EXPLORING AND EVALUATING THE EFFECTS OF USER-ENHANCED VIDEO BROWSING

by

ROIY SHPANER

A thesis submitted to the Faculty of Graduate Studies of
the University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada

ABSTRACT

In light of the massive growth of creation and consumption of video, in this thesis I explore the concept of user-enhanced video browsing and evaluate the quantitative and qualitative effects of this approach. For this purpose, I create a unique prototype based on the VLC media player. The player interprets user-generated video tagging and annotations in my designed format, and allows the viewing of multiple event layers to create a personalized video playback. I perform evaluations in two user studies for this work. Among my observations I find benefits in navigation, personalization, consumption, and comprehension. I then look at the way viewers behave when contributing data of annotations and tagging. I find their preferred tasks, their perceived quality of other contributions, as well as their opinions on this system. I conclude with a discussion of the results and list possible use-cases for this concept.

ACKNOWLEDGMENTS

I must admit that somehow the acknowledgments section seems to be my favorite part in every thesis I read. So I may go a bit overboard here.

I had a very eye opening experience during my Masters program, and I'm happy to have learned so much. Not only about my thesis topic, but also about research, science, and the scientific method. I found very talented people that spend a large part of their lives trying to improve the quality of life on this planet, and clearly not for the sake of money. I got a chance to see and try the very cutting edge of technology powered by the most recent human knowledge and capabilities.

There is one person that I need to thank most and that is my advisor, Pourang. You believed in me from the very beginning, starting with a blind videochat over two years ago, and your faith only grew stronger with time. I greatly appreciate your support in my path and passion, even when it meant stepping slightly out of the road you are familiar with. You have taught me a great deal, and I'm very glad to have been given the chance to join your exciting lab.

I would like to acknowledge Dr. Marcos Serrano who worked with me in the early stages of this project. We didn't always see eye to eye, but Marcos provided me with focus and guidance about research. I do thank him for his time, efforts and feedback.

My friends and lab-mates: Reza, Srikanth, Juan, Ashik, Amir, Hina, and all the rest that I can't fit in this page. The social environment always determines the experience, and mine was an incredible one. Thank you all for making the ride fantastic and my cold new home a happy one.

I also thank my committee members, Dr. Carson Leung and Dr. Herbert Enns for their time and constructive feedback on this work.

And finally, one Chinese girl, that has made me consider scenarios I never thought I would. While our journey ended, the memories are timeless and priceless. Yan, I wish on you what you brought for me: a whole lot of happiness.

This thesis is dedicated to myself. And I must say it's a great gift.

P.S.: At the time of writing this "This thesis is dedicated to myself" yields 6 results on google compared with 160,000 for "This thesis is dedicated to my parents"

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Video data has exploded in recent years with the rise of user-generated content through smartphones, tablets and the increasing use of video sharing websites. It is not only the sheer number of videos that has grown rapidly, but also their resolution, size and length. The average duration of an online video was 45 seconds in 1997, and has risen to 5.6 minutes in 2013 [9](Figure 1). As videos become longer users require more efficient ways of reaching the sections they want to see (and of skipping the uninteresting segments). Statistics from Wistia, a video hosting service that serves millions of files, show that on average, users watch only 30 percent or less of the content in videos that are over 30 minutes long. In contrast, viewers watch 60 percent and more of those that are up to 5 minutes long [28]. Whether the reason lies in frustration, lack of time, or impatience, we see users either give up watching long videos or abandon them midway. One promising way to improve retention is to allow viewers more control in navigating the content.

1.1 MOTIVATION

Many video browsing and navigation techniques have been studied over the years, but they have primarily focused on algorithms for

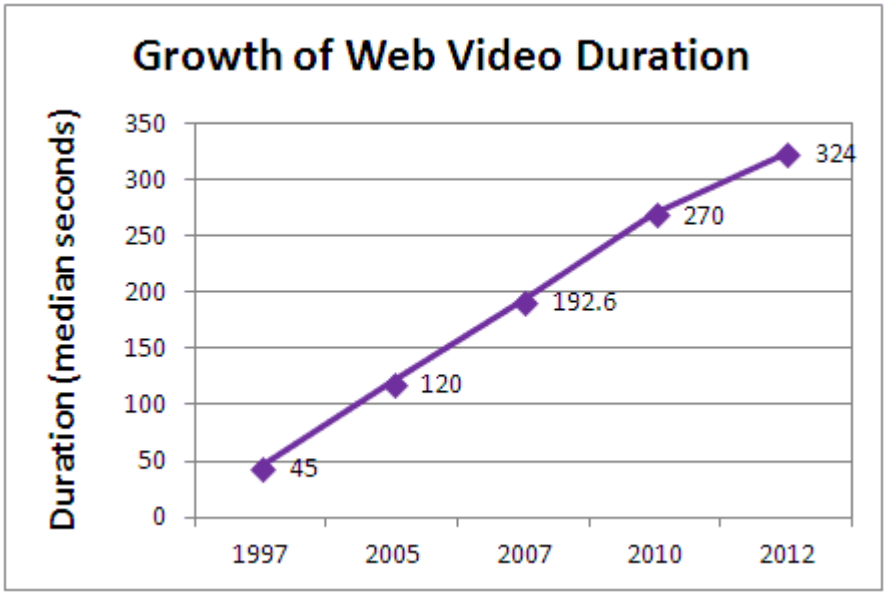

Figure 1: Web video length growth. Source: Comscore.com, WebsiteOptimization.com

automatic scene detection based on visual [8] or audio cues [25], or by analyzing a video transcript [16, 14]. These approaches ignore two important aspects of video browsing. First, interest is a subjective matter, and current automatic methods are not yet able to predict what a person would be interested in, or the kind of information they would be looking for. Second, the semantics of the video are often not extractable without the aid of a human who can offer a deeper understanding of the content. One promising way to solve these issues is to use the knowledge of previous viewers of the video contents to improve the browsing experience.

In this thesis I investigate if user-enhanced videos offer a solution to the browsing problem, and get supporting quantitative and qualitative data for other potential effects from two user studies. The user-enhanced browsing concept studied here has a unique personalization aspect (by combining multiple event layers) and uses

previously suggested viewing techniques. The system is based on tagging (i.e., segmenting parts of a video as events that are added to relevant event layers) and annotations (i.e., free form of text made by users or producers) of events. The resulting event layers can then be added according to preference during playback, in order to create a custom viewing sequence. This sequence is in essence a sub-video inside the video that skips unselected topics or events and plays only the subjectively relevant parts for the user.

1.2 CONTRIBUTIONS

My primary contribution in this thesis is the exploration and evaluation of the user-enhanced video concept. My process is structured in steps: first, I introduce a new simple and human-readable metadata file format that allows playing user-enhanced videos in a personalized way. I implement the concept and add support for the enhancement data format in the popular VLC media player (Figure 2). I then show in two user studies that the use of tagging and annotations provides significant advantages (and certain challenges) for video browsing. In my first study I ask subjects questions about the content of four different videos. I measure the completion time, error rate, and confidence levels. In the second study I look at the creation and consumption aspect of the concept. I measure the effects of having participants iteratively contribute and modify the tags and annotations. I also collect user feedback on the browsing experience in both of my studies. I then describe my efforts in integrating my work into

the official VLC media player. Finally, I summarize the effects that were found and suggest potential uses for this kind of system.

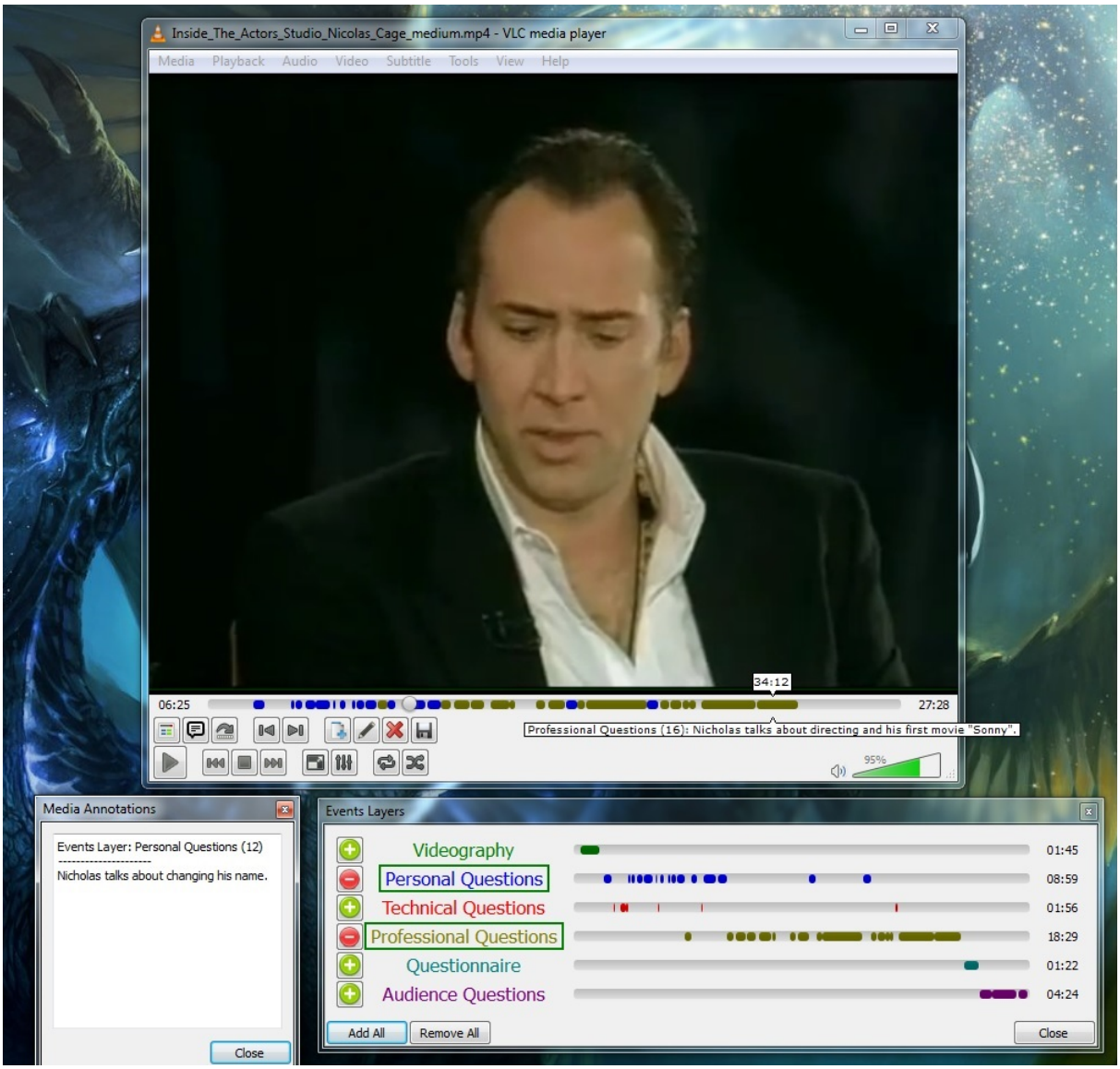

Figure 2: The modified VLC player (using the new MED format). The different event type layer names are visible, along with the related events that were tagged for them. Two layers are chosen by the user (in green boxes) for this interview video. This allows the viewer to only watch the relevant segments for a shorter and more focused viewing (other parts are automatically skipped). Annotations are displayed on the left window.

2

RELATED WORK

Video research has been extensive in the last few decades, and this work relates to several different areas. In this chapter I bring the main points of the most relevant topics.

2.1 VIDEO NAVIGATION AND FAST VIEWING

In recent years we have seen more effective and informative ways of watching videos. Cheng et al. [8] proposed a video player that increases the playback speed during periods when little or no motion is detected. The idea is that static parts of the video would be less interesting to viewers, which is often the case, but does present challenges with videos that are audio based like news clips. A similar approach was attempted by Kurihara [16] who proposed a player that speeds up during scenes that do not involve speech after analyzing a video transcript. This audio based solution works well for many types of videos, but cannot improve clips like surveillance, etc.. These techniques use cues about the video, which allow the player to react in a fairly limited way, and their results depend mostly on the suitability of the type of video being evaluated.

