CVPR 2016 Scene Understanding Workshop (SUNw)

Action and Interaction for Scene Understanding

Kristen Grauman

University of Texas at Austin

Dinesh Jayaraman

THE UNIVERSITY OF TEXAS AT AUSTIN

Chao-Yeh Chen

Outline

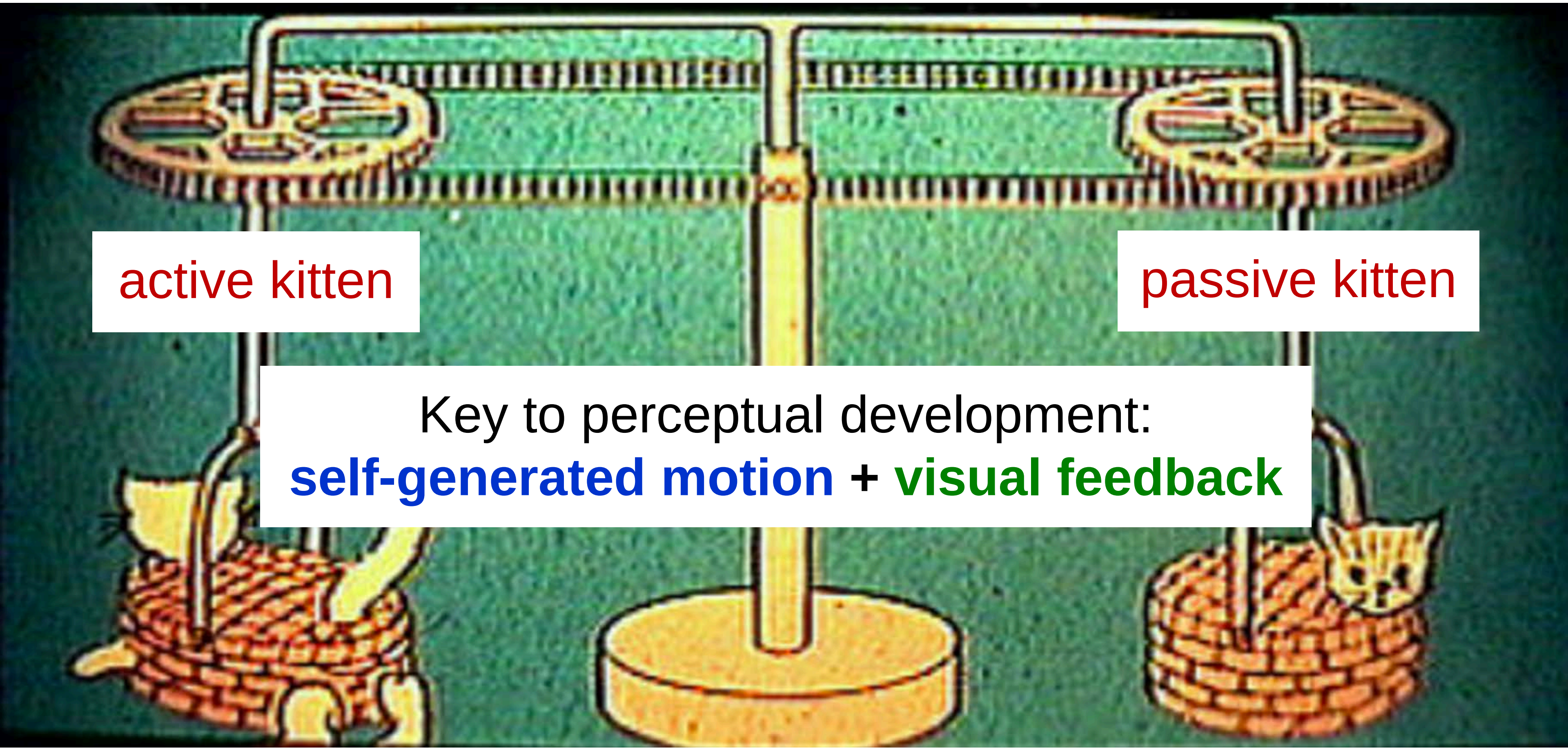
Action and interaction for scene understanding

1. Learning by moving about a scene

2. Learning how to best move about a scene

3. Open world “interactee” localization

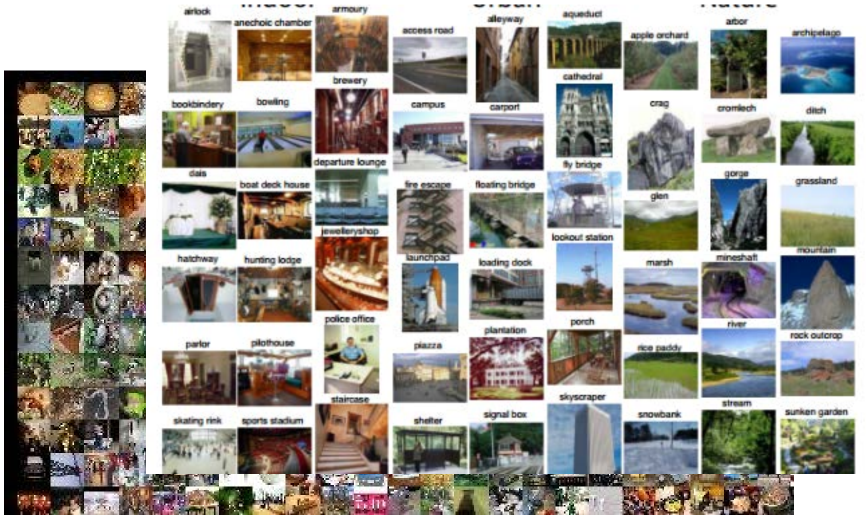
The kitten carousel experiment
[Held & Hein, 1963]



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.

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]

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"

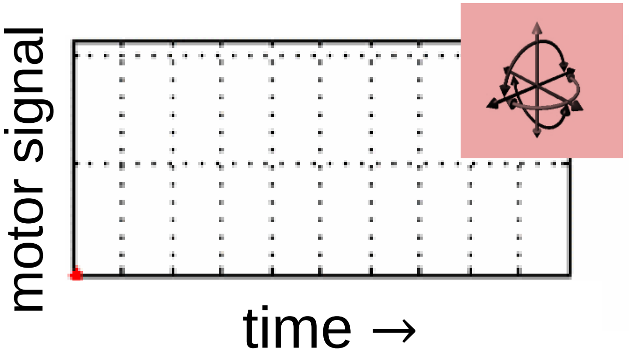
$$\mathbf{z}(g\mathbf{x}) \approx M_g \mathbf{z}(\mathbf{x})$$

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]

