CVPR 2016 Workshop: Moving Cameras Meet Video Surveillance: From Body-Borne Cameras to Drones

Summarizing Long First-Person Videos

Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Yong Jae Lee, Yu-Chuan Su, Bo Xiong, Lu Zheng, Ke Zhang, Wei-Lun Chao, Fei Sha

THE UNIVERSITY OF TEXAS AT AUSTIN

First person vs. Third person

Traditional third-person view

First-person view

UT TEA dataset

First person vs. Third person

First person “egocentric” vision:

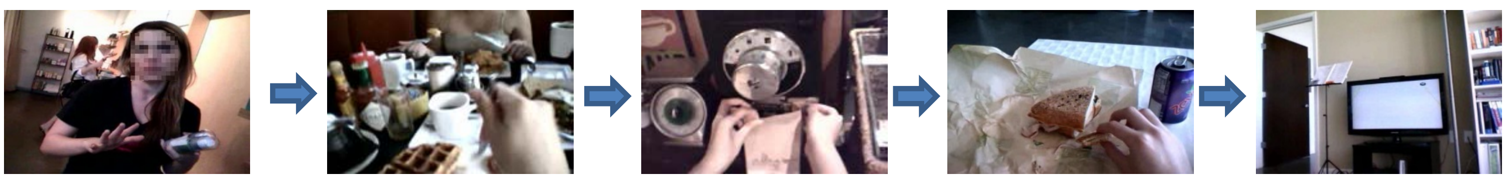
- Linked to ongoing experience of the camera wearer
- World seen in context of the camera wearer’s activity and goals

UT Interaction and JPL First-Person Interaction datasets

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

Why summarize egocentric video?

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: Video summarization

- Largely third-person
  - Static cameras, low-level cues informative
- Consider summarization as a *sampling* problem

[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,…]

Goal: Story-driven summarization

Characters and plot ↔ Key objects and influence

[Lu & Grauman, CVPR 2013]

Goal: Story-driven summarization

Characters and plot ↔ Key objects and influence

[Lu & Grauman, CVPR 2013]

