A First Person Perspective on Computational Vision

Kristen Grauman

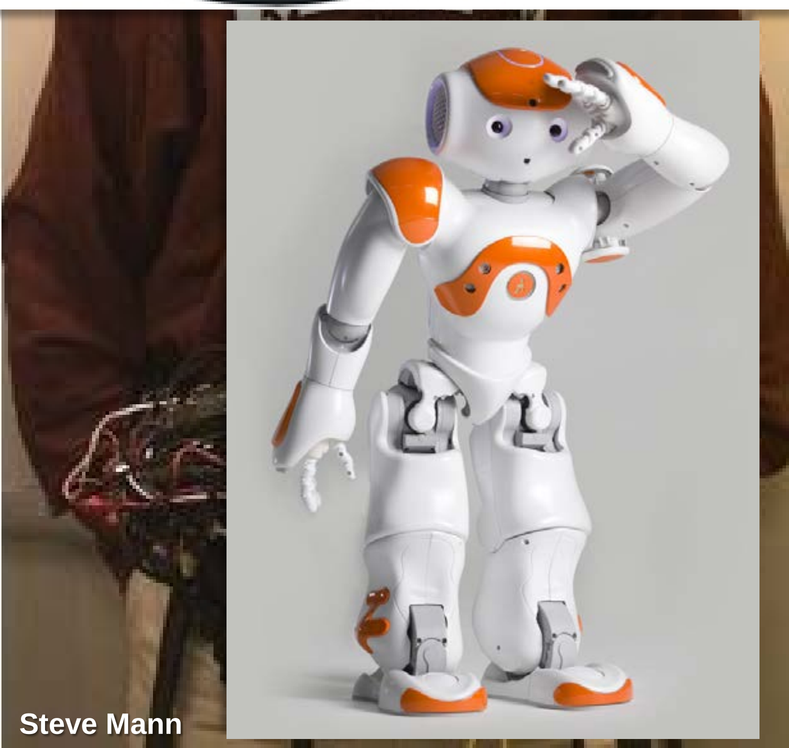
Department of Computer Science

University of Texas at Austin

~1990

Steve Mann

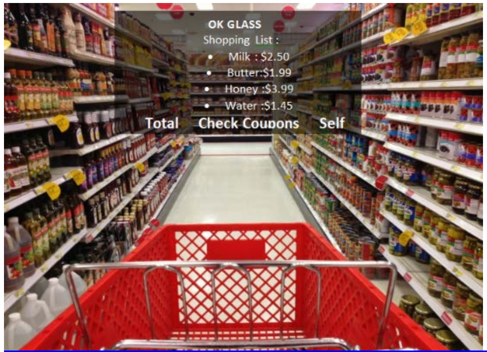
2016

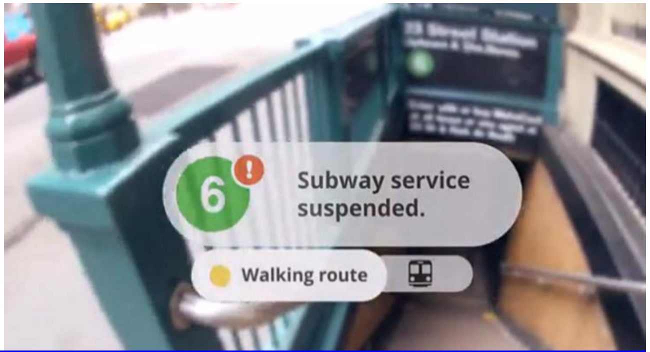


Steve Mann

2016

New era for first-person vision





Augmented reality

Health monitoring

Law enforcement


Science

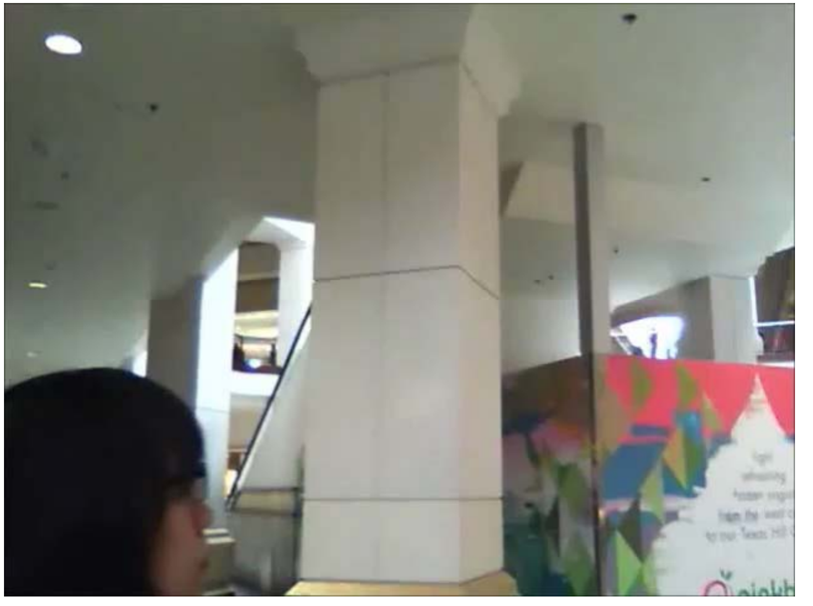
Robotics



Life logging

First person vs. Third person

