An Interactive Comic Book Presentation for Exploring Video

John Boreczky, Andreas Girgensohn, Gene Golovchinsky, and Shingo Uchihashi
FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue, Bldg. 4
Palo Alto, CA 94304, USA
{johnb, andreasg, gene, shingo}@pal.xerox.com

ABSTRACT

This paper presents a method for generating compact pictorial summarizations of video. We developed a novel approach for selecting still images from a video suitable for summarizing the video and for providing entry points into it. Images are laid out in a compact, visually pleasing display reminiscent of a comic book or Japanese manga. Users can explore the video by interacting with the presented summary. Links from each keyframe start video playback and/or present additional detail. Captions can be added to presentation frames to include commentary or descriptions such as the minutes of a recorded meeting. We conducted a study to compare variants of our summarization technique. The study participants judged the manga summary to be significantly better than the other two conditions with respect to their suitability for summaries and navigation, and their visual appeal.

KEYWORDS: Video summarization, video browsing, keyframe extraction.

INTRODUCTION

As video is used more and more as the official record of meetings, teleconferences, and other events, the ability to locate relevant passages or even entire meetings becomes important. To this end, we want to give users visual summaries and help them locate specific video passages quickly. Such a system is useful in settings that require a quick overview of video to identify potentially useful or relevant segments. Examples include recordings of meetings and presentations, home movies, and domain-specific video such as recordings used in surgery or in insurance. These techniques are also effective when applied to commercials and to films. These seemingly different forms of video are related because they consist of multiple shots, shot at different times, perhaps by different cameras or by a hand-held camera, but the segments are often not clearly separable from the user’s perspective (and thus are not readily accessible through an index or a table of contents).

We developed a novel approach for selecting still images from a video suitable for summarizing the video and providing entry points into it. Images are laid out in a compact, visually pleasing display reminiscent of a comic book or Japanese *manga*. For further exploration of the video, we created an interactive version of the pictorial summary. Playback can be started at a desired time by clicking on a particular frame; additional detail is available via hyperlinks. Textual information can be added to the presentation in the form of captions. Meeting minutes, for example, may be integrated with the video recording of the meeting.

The video summarization component has been used as part of a larger video database system [6] for over a year. The video database contains a large collection of video-taped meetings and presentations, as well as videos from other sources, and is used regularly by the employees of our company.

In another paper [12], we described the technical details for creating a manga summary of a video. In the meantime, we have added several new user interface features. This paper focusses on the user interface and its use for finding information in the video. We describe the user interface features and discuss how they help users find the information they are looking for.

To determine the effectiveness of our image selection and layout techniques, we conducted an experiment to test the different components of our system. To test the suitability for navigation within a video, we asked the study participants to perform several tasks of finding specific information in several videos. In addition, we also presented the participants with pair-wise comparisons of the different combinations of our techniques and asked them to rate them according to their suitability for summaries and navigation, their visual appeal, and to judge their overall preferences.

In the next section, we discuss the elements of our approach for presenting a compact summary of a video. After that, we present the user interface of the video summary and the interaction style for exploring the video. Next, we describe the setup and the results of the experiment. We conclude with a discussion of directions for future work.

CREATING VIDEO SUMMARIES

Typically, a video is summarized by segmenting it into shots demarcated by camera changes. The entire video can then be represented as a collection of keyframes, one for each shot. Although this can reduce the amount of information a user must go through to find the desired segment, it may still be too much data. Furthermore, the relative similarity of keyframes, coupled with a large and regular display, may make it harder to spot the desired one. In contrast, the system presented here abstracts video by selectively discarding or de-emphasizing redundant information (such as repeated or alternating shots). The goal of this approach is to present a concise summary of a video, a summary that can be used to get an overview of the video and one that also serves as a navigation tool.

One possible way of identifying the desired segment is to fast-forward through a video, producing a temporal rather than a spatial summary. Unfortunately, this approach does not work with streaming video, whereas the techniques described here are applicable to all types of video.[1]

Given an existing segmentation, we calculate shot importance for each segment. By thresholding the importance score, less important shots are pruned, leaving a concise and visually varied summary.

Elements in our summaries vary not only in their content, but also in the size of the keyframe. In our system, keyframes are displayed in different sizes according to their importance: less important keyframes are displayed in a smaller size. The algorithm used in our system packs different-sized keyframes into a compact representation, maintaining their time order. The result is a compact and visually pleasing summary reminiscent of a comic book or Japanese *manga*.

In this section, we provide a brief description of the steps for creating video summaries. More of the technical details are described in [12].

