CVPR 2016 Fourth Workshop on Egocentric (First-Person) Vision

Egomotion and Visual Learning

Kristen Grauman

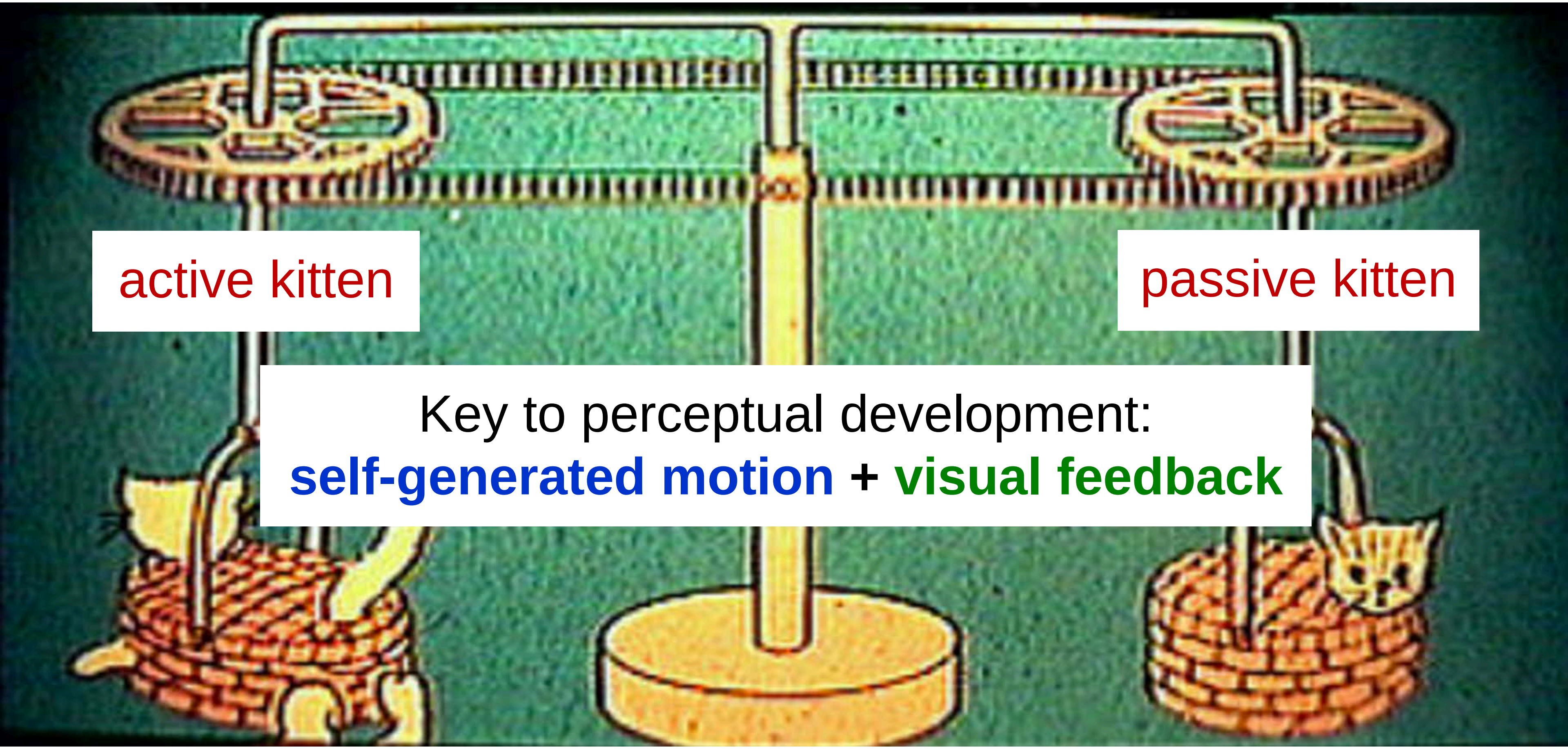
Department of Computer Science

University of Texas at Austin

THE UNIVERSITY OF TEXAS AT AUSTIN

Dinesh Jayaraman

The kitten carousel experiment
[Held & Hein, 1963]



Big picture goal: Embodied vision

Status quo:

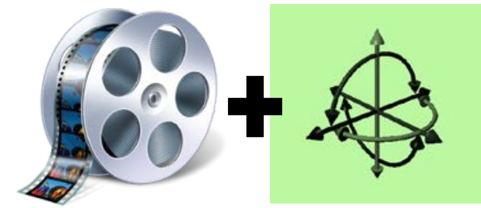
Learn from “disembodied” bag of labeled snapshots.

Our goal:

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

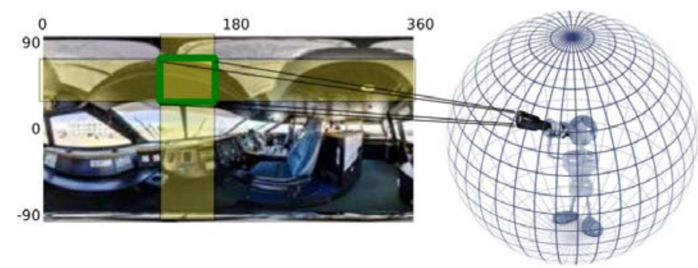
Talk overview

1. Learning representations tied to ego-motion



2. Learning representations from unlabeled video

3. Learning how to move and where to look



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**

[Jayarman & Grauman, ICCV 2015]

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]

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:

Ego-motion ↔ vision for recognition

Learning this connection requires:

- Depth, 3D geometry
- Semantics
- Context

} Also key to recognition!

Can be learned without manual labels!

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

[Jayaraman & Grauman, ICCV 2015]

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"

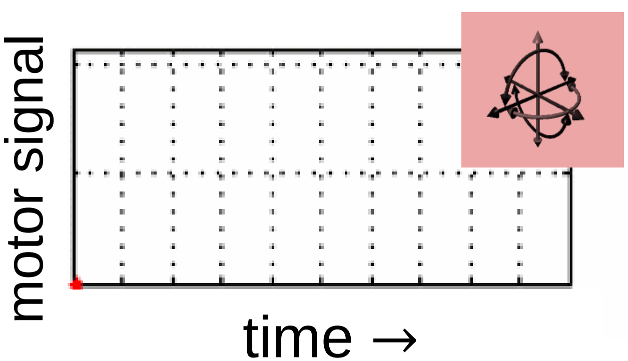
$$\mathbf{z}(g\mathbf{x}) \approx M_g \mathbf{z}(\mathbf{x})$$

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

Approach idea: Ego-motion equivariance

Training data

Unlabeled video +
motor signals



Learn

Equivariant embedding

organized by ego-motions

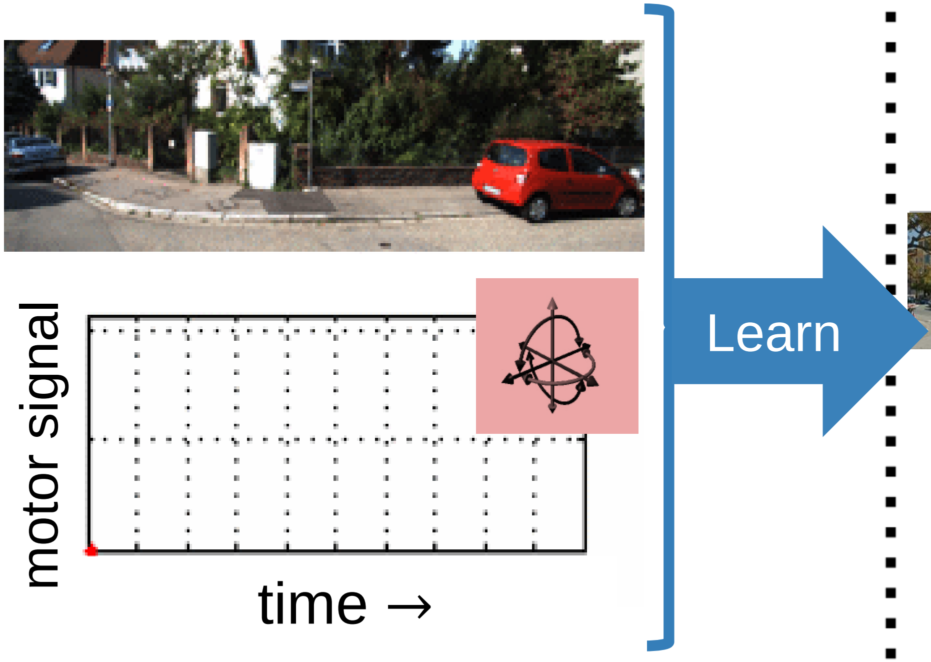
Pairs of frames related by similar ego-motion should be related by same feature transformation

[Jayaraman & Grauman, ICCV 2015]

