CREATING NAVIGABLE MULTI-LEVEL VIDEO SUMMARIES

Frank Shipman, Andreas Girgensohn, Lynn Wilcox

FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue, Palo Alto, CA 94304, USA
{shipman, andreasg, wilcox}@fxpal.com

ABSTRACT

We created an alternative approach to existing video summaries that gives viewers control over the summaries by selecting hyperlinks to other video with additional information. We structure such summaries as "detail-on-demand" video, a subset of general hypervideo in which at most one link to another video sequence is available at any given time. Our editor for such video, Hyper-Hitchcock, provides a workspace in which an author can select and arrange video clips, generate composites from clips and from other composites, and place links between composites. To simplify dealing with a large number of clips, Hyper-Hitchcock generates iconic representations for composites that can be used to manipulate the composite as a whole. In addition to providing an authoring environment, Hyper-Hitchcock can automatically generate multi-level hypervideo summaries for immediate use or as the starting point for author modification.

1. INTRODUCTION: INTERACTIVE VIDEO

Because watching video is very time-consuming, there have been many approaches for summarizing video. Several systems generate shorter versions of videos to support skimming [2, 3, 9]. Interfaces supporting access based on keyframe selection enable viewing particular chunks of video [10]. Video digital libraries use queries based on computed and authored metadata of the video to support the location of video segments with particular properties. We are exploring an alternative method that employs interactive video to allow viewers to watch a short summary of the video and to select additional detail on demand.

While many forms of interactive video have been envisioned, one variation — the inclusion of optional side trips — has gained acceptance in home playback. Some DVDs include options for viewers to follow links out of the currently playing video to watch other video clips. When a link is active, an icon appears on top of the playing video. The viewer can then press a button on the remote control to jump to the alternative video. For example, in *The Matrix*, links take the viewer to video segments explaining how the scene in the movie was filmed. Afterwards, the original video continues from where the viewer left. The growing acceptance of MPEG-4 implies that more varied forms of interactive video will be available in addition to existing DVD-based forms.

An obstacle to the broader adoption of interactive video is the overhead of authoring. Existing authoring tools primarily rely of scripting languages for defining a video's interactive behavior [6, 7]. This requires content authors to have some programming skills, reducing their number.

Expanding on the notion of optional side trips in video, we have designed an authoring and playing interface for "detail-on-demand" video. Our system is suitable for constructing videos where a viewer presses a button to get more information about the current topic or video sequence. Detail-on-demand video keeps the authoring and viewing interfaces as simple as possible while supporting a wide range of interactive video applications. This notion of an interlinked video stops short of being a full hypervideo, where multiple outgoing links may be available at once [5, 8]. At its simplest, the author selects a segment of the output video for which a link will be active and the video sequence that will be shown if the viewer takes that link.

The next section describes the detail-on-demand video model. This is followed by a description of Hyper-Hitchcock, with emphasis on its support for authoring hierarchically organized streams of video and links among the elements of these streams. We then describe the automated generation of detail-on-demand video for summarization and access purposes.

2. HIERARCHICAL VIDEO WITH LINKS

Detail-on-demand video has been designed to support the authoring and use of interactive video applications. Characteristics of video representations meeting this design goal are (1) a hierarchical structure where video clips are grouped into composites, and (2) links between elements in this hierarchy (see Figure 1).

Video editing involves the selection and sequencing of video clips into a linear presentation. Authoring detail-on-demand video is a process of authoring and interlinking one or more linear video presentations. In those presentations, individual video *clips* may be grouped into video *composites* as higher level building blocks. Thus, in designing the authoring tool we have emphasized how to combine existing video clips or composites into new composites, how to represent those composites to the author, and how to manipulate those composites once they are created.

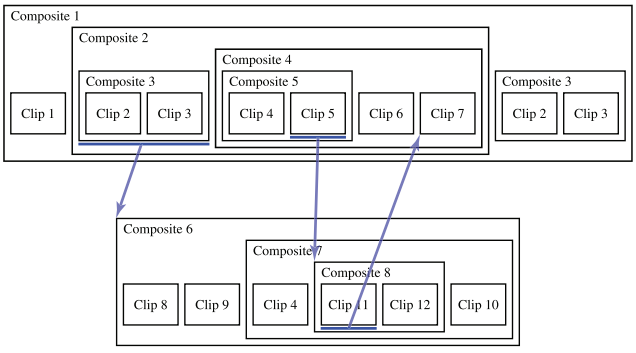

Figure 1: Links between elements in two video sequences.

