TVSum: Summarizing Web Videos using Titles

Yale Song*, Jordi Vallmitjana, Amanda Stent, Alejandro Jaimes
Computer Vision Group, Yahoo Labs NYC

YAHOO! LABS

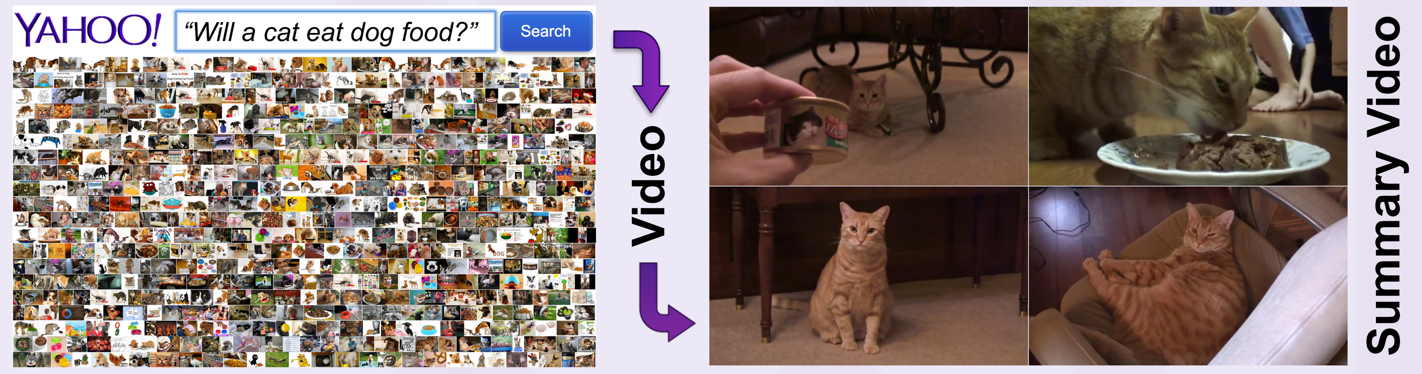
CVPR 2015 BOSTON JUNE 7-12

Contributions

- ❑ **TVSum**: **T**itle-based **V**ideo **sum**marization system guided by natural language description of the video
- ❑ **Co-archetypal Analysis**: a novel algorithm for **cross-dataset canonical visual concept learning**
- ❑ **TVSum50**: a novel, diverse dataset for video sum., available via Yahoo! Webscope Program

