Visual Learning with Unlabeled Video and Look-Around Policies

Kristen Grauman
Department of Computer Science
University of Texas at Austin
Visual recognition: significant recent progress

Big labeled datasets

Deep learning

GPU technology

ImageNet top-5 error (%)
How do systems typically learn about objects today?

- dog
- boat
Recognition benchmarks

A “disembodied” well-curated moment in time


Egocentric perceptual experience

A tangle of relevant and irrelevant multi-sensory information

Kristen Grauman, UT Austin
Big picture goal: Embodied visual learning

Status quo:
Learn from “disembodied” bag of labeled snapshots.

On the horizon:
Visual learning in the context of acting and moving in the world.

Kristen Grauman, UT Austin
Big picture goal: Embodied visual learning

Status quo:
Learn from “disembodied” bag of labeled snapshots.

On the horizon:
Visual learning in the context of acting and moving in the world.

Kristen Grauman, UT Austin
This talk

Towards embodied visual learning

1. Learning from unlabeled video and multiple sensory modalities

2. Learning policies for how to move for recognition and exploration
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:
Equivariant embedding organized by ego-motions

Training data
Unlabeled video + motor signals

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

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

Approach: Ego-motion equivariance

Motor signal
time \rightarrow

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]
Approach: Egomotion equivariance

Training data
Unlabeled video + motor signals

Equivariant embedding
organized by egomotions

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

Learn from *unlabeled* car video (KITTI)

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

30% accuracy increase when labeled data scarce

Geiger et al, IJRR ’13

Xiao et al, CVPR ’10
Passive $\rightarrow$ complete ego-motions

Pre-recorded video $\rightarrow$ Moving around to inspect

motor signal $\rightarrow$ time
One-shot reconstruction

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

Shape from many views
geometric problem

Shape from one view
semantic problem

[Snavely et al, CVPR ‘06]

[Shiva et al, ICCV’93]
Approach: ShapeCodes

- Implicit 3D shape representation
- No “canonical” azimuth to exploit
- Category agnostic

[Jayaraman & Grauman, arXiv 2017, ECCV 2018]
ShapeCodes for recognition

ModelNet
[Wu et al 2015]

ShapeNet
[Chang et al 2015]

Accuracy (%)

- Pixels
- Random wts
- DrLIM*
- Autoencoder**
- LSM^
- Ours

*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
Ego-motion and implied body pose

Learn relationship between egocentric scene motion and 3D human body pose

Input:

egocentric video

Output:

sequence of 3d joint positions

[Jiang & Grauman, CVPR 2017]
Ego-motion and implied body pose

Learn relationship between egocentric scene motion and 3D human body pose

Wearable camera video
Inferred pose of camera wearer

[Jiang & Grauman, CVPR 2017]
Towards embodied visual learning

1. Learning from unlabeled video and multiple sensory modalities
   a) Egomotion / motor signals
   b) Audio signals

2. Learning policies for how to move for recognition and exploration
Recall: Disembodied visual learning
Listening to learn
Listening to learn
Listening to learn

woof  meow  ring  clatter

**Goal**: A repertoire of objects and their sounds

Kristen Grauman, UT Austin
Visually-guided audio source separation

Traditional approach:
• Detect low-level correlations within a single video
• Learn from clean single audio source examples

[Darrell et al. 2000; Fisher et al. 2001; Rivet et al. 2007; Barzelay & Schechner 2007; Casanovas et al. 2010; Parekh et al. 2017; Pu et al. 2017; Li et al. 2017]
Learning to separate object sounds

Our idea: Leverage visual objects to learn from *unlabeled* video with *multiple* audio sources

[Image: Diagram of unlabeled video connected to disentangled object sound models for Violin, Dog, and Cat]

[Gao, Feris, & Grauman, arXiv 2018]
Our approach: learning

Deep multi-instance multi-label learning (MIML) to disentangle which visual objects make which sounds

Unlabeled video → Visual frames → Visual predictions (ResNet-152 objects) → Non-negative matrix factorization → Audio basis vectors

Output: Group of audio basis vectors per object class
Our approach: inference

Given a novel video, use **discovered object sound models** to guide audio source separation.

