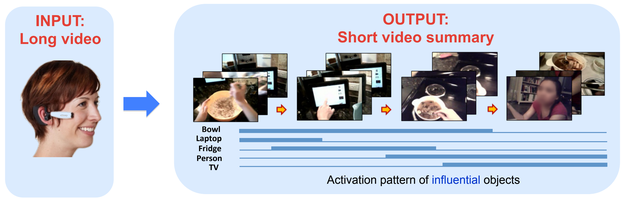
Story-Driven Summarization for Egocentric Video

Zheng Lu and Kristen Grauman
University of Texas at Austin

Our idea

Produce a short visual summary that contains the *story* of an egocentric video.



Key ideas

- ✧ Select a **coherent** chain of video subshots, beyond important persons and objects.
- ✧ Each subshot "***leads to***" the next through some subset of ***influential objects***.
- ✧ **Temporal segmentation** method uniquely designed for egocentric video.

Existing summarization methods often assume static cameras, and define sampling-based objectives that fail to model event influence. [Wolf ICASSP 1996, Ngo et al. ICCV 2003, Goldman et al. SIGGRAPH 2006, Pritch et al. ICCV 2007, Liu et al. PAMI 2010, Lee et al. CVPR 2012, …]

Potential applications

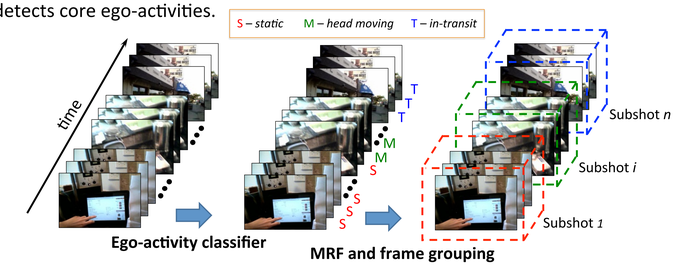
Memory aid

Law enforcement

Mobile robot discovery

Temporal subshot segmentation

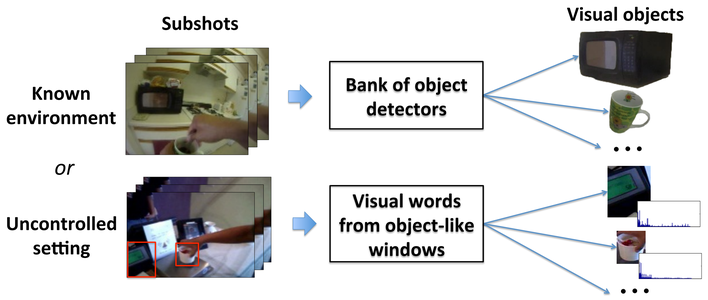
We propose a temporal segmentation method tailored to egocentric video that detects core ego-activities.



Provides an over-segmentation: a typical subshot lasts 15 seconds.

Subshot and object representation

For egocentric video, the story of activity is largely defined by **objects.**

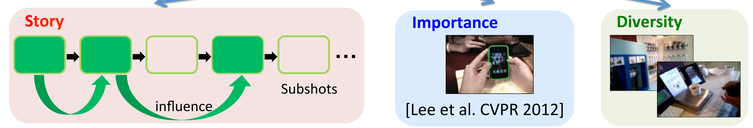


Subshot selection objective

Given a series of subshots segmented from the input video, our goal is to select the **optimal K-node chain** of subshots (or keyframes):

$$S^* = \arg\max_{S \subset \mathcal{V}} \; Q(S)$$

$$Q(S) = \lambda_s \mathcal{S}(S) + \lambda_i \mathcal{I}(S) + \lambda_d \mathcal{D}(S)$$



Story progress between subshots

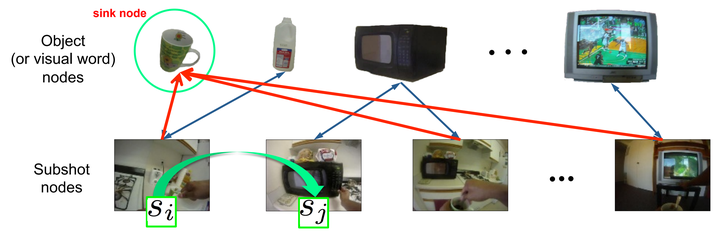
Good story: a ***coherent*** chain of subshots, where each strongly ***influences*** the next.

