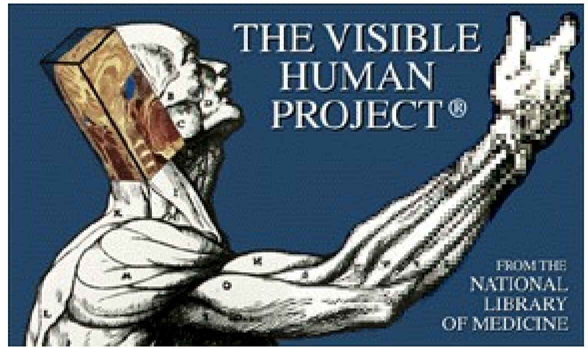
Summarizing video

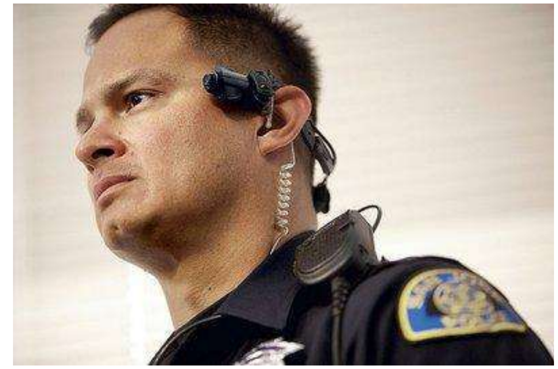
Kristen Grauman
Department of Computer Science
University of Texas at Austin

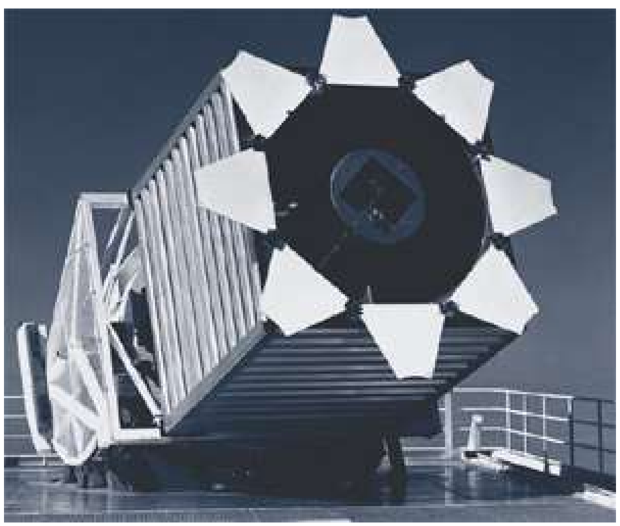
THE UNIVERSITY OF TEXAS AT AUSTIN

Presented at the TFI Communications Technology & Asset Valuation Conference, January 26-27, 2017, Austin, Texas

Explosion of visual data





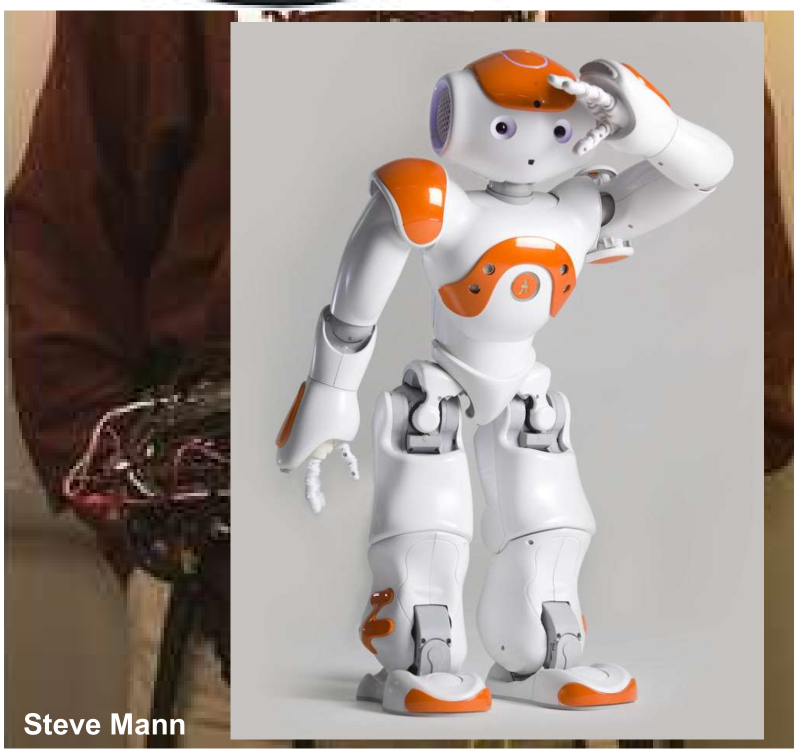


Kristen Grauman, UT Austin


~1990
Steve Mann

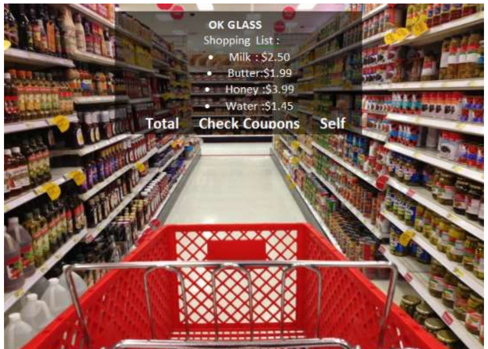

2017


360° Velodyne Laserscanner
Stereo Camera Rig
GPS
Monochrome
Color
KIT
KA EV 842


Steve Mann


2017

New era for first-person vision



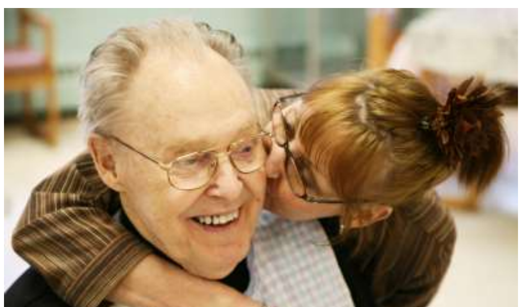
Augmented reality

Health monitoring

Law enforcement


Science

Robotics



Life logging

Kristen Grauman, UT 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:

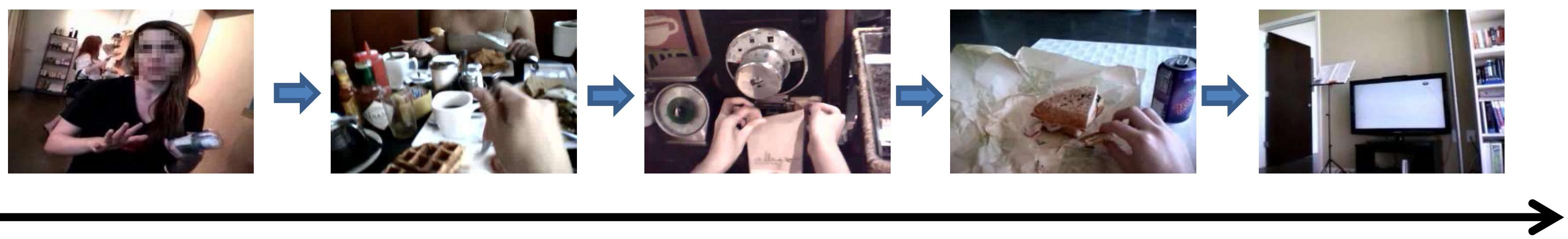
- Linked to ongoing experience of the camera wearer
- World seen in context of the camera wearer's activity and goals

UT Interaction and JPL First-Person Interaction datasets

Our 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 summary

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,…]

Idea: Story-driven summarization

Characters and plot ↔ Key objects and influence

[Lu & Grauman, CVPR 2013]

Idea: Story-driven summarization

Characters and plot ↔ Key objects and influence

[Lu & Grauman, CVPR 2013]

