Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

Collective intelligence within web video

Konstantinos Chorianopoulos

Correspondence: choko@ionio.gr
Department of Informatics, Ionian University, 7 Tsirigoti square, Corfu 49100, Greece

Abstract

We present a user-based approach for detecting interesting video segments through simple signal processing of users' collective interactions with the video player (e.g., seek/scrub, play, pause). Previous research has focused on content-based systems that have the benefit of analyzing a video without user interactions, but they are monolithic, because the resulting key-frames are the same regardless of the user preferences. We developed the open-source SocialSkip system on a modular cloud-based architecture and analyzed hundreds of user interactions within difficult video genres (lecture, how-to, documentary) by modeling them as user interest time series. We found that the replaying activity is better than the skipping forward one in matching the semantics of a video, and that all interesting video segments can be found within a factor of two times the average user skipping step from the local maximums of the replay time series. The concept of simple signal processing of implicit user interactions within video could be applied to any type of Web video system (e.g., TV, desktop, tablet), in order to improve the user navigation experience with dynamic and personalized key-frames.

Keywords: Video; Web; User-based; Key-frame; Collective intelligence; Signal processing; Analytics

Introduction

In this research, we examine the benefits of Web video platforms for the simplest type of user interaction, such as pause/play, skip/scrub. The convergence of diverse video and TV systems toward Web-based technologies has transformed the static conceptualization of the viewer, from consumer of content, to active participant. For example, IP-based video has become a popular medium for creating, sharing, and active interaction with video [1-3]. At the same time, IP-based video streaming has become available through alternative channels (e.g., TV, desktop, mobile, tablet). In the above diverse, but technologically converged scenarios of use, the common denominator is the increased interactivity and control that the user has on the playback of the video. For example, the users are able to pause and, most notably, to seek forward and backward within a video, regardless of the transport channel (e.g., mobile, web, broadcast, IPTV). In this work, we suggest that user-based video thumbnails that dynamically summarize and visualize the structure of a video are beneficial for all Web-based TV systems.

Before the emergence of Web video and TV systems, content-based research has established the need for video thumbnails [4], video summaries [5], and the usefulness of automatic detection of key-frames for user navigation [6,7], but has not regarded the benefits of user-based approaches. In this work, we explore the modeling of user interest based on simple user interactions that are common to any Web video platform, such as play/pause, seek/scrub. 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 the web-based video player [8]. 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 an extra effort. Moreover, the percentage of users leaving a comment is rather small when compared to the real number of viewers [3]. In this research, we have implemented and empirically evaluated a system that leverages seamless user interactions for extracting useful information about a video. In particular, we let the viewer browse the video, we store all the interactions with the player (e.g. play, pause, seek), and we model them as a continuous signal, which we analyze with simple signal processing techniques, in order to automatically generate key-frames of interesting video segments.

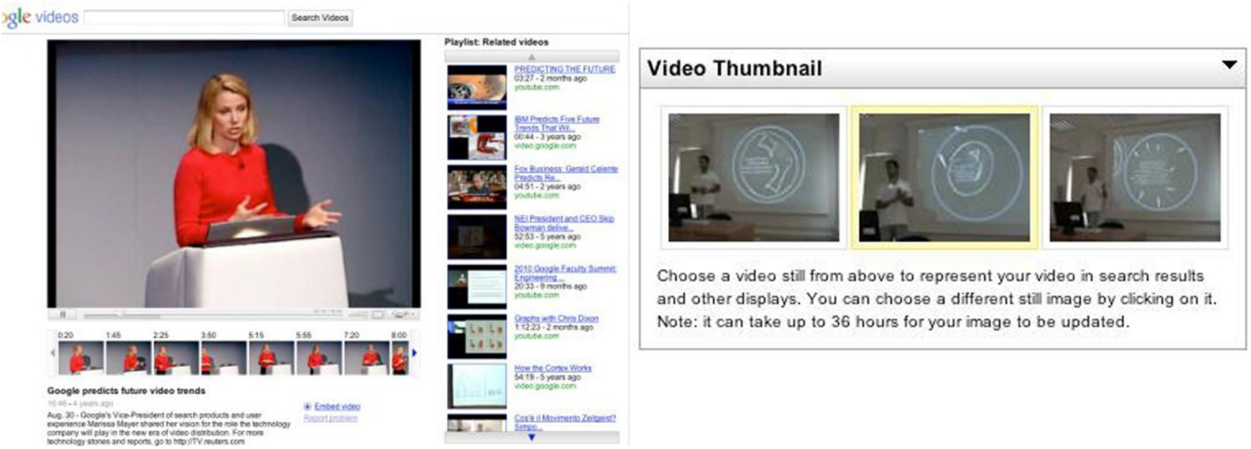
In the remaining of the paper, we examine the properties of the open source SocialSkip system, and we present the results of user-based key-frame extraction.

Related work

Previous research has explored several techniques in order to improve users' navigation experience. One of the major goals in multimedia information retrieval is to provide abstracts of videos. Abstraction techniques are a way for efficient and effective navigation in video clips [9]. Indeed, stationary images have proven an effective user interface in video editing, [10], as well as in video browsing [11]. According to Truong and Venkatesh [7] those techniques are classified in: 1) video skims, which provide moving images that stand for the important parts of the original video, and 2) key-frames, which provide stationary pictures of key moments from the original the video. According to Money and Agius [6], there is another interesting classification for video summarization techniques: 1) internal summarization techniques that analyse information sourced directly from the video stream, and 2) external ones that analyse information not sourced directly from the video stream. Notably, Money and Agius [6] suggest that the latter techniques hold the greatest potential for improving video summarization/abstraction, but there are rare examples of contextual and user-based works.

