

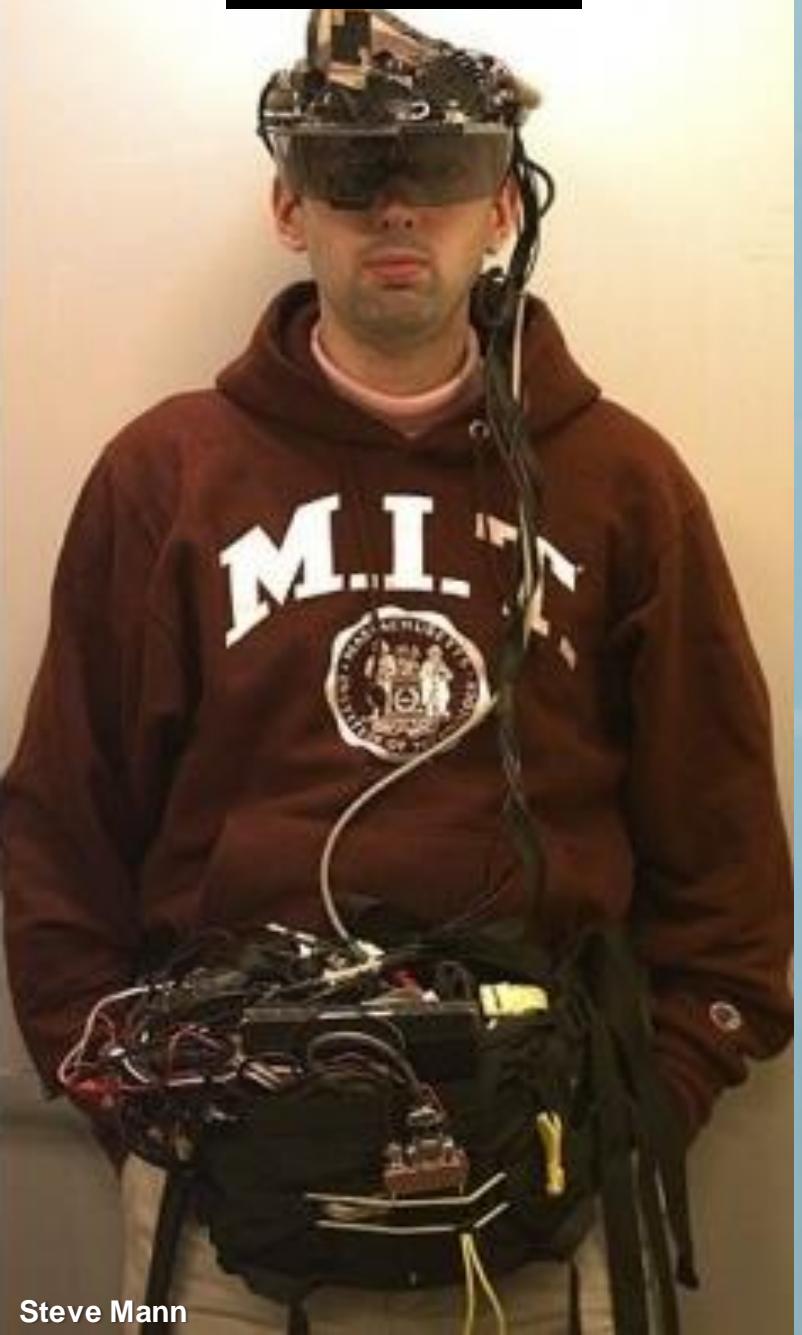
Action and Attention in First-Person Vision

Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Dinesh Jayaraman, Yong Jae Lee,
Yu-Chuan Su, Bo Xiong, Lu Zheng

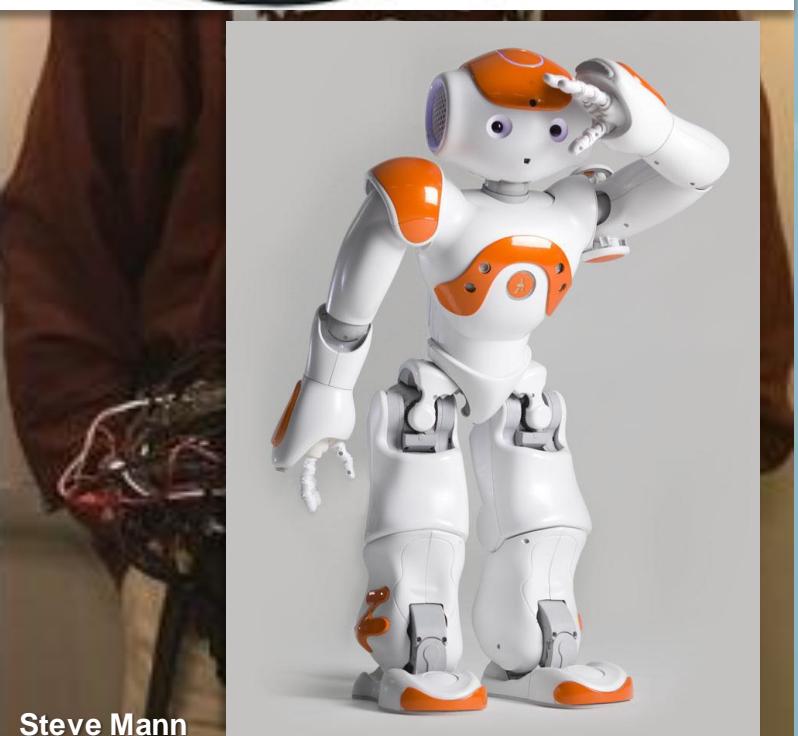


~1990



2015



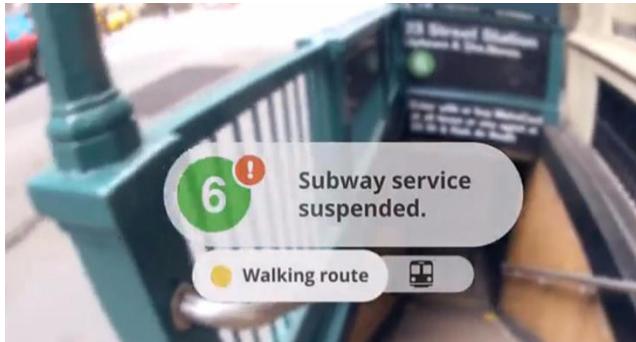


2015



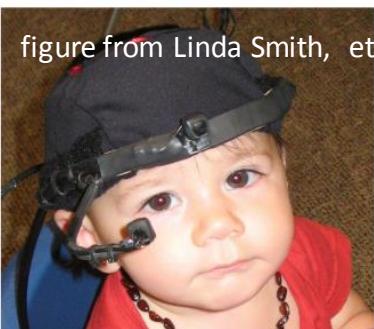
Steve Mann

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

First person vs. Third person



Traditional third-person view



First-person view

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

Recent egocentric work

- Activity and object recognition

[Spriggs et al. 2009, Ren & Gu 2010, Fathi et al. 2011, Kitani et al. 2011, Pirsiavash & Ramanan 2012, McCandless & Grauman 2013, Ryoo & Matthies 2013, Poleg et al. 2014, Damen et al. 2014, Behera et al. 2014, Li et al. 2015, Yonetani et al. 2015, ...]

- Gaze and social cues

[Yamada et al. 2011, Fathi et al. 2012, Park et al. 2012, Li et al. 2013, Arev et al. 2014, Leelasawassuk et al. 2015,...]

- Visualization, stabilization

[Kopf et al. 2014, Poleg et al. 2015]

Talk overview

Motivation

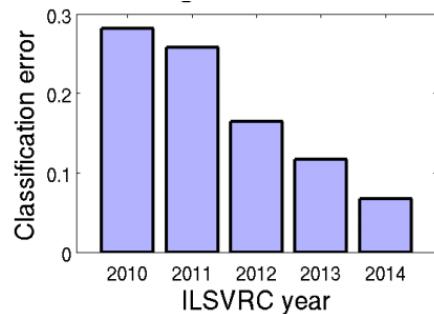
Account for the fact that **camera wearer is active participant** in the visual observations received

Ideas

1. **Action:** Unsupervised feature learning
 - How is visual learning shaped by ego-motion?
2. **Attention:** Inferring highlights in video
 - How to summarize long egocentric video?

