Egomotion and Visual Learning

Kristen Grauman

Department of Computer Science

The University of Texas at Austin

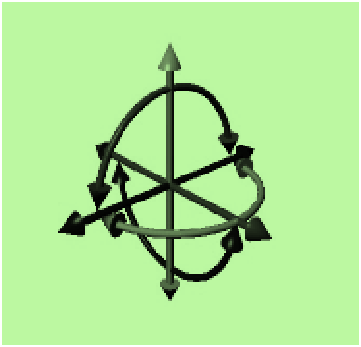
Dinesh Jayaraman, UT Austin

Hao Jiang, Boston College

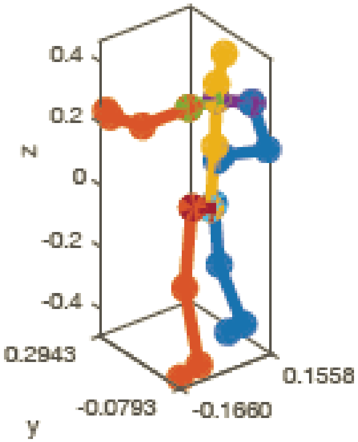
CVPR 2016 Tutorial on First Person Vision

What can a first person camera tell us about my motion?

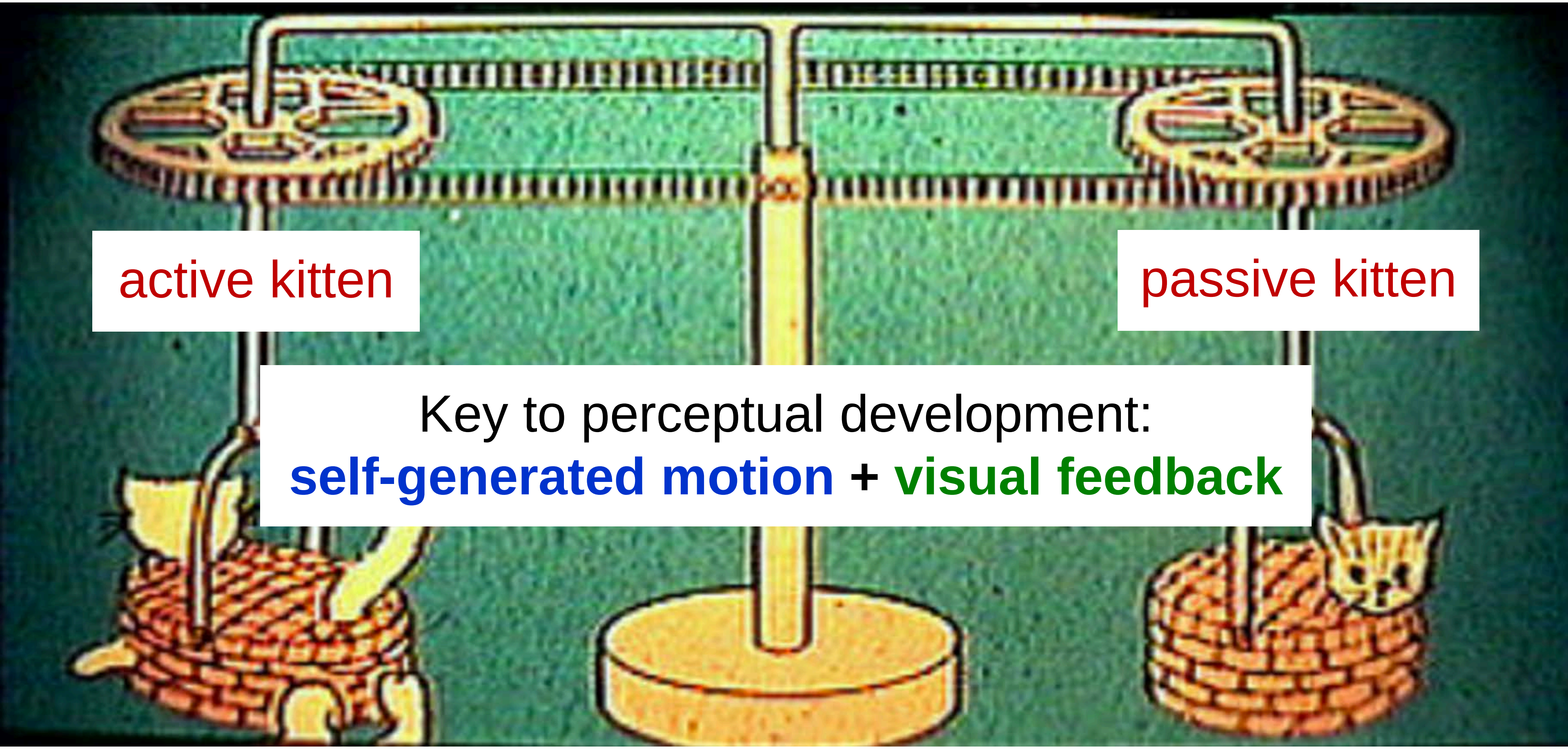
1. Learning representations tied to ego-motion



2. Estimating “invisible” articulated 3D body poses



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

[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"

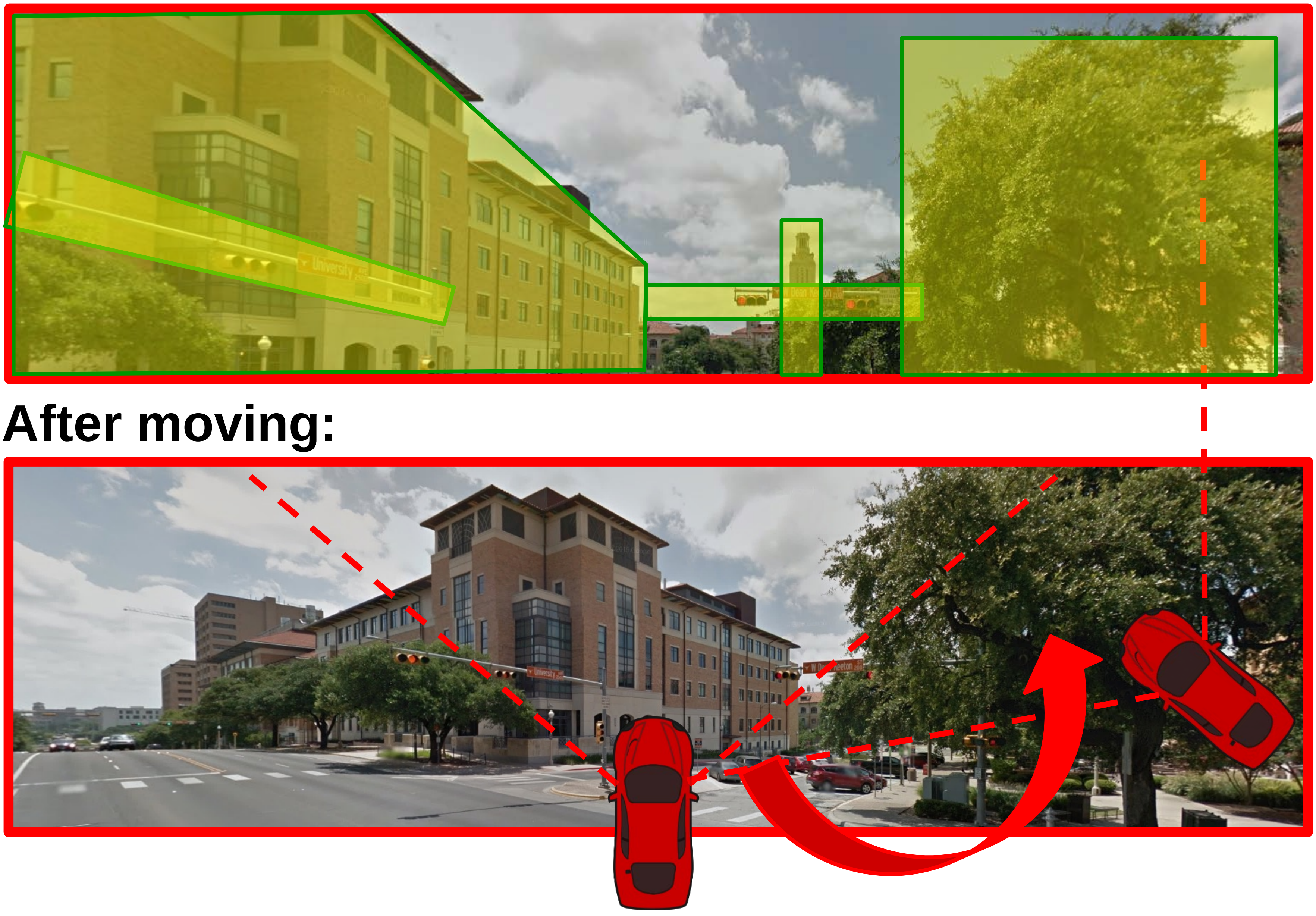
Ego-motion motor signals

\+

Unlabeled video

[Jayaraman & Grauman, ICCV 2015]