There are several research works on content-based key-frame extraction from videos, because a collection of still images is easier to deliver and comprehend when compared to a long video stream. Girgensohn et al. [12] found that clustering of similar colors between video scenes is an effective way to filter through a large number of key-frames. SmartSkip [13] is an interface that generates key-frames by analyzing the histogram of images every 10 seconds of the video and looking at rapid overall changes in the color and brightness. Li et al. [14] developed an interface that generates shot boundaries using a detection algorithm that identifies transitions between shots. Nevertheless, the techniques that extract thumbnails from each shot are not always efficient for a quick browse of video content, because there might be too many shots in a video (Figure 1, left).

In practical systems, web video players (e.g., Google Video, YouTube) provide thumbnails to facilitate user's navigation within a video and between related videos (Figure 1, left). Nevertheless, most of the existing content-based techniques that extract thumbnails at



Google predicts future video trends

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

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 Video-frames are an important part of user navigation within and between videos (left), but the research issue is that content-based techniques produce too many video thumbnails, which might not be representative, because they are selected by the video uploader (right).

regular time intervals, or from each shot are inefficient, because there might be too many shots in a video (e.g., how-to), or few (e.g., lecture). In the case of Google Video, there are so many thumbnails that a separate scroll bar has been employed for navigating through them. At the same time, search results and suggested links in popular video sites (e.g., YouTube) are represented with a thumbnail that the video authors have manually selected out of the three fixed ones (Figure 1, right). Moreover, by analogy to the early web-text search engines that were based on author definition 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 users (i.e., viewers), in order to reveal important video segments.

Previous research has already identified the benefits of user-based analysis of content (e.g., tags, comments, micro-blogs), but there is limited work on implicit indicators, such as seek/scrub within video. Social video interactions on web sites are very suitable for applying community intelligence techniques [15]. Levy [16] outlined the motivation and the social benefits of collective intelligence, but he did not provide particular technical solutions to his vision. In the seminal user-based approach to web video, Shaw and Davis [17] 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 [17], as well as of the re-use of video segments in community re-mixes and mash-ups [18] to understand media semantics. Nevertheless, the above approaches are not complete, because they are content-based, or they require increased user effort. Shamma et al. [19] has explored whether micro-blogs (e.g., Twitter) could structure a TV broadcast, but the timing of a micro-blog might not match the semantics of the respective cue point in a video, since there is common time duration for writing a short comment. Notably, Yew et al. [20] have recognized the importance of scrubs (fast forward and rewind), but they have only included counts in their classifier and not the actual timing of the scrub events. Thus, we propose to leverage implicit user activity (e.g., pause/play, seek/scrub), in order to dynamically identify video segments of interest.

In summary, content-based techniques, such as pattern recognition algorithms that focus on the contents of a video (e.g., detection of changes in shots, and scenes) are static. In contrast, the community (or crowd-sourced) intelligence of implicit user activity within web video is dynamic, because it continuously adapts to evolving users' preferences. In the following section, we describe the design and the implementation of a system that collects and analyses the collective intelligence of implicit user interactions within web video.

Design

SocialSkip is an open-source platform we developed to gather and analyze interactions of users while they browse a video. Based on these interactions, representative thumbnails of the video are dynamically generated, according to simple signal processing (Figure 2).

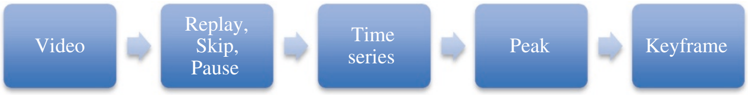


Figure 2 Overview of the framework for implicit user activity modeling.

Broadcast-, PC-, and web-based experimental systems

Researchers have developed various applications, in order to evaluate novel abstraction methods. Kim et al. [21] built a special-purpose system for their experimental environment. They wanted the subjects to believe that the content was being broadcast live. They used an interactive TV monitor, a TV encoder, a simulation server and an infrared remote control. Macromedia Director, a multimedia application platform, was used to develop SmartSkip [13]. The system was running on a desktop computer, it was connected to a television monitor, and a TV remote control was used by the participants for browsing. Crockford and Agius [22] designed a system as a wrapper around an ActiveX control of Windows Media Player. In summary, the majority of previous systems runs locally, needs special modification on software, and at the same time on video clips. Besides (broadcast and PC) stand-alone applications, there are few web-based systems. Fischlar [14] is a web-based system for capturing, storing, indexing and browsing broadcast TV material, but it only features content-based techniques. In the next sub-sections, we present a cloud-based system for user-based key-frame detection.

Cloud-based and open-source software architecture

In contrast to previous broadcast-, PC-, and Web-based systems, we used the Google App Engine (GAE) cloud platform (Platform-as-a-Service) and the YouTube Player API (YT API). At the time of writing, it is the first time that cloud-based technologies are used to build a system for key-frame extraction. SocialSkip (Figure 3) is a web application and has several advantages in contrast to stand-alone applications. Firstly, users do not have to go through an installation process, they just have to visit the link and if there is an updated version they just have to refresh the page. Secondly, the system architecture is modular and it allows re-use of the components. For example, a developer might decide to deploy a custom tablet video player and connect it to the cloud-based application logic, which tracks user activity and dynamically identifies interesting key-frames. Although we employed a simple Web-based video player, any player could connect to the application logic of the SocialSkip system.

