Social Video Retrieval: Research Methods in Controlling, Sharing, and Editing of Web Video

Konstantinos Chorianopoulos and David A. Shamma and Lyndon Kennedy

Abstract Content-based video retrieval has been a very efficient technique with new video content, but it has not regarded the increasingly dynamic interactions between users and content. We present a comprehensive survey on user-based techniques and instrumentation for social video retrieval researchers. Community-based approaches suggest there is much to learn about an unstructured video just by analyzing the dynamics of how it is being used. In particular, we explore three pillars of online user activity with video content: 1) Seeking patterns within a video is linked to interesting video segments, 2) Sharing patterns between users indicate that there is a correlation between social activity and popularity of a video, and 3) Editing of live events is automated through the synchronization of audio across multiple viewpoints of the same event. Moreover, we present three complementary research methods in social video retrieval: Experimental replication of user activity data and signal analysis, data mining and prediction on natural user activity data, and hybrid techniques that combine robust content-based approaches with crowd sourcing of user generated content. Finally, we suggest further research directions in the combination of richer user- and content-modeling, because it provides an attractive solution to the personalization, navigation, and social consumption of videos.

1 Introduction

Every second millions of users enjoy video streaming on a diverse number of terminals (TV, desktop, smart phone, tablet) and create billions of interactions within

Konstantinos Chorianopoulos
Ionian University, Greece, e-mail: choko@ionio.gr

David A. Shamma
Yahoo! Research, USA, e-mail: aymans@acm.org

Lyndon Kennedy
Yahoo! Research, USA, e-mail: lyndonk@yahoo-inc.com

video or between users. This amount of data might be converted into useful information for the benefit of all video users. In this chapter, we examine research methods for the main types of user interaction with video on the Web, such as controlling, sharing, and editing [3]. Indeed, Web-based video has become a popular medium for creating, sharing, and active interaction with video [4, 5, 20]. At the same time, Web-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 video. For example, the users are able to seek forward and backward within a video, to post comments, share with other users, and to post their own video recordings regardless of the transport channel (e.g., mobile, web, broadcast, IPTV). In this work, we suggest that user-based video retrieval techniques are beneficial for all Web-based video systems.

In the next Section, we present an outline of the most significant research findings in video retrieval. Moreover, we provide a summary of the research methods that have been employed by scientists in the exploration of video retrieval.

2 Related Work

Although online video availability is growing rather fast, there have been few research efforts to understand and leverage actual user behavior with video. 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 (summaries) of videos. According to Money and Agius [21], there is a classification for video summarization techniques: 1) internal summarization techniques that analyze information sourced directly from the video stream, and 2) external ones that analyze information not sourced directly from the video stream. Notably, Money and Agius 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.

Abstraction techniques are a way for efficient and effective navigation in video clips [17]. For example, stationary images have proven an effective user interface in video editing [1] as well as in video browsing [12]. According to Truong and Venkatesh [32] 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. The evaluation of a key-frame extraction and video summarization systems has been considered a very difficult problem [19].

In the following subsections, we present a comprehensive overview of the state-of-the-art in social video retrieval (Table 1), in order to create a context for the study of more detailed case Studies that follow immediately after.

2.1 Visual feature video retrieval

Content-based video retrieval has been concerned with signal analysis of audio and video content. Moreover, content-based research has regarded the meta-data that are being produced during the editing process of video content. In terms of research techniques, content-based researchers have defined a set of ground-truths that are used as benchmarks during the evaluation of systems that focus on the fixed data and meta-data of the video file. In this way, content-based systems have improved the quality of retrieving, adapting, and navigating in video content.

The main focus of content-based research has been the segmentation of video content by detecting key-frames, and important video segments. Content-based research has established the need for video thumbnails [10], video summaries [17], and the usefulness of automatic detection of key-frames for user navigation [32, 21]. 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 [11] 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. [16] developed an interface that generates shot boundaries using a detection algorithm that identifies transitions between shots.

The above research has found many practical applications in the industry of video retrieval. Nevertheless, in the case of Google Video, there are so many thumbnails that a separate scroll bar has been employed for navigating through them (Figure 1, left). 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). 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.

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. On the other hand, content-based approaches provide robust technologies for quickly analyzing large numbers of new items.

2.2 Audio feature video retrieval

A number of research efforts have addressed the domain of media from live music events. These efforts mostly looked at ways to present professionally produced or 'authoritative' video or audio content (e.g., a complete video capture provided by the event organizers). Naci and Hanjalic [22] provide a demo that utilizes audio analysis to help users find interesting moments from the concert. Detecting interest-

ing moments automatically is approximated by detecting applause, instrument solos and audio level of excitement; a browsing interface is provided that is supported by the extracted data. Snoek et al. [31], the authors use the visual signal of produced content from live concerts to create concept detectors including 'audience', 'drummer', stage' and so forth. Nevertheless, previous research on video recordings from concerts has not considered community-contributed content widely available on the Web.

A core audio identification technology known as audio fingerprinting [33, 13] is a rather robust and powerful technology that is already a reality in several commercial applications. In audio fingerprinting, audio recordings are characterized by local occurrences of particular structures. Given two recordings, one could rapidly identify if they were derived from the same original source material, despite a rather large amount of additive noise.

