Simplifying Video Editing with Intelligent Interaction

Juan Casares, Brad A. Myers, A. Chris Long, Rishi Bhatnagar,
Scott M. Stevens, Laura Dabbish, and Albert Corbett
Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
email@cmu.edu
http://www.cs.cmu.edu/~silver

ABSTRACT

Digital video is becoming increasingly ubiquitous. However, editing video remains difficult for several reasons. It is a time-based medium, it has dual tracks of audio and video, and current tools force users to work at the smallest level of detail. In this paper, we describe several visualization and interaction techniques that use video metadata, including a transcript, to mitigate the problems of editing in this domain. We implemented these techniques in Silver, an authoring tool designed to make it easier for novice users to edit video. To help users visualize video, Silver provides multiple views with different semantic content and at different levels of abstraction, including storyboard, editable transcript and timeline views. Silver offers intelligent editing operations that help users resolve the inconsistencies that arise because of the different boundaries in audio and video. We conducted a preliminary user study to investigate the effectiveness of the system. Participants successfully edited video after only a short tutorial, both with and without intelligent editing assistance. Our research suggests several ways in which video editing tools could use metadata to assist users in the reuse and composition of video.

Keywords

Digital video editing, multimedia authoring, intelligent user interfaces, Silver, Informedia.

INTRODUCTION

Digital video is becoming increasingly ubiquitous. Digital video equipment is more accessible than ever and there is an increasing amount of video material available on the World Wide Web and in digital libraries. As technology and content become available, users will shift from passive spectators to active creators of video. Forrester Research predicts that by 2005, 92% of online consumers will create personal multimedia content at least once a month [15]. However, although many exciting research projects are investigating how to search, visualize, and summarize digital video, there is little work on new ways to support the use of the video beyond just playing it.

This is unfortunate, because video editing has several unique challenges not found with other media. One is that digital video is a time-based medium. This property makes it difficult for users to browse and skim video. Users often must linearly search their source video to find the clip they desire.

Another challenge for editing video is that it is a dual medium. Most "video" actually consists not just of a video track but also an audio track. These tracks must be kept synchronized, but the user must also be able to overlay them when desired, for example during transitions from one shot to another. Further, when a shot is cut from a video for use elsewhere, the user must be able to disentangle overlaid audio and video.

A third problem is that the syntactic units that users want to edit are shots of video and words or sentences of audio, but current tools require users to examine video at the individual frame level and audio using a waveform. To perform most editing operations, such as cutting a shot, the user must manually pinpoint specific frames, which may involve zooming and numerous repetitions of fast-forward and rewind operations. Finding a specific word or sentence using a waveform is similarly tedious.

These problems make editing video a difficult, tedious, and error-prone activity. Commercially available tools, such as Adobe Premiere and Apple iMovie, allow creation of high-quality video, but they do not adequately address the issues raised above, which makes them harder to use, especially for novices.

We created a system, called Silver, to explore techniques to make video editing easier. The key innovations in the Silver editor include: providing an editable transcript view; coordinating the selection across all views including partial selections and different selections in audio and video; and intelligent editing through smart snap, smart selection and smart cut and paste. Silver uses video and metadata from the Informedia Digital Video Library [19].

This paper is structured as follows. First, we discuss related work on video systems. Next, we explain the Silver interface and intelligent video editing. Finally, we describe a pilot user study and conclude.

Submitted for Publication

RELATED WORK

In this section, we describe systems that use metadata for video editing. We look into systems that go further, automating the process of video editing to some degree. We discuss work on video visualization that addresses the issues of scale and time. Finally, we describe Informedia, which is our source of video and its corresponding metadata.

Metadata and Video Editing

The notion of using metadata for editing video is not new. For example, Mackay and Davenport examined the role of digital video in several interactive multimedia applications and concluded that video is an information stream that can be tagged, edited, analyzed and annotated [12]. Later, Davenport et al. proposed using metadata for home movie editing assistance [6]. However, they assumed this data would be obtained through manual logging or with a "data camera" during filming.

Currently, there is a large body of work on the extraction and visualization of information from digital video (e.g., [16, 7]) that make a data camera unnecessary.

