Collective Intelligence in Video User's Activity

Ioannis Karydis, Markos Avlonitis, and Spyros Sioutas

Dept. of Informatics, Ionian University, 49100, Kerkyra, Greece
{karydis,avlon,sioutas}@ionio.gr

Abstract. In this work, we study collective intelligence behavior of Web users that share and watch video content. We propose that the aggregated users' video activity exhibits characteristic patterns that may be used in order to infer important video scenes thus leading to collective intelligence concerning the video content. In particular, we have utilised a controlled user experiment with information-rich videos for which users' interactions (e.g., pause, seek/scrub) have been gathered. Modeling the collective information seeking behavior by means of the corresponding probability distribution function we argue that bell-shaped reference patterns are shown to significantly correlate with the predefined scenes of interest for each video, as annotated by the users. In this way, the observed collective intelligence may be used to provide a video-segment ranking tool that detects the importance of video scene. In practice, the proposed techniques might improve navigation within videos on the web and have also the potential to improve video search results with personalised video thumbnails.

Keywords: Video, Event detection, Semantics, Web, User-based, Interaction, User activity, Signal processing

1 Introduction

The Web has become one of the prominent media for sharing and watching video content [1] and as the volume of available content increases rapidly [2] video retrieval has already become a very important issue [3]. The identification of salient features in the content of a video offers information that will subsequently be used for analysis, indexing and retrieval of videos based on their content. Though, despite providing important information for the purposes of video retrieval, content-based techniques do not take into consideration the video-viewing pattern of the user that also includes valuable contextual/semantic information [4].

The aforementioned domination of the Web as a means for streaming video-watching offers the unique opportunity of monitoring the user's interaction with the video-player and thus inducing new and useful information concerning the viewing-pattern of a user as well as the content of the video. The user–video-player interaction, for example the press of the play, pause or move backwards buttons, provides information on scene viewing which has been shown [5] to relate to the emotive energy of the scene and thus to a wealth of semantic information.

In our research, we aim in harnessing such video-viewing interactions in order to identify high semantic value video intervals that may subsequently be used in video scene selection for the purposes of representing the video through a thumbnail. We claim that there is collective behavior of users watching a specific video which can be detected from bell-like patterns emerging in the corresponding users' activity distribution, where the users' activity distribution (Figure 1) is constructed from the number of the interactions with the corresponding buttons, such as play or pause, of the video-player. Moreover we show that this collective behavior can be employed to infer the most important scenes of a video which can then be used to automatically generate thumbnails, or even implement a summarisation feature, thus leading to collective intelligence[6].

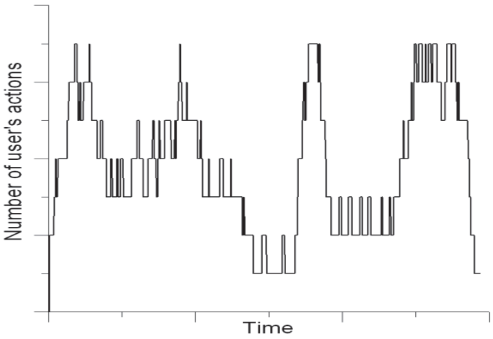

Fig. 1. The users' activity distribution: Video-player button clicks vs. time of click.

The rest of the paper is organised as follows. Section 2 describes related work while Section 3 presents two preliminarily approaches, a stochastic and a pattern matching, that detect patterns of collective behavior and reveal the behavior of the group exhibiting judgment on the importance of video scenes. Next, Section 4 details the setup of the experiment from which data where collected and Section 5 presents preliminarily results of the experiment conducted. Finally, the paper is concluded in Section 6.

2 Related Work

Research concerning video summarisation and, more generally, important scene selection in videos has mostly been based on content-based methodologies[1]. Nevertheless, as previously mentioned, such content-based methods often fail to capture high-level semantics that adhere to non-specialist users' navigation to videos [4].

