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 instructional videos is an uninteresting way to learn about any activity and 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 a notion of an ideal overview of the videos.

Author Keywords

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

INTRODUCTION

Watching long videos is uninteresting and users find it taxing to maintain uniform concentration levels throughout the playback time of the video. Therefore, creating shorter video summaries is valuable for better comprehension of the videos. Serving the goal to provide concise video understanding, video summaries in various domains are aimed at distilling key information content from long video recordings. For instance, summarization of lengthy egocentric videos [41, 53, 35, 17] are aimed at removing redundancy in the first person recordings of events (*e.g.* person going on a long cycling trip), overviews of sports videos detect the key events referred to as highlights of the sport archives [10, 30, 47], surveillance videos [21, 43, 36, 31] captured around the clock by the security cameras need to be summarized in order to retrieve and visualize anomalous events or critical activities than can be difficult to search for in the original video. On the other hand, informational videos (*e.g.* video lectures for massive online open courses (MOOCs) and classroom teaching) [8, 4, 40, 16] and instructional videos (*e.g. "how-to"* videos) [3] require creation of summaries to serve the purpose of conveying the full information content of the video in a concise manner even though the videos are not as lengthy as egocentric videos and possess relatively less redundancy in terms of content. In this work, we investigate videos related to instructional do-it-yourself (DIY) activities.

Use of web is commonplace for the users to acquire knowledge about certain pervasive DIY activities such as videos related to cooking, hairstyling, home management and repair. However, the crowdsourced nature of instructional video content available on the web poses several challenges to the users. First, temporal 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 involves lot of steps and the videos are longer. Second, depending on the source of the video content, high variation is observed in the focus of different steps of the instructions related to the activity in consideration. In particular, some steps span shorter duration of the video despite possessing same level of content detail as well as execution difficulty.

Given the fast-paced lives, users do not prefer to remember the details of pervasive activities and therefore, feel the need to review the steps involved in these activities using various sources of information such as online video streaming websites (*e.g.* YouTube, Dailymotion, etc) and forums [24]. 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. However, in case of long instructional videos, it becomes likely that the users would skip some steps with varying concentration levels leading to user dissatisfaction in terms of content relevance as well as learnability of these online videos. Therefore, it is important to understand the user requirements for a video summary in order to obtain concise instructions from the videos.

Considerable research efforts have been dedicated to video understanding and summarization in the fields of computer vision and multimedia. Conventional automated video summarization methods adopt either of the following principles: obtaining a sequence of important frames in the video referred to as *keyframes* [28, 33, 52, 15, 53, 50] or detecting key video segments known as *subshots* [20, 37, 42, 44, 26, 11]. In this work, we evaluate these design principles for the video summaries in the framework of usability of the summarized video content for instructional *"how-to"* videos.

The popularity of the machine learning based approaches and availability of significant amount of data have led to evolution of the automated video understanding algorithms. Nonethe-

less, one of the major shortcomings of supervised learning based 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 can be challenging because of the subjectivity involved in the video understanding. Moreover, annotations are likely to get biased with 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 difficult due to the absence of ground truth data. Against this background, our goal is to infer design recommendations for the video summaries of complex DIY videos which can serve as a useful feed for automated algorithms to create summaries with enhanced learnability of the content.

As an initial step towards this goal, in this project, we perform an observational study to get a detailed 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 quantify the content relevance of these in-the-wild videos perceived by the user. 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.

RELATED WORK

Recent Practices around Instructional Videos

Humans being commendable visual learners tend to seek technical information for learning in the videos. Further, several studies have illustrated the wide variety of purpose of instructional videos besides the technical description [45, 22, 49, 25]. In this work, we aspire to improve the user learnability and comprehensibility of instructional videos.

