

SILVER: An Intelligent Video Editor

Juan P. Casares

Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
juan.casares@cmu.edu

ABSTRACT

Silver is an authoring tool that allows novice users to edit digital video. A variety of AI techniques provide high-level metadata from the audio signal and video, including shot boundaries and a time-synchronized transcript. Silver uses this metadata to provide multiple, synchronized views of the content, including transcript, tree outline and hierarchical timeline views. These interface components are used to organize and edit the source material. The user can drag and drop representative frames or directly cut and paste in any view, including the transcript. Our work now focuses on supporting intelligent selection when there is a disparity between audio and video boundaries.

Keywords

Digital video editing, multimedia authoring, video library.

INTRODUCTION

Digital video is coming of age. The equipment required for video editing is increasingly affordable. For example, Apple claims that you can “turn your DV iMac into a personal movie studio” [1]. Also, digital video content is more accessible, as seen with the significant effort towards the creation of digital video libraries, and the amount of video available on the World Wide Web. However, tools for editing video still resemble analog professional consoles. Although they support the creation of high quality material, they are not easy for the casual user, especially when compared with applications such as text editors.

State of the Art

Current video editing software only operates at a low syntactic level, manipulating video as a sequence of frames and streams of audio. It does not take advantage of the content or structure of the video to assist this in the editing. Instead, it requires the user to pinpoint specific frames, which may involve zooming or numerous repetitions of fast-forward and rewind operations. For example, three-point editing is one option in professional editors such as Adobe Premier. In this kind of editing, the user must locate an “in point” and “out point” pair in the video source and a third point in the target to perform a copy operation (this

method can be traced to the use of physical videotape). However, this type of editing is very different from other consumer software that users might be familiar with.

Related Work

There is a large body of work on the extraction and visualization of information from digital video. However, most of it has focused on search and summarization, whereas we focus on authoring with the content once it is found.

Other systems that also use metadata to assist video editing include Hitchcock [2] and Impact [4]. The first automatically determines clips and their start and end points based on shot quality and desired duration. Impact creates a high level description of the structure of a video and allows users to directly modify it using different visualization methods.

SILVER

We are creating an intelligent video editor, called Silver, which has high-level tools for working with the material. Silver offers a set of views of a video, with different semantic content and levels of abstraction. It provides mechanisms to select and edit the material in any of these views.

Currently, we try to intelligently resolve the inconsistencies that arise as a consequence of the different semantic boundaries in audio and video.

Informedia

We obtain our source video and metadata through CMU's Informedia Digital Video Library [4]. Currently, this searchable multimedia library consists of 2,000+ hours of material. Informedia implements a fully automatic intelligent process to enable daily content capture, analysis and storage in on-line archives. Applying speech recognition, natural language processing and image analysis, Informedia is able to automatically provide titles, time-aligned transcripts, shot breaks, representative frames, summaries, identified faces, optically recognized text and geographical information.

Multiple Views

The basic unit in Silver is the clip, which represents a segment of video. The project view shows all clips the user is working with. Clips can be further organized using the tabbed list and tree outline views. These views display clip information such as the automatically extracted title and representative frame. The storyboard view supports multiple storylines for interactive video and hypermedia.

Two time-based views for the produced video are shown in Figure 1: a transcript and hierarchical timelines. The transcript view is a textual representation of the audio, and may be edited in a similar way to typical text editors.

To allow the user to work simultaneously with different levels of detail and context, we use a hierarchical timeline view, where the topmost level represents the whole video and the others offer increasing levels of detail. The user directly manipulates the position and degree of zoom. A similar approach is described in [3].

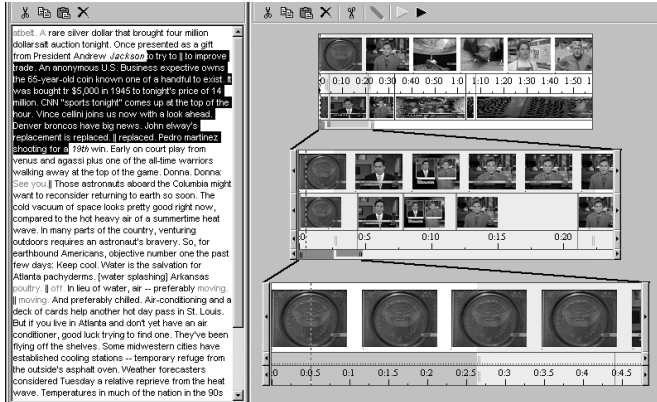


Figure 1. Hierarchical timelines and transcript views.