IMPACT [18] uses automatic cut detection and camera motion classification to create a high level description of the structure of the video. The user can organize the shots in a tree structure and then edit the composition by moving the branches [17]. IMPACT supports this process by recognizing objects across multiple shots.

Automation

Several systems edit video with varying degrees of automation. Fully automated systems may be used for news on demand [1] or quick skimming [4], but do not really support authoring.

The Video Retrieval and Sequencing System (VRSS) [5] semiautomatically detects and annotates shots for later retrieval. Then, a cinematic rule-based editing tool sequences the retrieved shots for presentation within a specified time constraint. Examples of cinematic rules include the parallel rule, which alternates two different sets of shots, and the rhythm rule, which selects longer shots for a slow rhythm and shorter shots for a fast one.

The Hitchcock system [8] automatically determines the "suitability" of the different segments in raw home video, based on camera motion, brightness, and duration. Similar clips are grouped into "piles." To create a custom video, the user drags segments into a storyboard and specifies a total desired duration and Hitchcock automatically selects the start and end points of each clip based on shot quality and total duration. Clips in the storyboard are represented with frames that can be arranged in different layouts, such as a "comic book" style layout [2].

While automation makes editing faster, it usually involves taking away power from the user and this is not always desirable. In fact, user studies led to changes in Hitchcock to give more information and control to the user [9].

Visualizing Time and Scale

Video editing is hard in part because it is hard to visualize time. Due to the semantic and syntactic nature of the video, where selecting a five-second piece involves fiddling with frames in $1/30^{th}$ of a second increments, it becomes important to present the information at different scales. No one scale is sufficient to efficiently perform all the operations needed.

The Hierarchical Video Magnifier [13] allows users to work with a video source at fine levels of detail while maintaining an awareness of the context. It provides a timeline to represent the total duration of the video source; a second timeline shows a detail of the first one and illustrates it with a frame sampling. The user can select which portion of the top timeline that is expanded on the next one. There is also a mechanism that can be used to add a new timeline with a higher level of detail. Successive timelines create an explicit spatial hierarchical structure of the video source. The Swim Hierarchical Browser [20] significantly improves on this idea by using metadata. Swim displays automatically detected shots in the higher level layers instead of frame samples. These tools were only used for top-down navigation, and not for editing. We also use a hierarchical view in our timeline.

Informedia

Silver obtains metadata from the Informedia Digital Video Library [19]. This is a growing searchable multimedia library that currently has over 2,000 hours of material, including documentaries and news broadcasts. Informedia adds two hours of additional news material every day.

For all of its video content, Informedia creates a textual transcript of the audio track using closed-captioning information and speech recognition [7, 10]. The transcript is time-aligned with the video using CMU's Sphinx speech recognition system [14]. It is important to note that no other system described here uses metadata derived from the audio track.

Informedia also performs image analysis to detect shot boundaries and extracts representative thumbnail images from each shot [4]. All of this metadata about the video is stored in a database. This metadata is used by Informedia to automatically create titles, representative frames and summaries for video clips, and to provide searching for query terms and visualization of the results. Silver takes advantage of the metadata provided by Informedia to enhance its visualization and editing capabilities.

THE SILVER INTERFACE

Text editors have long had multiple representations in order to support different styles of editing and different tasks. For example, Microsoft Word provides "outline", "normal", "print layout," and "print preview" views. Silver provides multiple views of video to address the problems of video editing described above. These views allow the user to browse the video by displaying it across space instead of time, to view and manipulate the audio and video separately, and to examine the video at multiple levels of

Figure 1. Overview of the Silver interface

detail. An overview of the Silver interface is shown in Figure 1.

The remainder of this section describes these views and how each supports video editing. The first subsections describe three important views, the storyboard, the timeline and the transcript. Following is a brief description of other parts of the Silver interface. The final subsection describes how Silver synchronizes across views.

Storyboard View
This view allows the user to arrange and rearrange the video composition at a large scale. It displays each *clip*[1] in the composition as a separate icon, containing a title and a key frame. The title is obtained from Informedia. Initially, we dynamically analyzed each clip to identify a representative key frame. However, early users were confused when the displayed key frame changed as they edited the clip and preferred using a static frame close to the start of the clip. The clips are sorted chronologically, so moving clips within the storyboard changes the order in which they appear in the final video.