Traditional third-person view



First-person view

UT TEA dataset

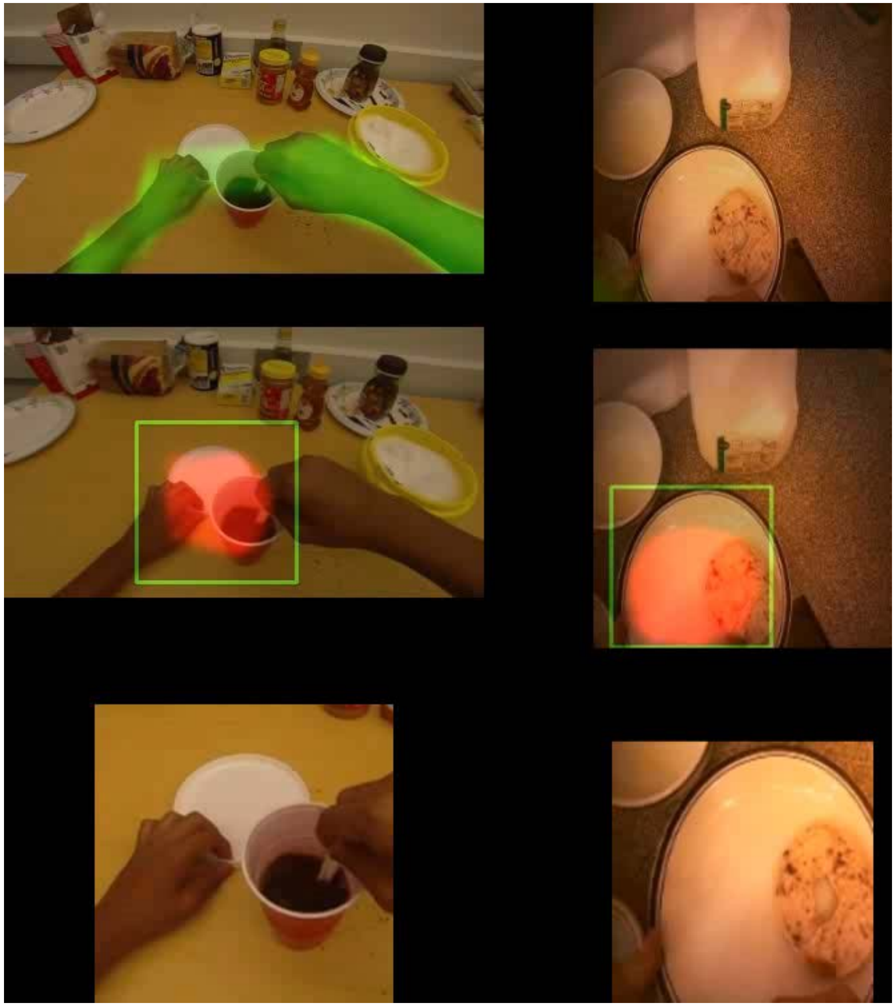
First person vs. Third person

First person “egocentric” vision:

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

RESULTS FROM THE FIELD

What am I doing?

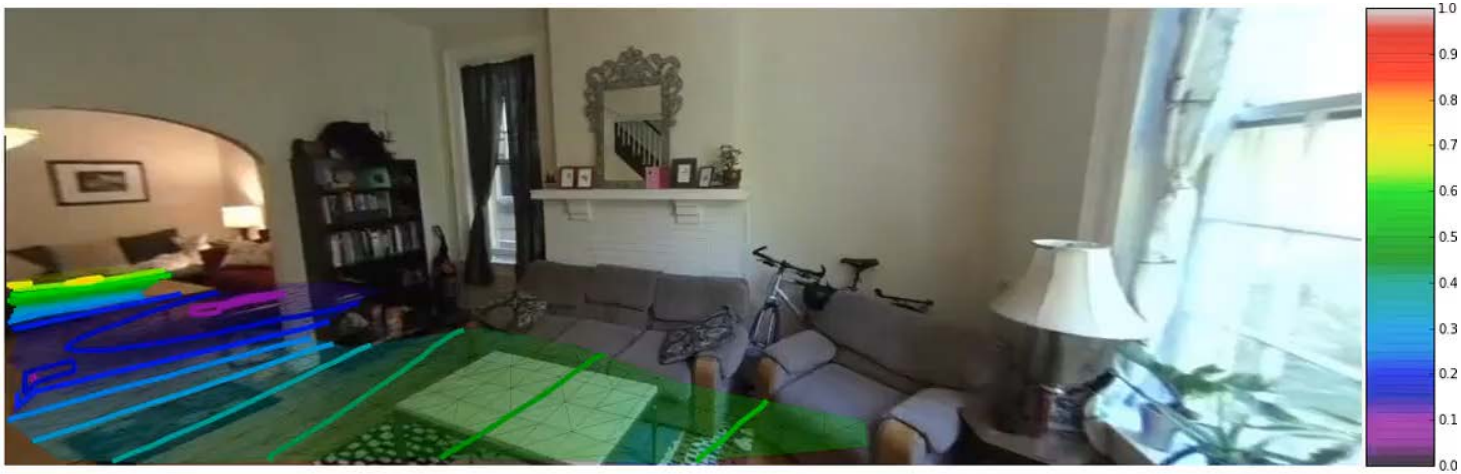


System learns where to pay attention to recognize activity.

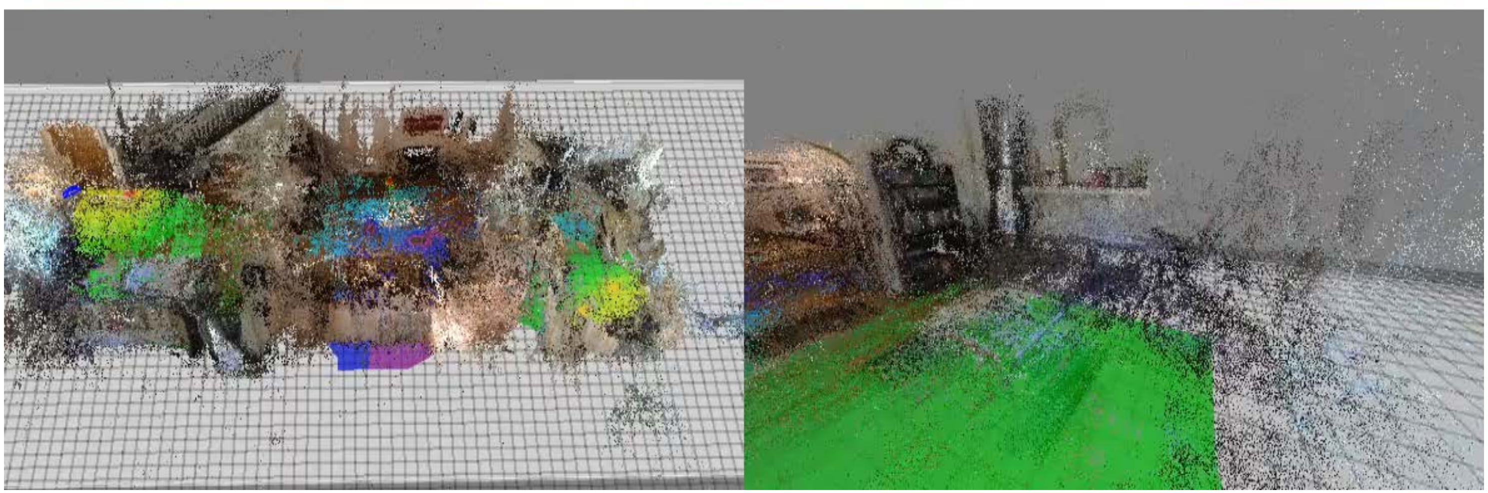
Going Deeper into First-Person Activity Recognition
M. Ma, H. Fan, K. Kitani. CVPR 2016

What could I do here?

