1

Basic Concepts

Markos Avlonitis
Department of Informatics, Ionian University, 49100 Kerkyra, Greece. avlon@ionio.gr

Ioannis Karydis
Department of Informatics, Ionian University, 49100 Kerkyra, Greece. karydis@ionio.gr

Konstantinos Chorianopoulos
Department of Informatics, Ionian University, 49100 Kerkyra, Greece. choco@ionio.gr

Spyros Sioutas
Department of Informatics, Ionian University, 49100 Kerkyra, Greece. sioutas@ionio.gr

CONTENTS

In this work, we study the collective intelligence behavior of Web users that share and watch video content. We discuss the aggregated users' video activity exhibiting characteristic patterns that may be used in order to infer important video scenes thus leading to collective intelligence concerning the video content. Initially, we review earlier works that utilize a controlled user exper-

iment with information-rich videos for which users' interactions are collected in a testing platform and modeled by means of the corresponding probability distribution function. It is shown that the bell-shaped reference patterns are significantly correlated 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 detection tool that identifies the importance of video scenes. Accordingly, we discuss both a stochastic and a pattern matching approach on the users' interactions information and report increased accuracy in identifying the areas indicated by users as having high importance information. Finally, in the last section, new insights in managing user interaction by means of a new stochastic algorithm are presented. In practice, the proposed techniques might improve navigation within videos on the web and have also the potential to improve video search results with personalized video thumbnails.

1.1 Introduction

The Web has become a very popular medium for sharing and watching video content [4]. In particular, many individuals, organizations, and academic institutions are making lectures, documentary, and how-to videos available online. Previous work on video retrieval has investigated the content of the video and has contributed a standard set of procedures, tools, and data-sets for comparing the performance of video retrieval algorithms (e.g., TRECVID), but they have not considered the interactive behavior of the users as an integral part of the video retrieval process. Besides watching and browsing video content on the web, people also perform other "social metadata" tasks, such as sharing, commenting videos, replying with other videos, or just expressing their preference/rating. Human-Computer Interaction (HCI) research has largely explored the association between commenting and micro-blogs, primarily tweets, or other text-based and explicitly user-generated content. Although there are various established information retrieval methods that collect and manipulate text, these could be considered burdensome for the users, in the context of video watching. In other cases, there is a lack of comment density when compared to the number of viewers of a video. All in all, there are a few research efforts to understand user-based video retrieval without the use of social metadata [3].

In recent research [10],video consumption activity was monitored in a well-instrumented environment that stores all the interactions with the player (e.g. play, pause, seek/scrub) for later research. Previous research [19, 18] has suggested that implicit interactions between the people and the video-player can be of great importance to video summarization. To this end, in [10] a web-video interface was constructed and a controlled user experiment was per-

formed with the goal of analyzing aggregate users' interactions with the video, through the respective player.

1.2 User-based & Content-based Approaches

1.2.1 Content-based semantics

Content-based information retrieval uses automated techniques to analyze actual video content. Accordingly, it uses images' colors, shapes, textures, sounds, motions, events, objects or any other information that can be derived from only the video itself. Existing techniques have combined the videos' metadata [23] with pictures [9], or sounds [17], while other researchers provide affective annotation [5], or navigation aids [15]. Even though content-based techniques have begun to emphasize the importance of users' content, still such approaches do not take into account peoples' browsing and sharing behavior. Moreover, low-level features (e.g. color, camera transitions) often fail to capture the high-level semantics (e.g. events, actors, objects) of the video content itself, yet such semantics are often what guide users, particularly non-specialist users, when navigating [7] within or between videos [15].

Since it is very difficult to detect scenes and extract meaning from videos, previous research has attempted to model video in terms of better-understood concepts, such as text and images [25]. To evaluate methods for understanding video content, researchers and practitioners have been cooperating for more than a decade on a large-scale video libraries 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 from videos [21].

Thus, content-based techniques facilitate the discovery of a specific scene, the comprehension of a video in a limited time and the navigation in multiple videos simultaneously. Again, the object of analysis remains the video content rather than the metadata associated with people or how people manipulated and consumed the video. Accordingly, content-based techniques are not applicable to some types of web video, such as lecture and how-to instruction that present, respectively, a visually flat-structure or complex schematic information.

1.2.2 User-based semantics

In comparison to the more so legacy content-based techniques, there are fewer works on user-based analysis of information retrieval for video content. One explanation for this imbalance is not the importance of content-based, but the relatively newer interest in the social web, sharing, and use of videos online.

Nevertheless, there is growing research and interest on user-based retrieval of video.

User interactions are one of the basic elements in user-based research. For this purpose, there is a need for detailed tracking of video browsing behavior. Syeda-Mahmood and Ponceleon [22] developed a media player-based learning system called the Media Miner. They tracked video browsing behavior, modeled users' states transition with Hidden Markov Model approaches and generated fast video previews to satisfy the "interestingness" constraint of them. MediaMiner featured the common play, pause and random seek into the video via a slider bar, fast/slow forward and fast/slow backward as well. Researchers tried to relate user activity to each user's browsing status, such as identifying if the user is bored or interested.

Besides stand-alone videos, few works perform user-based information retrieval from videos on the web. The principle example here is the work of Shamma et al. [19] and Yew and Shamma [26], who have highlighted the importance of implicit instrumentation and user-based semantic analysis of video on the web. In the former work, the authors have proposed a shift from semantics to pragmatics suggesting that content semantics follow the semantic utility of the interface. In the latter work, the authors have analyzed communicative and social contexts surrounding videos shared in synchronous environments as a means to determine a categorical genre, like Comedy, Music, etc. [27] and video virality [20].