Approach idea: Ego-motion equivariance

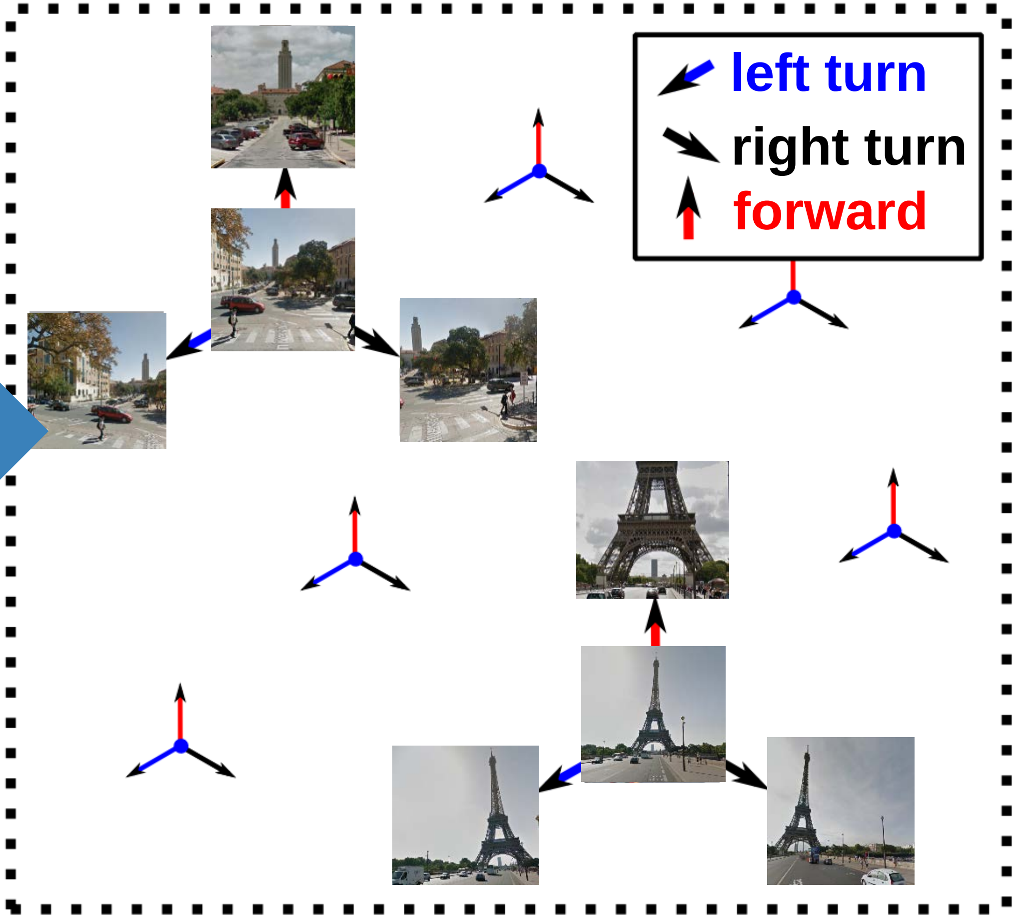
Training data

Unlabeled video +
motor signals



Equivariant embedding

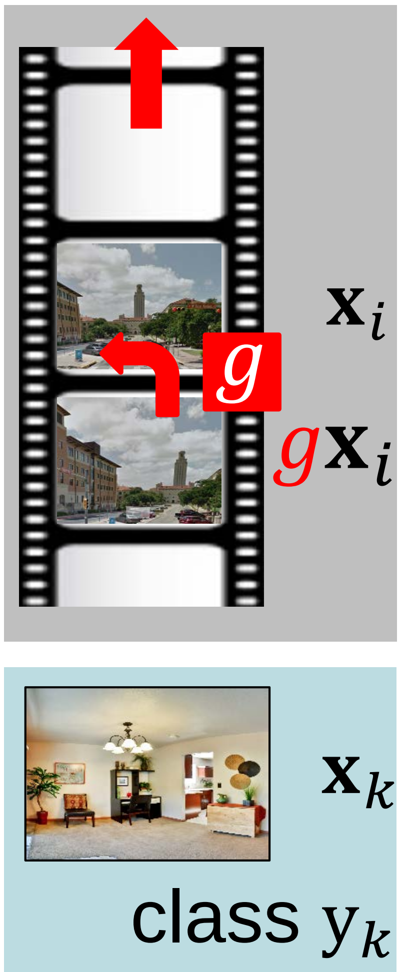
organized by ego-motions



[Jayaraman & Grauman, ICCV 2015]

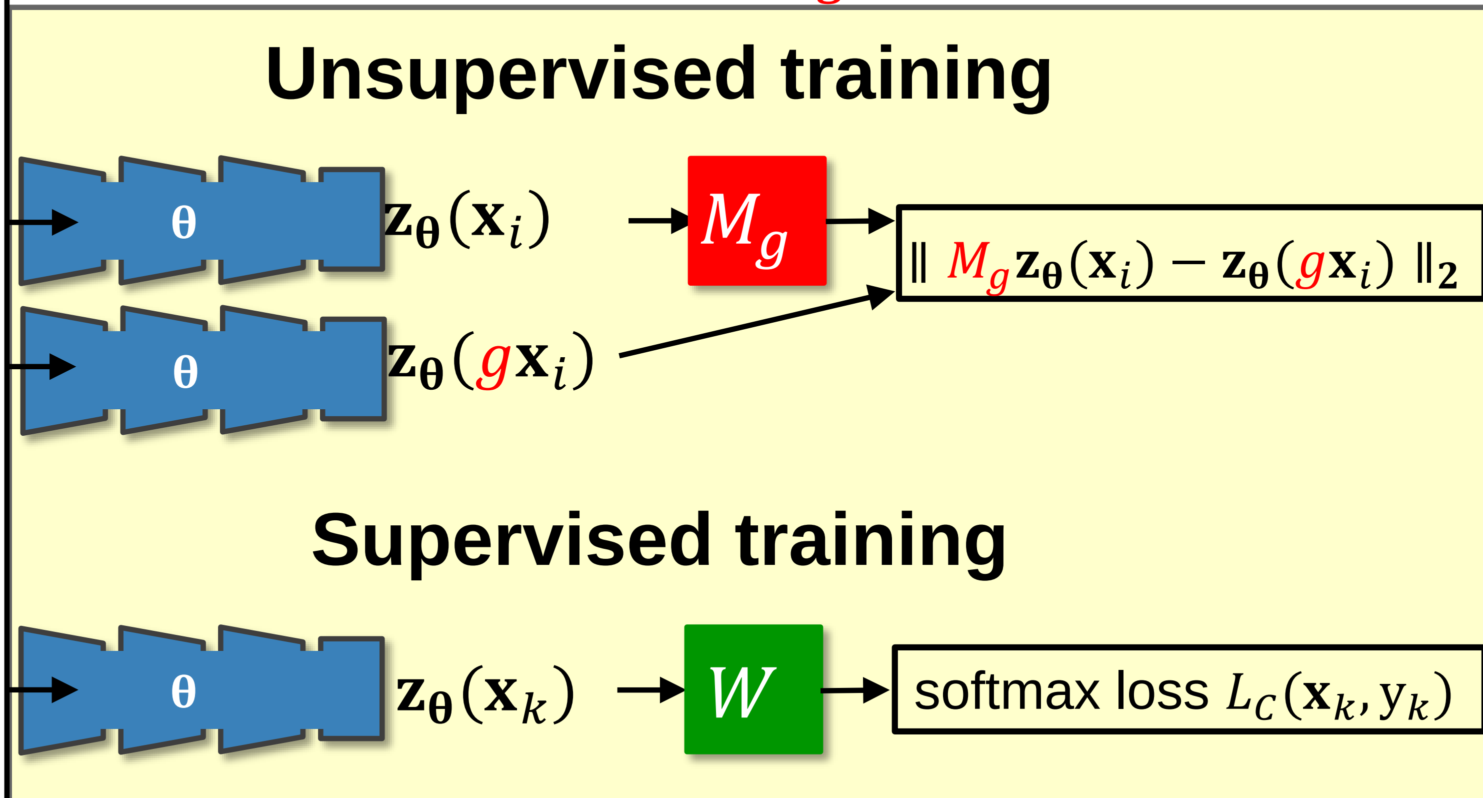
Ego-motion equivariant feature learning

Given:



Desired: for all motions g and all images $\mathbf{x}$,

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



θ, M_g and W jointly trained

[Jayaraman & Grauman, ICCV 2015]

Results: Recognition

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

Geiger et al, IJRR ’13

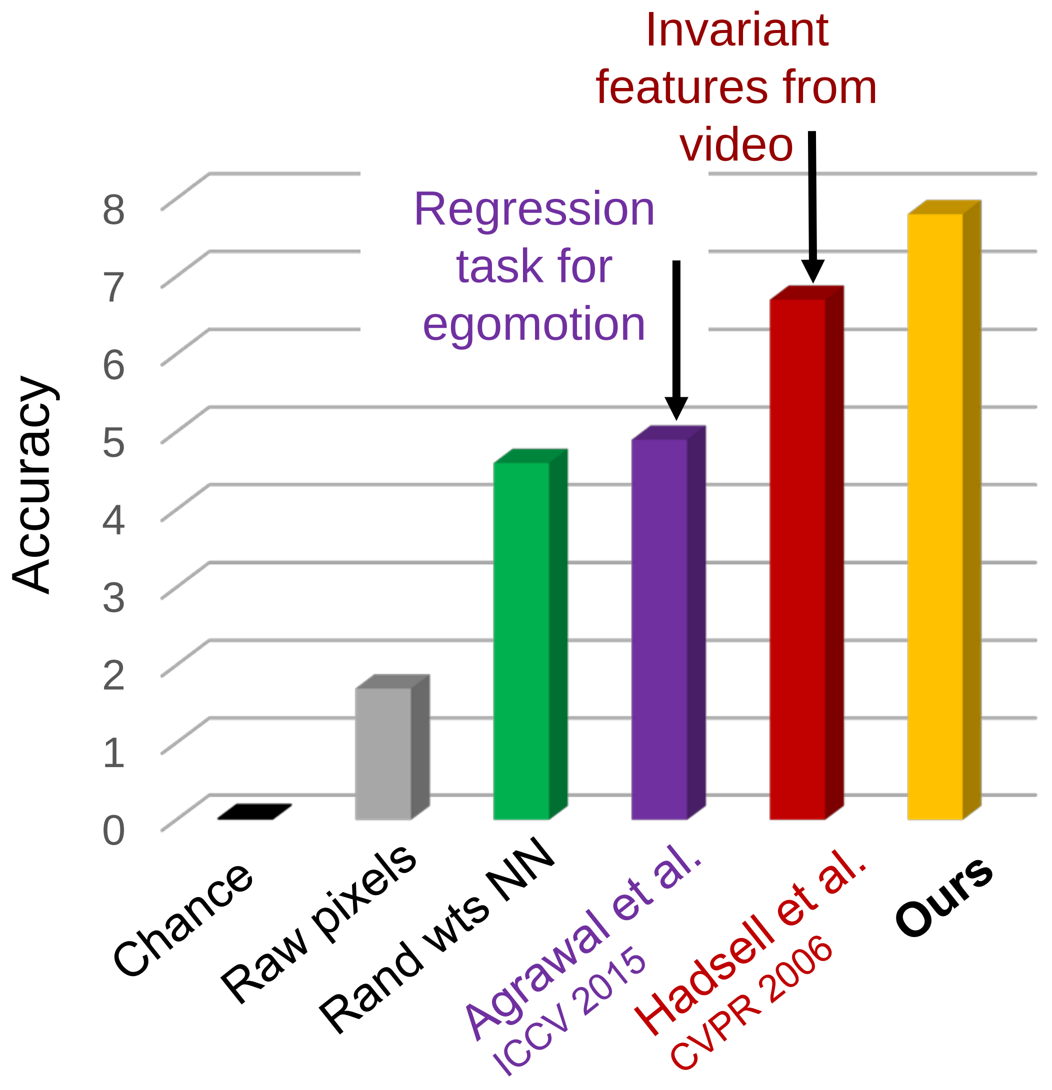
Exploit features for **static scene classification**
(SUN, 397 classes)