In addition to the composition, the storyboard can also contain a "scratchpad", or staging area. Clips in this area are not part of the composition but are easily available, which is convenient for pieces under consideration or being edited. A vertical line separates clips of video in the scratchpad from clips currently in use. Dragging a clip across this line adds or removes it from the composition.

Hierarchical Timeline View
Silver, like most other video editors, provides a timeline view as the primary view for detailed editing. Silver allows users to easily view and manipulate video at different scales by providing a three-level view—clip, shot and frame levels (see Figure 2). The timeline allows the user to work at a high level of detail without losing the context within the composition.

The top line is the clip level, which represents the entire video. At the top, it shows two colored bars, one for audio and one for video, that show clip boundaries and gaps in the tracks. The central section of the clip level shows a more detailed view of each clip than is shown in the storyboard. In addition to a representative frame, it shows a timeline and if space allows it shows the end frame, title, and duration.

Users control how much is displayed in the other two levels of the timeline using a sophisticated scrollbar at the bottom of the clip level. In addition to the usual dragging behavior, the scrollbar may also be resized to zoom the shot level in or out. It also contains a miniature version of the shot level scrollbar, which shows the position of the frame level within the shot level and can be dragged to scroll the frame level.

The shot level allows users to interact with video at a higher level than individual frames by displaying video as individual *shots*[2], as detected by Informedia. Algorithms that detect significant between-frame image changes identify the shot boundaries [11]. Each shot is displayed

---

[1] A clip is a continuous segment of video from a single source file.

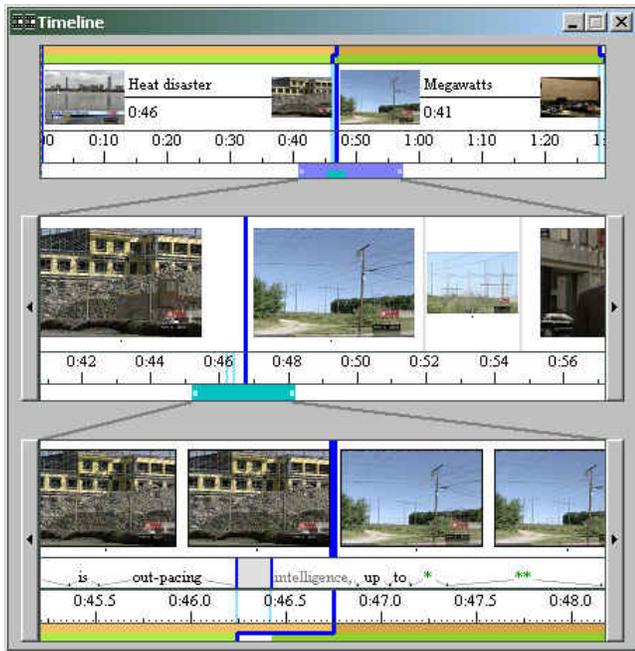[2] A shot is a continuous segment of video from a single camera.

Figure 2. Timeline window. The scrollbar in the clip view can be moved and resized to scroll and zoom the shot view. The shot view scrollbar can be moved and resized to scroll and zoom the frame view.

using the representative frame for the shot as chosen by Informedia. The size of the frame is proportional to the duration of the shot. As in the clip level, there is an augmented scrollbar at the bottom of the shot level used to scroll and zoom the frame level.

The frame level can display the individual frames of the video, so the user can quickly find particular cut points. The frame level addresses the dual medium issue by displaying the transcript, including the start and end times of every word. In other editors, there is little information regarding the audio, so users must repeatedly play the audio to find a desired part. The time-aligned transcript Silver provides allows the user to locate the desired part of the audio and video more quickly and easily than is possible with conventional editors.

At all levels, boundaries between clips are shown as blue lines. The user can resize a clip by dragging the clip boundary.