Overview

❑ **Idea: a video title hints towards the main story**

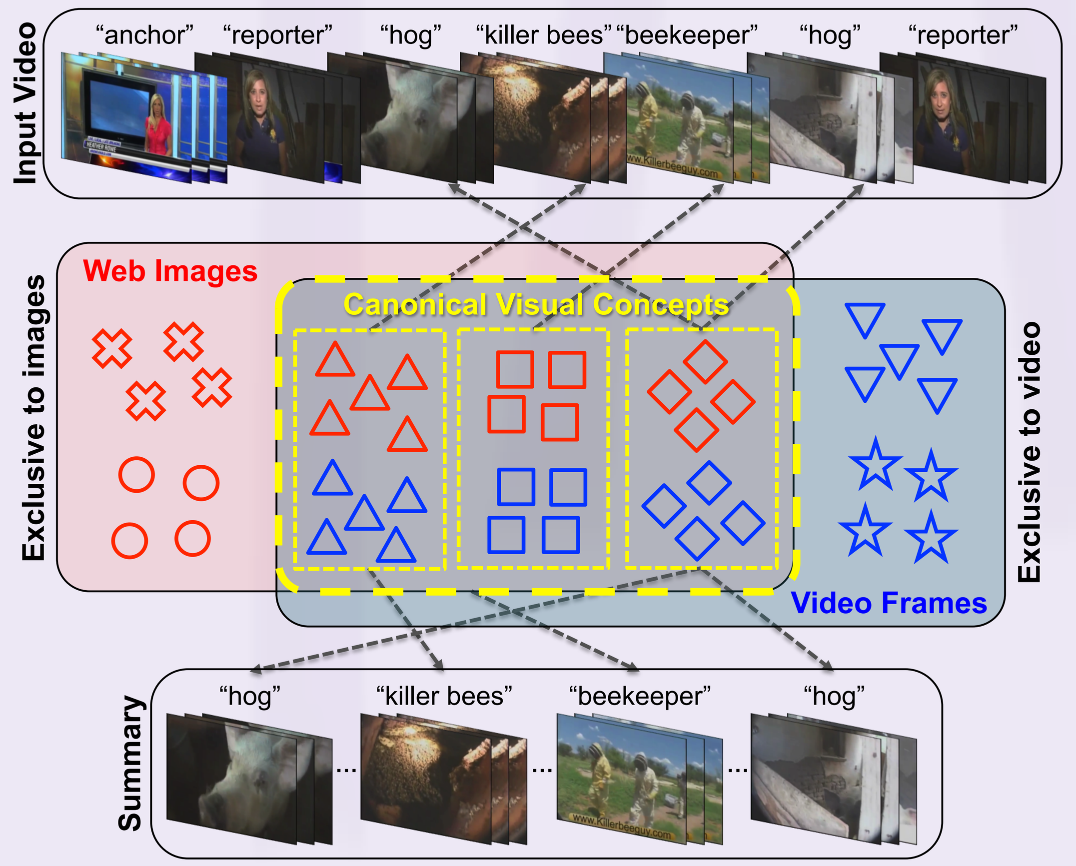


But real-world titles are free-formed, unconstrained, and ambiguous. This increases noise & variance in search results. Our work shows how to leverage title information effectively.

❑ **Solution: Co-archetypal Analysis**

- Learn canonical visual concepts from a combination of video frames and title-based image search results

Killer Bees Hurt 1000-lb Hog in Bisbee AZ



TVSum Pipeline

Web Image Crawling

- Video Title → **Query Builder**
  - Stop word & special character removal
  - Query expansion (n-gram, synonym)
- → **Image Search** → Web Images

Shot Boundary Detection

- Video → **Feature Extraction**
  - Color histogram pHOG, GIST, dSIFT
- → **Multiple CPD**
  - Convex formulation with $L_{2,1}$ group sparsity norm
- → Shot Boundaries

Canonical Visual Concept Learning

- Video, Web Images → **Feature Extraction**
  - Color histogram pHOG, GIST, dSIFT
- → **Co-arch. Analysis**
  - Two geometrical constraints
  - Block-wise convex formulation
- → Dictionary

Summary Generation

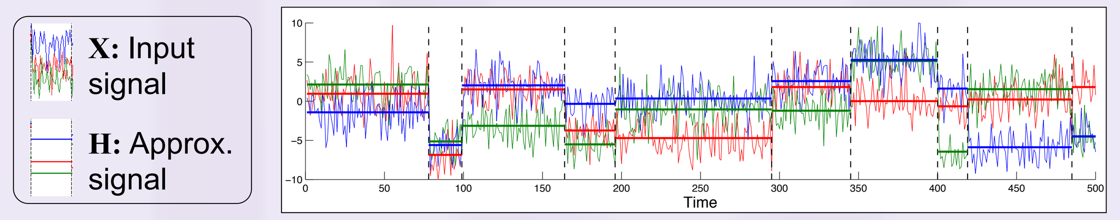
- Video, Dictionary → **Frame Scoring**
  - Two versions
- Shot Boundaries → **Shot Scoring**
  - Shot-wise average
- → **Shot Selection**
  - 0/1 Knapsack
- → Summary

Shot Boundary Detection

❑ **Multiple change point detection**

Convex formulation with an $L_{2,1}$ norm [Bleakley & Vert, 2011]

$$\min_{H} \overbrace{\frac{1}{2}\|X-H\|_F^2}^{\text{Reconstruction Error}} + \overbrace{\lambda\sum_{t=1}^{n-1}\|H_{\cdot,t+1}-H_{\cdot,t}\|_2}^{\text{Total Variation}}$$