In addition to video content, research has also been carried on the users' actions concerning their viewing and searching processes. Yu et al. [7] proposed

[1] Interested readers can refer to [3] for an extensive survey.

that users unintentionally show their understanding of the video content through their interaction with the viewing system. Their developed algorithm, *ShotRank*, is computed through a link analysis algorithm that utilises the voting of users on the subjective significance and "interestingness" of each shot. Moreover, in addition to user browsing log mining, *ShotRank* is also taking into consideration low-level content video analysis.

In their work [8], Syeda-Mahmood and Ponceleon, presented *MediaMiner*, a client-server-based media playing and data-mining system aiming at tracking video browsing behavior of users in order to generate fast video previews. In *MediaMiner*, users' interaction with video is recorded at the client side while gathered information is returned to the server for continuous learning and estimation of browsing states. Modeling users' states transition, while browsing through videos, is done with a Hidden Markov Model. *MediaMiner* features common video-browsing interaction buttons (e.g. play and pause) as well as random seek into the video via a slider bar, fast/slow forward and fast/slow backward.

Finally, Gkonela and Chorianopoulos [4], presented a user-centric approach, titled *VideoSkip*, wherein by analysis of implicit users interactions with a web video player (e.g. pause, play, thirty-seconds skip or rewind) semantic information about the events within a video are inferred. Using the simple heuristic concerning the local maxima identification on the accumulated information collected from user-activity, *VideoSkip* has been able to effectively detect the same video-events, as indicated by ground-truth manually annotated by the author of the videos.

Our work, in contrast to the hybrid solution proposed to [7], solely relies on user interaction with the player in order to identify high semantic value video intervals. As far as the work in [8] is concerned, our approach utilises a differentiated methodology than a Hidden Markov Model that does not necessarily require the assumption that the state of "interestingness" of a user is a function of the previous state of the user. Moreover, in contrast to *VideoSkip* presented in [4], our approach examines the information received from each button of the application separately, offering thus greater flexibility to the event identification.

3 User Activity Modeling

In order to extract pattern characteristics for each button distribution, i.e., scenes in which users exhibit high interaction with the video-player, the proposed methodology for user activity modeling and analysis consists of three distinct stages: (a) the smoothness procedure, (b) the determination of users' activity aggregates and (c) the estimation of pattern characteristics.

In the first stage, we use a simple procedure in order to average out user activity noise in the corresponding distribution. In the context of probability theory, noise removal can be treated with the notion of the moving average [9]: from a curve $S^{exp}(t)$ a new smoother curve $S_T^{exp}(t)$ may be obtained as shown in Equation 1,

$$S_T^{exp}(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} S^{exp}(t')dt' \tag{1}$$

where T denotes the averaging "window" in time. The larger the averaging window T, the smoother the curve will be. Schematically, the procedure is depicted in Figure 2. The procedure of noise removal of the experimentally recording distribution is of crucial importance for the following reasons: first, in order to reveal patterns of the corresponding signals (regions of high user's activity), and second in order to estimate local maxima of the corresponding patterns. It must be noted that the optimum size of the averaging window T is entirely defined from the variability of the initial signal. Indeed, T should be large enough in order to average out random fluctuations of the users' activities and small enough in order to avoid distortion of the bell-like localised shape of the users' signal which will in turn show the area of high user activity.

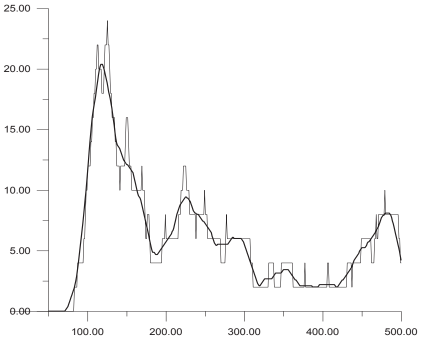

Fig. 2. The user's activity signal is approximated with a smooth signal: The y-axis shows the measured activity of the user while the x-axis shows the time in sec.