Ego-motion ↔ vision: view prediction



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"

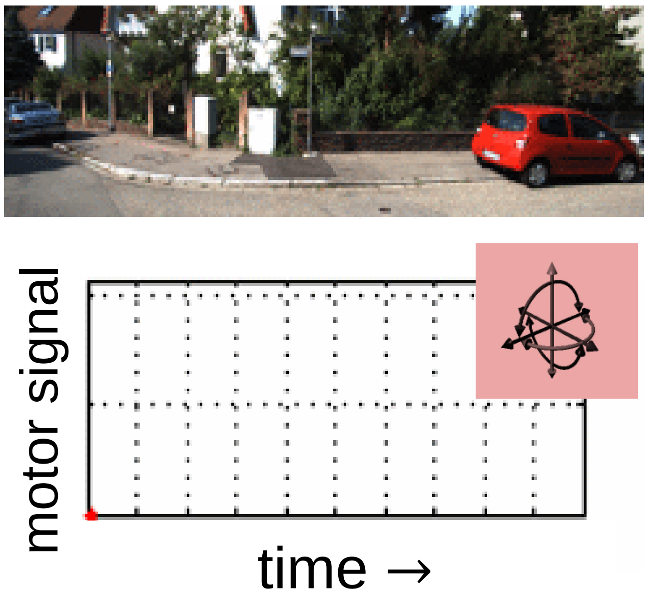
$$\mathbf{z}(g\mathbf{x}) \approx M_g \mathbf{z}(\mathbf{x})$$

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

Approach idea: Ego-motion equivariance

Training data

Unlabeled video +
motor signals



Learn

Equivariant embedding

organized by ego-motions

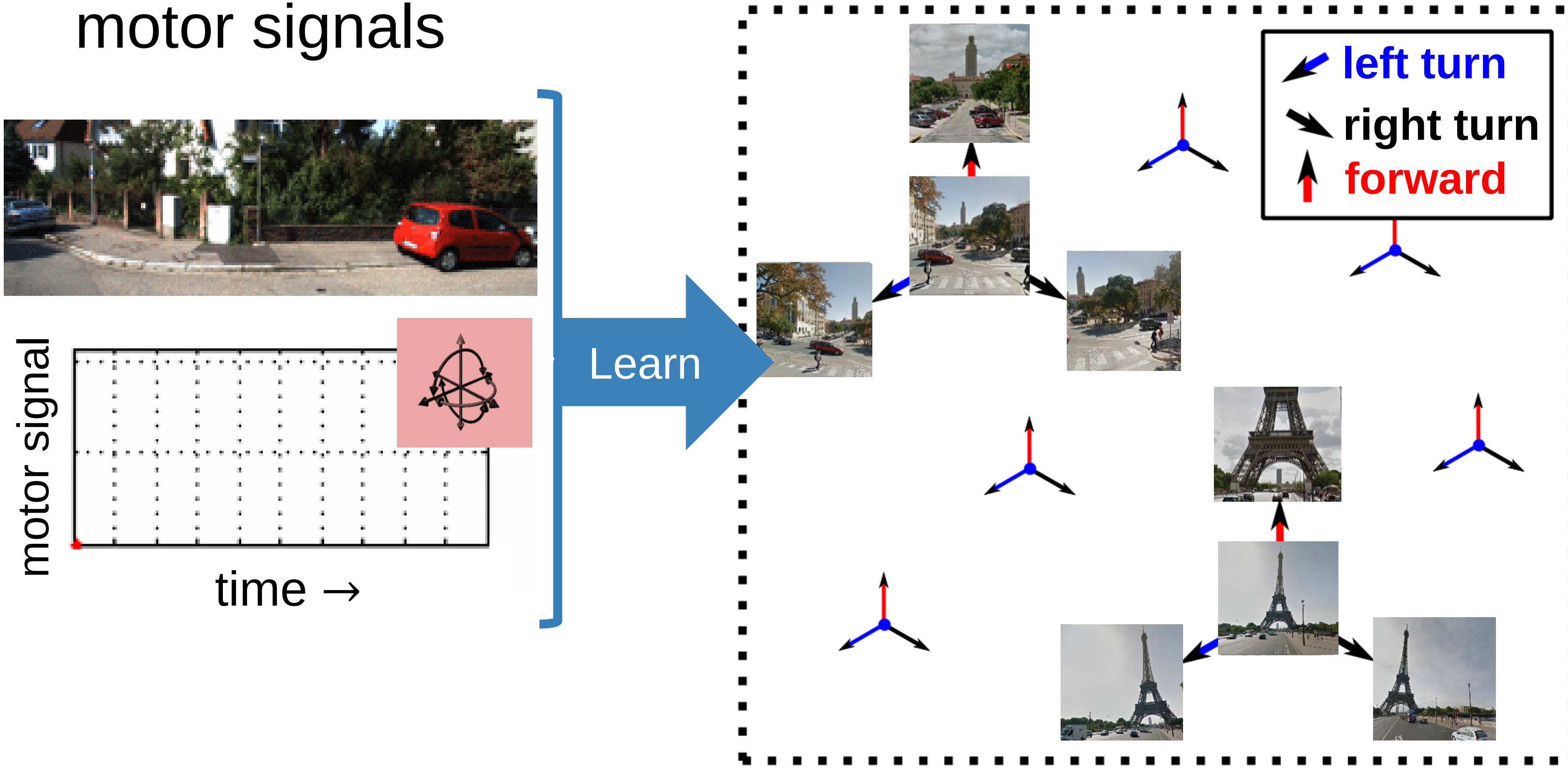
Pairs of frames related by
similar ego-motion should
be related by same
feature transformation

[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]

Approach overview

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

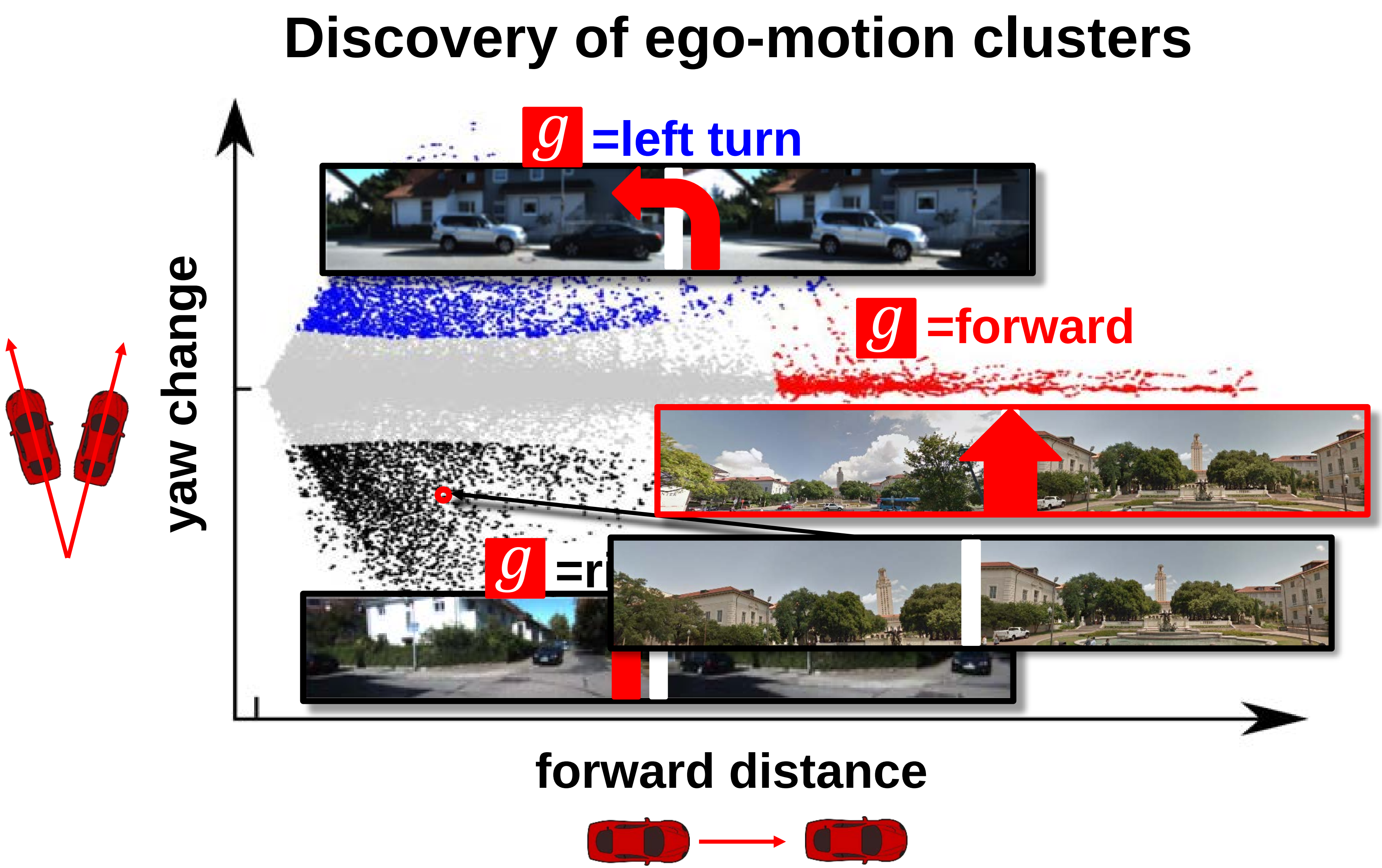
1. Extract training frame pairs from video