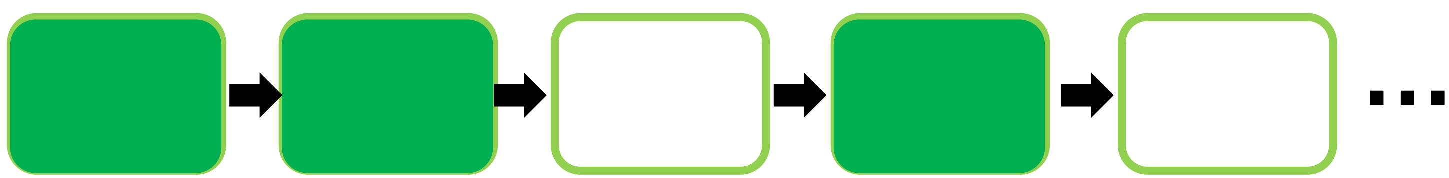
Summarization as subshot selection

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

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

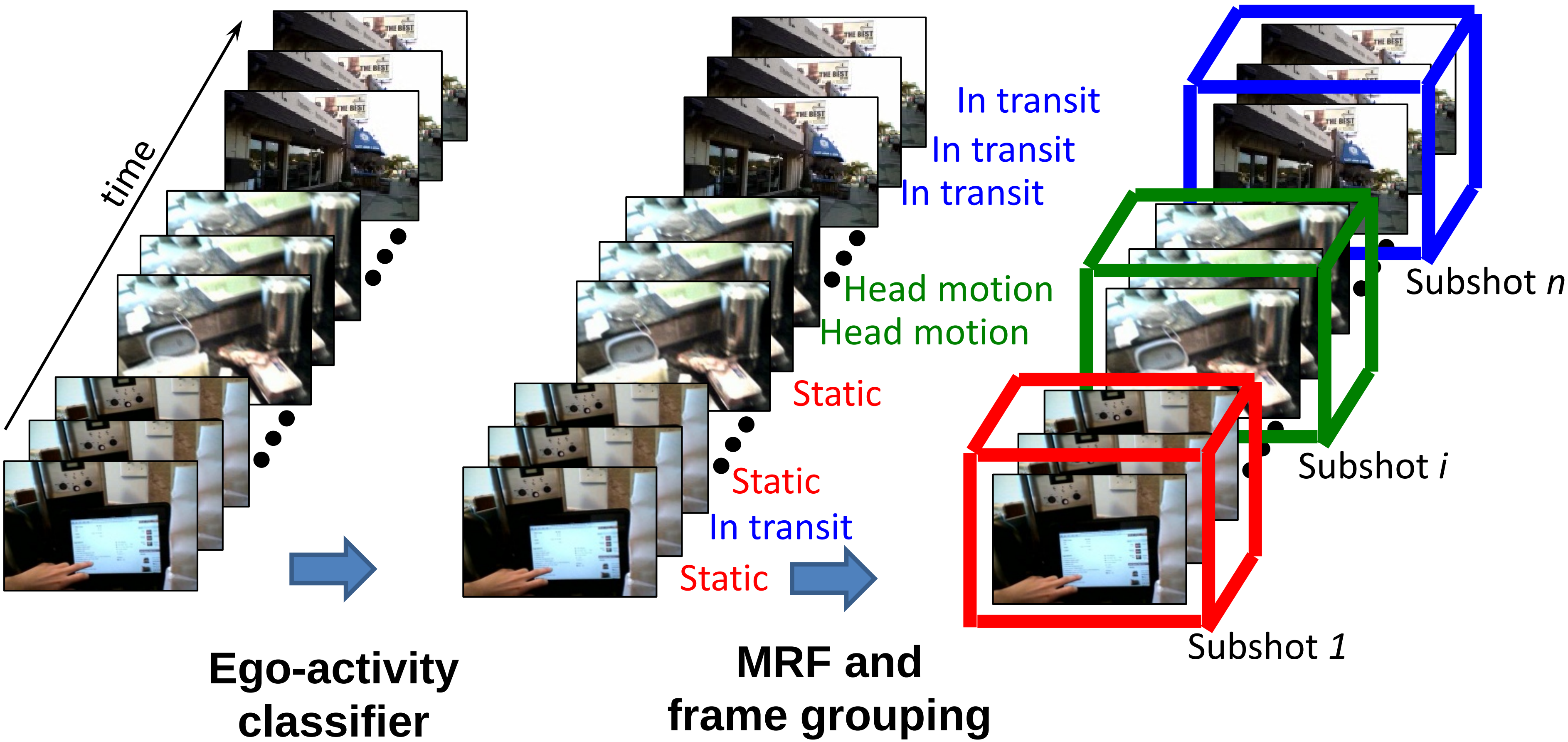
influence **importance** **diversity**

Subshots



[Lu & Grauman, CVPR 2013]

Egocentric subshot detection



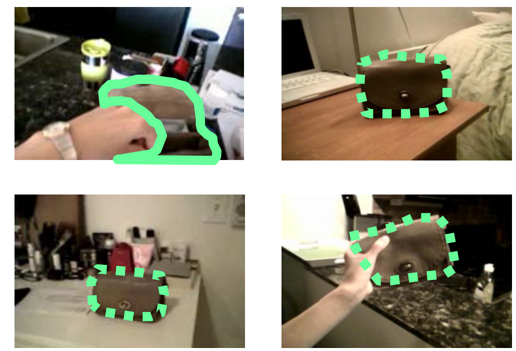
[Lu & Grauman, CVPR 2013]

Learning object importance

We learn to rate regions by their egocentric importance

distance to hand

distance to frame center



frequency

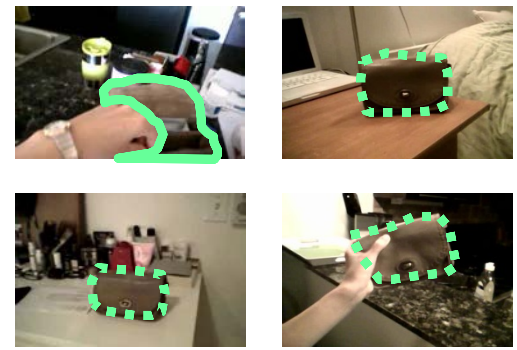
[Lee et al. CVPR 2012, IJCV 2015]

Learning object importance

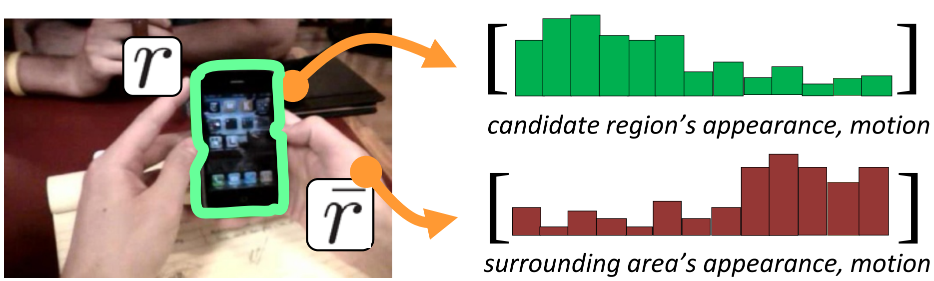
We learn to rate regions by their egocentric importance

distance to hand

distance to frame center

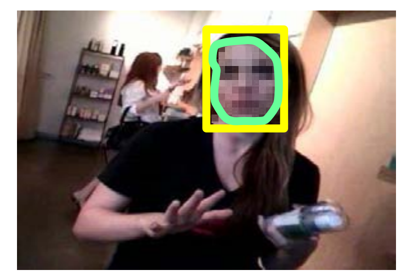


frequency



"Object-like" appearance, motion

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



overlap w/ face detection

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

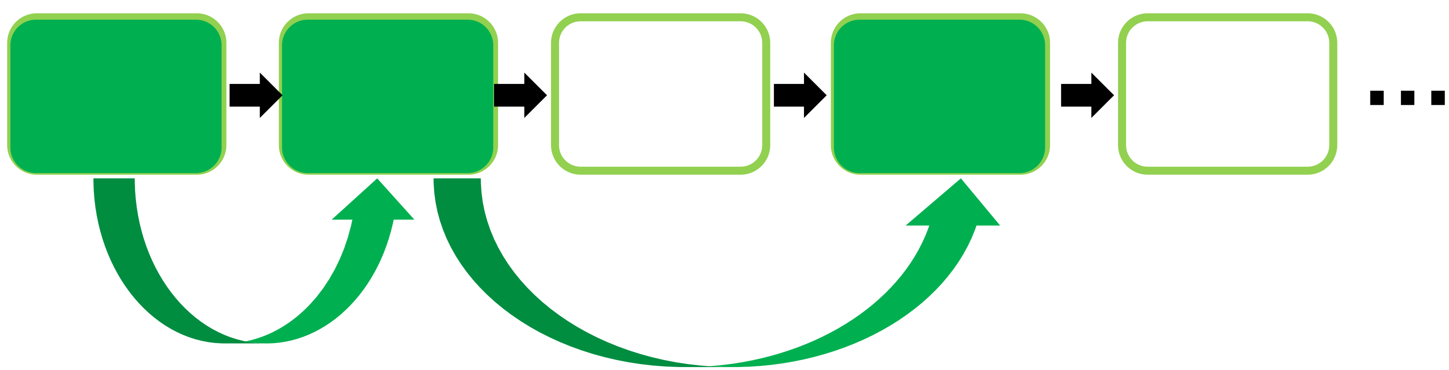
[Lee et al. CVPR 2012, IJCV 2015]