1.3 A Controlled Experiment on User-interaction

1.3.1 Event detection systems

According to [10]. several applications have been developed by the researchers, in order to evaluate novel event detection methods. Macromedia Director, a multimedia application platform, was used to develop SmartSkip [9]. The system used re-encoded videos in QuickTime format. Similarly, Emoplayer [5] was running locally on a laptop and participants used a pointing device to interact with it. The system was developed with VC++ and DirectShow and the annotated video clips were stored in XML files. In the case of Li et al. [17] Microsoft Windows Media Player had been modified to develop the enhanced browser with its special features, because its default playback features are not sufficient for video navigation. Crockford and Agius [7] designed a system as a wrapper around an ActiveX control of Windows Media Player. A video-recorder used to collect video, at first, and then it was encoded in MPEG-1. The majority of these systems run locally, need special modification on software, and at the same time on video clips. Another important procedural parameter of the aforementioned experiments was that subjects had to be at a specific place

where the experiment was conducted. Still, besides stand-alone applications, a number of web-based systems do exist. Hotstream [12] employed Java 2 Enterprise Edition (J2EE) to develop a multi-tier web based architecture system (web-tier, middle-tier, backend database-tier, streaming platform) in order to deliver personalized video content from streaming video servers. In the same direction, Shamma et al. [19] created different web-based platforms where the user can watch, browse, select and annotate video material.

1.3.2 VideoSkip system design

The VideoSkip player provides the main functionality of a typical VCR device [7]. The selection of the buttons was made to remind the main playing/browsing controls of VCR devices because these are familiar to users. ReplayTV system and TiVo provide the ability to replay segments, or to jump forward in different speeds. In this way, the classic forward and backward buttons were modified to *GoForward* and *GoBackward*. The first one goes backward 30 seconds and its main purpose is to replay the last viewed seconds of the video, while the *GoForward* button jumps forward 30 seconds and its main purpose is to skip insignificant video segments. The thirty-second step is an average time-step used in previous research and commercial work due to the fact that it is the average duration of commercials. Next to the player's buttons, the current time of the video is shown followed by the total time of the video in seconds. A seek thumb is not available in order to avoid random guesses as this would have made difficult to analyze users' interactions. Li et al. [17] observed that when seek thumb is used heavily, users have to make many attempts to find the desirable section of the video and thus causing significant delays. VideoSkip [16] is a web video player developed with Google App Engine and the YouTube API to gather interactions of the users while they watch a video. Based on these interactions, representative thumbnails of the video are generated. Users of VideoSkip should have a Google account in order to sign in and watch the uploaded videos. Thus, users' interactions are recorded and stored in Google's database alongside with their Gmail addresses. Google App Engine's database, the Datastore, is used to store users' interactions. Each time a user signs in the web video player application, a new record is created, while whenever a button is pressed, an abbreviation of the button's name and the time it occurred are added to the Text variable. The time is stored within a second's accuracy.

1.3.3 User heuristic for event detection

Every video is associated with an array of k cells, where k is the number of the duration of the video in seconds. The user activity heuristic consists of three distinct stages. In the first stage, every cell is initialized to the number of users who have watched the video. This initial value is used to avoid extremely large negative values, to increase viewing value of the whole video and to provide

TABLE 1.1
User activity heuristic provides a simple mapping between user action and value.

User action	Play	*GoForward* (30 s)	*GoBackward* (30 s)	Pause
Heuristic	+2	-2	+2	+2

a balance for random interactions. In the second stage, the value for each cell, that has been played by the user, is increased by two. Moreover, every interaction means something for the event detection scheme. Each time a user presses the *GoBackward* button, the cells' values matching the last 30 seconds of the video, are incremented by two again. On the other hand, each time the user presses the *GoForward* button, the cells' values matching the next 30 seconds of the video, are decreased by two. A set of different values have been tested for interactions leading to the values of Table 1.1. For example we used for play, *GoBorward*, *GoBackward* and pause "+1" or for play/pause "+1" and for *GoBorward*/*GoBackward* "+2". This combination was selected in order to made the results distinguishable while avoiding increased complexity. In the third stage, the highest values of the array are considered and at the same time the number of values (interactions) that are gathered in a specific cell area (i.e., the surface size). Moreover, a distance threshold of 30 s between the selected thumbnails was defined in order to avoid having consecutive cells as a result. These specific scenes can be used as proposed thumbnails and improve users' browsing experience. Each proposed thumbnail begins at the first second of the selected area.

1.3.4 Experimental methodology

1.3.4.1 Materials

One of the key points of the research in [10], was the exploration of methods for event detection, accordingly, the selection of the suitable video content is of high importance. The videos selected are as much visually unstructured as possible, because content-based algorithms have already been successful with videos having visually structured scene changes. Another key factor considered was the length of a video. In general, the YouTube service allows video uploading up to 15 minutes, while an option exists that allows to request for increase limit leading to file uploads greater than 20GByte [31]. Although there were videos that exceeded that limit, they decided not to use them, because it would be tiresome for the majority of users. Indeed, some early pilot user tests have revealed that user attention is reduced after they have watched more than 34 videos of 10 minutes each. Narrative and entertainment have been the most popular category, while according to He et al. [11] entertainment content is more likely to be watched in a leisurely manner and costs so much to produce that it is reasonable to have a human produce previews and summaries. More-

over, lecture and how-to videos were selected, as users are actively watching them to retrieve information about a specific topic. Documentary videos could be categorized as video or audio-centric, lectures have an audio-centric content and cooking videos have more video-centric features. The documentary video features a segment of a television program called "Protagonists" [29]. The selected segment refers to the use of internet by young people. The lecture video is a paper presentation from a local workshop [28] and the presentation's topic is "The acceptance of free laptops, that have been given to secondary education students". Finally, the how-to video is a segment of a cooking TV show for a souffle-cake [30]. Each one lasts ten minutes and is available on YouTube.

1.3.4.2 Measurement

The measuring process adopted was based on the assumption of Yu et al. [32] that 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. When enough user data is available, user behavior will exhibit similar patterns even if they are not explicitly asked to answer questions. In order to experimentally replicate user activity a questionnaire was developed that corresponds to several segments of the video. Scene selection was based on a combination of audio and video factors. Thus, each question corresponds to a visual and/or a structural cue that could be used as a hint to find an answer. Furthermore, some irrelevant questions were included in order to check that the users are searching for the answers and do not attempt guess the replies. The following parameters were under consideration: audio channel, speaker's channel, end-users' actions watching the talk to reveal the significant portions and video channel to select the thumbnails. Google Docs were used in order to create online forms for users' questionnaires and subsequently integrated in the user interface as presented in Figure 1.1.

