Temporal- and Spatial-Driven Video Summarization Using Optimum-Path Forest

Guilherme B. Martins, João Paulo Papa, Jurandy G. Almeida

UNESP - São Paulo State University
Bauru - SP, Brazil
papa@fc.unesp.br

October 7, 2016

Talk Outline

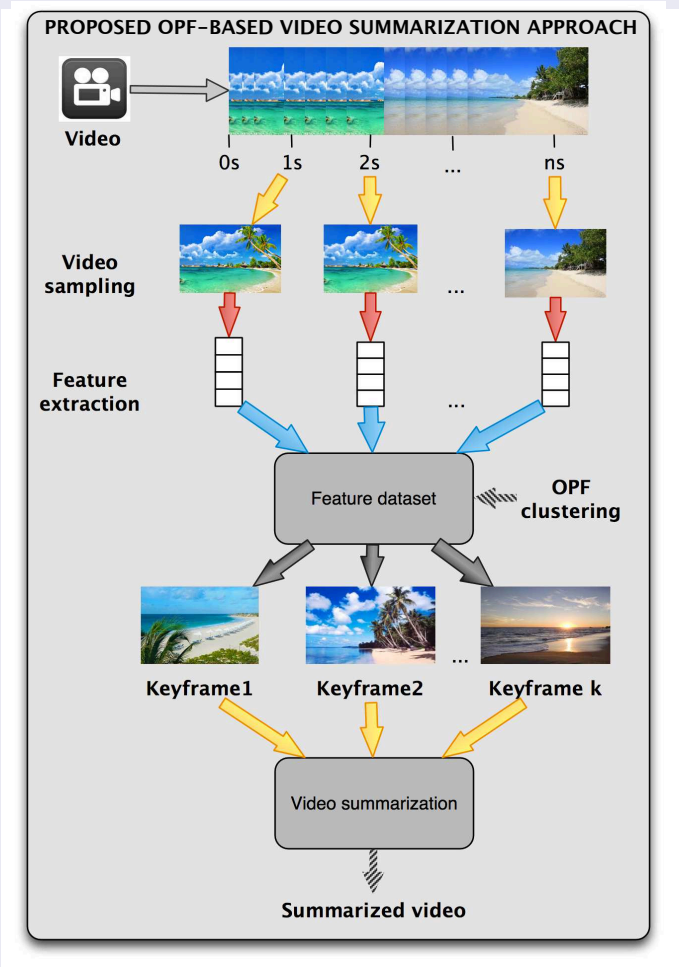
Introduction

Main concepts

- What is the meaning of video summarization?
- What is its importance? It can speed up media classification, retrieval and recommendation.
- It can be seen as a clustering task.

Introduction

Video Summarization at a glance



Introduction

Related Research

- Deep learning for both feature (CNNs/DBNs) and event learning (LSTMs).
- Different clustering-based algorithms have been used to learn the **key-frames** (mainly ANNs, GMMs and SVMs).
- Previous works[a b] have used the Optimum-Path Forest (OPF) clustering for video summarization. However, they did not account for the **temporal** information.

---

[a] G. B. Martins, L. C. S. Afonso, D. Osaku, Jurandy Almeida, J. P. Papa, "Static Video Summarization through Optimum-Path Forest Clustering", CIARP'2014

[b] C. Castelo-Fernández, G. Calderón-Ruiz, "Automatic Video Summarization Using the Optimum-Path Forest Unsupervised Classifier", CIARP'2015

Introduction

Main Goal

- To propose a new OPF-based video summarization approach that can take into account the temporal information for video summarization.
- The proposed work has been compared against some state-of-the-art video summarization approaches, obtaining promising results.

Talk Outline

Optimum-Path Forest

Theoretical Background

- Graph-based approach for pattern classification (unsupervised, semi-supervised and supervised learning).
- The whole idea is: dataset samples $\rightarrow$ feature vector $\rightarrow$ graph nodes $\rightarrow$ **reward-based competition process** $\rightarrow$ partitioned graph (clusters or groups of labeled samples).
- Depending on the configuration you want, a different OPF classifier you will have. We usually say OPF is a **framework** instead of a single classifier.

