

# Improving Sequential Determinantal Point Processes for Supervised Video Summarization

Aidean Sharghi<sup>1</sup>[0000000320051334], Ali Borji<sup>1</sup>, Chengtao Li<sup>2</sup>[0000000323462753], Tianbao Yang<sup>3</sup>[0000000278585438], and Boqing Gong<sup>4</sup>[0000000339155977]

<sup>1</sup> Center for Research in Computer Vision, University of Central Florida, Orlando, Florida

<sup>2</sup> Massachusetts Institute of Technology, Cambridge, Massachusetts

<sup>3</sup> University of Iowa, Iowa City, Iowa

<sup>4</sup> Tencent AI Lab, Seattle, Washington

**Abstract.** It is now much easier than ever before to produce videos. While the ubiquitous video data is a great source for information discovery and extraction, the computational challenges are unparalleled. Automatically summarizing the videos has become a substantial need for browsing, searching, and indexing visual content. This paper is in the vein of supervised video summarization using sequential determinantal point processes (SeqDPPs), which models diversity by a probabilistic distribution. We improve this model in two folds. In terms of learning, we propose a large-margin algorithm to address the exposure bias problem in SeqDPP. In terms of modeling, we design a new probabilistic distribution such that, when it is integrated into SeqDPP, the resulting model accepts user input about the expected length of the summary. Moreover, we also significantly extend a popular video summarization dataset by 1) more egocentric videos, 2) dense user annotations, and 3) a refined evaluation scheme. We conduct extensive experiments on this dataset (about 60 hours of videos in total) and compare our approach to several competitive baselines.

## 1 Introduction

It is now much easier than ever before to produce videos due to ubiquitous acquisition capabilities. The videos captured by UAVs and drones, from ground surveillance, and by body-worn cameras can easily reach the scale of gigabytes per day. In 2017, it was estimated that there were at least 2.32 billion active camera phones in the world [25]. In 2015, 2.4 million GoPro body cameras were sold worldwide [13]. While the big video data is a great source for information discovery and extraction, the computational challenges are unparalleled. Automatically summarizing the videos has become a substantial need for browsing, searching, and indexing visual content.

Under the extractive video summarization framework, a summary is composed of important shots of the underlying video. This notion of importance, however, varies drastically from work to work in the literature. Wolf defines the importance as a function of motion cues [42]. Zhao and Xing formulate it by reconstruction errors [48]. Gygli et al. learn a mixture of interestingness, representativeness, and uniformity measures to find what is important [12]. These differences highlight the complexity of video summarization. The criteria for summarizing vastly depend on the content, styles, lengths, etc.



