A key feature of the Silver timeline is that it allows different representations to be shown together, allowing the user to see the clip boundaries, the time, samples of frames, the cuts, the transcript, annotations, etc. In the future, other representations could be added, such as audio waveform or labels for recognized faces.

The Hierarchical Video Magnifier [23] and Swim [20] also provide multi-level views of video content. However, these systems are designed for browsing, and navigation is achieved drilling to finer levels of detail by adding levels to the timeline. In Silver the basic interaction for navigating in the timeline is scrolling and zooming, which is better suited

for editing tasks. In addition, Silver is different in that it combines multiple representations of the video within each level.

Transcript View

An important innovation in Silver is the provision of a time-aligned textual transcript of the video. Silver offers a text-editor-like window dedicated to the transcript (see Figure 3). The user can edit the *video* by cutting and pasting the text in the *transcript*. This view also provides an easy way to browse the video for specific content words. The transcript is also shown in the frame level of the timeline view, however, there it is harder to read and edit because it is shown mapped to time in a single line.

To display information outside the spoken words we insert special characters into the text. For example, the boundary between clips is shown as a blue double bar ("||"), and sections without audio are represented with dashes ("-"). This is shown near the top of Figure 3.

Because speech recognition can contain mistakes, Silver inserts a green star ("*") where there appears to be a gap, misalignment, or noise. Silver allows the user to manually correct mistakes in timings or recognition. Users can type the correct text and show what video is associated with it. In the future, we hope to use these corrections to guide the speech recognizer in the text alignment, so users will not need to align the text manually. We also plan to use the transcript view to support the authoring of new productions from scripts. The user could type or import a new script, and then later the system would automatically match the script to the video as it is shot.

Other Views

Silver provides the user with many other views. The tree view and the subject view permit the classification of clips in a hierarchical or flat taxonomy to help users organize their composition. The source view provides access to all the source material in one convenient place. The selection view shows the first and last frames of the selected video. The preview window is used for video playback.
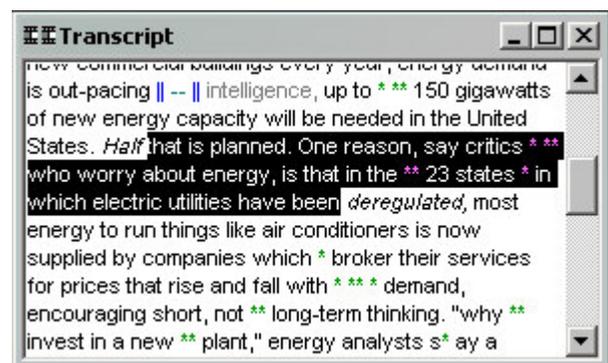

Figure 3. Transcript window. In the top line, we can see blue clip breaks ("||"), enclosing a gap in audio ("- -"). The word *intelligence* was cut in the middle. Green stars represent noise or recognition errors. At the edges of the selection, *half* and *deregulated* are partially selected.
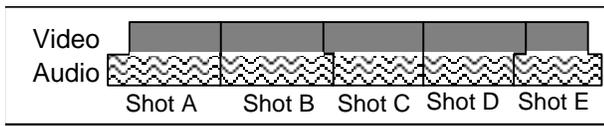
Silver is built on top of the Informedia client interface, so all the search, browse and summarization views provided by Informedia are available to the user. For example, the user can use the Informedia Search window to enter a keyword and then drag one of the clips displayed in the Search Results to one of the Silver windows.

These views are described in more detail in another paper [*anonymized*].

Many Views, One Model

All of Silver's views actually represent the same composition. Any change to the composition is reflected immediately in every view. For example, when the user drops a new clip into one of Silver's views, all others are instantly updated.

In addition, selection in one view is mirrored across the others. This feature brings up a number of interesting user interface design challenges.

First, the selection in video may begin and end at different times from the selection in audio. As is described below in the intelligent editing section, this feature is critical to achieving high quality when composing previously edited content. Some views only show the audio or the video track; these views only show their corresponding selection.

