Learning image representations from unlabeled video

Kristen Grauman
Department of Computer Science
The University of Texas at Austin

Work with Dinesh Jayaraman
Learning visual categories

• Recent major strides in category recognition

• Facilitated by large labeled datasets

[Images of classification error over years, showing improvements in accuracy.]

[Images of categories: red fox (100), hen-of-the-woods (100), ibex (100), goldfinch (100).]

[Images of datasets: ImageNet [Deng et al.], 80M Tiny Images [Torralba et al.], SUN Database [Xiao et al.].]

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.
Beyond “bags of labeled images”? 

Visual development in nature is based on:

- continuous observation
- multi-sensory feedback
- motion and action

… in an environment.

Inexpensive, and unrestricted in scope

Evidence from: psychology, evolutionary biology, cognitive science.

Talk overview

1. Learning representations tied to ego-motion

2. Learning representations from unlabeled video

3. Learning how to move and where to look

Kristen Grauman, UT Austin
The kitten carousel experiment
[Held & Hein, 1963]

Key to perceptual development:
self-generated motion + visual feedback

Kristen Grauman, UT Austin
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

Kristen Grauman, UT Austin
Ego-motion ↔ vision: view prediction

After moving:

Kristen Grauman, UT Austin
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
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
- …

Kristen Grauman, UT Austin
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)

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

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

Kristen Grauman, UT Austin
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

Kristen Grauman, UT Austin
Approach idea: Ego-motion equivariance

**Training data**
Unlabeled video + motor signals

**Equivariant embedding**
organized by ego-motions

Kristen Grauman, UT Austin
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
Training frame pair mining

Discovery of ego-motion clusters

$g = \text{left turn}$

$g = \text{right turn}$

$g = \text{forward}$

Kristen Grauman, UT Austin
Training frame pair mining

Discovery of ego-motion clusters

- $g = \text{left turn}$
- $g = \text{forward}$
- $g = \text{right turn}$

Kristen Grauman, UT Austin
Ego-motion equivariant feature learning

**Given:**
- \(\mathbf{x}_i\)
- \(g\)
- \(g\mathbf{x}_i\)

**Desired:** for all motions \(g\) and all images \(\mathbf{x}\),
\[
z_\theta(g\mathbf{x}) \approx M_g z_\theta(\mathbf{x})
\]

**Unsupervised training**
- \(z_\theta(\mathbf{x}_i)\)
- \(M_g\)
- \(\| M_g z_\theta(\mathbf{x}_i) - z_\theta(g\mathbf{x}_i) \|_2\)

**Supervised training**
- \(z_\theta(\mathbf{x}_k)\)
- \(W\)
- Softmax loss \(L_C(\mathbf{x}_k, y_k)\)

\(\theta, M_g\) and \(W\) jointly trained

Kristen Grauman, UT Austin
Method recap

**APPROACH**

- **Ego-motion training pairs**
- **Neural network training**
- **Equivariant embedding**

**RESULTS**

- Scene and object recognition
- Next-best view selection

- Football field?
- Pagoda?
- Airport?
- Cathedral?
- Army base?

- Cup
- Frying pan

Kristen Grauman, UT Austin
### Datasets

**KITTI video**
*Geiger et al. 2012*
- Car platform
- Egomotions: yaw and forward distance

<table>
<thead>
<tr>
<th>City</th>
<th>Residential</th>
<th>Road</th>
<th>Campus</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="City" /></td>
<td><img src="image2.png" alt="Residential" /></td>
<td><img src="image3.png" alt="Road" /></td>
<td><img src="image4.png" alt="Campus" /></td>
</tr>
</tbody>
</table>

**SUN images**
*Xiao et al. 2010*
- Large-scale scene classification task with 397 categories (static images)

**NORB images**
*LeCun et al. 2004*
- Toy recognition
- Egomotions: elevation and azimuth

Kristen Grauman, UT Austin
Results: Equivariance check

Visualizing how well equivariance is preserved

Query pair

Neighbor pair (our features)

Neighbor pair (pixel space)
Results: Equivariance check

How well is equivariance preserved?

