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 ≜ Gaze
- Efficient algorithms ≜ 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
  \[ M(V \setminus S; S) = H(V \setminus S) - H(V \setminus S | S) \]
- Entropy
  \[ H(S) = \frac{1 + \log(2e)}{2} |S| + \frac{1}{2} \log(\det(L_S)) \]

Maximizing \( M(V \setminus S; S) \) is equivalent to maximizing
\[ M(S) = \frac{1}{2} \log(\det(L_{V \setminus S})) + \frac{1}{2} \log(\det(L_S)) \]
as \( |S| + |V \setminus S| = n \), and \( H(V) \) is constant.

Relation to Determinantal Point Process:
Positive semidefinite kernel matrix \( L \) indexed by elements of \( V \)
\[ L_{ij} = \frac{v_i^T v_j}{||v_i|| ||v_j||} \]
Diversity score for \( S \in V \)
\[ D(S) = \log(\det(L_S)) \]

Partition Matroid Constraint:
- High level supervision: timeline
- Partition the video into \( b \) disjoint blocks \( P_1, P_2, \ldots, P_b \)
- Compactness: allocation bound for each block
  \[ I = \{A : |A \cap P_m| \leq f_m, m = 1, 2, \ldots, b\} \]

**Submodular Summarization**

Main Model:
\[ \max_S F(S) = M(S) + \lambda I(S) \quad \text{s.t. } S \in I \]

Corollary: \( F(S) \) is submodular.

**Algorithm 1** Local Search for Constrained Submodular Maximization

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

Proposition: Greedy local search achieves a \( \frac{1}{2} \) 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.

**Systematic Evaluation**

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

<table>
<thead>
<tr>
<th>Method</th>
<th>uniform</th>
<th>kmeans</th>
<th>uniform (our subshots)</th>
<th>kmeans (our subshots)</th>
<th>ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>0.167</td>
<td>0.215 ± 0.016</td>
<td>0.526</td>
<td>0.475 ± 0.026</td>
<td>0.621</td>
</tr>
</tbody>
</table>

Table 2: Comparisons 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

Computer Vision and Pattern Recognition, 2015

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

jiuxu@cs.wisc.edu