We use color to show the difference between selections in the different media: orange for video, green for audio, and yellow for segments selected in both (see Figure 4). Usually, users want to select audio and video together, but sometimes they need to separate the two. To select only one track, users use the right mouse button the same way they would normally use the left. For example, dragging the right button across the frame view of the timeline would select a piece of video without selecting audio.

The second problem that arises when having a synchronized selection is defining what the "equivalent" portion is. The different views show different levels of granularity, so it may not be possible to represent the selection accurately in some views. For example, if a short segment is selected in the timeline view, as shown in Figure 4, it is not obvious what should be shown in the storyboard view, since it shows only complete clips. Silver uses a dark shade to highlight a selection, and a lighter shade to highlight an item that is only partially selected.

Another example of this problem arises between video and audio views. Individual words in the transcript span multiple frames. When a word is selected in the transcript, it is straightforward to select all the corresponding frames. However, selecting only one of those frames in the video will correspond to only part of that word. This partial selection is shown in the transcript by making the word italic, for example the words "half" and "deregulated" at the edges of the selected text in Figure 3.[3] In this case, if

Figure 4. Smart selection. Audio and video can be selected separately, and video selection snaps to shot boundaries.

the selected video is moved, the word will be cut into two pieces. Silver represents this by repeating the word in both places, but showing them in a faded color. The word "intelligence" in Figure 3 was split.

INTELLIGENT EDITING

Silver tries to reduce the tedium of editing video by using metadata to make many interactions more intelligent. One large source of this tedium is L-Cuts, which are a common feature of edited video. In this section, we describe L-Cuts, followed by smart snap, selection, and cut and paste.

L-Cuts

When working with video that has been professionally edited, it often happens that the start and end times of a segment are different for the audio and the video. This is called an "L-cut". For example, the voice can start before the talking head is shown (Figure 4). This gives the video a continuous, seamless feel, but makes extracting pieces much harder because the video and audio portions must be adjusted separately, which requires much fine-tuning to remove extraneous video or audio portions.

We did an informal study to investigate the prevalence of these cuts, and found it to be significant, at least in recorded news shows. In looking at 238 transitions in 72 minutes of video clips recorded from CNN by Informedia, 61 (26%) were L-cuts. Of the 61 L-cuts, 44 (72%) were cases where the audio of the speaker came in before the video but ended at the same time. Most of these were interviews where the voice of the person being interviewed would come in first and then the video of the person after a few seconds. Most of the other ways to overlap the video and audio were also represented. Sometimes, the audio and the video of the speaker came in at the same time, but the audio continued past the video (Figure 5-Shot B) or the video ended after the audio (Shot D). When an interviewee was being introduced while he appeared on screen, the video might come in before the audio, but both end at the

---

[3] We would prefer the selection highlight to be green over fully selected words and light green over partially selected words, to

be consistent with other views. Unfortunately, the Visual Basic text component we use does not support this.

Figure 5. Different kinds of L-cuts

same time (Shot C). To copy or delete any of these would require many steps in other editors.

The Silver editor removes much of the tedium associated with editing such video by automatically adjusting the portions of the video and audio used for selection, cut, copy and paste, in the same way that text editors such as Microsoft Word adjust whether the spaces before and after words are selected. These capabilities are discussed in the following sections.

Smart Selection

One reason that video editing is much more tedious than text editing is that in video the user must select and operate on a frame-by-frame basis. Simply moving a section of video may require many minutes while the beginning and end points are laboriously located.

Silver makes some of these tasks much easier by providing a mechanism to select a meaningful segment based on the view. For example, in the shot row of the timeline, a double click selects the corresponding shot. In some cases, a larger segment is selected with a triple click, such as in the transcript view (where a double click selects a word and a triple click selects a sentence).

When the user selects a region, Silver may adjust the audio or video boundaries based on editing rules of thumb and the available metadata. This is especially useful when the source contains L-cuts.

For selections in an audio-based view, Silver may snap the edge of video selection to a shot break. To avoid flashing a few frames of another shot, we use a threshold of one second when the shot break is outside of the audio selection. For example, the right edge of the video selection in Figure 4 is snapped to a shot boundary. As shots become longer than a second, they no longer are objectionable. We use a threshold of 0.3 seconds when the shot break is inside the selected region. It is preferable not to jump to the middle of a shot.