1.3.4.3 Procedure

The goal of the user experiment was 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 [14]. Instead of mining real usage data, a controlled experiment was selected as it provides a clean set of data that might be easier to analyze. The experiment took place in a lab with Internet connection, general-purpose computers and headphones. Twenty-three university students, the characteristics of which are shown in Table 1.3, spent approximately 10 minutes to watch each video, while 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. Next, there was a

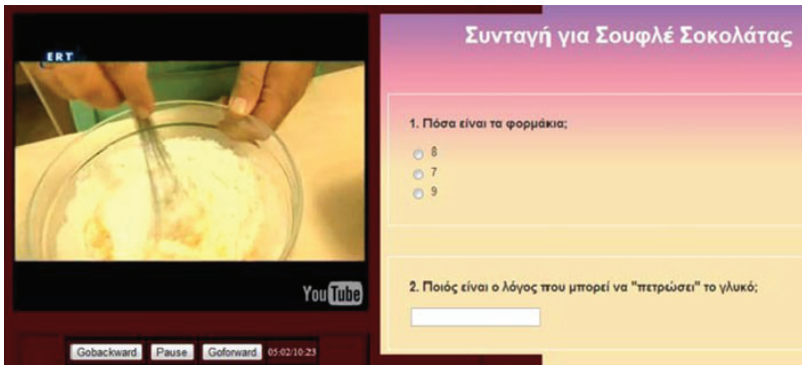

FIGURE 1.1
Screenshot of VideoSkip with the questionnaire.

TABLE 1.2
Example questions from each video.

Video	**Indicative questions**
Lecture video	Which are the main research topics? What the students did not like? What time does the first part of the talk end?
Documentary video	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?

time restriction of 5 minutes, in order to motivate the users to actively browse through the video and answer the respective questions. Example questions for each video are shown in Table 1.2. Users were informed that the purpose of the study was to measure their performance in finding the answers to the questions within time constraints.

Before the experimental procedure participants were introduced to the user interface of the video player and the questionnaire. The experimental session for each video consisted of two parts. Initially, the users had to watch a video and afterwards to answer the respective questionnaire. They could not see the

TABLE 1.3
Summary of users' characteristics.

Number	23
Age	18-35
Gender	13 Female, 10 Male
Occupation	Studying informatics
Motivation	Course credit

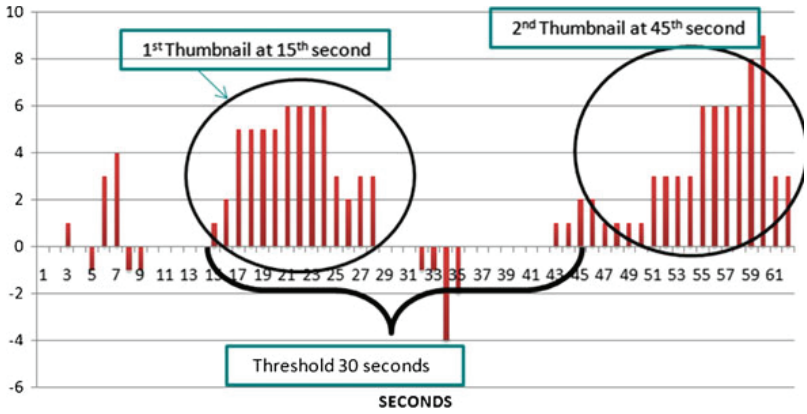

FIGURE 1.2
User activity graph with heuristic rules.

TABLE 1.4
Overview of the user activity modeling and analysis

Stage	User activity signal processing
1	Smoothness procedure
2	Determination of users' activity aggregates
3	Estimation of pattern characteristics

questions from the beginning and the video player's buttons were disabled during the first part. Buttons were re-enabled for use in the second part and participants could use them to browse video and search for the answer. Figure 1.1 portrays the second part of the experimental procedure. Furthermore, there was a time restriction of 5 minutes in this part, in order to motivate the users to actively browse through the video. The procedure was repeated in a random sequence for each video, in order to minimize possible learning effects. The result of this simple heuristic procedure is shown in Figure 1.2, where a coarse grained aggregation is evident.

1.4 Modeling User Interaction as Signals

The analysis to follow is based on the idea presented at [1]. Indeed, in order to extract pattern characteristics for each button distribution, i.e. scenes in which users exhibit high interaction with the video-player, three distinct stages (as shown in Table 1.4), are used.

In the first stage, a simple procedure is used in order to average out user

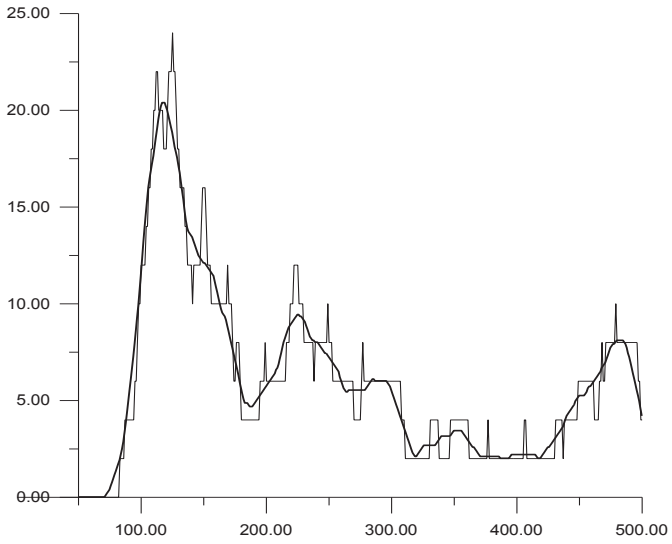

FIGURE 1.3
The user's activity signal is approximated with a smooth signal.

activity noise in the corresponding distribution. In the context of probability theory, noise removal can be treated with the notion of the moving average [24]: from a curve $S^{exp}(t)$ a new smoother curve $S_T^{exp}(t)$ may be obtained as shown in Equation 1.1,

$$S_T^{exp}(t) = \frac{1}{T}\int_{t-T/2}^{t+T/2} S^{exp}(t')dt' \tag{1.1}$$

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 1.3. 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 users' 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 localized shape of the users' signal which will in turn show the area of high user activity.

