Action and Interaction for Scene Understanding

Kristen Grauman
University of Texas at Austin

Dinesh Jayaraman
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]

active kitten
 passive kitten

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

Our idea: **Ego-motion ↔ vision**

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 $\leftrightarrow$ vision for recognition

Learning this connection requires:

- Depth, 3D geometry
- Semantics
- Context

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

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

- 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

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

**Equivariant features**: predictably responsive to some classes of transformations, through simple mappings (e.g., linear)

\[ z(gx) \approx \mathcal{M}_g z(x) \]

Invariance **discards** information; equivariance **organizes** it.
Approach idea: Ego-motion equivariance

Training data
Unlabeled video + motor signals

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

**Given:**
- $x_i$ (image)
- $g$ (motion)
- $z_\theta(gx_i)$

**Desired:** for all motions $g$ and all images $x$,

$$z_\theta(gx) \approx M_g z_\theta(x)$$

**Unsupervised training**
- $x_i$ is given
- $z_\theta(x_i)$ is produced
- $M_g$ is applied
- $\| M_g z_\theta(x_i) - z_\theta(gx_i) \|_2$ is minimized

**Supervised training**
- $x_k$ and class $y_k$ are given
- $z_\theta(x_k)$ is produced
- $W$ is applied
- Softmax loss $L_C(x_k, y_k)$

$\theta, M_g$ and $W$ jointly trained

[Jayaraman & Grauman, ICCV 2015]
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
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

397 classes
KITTImap → SUN

6 labeled training examples per class

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

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

**Mobahi et al., Deep Learning from Temporal Coherence in Video, ICML’09
Learning from arbitrary unlabeled video?
Our idea: Steady feature analysis

Learning from arbitrary unlabeled video

Equivariance \approx \text{“steadily” varying frame features!}

\frac{d^2 z_\theta(x_t)}{dt^2} \approx 0

[Jayaraman & Grauman, CVPR 2016]
Equivariance $\approx$ "steadily" varying frame features!

$$d^2z_\theta(x_t)/dt^2 \approx 0$$

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

[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!

Learning how to move for object recognition

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

NORB data

Accuracy (%)

- Random
- DrLim [Hadsell et al.]
- Temporal [Mobahi et al.]
- Ours

[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("Church"): (0.53)
Top 3 guesses:
- Forest
- Cave
- Beach

P("Plaza courtyard"): (0.28)
Top 3 guesses:
- Restaurant
- Train interior
- Shop

Active recognition: results
Jayaraman and Grauman, UT TR AI15-06
Active recognition: results

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
  

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

\[ y = [p_x, p_y, \alpha] \rightarrow \text{Interactee localization} \]
Approach: Learning to localize interactees

Interaction-guided embedding + locally weighted regression

CNN fine-tuned for interactees

Head/torso orientation
[Bordev et al. 11]
Results: interactee detection

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

System has no object detector for the highlighted objects!
Tasks leveraging interactees

Prior for “what to mention” about the scene

<table>
<thead>
<tr>
<th>Method</th>
<th>Mention rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth interactee</td>
<td>78.4 (0.6)</td>
</tr>
<tr>
<td>Ours-embedding</td>
<td><strong>70.5 (0.4)</strong></td>
</tr>
<tr>
<td>Importance (Berg et al 2012)</td>
<td>65.4 (0.4)</td>
</tr>
<tr>
<td>Ours-MDN</td>
<td>65.2 (0.5)</td>
</tr>
<tr>
<td>Near Person</td>
<td>67.5 (0.5)</td>
</tr>
<tr>
<td>Prior</td>
<td>64.6 (0.6)</td>
</tr>
<tr>
<td>Majority</td>
<td>51.7 (0.6)</td>
</tr>
</tbody>
</table>
Tasks leveraging interactees

Prior for “what to mention” about the scene

1. A little boy in a chair eating a cake.
2. A small boy is reaching up for a frisbee.
Tasks leveraging interactees

Prior for “what to mention” about the scene

Query

The men is flying a kite on a sunny day

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

[Devlin et al. 15]
Tasks leveraging interactees

Prior for “what to mention” about the scene

Men walking into the ocean with their surfboards

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

A man surfs on a surfboard on a lake

A man with a surf board walks across the beach

A young man carrying a surfboard next to a wave

[Ordonez et al. 11]
Tasks leveraging interactees

Image retargeting that preserves interactee region

Input

Ours

Baseline (objectness)
Tasks leveraging interactees

Focus object detector’s search

(a) Using computer

(b) Reading
Summary

– “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

CVPR 2016 Scene Understanding Workshop (SUNw)
Papers

• Egomotion and visual learning

• Interaction and scene understanding