2. Learn ego-motion-equivariant image features

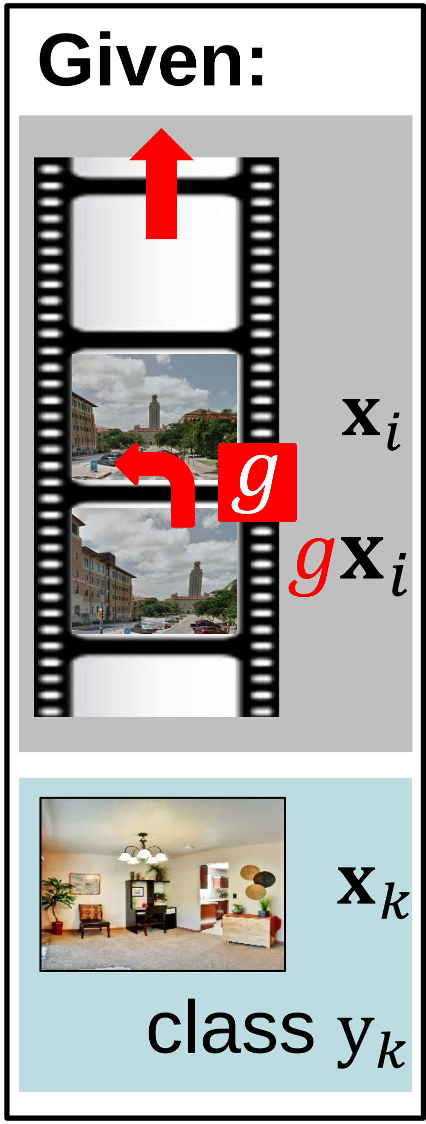
3. Train on target recognition task in parallel

[Jayaraman & Grauman, ICCV 2015]

Training frame pair mining

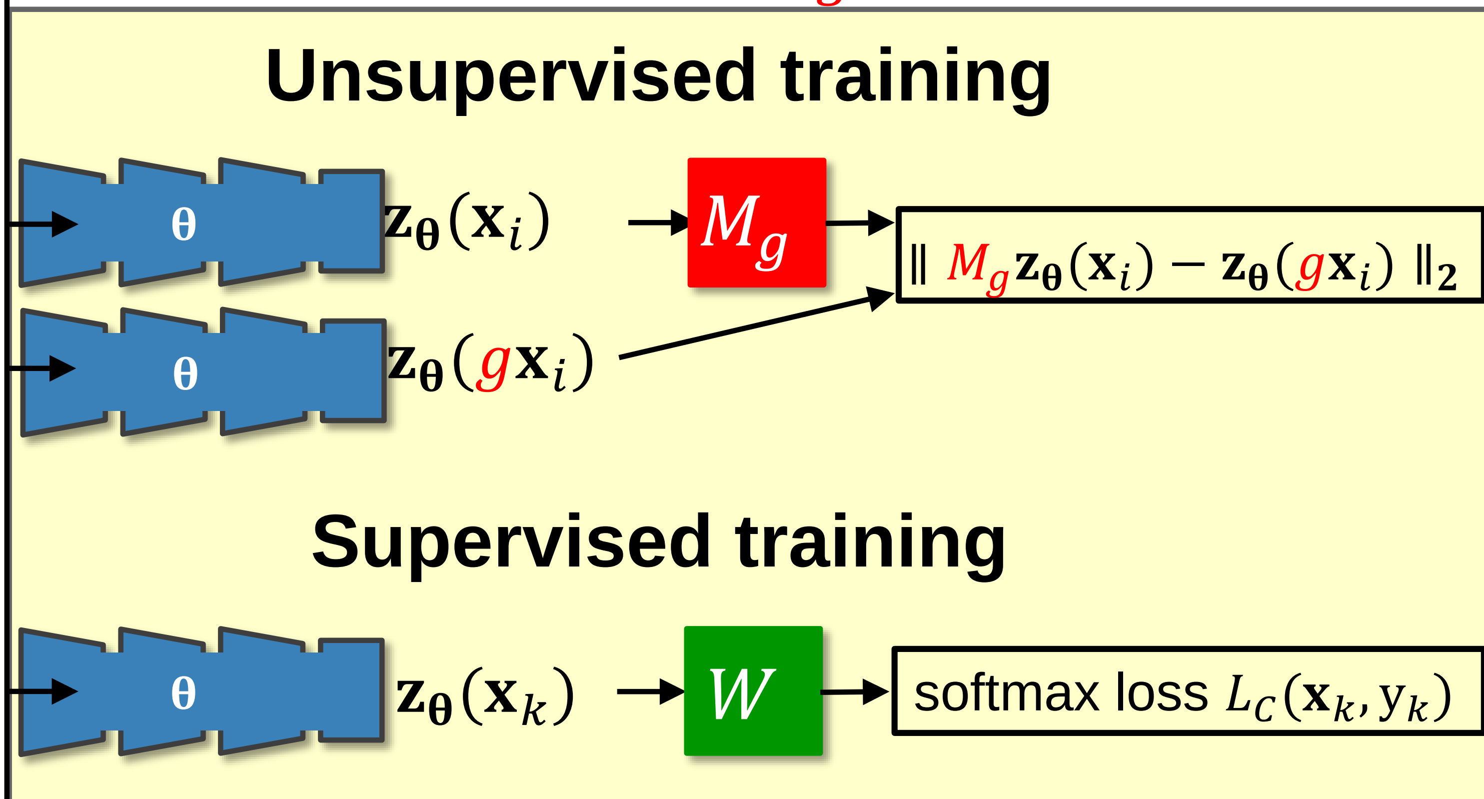


Ego-motion equivariant feature learning



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

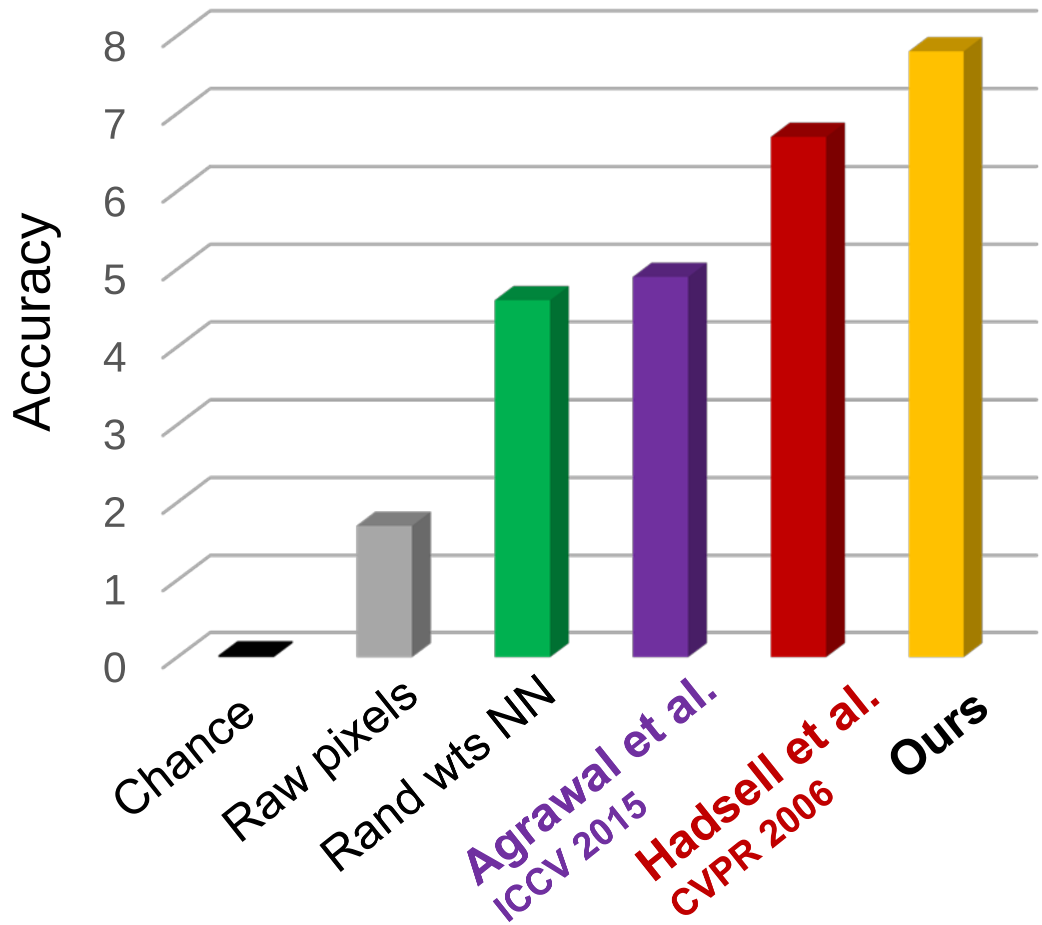
Exploit features for **static scene classification**
(SUN, 397 classes)

