FrameRank: Implicit User-based Video-frame Detection

First Author Name (Blank if Blind Review)
Affiliation (Blank if Blind Review)
e-mail address (Blank if Blind Review)

Second Author Name (Blank if Blind Review)
Affiliation (Blank if Blind Review)
e-mail address (Blank if Blind Review)

ABSTRACT

We present a method for user-based detection and ranking of important video key-frames. Instead of content-based analysis that detects object, shot, and scene changes, we analyze aggregate user interactions (e.g., pause, seek/scrub) within a web video. Moreover, we validated the proposed method in a controlled lab experiment with lecture videos that content-based approaches cannot structure meaningfully. In particular, we modeled the collective information seeking behavior as a time series of user interest. We assumed that replay of a video segment stands for increased user interest in that segment and skip stands for less interesting parts. We found that only the replay time series matches significantly well the semantics of the lecture video. In practice, user-based detection of interesting video frames might improve navigation within information-rich but visually unstructured videos (e.g., lectures) on the web and has also the potential to improve video search results with personalized video thumbnails.

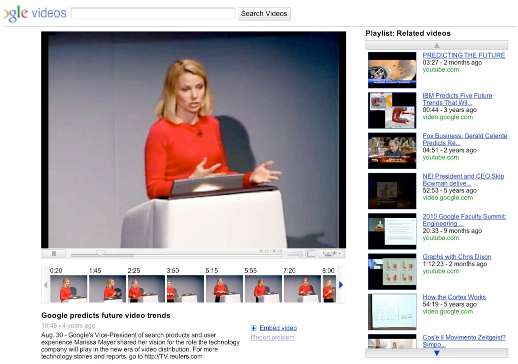
Author Keywords
User-based, implicit, video, key-frame.

ACM Classification Keywords
H.5.1 Multimedia Information Systems

General Terms
Human Factors, Design, Experimentation, Algorithms

INTRODUCTION

Web video players (e.g., Google Video in June 2011) provide thumbnails to facilitate user's navigation within a video (Figure 1, horizontal series of thumbnails) and between related videos (Figure 1, vertical series of thumbnails). 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). For example, in Figure 1, there are so many thumbnails that a separate scroll bar has been employed for navigating through them. Moreover, low-level features often fail to capture the high-level semantics of the video content itself, yet such semantics are often what guides users (Crockford and Agius 2006). By analogy to the early web-text search engines that were only based on the frequencies of search keywords in web pages, current navigation within video streams depends too much on the actual content of the video. Thus, there is a need to apply an implicit user-based approach for structuring the content within a web video.



Google predicts future video trends
16:46 - 4 years ago
Aug. 30 - Google's Vice-President of search products and user experience Marissa Mayer shared her vision for the role the technology company will play in the new era of video distribution. For more technology stories and reports, go to http://TV.reuters.com

PREDICTING THE FUTURE
03:27 - 2 months ago
youtube.com

IBM Predicts Five Future Trends That Wil...
00:44 - 3 years ago
video.google.com

Fox Business: Gerald Celente Predicts Re...
04:51 - 2 years ago
youtube.com

NEI President and CEO Skip Bowman delive...
52:53 - 5 years ago
video.google.com

2010 Google Faculty Summit: Engineering ...
20:33 - 9 months ago
youtube.com

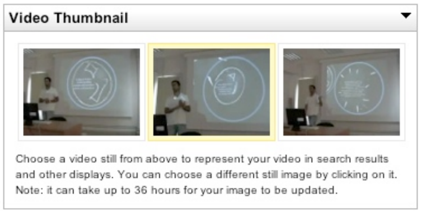
Graphs with Chris Dixon
1:12:23 - 2 months ago
youtube.com

How the Cortex Works
54:19 - 5 years ago
video.google.com

Cos'è il Movimento Zeitgeist? Simpo...

Figure 1 Video-frames are an important part of user navigation both within and between videos.

