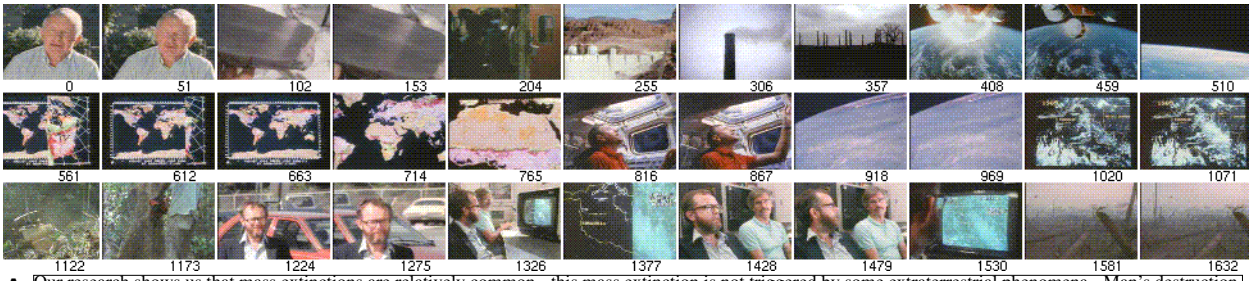
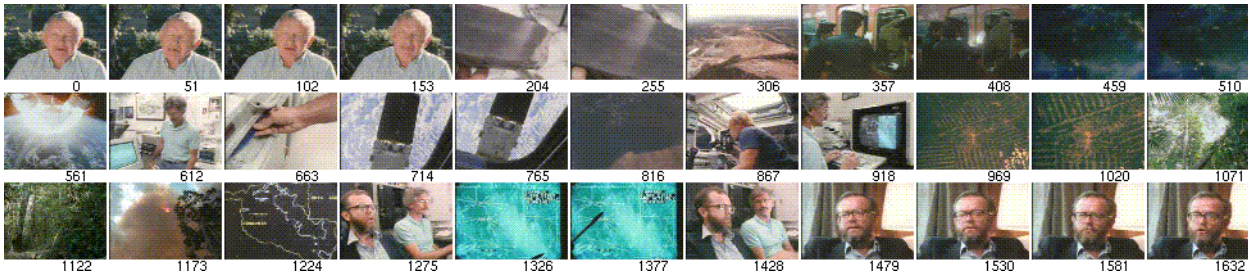
[7] Stevens, S., Christel, M., and Wactlar, H. "Informedia: Improving Access to Digital Video" *Interactions* 1 Oct.94.
[8] Zhang, H., *et a.l*, "Automatic Partitioning of Full-Motion Video," *Multimedia Systems* 1993 1, pp. 10-28.
[9] Arman, F., Hsu, A., and Chiu, M-Y. "Image Processing on Encoded Video Sequences," *Multimedia Systems* 1994.
[10] Arman, F., *et al*., "Content-Based Browsing of Video Sequences," *Proc. of ACM Multimedia '94*, Oct., 1994.
[11] "TREC 93," *Proceedings of the 2nd Text Retrieval Conference*, D. Harmon, Ed., sponsored by ARPA/SISTO, 1993.
[12] Hauptmann, A.G. and Smith, M, "Video Segmentation in the Informedia Project", IJCAI-95, *Workshop on Intelligent Multimedia Information Retrieval*. Aug. 1995.
[13] "MPEG-1 Video Standard", *Communications of the ACM*, April 1991.
[14] Smallman, K., "Creative Film-Making", 1st ed., Publisher Macmillan, New York 1970.

Acknowledgments

We thank Henry Rowley and Shumeet Baluja for the face detection routine; Michael Witbrock and Yuichi Nakamura for the keyword selection routines; and the Carnegie Mellon Drama Department for their insights on skim creation. This work is partially funded by NSF, NASA, and ARPA. Michael Smith is supported by Bell Laboratories.



