Video Skimming and Characterization through the Combination of Image and Language Understanding

Michael A. Smith   Takeo Kanade

Department of Electrical and Computer Engineering

Carnegie Mellon University

Pittsburgh, PA 15213

{msmith, tk}@cs.cmu.edu

Abstract

Digital video is rapidly becoming important for education, entertainment, and a host of multimedia applications. With the size of the video collections growing to thousands of hours, technology is needed to effectively browse segments in a short time without losing the content of the video. We propose a method to extract the significant audio and video information and create a "skim" video which represents a very short synopsis of the original. The goal of this work is to show the utility of integrating language and image understanding techniques for video skimming by extraction of significant information, such as specific objects, audio keywords and relevant video structure. The resulting skim video is much shorter, where compaction is as high as 20:1, and yet retains the essential content of the original segment. We have conducted a user-study to test the content summarization and effectiveness of the skim as a browsing tool.

1 Introduction

With increased computing power and electronic storage capacity, the potential for large digital video libraries is growing rapidly. These libraries, such as the Informedia^{TM} Project at Carnegie Mellon [7] [14], will make thousands of hours of video available to a user. For many users, the video of interest is not always a full-length film. Unlike video-on-demand, video libraries should provide informational access in the form of brief, content-specific segments as well as full-featured videos.

Even with intelligent content-based search algorithms being developed [5], [11], multiple video segments will be returned for a given query to insure retrieval of pertinent information. The users will often need to view all the segments to obtain their final selections. Instead, the user will want to "skim" the relevant portions of video for the segments related to their query.

1.1 Browsing Digital Video

Simplistic browsing techniques, such as fast-forward playback and skipping video frames at fixed intervals, reduce video viewing time. However, fast playback perturbs the audio and distorts much of the image information[2], and displaying video sections at fixed intervals merely gives a random sample of the overall content. Another idea is to present a set of "representative" video frames (e.g. keyframes in motion-based encoding) simultaneously on a display screen. While useful and effective, such static displays miss an important aspect of video: video contains audio information. It is critical to use and present audio information, as well as image information, for browsing. Recently, researchers have proposed browsing representations based on information within the video [8], [9], [10], [16]. These systems rely on the motion in a scene, placement of scene breaks, or image statistics, such as color and shape, but they do not make integrated use of image and language understanding.

Research at AT&T Research Laboratories has shown promising results for video summaries when closed-captions are used with statistical visual attributes [15]. A separate group at the University of Mannheim has proposed a system for generating video abstracts [17]. This work uses similar statistics to characterize images, and audio frequency analysis to detect dialogue scenes. These systems analyze the image and audio information, but they do not extract content specific portions of the video.

An ideal browser would display only the video pertaining to a segment's content, suppressing irrelevant data. It would show less video than the original and could be used to sample many segments without viewing each in its entirety. The amount of content displayed should be adjustable so the user can view as much or as little video as needed, from extremely compact to full-length video. The audio portion of this video should also consist of the significant audio or spoken words, instead of simply using
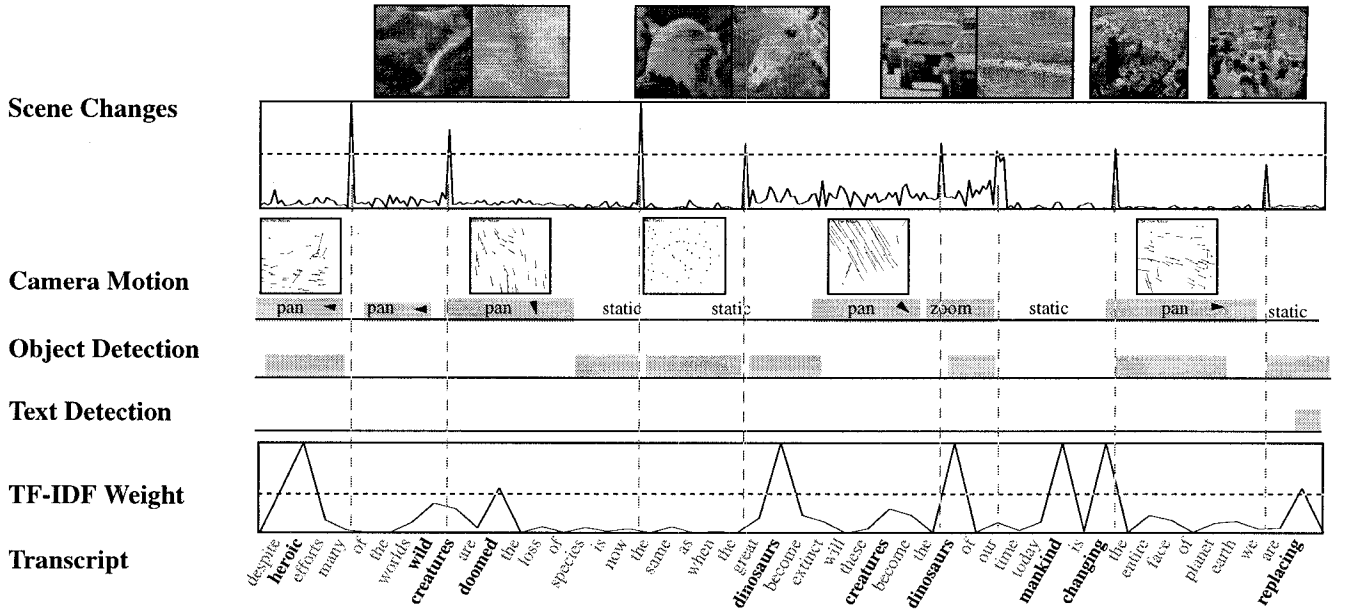
Figure 1: Video Characterization: keywords, scene breaks, camera motion, significant objects (faces and text).