Estimating visual influence

- Aim to select the *k* subshots that maximize the influence between objects (on the weakest link)

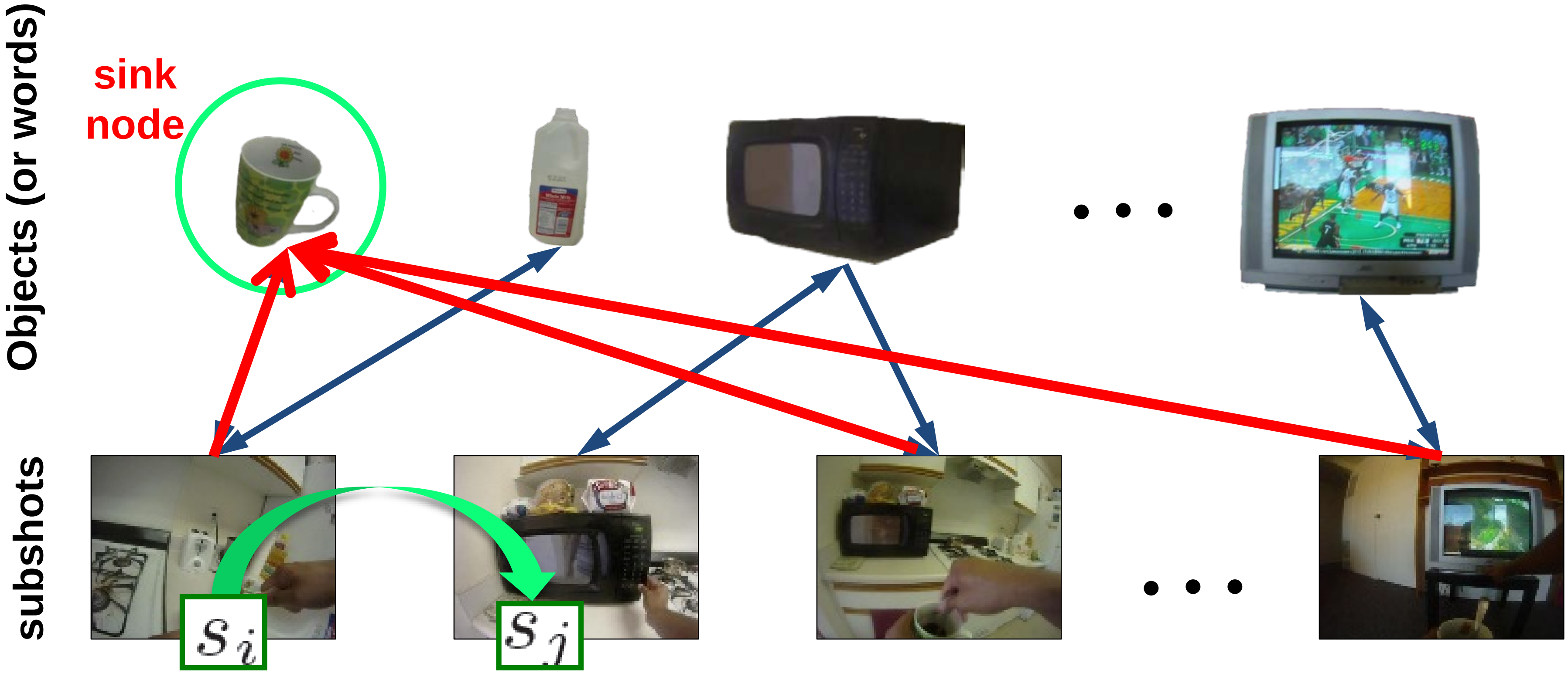
$$\mathcal{S}(S) = \max_{a} \min_{j=1,\dots,K-1} \sum_{o_i \in O} a_{i,j} \textsc{Influence}(s_j, s_{j+1} | o_i)$$

Subshots



[Lu & Grauman, CVPR 2013]

Estimating visual influence



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

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

[Lu & Grauman, CVPR 2013]

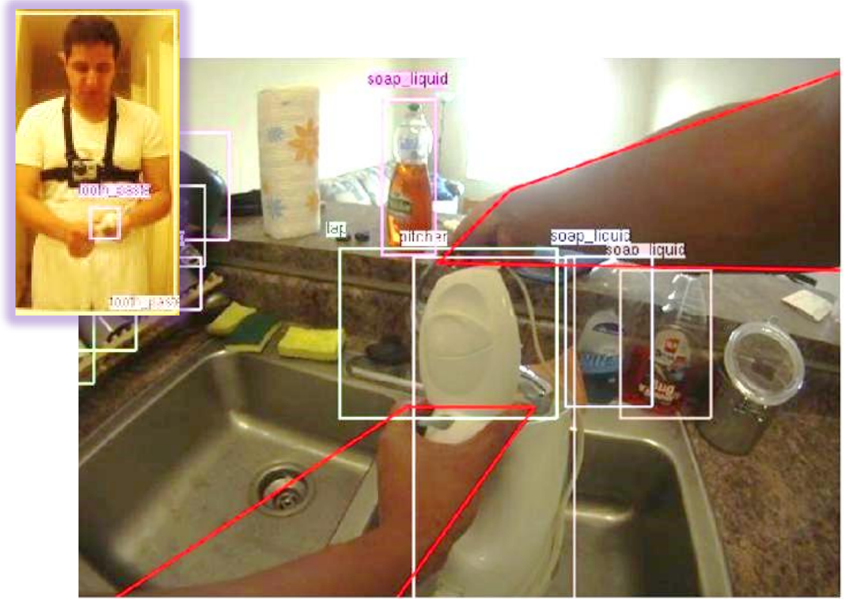
Datasets

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 2012]



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

We use object bounding boxes and keyframes.