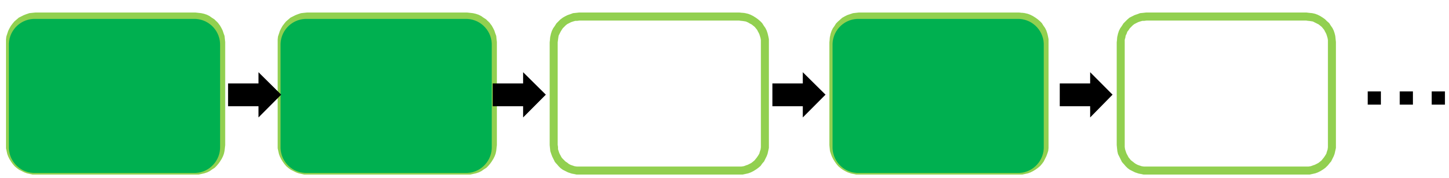
Summarization as subshot selection

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**

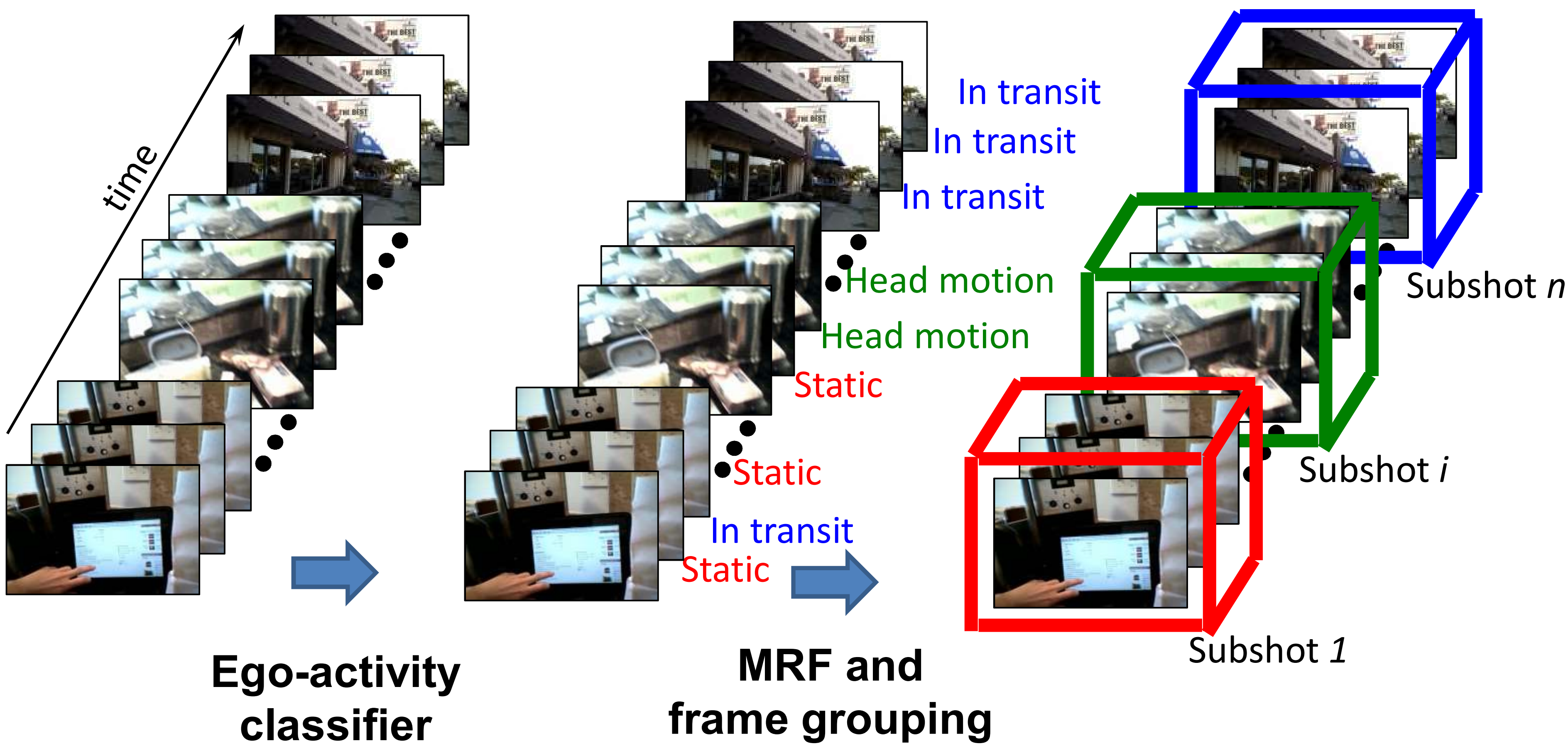
Subshots



Kristen Grauman, UT Austin

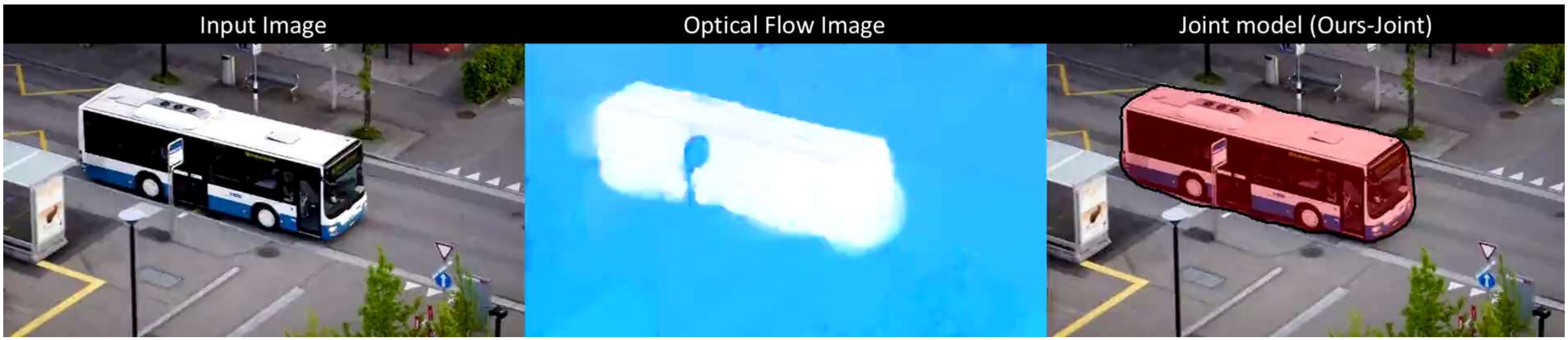
[Lu & Grauman, CVPR 2013]

Egocentric subshot detection



[Lu & Grauman, CVPR 2013]

Finding objects in video



Deep learning framework to automatically segment generic objects in video

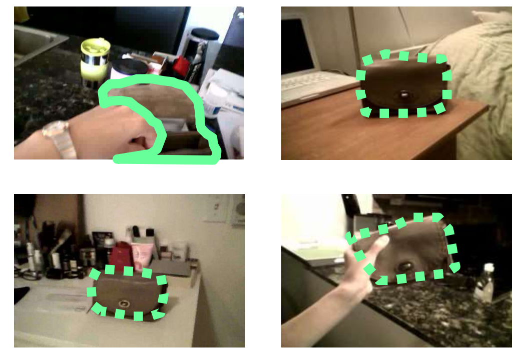
Jain, Xiong, and Grauman, 2017

Learning object importance

We learn to rate regions by their egocentric importance

distance to hand

distance to frame center



frequency

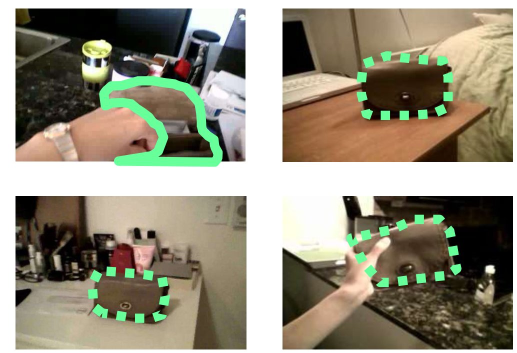
[Lee et al. CVPR 2012, IJCV 2015]

Learning object importance

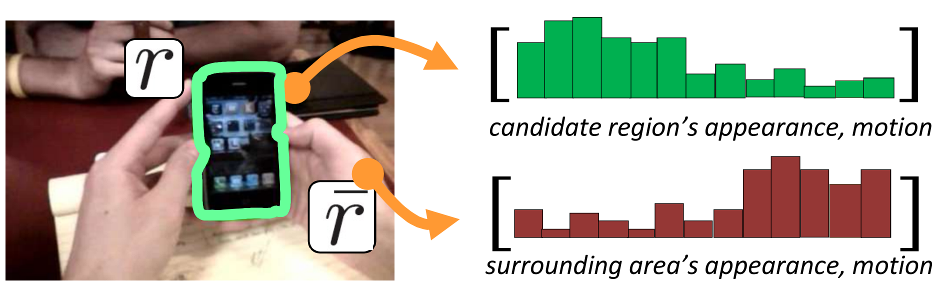
We learn to rate regions by their egocentric importance

distance to hand

distance to frame center

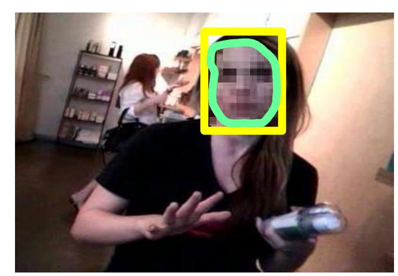


frequency



"Object-like" appearance, motion

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



overlap w/ face detection

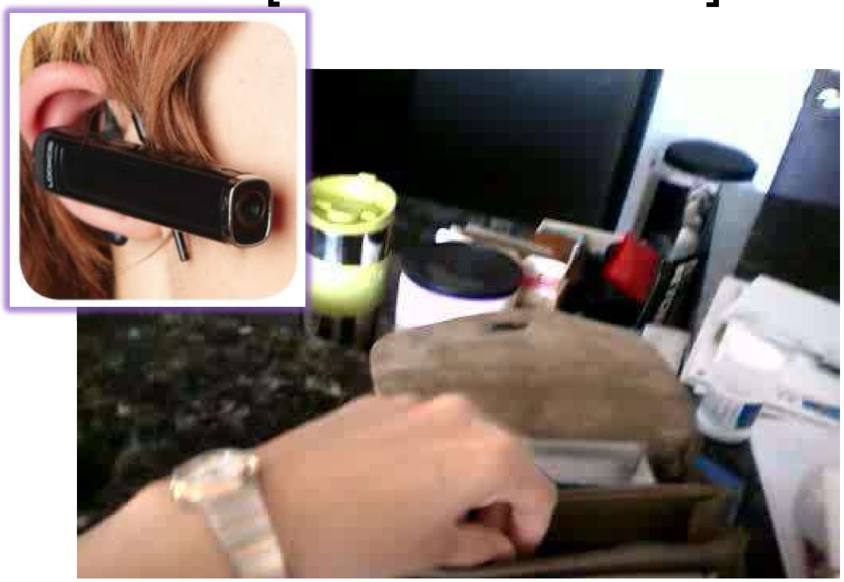
Region features: *size, width, height, centroid*