An approach that allowed viewers to navigate videos using a 3D summary cube was developed by Nguyen et al. [20]. This method

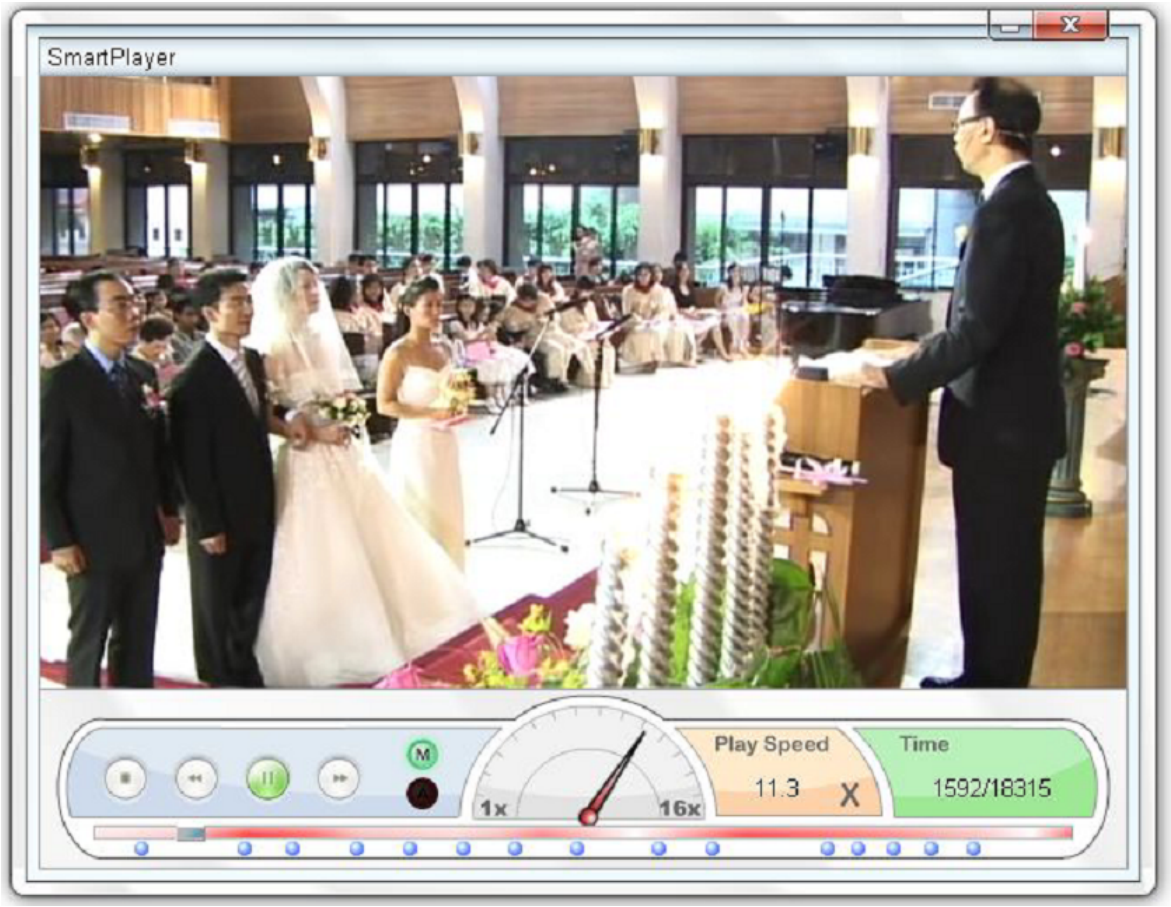

Figure 3: SmartPlayer [8] interface. Playback would speed up during static segments of the video.

relies mostly on the video being shot with specific attributes (static camera, panning motion, etc.) that allow the creation of a straightforward shot aggregation. For many years, one of the most common ways to navigate video has been the static storyboard style which uses key frames of the video as hotspots for navigation. This method and several others like motion visualization, and a moving storyboard were used in the recent Video Explorer by Schoeffmann et al. [10].

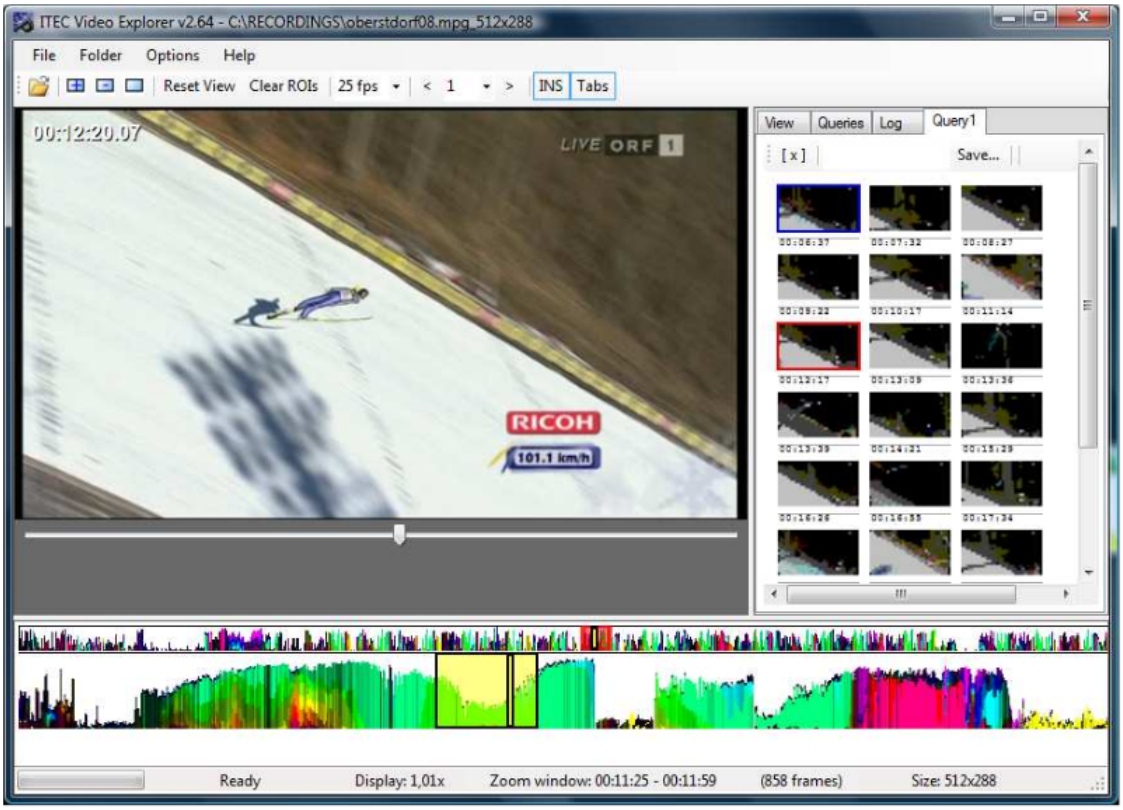

Figure 4: Video Explorer [10] interface. This is a query for a visual sequence in the video.

2.2 INTERACTIVE VIDEO

Attempts have been made to make a transition from passive consumption of media to an active or hybrid ways of control. The work by Smith [24] from the iCinema group looks into the categorization of video shots by properties like scene motion and displayed objects, and allows viewers to watch a looping sequence of available media that fits their desired query. Thus transforming traditional passive media viewers to active "viusers". This approach was implemented in the experimental artwork T_Visionarium [11] by Favero et al. This interactive VR theatre consists of a cylinder with a projector that presents the video according to the user's point of view, which

allows navigating the dataset by looking at a different part of the cylinder. A remote that accepts search keywords as input is used in order to select the desired stream. While this thesis does not directly use the work of iCinema, it builds on the concept of shifting the paradigm from a completely passive viewing experience to deeper involvement, customized browsing, and easier navigation.

Figure 5: Inside the T_Visionarium [11]

A recent example of interactive video is the 2013 JAM! [5] project by Caille et al. These applications focus the interactive analysis of recorded theater performances. The JAM! Dual Players application allows the user to play two simultaneous performances at the same

time, along with the annotated lines of the performance text. This helps the analyser compare different versions of the same production in an easy synchronized fashion, jump to the relevant scenes, and spot similiarities and differences in elements of the actors and cinematographers work.

Figure 6: The JAM! Dual Player [5] running two synchronized videos

2.3 USER-ENHANCED BROWSING

While the previously mentioned works do show navigation improvement, they only do so with very specific videos. What they lack is any kind of semantic analysis to inform the user of the different logical scenes in the video. Content in these videos is also rather one dimensional, and does not take into account the viewer's input. To overcome these issues, research has been done on enhancing videos with user-based tagging and annotations. CWaCTool [12]

is an example of such a tool; textual, audio, and even virtual ink drawing can be added on top of a video to create a richer experience for future viewers. One attempt to examine the effects of user annotations used as a means of communication was made by Mukesh et al. in their work on CollaboraTV [19]. Their system allowed users to watch videos asynchronously with friends. Viewers would add comments to certain points in the video, and these would later be displayed when their friends watch the video to create a more social experience.

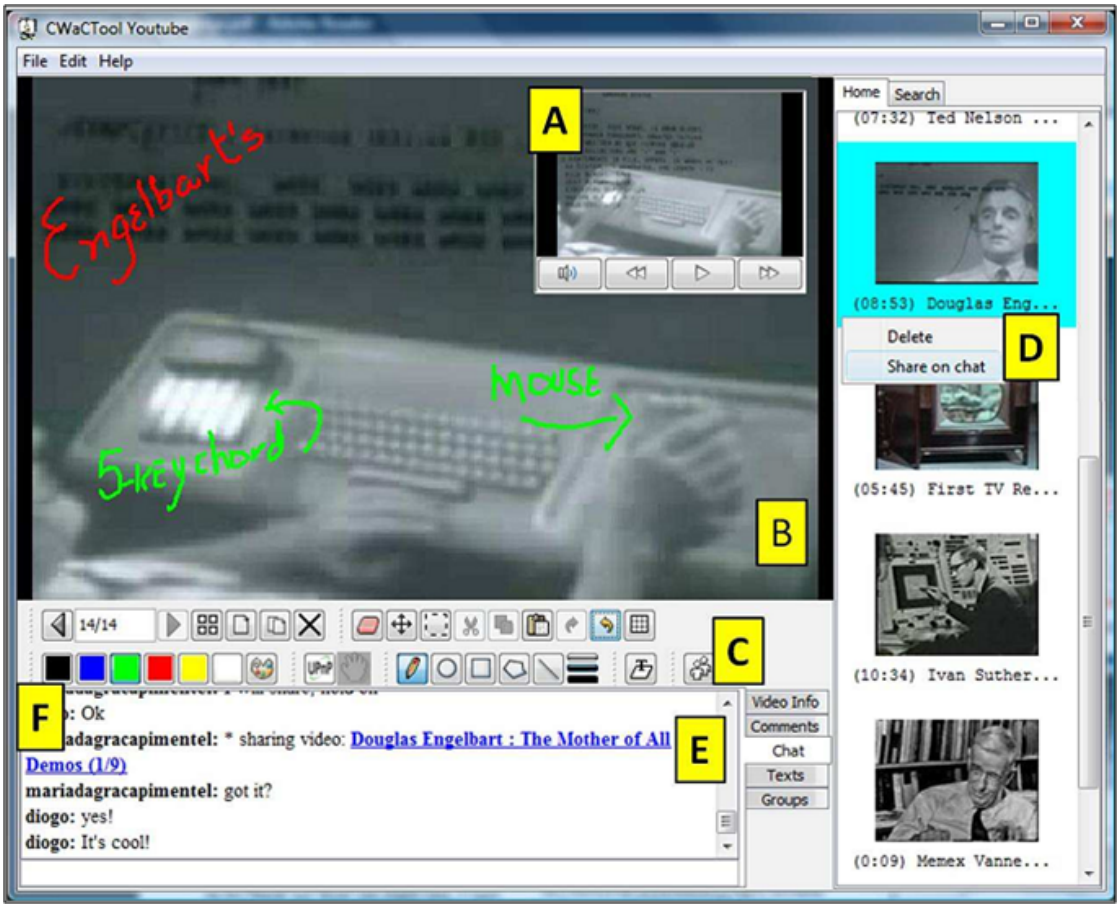

Figure 7: CWaCTool [12]. Allows for text, audio, and drawing on the playback.

