Action and Attention in First-Person Vision

Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Dinesh Jayaraman, Yong Jae Lee, Yu-Chuan Su, Bo Xiong, Lu Zheng
New era for first-person vision

Augmented reality

Health monitoring

Law enforcement

Science

Robotics

Life logging

Kristen Grauman, UT Austin
First person vs. Third person

Traditional third-person view

First-person view

UT TEA dataset

Kristen Grauman, UT Austin
First person vs. Third person

First person “egocentric” vision:

• Linked to ongoing experience of the camera wearer

• World seen in context of the camera wearer’s activity and goals
Recent egocentric work

• Activity and object recognition


• Gaze and social cues


• Visualization, stabilization

  [Kopf et al. 2014, Poleg et al. 2015]
Talk overview

Motivation

Account for the fact that camera wearer is active participant in the visual observations received

Ideas

1. **Action**: Unsupervised feature learning  
   • How is visual learning shaped by ego-motion?

2. **Attention**: Inferring highlights in video  
   • How to summarize long egocentric video?

Kristen Grauman, UT Austin
Visual recognition

• Recent major strides in category recognition

  ![Classification error chart](chart.png)

  red fox (100) hen-of-the-woods (100) ibex (100) goldfinch (100)

• Facilitated by large labeled datasets

  ![ImageNet](image.png)
  ![80M Tiny Images](image.png)
  ![SUN Database](image.png)

Problem with today’s visual learning

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

- …yet visual perception develops in the context of **acting** and **moving** in the world

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

Key to perceptual development:
Self-generated motions + visual feedback

Kristen Grauman, UT Austin
Our idea:
Feature learning with ego-motion

**Goal:** Learn the connection between
“how I move” ↔ “how visual surroundings change”

**Approach:** Unsupervised feature learning using motor signals accompanying egocentric video
Key idea: Egomotion 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, whereas equivariance *organizes* it.

Kristen Grauman, UT Austin
Key idea: Egomotion equivariance

Training data=
Unlabeled video
+ motor signals

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

Equivariant embedding organized by egomotions

Kristen Grauman, UT Austin
Key idea: Egomotion equivariance

Training data=
Unlabeled video + motor signals

Equivariant embedding organized by egomotions

Kristen Grauman, UT Austin
Approach

Ego motor signals + Observed image pairs

Deep learning architecture

Output embedding

Kristen Grauman, UT Austin

[Jayaraman & Grauman, ICCV 2015]
Approach

"Active": Exploit knowledge of observer motion

Kristen Grauman, UT Austin

[Jayaraman & Grauman, ICCV 2015]
Learning equivariance

ego-motion data stream

\( \forall x \in \mathcal{X} : z_\theta(gx) \approx M_g^* z_\theta(x) \)

Unlabeled video frame pairs

Class-labeled images

Kristen Grauman, UT Austin

[Jayaraman & Grauman, ICCV 2015]
Learning equivariance

Unlabeled video frame pairs

Class-labeled images

Kristen Grauman, UT Austin

[Jayaraman & Grauman, ICCV 2015]

Embedding objective:

\[
(\theta^*, (\theta^*, M^*)) = \arg \min_{\theta M} \sum_{g, i, j} d_g(z_\theta(x_i), M_g z_\theta(x_j), p_{ij}) + \lambda L_c(W, \mathcal{L})
\]
Datasets

KITTI video
[Geiger et al. 2012]
Autonomous car platform
Egomotions: yaw and forward distance

SUN images
[Xiao et al. 2010]
Large-scale scene classification task

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 (ours)

Pixel space neighbor pair

Kristen Grauman, UT Austin  [Jayaraman & Grauman, ICCV 2015]
Results: Equivariance check

Visualizing how well equivariance is preserved

Query pair

Neighbor pair (ours)

Pixel space neighbor pair

Kristen Grauman, UT Austin

[Jayaraman & Grauman, ICCV 2015]
Results: Recognition

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

Exploit features for large multi-way **scene classification** (SUN, 397 classes)

30% accuracy increase for small labeled training sets

Kristen Grauman, UT Austin

[Jayaraman & Grauman, ICCV 2015]
Results: Recognition

Do the learned features boost recognition accuracy?

**KITTI->SUN**

397 classes

6 labeled training examples per class

*Mobahi et al. ICML09; **Hadsell et al. CVPR06

Kristen Grauman, UT Austin
Leverage proposed equivariant embedding to predict next best view for object recognition.

Results: Active recognition


Kristen Grauman, UT Austin
Next steps

- Dynamic objects
- Multiple modalities, e.g., depth
- Active ego-motion planning
- Tasks aside from recognition
Talk overview

Motivation

Account for the fact that camera wearer is active participant in the visual observations received

Ideas

1. Action: Unsupervised feature learning
   • How is visual learning shaped by ego-motion?

2. Attention: Inferring highlights in video
   • How to summarize long egocentric video?
**Goal:** Summarize egocentric video

**Input:** Egocentric video of the camera wearer’s day

**Output:** Storyboard (or video skim) summary

Kristen Grauman, UT Austin
Potential applications of egocentric video summarization

- Memory aid
- Law enforcement
- Mobile robot discovery
Prior work: Video summarization

• Largely third-person
  – Static cameras, low-level cues informative
• Consider summarization as a *sampling* problem

What makes egocentric data hard to summarize?