For video-based views, if the selection cuts a word in half then Silver extends or reduces the audio selection to the nearest word edge. This is shown in Figure 4 where the left edge of the audio selection snaps to the word "yes". This might take the audio selection across a shot break.

Selecting in a mixed media view causes the selection to snap in both audio and video. However, the user is always in control: he or she can disable the smart selection by pressing a modifier key.

Smart Snap

When moving the cursor around in the timeline view, the cursor snaps differently to features of the composition depending on which row the mouse is moving over. In
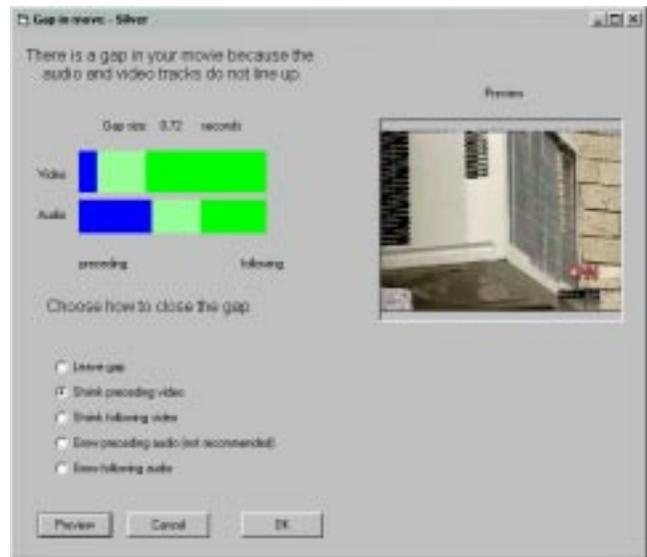

Figure 6. Close gap dialog. This dialog would appear if the user deleted the selection in Figure 4.

video-based rows, the cursor snaps to shot breaks if close by, or to the nearest frame otherwise. In the transcript row, the cursor snaps if close to a word boundary. In the case of silence, the cursor does not snap. In addition, the cursor always snaps to clip boundaries and the edges of the current selection.

Smart Cut and Paste

When a segment with different audio and video edges is deleted or pasted, the resulting ragged edges would typically require the user to again go down to frame level and adjust the edges or create a special effect to avoid gaps in the composition. For example, if a segment where the video is shorter than the audio (such as Shots A, B, or E in Figure 5) is inserted, the user would need to close the gap in video by either extending the video or shrinking the audio on either end of the gap.

Silver makes this task much easier by semi-automatically determining how to deal with the uneven ends. It presents a dialog box to the user with different options to close the gap. This dialog visualizes the current gap and how the different options would close it. For example, after deleting the segment that is shown selected in Figure 4, a dialog box like the one in Figure 6 would appear. This dialog is needed because more audio than video is selected in Figure 4, so deleting the selection would leave a gap in the audio. Deleting Shot E in Figure 5 would have the same effect.

If there is a gap in audio (e.g., if Shots C or D from Figure 5 are pasted), then Silver recommends leaving it because silence is generally not disruptive. However, it is not acceptable to leave a gap in video, which would show up as black or blank video. Silver recommends extending the edges of involved video as long as it doesn't take them across a shot boundary.

In cases where the editing operation leaves two gaps, such as when a clip with jagged edges on both ends is inserted,

Silver presents a double version of the dialog box where the user can look at both ends simultaneously.

We are exploring when these adjustments can be fully automated in a reliable way. We are also looking at other metadata that might be helpful in giving a recommendation. For example, we are considering mixing the audio of both edges when one audio track is music or other background sounds. We also want to increase the palette of options for handling gaps in the composition. For example, a third piece of video footage can be placed to cover the gap area, or a "special effect" like a dissolve can be used to smooth the transition over the gap.

USER STUDY

We conducted a pilot study to evaluate the editing features of Silver. Participants were asked to use Silver to perform a series of editing operations on a piece of video with intelligent editing enabled, and another series on a different video, this time with intelligent editing disabled. Our participants were seven paid subjects, graduate students from the School of Computer Science.