Kristen Grauman, UT Austin

[Lee et al. CVPR 2012, IJCV 2015]

Datasets

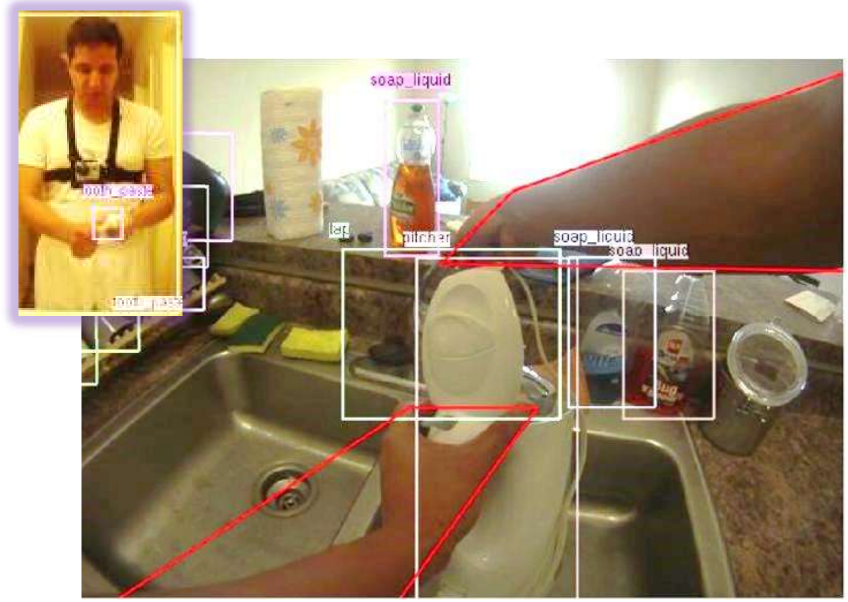
UT Egocentric (UT Ego)
[Lee et al. 2012]



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

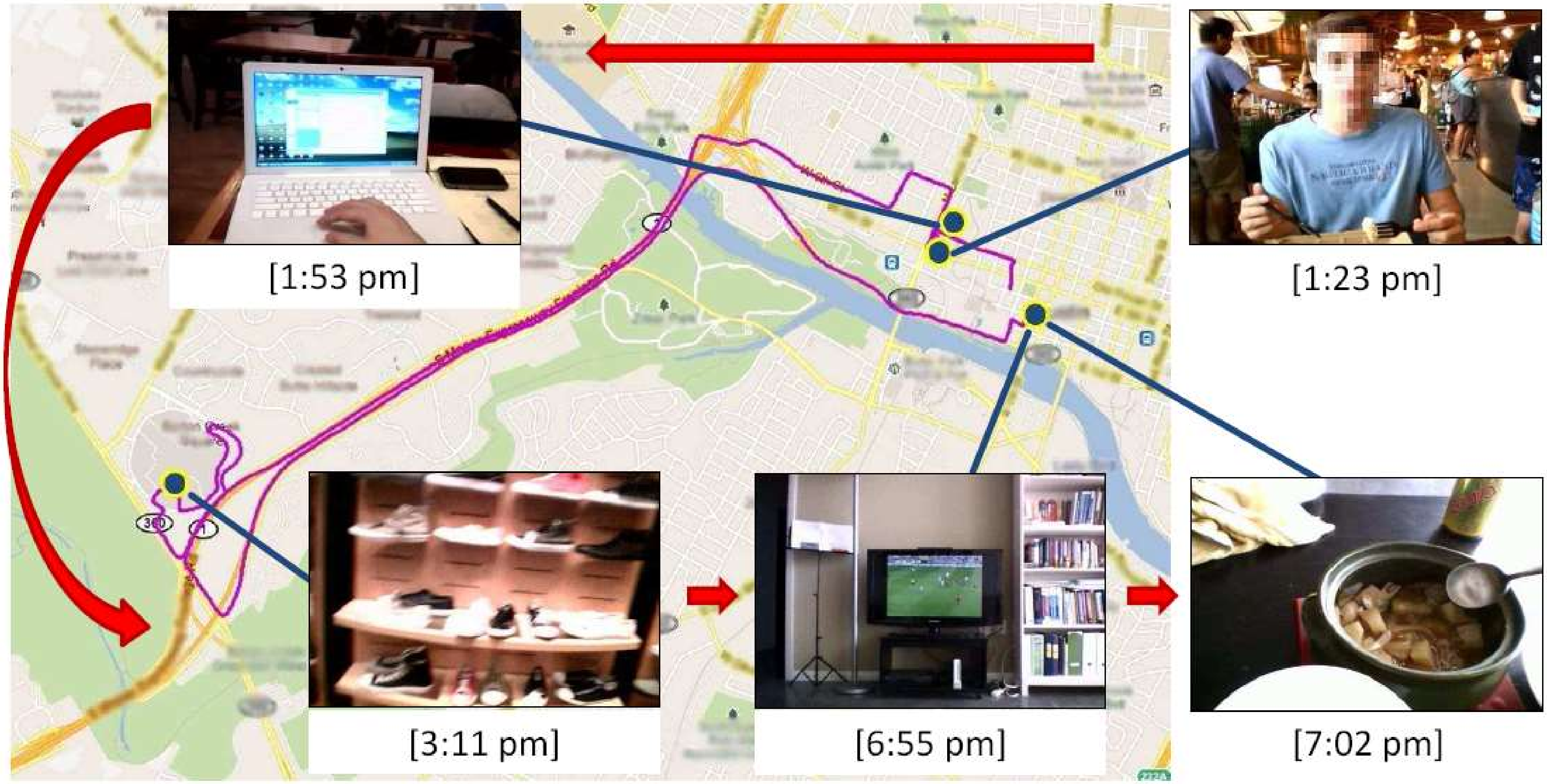
http://vision.cs.utexas.edu/projects/egocentric/

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 video

Key questions

– What is the story told by important objects?

– When is recorder engaging with scene?

– Where to look within a wide field of view?

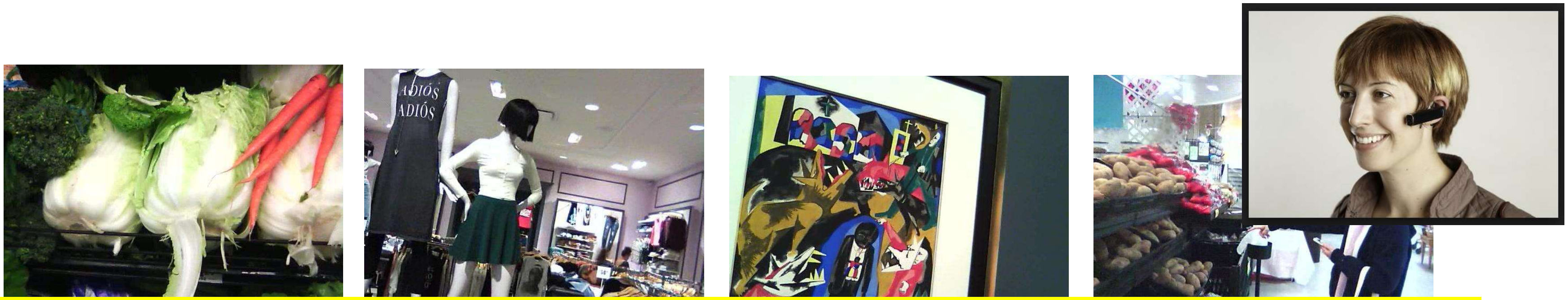
Goal: Detect engagement

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, 2015]