In the second stage, aggregates of users' activity are estimated by means of an arbitrary bell-like reference pattern. As a milestone of this work 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 users' actions are concentrated at a specific time interval (the

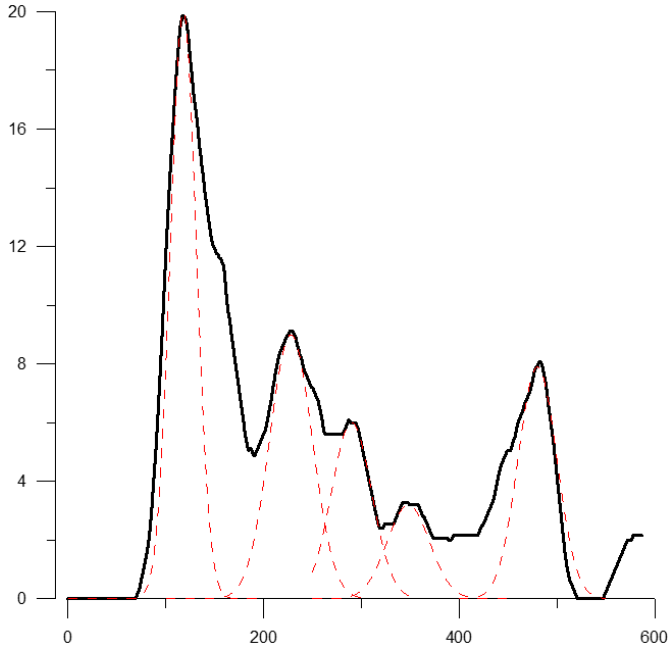

FIGURE 1.4
The users' activity signal is approximated with Gaussian bells in the neighborhood of user activity local maxima.

center of the bell) while this probability tends to zero quite symmetrically as we move away from this interval (Figure 1.4). 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. The same idea, in a rather premature form, was used in Karydis et al. [13]. The notion of the bell like characteristic pattern for users' aggregation is revisited and further detailed in Section 1.6.

The third step produces an estimation of the pattern characteristics, i.e. the number of users' aggregates for the specific signal and moreover their exact locations in time, by application of two different methodologies, a stochastic and a pattern matching.

1.5 Treating User Signals

The aforementioned stochastic and pattern matching methodologies for estimation of the pattern characteristics are herein detailed.

- In the stochastic approach, the estimation of the exact locations can be done via the estimation of the generalized local maxima. The term generalized local maxima in this context refers 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. This is possible by estimation of 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 the 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. Experimentation in [13] proposed 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 distance of the reference bell-shaped pattern to the accumulated user interaction signal is measured using 3 different distance measures. Initially, a Scaling and Shifting (translation) invariant Distance (SSD) measure (Equation 1.2), is adopted from [6]. Accordingly, for two time series x and y, the distance $\hat{d}_{SSD}(x, y)$ between the series is:

$$\hat{d}_{SSD}(x, y) = min_{a,q} \frac{\|x - \alpha y_{(q)}\|}{\|x\|} \tag{1.2}$$

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

  The second distance measure used is the Euclidean Distance (ED) measure (Equation 1.3) that has been shown to be highly effective [8] in many problems, despite its simplicity:

$$d_{ED}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1.3}$$

  Finally, the third distance measure utilized is a Complexity-Invariant Distance (CID) measure (Equation 1.4) for time series as discussed by Batista et al. [2]:

$$d_{CID}(x, y) = ED(x, y) \times CF(x, y) \tag{1.4}$$

where the two time series x and y are of length n, $ED(x, y)$ is the Euclidean distance (Equation 1.3), $CF(x, y)$ is the complexity correction factor defined in Equation 1.5:

$$CF(x, y) = \frac{max(CE(x), CE(y))}{min(CE(x), CE(y))} \tag{1.5}$$

and $CE(x)$ is a complexity estimate of a time series X, calculated as shown in Equation 1.6:

$$CE(x) = \sqrt{\sum_{i=1}^{n-1}(x_i - x_{i+1})^2} \tag{1.6}$$

The aforementioned distance measures produce another time series *dist* that describes the distance of the reference bell-shaped pattern to the accumulated user interaction signal and thus requires the identification the locations of *dist* where its value is minimal, indicating a close match of the the reference bell-shaped pattern to the accumulated user interaction signal. To avoid using a simplistic global cut-off threshold a local minima peak detection methodology is employed, where a point in *dist* is considered a minimum peak if it has the minimal value, and was exceeded, to the left of the signal, by a value greater by $DELTA$, the peak detection sensitivity value.

In the experimentation to follow, the focus has been 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 1.5 concerning 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 characterized 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. The following, present preliminary results demonstrating the results received from the aforementioned methodologies for detecting patterns of users' activity.

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 the lecture video are shown in Figure 1.6, 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

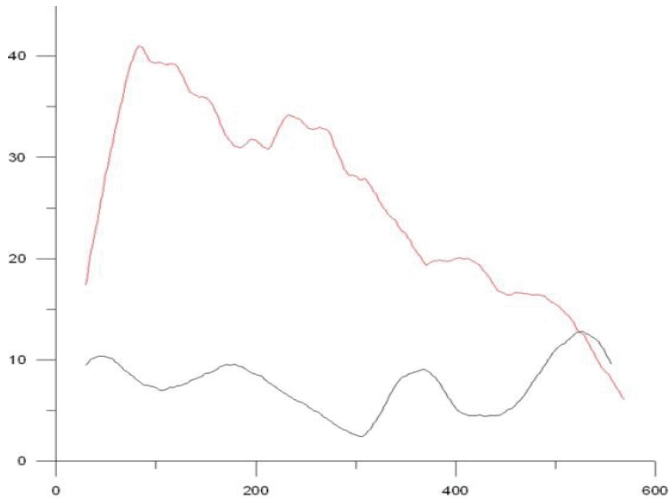

FIGURE 1.5
GoBackward signal (blue, bottom) compared to *GoForward* signal (red, top), in order to understand which one is closer to the semantics of the video: The y-axis shows the measured activity of the user while the x-axis shows the time in sec.