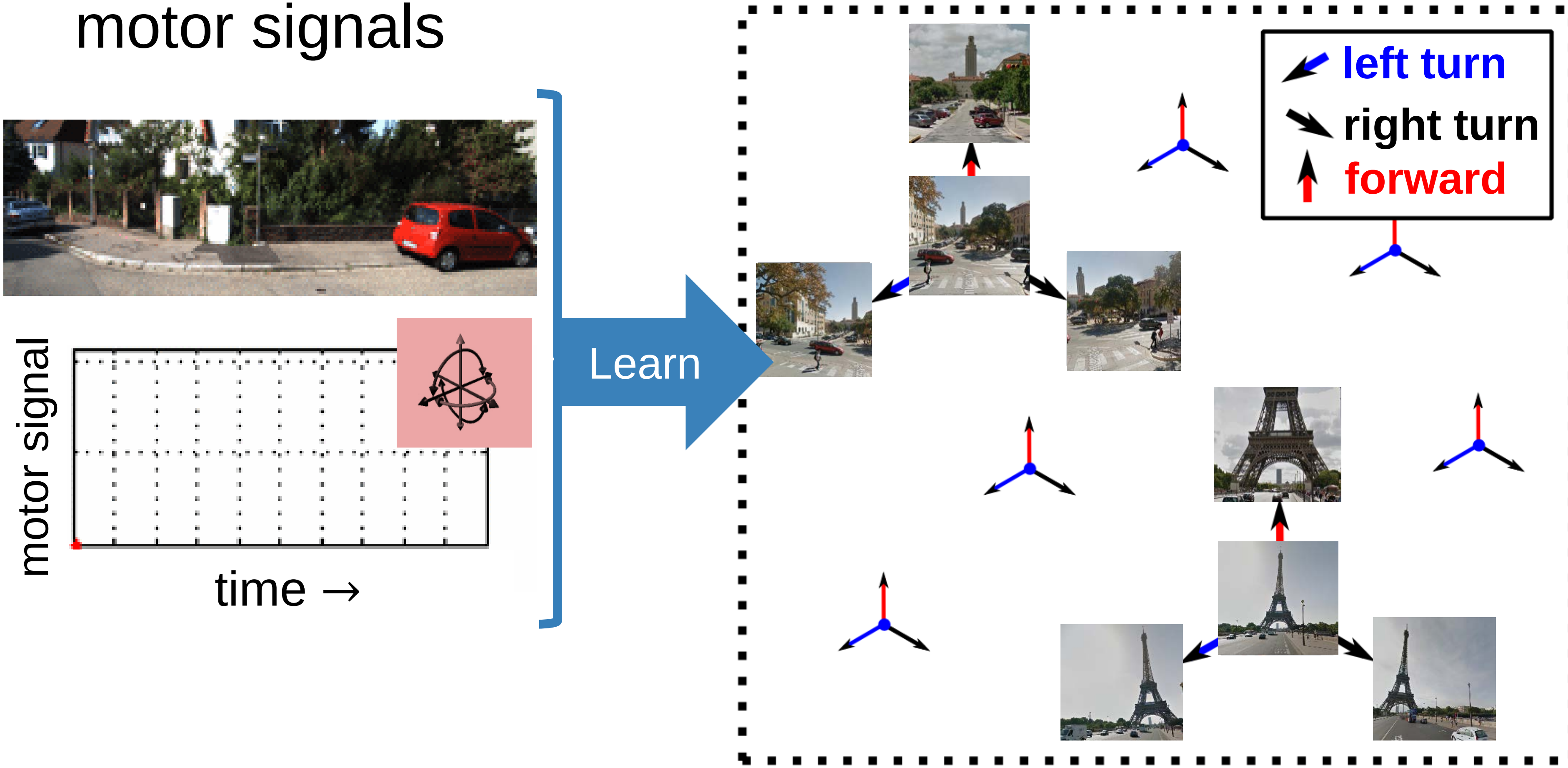
Approach idea: Ego-motion equivariance

Training data

Unlabeled video + motor signals

Equivariant embedding

organized by ego-motions

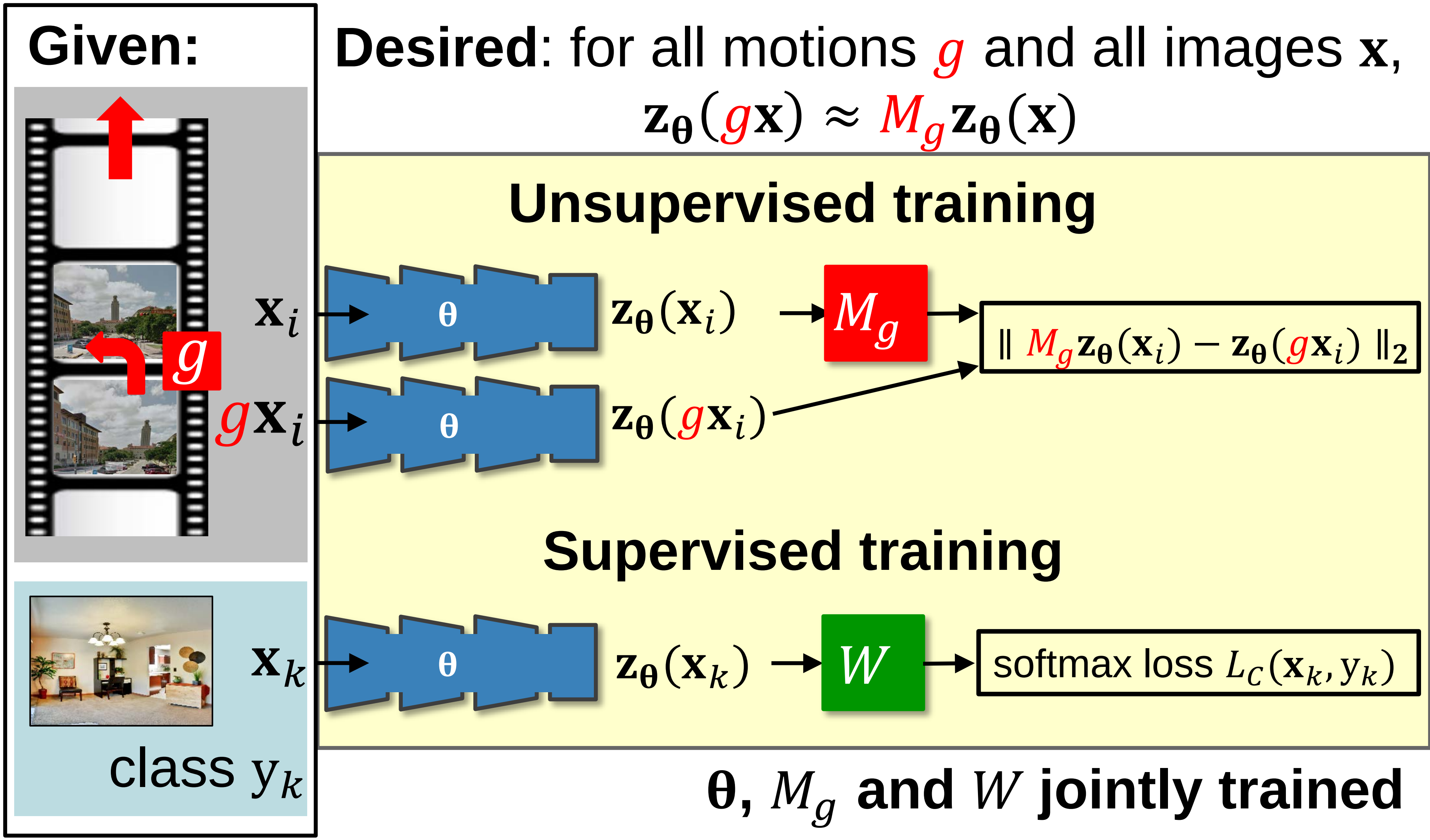


[Jayaraman & Grauman, ICCV 2015]

Ego-motion equivariant feature learning

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

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



[Jayaraman & Grauman, ICCV 2015]

Results: Recognition

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

Geiger et al, IJRR ’13

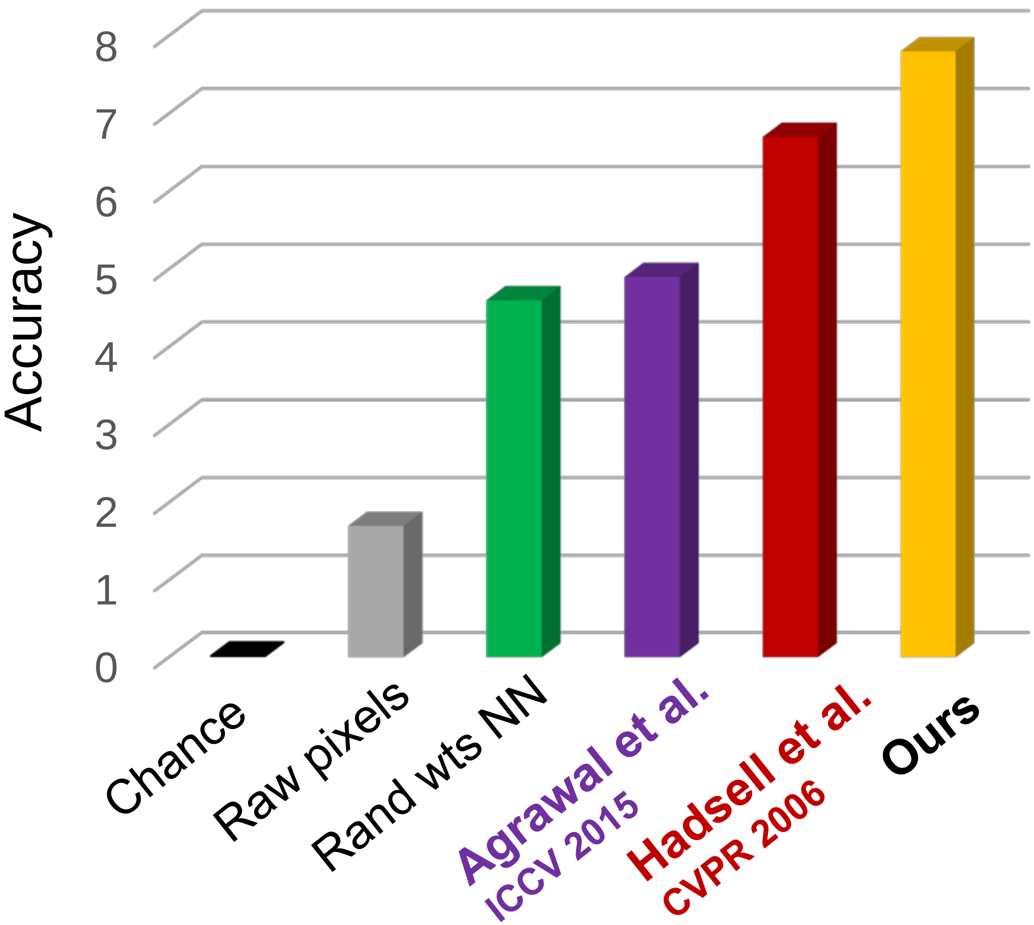
Exploit features for **static scene classification**
(SUN, 397 classes)

