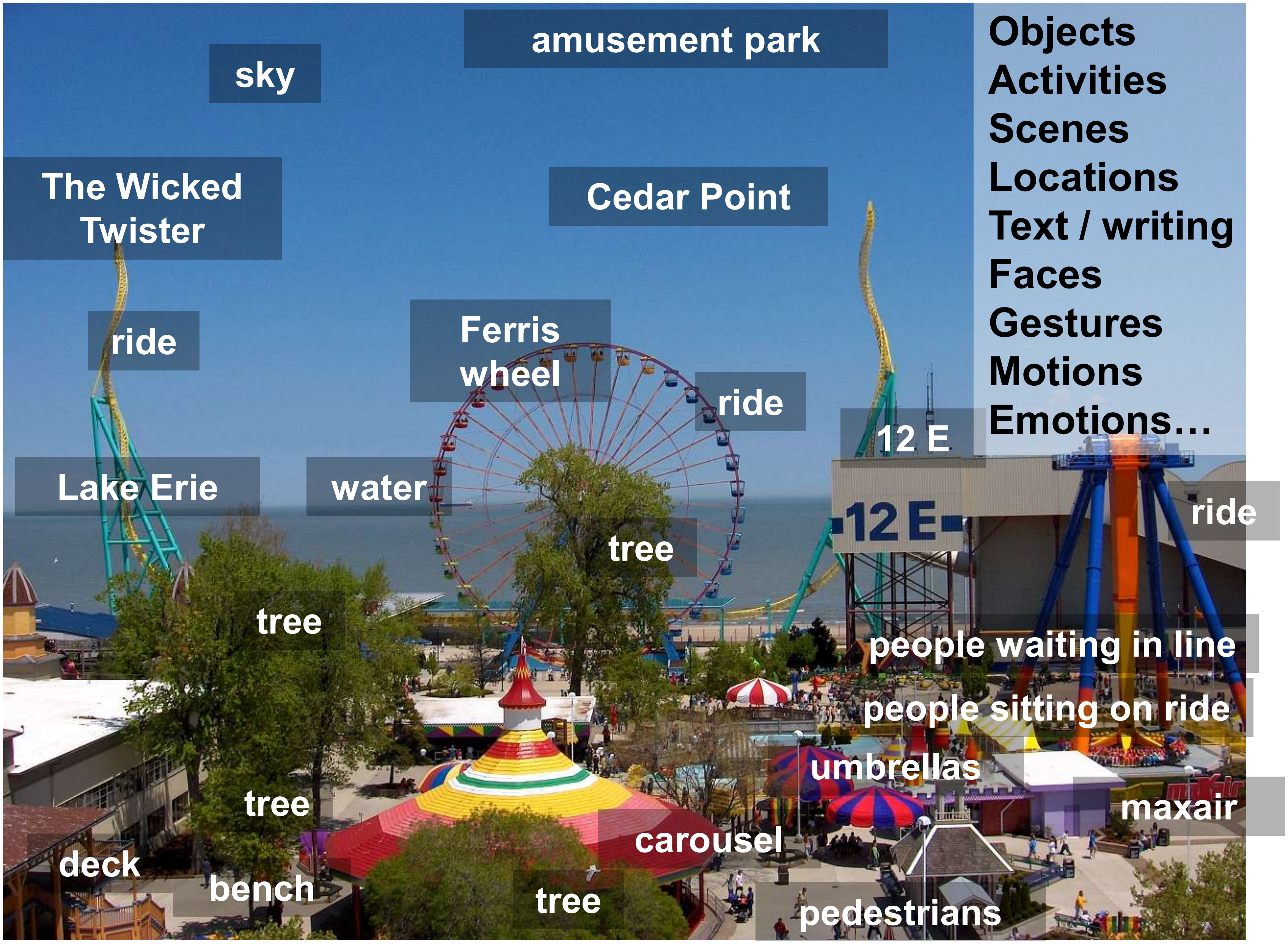
Learning from Unlabeled Video

Kristen Grauman
Department of Computer Science
University of Texas at Austin

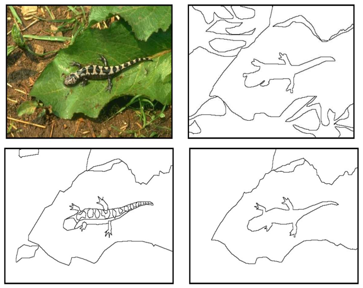
THE UNIVERSITY OF TEXAS AT AUSTIN

Visual recognition

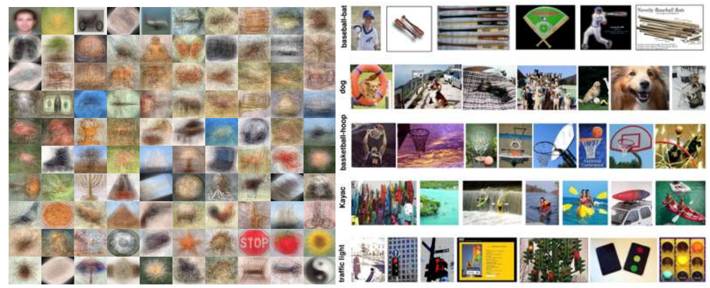


Objects
Activities
Scenes
Locations
Text / writing
Faces
Gestures
Motions
Emotions…

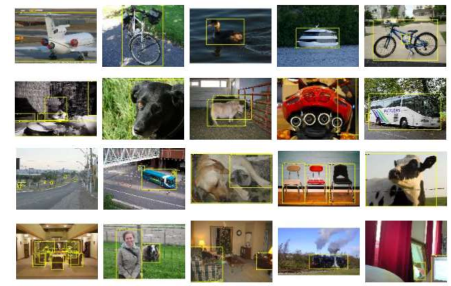
Recognition: as seen by its benchmarks



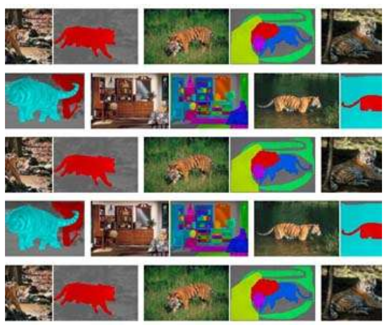
BSD (2001)



Caltech 101 (2004), Caltech 256 (2006)



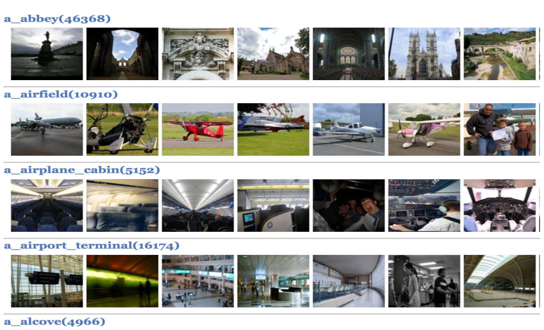
PASCAL (2007-12)



LabelMe (2007)

ImageNet (2009)

SUN (2010)



Places (2014)

MS COCO (2014)

Visual Genome (2016)

How do our systems learn about the visual world today?

Big picture goal: Embodied visual learning

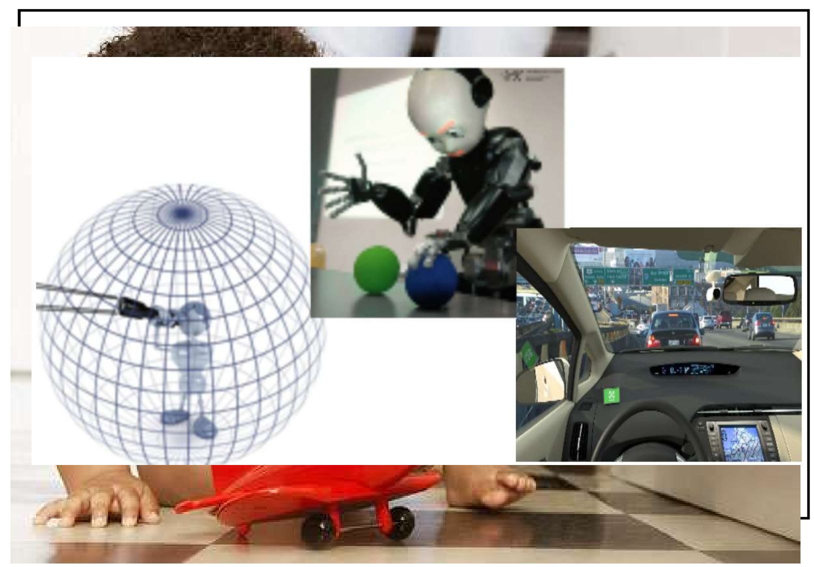
Status quo:

Learn from “disembodied” bag of labeled snapshots.

Our goal:

Visual learning in the context of acting and moving in the world.