2.3 Text-based video retrieval

Besides audio and visual features, researchers have leveraged existing techniques in text retrieval, in order to index and understand the contents of a video. Although text is not an inherent part of every video stream, there are several occasions that text complements a video [34]. For example, broadcast video usually includes closed-captions (text with time code synchronized to the video), which have been mined to assign meaning to the respective video segments. Moreover, in some video segments, text might be part of the image. Then, optical character recognition might be employed in order to understand the respective text. Despite the robustness of the existing text-based retrieval techniques, we cannot assume that the majority of commercially available video streams are coupled or embed text. In the next subsection, we are also describing those text-based video retrieval techniques that are generated by the users in social media.

Fig. 1 Video-frames are an important part of user navigation within and between videos (left). 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).

How a video is used, interacted with, and commented is often indicative of the nature of its content. Shamma et al. [27] has explored whether micro-blogs (e.g., Twitter) could structure a TV broadcast. Video sharing sessions leave behind digital traces in the form of server logs and metadata. The ability to share videos in real-time while in an Instant Messaging (IM) session is an attempt to replicate the social experience of watching videos in a virtual environments. 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, such as writing a micro-blog or posting a comment. Nevertheless, the percentage of users leaving a comment is rather small when compared to the real number of viewers [20].

2.4 User-based video retrieval

User-based techniques approach the problem of video retrieval differently to the established content-based ones. Rather than paying attention to the content of the video, its metadata, or its position in a network, they focus mainly on identifying particular video interaction patterns, such as video seeking and sharing between users.

Media is often experienced socially and a large part of that experience involves frequent commentary, backchannel conversations and starts/stops to review key moments in the media content. Social video interactions on web sites are very suitable for applying community intelligence techniques [37]. Levy [15] 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 [29] 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, as well as of the re-use of video segments in community re-mixes and mash-ups [30] to understand media semantics.

2.5 Research issues and methods in social video retrieval

2.5.1 Controlling video and controlled experiments

The concept of analyzing implicit user interaction in computing activities, in order to develop user models and to provide intelligent interactions is not new. Liu et al [18] have improved the personalization of news items by analyzing previous users interactions with news items. In the context of multimedia, previous research has considered both content- and user-based methods for video retrieval. The most generic user interaction with social video is the seeking behavior within video. Notably, the video seeking behavior has been employed as a research granule in key-frame detection. 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. [19] have argued that: 'Although the issues of key-frame extraction and 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. Chorianopoulos et al. [6] 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 experimentally replicated ground truths about the content (e.g., questions about specific parts of the video).

2.5.2 Sharing video and data-mining

In online multimedia sharing contexts one-to-one chatting provides a rich context for social exchange and data collection. While still emerging, several systems support various realtime multimedia sharing interactions, like TuVista, Zync, and Google Hang-outs. In effect, as people chat while sharing online videos, they leave traces of activity (clicks and chats) and inactivity (pauses), which can reveal more about the underlying multimedia, which is fueling the conversation. To discover the content categories of videos in one-to-one sharing systems such as Zync, Yew et al. [35] examined the types of non-content data available. They collected an aggregate volume of chat activity as number of characters typed during a playback moment of the video. Beyond chat, play, pause, and scrubs were also logged. Most importantly, they looked at the length of the chat session while the embedded video player was open. This is to be distinguished from the length of the video. Using the collected data, they modeled each video as a feature vector that was informed from qualitative surveys and semi-structured interviews. This vector was then used to predict the video's content category, like news, sports, film, or TV.

2.5.3 Editing video and hybrid techniques

In some cases, we can consider the video object itself to be a proxy for an interaction with an actual live event. In particular, many online video users commonly record videos of events that they are attending and later share those clips online in order to express presence to their friends and others. In a system by Kennedy and Naaman [14], this application was explored in the context of videos recorded at live concert events. This collection of videos from a concert event can tell us a lot about the relative importance of any particular moment in the event and give us some clues for understanding semantically why the given moment is important. To uncover these importance cues, one has to first discover which videos were actually recorded at the same time during a given concert event. This can be approached by

utilizing an audio fingerprinting system, which can detect replicated audio tracks under extreme noise conditions with very high precision. The insight here is that the videos are not just videos, but rather an expression of interest by an individual in a particular point in an event. Aggregating across many different individuals expressing their interest in various points in the media, we can arrive at a general level of interest spread temporally across the event. Furthermore, the words that many different individuals use to describe each point in the event can be aggregated to give us clues about why, semantically, the point in the event is of interest.

