mediacaptain – an interface for browsing streaming media

Florian Mueller
FX Palo Alto Laboratory
3400 Hillview Avenue, Bldg. 4
Palo Alto, CA 94304, USA
floyd@mediacaptain.com

ABSTRACT

The increase of bandwidth and streaming technology has made video on the Web the current "killer-app" of the dot-com world. However, users still face many problems. Users have to find the right video and the right segment within the video. Locally stored files provide easy (but still not very sophisticated) access to individual points in the video by utilizing a seek slider. If the video is streamed over the Internet, this slider loses much of its attraction. Every accessed point in the video requires the video player to buffer, which causes a time lag.
The mediacaptain is a system that addresses this issue by using supplementary material like text and graphics to provide indices. This time-aligned material is used to help the user make an informed decision on whether they want to watch a video and if so, what portions. This web-enabled prototype called mediacaptain emerged from user surveys and is demonstrated on several content types and represents an advanced experience with video on the Web.

Keywords

Streaming media, streaming video, buffering, supplementary text, supplementary graphics, video on the Web, video and text, user interface.

1. INTRODUCTION

Streaming media is "the" reason for a broadband Internet. Advancements in video technology and digital postproduction allow home users as well as traditional content providers in the television/movie market to produce quality video content easily, which they can distribute over the Web. Both parties use streaming technology to free the user from long waits. Once the streaming media player finishes downloading a predetermined amount of data, the video starts playing although the entire video data is not received so far. This so called "buffering" is necessary to intercept possible dropouts during the playback.

This greatly reduces the time for the user to wait due to bandwidth limitations.

Streaming technology facilities the user's experience, but is still far from easy -meaning instant- access to video on the Web. What if the user does not want to start watching from the beginning? Or quickly wants to jump to the next chapter? For each jump, buffering takes place making the user wait. These arbitrary jumps can be avoided by providing entry points in order to allow the user to make an informed assessment on where to jump within the video. Several solutions exist that create such index points, either authored in terms of annotations [1] or automatically generated [2]. The first utilizes textual content to retrieve index points, the later visual features within the video.

The mediacaptain is a system that tries to combine the advantages of both into a complete video-text or video-graphics solution that helps the user to determine, with contextual and visual support, the right point within a video to jump to. If this is achieved, it saves the user a lot of time waiting for the player to buffer and then realizing that this is not the point where the user wanted to start watching.

2. QUESTIONNAIRE

Due to the adolescent stage of video on the Web, it is important to see what the users' experience has been so far, what their needs are and where they see ways for improvement with streaming media.

A questionnaire was made available over the Internet to a selected audience, all of them very computer-literate. Of the 98 survey participants, only 66 had ever watched a video on the Web, and only 34 of them had actually searched for video at least once. None of the participants who watched video on the Web were satisfied with it: most of them criticized the poor quality and generally described their video on the Web experience as "too slow."

The users were asked further what they think would be the best representation for a video. They were introduced to the results in [3], which suggests a salient image with corresponding text as a good summary. If this could not be provided, the participants had to choose between three alternatives: just the image (23%), just the sentence (16%) or a combination of both, but the sentence would not really relate to the image (because it was done by a computer) (22%). Most of the participants were not able to make a decision (39%). Different users showed different needs, and this led to the design decision to incorporate text with a high level of user authoring possibilities.

The subjects also had to evaluate what they would see as a good representation of a sample video. Three possible solutions were presented:

1. An animated gif containing six keyframes extracted from a video in isochronous timeframes
2. A larger frame extracted from exactly the middle of the video, with the spot's title next to it
3. A sequence of three keyframes

The animated gif was favored by 66 participants out of the 92 who answered this question. Answer choices 2 and 3 were equally chosen by the remaining participants. These numbers changed little after the users were told what the video was about. This led to the design decision to emphasize the "moving images" character in the representation of a video, without utilizing a video skim, as also advised against in [4].

The study gave insights on the user's experience with streaming media, and it especially showed the current flaws of the technology. Two of the biggest complaints were the insufficient quality and the delays users had to face. Study participants complained that they sometimes do not want to start watching the video from the beginning, and this is where the annoyance of the buffering shows up repeatedly: Every time the seek slider is moved to an arbitrary point, the video has to buffer, most of the time just to show the user, that this is not the point the user wanted to jump to. So with the reduction of unnecessary jumps, which causes buffering, the delay time for the user can be dramatically reduced.

