Summary Transfer: Exemplar-based Subset Selection for Video Summarization

Ke Zhang*, Wei-Lun Chao*, Fei Sha, and Kristen Grauman*

1University of Southern California, 2University of California, Los Angeles, 3University of Texas at Austin

Introduction

Video summarization is indispensable: >300 hours of new Youtube video per min

Popular ways: key frame (shot selection)

Previous work:
- Unsupervised: hand-crafted criteria
- Supervised: (complex) parametric modeling

Motivation

- Similar videos ought to have similar compositional structures in their summaries (Wedding: bride entering, groom waiting...)
- Transfer summaries from human-annotated videos to new ones by selecting sequentially ordered frames w/ high visual similarity

Approach

Challenges of video summarization:
1) A structured prediction problem
2) Transfer summarization labels (selected vs. not selected) fails to consider relatedness of frames

Solution: Transfer the "underlying" summarization structure

Determinantal Point Process (DPP) for modeling the structure
DDPs define the probability of selecting a subset \( y \) from a N-tem ground set: given the similarity kernel \( \mathbf{L} \), diverse & representative subsets are highly probable

How to obtain the similarity (summarization) kernel \( \mathbf{L}_r \) for a human-annotated video \( r \)?

Summary Transfer constructing \( \mathbf{L} \) for the test video

| \( \begin{align*}
L_{r,y} &= \det(\mathbf{L}_r | y) \\
\delta(1 \in y, \ldots, 0) \\
L_r &= \alpha_r \\
0 \ldots \delta(N_r \in y)
\end{align*}\)

Learning: adjust parameters \( \mathbf{C}_r \) using MLE (leave-one-out on training videos)

Category-specific summary transfer: Videos from the same category have close high-level semantic cues

Solution: Learning for each category of videos a specific set of \( \mathbf{C}_r \)

\[ \begin{align*}
\text{Ours} &= \text{SegDPP (F = 62)} \\
\text{VSUMM} &= \text{Submodular (Gygli '15), and unsupervised methods}
\end{align*}\]

Experiments

Dataset: SumMe (50), OVP (50), Youtube (31), Kodak (18), and MED (160)

Evaluation: F-score, average or maximum over multiple human-created summaries

Feature: SIFT & Color histogram

Comparison: seqDPP [Gong '14], Submodular [Gygli '15], and unsupervised methods

<table>
<thead>
<tr>
<th>Setting</th>
<th>Kodak</th>
<th>OVP</th>
<th>Youtube</th>
<th>MED</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSUMM</td>
<td>69.5</td>
<td>70.3</td>
<td>59.9</td>
<td>28.9</td>
</tr>
<tr>
<td>seqDPP</td>
<td>79.9</td>
<td>77.7</td>
<td>60.0</td>
<td>28.9</td>
</tr>
<tr>
<td>Ours</td>
<td>82.3</td>
<td>76.5</td>
<td>61.8</td>
<td>30.7</td>
</tr>
</tbody>
</table>

Despite the variety of the datasets, we obtain state-of-the-art performance on most of them

Video category information helps summarization

Positive example:
- supervised learning helps identify representative contents
- non-parametric transfer leads to better kernel \( \mathbf{L}_r \), eliminating uninformative frames

Negative example:
- Fail to capture the relationship between frames within the test video

K.G. is partially supported by NSF IIS-1514118, Others are partially supported by USC Annenberg and Viterbi Graduate Fellowships, NSF IIS-1451412, 1513966, and CCF-1329144