VideoJot : a multifunctional video annotation tool

Michael Riegler, Mathias Lux, Vincent Charvillat, Axel Carlier, Raynor Vliegendhart, Martha Larson

VideoJot: A Multifunctional Video Annotation Tool

Michael Riegler
Klagenfurt University
Klagenfurt, Austria
michael.riegler@aau.at

Mathias Lux
Klagenfurt University
Klagenfurt, Austria
mathias.lux@aau.at

Vincent Charvillat
University of Toulouse
Toulouse, France
Vincent.Charvillat@enseeiht.fr

Axel Carlier
University of Toulouse
Toulouse, France
Axel.Carlier@enseeiht.fr

Raynor Vliegendhart
Delft University of Technology
Delft, The Netherlands
r.vliegendhart@tudelft.nl

Martha Larson
Delft University of Technology
Delft, The Netherlands
m.a.larson@tudelft.nl

ABSTRACT

Videos are becoming more and more a tool of communication. There are how-to videos, people are discussing actions of others based on their recorded performance, e.g., in soccer, or they simply record videos of great moments and show them to friends and family. In this paper we focus on very specific how-to videos and present a novel, web based annotation tool, that combines (i) zoom, (ii) drawing, and (iii) temporal social bookmarking in video streams. Moreover, we present a short study on the usefulness of the tool to communicate general concepts of a specific video game based on a captured game session.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*

Keywords

demonstration, video annotation, zoomable video, social heatmap

1. INTRODUCTION

Videos are currently hugely popular on the Internet. YouTube alone ingests 100 hours of new video content per minute [7]. However, sometimes consuming videos is not just watching them. Users do not necessarily fully comprehend a video just by watching the raw video data. This can be caused by several factors. Either, the video requires theoretical knowledge (e.g., an online class), or multiple things happen at the same time (e.g., games replay). Or, even simpler, the viewers are not familiar with the video's context, i.e., the culture, setting, time, or language spoken.

The annotation tool at hand is specifically designed for complex video content with entities and events that are rare, detailed, rapid, or not easy to distinguish. Without the support of annotations, the essential elements of complex video content are overlooked or misunderstood by the viewers, especially those with little experience in the content domain. We are specifically interested in scenarios in which video annotations are used to support a discussion between two people about the video material, for example, an analysis of a recorded event, or a training session, in which one person is teaching another. The medical domain is one domain that has been treated in the past by a previous annotation tool prototype [8]. Here, however, we are interested in also accommodating videos with a faster pace and more actors. For this reason, we have selected the area of game play recordings, i.e., videos that have been recorded by people playing a video game. We combine three different types of annotations, in order to create one strong video annotation tool, which is able to support the user best in communicating based on the video.

Our goal is to keep the tool as simple and intuitive as possible. Using our annotation tool should help to:

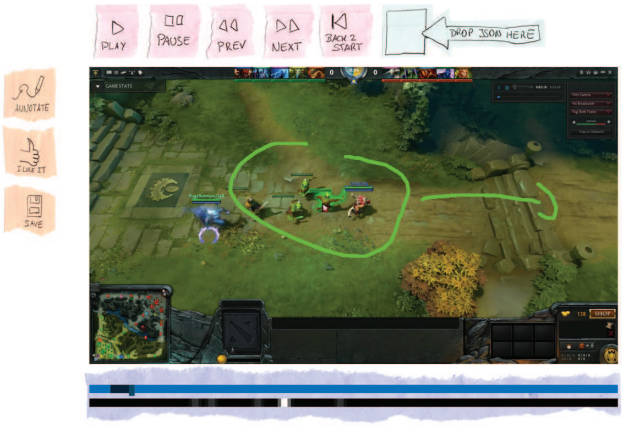

Figure 1: Annotation interface of VideoJot showing a video from the thinking aloud test. Buttons on top of the video allow for navigation while buttons left to the video are used for annotation. Below the video two time lines showing the annotations and the *likes* are given.

- Identify and share interesting events shown in a video (temporal information);
- Annotate and share interesting parts of these events (temporal and spatial information);
- Add information to enrich the content of the video (information enrichment).

We assume that annotations created with our tool do not simply add additional information to the video, but that those annotations also change the way how the video will be replayed and experienced by the viewers. Playback of annotated videos in our tool differs from conventional methods. The first difference is that the video player will not simply pause the video for replaying the annotator's annotation, but instead the annotation will be exactly synchronized to the video as it was at annotation time. That is, if the annotator added the annotation while the original video was playing, the annotation will be played during video playback as well. If the annotation was drawn when the annotator paused the video, video playback is paused as well during the replay of the annotation. The second difference is that the video player zooms into the video in the same way as the annotator did it at annotation time. Thereby, the tool creates a new and enriched version of the original video with an altered time line and viewport, which is an important advantage for the users of our tool.