Automated Video Summarization

Creating shorter summaries of videos in an automated manner has been studied for long in the research communities. While some of the summarization algorithms rely on visual cues to identify salient content in the video [28, 32, 15, 52], others aim at creating shorter videos using video segments corresponding to the key events in the video [29, 34, 47, 21]. Several machine learning algorithms tend to take a multimodal approach to create video summary by leveraging information such as subtitles and transcripts in addition to the visual content analysis [38, 13, 12]. Another line of work explores summarization using multiple related video sources that contribute to create a storyline of events in the video[35, 53]. More sophisticated computer vision algorithms focus at performing exemplar based summarization where the patterns in human annotated summary for few videos belonging to one category are used as templates to create summaries for other videos[54, 39]. Further, recent approaches have been tried to understand the relations between multiple events in the frames of video using the transcript [18]. It is important to mention that these algorithms have been experimented over videos that are simple in terms of number of events. In contrast, we evaluate the performance of these summarization algorithms in case of complex DIY activity videos that involve relatively higher number of events and possess referencing to previous events in the video.

Video Summarization based tools

Different video summarization based tools have been proposed in the past with a target to improve the experience of video authors as well as viewers. Democut [3] provides a video editing tool using a semi-automatic video summarization method to edit instructional videos. EgoScanning [17] provides the flexibility of fast forwarding the video to the help the users in finding important events from egocentric videos. In contrast, in our work we evaluate the system from the perspective of the viewers of the instructional video. VideoDigest [40] provide video summarization in order to improve the learnability of informational videos by leveraging the transcript of the video. Viewer interactions have been used for labeling of steps in instructional videos [23] and recognize highlights in sports videos using social media[48]. In contrast to these crowdsourced platforms, we monitor the interactions of individual viewers and infer design principles for summaries of instructional videos.

Types of Video Summaries

Video summaries has been of great interest to the Computer Vision and Multimedia community. Storyboard presentation of key frames is one of the common format of static video summaries [9, 27, 7]. On the other hand, dynamic summaries have been presented in formats such as storyboards of video skims [6, 46, 19] and video collage [5, 51, 14]. It has been seen in the prior research work that the format of video summaries are chosen based on the goal of video abstraction. By scoping down to instructional videos, we investigate the content and semantic structure in the instructional videos. Inspired by the work [2], we build on the idea of exploring the connections and hierarchy in the important events detected using visual information in the videos.

METHOD

In this project, we use YouTube videos related to cooking activities of duration more than 15 minutes. We select 4 different recipes which are similar in the number of steps involved. It is important to mention that we consider the number of steps involved in the process as an indicator for the difficulty of the recipe.

Users information

In our study, we recruit 6 participants from SFU student population with 3 participants being novice in cooking while 3 other participants being reasonably experienced in cooking. The diversity in experience with cooking would indicate the correlation of the browsing behavior and expertise.

Study Protocol

With the overall goal of observing the browsing and information seeking behavior of the viewers while watching the instructional video, we adopt the following protocol to perform our study.

- **Pre-study Questionnaire:** We gather demographic information of the viewers. Additionally, we collect information about usual video browsing behavior and their preferences for the description of DIY activities (*e.g* textual description, audio description or audio-visual content).
- **Task Description:** We request the users to watch four cooking videos in different conditions. The videos are to be watched subject to time constraints so as to simulate the scenario of getting a quick overview of the steps involved in an activity. In order to explore the role of different modes of information, we investigate four conditions listed as follows: (1) video without captions, (2) video with captions (3) only visual content with captions but without audio, (4) visual content with captions and audio content. As part of the task, also gather information about importance of an event by providing an option to mark important or irrelevant as perceived by them while watching the videos. We record these timestamps marked as important and irrelevant by the users and analyze the patterns in the browsing behavior.
- **Post-study Interview:** As a follow-up of the tasks being performed by the user, we conduct semi-structured interviews to get more insights about their browsing behavior and challenges they faced while watching these complex instructional videos.

Data Collection

We will have paper questionnaires and the user's identity will be anonymized while data collection by assigned each user a unique ID. For the data collection, we use a web interface with users given an option of marking certain events important or irrelevant and track the timestamps in a log file which will also be de-identified. Further, for qualitative analysis, we record the interview and the user interactions with the web interface using the recording software Camtasia.

Data Analysis

We report qualitative and quantitative results of the data collected. We perform coding on the qualitative data to infer the user requirements of an overview of a video. We quantify the content relevance by looking at the data and user reactions to the degree of irrelevance in the content and infer actionable insights from these correlations.