Table 1 Previous user-based research has established the significance of mapping user actions to video semantics, but there is no silver-bullet because each approach has some drawback

	Advantages	**Challenges**
Ma et al. [19]	Assumes that viewers are interested in particular well-defined and easy to retrieve content features (e.g., faces).	Content-based and relies on a limited, preset vocabulary of what is interesting.
Shaw and Davis [29]	User contributed comments and tags.	Most data lacks temporal indexing into the content.
Shaw and Schmidt [30]	Community remix of popular video revels salient segments.	Only a portion of users performs re-mixes of video.
Shamma et al. [27]	Micro-blogs are associated to many TV broadcasts.	Deep timing information might lag against the video cue time.
Kennedy and Naaman [14]	Audio fingerprinting on aggregated recordings of the same live event	Only a small portion of online content has been recorded and uploaded by multiple users.
Carlier et al. [2]	Zoom denotes areas of interest within a video frame	Zoom is not a common feature
Yew et al. [35]	User comments accurately predict the category of the video	Only a portion of users posts comments.
Olsen and Moon [24]	Interest function	Explicit ratings are required for training the system
Chorianopoulos et al. [6]	Implicit and generic user interactions with video player (seek/scrub).	Hard to capture the needed information from public video websites.
Peng et al. [25]	Eye tracking and face recognition	Requires an always on web camera

In the next Sections, we examine in more detail three indicative Case Studies in Social Video Retrieval. The selected Case Studies stand for the diversity of methodologies and research instrumentation found in the area of Social Video Retrieval. Therefore, we highlight the complementary research methods in each one the three case studies and we encourage the reader to elaborate into the detailed results by visiting the respective publication, which is indicated at the end of each case study.

3 Case Study 1: SocialSkip

SocialSkip [6] is an open-source Web-based system that collects and visualizes activity data from simple user interactions such as play/pause, seek/scrub. SocialSkip (Figure 2) employs few buttons, in order to be simpler to associate user actions with video semantics. We have simplified the standard random-seek bar to the Goforward and Gobackward buttons. 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 a popular seek window in previous research and commercial work due to the fact that it is the average duration of commercials. Furthermore we have observed replay functions and buttons in mobile devices such as Apple's iPhone and Safari QuickTime video players, which has the default time of 30 seconds as a replay.

We did not use a random seek timeline because it would be difficult to analyze users' interactions. Li et al. [16] observed that when seek thumb is used heavily, users had to make many attempts to find the desirable section of the video and thus caused significant delays. Drucker et al. [11] and Li et al. [16] tested different levels of speed for the functions of forward and rewind, too. User could make the choice of speed and locate more quickly the segment he wanted. For example, there have been commercial systems such as ReplayTV and TiVo that provide the ability to replay segments, or to jump forward in different speeds. Next to the player's button the current time of the video is shown followed by the total time of the video in seconds. Although we did not have a seek bar, we suggest that the data collected from the fixed skip could simulate the use of random seek, because any random seek activity can be modeled as a factor of fixed skipping actions (e.g., a random seek of 180 seconds is equal to 6 skips of 30 seconds).

In this case study, we selected three videos (lecture, how-to, documentary) that are as much visually unstructured as possible, because content-based approaches have already been successful with those videos that have visually structured scene changes.

In order to experimentally replicate user activity we added an electronic questionnaire that corresponds to a few segments of the video. According to Yu et al. [36] 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

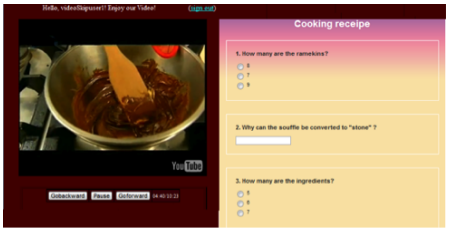

Fig. 2 SocialSkip Player is focused on skipping buttons, and questionnaire functionality

some interesting questions. In other words, it is expected that in a future field study, when enough user data is available, user behavior will exhibit similar patterns even if they are not explicitly asked to answer questions. The experiment took place in a lab with Internet connection, general-purpose computers and headphones. Twenty-five users spent approximately ten minutes to watch a video with all buttons muted, so they could not skip or pause. Next, there was a time restriction of five minutes, in order to motivate the users to actively browse through the video and answer the questions that corresponded to a few key-frames. 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.

In order to understand video pragmatics, we visualized the user activity data with a simple user heuristic. Firstly, we considered that every video is represented with an array with a length that equals the duration of the video in seconds. Next, we modified the value of each cell, depending on the type of interaction. For each Play, Pause and GoBackward, we increased the value. We decreased the value for each GoForward. In this way we have created the following activity graph (lecture video) that assist the understanding of video content, based on the pragmatics (user video browsing actions) rather than the content itself (Figure 3). The main benefit of this technique is that user interactions within a video have been transformed into a time-based signal, which might be further analyzed with techniques from signal processing.

In comparison to previous research, the proposed user activity heuristic is more malleable, because researchers can make various combinations and give different meaning to them. Yu et al. [36] experimental process used some questions to help mimic user interests and focus user behavior. Their algorithm should work with any video as long as it contains some commonly attractive content. SocialSkip has been developed with the same assumption. On the other hand, they have implemented a system with a custom video browsing applications. Peng et al. [25] have examined the physiological behavior (eye and head movement) of video users, in order to

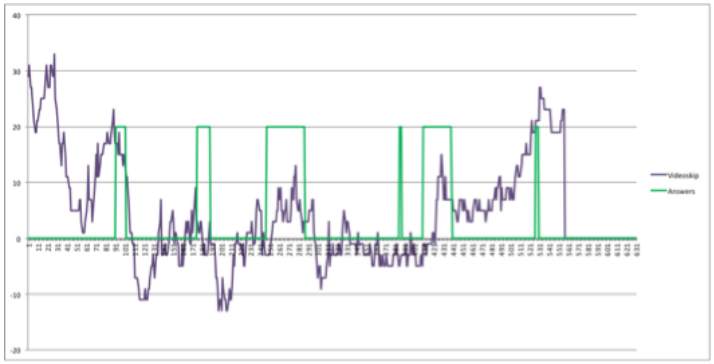

Fig. 3 The user activity graph provides a comprehensive visualization of cumulative user interactions and direct comparison to the experimentally defined ground-truth

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 contrast, the majority of users browses web video in more traditional ways that require no extra interactions or extra equipment, besides play, pause and seek, which are the main controls of SocialSkip.

