Learning How to Move and Where to Look from Unlabeled Video

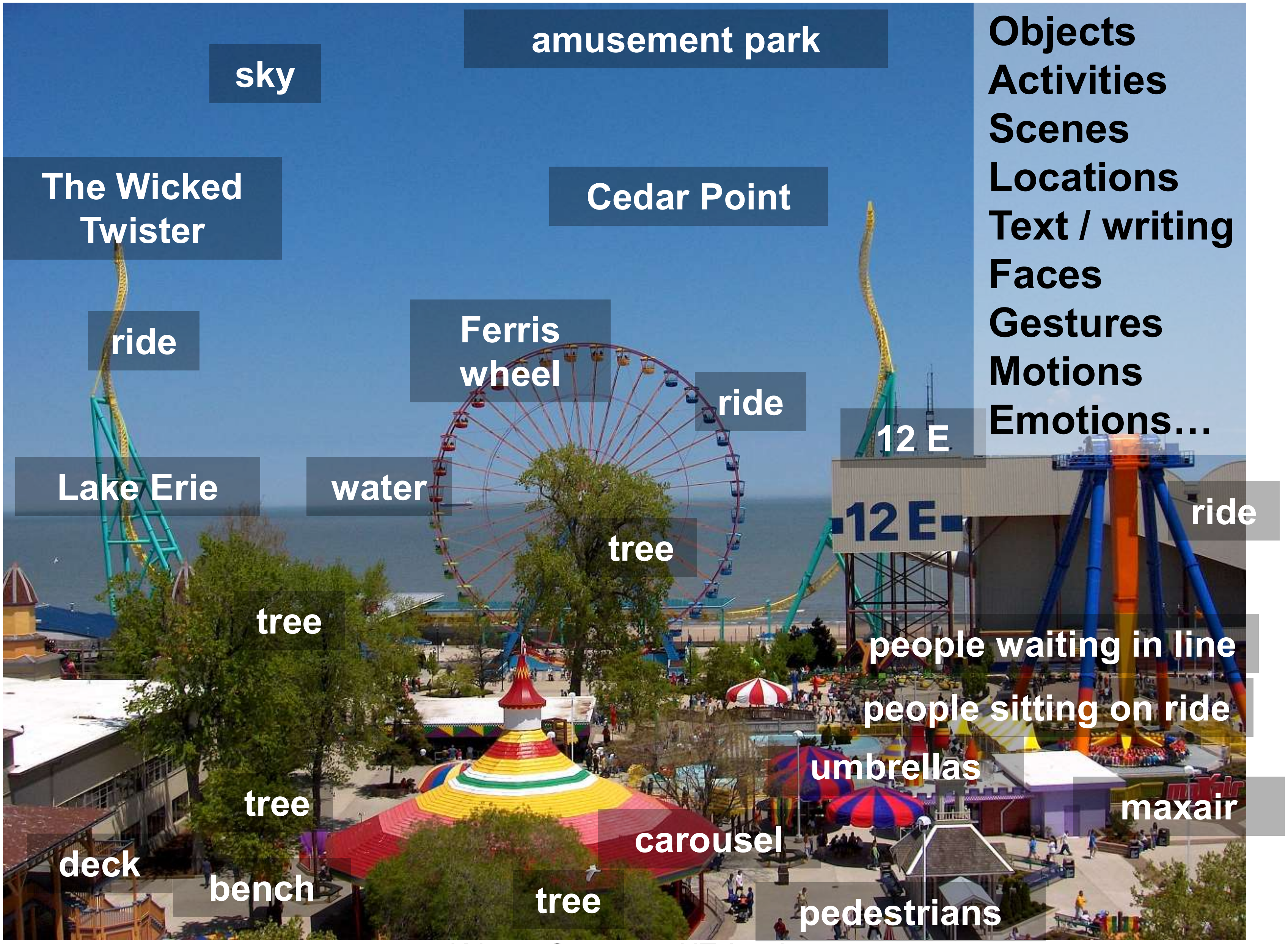
Kristen Grauman

Department of Computer Science

University of Texas at Austin

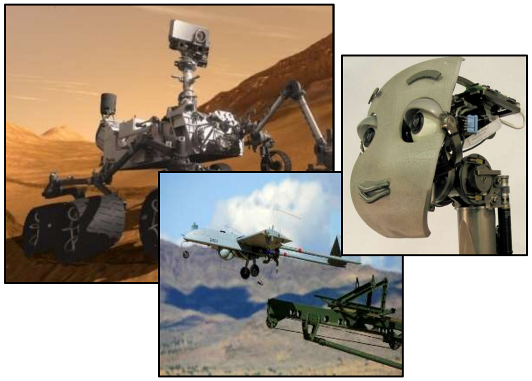
THE UNIVERSITY OF TEXAS AT AUSTIN

Visual recognition

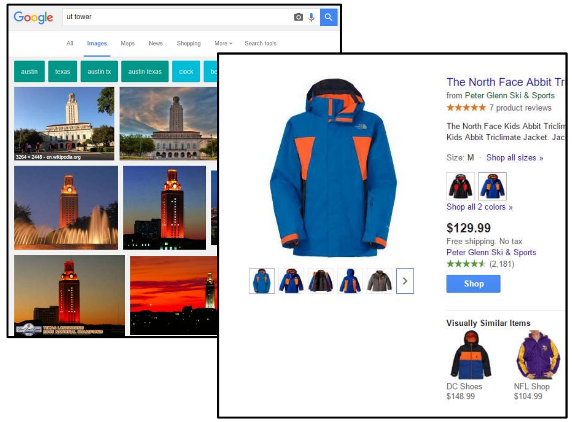


Kristen Grauman, UT Austin

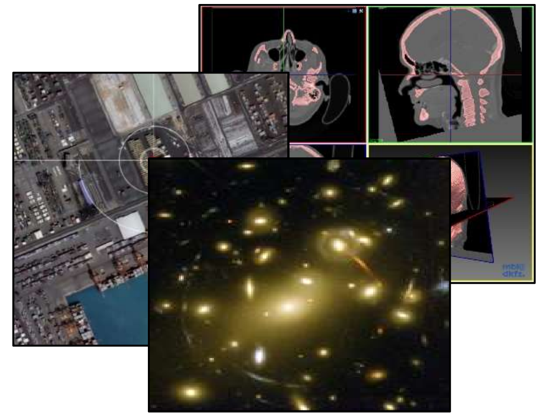
Visual recognition: applications



AI and autonomous robotics



Organizing visual content



Science and medicine

Gaming, HCI, Augmented Reality

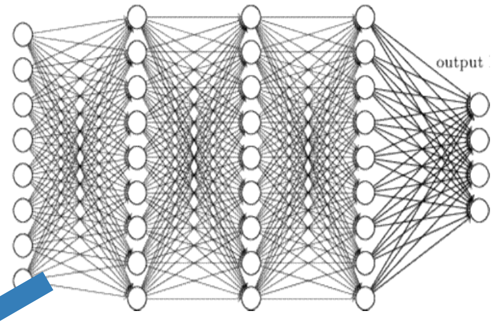
Surveillance and security

Personal photo/video collections

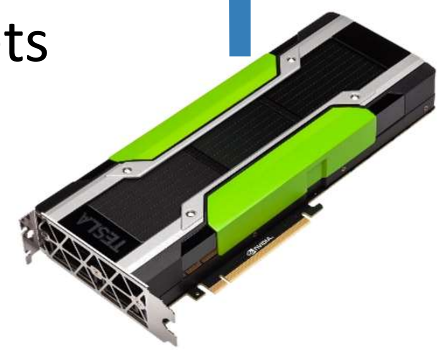
Kristen Grauman, UT Austin

Significant recent progress in the field

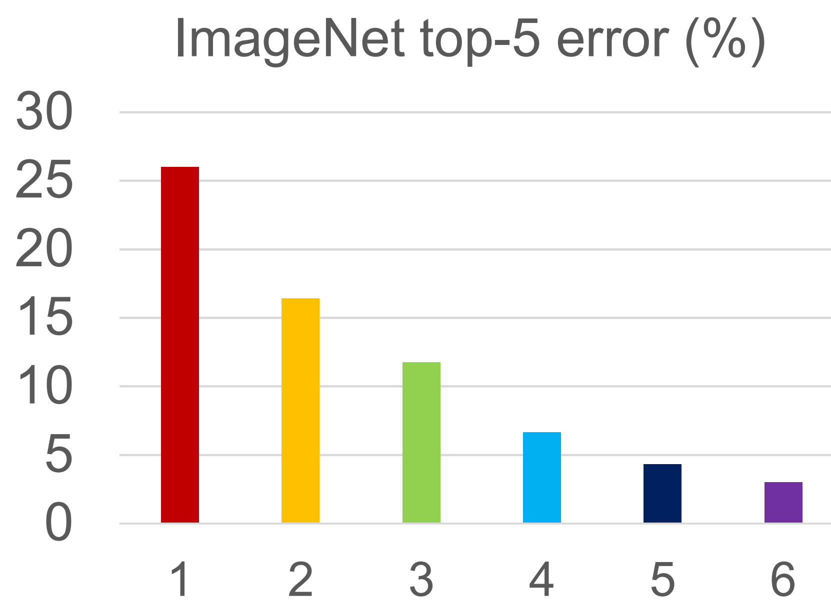
Big labeled datasets



Deep learning

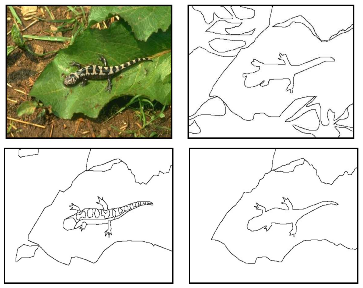


GPU technology

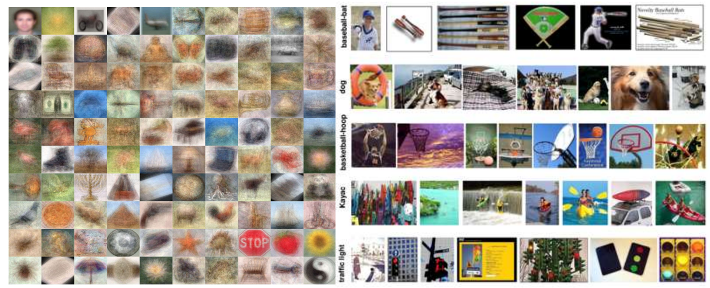


Kristen Grauman, UT Austin

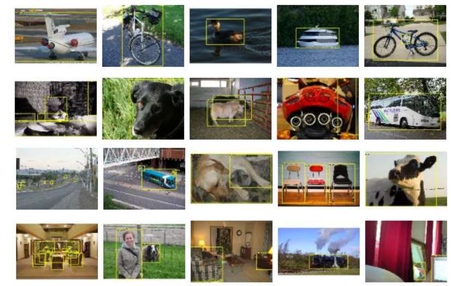
Recognition benchmarks



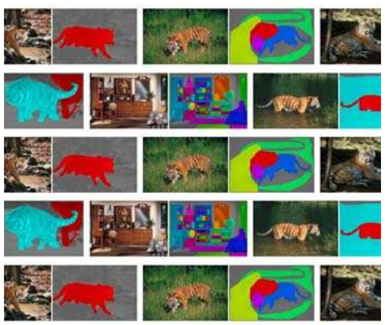
BSD (2001)



Caltech 101 (2004), Caltech 256 (2006)



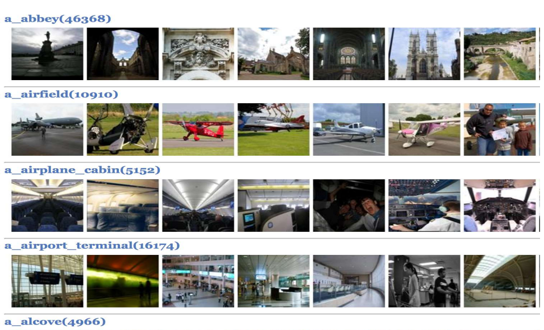
PASCAL (2007-12)



LabelMe (2007)

ImageNet (2009)

SUN (2010)



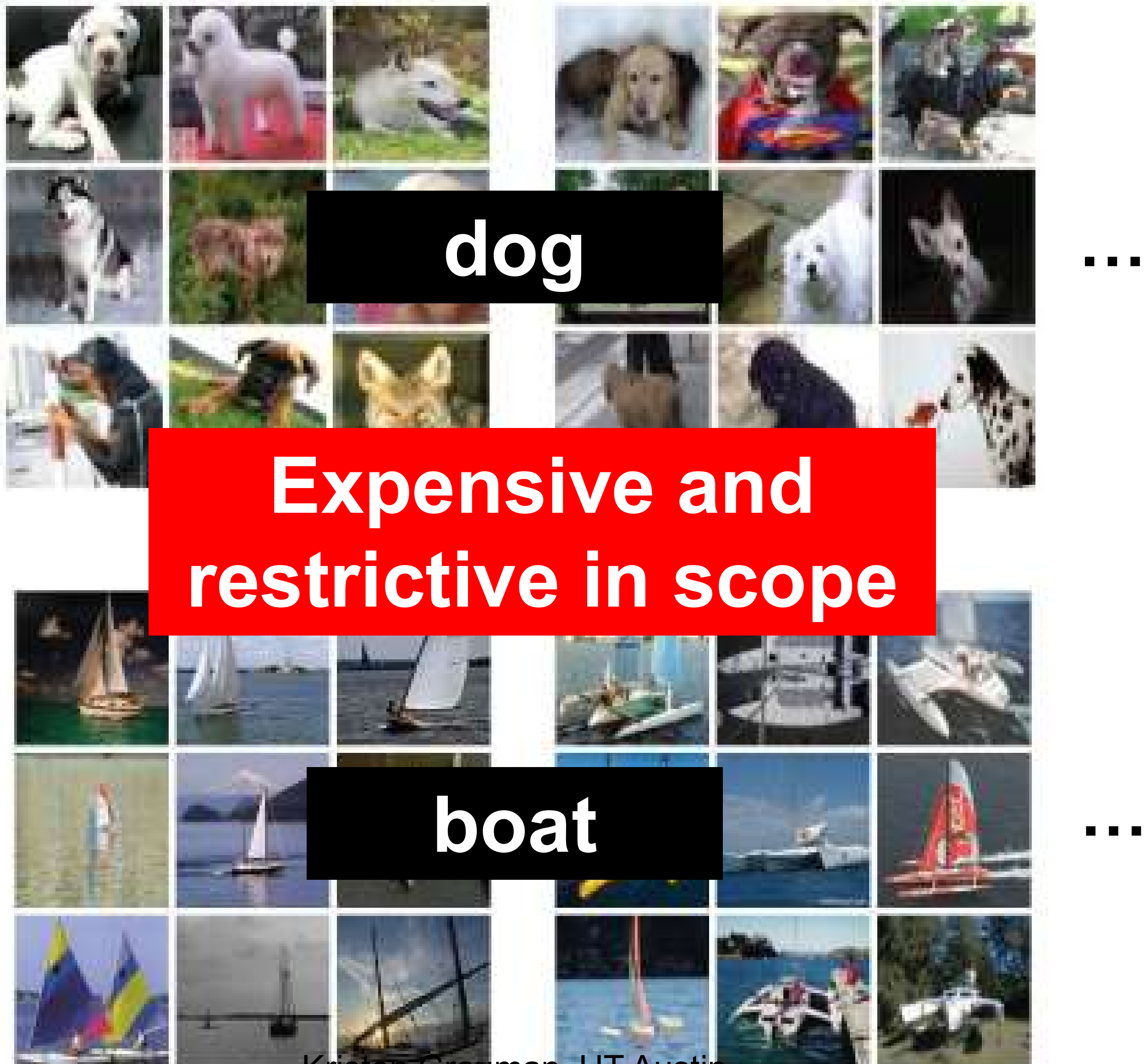
Places (2014)

MS COCO (2014)

Visual Genome (2016)

Kristen Grauman, UT Austin

How do our systems learn about the visual world today?