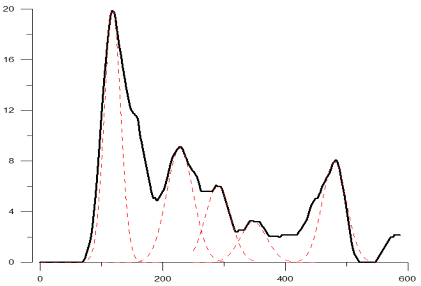

Fig. 3. The users' activity signal is approximated with Gaussian bells in the neighborhood of user activity local maxima: The y-axis shows the measured activity of the user while the x-axis shows the time in sec.

In the second stage, we estimate aggregates of users' activity by means of an arbitrary bell-like reference pattern. Accordingly, we propose that there is an aggregate of users' actions if within a specific time interval a bell-like shape of the distribution emerges in the sense that there is high probability that user's actions are concentrated at a specific time interval (the center of the bell) while this probability tends to zero quite symmetrically as we move away from this interval (Figure 3). Without loss of generality, the parameters of the width and height of the Gaussian function are set of the order of the averaging window and half of the number users' actions correspondingly.

In the third step, we estimate the pattern characteristics, i.e. we estimate the number of users' aggregates for the specific signal and moreover their exact locations in time by employing two different methodologies, a stochastic and a pattern matching:

- In the stochastic approach, the estimation of the exact locations can be done via the estimation of the generalised local maxima. By the term generalised local maxima we refer to the center of the corresponding bell-like area of the average signal, as the nature of the original signal under examination may cause more than one peaks at the top of the bell due to the micro-fluctuation. We claim that this is possible by estimating the well known correlation coefficient $r(x, y)$ between the two signals (time series), that is, the average experimental signal and the introduced aforementioned reference bell-like time signal.

  It should be noted that while the height of the reference bell-like pattern does not affect our results, the width of the bell D is a parameter that must be treated carefully. In particular, the variability of the average signal determines the order of the width D. Herein, we propose that the bell width should be equal to the average half of the widths of the bell-like regions of the signals. This estimation was found optimum in order to avoid overlap between different aggregates.
- In the pattern matching approach, the reference bell-shaped pattern is matched with the accumulated user interaction signal using a scaling and translation invariant distance measure (Equation 2), adopted from [10]. Accordingly, for two time series x and y, the distance $\hat{d}(x, y)$ between them is:

  $$\hat{d}(x, y) = min_{a,q} \frac{\|x - \alpha y_{(q)}\|}{\|x\|} \tag{2}$$

  where $y_{(q)}$ is the result of shifting the signal y by q time units, and $\| \cdot \|$ is the l_2 norm. In our case, and for simplicity, the shifting procedure is done by employing a window the size of which is empirically calculated to minimise the distance, while the scaling coefficient α is adjusted through the maximum signal value in the window context.

4 Experiment design

To explore the usefulness of the methodologies presented herein, we utilised the dataset collected in [4]. The goal of the user experiment was to collect activity data from the users but instead of mining real usage data, a controlled experiment was conducted as it provided a clean set of data that was easier to analyse.

The *VideoSkip* platform presentend therein, employs few buttons, in order to be simple in the association of a user's actions with video semantics. The common forward and backward buttons have been modified to *GoBackwards* and *GoForwards.*

GoBackwards 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 [11]. The selection of videos was based on their degree of visual structure, aiming at videos as much visually unstructured as possible (e.g., lecture, documentary), since content-based algorithms have already been successful with videos that have visually structured scene changes. In particular, *Video A* [12], which is a lecture video, includes typical camera pans and zooms from speaker to projected slides, while *Video B* [13], a documentary, includes a basic narrative and quick scene changes. In order to experimentally replicate users' activity, the experiment designers developed a questionnaire that draws questions from several segments of each video. According to Yu et al. [7] 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. Thus, the intuitive assumption of using of these videos in the field (e.g., YouTube) is that when enough user data are available, users' behavior will exhibit similar patterns even if they are not explicitly asked to answer questions.