Predict functionality/affordances for regions in environment



P(open door | scene)

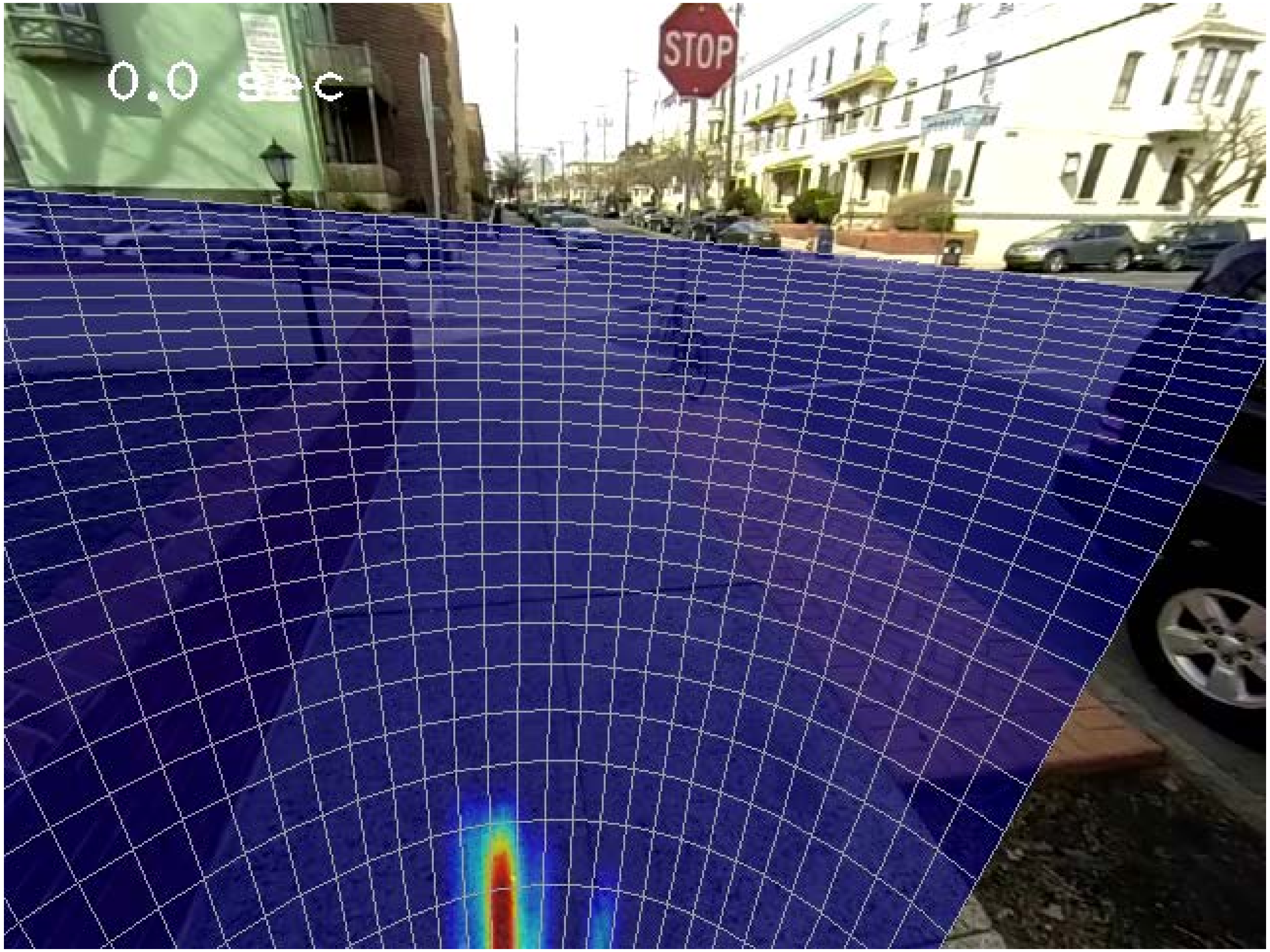


Learning Action Maps of Large Environments via First-Person Vision.
N. Rhinehart, K. Kitani. CVPR 2016

Carnegie Mellon University

Where will I go?