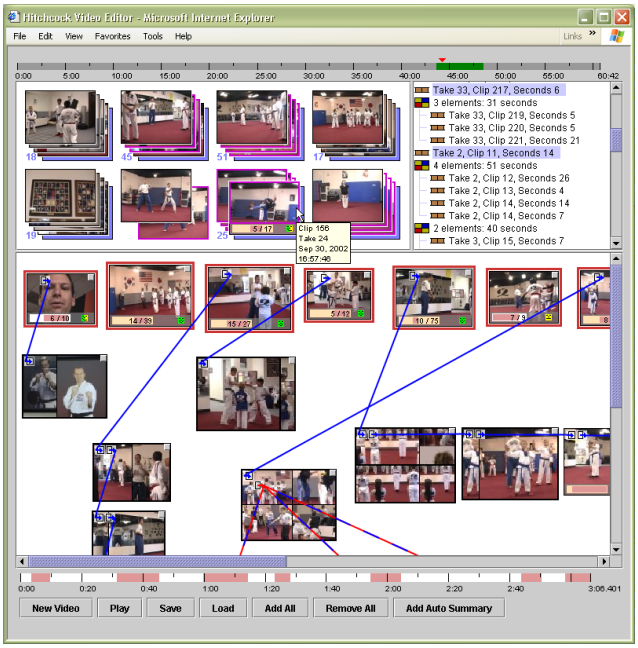

Figure 2: Martial arts training video authored in Hyper-Hitchcock.

Any video clip or composite can be the source or destination anchor for a navigational link. Source anchors specify the starting point and length for when a link is active and destination anchors specify the starting point and length of the video played as a result of the viewer traversing the link. Unlike hyperlinks in Web pages or in most hypervideo systems, the link destination is not just a starting point but an interval of content. Playback continues at the link source upon completion of the link or when the viewer aborts the playing of the link destination.

Figure 1 shows a diagram of two hierarchically organized videos and three links between them: the first being from "Composite 3" to "Composite 6", the second being from "Clip 5" to "Composite 8", and the third being from "Clip 11" to "Clip 7". If more than one link would be active at a particular time, which can happen if links are specified for multiple levels of the hierarchy, the lowest-level link has precedence. Links can be authored to form cycles, counterpoints, tangles, mirror worlds, and all the other common patterns of hypertext [1].

3. HYPER-HITCHCOCK

Hyper-Hitchcock is a prototype editing and viewing environment for detail-on-demand video. It builds on prior work on Hitchcock [4], a system designed for the editing of home video. To simplify video editing, we developed techniques for automatically segmenting takes, defined by camera on/off points, into useful clips. These clips can then be trimmed and ordered into a video sequence. Authors of detail-on-demand video can place links between elements in the video sequences.

3.1. Hyper-Hitchcock Authoring Tool

The strength of our existing technical infrastructure is the automatic segmentation of video takes (the video recorded contiguously) into smaller, more useful clips for use in the output video. The segmentation is performed by recognizing fast camera motion and lighting changes, which are used as clip boundaries. In addition, this information is used to select suitable default in/out points for the clip. The output video is constructed by selecting, modifying, and reordering these clips. Where Hitchcock includes an editing area consisting of a linear sequence of clips, Hyper-Hitchcock provides authors with a two dimensional workspace in which to collect, organize, and interlink clips. Figure 2 shows the Hyper-Hitchcock editing interface. The top-left pane groups the source video clips by criteria such as recording time or color similarity. Authors drag clips from this pane to the workspace below where they can be ordered and grouped into composites. Links can be placed between any two clips or composites in the workspace. As in Hitchcock, authors change the length of video clips and composites by resizing the video frames.

Figure 3: Five clips being grouped into composites.