A Our research shows us that mass extinctions are relatively common - this mass extinction is not triggered by some extraterrestrial phenomena - Man's destruction of the diversity of life - man has the technology to change the world - human waste are threatening the web of life - A tapestry of lights track the human presence - Fires in Africa fuel the struggle against famine - At NASA's Goddard Space Flight Center in Maryland Compton Tucker draws - past eight years Tucker has observed subtle shifts - Discovery returns to space on a mission to photograph the Earth - you see these incredibly large fires - Compton Tucker has been monitoring these fires - The destruction was triggered - homesteaders are transforming the wilderness - Norman Myers has voiced his concern - invited Myers to review his most recent findings - we could have 2 or 3 smaller fires burning - we're pushing species down the tubes on our own planet - What splendid creatures lived here



B Our research shows us that mass extinctions are relatively - and diversity lost - and if fact may be regarded as a human meteorite - stripped away for resources - against famine in the Third World - At NASA's Goddard Space Flight - Houston Discovery - Amazon they are stunned by what they see - images confirm that in one - roads - visible from space - as - is particularly disturbing - fires there at about 2:45 in the afternoon - more probes sent off to have a look at this little - in the long run. It matters enormously - a few becomes many becomes too many

Figure 11: Image and text output for the "Mass Extinction" segment: A) Skim creation using image and language understanding, B) Skim creation using fixed intervals for image and audio.

4 Conclusions

The emergence of high volume video libraries has shown a clear need for content-specific video-browsing technology. We have described an algorithm to create skim videos that consist of content rich audio and video information. Compaction of video as high as 20:1 has been achieved without apparent loss in content.

While the generation of content-based skims presented in this paper is very limited due to the fact that the true understanding of video frames is extremely difficult, it illustrates the potential power of integrated language, and image information for characterization in video retrieval and browsing applications.

5 References

[1] Akutsu, A. and Tonomura, Y. "Video Tomography: An efficient method for Camerawork Extraction and Motion Analysis," *Proc. of ACM Multimedia '94*, Oct.1994.

[2] Degen, L., Mander, R., and Salomon, G. "Working with Audio: Integrating Personal Tape Recorders and Desktop Computers," *Proc. CHI '92,* May 1992, Monterey, CA.

[3] Hampapur, A., Jain, R., and Weymouth, T. "Production Model Based Digital Video Segmentation," *Multimedia Tools and Applications* **1** March 1995.

[5] Mauldin, M. "Information Retrieval by Text Skimming," PhD Thesis, Carnegie Mellon University. Aug. 1989.

[6] Rowley, H., Baluja, S. and Kanade, K. "Neural Network-Based Face Detection," *Computer Vision and Pattern Recognition*, San Francisco, May 1996.

has captured 23 scenes, and skim B has captured 17 scenes. Studies involving different skim creation methods are discussed in the next section.

Skim A has only 1632 frames, while the first scene of the original segment is an interview that lasts 1734 frames. The scenes that follow this interview contain camera motion, so we select them for the keyphrases towards the end of the scene. Charts and figures interleaved between successive human subjects are selected for the latter scenes.

Table 3: Skim Compaction Data

Title	Original(sec)	Skim (sec)	Comments
K'nex, CNN Headline News	61.0	7.13	MC-AS
Species Destruction I	68.65	6.40	MC-AS
Species Destruction II	123.23	12.43	MS
International Space University	166.20	28.13	MS
Rain Forest Destruction	107.13	5.36	MS
Mass Extinction	559.4	55.5	AC-AS
Human Archeology	391.2	40.8	AC-AS
Planet Earth I	464.5	44.1	AC-AS
Planet Earth II	393.0	40.0	AC-AS

Comments
MC- Manually Assisted CharacterizationAC- Automated Characterization
MS- Manual Skim CreationAS- Automated Skim Creation

3.5 User Evaluation

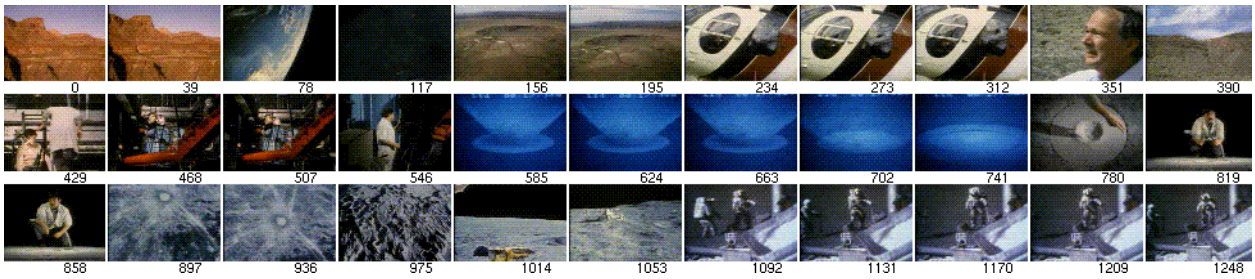
The results of several skims are summarized in Table 3. The manually created skims in the initial stages of the experiment help test the potential visual clarity and comprehension of skims. The compaction ratio for a typical segment is 10:1; and it was shown that skims with compaction as high as high as 20:1 still retain most of the content. Our results show the information representation potential of skims, but we must test our work with human subjects to study its effectiveness.