In this demo paper we first outline the annotation methods used in our tool, and give a short introduction on the actual implementation. Then we present the results of a small thinking aloud test and conclude our contribution.

An online video which shows the function of the tool can be found on YouTube.[1]

2. VIDEO ANNOTATION

Most of the applications for video annotation use elements, i.e., text, speech bubbles and other kinds of overlays, that are simply "put on top of the video" and "added to the time line". The usage of these tools is typically complicated for untrained users and requires a significant amount of time. Current applications that feature such annotations are for instance, YouTube, VideoWiki and Popcorn Maker. YouTube provides text boxes and speech bubbles for the annotation [2]. VideoWiki [5] is a tool which allows the users to draw text and still annotations on parts of the videos for educational purposes. Furthermore, Mozilla Popcorn can be used for annotations but is mostly used for creating new content [1]. Most of the available tools only support static annotation overlays.

Our tool gives the user simple but very efficient methods to annotate videos. For simple temporal annotations, we provide *LikeLines* [9]. For temporal and spatial annotations and enrichment, we let the user draw arbitrary shapes on the video, either zoomed in or not, or with paused video or running video. To the best of our knowledge, no tool exists which allows the user to annotate a video in various ways at the same time in a very intuitive and easy way.

3. METHODOLOGY

We implemented a tool which combines LikeLines [9], a zoomable video interface [4] and hand-drawn annotations [8]. The implementation was done in HTML5 and JavaScript because of its multi-platform compatibility, which allows for deploying the tool on many different platforms (Windows, Linux, iOS) and devices (PCs, tablets, mobile phones). A system overview can be seen in Figure 2. To test if and how well the components work together we performed a thinking aloud test as a user study.

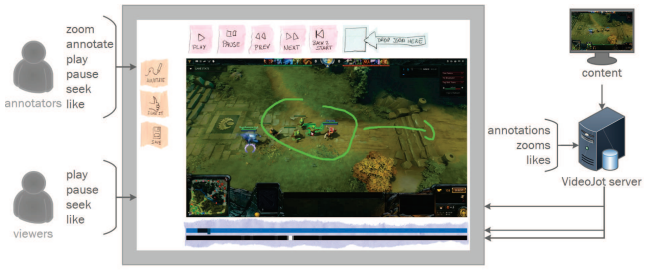

Figure 2: System overview for VideoJot showing the interaction between the users and the system.

3.1 Zoom

The technique used for zooming in VideoJot has been studied in [3, 4]. It consists of displaying users a scaled down version of an original video, but allowing them interactions in order to zoom in, zoom out, or pan on the video frame.

As shown in [3], zooming is also a way for the user to show its attention towards a particular spatial region on the video. Thus by zooming, users not only get more details on the video but also implicitly annotate as interesting a spatial sub-region of the frame.

In order to add a zooming interaction to VideoJot, we define the display size of the video as half the resolution of the original video (i.e., full HD, 1080p). Three modes of interacting are offered to the user:

- double-clicking on the video for zooming in;
- right-clicking for zooming out;
- when zoomed in, drag-and-drop can be used to move the viewport.

When users zoom in, only a subregion of the full video is displayed. When users zoom out, they are prompted once again with the scaled down version of the video.

The zoom browser in this tool should help the user to focus on interesting parts of scenes in a video. It can also be used during the drawn annotation process and can help to add more accurate annotations. Furthermore, the use of zoom in videos with very high resolutions can help the user to recognize important parts better.

3.2 LikeLines

With the LikeLines [9], a one-dimensional heatmap is presented below the video player. This "like line" visualizes the parts of the video which viewers like or most likely will like. The heatmap's intensity is driven by time-specific "likes" given by the viewers, but also by implicit information derived from their playback behaviour. This implicit information is captured when the viewer is interacting with the player and includes the parts that have been watched and those that have been skipped. Additionally, the initial state

[1] `http://youtu.be/cqF_1TWKSsQ`

of the heatmap can be seeded with content analysis techniques.

We adopt the idea of the LikeLines heat-map for VideoJot, but use it solely to display which parts of the video received explicit likes and do not include implicit playback information nor any output of content analysis methods. Furthermore, we display the heatmap as a secondary, separate time line below the annotations time line instead of merging the two into a single time line. As a result, a user can easily see the following at a glance: (i) annotations that are liked, (ii) annotations that have not been liked (yet), and (iii) parts of a video that have been liked but lack annotations.

