Video Summarization by Learning Submodular Mixtures of Objectives

Michael Gygli[1], Helmut Grabner[1] Luc Van Gool[1,2]
[1]Computer Vision Laboratory, ETH Zurich [2]PSI - VISICS, K.U. Leuven

Introduction. In this paper, we focus on reducing raw, casually captured videos to short, dynamic summaries. Automatically creating such a *skim* is challenging, as it must comply with at least two objectives [10]. Firstly, it should contain the most interesting parts of a video. Secondly, the summary should be representative in keeping the diversity of the original, while removing redundancy. Many recent methods predict a score per segment and ignore the structure of the video [1, 8], and therefore have difficulties to jointly optimize both objectives. Methods that go in this direction typically cluster the video into events and select the most important segment(s) per event [2, 5], following a kind of successive optimization of the objectives. Instead, our method optimizes for multiple objectives *globally*, avoiding hard decisions early on. Rather than using supervision only for some components [5] or making simplifying assumptions [1, 8], our method learns the importance of summarization objectives directly from reference summaries created by human annotators, as depicted in Fig. 1. Using supervision for the task of video summarization is crucial, since it is extremely complex and highly task-dependent – summaries from surveillance or live-logging data are expected to meet different criteria than summaries of short clips obtained by a mobile phone. Our approach is able to automatically adapt to the type of video and the desired output. It is therefore much more general and can be applied in all of these settings. In Tab. 1 we show a comparison to the most relevant related work in terms of summarization objectives.

Method formulation. We formulate the task of video summarization as a subset selection problem. We are given a video $\mathcal{V}$ and a budget B. Let $\mathcal{Y}_\mathcal{V}$ denote the set of all possible solutions $\mathbf{y} \subseteq \mathcal{V}$ given this constraint.

The task of our method is to select a summary $\mathbf{y}^*$, such that it optimizes an objective o:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}_\mathcal{V}} o(\mathbf{x}_\mathcal{V}, \mathbf{y}), \tag{1}$$

where $\mathbf{x}_\mathcal{V}$ are features extracted from the video $\mathcal{V}$. We define $o(\mathbf{x}_\mathcal{V}, \mathbf{y})$ as a linear combination of objectives $\mathbf{f}(\mathbf{x}_\mathcal{V}, \mathbf{y}) = [f_1(\mathbf{x}_\mathcal{V}, \mathbf{y}), ..., f_n(\mathbf{x}_\mathcal{V}, \mathbf{y})]^T$, each capturing a different aspect of a summary:

$$o(\mathbf{x}_\mathcal{V}, \mathbf{y}) = \mathbf{w}^\mathrm{T}\mathbf{f}(\mathbf{x}_\mathcal{V}, \mathbf{y}). \tag{2}$$

Since $\mathcal{Y}_\mathcal{V}$ is growing exponentially with the length of the video, optimally solving Eq. (2) quickly becomes intractable. Therefore, we restrict the objectives $\mathbf{f}(\mathbf{x}_\mathcal{V}, \mathbf{y})$ to be monotone submodular and $\mathbf{w}$ to be non-negative. This allows to find a near optimal solution for Eq. (1) in an efficient way [4, 7].

For the objectives $f_i(\mathbf{x}_\mathcal{V}, \mathbf{y})$, any kind of submodular function is possible. In this work, we use three objectives: (i) An interestingness predictor, as [1, 5], (ii) a k-medoid objective, which favors representative solutions and (iii) an objective that regularizes the summary to select segments at more uniform time intervals. In order to learn the weights of Eq. (2), we use a large-margin formulation that is optimized using stochastic gradient descent [6].

		Sun [9]	Gygli [1]	Patapov [8]	Lee [5]	Kim [3]	**Ours**
Obj.	Interesting	√	√	√	√	√	√
	Representative	-	-	-	√	√	√
	Uniform	-	-	-	(√)	-	√
Comb.	Learnt weights	-	-	-	-	-	√
	Optimized jointly	-	-	-	-	√	√

Table 1: Taxonomy of the most recent and relevant methods. We differentiate in terms of objectives they use and and how they combine them. Many methods score segment locally. Others combine multiple objectives, but do so based on a hand-defined sequential optimization. In opposition, we learn the importance of each objective from data and optimize them jointly.

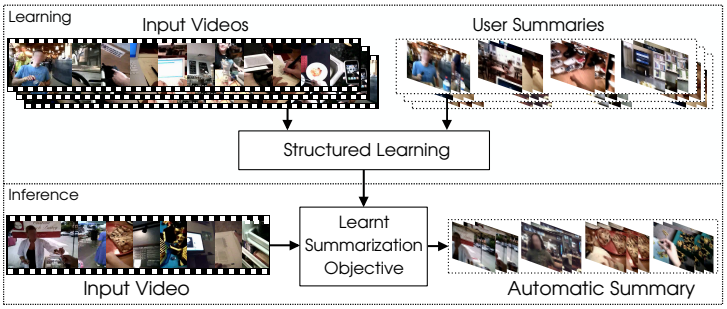

Figure 1: **Overview.** Our method consists of two parts: A supervised learning stage (training) and inference (testing). Given pairs of videos and their user created summaries as training examples, we learn a combined objective. Then, when given a new video as input, our method creates summaries that are both interesting and representative.

Results. We evaluate the performance of our method on two datasets: (i) a egocentric dataset [5] and (ii) the SumMe dataset [1]. These datasets are extremely diverse: While the SumMe dataset consists of short user videos, the egocentric dataset contains hour long life-logging data. Our supervised approach is able to learn that on short user videos [1], interestingness is dominant, with 97.5% of the weight. For life-logging videos [5], on the other hand, it learns that it is important to select summaries that are also representative (38%) and more uniform (9.6%). As a consequence, we are able to match [1] or outperform [5] previous methods on these two very diverse datasets.

[1] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating Summaries from User Videos. *ECCV*, 2014.
[2] Aditya Khosla, Raffay Hamid, CJ Lin, and Neel Sundaresan. Large-Scale Video Summarization Using Web-Image Priors. *CVPR*, 2013.
[3] Gunhee Kim, Leonid Sigal, and Eric P Xing. Joint Summarization of Large-scale Collections of Web Images and Videos for Storyline Reconstruction. *CVPR*, 2014.
[4] Andreas Krause and Daniel Golovin. Submodular Function Maximization. *Tractability: Practical Approaches to Hard Problems*, 2011.
[5] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. *CVPR*, 2012.
[6] Hui Lin and Jeff Bilmes. Learning mixtures of submodular shells with application to document summarization. In *Uncertainty in Artificial Intelligence (UAI)*, 2012.
[7] GL Nemhauser, LA Wolsey, and ML Fisher. An analysis of approximations for maximizing submodular set functions - I. *Mathematical Programming*, 14, 1978.
[8] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. *ECCV*, 2014.
[9] Min Sun, Ali Farhadi, and Steve Seitz. Ranking Domain-Specific Highlights by Analyzing Edited Videos. *ECCV*, 2014.
[10] Ba Tu Truong and Svetha Venkatesh. Video abstraction. *ACM TOMCCAP*, 2007.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.