A Topological Data Analysis Approach to Video Summarization

September 24, 2019

Chuan-Shen Hu
Department of Mathematics
National Taiwan Normal University

Mei-Chen Yeh
Department of Computer Science
and Information Engineering
National Taiwan Normal University

Overview

Problem Statement

Definition

Video Summarization is a technique that provides a condensed and interesting storyline of a video.

Why is it challenging?

- Unlimited contents;
- Subjective;
- "Relation" is abstract for computing.

Contributions

Contributions

For video analysis :

- Summarize videos by using TDA techniques.

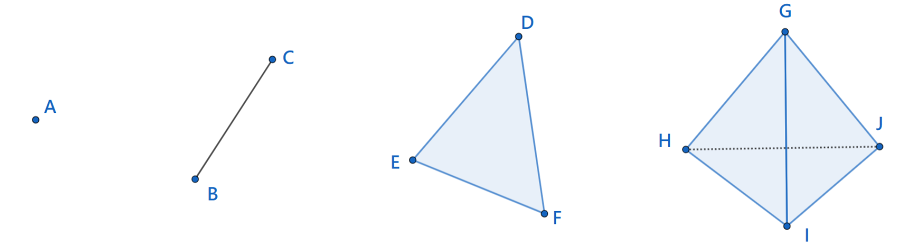
For TDA :

- Approach for sequential data;
- Not only Betti numbers but also locations of **1-homologies**.

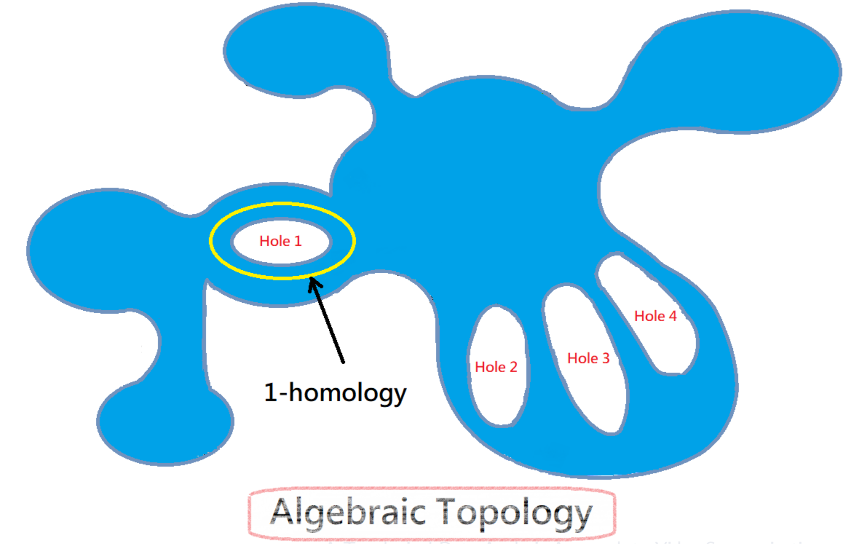
Why Simplicial Complexes?

Simplicial complexes

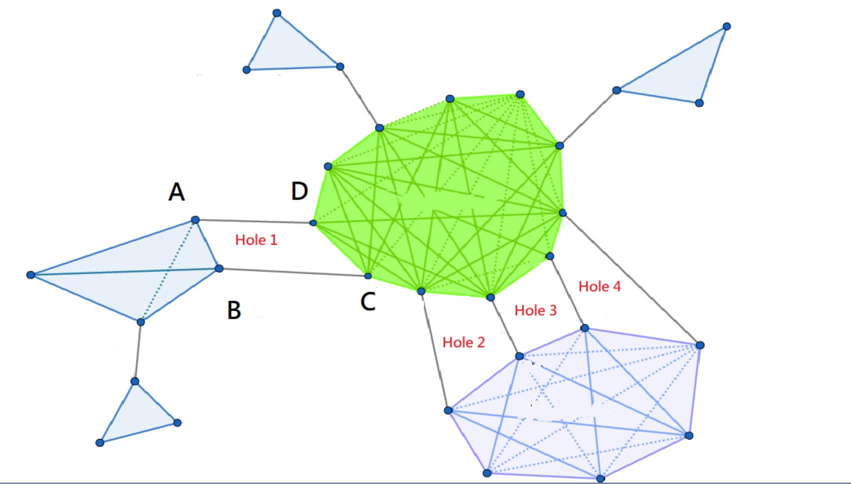
Simplicial complexes were widely used for approximating objects by using edges, triangles etc. (*e.g.* meshes in computer graphics)