Example keyframe summary – UT Ego data

`http://vision.cs.utexas.edu/projects/egocentric/`

Original video (3 hours)

Our summary (12 frames)

[Lee et al. CVPR 2012, IJCV 2015]

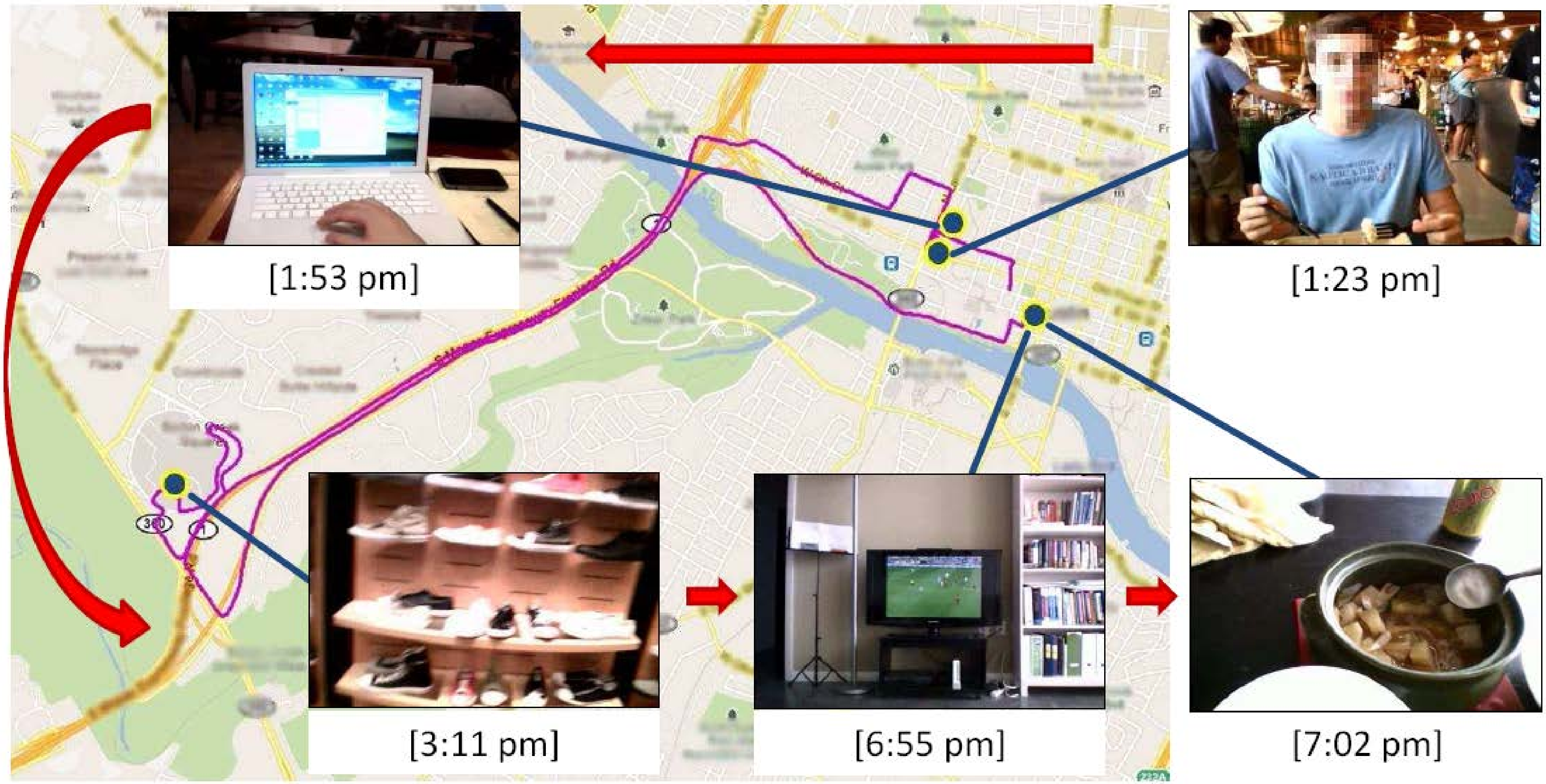
Example skim summary – UT Ego data

Ours

Baseline

[Lu & Grauman, CVPR 2013]

Generating storyboard maps



Augment keyframe summary with geolocations

[Lee et al., CVPR 2012, IJCV 2015]

Human subject results: Blind taste test

How often do subjects prefer our summary?

Data	Vs. Uniform sampling	Vs. Shortest-path	Vs. Object-driven Lee et al. 2012
UT Egocentric Dataset	**90.0%**	**90.9%**	**81.8%**
Activities Daily Living	**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]

Summarizing egocentric video

Key questions

– What objects are important, and how are they linked?

– When is recorder engaging with scene?

– Which frames look “intentional”?

– Can we teach a system to summarize?