3.2. Video Composites

Authoring involves generating sequences of video clips. Given the limited space of the screen, it is convenient to create an iconic representations of a video composite after authoring. Hyper-Hitchcock represents the composite as a single image consisting of a collage of images from individual video clips. As an alternative representation, Hyper-Hitchcock provides a tree view of the hierarchy of composites and clips (see top-right pane in Figure 2)

As an example, consider the video sequence of five clips shown at the top of Figure 3. The size of each image indicates the length of the corresponding video clip. The important characteristics for this sequence must be visible in the iconic representation. These include the starting and ending clips since those are important to the author who must generate transitions between this sequence and other video being edited. Other useful information includes the other clips in the sequence, the time length of the sequence, the number of component clips, and the depth (when the composite contains other composites).

We visualize the composite as a four-way representation shown at the bottom of Figure 3. Given the need to indicate the starting and ending clips in the sequence, the top-left keyframe will always be the first clip and the bottom right keyframe will be from the last clip in the sequence. We select the remaining two clips by determining the longest clips (thus larger size). Taking the areas of the keyframes of the four individual clips as a starting point, the composite image is constructed by moving a vertical and a horizontal divider such that the proportions of the image areas change as little as possible. Rather than scaling the individual images down, they are scaled to the size of the composite image and the center area of each individual image is extracted. That avoids scaling images down to a point where the content is unrecognizable. The top of Figure 3 shows a sequence of video clips in the workspace just prior to being placed in a video com-

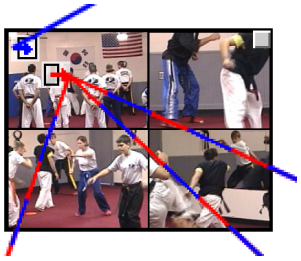

Figure 4: Links in and out of composite.

posite. The bottom of Figure 3 shows the representation of the resulting composite in the workspace in two different variants, either with a single vertical divider or with independent dividers.

An alternative compact video visualization is described by Yeung and Yeo [10]. It uses patterns of keyframes on a 4x4 grid to summarize a video sequence. This approach could be used to produce visualizations meeting the size constraints of our video editing system. However, their approach is not well suited to make best use of a small presentation space because images can only occupy multiples of grid cells. No extraction of part of a keyframe takes place to show more detail or to use layouts that change the aspect ratio of some of the used images.

Resizing of a composite causes the length of the component clips to be modified to meet the desired length of the composite. The individual clips are resized such that they contain the portion with the least camera motion and with sufficient brightness unless the author locked in/out points. Source clips retain their relative lengths while being resized unless they are locked.

3.3. Links in Hyper-Hitchcock

Any video clip or composite can be a link anchor or link destination. In the Hyper-Hitchcock workspace, links are represented as colored arrows into or out of video frames and iconic representations. Color and line placement provide information about whether the link is into or out of an element in the workspace and the color of the link indicates if the link is connected to the whole element or to a component of the element in the workspace. Figure 4 shows a close up of the arrows in and out of one of the composites in Figure 2. In this case there is one incoming link to the whole composite, represented by the blue arrow at the top left, and three red/blue links from elements of this composite to other elements in the workspace.

Links have different properties that control where the video continues after the playback of a link target ends or after the viewer presses a button to prematurely return to the link source. Returning to the point where the side trip started is the default but other options include the start or the end of the clip where the link originated. If the viewer followed a sequence of links, it is also possible to return to the main sequence rather than just returning to the immediate link source.

3.4. Detail-on-demand Video Player

Browsing video in Hyper-Hitchcock combines interaction characteristics from browsing the Web and changing channels on TV. As the viewer watches a video, the player indicates when links are available and presents labels for them. The viewer can follow the link to see the destination video or let the original video keep playing. The destination video will play until completion, at which time the original video will continue. If the destination

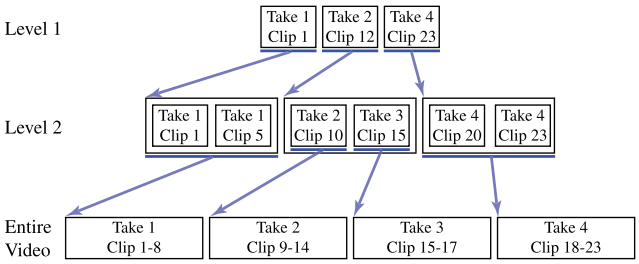

Figure 5: Links in three-level video summary.

video is not of interest, the viewer can press a "back" button to return to the source video similar to in a Web browser. With such a simple viewing interface, interactions could be performed with a DVD remote control.

