Investigating User Requirements for Video Summaries of Instructional Videos

Megha Nawhal
mnawhal@sfu.ca

Tatyana Mozgacheva
tmozgach@sfu.ca

ABSTRACT

With a plethora of videos available on the web, it is commonplace for users to look for online informal learning resources related to complex instructional activities. 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 viewers find it taxing to maintain uniform concentration levels throughout the playback time of the video. Therefore, creating video summaries is valuable to enhance the effectiveness 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 viewers' 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

Given the fast-paced lives, users do not prefer to remember the details of certain pervasive do-it-yourself (DIY) activities such as cooking, home management, craft, or repair. Therefore, users feel the need to review the steps involved in these activities before performing an activity. The internet is rife with textual descriptions of the procedural DIY tasks, however, activities referring to the domains such as cooking, craft, or home management rely on the look and feel of the objects or events involved in accomplishing the task and the visual information associated with activities have been observed to enhance the comprehension of descriptions of these activities. Use of the video streaming websites (*e.g.* YouTube, Dailymotion, Vimeo, etc) is commonplace for the users to look for instructional *how-to* videos to acquire knowledge about procedural tasks [24]. However, watching long instructional videos is uninteresting which prompts the viewers to skip around while watching a video. Further, users find it taxing to maintain uniform concentration levels throughout the playback of a long video. With varying attention, it becomes likely that the users would skip some steps with varying concentration levels leading to user dissatisfaction in terms of learnability of these online videos. This calls for a need to create summaries to provide an overview that conveys the key information in the videos.

Despite the evident need for concise summaries of instructional videos, the comprehensibility of an overview is associated with the requirements of the user watching the video. For instance, while a cooking enthusiast watching a video to get an overview of a recipe for *cooking perfect cinnamon panna cotta* may not be interested in the detailed video tutorial, the same person trying to *assemble the components of a table purchased from IKEA* may want to glance at specific details of tools required during the assembly. Therefore, it is important to understand the viewers' requirements for visual overviews of instructional videos. Further, lengthy videos contain huge amount of audio-visual content, therefore it is important to understand the perceived effectiveness of each component of the video, namely, visual content, textual content (*i.e.* captions), and audio.

Moreover, the high variability in the readily available instructional video content poses several challenges to the task of automatically creating overviews of instructional video content. First, depending on the source of the video, there exists high variation in the focus on the different steps involved in the instructions. In particular, some steps span shorter duration of the video despite possessing same level of content detail as well as execution difficulty. Moreover, some description might be completely irrelevant to the instructions pertaining to the activity (*e.g.* advertisement of products used in the video, sharing personal experiences, etc.). Second, the temporal semantic variation in the visual content when the steps are associated to earlier steps of the task, *i.e.*, earlier events in the video being referenced by the current one or two consecutive steps are independent implying two unrelated events in the video, make it difficult for the users to follow the video, especially when the video involves lot of steps. For simplicity, we refer to these instances of temporal variation in content as context switch. Further, we refer to the switch due to associated steps as *linked* and unrelated steps as *disjoint*. For instance, making a *classic margherita pizza* from scratch consists of disjoint or unrelated steps such as preparing the *pizza dough* and *tomato sauce* and connected steps such as rolling the dough to a *flat pizza base* which is associated with the outcome of dough preparation stage).

Serving the goal to provide concise video understanding, video summaries for several applications are aimed at distilling key information content from lengthy videos. Conventional automated video summarization methods adopt either of the following principles: obtaining a sequence of important frames in the video referred to as *keyframes* [27, 32, 54, 12, 55, 53] or detecting key video segments also known as *subshots* [17, 38, 44, 46, 26, 9]. Summarization of lengthy egocentric videos [43, 55, 34, 15] are aimed at removing redundancy in the first person recordings of events. Overviews of sports videos provide athe key events referred to as highlights of the sport archives [8, 29, 50]. The surveillance videos [18, 45, 36, 30] captured round-the-clock by the security cameras need to be summarized in order to retrieve and visualize anomalous events or critical activities than can be time consuming to trace in the original video. In contrast, informational videos (*e.g.* video lectures for massive online open courses (MOOCs) and classroom teaching) [7, 6, 42, 14] and instructional videos (*e.g. "how-to"* videos) [5] require creation of summaries to provide in a concise content information even though the videos are relatively less lengthy and possess considerably less redundancy in content. In this work, we focus on instructional videos pertaining to DIY activities that are rich in visual content and specifically, we choose cooking as an example domain.

