Summary Transfer: Exemplar-based Subset Selection for Video Summarization

Ke Zhang*[1], Wei-Lun Chao*[1], Fei Sha[2], and Kristen Grauman[3]

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

Highlights

- ★ **Propose a non-parametric method for supervised video summarization**
- ★ **Transfer the summarization structure from human-annotated videos to new ones**
- ★ **Exploit side information (e.g., video categories) for semantically guided transfer**

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

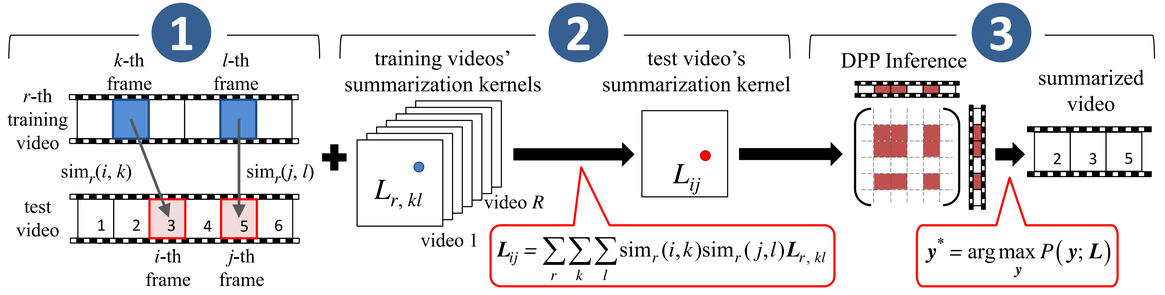
$$P(\boldsymbol{y};\boldsymbol{L}) = \frac{\det(\boldsymbol{L}\{\boldsymbol{y}\})}{\det(\boldsymbol{L}+\boldsymbol{I})}$$

DPPs define the probability of selecting a subset $\boldsymbol{y}$ from a N-item ground set: given the similarity kernel $\boldsymbol{L}$, diverse & representative subsets are highly probable

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

$$\boldsymbol{L}_r = \alpha_r \begin{bmatrix} \delta(1 \in \boldsymbol{y}_r) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta(N_r \in \boldsymbol{y}_r) \end{bmatrix}$$

Summary Transfer constructing $\boldsymbol{L}$ for the test video



Learning: adjust parameters α_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 α_r

Other details: sub-shot based summarization, sequential modeling, and complexity, etc.

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

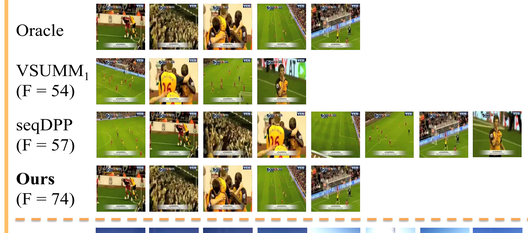
Setting	Kodak	OVP	YouTube	MED
VSUMM	69.5	70.3	59.9	28.9
seqDPP	78.9	**77.7**	60.8	-
Ours	**82.3**	76.5	**61.8**	**30.7**

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

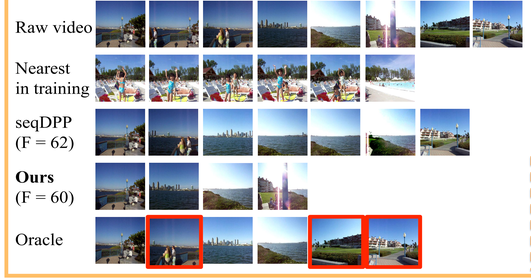
	VSUMM	SumMe	Submodular	Ours
SumMe	33.7	39.3	39.7	**40.9**

Setting	YouTube (2)	MED (10)	SumMe (2)
w/o cs	60.0	28.9	39.2
cs hard	**61.5**	30.4	**40.9**
cs soft	60.6	**30.7**	40.2

Video category information helps summarization
(cs stands for category-specific, # of categories in the brackets)



Positive example
- supervised learning helps identify representative contents
- non-parametric transfer leads to better kernel $\boldsymbol{L}$, 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 - 1139148