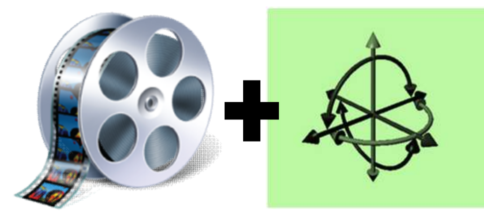
Inexpensive and unrestricted in scope



Talk overview

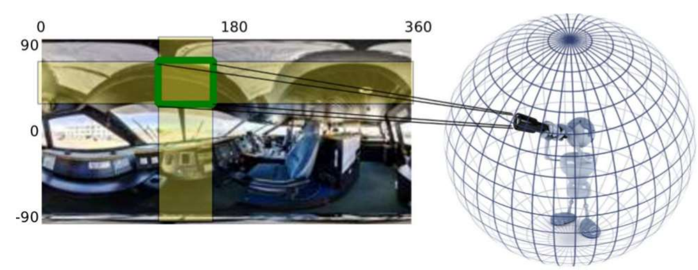
Towards embodied visual learning

1. Learning representations tied to ego-motion

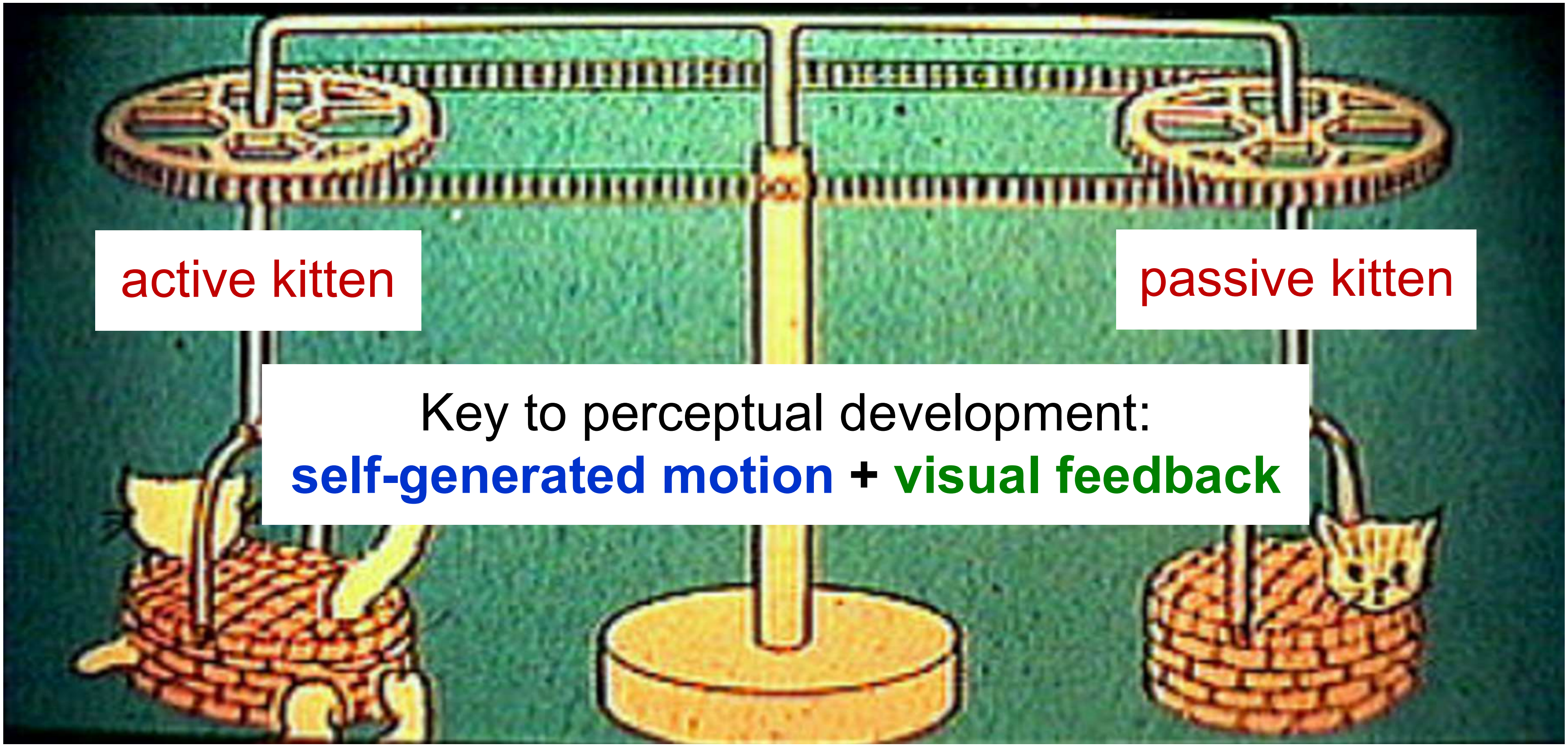


2. Learning representations from unlabeled video

3. Learning how to move and where to look



The kitten carousel experiment
[Held & Hein, 1963]



Our idea: Ego-motion ↔ vision

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

Ego-motion motor signals

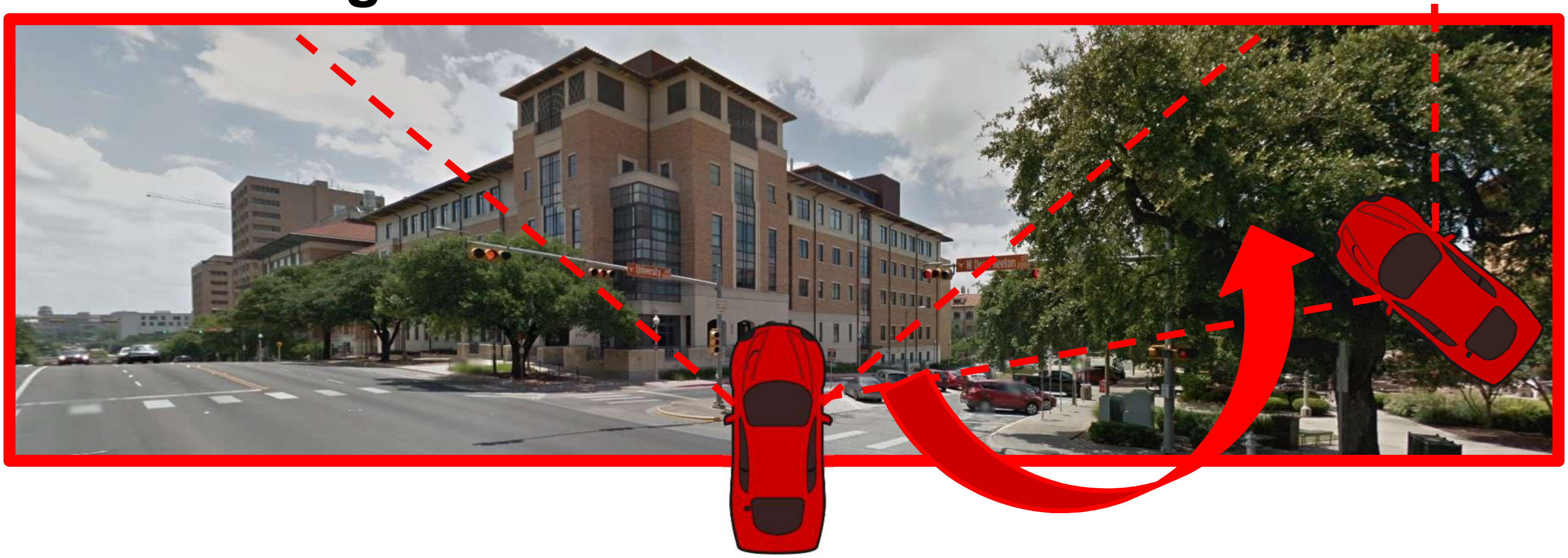
+

Unlabeled video

[Jayaraman & Grauman, ICCV 2015]

Ego-motion ↔ vision: view prediction

After moving:



Approach idea: Ego-motion equivariance

Invariant features: unresponsive to some classes of transformations

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

Simard et al, Tech Report, '98
Wiskott et al, Neural Comp '02
Hadsell et al, CVPR '06
Mobahi et al, ICML '09
Zou et al, NIPS '12
Sohn et al, ICML '12
Cadieu et al, Neural Comp '12
Goroshin et al, ICCV '15
Lies et al, PLoS computation biology '14
…

Approach idea: Ego-motion equivariance

Invariant features: unresponsive to some classes of transformations

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

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

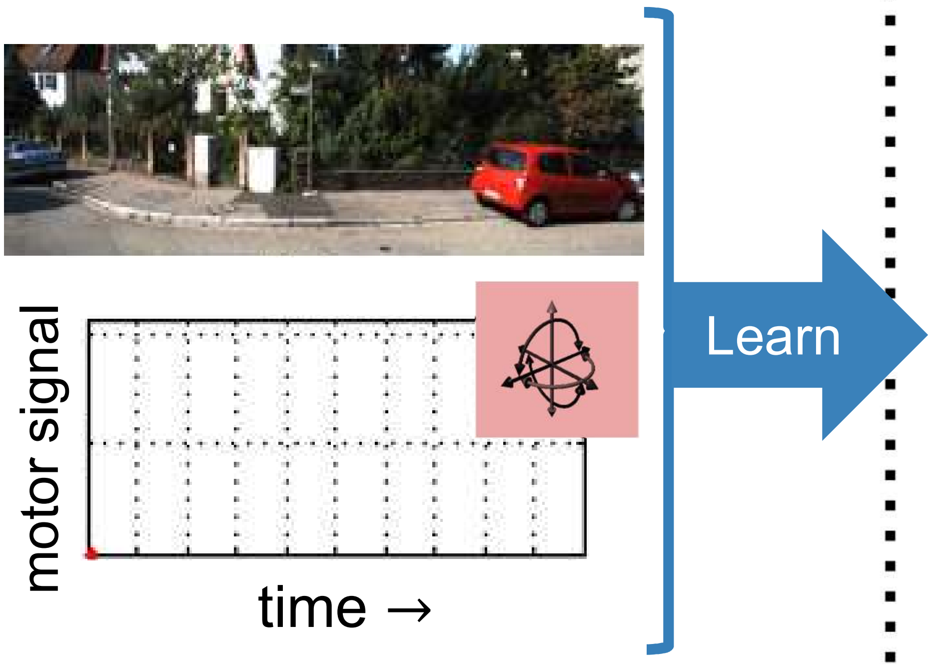
"equivariance map"

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

Invariance *discards* information;
equivariance *organizes* it.

Approach idea: Ego-motion equivariance

Training data
Unlabeled video +
motor signals



Equivariant embedding
organized by ego-motions

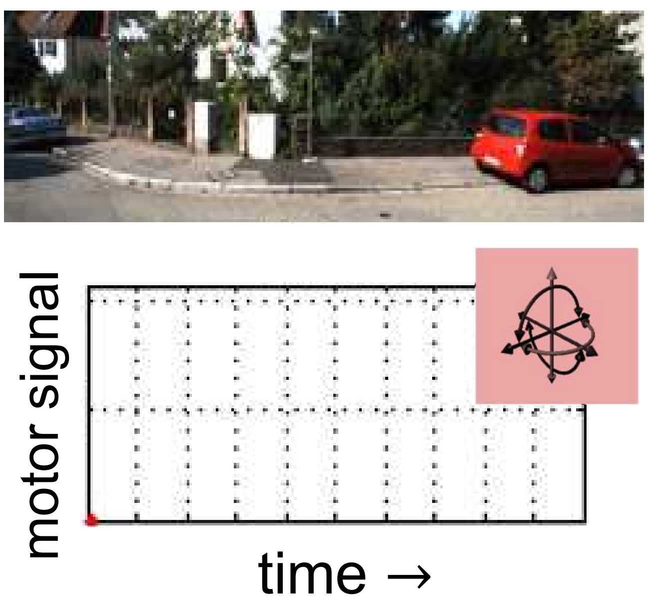
Pairs of frames related by similar ego-motion should be related by same feature transformation

[Jayaraman & Grauman, ICCV 2015]

