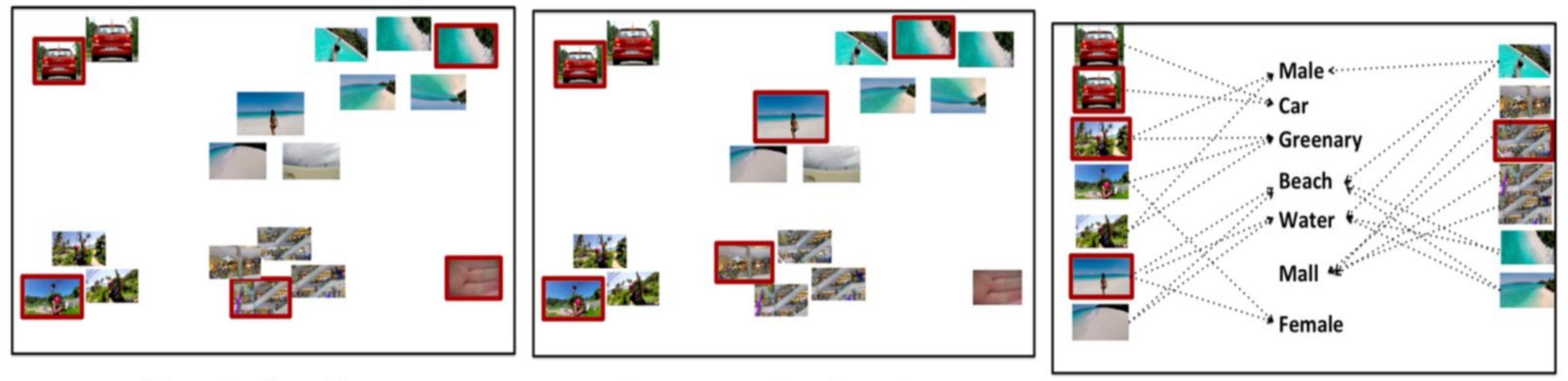
Demystifying Multi-Faceted Video Summarization: Tradeoff Between Diversity, Representation, Coverage and Importance

Vishal Kaushal, Rishabh Iyer, Khoshrav Doctor, Anurag Sahoo, Pratik Dubal, Suraj Kothawade, Rohan Mahadev, Kunal Dargan, Ganesh Ramakrishnan

IEEE 2019 Winter Conference on Applications of Computer Vision

Motivation

- Video Summarization attempts to provide a highlight of the most critical and important events in a video.
- Different summarization models capture different aspects of a video. Types of characteristics captured by summarization models:
  - Diversity
  - Representation
  - Importance
  - Coverage
- What comprises of the key aspects of a video, depends largely on the domain.
- Different domains, require different summarization models.
- Which summarization model to use for a domain?



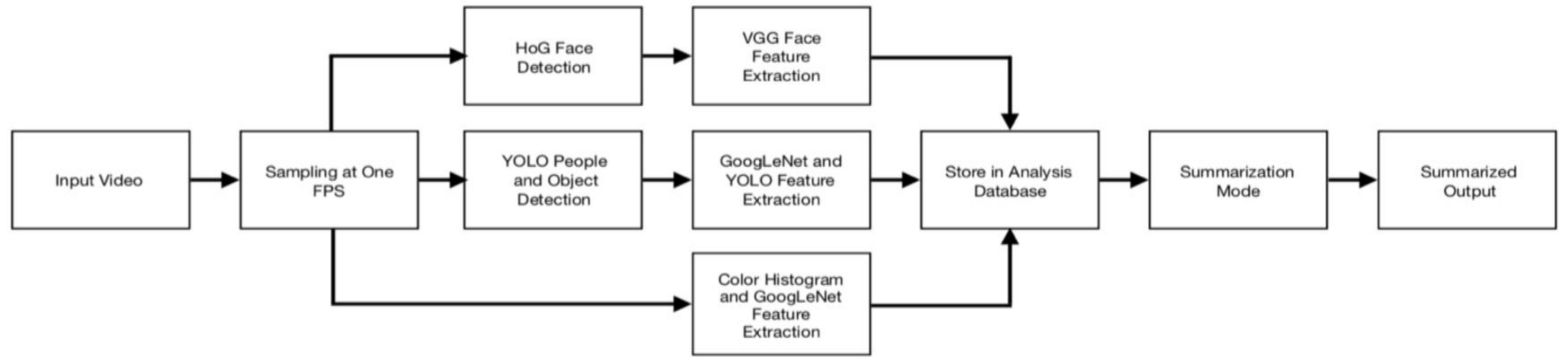
Diversity Functions　　Representation Functions　　Coverage Functions

Our Contributions

- Study the role and characteristics of various summarization models.
- Investigate several diversity, coverage and representation models, and demonstrate how different models are applicable in different kinds of summarization tasks.
- Empirically and quantitatively prove the behaviour of the functions on various videos across domains.
- Discuss computational scalability of the optimization algorithms and the usage of computational tricks, such as lazy evaluations and memoization.
- Explore various video summarization variants and study the details on how to design a video summarization system.