Egocentric Engagement Dataset

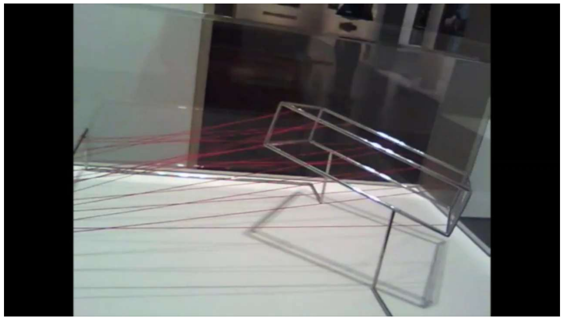
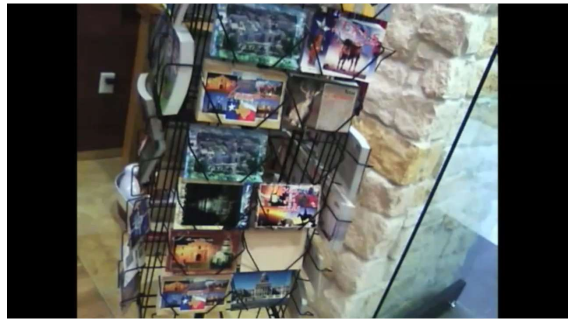
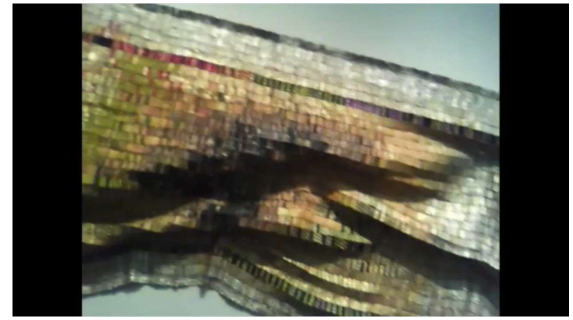


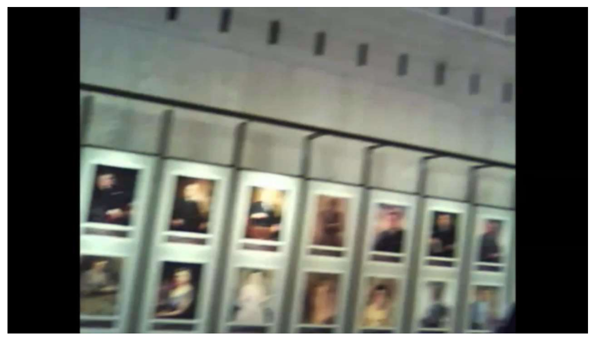
14 hours of labeled ego video

- “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



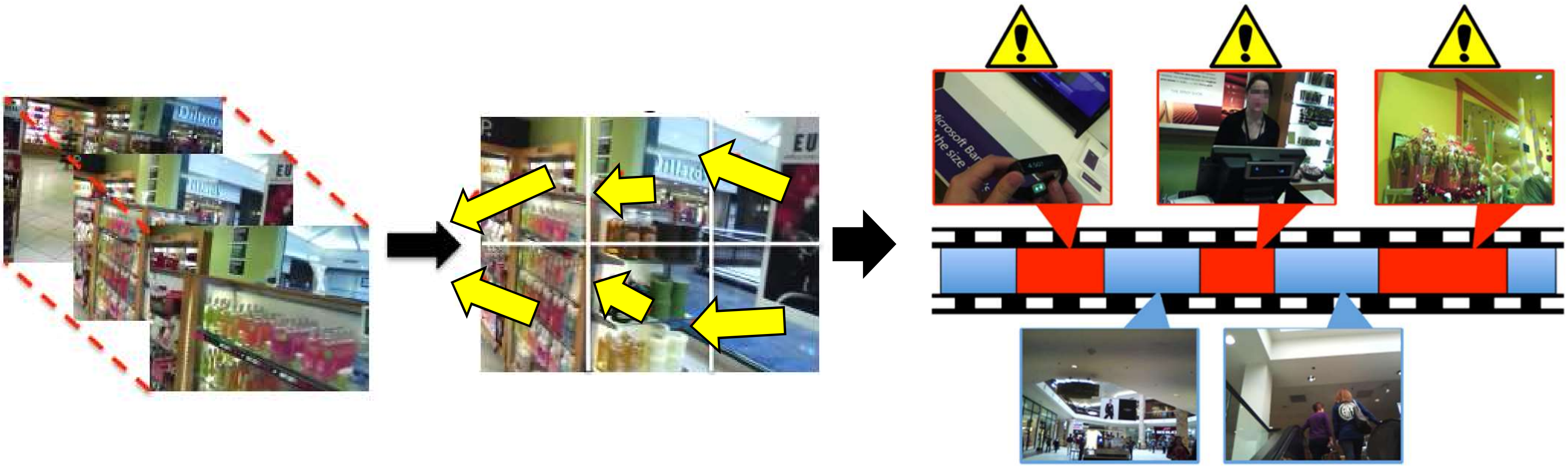

- Interesting things vary in appearance!
- Being engaged ≠ being stationary
- High engagement intervals vary in length
- Lack cues of active camera control

Kristen Grauman, UT Austin

[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]

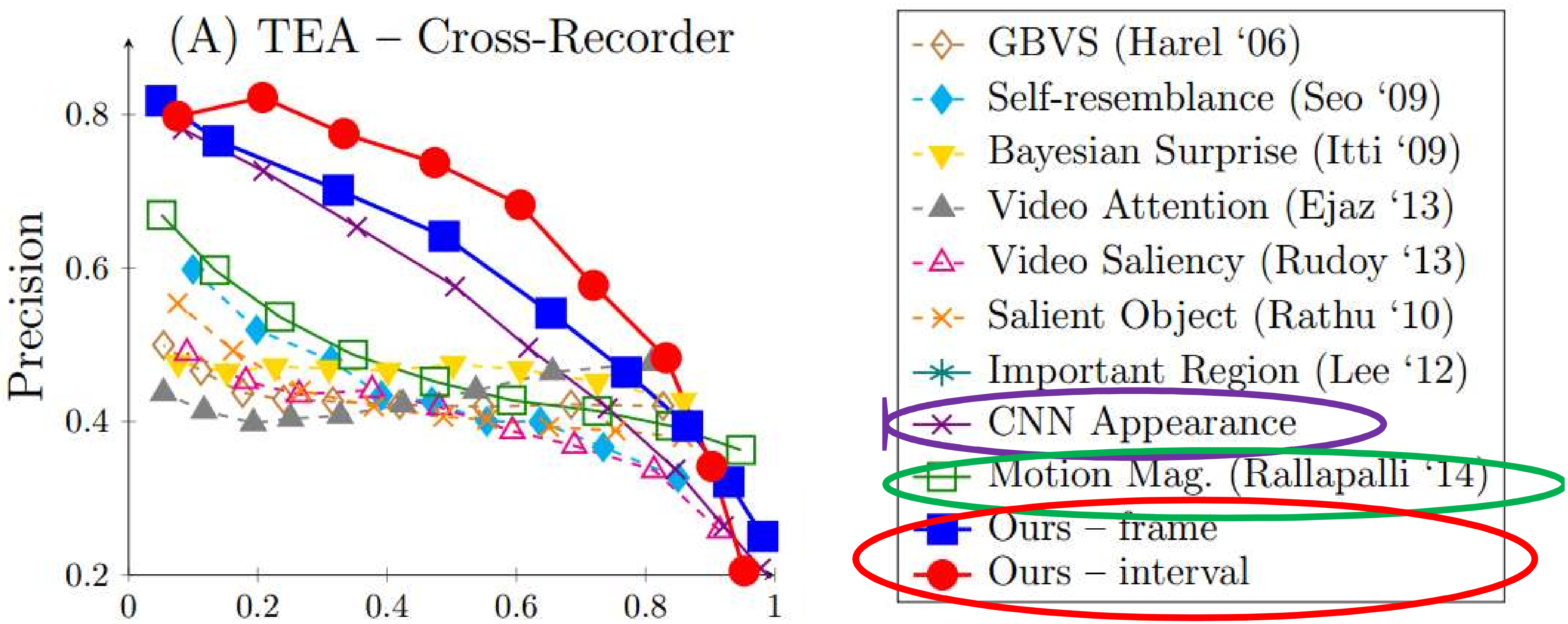
Results: failure cases

Blue=Ground truth
Red=Predicted

Kristen Grauman, UT Austin

[Su & Grauman, ECCV 2016]

Results: detecting engagement



- 14 hours of video, 9 recorders

[Su & Grauman, ECCV 2016]

Summarizing video

Key questions

- What is the story told by important objects?
- When is recorder engaging with scene?
- Where to look within a wide field of view?