Visual recognition

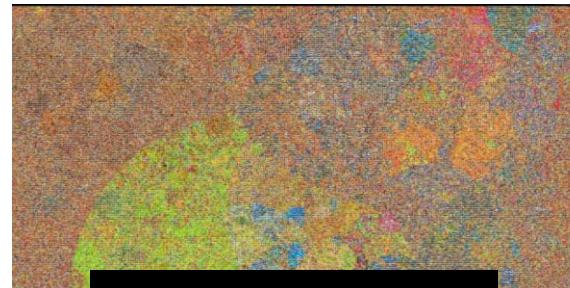
- Recent major strides in category recognition



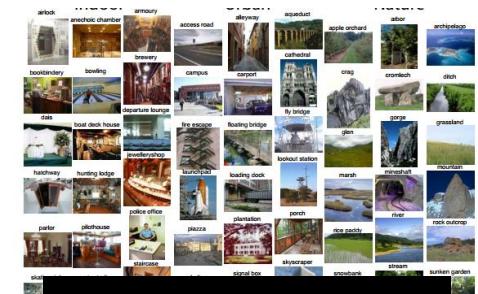
- Facilitated by large labeled datasets



ImageNet
[Deng et al.]



80M Tiny Images
[Torralba et al.]



SUN Database
[Xiao et al.]

[Papageorgiou & Poggio 1998, Viola & Jones 2001, Dalal & Triggs 2005, Grauman & Darrell 2005, Lazebnik et al. 2006, Felzenszwalb et al. 2008, Krizhevsky et al. 2012, Russakovsky IJCV 2015...]

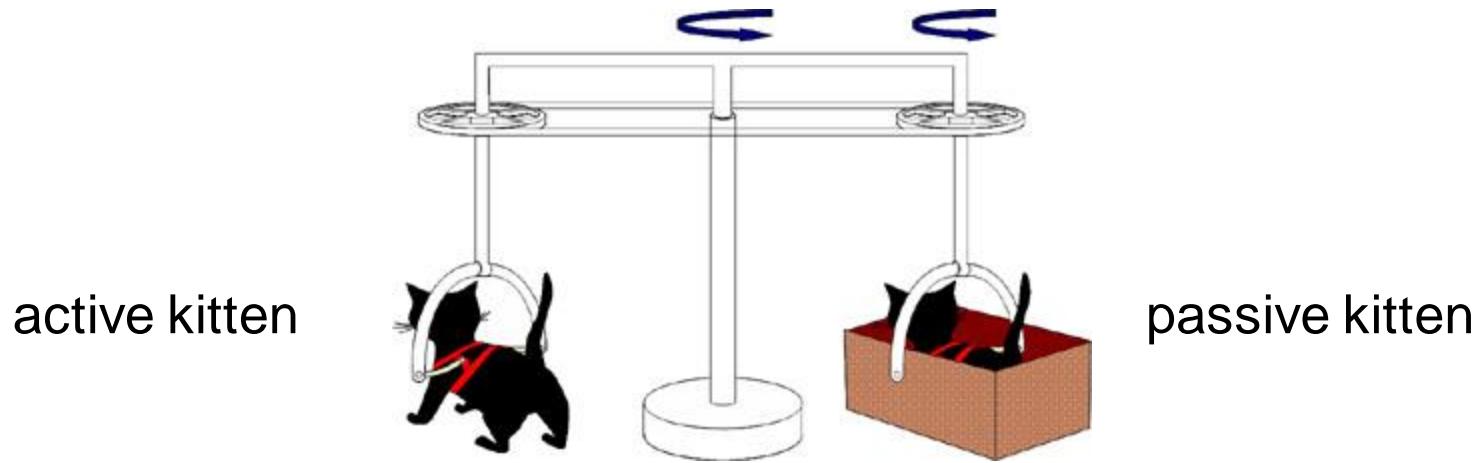
Problem with today's visual learning

- **Status quo:** Learn from “disembodied” bag of labeled snapshots
- ...yet visual perception develops in the context of **acting** and **moving** in the world



The kitten carousel experiment

[Held & Hein, 1963]

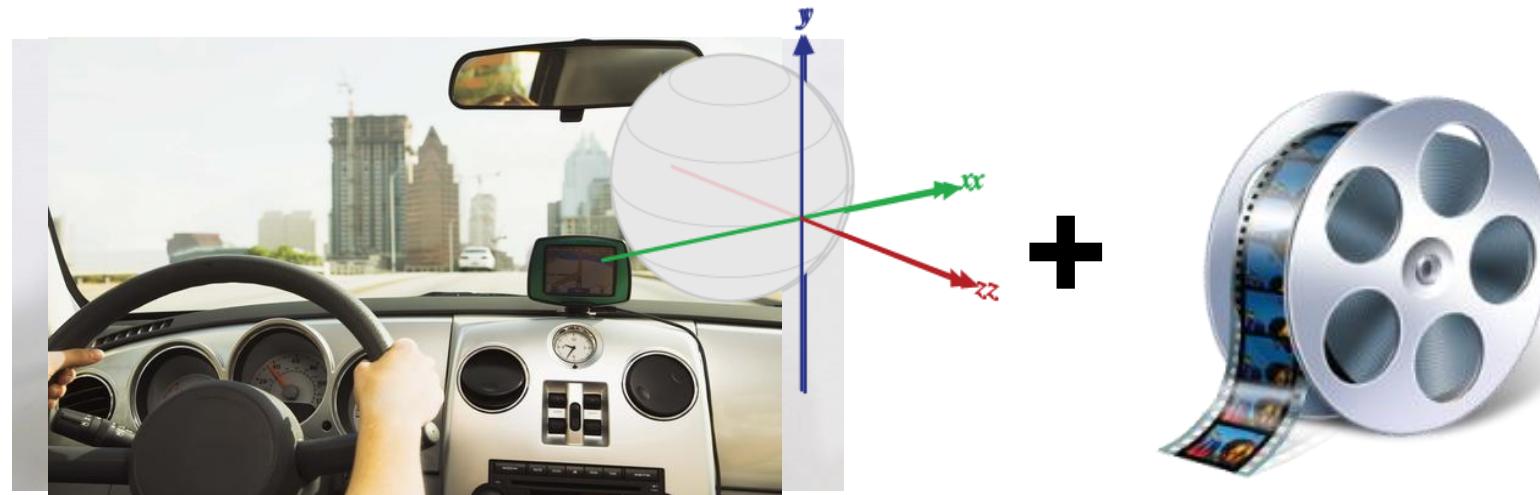


Key to perceptual development:
Self-generated motions + visual feedback

Our idea: Feature learning with ego-motion

Goal: Learn the connection between
“how I move” \leftrightarrow “how visual surroundings change”

Approach: Unsupervised feature learning using
motor signals accompanying egocentric video



Key idea: Egomotion equivariance

Invariant features: unresponsive to some classes of transformations

$$z(g\mathbf{x}) \approx z(\mathbf{x})$$

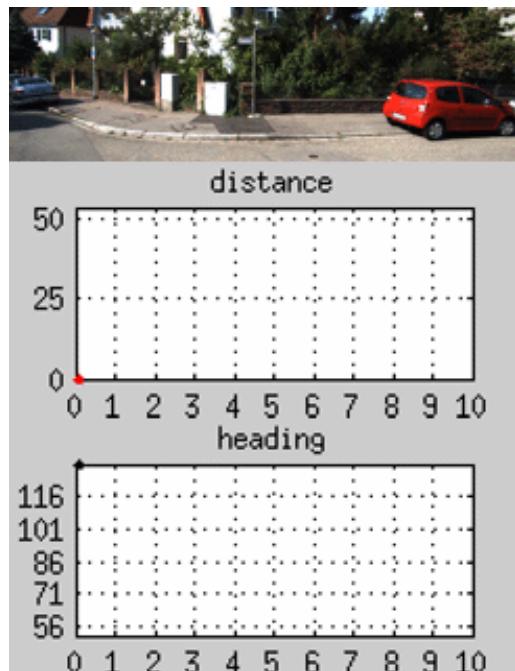
Equivariant features : *predictably* responsive to some classes of transformations, through simple mappings (e.g., linear)

$$z(g\mathbf{x}) \approx M_g z(\mathbf{x})$$

“equivariance map”

Invariance *discards* information,
whereas equivariance *organizes* it.