Xiao et al, CVPR ’10

Results: Recognition

Purely unsupervised feature learning

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

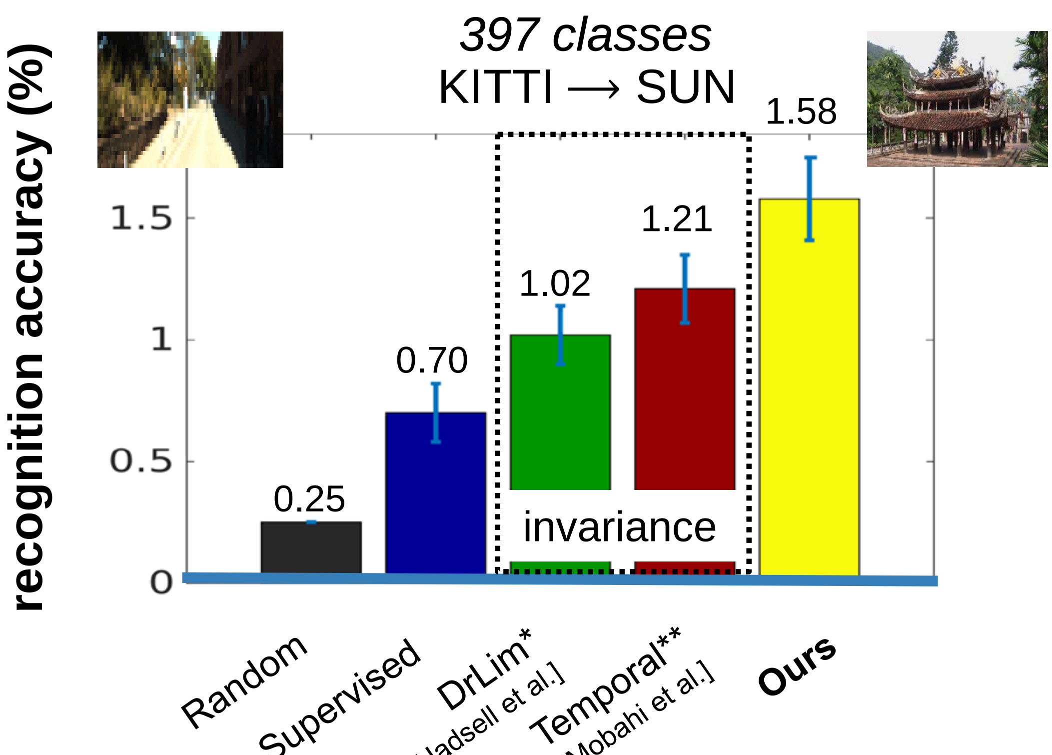


Agrawal, Carreira, Malik, Learning to see by moving. ICCV 2015

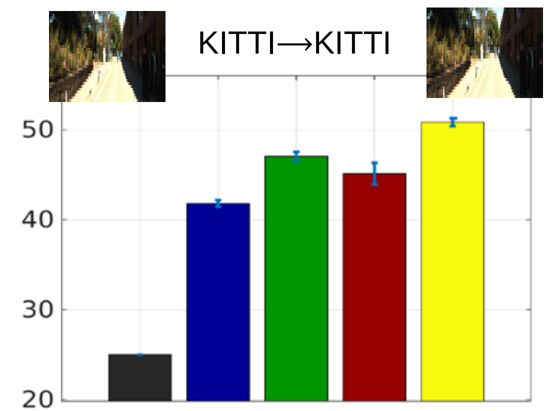
Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping. CVPR 2006

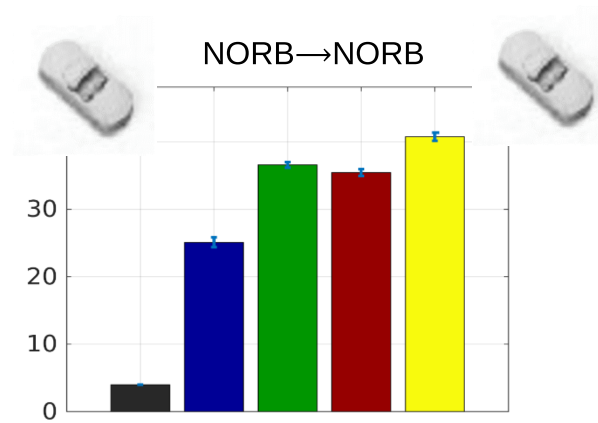
Results: Recognition

Ego-motion equivariance as a regularizer



6 labeled training examples per class





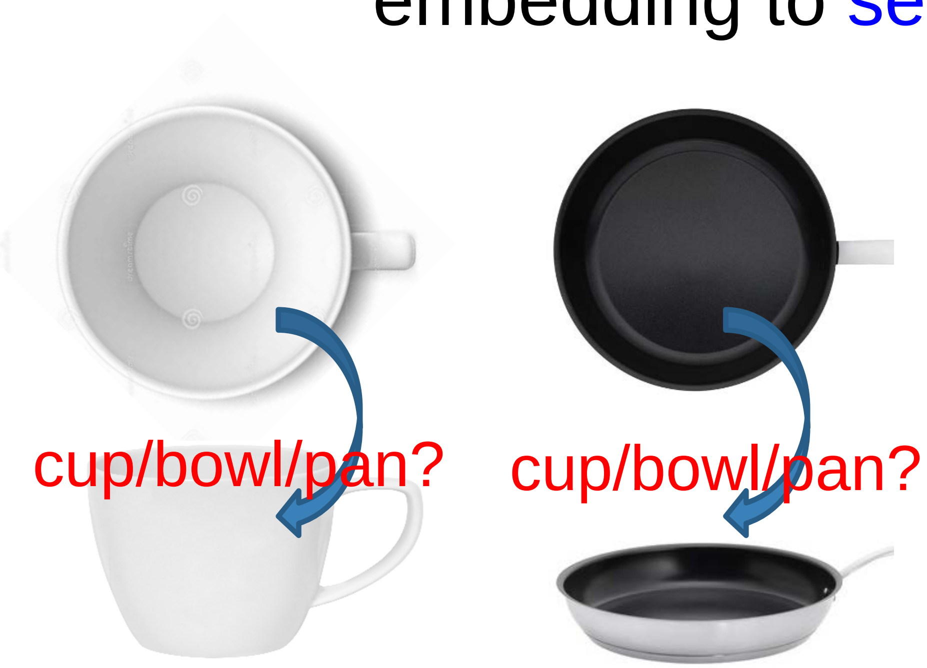
Up to 30% accuracy increase over state of the art!

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

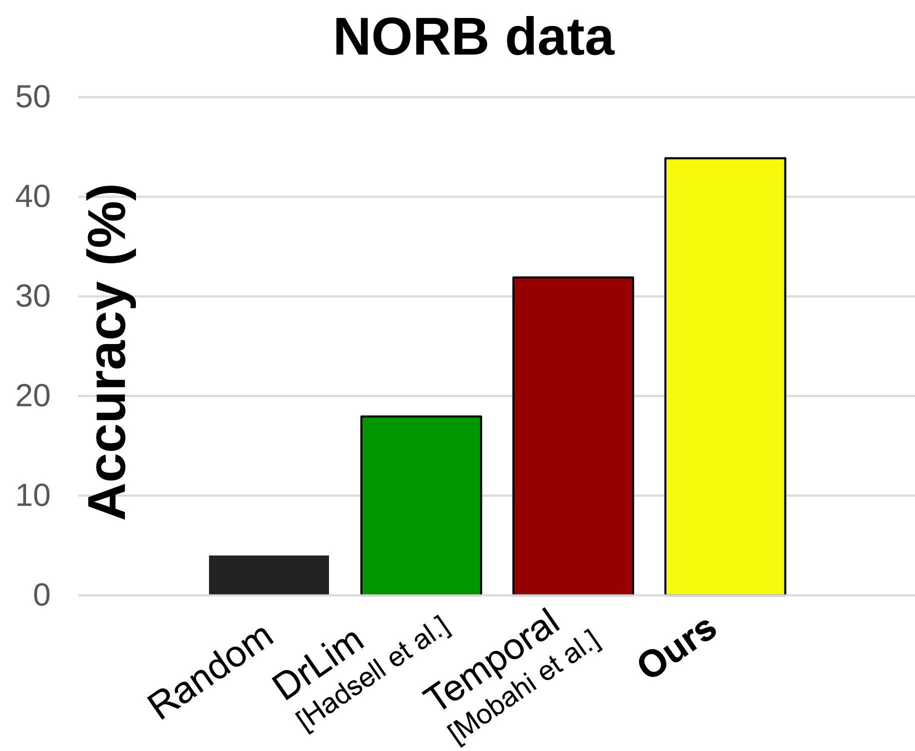
**Mobahi et al., Deep Learning from Temporal Coherence in Video, ICML'09