The utilization of user-based metadata in video navigation has been studied in work done on the Advene player by Aubert and Prie [2]. Advene, a heavyweight viewer which is targeted for professional video users, allows very rich information to be added to the video by users, such as HTML, graphical content, as well as simple text. Advene supports a layered approach to viewing the video; however it does not let the user watch events of more than one layer at a time, limiting its flexibility and personalization. Advene and other user-enhanced players [22, 18, 17, 7] have not been thoroughly evaluated to find the extent of the potential benefits this kind of system has over traditional players. I believe this exploration is critical for the widespread adoption of this concept, as well as its progression and development as a preferred method of video browsing.

Figure 8: CollaboraTV [19]. Allows for a more social viewing experience with time-based comments.

2.4 CROWDSOURCED VIDEO DATA

To create the information needed for user-enhanced browsing, the most suitable available method is crowdsourcing. With the adoption of the internet and user-generated content, the prospect of using the wisdom of the crowd to improve video viewing became much more attainable. One work that explored this direction is EpicPlay [26], in which Twitter activity of viewers was used to automatically create highlights of a Football game video footage. Another example was demonstrated by Carlier et al. [6], the authors asked users to mark the interesting areas in a zoomable video, so they can be more easily viewed by subsequent viewers. But the two previous papers utilized users in an indirect way; in the work of Riek et al. [23], users were directly asked about the content and context of the video as part of a game. The answers from the users helped to add information about the video and made it easier to tag or watch. Finally, user-generated content is still faced with the difficult problem of maintaining a standard of quality. However, data from a recent paper by Park et al. [21] on crowdsourced video annotations shed some new light on the issue. In this paper the authors investigated close-ended annotations and compared the quality of expert-created annotations versus those created by crowd workers. The authors reached the conclusion that given 3 or more crowdsourced annotations, the level of quality is roughly equal to an expert's contribution. This gives us more reason to think user-enhanced browsing is applicable, feasible and worth exploring further.

But despite the growing use of crowdsourcing in videos, it appears that we have not fully utilized all of the potential benefits with such methods. At the very least it is still not clear how users behave when creating metadata collaboratively, and how they perceive this type of viewing. This work tries to uncover these details and to support the foundation of related research in the field.

3

FORMAT AND IMPLEMENTATION

The solution I had in mind for the video challenges of navigation, information, and customization was a media player that took advantage of the knowledge of previous viewers. The ability to tag events in the video into relevant layers and play the combination of layers you wish to watch was a promising (but not very researched) way to handle the mentioned issues. However, no current player provided all of the features I needed, and I also needed to choose a way of saving the generated metadata. The file format had to allow the use of annotations, as well as the multi-layered playback mechanism for the events. There were existing formats with some of the necessary requirements (e.g., MPEG-7 [13] or SMIL [27]), but these formats were too cumbersome, general-purposed, and lacked features like Event Id and Event Layers. It is also interesting to note that none of the previously mentioned formats really took off with the public, and they are not supported in any major media player. To increase the odds of integration to existing media players, I decided to create a new media metadata format that is lean, simple, human-readable, similar to supported formats, and extensible.

3.1 MEDIA ENHANCEMENT DATA FORMAT

After considering the different options, I decided to use a separate file for the metadata instead of encoding it into the video file. The reason being that this data is meant to be modified by viewers, and it would not make sense to resend or upload an entire video file because of a small change to the metadata. In this design only a small text file needs to be resent so it can be quickly shared among viewers or on a website. I elected to base the structure principles on the popular subtitles format SubRip [1](figure 9 shows the comparison between SubRip and my format). SubRip is the most supported subtitles format by available media players, and is easy to edit even when using only a text editor. It is a simple text based description of time segments and the relevant subtitle data.

I have changed the SubRip format to include a line with the event layer name and the event id. Instead of 2 lines of subtitles I defined a reserved location for free-text annotations that may span multiple lines (and may also be omitted for a non-annotated event). The way to separate different events is by incorporating 2 consecutive empty lines between them. These changes enable the tagging of events to specific types with commonalities, and subsequently the control of entire layers of events in a media player. I chose to call this format Media Enhancement Data (MED file extension).

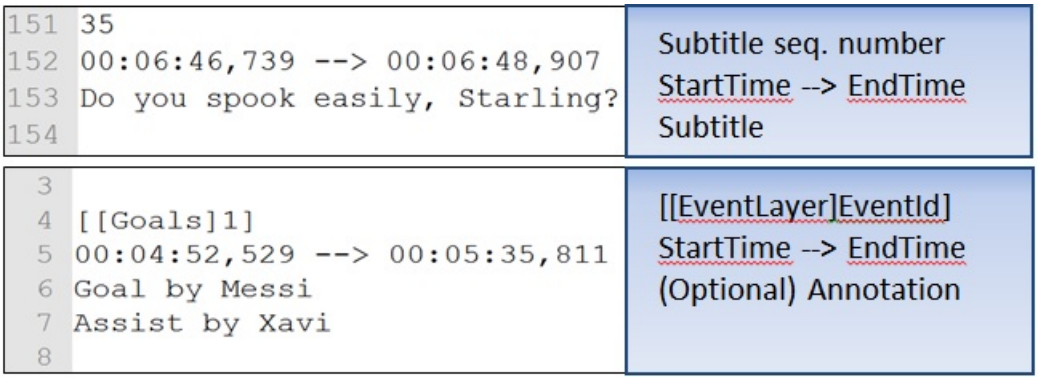

Figure 9: Top left: example of the SubRip (.srt) subtitles format. Top right: basic subtitle structure in the SubRip format. Bottom left: example of the Media Enhancement Data (.med) format. Bottom right: basic event structure in the MED format.

3.2 IMPLEMENTATION

For the playback requirements, I wanted to extend an existing open-source media player with the features we needed in order to realize the user-enhanced concept. I selected the VLC media player because of its extensibility, large development community, and high popularity in the industry. I started by adding support for the parsing of the MED format. I extended the VLC GUI with a window for event layers selection (Figure 11). This window shows all available event layers that were parsed from the MED file. Each layer has a button for either adding or removing itself from the current playback. The layer names are shown, and then a uniquely colored timeline containing a visual representation of the parsed events. Clicking on the events in these timelines will skip the playback to the relevant place in the video, and hovering over the events will bring up a preview of the annotation for that segment. The final element in each layer is the

useful "total time" this layer would take to play, calculated by adding up all event durations in the layer.

I then added a new mode for the VLC player that would change the playback behavior to the following rule set:

1. Check if there are any selected layers by the user.
2. If there are, play the earliest event of all selected layers.
3. Once the event has finished playing, skip to the next earliest event from all selected layers.

This behavior repeats until there are no more events to play. While playing a selected event, the annotation for it is displayed in a separate annotations window (Figure 13). It is possible to disable the automatic skipping behavior using a designated button. In this case the video plays as it would normally, but annotations will still appear in the appropriate window during the selected events. I also added navigation buttons to reach the start point of the previous or next event for quick control of the playback.

After the user control development was done, I needed a GUI editor for the creation of the events by users. I added the options to create, edit, and delete an event. The creation window can be seen in Figure 12. Two markers represent the start and end point of the event. These markers can be dragged, or alternatively clicked to activate and advance along with the video playback. The only information the user needs to input in the creation window is the relevant layer for this event (or create a new layer), and possibly add an annotation. This means that an event can be created with just three mouse clicks.

Finally, I added support for saving all enhancement data to a file in the .med format.

Figure 10: A sketch of the first concept for the media player implementation

3.3 PROTOTYPE WORKSHOP

To get preliminary feedback from users and to improve the design, I held separate workshops with 5 students from my HCI lab. After explaining about the new features, and allowing them to play with the new interface and watch a few annotated videos, I asked them a few questions. When asked what they liked the most about this player, 3/5 answered “I Like the time saving potential in this”. 4/5

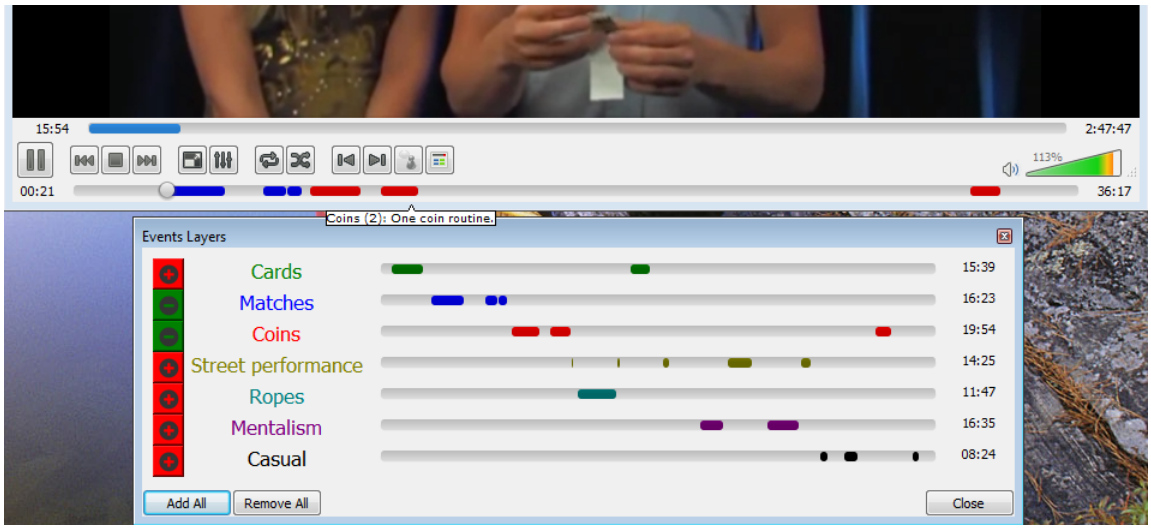

Figure 11: The resulting implemented prototype that was used by participants in the studies

said they would invest time to find and download an annotations file for a video, but it depends on the genre. 3 of the participants did mention they would prefer for the technology to be transparent and on a server. When asked what kind of videos would benefit the most out of this concept, 3/5 said longer videos that are at least 10 minutes long as well as sport videos. In all, the feedback was very positive, with all 5 participants saying they strongly prefer annotated videos, even if it might not be completely accurate. Some of the future design of the interface came from the feedback of these meetings, with the eventual unification of the 2 main sliders, and the reorganization of the buttons.

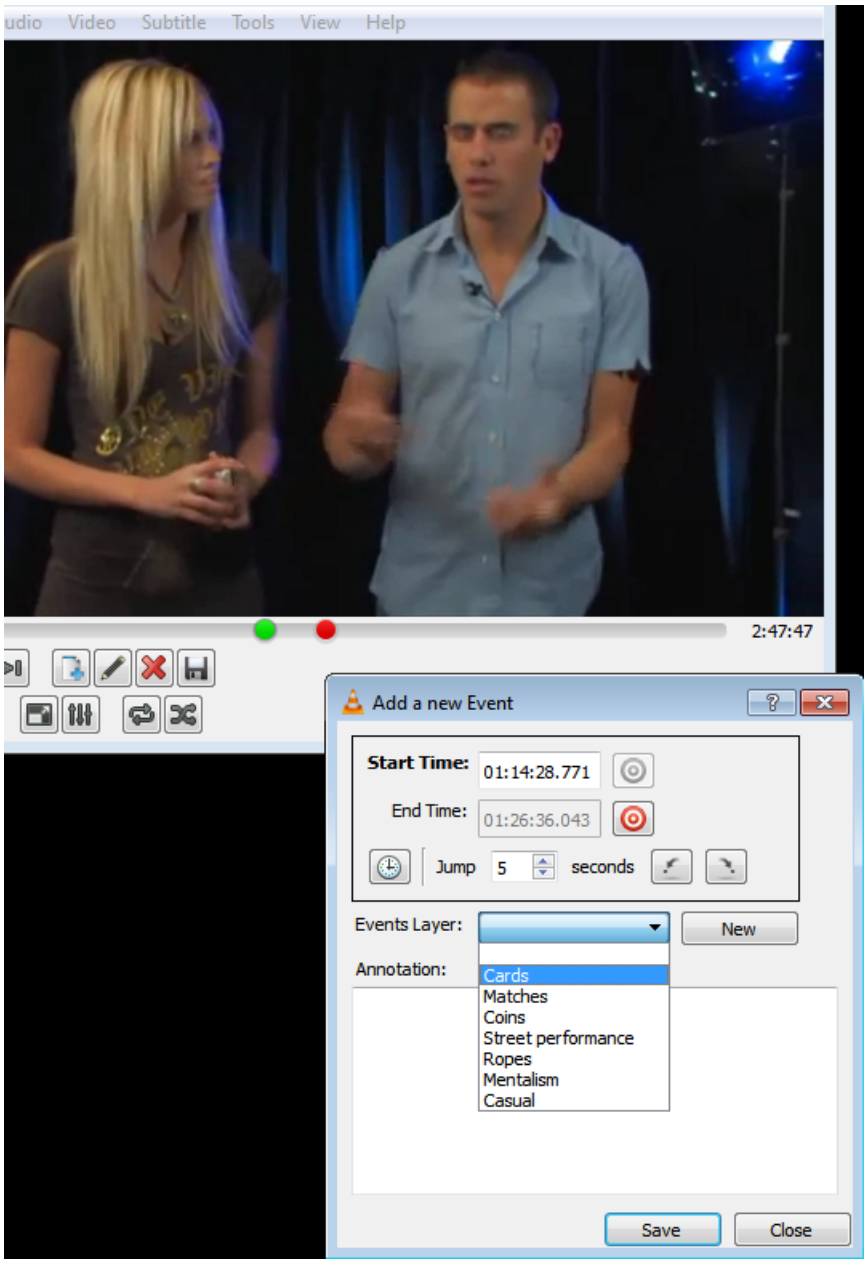

Figure 12: Creating an event. The green and red markers represent the start and end times of the event, the window shows the numeric time points, the list of available layers, and annotation field.