There are several benefits of the selected tools (GAE, YouTube, Google accounts). GAE enables the development of web-based applications, as well as maintenance and administration of the traffic and the data storage. YT API allows developers to use the

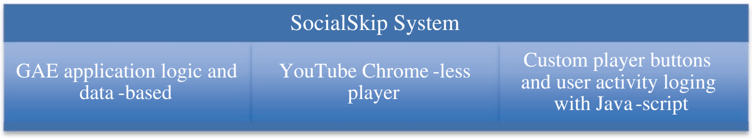

SocialSkip System
GAE application logic and data -based
YouTube Chrome -less player
Custom player buttons and user activity loging with Java -script

Figure 3 SocialSkip architecture is modular and cloud-based.

infrastructure of YouTube and therefore YouTube videos. In particular, the YT API provides a chrome-less user interface, which is a YouTube video player without any controls. This facilitates customization within Flash or HTML 5. In this way, we used JavaScript to create custom buttons and to implement their functions. Additionally, users of SocialSkip should have a Google account in order to sign in and watch the uploaded videos. In this way, we accomplish user authentication and we avoid the effort of implementing a user account system just for the application. Thus, users' interactions are recorded and stored in Google's database alongside with their Gmail addresses (Figure 4).

The Google App Engine database (Datastore) is used to store users' interactions. Each time a user signs in the web video player application, a new record is created. Whenever a button is pressed, an abbreviation of the button's name and the time it occurred are stored. This record includes four fields: a unique id, the username of the user's Google account, the date and a Text variable including all the interactions with the buttons of the web video player (Figure 5).

The SocialSkip video player (Figure 6) employs custom buttons, in order to be simple to associate user actions with video semantics. We have modified the classic forward and backward buttons to "GoBackward" and "GoForward." The first one jumps backwards 30 seconds and its main purpose is to replay the last 30 seconds of the video, while the GoForward button jumps forward 30 seconds and its main purpose is to skip insignificant video segments. Therefore, the player provides a subset of the main functionality of a typical VCR device [22]. We decided to use buttons that are similar to the main controls of VCR remote controls because they are familiar to users. Although we employed a fixed-step skip, we suggest that natural user interactions from any real system could be mapped to an average user skip-step. Thus, the assumption of the fixed-step skip should have external validity to field data, as well.

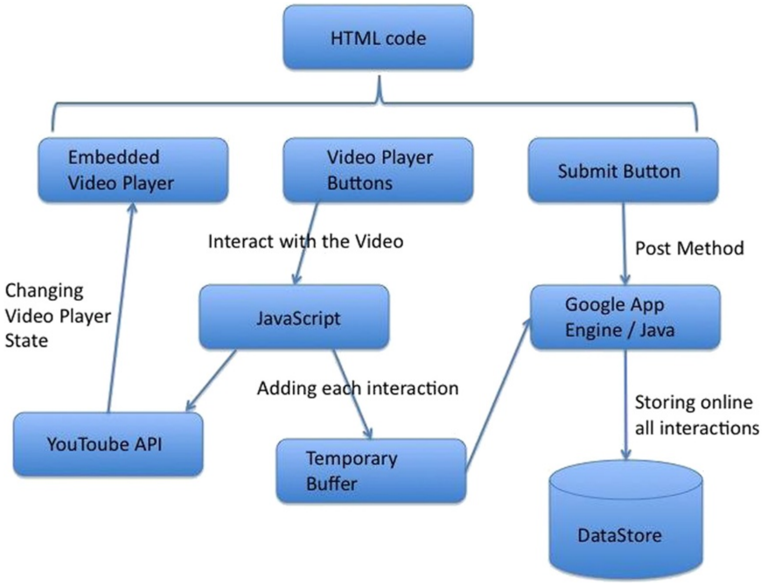
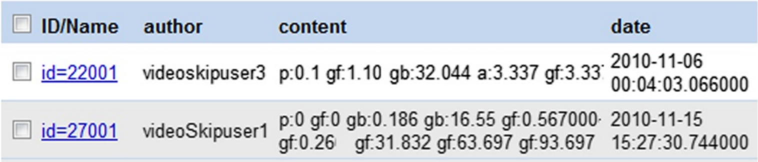


Figure 4 SocialSkip architectural components.



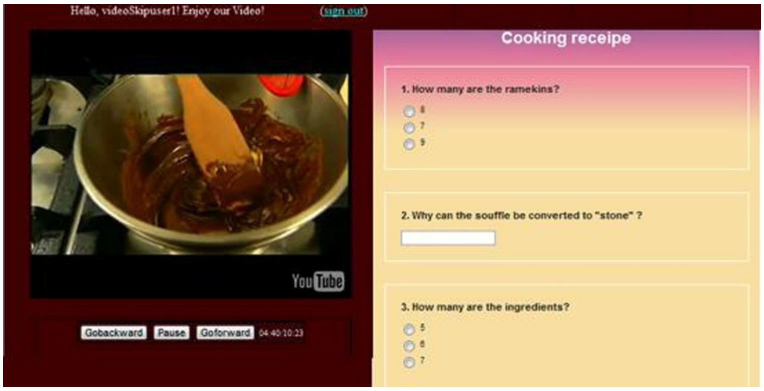
□ ID/Name	author	content	date
□ id=22001	videoskipuser3	p:0.1 gf:1.10 gb:32.044 a:3.337 gf:3.33	2010-11-06 00:04:03.066000
□ id=27001	videoSkipuser1	p:0 gf:0 gb:0.186 gb:16.55 gf:0.567000 gf:0.26 gf:31.832 gf:63.697 gf:93.697	2010-11-15 15:27:30.744000