Learning how to move for recognition

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

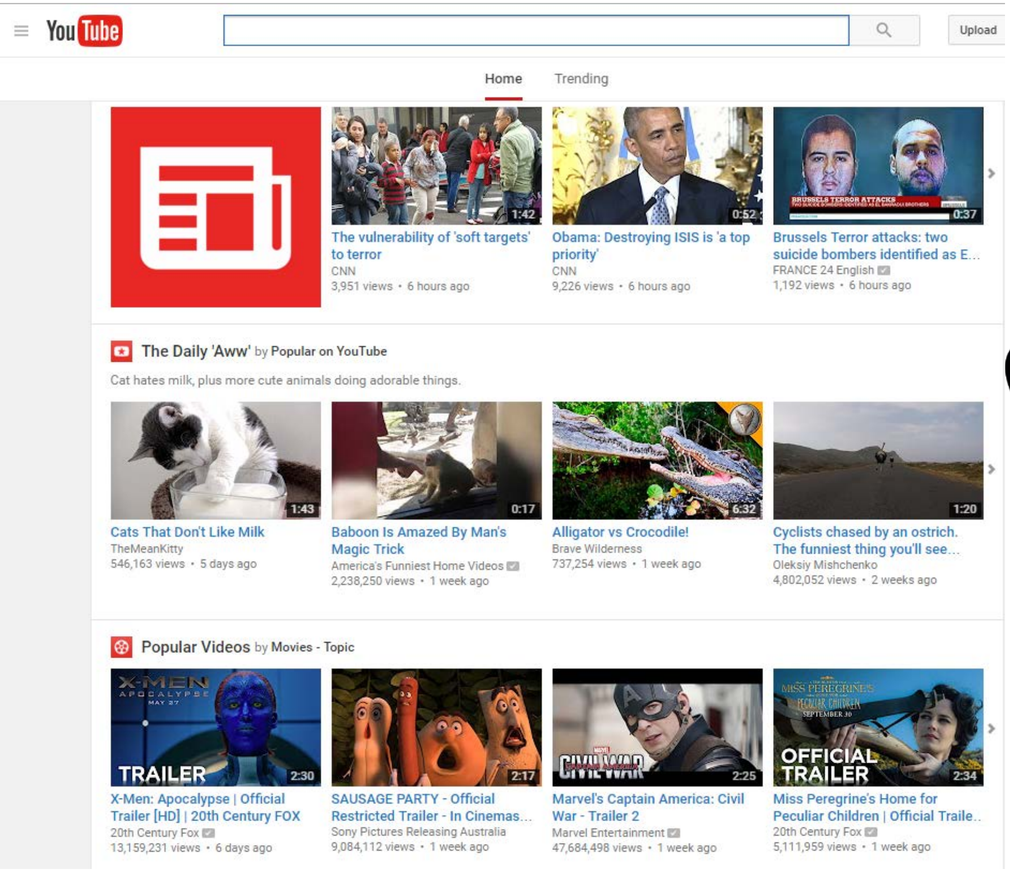


cup frying pan



[Jayarman & Grauman, ICCV 2015]

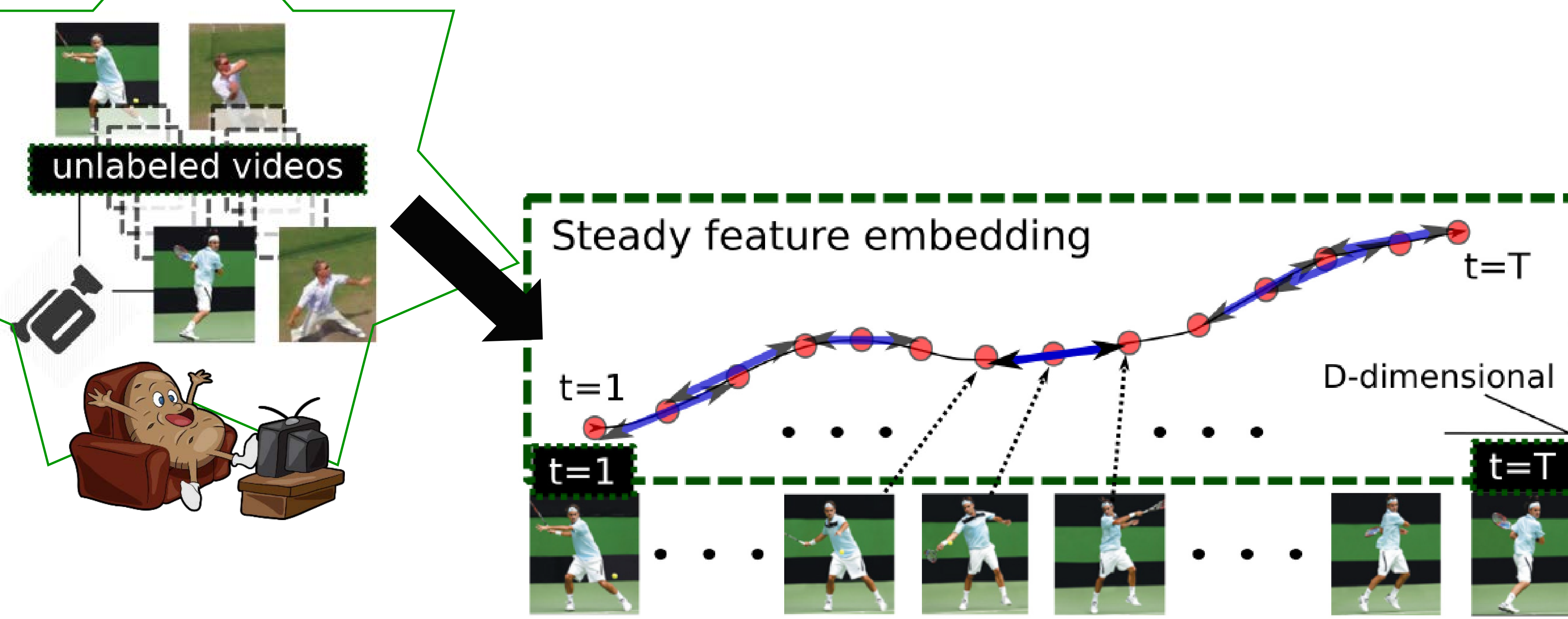
Learning from arbitrary unlabeled video?



Unlabeled video

Our idea: Steady feature analysis

Learning from arbitrary unlabeled video



Equivariance ≈ “steadily” varying frame features!

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

[Jayaraman & Grauman, CVPR 2016]

