Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization

Jia Xu[†], Lopamudra Mukherjee[§], Yin Li[‡], Jamieson Warner[†], James M. Rehg[‡], Vikas Singh[†]

[†]University of Wisconsin-Madison, [§]University of Wisconsin-Whitewater, [‡]Georgia Institute of Technology

Motivation

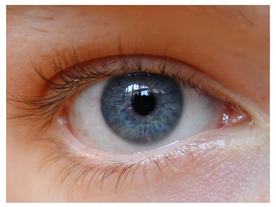
Egocentric video in real life:

- Life-logging with wearable cameras: GoPro, Google Glass
- Memory aid for the aging population (Alzheimer Disease)

Need for summarization:

- Personalization $\Longleftarrow$ Gaze
- Efficient algorithms $\Longleftarrow$ Submodular optimization

A submodular summarization model for egocentric videos, capturing common-sense properties of a good summary: relevance, diversity, compactness, and personalization



Problem Formulation

Relevance and Diversity Measurement:

- Mutual Information

$$M(\mathcal{V}\backslash\mathcal{S};\mathcal{S}) = H(\mathcal{V}\backslash\mathcal{S}) - H(\mathcal{V}\backslash\mathcal{S}|\mathcal{S}) = H(\mathcal{V}\backslash\mathcal{S}) + H(\mathcal{S}) - H(\mathcal{V})$$

- Entropy

$$H(\mathcal{S}) = \frac{1+\log(2\pi)}{2}|\mathcal{S}| + \frac{1}{2}\log(\det(L_{\mathcal{S}}))$$

where $L_{\mathcal{S}}$ is the principal submatrix of L indexed by $\mathcal{S}$

- Maximizing $M(\mathcal{V}\backslash\mathcal{S};\mathcal{S})$ is equivalent to maximizing

$$M(\mathcal{S}) = \frac{1}{2}\log(\det(L_{\mathcal{V}\backslash\mathcal{S}})) + \frac{1}{2}\log(\det(L_{\mathcal{S}}))$$

as $|\mathcal{S}| + |\mathcal{V}\backslash\mathcal{S}| = n$, and $H(\mathcal{V})$ is constant.

Relation to Determinantal Point Process:

Positive semidefinite kernel matrix L indexed by elements of $\mathcal{V}$

$$L_{ij} = \frac{\mathbf{v}_i^T \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}$$

Diversity score for $\mathcal{S} \in \mathcal{V}$

$$D(\mathcal{S}) = \log(\det(L_{\mathcal{S}}))$$

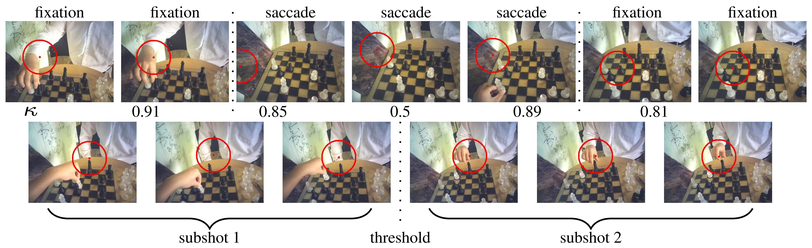
Partition Matroid Constraint:

- High level supervision: timeline
- Partition the video into b disjoint blocks $\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_b$
- Compactness: allocation bound for each block $\mathcal{I} = \{\mathcal{A} : \quad |\mathcal{A} \cap \mathcal{P}_m| \leq f_m, m = 1, 2, \cdots, b\}$

What can help the summarization?

Gaze in egocentric video summarization:

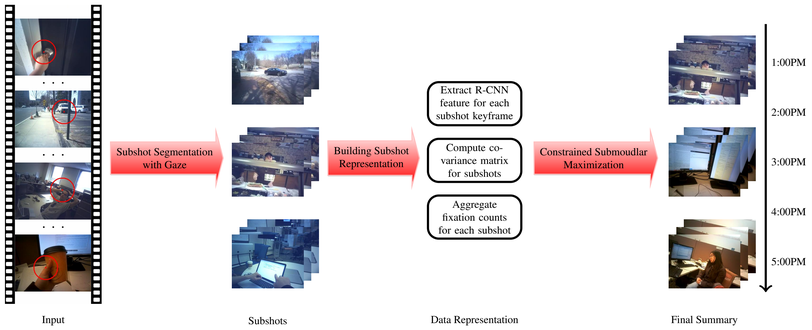
- Better temporal segmentation: egocentric video is continuous, but gaze is discrete
- Personalization: fixation tells intention



Attention Measurement using Gaze Fixations:

$$I(\mathcal{S}) = \sum_{i \in \mathcal{S}} c_i$$

Submodular Summarization



Main Model:

$$\max_{\mathcal{S}} \quad F(\mathcal{S}) = M(\mathcal{S}) + \lambda I(\mathcal{S})$$
$$\text{s.t.} \quad \mathcal{S} \in \mathcal{I}$$

Corollary: $F(\mathcal{S})$ is submodular.