Key idea: Egomotion equivariance

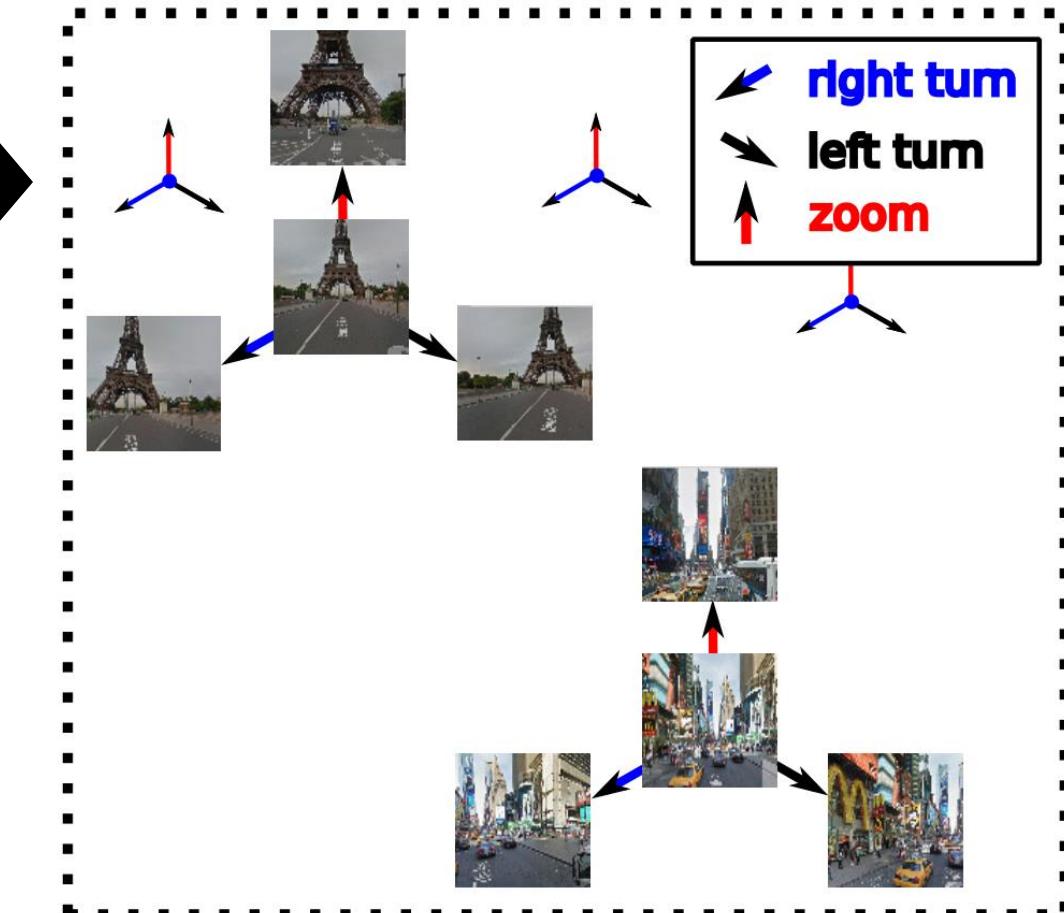


Training data=

Unlabeled video

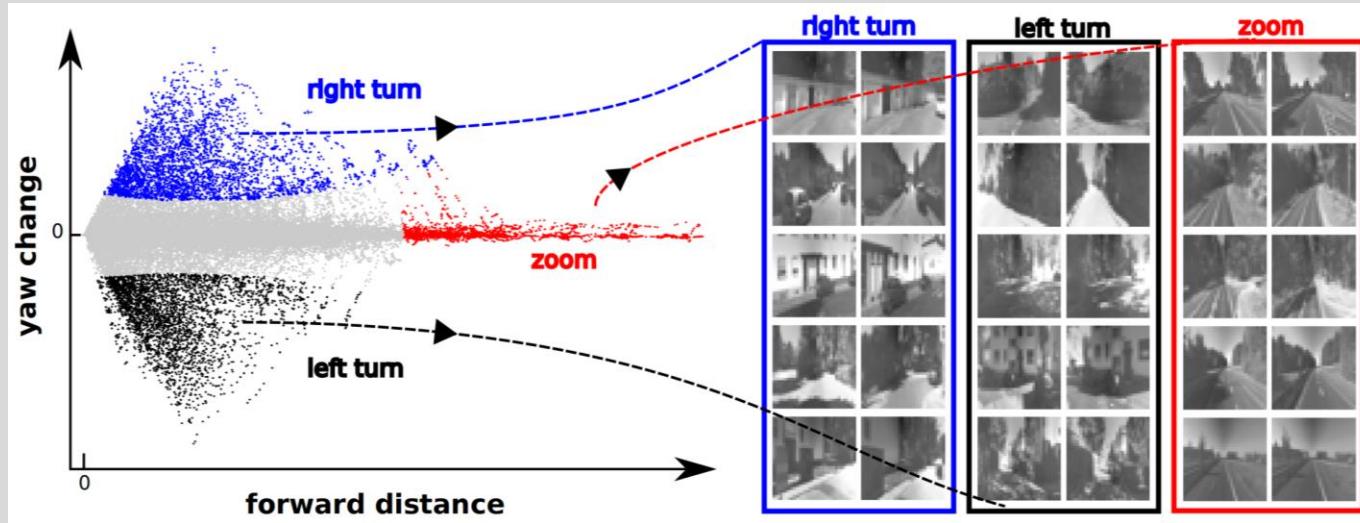
+

motor signals

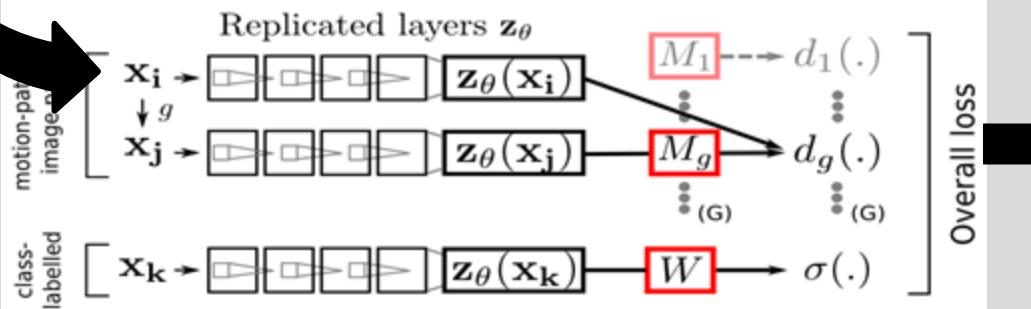


Approach

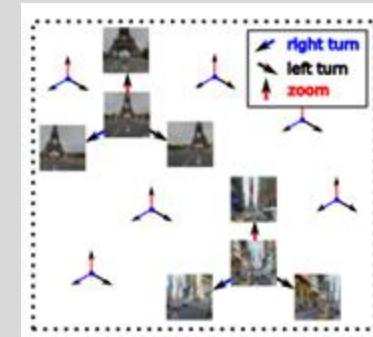
Ego motor signals + Observed image pairs