Predict future walking trajectory given video

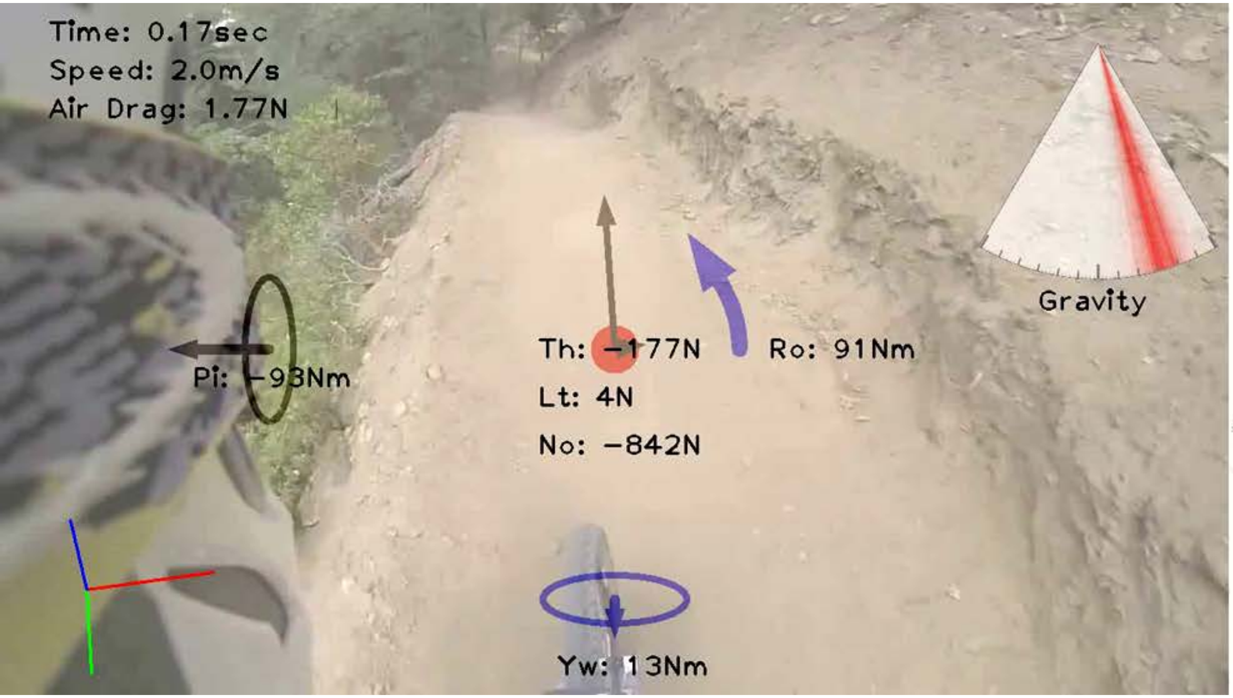


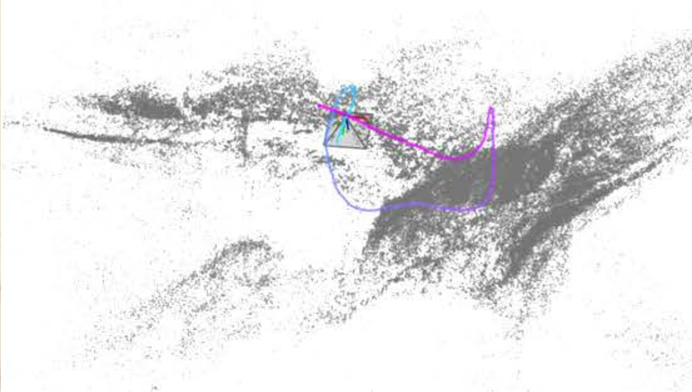
Egocentric Future Localization.
H. S. Park, J-J. Hwang, Y. Niu, and J. Shi. CVPR 2016

University of Pennsylvania

What am I experiencing?

First person video reveals physical interactions with surroundings





3D reconstruction

Force from Motion: Decoding Physical Sensation from a First Person Video
H. S. Park, J-J. Hwang, J. Shi, CVPR 2016

University of Pennsylvania

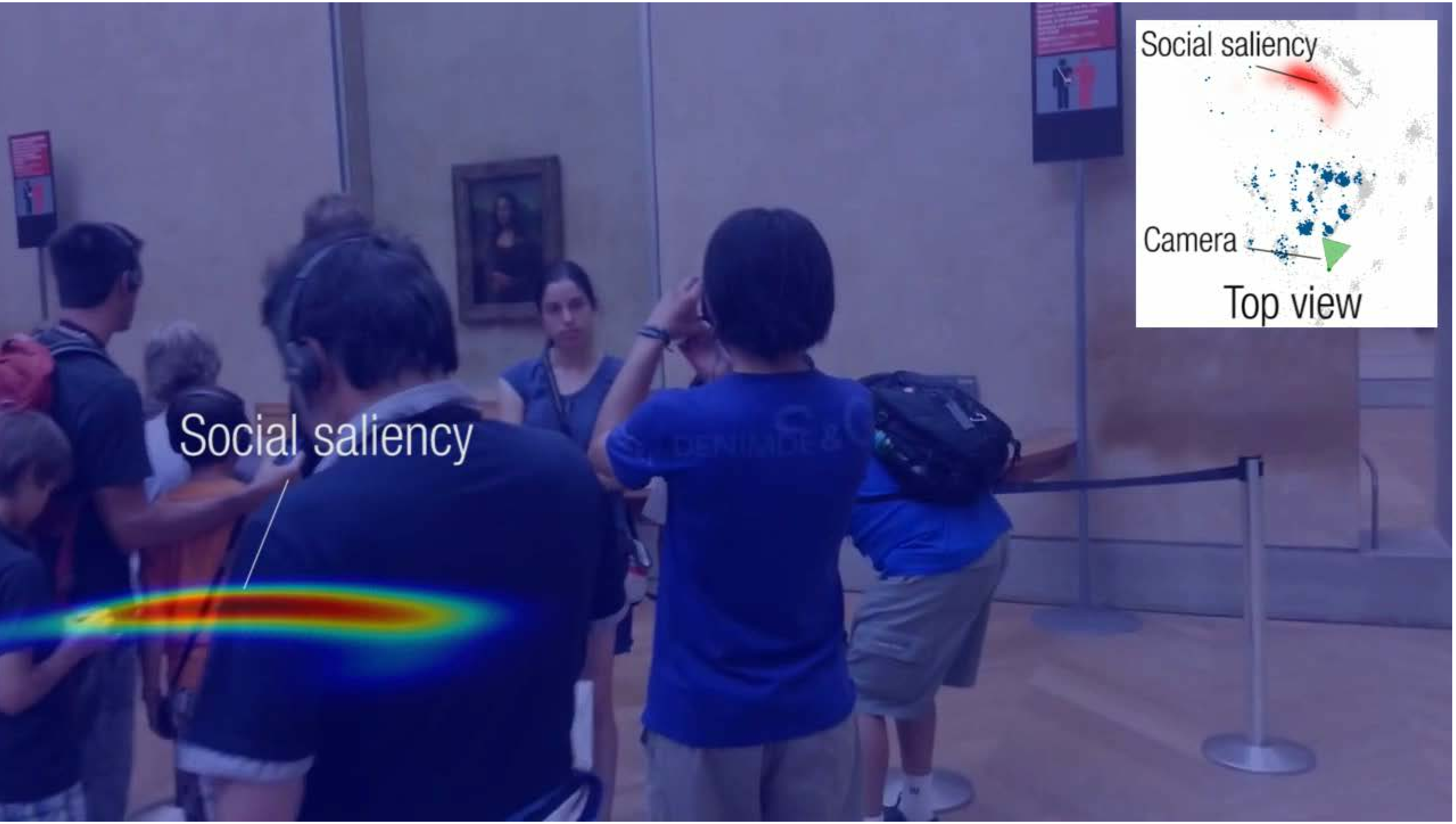
Where do I look?

Computational behavior: quantify moments of eye contact

Detecting Bids for Eye Contact Using a Wearable Camera.
Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. Rehg, F&G 2015

Georgia Tech

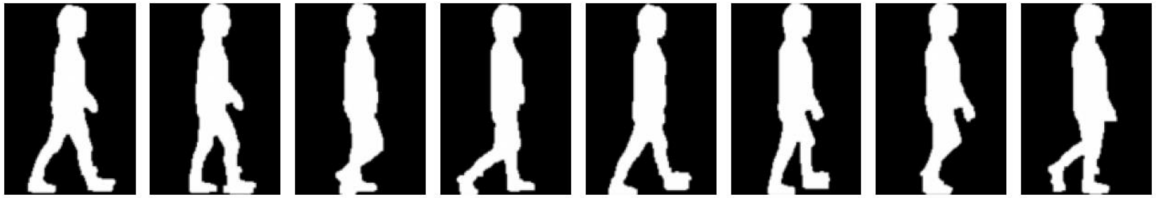
Where do we look?