Xiao et al, CVPR ’10

Results: Recognition

Purely unsupervised feature learning

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

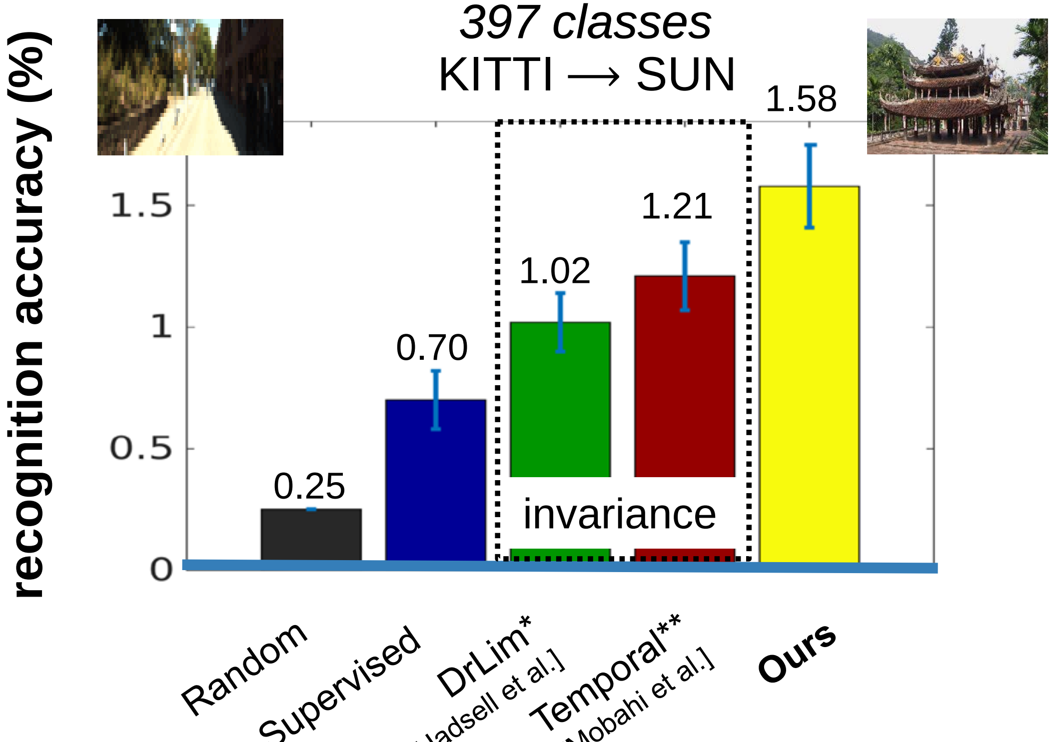


Agrawal, Carreira, Malik, Learning to see by moving. ICCV 2015

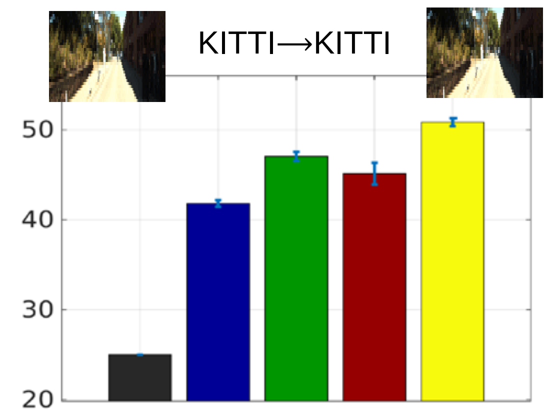
Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping. CVPR 2006

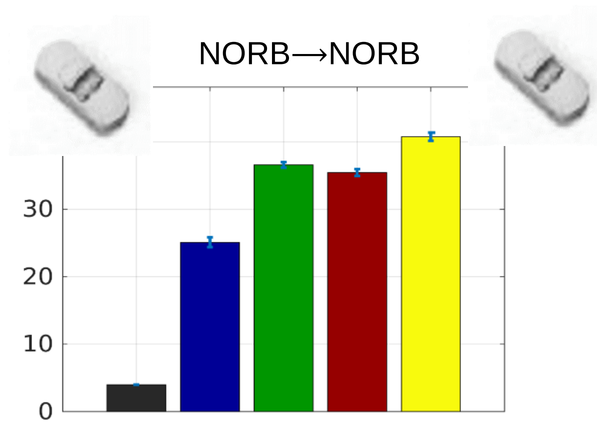
Results: Recognition

Ego-motion equivariance as a regularizer



6 labeled training examples per class



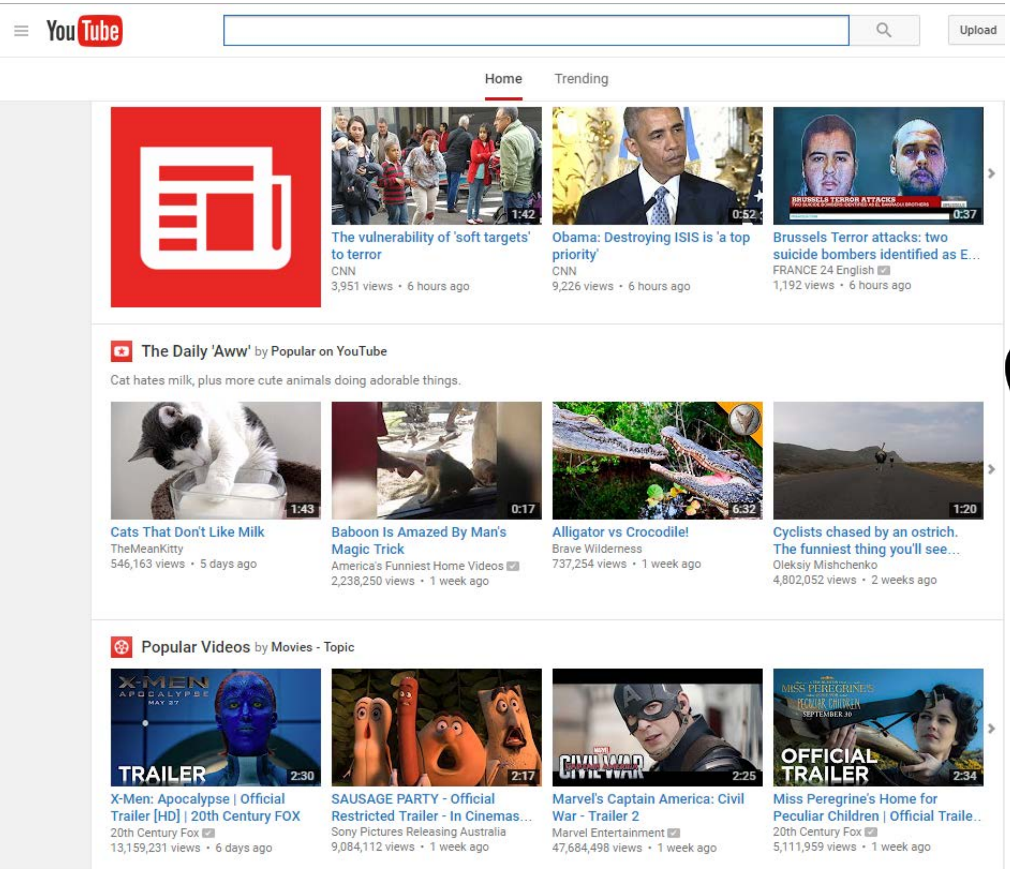


Up to 30% accuracy increase over state of the art!

