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

**Michael A. Smith and Takeo Kanade**

February 3, 1997

CMU-CS-97-111

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**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.

This research is sponsored by the National Science Foundation under grant no. IRI-9411299, the National Space and Aeronautics Administration, and the Advanced Research Projects Agency. Michael Smith is sponsored by AT&T Bell Laboratories. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either expressed or implied, of the United States Government or AT&T Bell Laboratories.

**Keywords:** video skimming, audio skim, image skim, keyphrases, characterization, integrated technology, video compaction

## 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™ Project at Carnegie Mellon [7], 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.

### 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]. 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.

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 the synchronized portion corresponding to the selected video frames.

### Video Skims

Figure 1 illustrates the concept of extracting the most representative video frames and audio information to create the skim. 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 lan-

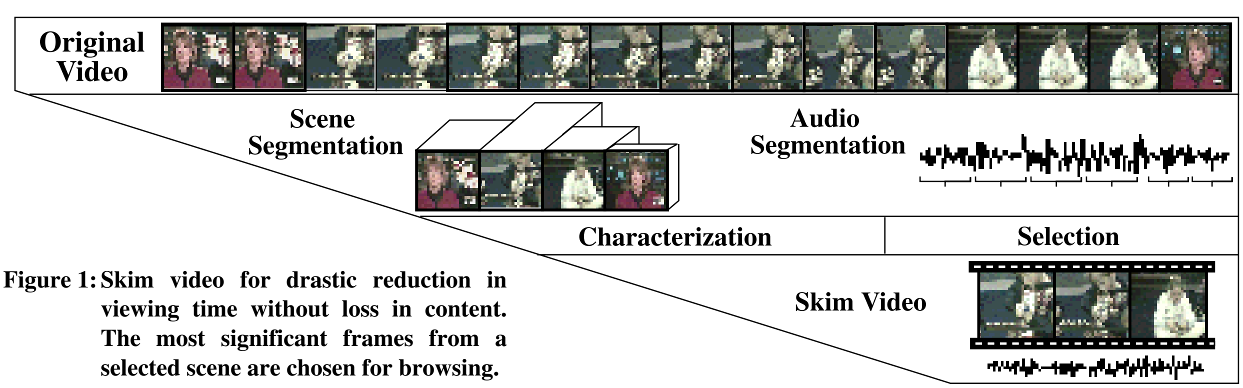



**Figure 1: Skim video for drastic reduction in viewing time without loss in content. The most significant frames from a selected scene are chosen for browsing.**

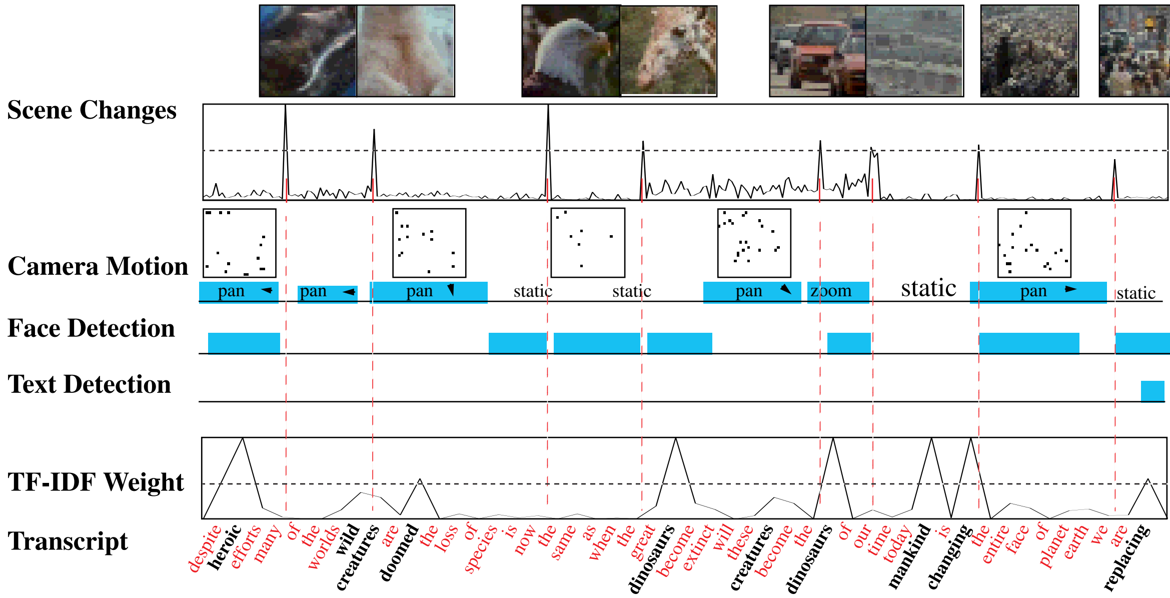



**Figure 2: Video Characterization Technology. Video is segmented into scenes, and camera motion is detected along with significant objects (faces and text). Bars show frames with positive results.**