Optimum-Path Forest

Theoretical Background

- How does the competition process work? We have to follow three main steps:
  - Adjacency relation (k-nn or complete graph)
  - Prototype estimation (MST, density)
  - Path-cost function (f_{sum}, f_{max}, f_{min})
- The good thing is: you can design your **own** classifier.

Optimum-Path Forest

OPF Clustering

- We have the following configuration:
  - Adjacency relation: k-nn graph
  - Prototype estimation: as we do not have labels, prototypes are located at the regions with the highest density.
  - Path-cost function: f_{min}, which is the minimum value between the cost of a given sample and the density of the another one.

Optimum-Path Forest

OPF-based Video Summarization

- In our previous approach, the key-frames were encoded by the prototype samples.
- After extracting features from frames, they are mapped to a graph and the OPF algorithm takes place.
- Since each prototype is the **root** of its cluster (optimum-path tree), it is usually placed at the center of the group, thus becoming a nice candidate to be a key-frame.
- However, we did not consider **temporal** information (up to date).

Talk Outline

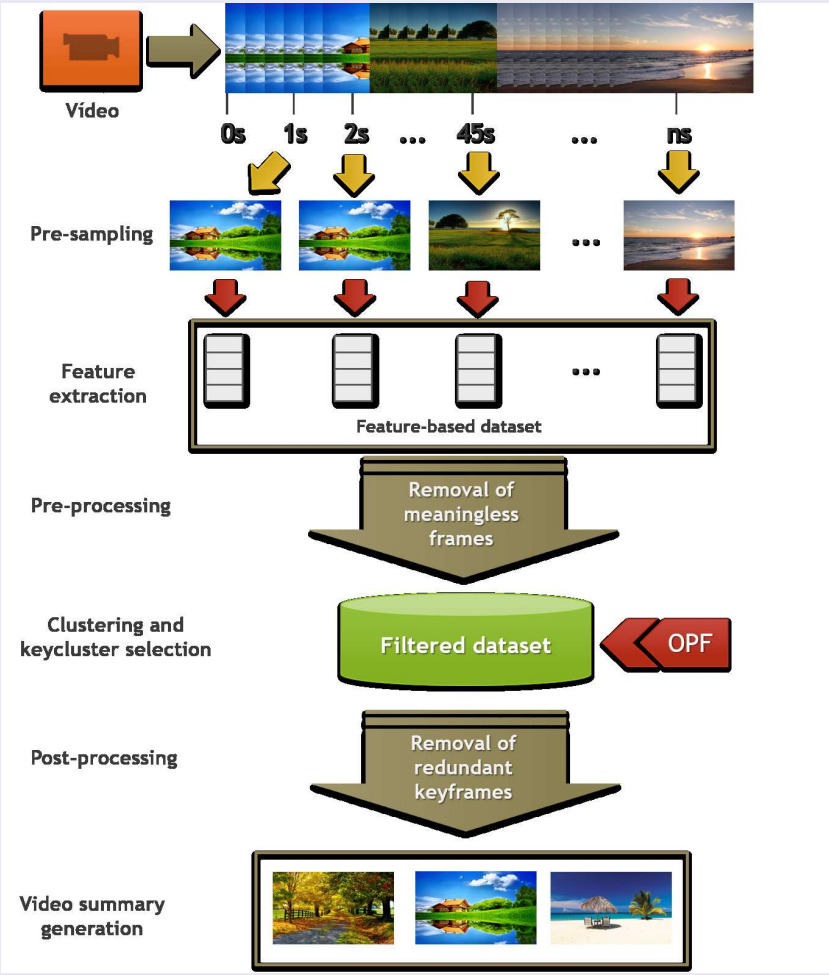
Video Summarization

Temporal Encoding

- The main question is: how can we encode temporal information with non-temporal features? That is an interesting challenge.
- We divided the whole problem in five steps:
  - sampling process.
  - pre-processing.
  - feature extraction.
  - clustering.
  - filtering.