Social Saliency Prediction. H. S. Park and J. Shi. CVPR 2015

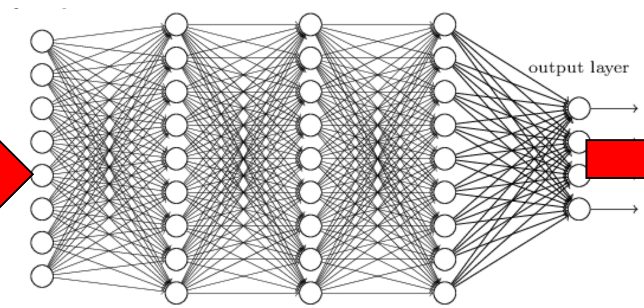
University of Pennsylvania

Who am I?



3rd person:
Gait

First person video: camera motion reveals camera wearer's identity



Predict
Person ID

An Egocentric Look at Video Photographer Identity, Y. Hoshen and S. Peleg, CVPR 2016

Hebrew University of Jerusalem

What am I doing?
What could I do here?
Where will I go?
What am I experiencing?
Where do I look?
Where do we look?
Who am I?

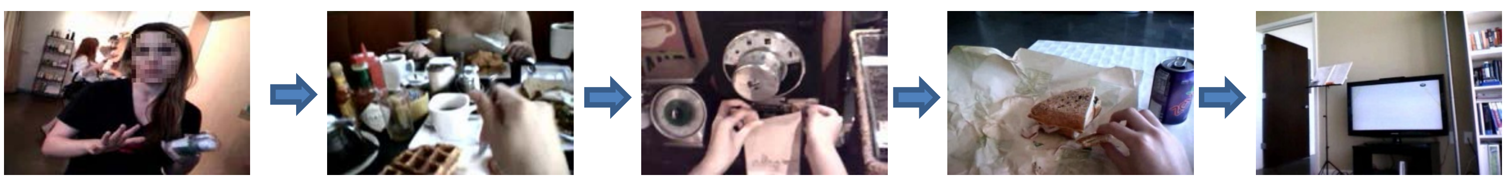
RESULTS FROM MY GROUP

WHAT HAVE I SEEN?

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

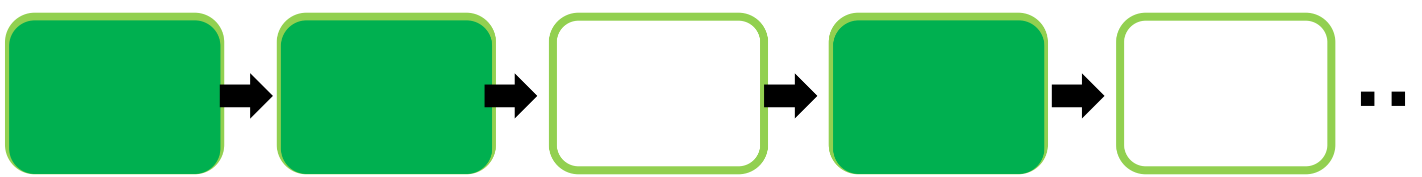
What have I seen?

Story-based summaries of first-person videos

Original video (3 hours)

Our summary (12 frames)

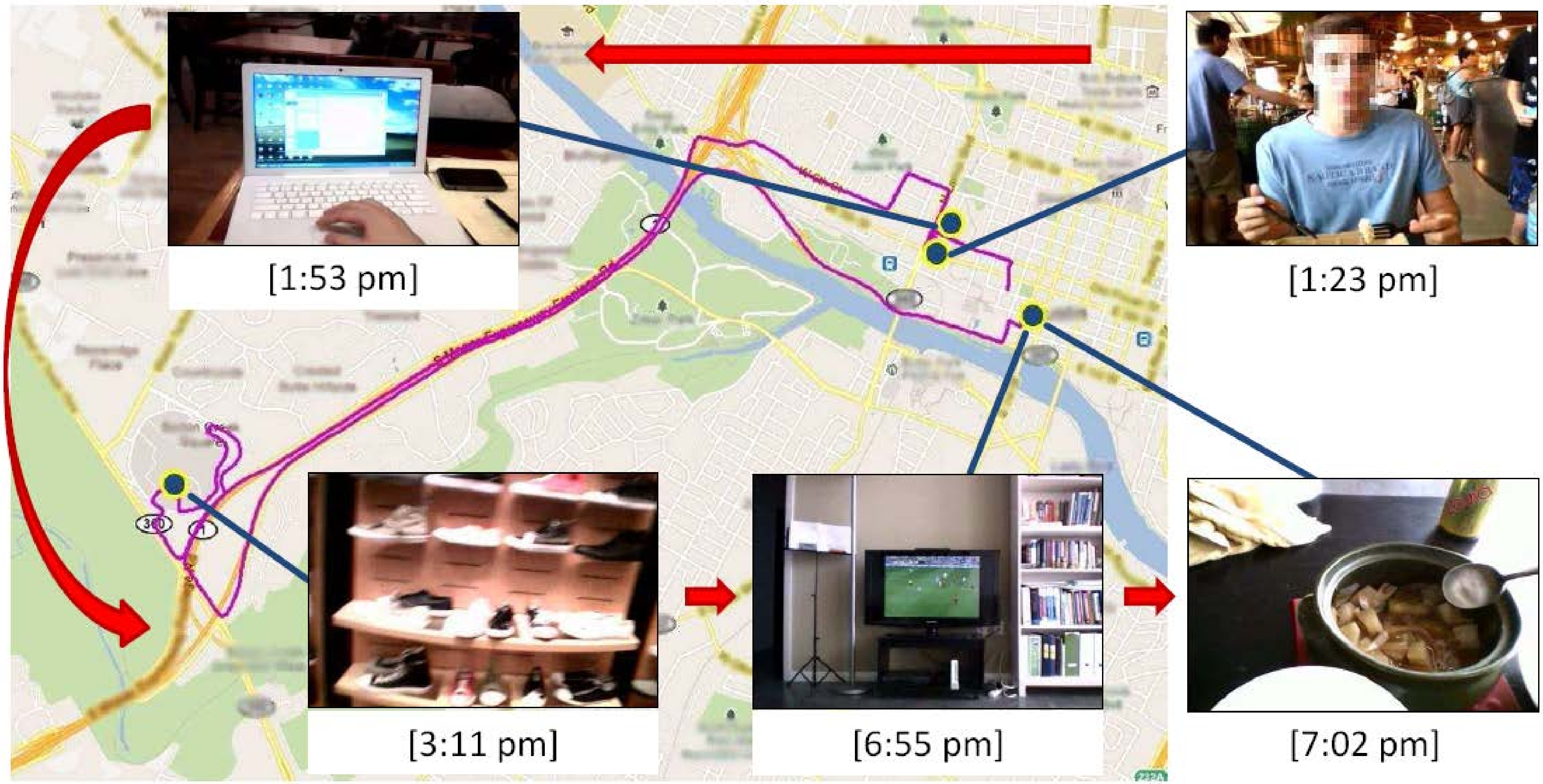
Subshots



$$S^* = \arg\max_{S \subset \mathcal{V}} \; \underbrace{\lambda_s \, \mathcal{S}(S)}_{\text{influence}} + \underbrace{\lambda_i \, \mathcal{I}(S)}_{\text{importance}} + \underbrace{\lambda_d \, \mathcal{D}(S)}_{\text{diversity}}$$

What have I seen?

