Intelligent look-around behavior: Learning to examine 3D objects and scenes

Kristen Grauman
Department of Computer Science
University of Texas at Austin
## Human-taken photos

A well-framed, well-curated moment in time

- BSD (2001)
- PASCAL (2007-12)
- LabelMe (2007)
- ImageNet (2009)
- SUN (2010)
- Places (2014)
- MS COCO (2014)
- Visual Genome (2016)
Machine-taken photos

Highly variable inputs, 3D viewing conditions
Learning to examine 3D objects/scenes

Recognition beyond momentary 2D appearance patterns:

1. **Look-around prediction**: How will a given 3D egomotion affect what is seen next?

2. **Look-around action**: What sequence of egomotions (or manipulations) will best reduce ambiguity?
The kitten carousel experiment
[Held & Hein, 1963]

Key to perceptual development:
self-generated motion + visual feedback
Goal: Teach computer vision system the connection: “how I move” ↔ “how my visual surroundings change”

Idea: Ego-motion ↔ vision

Ego-motion motor signals + Unlabeled video

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]
Ego-motion ↔ vision: view prediction

After moving:
Approach idea: Ego-motion equivariance

Training data
Unlabeled video + motor signals

Equivariant embedding organized by ego-motions

\[ z(gx) \approx M_g z(x) \]

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

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]
Approach idea: Ego-motion equivariance

Training data
Unlabeled video + motor signals

Equivariant embedding
organized by ego-motions

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]
Results: Recognition

Learn from *unlabeled* car video (KITTI)

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

Geiger et al, IJRR ’13

Xiao et al, CVPR ’10
Results: Recognition

Ego-equivariance for unsupervised feature learning

SUN scenes: 397 multi-class accuracy

- **random weights**
- **DrLIM+ (Hadsell et al. CVPR06)**
- **LSM* (Agrawal et al. ICCV15)**

* Agrawal, Carreira, Malik, “Learning to see by moving”, ICCV 2015

[Jayaraman & Grauman, IJCV 2017]
Passive → complete egomotions

Pre-recorded video

Comprehensive observation

motor signal

time →
One-shot reconstruction

Key idea: One-shot reconstruction as a proxy task to learn semantic features.
One-shot reconstruction

Shape from dense views
geometric problem

Shape from one view
semantic problem

[Snavely et al, CVPR ’06]

[Sinha et al, ICCV’93]
Approach: One-shot reconstruction for feature learning

- Implicit 3D shape representation
- Agnostic of category
- No “canonical” azimuth to exploit

[Jayaraman & Grauman, arXiv 2017]
One-shot reconstruction example
One-shot features for recognition

ModelNet

Accuracy (%)

[Chang et al 2015]

ShapeNet

[Chang et al 2015]

* Hadsell et al, Dimensionality reduction by Learning an invariant mapping, CVPR 2005
** Masci et al, Stacked Convolutional Autoencoders for Hierarchical Feature Extraction, ICANN 2011
^ Agrawal, Carreira, Malik, Learning to See by Moving, ICCV 2015
Learning to examine 3D objects/scenes

Recognition beyond momentary 2D appearance patterns:

1. **Look-around prediction**: How will a given 3D egomotion affect what is seen next?

2. **Look-around action**: What sequence of egomotions (or manipulations) will best reduce ambiguity?
Actively moving to recognize


Jayaraman and Grauman, ECCV 2016
Actively moving to recognize

Our idea: End-to-end active recognition + 3D motion look-ahead

Jayaraman and Grauman, ECCV 2016
End-to-end active recognition: results

1. Look around scene
2. Manipulate object
3. Move around object
Learned active policies outperform prior active recognition approaches.

Jayaraman and Grauman, ECCV 2016
End-to-end active recognition: example

[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]
Goal: Learn to “look around”

reconnaissance  vs.  recognition

task predefined  vs.  task unfolds dynamically

Can we learn look-around policies for visual agents that are curiosity-driven, exploratory, and generic?
Key idea: Active observation completion

**Completion objective:** Learn policy for efficiently inferring (pixels of) all yet-unseen portions of environment

Agent must choose where to look *before* looking there.

Jayaraman and Grauman, arXiv 2017
Key idea: Active observation completion

Completion objective: Learn policy for efficiently inferring (pixels of) all yet-unseen portions of environment

Agent must choose where to look before looking there.

Jayaraman and Grauman, arXiv 2017
Approach: Active observation completion

**Non-myopic:** Train to target a budget of observation time

Jayaraman and Grauman, arXiv 2017
Datasets: Two scenarios

Where to look next?

SUN 360 panoramas
[Xiao 2012]

How to manipulate?
Active “look around” results

*SUN360

*ModelNet (seen cls)

*ModelNet (unseen cls)

*Harel et al, Graph based Visual Saliency, NIPS’07

Jayaraman and Grauman, arXiv 2017
Active “look around” results

Learned active look-around policy: quickly grasp environment independent of a specific task

*Harel et al, Graph based Visual Saliency, NIPS’07

Jayaraman and Grauman, arXiv 2017
Active “look around” visualization

observed view

Ground truth

Visualized internal model over time

Jayaraman and Grauman, arXiv 2017
Active “look around” visualization

Ground truth

Visualized internal model over time

observed view

Jayaraman and Grauman, arXiv 2017
Active “look around” visualization

observed view

Ground truth

Visualized internal model over time

Jayaraman and Grauman, arXiv 2017
Motion policy transfer

Unsupervised observation completion

- Decoder
- Look-around encoder
- Look-around Policy

Supervised recognition

[Jayaraman et al, ECCV 16]

- Classifier
- Classification encoder
- Classification Policy

“beach”

Plug observation completion policy in for new task
Motion policy transfer

Unsupervised exploratory policy approaches supervised task-specific policy accuracy!
Summary

• Learning to examine 3D objects and scenes
  – Linking physical egomotion to visual features
  – One-shot viewgrid reconstruction for feature learning
  – Active recognition and exploration
  – Potential for pretraining motion policies

Dinesh Jayaraman
Papers


• Learning to look around, Dinesh Jayaraman, Kristen Grauman, arXiv Sept 2017