Deep learning architecture

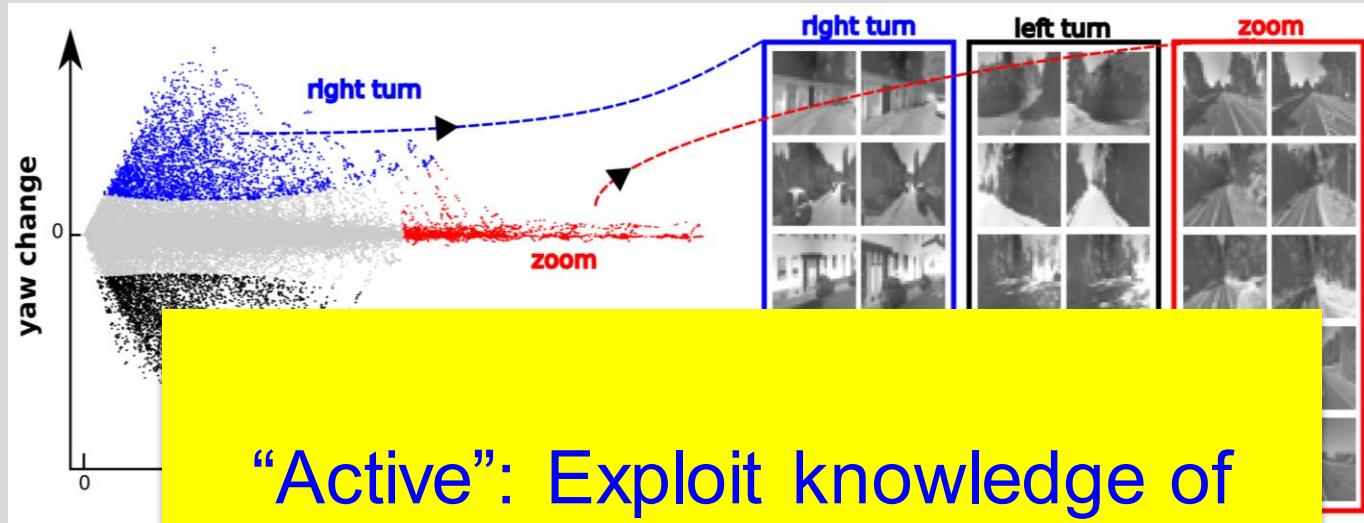


Output embedding

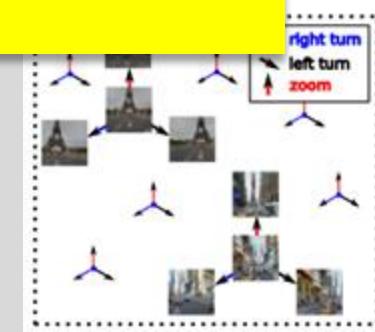
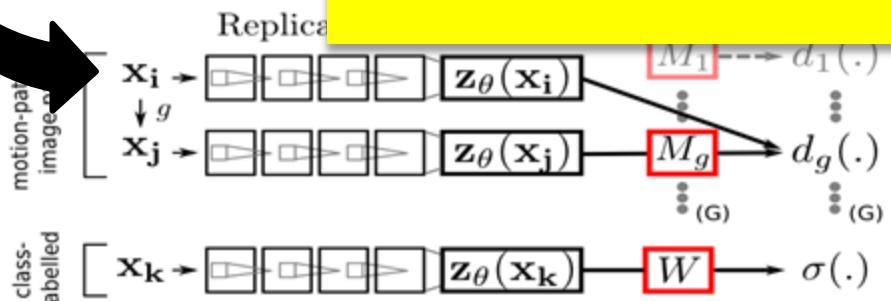


Approach

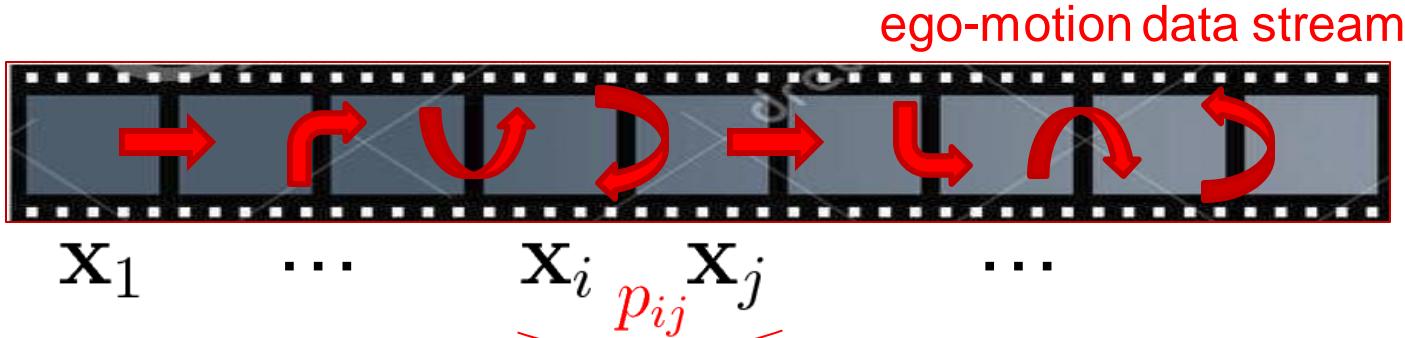
Ego motor signals + Observed image pairs



Deep I

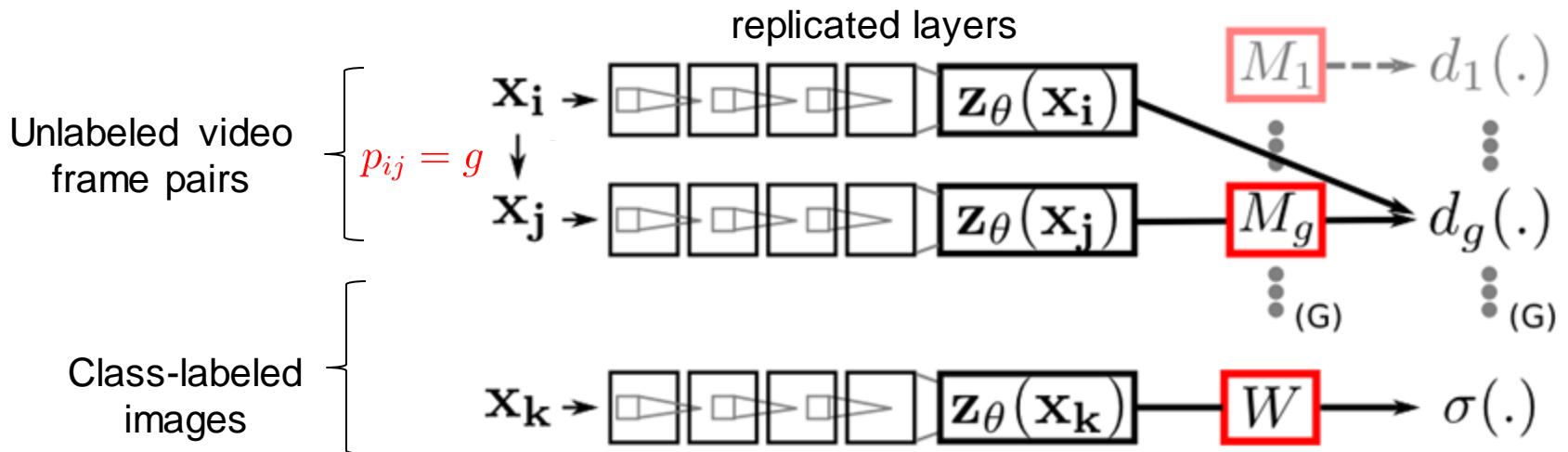


Learning equivariance



Embedding objective:

$$(\boldsymbol{\theta}^*, \mathcal{M}^*) = \arg \min_{\boldsymbol{\theta}, \mathcal{M}} \sum_{g, i, j} d_g (\mathbf{z}_{\boldsymbol{\theta}}(\mathbf{x}_i), M_g \mathbf{z}_{\boldsymbol{\theta}}(\mathbf{x}_j), p_{ij}) + \lambda L_c(W, \mathcal{L})$$



Datasets

KITTI video

[Geiger et al. 2012]

Autonomous car
platform

Egomotions: yaw and
forward distance



City

Residential

Road

Campus

SUN images

[Xiao et al. 2010]

Large-scale scene
classification task

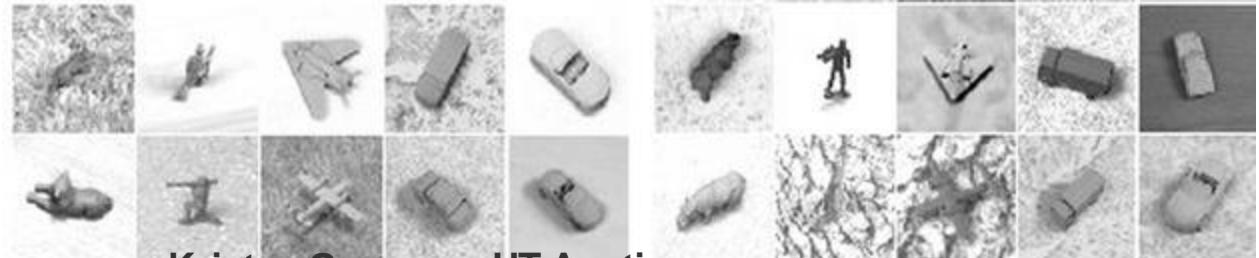


NORB images

[LeCun et al. 2004]

