[15] "TREC 93," *Proceedings of the Second Text Retrieval Conference*, D. Harmon, editor, sponsored by ARPA/SISTO, August 1993.

[16] Tonomura, Y., Akutsu A., Taniguchi, Y., and Suzuki, G. "Structured Video Computing," *IEEE Multimedia Magazine*, Fall 1994, pp. 34-43.

7 Acknowledgment

The authors would like to thank Henry Rowley and Shumeet Baluja for providing the routines for face detection; and Michael Mauldin for providing the routines for keyword selection. This work is partially funded by the National Science Foundation, the National Space and Aeronautics Administration, and the Advanced Research Projects Agency.

6 References

[1] Akutsu, A. and Tonomura, Y. “Video Tomography: An efficient method for Camerawork Extraction and Motion Analysis,” *Proc. of ACM Multimedia ‘94*, Oct. 15-20, 1994, San Francisco, CA, pp. 349-356.

[2] Arons, B. “SpeechSkimmer: Interactively Skimming Recorded Speech,” *Proc. of ACM Symposium on User Interface Software and Technology (UIST)’93*, November 3-5, 1993, Atlanta, GA, pp. 187-196.

[3] Degen, L., Mander, R., and Salomon, G. “Working with Audio: Integrating Personal Tape Recorders and Desktop Computers,” *Proc. CHI ‘92*, May 1992, Monterey, CA, pp. 413-418.

[4] Hampapur, A., Jain, R., and Weymouth, T. “Production Model Based Digital Video Segmentation,” *Multimedia Tools and Applications* **1** March 1995, pp. 9-46.

[5] Lucas, B.D., Kanade, T. “An Iterative Technique of Image Registration and Its Application to Stereo,” *Proc. 7th International Joint Conference on Artificial Intelligence*, pp. 674-679, Aug. 1981.

[6] Mauldin, M. “Information Retrieval by Text Skimming,” PhD Thesis, Carnegie Mellon University. August 1989. Revised edition published as “Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing, Kluwer Press, September 1991.

[7] Rowley, H., Baluja, S. and Kanade, K. “Human Face Detection in Visual Scenes,” Carnegie Mellon University, 1995. Computer Science Technical Report CMU-CS-95-158.

[8] Salton, G., and McGill, M.J. “Introduction to Modern Information Retrieval,” McGraw-Hill, New York, McGraw-Hill Computer Science Series, 1983.

[9] Stevens, S., Christel, M., and Wactlar, H. “Informedia: Improving Access to Digital Video”. *Interactions* 1 October 1994, pp. 67-71

[10] Tomasi, C., and Kanade, T. “Shape and Motion without Depth,” *ICCV 90*, Osaka, Japan.

[11] Zhang, H., Kankanhalli, A., and Smoliar, S. “Automatic partitioning of full-motion video,” *Multimedia Systems* 1993 1, pp. 10-28.

[12] Zhang, H., Low, C., and Smoliar, S. “Video parsing and indexing of compressed data,” *Multimedia Tools and Applications* 1 March 1995, pp. 89-111.

[13] Arman, F., Hsu, A., and Chiu, M-Y. “Image Processing on Encoded Video Sequences,” *Multimedia Systems* 1994 1, pp. 211-219.

[14] Arman, F., Depommier, R., Hsu, A., and Chiu, M-Y. “Content-Based Browsing of Video Sequences,” Proc. of ACM Multimedia ‘94, October 15-20, 1994, San Francisco, CA pp. 97-103.

4 Discussion

The skims have shown to provide adequate descriptions of full-length video segments in a relatively short time span without losing the essential content. Without any prior knowledge of the scenes, most users can interpret the content of figures 13b, and 14b. The actual video skim is even more informative during playback.

The final length of the skim is completely dependent on the user. The compaction level can be set to include information from as many or as few of the selected scenes as needed. For browsing of multiple segments, the amount of video needed to capture the content is typically very small.

The first testbed for the video skim will be a local K-12 school and the undergraduate community at Carnegie Mellon during the first release of the Informedia Library. From this, we hope to gain practical knowledge as to the effectiveness of the video skim as a browsing tool.

All video is processed with images digitized from VHS quality data. We are currently modifying the system to work with MPEG compressed data. It has been shown that some image analysis on encoded data can be more efficient and just as accurate as still image analysis [12], [13]. With the use of MPEG video, we can eliminate much of the overhead used in detecting scene breaks and camera motion. Monitoring the DCT coefficient can serve as an effective means to detect scene breaks. During encoding, subregions within each image are tracked over time. The resulting vectors accurately depict optical flow. This information is embedded in the compressed video and can be accessed with little computation. With extended work in optical flow analysis we will eliminate unnecessary computation by analyzing only the foreground objects of interest.

Audio segmentation is currently a manual process which will be automated. Since we only use individual words, the audio is fragmented and somewhat incomprehensible for some speakers. We will extend the language analysis to improve the audio skim segments.

We will broaden our scope of object detection to include outdoor and indoor scenes, synthetic and natural objects, and other items of interest such as automobiles, buildings, and animals. The ultimate goal of the detection technology is true semantic characterization of video images.

At present our selection rules are based purely on empirical tests. A film producer will often follow a set pattern in deciding which frames to use as the focus. Scene selection rules based on actual production standards may be a more accurate method to select significant video.

5 Conclusion

The emergence of high volume video libraries has shown a clear need for content specific video browsing technology. We have described an algorithm to automatically create video browsing data that incorporates content specific audio and video information. By viewing only the skimmed video segments, the content of hours of video is reduced into minutes. While this generation of content-based skims is still primitive and much room remains for improvement, it illustrates the potential power of integrated speech, language, and image information for characterization in video retrieval and browsing applications.

Figure 14b: Skim for the test set, “Destruction of Species”. Frames are displayed at 7.5 fps. The length of the above skim is 9.6 seconds. Total time of corresponding original segment is 56.35 seconds.