Video Summarization

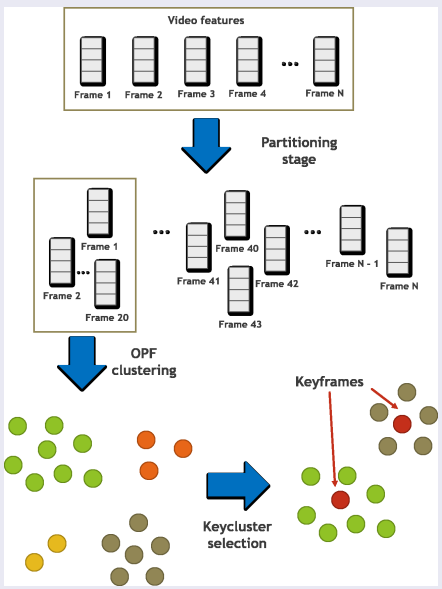
Proposed Approach



Video Summarization

Proposed Approach

- The most important contribution of this paper concerns the fourth step, which is the clustering process. Going back to the question: how to encode temporal information without temporal features?



Proposed Approach

Let T_{ij} be the **temporal term** between frames p_i and p_j given by:

$$T_{ij} = |p_i - p_j|, \tag{1}$$

where p_k means the normalized position (ratio between the frame number and the total number of frames) of frame k. Also, we define the **spatial term** S_{ij}, given by:

$$S_{ij} = \frac{d(i,j)}{d_{max}}, \tag{2}$$

where $d(i,j)$ stands for the Euclidean distance between frames i and j, and d_{max} denotes the maximum distance among any two frames in the dataset.

Proposed Approach

The main ideia is to compose a hybrid distance function D_{ij} that considers both **spatial** and **temporal** information:

$$D_{ij} = S_{ij} + \alpha T_{ij}, \tag{3}$$

where α weights the amount of temporal information considered during the distance computation. Distance D_{ij} will be used to weight the arcs between nodes i and j during OPF clustering process.

Talk Outline

Datasets

We used two well-known datasets:

- Open Video.
- YouTube.

How to choose the k-nn adjacency?

- OPF has the k_{max} parameter, which basically controls the maximum size of the neighborhood.
- We tried different percentages of the subset sizes (20%, 25% and 50%), and for each one we evaluated $k_{max} \in [5, 50]$ with steps of 5.
- Finally, we selected the subset size and k_{max} that maximized the F-measure (25% and $k_{max} = 5$).

Compared techniques

- OPF[a].
- OV[b].
- DT[c].
- STIMO[d].
- VSUMM[e].
- VISON[f].

---

[a] Martins et al., CIARP'14
[b] DeMenthon et al., ICM'98
[c] Mundur et al., IJDL'06
[d] Furini et al., MTA'10
[e] Avila et al., PRL'11
[f] Almeida et al., PRL'12