Auto-generating storyboard maps



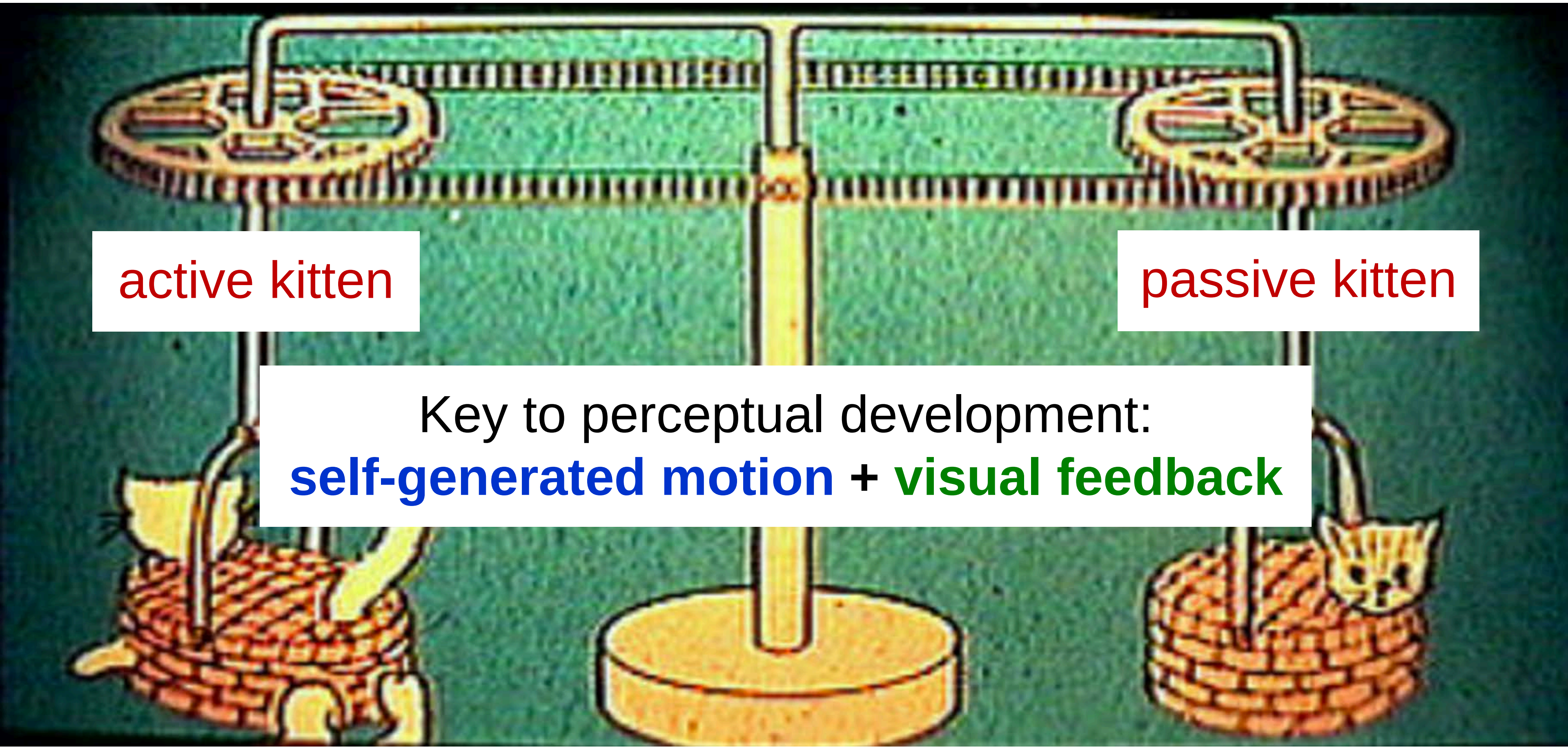
Predicting Important Objects for Egocentric Video Summarization.
Y J. Lee and K. Grauman. IJCV 2015

What am I doing?
What could I do here?
Where will I go?
What am I experiencing?
Where do I look?
Where do we look?
Who am I?
What have I seen?

RESULTS FROM MY GROUP

WHAT WILL I SEE, IF I MOVE?

The kitten carousel experiment
[Held & Hein, 1963]



Key to perceptual development:
self-generated motion + visual feedback

Big picture goal: Embodied vision

Status quo:

Learn from “disembodied” bag of labeled snapshots.

Our goal:

Learn in the context of acting and moving in the world.

Our idea: Ego-motion ↔ vision

Goal: Teach computer vision system the connection:
"how I move" ↔ "how my visual surroundings change"

+

Ego-motion motor signals

Unlabeled video

Learning Image Representations Tied to Ego-Motion.
D. Jayaraman and K. Grauman. ICCV 2015

Ego-motion ↔ vision: view prediction

After moving:

Kristen Grauman, UT Austin

Ego-motion ↔ vision for recognition

Learning this connection requires:

- Depth, 3D geometry
- Semantics
- Context

Also key to recognition!

And can be learned *without* manual annotations!

Our approach: unsupervised feature learning using egocentric video + motor signals

[Jayaraman & Grauman, ICCV 2015]