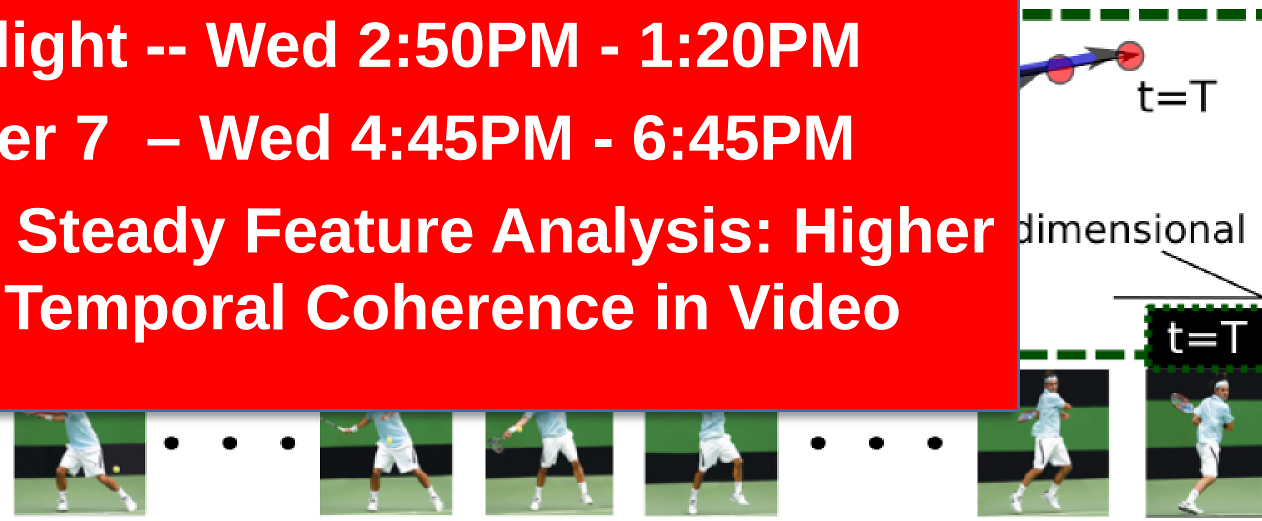
Our idea: Steady feature analysis

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

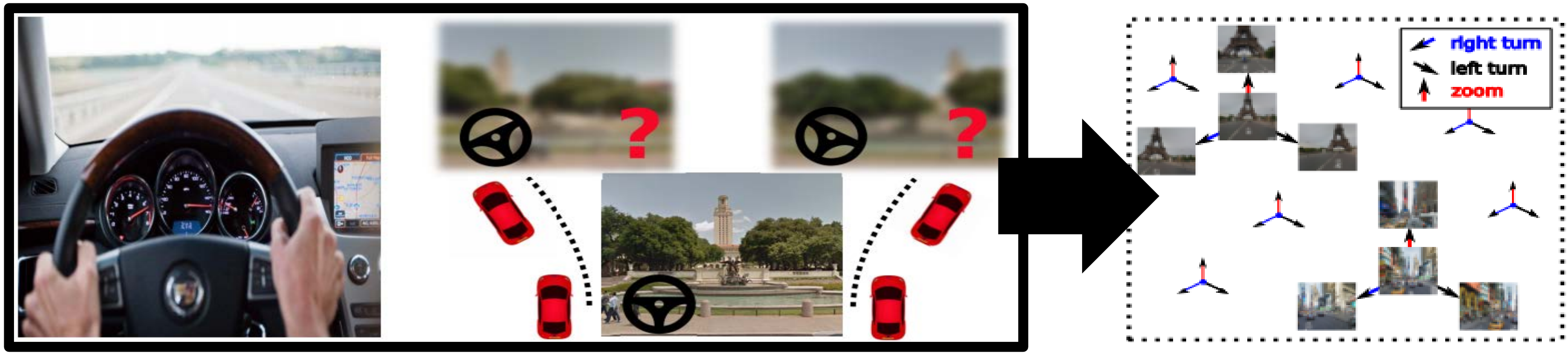


Equivalerance ≈ “steadily” varying frame features!

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

[Jayaraman & Grauman, CVPR 2016]

Recap so far

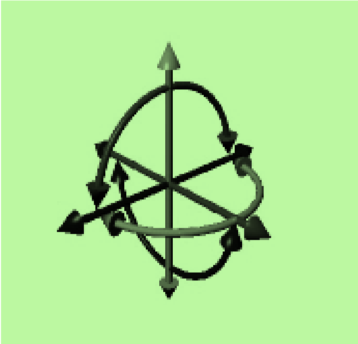


http://vision.cs.utexas.edu/projects/egoequiv/

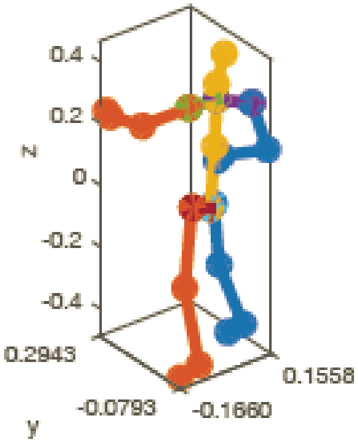
- New *embodied* visual feature learning paradigm
- Ego-motion equivariance boosts performance across multiple challenging recognition tasks
- Future work: volition at training time too

What can a first person camera tell us about my motion?

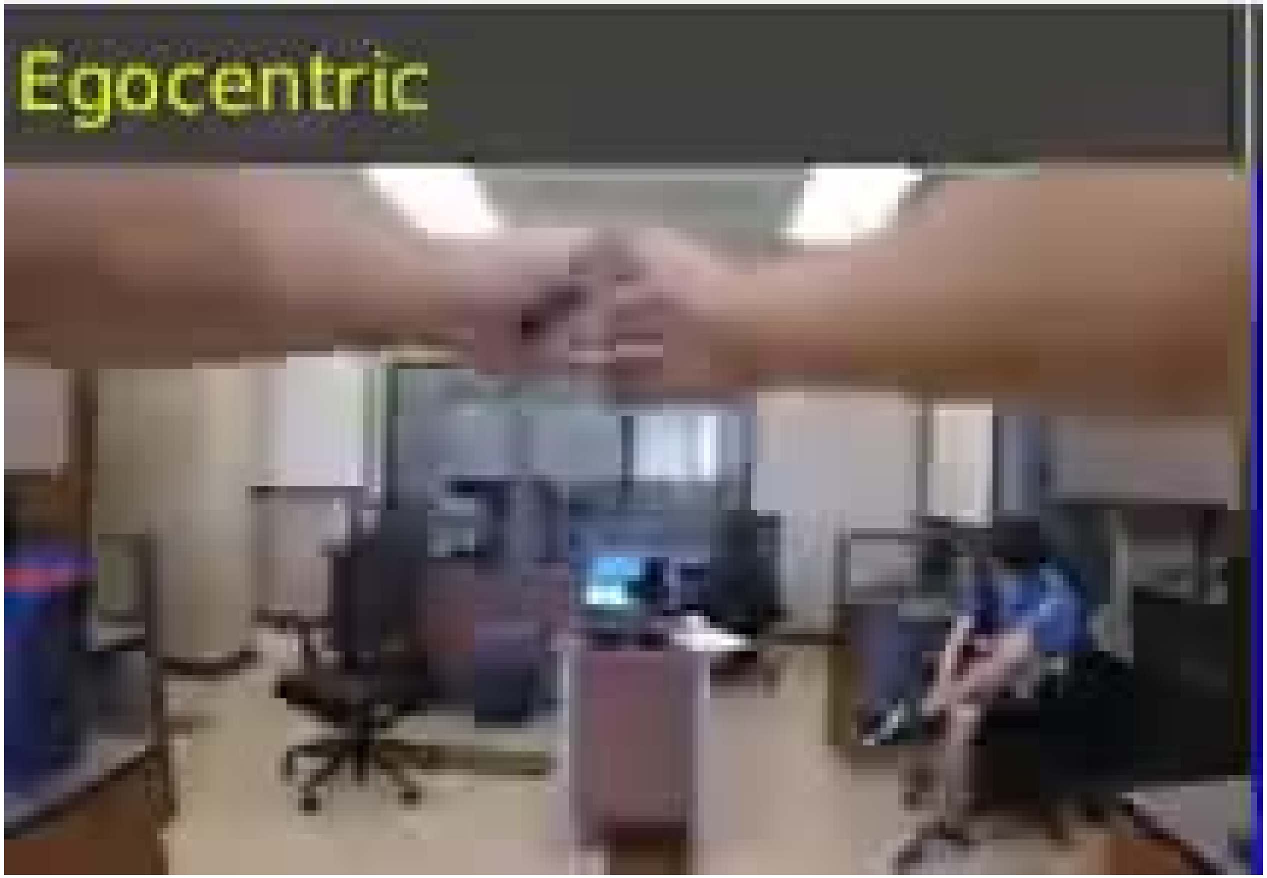
1. Learning representations tied to ego-motion



2. Estimating “invisible” articulated 3D body poses



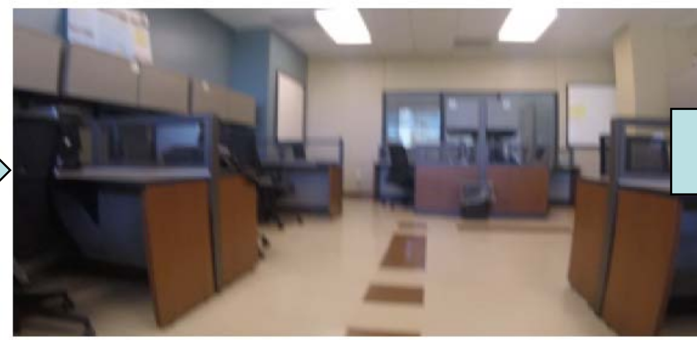
What’s on the other side of the camera?

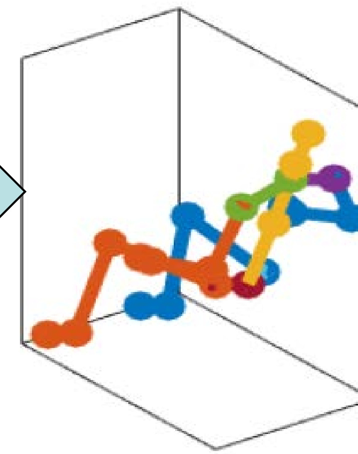


What does apparent ego-motion reveal about the person behind the camera?