4 Case Study 2: Viral Actions

In the case of Yahoo!'s Zync [28], two people in an IM conversation can watch an embedded video together in realtime; videos can be from Yahoo!, Flickr, or YouTube. The video stays in sync across two people and both individuals share control. In effect, the video becomes a reified synchronous context for the conversation. The chat and play, pause, and scrub behavior become one trace around the media object.

We argue that the implicit social sharing activity that occurs while sharing a video in a realtime IM conversation would result in more accurate predictions of a video's potential viewership. Implicit social sharing activity here refers to the number of times a video was paused, rewound, or fast-forwarded as well as the duration of the IM session while sharing a video. We believe that implicit social sharing activity is indicative of deeper and more connected sharing constructs, and hence better fidelity data to predict how much viewership a particular video is likely to attract. How a video is interacted with and shared between users is often indicative of how popular it is likely to be in the future. For instance, videos that have great appeal and potential to be popular will mostly likely be interacted with more and generate more conversation than others. Taken in aggregate across all users, patterns and 'signatures' [8] of interactions found in the implicit social sharing data can point to how popular and even viral a video is likely to be.

Viral videos are those that have gained outsized prominence and viewership as a result of an epidemic-like social transmission. In this case, we argue that the usage data that surrounds such viral videos can be used to predict the popularity of the video. Here we capitalize on the 'wisdom of the masses' by identifying patterns in the metadata to make predictions about the future popularity of that content [26].

New social media systems allow users to synchronously interact with each other and share videos simultaneously. These real-time interactions leave behind large amounts of contextual usage data that, we believe, are reflective of the deeper and more connected social interaction that accompanies synchronous content sharing. In this paper, we present a method of utilizing usage data from synchronously sharing videos to make predictions about the popularity of a particular video. In particular, we use play/pause behavior and chat volume pulled from a realtime video sharing environment, Zync (a plug-in for the Yahoo! Instant messaging (IM) client that allows participants to view and interact with a video simultaneously during a chat session). We argue that the usage data from synchronous video sharing tools provides robust data on which to detect how users are consuming and experiencing a

video. By extension, we can predict a video's popularity based on how it has been shared in a handful of sessions. To do this, we trained a Naïve Bayes classifier, informed by synchronous sharing features, to predict whether a video is able to garner 10 million views on it's hosting site. Our goal is to eventually predict a video's viral potential based on how its being shared.

The ability to socially share videos online has enabled select videos to gain a viewership of thousands in a very short period of time. Often, but not always, the-se videos take on a viral nature and gain tens of millions of views, while other videos only receive a fraction of the attention and viewing. These popular, viral videos also benefit from rich get richer dynamic where the more popular they be-come, the more views they are likely to attract. Viral videos attract not only a dis-proportionate amount of attention; they also consume greater amounts of resource and bandwidth as well. Thus, it would be helpful to be able to predict and identify which videos are most likely to go viral for recommendation, monetization, as well as, systems performance.

We acquired a 24-hour sample of the Zync event log for Christmas Day 2009. Zync allows two people to watch a video together in an instant message session; both participants share playback and control of the video and the video stays in sync across both participants IM windows, see Figure 4. The dataset provides a list of watched videos from YouTube as well as synchronous activity from the shared control of the video. These features are: anonymous user id hashes, session start/stop events, the session duration, the number of play commands, the number of pause commands, the number of scrubs (fast forwards or rewinds), and the number of chat lines typed as a character and word count. For the chat lines, the dataset contained no actual text content, only the aggregate count of characters, words and lines. The only textual content that is collected is video URLs and emoticons. Each activity collected is a row in the dataset and is associated with the time of the event and the playback time on the video.

The final test sample contained 1,580 videos with valid YouTube metadata and valid session data. The data collected from YouTube consisted of a video identifier, the video's title, its published date, its description, the genre category the uploader used, the tags the video was labeled with, the video's duration and the 5-star rating

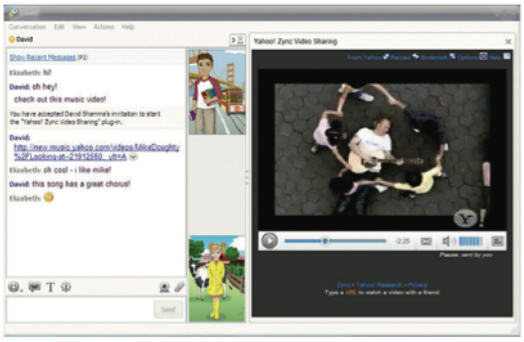

Fig. 4 The Zync plugin allows two Instant Messenger users to share a video in sync while they chat. Playback begins automatically; the users share control over the video

score it attained. Of these data, we only use the video's YouTube view count. For this case, the view count will be the predictive variable for the classifier and allows us to investigate if there's a match between YouTube view count and the synchronous social session actions.