Approach idea: Ego-motion equivariance

Training data

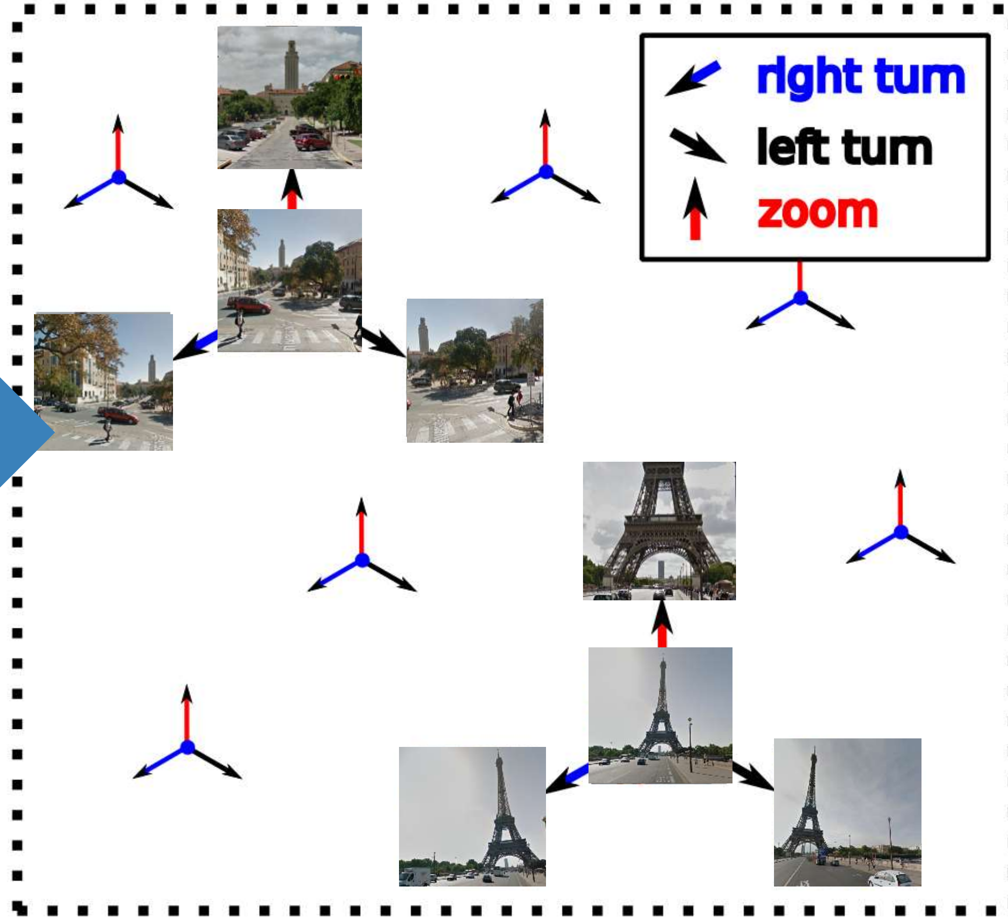
Unlabeled video + motor signals



Learn

Equivariant embedding

organized by ego-motions



[Jayaraman & Grauman, ICCV 2015]

Results: Recognition

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

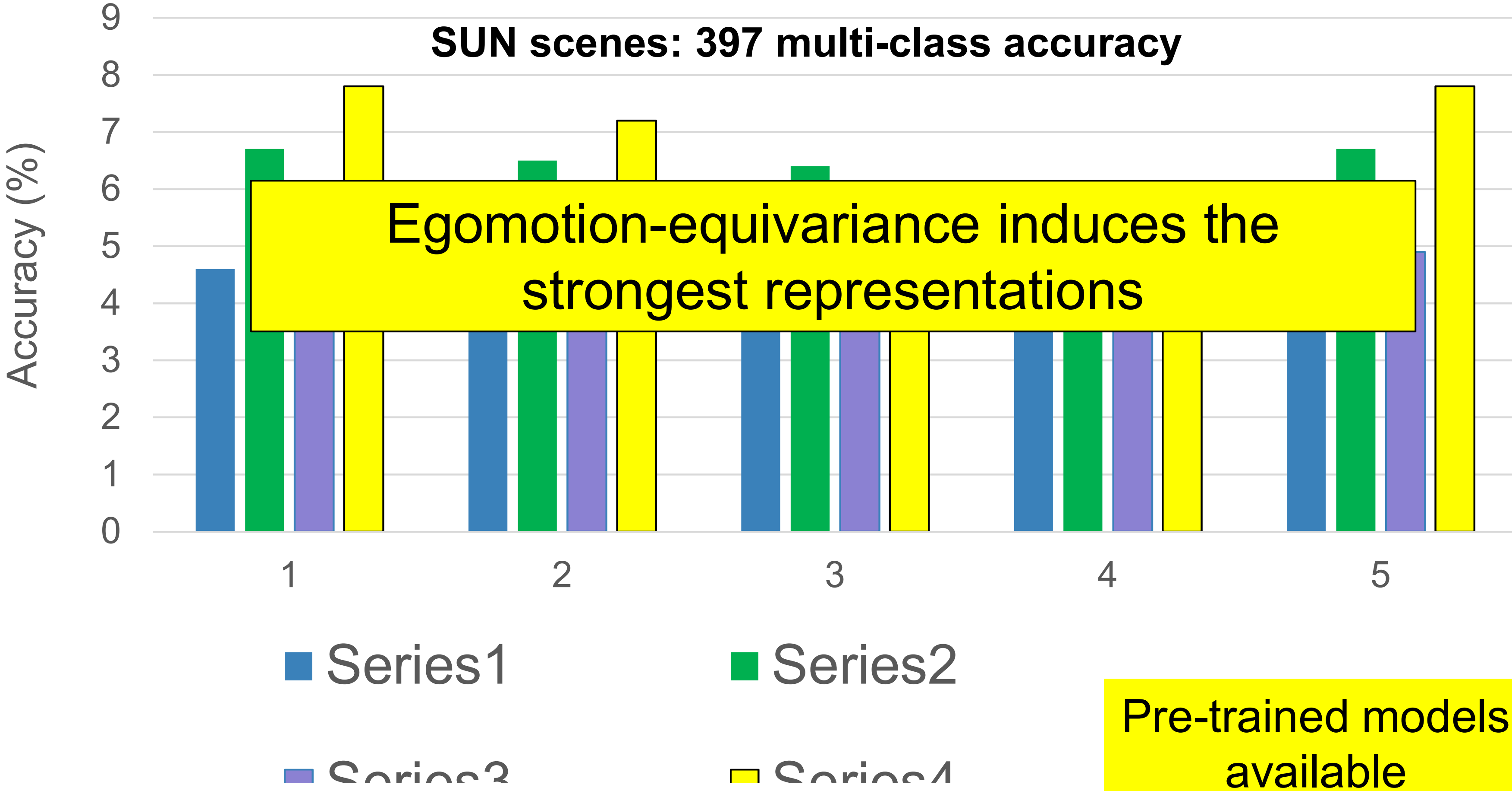
Geiger et al, IJRR ’13

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

Xiao et al, CVPR ’10

Results: Recognition

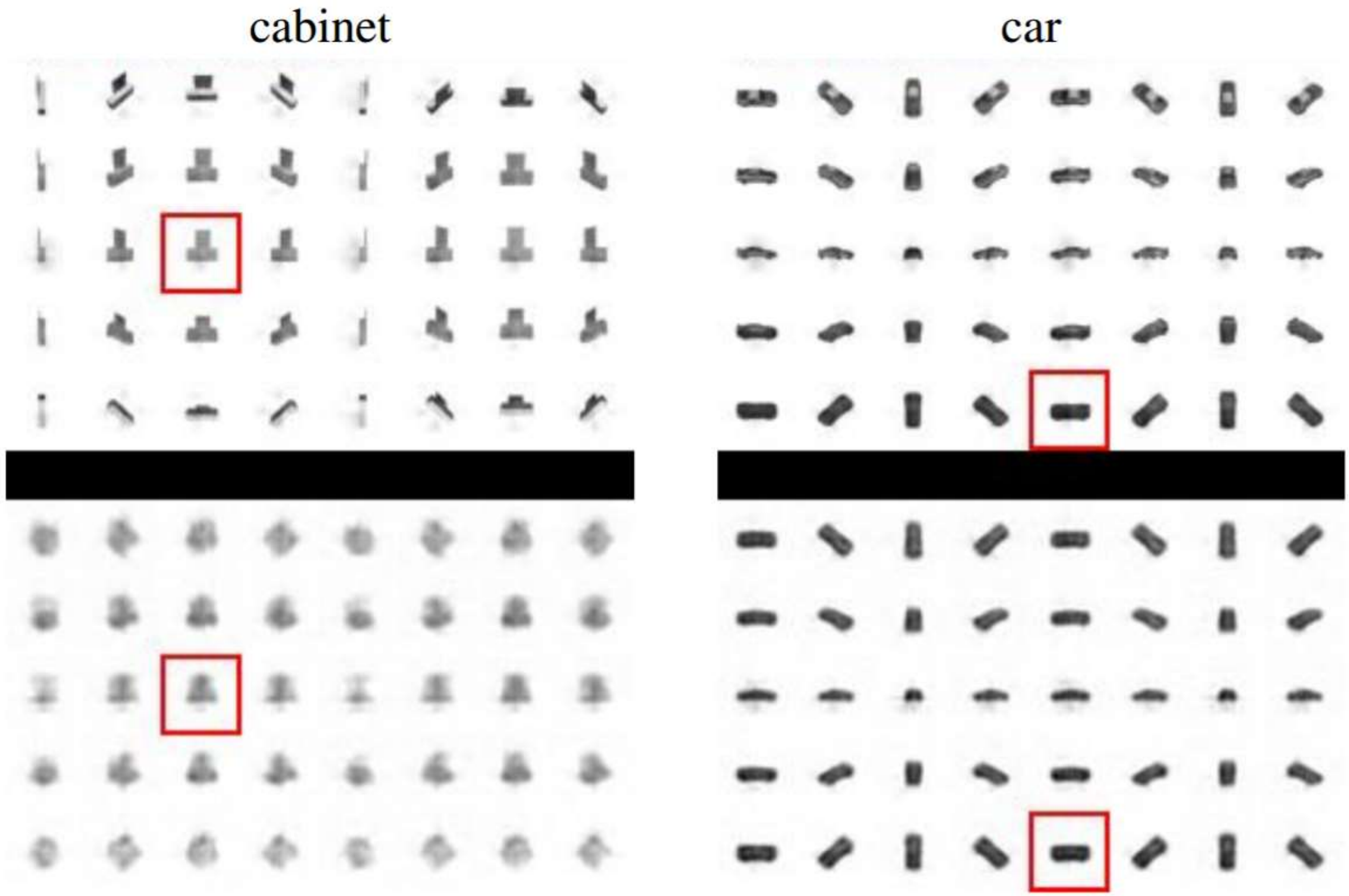
Ego-equivariance for unsupervised feature learning