Kristen Grauman, UT Austin

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

Kristen Grauman, UT Austin

Big picture goal: Embodied visual learning

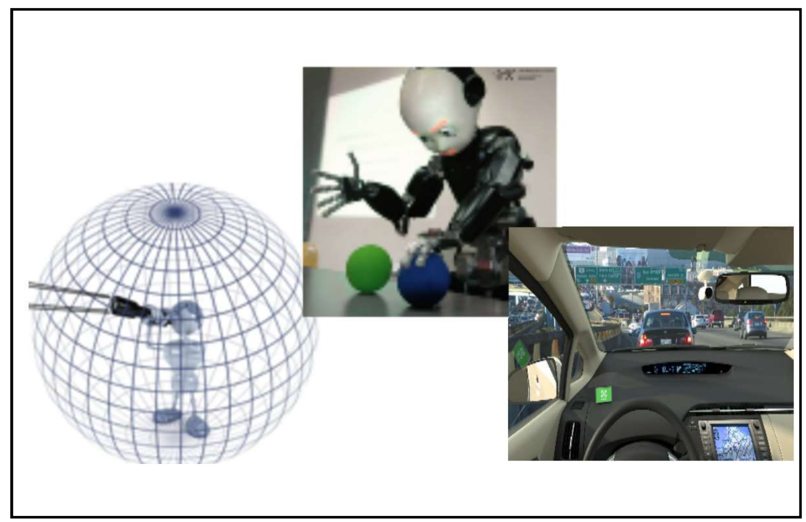
Status quo:

Learn from “disembodied” bag of labeled snapshots.

Our goal:

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

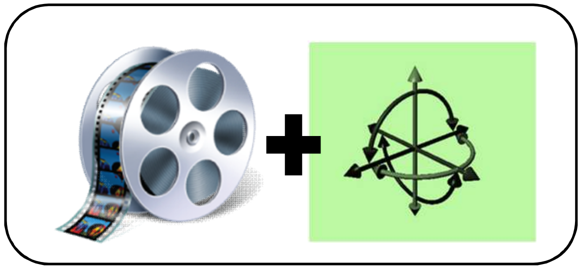
Inexpensive and unrestricted in scope



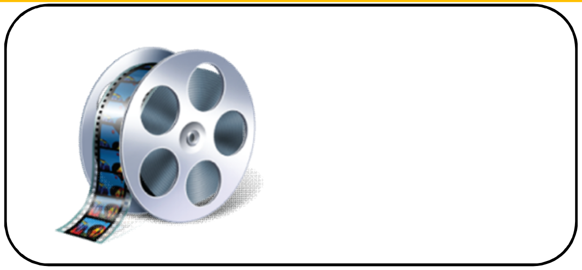
Talk overview

Towards embodied visual learning

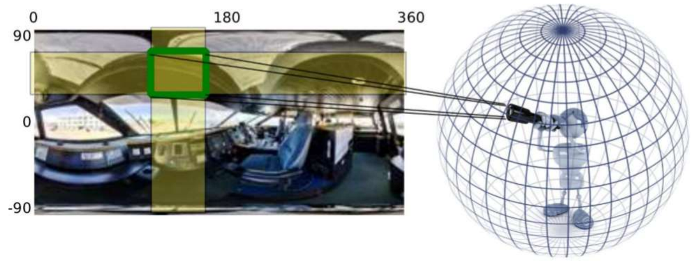
1. Learning representations tied to ego-motion



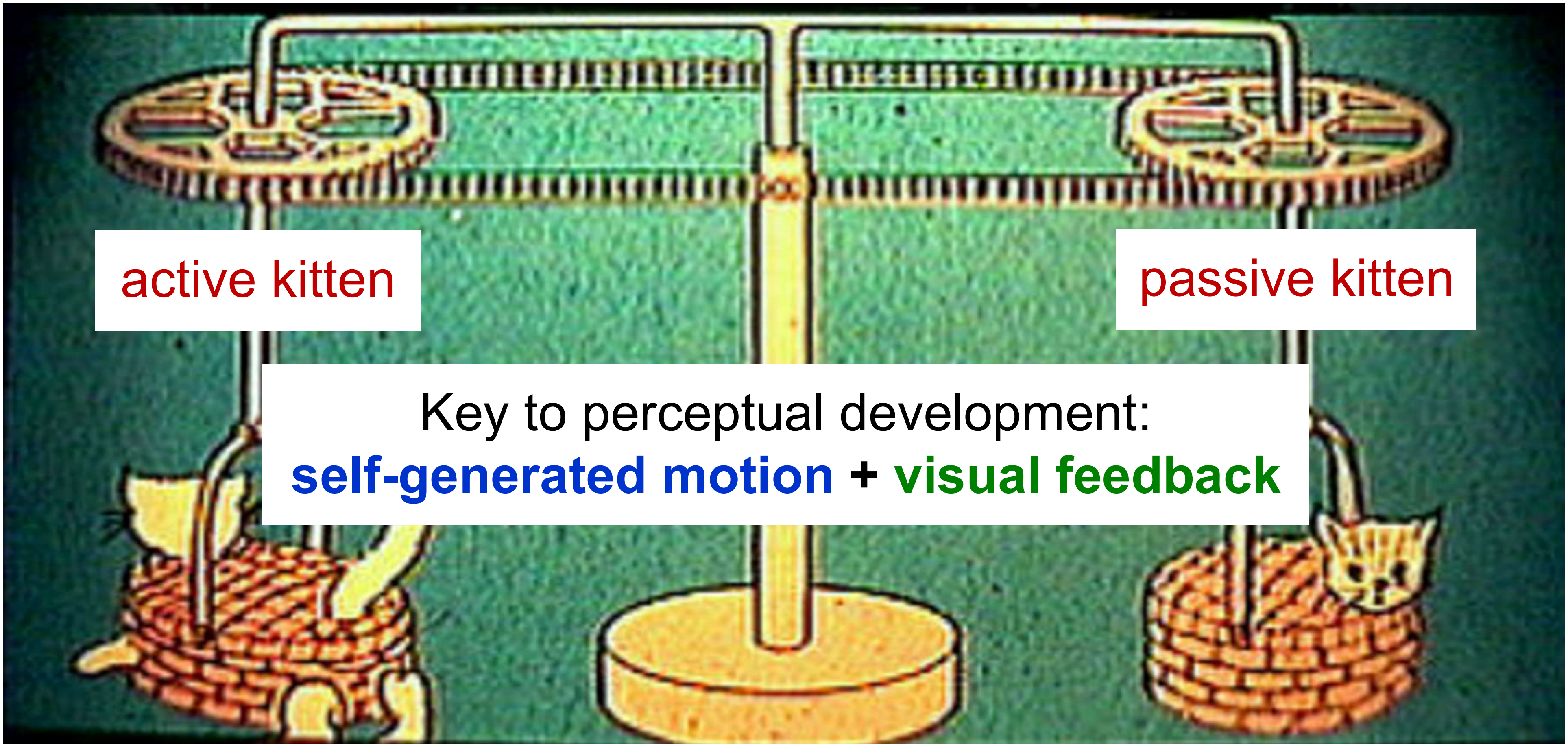
2. Learning representations from unlabeled video



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:

Kristen Grauman, UT Austin

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"

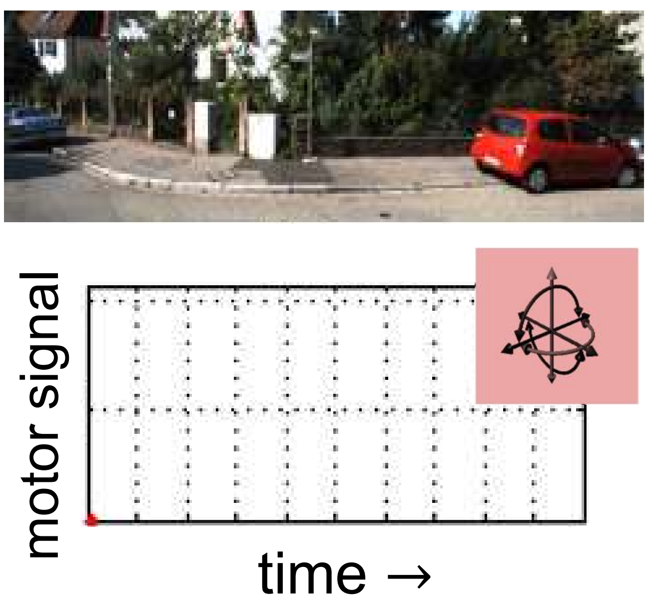
$$\mathbf{z}(g\mathbf{x}) \approx M_g\mathbf{z}(\mathbf{x})$$

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

Approach idea: Ego-motion equivariance

Training data

Unlabeled video + motor signals



Learn

Equivariant embedding

organized by ego-motions

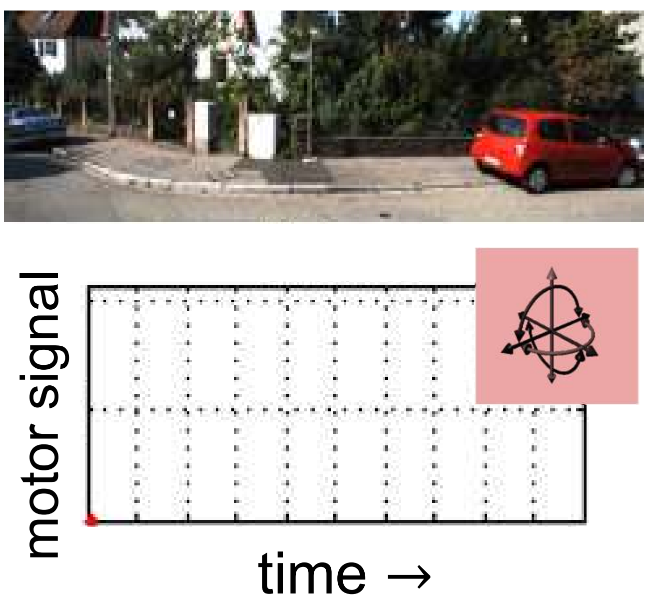
Pairs of frames related by similar ego-motion should be related by same feature transformation

[Jayaraman & Grauman, ICCV 2015]

Approach idea: Ego-motion equivariance

Training data

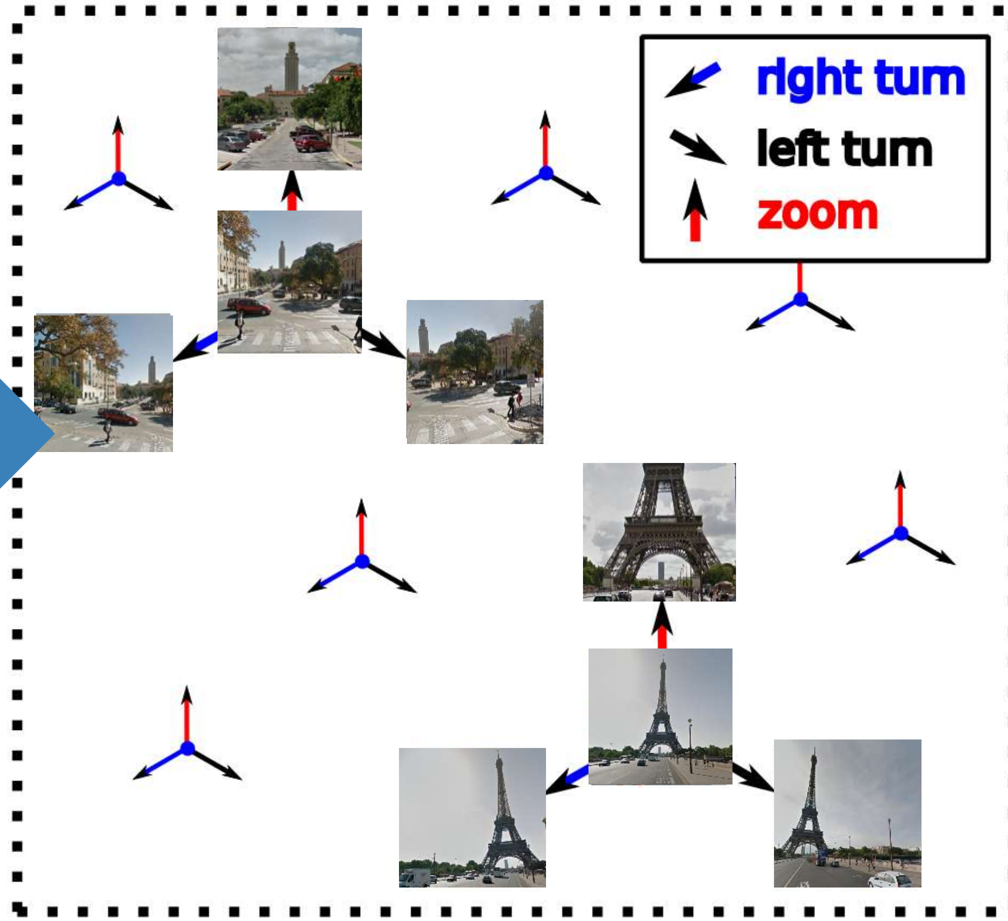
Unlabeled video + motor signals



Learn

Equivariant embedding

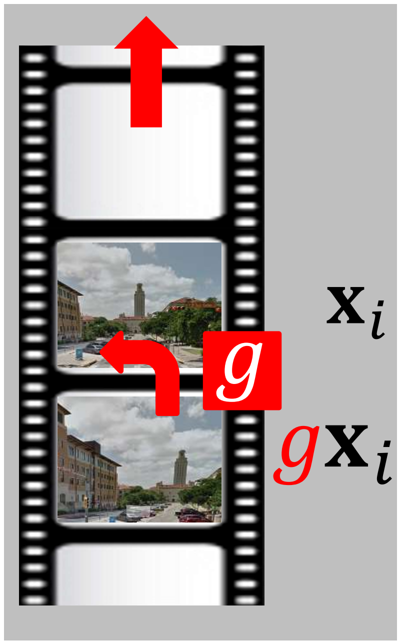
organized by ego-motions



[Jayaraman & Grauman, ICCV 2015]