As mentioned earlier, each single event from every session is a row in the dataset. This data needs to be aggregated into a feature vector for training a classifier. To do this, every session was divided into segments (sub-sessions) where a single video was viewed. This was necessary as many sessions contained multiple videos. The sub-sessions were then grouped by their representative video, mixing all the sessions that watched the same video. Finally, each event type and the overall sub-session durations were averaged into a single feature vector. Lastly, we assign a label indicating if the YouTube view count is over 10 million views; see Table 2 for some sample feature sets.

Table 2 Random and Naïve Bayes Prediction Accuracies. Random guess is calculated by using the distribution bias (6.3% of guessing Yes). The F1 score illustrates the overall performance accounting for both Precision and Recall. The Naïve Bayes predictions use crossfolded verification

Method	**Training Sample**	**Accuracy**	F_1 **Score**
Guessing	All Yes	6.3%	0.119
	Random	88.3%	0.041
	All No	93.7%	*NaN*[a]
Naïve Bayes	25%	89.2%	0.345
	50%	95.5%	0.594
	60%	95.6%	0.659
	70%	95.8%	0.778
	80%	96.6%	0.786

[a] Divide by zero.

In the model and results we have addressed our research question: we can predict the view count of a video based on how it is viewed in a rich, shared, synchronous environment. In total, 100 of the 1580 had over 10 million views. The Naïve Bayes classifier correctly identified 81 of these popular videos. There are far more videos that have less than 10 million views and thus higher prediction accuracy. It is important to note that our classifier produces a larger increase over a fair random prediction. We believe the session's duration to be the dominant feature in the predictive model as the correlation between Zync session duration and YouTube view count had the highest correlation ($p < 0.12$). While not quite significant, the average session duration in the feature vector is completely independent of the view count. Furthermore, there is no significant or near significant correlation between the session duration and the video's playback time. Similarly, no significant correlations were observed within the other meta-data from YouTube (ratings and playback time) and the YouTube view count.

5 Case Study 3: Less Talk, More Rock

In this case study, we highlight the methods that we employed in a study of user generated music video recordings [14]. The availability of video capture devices and the high reach and impact of social video sharing sites like YouTube make video content from live shows relatively easy to share and find [9]. Users of YouTube share millions of videos capturing live musical performances, from classical pianist Ivo Pogorelich to metal rockers Iron Maiden. Potentially, such an abundance of content could enable comprehensive and deeper multimedia coverage of captured events. However, there are new challenges that impede this new potential: the sample scenario above, for example, illustrates issues of relevance, find-ability, and redundancy of content.

The lack of detailed metadata associated with video content presents several interesting challenges. First, with no accurate, semantic event-based metadata, it is not trivial to automatically identify a set of video clips taken at a given event with high recall and precision. Second, with no dependable time-based metadata associated with the clips, aligning and synchronizing the video clips from the same event cannot be done using simple timestamps.

In this case study, we report on an approach for solving the synchronization problem, and how we leverage the synchronization data to extract additional metadata. The metadata would help us organize and present video clips from live music shows. We start by assuming the existence of a curated set of clips, having already identified the video clips from each event.

We use audio fingerprinting [33, 13] to synchronize the content. In other words, we use the clips' audio tracks to detect when the same moment is captured in two different videos, identify the overlap, and specify the time offset between any pair of overlapping clips. The synchronization of clips allows us to create a novel experience for watching the content from the event, improving the user experience and reducing the redundancy of watching multiple clips of the same moment. Figure 5 presents one possible viewing interface.

Once synchronized, we use both the relative time information and links between overlapping clips to generate important metadata about the clips and the event. First,

Fig. 5 A sample interface for synchronized playback of concert video clips

we show how we identify the level of interest [23] and significant moments in the show as captured by the users. Second, we mine the tags of videos associated with a single point in time to extract semantically meaningful descriptive terms for the key moments in the show; these terms can be used to represent or explain the aggregated content. Third, we use the link structure created by the audio fingerprinting when a clip matches another to find the highest-quality audio recording of any time segment, given multiple overlapping recordings.

The clip overlap structure, created by the community activity, can help identify moments in an event that are likely interesting to consumers of content [23]. In particular, we hypothesize that the segments of concerts that are recorded by more people might be of greater appeal to content consumers. Identifying these segments can be helpful for search, summarization, keyframe selection [7] or simple exploration of the event media. Videos of the most important segments or other aspects of the concert could be highlighted, while filtering lower-scoring clips that are either unrelated or, presumably, less interesting.

Our hypothesis is that larger clusters of matches between clips typically correspond to segments of the concert that are subjectively most 'interesting.' In the case of live music, these clusters could reflect significant moments in the show where a hit song is being played, or something particularly interesting is happening on stage.

We use the synchronization data to select the highest quality audio for each overlapping segment. The synchronization between video clips can be used for playback, remixing or editing content. Inevitably, given the nature of user-generated recordings, the video and audio quality and content can be highly variant between clips as well as from minute-to-minute within clips. Interestingly, low-quality audio tracks cause the audio fingerprinting method to fail in systematic ways that can be leveraged to point us towards higher-quality recordings.