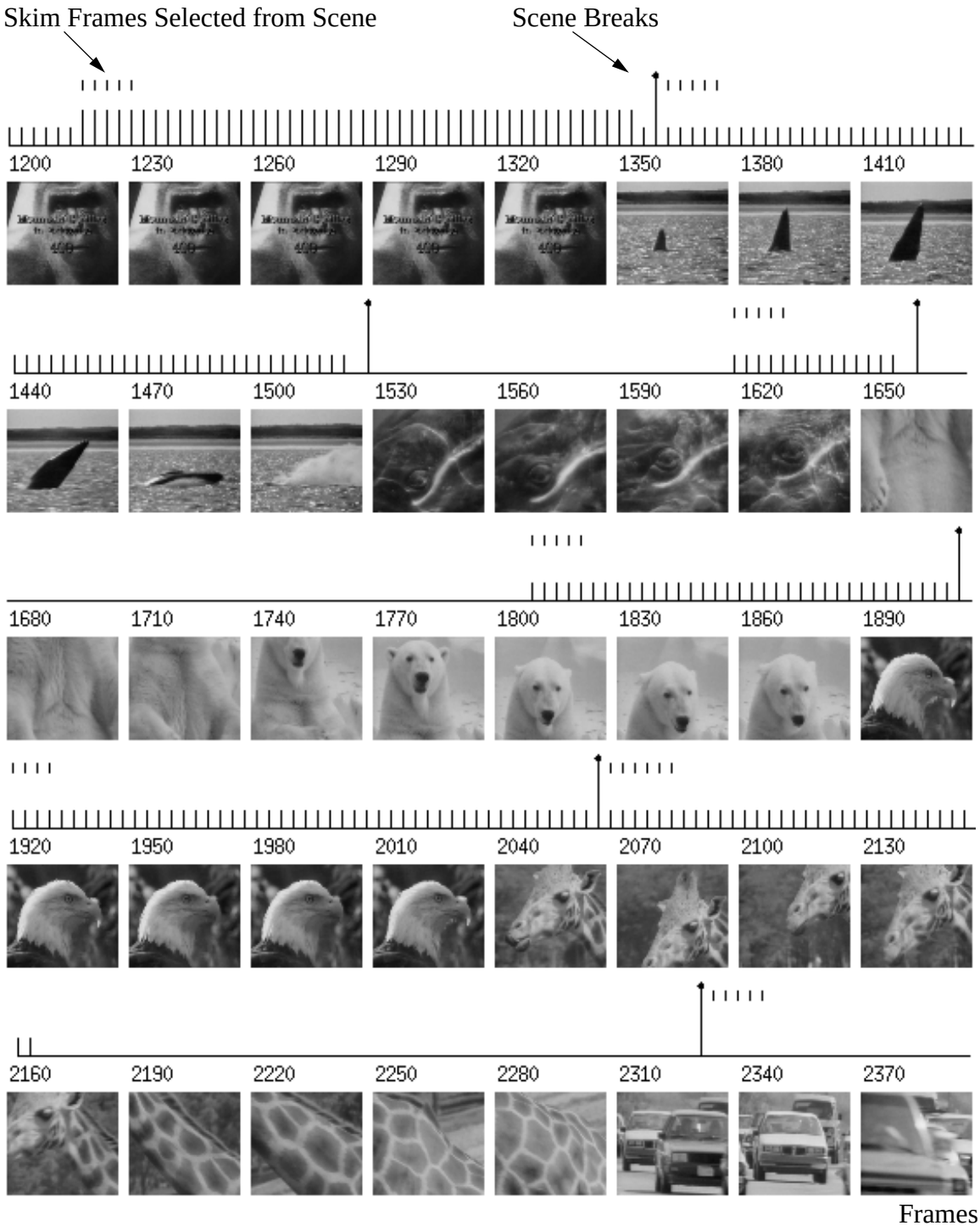


Figure 14a (cont.): Skim frame selection continued from figure 14b. Note the frames selected from the polar bear scene follow camera motion. Some scenes for this set contain no interesting motion or objects so we select the initial frames of the scene for the skim.