Figure 5 **A screenshot of the records' table showing the id, the username (author), the content of the interactions and the date.**

User interest modeling

The transition from the implicit user activity log data to a time series requires some form of user interest modeling. Previous research in user interest modeling has explored several functions that map user behavior to levels of interest. In a content-based approach, Ma et al. [5] made the assumption that video viewers are visually attracted by faces and sudden camera motion, as well as by sudden sounds. Olsen and Moon [23] have proposed that the interest function might be related to user interactions. Most notably, Peng et al. [24] have developed a system that connects the interest function to actual user attention, as measured through eye tracking and face recognition. In summary (Table 1), the main drawback of previous related works has been: 1) the inherent difficulty of modeling either user interest according to video cue time, and/or 2) the lack of any common infrastructure available to all users across Web-based video systems. As a remedy, we propose user interest modeling based on implicit user interaction with the video player buttons, which is common along any of the Web-based video systems (TV, mobile, desktop, tablet).

In this work, we consider that every video is associated with an array of k cells, where k is the duration of the video in seconds. Initially, the array has zero values. Each time the user presses the GoBackward/GoForward button the cells' values matching the last/next thirty seconds of the video, are incremented/decreased by one (Figure 7). We make the assumption that the user replays a video segment either because there is something interesting, or because there is something difficult to understand, while the user skips forward a video segment because there is nothing of interest. In this way, an experimental time series is constructed for each button and for each video—a depiction of collective users' activity over time.



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 6 **The SocialSkip video player has familiar buttons, as well as a questionnaire functionality, in order to experimentally simulate the collective interest to particular video segments.**

Table 1 Previous user interest modeling research has established the significance of mapping user actions to video semantics, but there are drawbacks in all approaches

User interest modeling	Advantages	Disadvantages
Ma et al. [5]	Assumes that viewers are interested in particular well defined and easy to retrieve content features (e.g., faces)	Content-based and thus static vocabulary of what is interesting
Shaw and Davis [17]	User comments and tags	Do not have time information
Shaw and Davis [17]	Remix of popular video segments	Only a portion of users perform re-mixes of video
Shamma et al. [19]	Micro-blogs are associated to TV broadcast	The timing information might not correspond to video cue time
Carlier et al. [25]	Zoom denotes areas of interest within a video frame	Zoom is not a common feature
Olsen and Moon [23]	Interest function	Explicit ratings
Peng et al. [24]	Eye tracking and face recognition	Web camera

In order to extract pattern characteristics from each time series a key-frame detection scheme is developed based on the proposed user interest model. Figure 7 shows a flowchart of the proposed scheme. In this scheme, the component user interest models are first computed; then, a composite user interest time series is generated by linear combination. The user interest is composed of a time series of the interest values associated with each second in a video sequence. After smoothing, we can identify a number of local maximums. According to the definition of user interest model, 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 the area that is close to those local maximums. 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 [19]). Next, we have to compute the exact location of the proposed key-frame in comparison to an established ground truth. Notably, the interest value of a key-frame can be used as the importance measure of the key-frame. Based on such a measure the most highly ranked key-frames can be used as representative frames of a video in search results and lists of related videos, instead of the fixed ones (Figures 1,2).

Evaluation

The evaluation of a key-frame extraction and video summarization systems has been considered a very difficult problem, as long as user-based systems are concerned. Notably, Ma et al. [5] have argued that: "Although the issues of key-frame extraction and

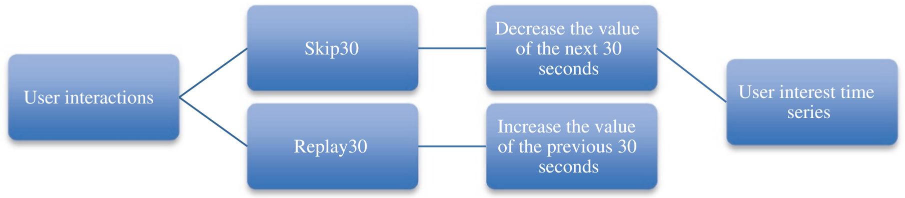

User interactions
Skip30
Replay30
Decrease the value of the next 30 seconds
Increase the value of the previous 30 seconds
User interest time series

Figure 7 Overview of the mapping between user interactions and user interest time series.

video summary have been intensively addressed, there is no standard method to evaluate algorithm performance. The assessment of the quality of a video summary is a strong subjective task. It is very difficult to do any programmatic or simulated comparison to obtain accurate evaluations, because such methods are not consistent with human perception." In content-based research (e.g., TRECVID), researchers have defined a set of ground-truths that are used as benchmarks during the evaluation of novel algorithms. In this work, we propose that the evaluation of user-based key-frame extraction systems could be transformed into an objective task as long as there is a set of ground truths about the content. In particular, we select videos that are relevant to the users and we ask the users to retrieve information from the video, in order to answer a set of questions in an experimental setting. In the following sub-sections, we are describing the selection of the videos, of the users, and of the questions.

Materials

We selected videos that are as much visually unstructured as possible, because content-based algorithms have already been successful with those videos that have visually structured scene changes. In particular, the lecture video included typical camera pans and zooms from speaker to projected slides, the documentary included a basic narrative and quick scene changes, and the how-to (cooking) video consisted of rapid changes of shots between the people and the cooking activity. In order to experimentally replicate user activity we developed a questionnaire that corresponds to several segments of each video. According to Yu et al. [26] 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. In this way, we can assume that during the experimental process the questions that the users are asked to answer stand for interesting topics and that the respective video segments are semantically interesting. In the field (e.g., YouTube), when enough user data is available, user behavior might exhibit similar patterns even if they are not explicitly asked to answer questions, at least for those videos that users browse for utilitarian purposes (e.g., lecture, how-to).

