Viewpoint-aware Video Summarization

Atsushi Kanehira[1], Luc Van Gool[3,4], Yoshitaka Ushiku[1], Tatsuya Harada[1,2]

[1] The University Tokyo, [2] RIKEN, [3] ETH Zurich, [4] KU Leuven

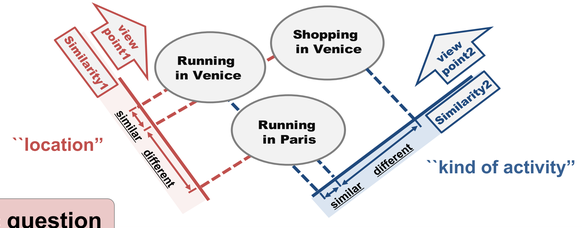
CVPR 2018 SALT LAKE CITY • JUNE 18-22

Introduction:

- Viewpoint changes how the video looks.
- This work treats *viewpoint* as …

✓ *A specific aspect of a video*
- e.g., kind of activity, location, …

✓ Different *viewpoint* has different semantic similarity.



Basic question

Can we estimate *viewpoint* from given *similarity*?

What is used as *similarity*?

▸ Pre-determined video groups (easy to obtain, such as by user preference)

How to show output ?

▸ Summarizing video (easy to see, contain much information)

Goal

Generating video summary that explains ``why videos are divided in the way they are?''

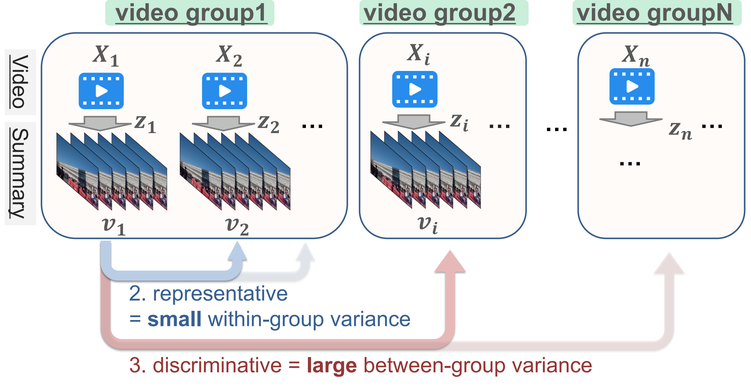
1. Why the video are in that group?
   ▸ summary of target and other videos in the same group
2. Why the video are not in other groups?
   ▸ summary other videos in the different groups

Video Co-summarization

Method:

Video co-summarization reflecting *similarity*

Requirements:
1. **diverse**
2. **representative** to videos in the same group
3. **discriminative** against videos in the different groups



- Visual feature of summaries with shot-indicator $\boldsymbol{z} = \{0,1\}^N, ||\boldsymbol{z}||_0 = s$

$$\underbrace{v = Xz}_{\text{summary-feature}} \qquad \underbrace{X = [x_1, x_2, \ldots, x_N]^T}_{\text{shot-features}}$$

- Minimizing the combination of three variances inspired by Fisher Criteria

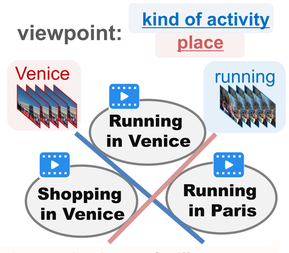
$$\underbrace{\lambda_1 Tr(S^V)}_{\text{inner-summery}} - \underbrace{\lambda_2 Tr(S^W)}_{\text{within-group}} + \underbrace{\lambda_3 Tr(S^B)}_{\text{between-group}}$$

diverse ↑ representative ↑ discriminative

- Continuously relaxed problem is DC (difference of convex) programming, which can be optimized with CCCP (concave-convex procedure)

Dataset:

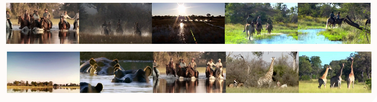
- Collect a set of videos using 3 keywords.
  ▷ Videos can be divided with 2 explicit viewpoints.
- Target videos have 2 different annotations.
  ▷ Each corresponds to one viewpoint.
- Compare predicted summary with ground-truth.
- 10 types of groups, 150 videos, 4 hours in total.



Examples of dataset

``riding helicopter in NewYork'' with NewYork(↑), helicopter(↓)

``riding horse in the safari'' with riding horse(↑), safari(↓)



Experiment:

Automatic evaluation (top5 mean Average Precision)

only representative | representative & discriminative

method	SMRS	CK	CS	MBF	CVS	WSVS	WSVS(large)	ours	ours (feature learning)
mAP	0.322	0.294	0.312	0.333	0.353	0.322	0.319	**0.379**	**0.385**

User study

summary quality

method	MBF	CVS	ours
accuracy	0.47	0.60	**0.76**

explanation quality

method	MBF	CVS	ours
score	1.07	1.22	**1.32**

example results

``recing in the desert'' with racing(↑), desert(↓)

``riding helicopter in NewYork'' with NewYork(↑), helicopter(↓)