Apse
Window seat
Art school
Library
Auditorium
Bus interior
Cathedral
Freeway
Guardhouse

Xiao et al, CVPR ’10

Results: Recognition

- Purely unsupervised feature learning
- *k*-nearest neighbor scene classification task in learned feature space
    - Unlabeled video: KITTI
    - Images: SUN, 397 classes
    - 50 labels per class

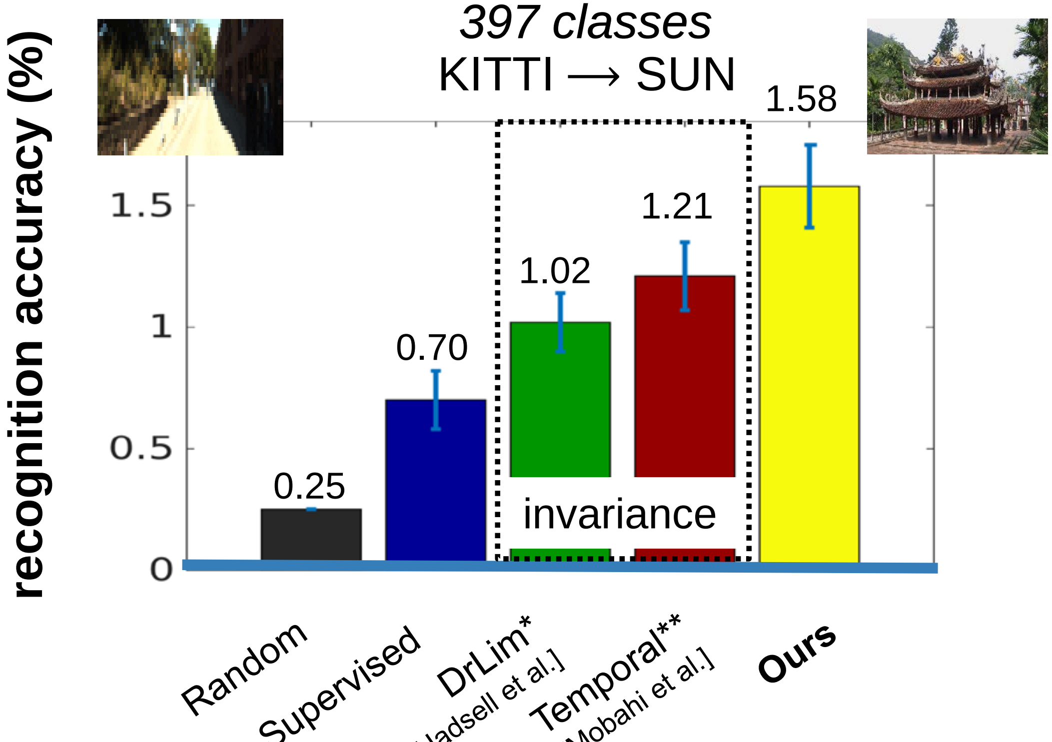


Agrawal, Carreira, Malik, Learning to see by moving. ICCV 2015

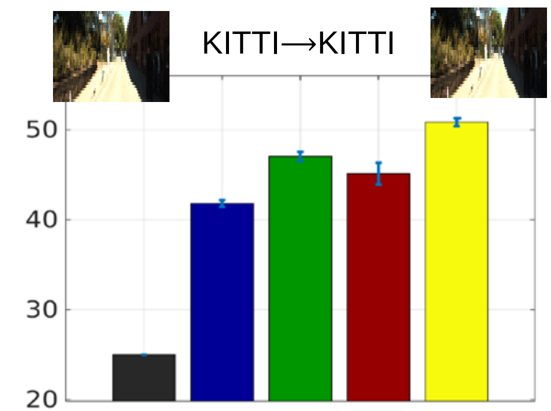
Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping. CVPR 2006

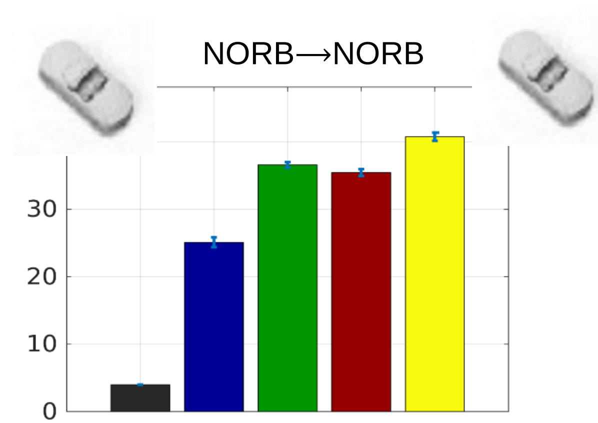
Results: Recognition

Ego-motion equivariance as a regularizer



6 labeled training examples per class





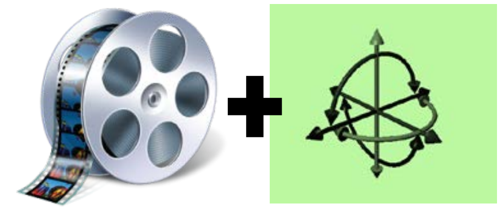
Up to 30% accuracy increase over state of the art!

*Hadsell et al., Dimensionality Reduction by Learning an Invaria

**Mobahi et al., Deep Learning from Temporal Coherence in Video, ICML'09

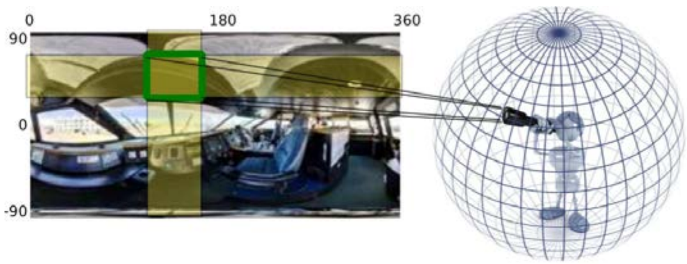
Talk overview

1. Learning representations tied to ego-motion

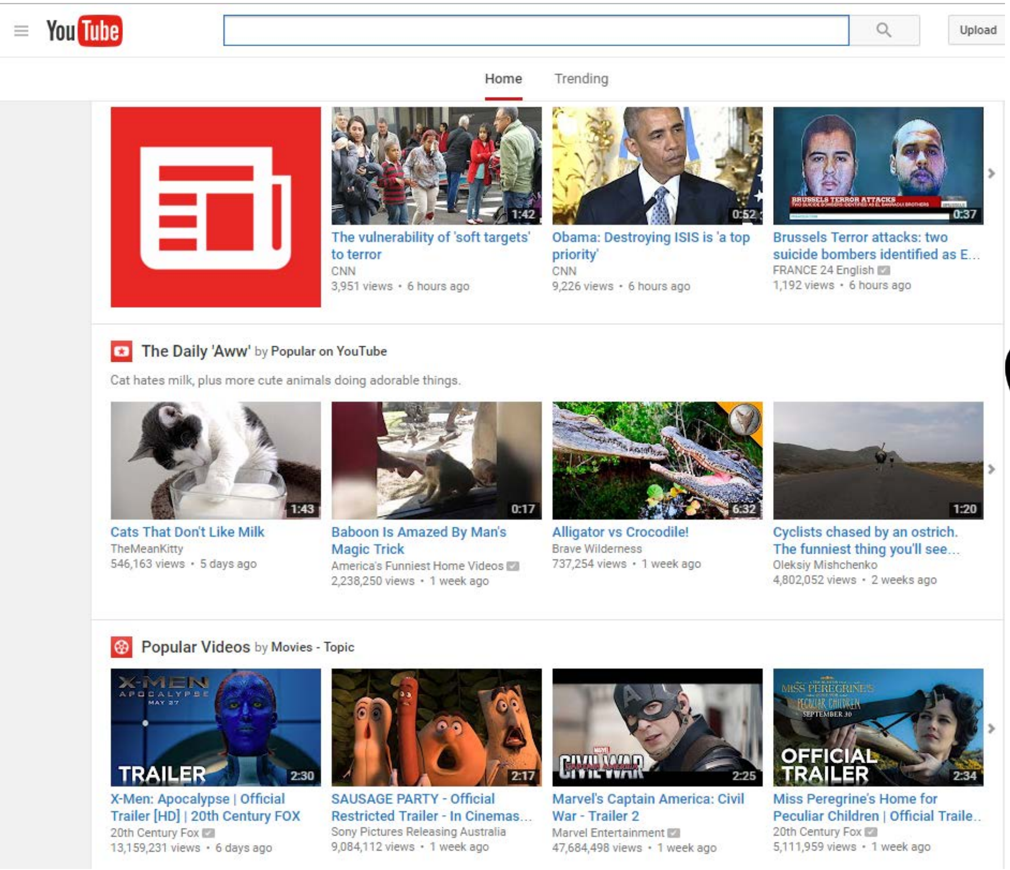


2. Learning representations from unlabeled video

3. Learning how to move and where to look



Learning from arbitrary unlabeled video?