Ego-motion equivariant feature learning

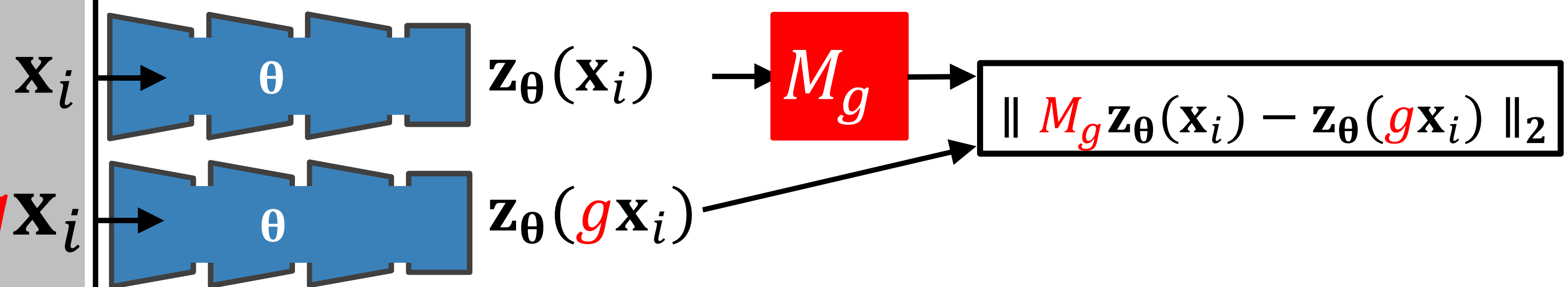
Given:



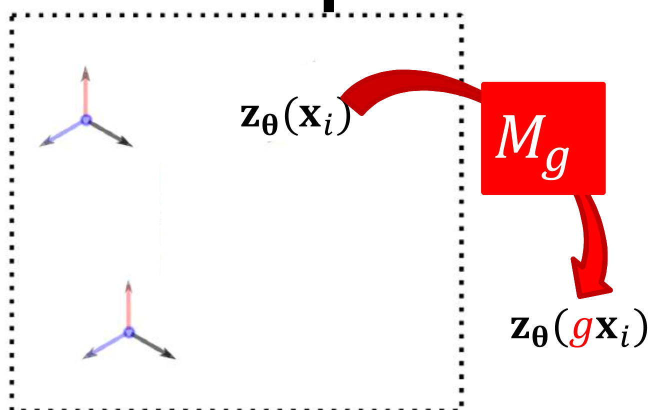
Desired: for all motions g and all images $\mathbf{x}$,

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

Unsupervised training



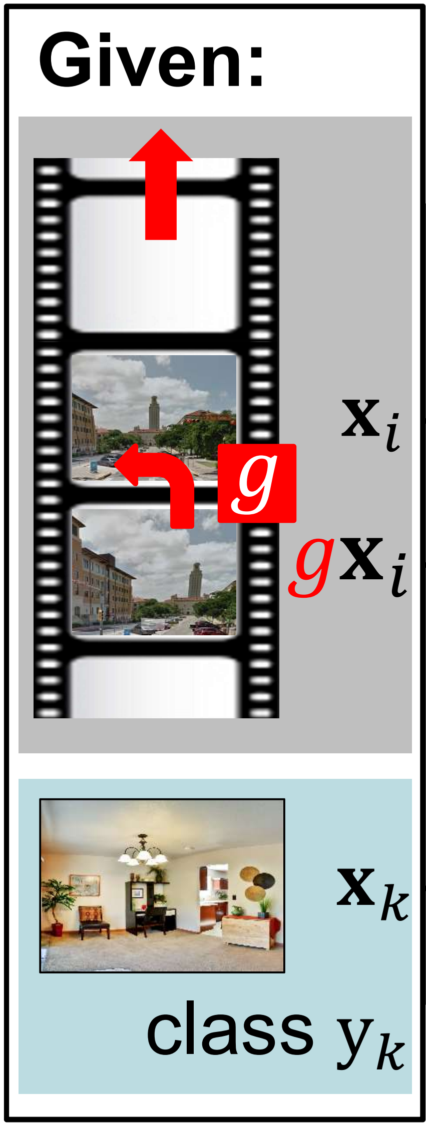
Feature space



Kristen Grauman, UT Austin

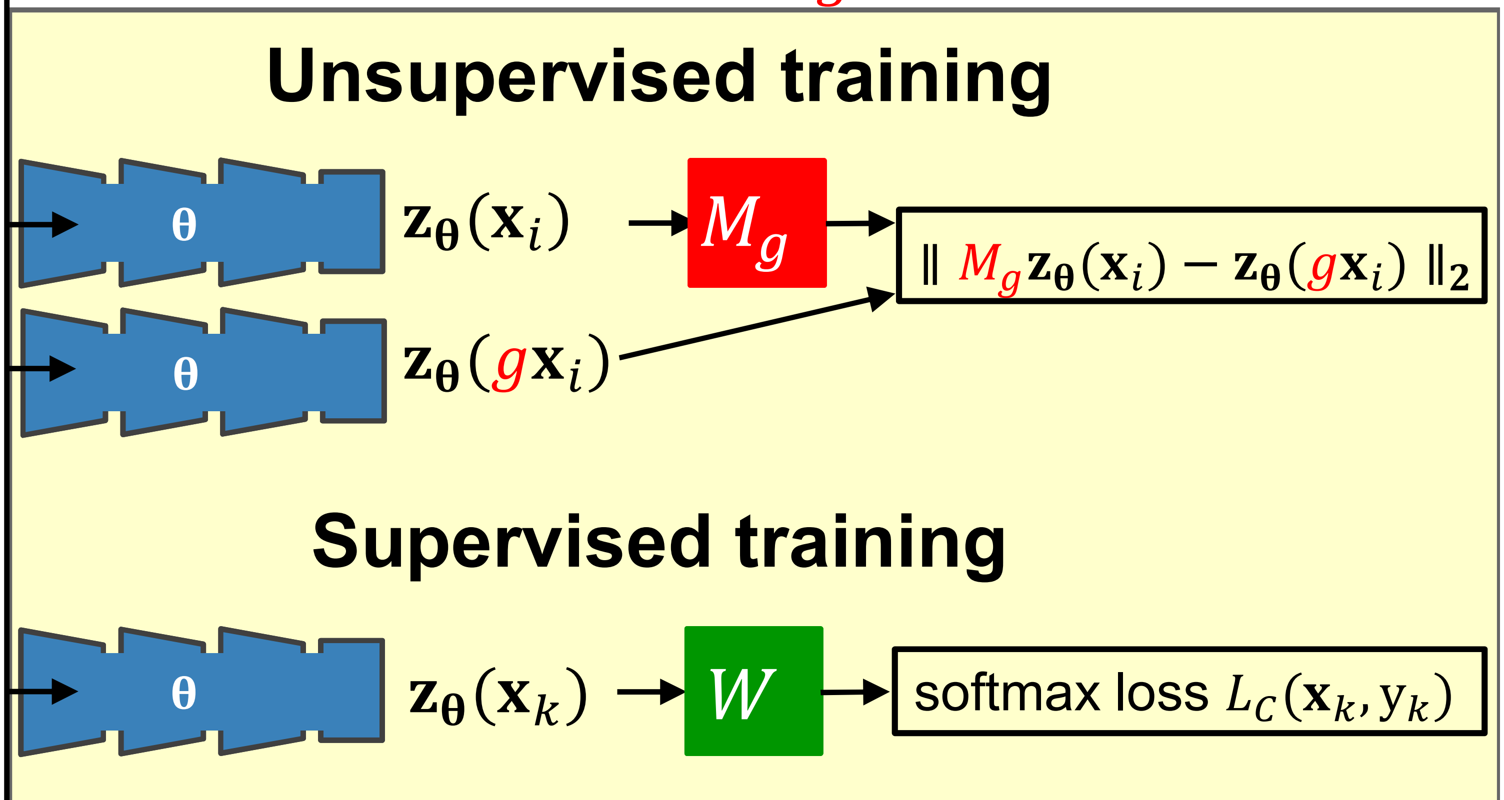
[Jayaraman & Grauman, ICCV 2015]

Ego-motion equivariant feature learning



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

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



$\boldsymbol{\theta}$, M_g and W jointly trained

[Jayaraman & Grauman, ICCV 2015]

Results: Recognition

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

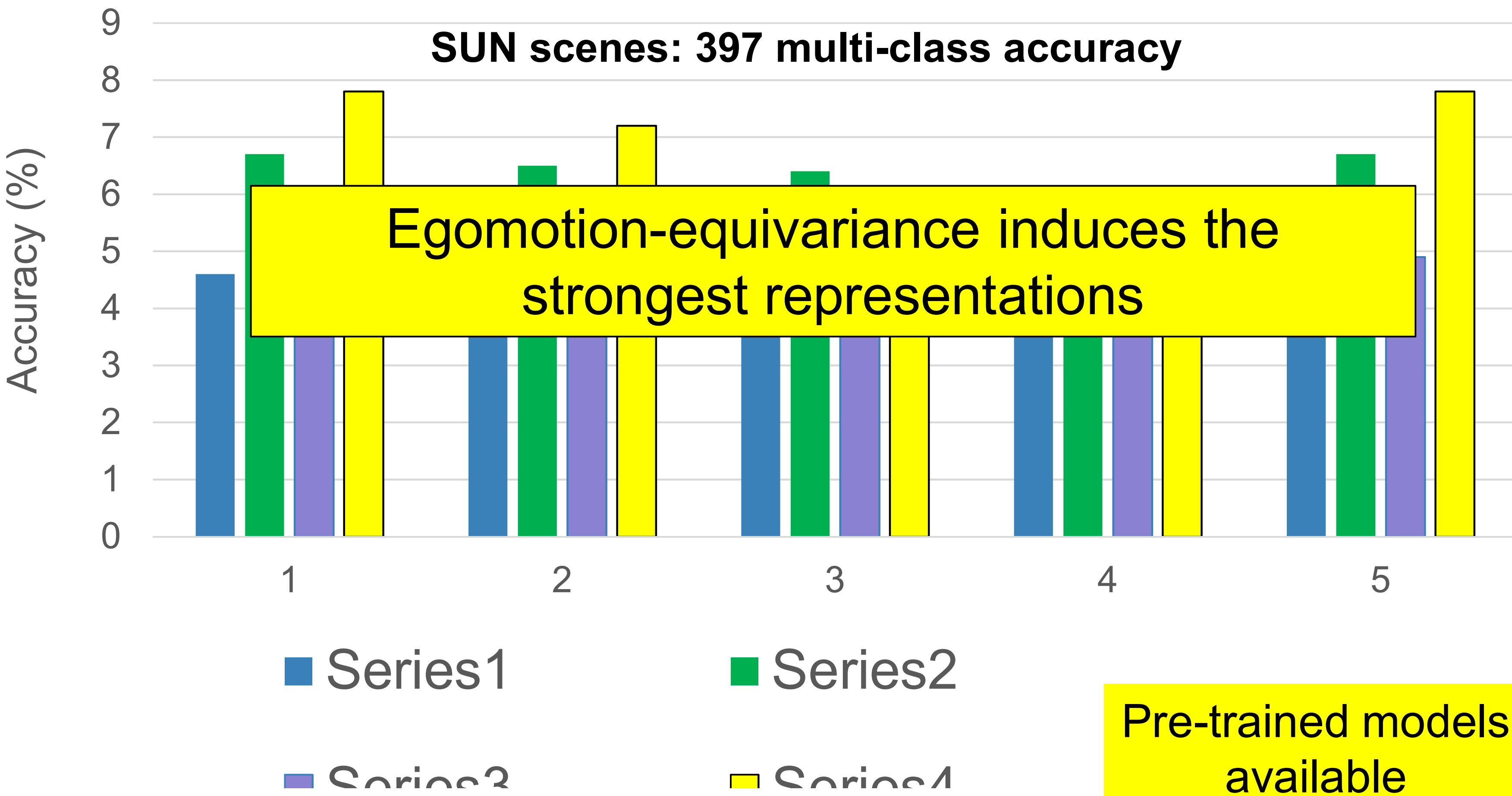
Geiger et al, IJRR ’13

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

Xiao et al, CVPR ’10

Results: Recognition

Ego-equivariance for unsupervised feature learning



Pre-trained models available

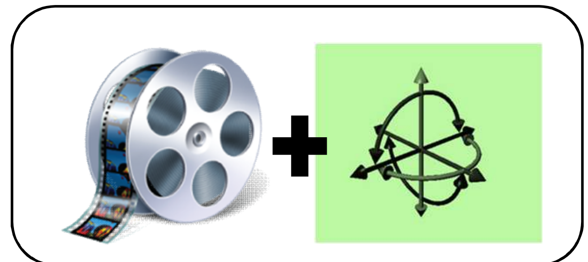
+ Hadsell, Chopra, LeCun, "Dimensionality Reduction by Learning an Invariant Mapping", CVPR 2006
* Agrawal, Carreira, Malik, "Learning to see by moving", ICCV 2015