Video Segmentation by Clustering

Many techniques exist to automatically segment video into its component shots, typically by finding large frame differences that correspond to cuts, or shot boundaries. Once detected, shots can be clustered by similarity such that related shots (e.g. similar camera angles or subjects) are considered to be one shot or cluster. For example, a film dialog where the camera repeatedly alternates between two actors would typically consist of two clusters, one for each actor. Many systems break videos into shots and use a constant number of keyframes for each shot. For example, Ferman *et al.* [4] cluster the frames within each shot. The frame closest to the center of the largest cluster is selected as the keyframe for that shot.

Other systems use more keyframes to represent shots that have more interesting visual content. Zhang *et al.* [15] segment the video into shots and select the first frame after the completed shot transition as a keyframe. The rest of the shot is examined, and frames that are sufficiently different from the last keyframe are marked as keyframes as well. Zhuang *et al.* [16] use a clustering approach to determine keyframes for a shot. They still extract at least one keyframe per shot. Cherfaoui *et al.* [1] segment the video into shots and determine if there is camera motion or zooming in each shot. Shots with camera motion are represented with three frames. Zooms and fixed shots are represented as single frames with graphical annotations that describe object motion or the zoom parameters.

These existing systems either provide only limited control over the number of keyframes or do not perform an adequate job of finding truly representative frames. Instead of segmenting the video into shots, we cluster all the frames of the video. Frames are clustered using the complete link method of the hierarchical agglomerative clustering technique [8]. This method uses the maximum of the pair-wise distances between frames to determine the intercluster similarity, and produces small, tightly bound clusters. For pictorial summaries described in this paper, we use smoothed three-dimensional color histograms in the YUV color space for comparing video frames. Once clusters have been selected, each frame is labeled with its corresponding cluster. Uninterrupted frame sequences belonging to the same cluster are considered to be segments.

Determining Keyframes and Their Sizes

To produce a good summary, we must still discard or de-emphasize many segments. To select appropriate keyframes for a compact pictorial summary, we use the importance measure described in [12]. This calculates an importance score for each segment based on its rarity and duration. Longer shots are preferred because they are likely to be important in the video. This preference also avoids the inclusion of video artifacts such as synchronization problems after camera switches. At the same time, shots that are repeated over and over again (for example wide-angle shots of the room) do not add much to the summary even if they are long. Therefore, repeated shots receive lower scores. The clustering approach discussed in the previous section can easily identify repeated shots.

Segments with an importance score higher than a threshold are selected to generate a pictorial summary. For each segment chosen, the frame nearest the center of the segment is extracted as a representative keyframe.[2] Frames are sized according to the importance measure of their originating segments, so that higher importance segments are represented with larger keyframes. This draws the attention to the more important portions of the video. The keyframe packing algorithm described in the next section creates a compact and attractive presentation that summarizes the whole video.

Keyframe Packing

Once frames have been selected, they need to be arranged in a logical order. A number of layouts have been proposed. Shahraray *et al.* [9] at AT&T Research use keyframes to represent an HTML presentation of video. One keyframe was selected for each shot; uniformly sized keyframes are laid out in a column along closed-caption text. Taniguchi *et al.* [11] summarize video using a 2-D packing of “panoramas” which are large images formed by compositing video pans. In this work, keyframes are extracted from every shot and used for a 2-D representation of the video content. Because frame sizes are not adjusted for better packing,

[1] A separate channel that delivers pre-computed keyframes may be used to construct a summary of the video.

[2] Although frames early in the shot may be preferable semantically in some cases, camera motion and noise due to camera switching may make these frames unreliable.

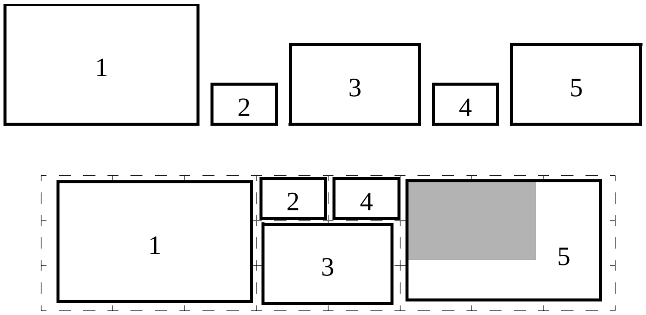


Figure 1: Packing keyframes into a row block

much white space appears in the result, making the summary inefficient.

Yeung *et al.* [14] use clustering techniques to create pictorial summaries of videos. They select one keyframe for each video shot and cluster those frames on the basis of visual similarity and temporal distance to groups of similar shots. They create pictorial summaries of video using a "dominance score" for each shot. Though they work towards a similar goal, their implementation and results are substantially different. The sizes and the positions of the still frames are determined only by the dominance scores, and are not time-ordered.

In our system, keyframes are laid out in two dimensions to form a pictorial abstract of the video similar to the panels in a comic book. Thresholding the importance score allows the desired number of frames to be displayed. To facilitate layout, frames can be displayed in smaller or larger sizes, depending on their importance score, and are resized to best fit the available space.