Our main interest is 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 has an active visual structure, which might have created too many false positives for a content-based approach. Finally, we employed a documentary video, which provides a baseline for evaluation with narrative-based videos. The questionnaire employed very simple questions that could not be answered by previous knowledge of the users Table 2.

Procedure

The goal of the user experiment is to collect activity data from the users, as well as to establish a flexible experimental procedure that can be replicated and validated by other researchers. There are several suggested approaches to the evaluation of interactive information retrieval systems [27]. Instead of mining real usage data, we have designed a controlled experiment, because it provides a clean set of data that might be easier to

Table 2 Example questions from each video

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?
Documentary B	What time do you see the message “coming next”?
	What is the purpose of hackers?
	What is the name of the girl in the video?
Cooking C	How many are the ramekins?
	How many are the ingredients?
	Which is the right order for mixing the ingredients?

analyze. 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 informed the users that the purpose of the study was to measure their performance in finding the answers to the questions within time constraints. After a basic understanding between the user behavior data and the key-frame detection is established, further research could progress to larger scale studies, or even to field studies and data mining of large data-sets.

Results

According to the proposed implicit user-based key-frame detection scheme (Figure 8), we created graphs that facilitated the visual comparison between the original user interest, the ground truth (interesting video segments), and smooth versions of the user interest time series (Figure 9). We explored alternative smoothing windows for each one of the three types of time series (Replay30, Skip30, Composite). We observed that in all cases the smoothing window should be at least 30 seconds (equal to the fixed skip-step), in order to provide a smooth signal with clear peaks. Moreover, we found that the ideal moving window ranges between 30, 45, and 60 seconds for each one of the three videos respectively (Documentary, Cooking, Lecture). Since we have a controlled experiment (equal number of users, time, total interactions), we suggest that the variability of the smoothing window might depend on the number and duration of the interesting video segments. We do not have conclusive results on this issue, because this was an unexpected finding that should be considered for elaboration in further research.

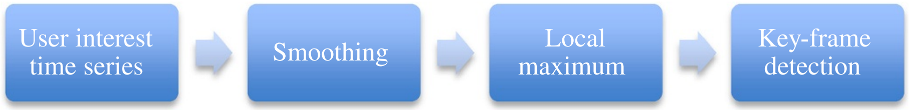


Figure 8 Key-frame detection through implicit user interest modeling.

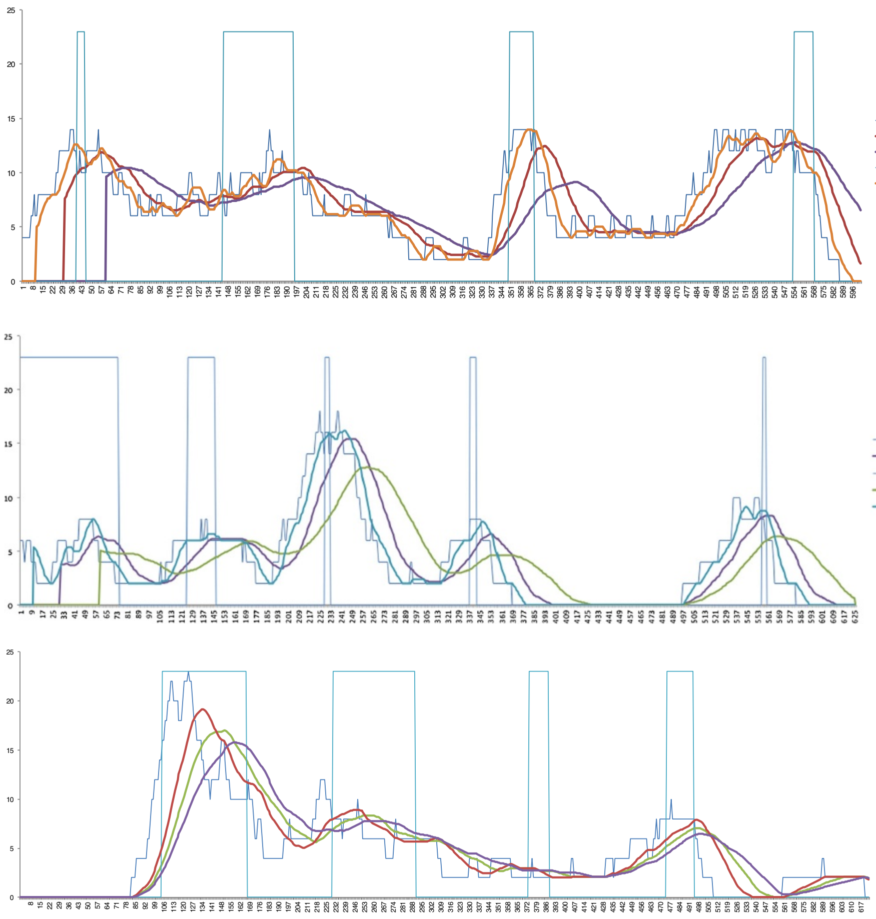


Figure 9 The original user interest time series (Replay30) is compared to alternative smoothed ones in order to select the most suitable moving window for further analysis. We observed that in all cases the smoothing window should be at least 30 seconds.

Next, we visually compared the smooth versions of the component and composite times series to the ground truth (Figure 10). We observed that in most cases the Replay30 time series closely matched the ground truth. Neither the Skip30, nor the composite time series seem to match the ground truth (Figure 10). Therefore, we computed the local maximums of the Replay30 time series for each one of the three videos.