The labels on the left of the figure read: Scene Changes, Camera Motion, Object Detection, Text Detection, TF-IDF Weight, Transcript.

the synchronized portion corresponding to the selected video frames.

1.2 Video Skims

The critical aspect of compacting a video is context understanding, which is the key to choosing the "significant images and words" that should be included in the skim video. We characterize the significance of video through the integration of image and language understanding. Segment breaks produced by image processing can be examined along with boundaries of topics identified by the language processing of the transcript. The relative importance of each scene can be evaluated by 1) the objects that appear in it, 2) the associated words, and 3) the structure of the video scene. The integration of language and image understanding is needed to realize this level of characterization and is essential to skim creation.

In the sections that follow, we describe the technology involved in video characterization from audio and images embedded within the video, and the process of integrating this information for skim creation.

2 Video Characterization

Through techniques in image and language understanding, we can characterize scenes, segments, and individual frames in video. Figure 1 illustrates characterization of a segment taken from a video titled "Destruction of Species", from WQED Pittsburgh. At the moment, language

understanding entails identifying the most significant words in a given scene, and for image understanding, it entails segmentation of video into scenes, detection of objects of importance (face and text) and identification of the structual motion of a scene.

2.1 Audio and Language Characterization

Language analysis works on the transcript to identify important audio regions known as "keywords". We use the well-known technique of TF-IDF (Term Frequency

$$TF\text{-}IDF = \frac{f_s}{f_c} \qquad (1)$$

Inverse Document Frequency) to measure relative importance of words for the video document [5]. The TF-IDF of a word is its frequency in a given scene, f_s, divided by the frequency, f_c, of its appearance in a standard corpus. Words that appear often in a particular segment, but relatively infrequently in a standard corpus, receive the highest TF-IDF weights. A threshold is set to extract keywords, as shown in the bottom rows of Figure 1.

Our experiments have shown that using individual keywords creates an audio skim which is fragmented and incomprehensible for some speakers. To increase comprehension, we use longer audio sequences, "keyphrases", in the audio skim. A keyphrase may be obtained by starting with a keyword, and extending its boundaries to areas of silence or neighboring keywords. Another method for extracting significant audio is to segment actual phrases. To detect breaks between utterances we use a modification
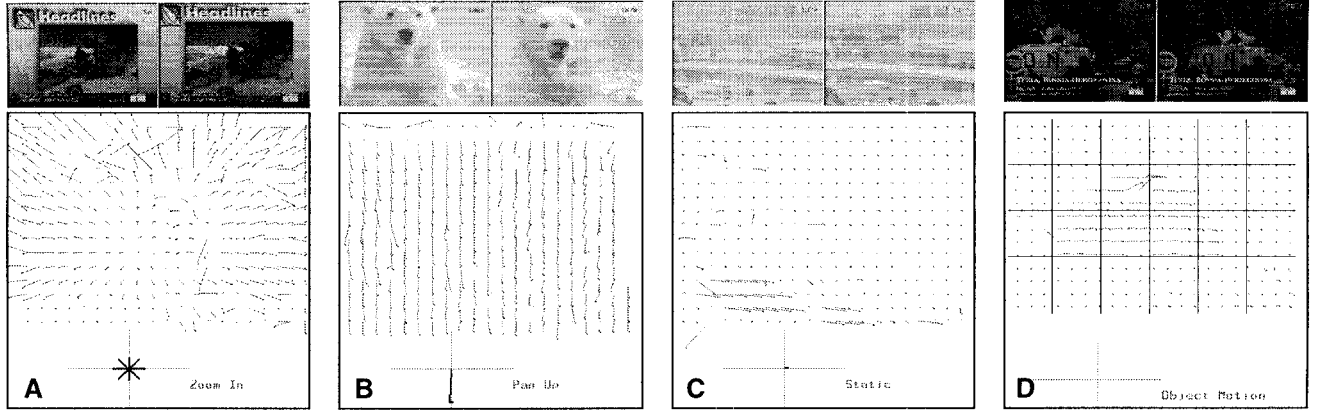
Figure 2: Camera motion from MPEG motion vectors: A) Zoom, B) Pan, C) Static, D) Object motion.