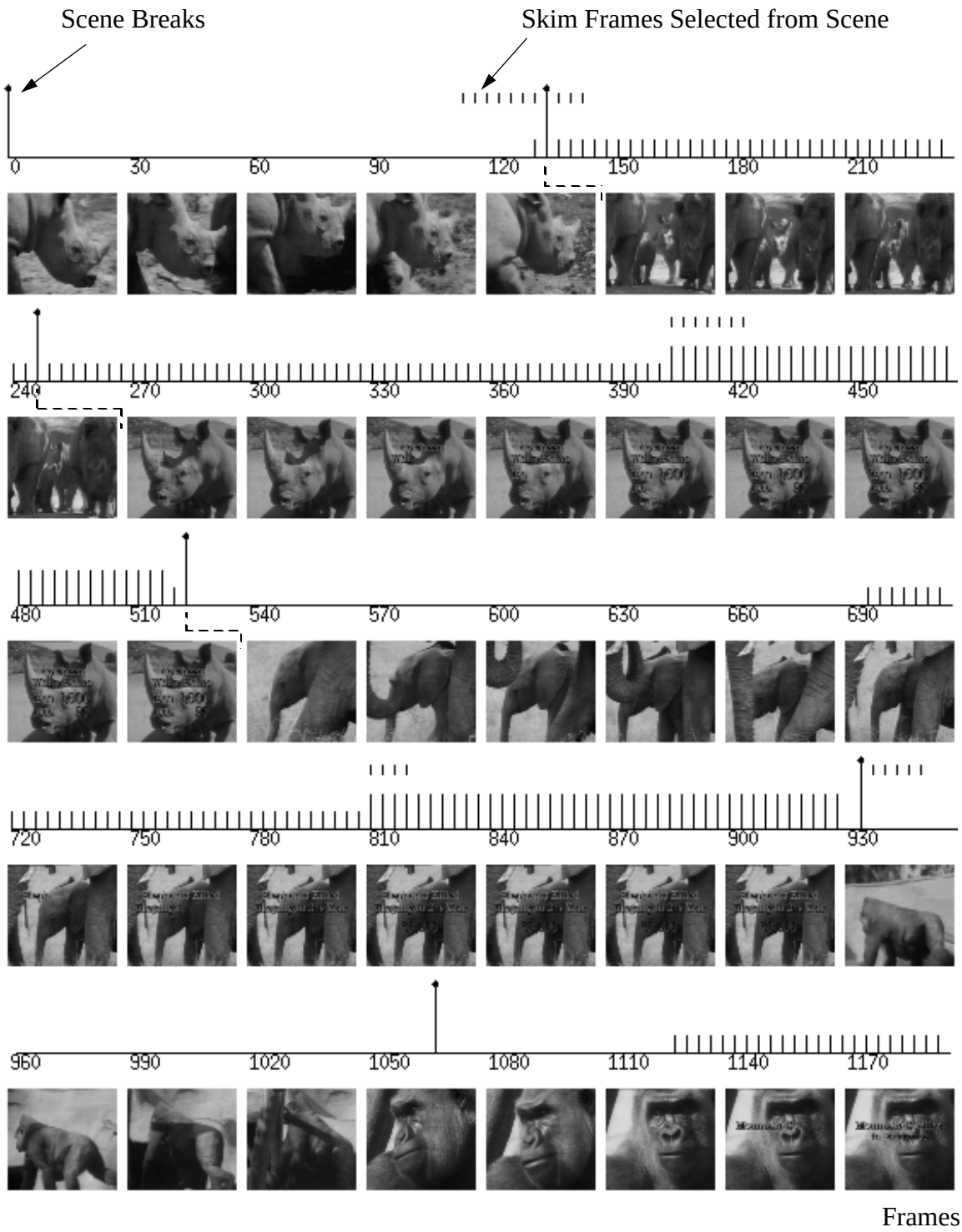


Figure 14a: Skim frame selection. Many frames contain captions, however, extraction peaks when the text content is at its highest. This can be seen in the rhinoceros (frames 400 - 510) and elephant (frames 810-930) scenes which contain titles. The first scene is primarily a panning sequence except for the final frame (125) which is somewhat static and used for skimming. Frames are displayed 1.0 fps.

Figure 13b: Skim video frames and audio keywords for the test set, "K'nex Toy". The word "toy" appears often in this segment and thus has a high TF-IDF weight. Frames with faces and captions have the highest priority. Frames are displayed at 6.0 fps. The length of this skim is 11.33 seconds with the original segment consisting of 61 seconds.

Figure 13a: Scene breaks with skim regions for each scene: (a) The highest level indicates human presence, captions and no camera motion; (b) The second level is for humans or captions in static frames or frames following camera motion; (c) The third level indicates static frames only. The number of frames for a skim scene correspond to the length of the corresponding keyword. Frames are displayed 1.5 fps.

Table 2: Skim Compaction Results

Video Segments	Original Video Length (seconds)	Skim with All Scenes	Skim with Select Scenes
K'nex Toy	61.0	11.33	7.13
Species Destruction(short)	68.65	9.83	6.40
* Species Destruction(full)	123.23	NA	12.43
* Space University	166.20	NA	28.13
* Rain Forest	107.13	NA	5.36
* Peru Forest Destruction	58.13	NA	5.30
* Underwater Exploration	119.50	NA	5.67

* Manual Skims

By limiting the number of keywords, we select which scenes to include in the skim. The level of compression determines the number of words in the audio skim, and thus, the number of scenes included. This level is typically set to 10:1, although levels as high as 20:1 have shown to offer sufficient comprehension.

3.4 Example Results

We have tested the automated skim on various videos. The results of two examples are shown below. Although the detection technology is extremely accurate, the face and text detection results have been corrected for these skims.

Figure 13a shows the process of selecting skim frames for each scene from the "K'nex toy" video, CNN Headline News. The number of frames selected for each scene correspond to the word length of the keyword selected from that scene. Frames with human-faces, text, and static frames are the most significant. The frames which contain faces and text have higher priority than frames with only faces, as seen in example (a), of figure 13a. Figure 13b shows the complete skim, with frames from all scenes, and the associated keyword. Although we limit the repetition of keywords in a skim, there is often a need to display a word more than once, as seen in Figure 13a with the word "toy". The subject of the segment is a new toy and thus the word appears quite often in the transcript creating an extremely high TF-IDF weight.

Figure 14a shows the process of selecting skim frames for each scene from the "Destruction of Species" video, WQED Pittsburgh. Although many of the frames contain captions, frames with the most text in a given scene received the highest priority. In the rhinoceros, elephant, and primate scenes, we see that frames with full captions have higher priority than the previous frames which contain only limited text. Even though our detection is limited to text and humans, we see a clear need for other methods of detection, such as animals and land structures. Figure 14b shows the complete skim for this video. Although the keyword "dinosaur" appears twice in the transcript, it's relative TF-IDF weight is not high enough to allow its presence in multiple scenes. The word "changing" actually appears at a false scene break. Although segmentation may fail with abrupt movement, the change is usually so significant that the visual information displayed is not similar to the previous frames. There is constant motion and no recognizable objects throughout the scenes which contain the words "protected" and "mankind". This is the default case so we simply use the initial frames of the scene for the skim. Since the skims in figures 13b and 14b are displayed at a relatively low number of frames per second, no additional frames from adjacent scenes are used with long words, such as "unusual" and "dinosaur".

The compaction results of several automatic and manual skims are shown in Table 2. We manually created skims for 5 hours of video in the initial stages of the experiment to test for visual clarity and comprehension. For some of the examples below, the pre-set compaction ratio is as high as 20:1.

With the extraction of frames from each scene, we now have a suitable representation for the image skim. We place particular importance on frames with captions and human faces (Example d) in Figure 10). For many scenes, camera motion will preceed frames of importance (Example a) in Figure 10). This example of desirable static frames and frames that follow excessive

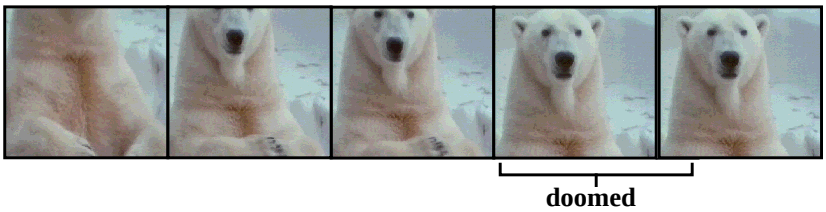


Figure 12: Skim frame selection based on minimal camera motion for the keyword "doomed". The static region is typically the focus of the scene. In this example, the initial frames offer little information as to the content of the scene.

camera motion is shown in detail in figure 12. The process of ordering the image and audio skims for the final skim video is described in the next section.