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 the documentary video are shown in Figure 1.7, 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 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.

For the same two videos, the application of the pattern matching approach is examined for each distance measure using the F1 score & Matthews Correlation Coefficient (MCC) value for varying peak detection sensitivity values for each of the three distance measures, SSD, ED and CID respectively. It should be noted that in the results to follow the F1 score is linearly transposed from $[0, 1]$ to $[0, 100]$ in order ensure ease of comparison.

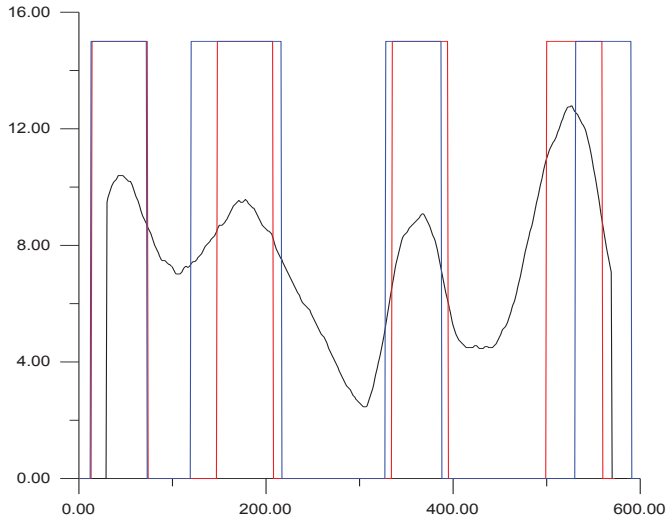

FIGURE 1.6
Lecture video: Cumulative users' interaction vs. time including results from stochastic approach: The y-axis shows the measured activity of the user while the x-axis shows the time in sec.

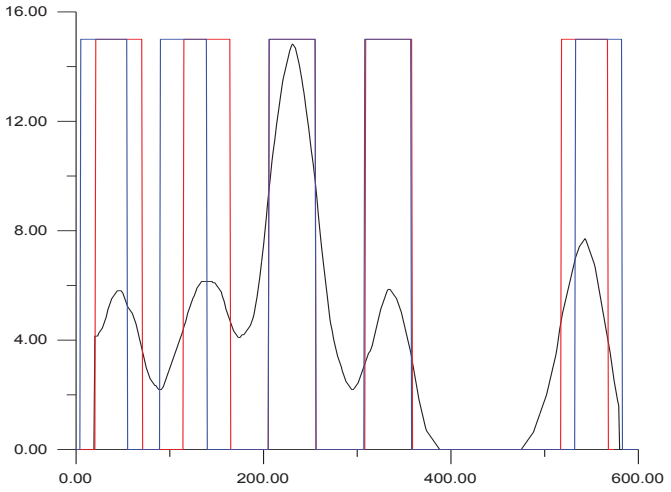

FIGURE 1.7
Documentary video: Cumulative users' interaction vs. time including results from stochastic approach: The y-axis shows the measured activity of the user while the x-axis shows the time in sec.

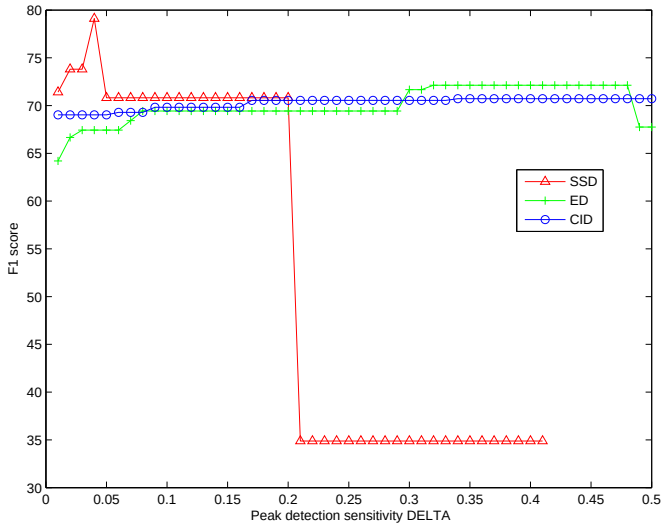

FIGURE 1.8
Lecture video, pattern matching approach, F1 score for SSD, ED and CID metrics.

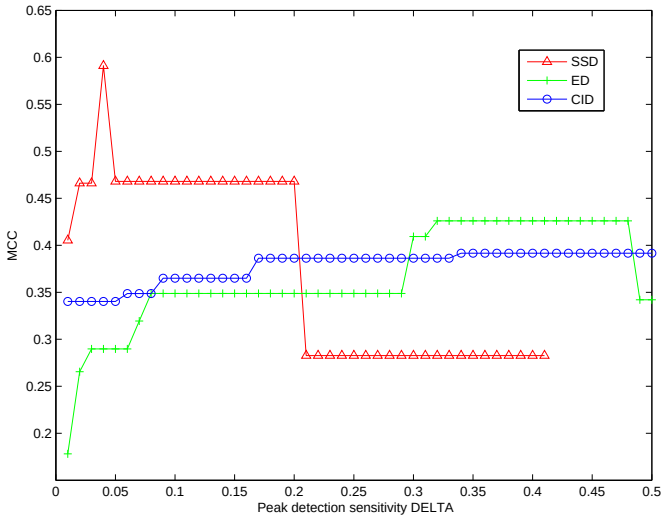

FIGURE 1.9
Lecture video, pattern matching approach, MCC for SSD, ED and CID metrics.