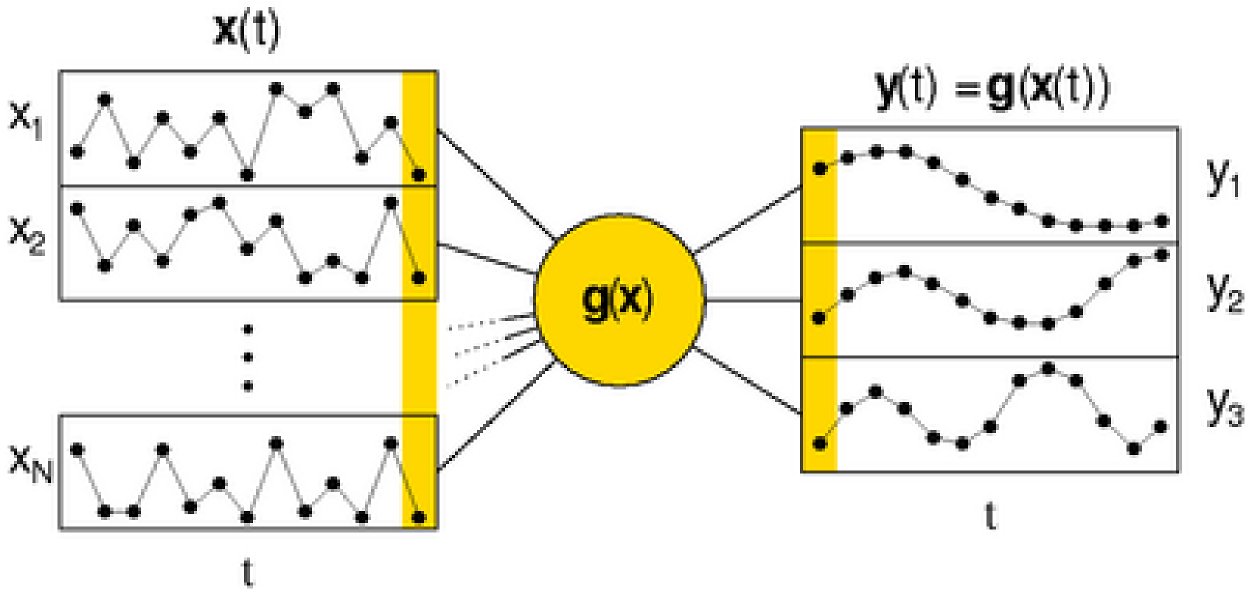


Unlabeled video

Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]

Find functions **g(x)** that map



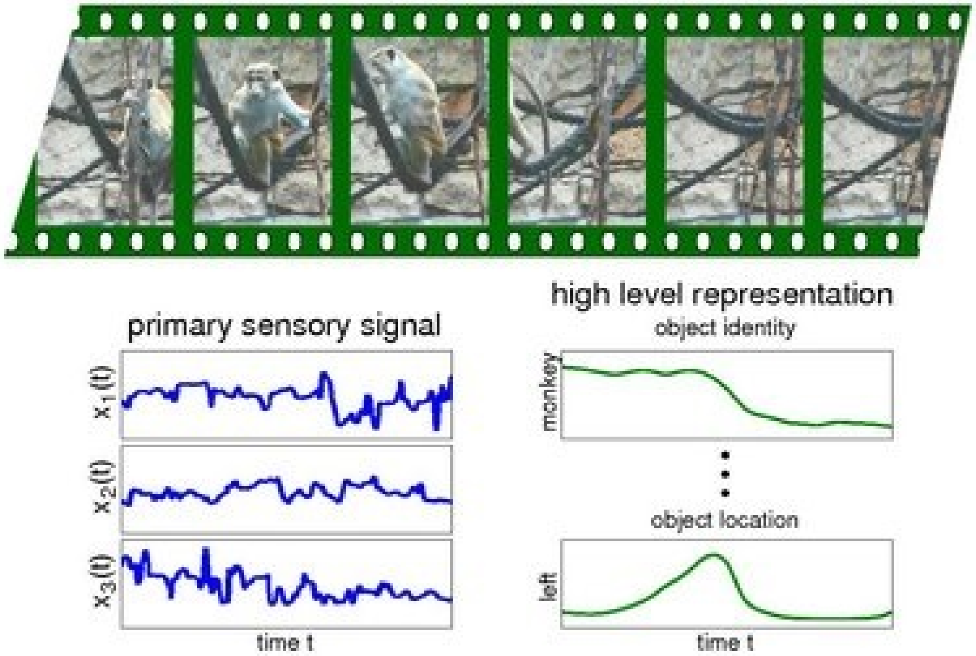
quickly varying input signal **x(**t**)** → slowly varying features **y(**t**)**

Figure: Laurenz Wiskott, http://www.scholarpedia.org/article/File:SlowFeatureAnalysis-OptimizationProblem.png

Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]

Find functions **g(x)** that map

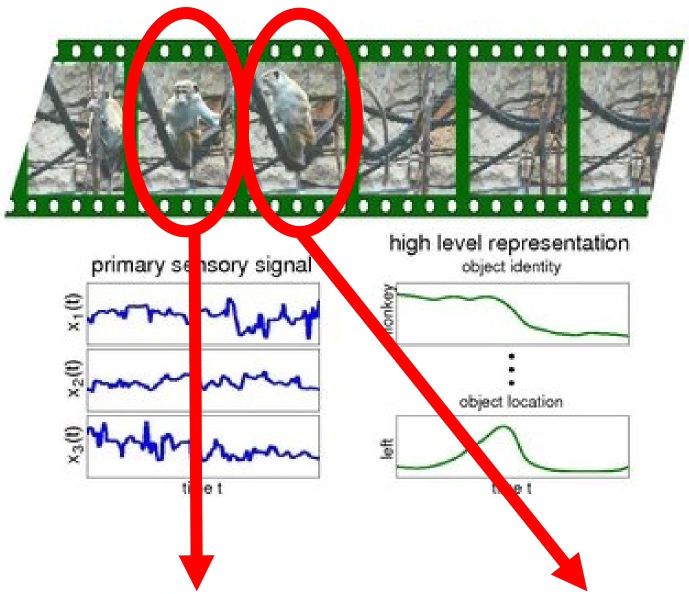


quickly varying input signal **x(**t**)** → slowly varying features **y(**t**)**

Figure: Laurenz Wiskott, http://www.scholarpedia.org/article/File:SlowFeatureAnalysis-OptimizationProblem.png

Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]



$$\mathbf{z}(\boldsymbol{a}) \approx \mathbf{z}(\boldsymbol{b})$$

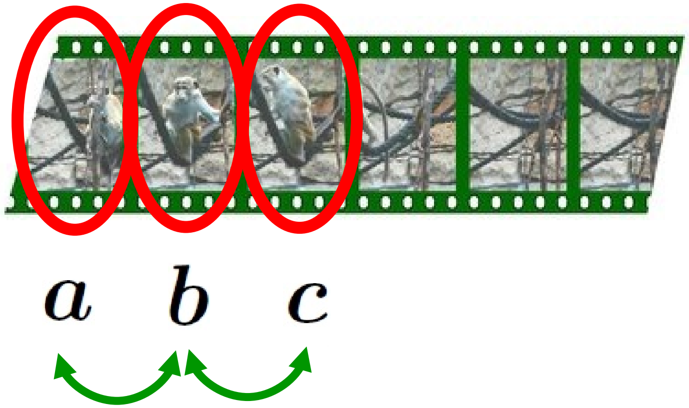
in learned embedding

- Existing work exploits "slowness" as temporal coherence in video → learn invariant representation

  [Hadsell et al. 2006; Mobahi et al. 2009; Bergstra & Bengio 2009; Goroshin et al. 2013; Wang & Gupta 2015,…]

- Fails to capture *how* visual content changes over time

Our idea: Steady feature analysis



- Higher order temporal coherence in video → learn equivariant representation

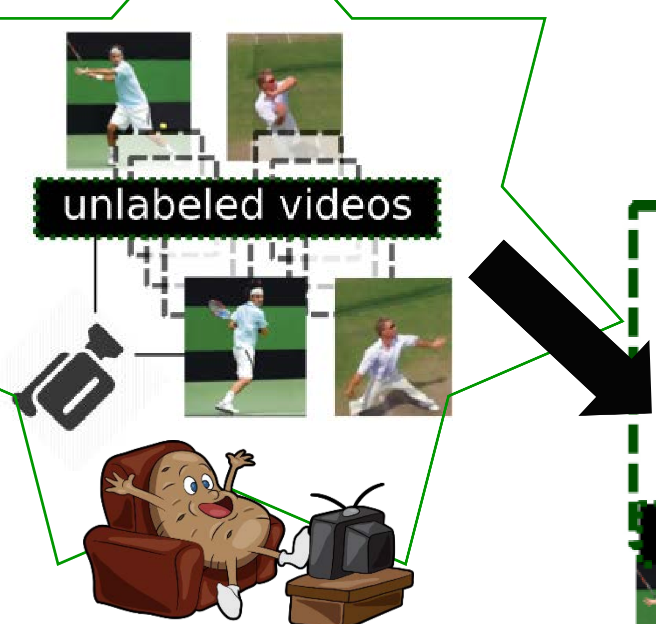
Second order slowness operates on frame triplets:

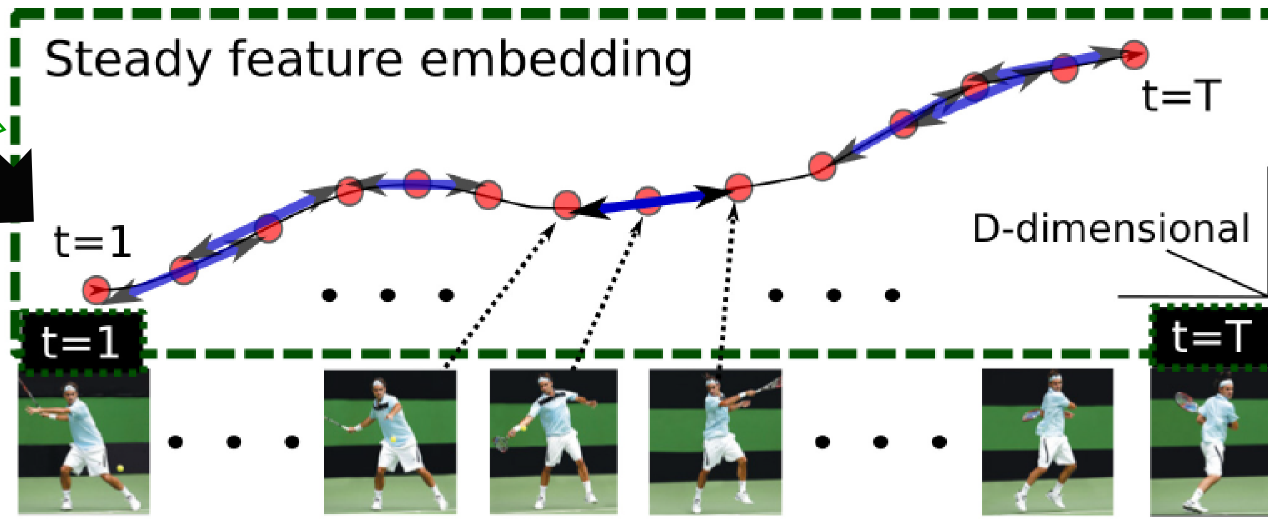
$$\mathbf{z}(b) - \mathbf{z}(a) \approx \mathbf{z}(c) - \mathbf{z}(b)$$

in learned embedding

[Jayaraman & Grauman, CVPR 2016]

Our idea: Steady feature analysis





Equivariance ≈ “steadily” varying frame features!

$$d^2\mathbf{z}_{\boldsymbol{\theta}}(\mathbf{x}t)/dt^2 \approx \mathbf{0}$$

[Jayaraman & Grauman, CVPR 2016]

Datasets

Unlabeled video

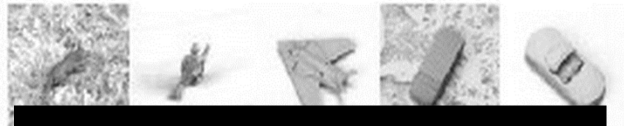
Target task (few labels)

Human Motion Database (HMDB)

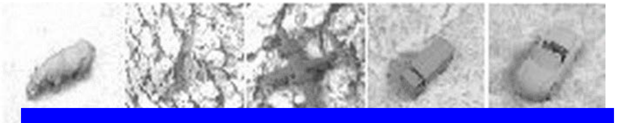
PASCAL 10 Actions

KITTI Video

SUN 397 Scenes



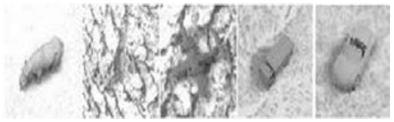
NORB



NORB 25 Objects

32 x 32 images or 96 x 96 images

Results: Steady feature analysis



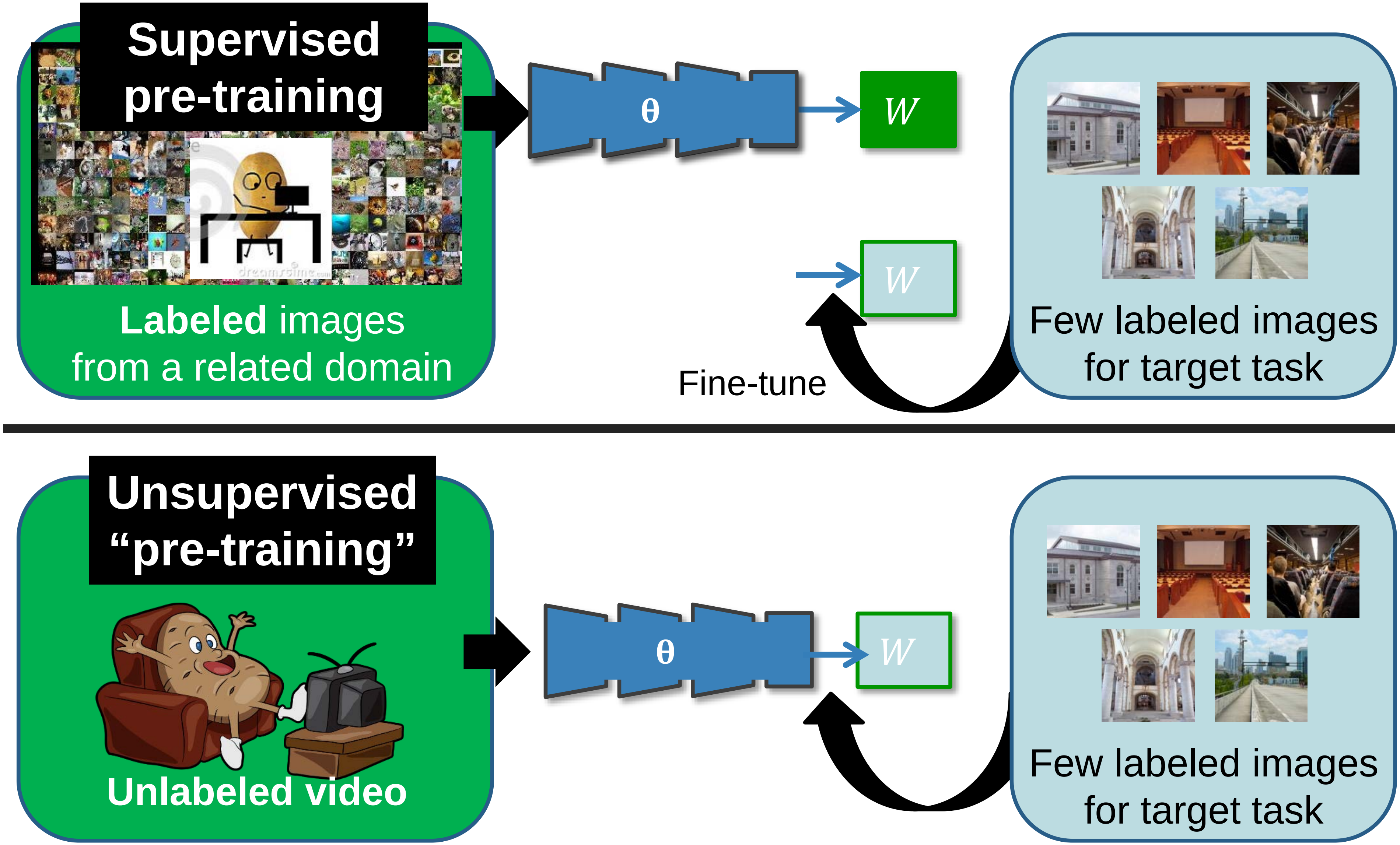
Task type→	Objects	Scenes		Actions
Datasets→	NORB→NORB	KITTI→SUN		HMDB→PASCAL-10
Methods↓	[25 cls]	[397 cls]	[397 cls, top-10]	[10 cls]
random	4.00	0.25	2.52	10.00
UNREG	24.64±0.85	0.70±0.12	6.10±0.67	15.34±0.28
SFA-1 [30]*	37.57±0.85	1.21±0.14	8.24±0.25	19.26±0.45
SFA-2 [14]**	39.23±0.94	1.02±0.12	6.78±0.32	19.04±0.24
SSFA (ours)	**42.83±0.33**	**1.65±0.04**	**9.19±0.10**	**20.95±0.13**

Multi-class recognition accuracy

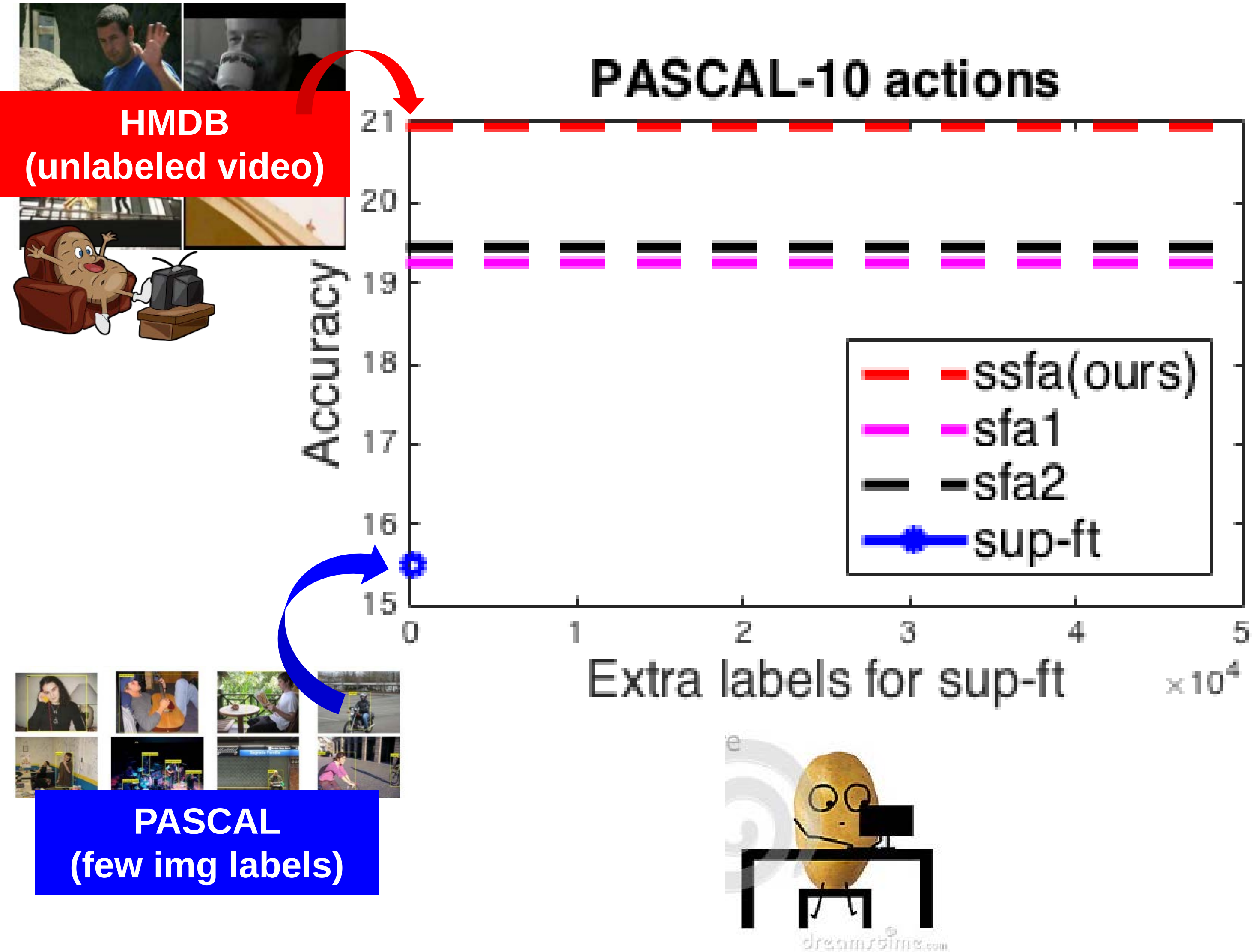
*Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR'06