guage 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 2 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 Language Characterization

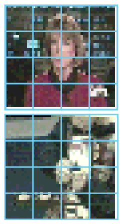
Language analysis works on the transcript to identify important audio regions known as “keywords”. We use the well-known technique of TF-IDF (Term Frequency Inverse Document Fre-

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

quency) 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 from the TF-IDF weights, as shown in the bottom rows of Figure 2.

## 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,



$$D(t) = \sum_{v=0}^{N} \left| H_t(v) - H_{t+1}(v) \right| \tag{2}$$

$H_t(v)$ : Histogram of Color in Image($t$)

video sequences are separated into scenes. Peaks in the difference, *D(t)*, are detected and an empirically set threshold is used to select scene breaks. We have found that 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). Examples of segmentation results are shown in the top row of Figure 2.

## 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 [13]. An affine model is used to approximate the flow patterns consistent with all types

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

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

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 $(x_0, y_0)$ does not exist and 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. Table 1 shows the statistics for detection

**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 |

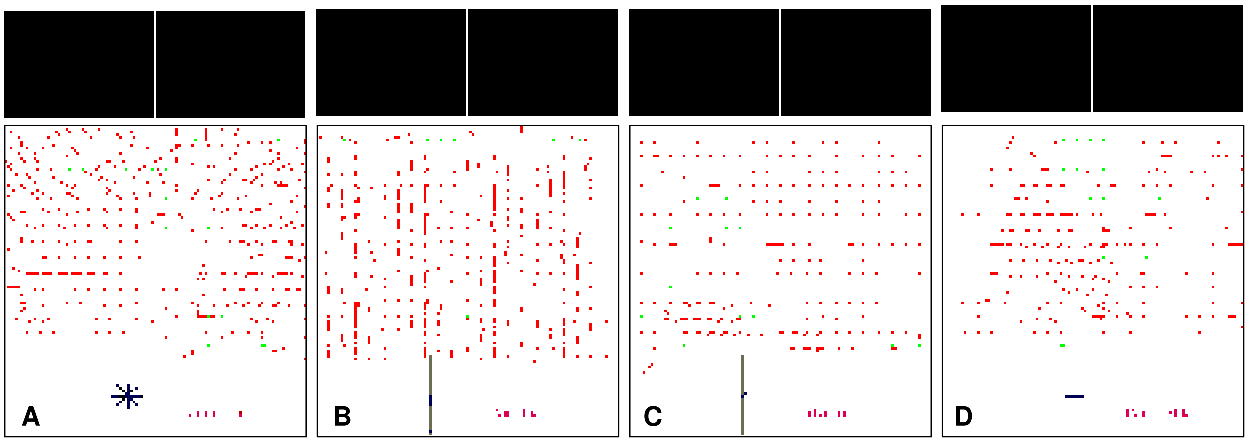



**Figure 3: Camera motion analysis from MPEG motion vectors: A) Zoom distribution, B) Upward pan with subtle object motion, C) Static, D) Significant object motion detected as pan.**

on various sets of images. Regions detected are either pans or zooms. Examples of the camera motion analysis results are shown in Figure 3.

## 2.4 Object Detection

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.

### Face Detection

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. Figure 4 shows examples of its output, illustrating the range of face sizes that can be detected.