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

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

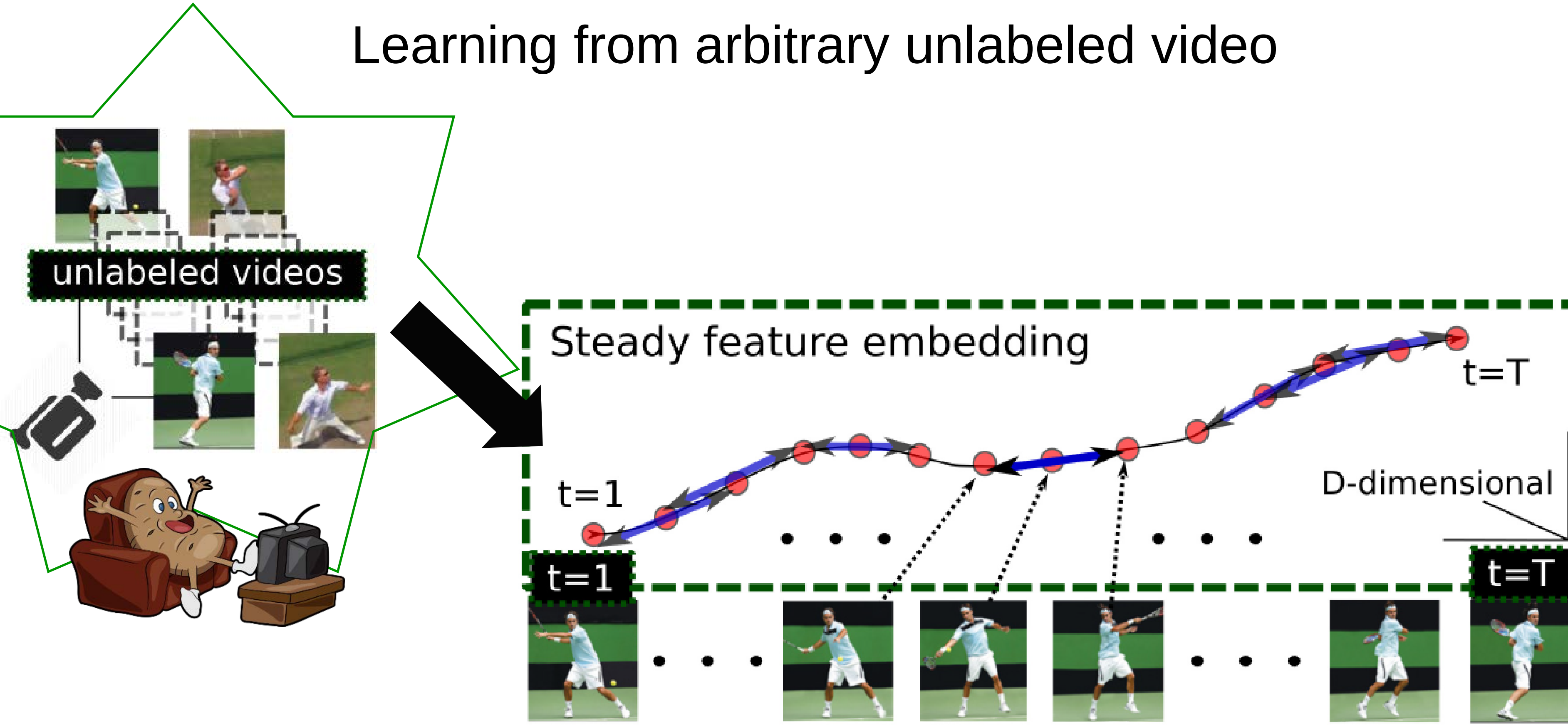
Learning from arbitrary unlabeled video?



Unlabeled video

Our idea: Steady feature analysis

Learning from arbitrary unlabeled video



Equivariance ≈ “steadily” varying frame features!

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

[Jayaraman & Grauman, CVPR 2016]

Our idea: Steady feature analysis

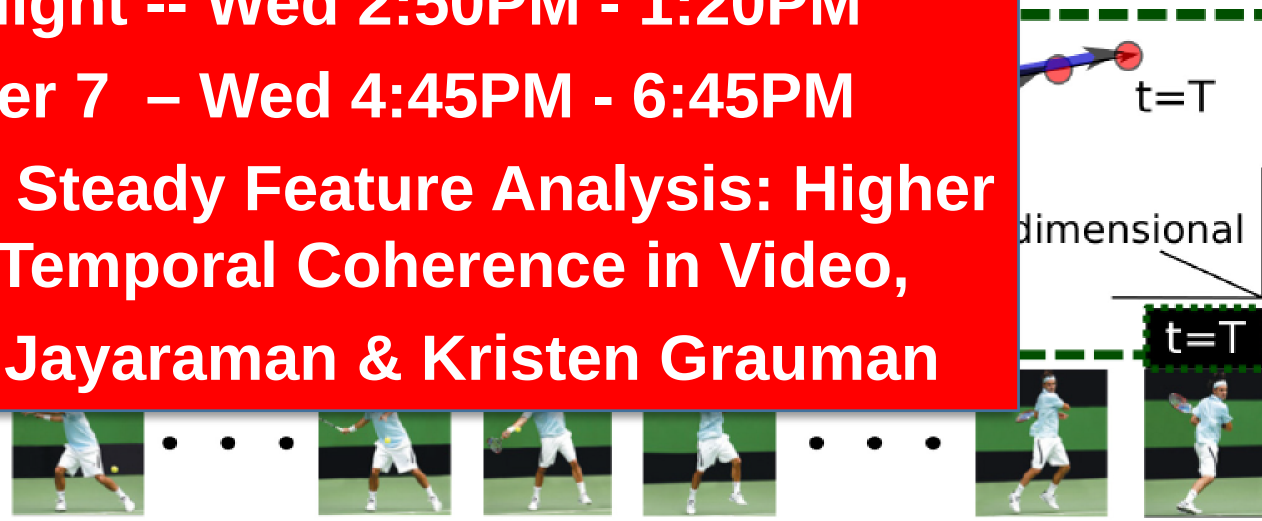
Learning from arbitrary unlabeled video

Spotlight -- Wed 2:50PM - 1:20PM

Poster 7 – Wed 4:45PM - 6:45PM

Slow and Steady Feature Analysis: Higher Order Temporal Coherence in Video,

Dinesh Jayaraman & Kristen Grauman



Equivariance ≈ "steadily" varying frame features!

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

[Jayaraman & Grauman, CVPR 2016]

Outline

Action and interaction for scene understanding

1. Learning by moving about a scene
2. Learning how to best move about a scene
3. Open world “interactee” localization

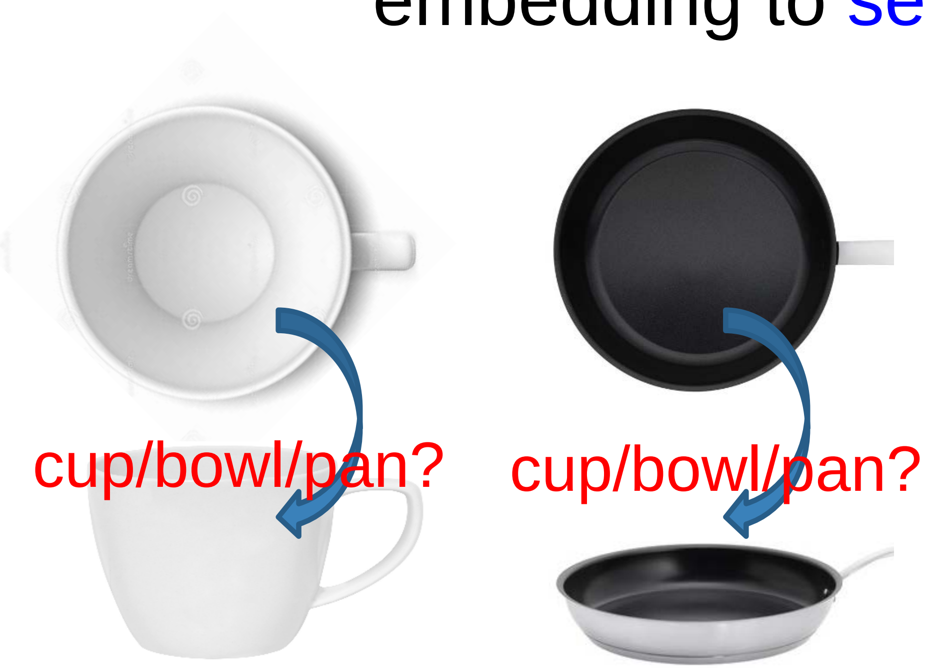
Learning how to move for recognition

Time to revisit active recognition in challenging settings!

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

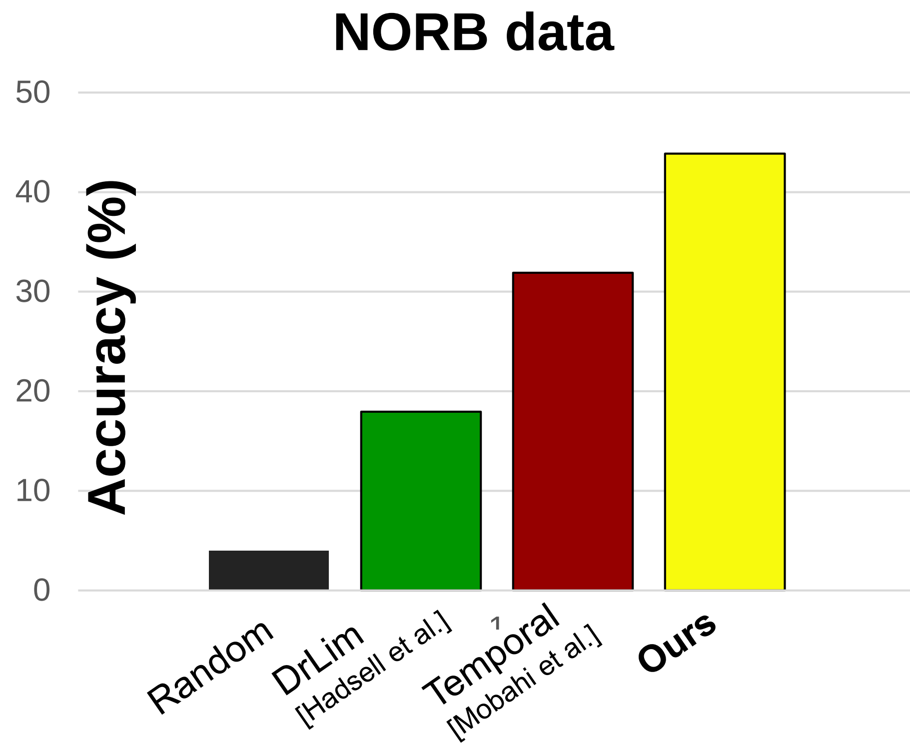
Learning how to move for object recognition