3. MEDIACAPTAIN

These findings led to the mediacaptain, which provides access points into the video. They are not only of textual or graphical content, but also provide visual aid from the video for the decision-making process of deciding whether this is the right point to start the video from, all without requiring the video to buffer.

A screenshot of the current implementation can be seen here:

Every screen consists of two main parts. On the left, the video window is embedded, on the right is the supplementary material, mostly text, positioned in an extra frame. The text is in HTML format and therefore allows easy formatting and supports all layout capabilities of the web browser.
In addition to that, the text exhibits additional functionality: If the mouse is moved over any parts of the text, the keyframe (a JPEG) for that exact point in the video is displayed overlaying the video. So instead of having multiple keyframes all displayed at once as mentioned earlier, they are displayed "on demand". If the user wants the visual information of what happened in the video at this point in the text, the user moves the mouse over and gets it right there.

The images, which are compressed JPEGs can be either preloaded or downloaded on demand. They are small in file size, so they can be quickly downloaded, and have the same dimensions as the video. The JPEGs are displayed overlaying the video, so it is clear for the user that they are part of the video, and before the user starts the movie, there is already a visual aid of what is going to be displayed upon a press on the start button.

For the user, it looks like the video would advance to the specific point and show the keyframe. In reality, it is not the video showing it (because this would require buffering), but pictures overlaying the video window.

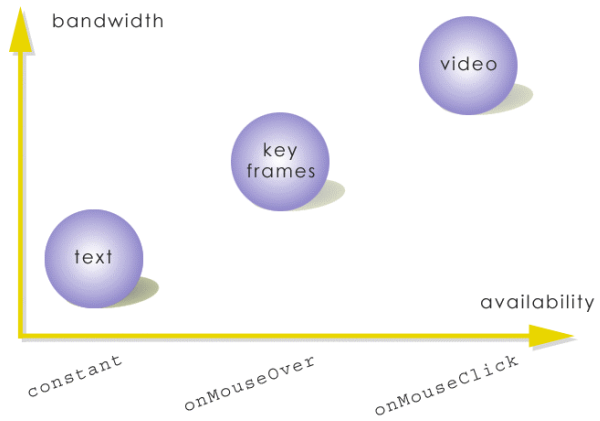
If the mouse is moved along while reading the text, it can become like a slideshow of changing images. The user can basically play the video, with a very small frame rate and no audio, at any speed. This is coherent with the results of the user survey, where the users favored the animated gif instead of static images.

By clicking on the text, the video starts playing at that particular point. While the video is playing, the active text segment is highlighted and scrolls into view. It is a bi-directional connection, which means interaction on one side triggers action on the other side.

In case of text as supplementary material, the user can also easily retrieve the desired index point by simply utilizing the search functionality of the web browser.

3.1 Interaction vs Buffering

The restrictions of streaming media require an interface design, which guides the user's interaction with the video. The following graphic shows a major concept of the mediacaptain:



The more bandwidth a medium requires, the more intervention from the user is necessary to access it. This way, the loading and buffering times are kept to an absolute minimum.

Text is almost instantly loaded, therefore, it is available immediately and displayed constantly on the web page. The user can start exploring the text and decide whether the video is of any interest. If the text does not give enough information to assess

relevance, the user can use the visual information of the keyframes to support this process. Keyframes require more bandwidth, so they are only available if the user requests them by moving the mouse over the point of interest. If the combination of text and keyframes provided sufficient information and the user decides to watch the movie at this point, a mouse click starts the video.