Design

This study was a within-subjects design. All participants did both tasks. Each performed one task with smart editing enabled and one with smart editing disabled. The order of the tasks and the order of smart editing were counterbalanced across participants.

Procedure

The experiment had three parts: tutorial, experimental tasks, and post-experiment questionnaire.

In the first part, participants read the Silver tutorial and performed training exercises in it. They learned how to scroll and zoom the timeline, select video, and play video. They also learned how to delete segments of video and clean up the cuts. They performed a practice task in which they manually cleaned up a cut and a practice task in which they used the smart cut and paste.

In the second part of the study, each participant performed the two experimental tasks. For each task, participants were given one of two video compositions and a corresponding list of edits to be performed. One video composition was about the Mars polar lander; the other was about forest fires in Russia. The source videos for each task were obtained from CNN footage. The videos and shots to cut for each of the two tasks were selected to be equally difficult. We also tried to select entertaining material.

Participants were asked to cut specific shots from the composition. After each cut, they were instructed to play the transition and, if they judged it to be "rough", to edit it to make it "smooth".

At the end, participants filled out a paper questionnaire about the study and demographic information. Users were asked about the system in general, the intelligent editing features, the tutorial, and each of the two tasks.

Results

On average, both experimental tasks together took participants 33 minutes. There were no statistically significant differences in time to complete either task due to task, order, or smart editing.

We did find significant correlations between two questions on the questionnaire and the time to complete the task with smart editing disabled. Participants who rated "exploration of features by trial and error" as more "encouraging" than "discouraging" (on a scale of 1-9) took longer to complete the task (correlation 0.82, $p < 0.046$). This may be due to participants who felt encouraged to explore the interface spending time with parts of the interface that were not directly related to the task.

Qualitatively, nearly all participants made a positive comment about Silver, such as describing it as "cool software, very intuitive to use". Four users noticed a learning effect; one said that one task was "easier because it was the second task". Another user mentioned, "after I could familiarize myself with it, it became easier to use and not as intimidating".

The unconventional scrolling and zooming of the timeline disconcerted five users. However, once the mechanisms were explained, they were able to use it. Users felt disoriented after deleting a shot. When a shot is deleted, timelines in other systems typically shrink, whereas in Silver it maintains its size by readjusting its zoom level. This sometimes led to participants having difficulty locating its former position. A brief animation might make this explicit to users.

Discussion

We were very encouraged by the fact that all of the study participants were able to succeed in both of their editing tasks, especially given the brevity of the tutorial.

However, we expected to find a greater effect of using the intelligent editing options. The few video editing experts that have been involved with this project were enthusiastic about incorporating these features. Apparently, our naïve users could not distinguish between the different options offered by the intelligent cut and paste dialog. This suggests that we should: a) automatically carry out the recommended action when the effect is small, and b) provide a better visualization of each of the options in the dialog.

We suspect that participant motivation was an important variable in performance. Some users seemed to try to finish as soon as possible. Time to complete a task might not be the best measurement of the effectiveness of a creative authoring tool, since it penalizes exploring different options and paying attention to detail. This might explain the significant correlation described in the previous section. In the future, we hope to study the quality of the produced video which might provide richer information.

CONCLUSIONS

Silver is a system designed to make quality video editing easier. It demonstrates several techniques to overcome some of the main difficulties of authoring in this domain. Namely, it uses multiple views that display not only the raw audio and video but also higher-level metadata. A single selection is coordinated across all views, even when the audio and video are unaligned. A textual transcript is linked to the video, providing a view that affords quick and accurate editing in the audio domain. Silver provides a three-level timeline, presenting both context and detail, which can be a useful tool for video editing.

Silver shows that video metadata can be used to create an intelligent interface for aiding novice users in the performance of otherwise difficult or tedious editing tasks. It demonstrates three forms of intelligent editing, with selection, snap, and cut and paste. Currently the metadata used is only the clip titles, the transcript, and shot breaks. However, as more metadata becomes available, the smart editing features can be improved. For example, we could adjust the selection to match speaker changes or to avoid cutting in the middle of a panning or zooming camera operation.