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



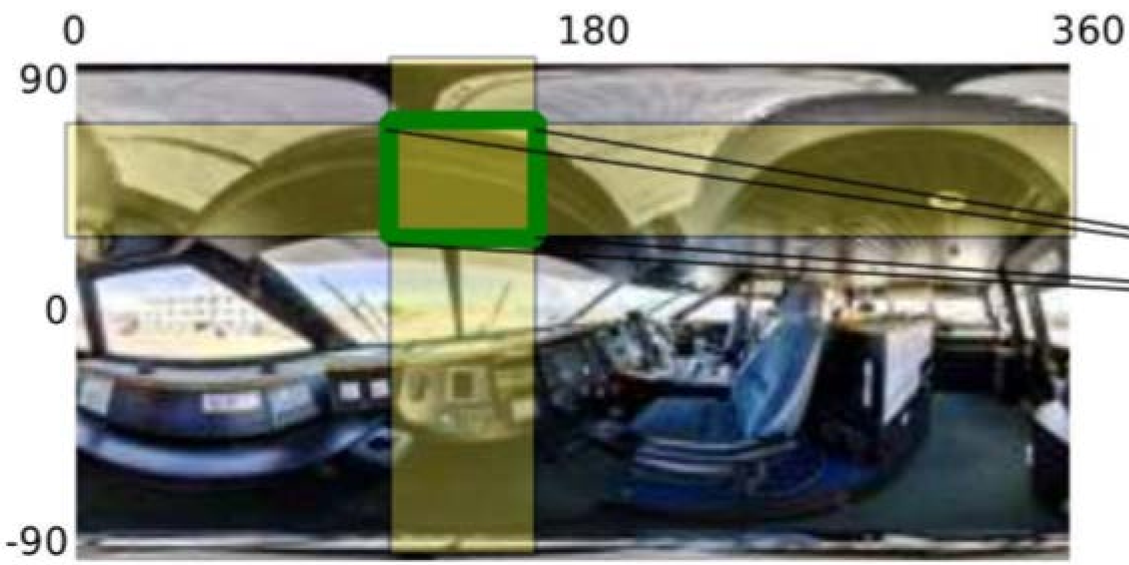
cup

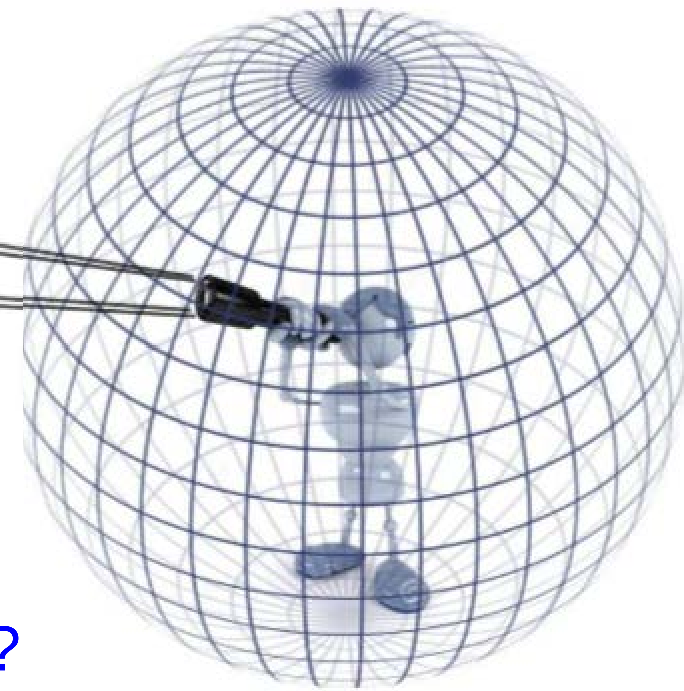
frying pan



[Jayaraman & Grauman, ICCV 2015]

Learning how to move for scene 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

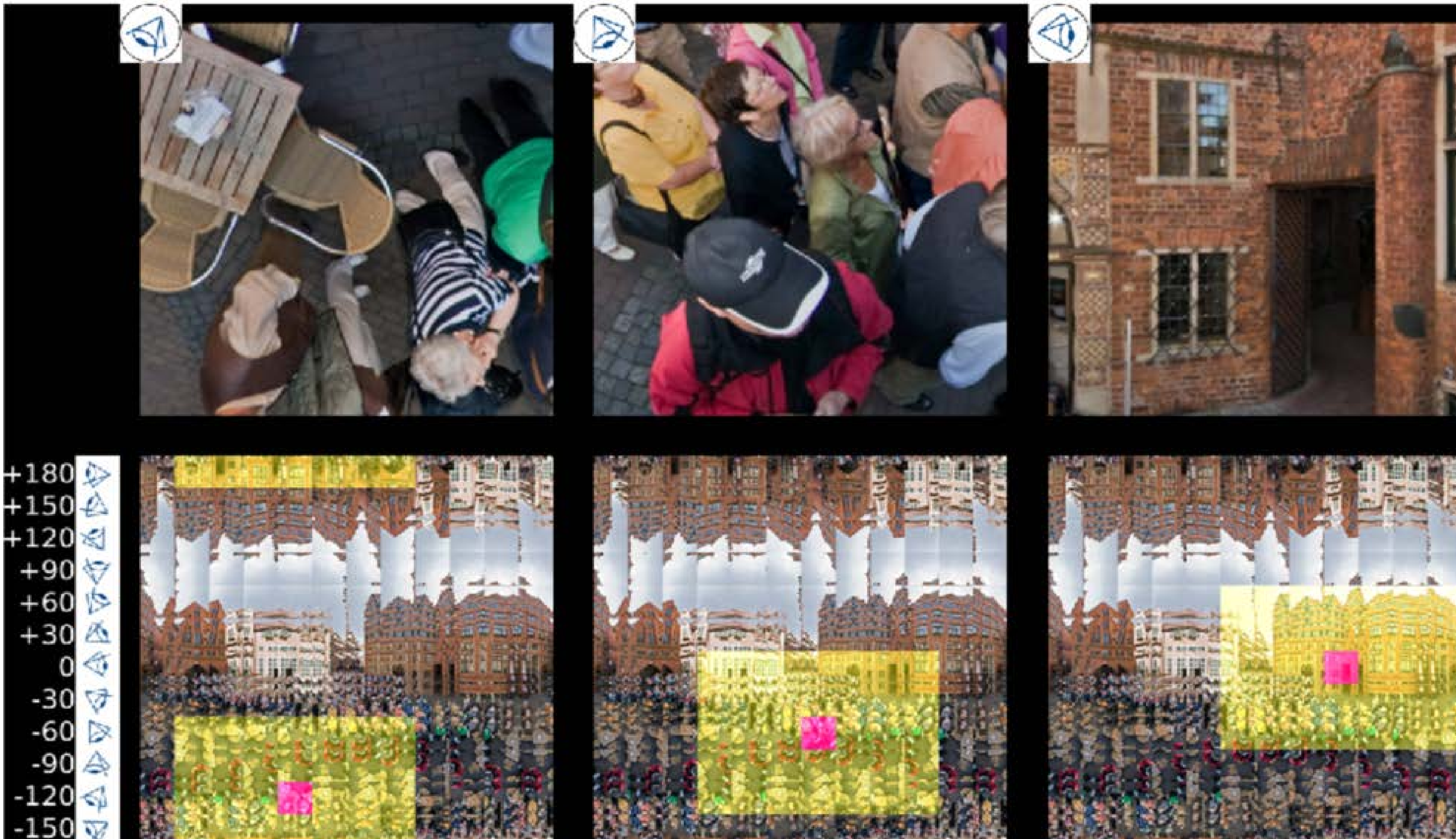
Active recognition: results

P("Plaza courtyard"): (6.28) (11.95) (67.38)

P("Church"): (0.58) (5.09) (38.89)