3.3 Skim Selection and Creation:

The final skim scenes are selected by analyzing word relevance and the structure of the prioritized audio and image skims. Several heuristic rules have been developed for the final selection and ordering of the video skim depending on various conditions, such as the duration of the words, scene contents and previously selected frames. The number of scenes used in the final skim depends on the compression rate. These scenes are selected according to the following constraints:

1: Final skim length, l_s, is computed from the skim compression rate, r_c, and the original video length, l_v.

$$l_s = r_c \times l_v \tag{6}$$

2: We select as many skim scenes needed to fill l_s by appropriately setting the threshold for allowable keywords.

3. The number of skim scenes with consecutive talking heads is limited to three.

4. We avoid keywords that repeat or appear in close proximity.

To avoid redundancy in the skim playback we reduce the number of sections with similar characteristics. For example, we only allow a fixed number of consecutive "talking-heads" in a skim. We include frames from other scenes when words are longer than 1.1 seconds, 33 frames (Example b) in Figure 10). For visual clarity, we display at least 18 frames per skim scene. Keywords that appear in close proximity or repeat throughout the transcript may create redundant skims and offer little insight to the global content of a scene. We avoid this by maintaining a minimum of 70 frames between keywords and limiting repetition for each word. For scenes containing no keywords we extract keywords from adjacent scenes.

3 Skim Creation:

We have segmented and characterized the video by camera motion, object appearance and keywords. In order to create the video skim, we attempt to interpret the intent of the video segments using the characterization results to extract and order the significant video frames and audio. Figure 10 illustrates a few examples of applying these results and the resultant video skim. The sections below describe the steps involved in skim creation.

3.1 Keywords and Skim Audio

The first level of analysis for the skim is creating the compressed audio track, which is based on the selected keywords. We create the skim audio track by simply appending each successive keyword. By varying the number of keywords we can control the size of the skim. The actual word length of audio for each keyword is isolated from the audio track to form the skimmed audio as shown in figure 11. Since the audio length is fixed, we will need to choose the corresponding

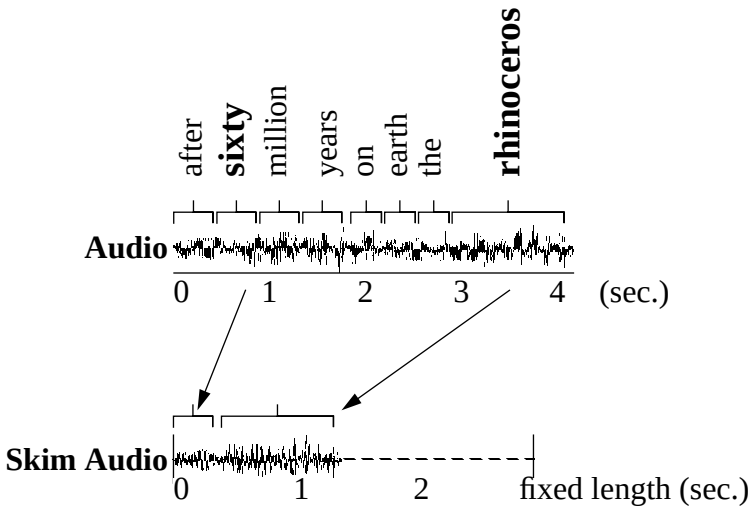


Figure 11: The word length for "rhinoceros" is 1.10 seconds which allows for 33 frames; the word "sixty" uses 19 video frames for 0.63 seconds of audio. Keywords are added until the audio skim length is filled.

number of video frames to fill the image skim. The frames for the image skim will not necessarily align to the words of the audio skim, as seen in figure 10.

3.2 Prioritizing Image Frames

We now select the image portion to combine with the skim audio for the complete video skim. For each scene we analyze the characterization results of every frame and select a set of frames most appropriate for skimming. Priority for each set of frames is based on the following ranking system:

1: Frames with faces or text
2: Static frames following camera motion
3: Frames with camera motion and human faces or text
4: Frames at the beginning of the scene (Default)

