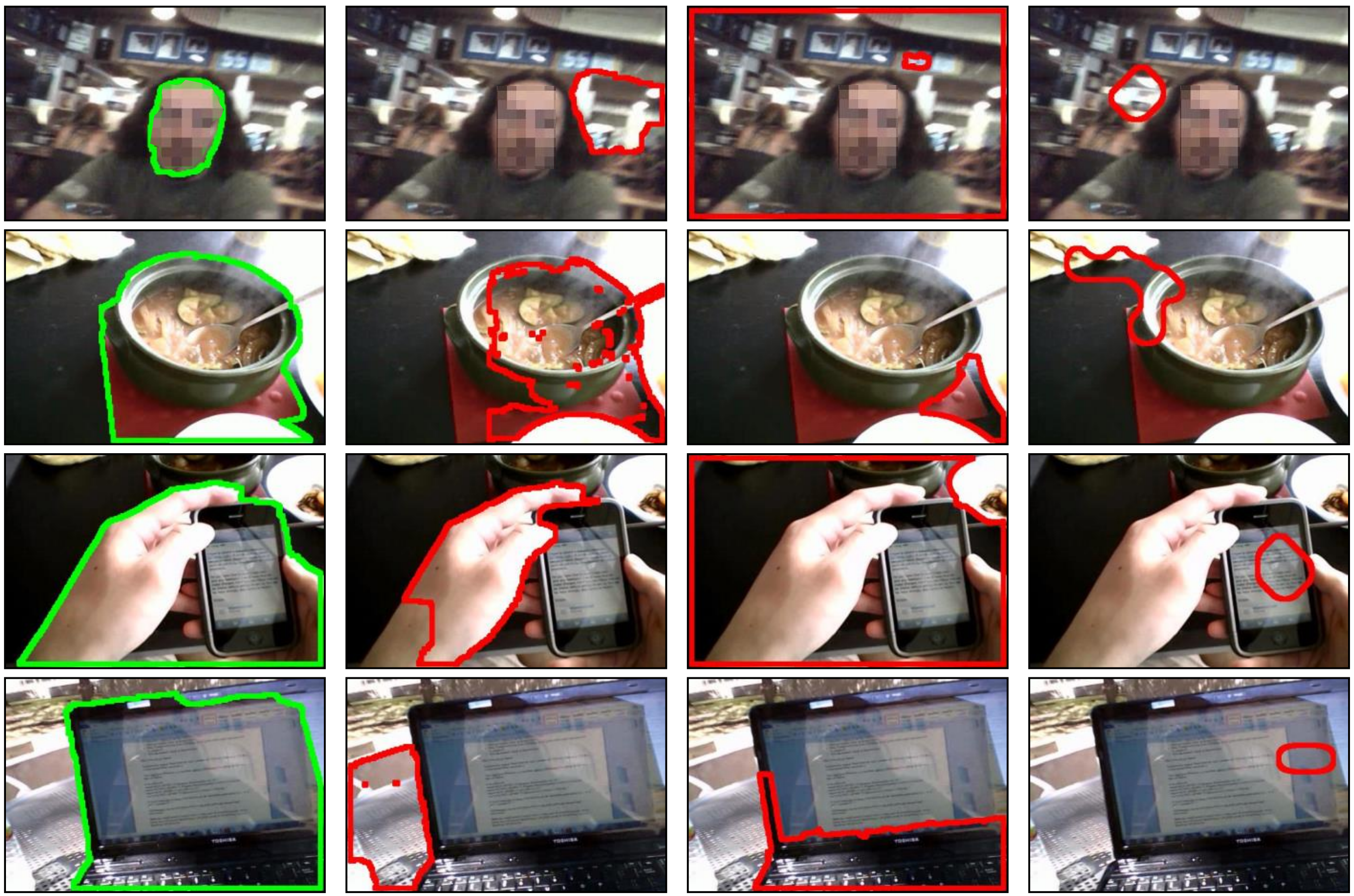


20 videos, each 20-60 minutes, daily activities in house.

