What to Keep?:
Summarizing Long Egocentric Videos

Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Yong Jae Lee, Bo Xiong, Lu Zheng
**Goal:** Summarize egocentric video

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

**Output:** Storyboard (or video skim) summary
Potential applications of egocentric video summarization

- Memory aid
- Law enforcement
- Mobile robot discovery
What makes egocentric data hard to summarize?

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

• **Egocentric recognition**
  

• **Video summarization**
  

→ **Low-level cues, stationary cameras**
→ **Consider summarization as a *sampling* problem**
Our idea:
Story-driven summarization

[Lu & Grauman, CVPR 2013]
Our idea:
Story-driven summarization

Good summary captures the progress of the story

1. Segment video temporally into subshots

2. Select chain of $k$ subshots that maximize both weakest link’s influence and object importance

[Lu & Grauman, CVPR 2013]
Egocentric subshot detection

Define 3 generic ego-activities:

- Static
- In transit
- Head moving

- Train classifiers to predict these activity types
- Features based on flow and motion blur

[Lu & Grauman, CVPR 2013]
Egocentric subshot detection

Ego-activity classifier

MRF and frame grouping

[Lu & Grauman, CVPR 2013]
Subshot selection objective

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

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

Subshots

[Lu & Grauman, CVPR 2013]
Learning region importance

- First task: watch a short clip, and describe in text the essential people or objects necessary to create a summary.

[Lee et al. CVPR 2012]
Learning region importance

- **Second task**: draw polygons around any described person or object obtained from the first task in sampled frames

[Lee et al. CVPR 2012]
Learning region importance

Video input

Generate candidate object regions for uniformly sampled frames

[Lee et al. CVPR 2012]
Learning region importance

Egocentric features:

- distance to hand
- distance to frame center
- frequency

[Lee et al. CVPR 2012]
Learning region importance

Egocentric features:

*distance to hand*

*distance to frame center*

*frequency*

Object features:

“Object-like” appearance, motion

- Candidate region’s appearance, motion
- Surrounding area’s appearance, motion

Overlap w/ face detection

Region features: *size, width, height, centroid*
Influence criterion

- Want the $k$ subshots that maximize the weakest link’s influence, subject to coherency constraints

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

[Lu & Grauman, CVPR 2013]
Estimating visual influence

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

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

[Lu & Grauman, CVPR 2013]
Estimating visual influence

- Prefer small number of objects at once, and **coherent** (smooth) entrance/exit patterns

<table>
<thead>
<tr>
<th>Microwave</th>
<th>Bottle</th>
<th>Mug</th>
<th>Tea bag</th>
<th>Fridge</th>
<th>Food</th>
<th>Dish</th>
<th>Spoon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Our method**

<table>
<thead>
<tr>
<th>Microwave</th>
<th>Bottle</th>
<th>Food</th>
<th>Kettle</th>
<th>Fridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Uniform sampling**

[Lu & Grauman, CVPR 2013]
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 2009]

20 videos, each 20-60 minutes, daily activities in house.

We use **object** bounding boxes and **keyframes**.
Results: Important region prediction


Good predictions

[Lee et al., CVPR 2012]
Results: Important region prediction

Ours

Object-like [Carreira, 2010]

Object-like [Endres, 2010]

Saliency [Walther, 2005]

Failure cases

[Lee et al., CVPR 2012]
Results: Important region prediction

- Ours
- Object-like [Carreira, 2010]
- Object-like [Endres, 2010]
- Saliency [Walther, 2005]

Failure cases

[Lee et al., CVPR 2012]
Example keyframe summary – UT Ego data

Original video (3 hours)  Our summary (12 frames)
Example keyframe summary – UT Ego data

Alternative methods for comparison

Uniform keyframe sampling
(12 frames)

[Liu & Kender, 2002]
(12 frames)
Example summary – UT Ego data

**Ours**

**Baseline**
Example summary – ADL data

Ours

Baseline 1
Generating storyboard maps

Augment keyframe summary with geolocations

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

How often do subjects prefer our summary?

<table>
<thead>
<tr>
<th>Data</th>
<th>Uniform sampling</th>
<th>Shortest-path</th>
<th>Object-driven Lee et al. 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTE</td>
<td>90.0%</td>
<td>90.9%</td>
<td>81.8%</td>
</tr>
<tr>
<td>ADL</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

[Lu & Grauman, CVPR 2013]
Next steps

• Personalization
• Object-centric → activity-centric?
• Additional sensors
• Evaluation for search tasks
• Summaries while streaming
Which photos were purposely taken by a human?

Incidental wearable camera photos

Intentional human taken photos

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

- Unsupervised data-driven approach to detect frames in first-person video that look intentional.

[Xiong & Grauman, ECCV 2014]
Example snap point predictions

Ego data

Snap points

Robot data

Ego v3

Ego v4

Robot

precision

precision

precision

recall

recall

recall

0 0.5 1

0 0.5 1

0 0.5 1

0.49

0.633

0.323

0.674

0.685

0.19

0.243

0.137

0.289

0.301

0.305

0.253

0.236

0.319

0.326

Blur

People Likelihood

Saliency

Web Prior (ours)

Web Prior + DA (ours)
Snap points can boost precision for object detection

Person detection in intentional photos vs. Person detection in first-person frames
Snap points can boost precision for object detection.

Person detection in first-person frames.
Summary

- Deluge of first-person video imminent
  → Need **summaries** to access and browse

- First person video summarization
  - Estimate **influence** between events given their objects
  - Category-independent region **importance** prediction
  - **Snap point detection** with a Web prior