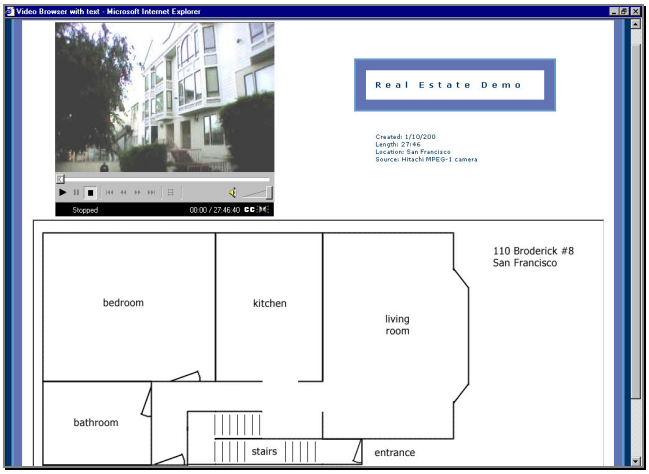
The more bandwidth that is needed, the more action is required from the user. The system leads the user to use text, keyframes and video in this order to assure that bandwidth is used optimally.

Users should not need to wait for a video to start. If they have to, the system functions as a guide to make sure that the user has chosen the right video and it starts playing at the right point.

3.2 Video and Graphics

"This sounds nice, but what if I don't have text? What if I have a graphic along with the video?"

The following shows a possible application with video content shot by a real estate agent. Important supplementary material to the video is not so much of a textual character, but more graphical, here a floor plan. With it, the potential buyer will get an idea of what the house will look like before even going there. The video is a taped tour the real estate agent made, which can be followed on the floor plan.



Below the video the floor plan is displayed. If the user moves the mouse over a room or any important marked area on the plan, the keyframe from the video will be displayed overlaying the video area. The keyframe is extracted at the point when the real estate agent in the video is entering the chosen room. If the mouse is clicked, the video starts playing at that point. While the video plays, the room that is being shown is highlighted in a different color. This way, the user knows which room on the floor plan the real estate agent is showing at the moment.

3.3 Only one file to author

The file format chosen to synchronize the video with the supplementary material is a text format that includes time-stamps, which is normally used for closed captioning for web content. It is called SAMI (Synchronized Accessible Media Interchange) and is a "…simple format optimized for authoring captions…in a single document." [5] It follows the XML specification and is therefore easily readable by human beings.

Sophisticated nesting of the tags needed for the mediacaptain and the SAMI specific XML tags allows creating one file: it is read in by the video, which controls the highlighting of the text, and is also read in by the text frame, and controls there the mouseOver and the display of the appropriate keyframes. This means, if anything in the text needs to be changed, there is only one file to alter.

4. CONCLUSION

The mediacaptain gives the user a better interface to browse streaming video. Through the implementation of text, graphics, keyframes and video, the user can make an informed choice on which medium to use for which purpose. Each medium has advantages in its area of representation and can be used interchangeably.

The user is guided along an interaction path that is required by bandwidth constraints. Text, keyframes and video are available to the user, but require different kind of interaction levels. These levels correspond to the bandwidth requirements of the text, keyframes and video.

A two-way connection between the text and the video allows the video to control the text, and vice versa, and also gives the user permanent feedback on the question, "Where am I?"

Users can use the textual index points to access specific parts in the video faster. They have a variable rather than a sequential way to browse the video, without losing the temporal aspect of the medium.

5. DEMO

A demo can be seen at http://www.mediacaptain.com, which is also mentioned in the demo section of the ACM Multimedia 2000 proceedings.

[1] Bargeron, D., Gupta, A., Grudin, J., and Sanocki, E. "Annotations for Streaming Video on the Web: System Design and Usage Studies," (1998), ftp://ftp.research.microsoft.com/pub/tr/tr-98-60.doc

[2] Foote, J., Boreczky, J., Girgensohn, A., and Wilcox, L. "An Intelligent Media Browser using Automatic Multimodal Analysis." In *Proceedings of Multimedia '98*, ACM Press, pp. 377 (1998), http://www.fxpal.xerox.com/PapersAndAbstracts/papers/foo98.pdf

[3] Ding, W., Marchionini, G., and Soergel, D. "Multimodal Surrogates for Video Browsing," http://www.clis.umd.edu/faculty/soergel/soergeldl99wdgmds.pdf

[4] Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. "Video Manga: Generating Semantically Meaningful Video Summaries." In *Proceedings ACM Multimedia*, ACM Press, pp. 384 (1999), http://www.fxpal.xerox.com/PapersAndAbstracts/papers/uch99b.pdf

[5] Closed Captions for Web Multimedia, http://www.microsoft.com/enable/sami/details.htm