We use object bounding boxes and keyframes.

Results: Important region prediction

Ours | **Object-like [Carreira, 2010]** | **Object-like [Endres, 2010]** | **Saliency [Walther, 2005]**

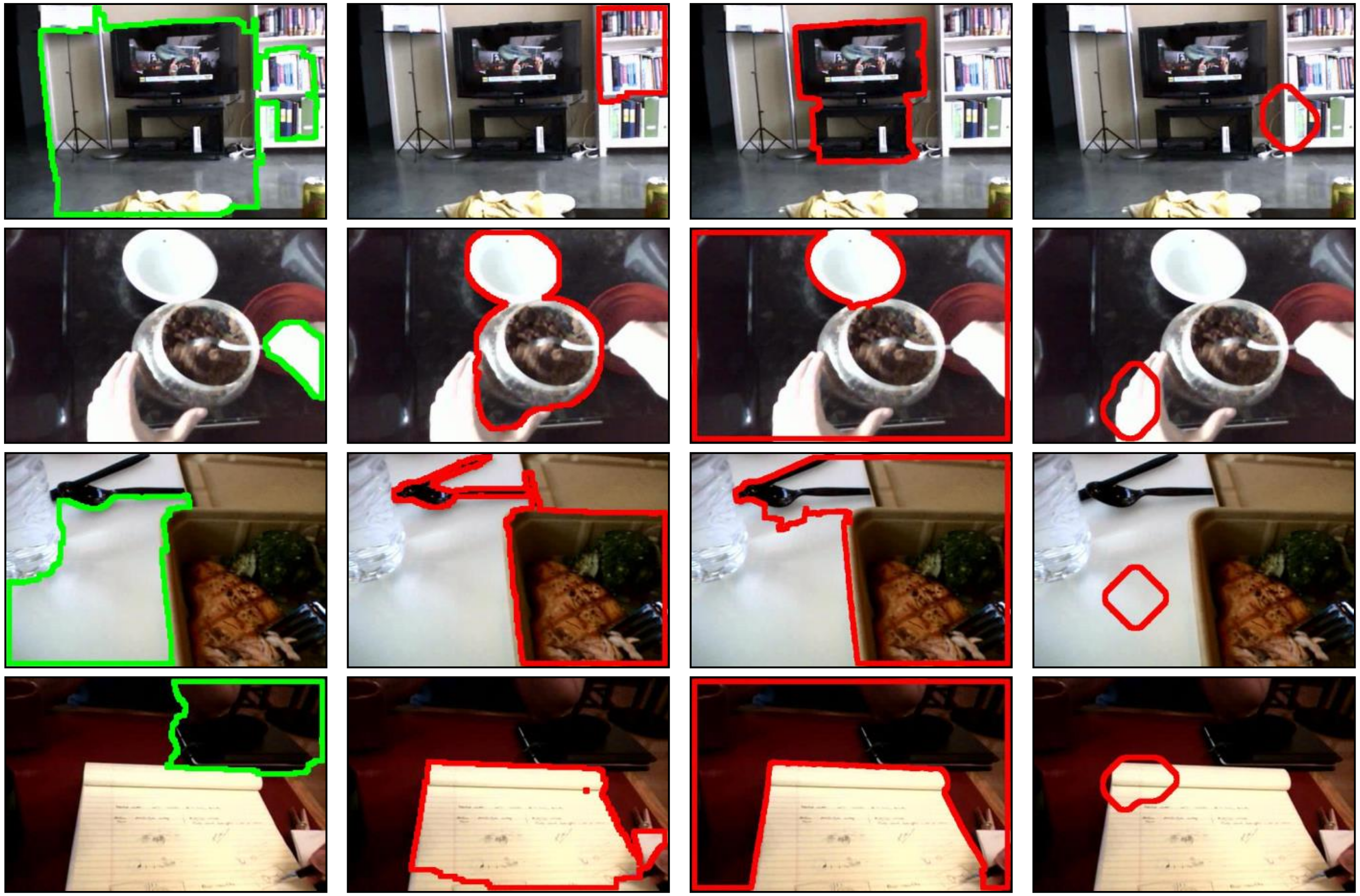


Good predictions

[Lee et al., CVPR 2012]

