Diverse Sequential Subset Selection for Supervised Video Summarization

Boqing Gong†*, Wei-Lun Chao†*, Kristen Grauman‡, and Fei Sha†

†University of Southern California, ‡University of Texas at Austin

Highlight

- **Pose video summarization as a supervised learning problem for subset selection**
- **Propose sequential determinantal point process (seqDPP) as the underlying probabilistic model**
- **Evaluate on three video summarization tasks and obtain state-of-the-art performance**

Introduction

Video summarization: *pressing need*

- 100 hours of new Youtube video per min
- 422,000 CCTV cameras in London 24/7

Summaries by three users

Challenges

- Heterogeneous subjects/categories
- Various temporal changing rates
- Subjective, disparate, and noisy labels

Previous work

- Criteria: representativeness vs. diversity
- Largely unsupervised, frame clustering
- Require sophisticated handcrafting

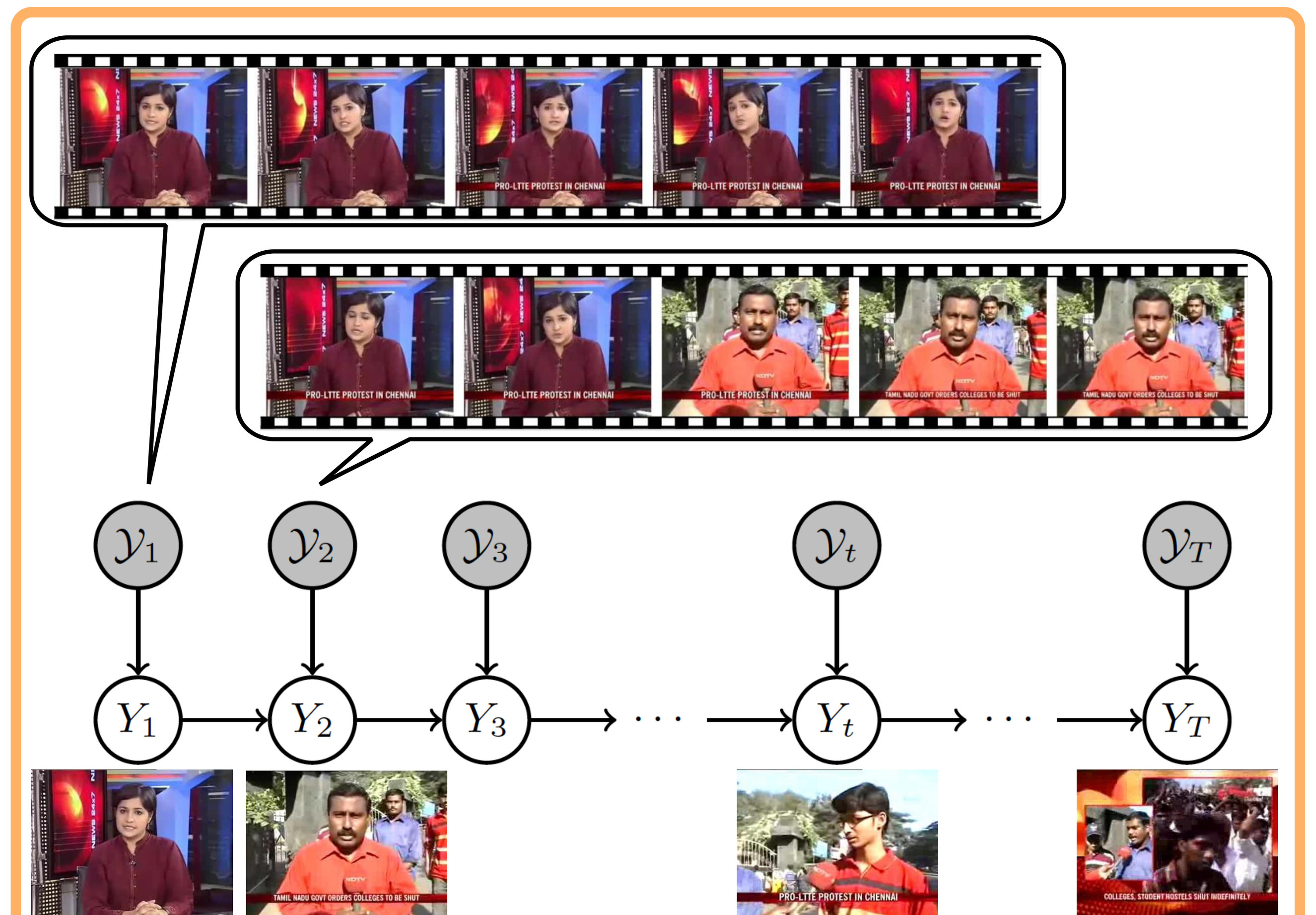
Our main idea

- Supervised learning from human supplied annotations
- Summarization as subset selection
- Modeling temporal cue & diversity

Approach

Sequential DPP (seqDPP)