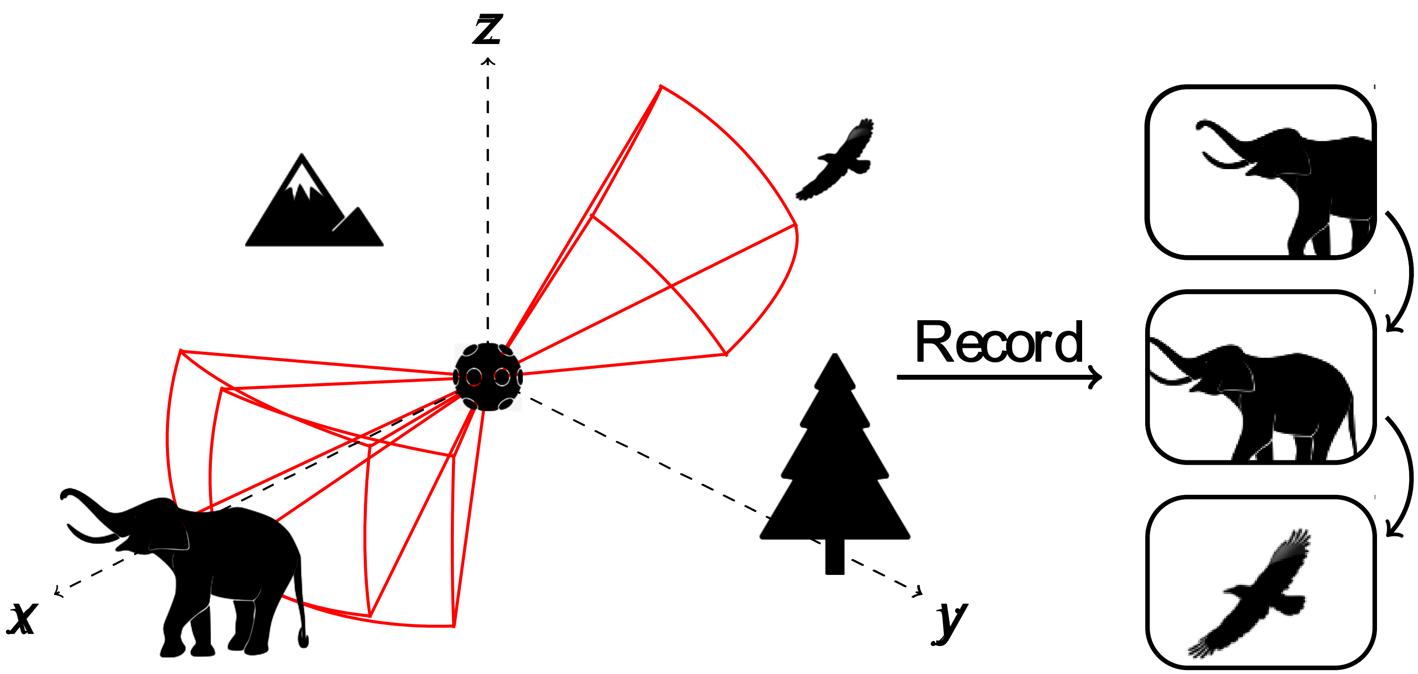
360° Cameras

Challenge of viewing 360° videos

Control by mouse

How to find the right direction to watch?

New problem: Pano2Vid



Pano2Vid Definition

Input: 360° video

Output: natural-looking normal-field-of-view video

Task: control the virtual camera direction

Kristen Grauman, UT Austin

[Su et al. ACCV 2016]

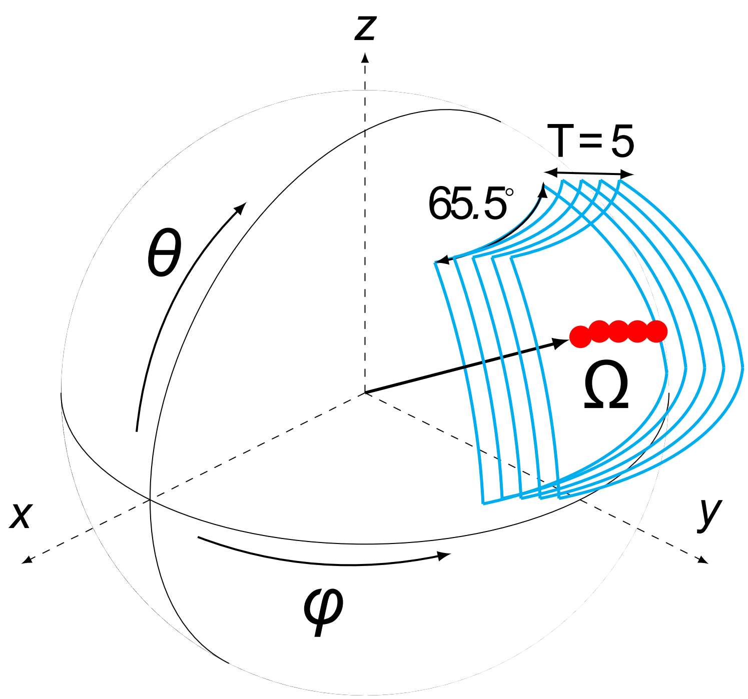
Our approach – AutoCam

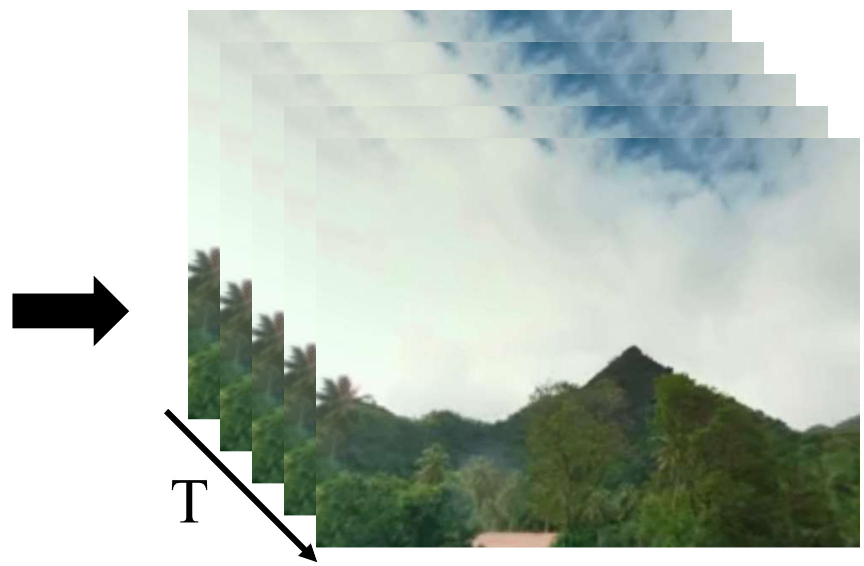
1. Handle unrestricted real 360° video
2. Learn videography tendencies from unlabeled Web videos
   - Diverse capture-worthy content
   - Proper composition

[Su et al. ACCV 2016]

Spatio-temporal glimpse

- Short NFOV video extracted from 360° video
- Makes 360° content comparable with NFOV videos





Scoring capture-worthiness

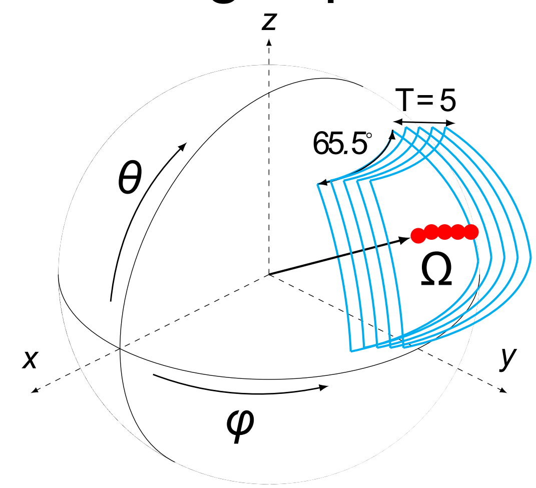
- Capture-worthiness
  - Does the spatio-temporal glimpse look human-captured?

Human-captured NFOV videos (“HumanCam”)

How close?

C3D features [Tran et al. 2015]

ST-glimpses

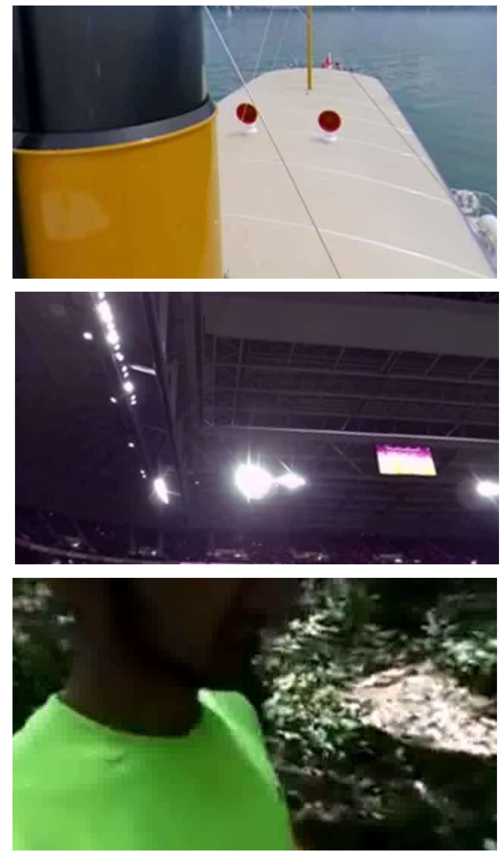


Unlabeled video

Example spatio-temporal glimpses

High capture-worthiness

Low capture-worthiness



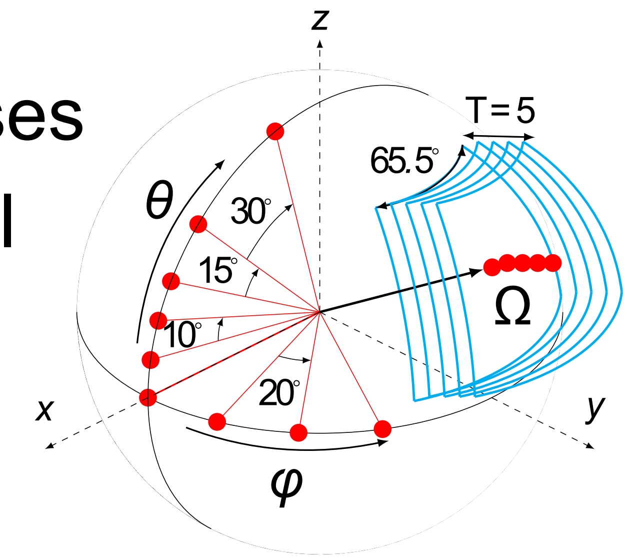
First frame of glimpses scored high/low by our approach