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 disabled). 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. In order to motivate users to actively browse through the video and answer the respective questions, a time restriction of five minutes was in effect during the experiment. Users were informed that the purpose of the study was to measure their performance in finding the answers to the questions within time constraints.

In the initialisation phase, every video was considered to be associated with four distinct distributions in the time interval of length k, where k is the number of the duration of the video in seconds. Each resulting series corresponds to the frequencies with which the four distinct buttons of Play/Pause, *GoForward* and *GoBackward* were used by the users at specific times. The users' activity distribution was created as follows: each time a user pressed the *GoBackward*/*GoForward* button, the interval matching the last or next, respectively, 30 seconds of the video, were incremented by a unit, meaning that during all these 30 seconds the corresponding button was assumed pushed. The underlying assumption in this case is that the user rewinds a video either because there is something interesting, or because there is something difficult to understand, while the user forwards a video because there is nothing of interest. In this way, a distribution was constructed for each button and for each video, a depiction of users' activity patterns over time.

5 Preliminary Results

In our experimentation, we have focused on the analysis of the video seeking user behavior, such as *GoBackward* and *GoForward* after the previously described smoothing procedure. An exploratory analysis with time series probabilistic tools, such as variance and noise amplitude, verified what is visually depicted in Figure 4 concerning *Video A*, the lecture video. While the *GoBackward* button signal has a quite regular pattern with a small number of regions with high users' activity, the *GoForward* button signal is characterised by a large number of seemingly random and abnormal local maxima of users' activity. This is due to the experiment design, where there was limited time for information gathering from the respective video and thus, usage of the *GoForward* shows users' tendency to rush through the video in order to remain within the time limit. We have also considered the use of the Play/Pause buttons, but for the current dataset, there were too few interactions. In the following, we present preliminary results demonstrating the proposed methodologies for detecting patterns of users' activity.

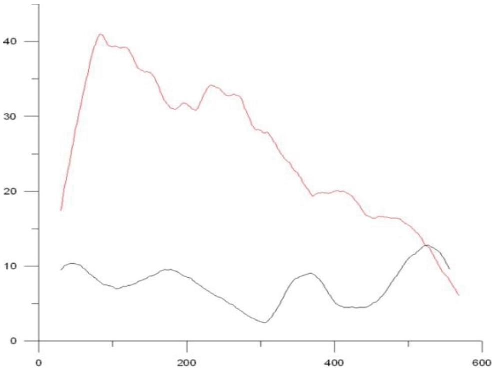

Fig. 4. The GoBackward signal (blue, at the bottom) was compared to the *GoForward* signal (red, at the top), in order to understand which one is closer to the semantics of the video: The y-axis shows the measured activity of the user while the x-axis shows the time in sec.

As far as the stochastic approach is concerned, the analysis of the users' activity distributions was based on an exploration of several alternative averaging window sizes. Results of the proposed modeling methodology for *Video A* are shown in Figure 5, and, in this case, the pulse width D is 60 seconds and the smoothing window T is 60 seconds. The results are depicted by means of pulses instead of the bell shapes in order to compare with the corresponding pulses of the ground-truth designated by the videos' authors. The mapping of between pulses and bells are based on the rule that the pulse width is equal to the width between the two points of the bell where the second derivative changes sign. Similarly, results of the proposed modeling methodology for *Video B* are shown in Figure 6, while in this case, the pulse width D is 50 seconds and the smoothing window T is 40 seconds. The smoothed signals are plotted with the solid black curve. Moreover, pulse signals were extracted from the corresponding local maxima indicating time intervals where aggregates were detected according to the definition given in Section 3. These pulses are depicted with the red line.