Notably, search results and suggested links in YouTube [6] are represented with a thumbnail that the video authors have manually selected out of the three fixed ones (Figure 2). By analogy to the early web-text search engines that were based on author declaration of important keywords, the current video search engine approach puts too much trust on the frames selected by the video author. Besides the threat of authors tricking the system, the author-based approach does not consider the variability of users' knowledge and preferences, as well as the comparative ranking to the rest of the video frames within a video. Thus, there is a need for ranking video-frames according to the collective action of video viewers, in order to reveal important video segments.



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 2 YouTube upload tool asks the user to manually select a video frame, which has been automatically generated

Related work on content-based video retrieval has contributed a standard set of procedures, tools, and data-sets

CHI 2012, May 5-10, 2012, Austin, TX, USA.

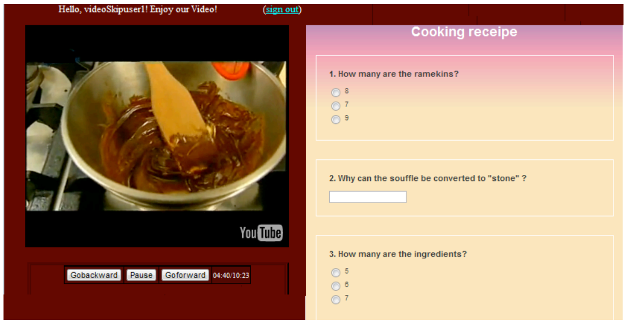
for comparing the performance of video retrieval algorithms (e.g., TRECVID [9]). Content-based work has established the importance of video key-frames as an important navigation mechanism and a summary of the video, either with thumbnails, or with video-skims. The optimum number of key-frames depends on several parameters, such as the type and length of the video. Therefore, it is unlikely that there are a fixed number of key-frames that describes a particular category of videos (e.g., lectures). If the ideal number of key-frames is different for each video, then, besides the key-frame detection technique, we need a ranking measurement to select the most important of them. Thus, the aim of this research is to detect and to rank the importance of key-frames within a web video through a user-based approach that is not burdensome to the users.

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 (2005) proposed that video representation might be better modeled after the actual use made by the users. In this way, they have employed analysis of the annotations (Shaw and Davis 2005) to understand media semantics. Peng et al. (2011) 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 home environment. Olsen and Moon (2011) have proposed a degree of interest (DOI) function, but it requires training by humans who rate their interest in a scene on a scale from 1-10. In summary, although there are various methods that collect and manipulate user-based data, the majority of them are considered burdensome for the users, because they require extra effort. In contrast, the proposed FrameRank method is only based on implicit user interactions, such as seek/scrub.

METHODOLOGY

The evaluation methodology of FrameRank consists of: 1) customized web video player that logs user interactions, 2) manually selected video segments of interest, 3) controlled experiment that produces a user interaction data-set, and 4) modeling and comparison of aggregated users' interactions and manually selected video segments.

The experimental web video player (Figure 3, left part) employs few buttons, in order to facilitate the association of user actions to user interest. There is the familiar pause/play button, but instead of the common video seek bar timeline, we employed 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. Next to the player's button there is the current cue-time and the total time of the video in seconds. Although we did not have a seek bar, we suggest that the data collected from the fixed skip could simulate the use of random seek, because any random seek activity can be modeled as a factor of fixed skipping actions (e.g., a random seek of 180 seconds is equal to 6 skips of 30 seconds). We did not use a random seek timeline because it would be difficult to analyze users' interactions. Moreover, the thirty-second step is a popular seek-step in previous research and commercial work. For example, we observed replay functions in Apple's iPhone and Safari QuickTime video players, which has the default time of 30 seconds as a replay. Finally, Li et al [8] 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.


Hello, videoSkipuser1! Enjoy our Video!

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

04:40/10:23

Figure 3 The experimental web video player includes skipping buttons and questionnaire functionality.

Instead of mining real video usage 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). In order to experimentally replicate user interest, we added an electronic questionnaire (Figure 3, right part) that corresponds to a few manually selected video segments (ground truth). According to Yu et al (2003) 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.