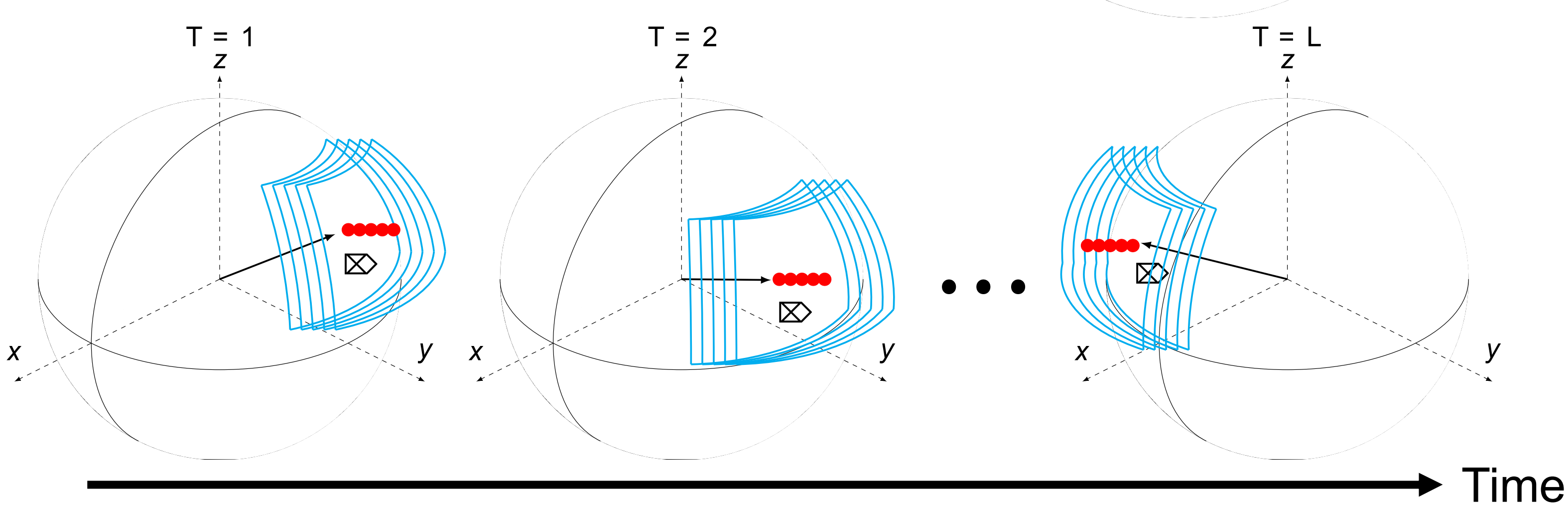
Kristen Grauman, UT Austin

Construct virtual camera trajectory

- Densely sample ST-glimpses
- Continuous camera control

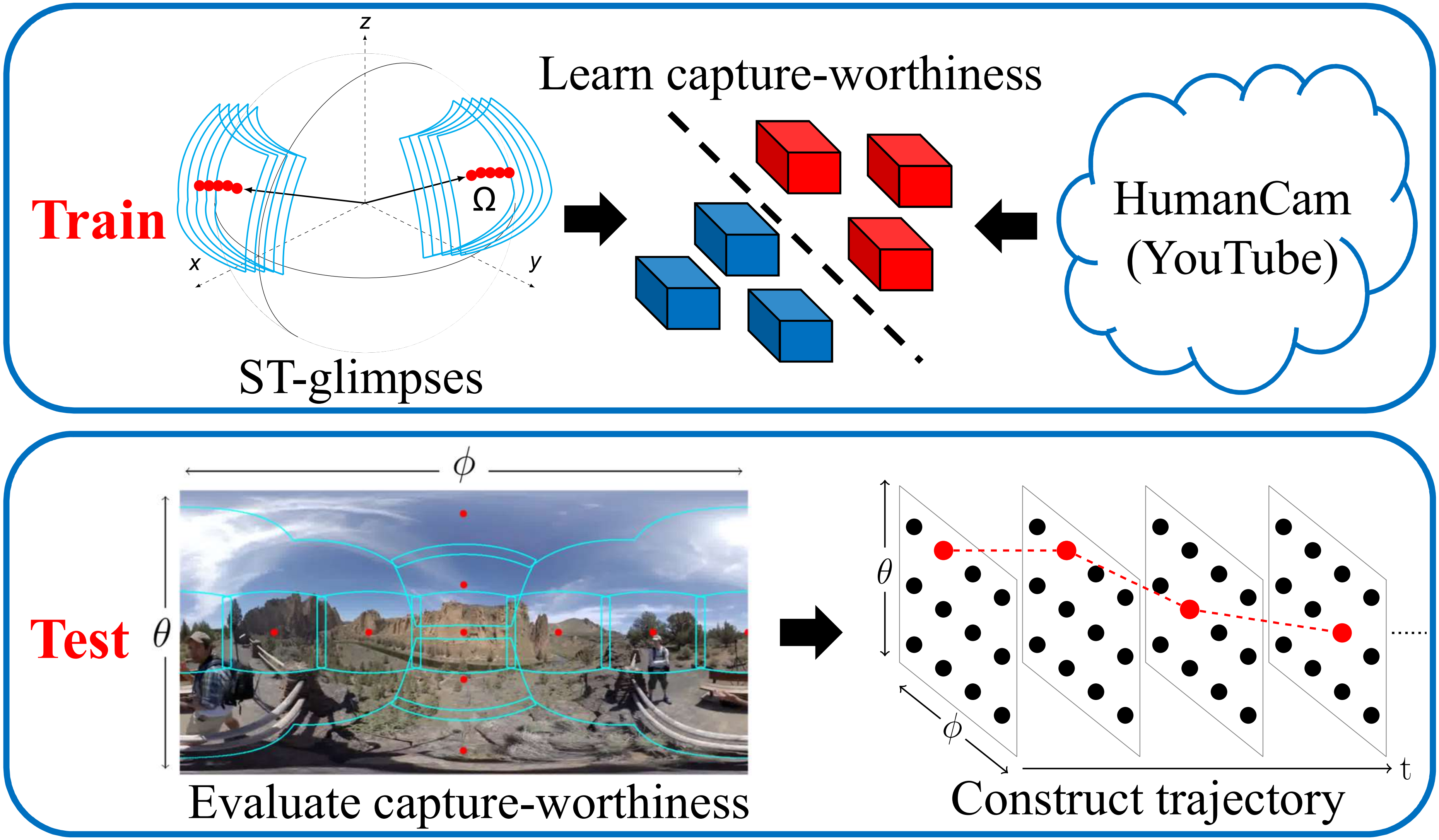
ST-glimpse selection





Pose as shortest path(s) problem

AutoCam Recap



Kristen Grauman, UT Austin

[Su et al. ACCV 2016]

Datasets

- All videos crawled from YouTube using keywords:

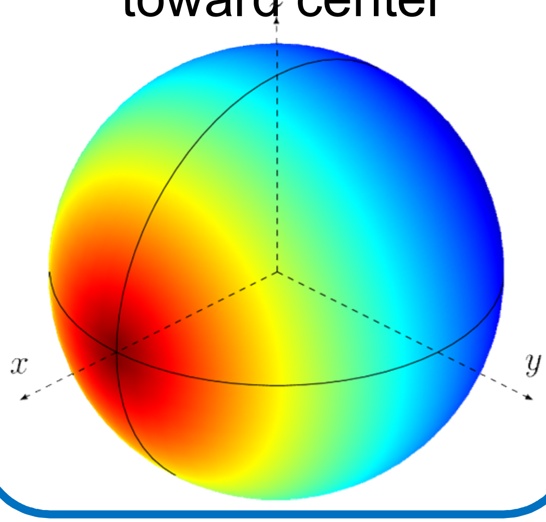
  “Hiking”, “Mountain climbing”, “Parade”, “Soccer”