4. AUTOMATIC AUTHORING / SUMMARIZATION

Detail-on-demand videos can provide an interactive summary for access into longer linear videos. Human authoring of such summaries is very time consuming and not cost effective if the summary will only be used a few times. Hyper-Hitchcock can automatically generate a multi-level summary by selecting short clips from the original video. Each level of the summary is of different length with the top level being a rapid overview of the content and the lowest level containing the entire source video. Links take the viewer from a clip at one level to the corresponding locations in the next lower level. Viewers navigate from clips of interest to more content from the same period. Figure 5 shows an example of a three-level summary created from 23 high-value clips identified in a four-take source video.

The generation of the multi-level video summary includes three basic decisions: how many levels to generate and of what lengths, which clips from source video to show in each summary, and which links to generate between the levels. The following describes our initial algorithm for automatically generating multi-level linked summaries.

The number of levels in the interactive summary is dependent on the length of the source video. For videos under five minutes in length, only one 30 second video summary is generated. For videos between 5 minutes and 20 minutes, two summaries are generated — the first level being 30 seconds in length and the second being 3 minutes in length. For videos over 20 minutes, three summaries are generated — one 30 seconds long, one three minutes long, and the last being one quarter the length of the total video to a maximum of 10 minutes.

Selection of clips to fill each video summary is based on the identification of an array of m high-quality video clips via an analysis of camera motion and lighting. Hyper-Hitchcock assumes an average length of each clip (currently 3.5 seconds) so the number of clips (n) needed for a summary is the length of the summary in seconds divided by 3.5. The first and last clip are guaranteed to be in each summary with the remainder of the clips being evenly distributed in the array of potential clips. Thus, Hyper-Hitchcock selects one clip every $(m - 1) / (n - 1)$ potential clips. The use of an estimate of average clip length generates summaries of approximately the desired length rather than exactly the requested length. The algorithm can easily be altered to support applications requiring summaries of exact lengths by modifying in/out points in the selected clips rather than accepting the in/out points determined by the video analysis.

Generating links between the levels of summary and the source video is based on the takes (on/off points from the record-

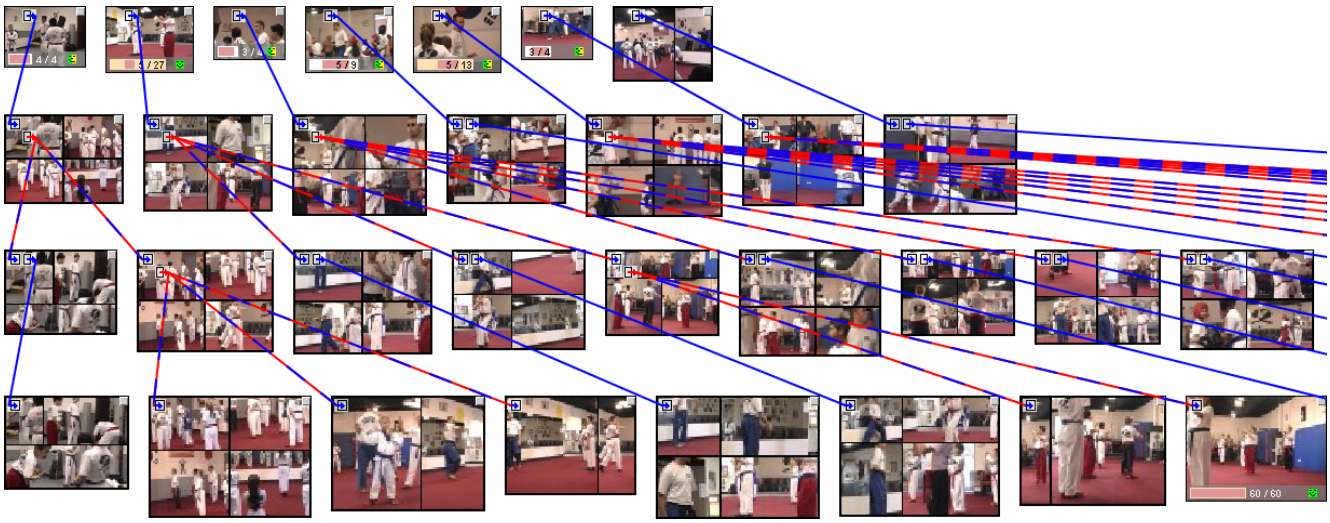

Figure 6: Automatically-generated interactive video interlinking three video summaries and complete source video.