Results: Important region prediction

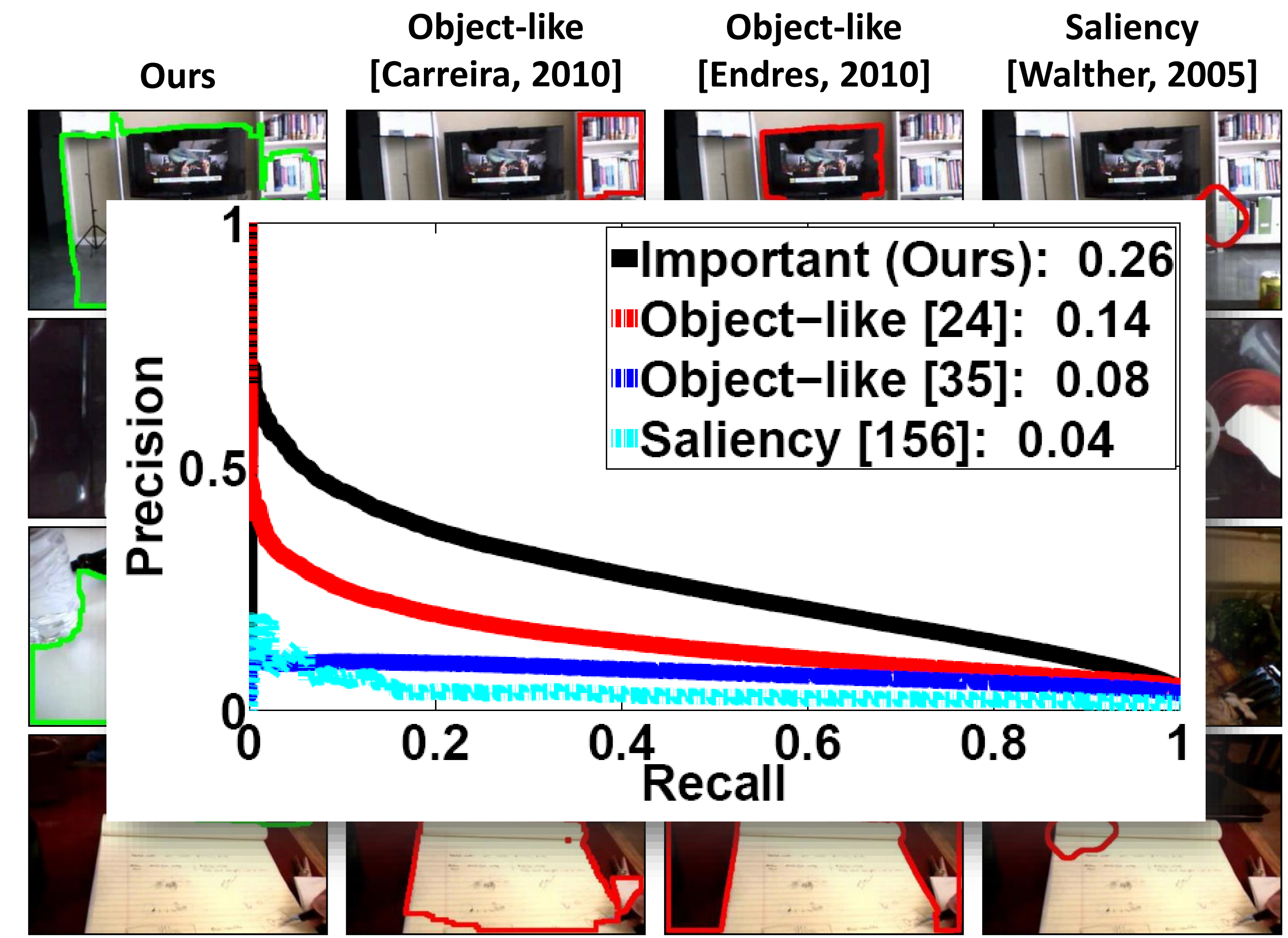
Ours	Object-like [Carreira, 2010]	Object-like [Endres, 2010]	Saliency [Walther, 2005]