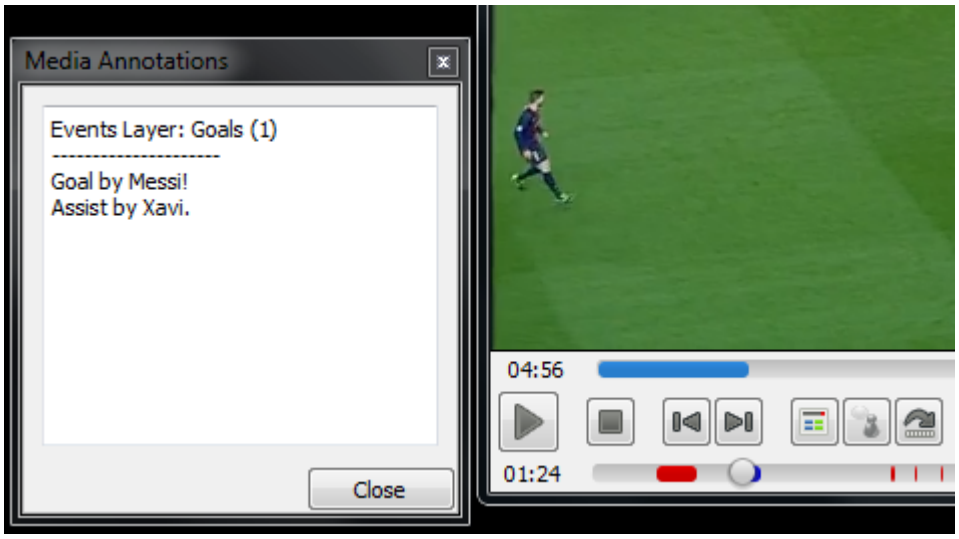

Figure 13: Annotation displayed during playback of a selected event

4

USER STUDY - EXPERT ENHANCEMENT

In the first study I investigated the benefits of the user-enhanced video player in terms of viewing efficiency (maximizing comprehension while minimizing navigation time) and user receptivity to the concept. Quality is a critical issue for user-generated tagging and annotations (challenges include weak non-informational annotations, wrong segmentations, etc.), and since we want to find the potential optimal gains of user-enhanced browsing, I left the exploration of quality to the second experiment. With that in mind, I created accurate tagging and annotations data and measured the differences in viewing time, content comprehension, and user experience to uncover the quantitative benefits of this system.

In the study, the performance of the prototype is compared to the baseline of a normal VLC Media player that is used by millions of users around the world. The reason for this being that any other type of video navigation system would be biased towards a specific kind of video, which would only represent a limited subset of the general solution. In contrast, the user-enhanced browsing concept is a potential solution for general video browsing, which currently correlates only to a normal media player. There is also the issue of resolving which video navigator is the best for producing relevant comparison results, but I found it is a very subjective matter and

determined that the most reasonable approach to this issue would be measuring the improvement over the baseline player in the fashion of recent papers [8, 15]. While we can expect to see improvements over the normal player with user-enhanced browsing, measuring the extent of these benefits compared to similar works can show us the efficiency of the system. I also decided to use 3 distinct video types in this study (videos with audio cues, visual cues, and mixed cues); with this decision I aimed to improve generalization and to check if there are specific genres that react better to user-enhanced browsing.

4.1 DESIGN

I recruited 12 participants (7 male) aged 18 to 53 (mean: 25.7, median: 22.5; 5 non-native English speakers) from the general population using an online ad in a temporary job posting site. All subjects had experience watching videos using offline and online players, but were not expert computer users. I explained to the participants that they will be asked questions about videos that we will provide using different players.

I then presented to the subjects the 4 videos explored in this study (Surveillance video: 14 minutes long, Review for a tablet device: 37 minutes long, Magic tricks tutorial: 2:48 hours long, Interview with Nicholas Cage: 42 minutes long) in a counter-balanced order. The subjects watched 2 videos in each player (the user-enhanced player and the normal VLC player) in an interlaced counter-balanced manner. This resulted in each video being watched 12 times, 6 in each player, once by each viewer. I asked the participants 4 different

questions about each video, recording the time it took them to answer, and the error rate for each question. I also asked the participants to rate how confident they are in each of their answers on a 1 to 10 scale. The questions were classified into 3 task types:

1. **Finding a scene or event**: "At what time in the video is a gun fired?", "At what point does the reviewer open the remote control application?"

2. **Getting information from the video** (requires finding a scene or event, watching at least a portion of it, and gathering the correct answer): "During the second card trick, which card does the instructor ask you to remember?", "What was the title of the movie in which Nicholas played his first leading role?"

3. **Getting scattered or fragmented information from the video** (requires finding and watching multiple parts): "How many cars enter the parking lot throughout the video?", "How many tricks with matches are taught in the video?"

Most navigation studies use the first task type, as the second and third tend to be more difficult to solve by existing technologies. The participants were instructed to answer these questions as fast as possible, but to keep in mind that they should not guess or rely solely on the annotations data. They must show the basis for their answers from the actual video to the study supervisor. To complete the experiment in one hour a maximum time limit of 5 minutes per question was set. I then asked subjects to fill a questionnaire about their browsing experience, as well as to verbally compare the players and state their feelings about the user-enhanced player and what

they liked and disliked about it. The independent variable in the study was the media player being used. The dependent variables were time to answer, correctness of the answer, and confidence in the answer. The control variables were the videos and questions.

4.2 RESULTS AND DISCUSSION

First, before the experiment started I asked each participant if they would prefer to seek information in a 3 hours long video, or a video that is 20 minutes long. Their answer did not affect the study structure. As expected, all subjects said they would prefer the 20 minutes long video. However, at the end of the study, after using my system, 83% of participants said they would prefer to watch a 3 hours video with the enhancement data over a 20 minute video in the normal player. While not a completely unexpected result, there are great implications to the fact that such lengthy videos can actually be made effective and even preferable for many different uses, and could be finally navigated in a manageable way with this technique.

Table 1 shows the rating participants gave to each player in terms of usefulness for the type of tasks they were asked to perform (a Friedman test showed statistical significance, $\chi^2(1) = 12.000, p = 0.001$) and frustration when looking for the answers ($\chi^2(1) = 8.333, p = 0.004$). I also asked the participant how easy it was to use and understand the modified player. The results show a very strong preference by the users for the modified player. It was deemed to be extremely suitable for navigation, and while the participants were still somewhat frustrated when looking for answers; it was still a

Table 1: Mean of participant rating for the two players. The Standard Error is shown in brackets

Question	Normal Player Rating	User-enhanced Player Rating
Usefulness for solving the tasks (1-10, 1: least useful)	4.67 (0.48)	9.17 (0.17)
Frustration level (1-10, 1: most frustrating)	4.33 (0.55)	8.67 (0.63)
How easy was it to understand and use the player's features? (1-10, 1: most difficult)	N/A	9.25 (0.28)

much improved experience over the normal player. Prior to the study I was concerned that some subjects will have a difficult time to learn and use the added functionality. Surprisingly, both the task completion time results and the user ratings show that the increased complexity was not an issue. After analyzing the time results, we see the extent of improvement when using the user-enhanced player (Figure 14). In total, subjects answered questions using the modified player in 43.18 seconds. The same questions took 146.8 seconds on average to solve in the normal player, over 3 times slower. The one-way repeated measures ANOVA test showed statistical significance for the player's effect on time ($F1, 191 = 89.975, p < 0.001$).

Note that the contrast in time results between the two players is likely even more acute, as the time means were bounded by the 5 minutes maximum time given to answer. If allowed to keep searching, some participants would have taken far longer to reach the

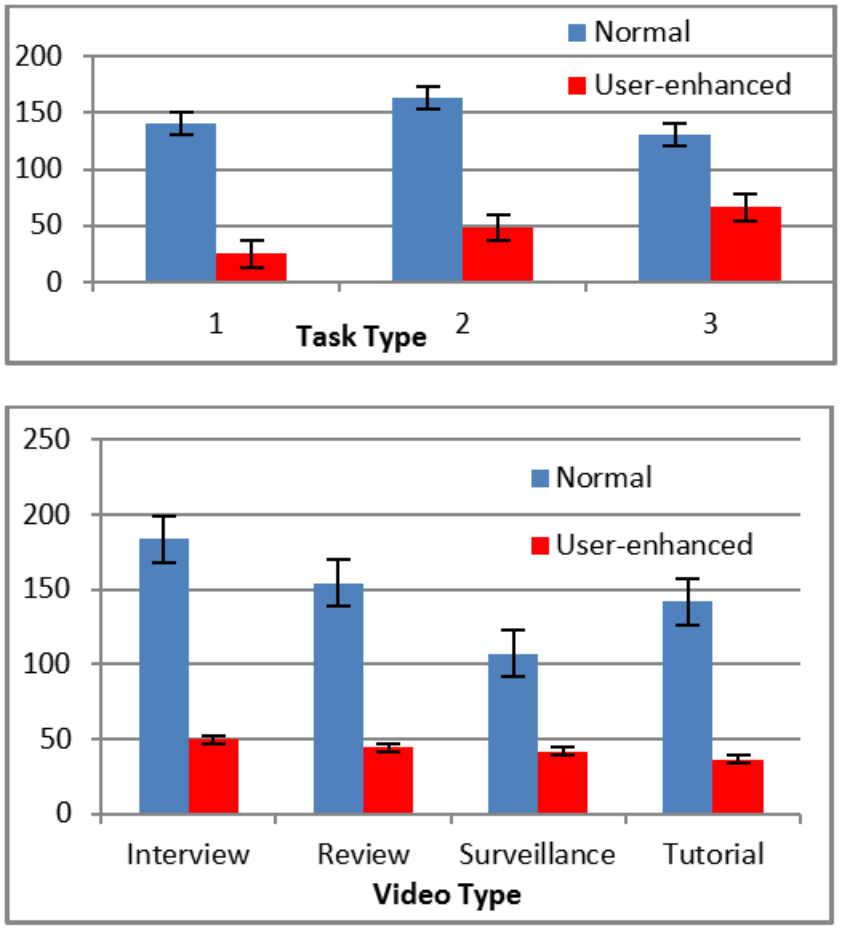

Figure 14: Mean time (in seconds) taken to answer the questions in each video. Top: By task type (1: Find scene, 2: Get information, 3: Get scattered information). Bottom: By video type.

answer, thus making the effective differences even larger. Regarding the correctness of the answers (figure 15), we see that on average, the user-enhanced player had 88.54% correct answers opposed to the 63.53% rate of the normal player. The participants found some questions to be very hard to solve with the normal player (e.g., 2: "How many tricks with matches are taught in the video?", 7: "How does Nicholas describe being directed by Martin Scorsese?", 15: "How does the reviewer describe the sound quality of the device?"), but quite easy with the user-enhanced player. Figure 15 also shows a

very interesting case with question 3 ("When does the instructor first demonstrate a trick in public?"). None of the viewers on the modified player have answered this question correctly, while all participants got it right using the normal player. The reason for this is in fact an accidental error in the tagging data provided for this video. The target event was mistakenly not tagged, which misdirected the participants (5 of 6 of which have answered with the supposedly correct answer). This example highlights one of the greatest weaknesses of this system – the blind reliance on user generated data. In reality, these kinds of mistakes are usually picked up and corrected, but until then many viewers may be misled.