ing). A clip in a higher-level summary (shorter) will be linked to the sequence of clips from the same take in the next level. If a take is not represented in a higher level, it will be included in the destination anchor for the link from the previous take.

This algorithm generates multi-level summaries with navigational links between the levels of summary to support video browsing. Figure 6 shows part of the resulting four-level summary of a one-hour, 33-take martial arts training video. In cases where the interactive summary will be used many times, such as in the case of an index into a training video, authors can refine the automatically-generated interactive summary in Hyper-Hitchcock.

5. CONCLUSIONS

Detail-on-demand video's hierarchy of video clips and navigational links allow viewers partial control over what video to skim and what to watch in detail. Detail-on-demand video is a subset of general hypervideo that supports a variety of tasks while maintaining a relatively simple authoring interface.

Hyper-Hitchcock supports the authoring of detail-on-demand video through the automatic decomposition of video takes into clips and through support for creation and manipulation of video composites. The iconic representation of composites enables authors to cope with the large number of clips necessary for longer works. Links between elements in the hierarchic representation of video provide viewers with optional video content. We are currently exploring how alternative link following and return behaviors will allow alternative rhetorical practices by authors.

Hyper-Hitchcock automatically generates multi-level summaries of source videos upon request. These summaries vary from two to four levels based on the length of the source video. Links are created from the clips in shorter summaries to the clips for the same take in longer summaries or to the whole take in the source video. Generated summaries can be edited in the workspace. A direction for future work is exploring alternative algorithms for generating interactive summaries.

Detail-on-demand video is a promising form of interactive video that simplifies authoring and viewing. We plan to investigate alternate forms of interactive video and to explore their uses for video summaries and other techniques that could aid access into existing video collections. Our goal is to create simple and intuitive user interfaces to good video summaries that can either be produced automatically or be easily refined by authors.

6. REFERENCES

[1] M. Bernstein. Patterns of Hypertext, *Proceedings of ACM Hypertext '98*, pp. 21-29, 1998.

[2] M.G. Christel, M.A. Smith, C.R. Taylor, and D.B. Winkler. Evolving Video Skims into Useful Multimedia Abstractions. *Proceedings of CHI'98*, ACM Press, pp. 171-178, 1998.

[3] R. Lienhart. Dynamic Video Summarization of Home Video, *SPIE 3972: Storage and Retrieval for Media Databases 2000*, pp. 378-389, 2000.

[4] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox. A Semi-Automatic Approach to Home Video Editing. *Proceedings of UIST '00*, ACM Press, pp. 81-89, 2000.

[5] K. Hirata, Y. Hara, H. Takano, and S. Kawasaki. Content-oriented Integration in Hypermedia Systems, *Hypertext '96 Proceedings*, ACM, New York. pp. 11-21, 1996.

[6] Macromedia Director, http://www.macromedia.com/software/director/.

[7] N. Sawhney, D. Balcom, and I. Smith. Authoring and Navigating Video in Space and Time, *IEEE Multimedia Journal*, Vol. 4, No. 4, pp. 30-39, 1997.

[8] J.M. Smith, D. Stotts, and S.-U. Kum. An Orthogonal Taxonomy of Hyperlink Anchor Generation in Video Streams Using OvalTine, *Proceedings of ACM Hypertext 2000*, pp. 11-18, 2000.

[9] H. Sundaram, S.-F. Chang. Condensing Computable Scenes Using Visual Complexity and Film Syntax Analysis. *Proceedings of ICME 2001*, pp. 389-392, 2001.

[10] M.M. Yeung and B.-L. Yeo. Video Visualization for Compact Presentation and Fast Browsing, *IEEE Transactions on Circuits and Systems for Video Technology*. Vol. 7, no. 5, 1997.