Toy recognition

Egomotions: elevation
and azimuth



Kristen Grauman, UT Austin

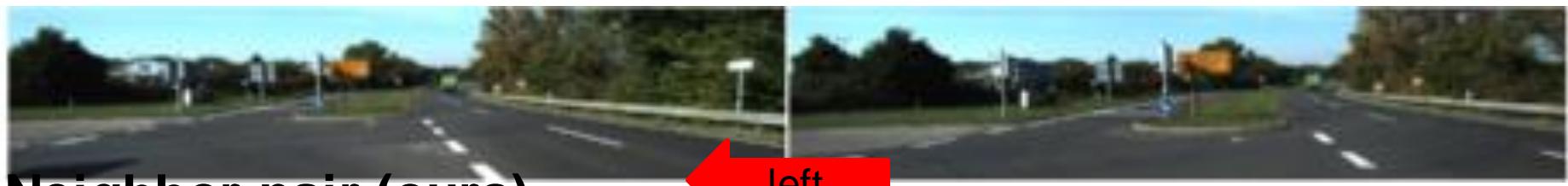
Results: Equivariance check

Visualizing how well equivariance is preserved



Query pair

← left



Neighbor pair (ours)

← left



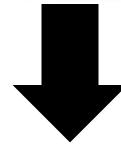
Pixel space neighbor pair zoom

Results: Recognition

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



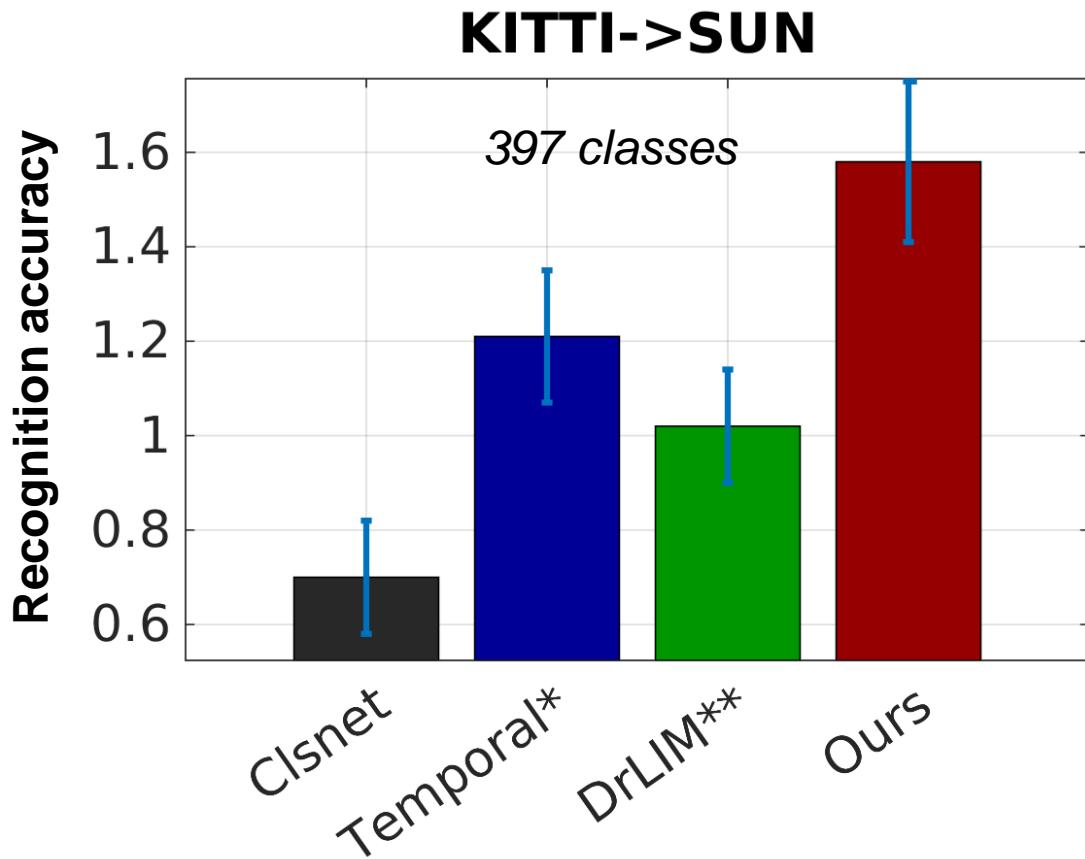
Exploit features for large multi-way
scene classification (SUN, 397 classes)



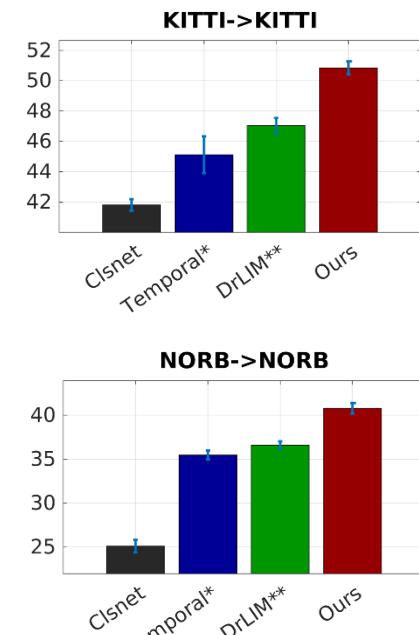
30% accuracy increase
for small labeled training sets

Results: Recognition

Do the learned features boost recognition accuracy?



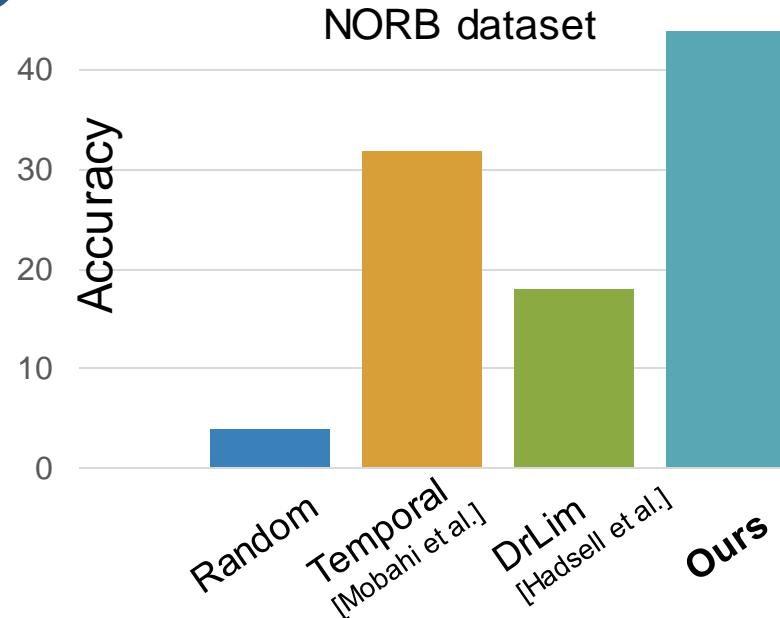
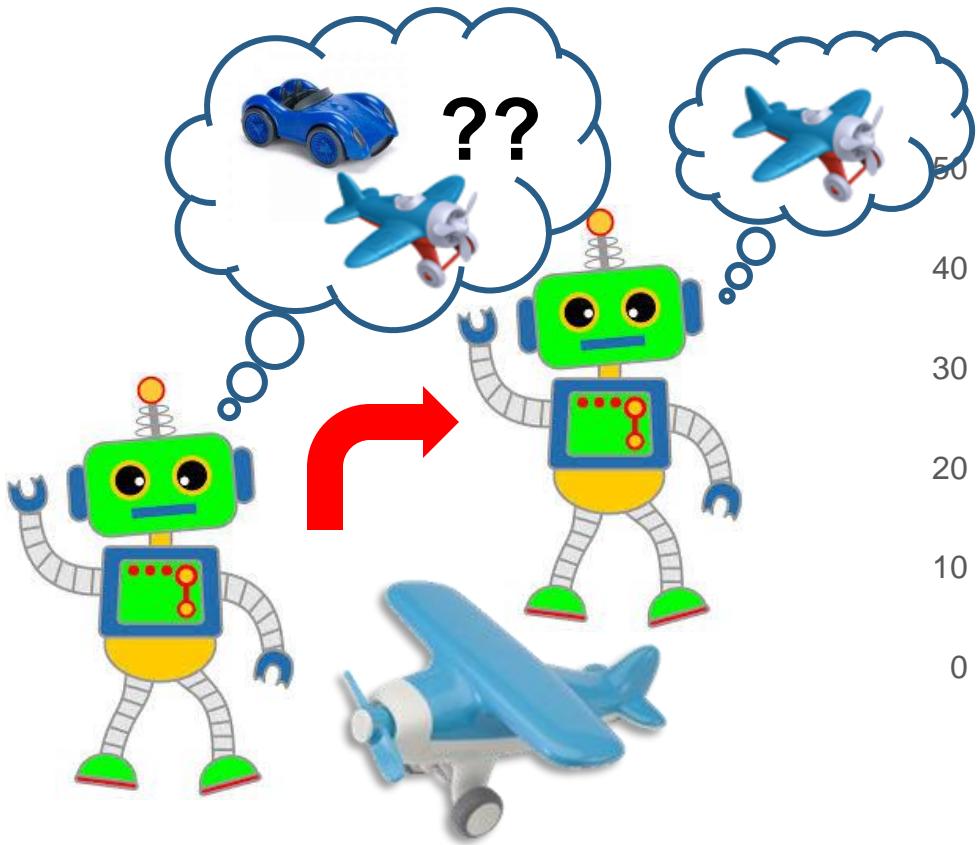
6 labeled training
examples per class