Frame Description

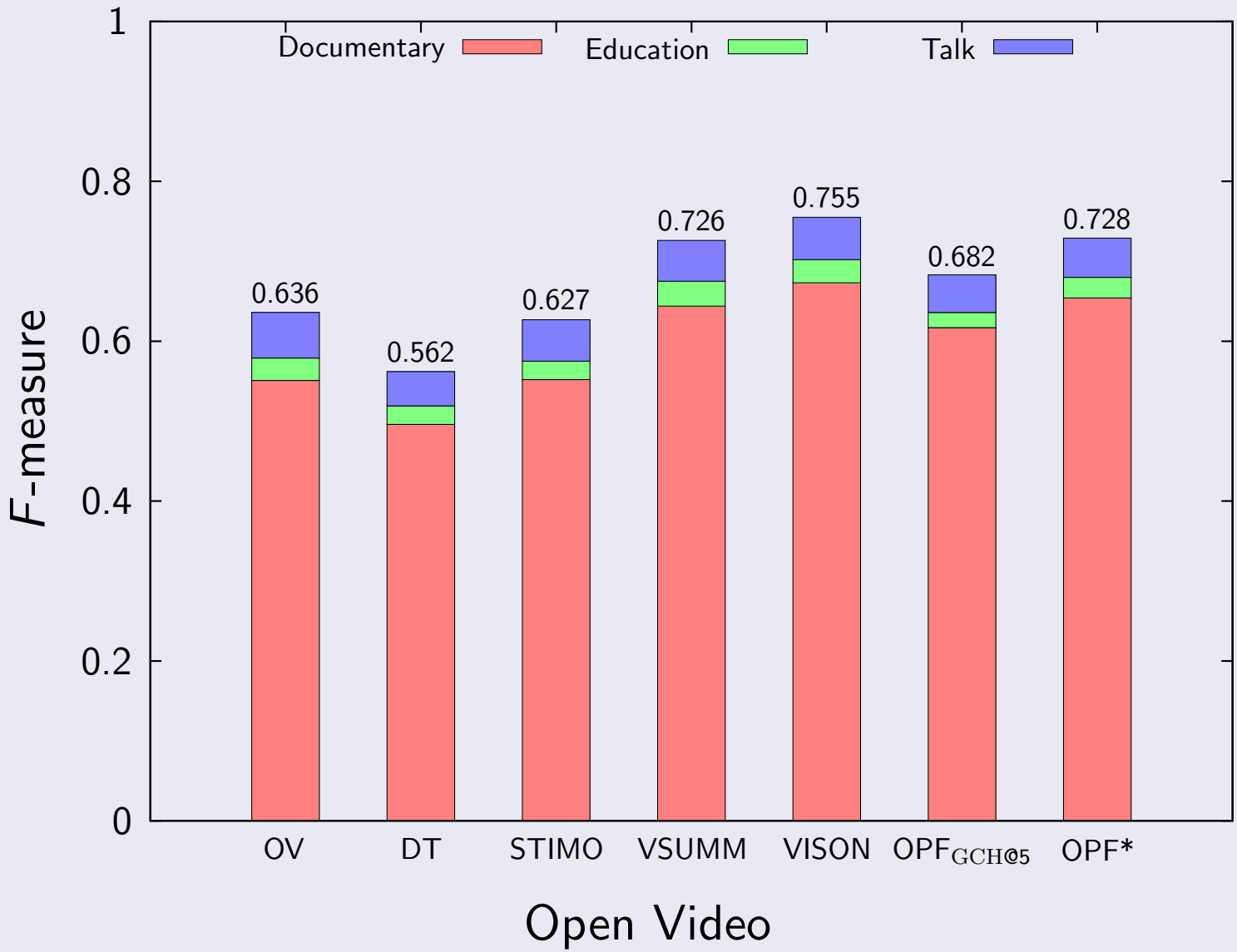
In this work, we considered two descriptors to encode color information:

- Global Color Histogram (GCH).
- Color Coherent Vector (CCV).