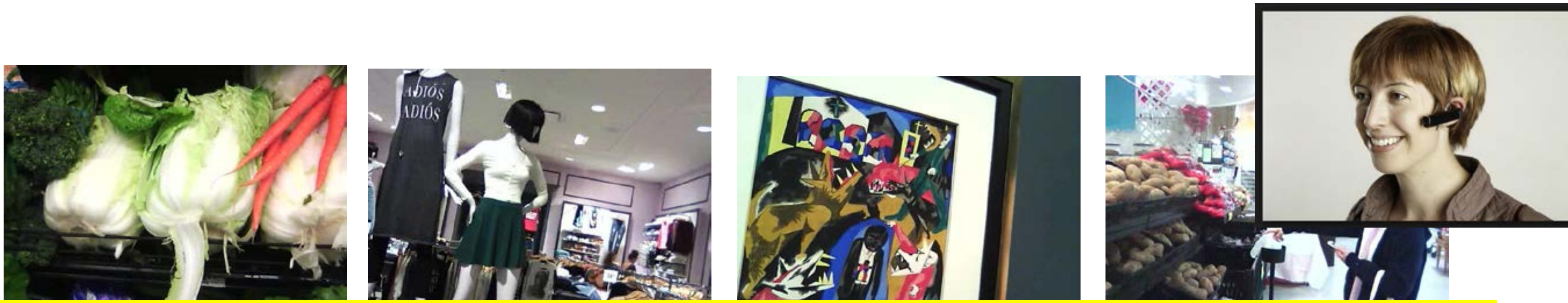
Goal: Detect engagement

Definition:

A time interval where the recorder is attracted by some object(s) and he interrupts his ongoing flow of activity to purposefully gather more information about the object(s)

[Su & Grauman, ECCV 2016]

Egocentric Engagement Dataset

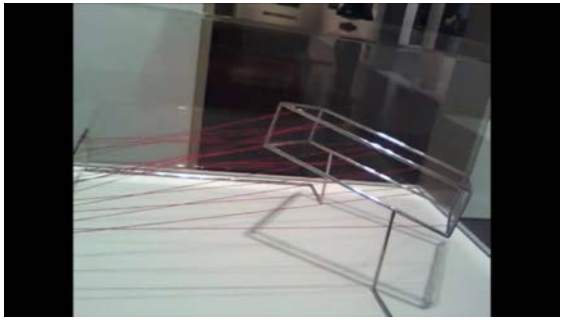


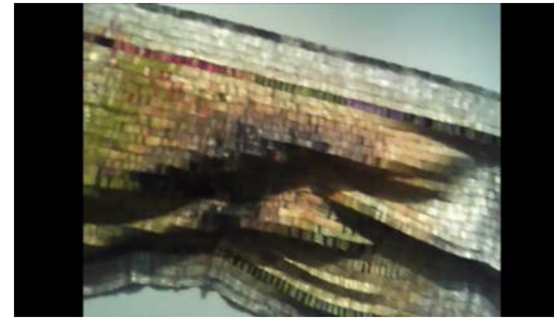
14 hours of labeled ego video

- “Browsing” scenarios, long & natural clips
- 14 hours of video, 9 recorders
- Frame-level labels x 10 annotators

[Su & Grauman, ECCV 2016]

Challenges in detecting engagement



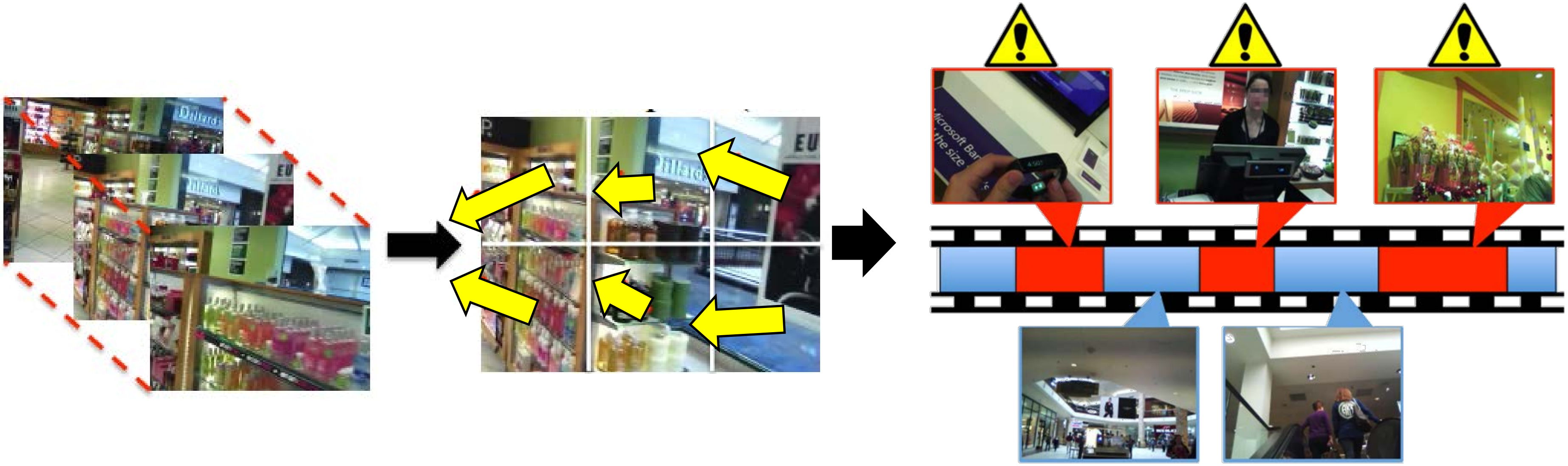



- Interesting things vary in appearance!
- Being engaged ≠ being stationary
- High engagement intervals vary in length
- Lack cues of active camera control

[Su & Grauman, ECCV 2016]

Our approach

Learn motion patterns indicative of engagement



[Su & Grauman, ECCV 2016]

