Crowdsourcing User Interactions with the Video Player

Konstantinos Chorianopoulos
Ionian University, Corfu, Greece
choko@acm.org

ABSTRACT

Every second millions of users enjoy content streaming on diverse video players (e.g., Web, Apps, social networks) and create billions of interactions within online video, such as *play*, *pause, seek/scrub*. This collective intelligence of video viewers might be leveraged into useful information for improved video navigation. For example, we can accurately detect and retrieve interesting scenes through the analysis of the aggregated users' *replay* interactions with the video player. Effective crowdsourcing of video interactions is grounded on previous work in multimedia, user modeling, and controlled user experiments. These research issues are described for the case of user-based detection of video thumbnails that stand for the semantics of the video. Moreover, we demonstrate the respective experimental environment with a focus on educational and user generated (e.g., how-to, lecture) videos.

Author Keywords

User-based, video, key-frame, replay, signal processing.

ACM Classification Keywords

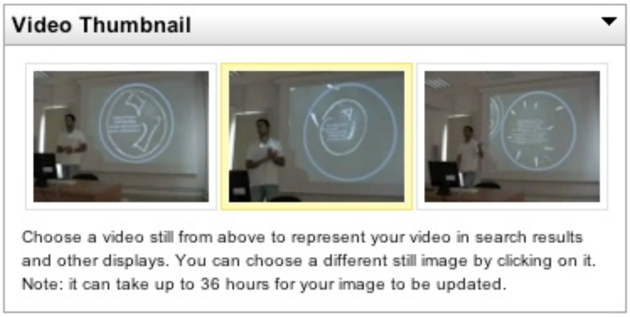
H.5.1 Multimedia Information Systems

INTRODUCTION

There is a growing research interest in user-based video modeling approaches of video interactions on the Web. The research interest has been largely motivated by the popularity of video with Web users, who have become familiar with video content, as an integral part of the Web experience in Web sites, smartphone and tablet Apps, and social networks. Notably, video search results and suggested videos are represented with a video thumbnail. In addition to the title and the short description of the video, the video thumbnail is an important information element that facilitates user navigation [2][3]. Moreover, video thumbnails have been useful as a summary and as a table of contents.

Previous multimedia research has considered content-based systems that have the benefit of analyzing a video without user interactions, but they are monolithic, because the resulting video thumbnails are the same regardless of the user preferences. Nevertheless, most of the existing content-based techniques that extract thumbnails at regular time intervals, or from each shot/scene are inefficient, because there might be too many shots in a video (e.g., how-to video), or rather few (e.g., lecture video).

In addition to the problem of detecting a set of descriptive video thumbnails for a given video, there is also the problem of selecting one of them for representing the video. For example, search results and suggested links in YouTube are represented with a thumbnail that the video authors have manually selected out of the three fixed ones (Figure 1). This approach puts too much trust on the video thumbnails selected by the video author or uploader. Besides the threat of authors tricking the system, the author-based approach does not consider the variability of users' knowledge and preferences. Thus, there is a need for selecting video thumbnails according to the collective wisdom of video viewers.



Video Thumbnail

Choose a video still from above to represent your video in search results and other displays. You can choose a different still image by clicking on it. Note: it can take up to 36 hours for your image to be updated.

Figure 1 The YouTube upload tool asks the user to manually select a video thumbnail, which has been randomly generated.

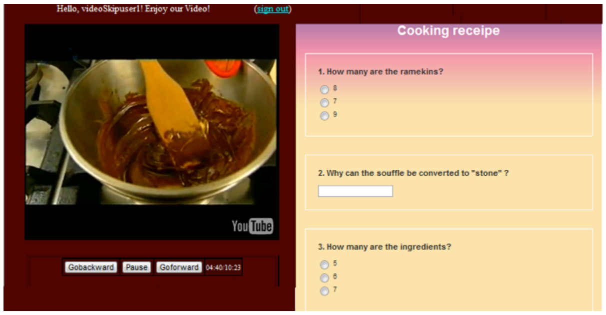
Previous user-based research on web video has focused on the meaning of the comments, tags, re-mixes, and micro-blogs, but has not examined simple user interactions with a web-based video player. In the seminal user-based approach to web video, Shaw and Davis [7] proposed that video representation might be better modeled after the actual use made by the users and they analyzed user-comments in order to understand media semantics. Peng et al. [5] have examined the physiological behavior (eye and head movement) of video users, in order to identify interesting key-frames, but this approach is not practical because it assumes that a video camera should be available and turned-on in the privacy-sensitive home environment. Shamma et al. [6] have created summaries of broadcasts (sports and political debate respectively) by analyzing the twitter stream of the respective real-time event. Although comments and tweets are very rich in meaning, they lack the real-time accuracy that is required in the generation of user-based video thumbnails, which we describe in the next sections.

