4-1-1998

Evolving Video Skims into Useful Multimedia Abstractions

Michael G. Christel
Carnegie Mellon University, christel@cs.cmu.edu

Michael A. Smith
Carnegie Mellon University

C. Roy Taylor
Carnegie Mellon University

David B. Winkler
Carnegie Mellon University

Evolving Video Skims into Useful Multimedia Abstractions

Michael G. Christel, Michael A. Smith, C. Roy Taylor, David B. Winkler

Carnegie Mellon University
Pittsburgh, PA 15213 USA
+1 412 268 7799
{christel, msmith, crt, dwinkler} @cs.cmu.edu

ABSTRACT

This paper reports two studies that measured the effects of different "video skim" techniques on comprehension, navigation, and user satisfaction. Video skims are compact, content-rich abstractions of longer videos, condensations that preserve frame rate while greatly reducing viewing time. Their characteristics depend on the image- and audio-processing techniques used to create them. Results from the initial study helped refine video skims, which were then reassessed in the second experiment. Significant benefits were found for skims built from audio sequences meeting certain criteria.

Keywords

Video abstraction, evaluation, digital video library, video browsing, video skim, empirical studies, multimedia

INTRODUCTION

With increasing computational power and storage capacity, the potential for large digital video libraries is growing rapidly. The World Wide Web has seen an increased use of digital video, and digital video remains a key component of many educational and entertainment applications. As the size of accessible video collections grows to thousands of hours, potential viewers will need abstractions and technology to help them browse effectively and efficiently through this new wealth of information.

A *multimedia abstraction* ideally preserves and communicates in a compact representation the essential content of a source video. Examples include brief titles and individual "thumbnail" images that, when selected appropriately, facilitate operations on the corresponding source [3]. Another common approach presents an ordered set of representative thumbnail images simultaneously on a computer screen [3, 5, 6, 9, 11, 15, 17, 18]. While these abstractions have proven useful in various contexts, their static nature ignores video's temporal dimension. In addition, they often concentrate exclusively on the image content and neglect the audio information carried in a video segment. Our preliminary investigations suggest that the opposite emphasis offers greater value.

We define "video skim" as a temporal, multimedia abstraction that incorporates both video and audio information from a longer source. A video skim is played rather than viewed statically, and a two-minute skim may represent a 20-minute original. Our goal for video skims goes beyond merely motivating a viewer to watch a full video segment; we seek to communicate the essential content of a video in an order of magnitude less time.

During the past few years, the Informedia Project has developed and integrated speech recognition, image processing, and natural language techniques for processing video automatically [7, 16]. We are applying these techniques to extract the most important content from a video, that is, its significant images and words, and using that content as components for its video skim [13]. This paper reports on two experiments that examined ways to improve upon simple, mechanistic skimming techniques.

GENERATING SKIMS

One straightforward method for creating skims would simply increase the frame rate across the whole video. This "fast forward" approach might achieve a tenfold decrease in viewing time, but would seriously degrade coherence [14], both perturbing the audio [4] and distorting image information.

The skims described here, however, all preserve the source video's frame rate and differ only in the rules used for selecting "important" audio and video components. Our skim-generating procedures automatically select and concatenate original video and audio data into new, shorter presentations, as Figure 1 shows.

The most basic of these methods "subsamples" a source video, skipping frames at fixed intervals and keeping, for example, the first 10 seconds of each 100. The selected pieces are then concatenated and played back at the original frame rate. Figure 2 illustrates how source components map to a skim. While dropping video at regular intervals will likely delete essential information

[16], this technique is trivial to implement and so serves as the default skim (DEF) in the studies reported here.

More ambitious methods analyze image, audio, and language information to differing degrees. An image-centric skim (IMG), for example, emphasizes visual content, decomposing the source into component *shots* [6, 11, 15, 17, 18], detecting "important" objects, such as faces and text [12, 13], and identifying structural motion within a shot [13]. Image heuristics, including weighting heavily those frame sequences with significant camera motion and those showing people or a text caption, prioritize shots for inclusion in the skim [13]. Metarules avoid overlapping regions and provide uniform coverage of the source video. Shots are repeatedly added to a skim until it reaches a threshold size, such as one-tenth of the full video.