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. Chang, M., Hare, V. M., Kim, J., and Agrawala, M. Recipescape: Mining and analyzing diverse processes in cooking recipes. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ACM (2017), 1524–1531.
3. Chi, P.-Y., Liu, J., Linder, J., Dontcheva, M., Li, W., and Hartmann, B. Democut: generating concise instructional videos for physical demonstrations. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM (2013), 141–150.
4. Choudary, C., and Liu, T. Summarization of visual content in instructional videos. *IEEE Transactions on Multimedia 9*, 7 (2007), 1443–1455.
5. Christel, M. G., Hauptmann, A. G., Wactlar, H. D., and Ng, T. D. Collages as dynamic summaries for news video. In *Proceedings of the tenth ACM international conference on Multimedia*, ACM (2002), 561–569.
6. Christel, M. G., Smith, M. A., Taylor, C. R., and Winkler, D. B. Evolving video skims into useful multimedia abstractions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co. (1998), 171–178.
7. Christel, M. G., Winkler, D. B., and Taylor, C. R. Multimedia abstractions for a digital video library. In *Proceedings of the second ACM international conference on Digital libraries*, ACM (1997), 21–29.
8. Curtis, K., Jones, G. J., and Campbell, N. Utilising high-level features in summarisation of academic presentations. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ACM (2017), 315–321.
9. Ding, W., Marchionini, G., and Soergel, D. Multimodal surrogates for video browsing. In *Proceedings of the fourth ACM conference on Digital libraries*, ACM (1999), 85–93.
10. Ekin, A., and Tekalp, M. Generic play-break event detection for summarization and hierarchical sports video analysis. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1, IEEE (2003), I–169.
11. Ellouze, M., Boujemaa, N., and Alimi, A. M. Im (s) 2: Interactive movie summarization system. *Journal of Visual Communication and Image Representation 21*, 4 (2010), 283–294.
12. Erol, B., Lee, D.-S., and Hull, J. Multimodal summarization of meeting recordings. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 3, IEEE (2003), III–25.
13. Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., Maragos, P., and Avrithis, Y. Video event detection and summarization using audio, visual and text saliency. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE (2009), 3553–3556.

14. Fang, F., Yi, M., Feng, H., Hu, S., and Xiao, C. Narrative collage of image collections by scene graph recombination. *IEEE transactions on visualization and computer graphics* (2017).

15. 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.

16. He, L., Sanocki, E., Gupta, A., and Grudin, J. Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, ACM (1999), 489–498.

17. Higuchi, K., Yonetani, R., and Sato, Y. Egoscanning: Quickly scanning first-person videos with egocentric elastic timelines. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM (2017), 6536–6546.

18. Huang, D.-A., Lim, J. J., Fei-Fei, L., and Niebles, J. C. Unsupervised visual-linguistic reference resolution in instructional videos. *arXiv preprint arXiv:1703.02521* (2017).

19. Hürst, W., and Klappe, G. Design parameters for storyboard-based mobile video browsing interfaces. In *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*, IEEE (2017), 405–410.

20. Jackson, D., Nicholson, J., Stoeckigt, G., Wrobel, R., Thieme, A., and Olivier, P. Panopticon: a parallel video overview system. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM (2013), 123–130.

21. Ji, Z., Su, Y., Qian, R., and Ma, J. Surveillance video summarization based on moving object detection and trajectory extraction. In *Signal Processing Systems (ICSPS), 2010 2nd International Conference on*, vol. 2, IEEE (2010), V2–250.

22. Käfer, V., Kulesz, D., and Wagner, S. What is the best way for developers to learn new software tools? an empirical comparison between a text and a video tutorial. *arXiv preprint arXiv:1704.00074* (2017).

23. Kim, J., Nguyen, P. T., Weir, S., Guo, P. J., Miller, R. C., and Gajos, K. Z. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2014), 4017–4026.

24. Kuznetsov, S., and Paulos, E. Rise of the expert amateur: Diy projects, communities, and cultures. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ACM (2010), 295–304.