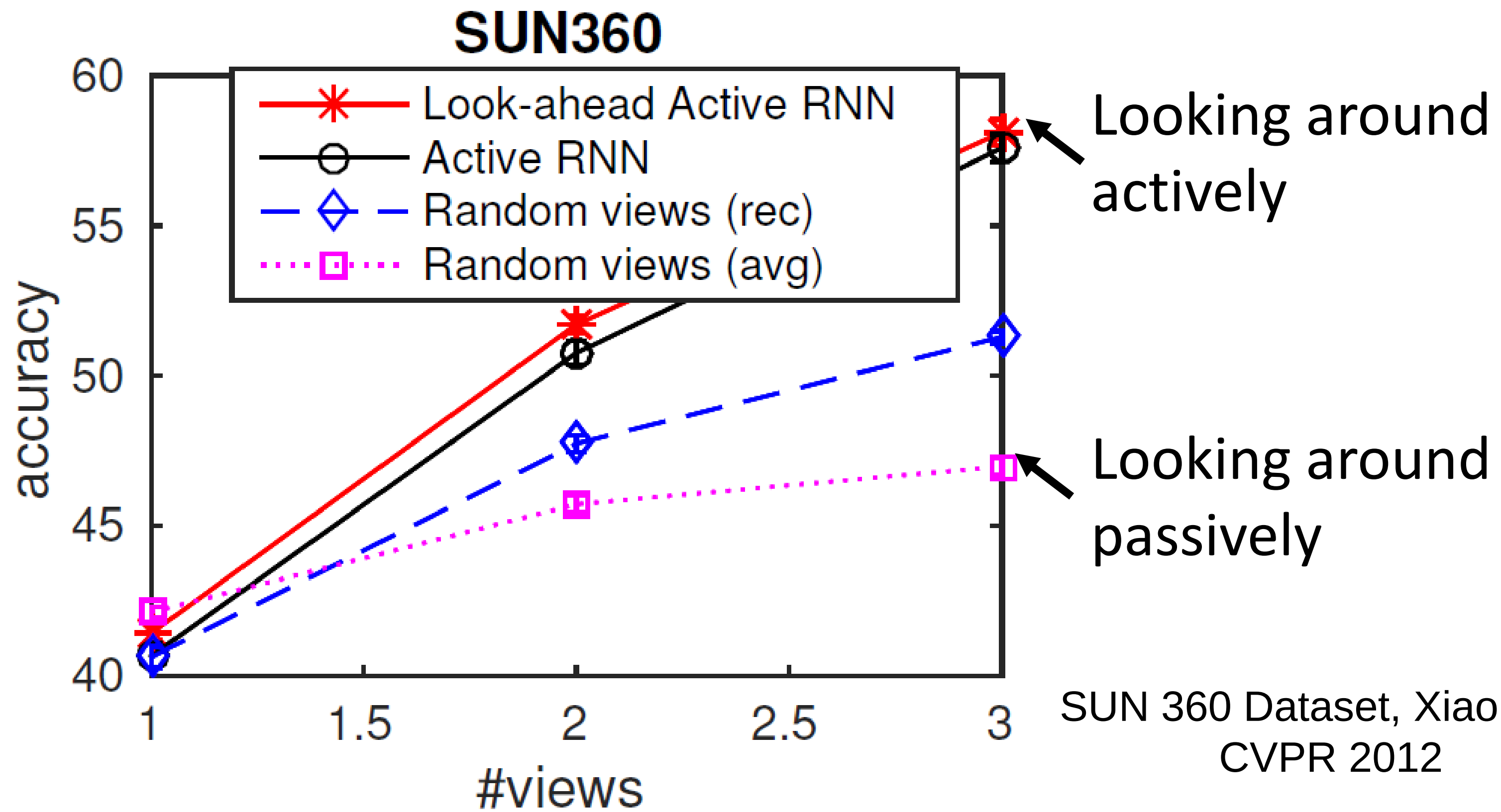
Top 3 guesses:

Restaurant / Forest	Theater / Street	Plaza courtyard / Church
Train interior / Cave	Restaurant / Cave	Lobby atrium / Street
Shop / Beach	Plaza courtyard	Theater / Street



Jayaraman and Grauman, UT TR AI15-06

Active recognition: results



SUN 360 Dataset, Xiao et al. CVPR 2012

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

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

Outline

Action and interaction for scene understanding

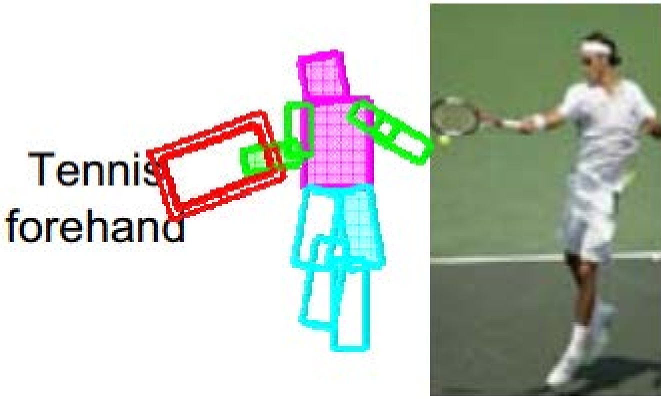
1. Learning by moving about a scene
2. Learning how to best move about a scene
3. Open world “interactee” localization

Understanding scenes with people

Prior work: human-object interactions

- Objects and actions/poses offer mutual context

  [Peursum et al 2005, Gupta et al 2009, Desai et al 2010, Yao and Fei-Fei 2010, Ikizler-Cinbis and Sclaroff 2010, Farhadi and Sadeghi 2011, Prest et al 2012, Delaitre et al 2012, Chao et al 2015]



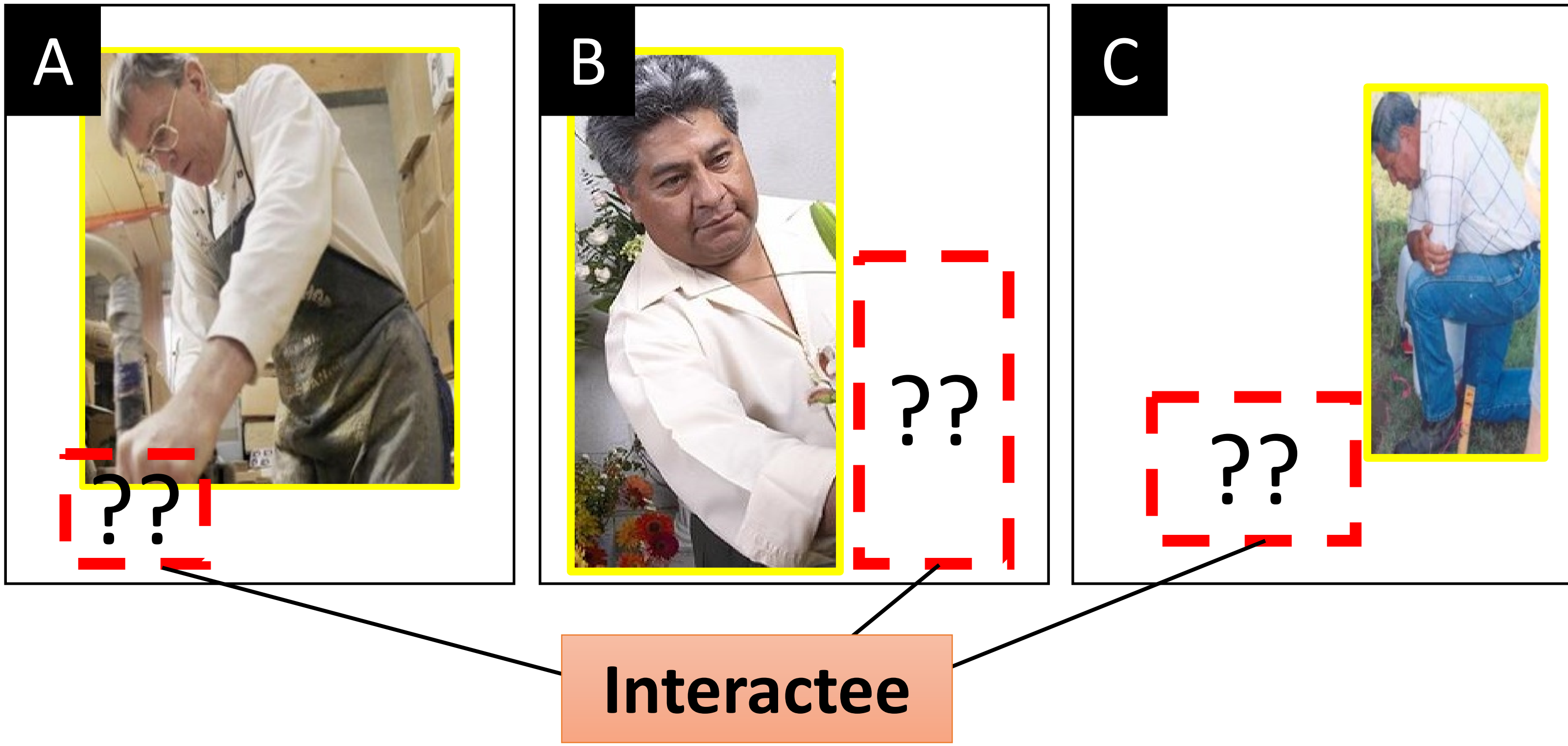
Yao et al. 2010

Desai et al. 2010

Closed-world models: learn about specific action/object pairings

Our goal: Interactee detection

Localize "interactee" object, in the open world setting



Definition:

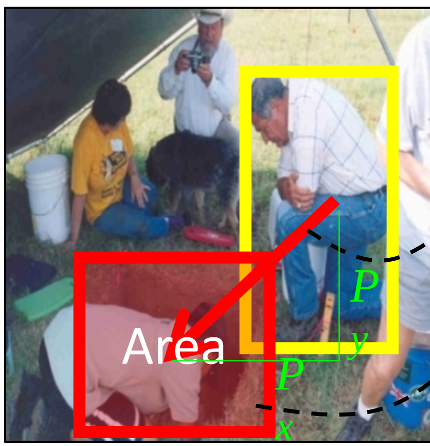
- Touched by the subject with a specific purpose.
- Watched by the subject with specific attention paid to it.