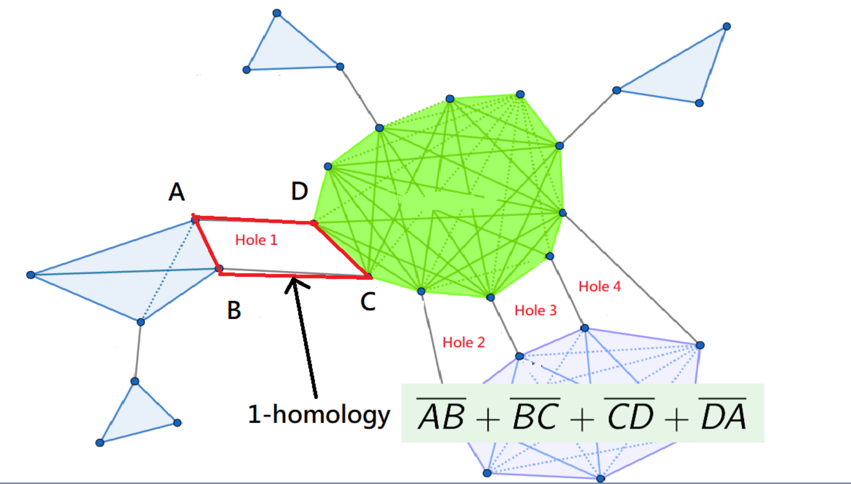
Why Simplicial Complexes?



Why Simplicial Complexes?



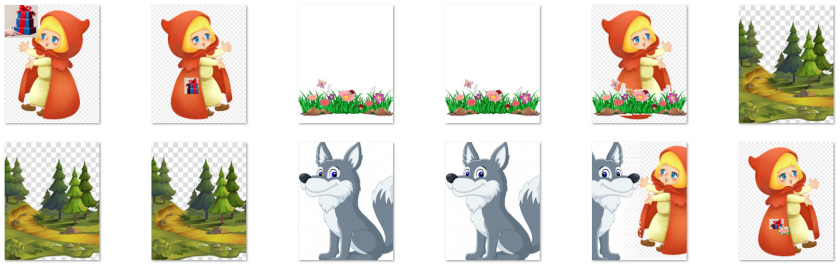
Why Simplicial Complexes?