WebMedia'12, October 15–18, 2012, São Paulo/SP, Brazil.
Copyright 2012 ACM 978-1-4503-1706-1/12/10...$15.00.

USER-BASED VIDEO THUMBNAIL DETECTION

In order to prove that video interactions stand for interesting video segments we have devised an experimental methodology. The evaluation methodology consists of the following parts: 1) customized web video player and cloud-based server that logs user interactions, 2) questions that can be only answered by retrieving information from particular video segments, 3) controlled experiment that produces a clean video interaction data-set, 4) data analysis for automatically generating video thumbnail. For the first couple of parts, the SocialSkip [1] video interaction logging and questionnaire system has been employed.

The experimental web video player (Figure 2, left part) employs few buttons. There is the familiar pause/play button, but instead of the common video seek bar timeline, we used two fixed-seek buttons. The GoBackward goes backward 30 seconds and its main purpose is to replay interesting parts of the video, while the Goforward button jumps forward 30 seconds and its main purpose is to skip insignificant video segments. We did not use a random seek timeline because it would be difficult to analyze users' interactions. Moreover, Li et al. [4] observed that when seek thumb is used heavily, users had to make many attempts to find the desirable section of the video and thus caused significant delays. Next to the buttons, there is the current cue-time and the total time of the video in seconds.


Hello, videoSkipuser1! Enjoy our Video! (sign out)

Cooking receipe

1. How many are the ramekins?
○ 8
○ 7
○ 9

2. Why can the souffle be converted to "stone" ?

3. How many are the ingredients?
○ 5
○ 6
○ 7

Figure 2 The experimental web video player includes skipping buttons and questionnaire functionality.

A user-centered approach identifies salient signals from human behavior, and not signals present in the video content itself. For the lecture and how-to videos we focus on, the video content tends to be either very static (usually a speaker at a podium), or very dynamic (multiple moving cameras provide alternative views of the same object and/or activity). Thus, signal processing of video content tends to fail, because it produces few or too many key-frames respectively.

In order to experimentally replicate user interest, we created an electronic questionnaire (Figure 2, right part) that corresponds to a few manually selected video segments, which stand for the semantics of the respective video. Indeed, according to Yu et al. [8] there are segments of a video clip that are commonly interesting to most users, and users might browse the respective parts of the video clip in searching for answers to some interesting questions. This assumption might be especially valid in the case of informational videos (e.g., lectures, how-to), when users seek to retrieve important information. The questions were relatively simple to answer, and did not depend on any previous knowledge, besides the information available within the video itself (Table 1). Therefore, the users had to seek/scrub through the video in order to answer those questions.

Video	**Indicative questions**
Lecture A	• Which are the main research topics? • What the students did not like? • What time does the first part of the talk end?
How-to B	• How many are the ramekins? • How many are the ingredients? • Which is the right order for mixing the ingredients?

Table 1 Example questions from each video. The questions are not supposed to be meaningful, but to direct the users towards a video segment.

The experiment took place in a lab with Internet connection, general-purpose computers, and headphones. Twenty-three university students (18-35 years old, 13 women and 10 men) spent approximately ten minutes to watch each video (buttons were muted). All students had been attending the Human-Computer Interaction courses at a post- or under-graduate level and received course credit in the respective courses. Next, the questionnaire appeared and there was a time restriction of five minutes, in order to motivate the users to actively browse through the video and answer the respective questions. We did not directly encourage the users to actively seek, but we informed the users that the purpose of the study was to measure their performance in finding the answers to the questions within time constraints.

Our goal is to examine whether video interactions are similar to the video semantics. For this purpose, we modeled the user interactions as a time series function and we compared it to the manually selected video semantics. For example, the *replay* function is based on the aggregation (crowdsourcing) of all *replay* interactions along the video cue time. In particular, we modified the value of the time series by one, depending on the type of interaction (e.g., for each GoBackward, we increased the value of the previous 30 cells). Next, we smoothed the signal and we calculated the local maximums. Finally, we constructed the corresponding semantic time series (pulse-like), which models the video semantics. In the next section, we present a comparison between user interactions and video semantics, in order to demonstrate that there is a close match between them.