Seeing invisible poses

- **Goal:** Learn to estimate 3D body pose of person behind the wearable camera



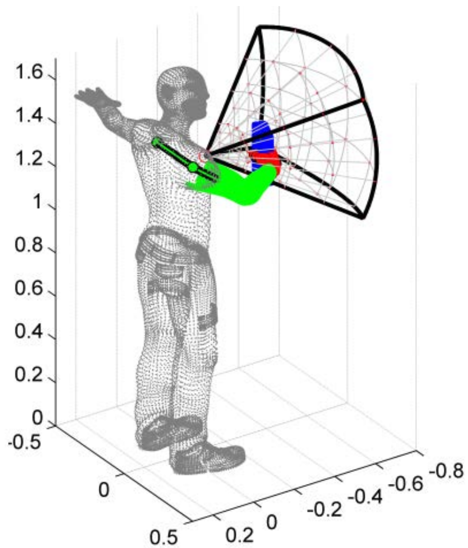


Input:
egocentric video

Output:
sequence of 3d joint positions

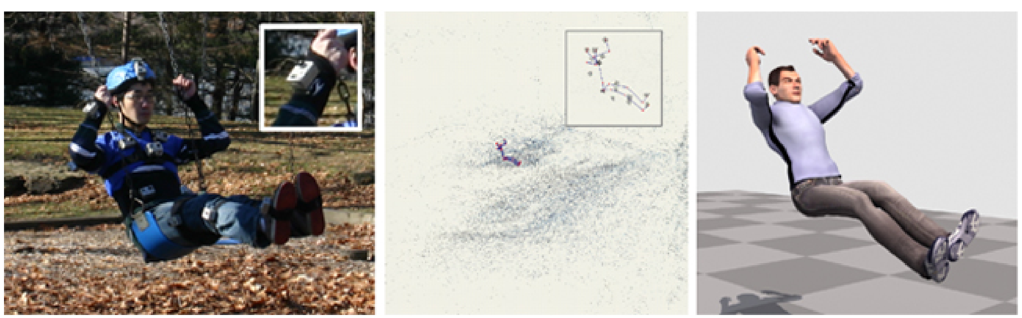
Jiang & Grauman, arXiv 2016

Prior work: Ego body pose



Rogez et al. 2015, Kitani et al. 2013,…

- Focus on hands and arms
- Assume visible body parts

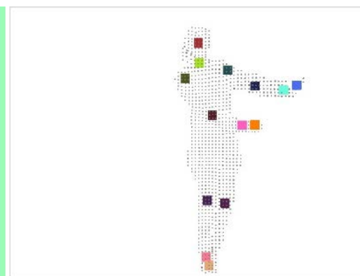


Shiratori et al., SIGGRAPH 2011

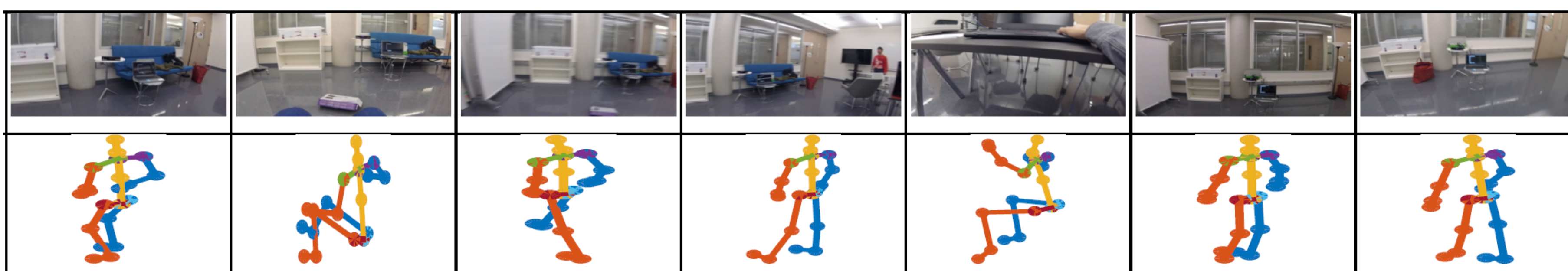
- Multiple cameras on joints
- Geometric solution
- Expensive (1.5 days for 1 min of capture)

Our approach:
Seeing invisible poses

- Training: Kinect for ground truth pose collection
  - Used only for training data and evaluation

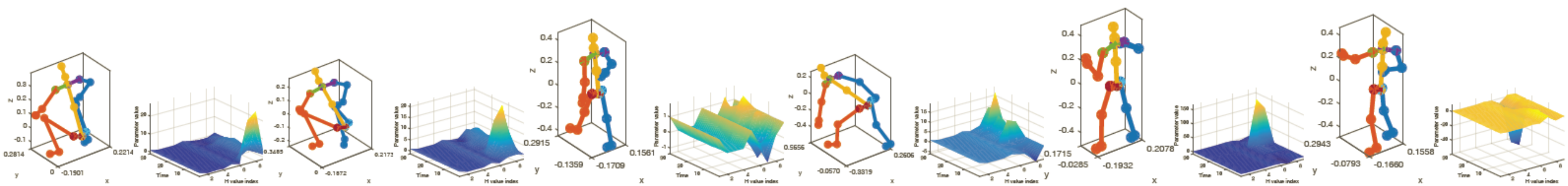




- 10 subjects
- ~1 hour video, 1-3 minute clips

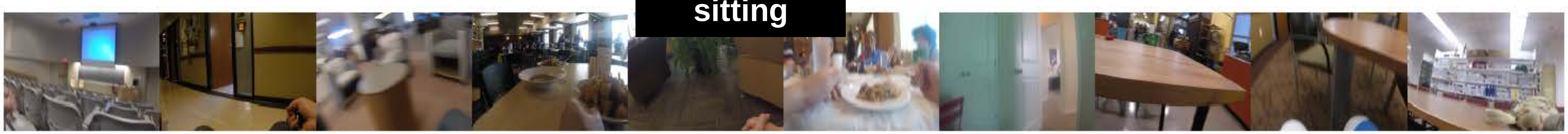


Our approach:
Seeing invisible poses

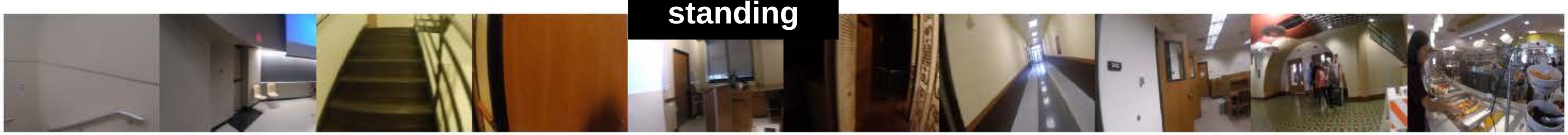
1. Instantaneous estimates based on
   - Dynamic motion signatures
     - Homographies between successive frames
   - Static scene structure



sitting



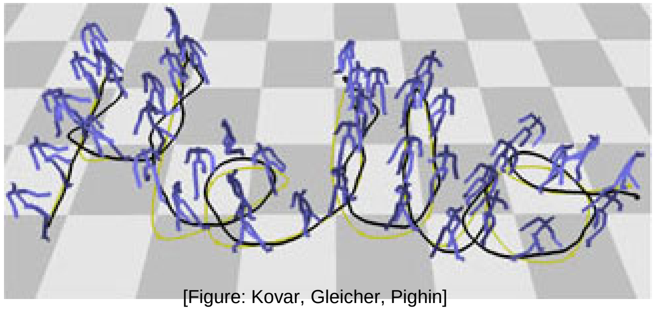
standing



Our approach: Seeing invisible poses

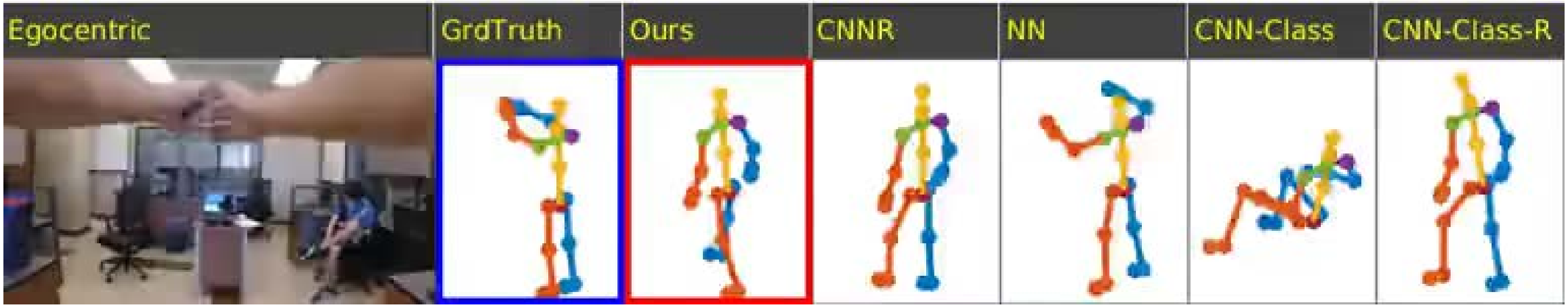
1. Instantaneous estimates based on
   – Dynamic motion signatures
     • Homographies between successive frames
   – Static scene structure