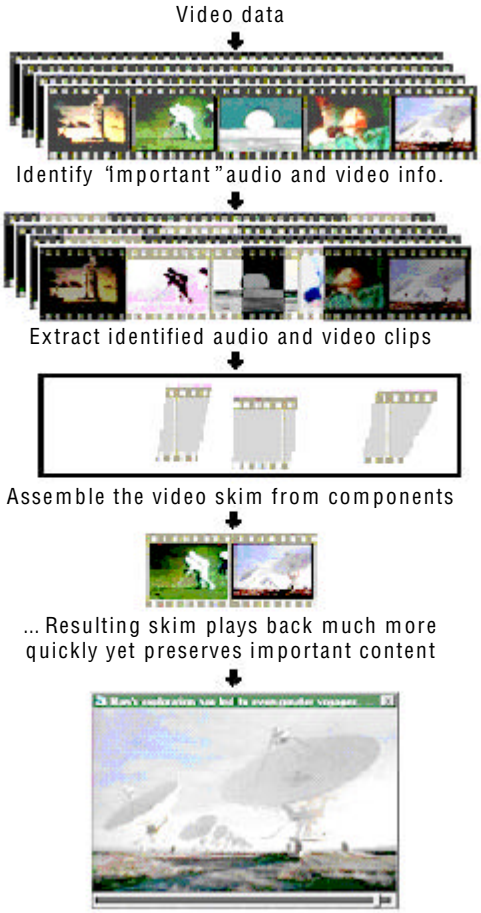

Figure 1. Generalized video skim creation process

Component shots for an image-centric skim may be quite abbreviated, however, and audio synchronized with brief shots will be short as well. Replaying audio components thus selected tends to yield a choppy presentation of partial words and noises. Pilot testing revealed such audio to be disconcerting and frustrating. The IMG skim design tested here maintains its singular emphasis on image analysis by incorporating an improved audio track, namely the same, subsampled audio as in the default (DEF) skim.

The structure of an audio-centric skim (AUD) derives solely from audio information. Automatic speech recognition and alignment techniques [7] register the audio track to the video's text transcript. A link-grammar parser developed at Carnegie Mellon identifies noun phrases within the transcript, and term-frequency, inverse-document-frequency (TF-IDF) scoring ranks them [13]. Words that appear often in a particular document but relatively infrequently in a standard corpus receive the highest weight. Noun phrases with many such words are judged "key phrases" and are assumed to contain the source video's most important audio information. Key phrases and their associated image frames are repeatedly added until the skim reaches a threshold size.

An "integrated best" skim (BOTH) merges the image-centric and audio-centric approaches while maintaining moderate audio/video synchrony. Top-rated audio regions are selected as in the AUD skim; hence, for a given source video, audio portions of AUD and BOTH skims are identical. The audio is then augmented with imagery selected — using IMG heuristics — from a temporal window extending five seconds before and after the audio region. This bounded window forces selection of visual components different from those in an IMG skim and aligns them more closely with the audio than in the IMG design. While the audio and video components of a BOTH skim may not be precisely synchronized, each attempts to capture the most significant information of its type.

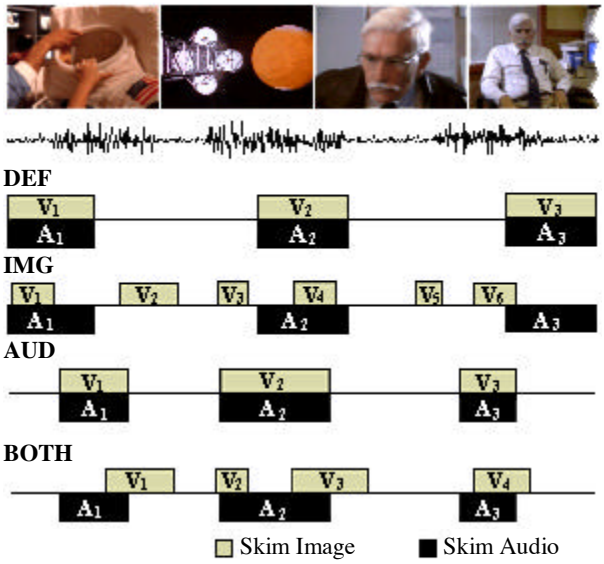

Figure 2. Audio/video alignment in skims (Exp. 1)

EXPERIMENT ONE

Our first experiment examined the use of these four different video skims on two tasks:

- Factfinding, where subjects used skims to locate video segments that answered specific, content-related questions.
- Gisting, grasping the essence of a video through its skim, where subjects matched skims of a longer video with representative text phrases and single-frame images for that longer video.