of Signal to Noise ratio (SNR) techniques which compute signal power. This algorithm computes the power of digitized speech samples where S_i is a pre-emphasized sample

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

of speech within a frame of 20 milliseconds. A low power level indicates that there is little active speech occurring in this frame (low energy). Segmentation breaks between utterances are set at the minimum power as averaged over a 1 second window. Empirical studies have shown that breaks at intervals between 2 and 12 seconds produce suitable keyphrases.

Each keyphrase is isolated from the original audio track to form the audio skim. The average keyphrase is roughly 7 seconds. Tests with smaller keyphrases produce skims with fragmented audio, as discussed in section 3.5.

2.2 Scene Segmentation

Many research groups have developed working techniques for detecting scene changes [8], [3], [9]. We choose to segment video by the use of a comparative color histogram difference measure. By detecting significant changes in the weighted color histogram of each successive frame, video sequences are separated into scenes. Peaks in the difference, are detected and an empirically set threshold is used to select scene breaks. This technique is simple, and yet robust enough to maintain high levels of accuracy for our purpose. Using this technique, we have achieved 91% accuracy in scene segmentation on a test set of roughly 495,000 images (5 hours). MPEG-1 video (352x240) is segmented at 36 fps on an SGI Indigo 2 workstation (MIPS R4400 200 MHz.). Examples of segmentation results are shown in the top row of Figure 1.

2.3 Camera Motion Analysis

One important aspect of video characterization is interpretation of camera motion. The global distribution of motion vectors distinguishes between object motion and actual camera motion. Object motion typically exhibits flow fields in specific regions of an image. Camera motion is characterized by flow throughout the entire image.

Motion vectors for each 16x16 block are available with little computation in the MPEG-1 video standard [12]. An affine model is used to approximate the flow patterns

$$u\left(x_i, y_i\right) = ax_i + by_i + c \quad (3)$$

$$v\left(x_i, y_i\right) = dx_i + ey_i + f \quad (4)$$

consistent with all types of camera motion. Affine parameters $a,b,c,d,e,$ and f are calculated by minimizing the least squares error of the motion vectors. We also compute average flow $\bar{v}$ and $\bar{u}$.

Using the affine flow parameters and average flow, we classify the flow pattern. To determine if a pattern is a zoom, we first check if there is the convergence or divergence point (x_0, y_0), where $u\left(x_i, y_i\right) = 0$ and $v\left(x_i, y_i\right) = 0$. To solve for (x_0, y_0), the following relation must be true:

$$\begin{vmatrix} a & b \\ d & e \end{vmatrix} \neq 0$$

If the above relation is true, and (x_0, y_0) is located inside the image, then it must represent the focus of expansion. If $\bar{v}$ and $\bar{u}$, are large, then this is the focus of the flow and camera is zooming. If (x_0, y_0) is outside the image, and $\bar{v}$ or $\bar{u}$ are large, then the camera is panning in the direction of the dominant vector.

If the above determinant is approximately 0, then (x_0, y_0)
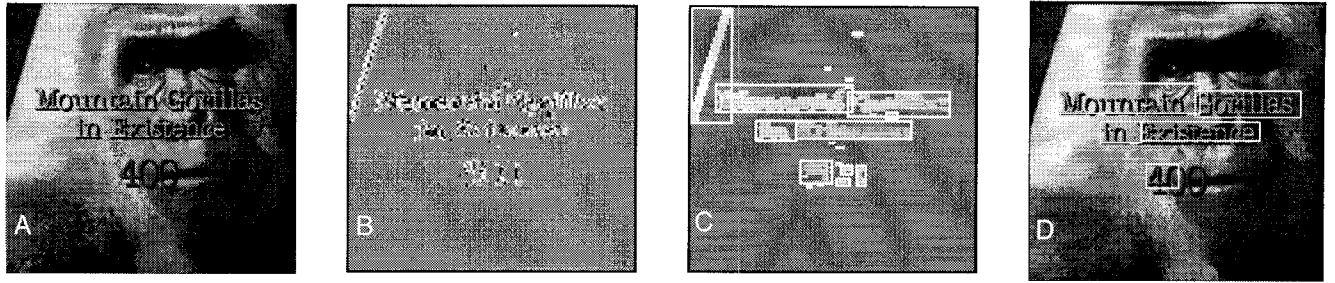
63