Finally, we found that a simple heuristic could provide key-frames 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 Replay30 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 = 13). Moreover, we observed that approximately 70% (9 out of 13) of the interesting video segments are within 30 seconds and before the local maximum. There is only one case that the local maximum is before the start of the interesting video segment (Cooking video, S3). It is notable that the observed 60 seconds distance is twice the duration of the fixed skip-step. Therefore, we suggest that the position of interesting key-frames can be automatically approximated for any video with informational content by locating the local maximums of the Replay30 time series and then moving back twice the size of the (average) replay step (Table 3).

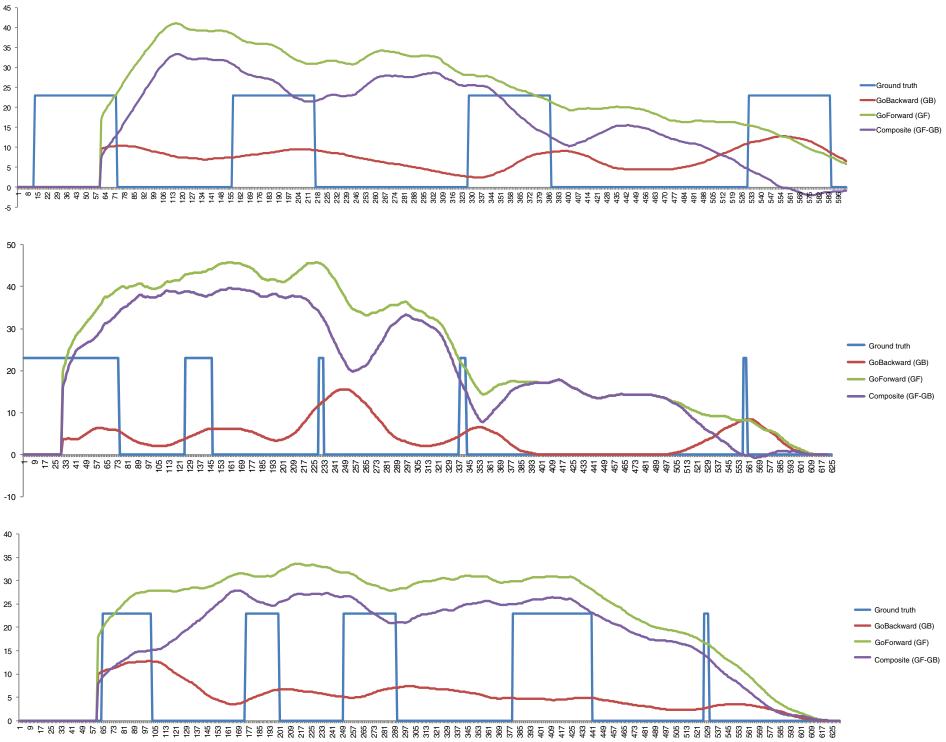


Figure 10 The comparison between the composite (Skip30-Replay30), the components Skip30, Replay30), and the reference time series reveals that the Replay30 time series is the most suitable for further analysis, because it closely matches the ground truth.

The experimental system also kept a log of the answers to the questions alongside the video interaction log, which was used to model the areas of interest. We considered separating the analysis of the user activity logs with correct answers of those with incorrect answers, but we realized that in many cases with incorrect answers the users did search for the answer at the right time of the video. Therefore, we decided that there is no reason to distinguish between correct and incorrect answers, because most users actively searched for interesting video segments.

Discussion

Key-frame detection system for research and practice

The open-source implementation of SocialSkip[a] is based on simple, modular, and well-established software components. SocialSkip is a cloud-based application, which uses cloud-based resources (bandwidth, processing, storage), open user-terminal software (any video streaming player), and videos provided by open video databases (e.g., YouTube). The SocialSkip architecture does not require any extra equipment beyond a computer and an internet connection. Previous efforts have introduced several applications in order to evaluate methods for understanding video content. The majority of related studies developed stand alone applications in order to avoid the elaborate installation, processing and streaming problems of broadcast systems. In terms of the user-based data, the most relevant work is the Hot-spots tool, 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

Table 3 The distance of the local maximum of the replay30 time series from the start of the respective pulse (inside parentheses) in the ground truth time series

Distance of local replay30 maximum (from ground truth start)	Lecture A	Documentary	Cooking
S1	33 (40)	58 (1)	45 (105)
S2	13 (145)	19 (126)	21 (230)
S3	48 (350)	16 (229)	−13 (374)
S4	1 (554)	15 (338)	21 (475)
S5		1 (557)	

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 SocialSkip is proposed as a back-end tool that might improve navigation for all video viewers. Most notably, researchers and practitioners have been cooperating for more than a decade on a large-scale video library and tools for analyzing the content of video. The TRECVID workshop series provides a standard-set of videos, tools, and benchmarks, which facilitate the incremental improvement of sense making for videos [28]. In similar way, we provide open access to both source code and the growing data-set of user interactions, which might facilitate further implementations, as well as alternative user-centric key-frame extraction algorithms.