Submodular Summarization Framework

- Our framework studies the following variants of Video Summarization:
  - Extractive Summarization
  - Query Summarization
  - Entity Summarization
- In all variants, we use Convolutional Neural Networks as feature extractors, before performing submodular optimization to generate the summary.



- The following table illustrates the various submodular functions studied by us, as a part of our framework.

Name	$f(X)$	$p_f(X)$	T_f^o	T_f^p
Facility Location	$\sum_{i \in V} \max_{k \in X} s_{ik}$	$[\max_{k \in X} s_{ik}, i \in V]$	$O(n^2)$	$O(n)$
Saturated Coverage	$\sum_{i \in V} \min\{\sum_{j \in X} s_{ij}, \alpha_i\}$	$[\sum_{j \in X} s_{ij}, i \in V]$	$O(n^2)$	$O(n)$
Graph Cut	$\lambda \sum_{i \in V} \sum_{j \in X} s_{ij} - \sum_{i,j \in X} s_{ij}$	$[\sum_{j \in X} s_{ij}, i \in V]$	$O(n^2)$	$O(n)$
Feature Based	$\sum_{i \in \mathcal{F}} \psi(w_i(X))$	$[w_i(X), i \in \mathcal{F}]$	$O(n\|\mathcal{F}\|)$	$O(\|\mathcal{F}\|)$
Set Cover	$w(\cup_{i \in X} U_i)$	$\cup_{i \in X} U_i$	$O(n\|U\|$	$\|U\|$
Prob. Set Cover	$\sum_{i \in \mathcal{U}} w_i[1 - \prod_{k \in X}(1 - p_{ik})]$	$[\prod_{k \in X}(1 - p_{ik}), i \in \mathcal{U}]$	$O(n\|\mathcal{U}\|)$	$O(\|\mathcal{U}\|)$
DPP	$\log\det(S_X))$	$\mathrm{SVD}(S_X)$	$O(\|X\|^3)$	$O(\|X\|^2)$
Dispersion Min	$\min_{k,l \in X, k \neq l} d_{kl}$	$\min_{k,l \in X, k \neq l} d_{kl}$	$O(\|X\|^2)$	$O(\|X\|)$
Dispersion Sum	$\sum_{k,l \in X} d_{kl}$	$[\sum_{k \in X} d_{kl}, l \in X]$	$O(\|X\|^2)$	$O(\|X\|)$
Dispersion Min-Sum	$\sum_{k \in X} \min_{l \in X} d_{kl}$	$[\min_{k \in X} d_{kl}, l \in X]$	$O(\|X\|^2)$	$O(\|X\|)$

Table 1. List of Submodular Functions used, with the precompute statistics $p_f(X)$, gain evaluated using the precomputed statistics $p_f(X)$ and finally T_o^f as the cost of evaluation the function without memoization and T_p^f as the cost with memoization. It is easy to see that memoization saves an order of magnitude in computation.