Algorithm 1 Local Search for Constrained Submodular Maximization

1: **Input:** $\mathcal{M} = (\mathcal{V}, \mathcal{I}), F, \epsilon \geq 0$
2: Initialize $\mathcal{S} \leftarrow \emptyset$;
3: **while** (Any of the following local operations applies, update $\mathcal{S}$ accordingly) **do**
4: **Add operation.** If $e \in \mathcal{V}\backslash\mathcal{S}$ such that $\mathcal{S} \cup \{e\} \in \mathcal{I}$ and $F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) > \epsilon$, then $\mathcal{S} = \mathcal{S} \cup \{e\}$.
5: **Swap operation.** If $e_i \in \mathcal{S}$ and $e_j \in \mathcal{V}\backslash\mathcal{S}$ such that $(\mathcal{S}\backslash\{e_i\}) \cup \{e_j\} \in \mathcal{I}$ and $F((\mathcal{S}\backslash\{e_i\}) \cup \{e_j\}) - F(\mathcal{S}) > \epsilon$, then $\mathcal{S} = (\mathcal{S}\backslash\{e_i\}) \cup \{e_j\}$.
6: **Delete operation.** If $e \in \mathcal{S}$ such that $F(\mathcal{S}\backslash\{e\}) - F(\mathcal{S}) > \epsilon$, then $\mathcal{S} = \mathcal{S}\backslash\{e\}$.
7: **end while**
8: **return** $\mathcal{S}$;

Proposition: Greedy local search achieves a $\frac{1}{4}$-approximation factor for our constrained submodular maximization problem.

Data Collection

- 5 subjects to record their daily lives
- 21 videos with gaze
- 15 hours in total

Annotation: subjects group subshots into events.

Experimental Evaluation

Systematic Evaluation:

Method	uniform	kmeans	uniform (our subshots)	kmeans (our subshots)	ours
F-measure	0.161	0.215 ± 0.016	0.526	0.475 ± 0.026	**0.621**

Table 1: Comparisions of average F-measure on GTEA-GAZE+.

Method	uniform	kmeans	uniform (our subshots)	kmeans (our subshots)	ours
F-measure	0.080	0.095 ± 0.030	0.476	0.509 ± 0.025	**0.585**

Table 2: Comparisions of average F-measure on our new EgoSum+gaze dataset.

Visual Results:

Figure 1: Results from GTEA-gaze+: pizza preparation.

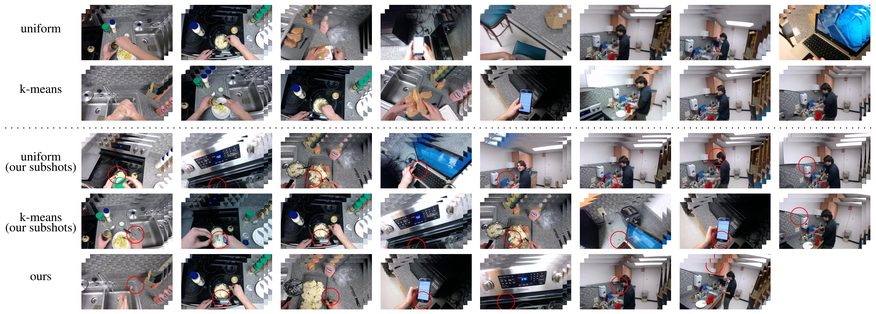


Figure 2: Results from our new EgoSum+gaze dataset: our subject mixes a shake, drinks it, washes his cup, plays chess and texts a friend.

Key Takeaways

- A very first study on the role of gaze in egocentric video summarization
- An efficient submodular summarization model for egocentric videos

Acknowledge to NSF RI 1116584, NSF CGV 1219016, NIH BD2K 1U54AI117924 and 1U54EB020404

jiaxu@cs.wisc.edu