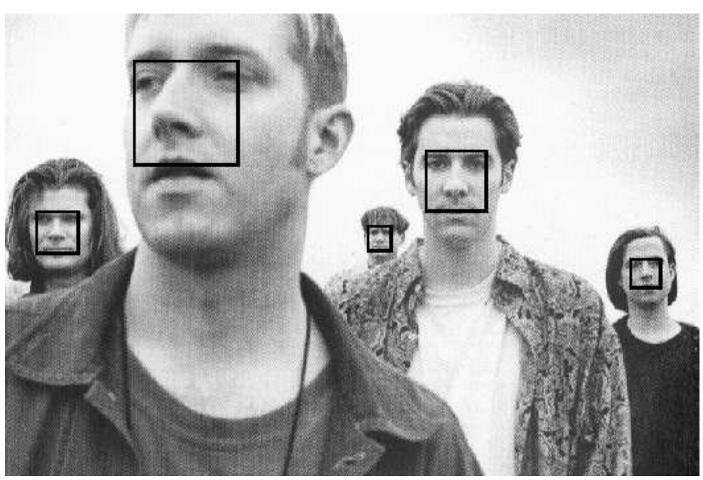


**Figure 4: Detection of human-faces.**

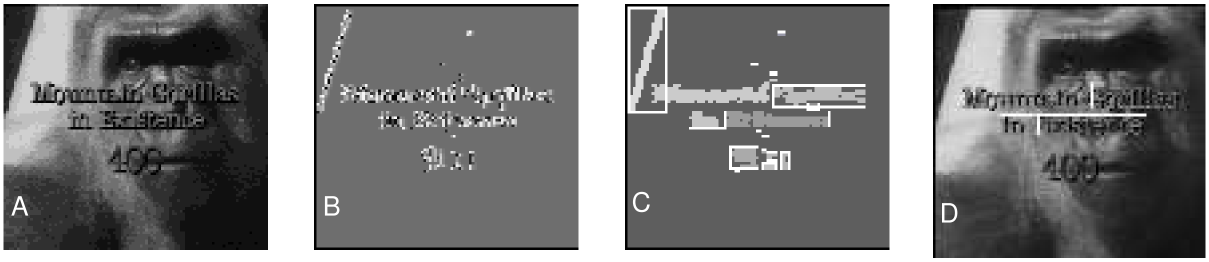



**Figure 5: Stages of text detection: A) Input, B) Filtering, C) Clustering, and D) Region Extraction.**

### Text Detection

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 detecting these properties we extract regions from video frames that contain textual information. Figure 5 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 > 70 pixels
Cluster Fill Factor ≥ 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. Figure 6 shows detection examples of words and subsets of a word. Table 2 presents statistics for detection on various sets of images.

**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 keywords. Skim creation involves selecting the appropriate keywords and choosing a corresponding set of images. Candidates for the image portion of a skim are chosen by two types of rules: 1)


**Figure 6: Text detection results with various images.**

Primitive Rules, independent rules that provide candidates for the selection of image regions for a given keyword, 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 keywords 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 keywords. Those words whose TF-IDF values are higher than a fixed threshold are selected as keywords. By varying this threshold, we control the number of keywords, and thus, the length of the skim. The length of the audio track is determined by a user specified compaction level.

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

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 is obtained by starting with a keyword, and extending its boundaries to areas of silence or neighboring keywords. Each keyphrase is isolated from the original audio track to form the audio skim. The average keyphrase lasts 2 seconds.

### 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 keyword or 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 [14], 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. A description of each primitive rule is given in order of priority below. The four rows above