We aggregate the textual information associated with the video clips based on the cluster structure to extract descriptive themes for each cluster. On many social media websites, users often provide lightweight annotations for the media in the form of titles, descriptions, or tags. Intuitively, if the overlapping videos within our discovered clusters are related, we expect the users to choose similar terms to annotate their videos such as the name of the song being captured or a description of the actions on stage. We can identify terms that are frequently used as labels within a given cluster, but used relatively rarely outside the cluster. These terms are likely to be useful labels / descriptions for the cluster, and can also be used as suggested metadata for unannotated clips in that cluster

We have applied our system to a large set of real user-contributed videos from three concerts crawled from the popular video sharing website, YouTube. Each concert collection contains several hundred video clips, providing for a total of just over 600 clips and more than 23 hours of video footage. The three concerts that we have investigated are: Arcade Fire in Berkeley, CA; Daft Punk in Berkeley, CA; and Iron Maiden in Bangalore, India. All three concerts occurred during the spring or summer of 2007.

We find that the proposed method is able to identify clusters of videos taken at the same point in time, with a near-perfect precision up to a recall in the range of

20% - 30%, which is sufficient for many discovery and browsing applications. We compare the number of people recording each song in a concert against the number of plays for the song on social music site last.fm and find a significant correlation between the number of plays and the size of the cluster ($r^2 \sim .44$, $p < 0.001$, $N = 41$), suggesting that the number of people recording is a reasonable estimate of the level of interest. We ask human subjects to score the relative audio quality of various segments and find that our system for scoring audio quality significantly correlates with human assessments ($r^2 \sim .26$, $p < 0.001$, $N = 50$). Finally, we find that repeated text terms in the videos associated with a moment in the concert often correspond to words from the names of the songs or to actions taking place on stage.

Our primary focus in this work is an in-depth exploration of the different methods, rather than building and evaluating a browsing system. We shift the focus of our system from developing matching algorithms and focus on mining the structure of the discovered overlaps and audio re-use to create compelling new ways of aggregating and organizing community-contributed Web data. The ideas above could be used in the design and implementation of a system for sharing live concert videos and content. We would also imagine such an application to elicit more accurate or structured metadata and contributions from users, contributions that might exceed and extend the social media tools available on YouTube.

6 Directions for Further Research

In this section, we make suggestion for further research, which has been organized according to the research instrumentation and method employed in the above Case Studies.

First, 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. The above research issues might be addressed by means of signal processing techniques on the user activity signal.

Secondly, we have demonstrated the possibility to a classifier, based on social synchronous sharing patterns, to predict if a video has a high view count. Our goal in this research is to predict if a video will go viral based on how it is shared within a conversation. The successful predictions in our classifier are based on most videos (85%) viewed once in only one session. The next step in this work is to collect vari-

ous data samples over time and investigate how a video's implicit sharing behaviors change as it becomes viral. In effect, this is somewhat of a fishing exercise over time; we need to collect data on videos as they turn viral to train a classifier on how to predict them. We expect the temporal changes between the feature vectors (the deltas and rate of change across our video feature vectors) to enable accurate viral predictions for recommendations. Additionally, when socially filtered, unique viral patterns found in some social groups and networks could bring socially targeted recommendations and content pro- motion.

Finally, further research in dynamic editing of live video events should consider the fusion of context provided by social media sites, and content analysis. In particular, the combination of content-based (e.g., audio or video) matching with metadata (e.g., tags, comments) from social media might create a better representation of the content. For example, audio fingerprinting features can be extracted only for clips identified as events using the social media context, and clips are pairwise compared only within a single show, significantly reducing the required scale as well as potentially improving precision. This approach for multimedia analysis leveraging social media contribution promises to change the way we consume and share media online.

It is important to underscore there is a use for traditional content analysis to discover the social interaction. While we suggest social interaction analysis can supersede many content analysis techniques, content techniques can identify and connect disassociated social actions, providing a reified context.

7 Conclusion

As long as the community of users watching videos on social video systems is growing, more and more interactions are going to be gathered and therefore, we are going to have a better understanding of a video according to evolving user interests. We also expect that the combination of richer user profiles and content metadata provide opportunities for personalization. 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 contrast to content-based video retrieval, we have employed few videos in the research methods of the case studies. Previous work on video retrieval has emphasized the large number of videos, because the respective algorithms treated the content of those videos. In this user-based work, we are not concerned with the content of the videos, but with the user activity on videos. Nevertheless, it is worthwhile to explore the effect of more videos and interaction types. Therefore, the small number of videos used in the case studies is not an important limitation, but further research has to elaborate on different genres of video (e.g., news, sports, comedy, music, lecture) and on the number of user interactions that are necessary to obtain meaningful user activity patterns.

The methodological approaches of the three case studies provide a balance between two very different research philosophies: the employment of big natural data

versus the design of controlled user experiments. We suggest that data mining on a large-scale web-video database is the most effective approach, because the data and the techniques have high external validity. Nevertheless, we found that the experimental approach is very flexible during the development phase of a new system. Moreover, the iterative and experimental approach is very suitable for user-based information retrieval, because it is feasible to associate user behavior to the respective data-logs.