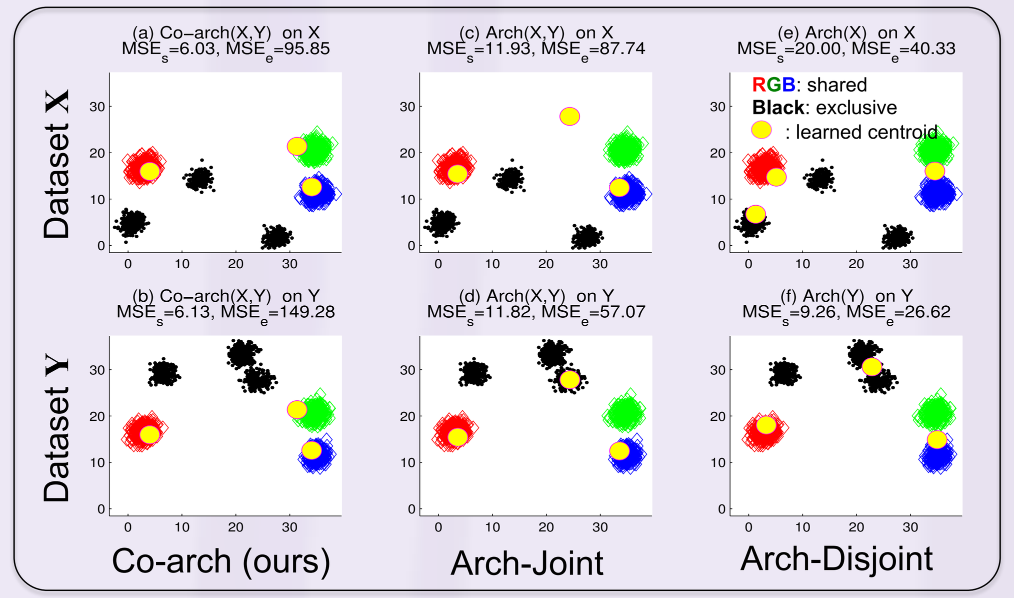
- **H** is **column-wise sparse** because of the $L_{2,1}$ norm
- Comparison to DP solution on NASA videos [TREC-01]
  - ☑ More accurate: mean F1 score of 0.60 vs. 0.35
  - ☑ Runs faster: 7.5s vs. 39.4s on a 7 minute video
  - ☑ More practical: no need to specify #CP a priori

Canonical Visual Concept Learning

❑ **Co-archetypal Analysis**

- Learn a dictionary of visual patterns **Z** shared *jointly* between video frames **X** and web images **Y**
- Two geometrical constraints:
  1. Each x_i and y_i should be well approx.-ed by a convex combination of **Z**, i.e., $X \approx ZA^X$ and $Y \approx ZA^Y$
  2. Each z_j should be well approx.-ed *jointly* by a C.C. of **X** and by a C.C. of **Y**, i.e., $Z \approx XB^X \approx YB^Y$

$$\min_{\Omega} \overbrace{\underbrace{\|X-ZA^X\|_F^2}_{\text{Video Factorization}} + \underbrace{\|Y-ZA^Y\|_F^2}_{\text{Web Image Factorization}}}^{\text{Geometrical constraint \#1}} + \gamma \overbrace{\underbrace{\|XB^X-YB^Y\|_F^2}_{\text{Video-Image Joint Regularization}}}^{\text{Constraint \#2}}$$



1. Co-arch finds **canonical patterns** (yellow circles) more closely aligned with the ground-truth
2. Co-arch results in the **lowest** MSE_s (shared) and the **highest** MSE_e (exclusive) scores