*Mohahi et al. ICML09; **Hadsell et al. CVPR06

Results: Active recognition

Leverage proposed equivariant embedding to predict **next best view** for object recognition



[Bajcsy 1988, Tsotsos 1992, Schiele & Crowley 1998, Tsotsos et al., Dickinson et al. 1997, Soatto 2009, Mishra et al. 2009,...]

Next steps

- Dynamic objects
- Multiple modalities, e.g., depth
- Active ego-motion planning
- Tasks aside from recognition

Talk overview

Motivation

Account for the fact that **camera wearer is active participant** in the visual observations received

Ideas

1. **Action:** Unsupervised feature learning
 - How is visual learning shaped by ego-motion?
2. **Attention:** Inferring highlights in video
 - How to summarize long egocentric video?

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

What makes egocentric data hard to summarize?

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

Existing summarization methods largely 3rd-person

[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,...]

Summarizing egocentric video

Key questions

- How to detect subshots in ongoing video?
- What objects are important?
- How are events linked?
- When is attention heightened?
- Which frames look “intentional”?

Goal: Story-driven summarization



Characters and plot \leftrightarrow Key objects and influence

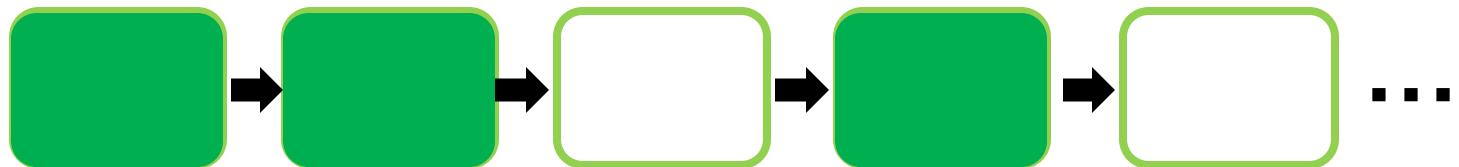
Summarization as subshot selection

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

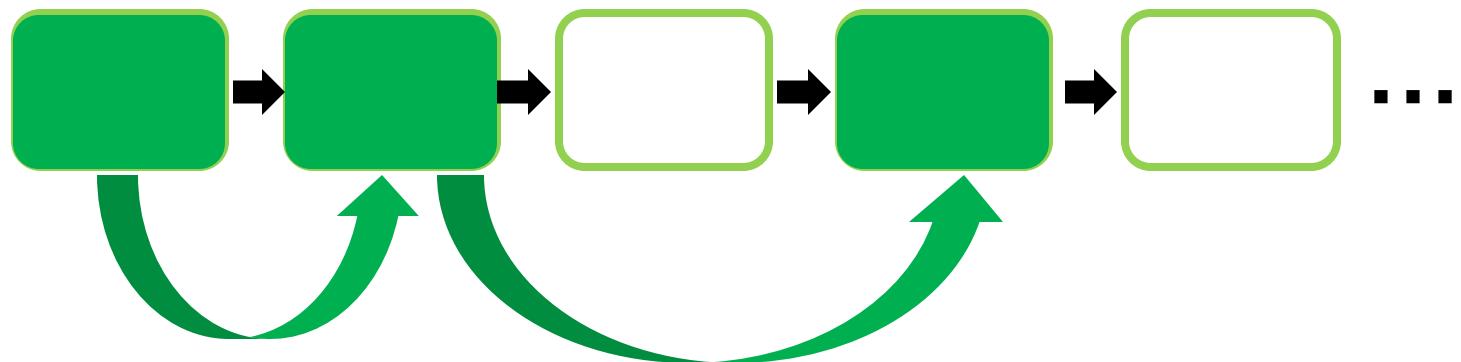


Estimating visual influence

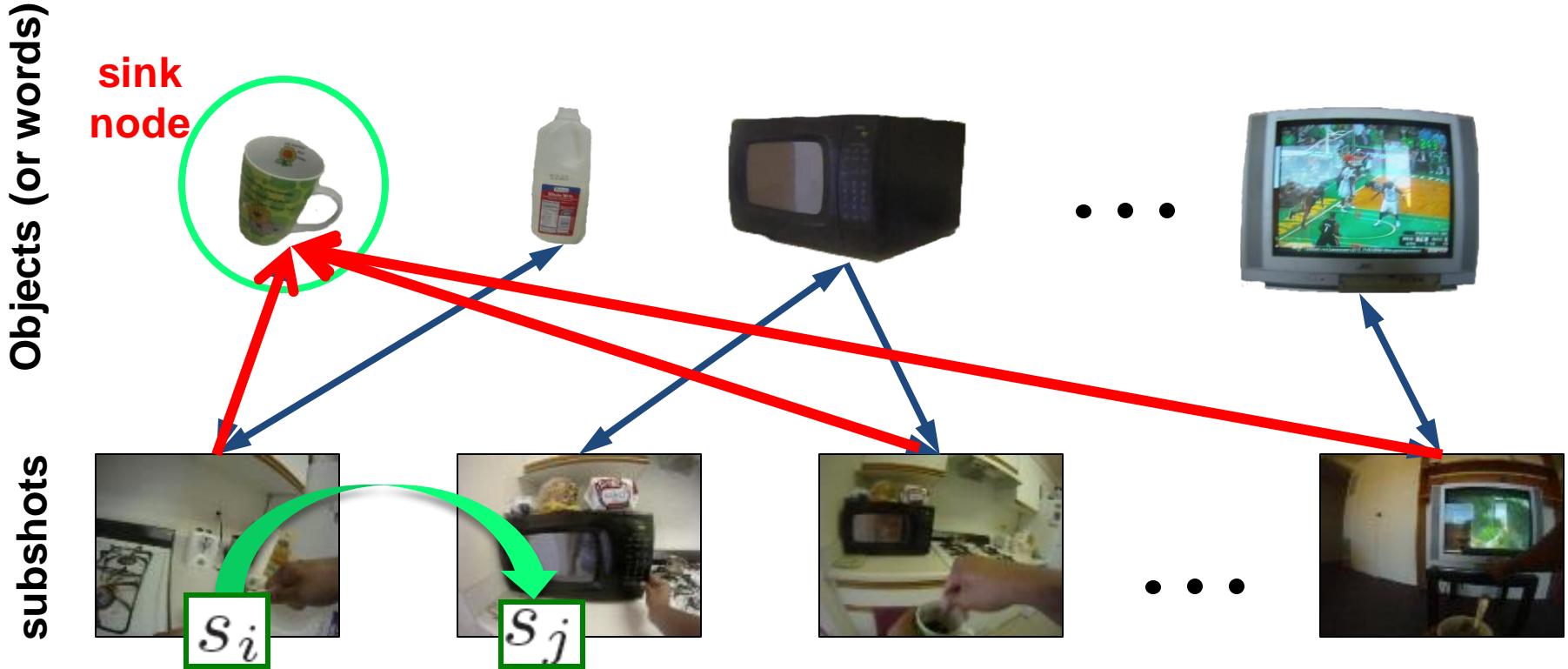
- Aim to select the k subshots that maximize the influence between objects (on the weakest link)

$$\mathcal{S}(S) = \max_a \min_{j=1, \dots, K-1} \sum_{o_i \in O} a_{i,j} \text{INFLUENCE}(s_j, s_{j+1} | o_i)$$