Failure cases

[Lee et al., CVPR 2012]

Results: Important region prediction



Failure cases

[Lee et al., CVPR 2012]

Example keyframe summary – UT Ego data

Original video (3 hours)

Our summary (12 frames)

Example keyframe summary – UT Ego data

Alternative methods for comparison

Uniform keyframe sampling (12 frames)

[Liu & Kender, 2002] (12 frames)

Example summary – UT Ego data

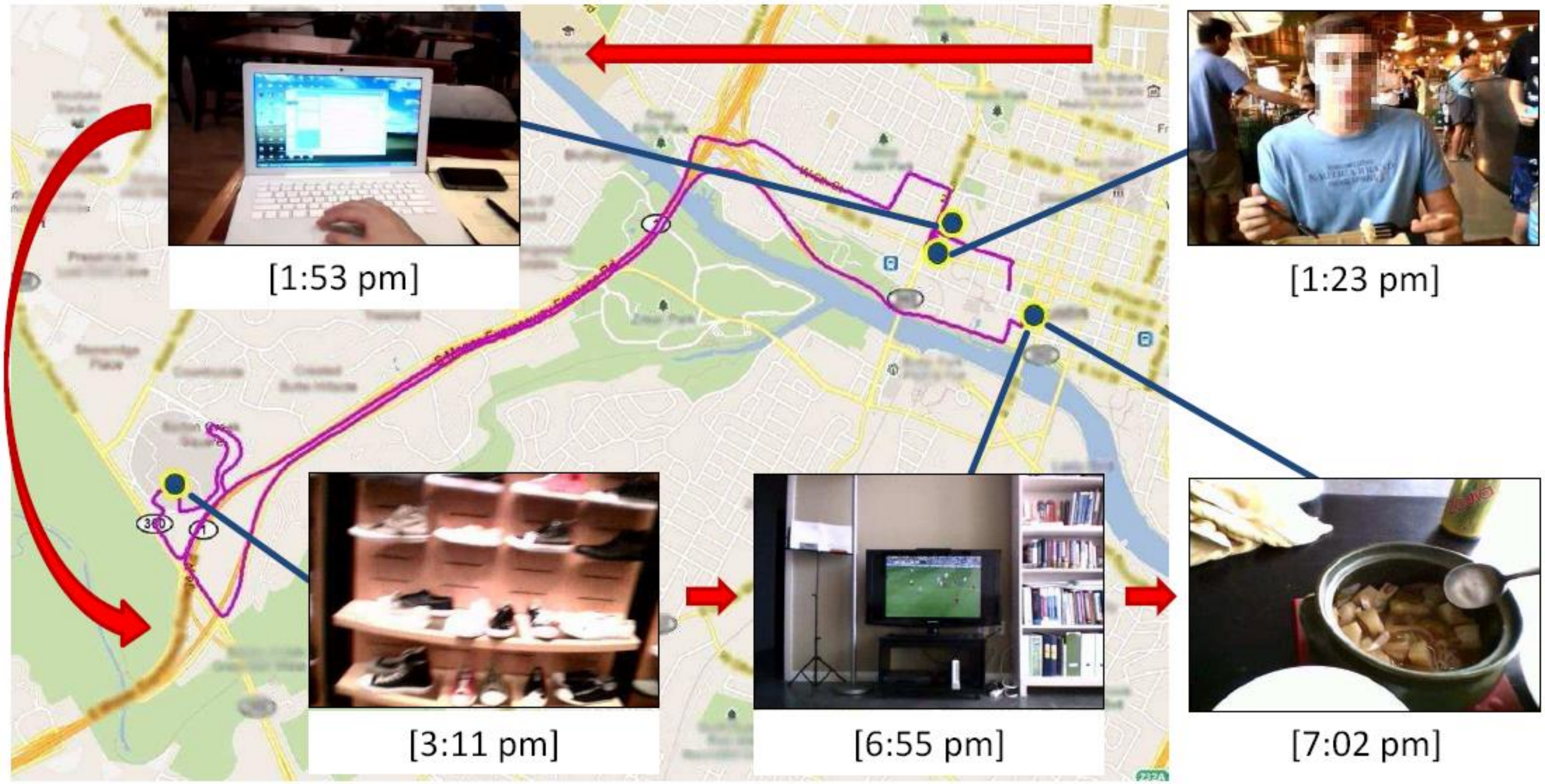
Ours

Baseline

Example summary – ADL data

Ours **Baseline 1**

Generating storyboard maps



Augment keyframe summary with geolocations

[Lee et al., CVPR 2012]

Human subject results: Blind taste test