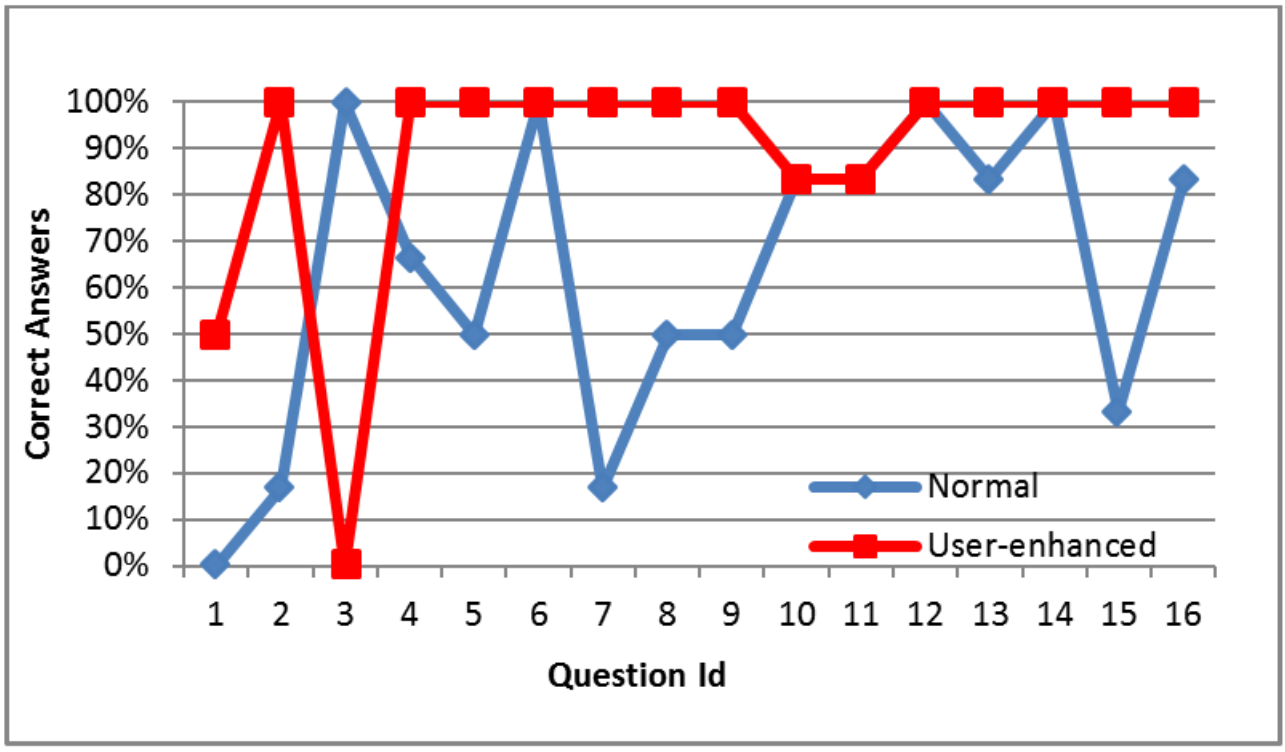

Figure 15: Percentage of correct answers for each question.

I also asked the subjects to rate their level of confidence between 1 and 10 (10 being absolutely certain) in each of their answers. I found

that on average, when using the modified player the answers had a higher confidence level altogether (ANOVA: $F1, 175 = 13.182, p < 0.001$) with 9.38 vs. 8.51 in favor of the user-enhanced player. I further discovered that when the participants were correct, the differences in confidence were rather small (9.41 vs. 9.05 in the normal player), but when they were wrong the difference became very noticeable (9.09 vs. 6.79 in the normal player). This shows more concretely the issue that question 3 presented – when using tags and annotations the viewers can be highly overconfident in their knowledge and understanding of the video content. The difference in confidence in correct answers is also worth noting as the participants answered using the same video segments, but it seems they feel that the enhancement data allows extra validation for their response. Finally, I asked the participants the question: "Would you use the user-enhanced player to watch videos in your day to day life? Why?". Only 5 out of 12 subjects answered "Yes" and gave some examples (skip to scenes in previously watched videos, show specific parts to friends, long videos can be seen in a short period of time, etc.). However, upon further inspection of the verbal explanation, I found that 5 out of the 7 participants that answered "No" actually gave scenarios in which they would want to use the player: "...but when watching videos for a second time I will only want to see certain scenes." (P2), "...If it is for learning I will use it." (P4), "...I can imagine that after the first watch I would really prefer the annotations to look for specific points." (P6), "...If I was watching lectures online I would probably use it" (P12). My interpretation is that while the

subjects clearly see the benefits of the concept and its potential uses, they are still not used to this idea, and it seems foreign to them.

5

USER STUDY - ITERATIVE ENHANCEMENT

The second study was focused on video consumption using enhancement data, as well as the process of iterative contribution of metadata by participants. To accomplish those goals I designed an atypical study, where participants have to rely on and use the output of the previous study-takers.

5.1 DESIGN

I recruited 12 participants (8 male) aged 19 to 35 from the general population using another online ad. I showed them the modified VLC player and demonstrated all of the new functionalities. I let them experiment with the player and an example video with premade enhancement data until they feel comfortable using it. I then presented them with the target video for the study: a one hour long TV special about four different scientific topics, released in 2013 [3]. The reasons for choosing this video were: 1) the fact that it is too long to watch to completion during the study, thus forcing participants to skip ahead and use additional information. 2) The video had very diverse content, with topics covering Space, Energy, Nanotechnology, and Biotechnology. This encourages possible categorization of the video by the viewers. Since I wanted to see the development of the

data coming from users, the study started off with no tagging or annotations for the first participant. Each subsequent participant could use the information added by the previous subjects to browse the video. The participant is allowed only 20 minutes to watch the video. Participants were encouraged to find the interesting parts out of the 1 hour show, and are also asked to try and familiarize themselves with all major topics of the video during this time. I then give the participants a questionnaire about their viewing experience. For the second part of the study, I asked the participant to contribute and improve the enhancement data for future viewers for another 20 minutes. I encouraged participants to add more events to existing layers, make sure events are correct, try to add topics (or scenes) as new layers, highlight important events, add comments on events, or to simply use their own ideas. I then gave participants another questionnaire concerning the creation process.

I instrumented logging facilities in the player in order to collect data about the layers each user selects for viewing, and the number and type of changes they make to the enhancement data. I retrieved user feedback and ratings using the questionnaire.

5.2 RESULTS AND DISCUSSION

This section describes the results for the two different major elements in the study: viewing and editing.

5.2.1 *Viewing*

I first asked the participants to rate how easy it was to find interesting parts in the video (1 being hardest, 10 being easiest). Figure 16 shows the results. We can see that the first and second participants had a very difficult time finding interesting events. The third participant already improved to mediocre difficulty, and already by the fourth subject we see that reaching the interesting places became very easy. This is a strong indication for a noticeable improvement in the viewing experience after only 2 to 3 enhancement contributions. This is an important point that proves it is not necessary to have an expert tag and annotate the video. Moreover, it is not necessary for a large number of viewers to contribute before seeing a positive effect. It is also worth noting that the added complexity of having more layers and more events (36 events in 21 layers after 12 participants) did not show a negative effect on this rating.

The participants were then asked "Do you think this tool improved your viewing experience? How?". 10 out of 11 (90.9%) participants answered "Yes". Among their detailed descriptions are: "I could skip around like using a map." (P2), "It helps sort out the different segments that are in the video." (P3), "I can easily find all the interesting parts that attract me." (P8), "It specifies which scenes viewers are more likely to watch. . . makes it more fun for the viewer as well as saves time by trimming down the video" (P9), "I only had 20 minutes to watch an hour long video and the player allowed me to watch what I wanted" (P5). The theme of personalization repeats in the feedback from users, and it is considered to be an important

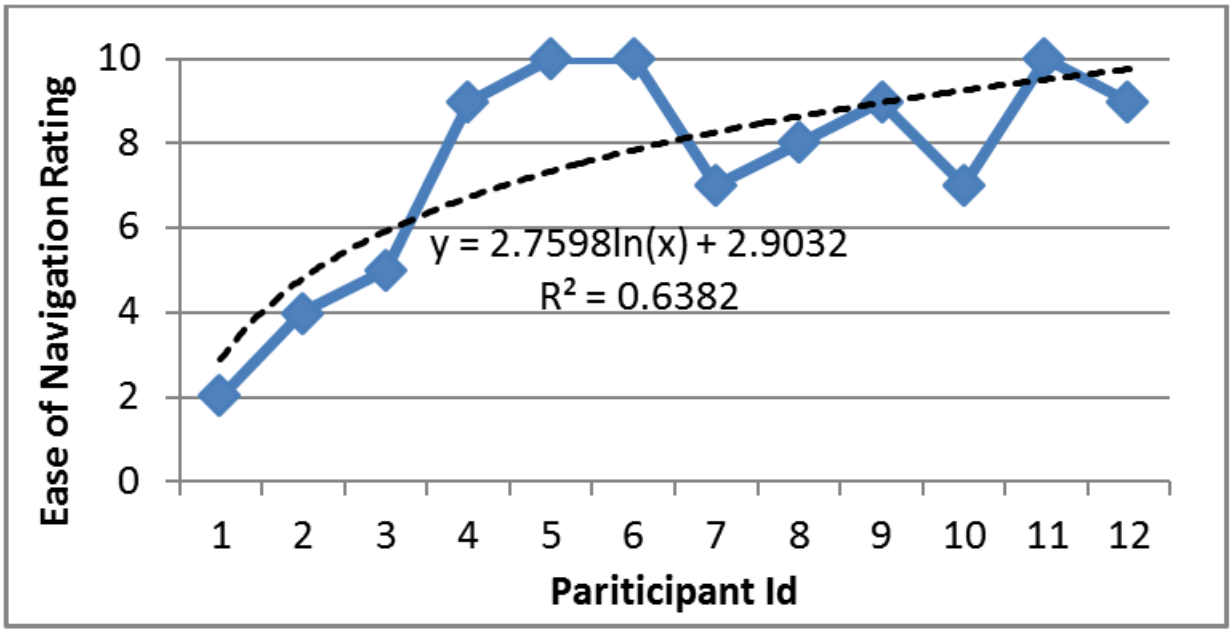

Figure 16: Participant rating of ease of reaching interesting parts in the video (1: hardest, 10: easiest). This pattern fits a logarithmic trend (black dotted line) showing how quickly the data becomes effective.

advantage of this system. I was also encouraged that again no subject mentioned difficulties in controlling the new functionalities of the tool. Over the course of the study I saw an interesting change in layer selection behavior (Figure 17). Up to the 7th participant, almost all layers were selected (mean of 94.43% of available layers) for playback, and were at least partially watched. However, after that point participants started skipping some layers and viewing only what seemed potentially interesting to them. It seems that more than 15 event layers might be too many to watch with a 20 minutes time limit. This is to be expected as the more layers viewers try out, the more time is spent on controlling than actually watching the video.