Kristen Grauman, UT Austin

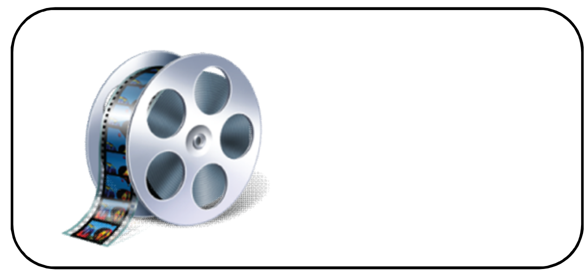
Talk overview

Towards embodied visual learning

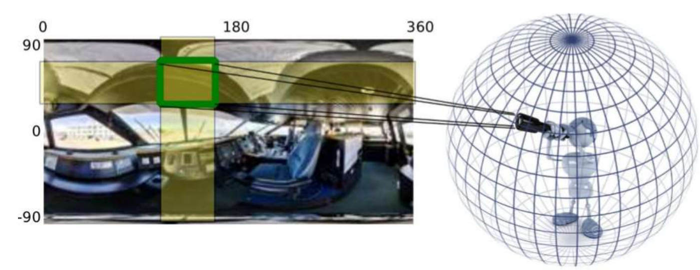
1. Learning representations tied to ego-motion



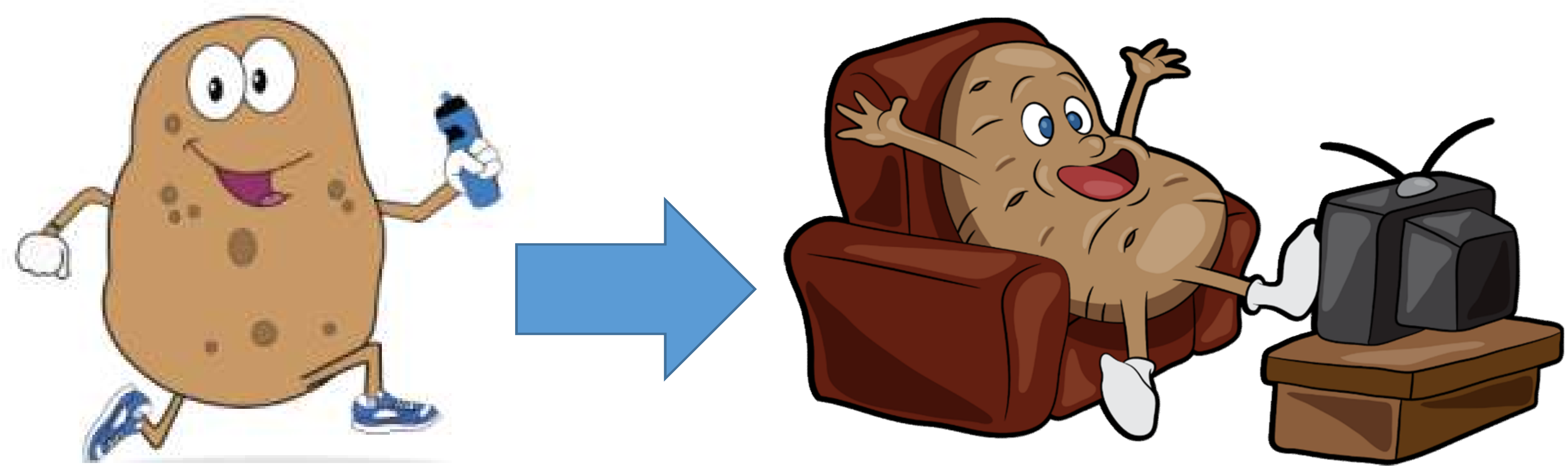
2. Learning representations from unlabeled video



3. Learning how to move and where to look



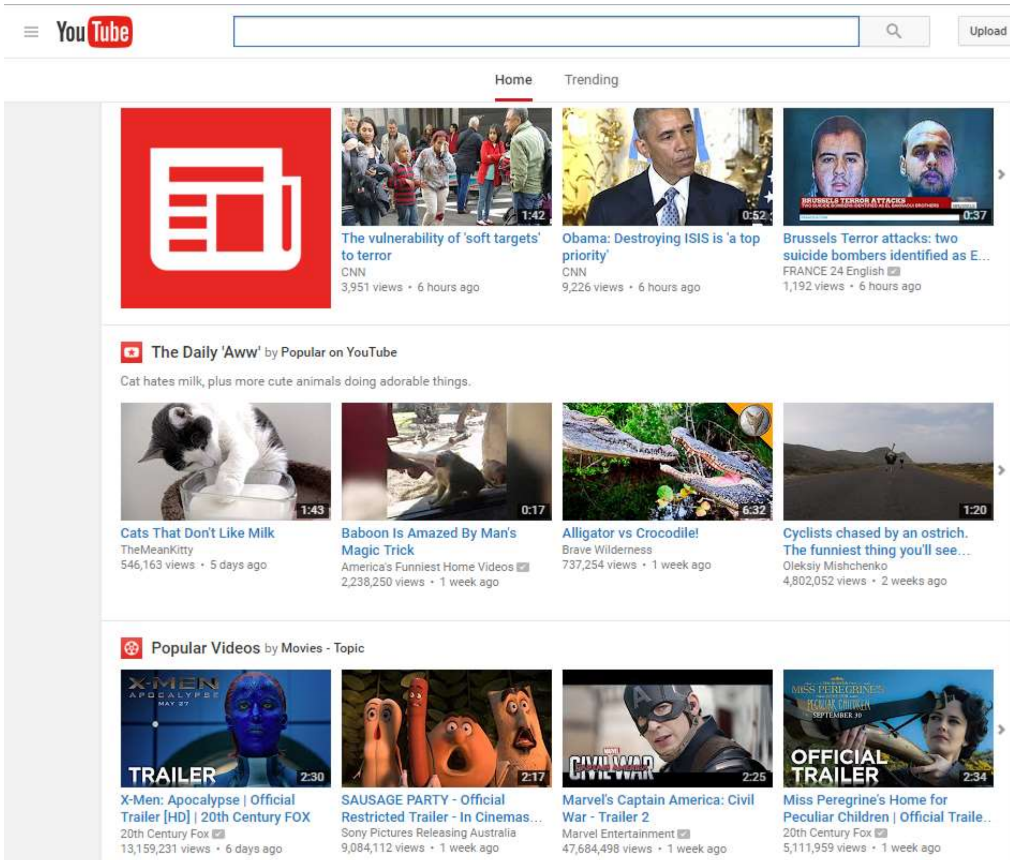
Learning from arbitrary unlabeled video?



Unlabeled video + ego-motion

Unlabeled video

Learning from arbitrary unlabeled video?



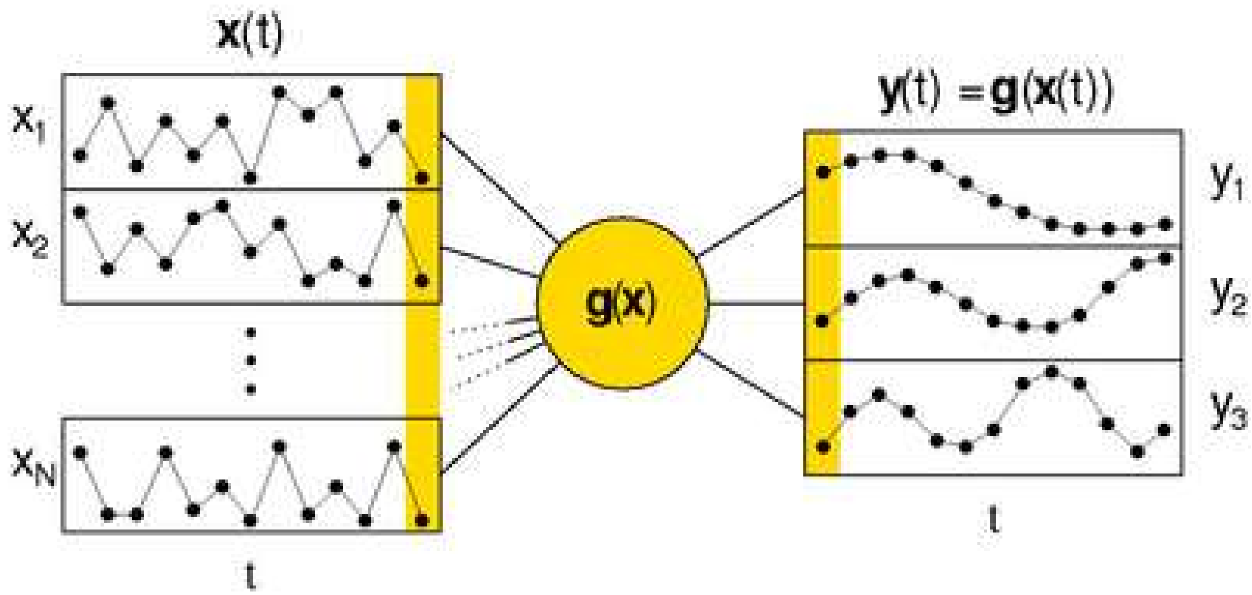
Unlabeled video

Kristen Grauman, UT Austin

Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]

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



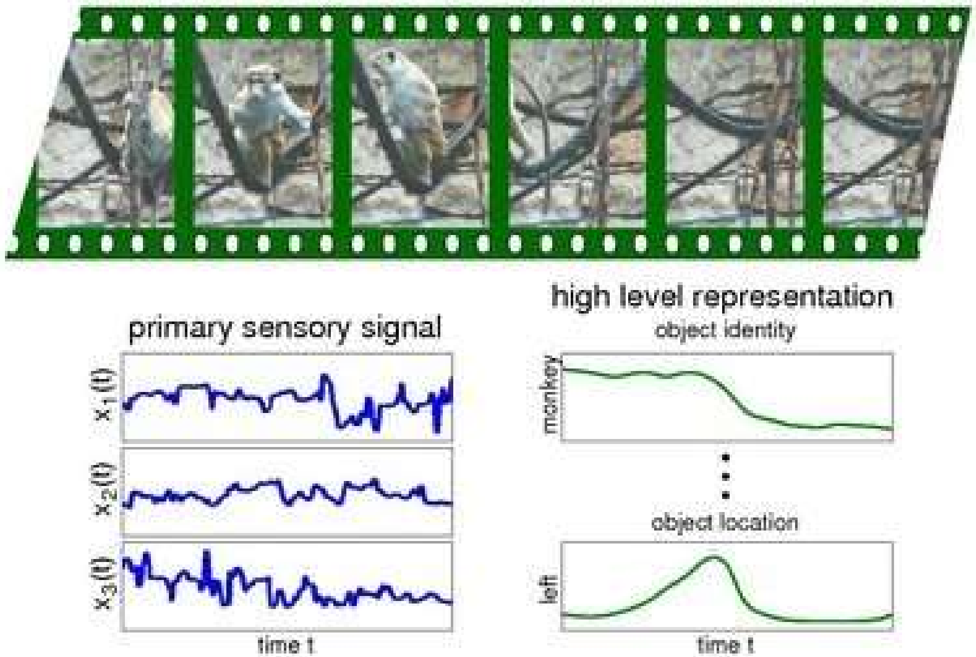
quickly varying input signal **x**(t) → slowly varying features **y**(t)

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

Prior work: Slow feature analysis

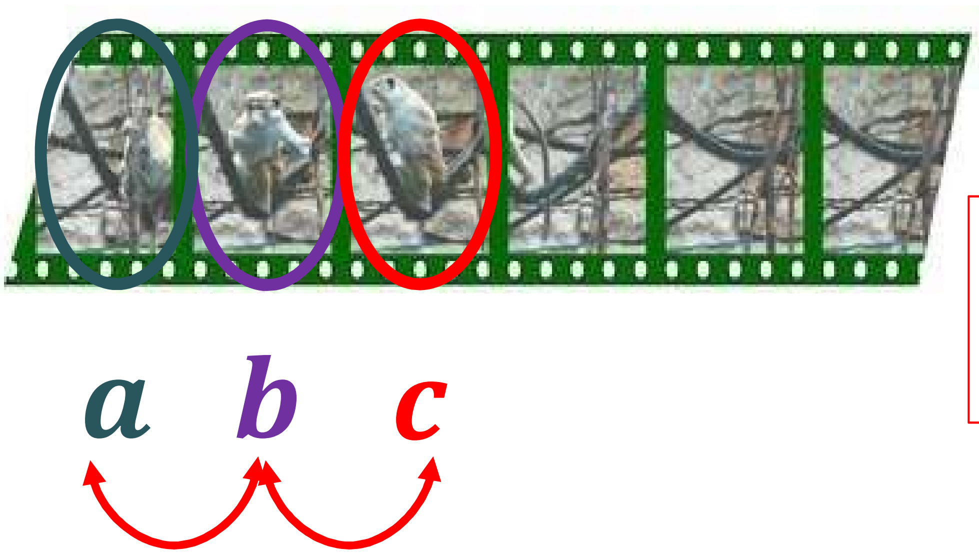
$\boldsymbol{a}$ $\boldsymbol{b}$

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

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

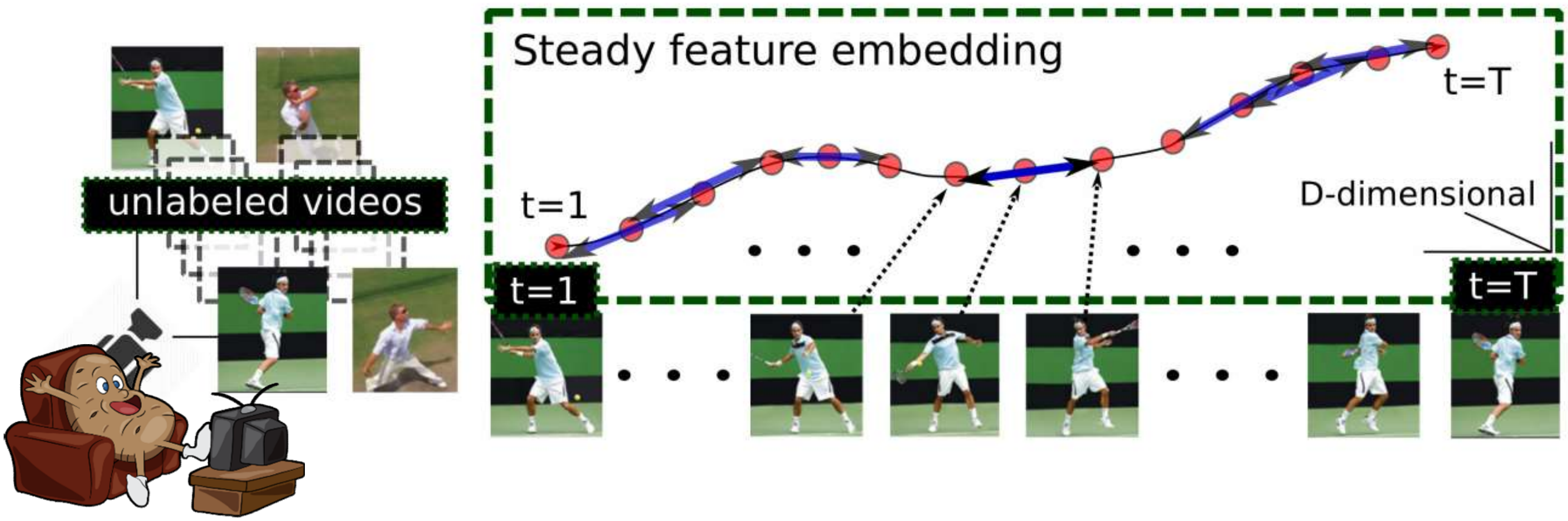
Learn feature map $\boldsymbol{z}(.)$ such that:

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

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

[Jayaraman & Grauman, CVPR 2016]

Our idea: *Steady* feature analysis



Learn feature map $z(.)$ such that:

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

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

[Jayaraman & Grauman, CVPR 2016]

Datasets

Unlabeled video

Target task (few labels)

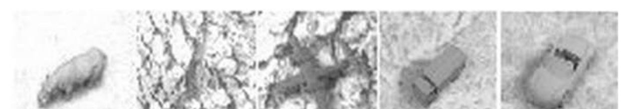
Human Motion Database (HMDB)

PASCAL 10 Actions

KITTI Video

SUN 397 Scenes

NORB

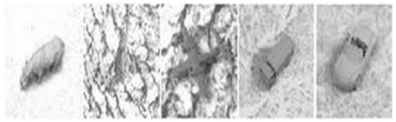


NORB 25 Objects

32 x 32 images or 96 x 96 images

Kristen Grauman, UT Austin

Results: Steady feature analysis



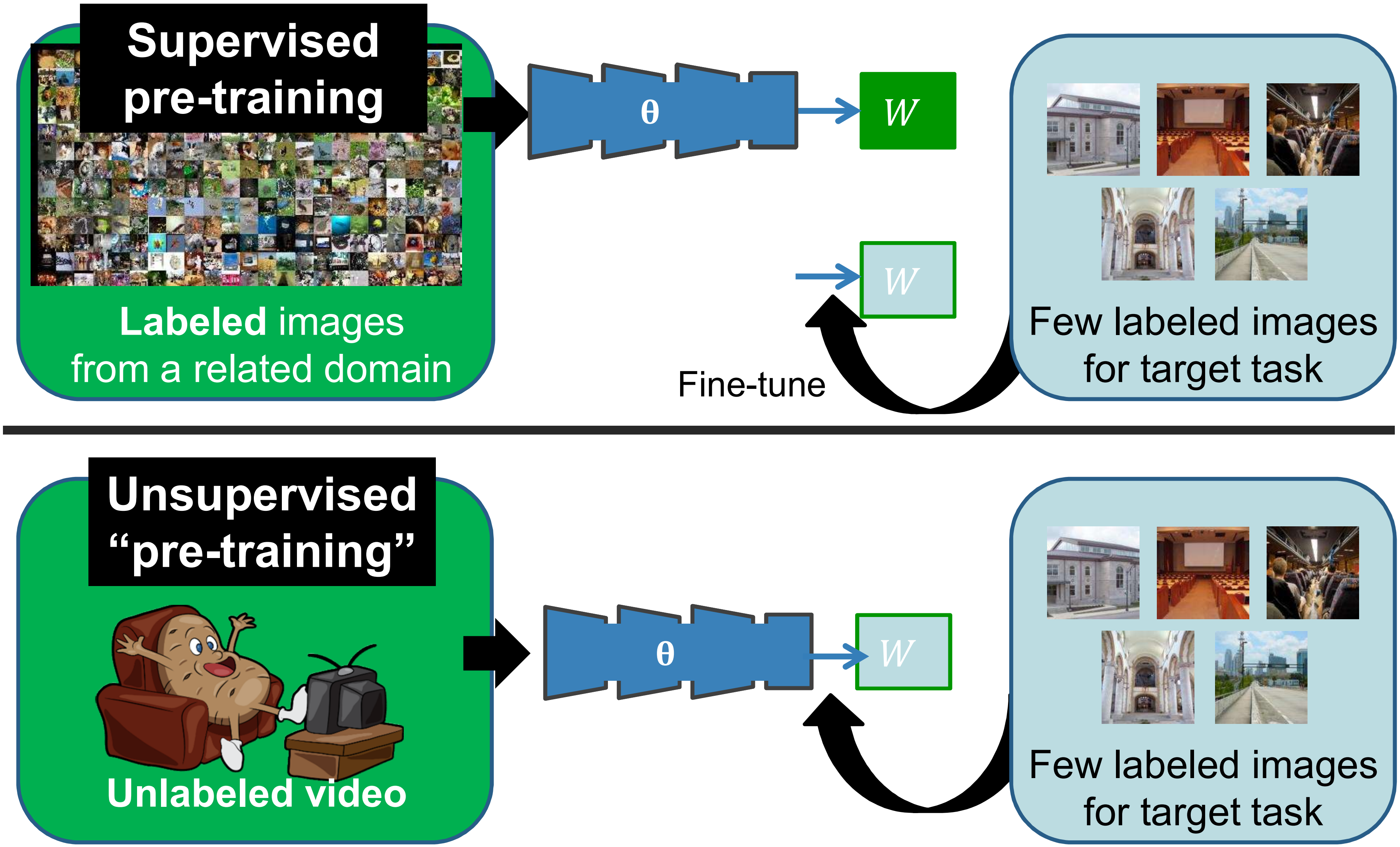

Task type→	Objects	Scenes		Actions
Datasets→	NORB→NORB	KITTI→SUN		HMDB→PASCAL-10
Methods↓	[25 cls]	[397 cls]	[397 cls, top-10]	[10 cls]
random	4.00	0.25	2.52	10.00
UNREG	24.64±0.85	0.70±0.12	6.10±0.67	15.34±0.28
SFA-1 [30]*	37.57±0.85	1.21±0.14	8.24±0.25	19.26±0.45
SFA-2 [14]**	39.23±0.94	1.02±0.12	6.78±0.32	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

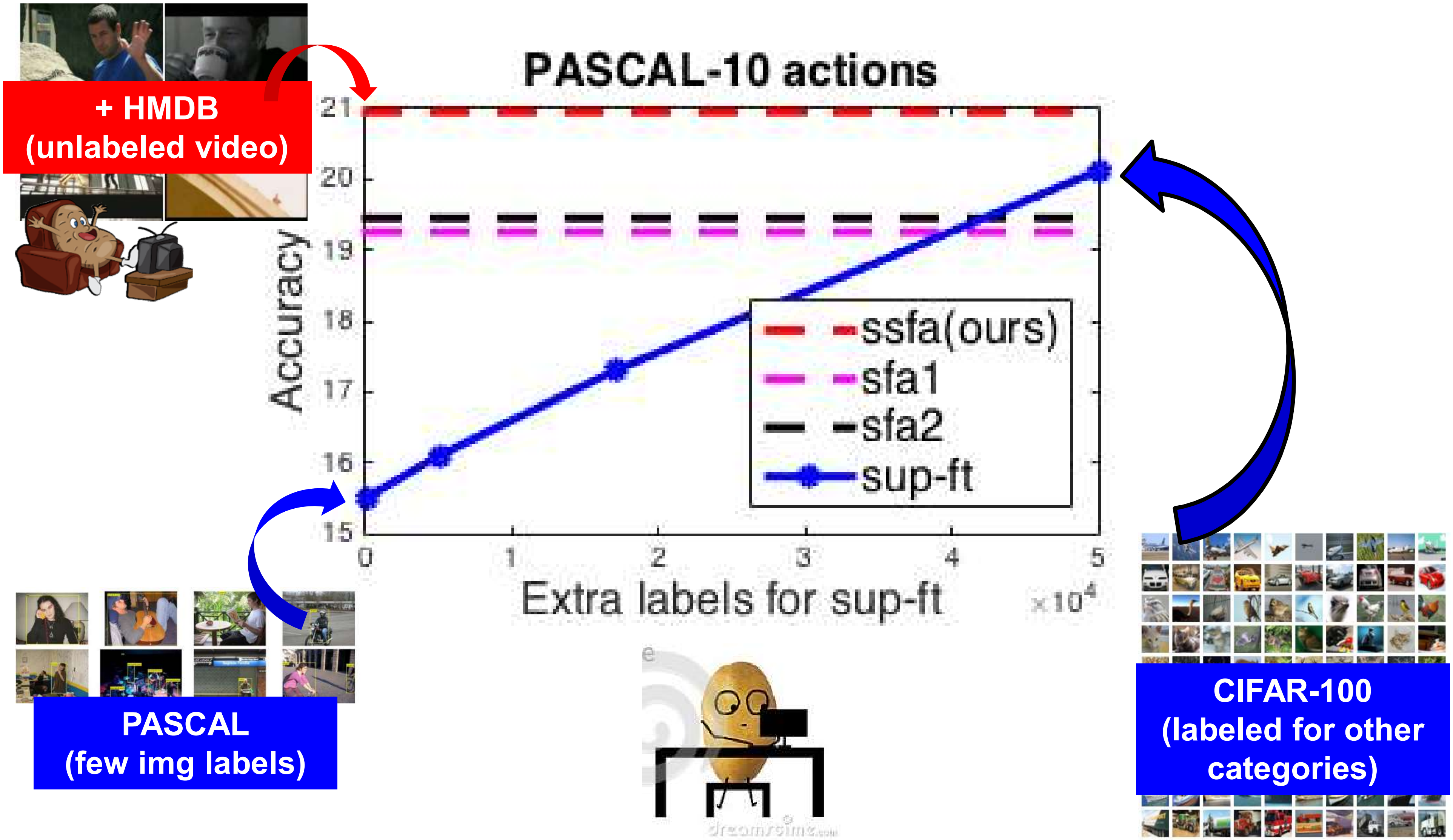
*Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR'06

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

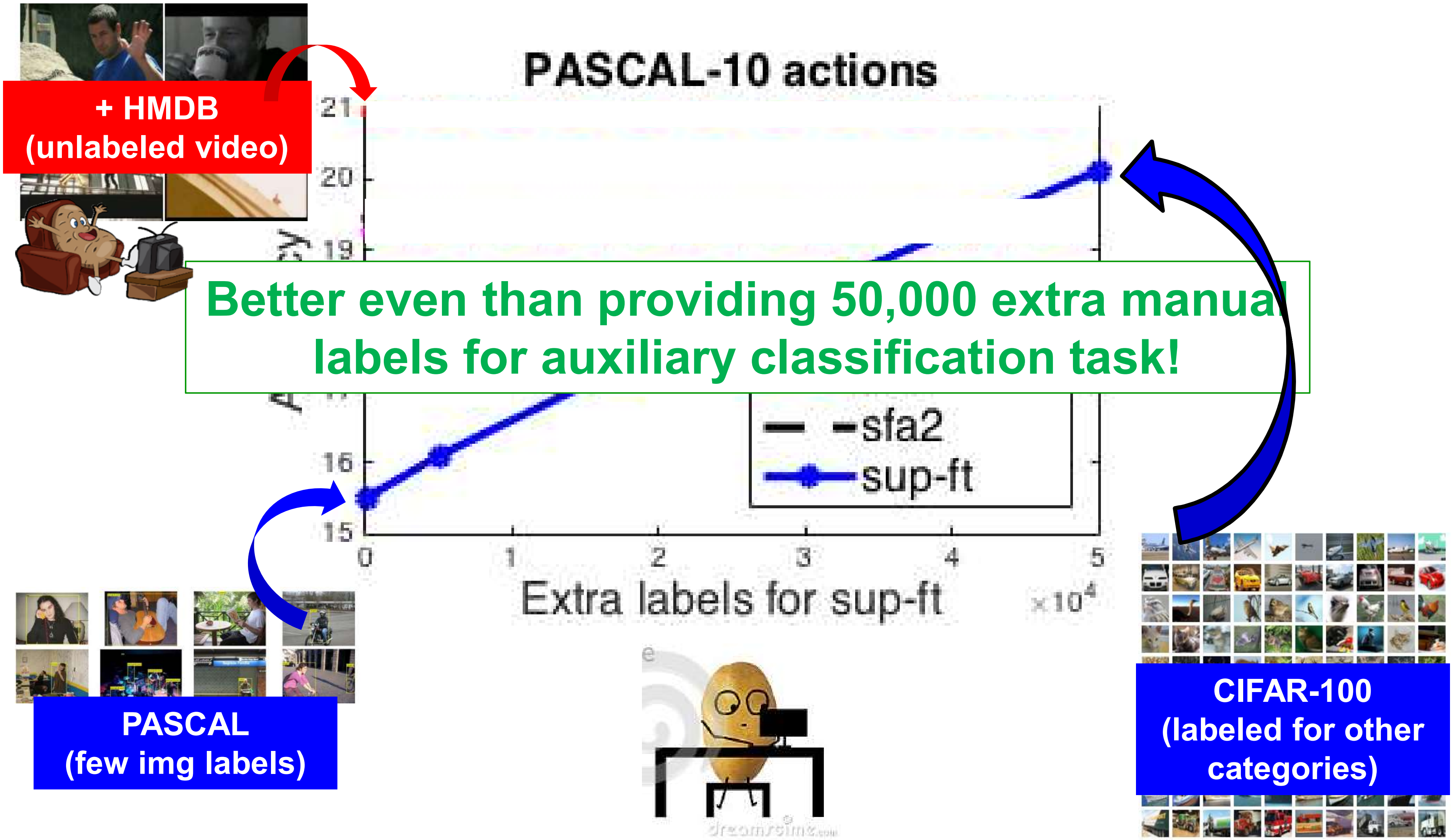
Pre-training a representation



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



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

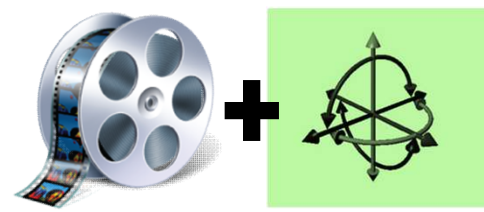


Kristen Grauman, UT Austin

Talk overview

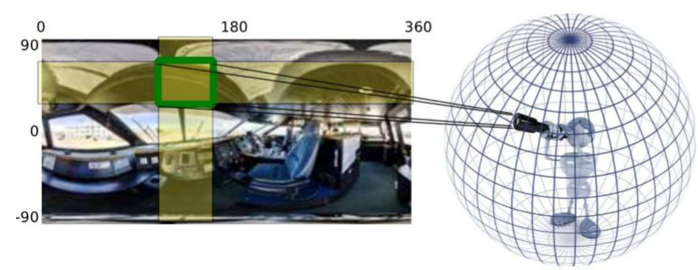
Towards embodied visual learning

1. Learning representations tied to ego-motion



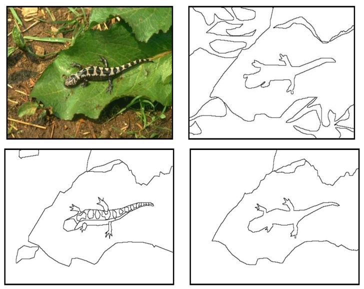
2. Learning representations from unlabeled video