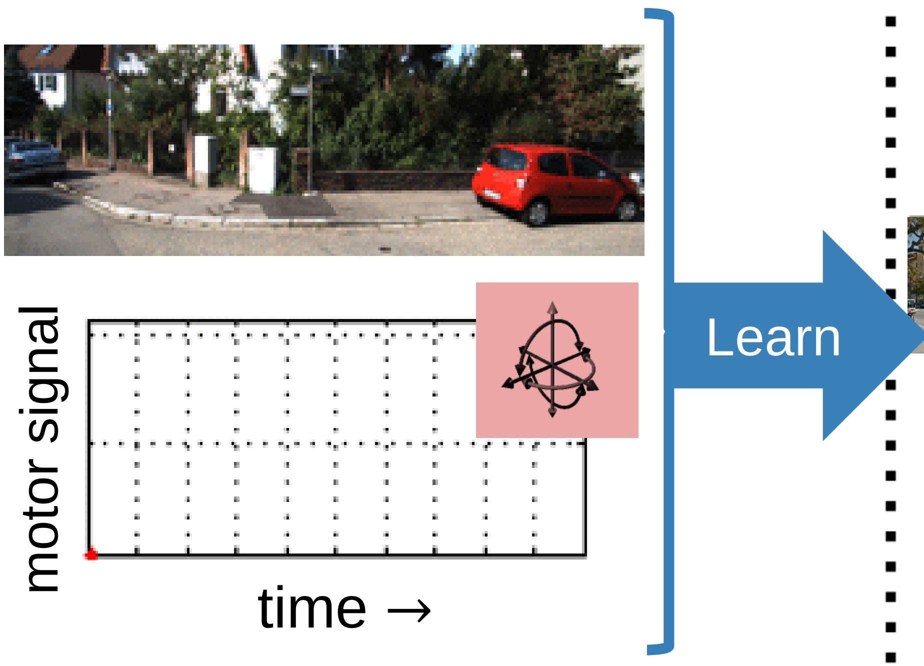
Approach idea: Ego-motion equivariance

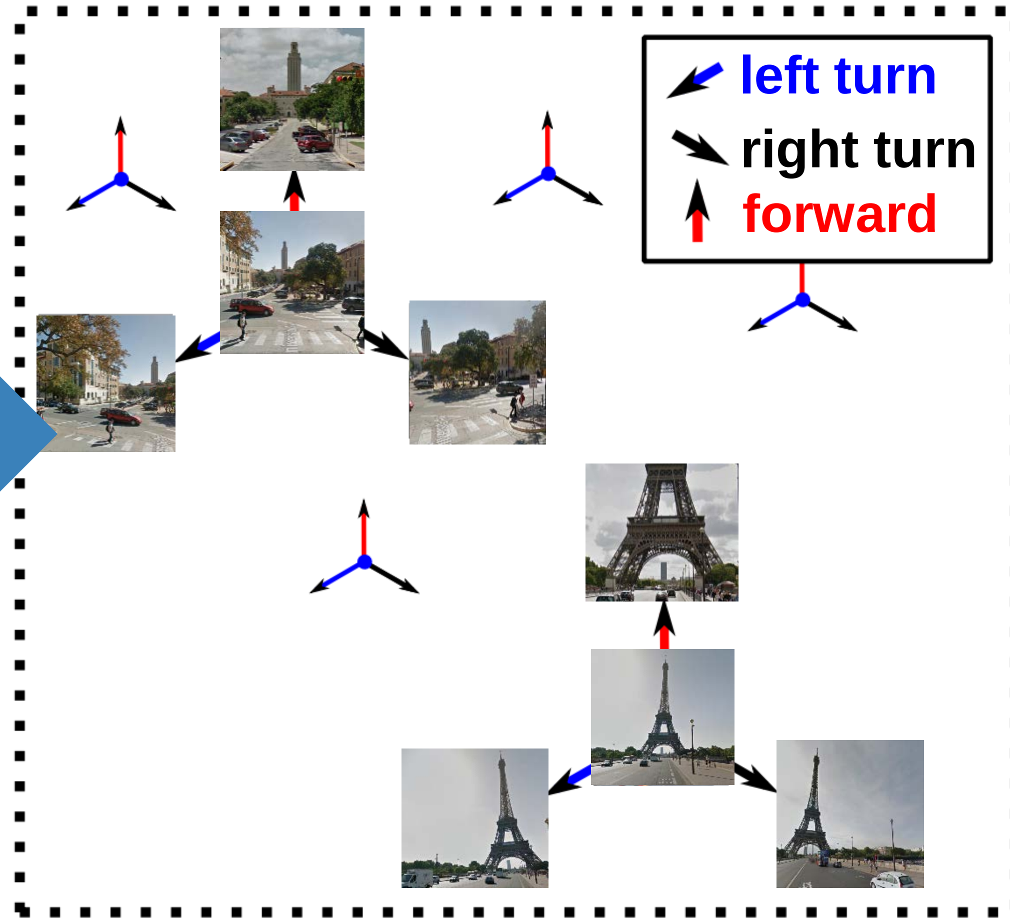
Training data

Unlabeled video +
motor signals

Equivariant embedding

organized by ego-motions





Result: Recognition

Learn from **unlabeled car video** (KITTI)

Geiger et al, IJRR ’13

Exploit features for **image scene classification**
(SUN, 397 classes)

Xiao et al, CVPR ’10

Result: Recognition

Learn from **unlabeled car video** (KITTI)

Geiger et al, IJRR ’13

Exploit features for **image scene classification**

Apse

Window seat

Guardhouse

Double the accuracy

vs.

Learning from labeled images alone

Xiao et al, CVPR ’10

Learning how to move for recognition

Time to revisit active recognition in challenging settings!

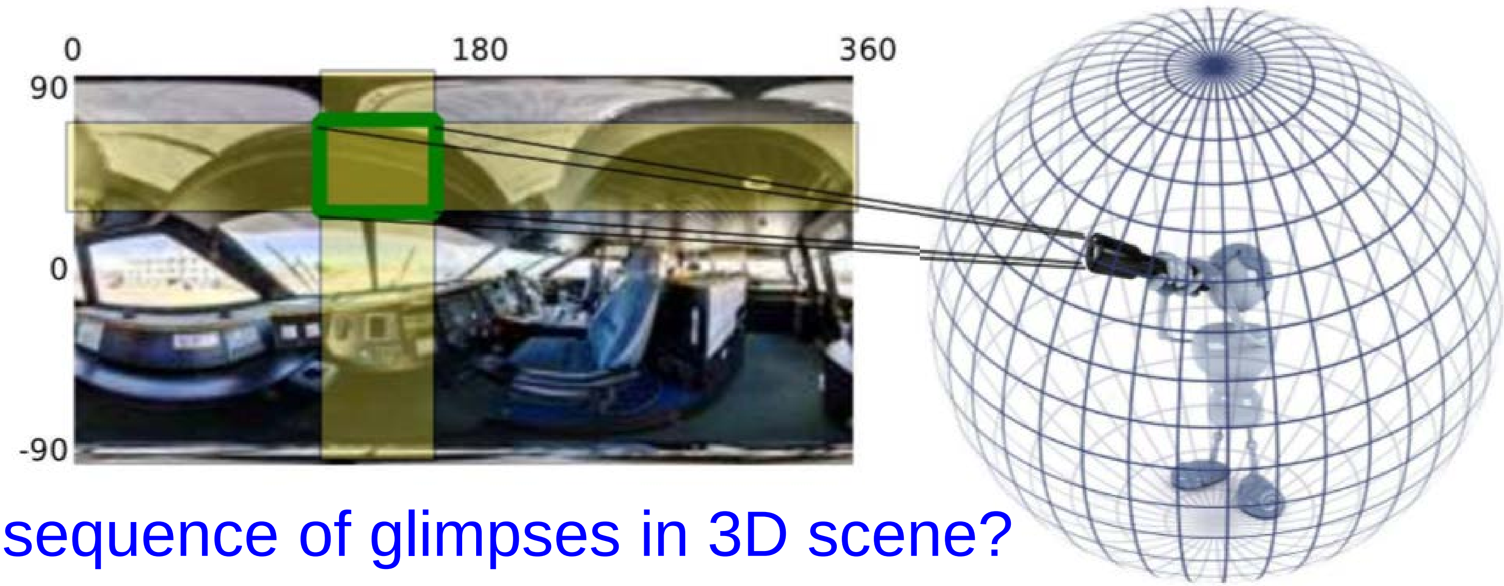
[Bajcsy 1988, Aloimonos et al. 1988, Schiele & Crowley 1998, Dickinson et al. 1997, Wilkes & Tsotsos 1992, Callari & Ferrie 2001,…]