Key-frame detection process through Implicit user-interest modeling

Although many corporations and academic institutions are making lecture videos and seminars available online, there have been few and scattered research efforts to understand and leverage actual user browsing behavior. He et al. [29] derive user activity (e.g., play, pause, random seek), but did not take advantage of them. Yu et al. [26] made the assumption that there is a shortest path in each video and evaluated user navigation among key-frames with link analysis. Syeda-Mahmood and Ponceleon [30] modeled implicit user activity according to the user’s sentiment (e.g., user is bored, or interested). In context of video editing in a studio environment [31], collective user behavior has been proven an effective way to understand and collaborate on video. The benefits of collective intelligence for web video have been noted by Carlier et al. [25], in the case of zoom-able video user interface. Yew et al. [20] have recognized the importance of scrubs (fast forward and rewind), but they have only included counts in their classifier and not the actual timing of the scrub events. Moreover, Martin and Holtzman [32] highlight the value of implicit interactions (views) on news items, but they did not explore this concept within a web video, in order to identify particular segments. Olsen and Moon [9] have devised a degree of interest (DOI) function for American football, which depends on the availability of different camera angles, on “plays”, and user ratings, but these features are not generic to all videos. Finally, Peng et al. [24] 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. In summary, SocialSkip proposes a very simple and generic approach that applies to any viewer and any video on Web-based TV systems.

Further research

In this work, we have focused on the design, development, and experimental evaluation of the system. Future work should consider the optimization of the key-frame-extraction algorithm and its adaptation to different users groups and video contents. For example, SocialSkip could also connect to other growing (lecture and how-to) video libraries, such as Vimeo, and khan academy.

Video key-frames provide an important navigation mechanism and a summary of the video, either with thumbnails, or with video-skims. There are significant open research issues with video-skims: 1) the number and relative importance of segments that are needed to describe a video, and 2) the duration of video-skims. 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). If the required number of segments is different for each video, then, besides the segment extraction technique, we need a ranking to select the most important of them. Moreover, 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.

Although the replay user activity seems suitable for modeling user interest, further research should consider the rest of the implicit user activities. We decided to ignore the "pause" interaction because, during the pilot tests, we noticed that the users paused the player to write down the answer to a question. Thus, the pause frequency distribution perfectly matched the ground truths, but this pattern might not have external validity. Nevertheless, in field data, 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 web-video database. Nevertheless, we suggest that the experimental approach might be more flexible than data mining for the development phase of the system. In particular, the incremental and experimental approach is very suitable for user-centric information retrieval, because it is feasible to connect user behavior with the respective data-logs. In contrast to data mining in large data-sets, a controlled experiment has the benefit of keeping a clean set of data that does not need several steps of frequency domain filtering, before it becomes usable for any kind of simple time-based signal processing.

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. In this way, video pragmatics is emerging as a new playing field for improving user experience.

Conclusion

We have developed an implicit user-based key-frame detection system and we have demonstrated that the collective intelligence of users' interactions with a familiar video player could be analyzed in order to generate user-based key-frames. Although we designed the SocialSkip system as a web-based one, the concept of mapping implicit user interactions to a time-series for further analysis has a much broader application. Every second millions of users enjoy video streaming on a diverse number of terminals

(TV, desktop, smart phone, tablets) and create billions of simple interactions. This amount of data might be converted into useful information for the benefit of all video users. As long as the community of users watching videos on Web-based video systems is growing, more and more interactions are going to be gathered and therefore, dynamic thumbnails would represent in a timely fashion the most important scenes of a video according to evolving user interests. We also expect that the combination of richer user profiles and content metadata provide opportunities for additional personalization of the thumbnails. Overall, our findings support the concept that we can learn a lot about an unstructured video just by analyzing how it is being used, instead of looking at the content item itself. In the end, we expect that a balanced mix of hybrid algorithms (content-based and user-based) might provide an optimal solution for navigating inside video content.

Endnotes

[a] Open-source project: http://code.google.com/p/socialskip/

Ethical approval

This research has been approved by Ionian University (Corfu, Greece) and it is in compliance with the Helsinki Declaration.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We are thankful to the participants of the user study, and to Markos Avlonitis, David Ayman Shamma, Ioannis Leftheriotis, and Chryssoula Gkonela for assisting in the implementation and evaluation of the system, as well as for providing feedback on early drafts of this paper. The work reported in this paper has been partly supported by project CULT (http://cult.di.ionio.gr). CULT (MC-ERG-2008-230894) is a Marie Curie project of the European Commission (EC) under the 7th Framework Program (FP7).

Received: 9 March 2013 Accepted: 4 June 2013
Published: 15 June 2013

References

1. Cha M, Kwak H, Rodriguez P, Ahn Y, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM Conference on internet Measurement (San Diego, California, USA, October 24–26, 2007). IMC '07. ACM, New York, NY, pp 1–14
2. Cheng X, Dale C, Liu J (2008) Statistics and social network of YouTube videos. In: Quality of service. IWQoS 2008. 16th International Workshop on, IEEE, pp 229–238
3. Mitra S, Mayank A, Amit Y, Niklas C, Derek E, Anirban M (2011) Characterizing Web-based video sharing workloads. ACM Trans Web 5(2):Article 8, May 2011
4. Davis M (1995) Media streams: an iconic visual language for video representation. In: Baecker RM, Jonathan G, Buxton WAS, Saul G (eds) Human-computer interaction. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp 854–866
5. Ma Y-F, Lu L, Zhang H-J, Li M (2002) A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on multimedia (MULTIMEDIA '02). ACM, New York, NY, USA, pp 533–542
6. Money AG, Agius H (2008) Video summarisation: a conceptual framework and survey of the state of the art. J Vis Comun Image Represent 19(2):121–143
7. Truong BT, Venkatesh S (2007) Video abstraction: A systematic review and classification. ACM Trans Multimedia Comput Commun Appl 3:1, Article 3 (February 2007)
8. Chorianopoulos K, Leftheriotis I, Gkonela C (2011) SocialSkip: pragmatic understanding within web video. In: Proceddings of the 9th international interactive conference on Interactive television (EuroITV '11). ACM, New York, NY, USA, pp 25–28
9. Lienhart R, Pfeiffer S, Effelsberg W (1997) Video abstracting. Commun ACM 40(12):54–62
10. Baecker R, Rosenthal AJ, Friedlander N, Smith E, Cohen A (1997) A multimedia system for authoring motion pictures. In: Proceedings of the fourth ACM international conference on Multimedia (MULTIMEDIA '96). ACM, New York, NY, USA, pp 31–42
11. Boreczky J, Girgensohn A, Golovchinsky G, Uchihashi S (2000) An interactive comic book presentation for exploring video. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI '00). ACM, New York, NY, USA, pp 185–192
12. Girgensohn A, Boreczky J, Wilcox L (2001) Keyframe-based user interfaces for digital video. Computer 34(9):61–67