Figure 3: Stages of text detection: A) Input, B) Filtering, C) Clustering, and D) Region Extraction.

does not exist and the camera is panning or static. If $\bar{V}$ or $\bar{u}$ are large, the motion is panning in the direction of the dominant vector. Otherwise, there is no significant motion and the flow is static. We eliminate fragmented motion by averaging the results in a 20 frame window over time. The processing rate is 26 fps on an SGI Indigo 2. Examples of the camera motion analysis results are shown in Figure 2. Table 1 shows the statistics for detection on various image sets (regions detected are either pans or zooms).

Table 1: Camera Motion Detection Results

Data(Images)	Regions Detected	Regions Missed	False Regions
Species I - II (20724)	23	5	1
PlanetEarthI-II (25680)	36	1	3
CNHAR News (30520)	14	1	2

2.4 Object Detection: Face and Text

Identifying significant objects that appear in the video frames is one of the key components for video characterization. For the time being, we have chosen to deal with two of the more interesting objects in video: human faces and text (caption characters). To reduce computation we detect text and faces every 15th frame. Figure 4 shows examples face detection, illustrating the range of face sizes that can be detected, and examples of words and subsets of a word that are detected.

The "talking head" image is common in interviews and news clips, and illustrates a clear example of video production focussing on an individual of interest. A human interacting within an environment is also a common theme in video. The human-face detection system used for our experiments was developed by Rowley, Baluja and Kanade [6]. It detects mostly frontal faces of any size and any background. Its current performance level is to detect over 86% of more than 507 faces contained in 130 images, while producing approximately 63 false detections. While improvement is needed, the system can detect faces of varying sizes and is especially reliable with frontal faces such as talking-head images.

Text in the video provides significant information as to the content of a scene. For example, statistical numbers and titles are not usually spoken but are included in the captions for viewer inspection. A typical text region can be characterized as a horizontal rectangular structure of clustered sharp edges, because characters usually form regions of high contrast against the background. By
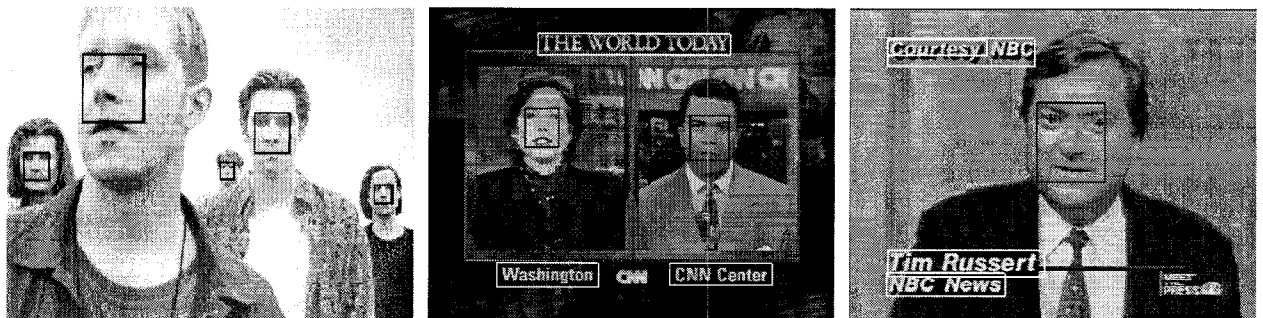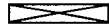

Figure 4: Detection of human-faces and text.

64

detecting these properties we extract regions from video frames that contain textual information. Figure 3 illustrates the process of detecting text; primarily, regions of horizontal titles and captions.