Learning how to move for recognition

Time to revisit active recognition in challenging settings!

[Bajcsy 1988, Aloimonos et al. 1988, Schiele & Crowley 1998, Dickinson et al. 1997, Wilkes & Tsotsos 1992, Callari & Ferrie 2001,…]

Learning how to move for recognition



Best sequence of glimpses in 3D scene?

Requires:
- Action selection
- Per-view processing
- Evidence aggregation
- Look-ahead prediction

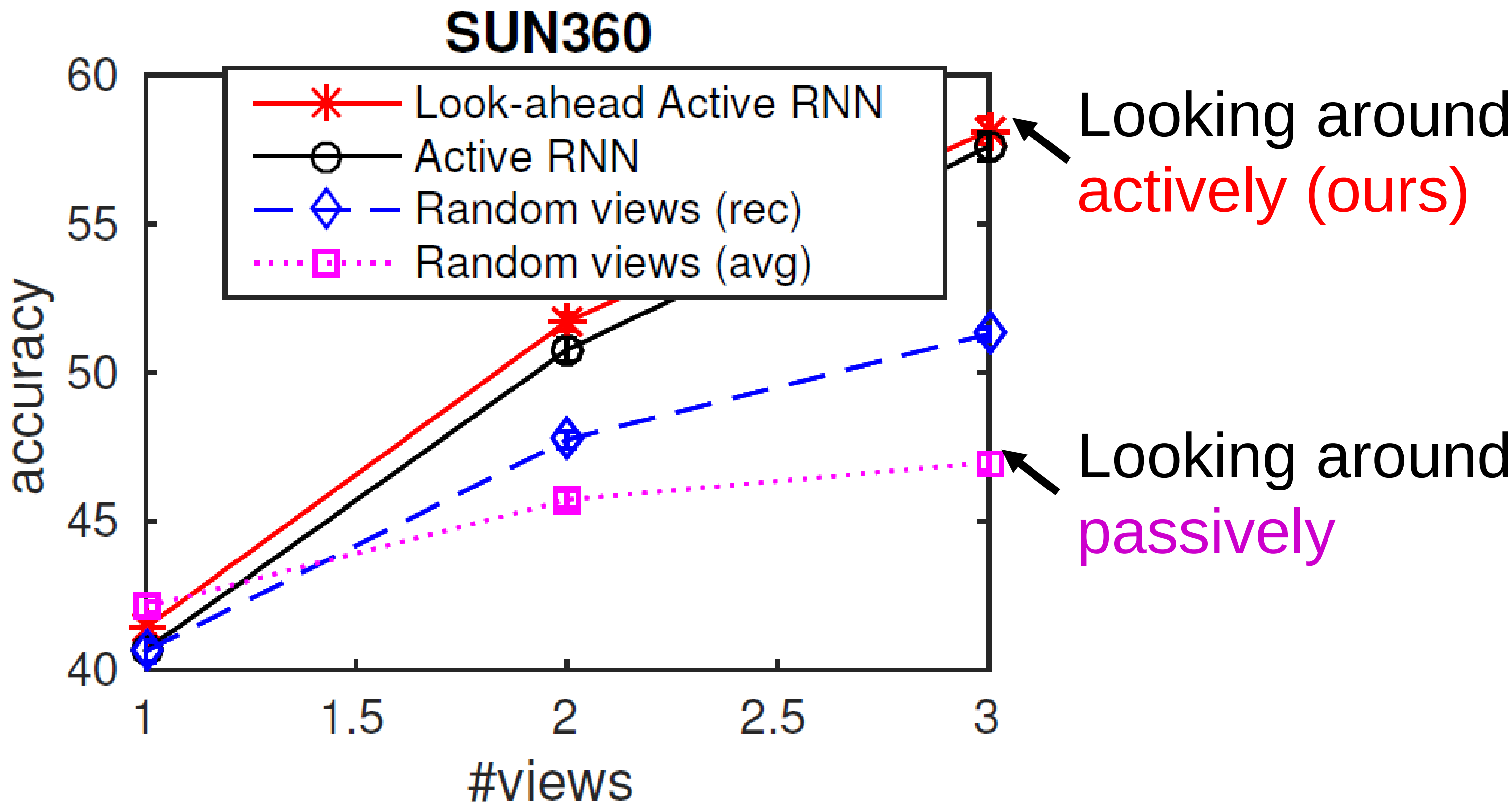
Learn all end-to-end

Look-Ahead Before You Leap: End-to-End Active Recognition by Forecasting the Effect of Motion. D. Jayaraman and K. Grauman. ECCV 2016

Active recognition: results

P("Plaza courtyard"):	(6.28)	(11.95)	(68.38)
Top 3 guesses:	Restaurant Train interior Shop	Theater Restaurant Plaza courtyard	Plaza courtyard Street Theater

Active recognition: results



Active selection + look-ahead → better scene categorization from sequence of glimpses in 360 panorama

Next steps

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

Kristen Grauman

Computer Vision Group

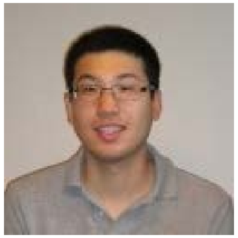
grauman@cs.utexas.edu

http://www.cs.utexas.edu/~grauman/

THE UNIVERSITY OF TEXAS AT AUSTIN

Summary

- Visual learning benefits from
  - context of action and motion in the world
  - continuous self-acquired feedback
- New ideas:
  - Story-like summaries for “always on” cameras
  - Embodied visual learning and recognition



Yong Jae Lee

Dinesh Jayaraman