Although we suggest the employment of user-based video retrieval techniques, we have also considered the benefits of content-based ones. 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, because they produce the same result all the time, but they are also very efficient in the analysis of new videos that do not have any interactions, such as pause, rewind, or sharing with other users. In contrast, the community (or crowd-sourced) intelligence of implicit user activity with web video is dynamic (e.g., scrubs, comments, remixes), because it continuously adapts to evolving users' preferences, but it is also more difficult to analyze and evaluate. In the end, we expect that a balanced mix of hybrid algorithms (content-based and user-based) might provide an optimal solution for editing, sharing, and navigating through video content on social media Web sites and applications.

References

1. Baecker, R., Rosenthal, A.J., Friedlander, N., Smith, E., Cohen, A.: A multimedia system for authoring motion pictures. In: Proceedings of the fourth ACM international conference on Multimedia, MULTIMEDIA '96, pp. 31–42. ACM, New York, NY, USA (1996). DOI 10.1145/244130.244142. URL http://doi.acm.org/10.1145/244130.244142
2. Carlier, A., Charvillat, V., Ooi, W.T., Grigoras, R., Morin, G.: Crowdsourced automatic zoom and scroll for video retargeting. In: Proceedings of the international conference on Multimedia, MM '10, pp. 201–210. ACM, New York, NY, USA (2010). DOI 10.1145/1873951.1873962. URL http://doi.acm.org/10.1145/1873951.1873962
3. Cesar, P., Chorianopoulos, K.: The evolution of tv systems, content, and users toward interactivity. Found. Trends Hum.-Comput. Interact. **2**(4), 373–95 (2009). DOI 10.1561/1100000008. URL http://dx.doi.org/10.1561/1100000008
4. 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: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, IMC '07, pp. 1–14. ACM, New York, NY, USA (2007). DOI http://doi.acm.org/10.1145/1298306.1298309. URL http://doi.acm.org/10.1145/1298306.1298309
5. Cheng, X., Dale, C., Liu, J.: Statistics and social network of youtube videos. In: Quality of Service, 2008. IWQoS 2008. 16th International Workshop on, pp. 229 –238 (2008). DOI 10.1109/IWQOS.2008.32
6. Chorianopoulos, K., Leftheriotis, I., Gkonela, C.: Socialskip: pragmatic understanding within web video. In: Proceddings of the 9th international interactive conference on Interactive television, EuroITV '11, pp. 25–28. ACM, New York, NY, USA (2011). DOI http://doi.acm.org/10.1145/2000119.2000124. URL http://doi.acm.org/10.1145/2000119.2000124

7. Christel, M.G., Hauptmann, A.G., Wactlar, H.D., Ng, T.D.: Collages as dynamic summaries for news video. In: Proceedings of the tenth ACM international conference on Multimedia, MULTIMEDIA '02, pp. 561–569. ACM, New York, NY, USA (2002). DOI 10.1145/641007.641120. URL http://doi.acm.org/10.1145/641007.641120
8. Crane, R., Sornette, D.: Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In: Proc. of AAAI symposium on Social Information Processing, Menlo Park, CA (2008)
9. Cunningham, S.J., Nichols, D.M.: How people find videos. In: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, JCDL '08, pp. 201–210. ACM, New York, NY, USA (2008). DOI 10.1145/1378889.1378924. URL http://doi.acm.org/10.1145/1378889.1378924
10. Davis, M.: Human-computer interaction. In: R.M. Baecker, J. Grudin, W.A.S. Buxton, S. Greenberg (eds.) Readings in Human-Computer Interaction: Toward the Year 2000, chap. Media Streams: an iconic visual language for video representation, pp. 854–866. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995). URL http://dl.acm.org/citation.cfm?id=212925.213009
11. Drucker, S.M., Glatzer, A., De Mar, S., Wong, C.: 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, CHI '02, pp. 219–226. ACM, New York, NY, USA (2002). DOI 10.1145/503376.503416. URL http://doi.acm.org/10.1145/503376.503416
12. Girgensohn, A., Boreczky, J., Wilcox, L.: Keyframe-based user interfaces for digital video. Computer **34**(9), 61–67 (2001). DOI 10.1109/2.947093. URL http://dx.doi.org/10.1109/2.947093
13. Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system with an efficient search strategy. Journal of New Music Research **32**(2), 211–221 (2003). DOI 10.1076/jnmr.32.2.211.16746
14. Kennedy, L., Naaman, M.: Less talk, more rock: automated organization of community-contributed collections of concert videos. In: Proceedings of the 18th international conference on World wide web, WWW '09, pp. 311–320. ACM, New York, NY, USA (2009). DOI http://doi.acm.org/10.1145/1526709.1526752. URL http://doi.acm.org/10.1145/1526709.1526752
15. Levy, P.: Collective Intelligence: Mankind's Emerging World in Cyberspace. Perseus Books, Cambridge, MA, USA (1997)
16. Li, F.C., Gupta, A., Sanocki, E., He, L.w., Rui, Y.: Browsing digital video. In: Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '00, pp. 169–176. ACM, New York, NY, USA (2000). DOI 10.1145/332040.332425. URL http://doi.acm.org/10.1145/332040.332425
17. Lienhart, R., Pfeiffer, S., Effelsberg, W.: Video abstracting. Commun. ACM **40**(12), 54–62 (1997). DOI 10.1145/265563.265572. URL http://doi.acm.org/10.1145/265563.265572
18. Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: Proceedings of the 15th international conference on Intelligent user interfaces, IUI '10, pp. 31–40. ACM, New York, NY, USA (2010). DOI 10.1145/1719970.1719976. URL http://doi.acm.org/10.1145/1719970.1719976
19. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on Multimedia, MULTIMEDIA '02, pp. 533–542. ACM, New York, NY, USA (2002). DOI 10.1145/641007.641116. URL http://doi.acm.org/10.1145/641007.641116
20. Mitra, S., Agrawal, M., Yadav, A., Carlsson, N., Eager, D., Mahanti, A.: Characterizing web-based video sharing workloads. ACM Trans. Web **5**(2), 8:1–8:27 (2011). DOI 10.1145/1961659.1961662. URL http://doi.acm.org/10.1145/1961659.1961662
21. Money, A.G., Agius, H.: Video summarisation: A conceptual framework and survey of the state of the art. J. Vis. Comun. Image Represent. **19**(2), 121–143 (2008). DOI 10.1016/j.jvcir.2007.04.002. URL http://dx.doi.org/10.1016/j.jvcir.2007.04.002