	# videos	**Total length**
360° videos	86	7.3 hours
HumanCam	9,171	343 hours

Baselines

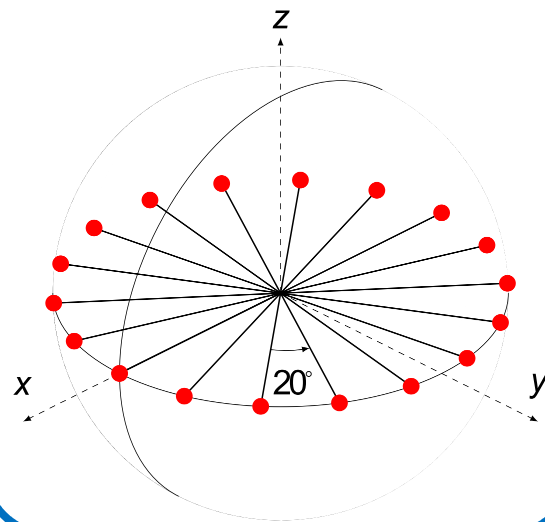
Center prior

Randomized trajectories biased toward center



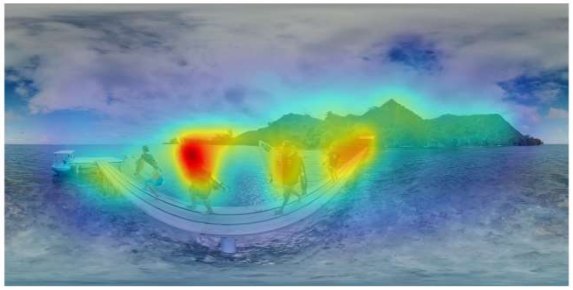
Eye-level prior

Static trajectories lying on the equator



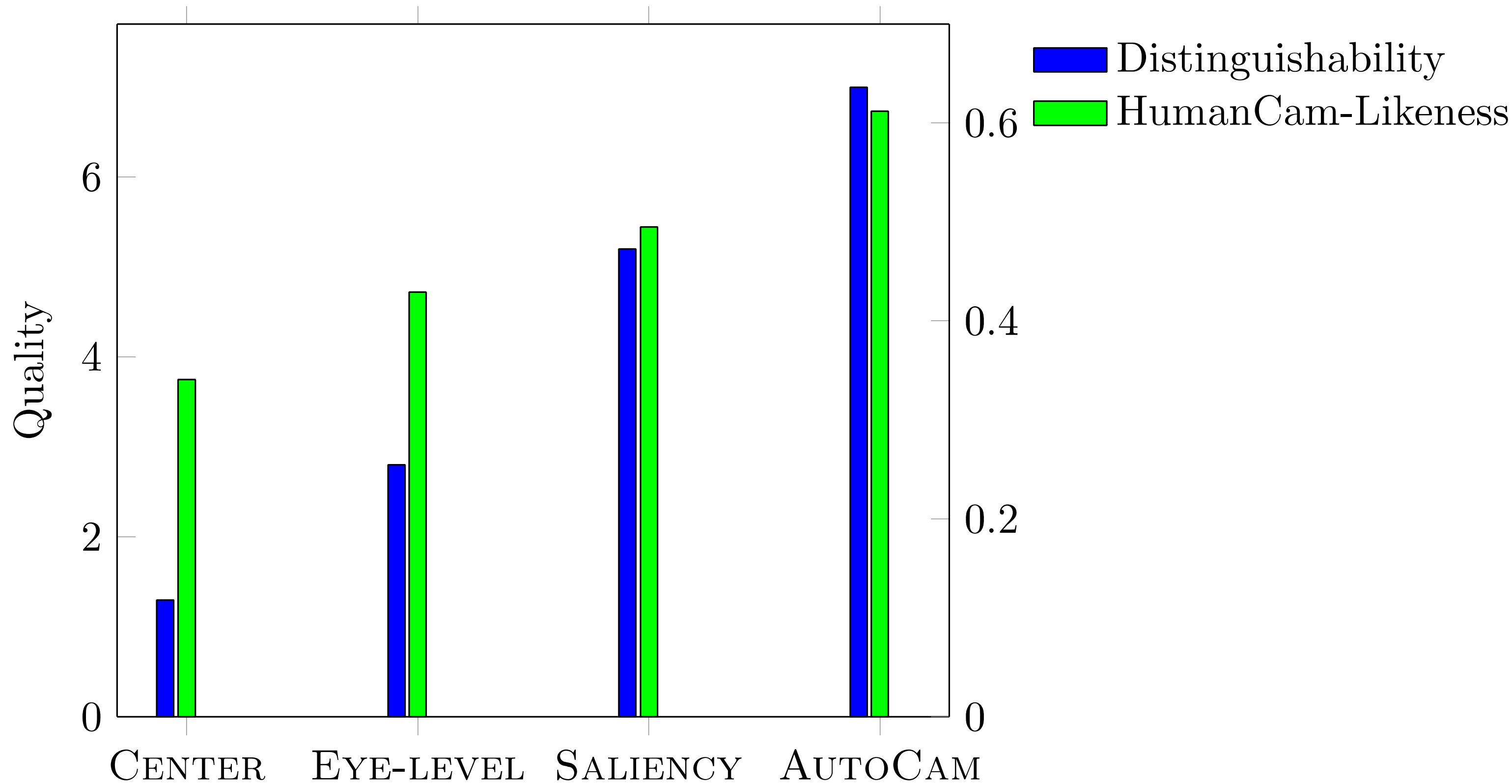
Saliency

Replace capture-worthiness with saliency score



[Harel 2006]

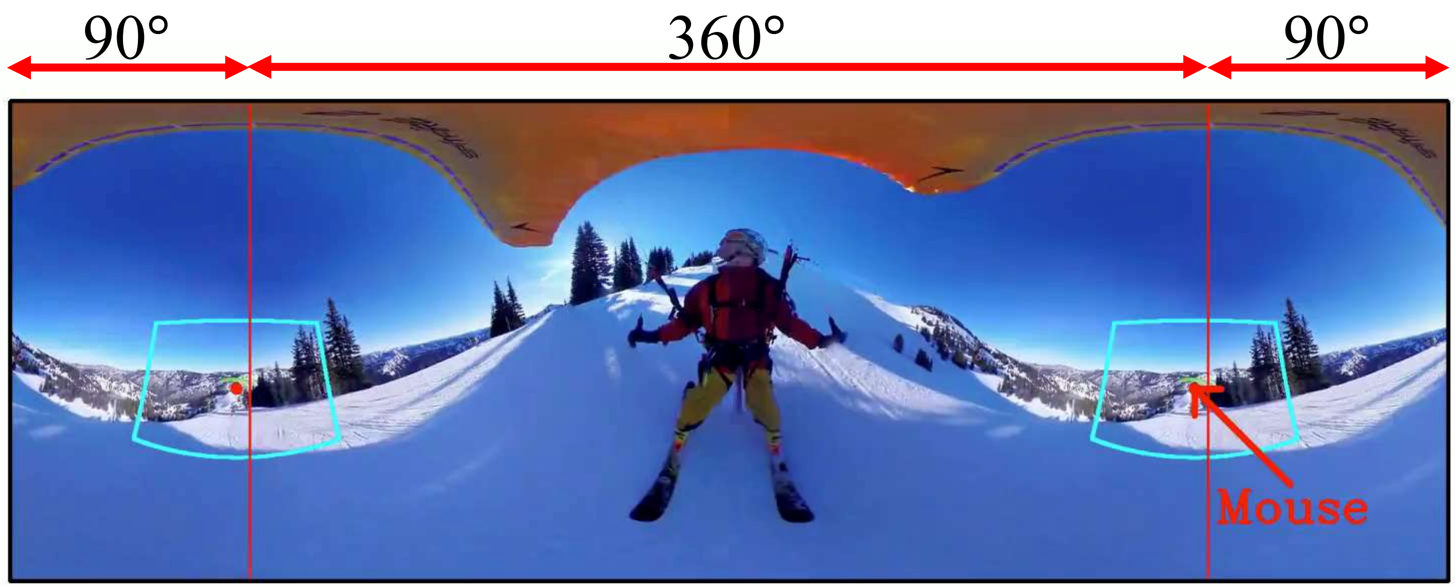
AutoCam results vs. Web videos



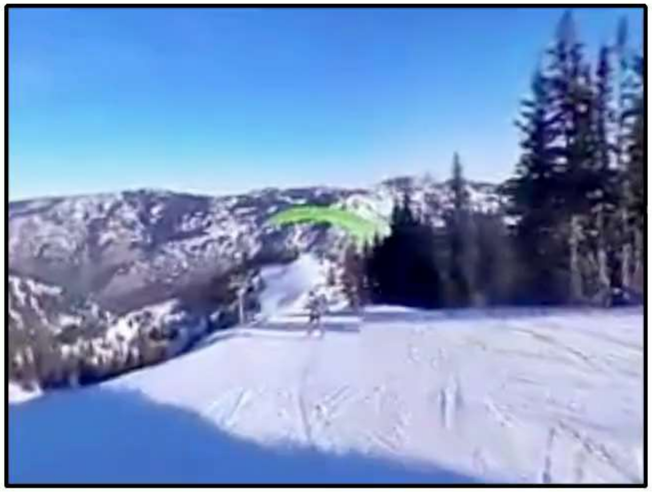
Quantify quality by how indistinguishable algorithm outputs are from human-taken video

What would a human editor select?