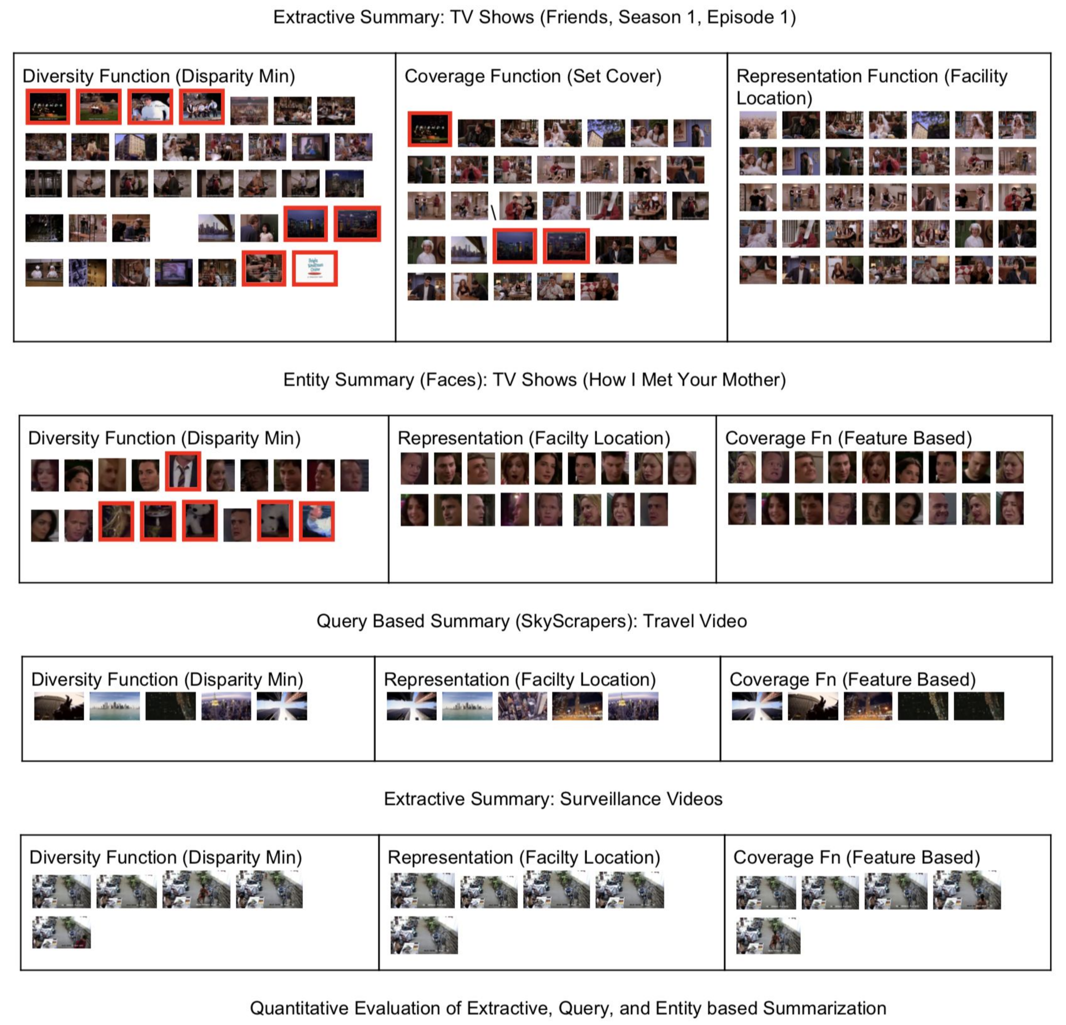
- In the framework, submodular video summarization is treated as two distinct optimization problems.
  - Budget Constrained Submodular Maximization

$$X^{t+1} = X^t \cup \operatorname{argmax}_{j \in V \setminus X^t} \frac{f(j|X^t)}{c(j)}$$

  - Submodular Cover Problem

$$X^{t+1} = X^t \cup \operatorname{argmax}_{j \in V \setminus X^t} f(j|X^t)$$, Until $f(X^t) = f(V)$

Experiments and Results



- In the illustrated results, we show the following:
  - Extractive summarization on TV shows.
  - Entity summarization on TV shows.
  - Query summarization for a query “skyscraper”.
  - Extractive summarization on surveillance videos.
- In each case, we compare Representation, Diversity and Coverage models.

Quantitative Evaluation of Extractive, Query, and Entity based Summarization

- The greedy nature of the optimization algorithm allows us to maintain precompute statistics for a set, using which the gain can be evaluated much more efficiently.
- With memoization, the order of the evaluation of gains is often an order of magnitude faster.

	Memoization			No Memoization		
Function	**5%**	**15%**	**30%**	**5%**	**15%**	**30%**
Fac Loc	0.34	0.4	0.71	48	168	270
Sat Cov	0.36	0.64	0.92	55	177	301
Gr Cut	0.39	0.52	0.82	41	161	355
Feat B	0.16	0.21	0.32	9	16	21
Set Cov	0.21	0.31	0.41	5	16	31
PSC	0.11	0.37	0.42	7	19	35
DPP	32	107	411	171	1003	4908
DM	0.11	0.61	0.82	21	125	221
DS	0.21	0.63	0.89	41	134	246

Table 2. Timing results in seconds for summarizing a two hour video for various submodular functions

Conclusions

- We present a unified picture of multi-faceted video summarization for extractive, query based and entity based summarization.
- Take a closer look at different summarization models and argue the benefits of these models in different domains by comparing qualitative and quantitative results.
- Implementation tricks, like memoization, can drastically improve the summary generation time, as compared to computational gains, such as usage of GPUs.