[Chen & Grauman, ACCV 2014]

Approach: Learning to localize interactees

Target output space:

Relative position and area of the interactee's bounding box



(p_x, p_y) ➔ Relative position to the person

a ➔ Area of the interactee

Normalize by person's height+width

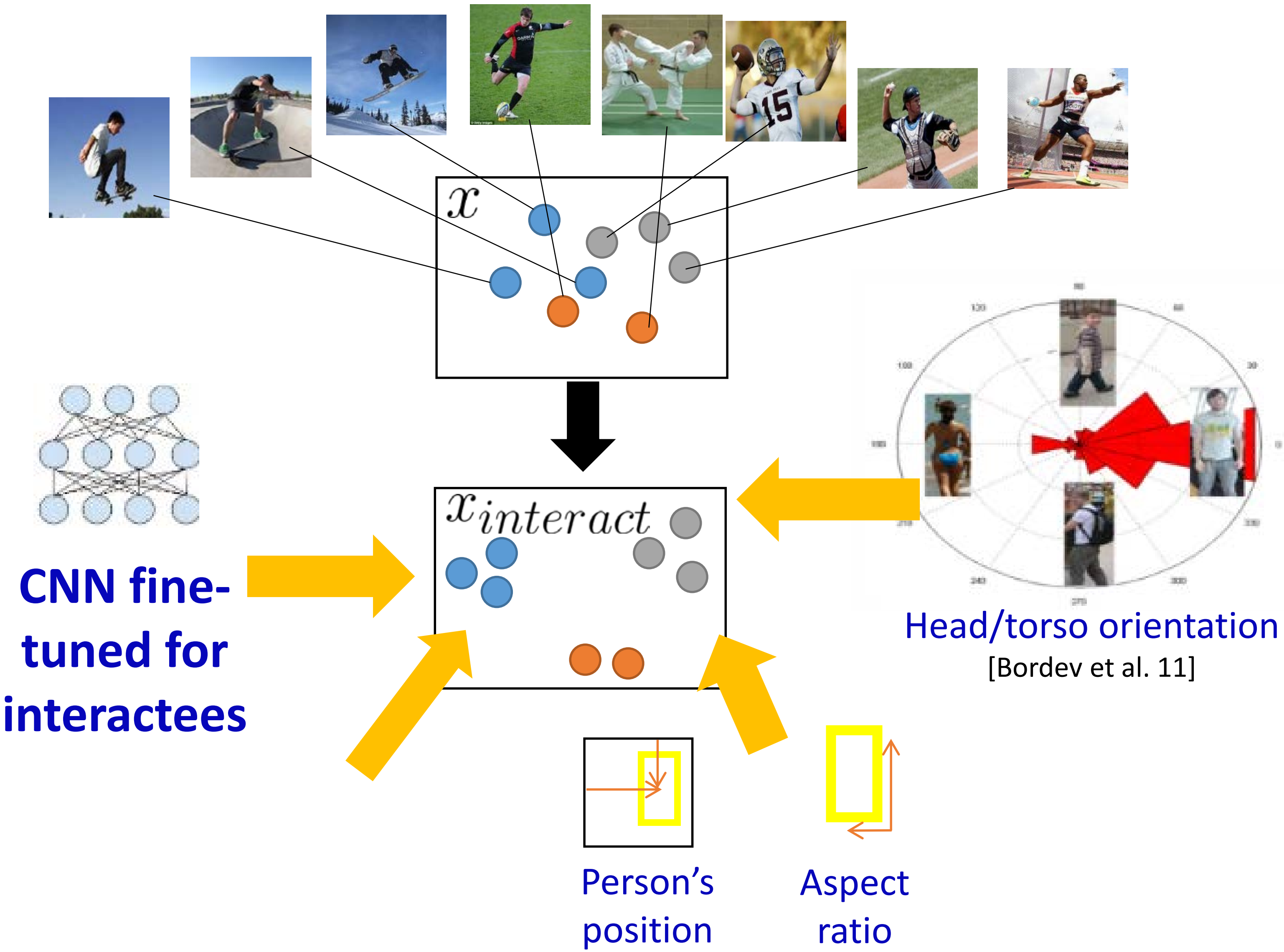
⬇

$$y = [p_x, p_y, a]$$ ➔ **Interactee localization**

[Chen & Grauman, ACCV 2014]

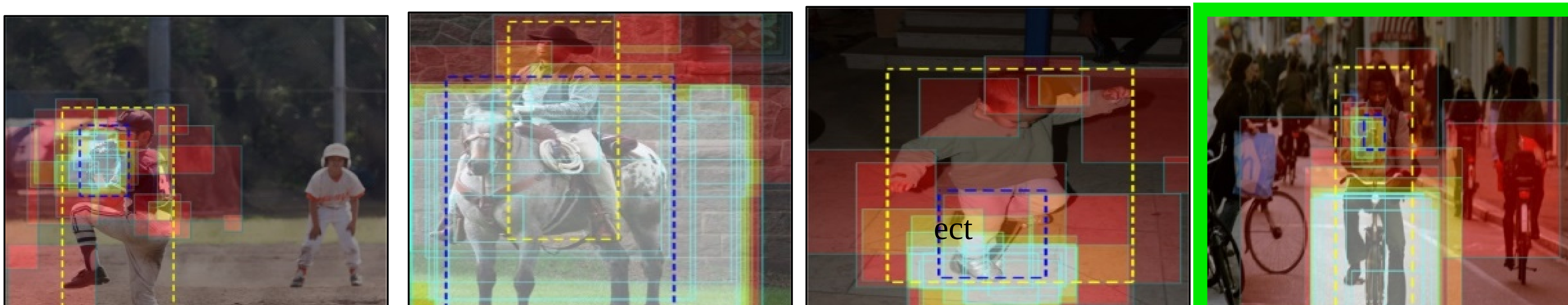
Approach: Learning to localize interactees

Interaction-guided embedding + locally weighted regression



Results: interactee detection

Failures



Metric	Dataset	Ours-embedding (w/CNN)	Obj (Alexe et al 2010)	Near Person	Random
Position error	COCO	**0.2256**	0.3569	0.2909	0.5760
	PASCAL	**0.1632**	0.2982	0.2034	0.5038
	SUN	0.2524	0.4072	0.2456	0.6113
Size error	COCO	**38.17**	263.57	65.12	140.13
	PASCAL	**27.04**	206.59	31.97	100.31
	SUN	**33.15**	257.25	39.51	126.64
IOU	COCO	**0.1989**	0.0824	0.1213	0.0532
	PASCAL	**0.2177**	0.0968	0.1415	0.0552
	SUN	**0.1710**	0.1006	0.1504	0.0523

System has no object detector for the highlighted objects!

Tasks leveraging interactees

1

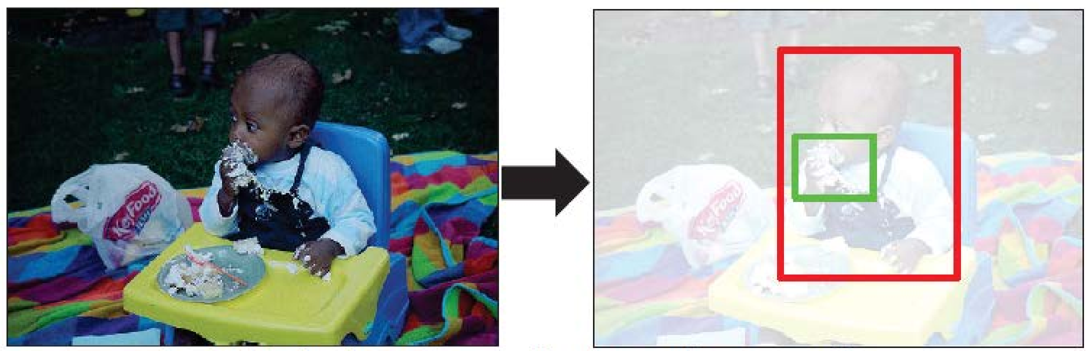
Prior for “what to mention” about the scene