3. Learning how to move and where to look

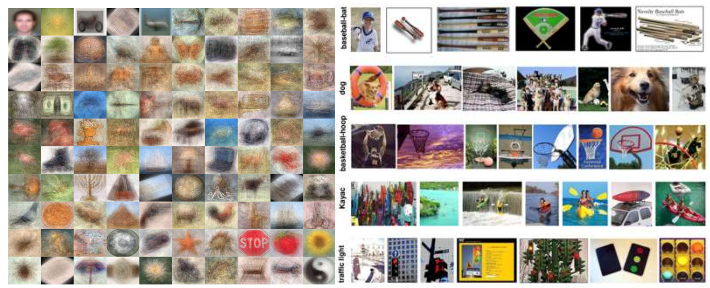


Current recognition benchmarks

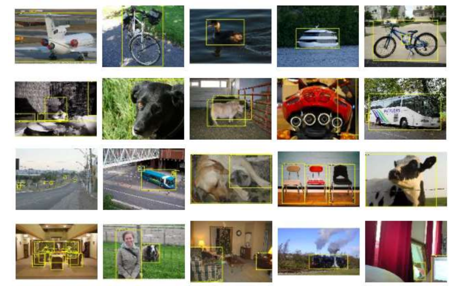
Passive, disembodied snapshots at *test* time, too



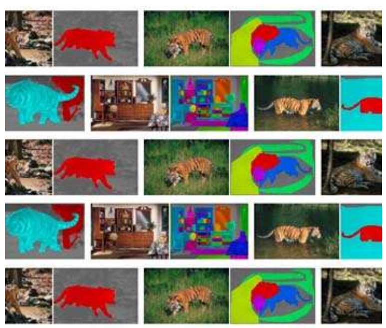
BSD (2001)



Caltech 101 (2004), Caltech 256 (2006)



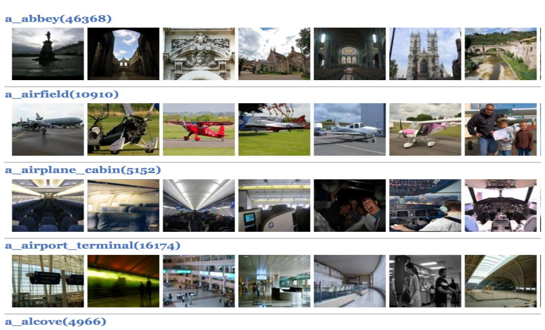
PASCAL (2007-12)



LabelMe (2007)

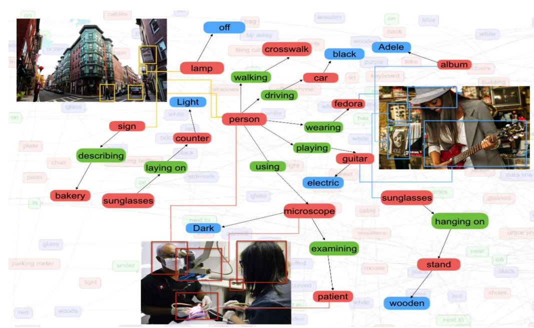
ImageNet (2009)

SUN (2010)



Places (2014)

MS COCO (2014)

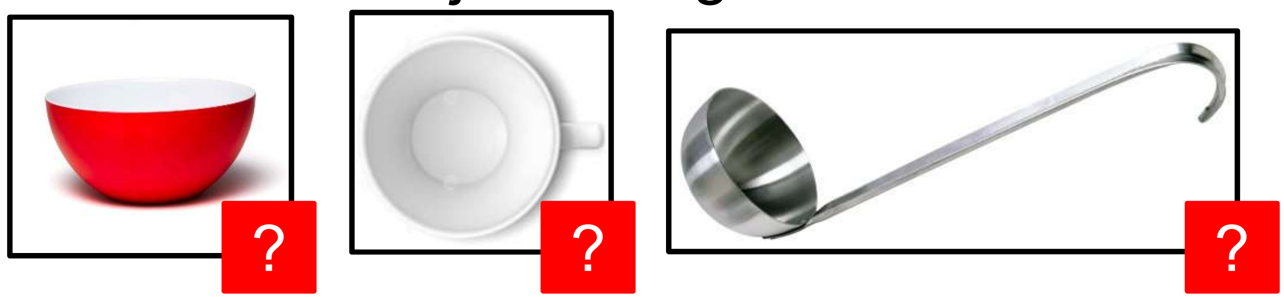


Visual Genome (2016)

Current recognition benchmarks

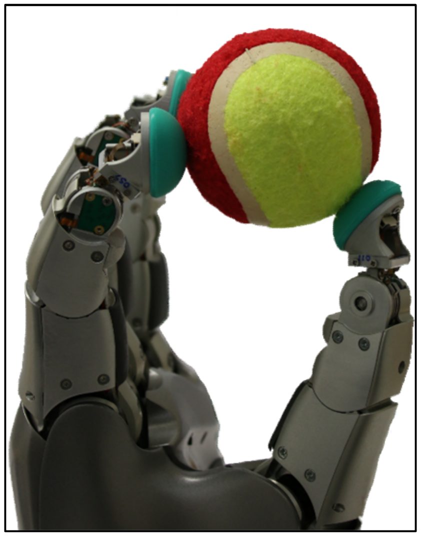
Passive, disembodied snapshots at *test* time, too

Object recognition



Scene recognition

Moving to recognize

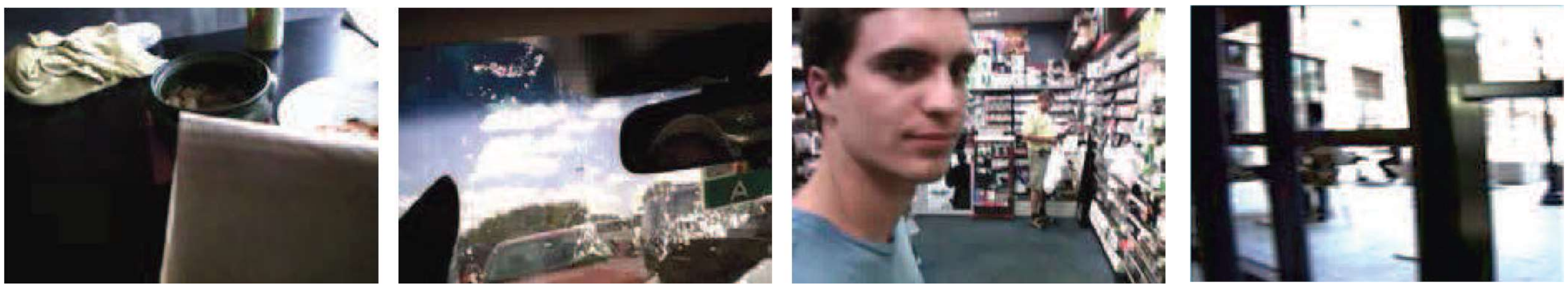


Time to revisit active recognition in challenging settings!

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



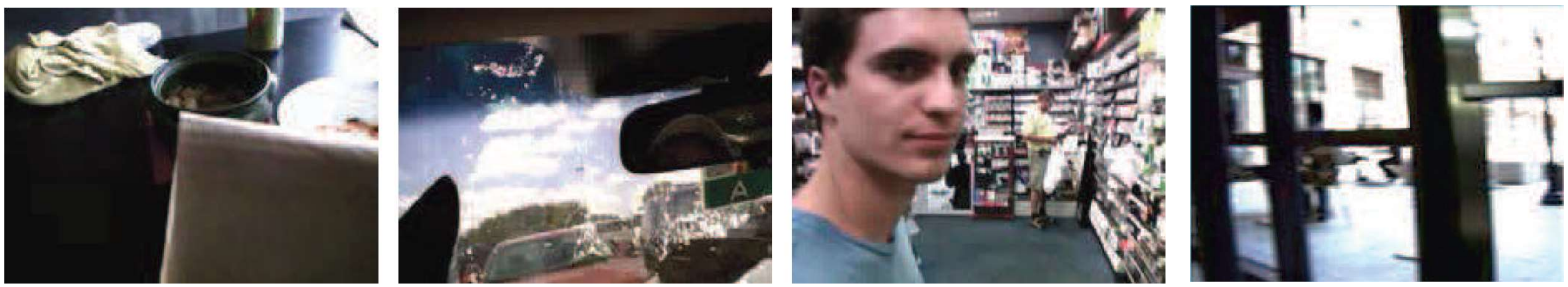
vs.

ImageNet Web images

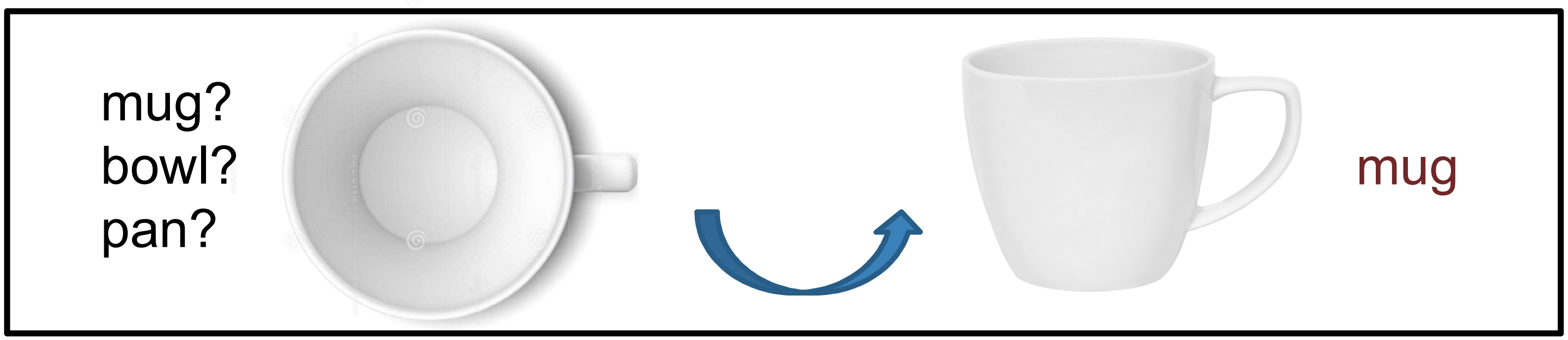
Kristen Grauman, UT Austin

Moving to recognize

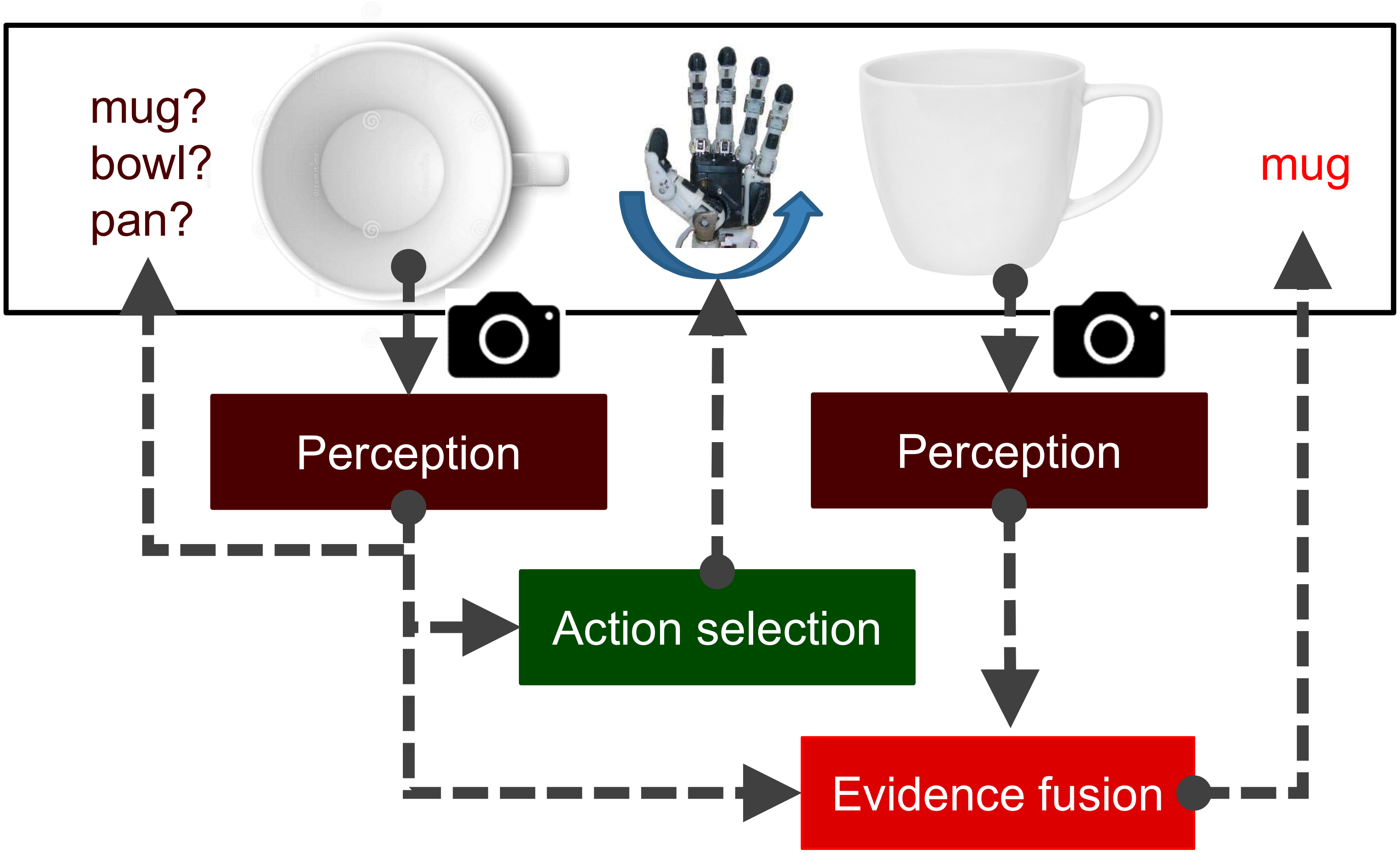
Difficulty: unconstrained visual input



Opportunity: ability to move to *change* input



Components of active recognition



Prior approaches to active recognition

Perception	Action selection	Evidence fusion

Perception

- Train for 1-view recognition

  Wilkes 1992
  Dickinson 1997
  Schiele
  Denzler
  Soatto 2
  Ramanathan 2011
  Aloimonos 2011
  Borotschnig 2011
  Wu 2015
  Jayaraman 2015
  Johns 2016