$$\overline{AB} + \overline{BC} + \overline{CD} + \overline{DA}$$

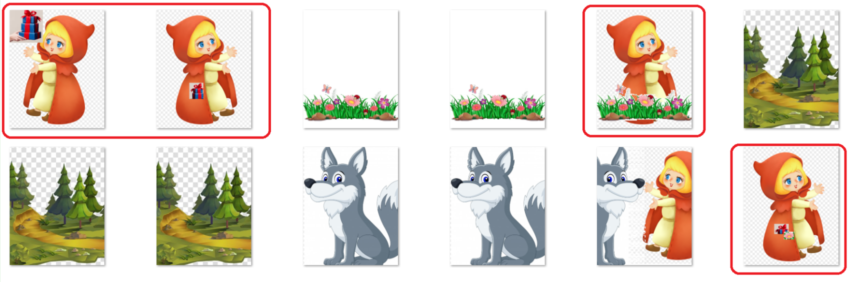
Motivation

Meaning of holes in videos



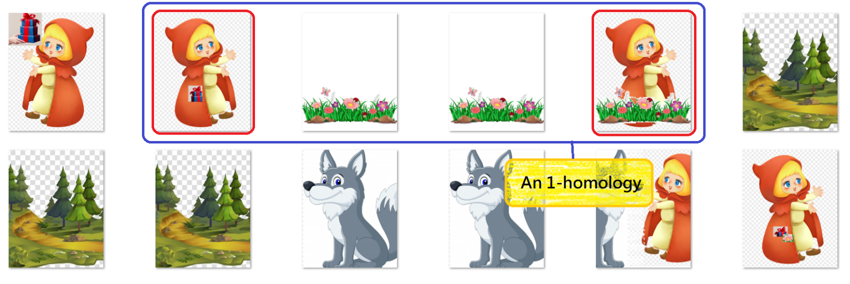
Motivation

Meaning of holes in videos (anchor frames)



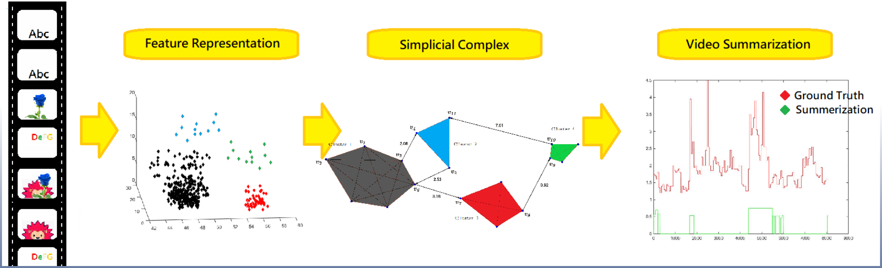
Motivation

Meaning of holes in videos (events as 1-homologies)



Pipeline

- **Step 1** : Sub-sampling of frames and extract CNN features;
- **Step 2** : **Simplicial complex** representation;
- **Step 3** : Summarize videos via **1-homologies**.



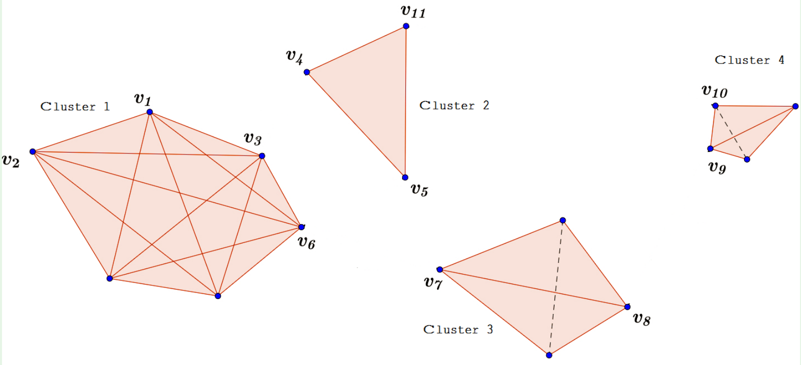
Step 1 : Frame Representation

Sub-sampling & Features

- Sample one frame per 400 milliseconds;
- Each sampled frame is summarized into a fixed length feature vector (fc7) using **ImageNet-trained AlexNet** in $\mathbb{R}^{4096}$.