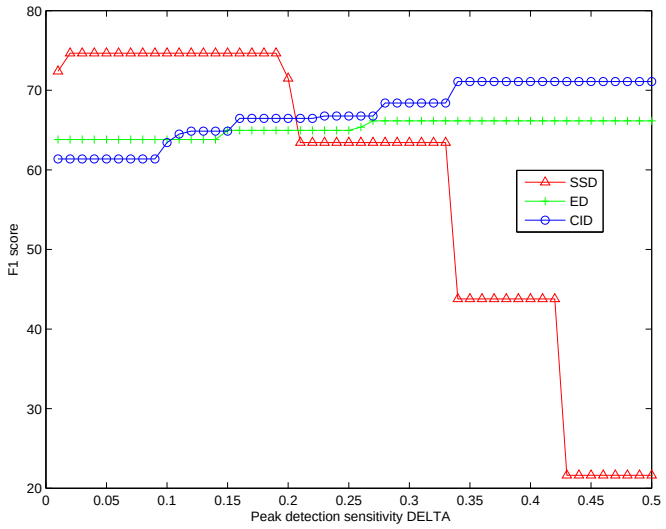

FIGURE 1.10
Documentary video, pattern matching approach, F1 score for SSD, ED and CID metrics.

Lecture video As shown in Figures 1.8 & 1.9, the SSD metric achieved an F1 score of 79 in a scale of $[0, 100]$, with 100 being the best value. Still as the F1 score does not take the true negative rate into account the MCC value has been computed leading to a 0.6 value on a scale of $[-1, 1]$, with 1 implying a perfect prediction. The claim of the ability of Euclidean Distance to be performing relatively high, despite its simplicity, is shown in this experiment where ED scored an F1 score of 72 and an MCC value of 0.42. Finally, the CID measure was outperformed by the other two measures having scored an F1 score of 70 and an MCC value of 0.39.

Documentary video As shown in Figures 1.10 & 1.11, the SSD metric achieved an F1 score of 75 and an MCC value of 0.56. The Euclidean Distance scored an F1 score of 66 and an MCC value of 0.34. Finally, the CID measure outperformed the ED measure having scored an F1 score of 71 and an MCC value of 0.45.

1.6 New Insights in Managing User-interactions

The stochastic and pattern matching methodologies for estimation of the pattern characteristics discussed in this work were tested on web videos under

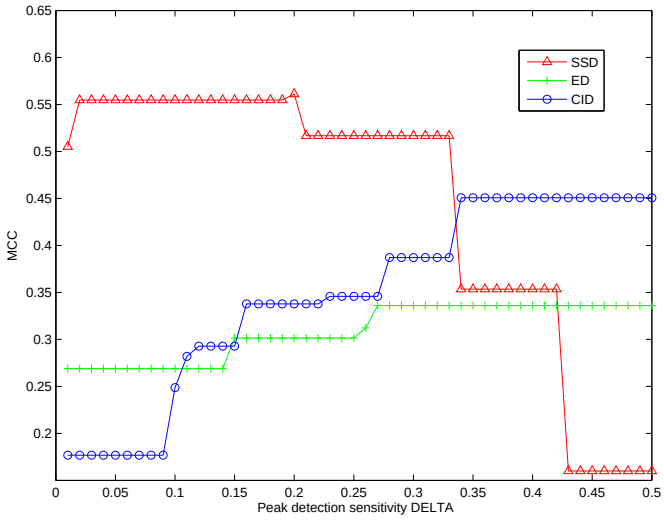

FIGURE 1.11
Documentary video, pattern matching approach, MCC for SSD, ED and CID metrics.

a controlled experiment and were shown to present interesting results. 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 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 personalization of content. For example, users might be able to apply other personalization techniques, e.g. collaborative filtering, to the user activity data.

According to the definition provided for the two approaches for aggregates of users' activity estimation, it has been shown 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 one may map different users' behavior to different patterns observed. Moreover, these observed patterns of users' actions may 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 a more general fashion, the methodology presented may treat general users' interactions for a specific (on line) content, by interpreting these interactions as an explicit time series. This could be a time series of clicks or plays of a video on YouTube, the number of times an article on a newspaper website was read, or even the number of times that a hash tag in Twitter was used.

Thus, this methodology can be applied for the detection of patterns emerging in the temporal variation of the corresponding time series indicating the importance of a segment of content at a specific time interval of its duration.

One may formally define this either as a problem of time series correlation based on the correlation between the shape of the (experimentally collected) time series with the shape of a reference time series indicating local maximization of users' activity or as pattern matching of time-series wherein different similarity measures can be utilized for the detection of local minima in the distance. In both cases a Gaussian function can be chosen as the optimum function for the reference time series.

Given that online content has large variation during its duration, i.e. users' actions occur at arbitrary times and with very different time intervals, a further extension of the proposed methodology is needed in order to adapt a time series metric that is invariant to scaling and shifting, i.e. to be able not only to detect the exact location of the local maxima of users' popularity but also to estimate the corresponding absolute importance as well as the corresponding time interval over which the specific piece of content was important enough.

To this end, it is possible, based to the aforementioned approach, to build a scale free similarity metric introducing the notion of the aforementioned reference bell-like time signal. Indeed, the final result of this extended algorithm would be the estimation of the maximum correlation coefficient in terms of the optimum time moment and optimum bell width.

Accordingly, we propose a two value correlation coefficient $r(t_c, w)$ where t_c is the time center of the Gaussian bell and w its width. In other words we construct a Gaussian time signal by shifting its center over the time domain of the experimental signal and for each position we create a number of different Gaussian time signals gradually increasing its width w (see Figure 1.12 from blue to red and to green solid Gaussian bell). For each Gaussian reference signal of different width we estimate the corresponding correlation coefficient with the experimental signal. In this way we produce a two dimension correlation coefficient for each time location and for different bell widths. We stated that whenever, for a specific time center of the Gaussian bell, a high correlation coefficient is identified during the time shifted process, a local maxima of the experimentally constructed time series is assumed. Indeed this can be seen in Figure 1.13 for an arbitrary time series (black solid line). With the red solid line, normalized to 10, the Figure depicts the corresponding correlation coefficient between the arbitrary time series and the shifted Gaussian bell. Initially