The Silver project is investigating some exciting ways to make quality video editing easier, and points the way to a future in which everyone can be as comfortable editing video as they are today editing text.

ACKNOWLEDGMENTS

REFERENCES

1. Ahanger, G. and Little, T. D. C., "Automatic Composition Techniques for Video Production", IEEE Transactions on Knowledge and Data Engineering, 10 (6), 1998.

2. Boreczky, J., et al. "An Interactive Comic Book Presentation for Exploring Video," in CHI 2000 Conference Proceedings. 2000. ACM Press. pp. 185-192.

3. Christel, M., Winkler, D., and Taylor, C. "Multimedia Abstractions for a Digital Video Library," in Proceedings of the 2nd ACM International Conference on Digital Libraries. 1997. Philadelphia, PA: pp. 21-29.

4. Christel, M., et al., "Techniques for the Creation and Exploration of Digital Video Libraries," in Multimedia Tools and Applications, B. Furht, Editor 1996, Kluwer Academic Publishers. Boston, MA.

5. Chua, T. and Ruan, L., "A video retrieval and sequencing system." ACM Transactions on Information Systems, 1995. 13(4): pp. 373-407.

6. Davenport, G., Smith, T.A., and Pincever, N., "Cinematic Primitives for Multimedia." IEEE Computer Graphics & Applications, 1991. 11(4): pp. 67-74.

7. Gauch, S., Li, W., and Gauch, J., "The VISION Digital Video Library." Information Processing & Management, 1997. 33(4): pp. 413-426.

8. Girgensohn, A., et al. "A semi-automatic approach to home video editing," in Proceedings of UIST'2000: The 13th annual ACM symposium on User interface software and technology. 2000. San Diego, CA: ACM. pp. 81-89.

9. Girgensohn, A., Bly, S., Shipman, F., Boreczky, J. and Wilcox, L., "Home Video Editing Made Easy — Balancing Automation and User Control", in Human-Computer Interaction INTERACT '01. 2001. IOS Press, pp. 464-471.

10. Hauptmann, A. and Smith, M. "Text, Speech, and Vision for Video Segmentation: The Informedia Project," in AAAI Symposium on Computational Models for Integrating Language and Vision. 1995.

11. Hauptmann, A.G. and Smith, M. "Video Segmentation in the Informedia Project," in IJCAI-95: Workshop on Intelligent Multimedia Information Retrieval. Montreal, 1995. Montreal, Quebec, Canada:

12. Mackay, W.E. and Davenport, G., "Virtual Video Editing in Interactive Multimedia Applications." CACM, 1989. 32(7): pp. 832-843.

13. Mills, M., Cohen, J., and Wong, Y. "A Magnifier Tool for Video Data," in SIGCHI '92 Conference Proceedings of Human Factors in Computing Systems. 1992. Monterey, CA: ACM. pp. 93-98.

14. Placeway, P., et al. "The 1996 Hub-4 Sphinx-3 System," in DARPA Spoken Systems Technology Workshop. 1997.

15. Schwartz, J., Rhinelander, T. and Dorsey, M., "Personal Rich Media Takes Off", *The Forrester Report*, Forrester Research Inc., October 2000.

16. Stevens, S.M., Christel, M.G., and Wactlar, H.D., "Informedia: Improving Access to Digital Video." Interactions: New Visions of Human-Computer Interaction, 1994. 1(4): pp. 67-71.

17. Ueda, H. and Miyatake, T. "Automatic Scene Separation and Tree Structure GUI for Video Editing," in Proceedings of ACM Multimedia '96. 1996. Boston:

18. Ueda, H., Tmiyatake, T., and Yoshizawa, S. "Impact: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System," in Proceedings of INTERCHI'93: Conference on human factors in computing systems. 1993. Amsterdam: ACM. pp. 137-141.

19. Wactlar, H.D., et al., "Lessons learned from building a terabyte digital video library.", IEEE Computer, 1999. 32(2): pp. 66 -73

20. Zhang, H., et al. "Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution," in ACM Multimedia 95: Proceedings of the third ACM international conference on Multimedia. 1995. San Francisco, CA: pp. 15-24.