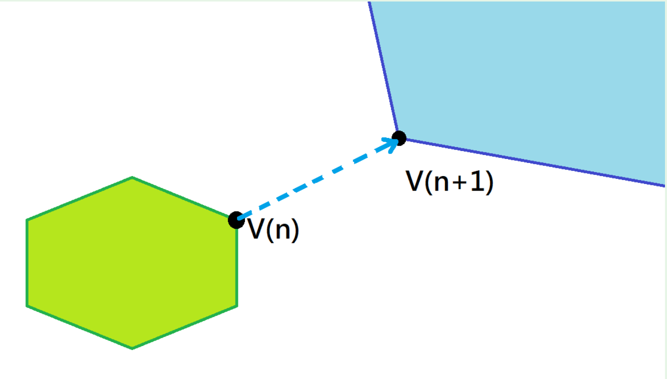
Step 2 : Simplicial Complex

Step 2-1 : (Topological) Vertex clustering



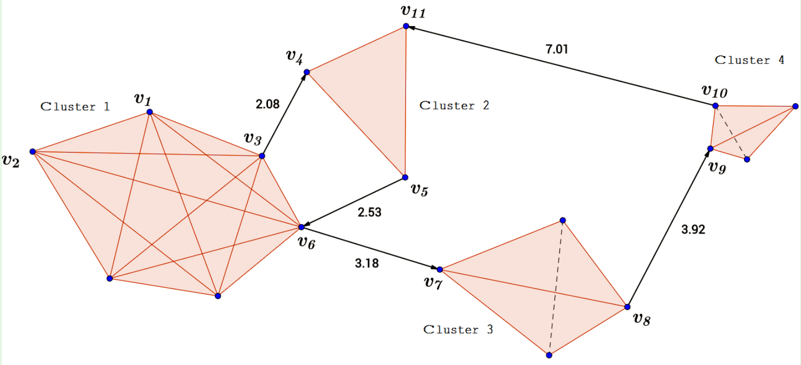
Step 2 : Simplicial Complex

Step 2-2 : Edge augmentation



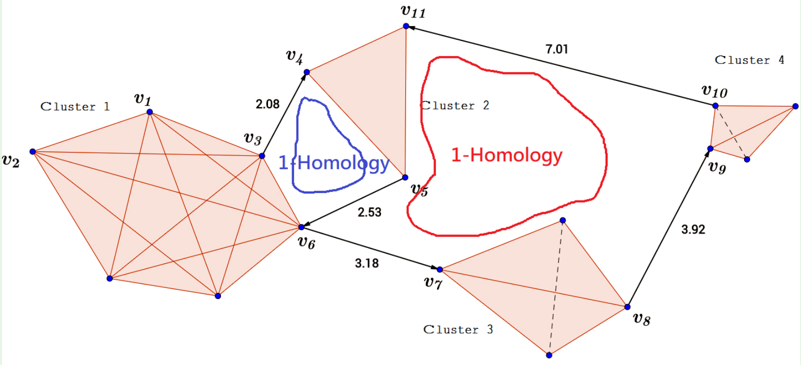
Step 2 : Simplicial Complex

Step 2-2 : Edge augmentation



Step 3 : Clustering Score & Summarization

Step 3-1 : 1-Homology detection (javaplex)



Step 3 : Clustering Score & Summarization

Definition (Anchor clusters)

After augmentation, define **repeat number** of a cluster σ by

$$r(\sigma) = \#\{v : v \text{ appears in a } 1-\text{homology}\}.$$

Cluster who have maximal repeat number are called an **anchor clusters**.

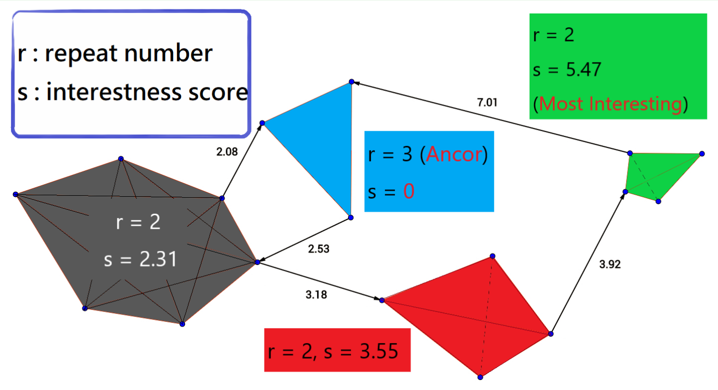
Definition (Cluster scores)

Given anchor clusters $\sigma_{i_1}, \ldots, \sigma_{i_k}$, define interestness score of σ by

$$s(\sigma) = \frac{1}{k} \cdot \left(\sum_{j=1}^{k} \mathrm{dist}(\sigma, \sigma_{i_j}) \right) \cdot 0.95^{r(\sigma)}$$