22. Naci, S.U., Hanjalic, A.: Intelligent browsing of concert videos. In: Proceedings of the 15th international conference on Multimedia, MULTIMEDIA '07, pp. 150–151. ACM, New York, NY, USA (2007). DOI 10.1145/1291233.1291264. URL http://doi.acm.org/10.1145/1291233.1291264
23. Nair, R., Reid, N., Davis, M.: Photo loi: browsing multi-user photo collections. In: Proceedings of the 13th annual ACM international conference on Multimedia, MULTIMEDIA '05, pp. 223–224. ACM, New York, NY, USA (2005). DOI 10.1145/1101149.1101187. URL http://doi.acm.org/10.1145/1101149.1101187
24. Olsen, D.R., Moon, B.: Video summarization based on user interaction. In: Proceddings of the 9th international interactive conference on Interactive television, EuroITV '11, pp. 115–122. ACM, New York, NY, USA (2011). DOI 10.1145/2000119.2000142. URL http://doi.acm.org/10.1145/2000119.2000142
25. Peng, W.T., Chu, W.T., Chang, C.H., Chou, C.N., Huang, W.J., Chang, W.Y., Hung, Y.P.: Editing by viewing: Automatic home video summarization by viewing behavior analysis. Multimedia, IEEE Transactions on **13**(3), 539 –550 (2011). DOI 10.1109/TMM.2011.2131638
26. Segaran, T.: Programming Collective Intelligence: Building Smart Web 2.0 Applications, 1 edn. O'Reilly Media (2007). URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0596529325
27. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: Understanding community annotation of uncollected sources. In: WSM '09: Proceedings of the international workshop on Workshop on Social Media. ACM, Beijing, China (2009)
28. Shamma, D.A., Kennedy, L., Churchill, E.F.: Viral actions: Predicting video view counts using synchronous sharing behaviors. In: ICWSM 11: Proceedings of the International Conference on Weblogs and Social Media Data. AAAI Press, Barcelona, Spain (2011)
29. Shaw, R., Davis, M.: Toward emergent representations for video. In: MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, pp. 431–434. ACM, New York, NY, USA (2005). DOI http://doi.acm.org/10.1145/1101149.1101244
30. Shaw, R., Schmitz, P.: Community annotation and remix: a research platform and pilot deployment. In: HCM '06: Proceedings of the 1st ACM international workshop on Human-centered multimedia, pp. 89–98. ACM Press, New York, NY, USA (2006). DOI http://doi.acm.org/10.1145/1178745.1178761
31. Snoek, C., Worring, M., Smeulders, A., Freiburg, B.: The role of visual content and style for concert video indexing. In: Multimedia and Expo, 2007 IEEE International Conference on, pp. 252–255. IEEE (2007)
32. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. **3**(1) (2007). DOI 10.1145/1198302.1198305. URL http://doi.acm.org/10.1145/1198302.1198305
33. Wang, A.: The shazam music recognition service. Commun. ACM **49**(8), 44–48 (2006). DOI 10.1145/1145287.1145312. URL http://doi.acm.org/10.1145/1145287.1145312
34. Yan, R., Hauptmann, A.G.: A review of text and image retrieval approaches for broadcast news video. Information Retrieval **10**, 445–484 (2007). DOI 10.1007/s10791-007-9031-y
35. Yew, J., Shamma, D.A., Churchill, E.F.: 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, pp. 297–306. ACM, New York, NY, USA (2011). DOI http://doi.acm.org/10.1145/1978942.1978984. URL http://doi.acm.org/10.1145/1978942.1978984
36. Yu, B., Ma, W.Y., Nahrstedt, K., Zhang, H.J.: Video summarization based on user log enhanced link analysis. In: Proceedings of the eleventh ACM international conference on Multimedia, MULTIMEDIA '03, pp. 382–391. ACM, New York, NY, USA (2003). DOI 10.1145/957013.957095. URL http://doi.acm.org/10.1145/957013.957095
37. Zhang, D., Guo, B., Yu, Z.: The emergence of social and community intelligence. Computer **44**(7), 21 –28 (2011). DOI 10.1109/MC.2011.65