How often do subjects prefer our summary?

Data	Uniform sampling	Shortest-path	Object-driven Lee et al. 2012
UTE	**90.0%**	**90.9%**	**81.8%**
ADL	**75.7%**	**94.6%**	**N/A**

34 human subjects, ages 18-60
12 hours of original video
Each comparison done by 5 subjects

Total 535 tasks, 45 hours of subject time

[Lu & Grauman, CVPR 2013]

Next steps

- Personalization
- Object-centric → activity-centric?
- Additional sensors
- Evaluation for search tasks
- Summaries while streaming

Which photos were purposely taken by a human?

Incidental wearable camera photos

Intentional human taken photos

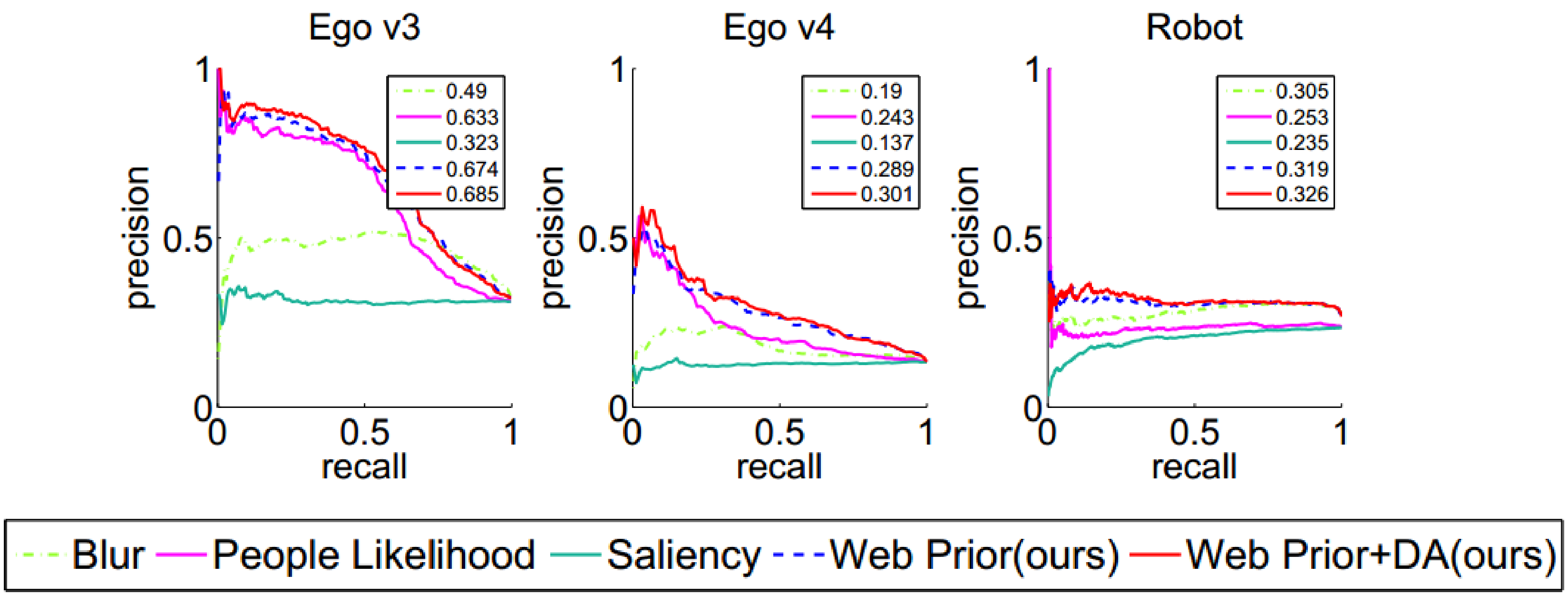
[Xiong & Grauman, ECCV 2014]