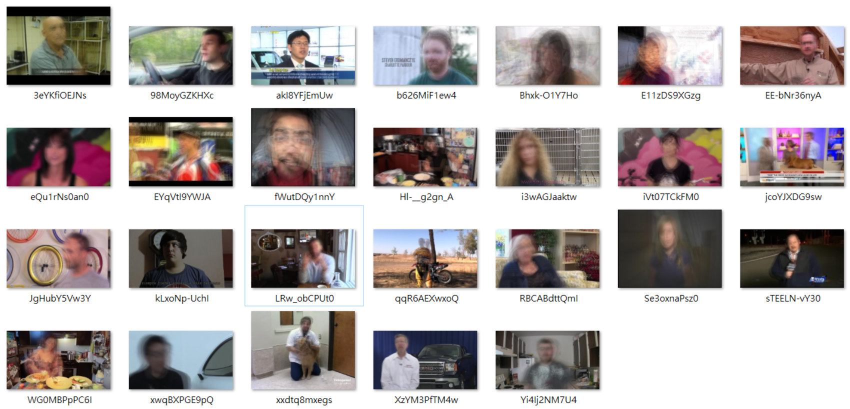
Step 3 : Clustering Score & Summarization

Step 3-2 : Cluster scoring



Visualization : Anchor Frames

Hosts, main interviewees or reporters



Visualization ： 1-homologies

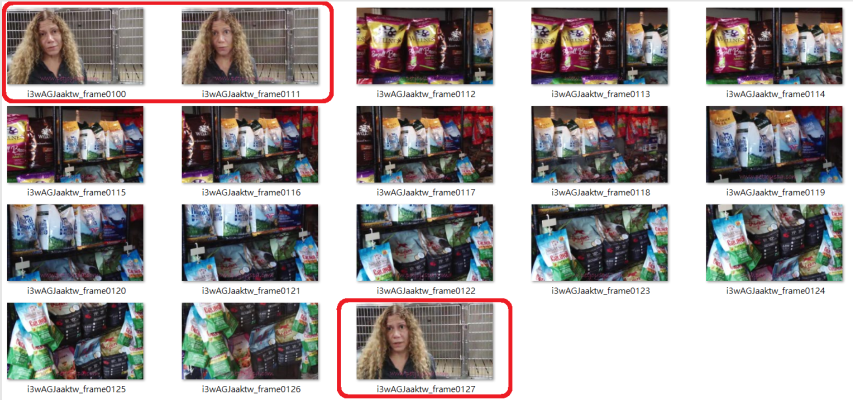
Examples ： Events, actions or motions

Visualization : 1-homologies

Examples : Events, actions or motions

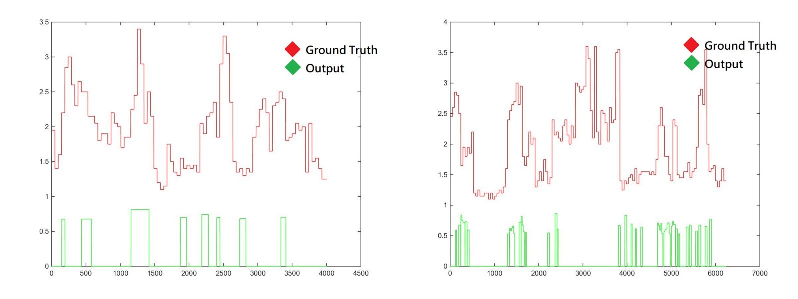
Visualization : 1-homologies

Examples : Events, actions or motions



Step 3 : Clustering Score & Summarization

Step 3-3 : Segmentation



Dataset

TVSum50 benchmark

- Provided by Y. Song *et al.* (2015)
- 50 videos collected from YouTube
- 10 categories (e.g. Dog Show (DS), Vehicle Tire (VT))
- Each category contains 5 videos
- In average, 20 summaries by people per video