We first apply a 3x3 horizontal differential filter to the entire image with appropriate binary thresholding for extraction of vertical edge features. Smoothing filters are then used to eliminate extraneous fragments, and to connect character sections that may have been detached. Individual regions are identified by cluster detection and their bounding rectangles are computed. Clusters with bounding regions that satisfy the following constraints are selected:

Cluster Size > 70pixels

Cluster FillFactor ≥ 0.45

Horizontal – Vertical Aspect Ratio ≥ 0.75

A cluster's bounding region must have a large horizontal-to-vertical aspect ratio as well as satisfying various limits in height and width. The fill factor of the region should be high to insure dense clusters. The cluster size should also be relatively large to avoid small fragments. An intensity histogram of each region is used to test for high contrast. This is because certain textures and shapes appear similar to text but exhibit low contrast when examined in a bounded region. Finally, consistent detection of the same region over a certain period of time is also tested since text regions are placed at the exact position for many video frames. We can process a 352x240 image in less than 0.8 seconds on an SGI Indigo 2 workstation. Table 2 lists the detection results for various segments.

Table 2: Text Region Detection Results

Data (Images)	Regions Detected	Regions Missed	False Detections
CNHAV News (1056)	26	1	3
CNHAR News (1526)	48	0	5
Species I (264)	12	2	0
Planet Earth I-II(1712)	0	0	2

3 Technology Integration and Skim Creation

We have characterized video by scene breaks, camera motion, object appearance and keyphrases. Skim creation involves selecting the appropriate keyphrases and choosing a corresponding set of images. Candidates for the image portion of a skim are chosen by two types of rules: 1) Primitive Rules, independent rules that provide candidates for the selection of image regions for a given key-

phrase, and 2) Meta-Rules, higher order rules that select a single candidate from the primitive rules according to global properties of the video. The subsections below describe the steps involved in the selection, prioritizing and ordering of the keyphrases and video frames.

3.1 Audio Skim

The first level of analysis for the skim is the creation of the reduced audio track, which is based on the keyphrases. Those phrases whose total TF-IDF values are higher than a fixed threshold are selected as keyphrases. By varying this threshold, we control the number of keyphrases, and thus, the length of the skim. The length of the audio track is determined by a user specified compaction level.

Keyphrases with words that appear in close proximity or repeat throughout the transcript may create skims with redundant audio. Therefore, we discard keyphrases which repeat within a minimum number of frames (150 frames) and limit the repetition of each keyword in a keyphrase.

3.2 Video Skim Candidates

In order to create the image skim, we might think of selecting those video frames that correspond in time to the audio skim segments. As we often observe in television programs, however, the contents of the audio and video are not necessarily synchronized. Therefore, for each keyphrase we must analyze the characterization results of the surrounding video frames and select a set of frames which may not align with the audio in time, but which are most appropriate for skimming.

To study the image selection process of skimming, we manually created skims for 5 hours of video with the help of producers and technicians in Carnegie Mellon's Drama Department. The study revealed that while perfect skimming requires semantic understanding of the entire video, certain parts of the image selection process can be automated with current image understanding. By studying these examples and video production standards [13], we can identify an initial set of heuristic rules.

The first heuristics are the primitive rules, which are tested with the video frames in the scene containing the keyword/keyphrase, and the scenes that follow within at least a 5 second window. The four rows above "Skim Candidates", in Figure 5, indicate the candidate image sections selected by various primitive rules. A description of each primitive rule is given in order of priority below.

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
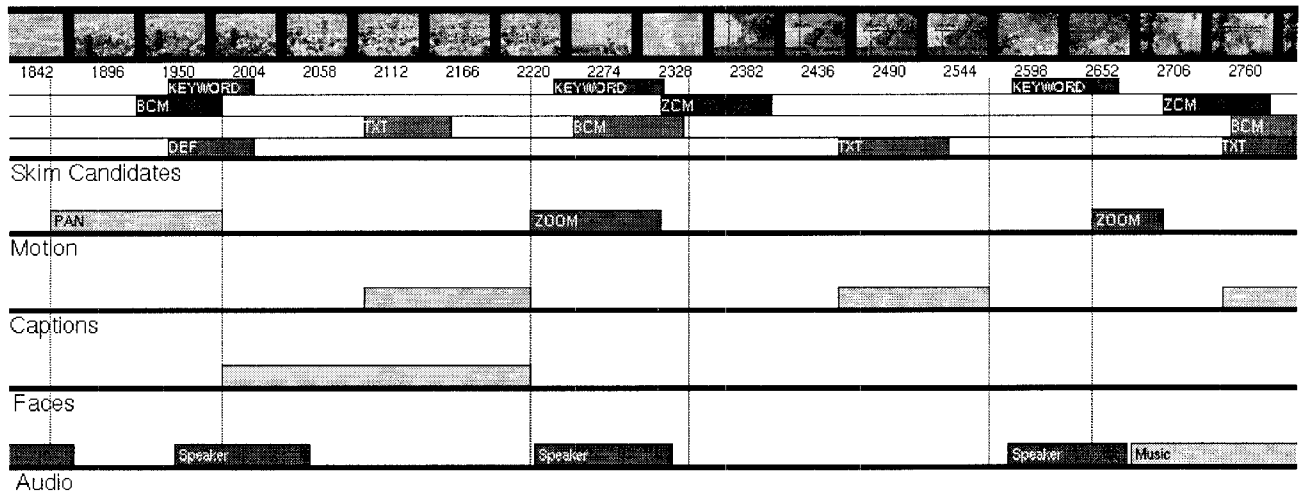
Figure 5: Characterization data with skim candidates and keyphrases for "Destruction of Species". The skim candidate symbols correspond to the following primitive rules: BCM, Bounded Camera Motion; ZCM, Zoom Camera Motion; TXT, Text Captions; and DEF Default. Vertical lines represent scene breaks.