In this work, we explore the following research questions. Based on the observation that the instructional videos comprise of rich visual content along with other modalities of information such as audio content and textual content (*e.g.* subtitles), what are the user requirements of an overview of these videos rich in visual content? How do the other modalities of the information, *i.e.*, text and audio affect the perception of the visual content? To what extent, these additional information modalities improve the comprehensibility of instructional activity videos?

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. With the data collected from the study, we attempt to quantify the content relevance of these in-the-wild videos as per the user perception. 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

In this project, we focus on investigating the user requirements for overviews of long instructional videos. Our work builds on the prior work focusing on presentation of visual summaries for videos and navigation of these videos. Additionally, our work also draws inspiration from [4] that explores the structure in cooking recipes. In contrast to their tagging based approach using the textual descriptions of the cooking recipes, we leverage the structure in the cooking videos to enhance the effectiveness of the cooking videos.

Browsing Instructional Videos

Humans being commendable visual learners tend to seek technical information for learning in the videos. Further, in addition to acquiring the technical description of DIY activities, several studies have illustrated the wide variety of user needs satisfied by instructional videos ranging from drawing motivation for their own activity to validation of their knowledge about certain activities [47, 19, 52, 25]. In contrast, we aspire to understand the viewers' requirements associated with an overview of instructional videos for all types of users.

Specific to cooking activities, there have been studies in the past aimed at understanding persona-based user needs in the kitchen [21]. Additionally, past studies have also established step wise analysis of textual description of a recipe [3, 2]. In our work, we focus on visual descriptions and understand how users perceive the additional information modalities.

Video Summarization and Navigation Tools

Automated video summaries equip the viewers with a higher level of abstraction of the content in a video and aids in navigation of the video. Nonetheless, the nature of video browsing depends on the goals of the users. Conventional automated video summarization methods adopt either of the following principles: obtaining a sequence of important frames in the video referred to as *keyframes* [27, 32, 54, 12, 55, 53] or detecting key video segments also known as *subshots* [17, 38, 44, 46, 26, 9].

Systems employing *keyframes*-based video navigation focus on obtaining major key instances in videos. These systems facilitate browsing long entertainment videos [35], supporting textbook-style navigation for educational videos (*e.g.* lecture videos from MOOCs [56, 37], generating minutes of lengthy egocentric videos [55, 49, 15] and meeting recordings [53], and faster detection of events in surveillance videos [17]. Moreover, multimodal approaches are used to determine key instances in video by leveraging other modalities of information besides visual content such as transcripts in informational videos [42]. Additionally, markers and anchor points based approaches are also helpful for video authors to accelerate in video editing and manipulation process [5].

On the other hand, video navigation tools utilizing the summaries in *subshots* format are aimed at generating shorter clips of the video by removing the non-essential parts of the video. These summaries can accelerate the segmentation of videos to procure the key components in surveillance videos using objects and associated motion trajectories [20], [23] and generating sports video highlights [51]. Visualizing the videos in a space-time cube format are also used for *subhot*-detection of in the videos [39].

Another line of work uses viewer interactions for labeling of steps in videos using crowdsourcing platforms [22] and social media [51]. In contrast to these crowdsourced platforms, we monitor the interactions of individual viewers and infer design principles for summaries of instructional videos.

