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

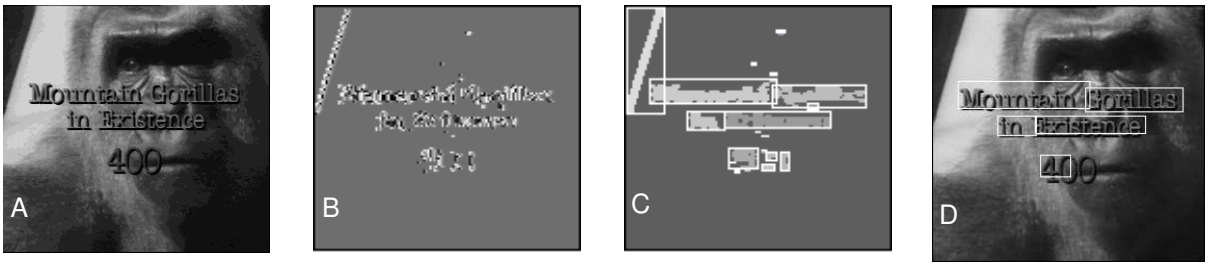
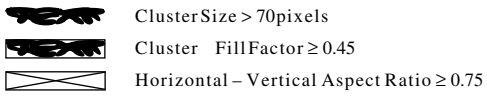


Figure 5: Stages of text detection: A) Input, B) Filtering, C) Clustering, and D) Region Extraction.

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:



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)

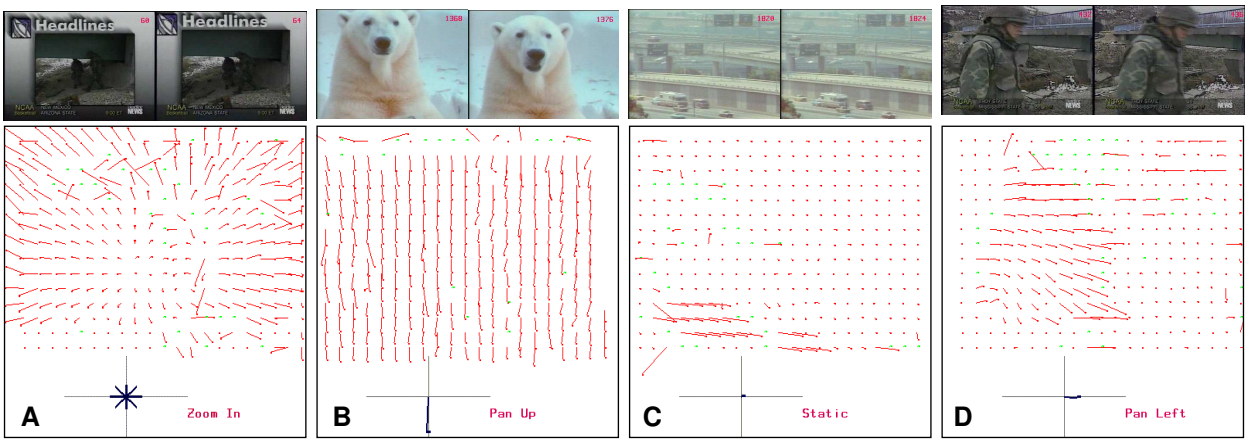


Figure 3: Camera motion analysis from MPEG motion vectors: A) Zoom distribution, B) Upward pan with subtle object motion, C) Static, D) Significant object motion detected as pan.

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.

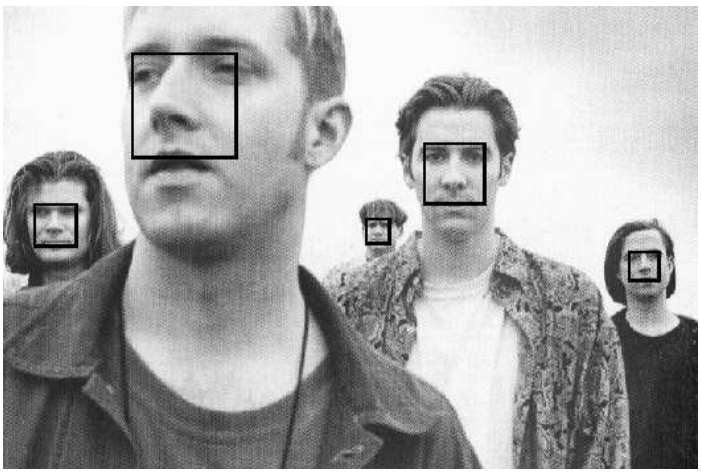
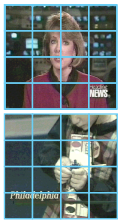


Figure 4: Detection of human-faces.