We are conducting a user-study to test the content summarization and effectiveness of the skim as a browsing tool in a video library. Subjects must navigate a video library to answer a series of questions. The effectiveness of each skim is based on the time to complete this task and the number of correct items retrieved. Although our evaluation results are tentative, the skim does appear to be an effective tool for browsing, as evident by the difference of time that subjects spend in skim mode versus regular playback mode.

We use various types of skims to test the utility of image and language understanding in skim creation. The following creation schemes are presently being tested:

A - Image and Language Characterization
B - Fixed Intervals (Default)
C - Language Characterization Only
D - Image Characterization Only

Figure 11 shows examples of skim type A and B. The visual information in skim A is less redundant and provides a greater variety of scenes. The audio for skim B is incoherent and considerably smaller. Although our skim does appear to provide more information, additional testing is needed.



Keyphrases

its restless geology erased the record - our long-dead Moon - by still more violent events - event of unimaginable force - Gene Shoemaker a scientist with the U.S. Geological Survey helped prove - another part of the evidence we can see - collisions with Earth are extremely rare - they reconstruct these great collisions - O.K., Earl, let's shoot it - and we can see the ejecta curtain simply move out - these clumps eventually will impact the surface - now this is the edge of the target - if we step back we find - this is what makes up the crater rays - and the Moon's dark plains - as astronauts roam the Moon's surface - most Moon rocks were between three and four billion - the samples gathered so carefully were not part - the early chapters of the Moon's history began to emerge

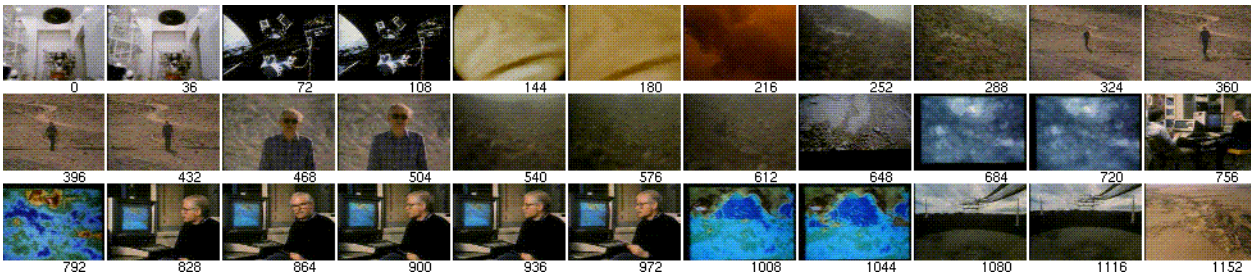
Figure 9: Skim video frames and keyphrases for "Planet Earth - I" (10:1 compaction).

3.4 Example Results

Figure 9 shows the video frames and audio from the "Planet Earth" video. The image portion of the skim has captured information from 18 of the 64 total scenes in the video. With the exception of the scene at frame 585, which lasts over 1,300 frames in the original video, most scenes are small and provide maximum visual information. An error in scene segmentation, near frame 702, causes this scene to split and, therefore, it is used twice for separate keyphrases. The final scene in the original video is long and contains two keyphrases. In this case, the search window cannot extend to other scenes and these keyphrases must share image frames from the final scene. Introduction scenes, bounded camera motion and human faces dominate the image skims for this segment.

Figure 10 shows another example from the "Planet Earth" video with 16 of the 37 scenes represented. This segment contains many long outdoor scenes that provide little information. However, most primitive rules do not match these scenes so the search window is extended and they appear less frequently in the image skim. The scene at frame 828 is an interview scene which contains 3 keyphrases and lasts several frames. Even with an extended search window, the scenes that follow do not match any of the primitive rules so the image skim is rather long for this scene.