We chose to work with lecture videos and we selected a reference set of video segments (ground truths), which we represented with the respective questions. The questions were relatively simple to answer, and did not depend on any previous knowledge, besides the information available within the video itself (e.g., "Which are the main topics of this talk?", "What time does the first part of the talk end?", etc). Therefore, the users had to seek/scrub through the video in order to answer those questions. It is expected that in a future field study, when enough user data is available, user behavior will exhibit similar patterns even if they are not explicitly asked to answer questions. This assumption might be especially valid in the case of informational videos (e.g., lectures, how-to), when users seek to find important information.

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 the Department of Informatics (…) at a post- or under-graduate level and received course credit in the respective courses. Next, 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.

It is our main aim to examine whether the user interest and the ground truths are correlated, e.g., whether the patterns revealed from the user's activity are correlated with objective regions (ground truth) of interest for each video. In order to evaluate the FrameRank performance, we modeled the user interest in a video frame with a simple heuristic and we compared the observed user interest to the manually selected video segment (ground truth) that contained the answer to the question the users' were seeking for.

Firstly, we considered that every video is associated with an array of k cells, where k is the duration of the video in seconds. Next, we modified the value of each cell by one, depending on the type of interaction. For each GoBackward/GoForward, we increased/decreased the value of the previous/next 30 cells. A similar approach (i.e., activity graph, smoothing window, local maximum) to the construction of time series from micro-blogs (e.g., Twitter) has been followed by a growing number of researchers (e.g., see citations to Shamma et al. 2009). Finally, we construct the corresponding pulse time series, which models the regions of interest of a video. The exact locations of the pulses are defined as the center of the corresponding regions of interest as defined initially from the experimental setting.

In summary, the following methodology is used: 1) smoothness procedure, 2) pulse construction at local maximums, 3) construction of reference pulses, and 4) determination of correlation between pulse signals.

RESULTS

An exploratory analysis of the time series revealed that the GoBackward button signal has a regular pattern that matches the ground truths, but the GoForward button signal is characterized by a large number of abnormal local maxima.

A	Reference
S1	43
S2	150
S3	187
S4	358
S5	561

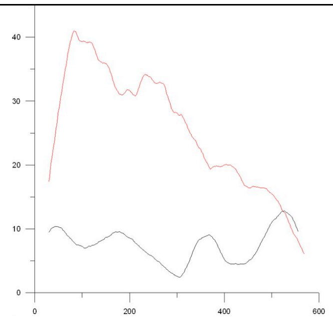


Figure 4 The GoBackward time series (black line) matches the ground truth better than the GoForward one (red line)

According to the definition of the user interest model for the GoBackward type of user interaction, the video segments with peaks are most likely to attract the viewers' interest. Therefore, it is reasonable to assume that key-frames should be extracted from these peaks. 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 order to produce the final user-based time series we explored alternative values for the smoothing window, which are reported alongside each of the videos. Moreover, we noticed that the width of the reference pulse depends on the size of the smoothing window. In this way, all key-frames and their ranking in a video sequence can be identified without the need of any content-based detection.

A	Reference	User
S1	43	44
S2	150	
S3	187	178
S4	358	365
S5	561	530

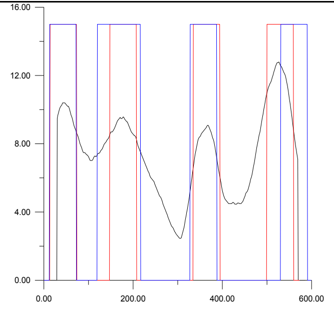


Figure 5 Video A is a lecture video. The pulse width D is 60 seconds and the smoothing window T is 60 seconds (r=0.67)

The user time series are plotted with the solid black curve. The user pulse signals were extracted from the corresponding local maxima are depicted with the red discontinued pulse signal while the reference signals are depicted with the blue solid pulse. The pulse modeling is reported with respect to the center of each pulse. Although the correlation of the constructed pulse signals for each video is visually evident in the graphs, the cross correlation coefficient was used in order to establish the respective quantitative measures. Indeed, the cross correlation coefficients that we estimated were 0.67, and 0.76 correspondingly, indicating strong correlation between the two signals (reference and user signal). The pulse modeling process has identified the majority of the manually selected video scenes with high accuracy (only one scene was not detected).