I asked the subjects to mention event layers they thought were most useful for them when watching the video. The most mentioned

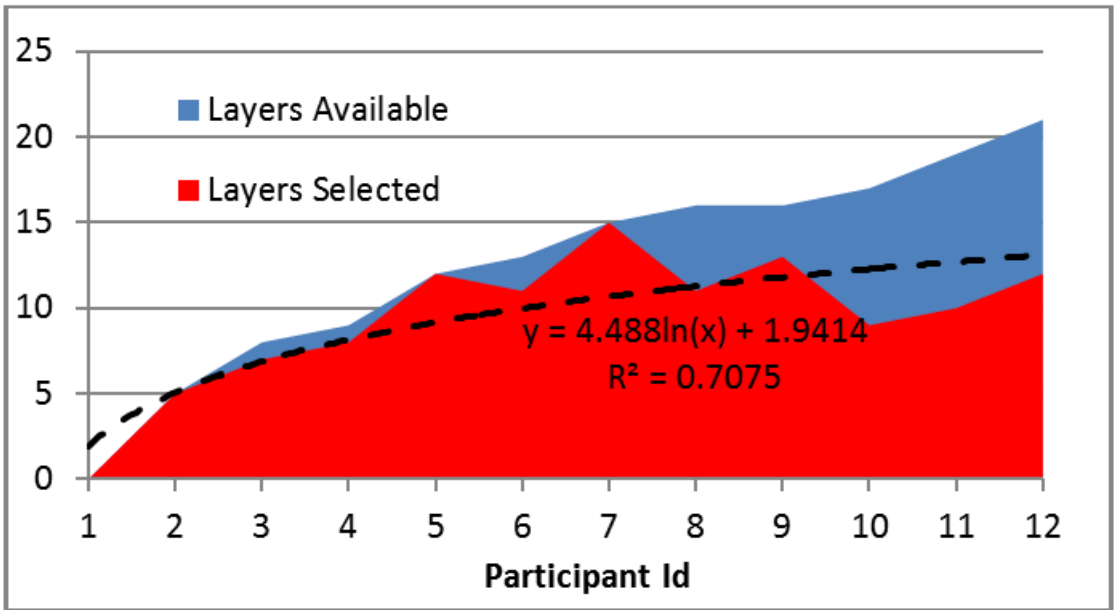

Figure 17: Number of layers each participant selected to play compared to the available number of layers. The black dotted line shows the logarithmic trend.

layers were: "Space" (6 mentions), "Nanotechnology" (6), "synthetic biology" (4), "Interesting fact" (3), and "Important Stat" (3). Interestingly, 4 out of these 5 layers were tagged by the first participant, while the remaining one ("synthetic biology") was added by the second participant. This is encouraging, and may mean that with only a handful of contributors, the main and interesting ideas in a video can be sufficiently tagged and significantly helpful to the next viewers. One may also suggest that the participants mentioned the layers above because they were the main topics of the video, but in fact, an entire section about renewable energy was not very popular despite taking up an equal share of the show. This strengthens the evidence that the ability to find highly focused content is improved.

I directly asked participants to rate the overall events usefulness and accuracy. Both accuracy and usefulness stay at a constant level:

the mean for accuracy is 7.36 (S.E. 0.527); the mean for usefulness is 8.00 (S.E. 0.632). These are good results for user-generated content with no real guidance, and the consistency is to be expected as not many quality corrections were made. An encouraging element is that the ratings do not fall due to complexity.

5.2.2 *Editing*

I asked the participants to rank the contribution tasks from most liked to least liked in order to find the appeal of different enhancement activities. The most liked task was 1) "Add missing events to existing layers", followed by 2) "Highlight events that seem important to you", 3) "Make sure existing events are correct and accurate", 4) "Comment on events", 5) "Add new layers of topics or scenes", 6) "Tag actors or objects in the video". It is worth noting that while correcting events ranked 3rd, an actual event correction occurred only twice throughout the study. Also, even though adding a new layer ranked only 5th, 11 out of 12 participants chose to do it regardless. Tagging people or objects in the video is perceived to be a very tedious task by participants, and as such it is likely to only be done by enthusiasts and professionals. I also asked to rate the overall annoyance from performing these tasks (ranging from 1: not annoying, to 10: most annoying). The mean rating was 2.42 (S.E. 0.48), which suggests that overall, tagging and annotating may not be the excruciating activities they are often made out to be, especially when done collaboratively. Figure 18 contains the number of events and the number of layers created as the study progresses. We can

see that both creation occurrences show a logarithmic trend, more strongly for the layers (R^2 = 0.5716) than for the events (R^2 = 0.3926). From the data we can conclude that viewers are less likely to add new layers when they are abundant, and would instinctively try to add their new events to the existing layers. This reduces clutter and overlapping, makes playback control easier, and improves existing layers. Events do seem to drop off as well as it gets harder to add valuable and non-redundant enhancing content.

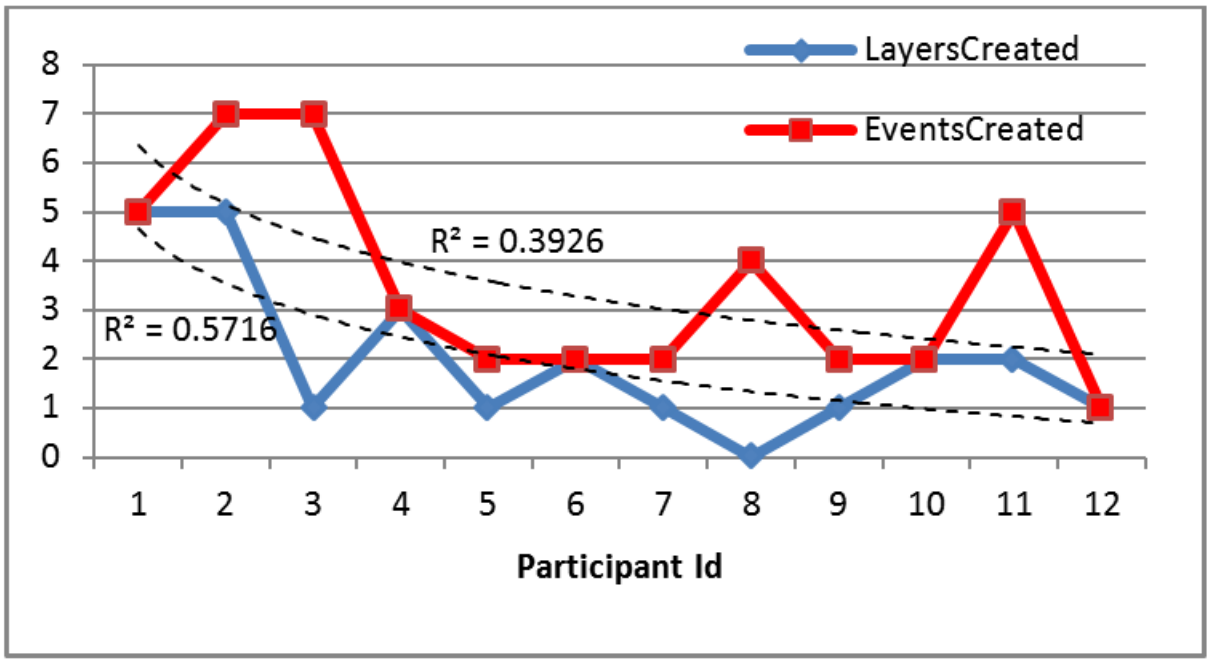

Figure 18: Number of events and layers created by each participant throughout the study. Black dashed lines are logarithmic trend lines.

We can see certain saturation in the tagging of content in Figure 19. As the study progressed, less and less new content was being tagged, as to be expected. Each participant tagged on average 9.99% of the video (including overlapping of other events) in the allocated 20 minutes. By the end of the study, and after 12 contributions, 74.15% of the video had been tagged. The declining number of new content

tagging per session is apparently related to the fact that most of the interesting parts were already added in another layer, and there was not much value in re-tagging them elsewhere. This is likely a positive point, as it is safe to assume overlapping content would be frustrating in many cases.

Finally, I asked the subjects "Would you perform these tasks in videos you watch in your daily life for other viewers?". I was surprised to find 9 out of 12 (75%) answers to be "Yes". Participants elaborated with: "...I would use these tools to highlight what I want to show people that are important to me." (P5), "As a study aid this would be useful, videos for classwork or instruction videos. I would do it for something I'd want to teach others." (P7), "because it will help people understand the video and spot the highlights much better" (P4), "Creative way of sharing." (P1), "...I would if there are multiple perspectives on it. I would probably only take the time to do this for my own work or something I felt strongly about" (P2). The participants that answered "No" explained that they enjoy watching the videos passively and do not wish to spend the time for these tasks except in specific cases.

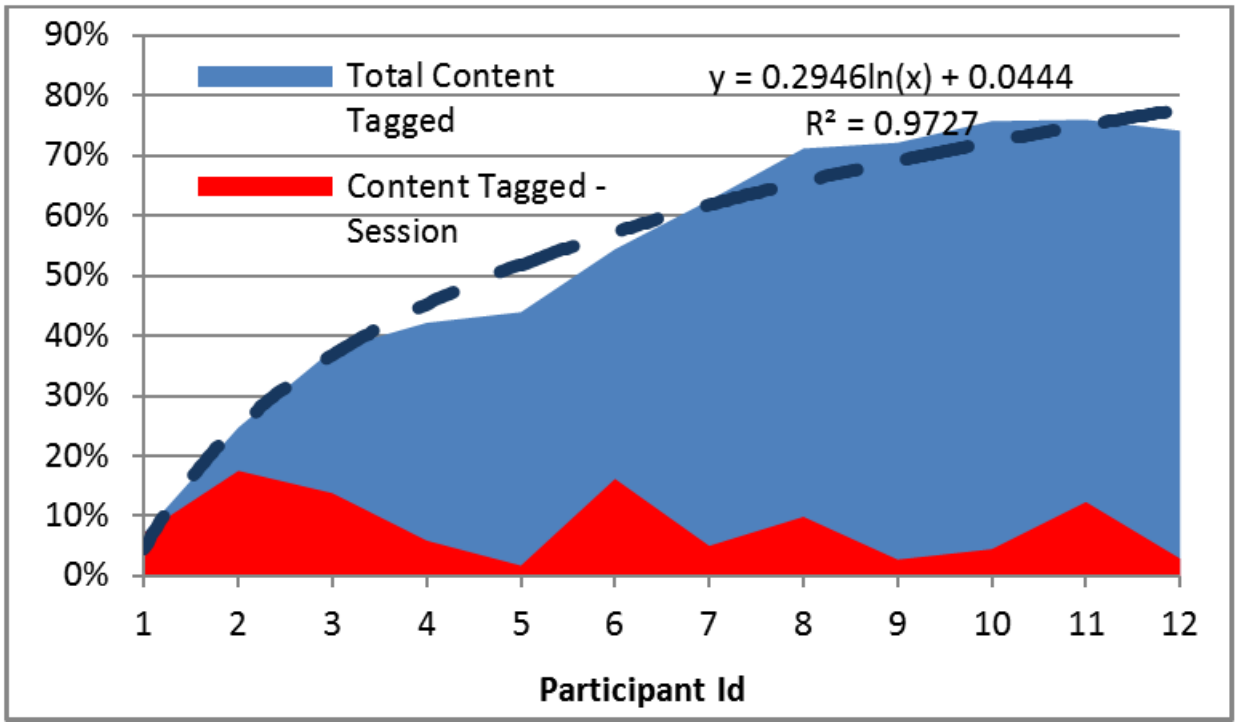

Figure 19: Percentage of content tagged each session compared to the ratio of total content tagged (no overlaps). The dashed line shows the logarithmic trend of the total content tagged.

6

INTEGRATION WITH THE OFFICIAL VLC VERSION

This chapter will discuss the process of integrating this work with the official version of the open source VLC Media Player. As mentioned previously, I was invited to present the independent work I had done with VLC in the yearly conference for VLC developers (VLC Dev Days) in 2013. I have shown my prototype and the results I got from my first study, and received very good feedback as well as a lot of interest, as indicated by having the most questions asked after any presentation. I was encouraged to finish the features I was working on and send the code to the VLC team for review.

6.1 OPEN SOURCE SOFTWARE

Before moving on, the concept of open source software needs to be clarified. The idea behind open source is a free release of the software product as well as the source code that compiles it. This allows anyone to use it (but not sell it), and make any kind of modifications and additions to the original code based on their own requirements. In VLC's case, the license of using the code is under LGPL, which means commercial companies can incorporate the VLC library in their products but the LGPL relevant code must be

made public freely. Since I've created my prototype with the latest VLC code branch, I had the option of trying to insert it into the official version of VLC. According to the statistics on the official website (http://www.videolan.org/vlc/stats/downloads.html), over 1.3 billion downloads of various versions for different operating systems have been logged over the years. This makes VLC one of the most popular media players in the world and would mean a huge user-base and impact for any new technology integrated in it. However, the process of integration may not be so simple. In most open source software communities there is an owner which can be a person or an organization that started the project or received ownership of it. There are also maintainers (who may or may not be the owners); individuals in these positions receive code changes (also called patches) from the public, and make sure they are useful, correct, and safe before applying the new code to the official version.