1. Partition video into T disjoint segments
2. Introduce subset selection (of frames) variable Y_t for each segment
3. Condition Y_t on $Y_{t-1} = y_{t-1}$ by DPP



$$P(Y_t = \boldsymbol{y}_t | Y_{t-1} = \boldsymbol{y}_{t-1}) = \frac{\det \boldsymbol{\Omega}_{\boldsymbol{y}_{t-1} \cup \boldsymbol{y}_t}}{\det(\boldsymbol{\Omega}_t + \boldsymbol{I}_t)}$$

$\boldsymbol{\Omega}_t$: kernel over ground set $\mathcal{Y}_t \cup \boldsymbol{y}_{t-1}$

Parameterization of DPP kernel

- Linear embedding (**L**): $\boldsymbol{f}_i^T \boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{f}_j$
- Neural networks (**NN**)

Inference

$$\boldsymbol{y}_1^* = \arg\max_{\boldsymbol{y} \in \mathcal{Y}_1} P(Y_1 = \boldsymbol{y})$$
$$\boldsymbol{y}_2^* = \arg\max_{\boldsymbol{y} \in \mathcal{Y}_2} P(Y_2 = \boldsymbol{y} | Y_1 = \boldsymbol{y}_1^*)$$

Learning via MLE

- through gradient descent

In contrast, bag DPPs:

Model permutable items (no temporal info)
Often use quality-diversity kernel (limited)
Inference NP hard

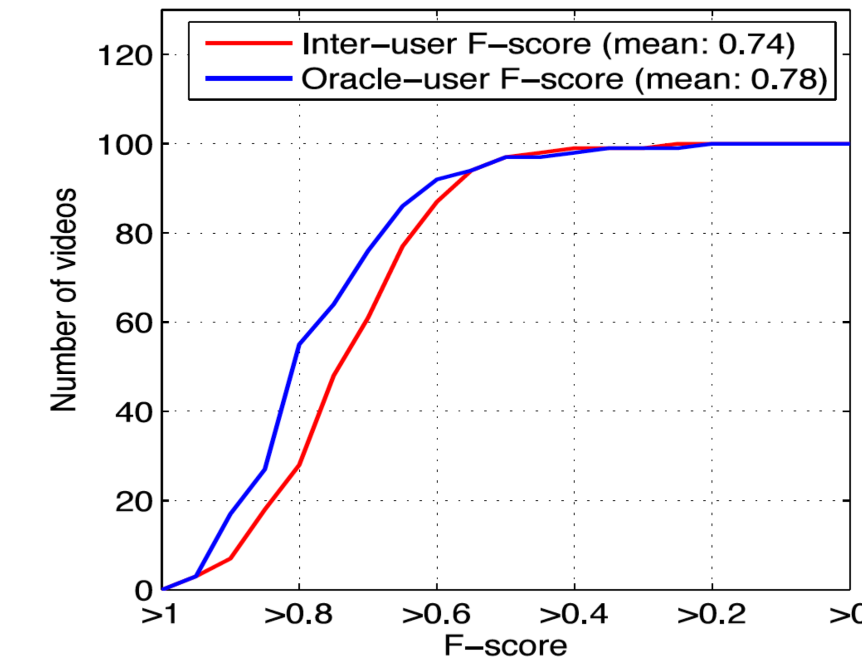
Generating target summaries

User study on inter-annotator agreement

- Data: 100 videos from Open Video Project and Youtube
- Annotation: 5 user summaries per video
- **Observation:** high inter-annotator agreement

Generate target summaries by greedy search

$$\boldsymbol{y}^* \leftarrow \boldsymbol{y}^* \cup \arg\max_i \sum_u F_{\boldsymbol{y}^* \cup i, \boldsymbol{y}_u}$$



Experiments

Setup

- Data: OVP (50), Youtube (39), Kodak (18)
- Feature: Fisher vector, saliency, context
- Evaluation: **P**recision, **R**ecall, **F**-score
- Comparison: bag DPP and previous (unsupervised) DT, STIMO, VSUMM

Results on Youtube and Kodak

	VSUMM2			Ours (L)			Ours (NN)		
	F	P	R	F	P	R	F	P	R
Youtube	55.7	59.7	58.7	57.8	54.2	69.8	**60.3**	59.4	64.9
Kodak	68.9	75.7	80.6	75.3	77.8	80.4	**78.9**	81.9	81.1

Results on OVP

	F	P	R
DT	57.6	67.7	53.2
STIMO	63.4	60.3	72.2
VSUMM1	70.3	70.6	75.8
VSUMM2	68.2	73.1	69.1
bag DPP	70.8±0.3	71.5±0.4	74.5±0.3
Ours + Q/D	68.5±0.3	66.9±0.4	75.8±0.5
Ours (L)	75.5±0.4	**77.5**±0.5	78.4±0.5
Ours (NN)	**77.7**±0.4	75.0±0.5	**87.2**±0.3

Target summary

Time-varying visual diversity

seqDPP (ours)
(F=70, P=60, R=88)

VSUMM1
(F=59, P=65, R=55)

Coping with time-varying diversity: seqDPP better than VSUMM

[1] S. Avila, A. Lopes, A. Luz Jr, A. Araujo. "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method". Pattern Recognition Letters, 32(1):56–68, 2011.
[2] A. Kulesza and B. Taskar. "Determinantal point processes for machine learning". Foundations and Trends® in Machine Learning, 5(2-3):123–286, 2012.