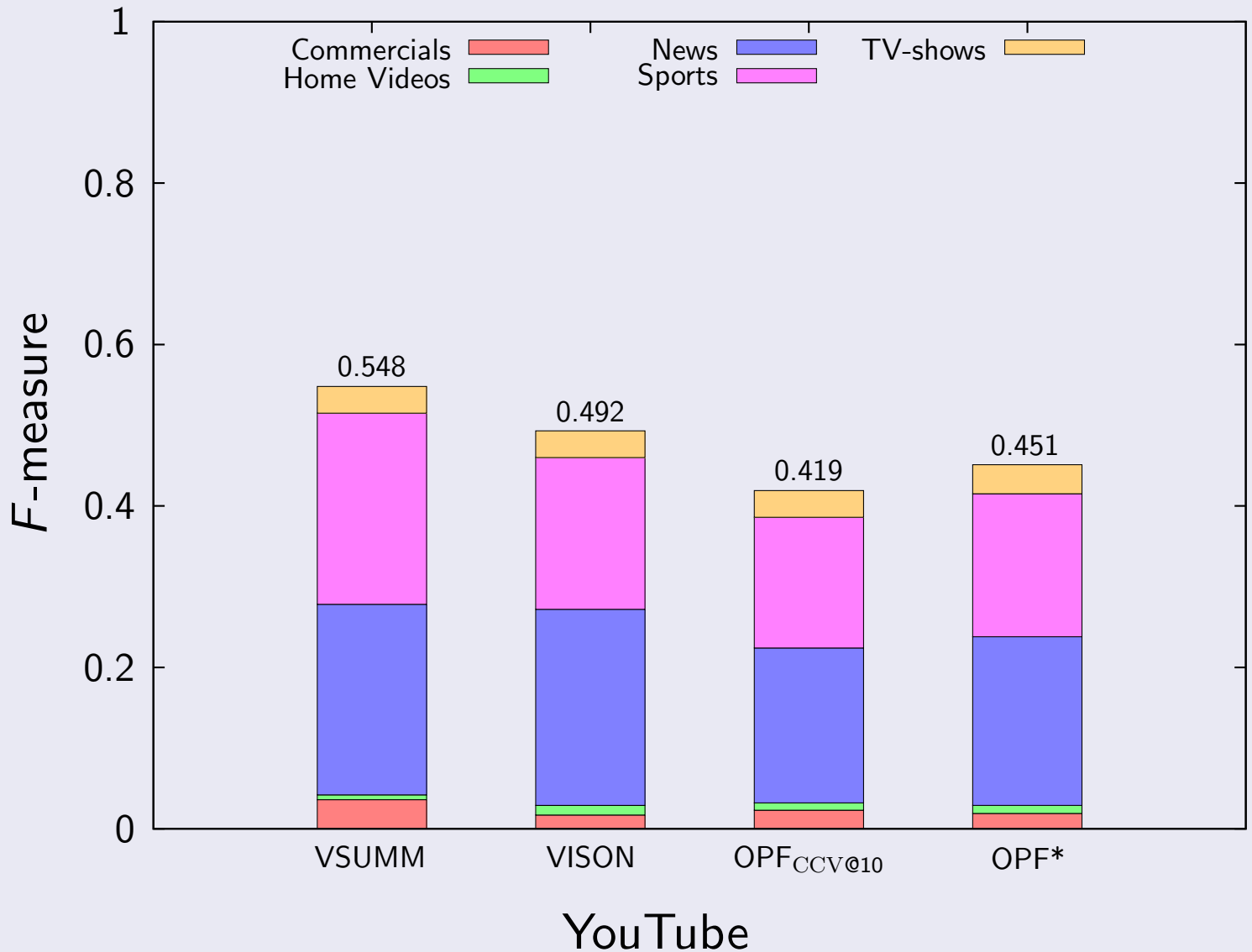
Fine-tuning α

We observed the proposed OPF (OPF*) seems to work better with smaller subsets, since larger ones do not favor the temporal information. Additionally, $\alpha = 0.86$ worked well for both datasets. In our experiments, we observed that small values concerning α did not contribute a lot for the final results.

Results


Documentary
Education
Talk
F-measure
1
0.8
0.6
0.4
0.2
0
0.636
0.562
0.627
0.726
0.755
0.682
0.728
OV
DT
STIMO
VSUMM
VISON
OPF$_{GCH@5}$
OPF*
Open Video

Results


Commercials
Home Videos
News
Sports
TV-shows
F-measure
1
0.8
0.6
0.4
0.2
0
0.548
0.492
0.419
0.451
VSUMM
VISON
OPF$_{CCV@10}$
OPF*
YouTube

Talk Outline

Final Remarks

- A new approach for video summarization has been proposed.
- It now consisders **temporal** information.
- It has outperformed our previous approach.
- Future work: class-specific OPF.