Action selection

- Navigate to a pre-selected viewpoint

  Dickinson 1997
  Schiele 1998

  Borotschnig 1998
  Ramanathan 2011
  Wu 2015
  Jayaraman 2015

- Reinforcement learning

  Paletta 2000,
  Malmir 2015

Evidence fusion

- Verification

  Dickinson 1997
  Schiele 1998

  Averaging

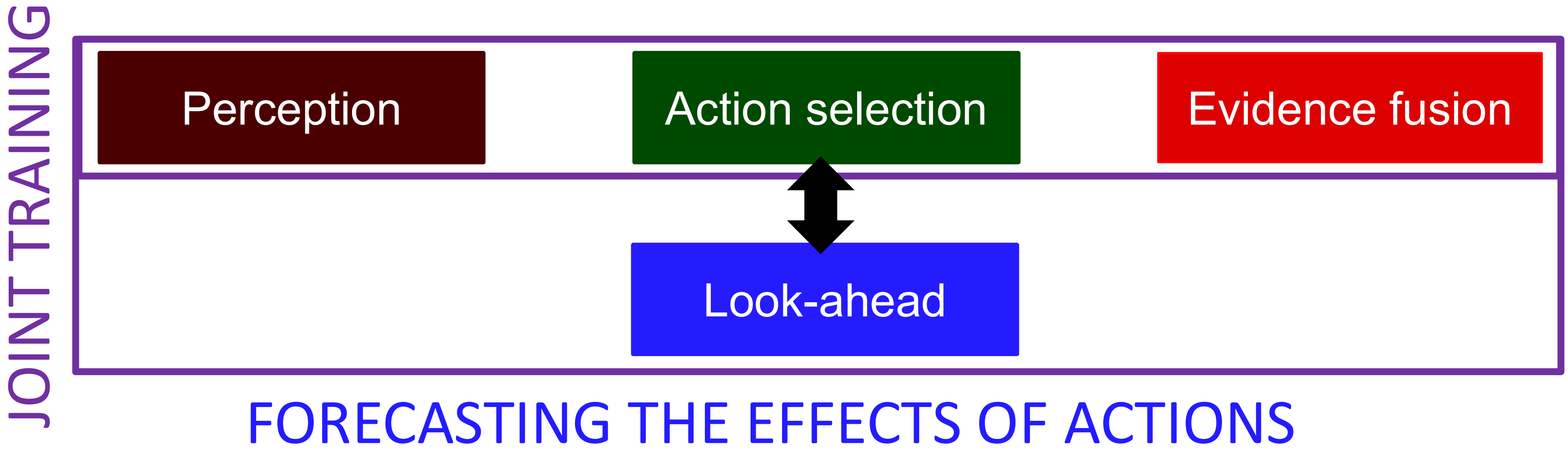
  6

- Bayes / Naïve Bayes

  Paletta 2000
  Denzler 2002
  Ramanathan 2011
  Malmir 2015

Independent solutions for the three components

Our idea: end-to-end active recognition



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

Jayaraman and Grauman, ECCV 2016

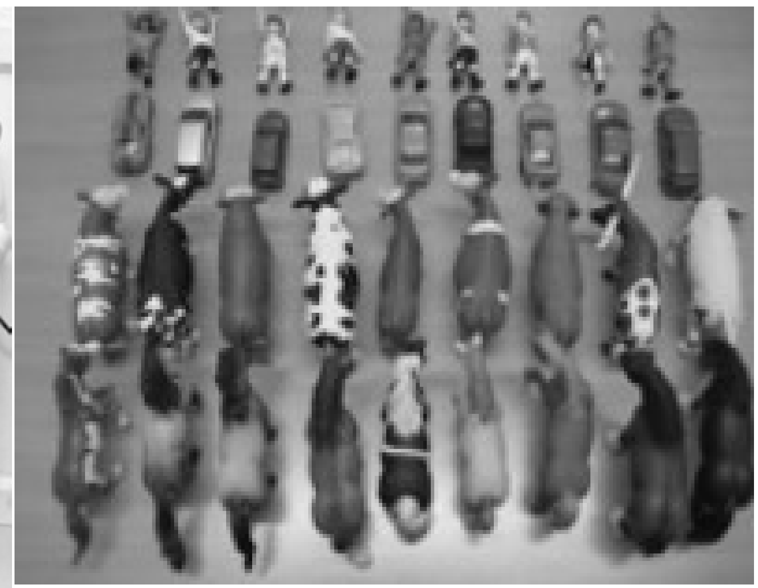
Experiments

How to **evaluate** active recognition?

Previously…

Instances, turntables

Custom robot setting

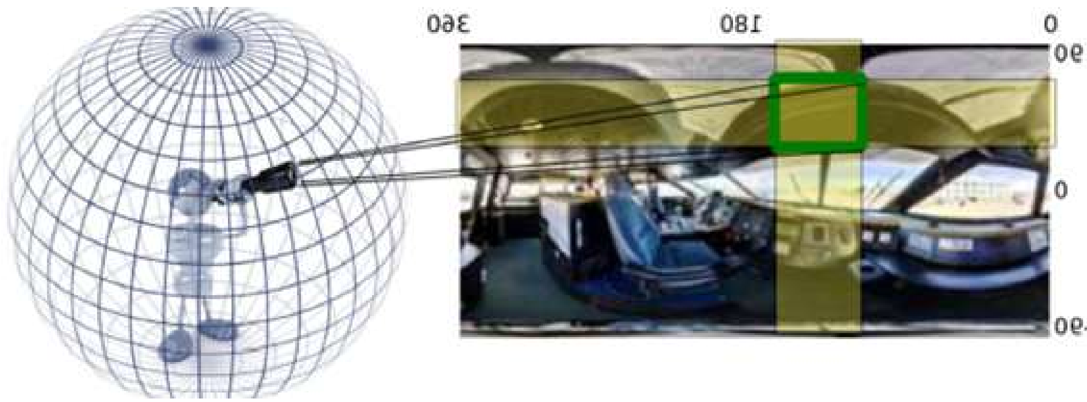


[Nene 1996, Schiele 1998, Denzler 2003, Ramanathan 2011…]

Jayaraman and Grauman, ECCV 2016

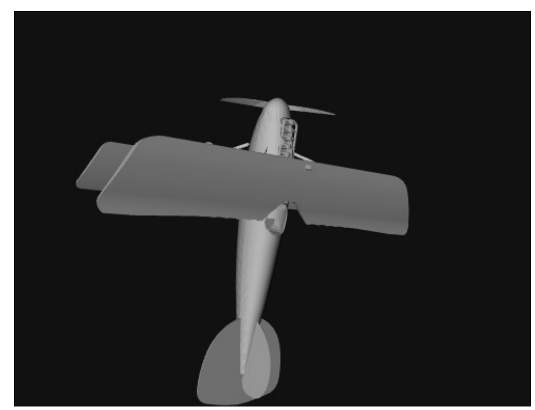
Experiments

SUN 360 panoramas [Xiao 2012]



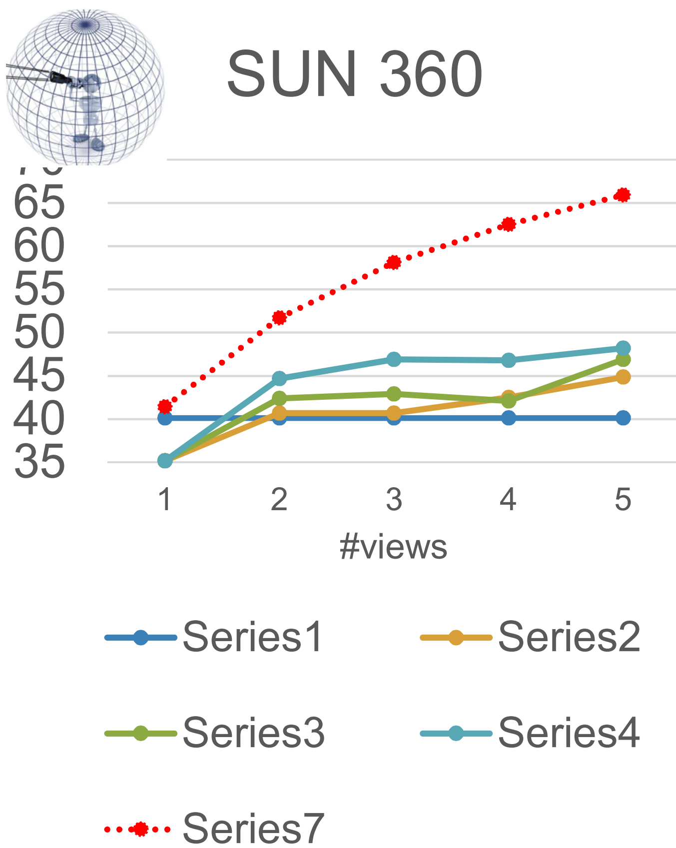
GERMS toy manipulation [Malmir 2015]

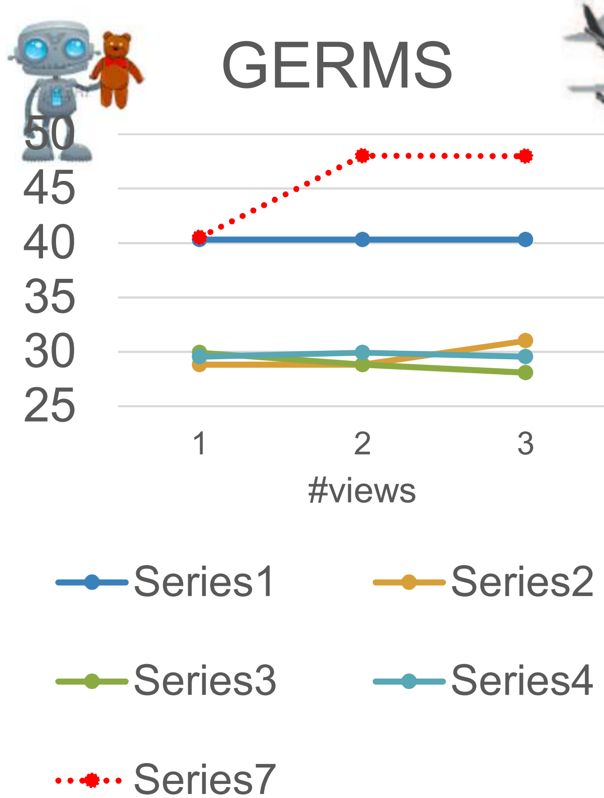
ModelNet-10 CAD models [Wu 2015]

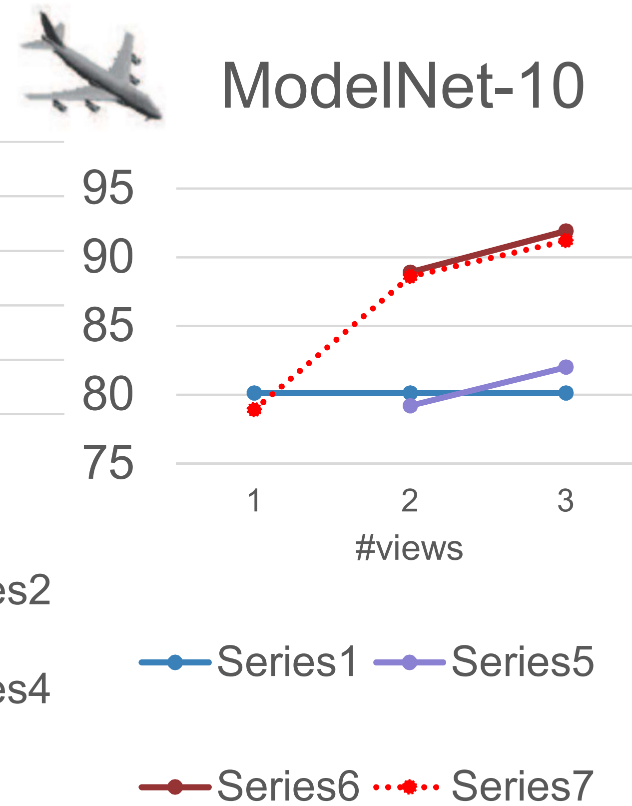


Jayaraman and Grauman, ECCV 2016

End-to-end active recognition: results







Strongly outperform traditional active recognition approaches.