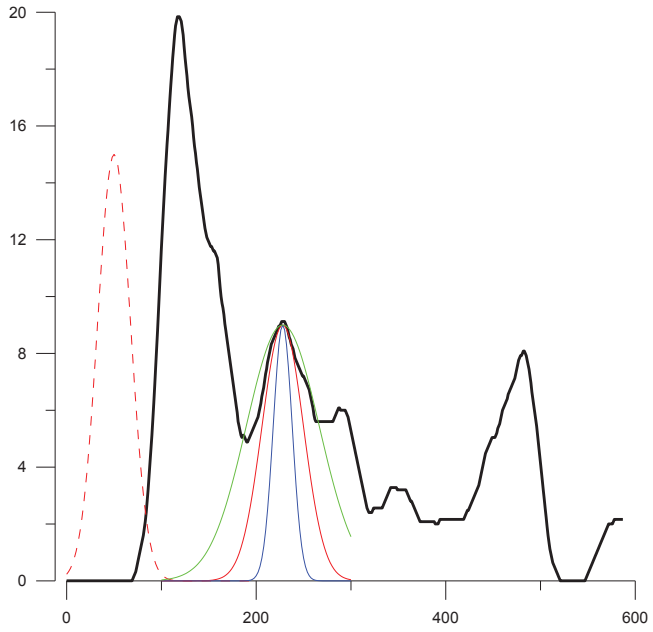

FIGURE 1.12
Gaussian bell is shifted over the time domain. When a local maximum of the correlation coefficient is detected a series of variables widths is created in order to estimate the optimum width.

we keep the width of the bell constant while a robust alternative measure for the initial width could be the variance of the smoothed experimental signal.

It is evident that there is a very clear maximum of the correlation coefficient exactly when the center of the Gaussian bell coincides with the maximum of the experimental series. As a result, the exact location of the experimental series is detected as the point of the local maximum of the corresponding correlation coefficient. Then, we relax the assumption of the constant bell width: keeping constant the center of the bell we built Gaussian bells of different widths (as depicted in Figure 1.12). For each bell of variable width a new correlation coefficient is computed. The maximum value of this second set of correlation coefficients is estimated completing thus our process. The final result is the estimation of the maximum correlation coefficient in terms of the optimum time moment and optimum bell width. We argue that the optimum time moment coincides with the local maximum of the online media popularity while the optimum Gaussian bell width coincides with the corresponding time interval over which popularity is important enough.

Summarizing, the proposed algorithm, r-algorithm 1 performs the follow-

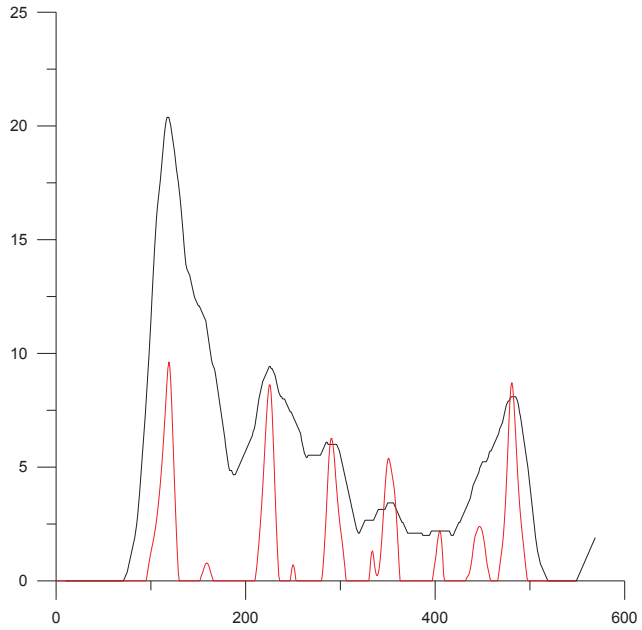

FIGURE 1.13
Local maxima of the correlation coefficient (red curve) coincide with local maxima of user's activity signal (black curve).

ing steps: we begin with an initial Gaussian bell the center of which is located at the time origin of the content and its width coinciding with the variance of the smoothed experimental signal. Then follows a two step procedure, the detection step and the refinement or characterization step. In the detection step the bell is shifted along the time domain computing the corresponding correlation coefficient between the Gaussian bell and the experimental signal. The local maxima of users' activity are identified as the time moments where the computed correlation coefficient reaches local maximum, with the local maximum being above a specific threshold.

In the characterization step, for each local maximum of the correlation coefficient a series of Gaussian bells with variable widths is generated (beginning from a value of few seconds to a fraction of the overall duration of the content) and the corresponding correlation coefficients are computed again. The calculated optimal bell width gives an estimation of the time interval over which the content was important enough for the users.

Algorithm 1 The r-algorithm

Require: Experimental time series, upper part of Gaussian time series $g(ct, w)$ of center ct and width w.
for $ct = 1$ **to** L **do** {detection step}
r_ct {the correlation coefficient for different centers}
if $r_ct > thress$ **then** {critical threshold of correlation}
for $w = 1$ **to** $L/10$ **do** {characterization step}
r_ct_w {correlation coefficient for variable widths}
end for
end if
end for
return r_ct_w {returns seconds of maximum user's activity and the corresponding time interval of popularity}

1.7 Epilogue

In this work we present a method that detects collective behavior of users' activity via the detection of characteristic patterns in the corresponding signal monitoring users' activity. The methodology has been verified with web videos and user interaction data from a controlled experiment.

An algorithm for real time detection of collective activity was presented, at the basis of which is the notion of a two parameter arbitrary Gaussian bell acting as a reference pattern for aggregation. Accordingly, the aggregation of users' actions coincides to the upper part of a bell-like shape of the corresponding distribution. The users' interactions pattern is defined by means of two parameters: the exact location of the center of the Gaussian bell, as well as the corresponding width. In this way, one may map different users' behavior to different patterns observed.

Moreover, within the discussed methodology the exact height of each local maxima of users' activity can also be addressed. Indeed, as soon as the exact locations of each local maximum are estimated then, the corresponding heights coincide with the respective value for that time instance of the smoothed experimental signal. It should be noted that alternatively the use of a three-parameter value correlation coefficient could be used, i.e. for the determination of the users' maximum height, the corresponding computational cost is very high.

Further on, we need to stress that the robust determination of the relative heights of each maximum of collective activity is a very crucial parameter since it scores the importance of each maxima. As a result, within the methodology presented a ranking procedure can be built: the relative height of each maximum indicates the relative importance of each scene.