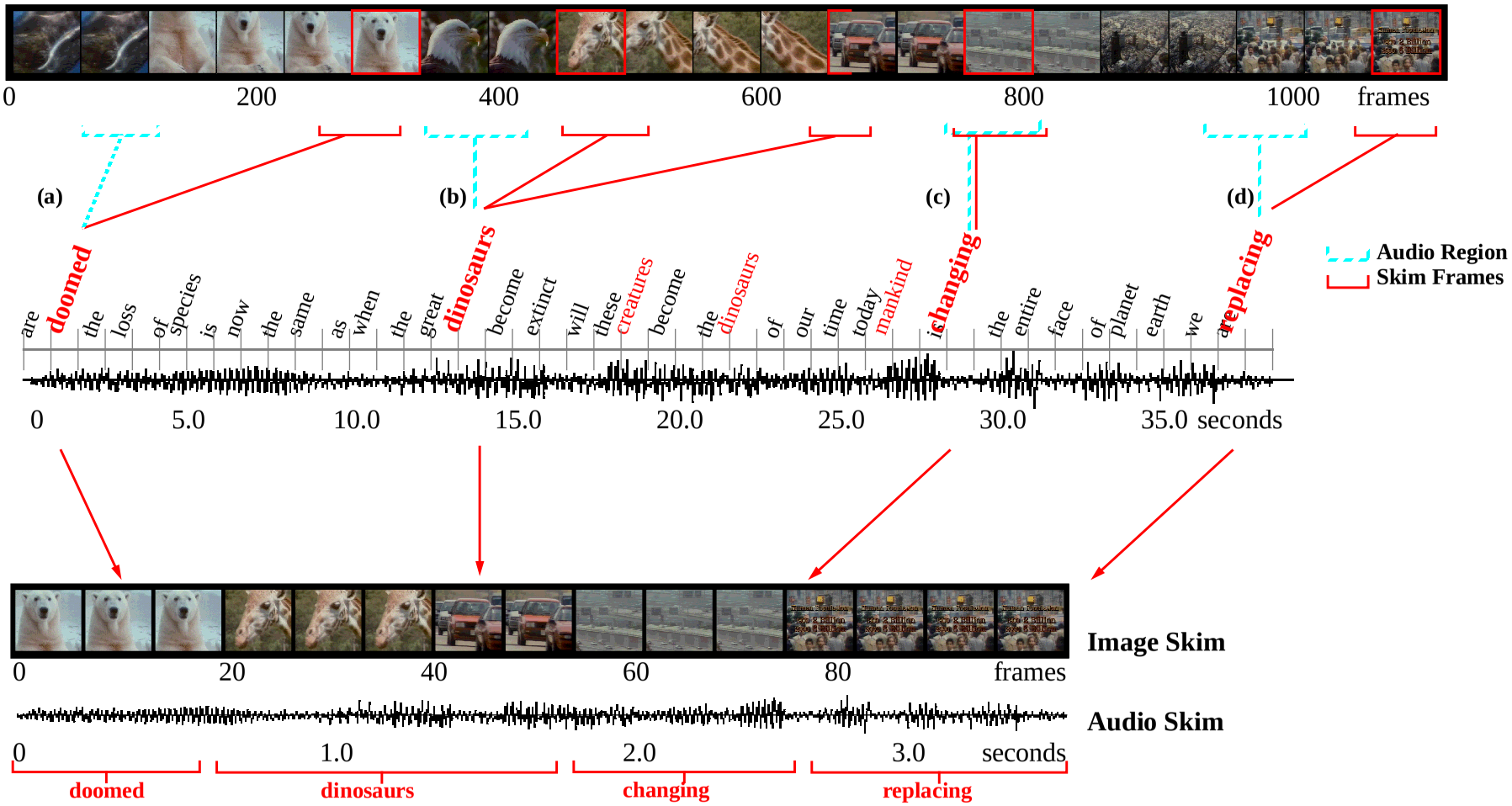


Figure 10: Skim creation from original video incorporating word relevance in the transcript, objects in video (humans and text), and camera motion. The examples above illustrates: (a) For the word "doomed", the portion of the scene with little or no motion is selected, since typically the static region is the focus of the scene; (b) The narrator uses 1.13 seconds (34 frames) to utter the word "dinosaur" so a portion of the next scene is included for more content; (c) With no significant motion or object, we use the initial portion of the scene for the word "changing"; (d) For the word "replacing" the latter portion of the scene which contains both text and humans is chosen.

angles are computed. We now extract clusters with bounding regions that satisfy the following constraints:

Bounding Aspect Ratio ≥ 0.75

Cluster Fill Factor ≥ 0.45

Cluster Size > 70pixels

A cluster's bounding region must have a small vertical-to-horizontal 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. Finally, we examine the intensity histogram of each region to test for high contrast. This is because certain textures and shapes are similar to text but exhibit low contrast when examined in a bounded region. This method works best with horizontal titles and captions. Table 1 shows the statistics for

Table 1: Text Detection Results

Data	Images	Text Detected	Text Missed	False Detection
News1	20	11	1	4
News2	23	7	0	3
Species	20	12	1	0

detection on various sets of images. Figure 9 shows several detection examples of words and subsets of a word.

Figure 9: Text detection results with various images.

in Carnegie Mellon. Its current performance level is to detect over 90% of more than 300 faces contained in 70 images, while producing approximately 60 false detections. While much improvement is needed, the system can detect faces of varying sizes and is especially reliable with frontal faces such as talking-head images. Figure 7 shows an example of its output, illustrating the range of face sizes that can be detected.

Figure 7: Detection of human-faces.

Text Detection

Text in the video provides significant information as to the content of a scene. For example, statistical numbers are not usually spoken but are included in the captions for viewer inspection. Names and titles are attached to close-ups of people. A text region is a horizontal rectangular structure of clustered sharp edges, due to characters with high contrast color or intensity, against the background. By detecting these properties we can extract regions from video frames that contain textual information. Figure 8 illustrates the process. We first apply a 3x3 horizontal differen-

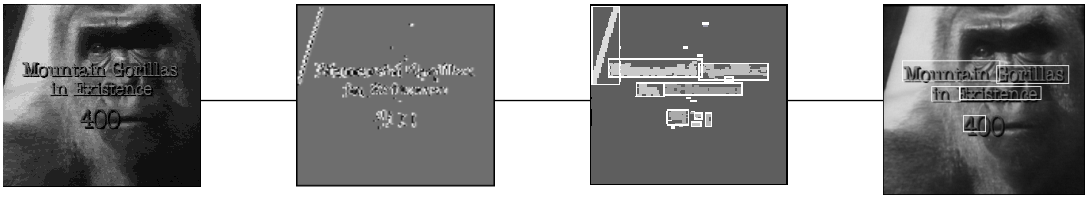


Figure 8: Output at various stages for text detection algorithm.

tial filter and appropriate binary thresholding to the entire image to extract vertical edge features. Then smoothing is applied to eliminate extraneous fragments, and connect edge elements that may have been detached. Individual regions are identified by cluster detection and bounding rect-