- ☑ Compared to Archetypal Analysis [Cutler & Brieman 1994], our Co-archetypal Analysis learns a dictionary Z that captures canonical patterns that appear both in X (video) and Y (images), but not in either alone
- ☑ Effectively deals with noise (images irrelevant to video) and variance (multiple, unknown numbers of visual concepts)

Summary Generation

❑ Compute shot scores by averaging frame scores

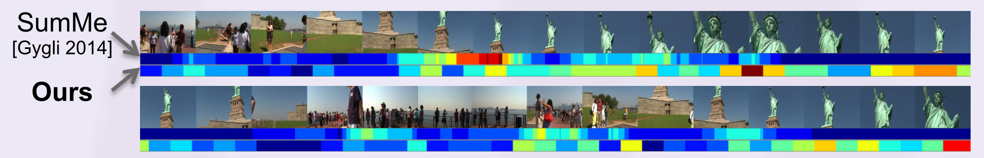
$$score_{ver1}(x_i) = \|x_i - Z\alpha_i\|_2, \quad score_{ver2}(x_i) = \sum_{j=1}^{n} B_i\alpha_j$$

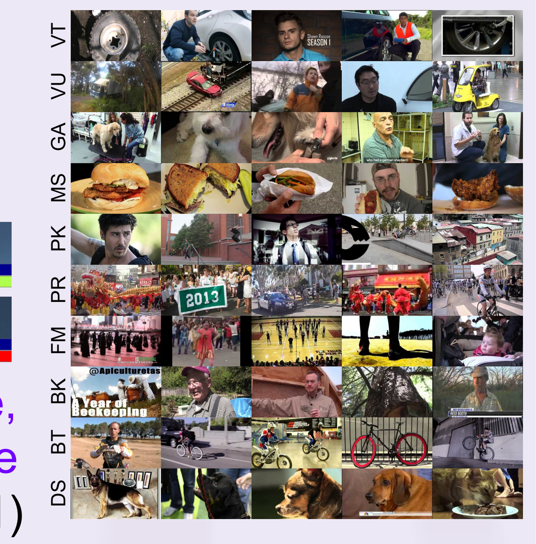
❑ Given a summary length budget, solve 0/1 knapsack

$$\max \sum_{i=1}^{\#shots} u_i \times score(s_i) \ \text{ s.t. } u_i \in \{0,1\}, \sum_i u_i \times length(s_i) \le budget$$

TVSum50 Benchmark Dataset

- ❑ 50 videos, 10 categories, 3.5 hours
- ❑ 1000 crowd ratings, 20 per video
- ❑ Avoiding chronological bias





- Rating done by visual importance & relevance, without being affected by temporal precedence
- Highly consistent ratings (Cronbach's α = 0.81)

❑ Available via Yahoo! Webscope Program

Experiments

- ❑ Task: summarize video with 15% length budget
- ❑ Metric: mean pairwise F1 scores w.r.t. human summaries
- ❑ Features: color histograms, pHOG, GIST, dSIFT

SumMe (25 videos) [Gygli et al '14]