Subshots



Estimating visual influence



$$\text{INFLUENCE}(s_i, s_j | o) = \prod_i(s_j) - \prod_i^o(s_j)$$

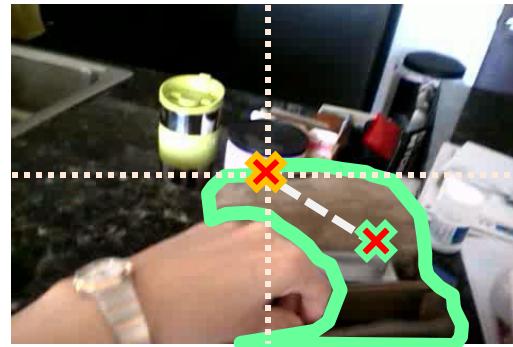
Captures how reachable subshot j is from subshot i , via any object o

Learning object importance

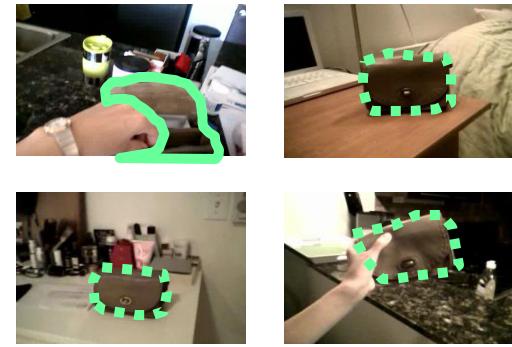
We learn to rate regions by their egocentric importance



distance to hand



distance to frame center



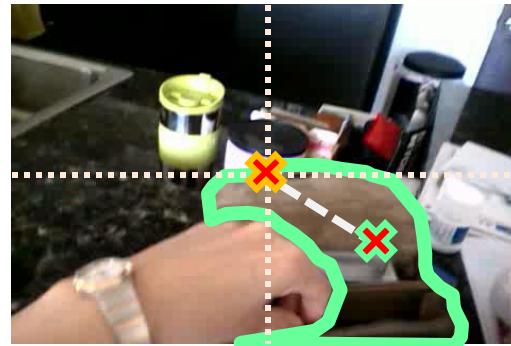
frequency

Learning object importance

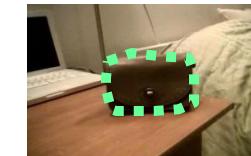
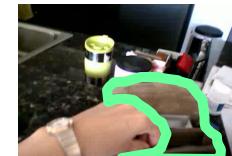
We learn to rate regions by their egocentric importance



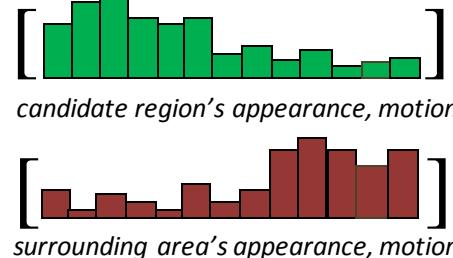
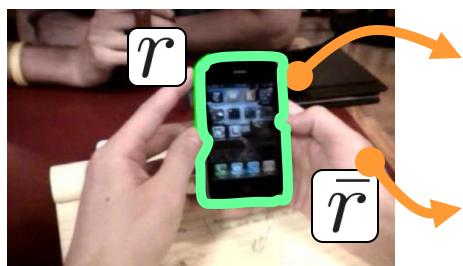
distance to hand



distance to frame center

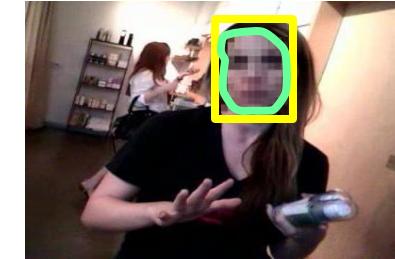


frequency



“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

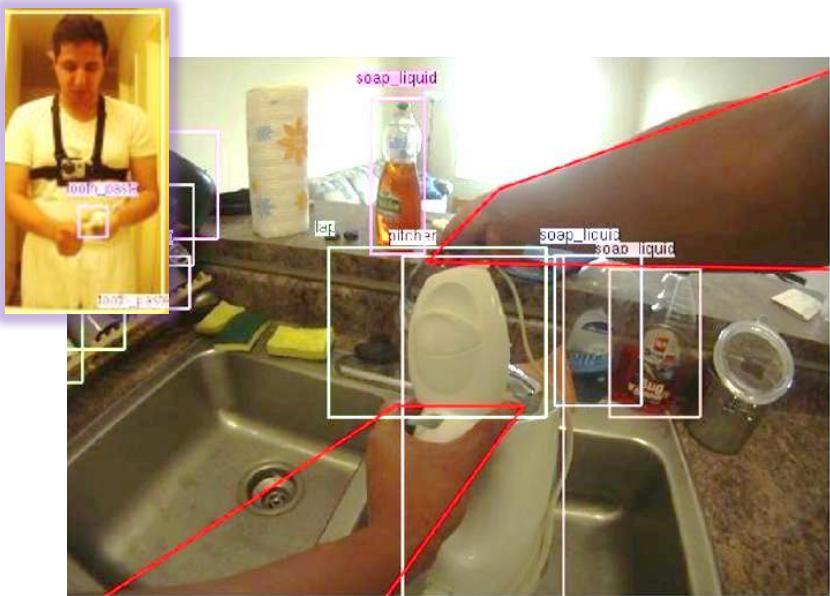
UT Egocentric (UT Ego) [Lee et al. 2012]



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)

Example keyframe summary – UT Ego data

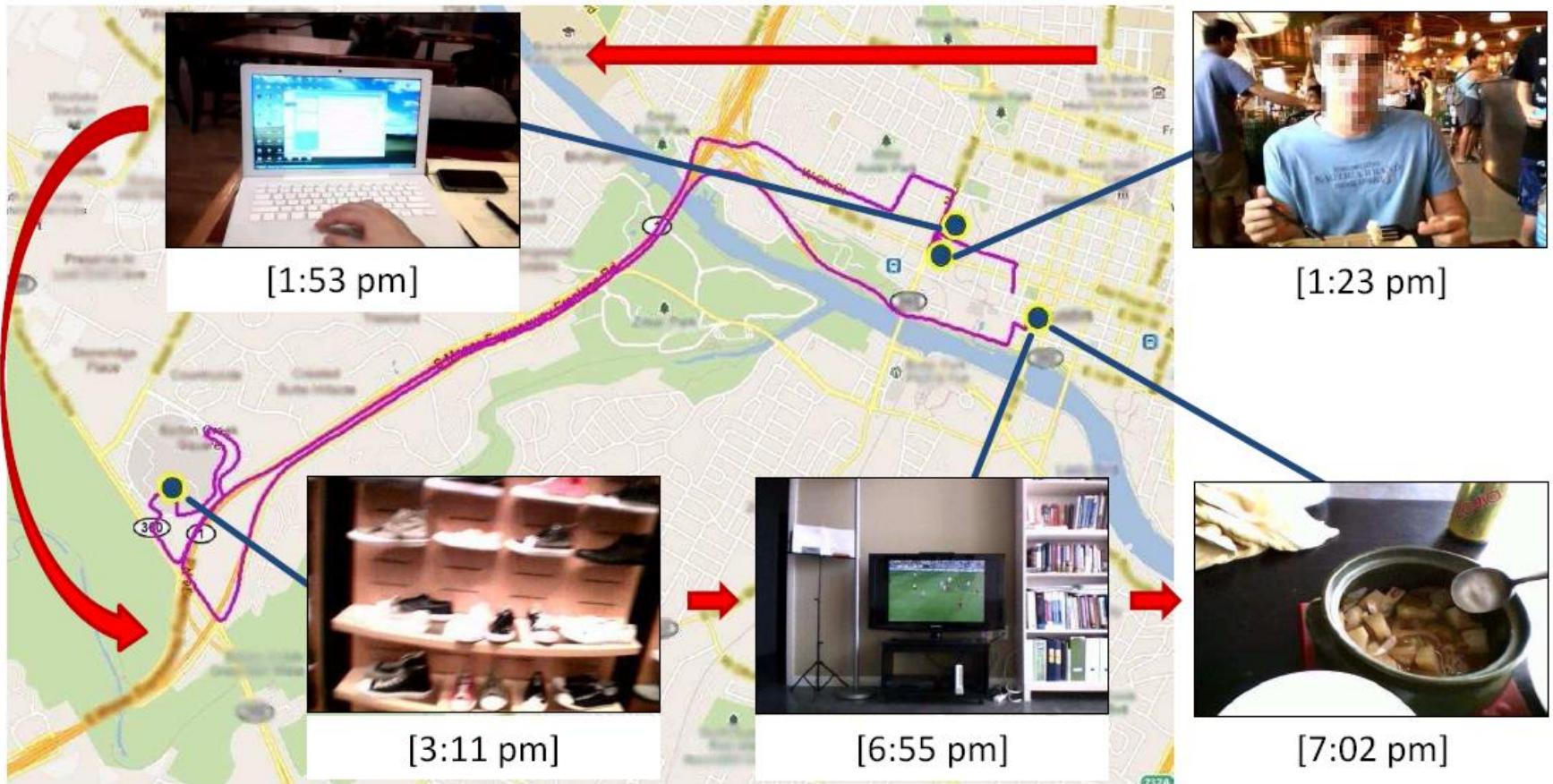
Alternative methods for comparison



**Uniform keyframe sampling
(12 frames)**

**[Liu & Kender, 2002]
(12 frames)**

Generating storyboard maps



Augment keyframe summary with geolocations

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 Ego-centric 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

Which photos were purposely taken by a human?