<table>
<thead>
<tr>
<th>Methods</th>
<th>atomic</th>
<th>composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>CLSNET</td>
<td>0.9239</td>
<td>0.9145</td>
</tr>
<tr>
<td>TEMPORAL [19]</td>
<td>0.7587</td>
<td>0.8119</td>
</tr>
<tr>
<td>DRLIM [7]</td>
<td>0.6404</td>
<td>0.7263</td>
</tr>
<tr>
<td>EQUIV</td>
<td>0.6082</td>
<td>0.6982</td>
</tr>
<tr>
<td>EQUIV+DRLIM</td>
<td>0.5814</td>
<td>0.6492</td>
</tr>
</tbody>
</table>

Normalized error:

$$\rho_g = E \left[ \frac{\|z_\theta (x) - M_g' z_\theta (g x)\|_2}{\|z_\theta (x) - z_\theta (g x)\|_2} \right]$$


Kristen Grauman, UT Austin
Results: Recognition

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

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

Kristen Grauman, UT Austin

Geiger et al, IJRR ’13

Xiao et al, CVPR ’10
Do ego-motion equivariant features improve recognition?

6 labeled training examples per class

397 classes
KITTI ⟷ SUN

0.25
0.70
1.02
1.21
1.58

recognition accuracy (%)

Random
Supervised
DrLi* [Hadsell et al.]
Temporal** [Mobahi et al.]
Ours

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

Kristen Grauman, UT Austin
Recap so far

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

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

Kristen Grauman, UT Austin
Talk overview

1. Learning representations tied to ego-motion

2. Learning representations from unlabeled video

3. Learning how to move and where to look

Kristen Grauman, UT Austin
Learning from arbitrary unlabeled video?

Unlabeled video + ego-motion

Unlabeled video

Kristen Grauman, UT Austin
Learning from arbitrary unlabeled video?

Kristen Grauman, UT Austin
Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]

Find functions $g(x)$ that map

quickly varying input signal $x(t)$  $\rightarrow$  slowly varying features $y(t)$

Figure: Laurenz Wiskott, http://www.scholarpedia.org/article/File:SlowFeatureAnalysis-OptimizationProblem.png

Kristen Grauman, UT Austin
Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]

Find functions \( g(x) \) that map

quickly varying input signal \( x(t) \)  

\arrow{longrightarrow}  

slowly varying features \( y(t) \)

Figure: Laurenz Wiskott, http://www.scholarpedia.org/article/File:SlowFeatureAnalysis-OptimizationProblem.png  
Kristen Grauman, UT Austin
**Background: Slow feature analysis**

[Wiskott & Sejnowski, 2002]

- Existing work exploits “slowness” as **temporal coherence** in video → learn invariant representation


- Fails to capture **how** visual content changes over time

Kristen Grauman, UT Austin
Our idea: **Steady** feature analysis

- Higher order temporal coherence in video $\rightarrow$ learn equivariant representation

Second order slowness operates on frame triplets:

$$z(b) - z(a) \approx z(c) - z(b)$$

in learned embedding

[Jayaraman & Grauman, CVPR 2016]

Kristen Grauman, UT Austin
Approach: Steady feature analysis

Learn classifier $W$ and representation $\theta$ jointly,

$$(\theta^*, W^*) = \arg\min_{\theta, W} L_s(\theta, W, S) + \lambda L_u(\theta, U)$$

with unsupervised regularization loss:

$$L_u(\theta, U) = R_2(\theta, U) + \lambda' R_3(\theta, U)$$

Contrastive loss that also exploits “negative” tuples

Kristen Grauman, UT Austin
Approach: Steady feature analysis

supervised

slow

unsupervised

steady

[Jayaraman & Grauman, CVPR 2016]
Recap: Steady feature analysis

Equivariance \( \approx \) “steadily” varying frame features!