These tasks represent two complementary facets of video information retrieval. Factfinding tends to emphasize analytic processing, focusing attention to find a specific image or phrase within a larger context. Gisting, on the other hand, emphasizes synthetic processing to distill content, as in scanning a set of search results to narrow the scope of further, more detailed examination.

Subjects

Forty-eight Carnegie Mellon students (31 male, 17 female) from several majors and programs volunteered for the study, responding to an electronic call for participation in the spring of 1997. Each received $5 and spent about an hour with the system. A background questionnaire found that the subjects were, in general, "very comfortable" with computers but had little prior experience with digital video.

Materials

The video material was drawn from three public television series: "The Infinite Voyage", "Planet Earth", and "Space Age." This material was carefully documented to a fine level of granularity for delivery with the Informedia system to a high school in the Pittsburgh area [2]. The documentation was done manually and then checked and corrected for accuracy, without any knowledge that it would be later used in skim experiments. For these documentaries, every 1.6 minutes of video, on average, are represented by a short text phrase and a thumbnail image. These manually chosen representative images and manually generated text phrases serve as the gist of a video for our experiment. Ideally, after watching a video skim that fully communicates the gist of a longer video, the subject would be able to select all text phrases and representative images that belong to that longer video.

Design

Each of the skims was one-tenth the length of its source video and built from segments averaging three seconds duration (Figure 2). This 3-second "grain size" equals the average duration of key phrases used in the AUD and BOTH skims. The study compared the following four types of skims:

- DEF, the default, subsampled skim, comprising seconds 0-3 of the source video, then seconds 30-33, seconds 60-63, 90-93, etc.
- IMG, "best video" skim
- AUD, "best audio" skim
- BOTH, "integrated best" skim

Procedure

Subjects participated in the study individually. Each used a computer with a 17-inch color monitor, keyboard, mouse, and headphones. Each subject completed the factfinding task four times, once for each skim type, and the gisting task eight times, viewing each skim type twice. Subjects thus viewed skims of 12 different videos. We used a repeated measures design in a 4×4 Latin Square configuration to balance any learning effect between treatments [8].

In the factfinding task subjects were given a question and asked to navigate to that region of a video presenting the answer. While seeking the answer region, they could toggle between the skim and the full video. A potential $25 bonus encouraged them to work quickly.

After each factfinding exercise with a skim type, we asked subjects to evaluate the interface using a subset of the QUIS instrument [10], including such nine-point Likert scales as "terrible-wonderful" and "dull-stimulating." We also invited the subjects to type open-ended comments.

In the gisting task subjects watched a video skim without the option of switching to the normal video presentation. After watching each skim, they chose from text-phrase and thumbnail-image menus those items that best represented the material covered by the skim. The menus were populated with the independently validated text phrases and representative images.

Results

At the 0.05 level of significance, the four skim types yielded no differences in mean accuracy or speed on either factfinding or gisting. This result was surprising to us, since we expected the default skim to produce slower and less accurate performances than the other three skims. Pilot studies had shown us that users found the default skim "jerky," "too hard to follow," and "too jumpy."

There were also no significant (0.05 level) differences between the QUIS answers concerning user satisfaction for the four skim types.

EXPERIMENT ONE ANALYSIS

Several factors may have contributed to the lack of observed differences among skim types:

- All the tested skims used a small grain size. Even if the IMG, AUD, and BOTH skims had successfully identified and incorporated "important" segments, those components may have been too brief to communicate content effectively. Thus fine granularity may have masked differences among skim designs, leading subjects to consider all the skims essentially equivalent.
- The source videos were fairly short, 8 to 12 minutes, so skims ran 48 to 72 seconds. While these skims

reduced viewing time by 7 to 11 minutes over watching the full video, perhaps the benefits of compaction become significant only for longer source video segments. Maybe 30-minute videos and 3-minute skims, for example, would work better for showcasing skim benefits.

- Two of the skim designs failed to preserve audio/video synchrony, a lack that may have distracted users enough to offset any benefits these skims provided over the other types.
- Users expressed difficulty in seeing the low-resolution thumbnail images, which occupied only 1/16 the area of the skim and source-video images.

REDESIGNING SKIMS

We addressed these shortcomings by:

- Modifying the key-phrase selection process and extending the average grain size from three to five seconds.
- Generating skims from longer, half-hour source videos.
- Improving audio/video synchronization in the "integrated best" skim.
- Using larger images in our gisting instrument (352×240 pixels, the same resolution as the MPEG-I video).

Our main concern in redesigning skims was granularity. User feedback in the first study indicated that all skims appeared disjointed and that the audio, in particular, was too choppy for easy comprehension.