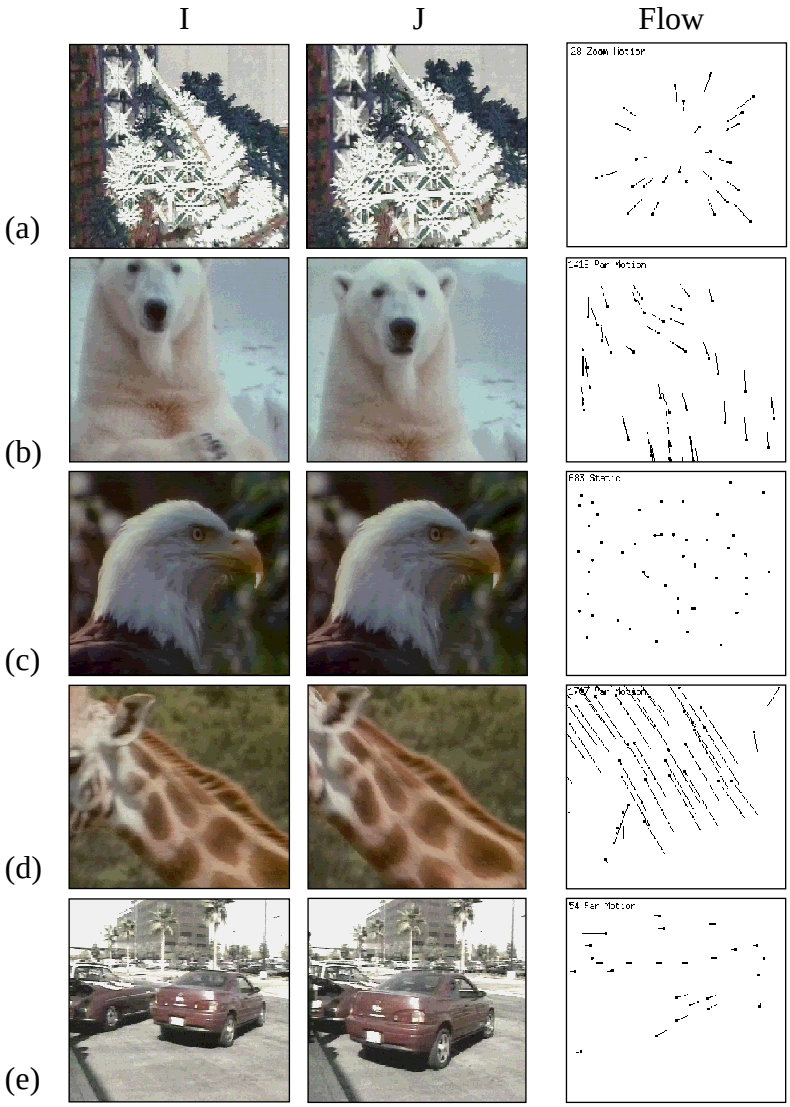


Figure 6: Camera motion analysis using optical flow: (a) Zoom distribution; (b) Downward pan with subtle object motion; (c) Static frames; (d) Significant object and panning motion; (e) Subtle pan with significant object motion. Flow vectors are amplified for visibility.

2.4 Object Detection

We identify significant objects by searching and matching known templates to individual regions in the frame. For the time being, we have chosen to deal with two of the more interesting objects in video, human faces and text (caption characters).

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, *et al* [7], at the Vision and Autonomous Systems Center

A multi-resolution structure is used to accurately track regions over large areas and reduce the time needed for computation. A motion representation of the scene is created by measuring the velocity that individual regions show over time. Velocity vectors for pans and zooms have distinct statistical characteristics for vector directions. Figure 5 describes the characterization of camera motion through statistics of the optical flow vectors. The angular distribution of the pan will peak at a single region, whereas the distribution of a zoom sequence is relatively flat.

Global motion analysis distinguishes between object motion and actual camera motion. Object motion typically exhibits flow fields in specific regions of the image. Camera motion is characterized by flow throughout the entire image. Frames with minimal camera motion are often suitable for descriptive representation.

For object motion description, trackable features must be identified. They must be features of an object, such as corners, or areas rich in texture, so that they do not show ambiguities in tracking. Such trackable features can be identified as the regions with large, well conditioned eigenvalues in the 2x2 gradient derivative matrix, *G*, that appears on the left side of equation (4) [10].

$$G = \begin{bmatrix} \sum \frac{\partial^2}{\partial x} I & \sum \frac{\partial}{\partial x} I \frac{\partial}{\partial y} I \\ \sum \frac{\partial}{\partial x} I \frac{\partial}{\partial y} I & \sum \frac{\partial^2}{\partial y} I \end{bmatrix} \qquad (5)$$

Since we are primarily interested in distinguishing static frames from motion frames, it was sufficient to track only the top 30 features. Examples of the motion analysis are shown in figure 6.

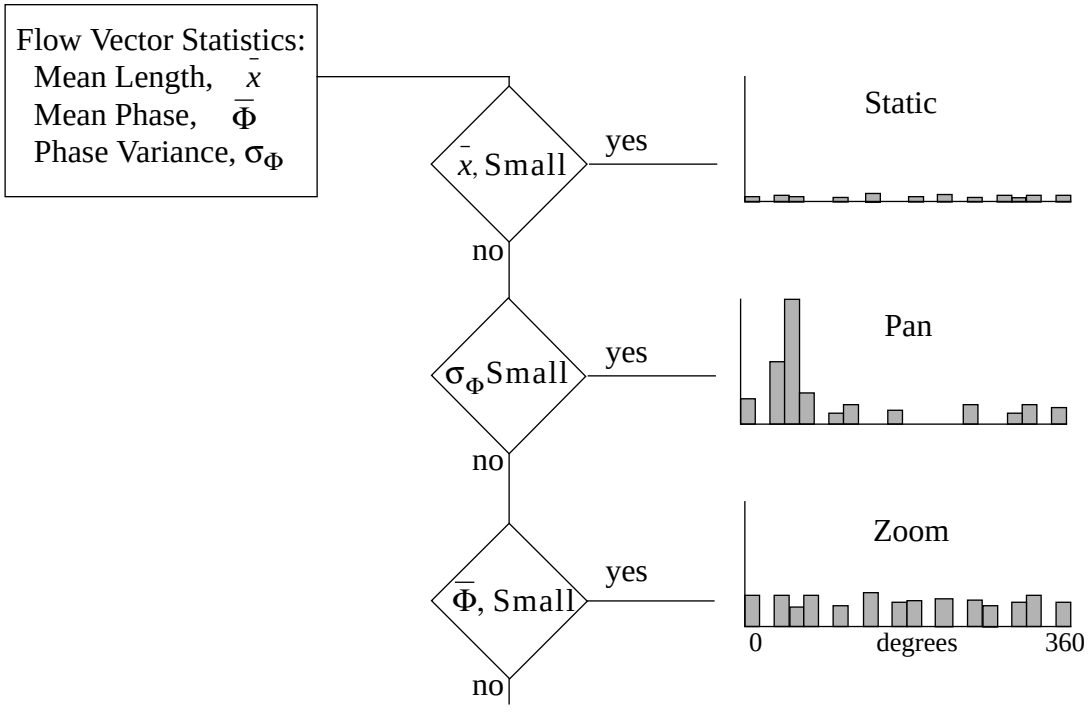


Figure 5: Flow diagram and angular distributions for motion analysis.

cessive frame, image sequences can be separated into scenes. In the difference, *D(t)*, peaks are detected and an empirical threshold is used to select scene breaks. Using only the histogram difference, we have achieved 90% accuracy on a test set of roughly 200,000 images (2 hours). An example of the scene detection result is shown in figure 4.