13. Drucker SM, Glatzer A, De Mar S, Wong C (2002) SmartSkip: consumer level browsing and skipping of digital video content. In: Proceedings of the SIGCHI conference on human factors in computing systems: changing Our world, changing ourselves (Minneapolis, Minnesota, USA, April 20–25, 2002). CHI '02. ACM, New York, NY, pp 219–226
14. Li FC, Gupta A, Sanocki E, He L, Rui Y, Rui Y (2000) Proceedings of the SIGCHI conference on human factors in computing systems (the Hague, the Netherlands, April 01–06, 2000). CHI '00. In: Proceedings of the SIGCHI conference on human factors in computing systems (the Hague, the Netherlands, April 01–06, 2000). ACM, New York, NY, pp 169–176
15. Zhang D, Guo B, Yu Z (2011) The emergence of social and community intelligence. Computer 44(7):21–28
16. Levy P (1997) Collective intelligence: Mankind's emerging world in cyberspace. Perseus Publishing
17. Shaw R, Davis M (2005) Toward emergent representations for video. In: Proceedings of the 13th annual ACM international conference on multimedia (MULTIMEDIA '05). ACM, New York, NY, USA, pp 431–434
18. Shaw R, Schmitz P (2006) Community annotation and remix: a research platform and pilot deployment. In: Proceedings of the 1st ACM international workshop on human-centered multimedia (HCM '06). ACM, New York, NY, USA, pp 89–98
19. Shamma DA, Lyndon K, Elizabeth F (2009) Tweet the debates: understanding community annotation of uncollected sources, Proceedings of the first SIGMM workshop on Social media (WSM '09). ACM, New York, NY, USA, Churchill, pp 3–10. doi:10.1145/1631144.1631148
20. Yew J, Shamma DA, Churchill EF (2011) Knowing funny: genre perception and categorization in social video sharing. In: Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11). ACM, New York, NY, USA, pp 297–306
21. Kim J, Kim H, Park K (2006) Towards optimal navigation through video content on interactive TV. Interact Comput 18(4):723–746
22. Crockford C, Agius H (2006) An empirical investigation into user navigation of digital video using the VCR-like control set. Int J Hum-Comput Stud 64(4):340–355
23. Olsen DR, Moon B (2011) Video summarization based on user interaction. In: Proceddings of the 9th international interactive conference on interactive television (EuroITV '11). ACM, New York, NY, USA, pp 115–122
24. Peng W-T, Chu W-T, Chang C-H, Chou C-N, Huang W-J, Chang W-Y, Hung Y-P (2011) Editing by viewing: automatic home video summarization by viewing behavior analysis. Multimedia, IEEE Transactions on 13(3):539–550
25. Carlier A, Charvillat V, Ooi WT, Grigoras R, Morin G (2010) Crowdsourced automatic zoom and scroll for video retargeting. In: Proceedings of the international conference on multimedia (MM '10). ACM, New York, NY, USA, pp 201–210
26. Yu B, Ma W-Y, Nahrstedt K, Zhang H-J (2003) "Video summarization based on user log enhanced link analysis", Proceedings of the eleventh ACM international conference on Multimedia - MULTIMEDIA'03. ACM Press, New York, New York, USA, p 382
27. Kelly D (2009) Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval: 3(1–2):1–224
28. Snoek CGM, Worring M (2009) Concept-based video retrieval. Foundations and Trends in Information Retrieval 2 (4):215–322
29. He L, Sanocki E, Gupta A, Grudin J (1999) "Auto-summarization of audio-video presentations", Proceedings of the seventh ACM international conference on Multimedia (Part 1) - MULTIMEDIA'99. ACM Press, New York, New York, USA, pp 489–498
30. Syeda-Mahmood T, Ponceleon D (2001) "Learning video browsing behavior and its application in the generation of video previews", Proceedings of the ninth ACM international conference on Multimedia - MULTIMEDIA'01. ACM Press, New York, New York, USA, p 119
31. Cohen J, Withgott M, Piernot P (1999) Logjam: a tangible multi-person interface for video logging. In: Proceedings of the SIGCHI conference on human factors in computing systems: the CHI is the limit (CHI '99). ACM, New York, NY, USA, pp 128–135
32. Martin R, Holtzman H (2010) Newstream: a multi-device, cross-medium, and socially aware approach to news content. In: Proceedings of the 8th international interactive conference on interactive TV\&\#38; video (EuroITV '10). ACM, New York, NY, USA, pp 83–90

doi:10.1186/2192-1962-3-10
Cite this article as: Chorianopoulos: **Collective intelligence within web video.** *Human-centric Computing and Information Sciences* 2013 **3**:10.