Where our initial approach to audio segmentation relied solely on transcript analysis, for our second study we grouped words into phrases using signal power. Other researchers have similarly used speech signal characteristics to produce compact audio representations [1]. This analysis calculates the power of an audio sample as:

$$Power = \log\left(\left(\frac{1}{n}\right) \bullet \sum \left(Si^2\right)\right)$$

where *Si* is the signal intensity — low frequencies pre-emphasized — within a 20 ms frame, and *n* is the count of frames averaged. A low power level indicates little active speech, and we inserted segment breaks where power dropped to the average minimum in a trailing 0.5-second window. Thus the audio signal itself delineates utterance beginnings and endings.

The resulting phrases are typically longer than those selected in our first study. For example, Figure 3 shows a case where the noun-phrase approach would isolate the first eight words of a seventeen-word sentence. The power method, however, selects the full sentence.

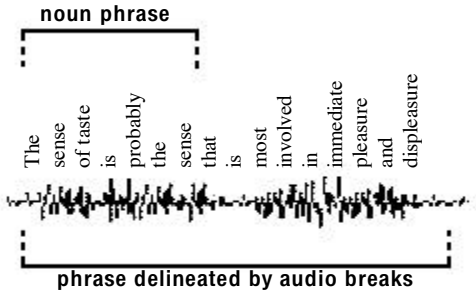

Figure 3. Audio regions based on parsed noun-phrases (Experiment 1) and signal-power segmentation (Exp. 2)

Detected utterances shorter than eight seconds are included unchanged. Words within a selected utterance constitute a candidate phrase, and phrases are scored using TF-IDF weighting, as before. Skims are built from the highest scoring phrases and, for the documentary videos used in these studies, average five seconds in length.

Another major concern with skims from our first study was synchrony between audio and video in the image-centric (IMG) and integrated best (BOTH) skims. For a given audio phrase these skims included images from a window considerably broader than that corresponding to the audio. For our second study we limited image adjustments to substitute only neighboring video for audio regions near shot breaks or blank video frames.

Pilot testing of our revised skims revealed that:

- People questioned the benefits of skims relative to using the full video.
- People found choppy audio more annoying than choppy video.
- Some people took the extreme position that the audio carries the entire gist for a movie and that two skims with the same audio track will produce similar results, regardless of the video content.

This feedback directly affected the design of the subsequent skim study conducted in September 1997.

RECONSIDERING TASKS

The factfinding task in our first experiment may have failed to distinguish among skims because it underutilized their temporal aspects. Our goal with skims is to communicate essential imagery and narrative. However, for locating a particular fact within a video, a skim's coverage may be more critical than how well it captures important parts. Sophisticated skim designs may offer little inherent advantage over more mechanistic ones that provide uniform coverage, abstractions such as our default skim (DEF) or simultaneous, static displays of representative thumbnail images [6, 9, 11, 15, 17, 18].

Showing where query-matching words occur within a video's transcript also may aid navigation to a point within a video more directly than a skim [3]. Rather than attempt to justify the use of skims for navigation, we decided to address only the issue of gisting in our subsequent skim study.

Our first experiment measured gisting through both text and image questionnaires. The text was not taken verbatim from the video, but rather was composed by Informedia staff members to serve as video descriptors for library delivery and use [2]. The same text representations are used again to measure gisting in Experiment Two.

For gisting images to complement the text phrases, we might have, ideally, developed a pictorial questionnaire that summarized a video without explicitly including any of its contents. Since this goal presented significant practical difficulties, we chose instead to use representative images carefully selected *from* the video and independently validated.

In our first experiment some skims incorporated such representative images while others, in fact, did not. For example, the DEF skim of a 12-minute video may have contained eight such images, while the AUD skim of the same source may have omitted them all. Viewers of this DEF skim have essentially an image recognition task, since they have seen the images being presented to them in the just-watched skim video. Viewers of this AUD skim would face the more difficult, although more authentic gisting task of determining whether the eight images could be part of the source video condensed in the skim. We wished to eliminate that variance in tasks.

For the image-gist instrument in our second experiment, we used only representative images that appeared in all skim treatments, so that, across all treatments, our pictorial questionnaire tested image recognition only. Subjects were asked to determine whether the image was part of the video they had just seen.

EXPERIMENT TWO

Our second study employed five experimental treatments: four skim types — each 7.5 times shorter than the associated video — and a fifth treatment that showed the full source video itself. The level of compaction in these skims extracted approximately eight seconds per minute of video, a capture ratio essentially determined by our power-based audio segmentation technique.