Figure 11 shows two types of skims for the "Mass Extinction" segment. Skim A was produced with our method of integrated image and language understanding. Skim B was created by selecting video and audio portions at fixed intervals. This segment contains 71 scenes, of which, skim A



Keyphrases

to explore these neighboring planets - the first planet we would visit was Venus - it resisted exploration - our voyages revealed it - through the dense atmosphere winds up to two hundred miles an hour whip the clouds - no place can replicate the climate - but Death Valley, on a hot afternoon - how did it become the hot desolate planet - Jim Head, a planetary scientist at Brown University, is an authority - it's hot enough on the surface of Venus - the heat that comes into the surface can actually be - they radioed back images - and Jim Head and his colleagues had a global - we can also see some very distinctive highland - this one, Maxwell Montes - we have the same problem the resolution - Jim Head and his colleagues use a very special - the extensional forces are pulling

Figure 10: Skim video frames and keyphrases for "Planet Earth - II" (10:1 compaction).

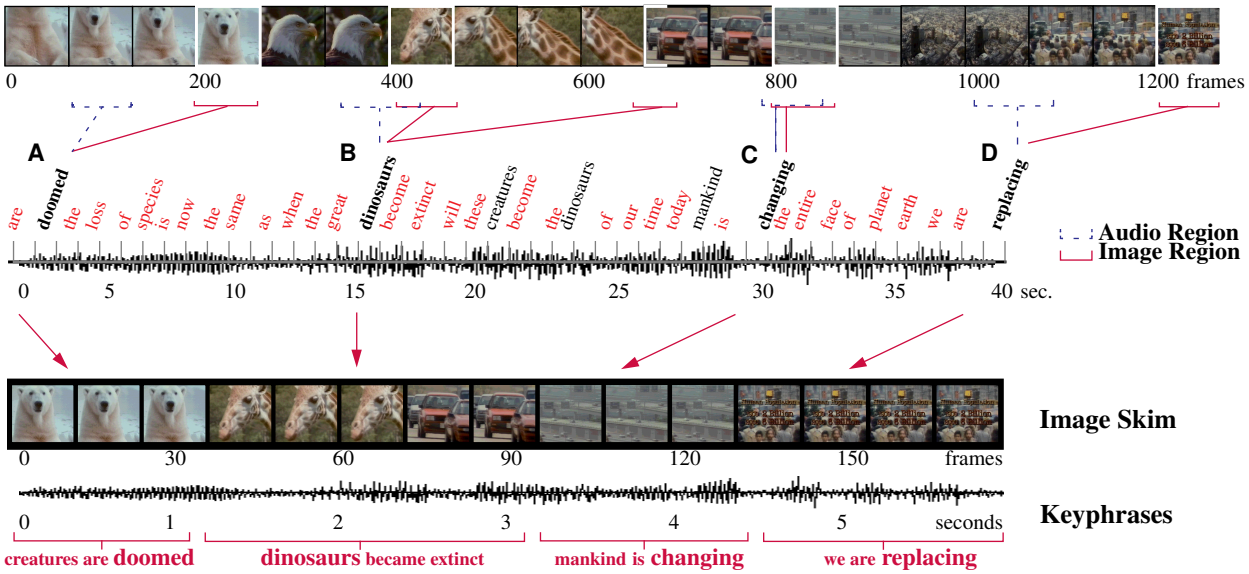


Figure 8: Skim creation incorporating word relevance, significant objects (humans and text), and camera motion: A) For the word "doomed", the images following the camera motion are selected, B) The keyphrase for "dinosaur" is long so portions of the next scene are used for more content, C) No significant structure for the word "changing", D) For the word "replacing" The latter portion of the scene contains both text and humans.

7. Significant Audio(AUD)

If the audio is music, then the scene may not be used for skimming. Soft music is often used as a transitional tool, but seldom accompanies images of high importance. High audio levels (e.g. loud music, explosions) may imply an important scene is about to occur. The skim region will start after high audio levels or music.

8. Default Rule(DEF)

Default video frames align to the audio keyphrases.

3.3 Image Adjustments

With prioritized video frames from each scene, we now have a suitable representation for combining the image and audio skims for the final skim. A set of higher order Meta-Rules are used to complete skim creation.

For visual clarity and comprehension, we allocate at least 30 video frames to a keyphrase. The 30 frame minimum for each scene is based on empirical studies of visual comprehension in short video sequences. When a keyphrase is longer than 60 video frames, we include frames from skim candidates of adjacent scenes within the 5 second search window. The final skim borders are adjusted to avoid image regions that overlap or continue into adjacent scenes by less than 30 frames.