Within the same figures, time intervals that were annotated as ground-truth by the author of the video to contain high semantic value information are also depicted with the blue line.

For the stochastic approach, the correlation of the estimated high-interest intervals and the ground-truth annotated by the author of the video, is visually evident. Cross correlation, between the two intervals, was calculated at 0.673 and 0.612 correspondingly, indicating strong correlation between the two pulses.

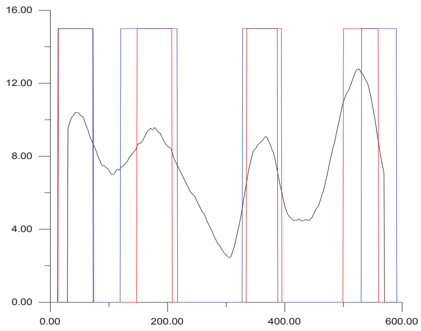

Fig. 5. *Video A*: Cumulative users' interaction vs. time including results from stochastic approach: The y-axis shows the measured activity of the user while the x-axis shows the time in sec.

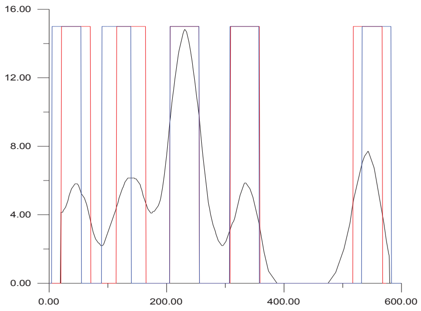

Fig. 6. *Video B*: Cumulative users' interaction vs. time including results from stochastic approach: The y-axis shows the measured activity of the user while the x-axis shows the time in sec.

For the same two videos, the application of the pattern matching approach is examined in Figures 7 and 8. In this case, both figures show additionally the calculated distance of the accumulated users' interaction with the scaled reference bell-shaped pattern (shown using the red dotted line). The bold parts of the red dotted line indicate the windows/intervals at which the calculated distance was locally minimum, thus representing the estimated intervals of the signal containing high semantic value video.

For the pattern matching apporach, the overlapping of the estimated high-interest intervals and the ground-truth annotated by the author of the video, is, as before, visually evident. The calculated overlapping percentage, between the two intervals, was calculated at 65.6% and 76.8% correspondingly.

6 Conclusion

In this research, we propose two methods that detect collective behavior of users via the detection of aggregates within the corresponding distribution of users' activity. The proposed methodologies are tested on web videos under a controlled experiment. Collective intelligence is attributing to the claim of being able to understand the importance of video content from users' interactions with the player. The results of this study can be used to understand and explore collective

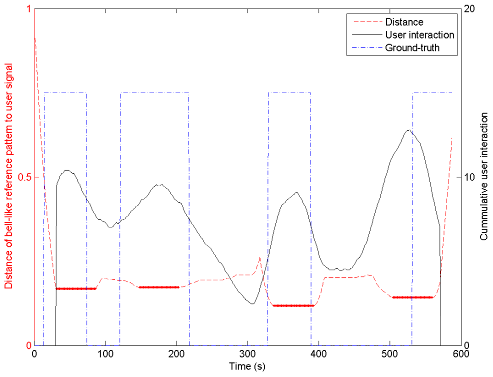

Fig. 7. *Video A*: Cumulative users' interaction vs. time including results from pattern matching approach.

