TVSum: Summarizing Web Videos Using Titles

Yale Song, Jordi Vallmitjana, Amanda Stent, Alejandro Jaimes
Yahoo Labs, New York

Video summarization is a challenging problem in part because knowing which part of a video is important, and thus is "summary worthy," requires prior knowledge about its main topic. Motivated by the observation that a video title is often carefully chosen to describe its main topic, and thus serves as a strong prior on the expected summary, we present *Title-based Video Summarization* (TVSum), an unsupervised video summarization framework that uses title-based image search results to generate a summary. Previous works have explored a similar direction with web images [4, 5], but with an assumption that an input query is in the form of carefully chosen topical keywords, making it easier to obtain a compact set of images representing the main topic. This, however, is far from the case with the titles of online videos: They are free-formed, unconstrained, and often written ambiguously. Consequently, images searched using the title can contain noise (images irrelevant to video content) and variance (images of different topics), which makes it particularly difficult to leverage web images for video summarization. To deal with this challenge, we developed a novel *co-archetypal analysis* technique that learns a shared representation of video and images; we call such representation *canonical visual concepts* (see Figure 1). Unlike the standard archetypal analysis [1, 2], our technique finds a joint-factorial representation of two data sets by incorporating a regularization term that penalizes the discrepancy between the factorizations of the two sets with respect to co-archetypes. This enforces canonical visual concepts to capture only those patterns that appear *jointly* in video and images, but not in either alone. As a result, our approach can extract important shots in a video given only its title information (or any other form of textual metadata) in a fully unsupervised manner. This makes our approach practical and generalizable in the real-world setting with a wide range of videos.

System Overview: We first segment a video into shots, grouping a sequence of visual coherent frames. After collecting topical images online using a set of query terms derived from the title, we learn canonical visual concepts shared between video frames and web images using our co-archetypal analysis. We then measure an importance score of each video frame using the learned representation, and generate a summary by combining shots that maximize the total importance score under a length budget.

Co-archetypal Analysis: Our main technical contribution is the development of a novel shared representation learning scheme, which we refer to as *co-archetypal analysis*. Given n video frames $\mathbf{X} \in \mathbb{R}^{d\times n}$ and m images $\mathbf{Y} \in \mathbb{R}^{d\times m}$, the goal of co-archetypal analysis is to obtain p co-archetypes $\mathbf{Z} \in \mathbb{R}^{d\times p}$ that satisfies two geometrical constraints: (i) each video frame $\mathbf{x}_i$ (and image $\mathbf{y}_i$) should be well approximated by a convex combination of co-archetypes $\mathbf{Z}$, i.e., $\mathbf{x}_i \approx \mathbf{Z}\alpha_i^X$ (and $\mathbf{y}_i \approx \mathbf{Z}\alpha_i^Y$), where α_i^X (and α_i^Y) is a coefficient vector in the p-dimensional unit simplex Δ^p; (ii) each co-archetype $\mathbf{z}_j$ should be well approximated *jointly* by a convex combination of $\mathbf{X}$ and of $\mathbf{Y}$, i.e., $\mathbf{z}_j \approx \mathbf{X}\beta_j^X \approx \mathbf{Y}\beta_j^Y$, where β_j^X and β_j^Y are coefficient vectors in the unit simplex Δ^n and Δ^m, respectively. We encode the two constraints by formulating co-archetypal analysis as an optimization problem:

$$\min_{\Omega} \|\mathbf{X} - \mathbf{Z}\mathbf{A}^X\|_F^2 + \|\mathbf{Y} - \mathbf{Z}\mathbf{A}^Y\|_F^2 + \gamma\|\mathbf{X}\mathbf{B}^X - \mathbf{Y}\mathbf{B}^Y\|_F^2 \qquad (1)$$

where $\Omega = \{\mathbf{A}^X, \mathbf{A}^Y, \mathbf{B}^X, \mathbf{B}^Y\}$. The first two terms encode the first geometrical constraint; the third encodes the second constraint. We solve this problem using block-coordinate descent, cycling through variables in a deterministic order $(\mathbf{A}^X, \mathbf{A}^Y, \mathbf{B}^X, \mathbf{B}^Y)$.