Automated Video Summarization

Various research communities have studied automatic creation of video summaries. While some of the summarization algorithms rely on visual cues to identify salient content in the video [27, 31, 12, 54], others aim at creating shorter videos using video segments corresponding to the key events in the video [28, 33, 50, 18]. Several machine learning algorithms tend to take a multimodal approach to create video summary by utilizing information such as subtitles and transcripts in addition to the visual content analysis [40, 11, 10]. Another

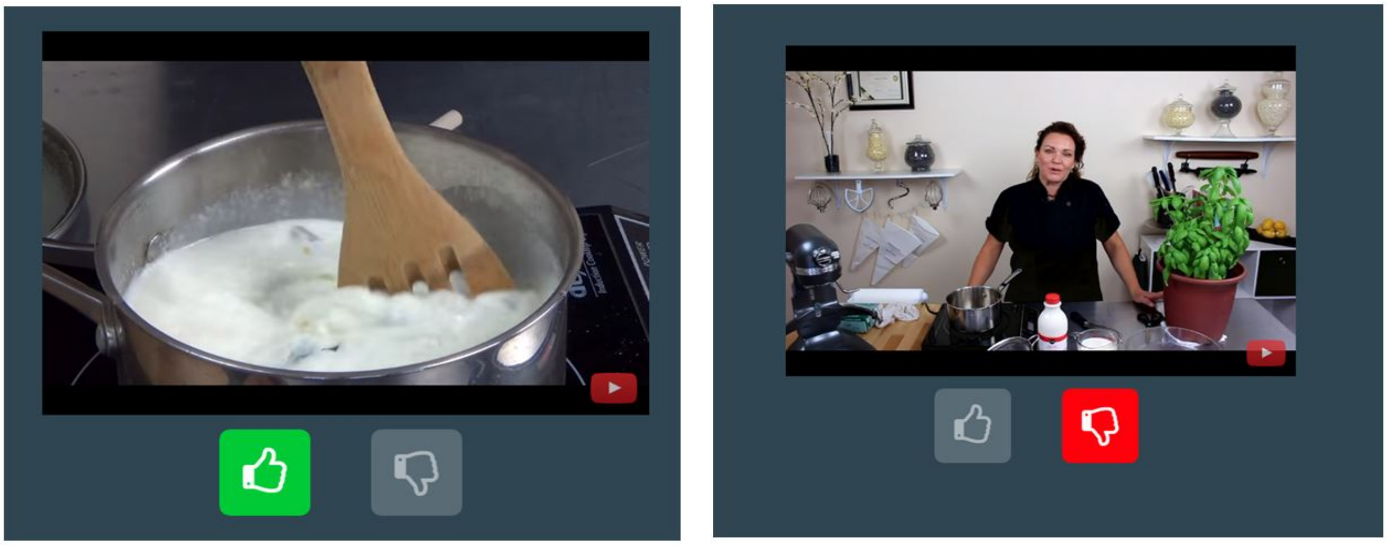


Figure 1. Participants mark the instance as relevant using the button shown in *left* image. All the instances from this point are logged as relevant until the participant marks some event as irrelevant using the button as shown in the *right* image

line of work explores summarization using multiple related video sources that contribute to create a storyline of events in the video [34, 55]. Recent computer vision algorithms focus on 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 [57, 41]. Furthermore, some recent approaches try to understand the relations between multiple events in the frames of video using the transcript [16]. However, these algorithms have been experimented with simple instructional videos having few number of steps [13, 48]. With this research, we provide design implications that can serve as useful feed for automated summarization algorithms for complex instructional videos that involve relatively higher number of events and several convoluted (*linked* and/or *disjoint context switch*) instances.

Alternate approach to algorithm based summarization is to crowdsource the summarization task. User interactions have been used for labeling of steps in instructional videos [22] and also to recognize highlights in sports videos using social media [51]. Crowdsourcing platforms have also been used to obtain representative keyframes from a video. In contrast, in this work we cater to individual user needs to browse an instructional video in a more effective manner.

OBSERVATIONAL STUDY

To better understand what specific information viewers seek while navigating through cooking videos and their preferences for different summary types, we carried out an observational user study. Moreover, we evaluated the effect of having other modalities of information, *i.e.*, captions and audio on the video browsing behavior.