3.3 Hand-Drawn Annotation

The hand-drawn annotations are based on an Android prototype which allows the user to annotate medical videos with drawn annotations, speech, speech to text and shakes (for important events) [8]. In this medical use case the usage of a tablet pc and multi-touch capabilities was convenient. Nevertheless, for multi-platform support we focus on HTML5 and JavaScript.

Basically, a user can draw on a video at any given time. When replaying the video the annotation and the process of drawing are replayed as created by the annotator. Beside this original functionality, we added the possibility to pause a video, then draw an annotation and then play the video again. This type of annotation is more convenient, especially when a lot of things happen at the same time. Replay then will pause the video and draw the annotation just like it was created by the annotator. Therefore, the overall time of the annotated video gets longer compared to the raw video stream.

Technically, the annotations themselves are drawn on a canvas, which is positioned over the video. For each annotation, the points of the annotation, the coordinates and the time, as well as the zooming information, are saved in a JSON (JavaScript Object Notation) data file on a server. For time synchronization, the time ticks of the HTML5 video player are used.

We did not implement the speech annotation and the speech-to-text annotation of the prior prototype because of (i) technical difficulties, (ii) the hypothesis that they might not be missed if not there, and (iii) PCs are more diverse than tablets and might not have a microphone. We also did not implement the shake annotation because common PCs lack movement sensors we can read easily from the web browser and are rather inconvenient to shake due to their size and weight.

We assume that the hand-drawn annotation provides the user with a tool to annotate interesting parts of scenes in a video and, therefore, to enrich the video with additional information. For example, it can be used to support teaching tasks, or for simple comments as someone would comment an image on Facebook. However, we also assume this type of annotations could support a more complex type of commenting, like what a sport commentator does during a soccer game.

4. USER EXPERIMENTS

While tools for annotating videos have already been tested in different domains, we here focus on the combination of different tools in one single annotation application and the applicability of it to one specific use case. Typically, either domains where annotation is needed by requirement or domains with a large audience are considered. Annotation has been applied to sports, e.g., soccer, basketball, baseball, football, and alike, to medical scenarios, to teaching, how-to and educational videos, and so on. However, recently web sites featuring live and recorded video streams from people playing video games have gained attention. Tens of thousands of people [6] are watching along when other play and discuss game play and strategy in a chat window, just beside the video player.

Computer games is a very broad area, as there are many games of different type around. In our study we investigated specifically one game, which is currently very popular: *DOTA 2*, published and maintained by Valve Corporation.[2] It is a fantasy action and strategy game, where five players encounter a team of five other players in an rectangular arena, called a map. The game is round-based, so two teams gather for a round, which takes roughly an hour, and work cooperatively against the other players in a king-of-the-hill like game. This game community has a strong notion of novice, advanced and expert users, and a lot of experience is needed to compete in the expert community, which also features leagues and prize money.

Many of the DOTA 2 players watch streams of live games to learn from the experts and increase their skills. Even more so, the game software records each and every game on the game servers and allows for replay and discussion of each game round. However, annotation of recorded games is not possible.

For evaluation purposes we chose to employ a *thinking aloud test*, where we ask DOTA players with a considerable amount of play time to save a short, interesting part of their choosing as a video and to annotate this video with our tool. While we only tested with two participants, the thinking aloud method allows for detailed inspection of the interface use and qualitative investigation of the annotation session. So while the results are in a way anecdotal, they still indicate the direction of results and provide a good start for a large study. Both session were recorded showing the interface of the annotation tool, the face of the participant (in a mirror) and the mouse and keyboard interaction. Both participants are PhD students at Klagenfurt University, play DOTA 2 significantly often and longer than a year, and are not participating in this research besides being study participants. The follow-up interview was also recorded. All recordings have been investigated by a group of three people to learn the most of the two session. The actual task of the participants was formulated as: *Use the annotation tool to teach a novel DOTA 2 Player a basic concept like gold mining, denial, harassing, lane dominance / distribution, managing creeps, jungle, etc. Please voice your thoughts while working with the tool.*

Both participants agreed that the free hand drawing based annotation helps a lot in trying to explain a basic concept of the game. But, they also agreed on the fact that while it is an intuitive way, one has to train to get "really good annotations" [sic!]. Both participants also agreed that a text annotation tool is necessary in addition to the free hand drawing. Whereas one participants would suggest either text or audio (speech based) annotation, the other participants suggested text in favour to audio, as he thought of the prototype as a

[2] `http://www.valvesoftware.com/`

Table 1: Data captured at the thinking aloud sessions.