Subjects

Twenty-five Carnegie Mellon students (16 male, 9 female) from several majors and programs volunteered for the study, responding to an electronic call for participation. Each received $7 for spending about eighty minutes with the system. As in Experiment One, a background questionnaire revealed that the subjects were, in general, "very comfortable" with computers but had little prior experience with digital video.

Materials

The video material was drawn from the same three public television series as used in the first study, with manually generated text phrases and chosen representative images again serving as the gist of a video for our experiment.

Design

The five treatments in this experiment were:

- DFS: a default skim using short components and comprising seconds 0-2.5 from the full source video, then seconds 18.75-21.25, seconds 37.5-40, etc.
- DFL: a default skim using long components and consisting of seconds 0-5, then seconds 37.5-42.5, seconds 75-80, etc.
- NEW: our redesigned, "integrated best" skim
- RND: "best audio" with reordered video
- FULL: complete source video, with no information deleted or modified

Figure 4 shows how source components map to these four skims. Two variants of our default skim (DFS and DFL) tested grain-size effects. DFS components were 2.5 seconds, and DFL used segments twice as long. The "new integrated best" design (NEW) had the same average granularity as DFL, constrained image regions to contiguous frames, and limited synchronization offsets to minor shifts between associated video and audio regions. A fourth skim type (RND) addressed the effects of extreme synchronization differences. It used the same audio and video as NEW but reversed video-component ordering, so that audio and video corresponded only at mid-skim.

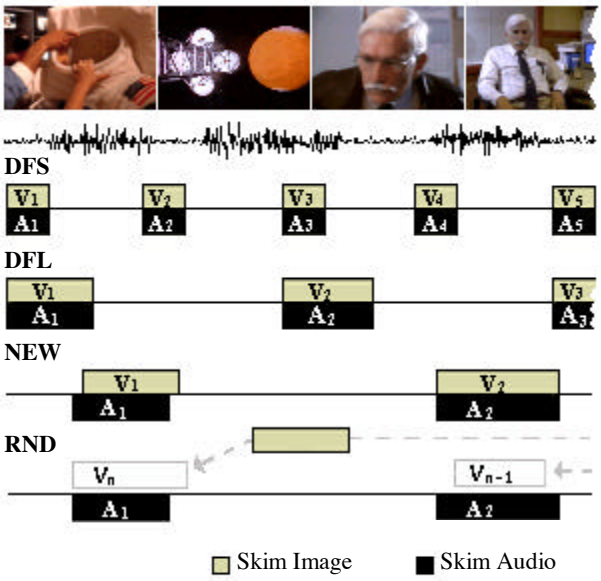

Figure 4. Audio/video alignment in four skims (Exp. 2)

Procedure

Subjects participated in the study individually, as in the first experiment. Each used a computer with a 17-inch color monitor, hardware support for smooth full-motion video playback, and headphones. All materials were presented online. After entering background information and reading the instructions, the subject viewed a short video skim of a popular movie to get acclimated to skim presentations. We used a 5×5 Latin Square configuration to balance any learning effect between treatments [8]. Thus we repeated the following procedure five times, using a different source video and treatment on each iteration:

1. The subject watched a video without interruption. For DFS, DFL, NEW, and RND, a presentation lasted about four minutes; the FULL video ran approximately 30 minutes. One-fifth of the subjects saw DFS first, one-fifth saw DFL first, etc.
2. The subject answered three Likert-scale questions taken from QUIS [10] plus three subjective questions concerning opinions about the just-completed video.
3. The interface then presented ten images, one at a time, each at the same resolution as the video. Subjects selected "yes" or "no" based on whether they recognized the image as one from the video.
4. The interface presented 15 text phrases, one at a time, and for each the subject selected "yes" or "no" to indicate whether that text phrase summarized information that would be part of the full source video. This is the same metric used in the first experiment.

Finally, we asked each subject how well the video had prepared him or her for the just-completed questions and invited them each to type comments concerning this particular video treatment.

Results

Analysis revealed significant ($p < 0.01$) differences in mean performance on text gisting and image recognition among the five video treatments. A Student-Newman-Keuls test (SNK) subsequently examined whether differences between specific means were significant [8], thus enabling us to evaluate the relative merits of the various skim treatments and the original source video treatment (FULL).