Given a set of frames, we have to find a sequence of frame sizes that both fills space efficiently and provides a good representation of the original video. We use a near-optimal "row block" packing procedure that we described in more detail in [12]. An example of this packing method is depicted in Figure 1. The rectangles at the top are the original frame sequence, sized by importance. The bottom picture illustrates the frames packed into a row block. Note that the size of Frame 5 has been increased from the original size (indicated as a gray rectangle) for a better packing with minimal white space.

Figure 2 shows the algorithm applied to the video of a staff meeting. A larger gap between rows graphically emphasizes that the temporal order is row by row. Within a row, frames go from left to right with occasional excursions up and down. Lines are provided between frames to unambiguously indicate their order.

The smallest image size and row width in the manga display default to useful values (64x48 pixels, 8 columns). These can be changed on-the-fly if users prefer a different-size presentation, perhaps to better fill their web browser window. The packing algorithm is very efficient so that a new layout can be produced on request.

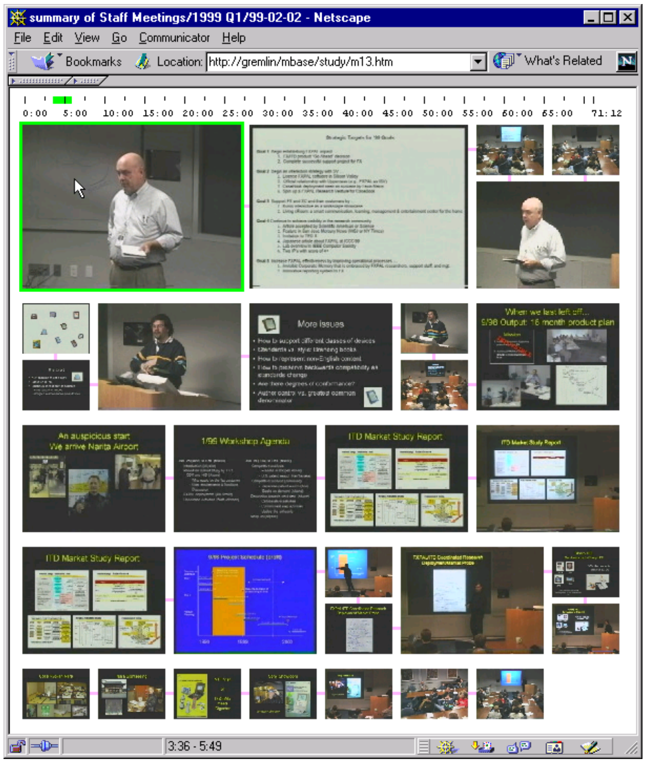

summary of Staff Meetings/1999 Q1/99-02-02 - Netscape
File Edit View Go Communicator Help
Bookmarks
Location: http://gremlin/mbase/study/m13.htm
What's Related
0:00 5:00 10:00 15:00 20:00 25:00 30:00 35:00 40:00 45:00 50:00 55:00 60:00 65:00 71:12
3:36 - 5:49

Figure 2: Pictorial summary of a video

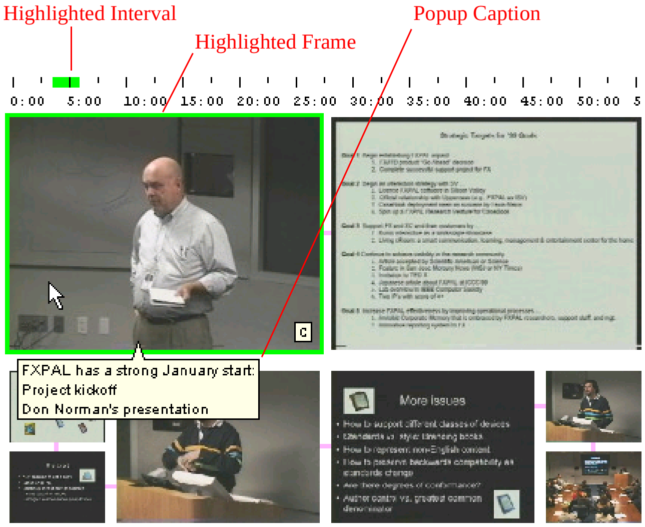
EXPLORING VIDEOS

A video summary may help a user identify the desired video simply by showing relevant images. Often, that is not enough, however. Even after finding an appropriate video, it may be difficult to locate desired passages within it simply by inspection. As video is used more often as a permanent, often official, record of events, it becomes more important to be able to locate relevant passages. As video collections grow, the time spent looking for material becomes a scarce commodity.