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.

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 just before areas of high audio levels and after areas of music.

8.Grayscale Video (GRY)

Grayscale video is often used to provide historical perspective. If the ratio of grayscale to color video is low for a particular segment, image regions containing grayscale video are selected for skimming. If this ratio is high, the segment is likely composed entirely of grayscale video and this rule does not apply.

9. Default Rule(DEF)

As a default, video frames will 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
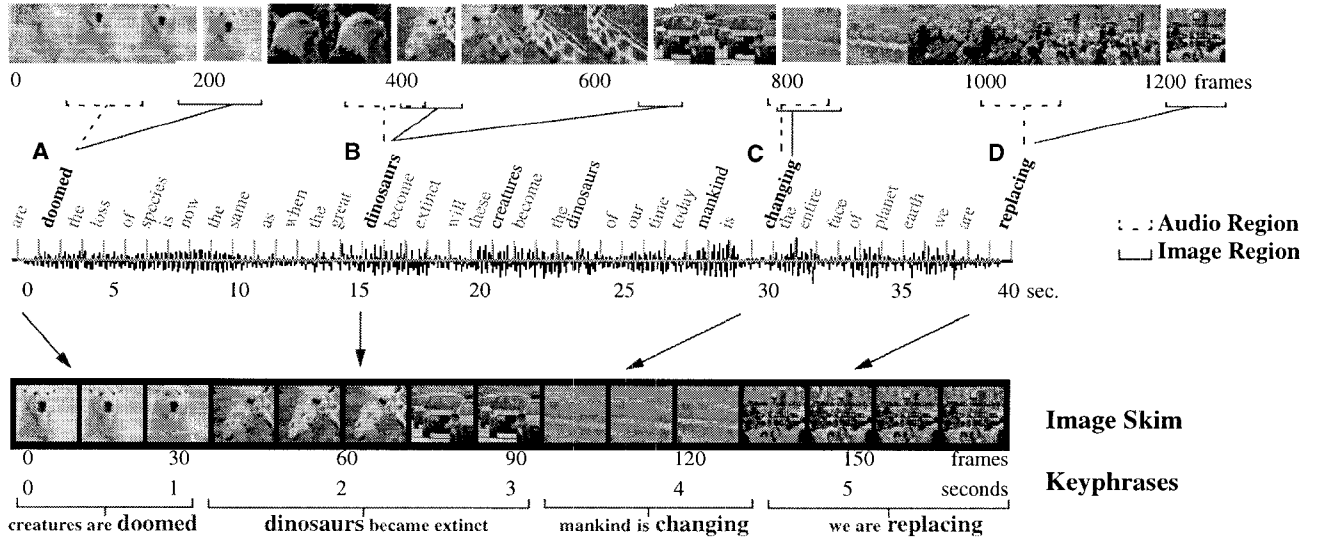
Figure 6: Skim creation incorporating word relevance, significant objects (humans and text), and camera motion: A) For the word "doomed", the images following the camera motion are selected, B) The keyphrase for "dinosaur" is long so portions of the next scene are used for more content, C) No significant structure for the word "changing", D) For the word "replacing" The latter portion of the scene contains both text and humans.

Meta-Rules are used to complete skim creation.

For visual clarity and comprehension, we allocate at least 60 video frames (2 seconds) to a keyphrase. The 60 frame minimum for each scene is based on empirical and user studies of visual comprehension in short video sequences. When a keyphrase is longer than 120 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 60 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 faces with respect to the total segment length. If a scene containing a face is very long, such as an interview, we look for other image candidates in an extended 10 second search window.

Figure 6 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.

3.4 Example Results

Figure 7 shows two types of skims for the "Mass Extinction" segment. The ALL skim was produced with

our method of integrated image and language understanding. The DEF skim was created by selecting video and audio portions at fixed intervals. This segment contains 71 scenes, of which, the ALL skim has captured 23 scenes, and the DEF skim has captured 17 scenes. Studies involving different skim creation methods are discussed in the next section.

