Investigating User Requirements for Video Summaries of Instructional Videos

Megha Nawhal
Student #: 301332706
mnawhal@sfu.ca

Tatyana Mozgacheva
Student #:
tmozgach@sfu.ca

ABSTRACT

With a plethora of videos available on the web, it is commonplace for users to search for information resources online. The crowdsourced nature of instructional video content available on the web poses a challenge in retrieval of relevant content from long or redundant videos. Watching long videos is an uninteresting way to learn about any activity as a result of which users find it taxing to maintain uniform concentration levels throughout the playback time of the video. Therefore, creating video summaries is valuable to enhance the learnability of the multimedia content. Significant amount of work has been dedicated to video understanding and video summarization in the field of Computer Vision, however, these methods are usually designed for the videos recorded in a controlled setting and do not incorporate the user requirements.

In this project, we perform an observational study to understand the requirements of users from visual summaries for instructional videos and infer design recommendations in order to create an ovmof an ideal overview of the videos.

Author Keywords

Video Summarization; Usability; Computer Vision; Instructional Videos; Learning

INTRODUCTION

Watching long videos is uninteresting as a result of which users find it taxing to maintain uniform concentration levels throughout the playback time of the video. Given the fast-paced lives, users do not prefer to remember the details of such activities and therefore, feel the need to review the steps involved in performing a task before performing one. These users might be familiar of the components involved in the activity and wish to know more about processing of those components to complete the activity (*e.g.* a cooking enthusiast will be aware of the ingredients used for new recipe but wants to review the process of preparing a new dish). As a result, they seek a quick overview of the steps involved in performing a task. For example, a cooking enthusiast watching a video to get an overview of a recipe such as *"cooking perfect cinnamon panna cotta* may not be interested in the detailed video tutorial. In particular, it becomes likely to skip some steps in case of instructional videos with varying concentration of the users leading to dissatisfaction of the users in terms of content relevance as well as learnability of the online videos.

Use of web is commonplace for the users to acquire knowledge about instructions for different types of activities. However, the crowdsourced nature of instructional video content available on the web poses several challenges to the users. High variation in the focus of different stages of the set of instructions is dependent on the tutorial source. For instance, some steps span shorter duration of the video despite possessing same level of content detail as well as execution difficulty. Context variation in the videos when the steps are often linked to earlier steps of the task make it difficult for the users to follow the task, especially when the task involved lot of steps and the videos are too long. Therefore, it is important to understand the user requirements and summarize these in-the-wild videos in a succinct manner.

Considerable research efforts have been dedicated to video understanding and video summarization in the field of Computer Vision [2],[3],[4],[6],[5],[7], however, these methods are usually designed for the videos recorded in a controlled setting and do not incorporate the user requirements. Conventional approaches of creating visual summaries of the videos in an automated fashion are based either of the following underlying paradigms: keyframe selection and subshot selection. With the popularity of the machine learning based approaches and availability of significant amount of data, the Computer Vision algorithms have evolved to become more sophisticated. Nonetheless, one of the major shortcomings of these approaches are that getting accurate annotations is expensive and in case of videos, the problem gets magnified because of the high dimensional nature of video data. Further, annotation of video content for instructional videos is challenging because the task is subjective. It is likely to involve the preferences of the human annotator since the ranking of relevance might vary from one person to another. On the other hand, the unsupervised learning based approaches eliminate the annotation cost, however, the evaluation of these performance of unsupervised approaches is defined based on some heuristic measures. These measures are often not able to incorporate all aspects of the task. Thus, it is valuable to evaluate these methods in the framework of usability of the summarized video content generated using the underlying principles of the current video summarization methodologies.

In this project, we perform an observational study to get a better understanding of the user requirements for visual summaries. We obtain the data from YouTube-8M dataset [1] for *Cooking show* categories and show these videos to the users. We utilize the formative evaluation of the way users watch the videos with a goal to perform that activity after looking at the video description. In this study, we attempt to study the content relevance of the in-the-wild videos that and quantify the amount of redundancy in the long videos. The goal of this study is to understand the requirements of users watching instructional videos and infer a notion of an ideal overview of the videos. Further, we derive design recommendations for obtaining summaries for the instructional videos based on the feedback of the users.

REFERENCES

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

2. Goldman, D. B., Curless, B., Salesin, D., and Seitz, S. M. Schematic storyboarding for video visualization and editing. In *ACM Transactions on Graphics (TOG)*, vol. 25, ACM (2006), 862–871.

3. Ji, Z., Xiong, K., Pang, Y., and Li, X. Video summarization with attention-based encoder-decoder networks. *arXiv preprint arXiv:1708.09545* (2017).

4. Laganière, R., Bacco, R., Hocevar, A., Lambert, P., Païs, G., and Ionescu, B. E. Video summarization from spatio-temporal features. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, ACM (2008), 144–148.

5. Lee, Y. J., Ghosh, J., and Grauman, K. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE (2012), 1346–1353.

6. Liu, D., Hua, G., and Chen, T. A hierarchical visual model for video object summarization. *IEEE transactions on pattern analysis and machine intelligence 32*, 12 (2010), 2178–2190.

7. Panda, R., Das, A., Wu, Z., Ernst, J., and Roy-Chowdhury, A. K. Weakly supervised summarization of web videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE (2017), 3677–3686.