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

Maximize influence of weakest link — Object activation variables — Influence between two subshots conditioned on an object

Predicting influence between subshots

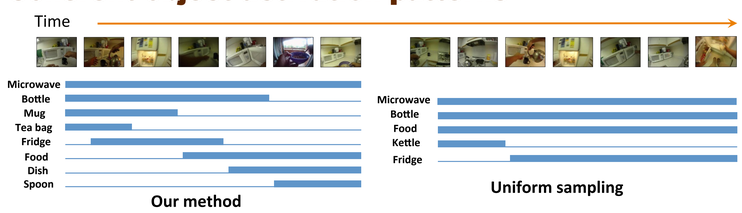
Construct a weighted bipartite graph over objects and the subshots that contain them.



$$\textsc{Influence}(s_i, s_j | o) = \prod\nolimits_i(s_j) - \prod\nolimits_i^o(s_j) \quad \text{[Shahaf \& Guestrin, KDD 2010]}$$

- Random walk: how reachable is subshot *j* from subshot *i*, **via each object *o***?
- Captures which objects "lead to" which---beyond simple object co-occurrence.

Coherent object activation patterns



Prefer activating few objects at once, and ***coherent*** (smooth) entrance/exit patterns.

Optimization: Relaxed activation variables optimized though linear programming. Priority queue to efficiently search for optimal K-node chain.

Results

Datasets

UT Egocentric (UTE)
[Lee et al. CVPR 2012]

4 videos, each 3-5 hours long, uncontrolled environment.

We use visual words and subshots.

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

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

We use object bounding boxes with keyframes.

Evaluating summary quality

Large-scale user study

- UTE: **5** hours and **11** events.
- ADL: **7** hours and **37** events for ADL.
- **34** subjects, from 18 – 60 years old.
- **5** users per comparison. Total **535** tasks, **45** hours of user time.

Data	Uniform sampling	Shortest-path	Lee et al. CVPR 2012
UTE	90.0%	90.9%	81.8%
ADL	75.7%	94.6%	N/A

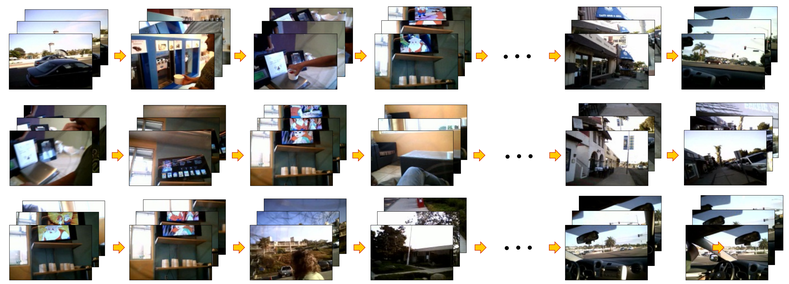
% of subjects who prefer our method's summary to the baseline

Blind taste test:

- Show sped up original video.
- Show our summary and one of baselines'.
- *Which better shows the progress of the story?*

Our failure cases: multi-thread or uneventful.

UTE example



Our method captures the progress of the story: serving ice cream leads to weighing the ice cream, which leads to watching TV in the ice cream shop, then driving home. While an existing object-driven method [Lee et al. CVPR 2012] finds important objects (e.g., TV, person), its summary fails to suggest the links between them.

ADL example

While here uniform sampling produces a plausible result, ours appears more coherent. Objects such as milk and cup connect the selected keyframes and show the progress of the story—preparing a hot drink and enjoying it by the TV. Shortest-path produces the weakest result due to its redundant keyframes.

Discovering influential objects

- Ours: objects ranked by **influence**.
- Baseline: objects ranked by **frequency**.
- Ground truth: workers on MTurk identify central objects to the story.

Results show our method's advantage; the most influential objects need not be the most frequent.