The ALL skim 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.

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 automated 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 have conducted two user studies to test video skims. In the first study, we used various types of skims in two tasks: 1. Content summarization and 2. Browsing for Fact-

ALL

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

DEF

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
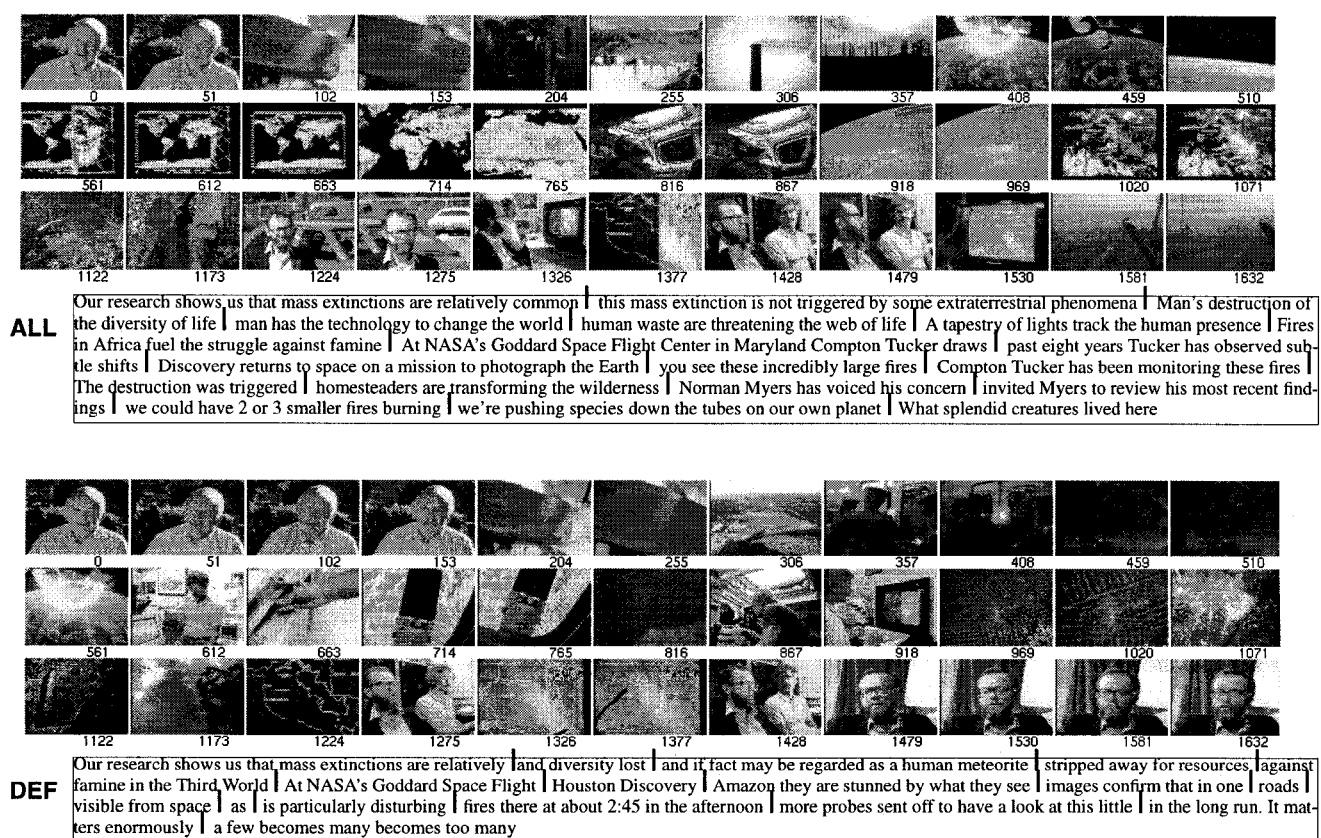
Figure 7: Image and audio output for the "Mass Extinction" segment: ALL) Skim creation using image and language understanding, DEF) Skim creation using fixed intervals for image and audio.

Finding in a video library. The following skim creation schemes were tested:

ALL   -   Image and Language Characterization
DEF   -   Fixed Intervals of 72 frames (Default)
AUD   -   Language Characterization Only
IMG   -   Image Characterization Only

Figure 7 shows examples of the ALL and DEF skim. The visual information in the ALL skim is less redundant and provides a greater variety of scenes. The audio for the DEF skim is incoherent and considerably smaller.