**Mobahi et al., Deep Learning from Temporal Coherence in Video, ICML'09

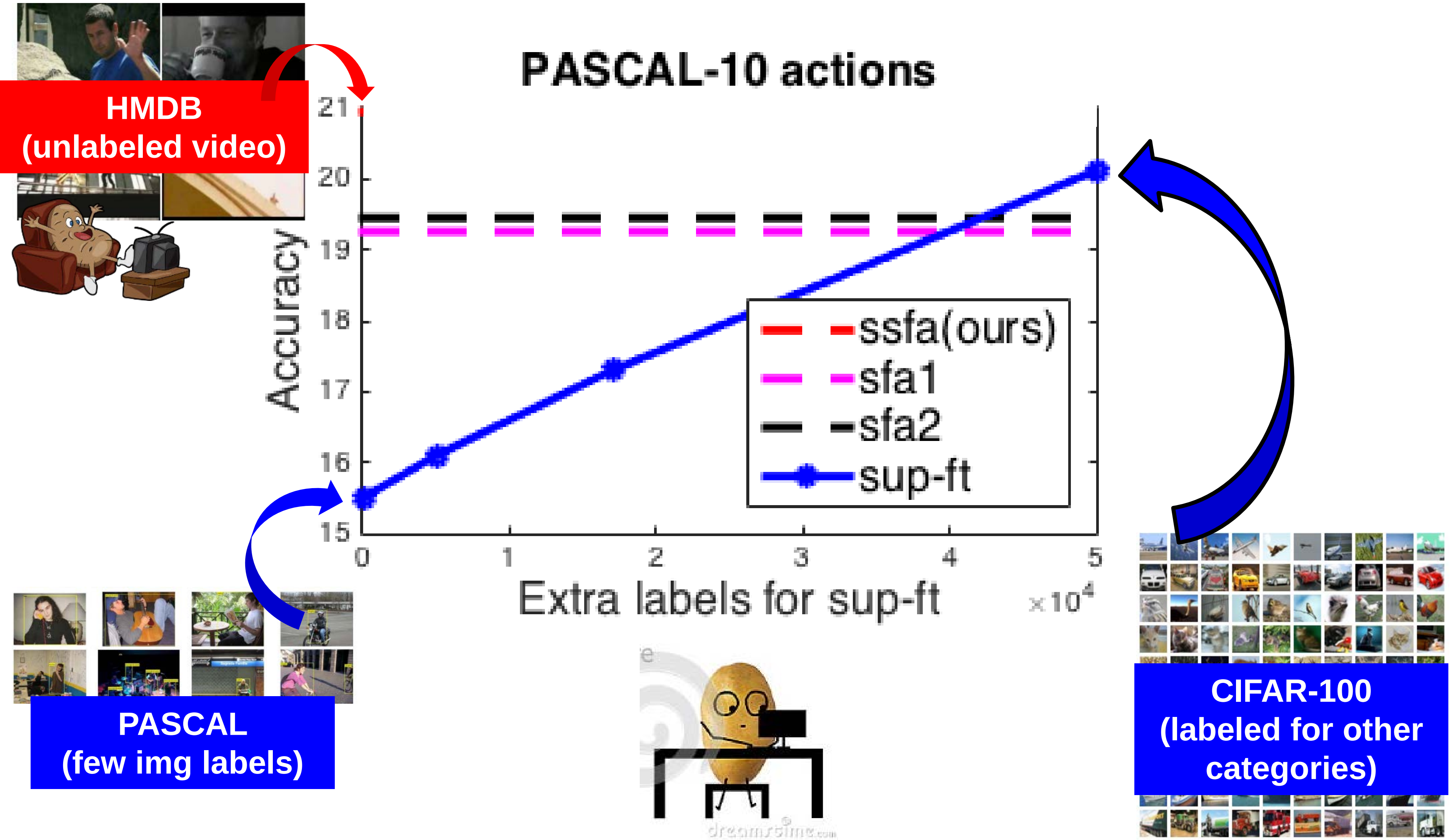
Pre-training a representation



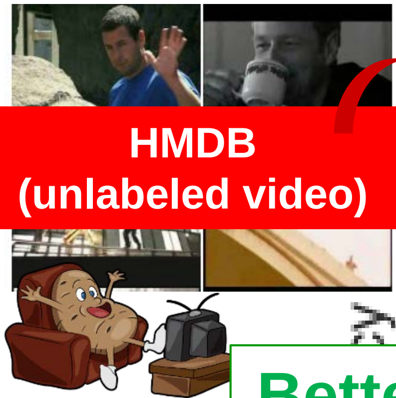
Results: Can we learn *more* from unlabeled video than “related” labeled images?



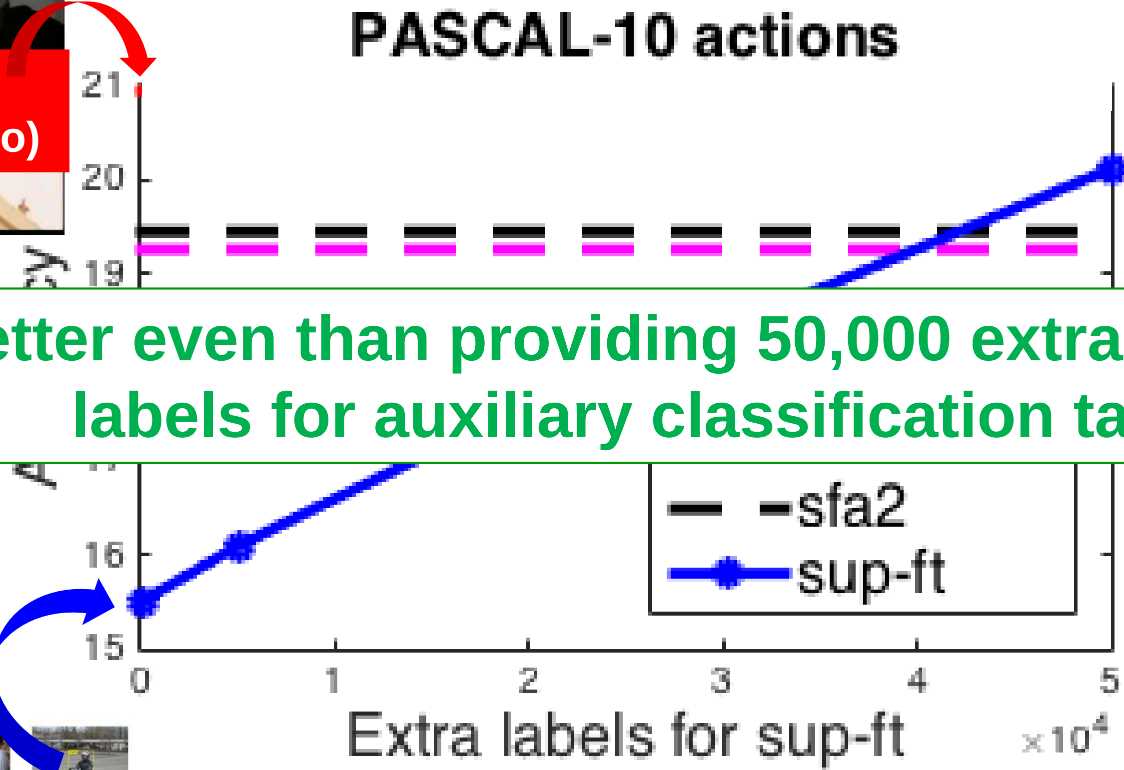
Results: Can we learn *more* from unlabeled video than "related" labeled images?



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



HMDB (unlabeled video)



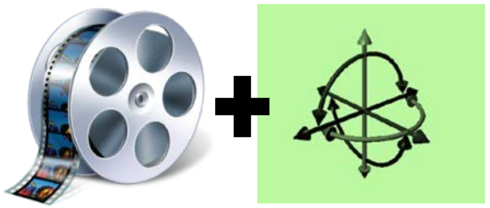
Better even than providing 50,000 extra manual labels for auxiliary classification task!