We believe that interactive browsing of video summaries can shorten the time required to find the desired segment. Guided by this principle, we have implemented a web-based interactive version of the pictorial summary. Users can play the video corresponding to a keyframe by clicking on it. They can request more information in the form of a "tool tips" pop-up (see Figure 3). Finally, they can request a more detailed view that shows additional shots in the selected segment and select one of those shots for playback (Figure 5). These features are described in more detail in the following sections.

Interactive Video Summary

Users may browse the video based either on the keyframes or on the video's timeline. The two views are always synchronized: the timeline shows the duration of the segment that corresponds to the frame under the cursor (see Figure 3). Similarly, when the cursor is over the timeline, the corresponding keyframe is highlighted. This display allows users to explore the temporal properties of a video. At a glance, they can see both the visual representation of an important segment and its corresponding time interval, and can inter-



FXPAL has a strong January start:
Project kickoff
Don Norman's presentation

Figure 3: Highlighted frames and embedded captions

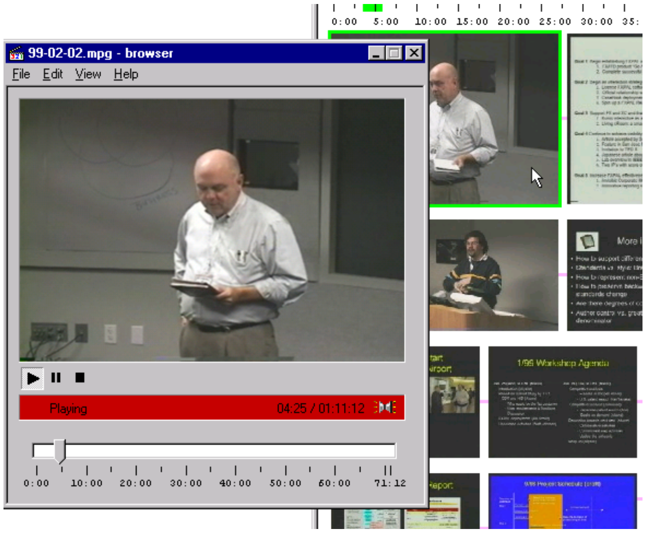


Figure 4: Playing the video

act with the system in manner that best suits their particular needs.

Once an interesting segment has been identified, clicking on its keyframe starts video playback from the beginning of that segment (see Figure 4). The ability to start the video playback at the beginning of a segment is important for exploring a video in-depth. It is not likely that the informal video material such as captured meetings would be segmented by hand, as is common with more formal material such as feature films. Our automatic clustering approach, combined with importance scores, yields good enough segment boundaries to aid the exploration of a video. This interface makes it easy to check promising passages of a video. If a passage turns out to be uninteresting after all, other segments can be easily reviewed just by clicking on their keyframes.

Several other systems use keyframes to support browsing of videos. The system described in Christel *et al.* [2] produces video highlights by selecting short segments of video. For each shot they select a keyframe that emphasizes moving objects, faces, and text. Shot keyframes are selected in rank order and a short segment of video surrounding the selected keyframe is added to the video highlight until the desired length of video is reached. Other tools have been built for browsing video content [13, 15]. These do not attempt to summarize video but rather present video content “as is.” Therefore, keyframes are typically extracted from every shot and not selected to reduce redundancy. In most approaches, the frame presentation is basically linear, although some approaches occasionally use other structures [13].

Video Captions

Many videos in our video database depict meetings. If meeting minutes exist, they can be included in the manga display as captions. Such captions are expected to increase the value of video summaries. Ding *et al.* [3], for example, report that summaries of videos consisting of keyframes and coordinated captions were preferred by participants, and led to better predictions of video content. Huang *et al.* have created summaries of news broadcasts, as reported in [7]. Story boundaries were pre-determined based on audio and visual characteristics. For each news story, a keyframe was extracted from a portion of video where keywords were detected the most. Their method nicely integrated information available for news materials, but relies heavily on the structured nature of broadcast news and would not apply to general videos.

Initially, we chose to place captions on top of the corresponding images in the summary. This caused problems because small images did not provide enough room for captions. Also, if there were a large number of captions (e.g., extracted from the close caption track of a video), many of the images were partially obscured. To address these issues, we decided to pop up captions as the mouse moves over an image (see Figure 3). Small icons (the letter “C”) indicate which images have attached captions.

Textual annotations can enhance the quality of the video summary [3]. The pictorial layout captioned with text from minutes is a better summary than either of the individual parts. If the meeting minutes are taken on a computer and time-stamped, they can be automatically time-aligned with the video. Otherwise, the minutes may be aligned by hand. Such video captions provide additional information about the scenes represented by the keyframes. For the study described in the next section, we did not provide any captions.