- Subtle event boundaries
- Subtle figure/ground
- Long streams of data

Kristen Grauman, UT Austin
Summarizing egocentric video

Key questions

– What objects are important, and how are they linked?
– When is attention heightened?
– Which frames look “intentional”?
Goal: Story-driven summarization

Characters and plot ↔ Key objects and influence

Kristen Grauman, UT Austin  [Lu & Grauman, CVPR 2013]
Goal: Story-driven summarization

Characters and plot ↔ Key objects and influence

Kristen Grauman, UT Austin  
[Lu & Grauman, CVPR 2013]
Summarization as subshot selection

Good summary = chain of $k$ selected subshots in which each influences the next via some subset of key objects

$$S^* = \arg \max_{S \subset \mathcal{V}} \lambda_s \mathcal{S}(S) + \lambda_i \mathcal{I}(S) + \lambda_d \mathcal{D}(S)$$

influence importance diversity

Kristen Grauman, UT Austin

[Lu & Grauman, CVPR 2013]
Estimating visual influence

- Aim to select the $k$ subshots that maximize the influence between objects (on the weakest link)

$$S(S') = \max_a \min_{j=1,...,K-1} \sum_{o_i \in O} a_{i,j} \text{INFLUENCE}(s_j, s_{j+1}|o_i)$$
Estimating visual influence

\[ \text{INFLUENCE}(s_i, s_j | o) = \prod_i(s_j) - \prod_i^o(s_j) \]

Captures how reachable subshot \( j \) is from subshot \( i \), via any object \( o \)

Kristen Grauman, UT Austin  

[Lu & Grauman, CVPR 2013]
Learning object importance

We learn to rate regions by their egocentric importance

Kristen Grauman, UT Austin

[Lee et al. CVPR 2012, IJCV 2015]
Learning object importance

We learn to rate regions by their egocentric importance

Distance to hand
Distance to frame center
Frequency

"Object-like" appearance, motion
[Endres et al. ECCV 2010, Lee et al. ICCV 2011]

Region features: size, width, height, centroid
[Lee et al. CVPR 2012, IJCV 2015]
Datasets

UT Egocentric (UT Ego) [Lee et al. 2012]

4 videos, each 3-5 hours long, uncontrolled setting.
We use visual words and subshots.

Activities of Daily Living (ADL) [Pirsiavash & Ramanan 2012]

20 videos, each 20-60 minutes, daily activities in house.
We use object bounding boxes and keyframes.

Kristen Grauman, UT Austin
Example keyframe summary – UT Ego data

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

Original video (3 hours)  Our summary (12 frames)

Kristen Grauman, UT Austin  [Lee et al. CVPR 2012, IJCV 2015]
Example keyframe summary – UT Ego data

Alternative methods for comparison

Uniform keyframe sampling (12 frames)

[Liu & Kender, 2002]
(12 frames)

Kristen Grauman, UT Austin

[Lee et al. CVPR 2012, IJCV 2015]
Generating storyboard maps

Augment keyframe summary with geolocations

Kristen Grauman, UT Austin

[Lee et al., CVPR 2012, IJCV 2015]
Human subject results: Blind taste test

How often do subjects prefer our summary?

<table>
<thead>
<tr>
<th>Data</th>
<th>Vs. Uniform sampling</th>
<th>Vs. Shortest-path</th>
<th>Vs. Object-driven Lee et al. 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>UT Egocentric Dataset</td>
<td>90.0%</td>
<td>90.9%</td>
<td>81.8%</td>
</tr>
<tr>
<td>Activities Daily Living</td>
<td>75.7%</td>
<td>94.6%</td>
<td>N/A</td>
</tr>
</tbody>
</table>

34 human subjects, ages 18-60
12 hours of original video
Each comparison done by 5 subjects

Total 535 tasks, 45 hours of subject time

Kristen Grauman, UT Austin

[Lu & Grauman, CVPR 2013]
Summarizing egocentric video

Key questions

– What objects are important, and how are they linked?
– When is attention heightened?
– Which frames look “intentional”? 
Goal: Detect heightened attention

Definition:
A time interval where the recorder is attracted by some object(s) and he interrupts his ongoing flow of activity to purposefully gather more information about the object(s).

[Su & Grauman, 2015]
Temporal Ego-Attention Dataset

- "Browsing" scenarios, long & natural clips
- 14 hours of video, 9 recorders
- Frame-level labels x 10 annotators

14 hours of labeled ego video

Kristen Grauman, UT Austin

[Su & Grauman, 2015]
Challenges in temporal attention

- Interesting things vary in appearance!
- Attention ≠ stationary
- High attention intervals vary in length
- Lack cues of active camera control

Kristen Grauman, UT Austin

[Su & Grauman, 2015]
Our approach

Learn motion patterns indicative of attention

Kristen Grauman, UT Austin

[Su & Grauman, 2015]
Results: detecting temporal attention

Blue=Ground truth
Red=Predicted

[Su & Grauman, 2015]
Results: detecting temporal attention

- 14 hours of video, 9 recorders

Kristen Grauman, UT Austin

[Su & Grauman, 2015]
Summarizing egocentric video

Key questions

– What objects are important, and how are they linked?
– When is attention heightened?
– Which frames look “intentional”?
Which photos were purposely taken by a human?

- Incidental wearable camera photos
- Intentional human taken photos

Kristen Grauman, UT Austin

[Xiong & Grauman, ECCV 2014]
Idea: Detect “snap points”

• Unsupervised data-driven approach to detect frames in first-person video that look intentional.
Example snap point predictions

[Xiong & Grauman, ECCV 2014]

Kristen Grauman, UT Austin
Snap point predictions

Kristen Grauman, UT Austin

[Xiong & Grauman, ECCV 2014]
Next steps

• Video summary as an index for search
• Streaming computation
• Visualization, display
• Multiple modalities – e.g., audio
Summary

- New opportunities with “always on” ego cameras
- Towards active first-person vision:
  - **Action**: “Embodied” feature learning from ego-video using both visual and motor signals
  - **Attention**: Egocentric summarization tools to cope with deluge of wearable camera data