HumanEdit
Screenshot

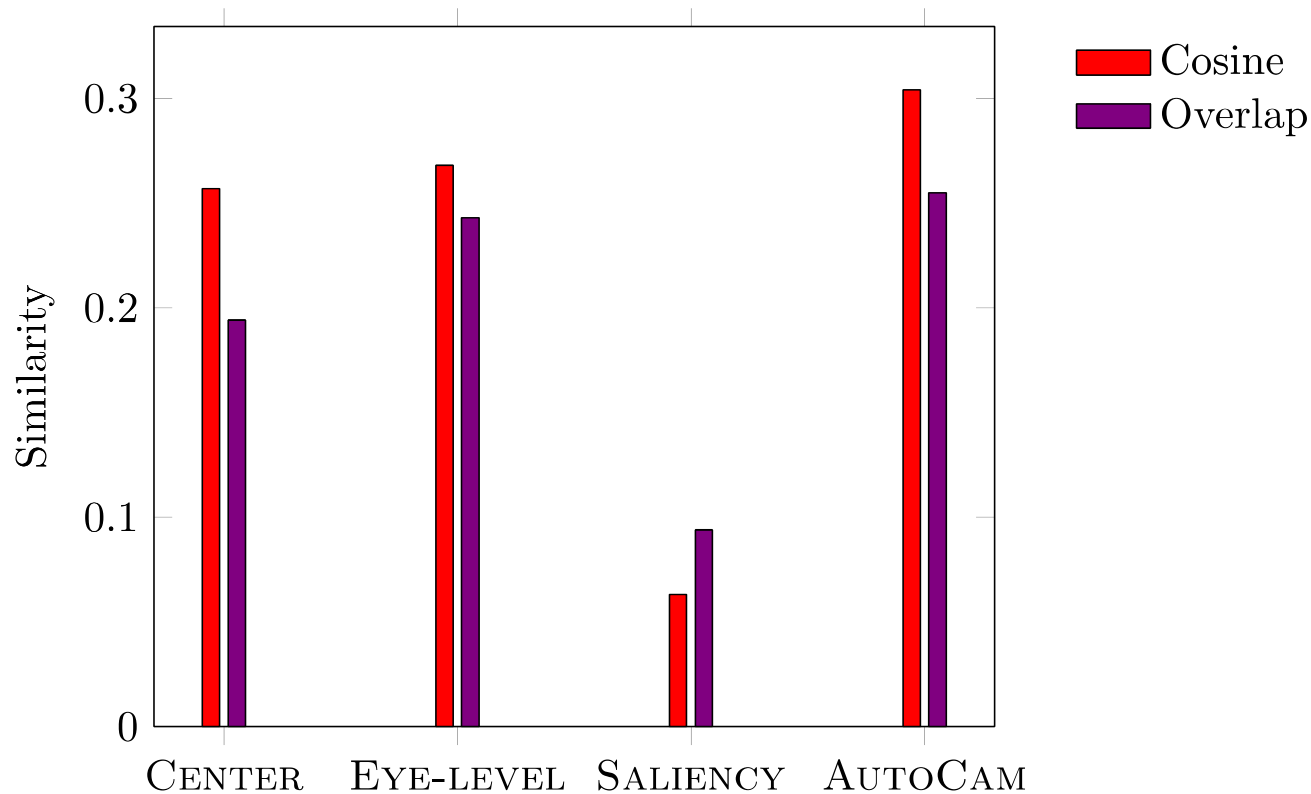


HumanEdit
NFOV Video



- 8 editors
- 120 HumanEdit trajectories
- 3.3 hours total length

AutoCam results vs. Human editors



AutoCam best matches the human-controlled camera

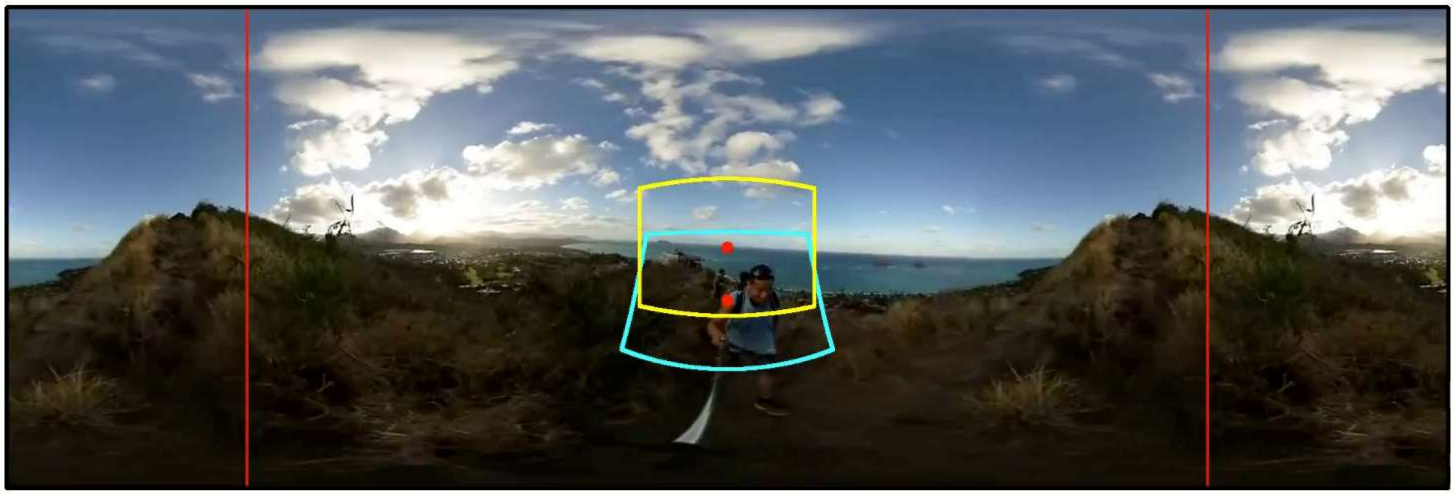
Kristen Grauman, UT Austin

Example AutoCam Output 1

Input 360° Video + Camera Trajectory

AutoCam
Output Video

Example AutoCam Output 2

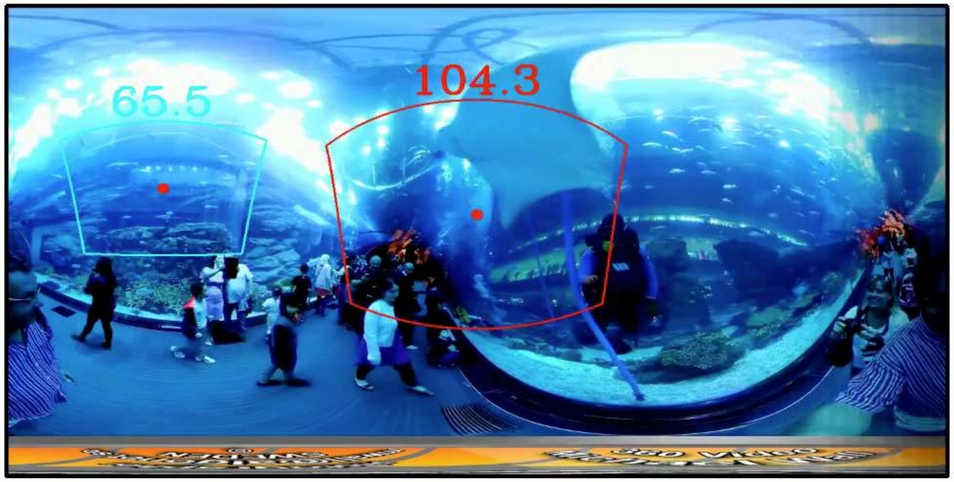


AutoCam

Eye-level Prior

Example AutoCam Output 3

Input 360° Video
+
Camera Trajectories



With Zooming

Without Zooming

Next steps

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

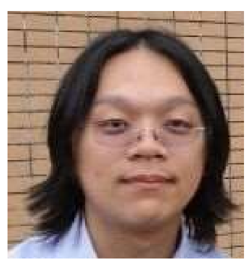
Kristen Grauman
grauman@cs.utexas.edu

Summary

THE UNIVERSITY OF TEXAS AT AUSTIN

- Summarization algorithms are urgently needed to cope with deluge of video data
- New ideas
  - Story-like summaries
  - Detecting *when* engagement occurs
  - Predicting “*where* to look” in 360 video

Yong Jae Lee



Yu-Chuan Su

Bo Xiong

Lu Zheng