Exploring Additional Details

While the initial display provides a good summary of the video, it is helpful to be able to explore additional details without having to play back all likely segments. Furht *et al.* [5, p. 311] describe an approach for showing additional detail while exploring a video. Equal numbers of shots are grouped and represented by a keyframe. Expanding a keyframe recursively reveals keyframes for each shot in the subgroup until each group consists only of a single shot.

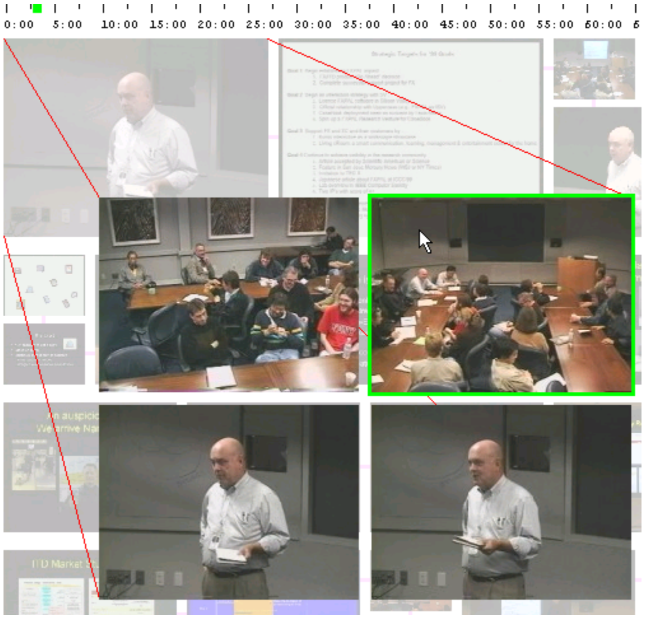


Figure 5: Drilling down

Their approach treats all shots equally and does not make any judgements about shot importance.

Our approach for exploring a video takes the structure of the video into account. We first look at the neighboring segments of the segment being explored. There might be several neighbors that have importance scores too low to be represented by keyframes. Showing keyframes for such segments can provide additional context for the segment being explored. If a direct neighbor of the segment is already represented by a keyframe, a different algorithm (described below) can be used to show additional structure within the segment.

The overlaid detailed view of a segment consists of some frames selected from shots in the selected segment. We show at least one image from the segment being explored. If, after showing the neighbors, there is space for additional images, we select subclusters of the cluster the segment belongs to by descending into the hierarchical cluster tree. As before, cluster membership is used for segmentation. We increase the number of clusters until we have as many subsegments as required. As before, we choose one keyframe from each subsegment.

Once identified, the new, detailed images are overlaid on a faded-out manga display. Lines are drawn from its representative frames to the four newly shown frames to show the connection to the segment being explored. The new frames have the same relation to the timeline as the original ones, and video playback can be started from them as well.

Figure 5 shows the results of exploring the segment shown in the top-left of Figure 3. The top two frames represent the two segments that precede the expanded segment. They provide context that may help the user understand the selected segment. The remaining two frames show additional detail about the expanded segment.

The interaction technique for exploring video segments in greater detail was used extensively in the experiment described in the next section. It is an effective means for exploring a video without having to wait for playback.

EXPERIMENT

We conducted an experiment to test the different components of our system. Participants performed several tasks of finding specific information in several videos. In the second part of the study, we asked the participants for their judgements regarding the suitability of combinations of our techniques for summaries and navigation, and regarding their visual appeal, and also asked them to report their overall preferences. We expected that our summarization techniques would provide an advantage in terms of time to complete the task, but we wanted to assess the contributions of individual components. In addition, the study was intended to provide general insights for how users use the summaries, and to identify potential usability problems.

Participants

Twenty-four participants (17 male, 7 female) participated in the study. Participants included researchers, support staff, and administrative staff at our company; they had varying degrees of expertise with using digital video. The videos used in the study were recordings of staff meetings so that the participants were exactly our target users who might have missed a staff meeting or who need to go back to find some information. For their efforts and time, the participants were rewarded with a free lunch.