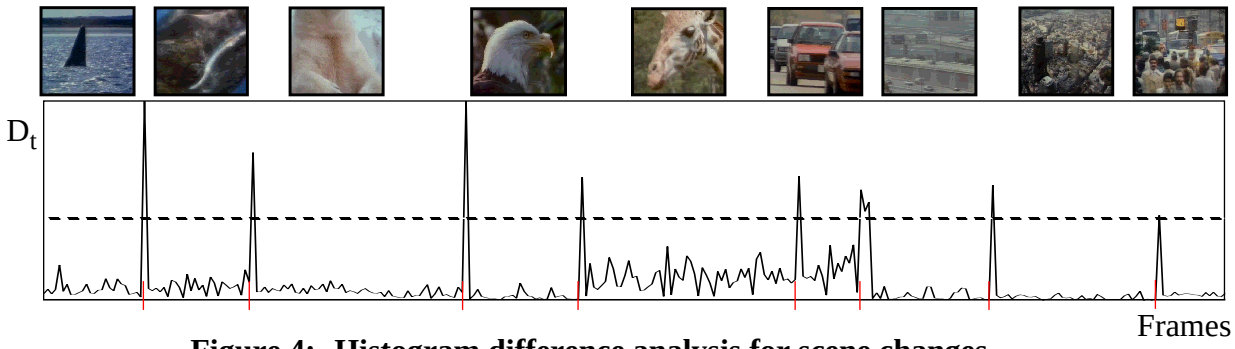


Figure 4: Histogram difference analysis for scene changes. Icons represent the first frame of each scene.

2.3 Camera Motion Analysis

One important method of video characterization is based on interpreting camera motion. Video contains a high level of redundancy in terms of visual information. Many scenes have beautiful poses and visual effects, but offer little in the description of a particular segment. A static scene may appear for several seconds when in fact less than 2 seconds is necessary for mere visual comprehension. Since the skim scene will consist of a small number of frames, we avoid frames in scenes with excess camera motion and visual redundancy to insure comprehension in a short time. We can interpret camera motion as a pan or zoom by examining the geometric properties of the optical flow vectors[1]. Using the Lucas-Kanade gradient descent method for optical flow[5], we can track individual regions from one frame to the next and create a vector representation for all associative camera motion. *I* and *J* represent features in successive images. A feature in *I*, displaced by $\boldsymbol{d} = (\Delta x, \Delta y)$, will be approximately equivalent to the same feature in *J*. An L_2 norm difference is used as the basis for region comparison. When assuming small feature motion $\Delta x, \Delta y$

$$E = \sum_{x, y \in \overline{W}} [I(x + \Delta x, y + \Delta y) - J(x, y)]^2 \qquad (3)$$

between frames, minimizing this difference with respect to $\boldsymbol{d}$ reduces to solving the following equation:

$$\begin{bmatrix} \sum \frac{\partial^2}{\partial x} I & \sum \frac{\partial}{\partial x} I \frac{\partial}{\partial y} I \\ \sum \frac{\partial}{\partial x} I \frac{\partial}{\partial y} I & \sum \frac{\partial^2}{\partial y} I \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} \sum (J - I) \frac{\partial}{\partial x} I \\ \sum (J - I) \frac{\partial}{\partial y} I \end{bmatrix} \qquad (4)$$

dard corpus, f_c. A high TF-IDF signifies relative importance. Words that appear often in a particular segment, but appear relatively infrequently in the standard corpus receive the highest weights. An example of the keyword selection results is shown in figure 3.

While we plan to automate the transcript creation process through speech recognition, we currently rely on manual transcription and closed captions. Techniques in speech recognition will also be used to segment video based on transitions between speakers and topics which are usually marked by silence or low energy areas in the acoustic signal[2].

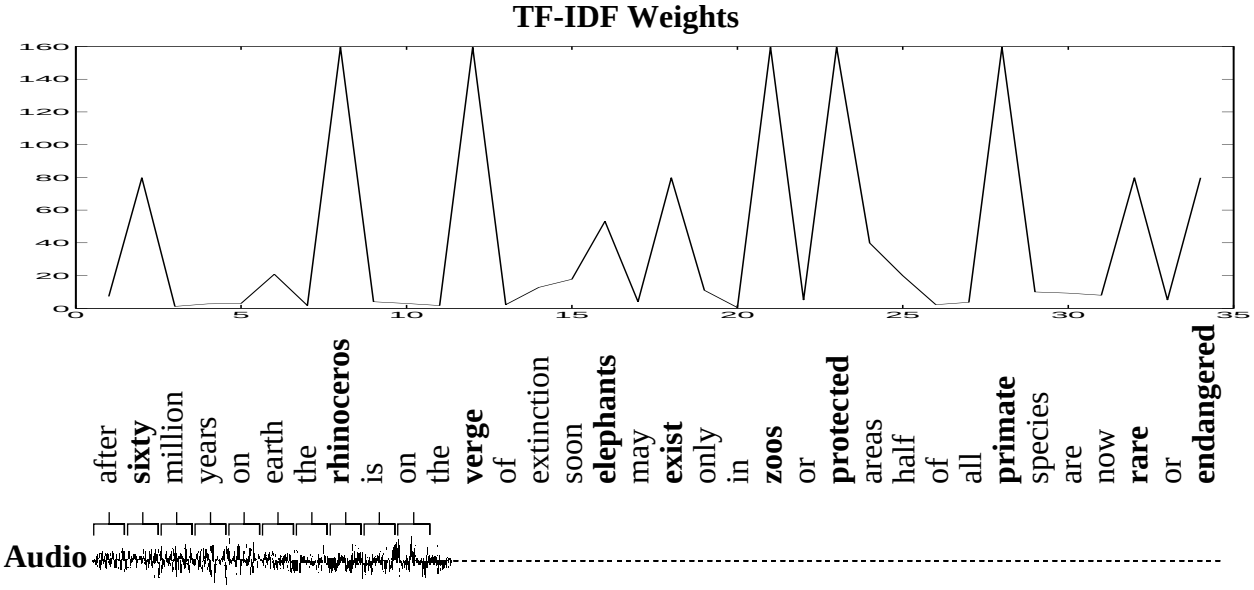
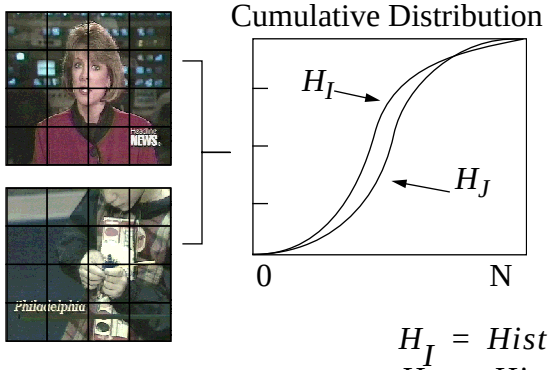