USER INTERACTIONS STAND FOR VIDEO SEMANTICS

We observed that the interaction counts are similar between the different types of videos, because user behavior was rather motivated by the controlled conditions of the experiment than their own preferences about the content types. Nevertheless, we expect that in a natural setting (e.g., data mining of video logs from a web site) there will be variability depending on the type of the video content. The most popular button was the *skip-forward*, because the users were under time pressure to retrieve the required information in order to answer the questions.

Next, we compared the smooth versions of the *replay* and *skip-forward* time-series to the semantic ones. We observed that the *replay* closely matched the semantics of the videos. Therefore, we computed the local maximums of the *replay* time series for each one of the videos. The video segments with peaks are most likely to attract the viewers' interest. In order to determine the precise position of the peak, a derivative curve is computed. The zero-crossing points from positive to negative on derivative curve are the locations of the peaks. In this way, all key-frames in a video sequence can be identified without the need of any content-based detection. In the following figures, the user time series are plotted with the solid red curve and the experimentally defined ground truths are plotted with the pulse-like solid blue line.

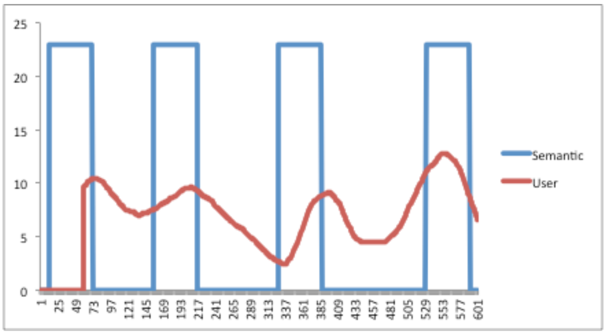


Figure 3 The user interactions function is similar to the semantic one for the lecture video

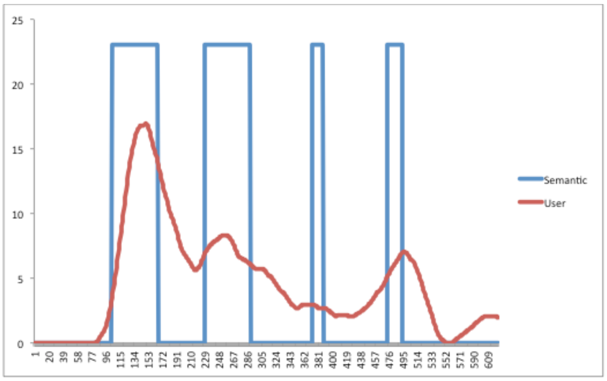


Figure 4 The user interactions function is similar to the semantic one for the how-to video

Although there is an obvious visual matching between the semantics and the user interaction signal, we have also calculated the user interaction signal peaks and the distance from the semantics one. In particular, the local maximum values of the user interaction signal can be used as the rank of the respective key-frame. Based on such a measure, it is convenient to generate a ranking of importance of key-frames. Then, the global maximum of the user interaction value in a video could be used as its representative video thumbnail (Table 2).

Semantics/	Lecture A	How-to B
Video(secs)	(81 interactions)	(68 interactions)
Semantic 1	33 (40) [10]	45 (105) **[16]**
Semantic 2	13 (145) [10]	21 (230) [8]
Semantic 3	48 (350) [9]	-13 (374) [3]
Semantic 4	1 (554) **[13]**	21 (475) [7]

Table 2 We have calculated the distance of the replay maximum from the semantic start (inside parentheses) and we provide in bold the user activity peak values

Finally, we found that a simple heuristic could automatically generate video thumbnails that are positioned at the start of each interesting video segment. In order to calculate this heuristic we observed that in all cases the distance of the local maximum of the *replay* time series from the start of the respective ground truth is less than 60 seconds. This simple heuristic detects 100% of the interesting video segments (n=8). There is only one case that the local maximum is before the start of the interesting video segment (how-to video, S3). Therefore, we suggest that the position of user-based video thumbnails can be automatically generated for any video by locating the local maximums of the *replay* time series and then selecting the one with the greatest value (Table 2).

In summary, the central contribution of this work is a novel conceptualization of video data-logs that holds the following unique properties: 1) implicit from people's action, 2) video signal/content-free, 3) adaptive based on consumption. The proposed heuristic ("sixty-seconds from local maxima of user activity") explains a methodology to support our core contribution and might hold different values depending on the video and on the distribution of video interactions. Moreover, we present a reproducible method that is verified and explained via qualitative and quantitative sources. Reimplementation of this system would only require a properly instrumented video player.