Materials

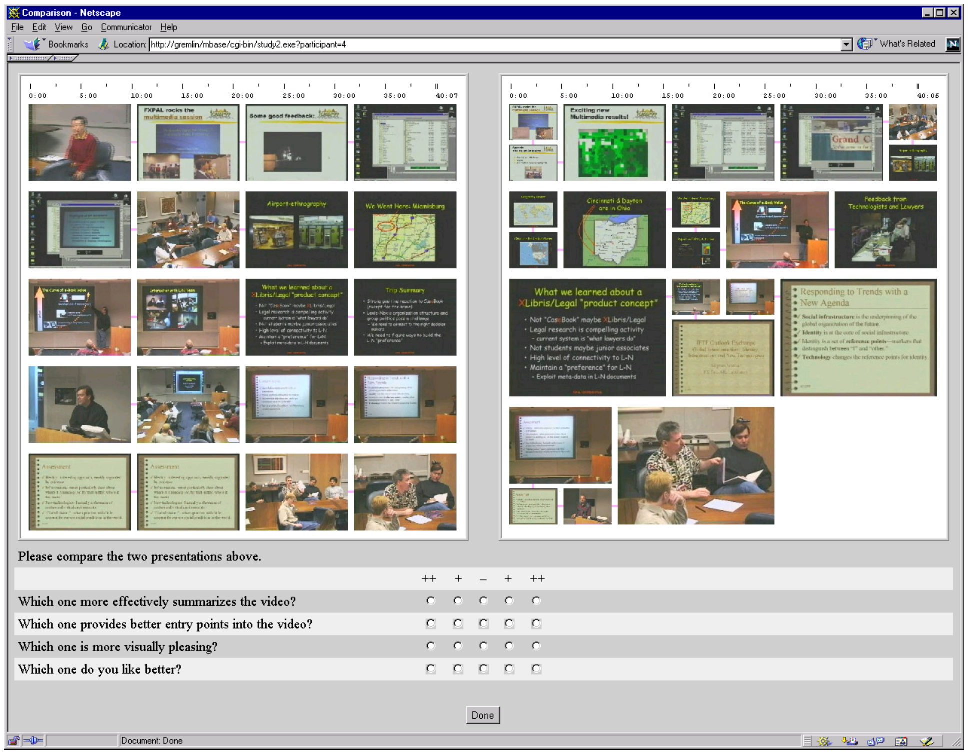
Visual summaries of three videos of staff meetings were created. Three different styles of summaries were created for each video. The different styles were designed to explore two manga features: image selection by importance score and variable-size image layout. All styles were displayed in the same screen area. The first style (control) used neither feature: the display consisted of fixed-size images sampled at regular time intervals. The second style (selected) used manga image selection, but retained fixed-size images. The third style (manga) used both features.

The image size of the video player was 352x240 pixels. Images were presented in three different sizes (64x48, 136x102, and 208x156 pixels, respectively). The first two styles only presented images in the intermediate size. They both placed 20 images in a 4x5 grid. For the third style, the largest number of images was used that still fit in the same area as used for the other two summaries. That led to summaries for the three videos using 20, 23, and 30 images, respectively.

All summaries used the same user interface with the same size popup images (208x156 pixels). For the control condition, popup images were determined by sampling in smaller regular intervals. The selected and manga conditions used the normal manga algorithm for determining popup images.

Procedure

The experiment used a within-subject design in which each participant saw all videos and all summary styles (just not the same combinations of the two). Each participant saw



Please compare the two presentations above.

	++	+	–	+	++
Which one more effectively summarizes the video?	○	○	○	○	○
Which one provides better entry points into the video?	○	○	○	○	○
Which one is more visually pleasing?	○	○	○	○	○
Which one do you like better?	○	○	○	○	○

Figure 6: Comparison of visual summaries

each video and each summary style once. We went through all permutations across the participants for the order of the summary styles, the assignment of videos to summary styles, and the order of videos so that the different conditions were balanced for multiples of 12 participants. Each participant was given a training session with a manga summary of a fourth video.

For each video, participants had to read three questions, find the relevant video segments, and then write the answers on a sheet of paper. If a question could not be answered within three minutes, the participants were shown the answer and asked to move on. All mouse actions were recorded and time-stamped. An experimenter noted the strategy used to discover the information required for each task. The completion time for each video was measured.

For the second part of the experiment, participants were asked to report preferences for the different summary styles. They were presented with pair-wise comparisons of different summaries for the same video. For all three videos, all three summaries were compared to each other for a total of nine pair-wise comparisons. Each summary was presented the same number of times on the left and the right side. The order of videos and summaries within each video was randomized across the participants. For each pair of presentations, participants answered the following questions:

1. Which one more effectively summarizes the video?
2. Which one provides better entry points into the video?
3. Which one is more visually pleasing?
4. Which one do you like better?

For each question, a preference could be indicated by selecting “++” or “+” for the summary on the left or on the right. A subject could also select “–” (no difference) if there was no preference (see Figure 6).

Results

The average task completion time of 273 seconds across all summaries is much shorter than the average length of the three videos (3184 seconds) so that the tasks were completed in a small fraction of the time for watching the whole videos. This corresponds to watching the whole video at 12-times speed without ever stopping. Of the 216 total questions, participants were unable to answer eight, of which five were in the control condition. We do not have task completion times for people just using a video player because we assumed that those times would be much longer and we did not want to impose too much on the study participants.