The results of this study could facilitate the understanding of collective

intelligence in online media i.e., how to detect collective behavior as well as how the detected collective behavior leads to judgment about the importance of fragments in time-based content. As a future work, the methods presented may be improved in order to capture not only quantitative measures of the Gaussian bells but also qualitative features such that specific symmetries of the bells corresponding to the specific content over which actions are reported, thus leading to collective intelligence.

Bibliography

[1] Markos Avlonitis, Konstantinos Chorianopoulos, and David Ayman Shamma. Crowdsourcing user interactions within web video through pulse modeling. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, Proc. ACM Multimedia Workshop on Crowdsourcing for multimedia, pages 19–20, 2012.

[2] Gustavo E. A. P. A. Batista, Xiaoyue Wang, and Eamonn J. Keogh. A complexity-invariant distance measure for time series. In *Proc. SIAM Conference on Data Mining*, pages 699–710, 2011.

[3] Tamara L. Berg, Alexander Sorokin, Gang Wang, David A. Forsyth, Derek Hoiem, Ian Endres, and Ali Farhadi. It's all about the data. *Proceedings of the IEEE*, 98(8):1434–1452, 2010.

[4] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. ACM SIGCOMM Conference on Internet Measurement*, pages 1–14, 2007.

[5] Ling Chen, Gen-Cai Chen, Cheng-Zhe Xu, Jack March, and Steve Benford. Emoplayer: A media player for video clips with affective annotations. *Interact. Comput.*, 20(1):17–28, January 2008.

[6] K. K. W. Chu and M. H. Wong. Fast time-series searching with scaling and shifting. In *PODS*, pages 237–248, 1999.

[7] Chris Crockford and Harry Agius. An empirical investigation into user navigation of digital video using the vcr-like control set. *Int. J. Hum.-Comput. Stud.*, 64(4):340–355, 2006.

[8] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proc. VLDB Endowment*, 1(2):1542–1552, 2008.

[9] Steven M. Drucker, Asta Glatzer, Steven De Mar, and Curtis Wong. Smartskip: consumer level browsing and skipping of digital video content. In Dennis R. Wixon, editor, *CHI*, pages 219–226. ACM, 2002.

[10] Chrysoula Gkonela and Konstantinos Chorianopoulos. Videoskip: event detection in social web videos with an implicit user heuristic. *Multimedia Tools and Applications*, pages 1–14.

[11] Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. Auto-summarization of audio-video presentations. In *Proc of ACM International Conference on Multimedia*, pages 489–498, 1999.

[12] Rune Hjelsvold, Subu Vdaygiri, and Yves Léauté. Web-based personalization and management of interactive video. In *Proc. of International Conference on World Wide Web*, pages 129–139, 2001.

[13] Ioannis Karydis, Markos Avlonitis, and Spyros Sioutas. Collective intelligence in video user's activity. In *Artificial Intelligence Applications and Innovations (2)*, pages 490–499, 2012.

[14] Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3(1–2):1–224, 2009.

[15] Jinwoo Kim, Hyunho Kim, and Kyungwook Park. Towards optimal navigation through video content on interactive tv. *Interact. Comput.*, 18(4):723–746, 2006.

[16] Ioannis Leftheriotis, Chrysoula Gkonela, and Konstantinos Chorianopoulos. Efficient video indexing on the web: A system that crowdsources user interactions with a video player. In *UCMedia*, pages 123–131, 2010.

[17] Francis C. Li, Anoop Gupta, Elizabeth Sanocki, Li-wei He, and Yong Rui. Browsing digital video. In *Proc. of SIGCHI Conference on Human factors in Computing Systems*, pages 169–176, 2000.

[18] Arthur G. Money and Harry W. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *J. Visual Communication and Image Representation*, pages 121–143, 2008.

[19] David A. Shamma, Ryan Shaw, Peter L. Shafton, and Yiming Liu. Watch what i watch: using community activity to understand content. In *Proc. of International Workshop on Multimedia Information Retrieval*, pages 275–284, 2007.

[20] David A. Shamma, Jude Yew, Lyndon Kennedy, and Elizabeth F. Churchill. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *Proc. of ICWSM*, 2011.

[21] Cees G. M. Snoek and Marcel Worring. Concept-based video retrieval. *Found. Trends Inf. Retr.*, 2(4):215–322, April 2009.

[22] Tanveer Syeda-Mahmood and Dulce Ponceleon. Learning video browsing behavior and its application in the generation of video previews. In *Proc. of ACM International Conference on Multimedia*, pages 119–128, 2001.

[23] Y. Takahashi, N. Nitta, and N. Babaguchi. Video summarization for large sports video archives. *Multimedia and Expo, IEEE International Conference on*, 0:1170–1173, 2005.

[24] Erik Vanmarcke. *Random fields, analysis and synthesis*. MIT Press, 1983.

[25] Rong Yan and Alexander Hauptmann. A review of text and image retrieval approaches for broadcast news video. *Information Retrieval*, 10:445–484, 2007.

[26] Jude Yew and David A. Shamma. Know your data: Understanding implicit usage versus explicit action in video content classification. In *Proc. of IS&T/SPIE Symp. on Electronic Imaging: Science & Technology*, pages 297–306, 2011.

[27] Jude Yew, David A. Shamma, and Elizabeth F. Churchill. Knowing funny: genre perception and categorization in social video sharing. In *Proc. of Annual Conference on Human Factors in Computing Systems*, pages 297–306, 2011.

[28] YouTube. The acceptance of free laptops, that have been given to secondary education students, 2012. http://www.youtube.com/watch?v=Z09ythJT9Wk.

[29] YouTube. Protagonists tv series, 2012. http://www.youtube.com/watch?v=GOQfIXxbjlE.

[30] YouTube. Soufle sokolatas - cooking lesson, 2012. http://www.youtube.com/watch?v0LzkYvtqlT5I.

[31] YouTube. Upload videos longer than 15 minutes, 2013. http://support.google.com/youtube/bin/answer.py?hl=en-US&answer=71673&rd=1.

[32] Bin Yu, Wei-Ying Ma, Klara Nahrstedt, and Hong-Jiang Zhang. Video summarization based on user log enhanced link analysis. In *Proc. of ACM International Conference on Multimedia*, pages 382–391, 2003.