Pre-trained models available

+ Hadsell, Chopra, LeCun, "Dimensionality Reduction by Learning an Invariant Mapping", CVPR 2006
* Agrawal, Carreira, Malik, "Learning to see by moving", ICCV 2015

Next steps: One-shot shape reconstruction for feature learning

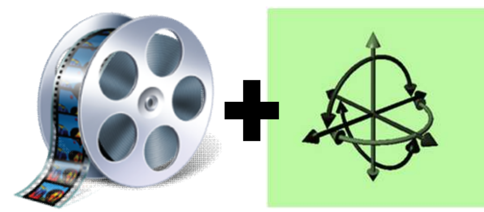


Jayaraman et al. 2017

Talk overview

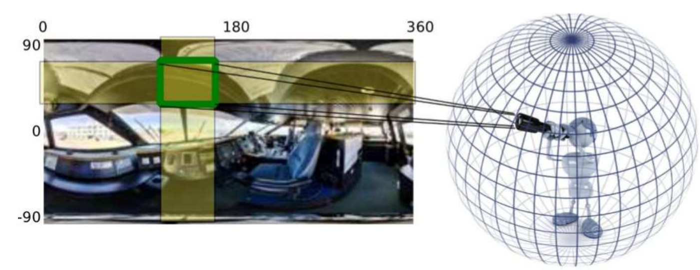
Towards embodied visual learning

1. Learning representations tied to ego-motion

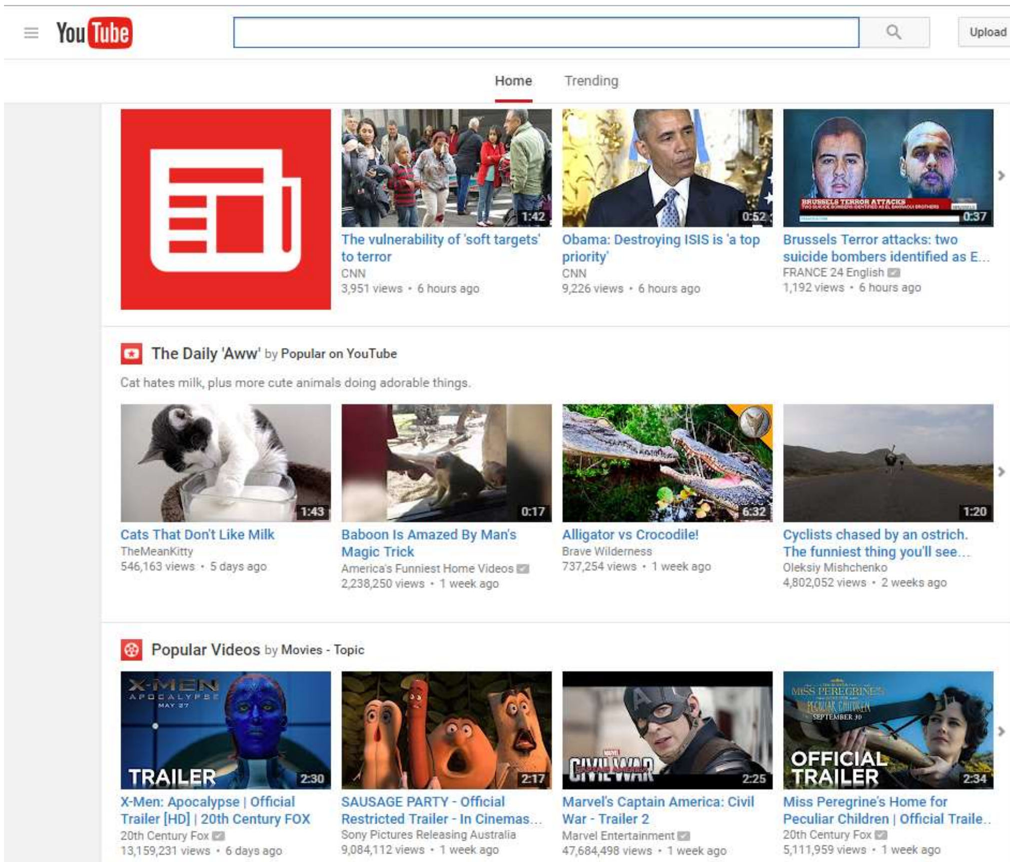


2. Learning representations from unlabeled video

3. Learning how to move and where to look



Learning from arbitrary unlabeled video?



Unlabeled video

Prior work: Slow feature analysis

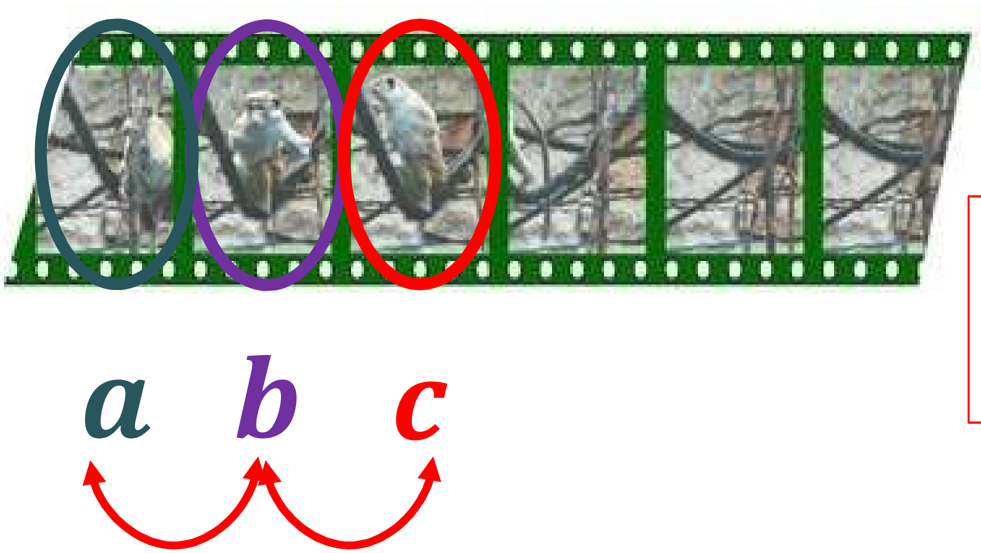
$\boldsymbol{a}$ $\boldsymbol{b}$

Wiskott et al, 2002
Hadsell et al. 2006
Mobahi et al. 2009
Bergstra & Bengio 2009
Goroshin et al. 2013
Wang & Gupta 2015
Gao et al. 2016
…

Learn feature map $\boldsymbol{z}(.)$ such that:

$$\boldsymbol{z}(\boldsymbol{a}) \approx \boldsymbol{z}(\boldsymbol{b})$$ *(invariance)*

Our idea: *Steady* feature analysis



Higher order temporal coherence

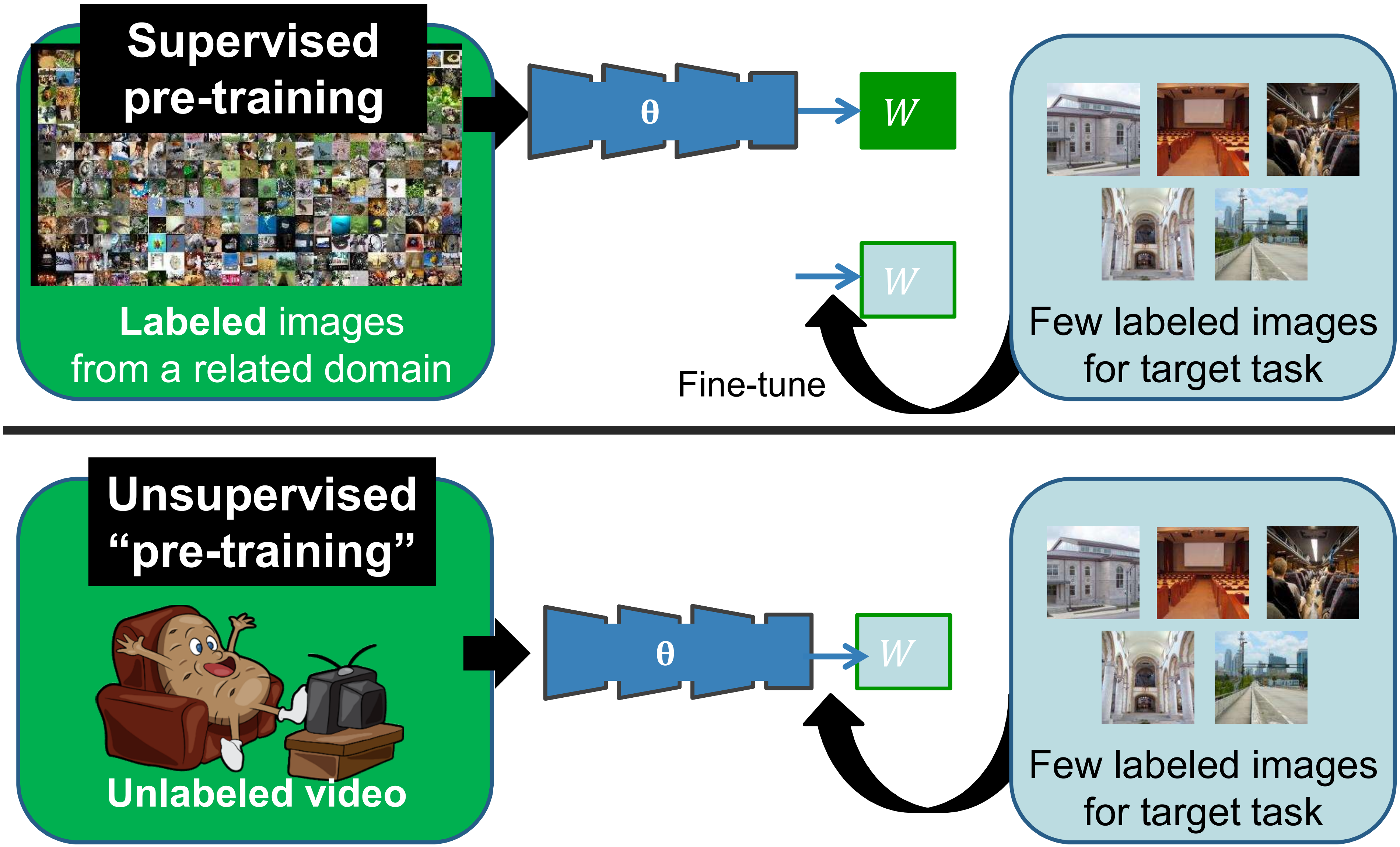
Learn feature map $\boldsymbol{z}(.)$ such that:

$$\boldsymbol{z}(\boldsymbol{a}) \approx \boldsymbol{z}(\boldsymbol{b}) \quad \textit{(invariance)}$$