For the user study task, we selected 4 recipe videos (Table 1) from the *Cooking_show* category of YouTube-8M dataset [1] to represent a diverse set of videos in terms of playback duration (6-8 minutes). During the selection process, we consulted 2 experts with more than 20 years of routine cooking experience to provide the step-wise breakdown of the recipe. With an objective to evaluate complex instructional videos, in consultation with experts, we selected recipe videos with multiple steps that possess both *disjoint* and *linked context switches* in the recipe instructions as per the description provided by experts.

Method

We conducted the study with 5 participants (3F/2M) who had varied levels of routine cooking experience ranging from less than 1 year to 10 years(refer to Table 2). Our study was divided in two consecutive components: Part (1): Task-specific session, and Part (2): Follow-up interviews. Each study session took approximately 40 minutes; 30 minutes for Part (1); and 10 minutes for Part(2).

Videos typically contain three modes of information, namely, visual content, audio, and captions or transcripts. It is important to mention that we focus on the viewers' perception of visual content and investigate the effect of additional modes of information. Hence, we designed 4 different conditions corresponding the the 4 videos described in Table 3. We followed a within-subjects design for our study, *i.e.*, all the participants were subjected to the same 4 conditions. Further, we randomized the order in which the participants were shown these video conditions.

In the task-specific session, we requested our participants to watch 4 cooking videos (Table 1) end-to-end. We asked our participants to mark the instances that are relevant and the ones that are irrelevant as they were watching the video. We developed an interactive tool for obtaining the timestamps in the video playback duration perceived as relevant and irrelevant by the participant. The interface is shown in Figure 1. To minimize the learning effects, we presented a commonplace

Recipe ID	**YouTube Video ID**	**Recipe name**	**Duration**	**# Steps**	**# Context Switches**
R1	yx75D8lAnqk	Basil icecream	7:19	21	11
R2	3DtEyQF5eAA	Carrot cake	7:24	14	5
R3	xErJVdZQP3M	Spanish rice with cilantro	6:38	20	9
R4	ELnNKcFo22E	Pumpkin Ravioli	6:34	18	8

Table 1. YouTube videos used for the study. The videos can be obtained from the website using the corresponding Video ID(column 2) in the following URL: `https://www.youtube.com/watch?v=`<Video ID>

.

ID	**Cooking experience**	**ID**	**Cooking experience**
	(# years)		(# years)
P1	8	P4	0.5
P2	10	P5	8
P3	4		

Table 2. Background information of participants of the formative study. Our participants have several levels of cooking experiences.

Condition ID	**Recipe ID**	**Visuals**	**Audio**	**Captions**
C1	R1	Yes	No	No
C2	R2	Yes	Yes	No
C3	R3	Yes	No	Yes
C4	R4	Yes	Yes	Yes

Table 3. The different conditions used in the study design showing the availability of the information modalities available for the recipe video

browsing interface for the study provided YouTube IFrame Player API [1].

We encouraged the users to think-aloud during the session. After browsing through each video, we requested the participants to provide ratings (on a Likert scale of 7) for the following facets of their experience of watching the video to learn the recipe: perceived difficulty level of the recipe, their familiarity with the recipe before watching the video, level of confidence in preparing the dish after watching the recipe video , and ease of following the instruction.

During the interviews, we asked our participants: (1) what specific information do they look for while browsing a cooking video, (2) to what extent did the captions and audio aid in comprehension of the video, (3) what type of information was perceived irrelevant or redundant to them and (3) what type of breakdown of the cooking video would help in better understanding of the instructions in the video.

RESULTS

In this section, we present the analysis of the recorded clicks and the ratings provided to us by the participants. We also describe the qualitative insights from the interviews conducted during the study.