Evaluation Metric

Evaluation

- Input datatype : segments and their interesting scores
- Metric : pairwise F_β-measure, defined by

$$\widetilde{F}_\beta = \frac{1}{N} \cdot \sum_{i=1}^{n} \frac{(1+\beta^2) \cdot p_i \cdot r_i}{(\beta^2 \cdot p_i) + r_i}$$

- n : gold-standard summaries
- p_i : precision, r_i : recall, $\beta = 1$
- Code : evaluation toolkit provided by TVSum50

Result

Result

non-deep learning	
Song [5]	0.50
deep learning	
LSTM [6]	0.55
Supervised tessellation [7]	0.64
Unsupervised tessellation [7]	0.63
Sup. adversarial LSTM [8]	0.56
Unsup. adversarial LSTM [8]	0.52
non-learning	
LiveLight [4]	0.46
Ours	**0.66**

Conclusion & Future Directions

Conclusion

- Videos may contain topological structures and those features can benefit video summarization task.
- TDA may provide additional information in ML models.

Future directions

- We currently try to improve the scoring method, and so far get promising result (a lifting from 0.66 to $\widetilde{F}_1$-measure **0.73**).
- Determine anchor frames by distribution of 1-homologies;
- Combine these features to machine learning models.

Selected references

Selected references

1 Ngo, Ma, and Zhang, IEEE TCSVT, 2005.

2 Gygli, Grabner, Riemenschneider and Gool, ECCV, 2014.

3 Panda, Kuanar, and Chowdhury, ICPR, 2014.

4 Zhao and Xing, CVPR, 2014.

5 Song, Vallmitjana, Stent and Jaimes, CVPR, 2015.

6 Zhang, Chao, Sha and Grauman, ECCV, 2016.

7 Kaufman, Levi, Hassner and Wolf, in arXiv:1612.06950, 2017.

8 Mahasseni, Lam and Todorovic, CVPR, 2017.

9 Zhu, IJCAI, 2013.

Thank you for your attention!

Vertex Clustering Algorithm (Modification for Witness Complex)

Algorithm

- (Input) Given feature vectors $v_1, v_2, ..., v_m$ an $R > 0$;
- 1. Exhibit cluster $\{v_1\}$.
- 2-0. Given $\sigma = \{v_1, ..., v_k\}$ and v;
- 2-1. If more than half of $\{v_1, ..., v_k\}$ satisfies $\|v_i - v\| \leq R$ $(*)$, then set $\sigma \leftarrow \sigma \cup \{v\}$.
- 3. Given v_k, $k \in \{2, ..., m\}$ and clusters $\sigma_1, \sigma_2, ...\sigma_s$. If no cluster satisfies $(*)$, construct new cluster $\sigma_{s+1} = \{v_k\}$.
- (Output) Collection $\widetilde{\mathcal{K}} = \{\sigma_1, ..., \sigma_l\}$ of disjoint clusters.

Vertex Clustering Algorithm

Remark

- Radius R can be determined automatically by WCC (Witness Complex Construction) algorithm;
- By viewing each $\sigma \in \widetilde{\mathcal{K}}$ as a sorted array of integers, $\sigma[0]$ denotes the moment that σ was born. This index also denotes a moment of a new scene occurs;
- Each cluster σ was viewed as a high dimensional simplex of dimension $|\sigma| - 1$.

Segmentation Algorithm

Segmentation

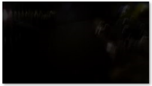
- For each cluster σ in $\widetilde{\mathcal{K}}$, choose sample frame $\sigma[0]$ as the representative of the cluster;
- extend it bidirectionally (in time) with $\lfloor \mathcal{F}/F \rfloor$ frames;
- $\mathcal{F} =$ number of total frames;
- $F =$ number of sampled frames.

Visualization : Anchor Frames

Opening/Closing credits, TV-show logos


gzDbaEs1Rlg


oDXZc0tZe04


PJrm840pAUI

VuWGsYPqAX8


XkqCExn6_Us

(Failure cases) Important events or objects


0tmA_C6XwfM


4wU_LUjG5lc


esJrBWj2d8


J0nA4VgnoCo


WxtbjNsCQ8A