PASCAL (few img labels)

CIFAR-100 (labeled for other categories)

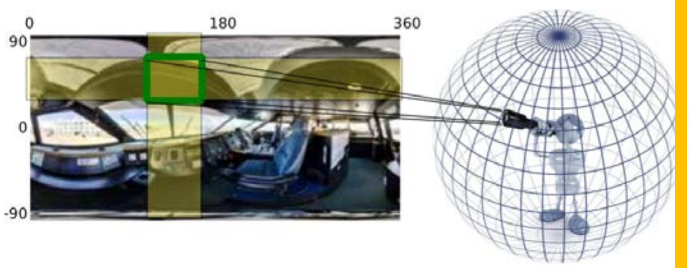
Talk overview

1. Learning representations tied to ego-motion



2. Learning representations from unlabeled video

3. Learning how to move and where to look



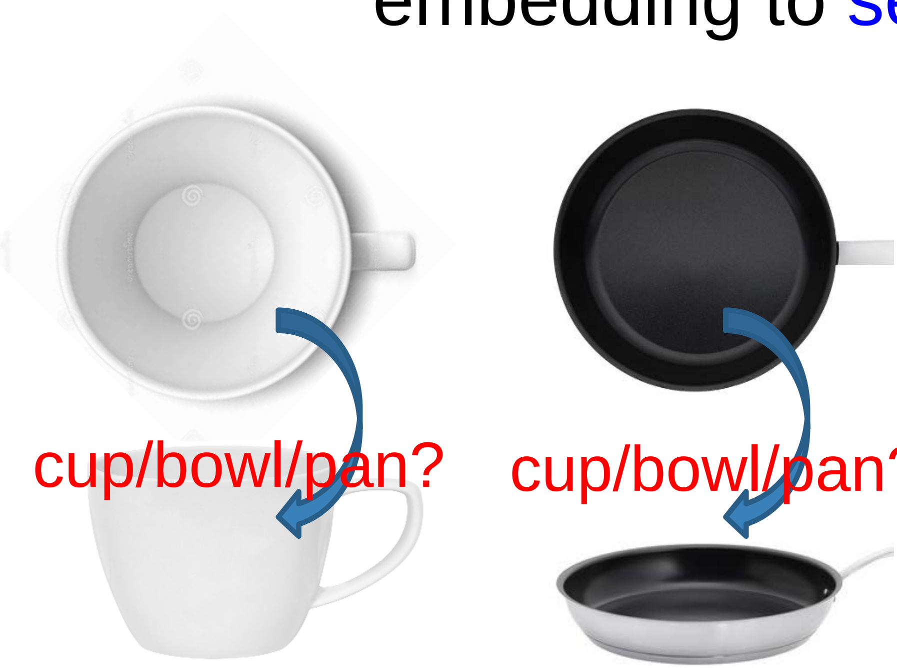
Learning how to move for recognition

Time to revisit active recognition in challenging settings!

[Bajcsy 1985, Schiele & Crowley 1998, Dickinson et al. 1997, Tsotsos et al. 2001, Soatto 2009,…]

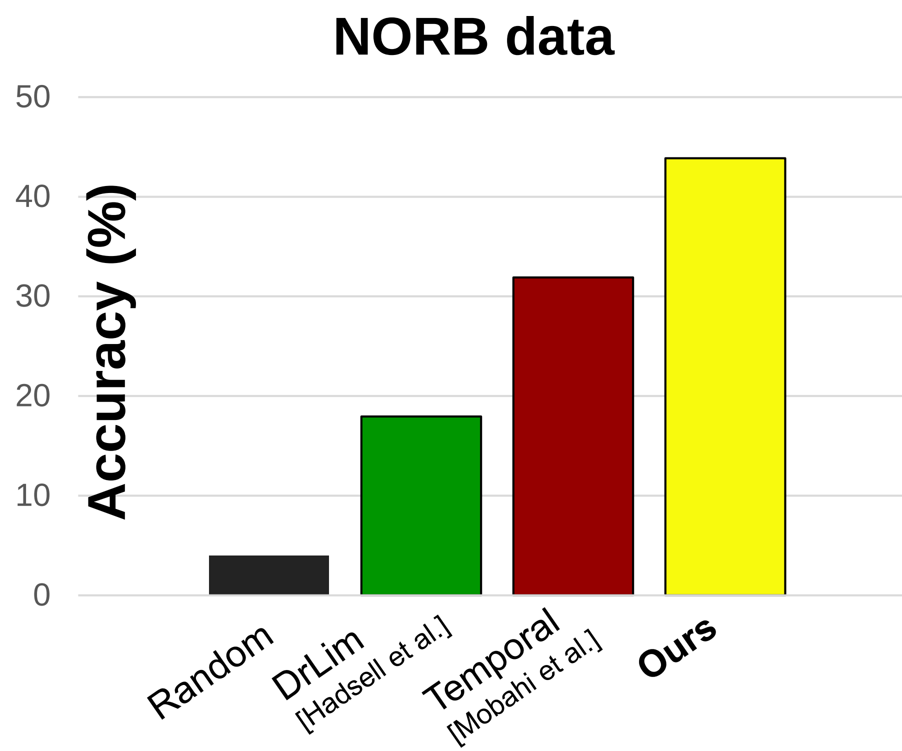
Learning how to move for recognition

Leverage proposed ego-motion equivariant embedding to select next best view



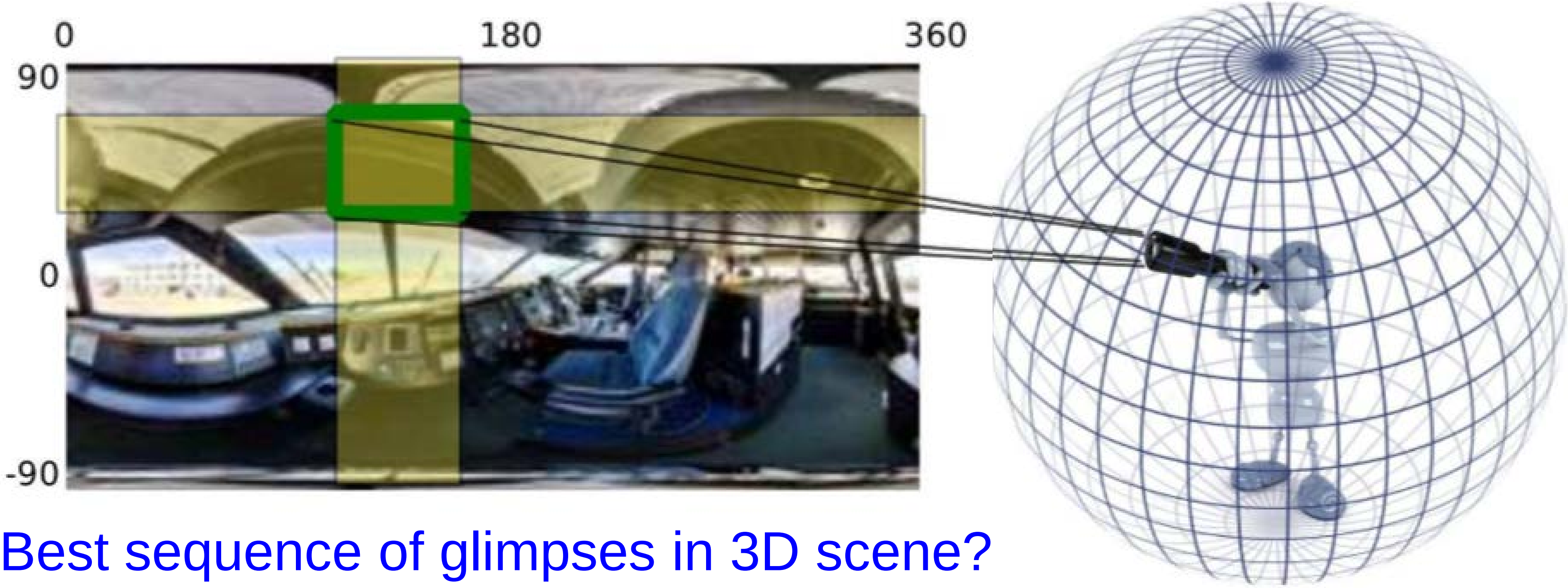
cup

frying pan



[Jayaraman & Grauman, ICCV 2015]

Learning how to move for recognition



Best sequence of glimpses in 3D scene?