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

[Jayaraman & Grauman, CVPR 2016]

Kristen Grauman, UT Austin
Datasets

Unlabeled video

Human Motion Database (HMDB)

KITTI Video

NORB

Target task (few labels)

PASCAL 10 Actions

SUN 397 Scenes

NORB 25 Objects

32 x 32 images or 96 x 96 images
Results: Sequence completion

Given sequential pair, infer next frame (embedding)

\[ \tilde{z}_\theta(x_3) = 2z_\theta(x_2) - z_\theta(x_1) \]

\( x_1 \quad x_2 \quad \text{Our top 3 estimates for } x_3 \)

KITTI dataset

Kristen Grauman, UT Austin
### Results: Sequence completion

Given sequential pair, infer next frame (embedding)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>NORB</th>
<th>KITTI</th>
<th>HMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>slow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFA-1 [30] *</td>
<td>0.95</td>
<td>31.04</td>
<td>2.70</td>
</tr>
<tr>
<td>slow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFA-2 [14] **</td>
<td>0.91</td>
<td>8.39</td>
<td>2.27</td>
</tr>
<tr>
<td>slow &amp; steady</td>
<td>SSFA (ours)</td>
<td>0.53</td>
<td>7.79</td>
</tr>
</tbody>
</table>

Percentile rank of correct completion (lower is better)

---

*Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR’06*

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

Kristen Grauman, UT Austin
## Results: Recognition

<table>
<thead>
<tr>
<th>Task type→</th>
<th>Objects</th>
<th>Scenes</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datasets→</td>
<td>NORB→NORB</td>
<td>KITTI→SUN</td>
<td>HMDB→PASCAL-10</td>
</tr>
<tr>
<td>Methods↓</td>
<td>[25 cls]</td>
<td>[397 cls]</td>
<td>[397 cls, top-10]</td>
</tr>
<tr>
<td>random</td>
<td>4.00</td>
<td>0.25</td>
<td>2.52</td>
</tr>
<tr>
<td>UNREG</td>
<td>24.64±0.85</td>
<td>0.70±0.12</td>
<td>6.10±0.67</td>
</tr>
<tr>
<td>SFA-1 [30]*</td>
<td>37.57±0.85</td>
<td>1.21±0.14</td>
<td>8.24±0.25</td>
</tr>
<tr>
<td>SFA-2 [14]**</td>
<td>39.23±0.94</td>
<td>1.02±0.12</td>
<td>6.78±0.32</td>
</tr>
<tr>
<td>SSFA (ours)</td>
<td>42.83±0.33</td>
<td>1.65±0.04</td>
<td>9.19±0.10</td>
</tr>
</tbody>
</table>

Multi-class recognition accuracy

*Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR’06

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

Kristen Grauman, UT Austin
Pre-training a representation

**Supervised pre-training**
- Labeled images from a related domain
- Few labeled images for target task
- Fine-tune

**Unsupervised “pre-training”**
- Unlabeled video
- Few labeled images for target task

Kristen Grauman, UT Austin
Results: Can we learn more from unlabeled video than “related” labeled images?
Results: Can we learn more from unlabeled video than “related” labeled images?

Kristen Grauman, UT Austin
Results: Can we learn more from unlabeled video than “related” labeled images?

Better even than providing 50,000 extra manual labels for auxiliary classification task!
Talk overview

1. Learning representations tied to ego-motion

2. Learning representations from unlabeled video

3. Learning how to move and where to look

Kristen Grauman, UT Austin
Learning how to move for recognition

Time to revisit active recognition in challenging settings!


Kristen Grauman, UT Austin
Learning how to move for 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

0 10 20 30 40 50

[Jayarman & Grauman, ICCV 2015]

Kristen Grauman, UT Austin
Learning how to move for 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

Kristen Grauman, UT Austin
Active visual recognition

Requires several separate functionalities:
- Action selection
- Per-view processing
- Across-view evidence aggregation
- Next-view prediction
- Final class belief prediction

Learn all end-to-end

Kristen Grauman, UT Austin
P(“Plaza courtyard”):  
Top 3 guesses:  
Restaurant  
Train interior  
Shop  
(6.28)  
Restaurant  
Theater  
Plaza courtyard  
(11.95)  
Plaza courtyard  
Street  
Theater  
(68.38)
Active recognition: Results

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

Kristen Grauman, UT Austin
Summary

• Visual learning requires
  – context of action and motion in the world
  – with continuous self-acquired feedback

• New ideas:
  – “Embodied” feature learning using both visual and motor signals
  – Feature learning from unlabeled video via higher order temporal coherence
  – Steps towards active view selection in 360 scenes

Kristen Grauman, UT Austin
References