Jayaraman and Grauman, ECCV 2016

Kristen Grauman, UT Austin

End-to-end active recognition: example

Top 3 guesses:

Restaurant / Forest	(15.05) / (1.90)	(37.88) / (68.89)
Train interior / Cave	Theater / Street	Plaza courtyard / Church
Beach / Shop	Restaurant / Cave	Lobby atrium / Street
	Plaza courtyard	Theater / Street

[Jayaraman and Grauman, ECCV 2016]

End-to-end active recognition: example

Predicted label:

T=1 T=2 T=3

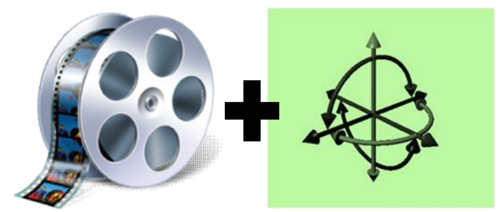
GERMS dataset: Malmir et al. BMVC 2015

[Jayaraman and Grauman, ECCV 2016]

Talk overview

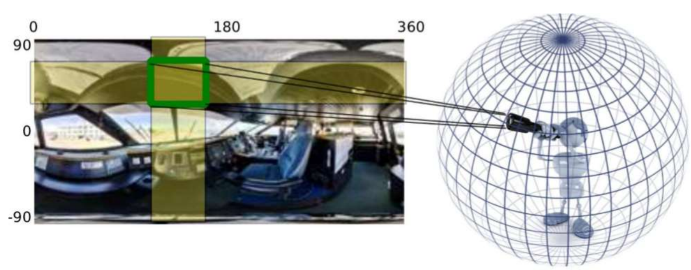
Towards embodied visual learning

1. Learning representations tied to ego-motion



2. Learning representations from unlabeled video

3. Learning how to move and where to look



360° cameras and panoramic video

Kristen Grauman, UT Austin

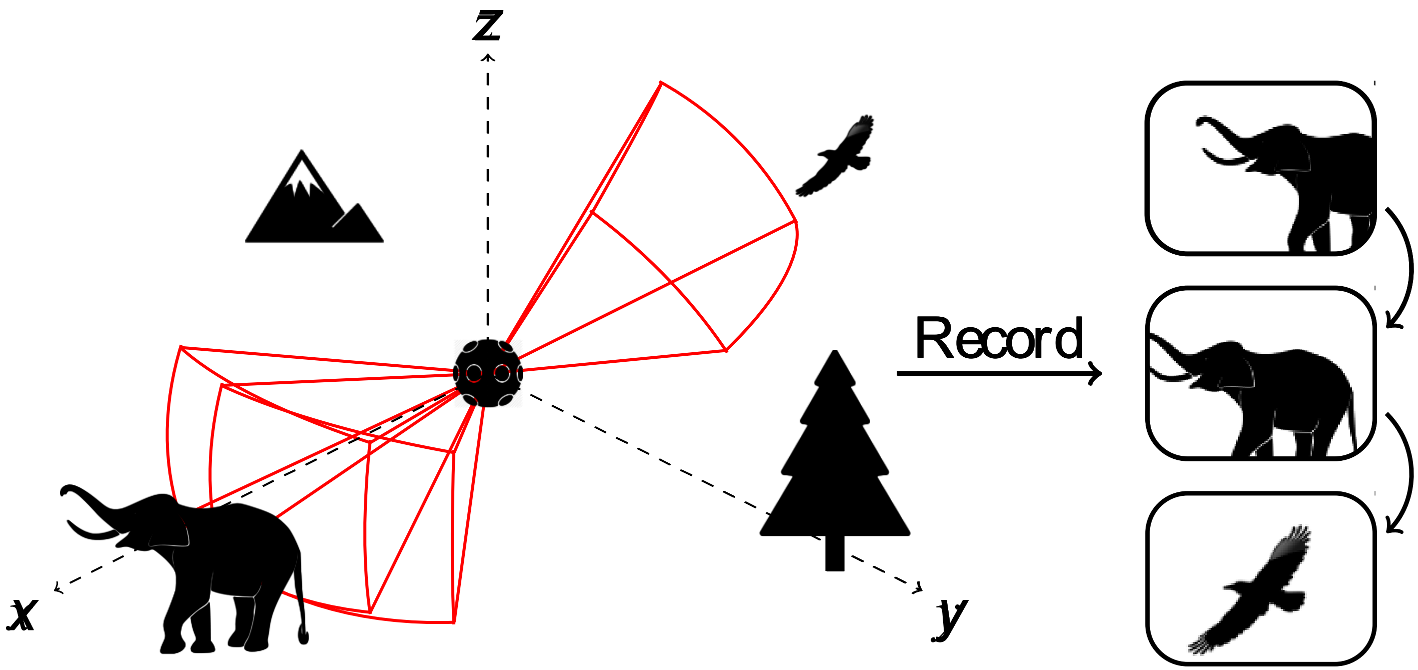
Challenge of viewing 360° videos

Control by mouse

How to find the right direction to watch?

Kristen Grauman, UT Austin

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]

New problem:
Pano2Vid automatic videography

Virtual camera direction

Input:
360° Video

Output:
normal-field-of-view
(NFOV) Video

[Su et al. ACCV 2016]

Our approach – AutoCam

Learn videography tendencies from unlabeled Web videos

- Diverse capture-worthy content
- Proper composition

Human-captured NFOV videos (“HumanCam”)

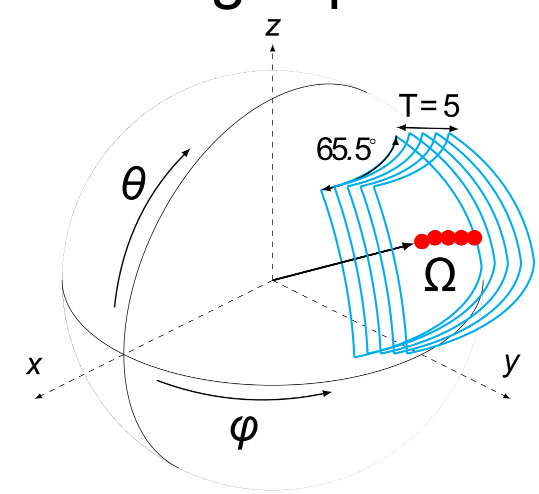
YouTube

vimeo

LIVE house.in

How close?

ST-glimpses



Unlabeled video

[Su et al. ACCV 2016]

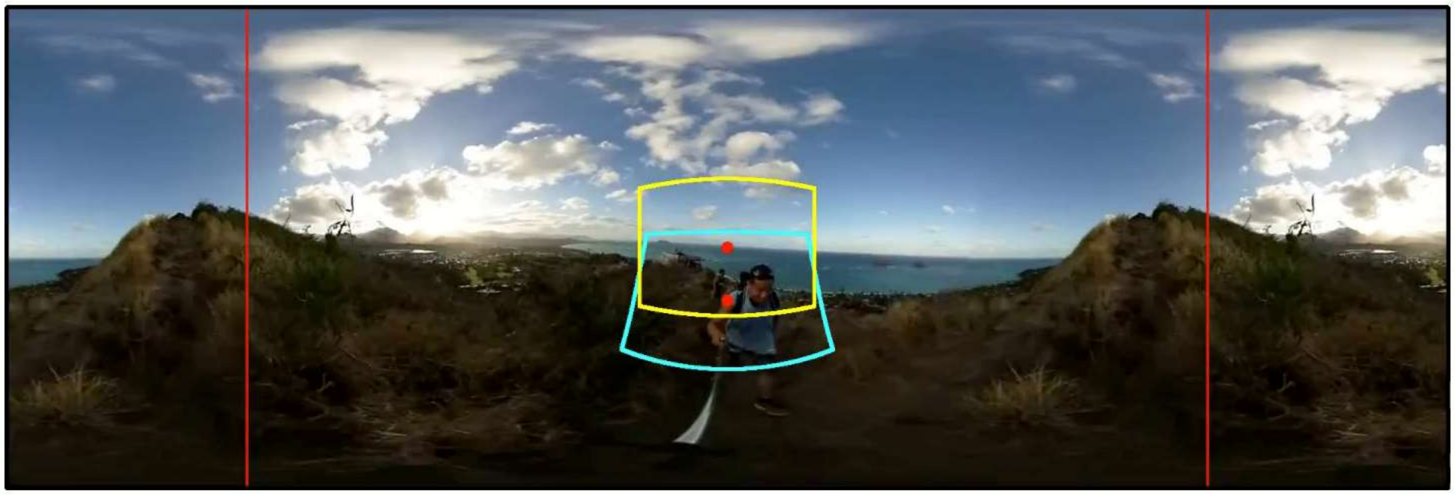
Example AutoCam Output 1

Input 360° Video + Camera Trajectory

AutoCam
Output Video

[Su et al. ACCV 2016]

Example AutoCam Output 2



AutoCam

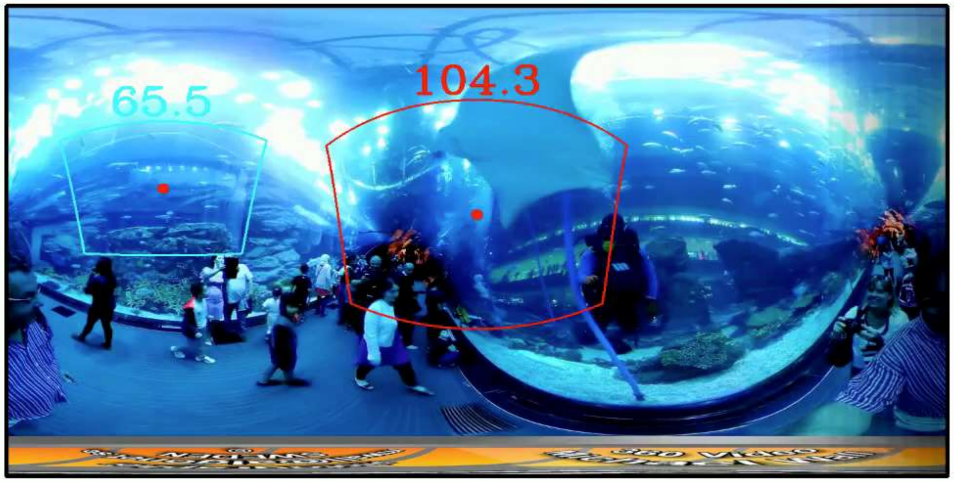
Eye-level Prior

Kristen Grauman, UT Austin

[Su et al. ACCV 2016]

Example AutoCam Output 3

Input 360° Video
+
Camera Trajectories



With
Zooming

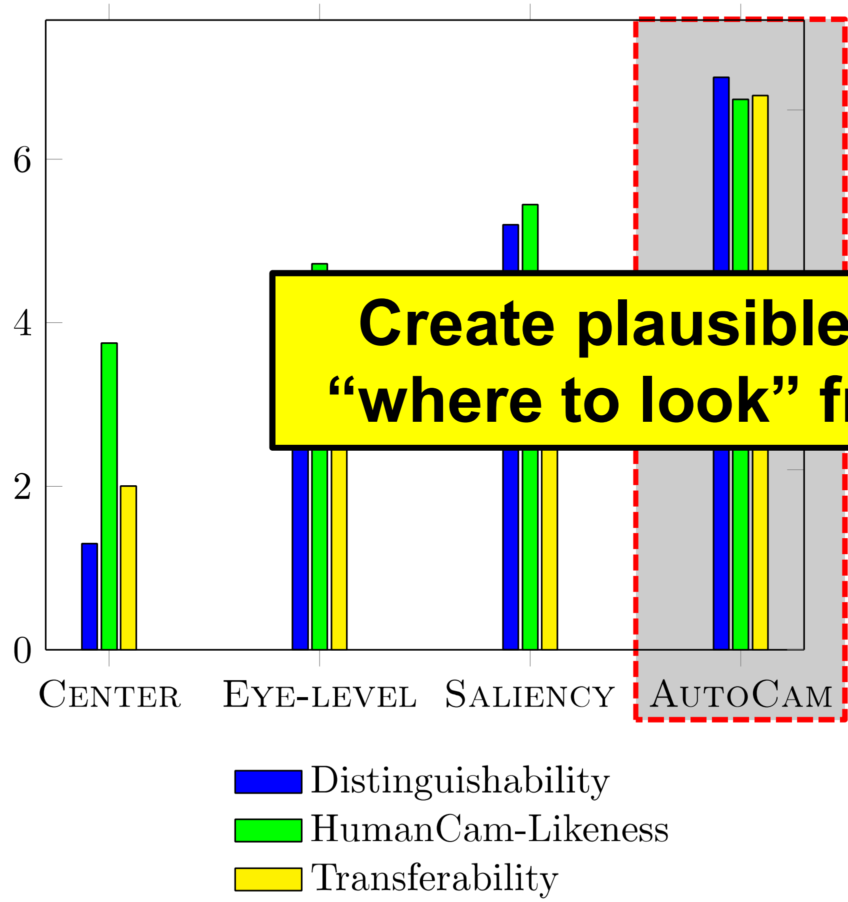
Without
Zooming

Kristen Grauman, UT Austin

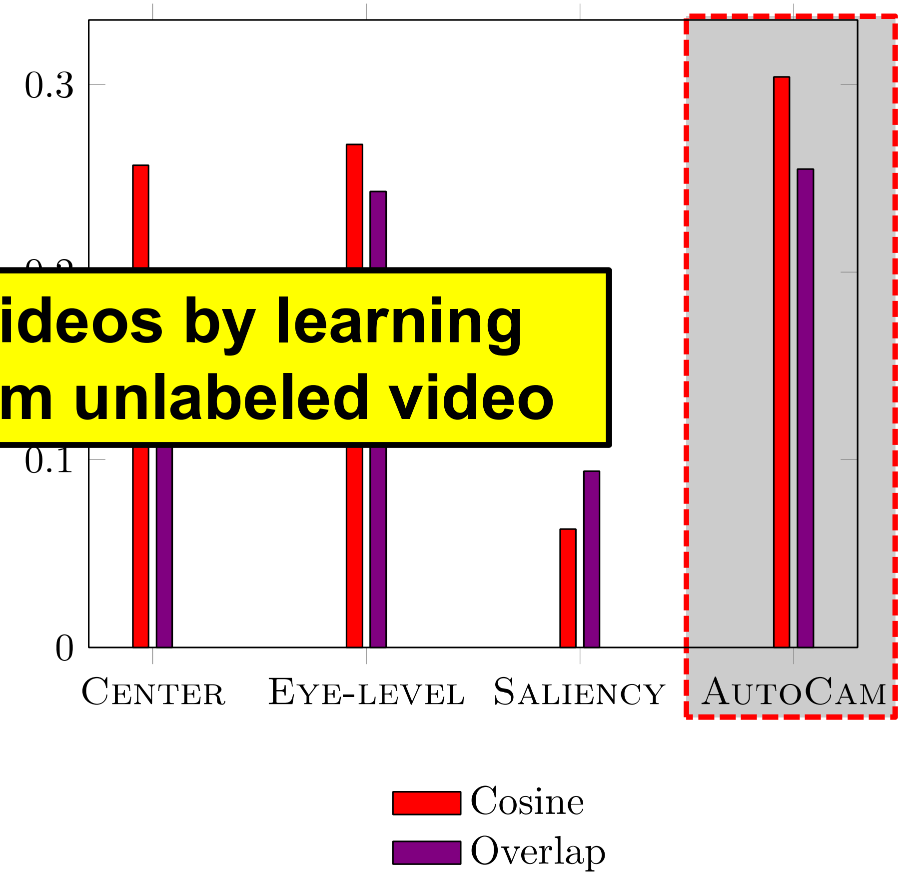
[Su et al. ACCV 2016]

Results: Quantitative evaluation

Similarity to user-uploaded standard web videos



Similarity to human-selected camera trajectories



Create plausible videos by learning "where to look" from unlabeled video

Kristen Grauman, UT Austin

[Su et al. ACCV 2016]

Next steps

- Active observations for representation learning
- Explore varied space of egomotions
- Multi-sensor active recognition
- Learning where to look +/- recognition
- 360 video summaries

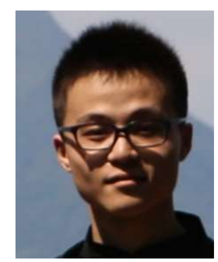
Summary

- 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`