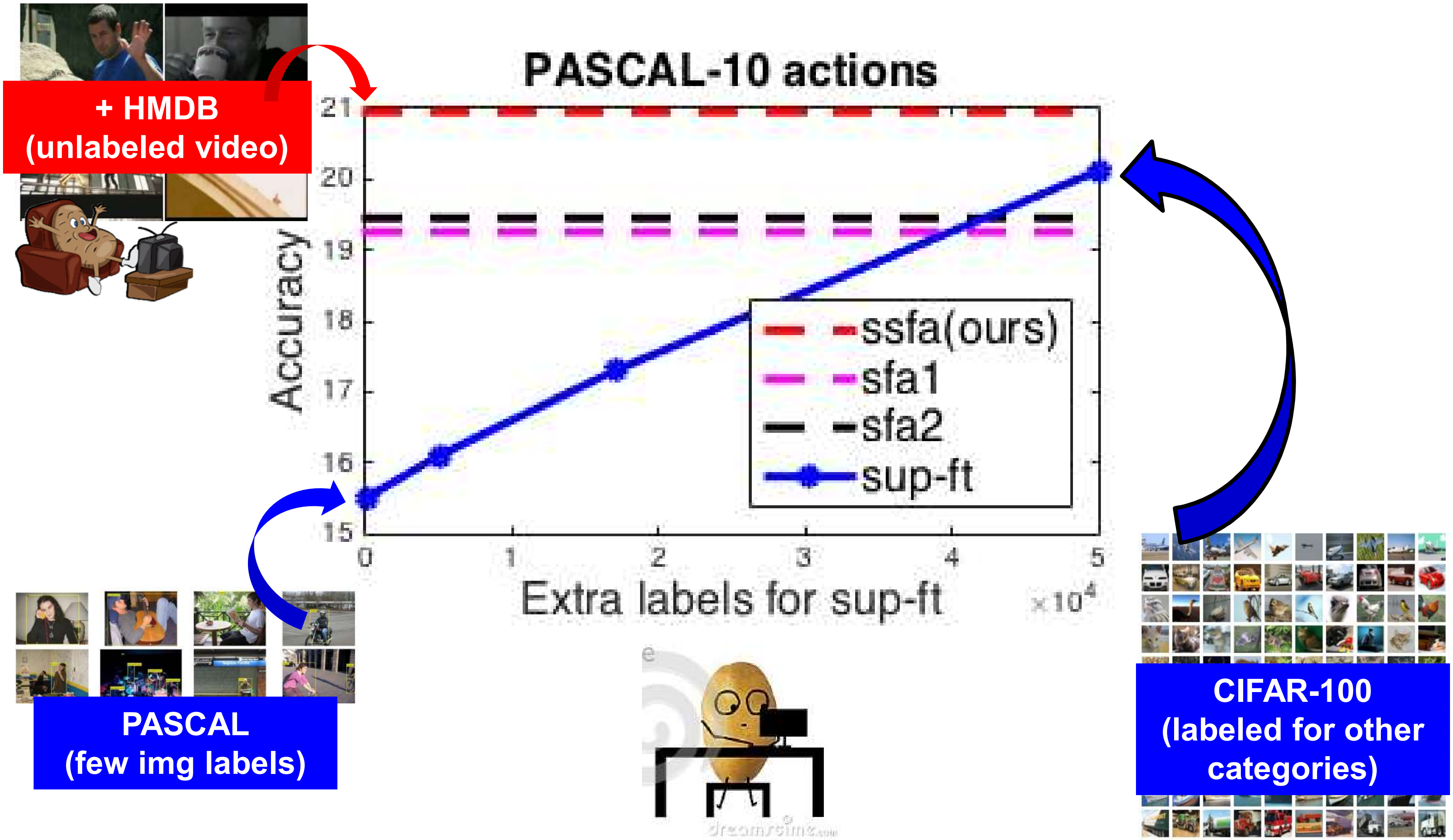
$$\boldsymbol{z}(\boldsymbol{a}) - \boldsymbol{z}(\boldsymbol{b}) \approx \boldsymbol{z}(\boldsymbol{b}) - \boldsymbol{z}(\boldsymbol{c}) \quad \textit{(equivariance)}$$

[Jayaraman & Grauman, CVPR 2016]

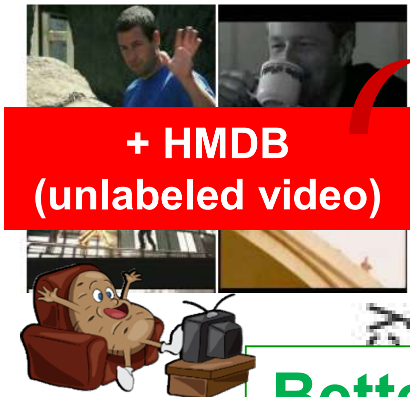
Pre-training a representation



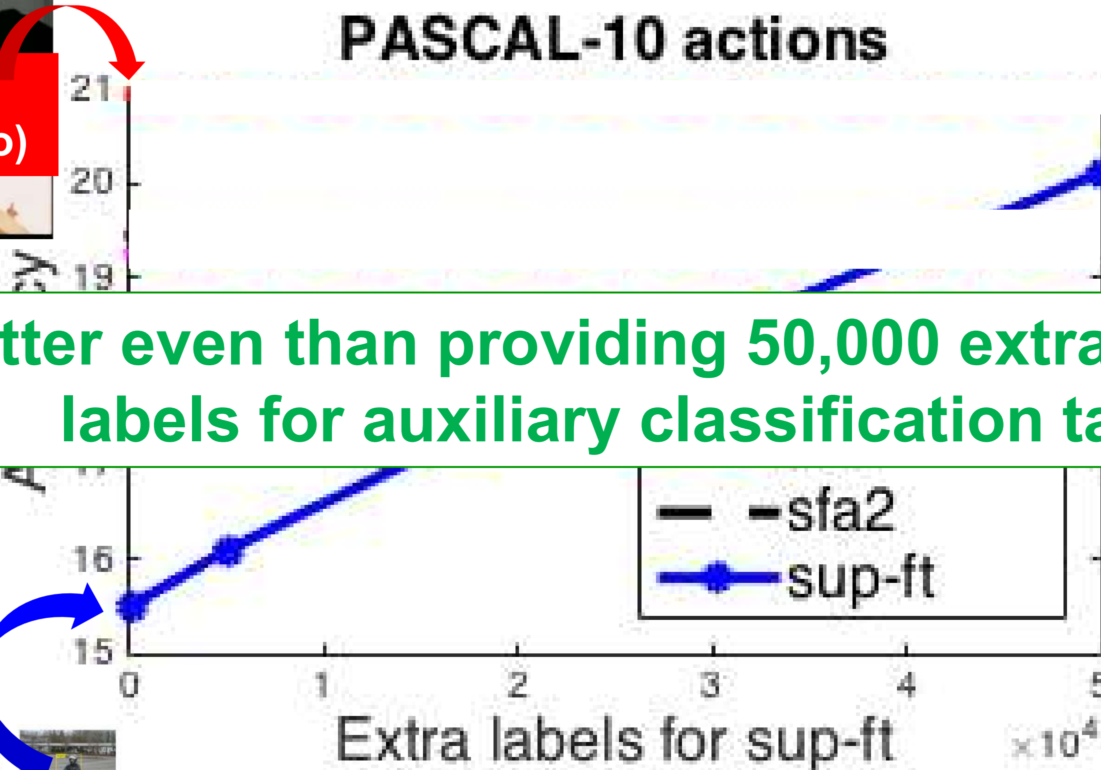
Results: Can we learn *more* from unlabeled video than “related” labeled images?



Results: Can we learn *more* from unlabeled video than "related" labeled images?



+ HMDB
(unlabeled video)



Better even than providing 50,000 extra manual labels for auxiliary classification task!

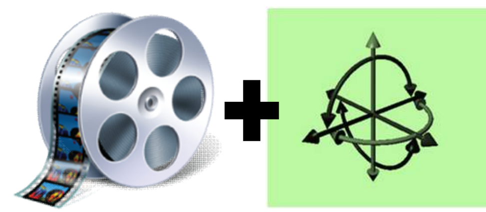
PASCAL
(few img labels)

CIFAR-100
(labeled for other categories)

Talk overview

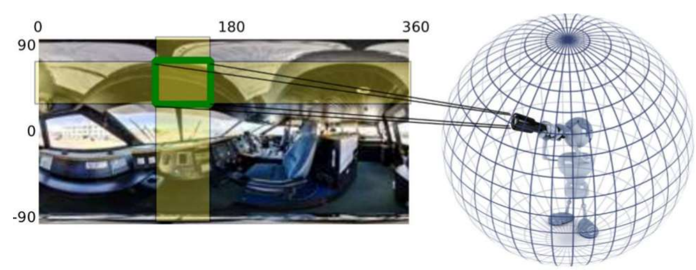
Towards embodied visual learning

1. Learning representations tied to ego-motion



2. Learning representations from unlabeled video

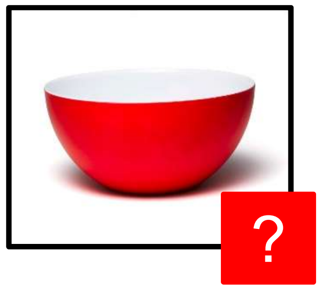
3. Learning how to move and where to look