Method	Mention rate (%)
Ground truth interactee	78.4 (0.6)
Ours-embedding	**70.5 (0.4)**
Importance (Berg et al 2012)	65.4 (0.4)
Ours-MDN	65.2 (0.5)
Near Person	67.5 (0.5)
Prior	64.6 (0.6)
Majority	51.7 (0.6)

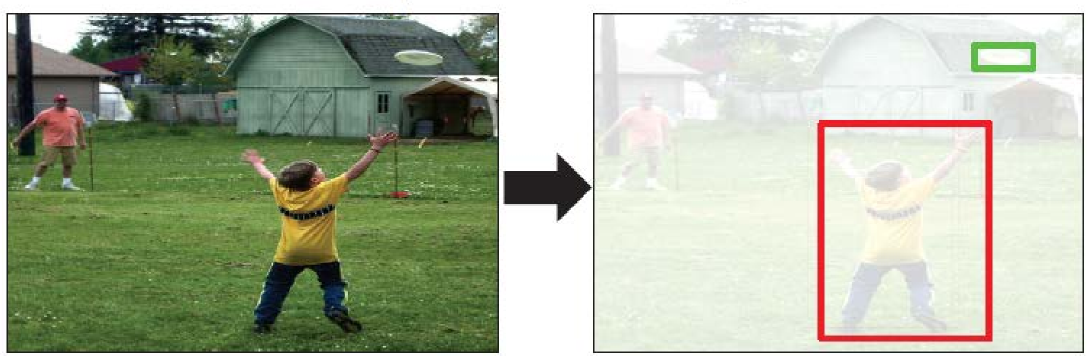
Tasks leveraging interactees

1

Prior for "what to mention" about the scene



A little boy in a chair eating a cake.



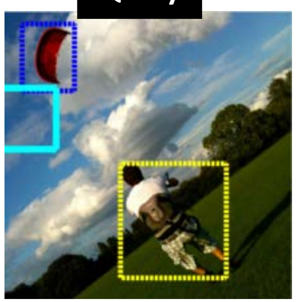
A small boy is reaching up for a frisbee.

Tasks leveraging interactees

1

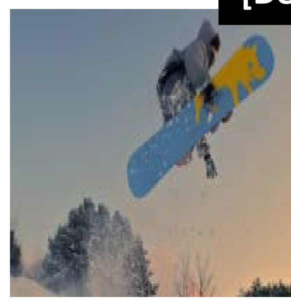
Prior for “what to mention” about the scene

Query

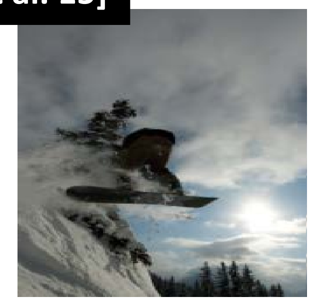


The men is flying a kite on a sunny day

[Devlin et al. 15]

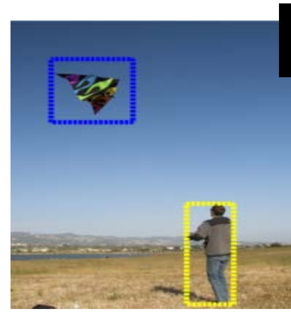


A person doing tricks in the air on a snowboard

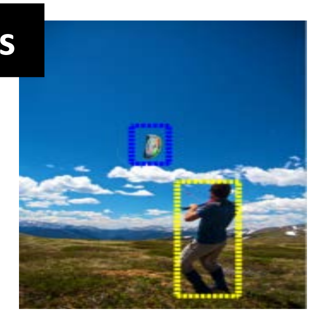


A man on a snowboard comes off the mountain

Ours



A man is flying a kite in a grassy field



A man flies a kite against a blue sky

Tasks leveraging interactees

1

Prior for “what to mention” about the scene

Query

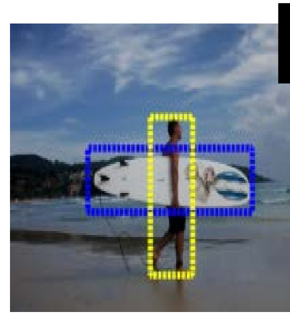
Men walking into the ocean with their surf boards

[Ordonez et al. 11]

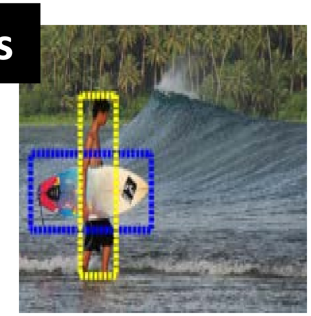
A man riding a board on top of a wave in the ocean

A man surfs on a surfboard on a lake

Ours



A man with a surf board walks across the beach



A young man carrying a surfboard next to a wave

Tasks leveraging interactees

2

Image retargeting that preserves interactee region

Input

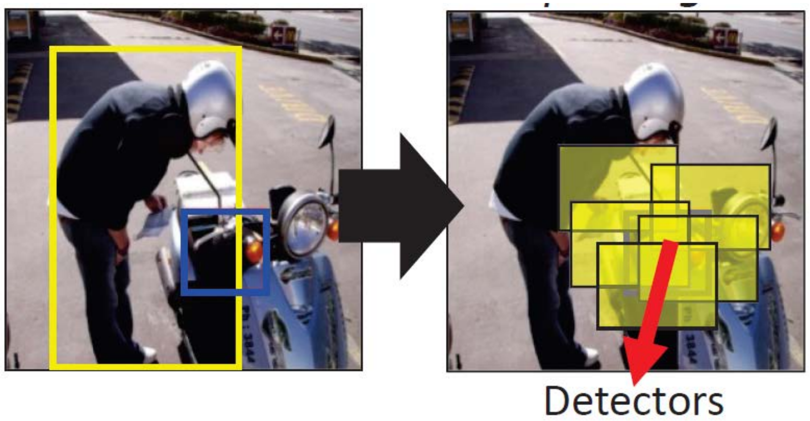
Ours

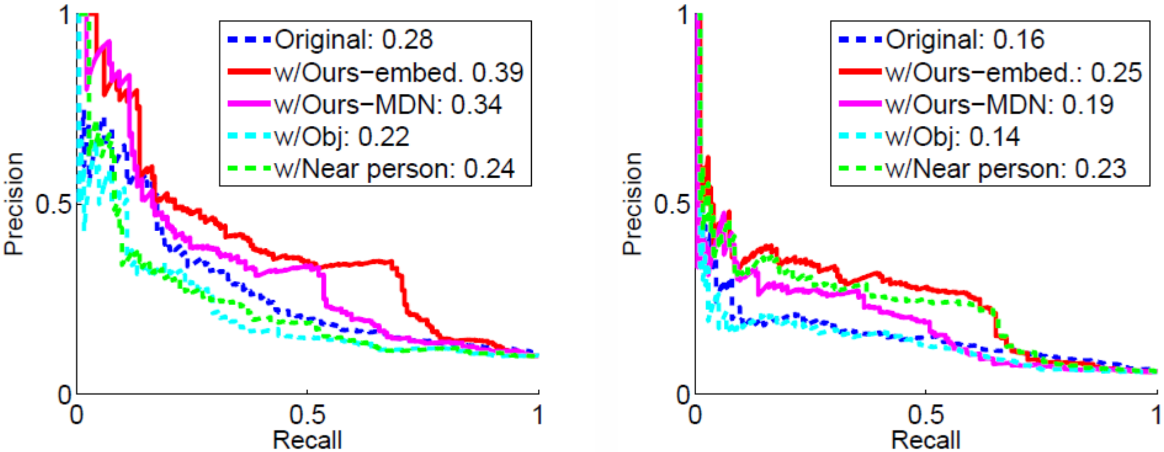
Baseline (objectness)

Tasks leveraging interactees

3

Focus object detector's search





(a) Using computer (b) Reading

Summary

THE UNIVERSITY OF TEXAS AT AUSTIN

– "Embodied" feature learning
  - Learn the link between egomotion and how the surrounding scene changes.
– End-to-end active recognition
  - Learn a policy for how to move, where to point camera within a 360 scene
– Interactee localization
  - Person-centric cues of saliency and open world human-object interactions

Dinesh Jayaraman

Chao-Yeh Chen

CVPR 2016 Scene Understanding Workshop (SUNw)

Papers

- **Egomotion and visual learning**
  - **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 Recognition by Forecasting the Effect of Motion**. D. Jayaraman and K. Grauman. To appear, ECCV 2016. arXiv:1605.00164
- **Interaction and scene understanding**
  - **Predicting the Location of "Interactees" in Novel Human-Object Interactions**. C-Y. Chen and K. Grauman. In Proceedings of the Asian Conference on Computer Vision (ACCV), Singapore, Nov 2014.
  - **Subjects and Their Objects: Localizing Interactees for a Person-Centric View of Importance**. C-Y. Chen and K. Grauman. arXiv: :1604.04842v1, April 2016.