Quantitative analysis

Saliency Maps

We analyzed the timestamps marked as salient for accomplishing the cooking task. We define the metric saliency percentage of the video as the percentage of time duration recorded as relevant by the participant for the video. Moreover, we define mean saliency percentage as the saliency percentage averaged across all the participants. The saliency graphs in Figure 2 represent the perceived relevance of content across all users. The saliency maps in Figure 2(b) and Figure 2(c) evaluate the role of captions and audio content respectively in the understanding of the visual content thereby evaluation of the effectiveness of the content. We observed that on addition of captions without any audio as well as audio without captions, the participants spent more time watching the video to understand the visual content with the help of text. This clearly illustrates that captions indeed aid in understanding of the content. However, when the participants were presented with a video containing both audio and captions, the mean saliency percentage dropped to 51.4% from the case when participants were presented Condition 2 (67.5%) and Condition 3 (64.4%).

In summary, we observed that on an average only 50-70% of the information is perceived as relevant for all the 4 conditions with minimum mean saliency percentage as 51.8% for the condition C4 and maximum mean saliency percentage being 67.5% for the condition C2.

User perception of saliency in different conditions

The presence of only one extra information modality was preferred by all the participants. We observed that the content saliency is perceived differently in different conditions. The saliency maps for the participant P4 is shown in Figure 3.

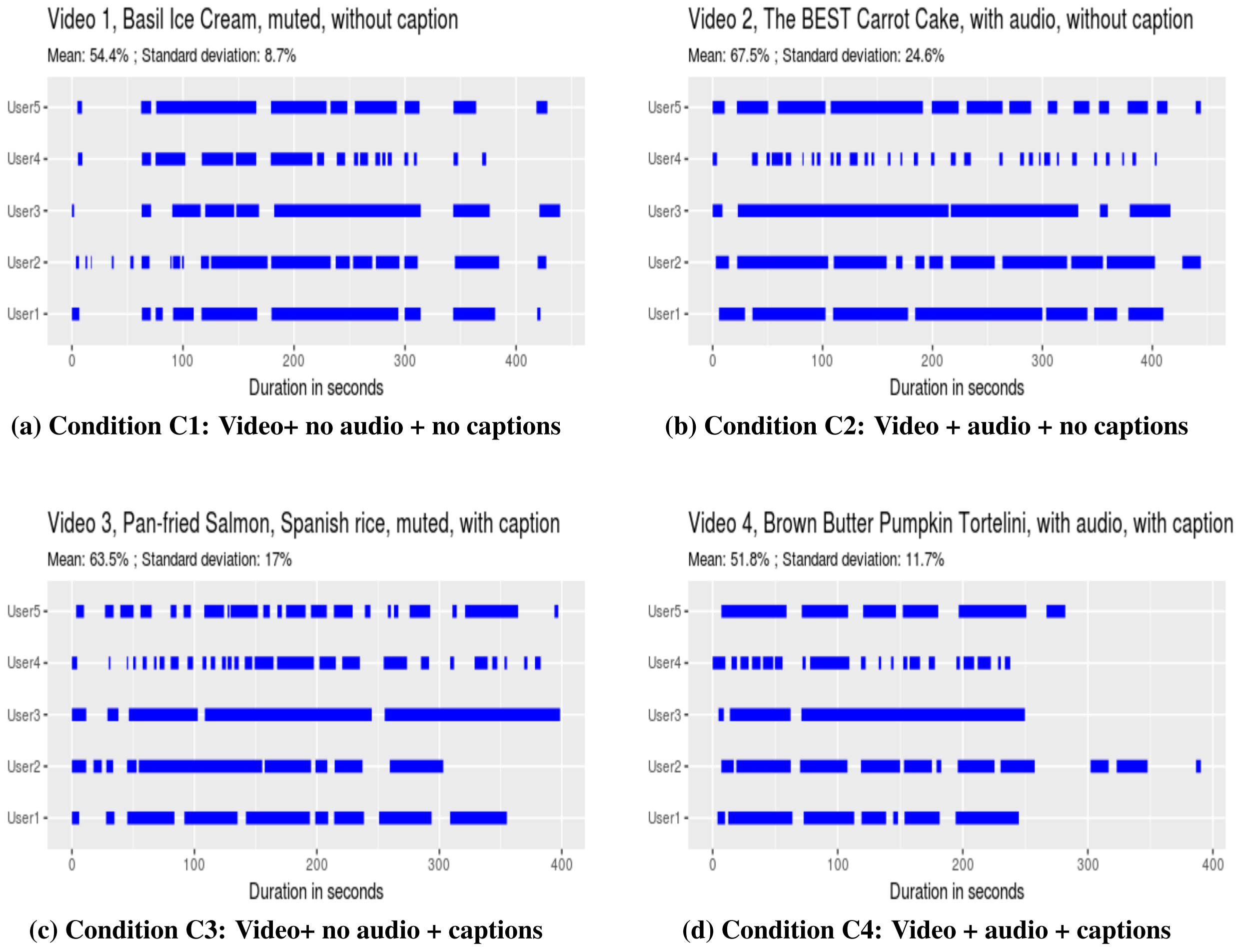
We observed that when there is no audio, the participants have to infer the content themselves from the visuals or the inaccurate subtitles which leads to significant content being not understood clearly such as list of ingredients and some information about the tools used in the recipe. On the other hand, when the participants were subjected to Condition 2 which contains the visual content along with only audio, the participants could infer more information about the actions performed in the video (*e.g.* folding of ravioli).