**Visual predictions** (ResNet-152 objects) 

**Initialize audio basis matrix**

**Estimate activations**

**Frames**

**Novel video**

**Audio**

**Visual predictions** (ResNet-152 objects)

**Violin bases**

**Piano bases**

**Semi-supervised source separation using NMF**

**Violin Sound**

**Piano Sound**
Results: Separating object sounds

Train on 100,000 unlabeled video clips, then separate audio for novel video


[Gao, Feris, & Grauman, arXiv 2018]
Results: Separating object sounds

Train on 100,000 unlabeled video clips, then separate audio for novel video

Failure case
original video
(before separation)

visual predictions:
acoustic guitar & electric guitar

Failure cases

[Gao, Feris, & Grauman, arXiv 2018]
Results: Separating object sounds

<table>
<thead>
<tr>
<th>Upper-Bound</th>
<th>Instrument Pair</th>
<th>Animal Pair</th>
<th>Vehicle Pair</th>
<th>Cross-Domain Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means Clustering</td>
<td>-2.85</td>
<td>-3.76</td>
<td>-2.71</td>
<td>-3.32</td>
</tr>
<tr>
<td>MFCC Unsupervised [65]</td>
<td>0.47</td>
<td>-0.21</td>
<td>-0.05</td>
<td>1.49</td>
</tr>
<tr>
<td>Visual Exemplar</td>
<td>-2.41</td>
<td>-4.75</td>
<td>-2.21</td>
<td>-2.28</td>
</tr>
<tr>
<td>Unmatched Bases</td>
<td>-2.12</td>
<td>-2.46</td>
<td>-1.99</td>
<td>-1.93</td>
</tr>
<tr>
<td>Gaussian Bases</td>
<td>-8.74</td>
<td>-9.12</td>
<td>-7.39</td>
<td>-8.21</td>
</tr>
<tr>
<td>Ours</td>
<td>1.83</td>
<td>0.23</td>
<td>0.49</td>
<td>2.53</td>
</tr>
</tbody>
</table>

Visually-aided audio source separation (SDR)

<table>
<thead>
<tr>
<th></th>
<th>Wooden Horse</th>
<th>Violin Yanni</th>
<th>Guitar Solo</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse CCA (Kidron et al. [43])</td>
<td>4.36</td>
<td>5.30</td>
<td>5.71</td>
<td>5.12</td>
</tr>
<tr>
<td>JIVE (Lock et al. [50])</td>
<td>4.54</td>
<td>4.43</td>
<td>2.64</td>
<td>3.87</td>
</tr>
<tr>
<td>Audio-Visual (Pu et al. [56])</td>
<td>8.82</td>
<td>5.90</td>
<td>14.1</td>
<td>9.61</td>
</tr>
<tr>
<td>Ours</td>
<td>12.3</td>
<td>7.88</td>
<td>11.4</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Visually-aided audio denoising (NSDR)

Train on 100K unlabeled video clips from AudioSet [Gemmeke et al. 2017]
This talk

Towards embodied visual learning

1. Learning from unlabeled video and multiple sensory modalities

2. Learning policies for how to move for recognition and exploration
Moving to recognize

Time to revisit active recognition in challenging settings!

End-to-end active recognition

[Jayaraman and Grauman, ECCV 2016, PAMI 2018]
End-to-end active recognition

Look around scene  Manipulate object  Move around an object

[Jayaraman and Grauman, ECCV 2016, PAMI 2018]
Agents that learn to look around intelligently can recognize things faster.

[Jayaraman and Grauman, ECCV 2016, PAMI 2018]
End-to-end active recognition: example

P("Plaza courtyard"): (6.28)
Top 3 guesses: Restaurant
Train interior
Shop

(11.95)
Theater
Restaurant
Plaza courtyard

(68.38)
Plaza courtyard
Street
Theater

[Jayaraman and Grauman, ECCV 2016, PAMI 2018]
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, PAMI 2018]
Goal: Learn to “look around”

recognition vs. reconnaissance vs. search and rescue

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, CVPR 2018
Approach: Active observation completion

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

Jayaraman and Grauman, CVPR 2018
Active “look around” results

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

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

Jayaraman and Grauman, CVPR 2018
Agent’s mental model for 360 scene evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018
Active “look around” visualization

Agent’s mental model for 3D object evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018
Active “look around” visualization

Agent’s mental model for 3D object evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018
Active “look around” visualization

Agent’s mental model for 3D object evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018
Motion policy transfer

Unsupervised observation completion

Supervised active recognition

Plug observation completion policy in for new task
Motion policy transfer

SUN 360 Scenes

ModelNet Objects

Unsupervised exploratory policy approaches supervised task-specific policy accuracy!

Jayaraman and Grauman, CVPR 2018
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
  – Learning to separate object sound models from unlabeled video
  – Active policies for view selection and camera control

Kristen Grauman, UT Austin
Papers/code/videos


• **Learning Image Representations Tied to Egomotion from Unlabeled Video.** D. Jayaraman and K. Grauman. International Journal of Computer Vision (IJCV), Special Issue for Best Papers of ICCV 2015, Mar 2017. [pdf] [preprint] [project page, pretrained models]