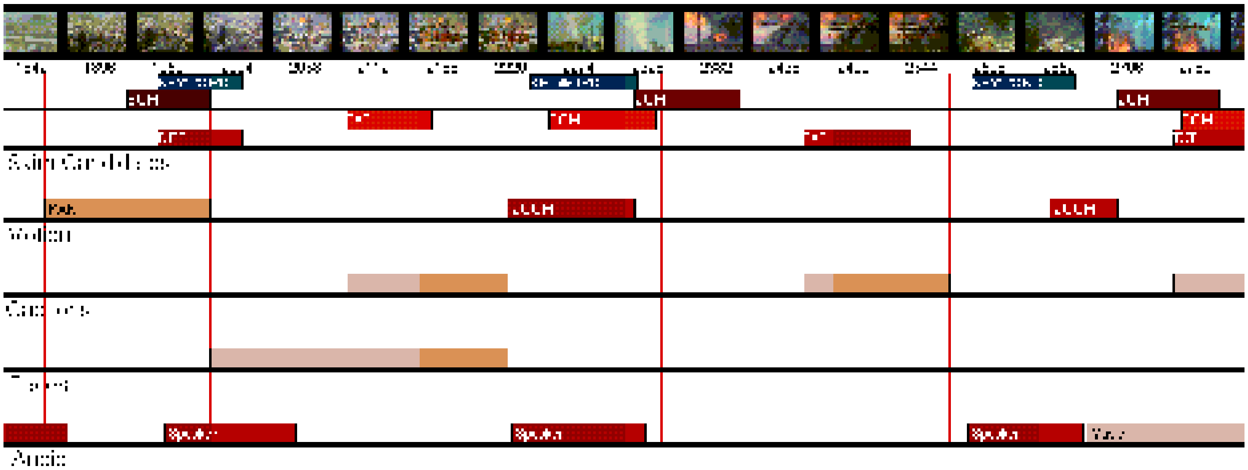


**Figure 7:Characterization data with skim candidates and keyphrases for "Destruction of Species". The skim candidate symbols correspond to the following primitive rules: BCM, Bounded Camera Motion; ZCM, Zoom Camera Motion; TXT, Text Captions; and DEF Default. Vertical lines represent scene breaks.**

"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.

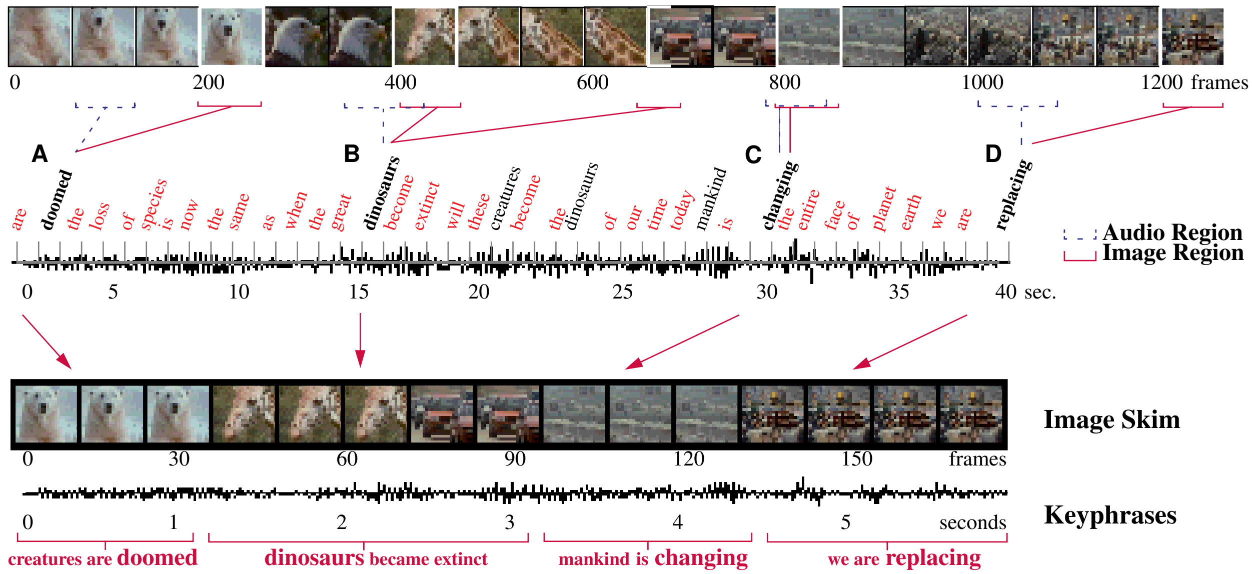



**Figure 8: Skim creation incorporating word relevance, significant objects (humans and text), and camera motion: A) For the word "doomed", the images following the camera motion are selected, B) The keyphrase for "dinosaur" is long so portions of the next scene are used for more content, C) No significant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

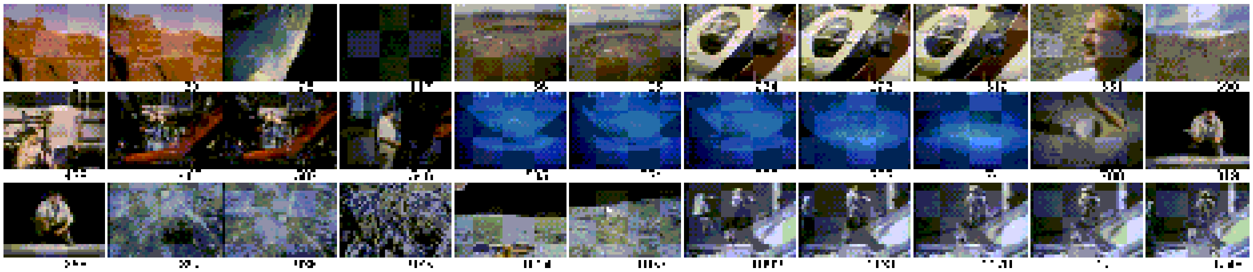
### 3.3 Image Adjustments

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.