6.2 CODE MODIFICATIONS

The prototype used to run the studies had all the necessary features. Playback skipped non-selected parts of the media, all layers were visible for preview, navigation between events was easily managed with two buttons, and users could quickly add, edit, delete and save the enhancement data. However, because it was not an actual released product, certain bugs were tolerated, stability didn't have priority, efficiency was not a factor, and there was no need to turn the feature on and off. This meant that in order to integrate to the official VLC version, that is used by millions of users, I had to fix every

single issue I knew about, add management features, and make the player more polished.

First, I had to make sure the design architecture fit the VLC logical system structure. VLC is an incredibly extensive project that is spread over thousands of files of code. It can run on Windows, Mac, Android, Linux, iOS and several other operating systems. It supports dozens of codecs and containers, and it has several interfaces that can run simultaneously. Designing a feature to work well on all of these is quite a challenging task, so I decided to start with the Qt4 interface that runs mostly on Windows and Linux. I had to redo my original feature design, and insert the logic of MED into the main loop of the running media thread. I've discovered a surprising race condition problem in VLC and had to design a solution in order to get my feature to work in a stable way. I then fixed the various bugs I encountered throughout my work and studies, and cleaned up my code. I unified the original two sliders into one that changed modes according to the media being played - non MED files were played normally and MED files caused the feature interface to pop up, this behavior allows the feature to be released enabled by default with no implication for a normal user. I added an option to disable the feature entirely, and customize the interface.

6.3 SUBMISSION

My final submission of the MED support patch to VLC had 80 files that were changed or added, and thousands of lines of new code. My original solution for the race condition problem in VLC was not

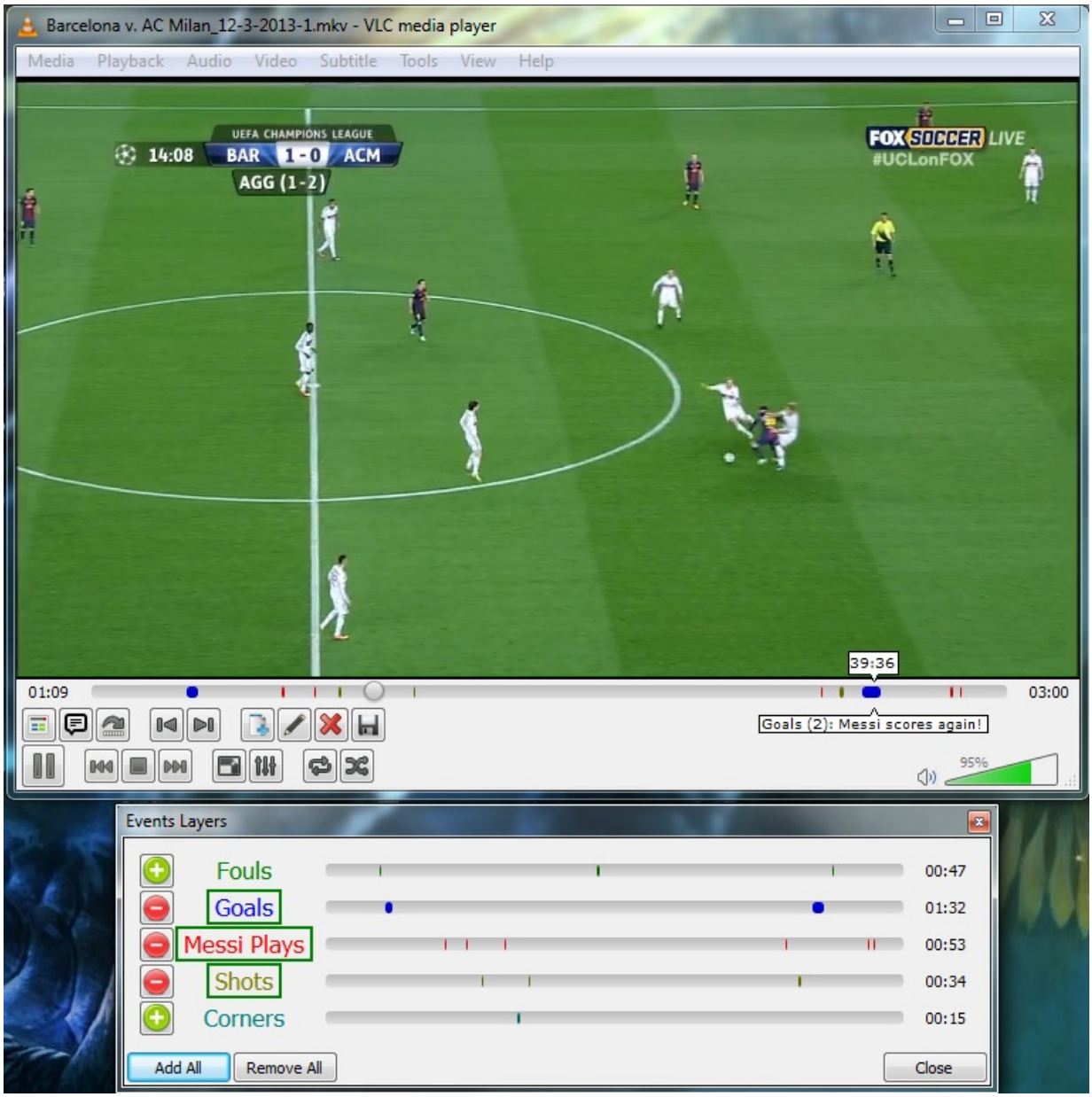

Figure 20: The final look of VLC integrated with MED

accepted, and I had to redesign it and test the result again. Currently, my feature is still under review (it takes a long time to go over all changes, and this is done by only a handful of volunteers), and feedback is being sent back to me, which leads to me fixing the new issues and resubmitting the new version. It takes a large effort on my part, but because of the very significant potential impact of the

integration to the official version of VLC, I consider it a worthwhile task.

7

CONCLUSION

7.1 DISCUSSION

I presented two user studies and results that show different effects for user-enhanced videos browsing. For the purpose of summary, in this section I focused the main effects and divided them into advantages and challenges.

7.1.1 *Advantages*

Navigation Speed

The results clearly show that this method is far preferable to existing normal players. While other navigation players show more modest improvements, participants reached content and answered questions from 250 to 388 percent faster with 4 different video types. This is strong evidence that utilizing user-enhanced media is a viable strategy for addressing the challenge of efficient video navigation.

Personalization

Each participant in the second study chose a different way to watch the video, with different layers of events. This is a very natural

process as each person would have specific interests and disinterests. But the implications are significant in terms of browsing. There are no other players to my knowledge that allow users to customize their viewing experience so extensively. This approach allows for the combination of layers, and the addition of new events (or the correction of events) to the existing data. If you are not happy with a certain layer, it could easily be removed from the playback and replaced with another one.

Content Description

In both studies we have repeatedly seen the participants appreciate and mention the fact they can tell what topics are covered in the video, and what type of video it is, even before watching a single frame. The fact that users can hover over events to see the annotations allows them to quickly decide whether or not the segment is one they are interested in.

Video Consumption

As pointed out previously, 10 out of 11 (the first subject watched the video unaided) participants replied that this system improved their browsing experience. The navigation, personalization, and content description add up to make a significant positive impact on the consumption of a long video in a short time. This could mean that long videos can be made effective and viable for many purposes, when they are enhanced by viewer metadata.

Content Comprehension

The first study clearly shows that using the enhancement data, viewers can understand the video and information in it much better, with a 29% increase in correct answers to questions about the video compared with watching in a normal media player. The added information and context that the metadata offers improve the experience and may help in education related videos, lectures, and tutorials.

7.1.2 *Observations*

Event and layer creations

The second study suggests that as a larger portion of the video is tagged, viewers will add fewer new layers and events. Potentially, the contribution focus could eventually shift to improving quality instead of extending the descriptions and adding events to make layers more robust.

7.1.3 *Challenges*

Overestimating Quality

The confidence results from the first study show us that viewers give significant weight to the credibility of the tagging and annotations. At times, this can negatively affect their knowledge and mislead them. This phenomenon may occur because the concept is a foreign

idea to viewers and I assume the estimation will improve as these systems become more common.

Maintaining Quality

Directly related to the previous challenge, as with any content that is produced by users we must consider malicious or simply low-quality contributions. Currently I see no other way to sort out the "bad apples" in the enhancement data other than a rating system for the metadata files. Users should be able to increase the ratings for good MED files and hide the poor quality ones. This is similar to the way third-party subtitle files are currently being controlled online. Future versions of the format may support ratings for smaller scale objects like a layer or a single event.

Sharing

While I have not touched much on the topic of sharing in this thesis, it is still an important issue with this type of data. The offline implementation evaluated here makes it a little difficult to collaborate with other viewers. A user would need to manually send the MED file to friends, or alternatively use a server to create a single sharing point. Optimally, this concept can be incorporated in video hosting websites. In this case, the metadata file would be invisible to the user, and any change in data will be immediately received by the next user. Instead of having to download the exact same video file, the shared URL of the video will take care of the synchronization issue.

7.1.4 *Use Cases*

There are many potential instances where this method would be useful but they may not be evident at first glance, so I included a list of examples here:

- Summaries and aggregation of summaries: Use versions from different sources to check for agreement or create a meta-summary.
- Creator or viewer commentary for the video: A way to display extra content or updated information while the video is running.
- Multiple versions of a video in the same file (e.g., Director's cut): Instead of several large files that have nearly the same content, enhance the largest file with different viewing options.
- Advanced bookmarking: Enables bookmarking for different users. Moreover, instead of highlighting a point in the video you can mark whole segments
- Tagging people or locations in the video: Videos of gatherings or important personal events could be tagged with the times people appear in them.
- Quick skimming of the video contents (simply by looking at layer names and annotations): Allows viewers to immediately see what the video contains before even watching it.

This is of course a partial list, but it still provides solutions in many different cases with just a single system.

7.2 LIMITATIONS

There are several potential limitations with this work. First, it requires a great deal of people contributing for the benefit of others. Since there are many other user-generated contributions on the internet these days (e.g. Wikipedia, Comments, Subtitles, Reviews, etc.) it is fairly safe to assume that if and when a metadata format is widespread enough, the users will want to contribute back to the community.

Another limitation of this work is the current lack of quality control. Although there is a way to edit and delete the data in the VLC implementation; there is no way of identifying the better or worse metadata before actually using it. This kind of concern is prevalent throughout all types of user-generated content, and can be resolved through proper management of the data in an external central database or website.

A limitation for the first study is the usage of expert data. The extent of experience improvements when average data is used is still unknown. Though it should be mentioned that it is a difficult situation to generalize: it is likely to be highly dependent in the specific data input. The goal of that study was to uncover the maximum improvement, and the final results seem to be consistent.

7.3 FUTURE WORK

I have reached some interesting conclusions and important evaluations in this work. But just as important, this concept opens the door to endeavors and research options in many other different directions. One example would be the search capabilities that can be gained from the abundant information in the layer names and annotations. One possible way to utilize it would be to select events that have been annotated with specific search keywords to the playback, thus creating a sub-video related to a personalized term.

There is also the possibility of creative extensions for future versions of the format. The data could be used to tell the media player to pause the video and ask for user input. Alternatively, the player could increase or decrease the volume or playback speeds in certain segments in the video. It allows for the creation of a language that can be used to control the video in ways we have rarely seen before: playing the video in non-chronological order, switching between linked media files during playback, etc.

Another exciting aspect is the use of this control concept to browse different types of documents altogether, such as educational books. Instead of events, specific sections of text could be tagged to a segment in a layer. After which, in addition to a table of contents, the reader could select layers of segments related to a topic or category (e.g., all of the images in a book, or all segments tagged to the "important" and "exciting" layers by other readers) which would personalize and enrich the reading experience.

7.4 SUMMARY

To summarize, the work in this thesis explores the main effects that are to be expected from creating and using user-enhanced content for video browsing. It is true that there are several obvious challenges with this concept. However, the potential to save countless human-hours and the comprehension and consumption advantages are, in my eyes, more than enough to justify the continued development and research of this approach.