B	Reference	User
S1	80	71
S2	181	175
S3	269	281
S4	386	396
S5	528	524

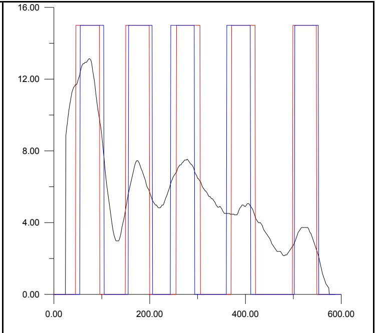


Figure 6 Video B is a lecture video. The pulse width D is 50 seconds and the smoothing window T is 50 seconds (r=0.76).

The user interest value of a key-frame can be used as the FrameRank of the key-frame. Based on such a measure, it is convenient to generate a ranking of importance of key-frames. Then, the maximum user interest value of the key-frames in a video could be used as its representative video thumbnail.

CONCLUSION AND FURTHER RESEARCH

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. The FrameRank system explores the application of an implicit user-based technique. We simply record users' interactions with video player buttons. In terms of the user activity data, the most relevant work is the Hot-spots tool [10], which is part of the YouTube Insight video account. The Hot-spots 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. Moreover, Hot-spots has been designed as a tool for video authors, but FrameRank is proposed as a back-end tool that might improve navigation for all video viewers.

Finally, the controlled experiment of the FrameRank system provided a rather clean set of user activity data, which is not the case in the analysis of natural user interactions within web video. In future research, we aim to provide an updated FrameRank algorithm that includes filters for detecting "noise" from field data (e.g., a "pause" might signify an important moment, but a pause that is too long might mean that the user is away). Moreover, there are significant open research issues with video-thumbnails and video-skims: 1) the relative number of segments that are needed to describe a video, 2) the duration of video-skims, and 3) evaluation of user activity on thumbnails. The number of segments depends on several parameters, such as the type and length of the video. Therefore, it is unlikely that there are a fixed number of segments (or a fixed video skim duration) that describes a particular category of videos (e.g., lectures). In addition, the duration of each video skim should not be fixed, but should depend on the actual duration of user interest for a particular video segment. Finally, we expect that the attributes (number of key-frames, skim duration, smoothing window) of the FrameRank algorithm are dependent on the type and the length of the video.

ACKNOWLEDGMENTS

The work reported in this paper has been partly supported by project (…) of the European Commission (EC) under the 7th Framework Program (FP7).

REFERENCES

1. C. Crockford and H. Agius, "An empirical investigation into user navigation of digital video using the VCR-like control set," International Journal of Human-Computer Studies, 2006, pp. 340-355.
2. 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.
3. Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A user attention model for video summarization. In Proceedings of MULTIMEDIA '02. 533-542.
4. Dan R. Olsen and Brandon Moon. 2011. Video summarization based on user interaction. In Proceddings of EuroITV '11. ACM, 115-122.
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. Picking a video thumbnail (YouTube Help): http://www.google.com/support/youtube/bin/answer.py?answer=72431 (September 2011)
7. 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.
8. Ryan Shaw and Marc Davis. 2005. Toward emergent representations for video. In Proceedings of MULTIMEDIA '05, 431-434.
9. TRECVID: http://trecvid.nist.gov (August 2011)
10. Your YouTube video: Hot or Not? http://googleblog.blogspot.com/2008/09/your-youtube-video-hot-or-not.html (June 2011)
11. B. Yu, W.-Y. Ma, K. Nahrstedt, and H.-J. Zhang, "Video summarization based on user log enhanced link analysis," Proceedings of MULTIMEDIA '03, 2003, p. 382.