Analysis of variance found no statistically-significant differences in time to complete the tasks among the three conditions ($p > 0.10$). We also calculated the fraction of total time spent watching the video, and the frequency of selecting video to watch. No differences between conditions were found for any of these measures. We may attribute this lack of difference in part due to a bug in the algorithm that made the start of a desired segment of one of the videos unavailable when a detailed view for that segment was displayed. The start of the segment was required to answer the question. As a result, participants spent quite some time watching the video, although it was already past the required point.

In addition to the task completion time data described above, we collected subjects’ comparisons of the interfaces (control, selected, manga) for each video. Subjects answered four questions, and their pair-wise judgments on a five point scale were binned into three categories: prefers interface shown on the left, neutral, prefers interface on the right. (Interface condition was counterbalanced across presentations.) We then counted responses in each bin, and classified responses as neutral either if the neutral category was selected more than either interface, or if opinion was split between the two interfaces. As shown in Table 1, the selected condition was indistinguishable from the control: the χ^2 test yielded $p > 0.10$ for Questions 1 and 2, and the neutral response dominated in Questions 3 and 4. Subjects preferred the manga condition to the others for all questions ($\chi^2(2) \geq 9.1$, $p \leq 0.01$), the manga condition being preferred from 1.9 to 5.8 times more frequently than either control or selected. Based on these results, we can conclude that subjects believe the manga condition to be a more effective summary, to provide better entry points, and to be more visually pleasing. Overall, they preferred it to the other two categories by a count of 97 to 24, with 23 neutral judgments.

Comparison	Q	$\chi^2(2)$	– %	0 %	+ %
Control vs. Selected	1	3.1	23	38	39
	2	3.3	24	40	36
	3	42.6	7	**68**	25
	4	10.3	17	**47**	36
Control vs. Manga	1	50.3	19	8	**72**
	2	25.8	24	15	**61**
	3	50.3	19	8	**72**
	4	49.1	13	15	**72**
Selected vs. Manga	1	18.6	25	18	**57**
	2	9.1	26	24	**50**
	3	40.1	26	7	**67**
	4	27.8	21	17	**62**

Table 1: Results of χ^2 test[1]

Discussion

The lack of significant differences among the conditions in time to complete tasks may be due in part to the tasks we assigned. We expect that more naturalistic use (when subjects have their own search needs and are better able to understand when the desired passage has been found) will reveal differences in the interface. Furthermore, addressing some of the usability issues discovered during the experiment should improve performance of the manga interface.

In addition to collecting performance and opinion data in the experiment, we observed subjects in their use of the interface, looking for usability defects. Some subjects also volunteered their impressions about the interfaces. Finally, a preliminary analysis of the log also indicated some usability issues. The problems fell into two categories: input and output.

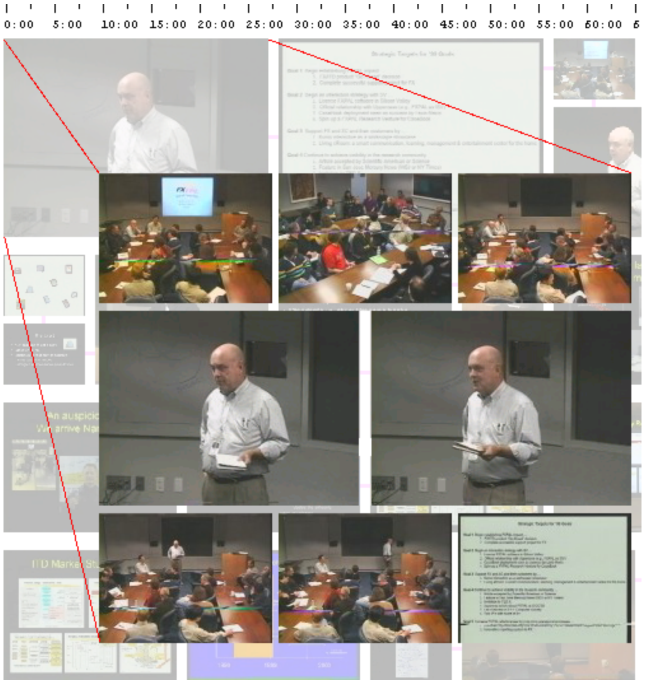


Figure 7: Mockup showing details in a focus+context view

On the input side, many participants were confused by the right/left mouse button (drill down vs. play video). Although some people only made the mistake initially, two participants had the problem throughout their sessions. A related suggestion was to expand the view when the mouse hovered over an image, rather than requiring a click. There was also an inconsistency in the way the timeline responded to mouse events: both left and right clicks caused playback, which was not consistent with the keyframe interface.