Results: detecting engagement

Blue=Ground truth
Red=Predicted

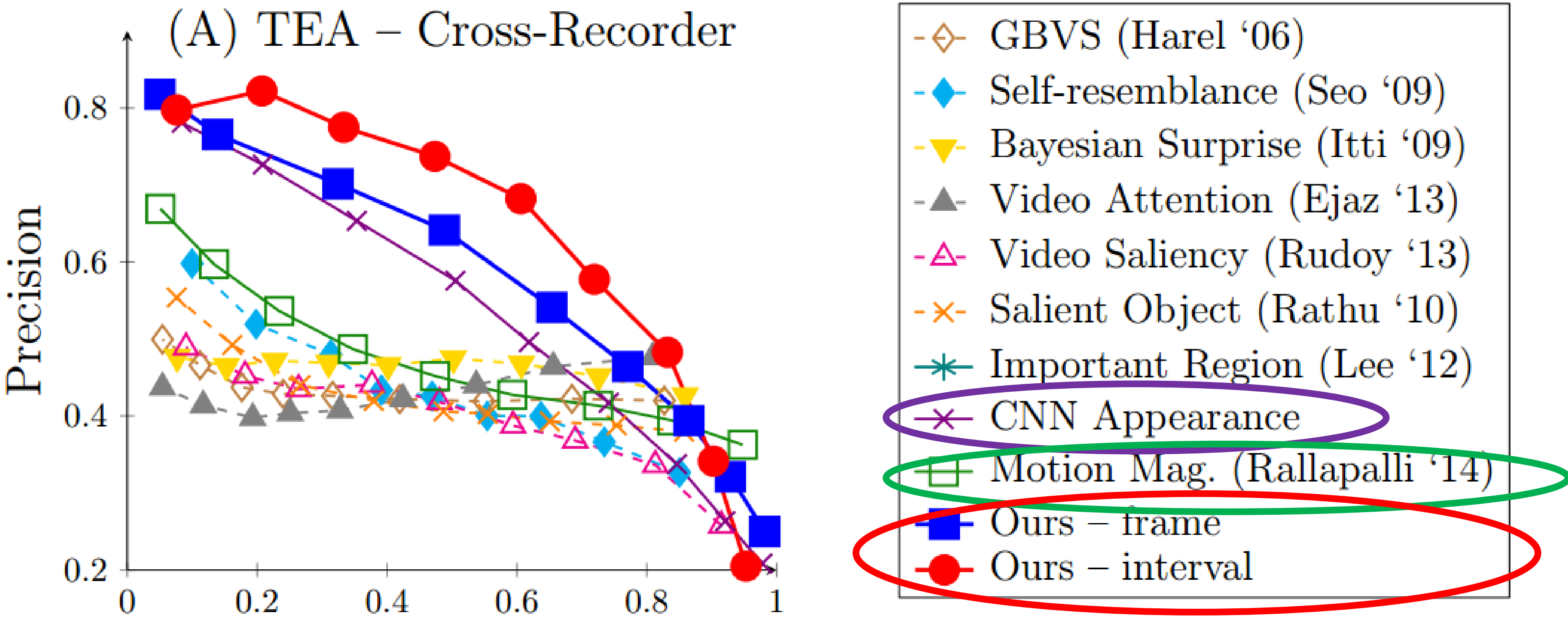
[Su & Grauman, ECCV 2016]

Results: failure cases

Blue=Ground truth
Red=Predicted

[Su & Grauman, ECCV 2016]

Results: detecting engagement



- 14 hours of video, 9 recorders

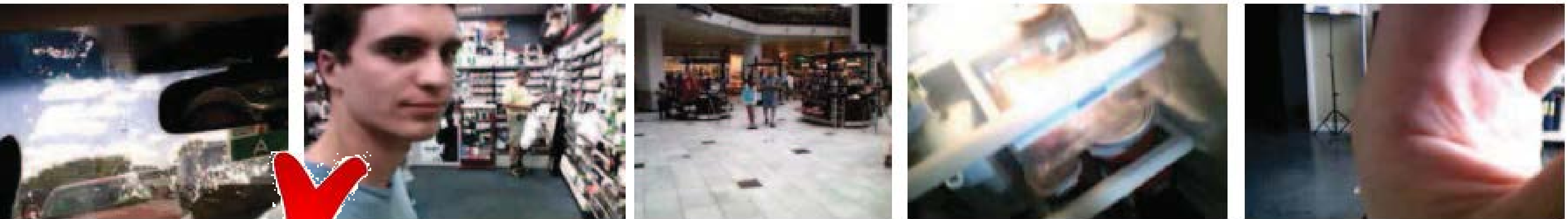
[Su & Grauman, ECCV 2016]

Summarizing egocentric video

Key questions

- What objects are important, and how are they linked?
- When is recorder engaging with scene?
- Which frames look “intentional”?
- Can we teach a system to summarize?

Which photos were purposely taken by a human?