Idea: Detect "snap points"

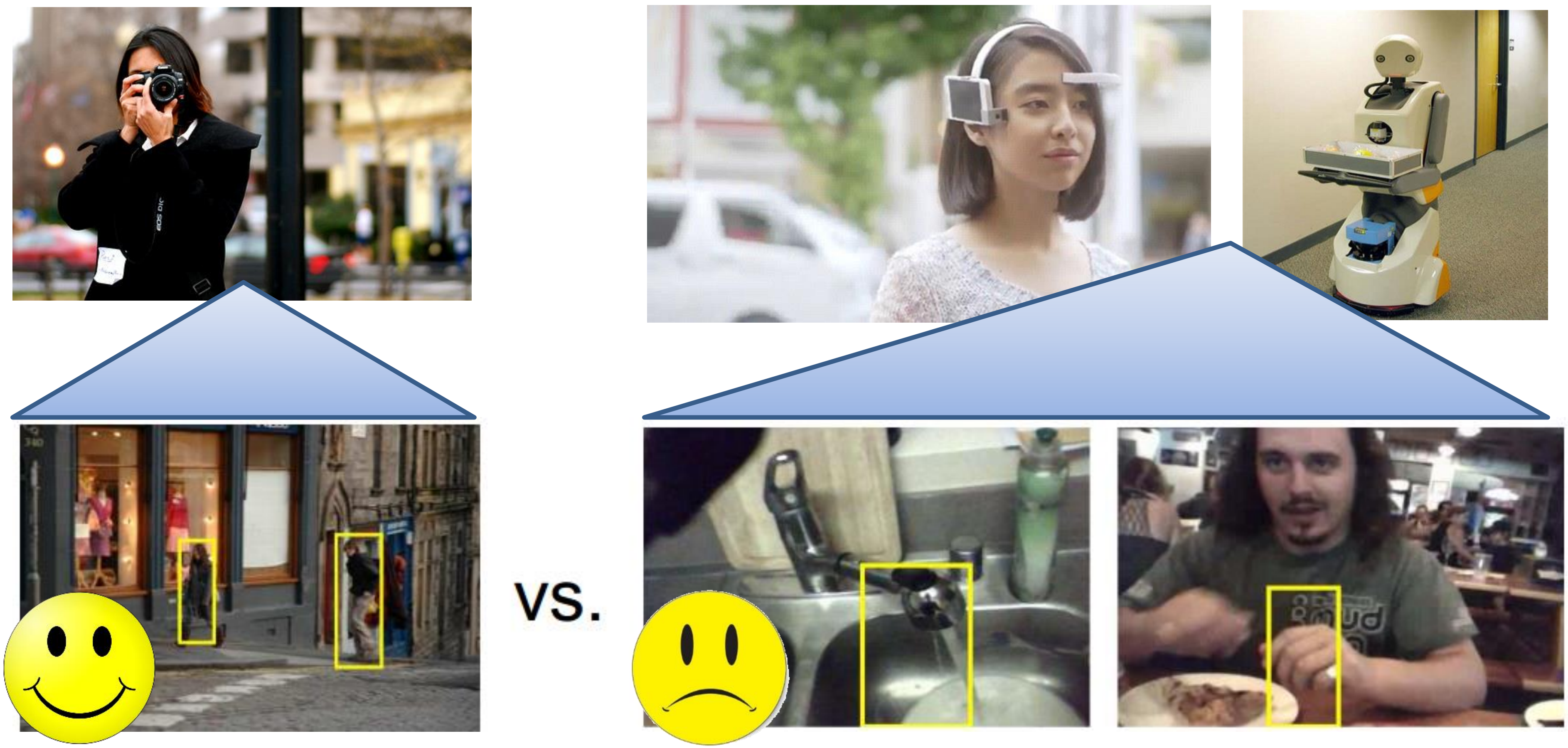
- Unsupervised data-driven approach to detect frames in first-person video that look intentional