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

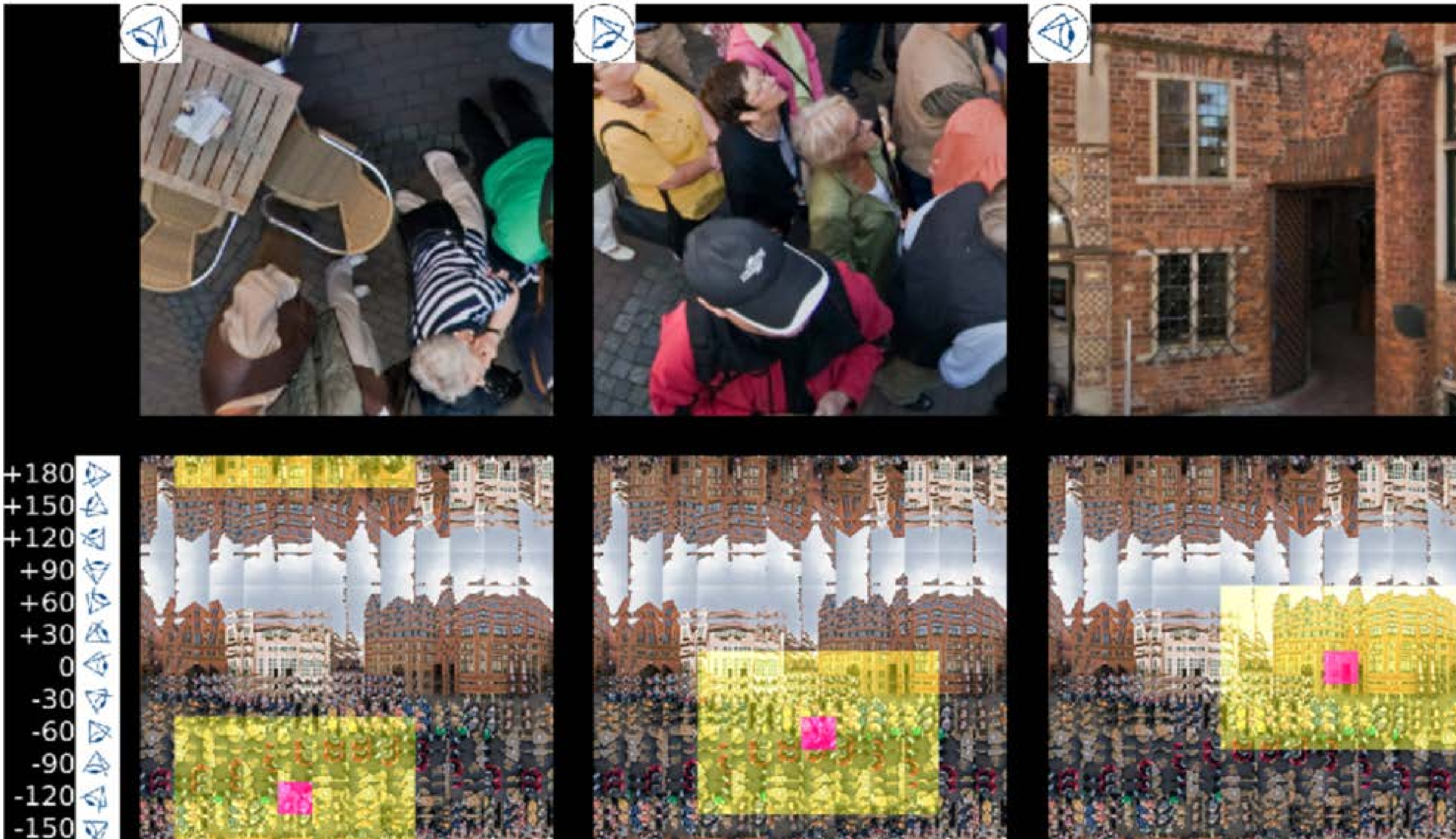
Learn all end-to-end

Jayaraman and Grauman, UT TR AI15-06

Active recognition: results

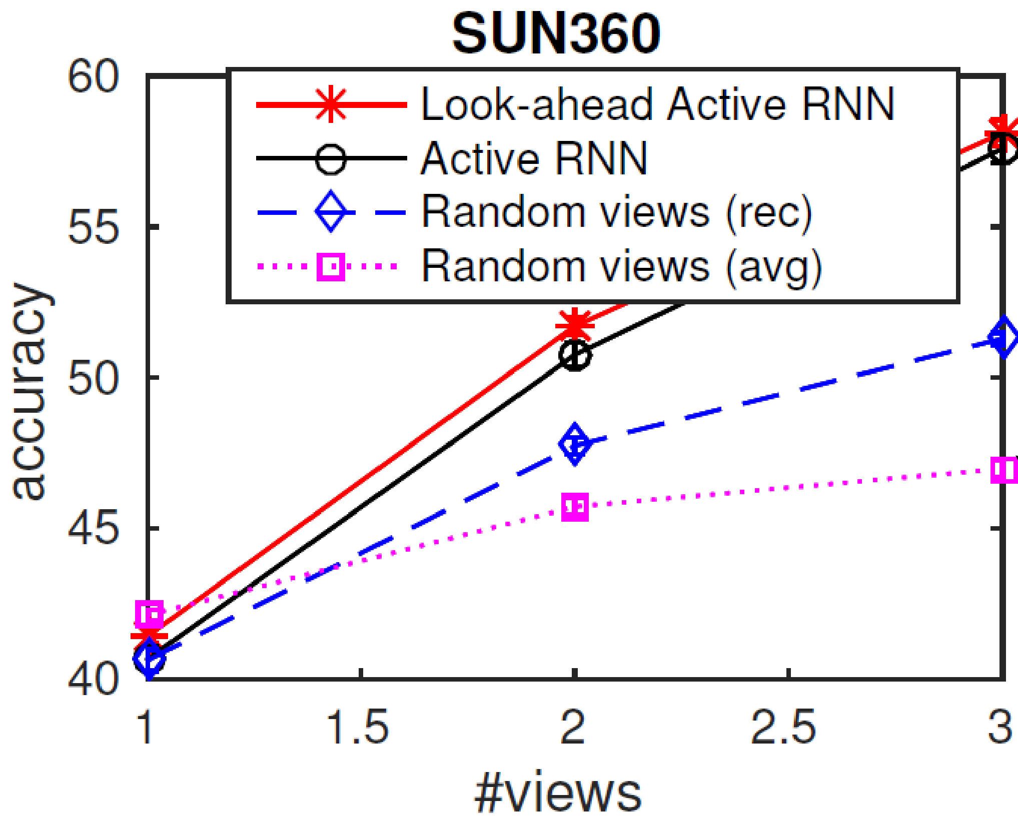
P("Plaza courtyard"): (6.28) (11.95) (67.89)
Top 3 guesses: Restaurant, Train interior, Shop | Theater, Restaurant, Plaza courtyard | Plaza courtyard, Lobby atrium, Theater

P("Church"): (0.53) (5.00) (38.38)
Top 3 guesses: Forest, Cave, Beach | Street, Cave, Plaza courtyard | Church, Street, Street



Jayaraman and Grauman, UT TR AI15-06, ECCV 2016

Active recognition: results



Looking around **actively**

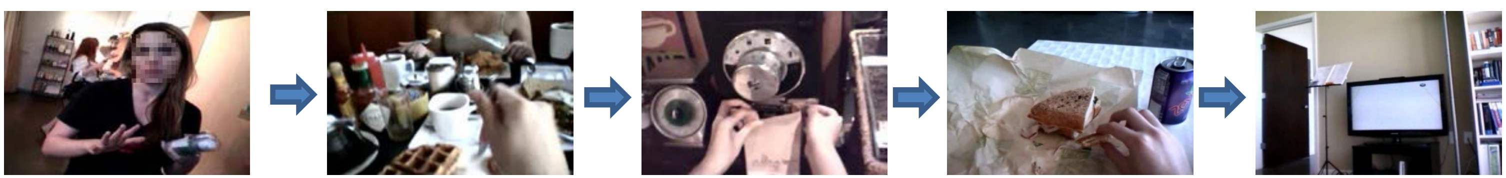
Looking around **passively**

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

Jayaraman and Grauman, UT TR AI15-06, ECCV 2016

Summarizing egocentric video

Input: Egocentric video of the camera wearer's day



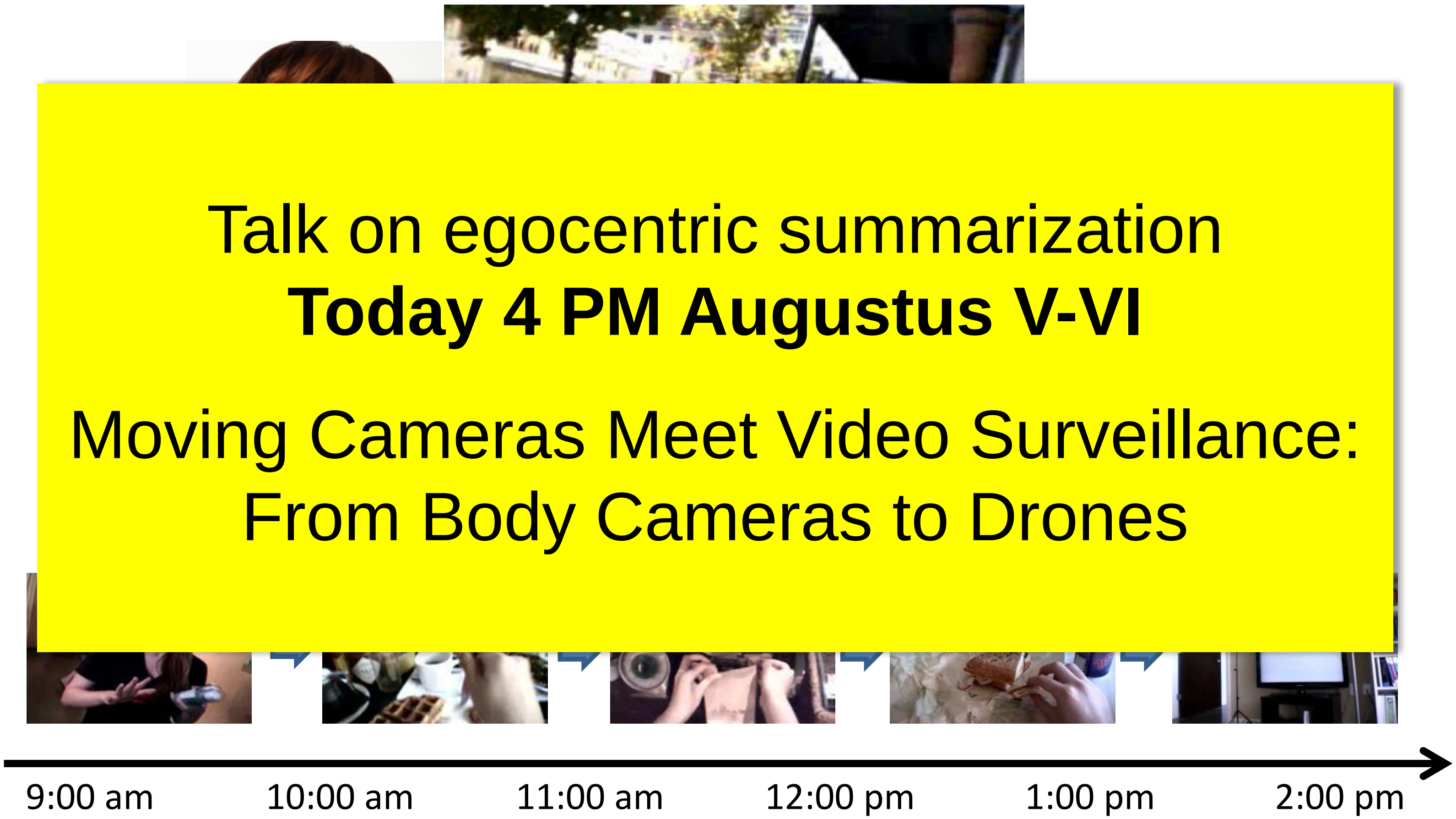
9:00 am 10:00 am 11:00 am 12:00 pm 1:00 pm 2:00 pm

Output: Storyboard (or video skim) summary

Summarizing egocentric video

Talk on egocentric summarization
Today 4 PM Augustus V-VI

Moving Cameras Meet Video Surveillance:
From Body Cameras to Drones



Output: Storyboard (or video skim) summary

Summary

THE UNIVERSITY OF TEXAS AT AUSTIN

Dinesh Jayaraman

- Visual learning benefits from
  - context of action and motion in the world
  - continuous self-acquired feedback
- New ideas:
  - "Embodied" feature learning using both visual and motor signals
  - Feature learning from unlabeled video via higher order temporal coherence
  - Steps towards active policies for view selection

Papers

- **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.
- **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.
- **Look-ahead before you leap: end-to-end active vision by forecasting the effect of motion.** D. Jayaraman and K. Grauman. To appear, European Conference on Computer Vision (ECCV), Oct 2016.