A

MEDIA ENHANCEMENT DATA FORMAT SPECIFICATIONS

I have attached here the format specifications in their current version.

Author: Roiy Shpaner

Last Updated: 4/30/2014

Media Enhancement Data Format Specifications

Version 0.5

1. Introduction

This format was created to provide a simple and lightweight way to direct media players on how to run files according to the user's preferences.
The MED file should guide the application to skip media segments automatically, and show additional information that is time-related to the file.

2. Format

2.1 General notes:

Timecode format:
HH:MM:SS,MIL *(hours, minutes, seconds, milliseconds)*

*2.2 **Properties:***

2.2.1

TargetName *(Followed by line terminator, Case insensitive)*

Optional field, this will list media filenames that should play with this enhancement file if it is in a known location.
All filenames should be aggregated until a blank line is found.

Example:

TargetName
Men.In.Black.avi
Men-In-Black.mp4

2.2.2

TargetHash *(Followed by line terminator, Case insensitive)*

Optional field, this will list hash values of videos that should play with this annotation file if it is in a known location.
All values should be aggregated until a blank line is found.

Example:

TargetHash
3hjann3c9rf03ft
1ma9ckan22dks

2.2.3

TargetSize *(Followed by line terminator, Case insensitive)*

Optional field, this will list byte count sizes of media files that should play with this enhancement file if it is in a known location. All sizes should be aggregated until a blank line is found.

Example:

TargetSize
26648211
28917578

2.2.4

Ordering notes: When playing a video file, the folder should be searched for a MED file with the following priorities:

(a) Matching name to the video file with the MED extension.

(b) Matching name in the TargetName section.

(c) Matching hash value in the TargetHash section.

(d) Matching size in the TargetSize section.

If there is more than one match, alert the user, and look for the next priority to disambiguate the options.

2.2.5

Event format:

Line 1 MUST be the unique event ID, which is the combination of the name of the event layer and the numeric Id of that particular event in the layer. It is contained in two square brackets, one for the event layer name and one for the numeric id.

[[*Layer Name*]*Event Id*]

Event name should be between 1 and 255 characters in length. Event Id should be an integer between 1 and 2,147,483,647.

Note: Event Id MUST be ignored by the application, and calculated independently when ordering the events. This data is saved for convenience.

Example: – For layer “Highlights”, event number 3*:*

[[Highlights]3]

Line 2 MUST be the start timecode, followed by the string “-->”, followed by the end timecode.

HH:MM:SS,MIL --> HH:MM:SS,MIL

Example:

01:07:33,529 --> 03:02:14,811

Line 3 is the optional start of the textual annotation. It will continue until a double blank line is found. If lines 3 and 4 are blank lines, there will be no annotation for this event. One blank line does not finish the annotation.
To put text as a spoiler, text should be between opening <*> and closing </*>.

Example:

This is surely a normal situation and not a fake one…

I only wish she would try a little harder.
<*>At least she leaves in the end of the movie</*>

Unified example:

[[Highlights]3]
01:07:33,529 --> 03:02:14,811
This is surely a normal situation and not a fake one…

I only wish she would try a little harder.
<*>At least she leaves in the end of the movie</*>

Possible Future extension: Play segment with modified parameters (volume / playback speed / pause / repeat, etc.)
[[Highlights]3] | vol 1.5 | x2

{Possible Future extension: User input at particular places. (Selecting ending / affecting the video while it plays)

2.3 ***File Format***

MED *(Case insensitive)*

TargetName (optional)

TargetHash (optional)

TargetSize (optional)

Event

Event

Event

...
...
...

MED *(Case insensitive)*

3. Playback

3.1 Player applications must start playing the file at the first timecode that appears in the selected layer by the user. If no layer is selected play the file normally.

3.2 If the automatic jumping behavior is enabled, any segment of the video that does not appear in the selected layer(s) must be skipped. Otherwise, play normally.

3.3 The user should be able to enable or disable a way to see the relevant annotations (separate window / overlay) for the current playback.

3.4 *Future feature*: Spoiler text must only appear if specifically asked by the user, either specifically for the segment, for the video, or for the viewing session.
Players should show SPOILER in places where spoiler text was marked.

B

MEDIA ENHANCEMENT DATA EXAMPLE FILE

Here is an example file that follows the MED format:

med

[[Cards]1]
00:03:21,529 −−> 00:12:53,811

[[Matches]1]
00:15:33,000 −−> 00:25:23,811
Ghost match trick.
Lighting a single match, then returning it to the box.

[[Matches]2]
00:31:42,000 −−> 00:35:40,811
Lighting all the matches except a hidden one.

[[Matches]3]
00:35:42,000 −−> 00:38:15,811
Matches trick - no fire.

[[Coins]1]
00:39:38,200 −−> 00:48:10,000
This is the most useful trick in the entire video.

[[Coins]2]
00:51:25,200 −−> 00:57:44,000
One coin routine.

[[Street performance]1]
00:57:45,200 −−> 00:58:23,000
One coin routine.

[[Ropes]1]
00:59:45,200 −−> 01:11:33,000
Shoelace escape.

[[Street performance]2]
01:11:34,000 −−> 01:12:33,000
Shoelace escape live.

[[Cards]2]
01:15:53,200 −−> 01:22:00,000
Guess someone's card in advance.

med

BIBLIOGRAPHY

[1] SubRip Subtitles Format. http://matroska.org/technical/specs/subtitles/srt.html, 2009. [Online; accessed 04-May-2014]. (Cited on page 15.)

[2] O. Aubert and Y. Prié. Advene: active reading through hypervideo. In *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 235–244. ACM, 2005. (Cited on page 11.)

[3] BBC. Tomorrow's World. http://www.telegraph.co.uk/culture/tvandradio/9988345/Tomorrows-World-a-Horizon-Special-BBC-Two-review.html, 2013. [Online; accessed 04-May-2014]. (Cited on page 31.)

[4] Robert Bringhurst. *The elements of typographic style*. Hartley & Marks, 1999. (Cited on page 59.)

[5] Marc Caillet, Cécile Roisin, and Jean Carrive. Multimedia applications for playing with digitized theater performances. *Multimedia Tools and Applications*, pages 1–17, 2013. (Cited on pages vii, 8, and 9.)

[6] A. Carlier, G. Ravindra, V. Charvillat, and W.T. Ooi. Combining content-based analysis and crowdsourcing to improve user interaction with zoomable video. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 43–52. ACM, 2011. (Cited on page 12.)

[7] Pablo Cesar, Dick CA Bulterman, Jack Jansen, David Geerts, Hendrik Knoche, and William Seager. Fragment, tag, enrich, and send: Enhancing social sharing of video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5(3):19, 2009. (Cited on page 11.)

[8] K.Y. Cheng, S.J. Luo, B.Y. Chen, and H.H. Chu. Smartplayer: user-centric video fast-forwarding. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 789–798. ACM, 2009. (Cited on pages vii, 2, 5, 6, and 23.)

[9] ComScore. ComScore online video rankings. http://www.comscore.com/Insights/Press_Releases/2013/6/comScore_Releases_May_2013_US_Online_Video_Rankings, 2013. [Online; accessed 04-May-2014]. (Cited on page 1.)

[10] M. Del Fabro, K. Schoeffmann, and L. Böszörmenyi. Instant video browsing: a tool for fast non-sequential hierarchical video browsing. *HCI in Work and Learning, Life and Leisure*, pages 443–446, 2010. (Cited on pages vii, 6, and 7.)

[11] Dennis Del Favero, Neil Brown, Jeffrey Shaw, and Peter Weibel. T_visionarium: the aesthetic transcription of televisual databases. In *Present Continuous Past (s)*, pages 132–141. Springer, 2005. (Cited on pages vii, 7, and 8.)

[12] Roberto Fagá Jr, Vivian Genaro Motti, Renan Gonçalves Cattelan, Cesar Augusto Camillo Teixeira, and Maria da Graça Campos Pimentel. A social approach to authoring media annotations. In *Proceedings of the 10th ACM symposium on Document engineering*, pages 17–26. ACM, 2010. (Cited on pages vii, 9, and 10.)

[13] The Moving Picture Experts Group. Mpeg-7 Metadata Protocol. `http://mpeg.chiariglione.org/standards/mpeg-7`, 2002. [Online; accessed 04-May-2014]. (Cited on page 14.)

[14] Adam Janin, Luke Gottlieb, and Gerald Friedland. Joke-o-mat hd: browsing sitcoms with human derived transcripts. In *Proceedings of the international conference on Multimedia*, pages 1591–1594. ACM, 2010. (Cited on page 2.)

[15] J. Kim. Toolscape: enhancing the learning experience of how-to videos. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 2707–2712. ACM, 2013. (Cited on page 23.)

[16] K. Kurihara. Cinemagazer: a system for watching videos at very high speed. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 108–115, New York, NY, USA, 2012. ACM. (Cited on pages 2 and 5.)

[17] Rodrigo Laiola Guimarães, Pablo Cesar, and Dick CA Bulterman. Creating and sharing personalized time-based annotations of videos on the web. In *Proceedings of the 10th ACM symposium on Document engineering*, pages 27–36. ACM, 2010. (Cited on page 11.)

[18] Gregor Miller, Sidney Fels, Abir Al Hajri, Michael Ilich, Zoltan Foley-Fisher, Manuel Fernandez, and Daesik Jang. Mediadiver: viewing and annotating multi-view video. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1141–1146. ACM, 2011. (Cited on page 11.)

[19] N. Mukesh, C. Harrison, S. Yarosh, L. Terveen, L. Stead, and B. Amento. Collaboratv: making television viewing social again. In *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*, pages 85–94. ACM, 2008. (Cited on pages vii, 10, and 11.)

[20] C. Nguyen, Y. Niu, and F. Liu. Video summagator: an interface for video summarization and navigation. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 647–650. ACM, 2012. (Cited on page 5.)

[21] S. Park, G. Mohammadi, R. Artstein, and L.P. Morency. Crowdsourcing micro-level multimedia annotations: the challenges of evaluation and interface. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, pages 29–34. ACM, 2012. (Cited on page 12.)

[22] Herwig Rehatschek and Gert Kienast. Vizard-an innovative tool for video navigation, retrieval, annotation and editing. In *Proceedings of the 23rd Workshop of PVA: Multimedia and Middleware*, 2001. (Cited on page 11.)

[23] L. Riek, M. O'Connor, and P. Robinson. Guess what? a game for affective annotation of video using crowd sourcing. *Affective Computing and Intelligent Interaction*, pages 277–285, 2011. (Cited on page 12.)

[24] Keir Smith et al. Rewarding the viuser: A human-televisual data interface application. (Cited on page 7.)

[25] Savitha Srinivasan, Dulce Ponceleon, Arnon Amir, and Dragutin Petkovic. "what is in that video anyway?": in search of better browsing. In *Multimedia Computing and Systems, 1999. IEEE International Conference on*, volume 1, pages 388–393. IEEE, 1999. (Cited on page 2.)

[26] A. Tang and S. Boring. #epicplay: crowd-sourcing sports video highlights. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1569–1572, New York, NY, USA, 2012. ACM. (Cited on page 12.)

[27] W3C. SMIL Protocol. `http://www.w3.org/AudioVideo/`, 2008. [Online; accessed 04-May-2014]. (Cited on page 14.)

[28] Wistia. Does Length Matter? It Does For Video: 2K12 Edition. `http://wistia.com/blog/does-length-matter-it-does-for-video-2k12-edition`, 2012. [Online; accessed 04-May-2014]. (Cited on page 1.)

COLOPHON

This thesis was typeset with the `pdflatex` LaTeX 2_ε interpreter using Hermann Zapf's *Palatino* type face for text and math and *Euler* for chapter numbers. The listings were set in *Bera Mono*.

The typographic style of the thesis was based on André Miede's wonderful `classicthesis` LaTeX style available from CTAN. My modifications were limited to those required to satisfy the constraints imposed by my university, mainly 12pt font on letter-size paper with extra leading. Miede's original style was inspired by Robert Bringhurst's classic *The Elements of Typographic Style* [4]. I hope my naïve, yet carefully considered changes are consistent with Miede's original intentions.

Final Version as of June 17, 2014 at 16:24.