Incidental wearable camera photos

Intentional human taken photos

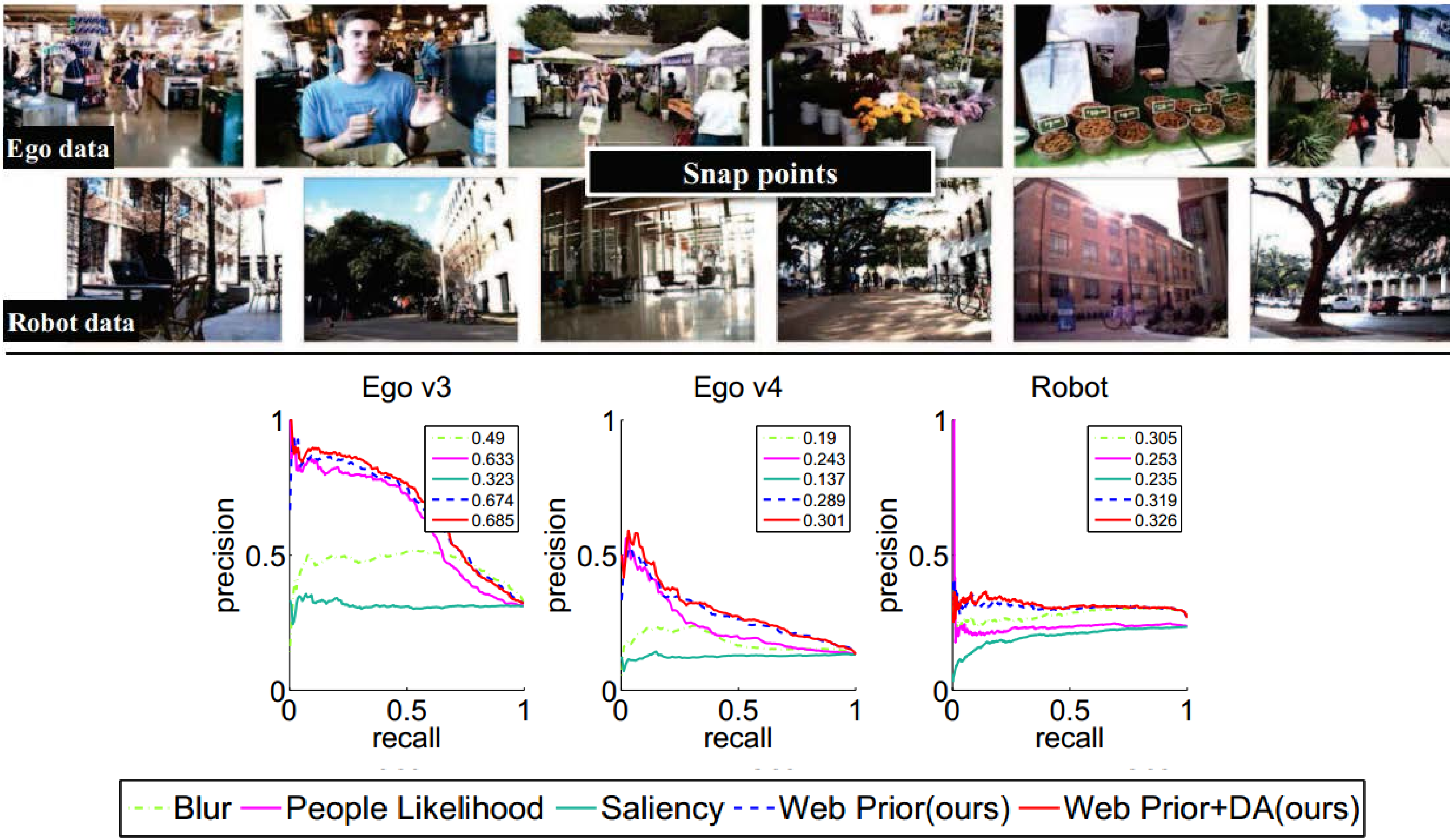
[Xiong & Grauman, ECCV 2014]

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 point predictions

[Xiong & Grauman, ECCV 2014]

Summarizing egocentric video

Key questions

- – What objects are important, and how are they linked?
- – When is recorder engaging with scene?
- – Which frames look “intentional”?
- – Can we teach a system to summarize?

Supervised summarization

- Can we *teach* the system how to create a good summary, based on human-edited exemplars?

[Zhang et al. CVPR 2016, Chao et al. UAI 2015, Gong et al. NIPS 2014]

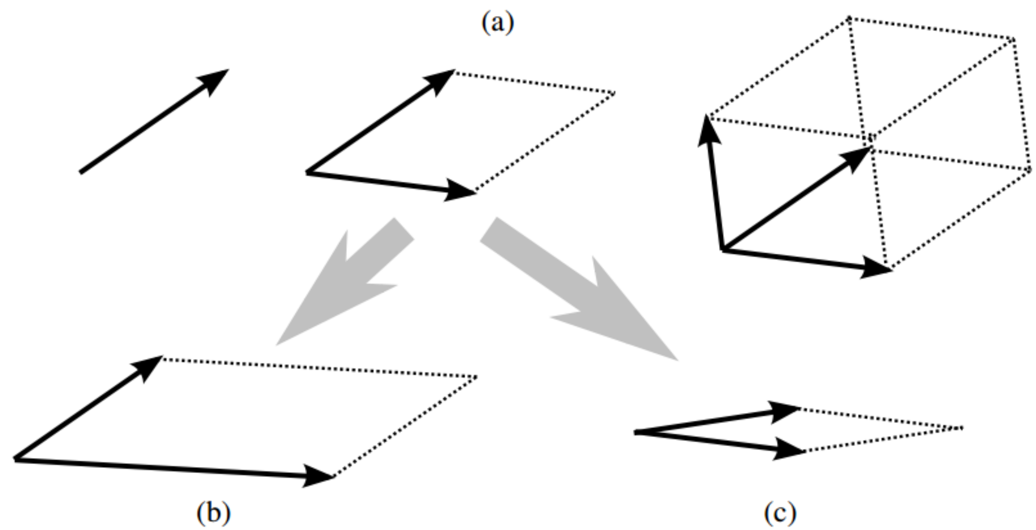
Determinantal Point Processes for video summarization

- Select subset of items that maximizes diversity and “quality”

$$P(\boldsymbol{y} | \boldsymbol{L}) = \frac{\det(\boldsymbol{L}\{\boldsymbol{y}\})}{\det(\boldsymbol{L} + \boldsymbol{I})}$$

subset indicator

$N \times N$ similarity



“quality” items

diverse items

Figure: Kulesza & Taskar

[Zhang et al. CVPR 2016, Chao et al. UAI 2015, Gong et al. NIPS 2014]