25. Lafreniere, B., Grossman, T., and Fitzmaurice, G. Community enhanced tutorials: improving tutorials with multiple demonstrations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 1779–1788.

26. 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.

27. Lee, H., and Smeaton, A. F. Designing the user-interface for the físchlár digital video library. *Journal of Digital information 2*, 4 (2002).

28. 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.

29. Lee, Y. J., Kim, J., and Grauman, K. Key-segments for video object segmentation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE (2011), 1995–2002.

30. Li, B., and Sezan, M. I. Event detection and summarization in sports video. In *Content-Based Access of Image and Video Libraries, 2001.(CBAIVL 2001). IEEE Workshop on*, IEEE (2001), 132–138.

31. Li, C., Wu, Y.-T., Yu, S.-S., and Chen, T. Motion-focusing key frame extraction and video summarization for lane surveillance system. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, IEEE (2009), 4329–4332.

32. 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.

33. Liu, T., and Kender, J. R. Optimization algorithms for the selection of key frame sequences of variable length. In *European Conference on Computer Vision*, Springer (2002), 403–417.

34. Lu, Y.-J., Zhang, H., de Boer, M., and Ngo, C.-W. Event detection with zero example: select the right and suppress the wrong concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ACM (2016), 127–134.

35. Lu, Z., and Grauman, K. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, IEEE (2013), 2714–2721.

36. Meghdadi, A. H., and Irani, P. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2119–2128.

37. Ngo, C.-W., Ma, Y.-F., and Zhang, H.-J. Automatic video summarization by graph modeling. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, IEEE (2003), 104–109.

38. 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.

39. Panda, R., and Roy-Chowdhury, A. K. Collaborative summarization of topic-related videos. In *CVPR*, vol. 2 (2017), 5.

40. Pavel, A., Reed, C., Hartmann, B., and Agrawala, M. Video digests: a browsable, skimmable format for informational lecture videos. In *UIST* (2014), 573–582.

41. Poleg, Y., Arora, C., and Peleg, S. Temporal segmentation of egocentric videos. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, IEEE (2014), 2537–2544.

42. Pongnumkul, S., Wang, J., and Cohen, M. Creating map-based storyboards for browsing tour videos. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, ACM (2008), 13–22.

43. Pritch, Y., Ratovitch, S., Hendel, A., and Peleg, S. Clustered synopsis of surveillance video. In *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, IEEE (2009), 195–200.

44. Pritch, Y., Rav-Acha, A., Gutman, A., and Peleg, S. Webcam synopsis: Peeking around the world. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, IEEE (2007), 1–8.

45. Riboni, G. The youtube makeup tutorial video. a preliminary linguistic analysis of the language of makeup gurus. *Lingue e Linguaggi 21* (2017), 189–205.

46. Smith, M. A., and Kanade, T. Video skimming and characterization through the combination of image and language understanding. In *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, IEEE (1998), 61–70.

47. Takahashi, Y., Nitta, N., and Babaguchi, N. Video summarization for large sports video archives. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, IEEE (2005), 1170–1173.

48. Tang, A., and Boring, S. # epicplay: Crowd-sourcing sports video highlights. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 1569–1572.

49. Torrey, C., McDonald, D. W., Schilit, B. N., and Bly, S. How-to pages: Informal systems of expertise sharing. In *ECSCW 2007*. Springer, 2007, 391–410.

50. Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. Video manga: generating semantically meaningful video summaries. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, ACM (1999), 383–392.

51. Wang, T., Mei, T., Hua, X.-S., Liu, X.-L., and Zhou, H.-Q. Video collage: A novel presentation of video sequence. In *Multimedia and Expo, 2007 IEEE International Conference on*, IEEE (2007), 1479–1482.

52. Wolf, W. Key frame selection by motion analysis. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, IEEE (1996), 1228–1231.

53. Xiong, B., Kim, G., and Sigal, L. Storyline representation of egocentric videos with an applications to story-based search. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), 4525–4533.

54. Zhang, K., Chao, W.-L., Sha, F., and Grauman, K. Summary transfer: Exemplar-based subset selection for video summarization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, IEEE (2016), 1059–1067.