[Xiong & Grauman, ECCV 2014]

Example snap point predictions



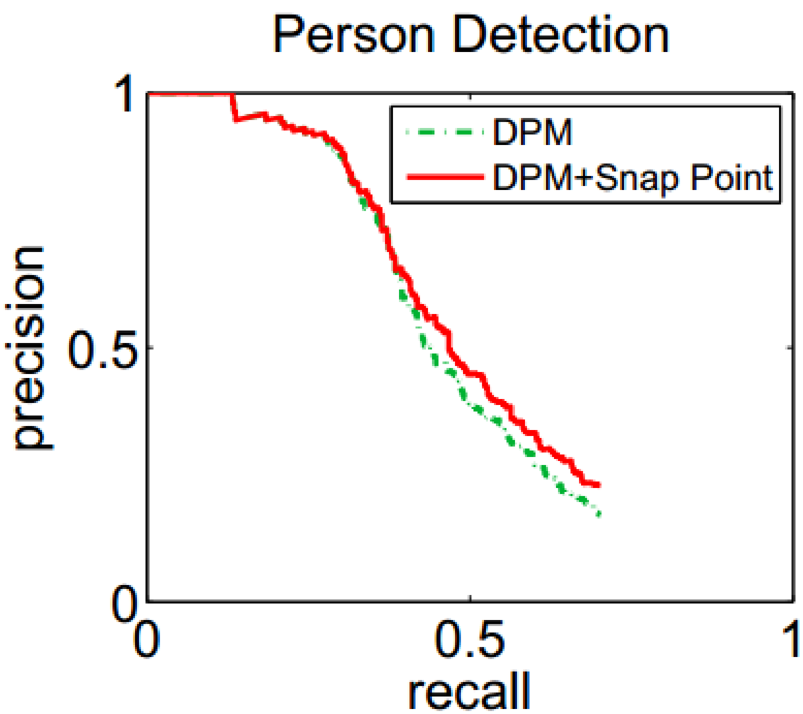
Snap points can boost precision for object detection

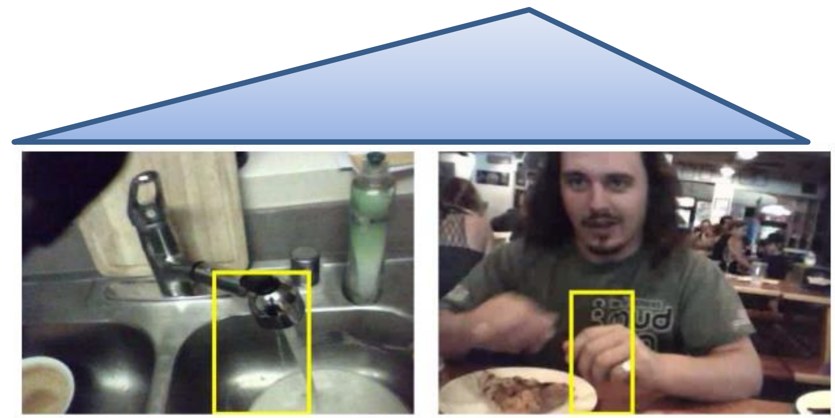


Person detection in intentional photos

Person detection in first-person frames

Snap points can boost precision for object detection





Person detection in first-person frames

Summary

- Deluge of first-person video imminent

  → Need **summaries** to access and browse

- First person video summarization
  - Estimate influence between events given their objects
  - Category-independent region importance prediction
  - Snap point detection with a Web prior