2. Longer term sequence estimate
   – Non-parametric model of dynamics
   – Identify least-cost “pose path” in exemplars

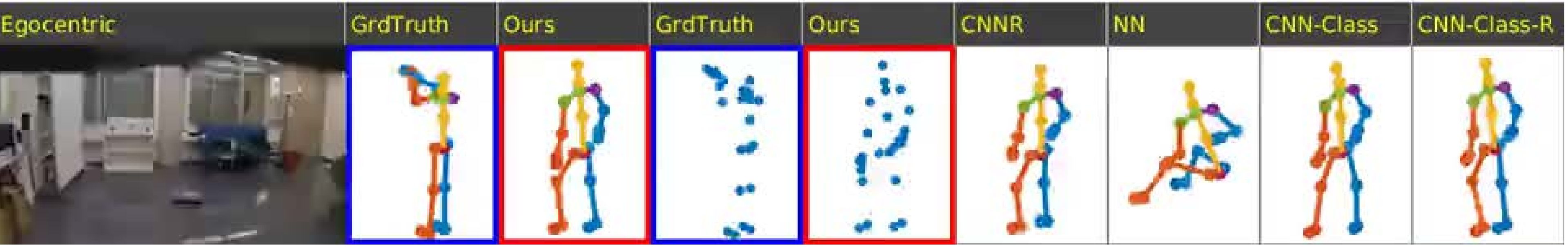


[Figure: Kovar, Gleicher, Pighin]

Results: Ego-video → body pose



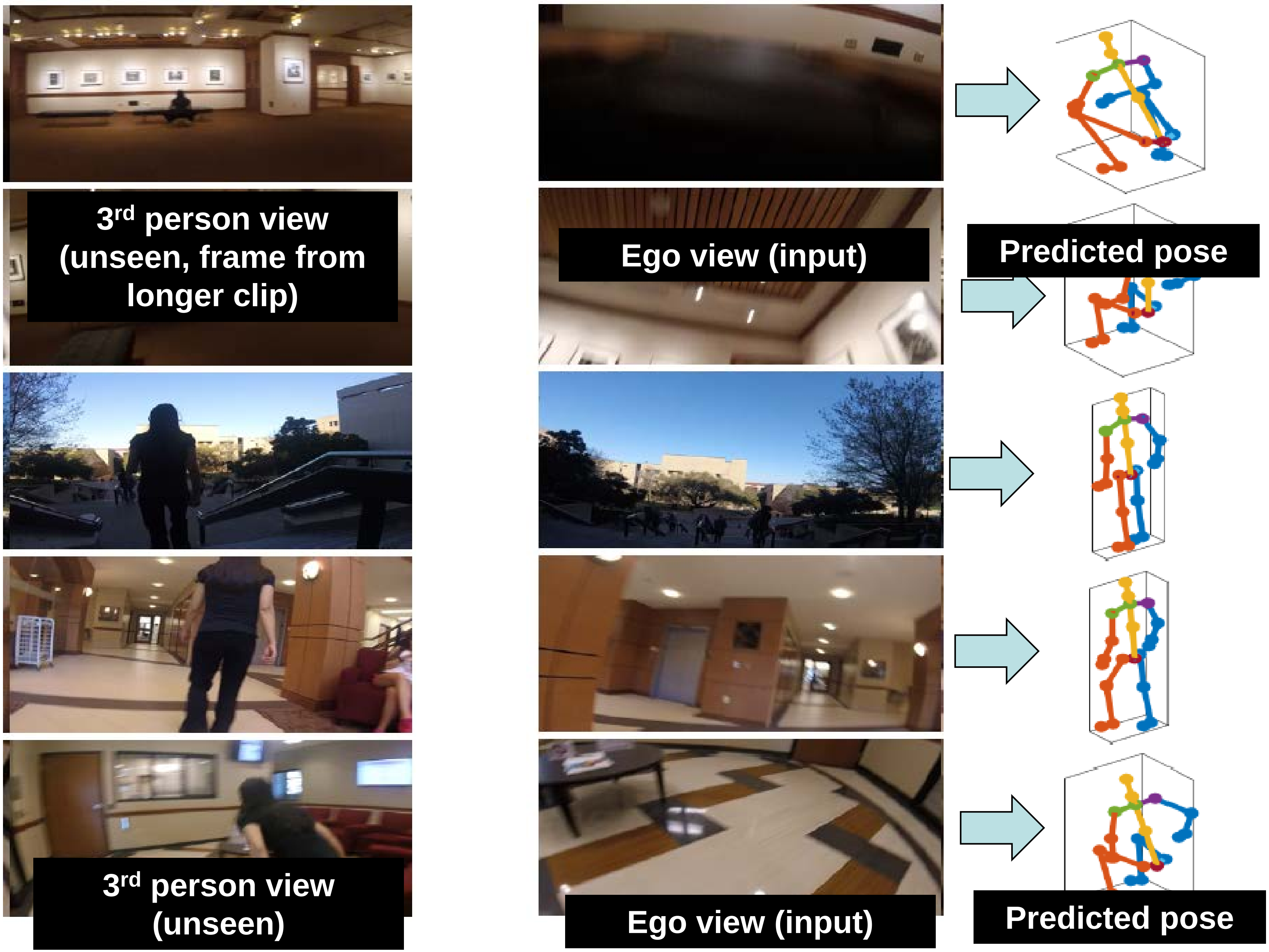
Train/test: Person repeats, but environment differs



Train/test: Person differs AND environment differs

Jiang & Grauman, arXiv 2016

Results: broader test settings



Results: Ego-video → body pose

- Joint errors (cm), ~40 minutes total test video

“DeepPose” [Toshev & Szegedy, 2014] trained for our task

	Path (Ours)	Path-Cluster	Path-CNN	Path-CNN-R	KdTree	CNN-Regr	AwaysStanding	AwaysSitting
Head	15.8(0.08)	16.5(0.08)	21.6(0.14)	22.9(0.14)	18.1(0.11)	16.2(0.10)	**15.1**(0.08)	32.5(0.09)
Elbow	**14.4**(0.07)	15.4(0.07)	18.6(0.12)	19.4(0.12)	15.8(0.10)	14.4(0.09)	14.5(0.08)	20.7(0.08)
Wrist	**19.1**(0.09)	20.6(0.10)	26.5(0.17)	27.1(0.17)	21.3(0.13)	22.0(0.14)	22.9(0.12)	21.3(0.08)
Knee	**15.4**(0.09)	17.2(0.09)	27.3(0.17)	26.2(0.17)	22.0(0.14)	21.3(0.13)	21.2(0.11)	40.0(0.11)
Ankle	**20.7**(0.10)	22.9(0.10)	33.8(0.21)	33.3(0.21)	28.4(0.18)	26.4(0.17)	26.7(0.13)	37.9(0.09)
NAvgAll	**17.2**	19.1	48.1	48.7	32.8	29.7	24.6	31.9
NAvg(W+A)	**19.9**	22.6	60.0	60.2	40.8	38.7	32.4	27.1

Train/test: Person repeats, but environment differs

	Path (Ours)	Path-Cluster	Path-CNN	Path-CNN-R	KdTree	CNN-Regr	AwaysStanding	AwaysSitting
Head	16.6(0.07)	18.0(0.07)	19.4(0.09)	21.3(0.10)	20.1(0.09)	15.8(0.07)	**14.3**(0.07)	29.1(0.07)
Elbow	15.3(0.06)	16.9(0.06)	19.1(0.09)	19.5(0.09)	18.0(0.08)	15.8(0.07)	**14.9**(0.06)	20.9(0.06)
Wrist	**22.2**(0.08)	24.2(0.08)	29.7(0.14)	29.4(0.14)	24.9(0.12)	24.3(0.1[illegible])	23.8(0.09)	22.9(0.07)
Knee	**18.9**(0.07)	24.4(0.09)	21.6(0.10)	21.8(0.10)	31.9(0.15)	27.6(0.13)	21.7(0.08)	45.7(0.09)
Ankle	**24.9**(0.09)	29.9(0.10)	29.2(0.14)	29.2(0.14)	38.1(0.18)	33.3(0.15)	28.2(0.10)	43.0(0.09)
NAvgAll	**19.9**	24.6	35.4	36.4	44.5	34.6	22.4	32.9
NAvg(W+A)	**23.6**	28.4	46.6	46.3	53.3	44.6	28.9	30.7

Train/test: Person differs AND environ

Generic body posture priors

Summary

- Visual learning benefits from
  - context of action and motion in the world
  - continuous self-acquired feedback
  - cues from ego-motion on multiple levels

Dinesh
Jayaraman

- Main ideas:
  - “Embodied” feature learning using both visual and motor signals
  - Learning to estimate articulated body pose from first person video

Hao Jiang
Boston College

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.
- **Seeing Invisible Poses: Estimating 3D Body Pose from Egocentric Video.** H. Jiang and K. Grauman. March 2016. arXiv:1603.07763