Session	#1	#2
Participants age	24	29
Participants gender	male	male
Avg. play time per day	2 h	1 h
Watch game play videos	rarely	2-3 per week
Video length (h:mm)	1:19	1:22
Annotation time (h:mm)	5:05	2:38

live annotation (annotate while talking) tool. Other means for annotation mentioned only by one participant were slow motion, e.g., down to $\frac{1}{8}$th of the original playback speed and a magnification lens to increase the zoom on only a portion of the video, e.g., the mini map. Both participants agreed that editing capabilities for existing annotations would be beneficial. Moreover, the participants agreed that annotation of still images is more useful than the annotation on a running video. This is due to the fact that in DOTA 2 players move the camera rapidly, and then the annotation is immediately out of context when the annotated objects or actors are moved away. Possible reactions to overall movement where suggested by the participants including object tracking and automatic fading of annotations on global movement. The zoom function was received differently by the participants, being on the one hand considered a useful function, but on the other hand deemed unnecessary if he display size could be enlarged, e.g., by using overlay buttons. The LikeLines were considered useful if and only if many users were accessing the video and a full length DOTA 2 game round was considered, as then the hot spots would emerge from the one hour long video stream. Finally, both participants considered annotation functions like the ones they have tested in the study as useful for a large amount of game play viewers on the internet, if they are made available at the right place, e.g., on web based game play streaming portals.

All in all, the study indicated that the participants of the study were able to use the tool after a short tutorial, and annotation didn't take an extensive amount of time, as can be seen by the annotation times in Table 1. Moreover, while there is room for improvement, the users, themselves DOTA 2 players, considered such a tool helpful and of great potential within the game play streaming community.

5. CONCLUSION

We presented a tool which combines different ways to annotate a video, i.e., zoom, hand-drawn and "like" annotations. To investigate the interplay of these different annotation approaches we performed a thinking allowed test in a more and more important area, namely the gaming sector. These tests showed us that the combination of the tools works well. Nevertheless, there are still some improvements necessary. It is important to point out that the participants did find that some training is needed to make good annotations. Another important fact is that the users want to have at least text information. This shows us that drawings alone can not fulfil the information need for video annotations completely. Moreover, it seems that it is important to give the users the possibility to slow down the video during the annotation process. This will lead to more precise and better annotations, because they can manage the start and end time of events in a better way. Another important fact is that still annotations are more suitable for complex scenes in the videos compared to the moving annotations. The zooming function was not seen as necessary, but we think that it can be important for videos with high resolution or video games with a lot of details on the screen, e.g., a map. Finally, the LikeLines were considered as well suited for bookmarking interesting parts of very long videos. All in all, concerning to the user tests, the tool can be seen as innovative and well suited for the use case.

For future work, we will extend our prototype as recommended by the participants (text or spoken annotation and slow motion replay as well as annotation fading on global movement). Then we plan to set up a larger user test and a creation of a database, which contains annotated videos for further investigations like, e.g., do the annotations have a correlation with the likes in the LikeLines, etc. This will give us further important insights in the annotation and experience process of videos by users which is an important field in multimedia research.

6. ACKNOWLEDGEMENTS

We wish to thank Wei Tsang Ooi for working together on the original idea of the zoomable video player. This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and has received funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214/25557/37319 and the European Commission's 7th Framework Programme under grant agreement N° 287704 (CUbRIK).

7. REFERENCES

[1] Mozilla Popcorn Maker. `https://webmaker.org/en/tools`.

[2] Youtube Video Annotations. `http://www.youtube.com/yt/playbook/annotations.html`.

[3] A. Carlier, V. Charvillat, W. T. Ooi, R. Grigoras, and G. Morin. Crowdsourced automatic zoom and scroll for video retargeting. In *Proceedings of the international conference on Multimedia*, pages 201–210. ACM, 2010.

[4] 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.

[5] A. Cross, M. Bayyapunedi, D. Ravindran, E. Cutrell, and W. Thies. Vidwiki: Enabling the crowd to improve the legibility of online educational videos.

[6] M. Kaytoue, A. Silva, L. Cerf, W. Meira Jr, and C. Raïssi. Watch me playing, i am a professional: a first study on video game live streaming. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 1181–1188. ACM, 2012.

[7] A. Kokaram. Challenges in video ingest at YouTube. Keynote Talk at the Multimedia Modeling Conference (MMM) 2014, Dublin, IE, Jan 2014.

[8] M. Lux and M. Riegler. Annotation of endoscopic videos on mobile devices: a bottom-up approach. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 141–145. ACM, 2013.

[9] R. Vliegendhart, M. Larson, and A. Hanjalic. LikeLines: collecting timecode-level feedback for web videos through user interactions. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1271–1272. ACM, 2012.