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,



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

$H_t(v)$: 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

$$u(x_i, y_i) = ax_i + by_i + c \qquad (3)$$

$$v(x_i, y_i) = dx_i + ey_i + f \qquad (4)$$

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 $u(x_i, y_i) = 0$ and $v(x_i, y_i) = 0$. 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

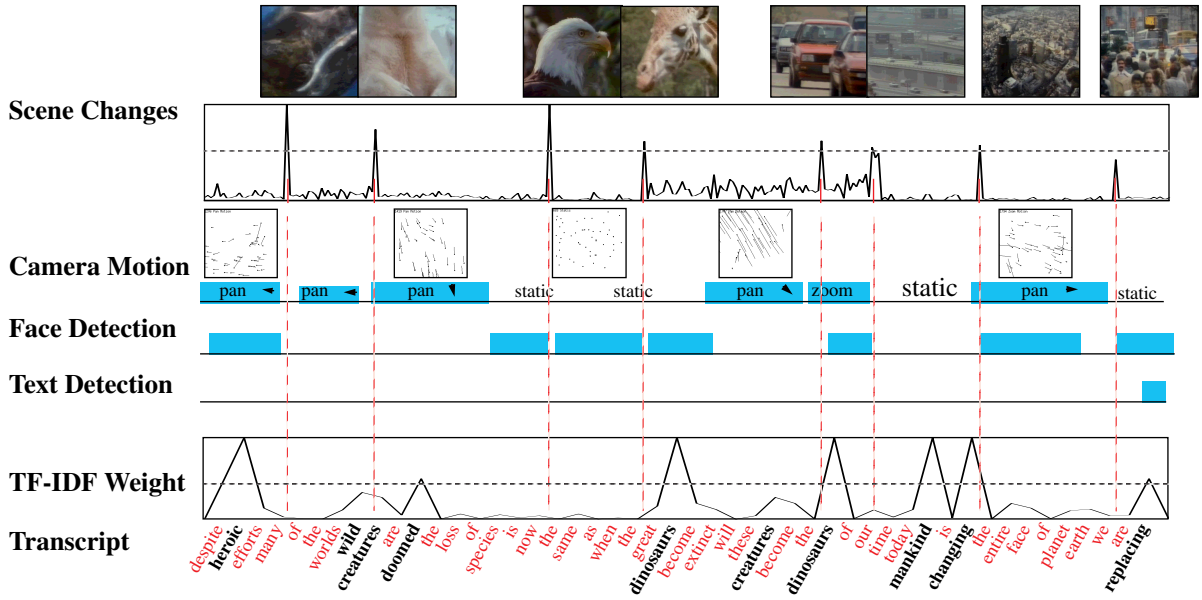


Figure 2: Video Characterization Technology. Video is segmented into scenes, and camera motion is detected along with significant objects (faces and text). Bars show frames with positive results.

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

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

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-

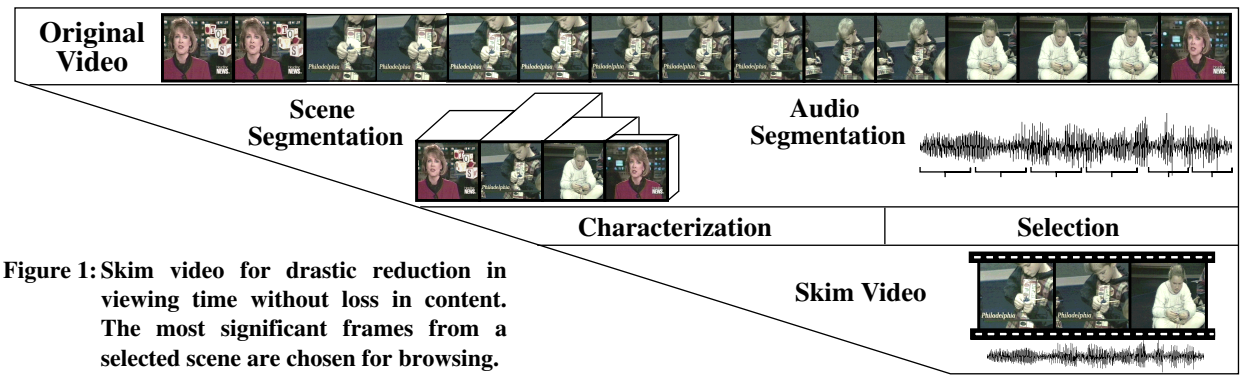


Figure 1: Skim video for drastic reduction in viewing time without loss in content. The most significant frames from a selected scene are chosen for browsing.

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

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 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 Bell Laboratories.