DISCUSSION AND FURTHER RESEARCH

The concept of crowdsourcing video interactions could be applied to any video, in order to generate user-based and dynamic video thumbnails in web search results, in Apps, and in social networks. In this work, instead of mining real video interaction data, we have devised a controlled experiment, because it provides a clean set of data that might be easier to model and understand. We focused on videos that are as much visually unstructured as possible (e.g., lecture, how-to), because content-based algorithms have already been successful with those videos that have visually structured scene changes (e.g., movies, series).

The majority of previous approaches employed content-based (e.g., detection of object, shot, and scene change) or explicit user-based methods (e.g., comments, tags, re-mix) to improve users' watching and browsing experience. In contrast, we suggest recording users' interactions with video player buttons. In terms of the user activity data, the most relevant work is the Audience Retention tool, which is part of the YouTube Analytics video account. The Audience Retention tool is employing the same set of data as suggested here, but there is no open documentation on the technique employed to map user interactions to a graph and only the video author has access to the data.

Previous work on content-based information retrieval from videos has emphasized the *number of videos* employed in similar experiments, because the respective algorithms treated the content of those videos. In this user-based work, we are not concerned with the content of the videos, but with the user activity *within* a video, thus the granule of analysis is the *number of video interactions*. Nevertheless, it is worthwhile to explore the effect of more videos and interaction types. Therefore, the small number of videos used in the preliminary studies is not an important limitation, but further research has to elaborate on different genres of video (e.g., news, sports, comedy). Notably, video interactions on educational videos might be analyzed in order to understand learning patterns of students.

Our main interest has been with lecture videos for two reasons: 1) they lack any meaningful visual structure that might have been helpful in the case of a content-based system, and 2) they contain lots of audio-visual (verbal and non-verbal) information that a user might actively seek to retrieve. In addition to video lecture, we employed a how-to (cooking) video because it has a rather complicated and active visual structure, which might have created too many false positives for a content-based approach. Moreover, the lecture and how-to video genres represent a growing and useful category of online videos, which facilitate life-long learning in many diverse topics, both theoretical and practical.

Although the *replay* user activity seems suitable for modeling user interest and video semantics, further research should consider the rest of the implicit user activities with the video player. For example, a *pause* might signify an important moment, but a pause that is too long might mean that the user is away. Another direction for further research would be to perform data mining on a large-scale database of video interactions. Nevertheless, we found that the experimental approach is very flexible during the development phase of a system, because it is straightforward to associate user behavior to the respective video interaction logs. Thus, this technique could be applied to novel video players with domain-specific user interfaces.

Finally, we suggest that user-based content analysis has the benefits of continuously adapting to evolving users' preferences, as well as providing additional opportunities for the personalization of content. For example, researchers might be able to apply several personalization techniques, such as collaborative filtering, to the user activity data. Moreover, further research should combine user-based methods in order to link user-comments to finely timed video interactions. In this way, crowdsourcing of video interactions is emerging as a new playing field for improving user experience with online video.

ACKNOWLEDGMENTS

The work reported in this paper has been partly supported by a Marie Curie European Reintegration Grants (CULT (MC-ERG-2008-230894, http://cult.di.ionio.gr) fellowship.

REFERENCES

1. Konstantinos Chorianopoulos, Ioannis Leftheriotis, and Chryssoula Gkonela. 2011. SocialSkip: pragmatic understanding within web video. In Proceddings of EuroITV '11, pages 25-28.
2. Marc Davis. 1995. Media Streams: an iconic visual language for video representation. In Human-computer interaction, Ronald M. Baecker et al. (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
3. Girgensohn, A. Boreczky, J., Wilcox, L, (2001). Keyframe-based user interfaces for digital video, Computer, 34(9):61–67
4. F.C. Li, A. Gupta, E. Sanocki, L.-wei He, and Y. Rui, "Browsing digital video," Proceedings CHI '00, vol. 2, 2000, pp. 169-176.
5. Peng, W.-T., Chu, W.-T., Chang, C.-H., Chou, C.-N., Huang, W.-J., Chang, W.-Y., and Hung, Y.-P. (2011). Editing by viewing: Automatic home video summarization by viewing behavior analysis. Multimedia, IEEE Transactions on, 13(3):539-550.
6. David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. 2009. Tweet the debates: understanding community annotation of uncollected sources. In Proceedings of WSM '09. ACM, 3-10.
7. Ryan Shaw and Marc Davis. 2005. Toward emergent representations for video. In Proceedings of MULTIMEDIA '05, 431-434.
8. Bin Yu, Wei-Ying Ma, Klara Nahrstedt, and Hong-Jiang Zhang. 2003. Video summarization based on user log enhanced link analysis. In Proceedings MULTIMEDIA '03. ACM, 382-391.