Output problems were related to the design of the detailed view. The detailed (overlaid) view did not contain the selected image, but did contain other, seemingly unrelated, keyframes (see Figure 5). The main problem with our design was that it was difficult to distinguish context frames from the actual expansion. One way to solve the problem is to make the context keyframes smaller, as shown in Figure 7. Finally, a number of comments centered around image size: some thought that the small (64x48 pixel) images were too small, and others wanted slide images to be large enough to be legible.

CONCLUSIONS

In this paper we described an approach for summarizing videos and for navigating the summaries. We presented a method for generating compact pictorial summarizations of video based on scores that determined the most important shots in a video and assigned sizes to extracted keyframes. Different-sized keyframes were packed into a compact, comic-book-like visual summary. Links from each keyframe allowed video playback and the exploration of additional detail.

[1] Although we show percentages, raw frequencies were used to calculate the χ^2 values.

We tested our approach in an experiment in which we had participants perform tasks using different summary styles with the manga interface. We also asked the participants to compare the different summary styles to each other with respect to different criteria. The study participants judged the manga summary to be significantly better than the other two conditions with respect to their suitability for summaries and navigation, and their visual appeal.

A system like the one described here can be useful in situations where large collections of video must be accessed. Examples include an application that presents a library of videos such as our staff meetings, or a search engine application that returns a large list of videos that must be examined. The interactive summary described here is part of a larger system for browsing video collections [6] used for the past two years in our laboratory.

Future work will improve summaries by including automatically generated information, such as face recognition data, and automatic captioning from different sources. We also plan to use other sources such as speaker identification and notification of slide changes to improve the segmentation of the videos.

ACKNOWLEDGMENTS

Thanks to John Doherty for producing the meeting videos in our corpus.

REFERENCES

1. Cherfaoui, M. and Bertin, C. (1994). Two-Stage Strategy for Indexing and Presenting Video. In *Proc. Storage and Retrieval for Still Image and Video Databases II, SPIE 2185,* San Jose, CA, pp. 174-184.

2. Christel, M.G., Smith, M.A., Taylor, C.R., and Winkler, D.B. (1998). Evolving Video Skims into Useful Multimedia Abstractions. In *Proc. ACM CHI 98,* Los Angeles, CA, pp. 171-178.

3. Ding, W., Marcionini, G., and Soergel, D. (1999). Multimodal Surrogates for Video Browsing. In *Proceedings of Digital Libraries 99*, ACM, pp. 85-93.

4. Ferman, A.M. and Tekalp, A.M. (1997). Multiscale Content Extraction and Representation for Video Indexing. In *Proc. Multimedia Storage and Archiving Systems II, SPIE 3229, (Dallas, TX)*, pp. 23-31.

5. Furht, B., Smoliar, S.W., and Zhang, H.J. (1995). *Video and Image Processing in Multimedia Systems*. Boston: Kluwer Academic Publishers.

6. Girgensohn, A., Boreczky, J., Wilcox, L., and Foote, J. (1999). Facilitating Video Access by Visualizing Automatic Analysis. In *Proc. INTERACT'99,* pp. 205-213.

7. Huang, Q., Liu, Z. and Rosenberg, A. (1999). Automated Semantic Structure Reconstruction and Representation Generation for Broadcast News. In *Proc. IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII*, Vol. 3656, pp. 50-62.

8. Rasmussen, E. (1992). Clustering Algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, pp. 419-442.

9. Shahraray, B. and Gibbon, D.C. (1995). Automated Authoring of Hypermedia Documents of Video Programs. In *Proc. ACM Multimedia 95*, pp. 401-409.

10. Snedecor, G.W. and Cochran, W.G. (1989). Statistical Methods. Ames: Iowa State University Press.

11. Taniguchi, Y., Akutsu, A., and Tonomura, Y. (1997). PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing. In *Proc. ACM Multimedia 97*, pp. 427-436.

12. Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. (1999). Video Manga: Generating Semantically Meaningful Video Summaries. In *Proc. ACM Multimedia 99*, pp. 383-392.

13. Yeo, B.-L. and Yeung, M. (1998). Classification, Simplification and Dynamic Visualization of Scene Transition Graphs for Video Browsing. In *Proc. IS&T/SPIE Electronic Imaging '98: Storage and Retrieval for Image and Video Databases VI*, pp. 60-70.

14. Yeung, M. and Yeo, B.-L. (1997). Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content. *IEEE Trans. Circuits and Sys. for Video Technology*, 7(5), pp. 771-785.

15. Zhang, H.J., Low, C.Y., Smoliar, S.W., and Wu, J.H. (1995). Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution. In *Proc. ACM Multimedia 95*, pp. 15-24.

16. Zhuang, Y., Rui, Y., Huang, T.S., and Mehrotra, S. (1998). Adaptive Key Frame Extraction Using Unsupervised Clustering. In *Proc. ICIP '98, Vol. I,* pp. 866-870.