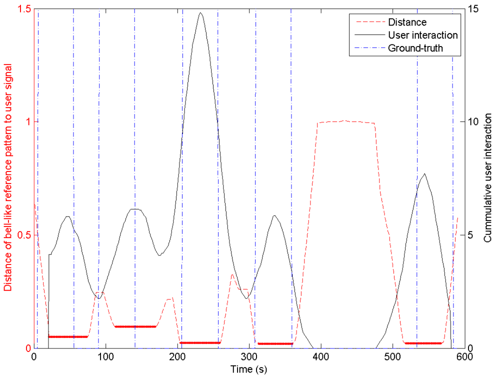

Fig. 8. *Video B*: Cumulative users' interaction vs. time including results from pattern matching approach.

intelligence in general i.e., how to detect users' collective behavior as well as how the detected collective behavior leads to judgment about the content from which users' activity was gathered. Moreover, collective intelligence may be used as a tool of user-based content analysis having the benefits of continuously adapting to evolving users' preferences, as well as providing additional opportunities for the personalisation of content. For example, users might be able to apply other personalisation techniques, e.g. collaborative filtering, to the user activity data.

Moreover, we have shown two approaches for aggregates of users' activity estimation, by means of an arbitrary bell-like reference pattern. According to the definition provided, we argue that the aggregate of users' actions locally coincides, to a large degree, with a bell-like shape of the corresponding distribution. The complete pattern of users' interactions is defined by the exact location of the center of bells of the total number of the bell-like patterns detected. In this way we map different users' behavior to different patterns observed. Moreover we have found that these observed patterns of users' actions can reveal specific judgment about the content for which actions were collected, leading thus to collective intelligence. Indeed, for the case study presented herein, the exact locations of the bell-like patterns detected can be mapped to the most important parts as was shown by experimentation. On the other hand, collective intelligence could reveal new unexpected results, i.e. important intervals of users' behavior that were unexpected.

In any case, the scope of our work is to report the large area in which collective intelligent can be applicable, to provide some initial hints as to how one can treat these phenomena as well as how to detect and define patterns which interpret collective intelligence. It is our aim to evolve the methodology presented herein and explore its applicability to more complex cases where interactions between users as well interactions of users with their environment come into play.

References

1. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proc. ACM SIGCOMM Conference on Internet Measurement. (2007) 1–14
2. YouTube: Statistics (2012) http://www.youtube.com/t/press_statistics.
3. Geetha, P., Narayanan, V.: A survey of content-based video retrieval. Journal of Computer Science **4**(6) (2008) 474–486
4. Gkonela, C., Chorianopoulos, K.: Videoskip: event detection in social web videos with an implicit user heuristic. Multimedia Tools and Applications 1–14
5. Shamma, D.A., Shaw, R., Shafton, P.L., Liu, Y.: Watch what i watch: using community activity to understand content. In: Proc. of International Workshop on Multimedia Information Retrieval. (2007) 275–284
6. Diplaris, S., Sonnenbichler, A., Kaczanowski, T., Mylonas, P., Scherp, A., Janik, M., Papadopoulos, S., Ovelgonne, M., Kompatsiaris, Y. In: Emerging Collective Intelligence for personal, organisational and social use. Springer (2011)
7. Yu, B., Ma, W.Y., Nahrstedt, K., Zhang, H.J.: Video summarization based on user log enhanced link analysis. In: Proc. of ACM International Conference on Multimedia. (2003) 382–391
8. Syeda-Mahmood, T., Ponceleon, D.: Learning video browsing behavior and its application in the generation of video previews. In: Proc. of ACM International Conference on Multimedia. (2001) 119–128
9. Vanmarcke, E.: Random fields, analysis and synthesis. MIT Press (1983)
10. Chu, K.K.W., Wong, M.H.: Fast time-series searching with scaling and shifting. In: PODS. (1999) 237–248
11. Crockford, C., Agius, H.: An empirical investigation into user navigation of digital video using the vcr-like control set. Int. J. Hum.-Comput. Stud. **64**(4) (2006) 340–355
12. xrgk: Multiple input devices (2010) http://youtu.be/8LebAtvulIY.
13. Mega tv: Protagonists tv series - use of internet by young people (2010) http://www.youtube.com/watch?v=GOQfIXxbjlE.