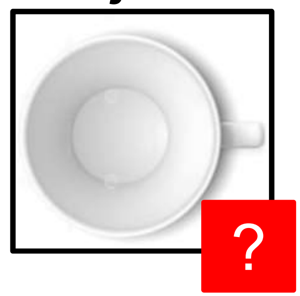


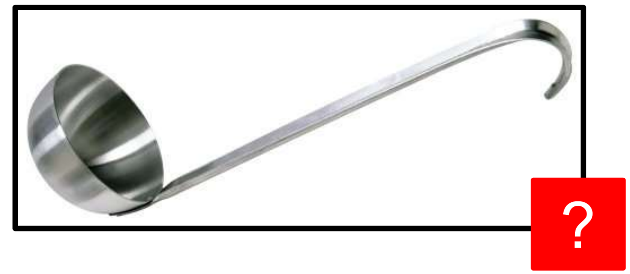
Current recognition benchmarks

Passive, disembodied snapshots at *test* time, too

Object recognition

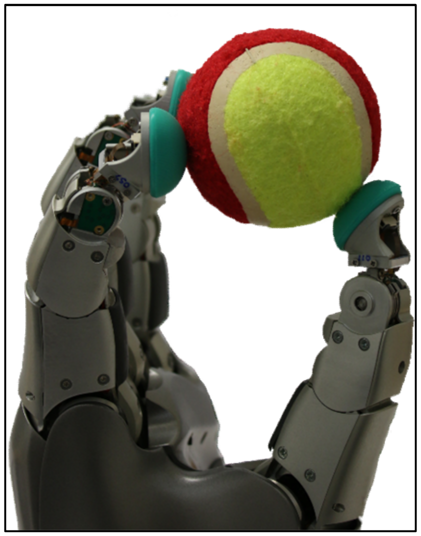






Scene recognition

Moving to recognize

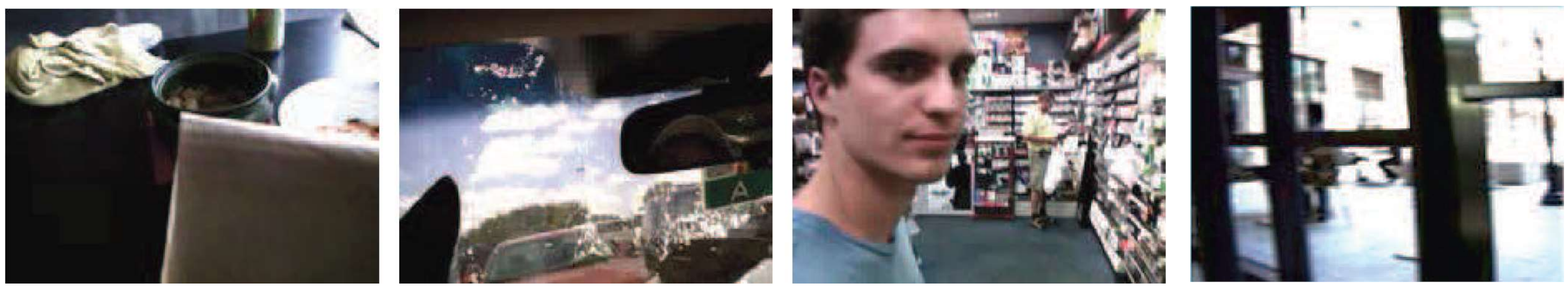


Time to revisit active recognition in challenging settings!

Bajcsy 1985, Aloimonos 1988, Ballard 1991, Wilkes 1992, Dickinson 1997, Schiele & Crowley 1998, Tsotsos 2001, Denzler 2002, Soatto 2009, Ramanathan 2011, Borotschnig 2011, …

Moving to recognize

Difficulty: unconstrained visual input

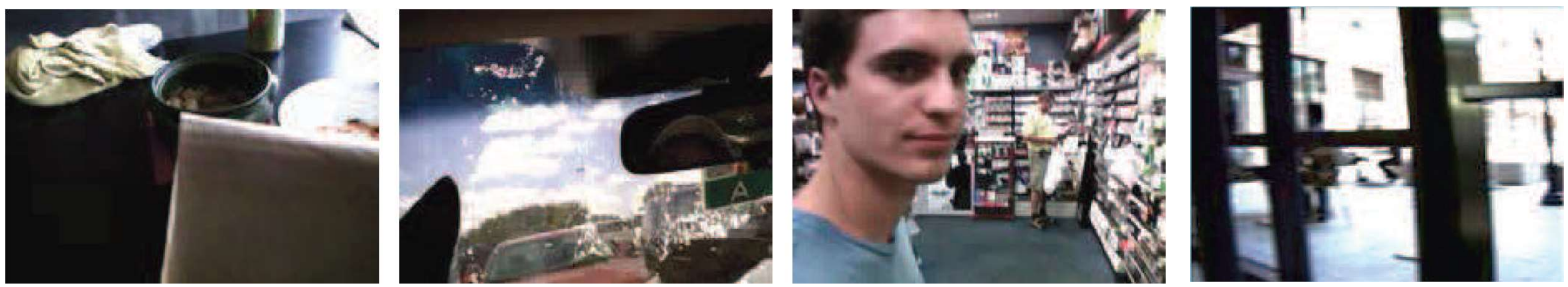


vs.

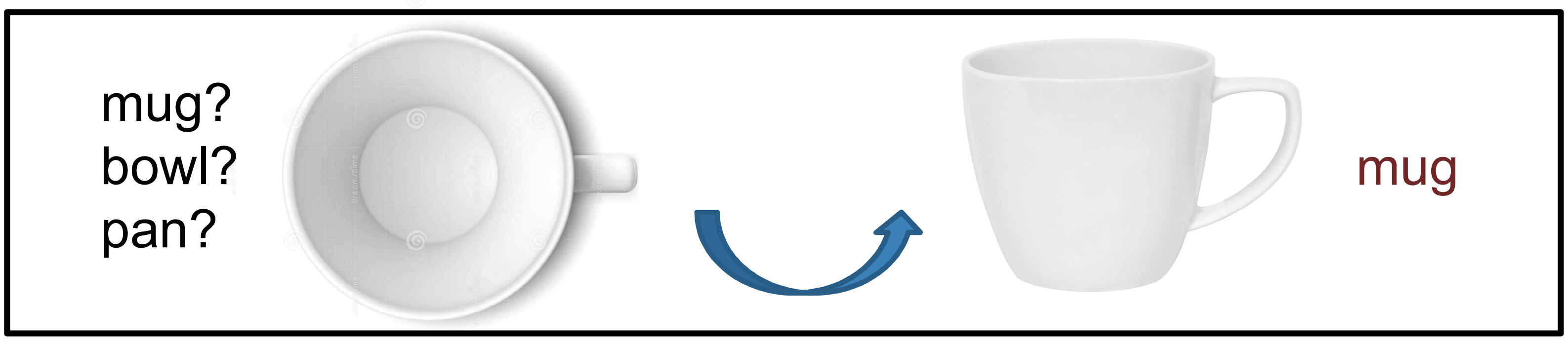
ImageNet Web images

Moving to recognize

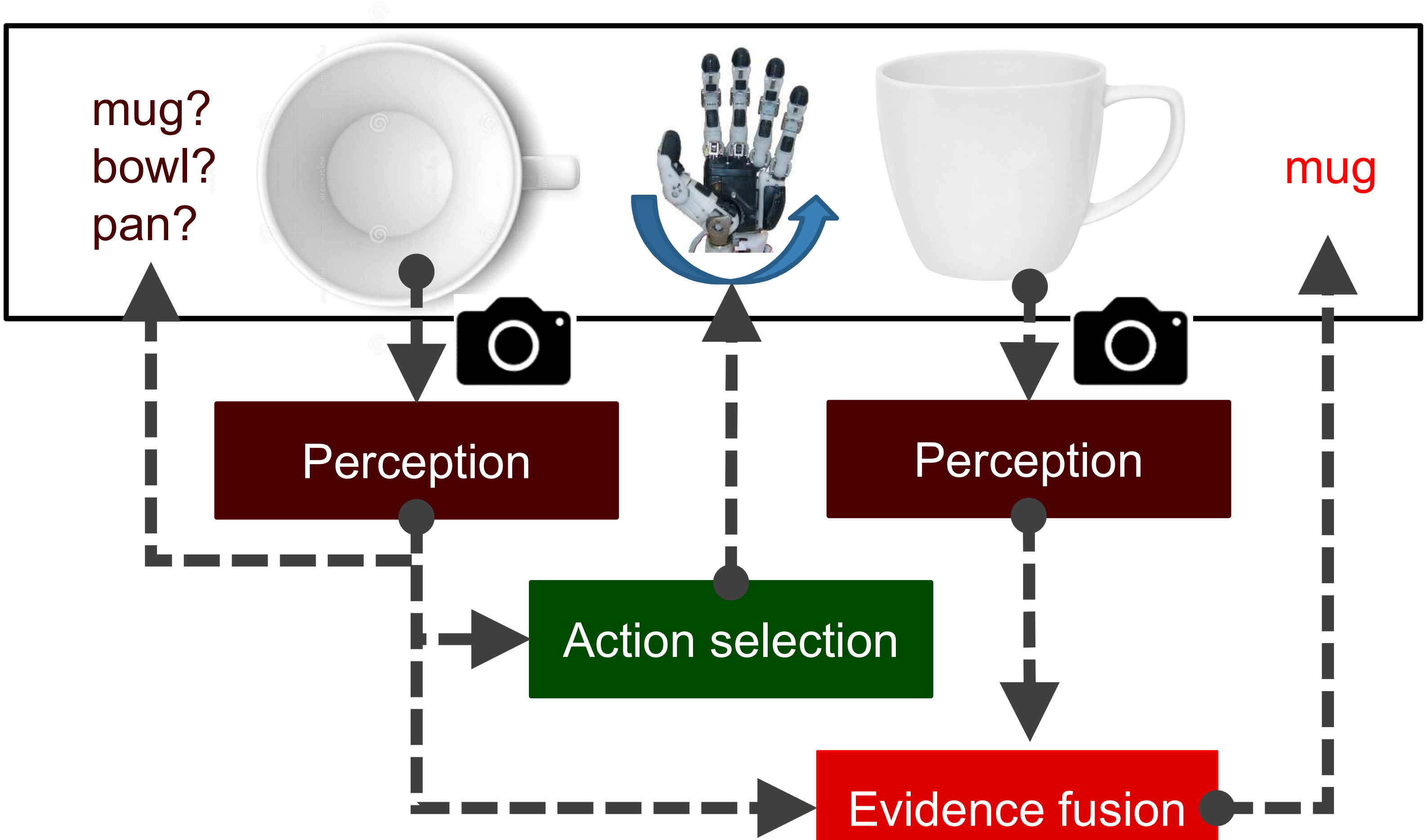
Difficulty: unconstrained visual input



Opportunity: ability to move to *change* input



Components of active recognition



Our idea: Multi-task training of active recognition components + look-ahead.