Shot Importance Scoring: We first measure the frame-level importance score using the factorization of $\mathbf{X}$ into $\mathbf{XBA}$, by aggregating the total contribution of the corresponding elements of $\mathbf{BA}$ in reconstructing the original signal, i.e., $\texttt{score}(\mathbf{x}_i) = \sum_j \mathbf{B}_i\alpha_j$. We then compute shot-level importance scores by taking an average of the frame scores within each shot. We show empirically that this scoring function provides better performance than the conventional method based on the reconstruction error, e.g., [6].

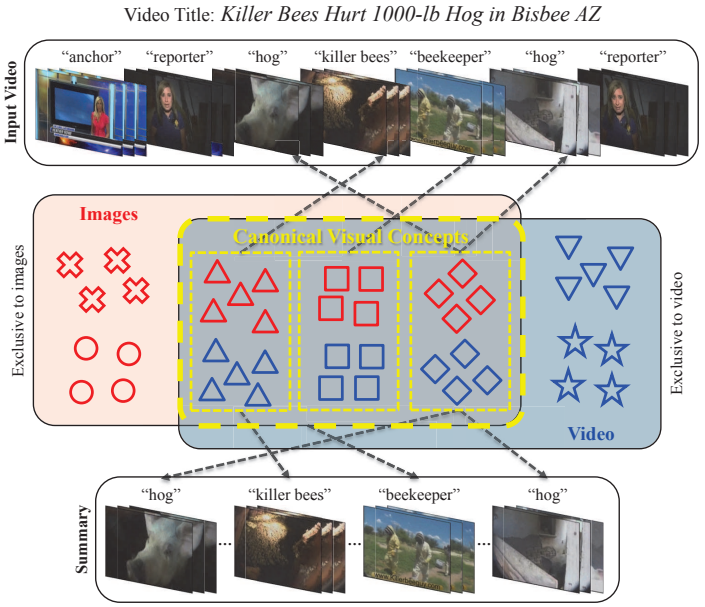

Figure 1: We use title-based image search results to summarize a video, selecting shots that are the most relevant to, and representative of, *canonical visual concepts* shared between video frames and images. We developed a novel *co-archetypal analysis* that learns canonical visual concepts by focusing on the shared region (yellow dotted rectangle area), singling out patterns that are exclusive to either set.

TVSum50 Dataset: We introduce a new benchmark dataset, *TVSum50*, that contains 50 videos representing various genres of online videos and their shot-level importance scores annotated using crowdsourcing. The videos are collected from YouTube using 10 video categories of the TRECVid MED task as the search query term (5 videos per category); all videos come with the Creative Commons license. The shot importance scores include 1,000 responses (20 per video) collected via Amazon Mechanical Turk, rated on a five-point Likert scale. Our design of the annotation interface, which avoids the *chronological bias* by pre-clustering and randomization, resulted in a high degree of label consistency, i.e., Cronbach's alpha of 0.81.

Experiments: We evaluated our approach on two real-world datasets, SumMe [3] and TVSum50, comparing against eight baselines including several recent approaches [1, 3, 4, 6]. The results showed that our approach outperforms all other baselines, achieving the mean pairwise F_1 score of 0.2655 on the SumMe dataset (previous state-of-the-art was 0.234 [3]) and of 0.4979 on our TVSum50 dataset. We also compared our co-archetypal analysis to the standard archetypal analysis [1, 2] on synthetic data, and showed that our approach correctly finds patterns shared between two data sets more accurately than archetypal analysis. The results suggest that our TVSum approach is able to produce a summary by learning a shared representation of video and title-based image search results.

[1] Y. Chen, J. Mairal, and Z. Harchaoui. Fast and robust archetypal analysis for representation learning. In *CVPR*, 2014.
[2] A. Cutler and L. Breiman. Archetypal analysis. *Technometrics*, 36(4), 1994.
[3] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *ECCV*, 2014.
[4] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013.
[5] G. Kim, L. Sigal, and E. P. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, 2014.
[6] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014.

This is an extended abstract. The full paper is available at the CVF webpage.