To avoid visual redundancy, we reduce the presence of human faces and default image regions in the skim. If the highest ranking skim candidate for a keyphrase is the default, we extend the search range to a 10 second window and look for other candidates. The human face rule is limited if the segment contains several interviews. Interview scenes can be extremely long, so we look for other candidates in a 15 second search window.

Figure 8 illustrates the adjustment and final selection of video skims. It shows how and why the image segments, which do not necessarily correspond in time to the audio segments, are selected.

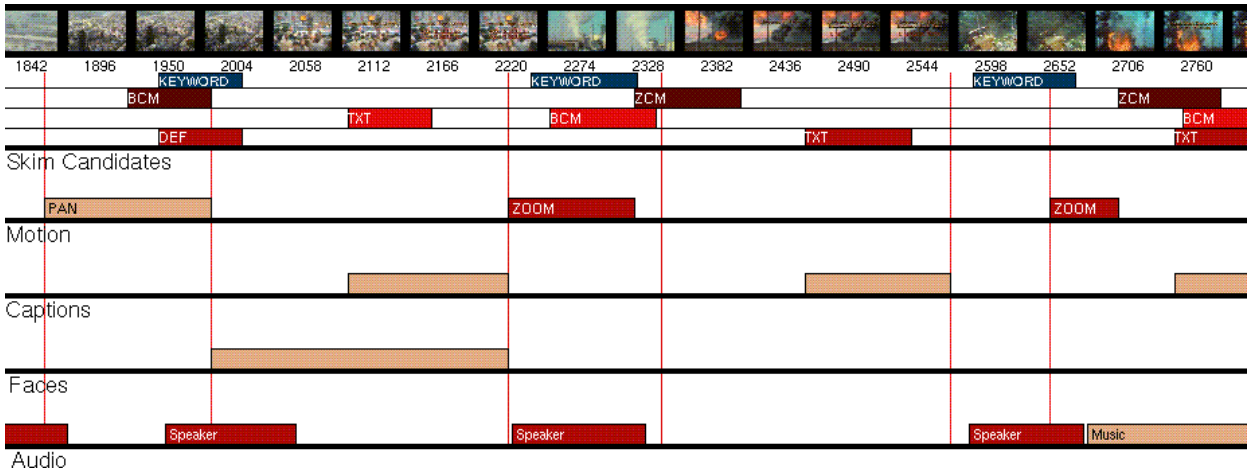


Figure 7:Characterization data with skim candidates and keyphrases for "Destruction of Species". The skim candidate symbols correspond to the following primitive rules: BCM, Bounded Camera Motion; ZCM, Zoom Camera Motion; TXT, Text Captions; and DEF Default. Vertical lines represent scene breaks.

"Skim Candidates", in Figure 7, indicate the candidate image sections selected by various primitive rules.

1. Introduction Scenes(INS)

The scenes prior to the introduction of a proper name usually describe a person's accomplishment and often precede scenes with large views of the person's face. If a keyphrase contains a proper name, and a large human face is detected within the surrounding scenes, then we set the face scene as the last frame of the skim candidate and use the previous frames for the beginning.

2. Similar Scenes(SIS)

The histogram technology in scene segmentation gives us a simple routine for detecting similarity between scenes. Scenes between successive shots of a human face usually imply illustration of the subject. For example, a video producer will often interleave shots of research between shots of a scientist. Images between similar scenes that are less than 5 seconds apart, are used for skimming.

3. Short Sequences(SHS)

Short successive shots often introduce a more important topic. By measuring the duration of each scene, we can detect these regions and identify "short shot" sequences. The video frames that follow these sequences and the exact sequence are used for skimming.

4. Object Motion(OBM)

Object motion is import simply because video producers usually include this type of footage to show something in action. We are currently exploring ways to detect object motion in video.

5. Bounded Camera Motion(BCM/ZCM)

The video frames that preceed or follow a pan or zoom motion are usually the focus of the segment. We can isolate the video regions that are static and bounded by segments with motion, and therefore likely to be the focal point in a scene containing motion.

6. Human Faces and Captions(TXT/FAC)

A scene will often contain recognizable humans, as well as captioned text to describe the scene. If a scene contains both faces and text, the portion containing text is used for skimming. A lower level of priority is given to the scenes with video frames containing only human-faces or text. For these scenes priority is given to text.