**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

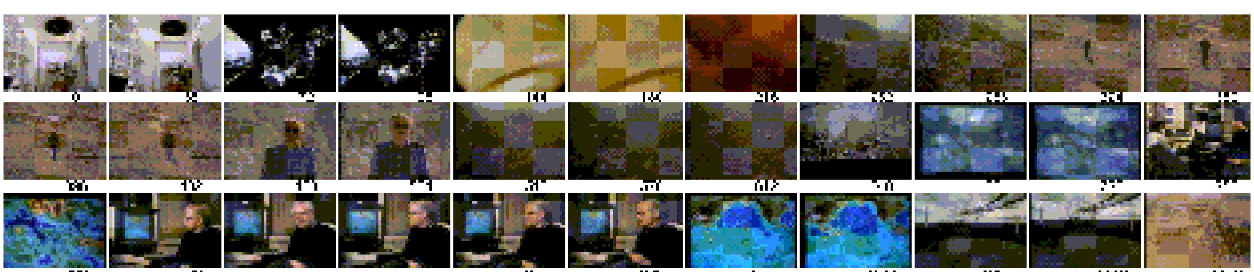
**Figure 9: Skim video frames and keyphrases for "Planet Earth - I" (10:1 compaction).**

### 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).**

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

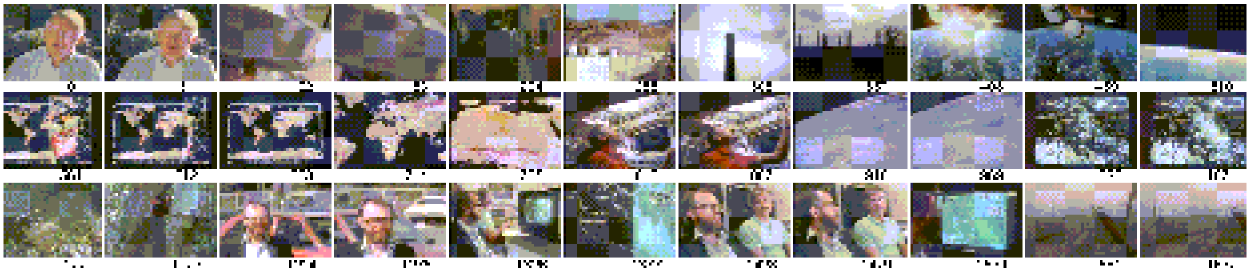
The results of several skims are summarized in Table 3. The manually created skims in the initial stages of the experiment help test the potential visual clarity and comprehension of skims. The compaction ratio for a typical segment is 10:1; and it was shown that skims with compaction as high as high 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.

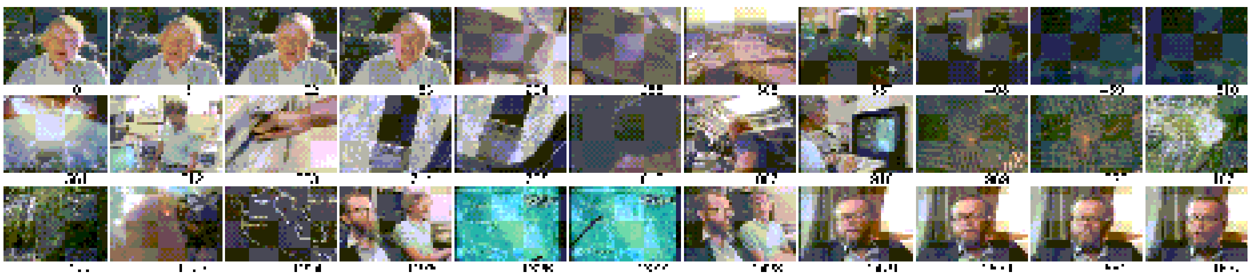
We use various types of skims to test the utility of image and language understanding in skim creation. The following creation schemes are presently being tested:

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.



**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.

[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 AT&T Bell Laboratories.