The first test was the Fact-Finding exercise in which a subject would search a video segment with a play-skim toggle and answer a series of questions. The effectiveness of each skim was based on the time to complete this task and the number of correct items retrieved. The results from the Fact-Finding test suggest that users can answer questions at roughly the same speed and accuracy, regardless of the skim type. However, users spent more time

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 Characterization
AC - Automated Characterization
MS - Manual Skim Creation
AS - Automated Skim Creation

viewing the processed skims (ALL, AUD, and IMG) and their comments indicate that the default skim was more frustrating to watch. The second task was an exercise in summarization. Users would view a skim and then answer a series of questions. The difference in scores was again very small, but the processed skims received the highest scores.

Although the first study found no significant differences between our processed skims and the default skim, we found several areas for improvement. There was more emphasis placed on the audio portion of the skim, as discussed in section 2.1. The average size of a single key-phrase was increased from 5 to 7 seconds and the minimum image search window was reduced from 15 to 5 seconds. We tested longer audio segments on the order of 30 minutes in duration, as opposed to 10 minutes for the first study. In the second user-study we focused solely on video summarization as the task. The following skim creation schemes were tested:

    NEW - Enhanced Audio Skim
    NRI - Same as NEW with Reordered Images
    DFL - Default Long (144 frames)
    DFS - Default Short (72 frames)
    FULL - Original Segment

The NRI skim was included to test the utility of a skim with little audio and image correspondence. We tested two types of default skims to test for differences in audio phrase duration. In this study we tested the full video along with the various skim videos. After viewing a particular treatment, the subjects were tested with text and image questions. Analysis revealed highly significant differences (p < 0.01) in mean performance on text summarization and image recall among the five video treatments. The FULL segment and NEW skim received the highest scores. An informal classification of 59 open-ended comments offering a favorable or critical opinion produced the distribution shown in Figure 8.

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.

Our user-studies have shown that the improvements to the skim provided a more pleasing video presentation and higher visual and audio retention over the default skims. The improvements in the audio show its importance in skim creation. Further testing is planned with other types

of video, such as, feature-films and broadcast news.

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.
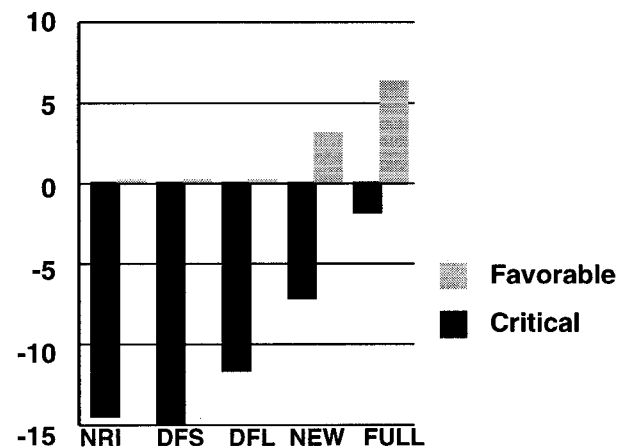
Acknowledgments

Figure 8: Count of open-ended comments by treatment.

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," *CVPR*, San Francisco, May 1996.

[7] Wactlar, H., *et al.* "Intelligent Access to Digital Video: The Informedia Project" IEEE Computer, Vol. 29(5), May 1996.

[8] Zhang, H., *et a.l*, "Automatic Partitioning of Full-Motion Video," *Multimedia Systems* 93 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] "MPEG-1 Video Standard", *Comm. of the ACM*, April 1991.

[13] Smallman, K., "Creative Film-Making", 1st ed., Publisher Macmillan, New York 1970.

[14] Wactlar, H., Hauptmann, A., Smith, M., and Pendyala, K. Garlington, D. "Automated Video Indexing of Very Large Video Libraries." *SMPTE Journal*, August, 1997.

[15] Shahraray, B., Gibbon, D., "Authoring of Hypermedia Documents of Video Programs," *Proc. Third ACM Conference on Multimedia*, pp. 401-409, San Francisco, CA, November, 1995.

[16] Yeung, M., Yeo, B.L., Wolf, W., and Liu B. "Video Browsing Using Clustering and Scene Transitions on Compressed Sequences," *Multimedia Computing and Networking 1995*, Proc. SPIE, Vol. 2417, San Jose, CA, February, 1995.

[17] Pfeiffer, S. Lienhart, R., Fischer, S., Effelsberg, W., "Abstracting Digital Movies Automatically," *Journal of Visual Communication and Image Representation*, Vol. 7, No. 4, pp. 345-353, December, 1996.