In summary, the participants preferred only one additional information modality. However, the participants use the information modalities differently to perceive the visual content saliency. In particular, while the participants use captions only to look for specific information, the participants pay significant attention to the audio to learn about specific actions.

Qualitative insights

All of the 5 participants agreed that the instructions pertaining to the cooking activity could be made more concise. Participant P1 expressed his annoyance by saying *"I prefer straight to the point, not too much personal stories, just how to do the*

[1] `https://developers.google.com/youtube/iframe_api_reference`



(a) Condition C1: Video+ no audio + no captions

(b) Condition C2: Video + audio + no captions

(c) Condition C3: Video+ no audio + captions

(d) Condition C4: Video + audio + captions

Figure 2. Video saliency graphs for all 4 conditions shown for all participants. The graphs also shown the mean saliency percentage along with standard deviation in the perceived saliency of the content.

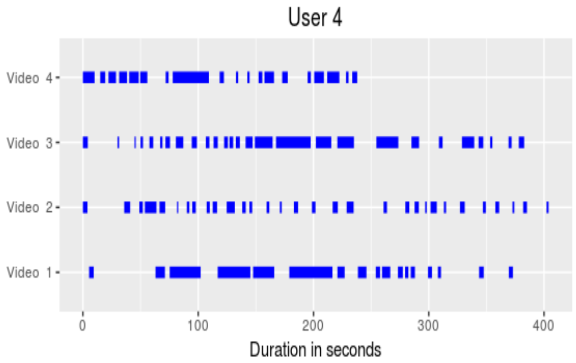


Figure 3. Video saliency graphs for all 4 conditions for participant P4.

cooking task." All our participants mentioned that the recipe instructions would be more useful and effective for them if they were major steps involved in the recipe video along with the ingredients. For example, participant P4 explained the notion of relevance of the video content as follows.*"It would be enough to see what I need to do. If you have idea how to cook in general, it is enough information that I had clicked."*

Most of the participants(4 of 5) preferred audio along with visual content (condition C2) over the other conditions. Nonetheless, they mentioned that they prefer explicit in textual description of information such as ingredients of the recipe, the tools used in the recipe, and the steps involving timing constraints (*e.g.* baking). Moreover, the participants found the long subtitles along with visual content (condition C3) distracting and mentioned that the amount of inference is the same as the condition with only visual content (condition C4). However, they acknowledged that short tags would be more helpful in comprehension of instructions. For instance, participant P5 summarized her needs and said *"I like when people talk and explain but also the are showing you how to do it.I usually dont read caption/subtitles, but I like when they put text explaining how much ingredients you need. It is helpful."*

DISCUSSION

In this paper, we investigated user requirements of content in instructional videos and to what extent other modes of information aid in understanding of the visual content. Our findings illustrated that the visual summaries showing key steps and the prerequisites for the task are more effective for the understanding of the complex instructional videos. Moreover, a multimodal summary with terse tags for key steps is more informative for viewers and would help the viewers to spend their video navigation time wisely. Additionally, our study also demonstrated that including audio with the visual content helps viewers in better understanding of continuous actions involved in the set of instructions.