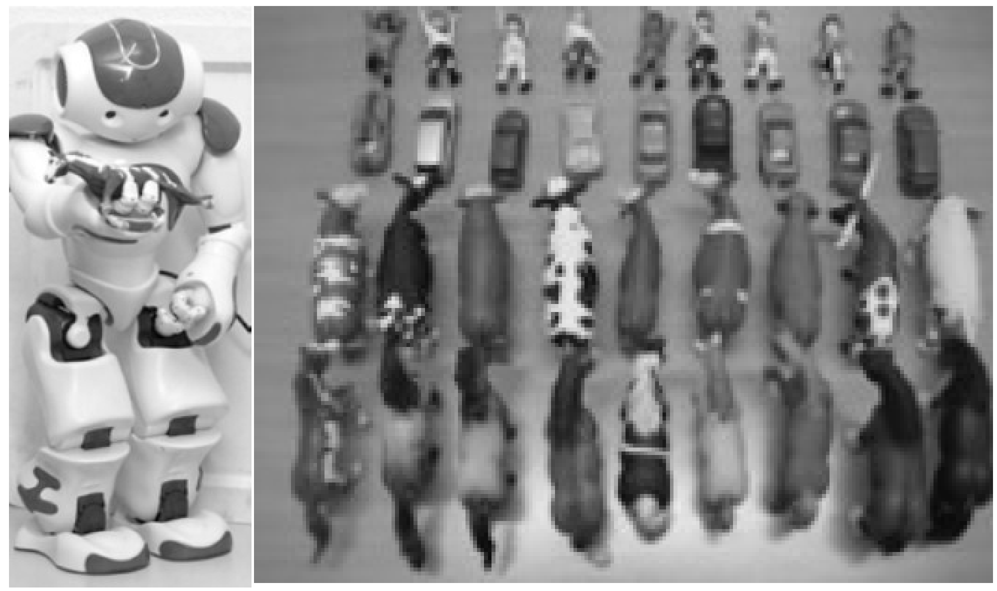
Jayaraman and Grauman, ECCV 2016

Experiments

How to **evaluate** active recognition?

Instances, turntables

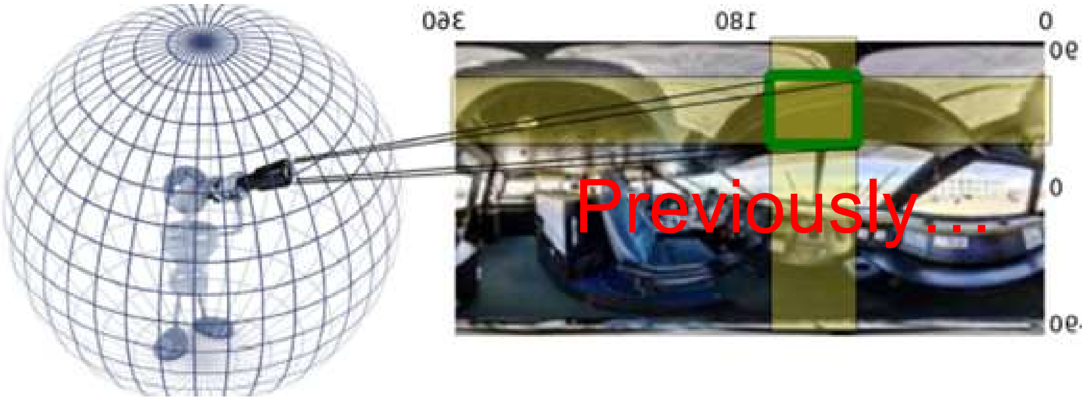
Custom robot setting



[Nene 1996, Schiele 1998, Denzler 2003, Ramanathan 2011…]

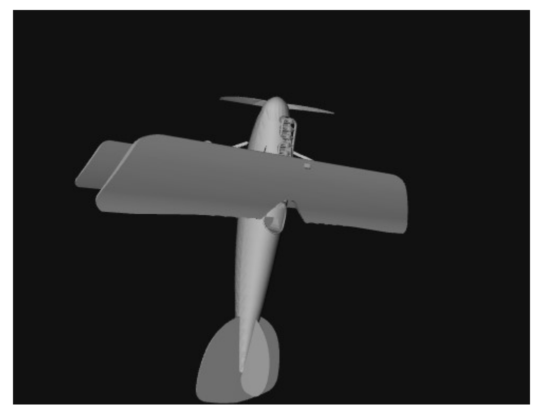
Experiments

SUN 360 panoramas [Xiao 2012]



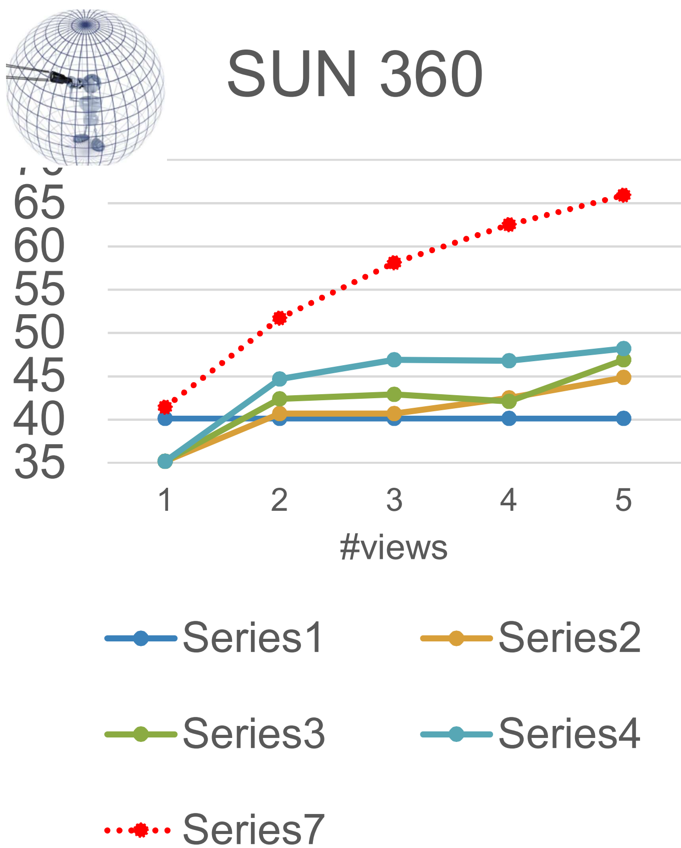
GERMS toy manipulation [Malmir 2015]

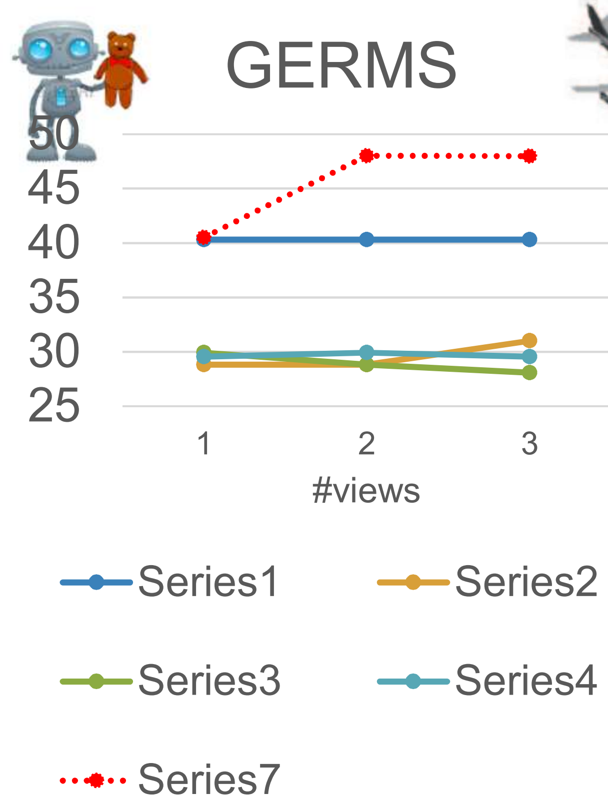
ModelNet-10 CAD models [Wu 2015]

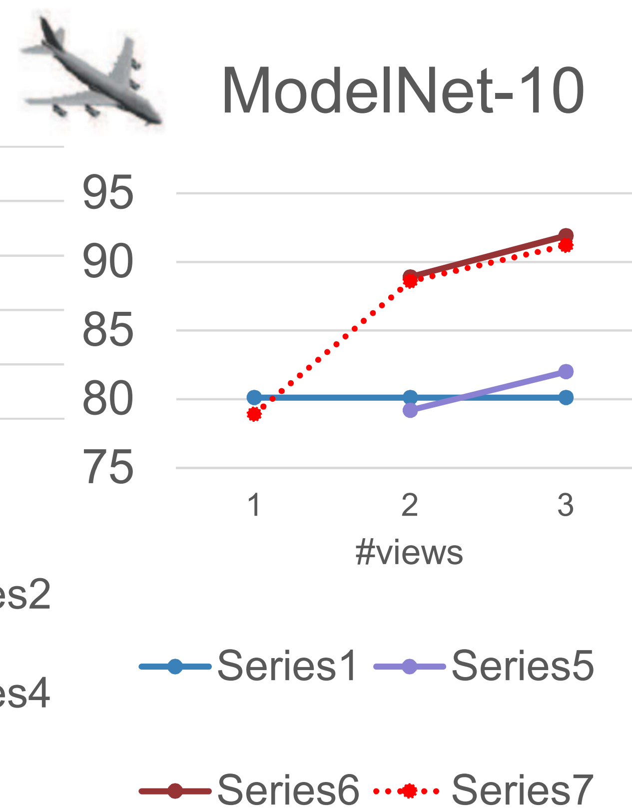


Jayaraman and Grauman, ECCV 2016

End-to-end active recognition: results







Strongly outperform traditional active recognition approaches.

Jayaraman and Grauman, ECCV 2016

End-to-end active recognition: example

Top 3 guesses:		(1.05) / (5.90)	(37.88) / (63.89)
	Restaurant / Forest	Theater / Street	Plaza courtyard / Church
	Train interior / Cave	Restaurant / Cave	Lobby atrium / Street
	Beach / Shop	Plaza courtyard	Theater / Street

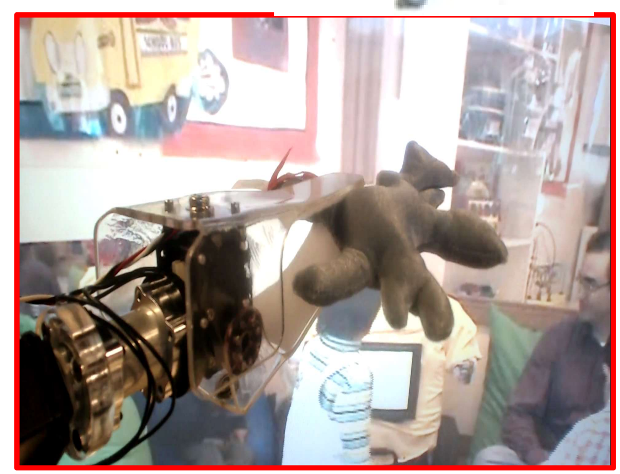
[Jayaraman and Grauman, ECCV 2016]

End-to-end active recognition: example

Predicted label:

T=1

T=2

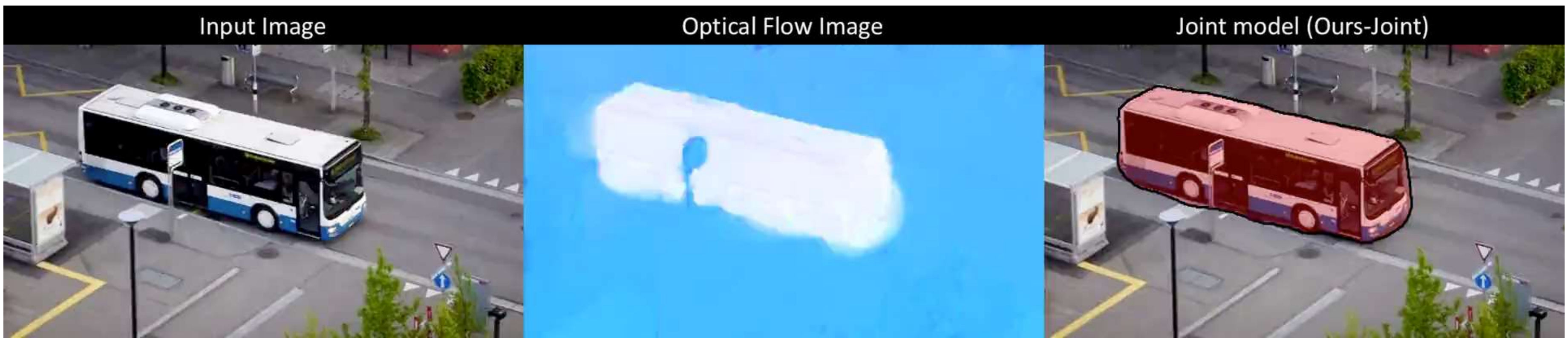


T=3

GERMS dataset: Malmir et al. BMVC 2015

[Jayaraman and Grauman, ECCV 2016]

FusionSeg: Pulling objects out of video

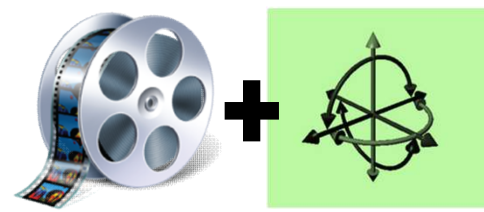


h

Talk overview

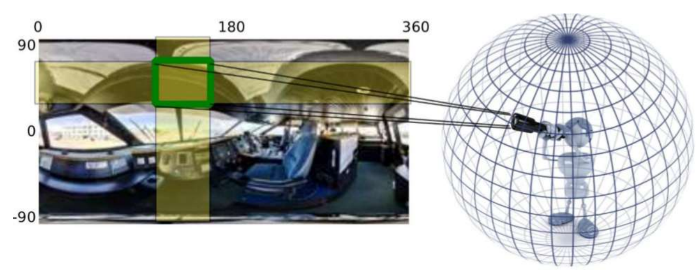
Towards embodied visual learning

1. Learning representations tied to ego-motion



2. Learning representations from unlabeled video

3. Learning how to move and where to look

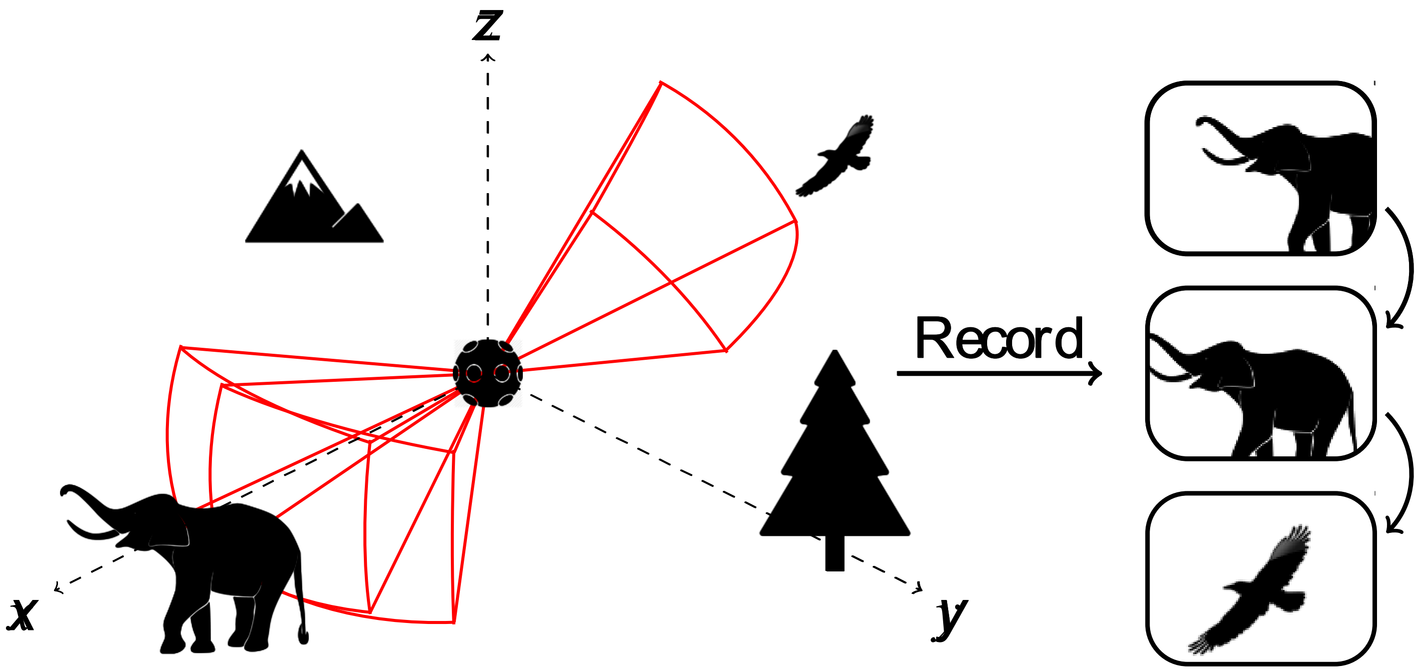


Challenge of viewing 360° videos

Control by mouse

How to find the right direction to watch?

New problem:
Pano2Vid automatic videography



Pano2Vid Definition

Input: 360° video

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

Task: control the virtual camera direction

[Su et al. ACCV 2016, Su & Grauman CVPR 2017]

Our approach – AutoCam

Learn videography tendencies from unlabeled Web videos

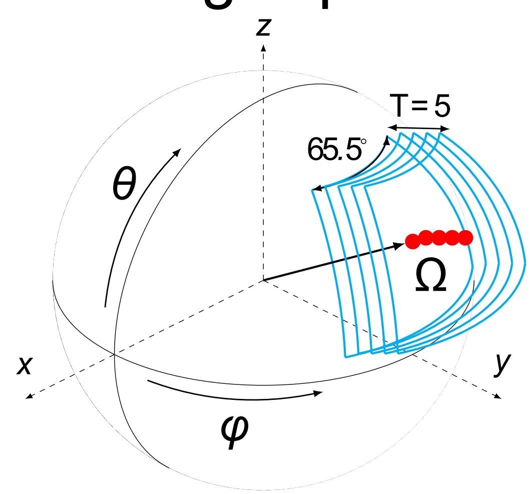
- Diverse capture-worthy content
- Proper composition

Human-captured NFOV videos (“HumanCam”)

Unlabeled video

How close?

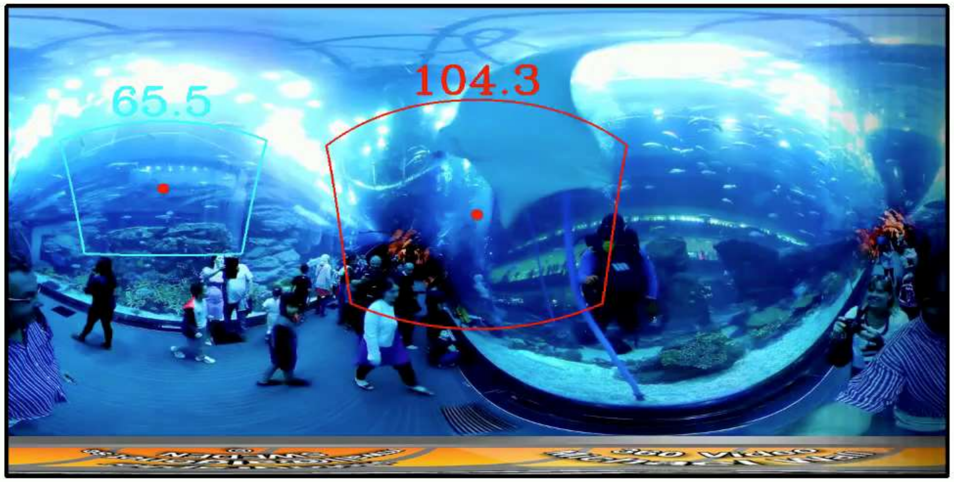
ST-glimpses



[Su et al. ACCV 2016, Su & Grauman CVPR 2017]

Example AutoCam Output 2

Input 360° Video
+
Camera Trajectories



AutoCam
Output Video

With
Zooming

Without
Zooming

[Su & Grauman CVPR 2017]

Summary

- Visual learning benefits from
  - context of action and motion in the world
  - continuous unsupervised observations
- New ideas:
  - "Embodied" feature learning via visual and motor signals
  - Feature learning from unlabeled video via higher order temporal coherence
  - Active policies for view selection and camera control

Dinesh Jayaraman

Yu-Chuan Su

Ruohan Gao

Code and pre-trained models available

`http://www.cs.utexas.edu/~grauman/research/pubs.html`

Relevant papers

- **Making 360 Video Watchable in 2D: Learning Videography for Click Free Viewing**. Y-C. Su and K. Grauman. To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, July 2017.

- **Learning Image Representations Tied to Egomotion from Unlabeled Video**. D. Jayaraman and K. Grauman. International Journal of Computer Vision (IJCV), Special Issue for Best Papers of ICCV 2015, 2017.

- **Pano2Vid: Automatic Cinematography for Watching 360◦ Videos**. Y-C. Su, D. Jayaraman, and K. Grauman. Proceedings of the Asian Conference on Computer Vision (ACCV), Taipei, November 2016.

- **FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos**, S. Jain, B. Xiong, K. Grauman, CVPR 2017

- **Look-Ahead Before You Leap: End-to-End Active Recognition by Forecasting the Effect of Motion**. D. Jayaraman and K. Grauman. Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, October 2016.

- **Slow and Steady Feature Analysis: Higher Order Temporal Coherence in Video.** D. Jayaraman and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, June 2016.

- **Learning Image Representations Tied to Ego-Motion**. D. Jayaraman and K. Grauman. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec 2015.