Each timeline offers a different view of the video, such as clip, shot, time, frame and annotation. Snapping and unit selection (by double-clicking) are specific to each view.

Synchronized Selection

Silver maintains a single selection coherent across all of the multiple views. However, the objects manipulated by the system, from clip, shot, word, to frame, all have different granularities and usually correspond to overlapping periods in time. We denote this with different selection highlights for “wholly selected” and “partially selected”. For example, selecting a few frames of video may only select part of a word.

In the transcript view this is represented using a different font for entirely and partially selected words. Editing can turn this into a bigger issue. If the user splits a clip in the middle of a word, corresponding parts of the word end in each resulting clip. Thus, Silver duplicates the word in both clips, and uses font to represent their partial state.

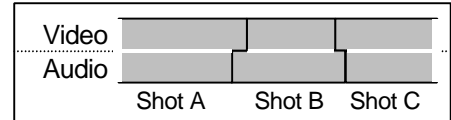
Intelligent Selection

Often, a shot in video doesn’t match exactly the corresponding audio. For example, voice can start a second before the talking head is shown. This gives the video a continuous, seamless feel, but makes extracting pieces much harder. Figure 2 shows this situation: the audio corresponding to Shot B extends before and after the corresponding video.

We are working on an intelligent editing feature that can take advantage of the metadata about the video to use heuristic rules and anticipate the user’s desired selection in

both audio and video. When a user selects a segment, we snap video and audio differently, based on their content.

Figure 2. Shots with differing audio and video.



A cut and paste operation with differentiated selection is indefinite. However, the interface can use heuristics to provide different levels of automation. For example, it would automatically overlap two streams of audio if one of them contained only silence; if it believes the audio is not synchronized with the video, it might ask the user if it is OK to shift it. The user would manually handle cases where no rule applies.

FUTURE WORK

Some informal user testing with experienced video editors has already provided a wealth of data. We plan to test Silver with a broader user set, including middle-school students.

We plan to incorporate more of the metadata that Informedia can provide, such as recognized faces. Incorporating the Informedia analysis pipeline will allow the use of Silver to edit home video and other camcorder content.

We are devising an agent that could provide feedback on the quality of the authored material. We intend to use heuristics such as “avoid shaky footage” (as in [2]) or “avoid cutting in the middle of a camera pan”.

We expect the result to be a video editor that is as easy to use as a text editor while helping novice video editors create better quality productions.

ACKNOWLEDGEMENTS

The Silver Project is funded in part by the National Science Foundation under Grant No. IIS-9817527, as part of the Digital Library Initiative-2. I thank Albert Corbett, Laura Dabbish, Scott Stevens and specially Brad Myers for the invaluable help in guiding the work described in this paper. Silver is built upon previous work by Dan Yocum.

REFERENCES

1. Apple. “iMac.” <http://www.apple.com/imac/>
2. Girgensohn, A., Boreczky, J., et al., A Semi-automatic Approach to Home Video Editing, *UIST '00 Conference Proceedings*, ACM Press, pp. 81-89, 2000.
3. Mills, M., Cohen, J. and Wong, Y., A Magnifier Tool for Video Data. *CHI '92 Conference Proceedings*, ACM Press, pp. 93-98, 1992.
4. Ueda, H., Miyatake, T., Sumino, S. and Nagasake, A. Automatic Structure Visualization for Video Editing, *INTERCHI '93*, ACM Press, pp. 137-141, 1993.
5. Wactlar, H., Christel, M.G., et al., Lessons Learned from Building a Terabyte Digital Video Library, *IEEE Computer*, (32) 2, pp. 66-73, 1999