In this work, we limited our scope to cooking videos. These inferences for instructional videos would be more conclusive when evaluated over videos in other DIY activity domains such as hairstyling, repair, and home management. It is also important to mention that we include only two modes of information presented in the standard format presented by YouTube. Nonetheless, there can be other way of presenting the content and exploring those presentation is also a promising direction.

We envision this work as an initial step towards the understanding of user requirements of the content of instructional videos. The inference from the study illustrate the multimodal nature of the viewers' requirements for visual summaries of the complex instructional videos. Further, the tool developed for obtaining the saliency maps of a video can be used for annotation of video content and these annotations can serve as a useful feed for the automated video summarization algorithms.

CONCLUSION

In this project, we have conducted an observational study to understand the user requirements of content in instructional videos. We have also evaluated the presence of additional information modality along with visual content in the context of video comprehensibility. We observed that only 50-60% of the content is perceived as important and the viewers prefer having short tags for the steps instead of having long subtitles. Moreover, we noted that only one additional mode of information was required with audio being preferred over the lengthy text. Our overall findings illustrate that video summaries are desirable to enhance the effectiveness of instructional videos.

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. Buykx, L., and Petrie, H. What cooks needs from multimedia and textually enhanced recipes. In *Multimedia (ISM), 2011 IEEE International Symposium on*, IEEE (2011), 387–392.
3. Buykx, L., and Petrie, H. Recipe sub-goals and graphs: an evaluation by cooks. In *Proceedings of the ACM multimedia 2012 workshop on Multimedia for cooking and eating activities*, ACM (2012), 59–64.
4. 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.
5. 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.
6. Choudary, C., and Liu, T. Summarization of visual content in instructional videos. *IEEE Transactions on Multimedia 9*, 7 (2007), 1443–1455.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. The something something video database for learning and evaluating visual common sense. In *Proc. ICCV* (2017).
14. 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.

15. 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.

16. 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).

17. 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.

18. 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.

19. 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).

20. Karrer, T., Weiss, M., Lee, E., and Borchers, J. Dragon: a direct manipulation interface for frame-accurate in-scene video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2008), 247–250.

21. Kerr, S. J., Tan, O., and Chua, J. C. Cooking personas: Goal-directed design requirements in the kitchen. *International Journal of Human-Computer Studies 72*, 2 (2014), 255–274.

22. 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.

23. Kimber, D., Dunnigan, T., Girgensohn, A., Shipman, F., Turner, T., and Yang, T. Trailblazing: Video playback control by direct object manipulation. In *Multimedia and Expo, 2007 IEEE International Conference on*, IEEE (2007), 1015–1018.

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, 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. Matejka, J., Grossman, T., and Fitzmaurice, G. Swifter: improved online video scrubbing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 1159–1168.

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. Monserrat, T.-J. K. P., Zhao, S., McGee, K., and Pandey, A. V. Notevideo: facilitating navigation of blackboard-style lecture videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 1139–1148.

38. 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.

39. Nguyen, C., Niu, Y., and Liu, F. Video summagator: an interface for video summarization and navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 647–650.

40. 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.

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

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

43. 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.

44. 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.

45. 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.

46. 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.

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

48. Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., and Schiele, B. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision 119*, 3 (2016), 346–373.

49. Soran, B., Farhadi, A., and Shapiro, L. Generating notifications for missing actions: Don't forget to turn the lights off! In *Proceedings of the IEEE International Conference on Computer Vision* (2015), 4669–4677.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

56. Yadav, K., Gandhi, A., Biswas, A., Shrivastava, K., Srivastava, S., and Deshmukh, O. Vizig: Anchor points based non-linear navigation and summarization in educational videos. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, ACM (2016), 407–418.

57. 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.