Mean performance on the ten image questions is given in Figure 5. An SNK analysis revealed that RND's mean was significantly ($\alpha = 0.05$) different from all other treatment means. No other significant differences were found between the treatment means, that is, the other three skim treatments promoted image recognition as well as the full video. Only when synchronization was extremely poor (the RND treatment) did image recognition performance diminish significantly.

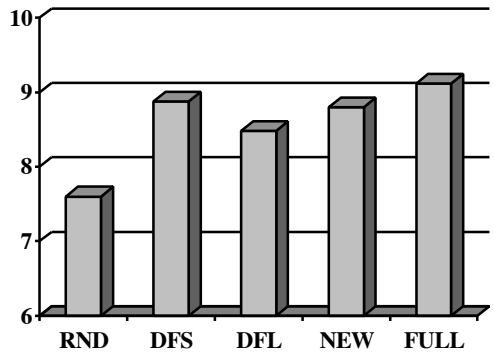

Figure 5. Mean scores for image recognition

The mean performance for the 15 text-gist questions, given in Figure 6, was generally worse than that for the image recognition questions. This difference is likely due to the fact that while the images in question were actually shown during the presentation, subjects neither saw nor heard the text.

Testing the text-gisting means with SNK revealed that FULL's mean was significantly ($\alpha = 0.05$) different from the other four treatment means. The subjects understood the essence of a video better when they saw the full version rather than a video skim. The NEW mean was also significantly different from the RND mean, with no other significant differences found between the treatment means.

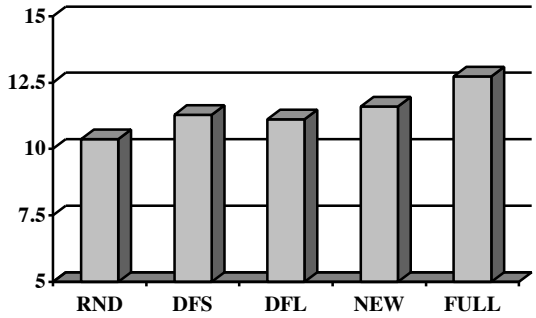

Figure 6. Mean scores for text phrase identification

Figure 7 shows mean subjective responses to the three QUIS questions used in this experiment, each with different shading. On these nine-point scales "1" mapped to "terrible," "frustrating," or "dull" and "9" to "wonderful," "satisfying," or "stimulating," respectively. The trend revealed here shows the FULL treatment to be the most preferred, followed in order by NEW, DFL, and then DFS or RND.

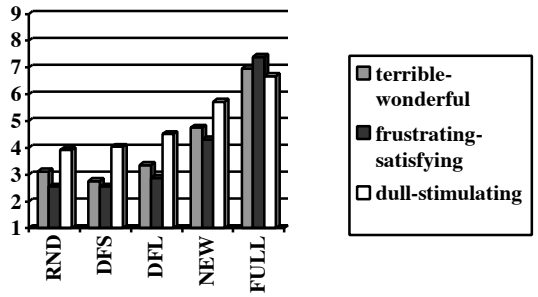

Figure 7. Mean scores for three QUIS subjective ratings

We added two nine-point scales to measure the subject's perception of audio and video quality ("1" = "poor audio" and "poor video"), with mean scores presented in Figure 8.

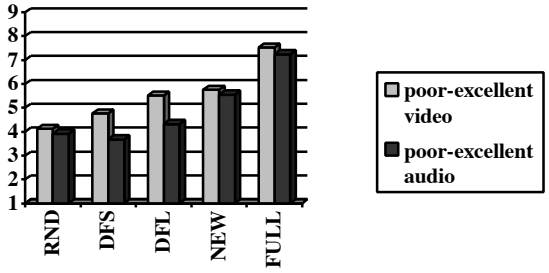

Figure 8. Mean scores for audio/video subjective ratings

The subjects were directly asked how well they felt the video skim did in communicating the essence of a longer video segment. This question was only asked following the viewing of one of the skim treatments, and the mean results from the nine-point scale ("1" = "inadequately") are shown in Figure 9. The subjects were also asked how well they felt the video treatment informed them for answering the text and image questions. These mean results ("1" = "poorly informed") are shown in the figure as well.