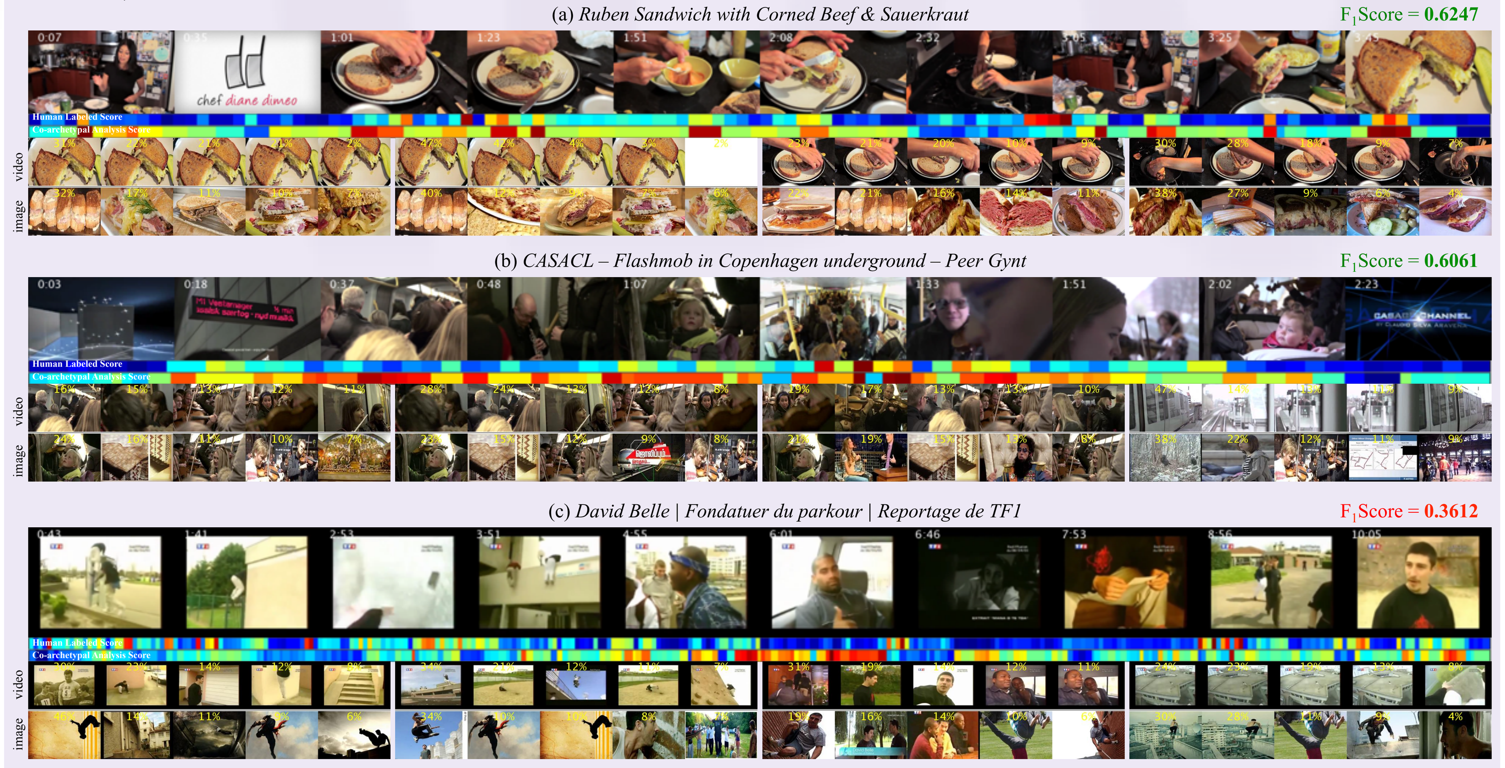
Methods	mpF1
Uniform Sampling*	0.14
K-means Clustering*	0.16
Attention* [Ejaz '13]	0.17
Interestingness* [Gygli '14]	0.23
LiveLight [Zhao & Xing '14]	0.24
Web Image Prior [Khosla '13]	0.24
Co-archetypal Analysis	**0.27**

* Results are from [Gygli et al. 2014]

Our TVSum50

Methods	mpF1
Uniform Sampling	0.36
Random Sampling	0.32
K-means Clustering	0.35
Spectral Clustering	0.39
LiveLight [Zhao & Xing '14]	0.46
Web Image Prior [Khosla '13]	0.36
Arch (Video Only)	0.33
Arch (Video + Web Image)	0.35
Co-archetypal Analysis	**0.50**

❑ Qualitative Results



(a) Ruben Sandwich with Corned Beef & Sauerkraut — F_1Score = 0.6247

(b) CASACL – Flashmob in Copenhagen underground – Peer Gynt — F_1Score = 0.6061

(c) David Belle | Fondateur du parkour | Reportage de TF1 — F_1Score = 0.3612