Incidental wearable camera photos



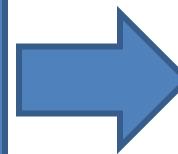
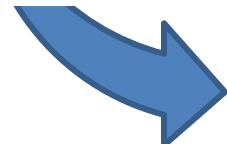
Intentional human taken photos

Idea: Detect “snap points”

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



Domain
adapted
similarity

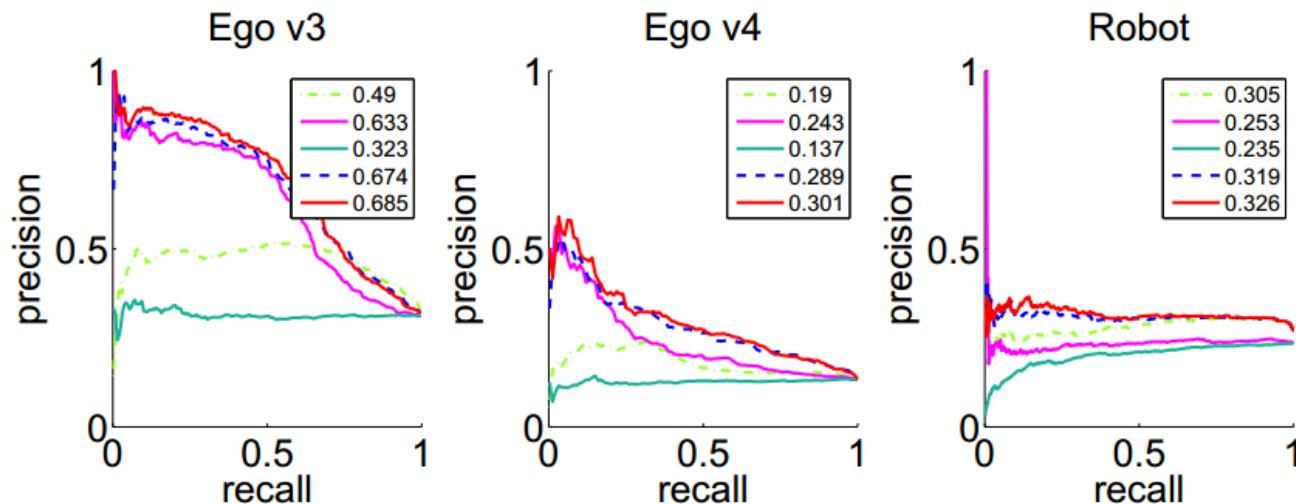
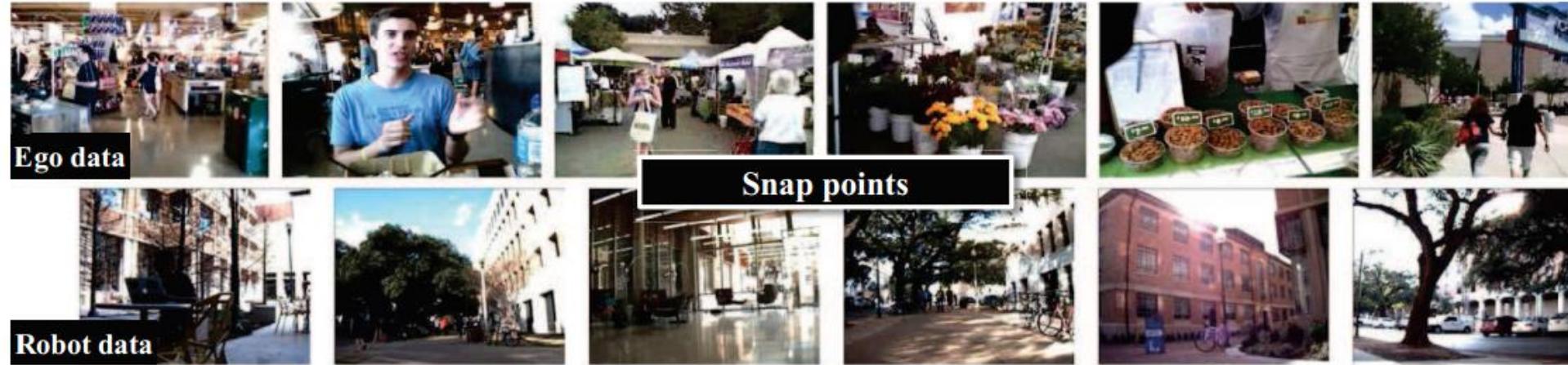


Snap point
score

Example snap point predictions



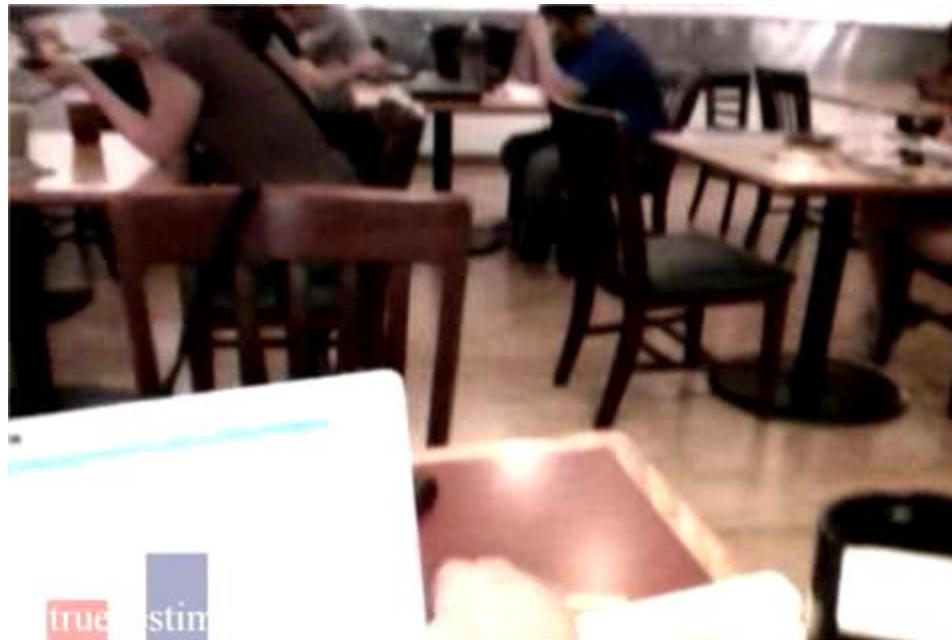
Example snap point predictions



Blur People Likelihood Saliency Web Prior(ours) Web Prior+DA(ours)

Snap point predictions

Intentional photos from an unintentional photographer



Next steps

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

Summary



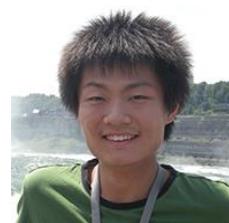
Dinesh
Jayaraman



Yong Jae
Lee



Yu-Chuan
Su



Bo
Xiong



Lu
Zheng

- New opportunities with “always on” ego cameras
- Towards active first-person vision:
 - **Action:** “Embodied” feature learning from ego-video using both visual and motor signals
 - **Attention:** Egocentric summarization tools to cope with deluge of wearable camera data