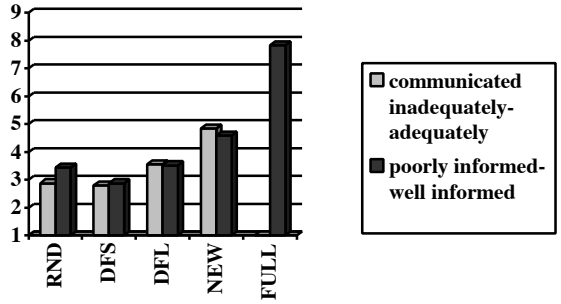

Figure 9. Mean scores for additional subjective ratings

Analysis of variance for the Latin Square design for all seven subjective questions showed significant ($p < 0.01$) differences in mean ratings among the five video treatments. Testing the means with SNK revealed that FULL's mean was significantly ($\alpha = 0.05$) different from all other treatment means and that, for six of seven cases, NEW's mean was significantly different from all other skim treatment means. For the seventh case ("poor-excellent video") NEW's mean was still the greatest of the skim treatment means and significantly different from all but the DFL treatment mean.

The subjects' open-ended comments supported these results as well. An informal classification of the 59 open-ended comments offering a favorable or critical opinion produced the distribution shown in Figure 10.

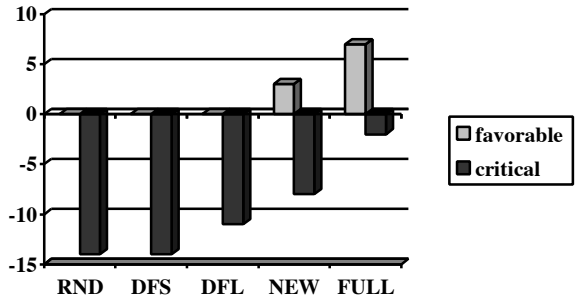

Figure 10. Count of open-ended comments by treatment

DISCUSSION

Clearly, the subjects preferred the full video to any of the skim types in the second experiment. However, subjects favored the NEW skim over the other skim treatments, as indicated by the subjective ratings. These results are encouraging, showing that incorporating speech, language, and image processing into skim video creation produces skims that are more satisfactory to users.

The RND skim distinguished itself as significantly poorer than NEW on the text-phrase gisting instrument, despite the fact that both RND and NEW use identical audio information. This result shows that the visual content of a video skim does have an impact on its use for gisting and so addresses a point raised during earlier pilot studies.

The DFS and DFL skim treatments did not particularly distinguish themselves from one another, leaving open the question of the proper grain size for video skims. The larger grain size, when used with signal-power audio segmentation, produced the NEW skim that *did* distinguish itself from the other skims. If the larger grain size is used only for subsampling, however, it yields no clear objective or subjective advantage over short grain size skims, such as DFS. In fact, both DFS and DFL often rated similarly to RND, indicating perhaps that *any* mechanistically subsampled skim, regardless of granularity, may not do notably well.

While our first study found no significant differences between a subsampled skim and a "best" audio and video skim, the second study uncovered numerous statistically significant differences. The primary reasons for the change can be traced to the following characteristics of the audio data in the latter experiment:

- Skim audio is less choppy due to setting phrase boundaries with audio signal-processing rather than noun-phrase detection.
- Synchronization with video is better preserved.
- Grain size has increased from three seconds to five.

Although the NEW skim established itself as the best design under study, considerable room for improvement remains. It received mediocre scores (4-6) on most of the subjective questions, and its improvement over the other skims may reflect more on their relatively poor evaluations than on its own strengths. NEW did distinguish itself from RND for the image recognition and text-phrase gisting tasks, but not from the DFS and DFL skims.

FUTURE WORK

Image components for skims merit further investigation. Our NEW skim achieved smoother audio transitions but still suffered abrupt visual changes between image components. Perhaps transitions between video segments should also be smoothed — through dissolves, fades, or other effects — when they are concatenated to form a skim.

Other researchers have focused exclusively on image-based video abstractions [5]. Such strategies typically decompose video into shots and represent each shot with a selected image [6, 11, 15, 17, 18]. Concatenating these representative images yields a form of video skim that provides full coverage of all component shots with duration proportional to the time for displaying each image [5]. Such skims resemble "automatic slide shows" where still images appear sequentially. The skims investigated here, however, more resemble "video digests" that are "played." Combining the two approaches would produce a visually dense representation with complementary audio. Further study is required to determine whether more uniform coverage offsets the loss of temporal flow and audio/video synchrony.

Finally, we have focused on general-purpose skims. Work on other multimedia abstractions has shown the benefits of exploiting context to tailor representations [3]. For example, given a query, a skim emphasizing target regions that contain matches may prove more effective than a general-purpose skim. We intend to explore such context-based skims in future work.

ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation, DARPA, and NASA under NSF Cooperative Agreement No. IRI-9411299. Michael Smith is supported by Bell Laboratories. Numerous Informedia Project members contributed to this work, including Howard Wactlar, Takeo Kanade, Alex Hauptmann, Michael Witbrock, Craig Marcus, Naomi Dambacher, Jayshree Ranka, and Bruce Cardwell. Special thanks go to Ellen Hughes, Yuichi Nakamuri, Bryan Maher, Ricky Houghton, and Laurel Margulis for their invaluable assistance. Finally, we thank Informedia Project partner QED Communications for the video source material.

REFERENCES

1. Arons, B. SpeechSkimmer: Interactively Skimming Recorded Speech. In *Proceedings of the ACM Symposium on User Interface Software and Technology*. (Atlanta, GA, Nov 1993), 187-196.
2. Christel, M.G., and Pendyala, K. Informedia Goes To School: Early Findings from the Digital Video Library Project. *D-Lib Magazine* (September 1996). URL http://www.dlib.org/dlib/september96/09contents.html.
3. Christel, M.G., Winkler, D.B., and Taylor, C.R. Improving Access to a Digital Video Library. In *Human-Computer Interaction: INTERACT97, the 6th IFIP Conf. On Human-Computer Interaction*. (Sydney, Australia, July 14-18, 1997).
4. Degen, L., Mander, R., and Salomon, G. Working with Audio: Integrating Personal Tape Recorders and Desktop Computers. In *Proceedings of the ACM CHI'92 Conference on Human Factors in Computing Systems*. (Monterey, CA, May 1992), 413-418.
5. Ding, W., Marchionini, G., & Tse, T. Previewing Video Data: Browsing Key Frames at High Rates Using a Video Slide Show Interface. In *Proceedings of the International Symposium on Research, Development & Practice in Digital Libraries*. (Tsukuba Science City, Japan, November 1997), 151-158.
6. Hampapur, A., Jain, R., and Weymouth, T. Production Model Based Digital Video Segmentation. *Multimedia Tools and Applications*, *1* (March 1995), 9-46.
7. Hauptmann, A.G., and Witbrock, M.J. Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. In *Intelligent Multimedia Information Retrieval*, M. Maybury, Ed. 1997, AAAI Press/MIT Press: Menlo Park, CA.
8. Lee, W. *Experimental Design and Analysis*. 1975, W.H. Freeman & Co.: San Francisco, CA.
9. Mills, M., Cohen, J., and Wong, Y.Y. A Magnifier Tool for Video Data. In *Proceedings of the ACM CHI'92 Conference on Human Factors in Computing Systems*. (Monterey, CA, May 1992), 93-98.

10. *QUIS 5.5b*. University of Maryland at College Park, 1994. Available through http://www.lap.umd.edu/QUISFolder/quisHome.html.

11. Pfeiffer, S., Lienhart, R., Fischer, S., and Effelsberg, W. Abstracting Digital Movies Automatically. *Journal of Visual Communication and Image Representation*, *7*, 4 (Dec 1996), 345-353.

12. Rowley, H., Baluja, S. and Kanade, T. Human Face Detection in Visual Scenes. Carnegie Mellon University, School of Computer Science Technical Report CMU-CS-95-158 (Pittsburgh, PA, 1995).

13. Smith, M., Kanade, T. Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques. In *Computer Vision and Pattern Recognition*. (San Juan, PR, 1997).

14. Stevens, S. Next Generation Network and Operating System Requirements for Continuous Time Media. In *Network and Operating System Support for Digital Audio and Video,* R. Herrtwich, Ed. 1992, Springer-Verlag; New York.

15. Taniguchi, Y., Akutsu, A., Tonomura, Y., and Hamada, H. An Intuitive and Efficient Access Interface to Real-Time Incoming Video Based on Automatic Indexing. In *Proceedings of the ACM Multimedia Conference*. (San Francisco, CA, November 1995), 25-33.

16. Wactlar, H.D., Kanade, T., Smith, M.A., and Stevens, S.M. Intelligent Access to Digital Video: Informedia Project. *Computer*, *29*, 5 (May 1996), 46-52.

17. Yeung, M., Yeo, B., Wolf, W., and Liu, B. Video Browsing Using Clustering and Scene Transitions on Compressed Sequences. In *Proceedings of IS&T/SPIE Multimedia Computing and Networking* (1995).

18. Zhang, H.J., Tan, S., Smoliar, S., and Yihong, G. Automatic Parsing and Indexing of News Video. *Multimedia Systems*, *2*, 6 (1995), 256-266.