Summary Transfer

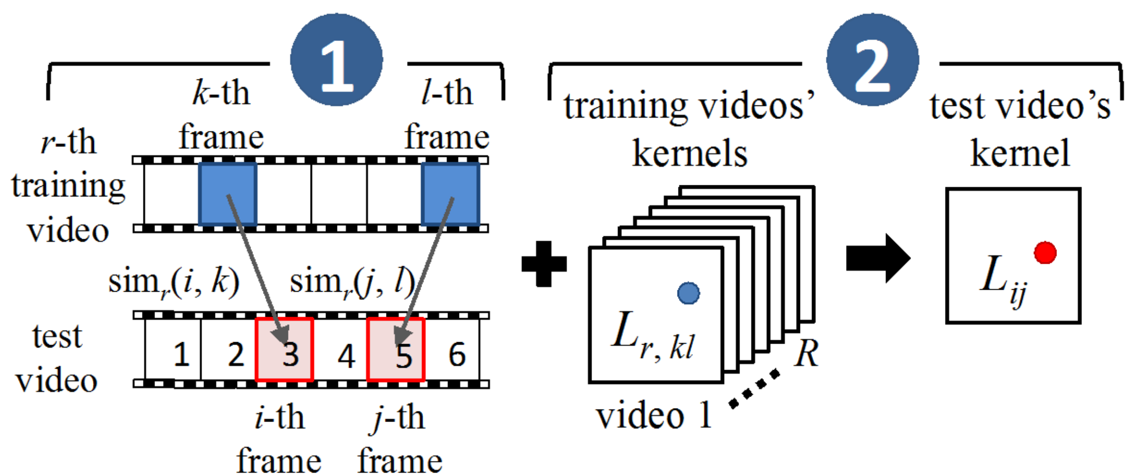
Ke Zhang (USC), Wei-Lun Chao (USC), Fei Sha (UCLA), Kristen Grauman (UT Austin)

- **Idea:** Transfer the underlying summarization structures

Training kernels:
"idealized"

$$L_r = \alpha_r \begin{bmatrix} \delta(1 \in \mathbf{y}_r) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta(N_r \in \mathbf{y}_r) \end{bmatrix}$$

Test kernel:
Synthesized from *related* training kernels



Zhang et al. CVPR 2016

Summary Transfer

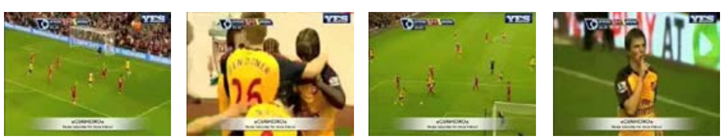
Ke Zhang (USC), Wei-Lun Chao (USC), Fei Sha (UCLA), Kristen Grauman (UT Austin)

Promising results on existing annotated datasets

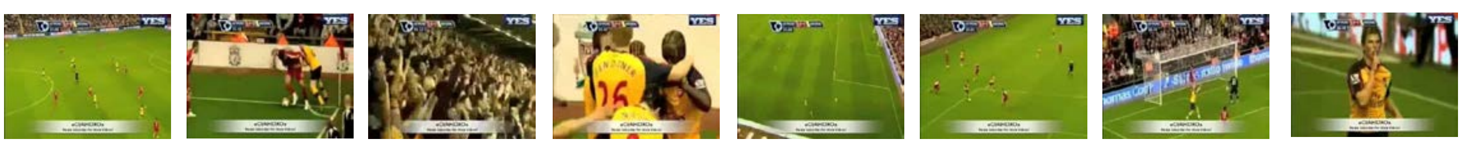
	Kodak (18)	OVP (50)	YouTube (31)	MED (160)
VSUMM [Avila ’11]	69.5	70.3	59.9	28.9
seqDPP [Gong ’14]	78.9	**77.7**	60.8	-
Ours	**82.3**	76.5	**61.8**	**30.7**

	VidMMR [Li ’10]	SumMe [Gygli ’14]	Submodular [Gygli ’15]	**Ours**
SumMe (25)	26.6	39.3	39.7	**40.9**

VSUMM_1 (F = 54)



seqDPP (F = 57)



Ours (F = 74)

Zhang et al. CVPR 2016

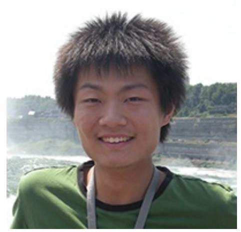
Next steps

- Video summary as an index for search
- Streaming computation
- Visualization, display
- Multiple modalities – e.g., audio, depth,…

Summary

Yong Jae Lee

Yu-Chuan Su

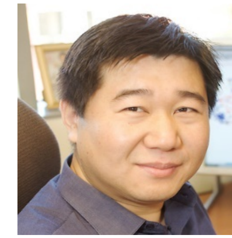


Bo Xiong

Lu Zheng

Ke Zhang

Wei-Lun Chao



Fei Sha

- First-person summarization tools needed to cope with deluge of wearable camera data
- **New ideas**
  - Story-like summaries
  - Detecting *when* engagement occurs
  - Intentional=looking snap points from a passive camera
  - Supervised summarization learning methods

Papers

- **Summary Transfer: Exemplar-based Subset Selection for Video Summarization**. K. Zhang, W-L. Chao, F. Sha, and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, June 2016.
- **Detecting Snap Points in Egocentric Video with a Web Photo Prior**. B. Xiong and K. Grauman. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, Sept 2014.
- **Detecting Engagement in Egocentric Video**. Y-C. Su and K. Grauman. To appear, Proceedings of the European Conference on Computer Vision (ECCV), 2016.
- **Predicting Important Objects for Egocentric Video Summarization**. Y J. Lee and K. Grauman. International Journal on Computer Vision, Volume 114, Issue 1, pp. 38-55, August 2015.
- **Story-Driven Summarization for Egocentric Video**. Z. Lu and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, June 2013.
- **Discovering Important People and Objects for Egocentric Video Summarization**. Y. J. Lee, J. Ghosh, and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 2012.