Figure 3: keywords isolated from transcript through TF-IDF weights. Words of highest relevance are indicated in bold.

2.2 Scene Segmentation

To analyze each segment as individual scenes, we must first identify frames where scene changes occur. Several techniques have been developed for detecting scene breaks [11], [4], [13]. We choose to segment video through the use of a comparative histogram difference measure. For our purpose we have found that this technique is simple, and yet robust enough to maintain high levels of accuracy. By detecting significant changes in the weighted color histogram of each suc-



$$D(t) = \sum_{i=0}^{N} \left| H_I(i) - H_J(i) \right| \qquad (2)$$

$$H_I = Hist(t)$$
$$H_J = Hist(t + \tau)$$

2 Video Characterization

Through techniques in image and language understanding, we can characterize scenes, segments, and individual frames in video. Figure 2 illustrates an example of analyzing a video

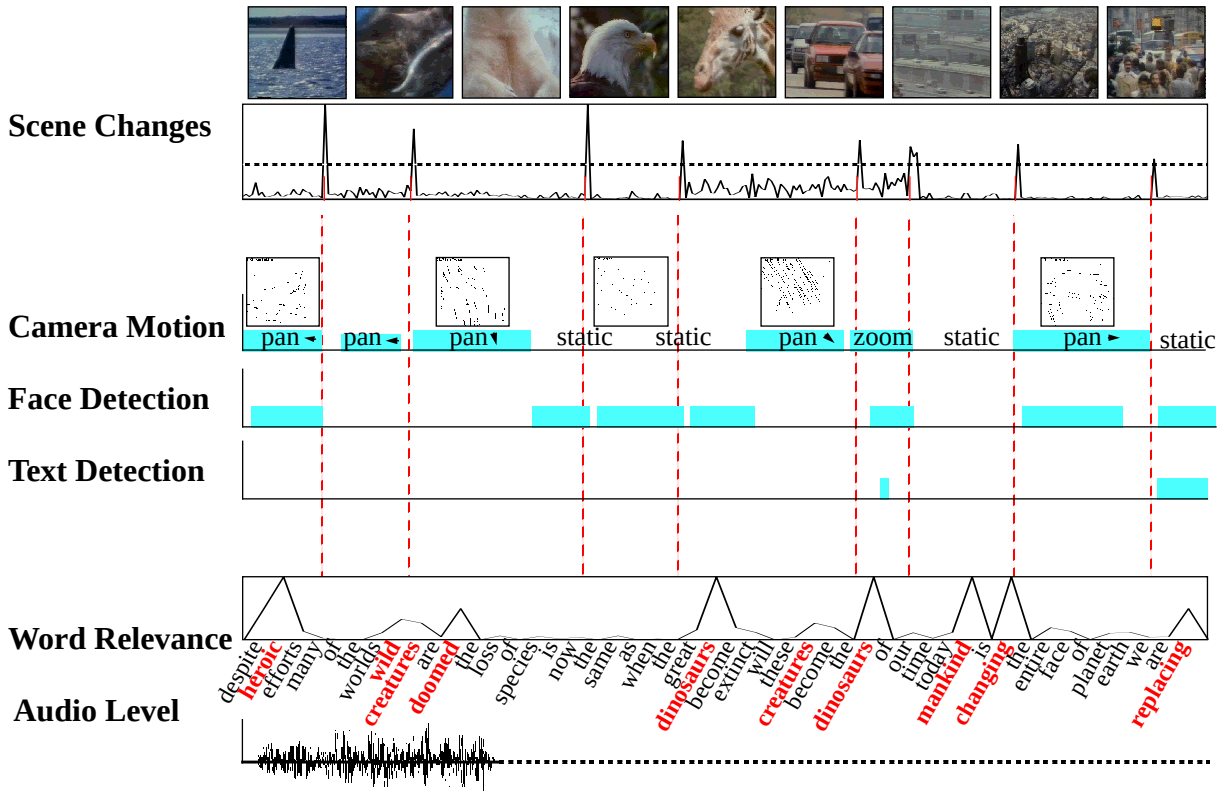


Figure 2: Characterization Technology for Skim Creation. The video is segmented into scenes. Camera motions are detected along with significant objects (faces and text). Bars indicate frames with positive results. Word relevance is evaluated in the transcript.

clip by various speech, language and image understanding techniques. For language understanding, this entails identifying the most significant words in a given scene. For image understanding, we identify frames which contain objects of importance as well as segmenting and identifying the structual motion of the scene.

2.1 Keyword Selection

Language analysis works on the audio transcript to identify keywords in it. We use the well-known technique of TF-IDF (Term Frequency Inverse Document Frequency) to identify critical words and their relative importance for the video document [6], [8]. The TF-IDF of a word is

$$\text{TF-IDF} = \frac{f_c}{f_s} \tag{1}$$

the frequency of a word in a given scene, f_s, divided by the frequency of its appearance in a stan-

words and images from a segment, and produce a skimmed video. Figure 1 illustrates the concept of extracting the most representative information to create the skim.

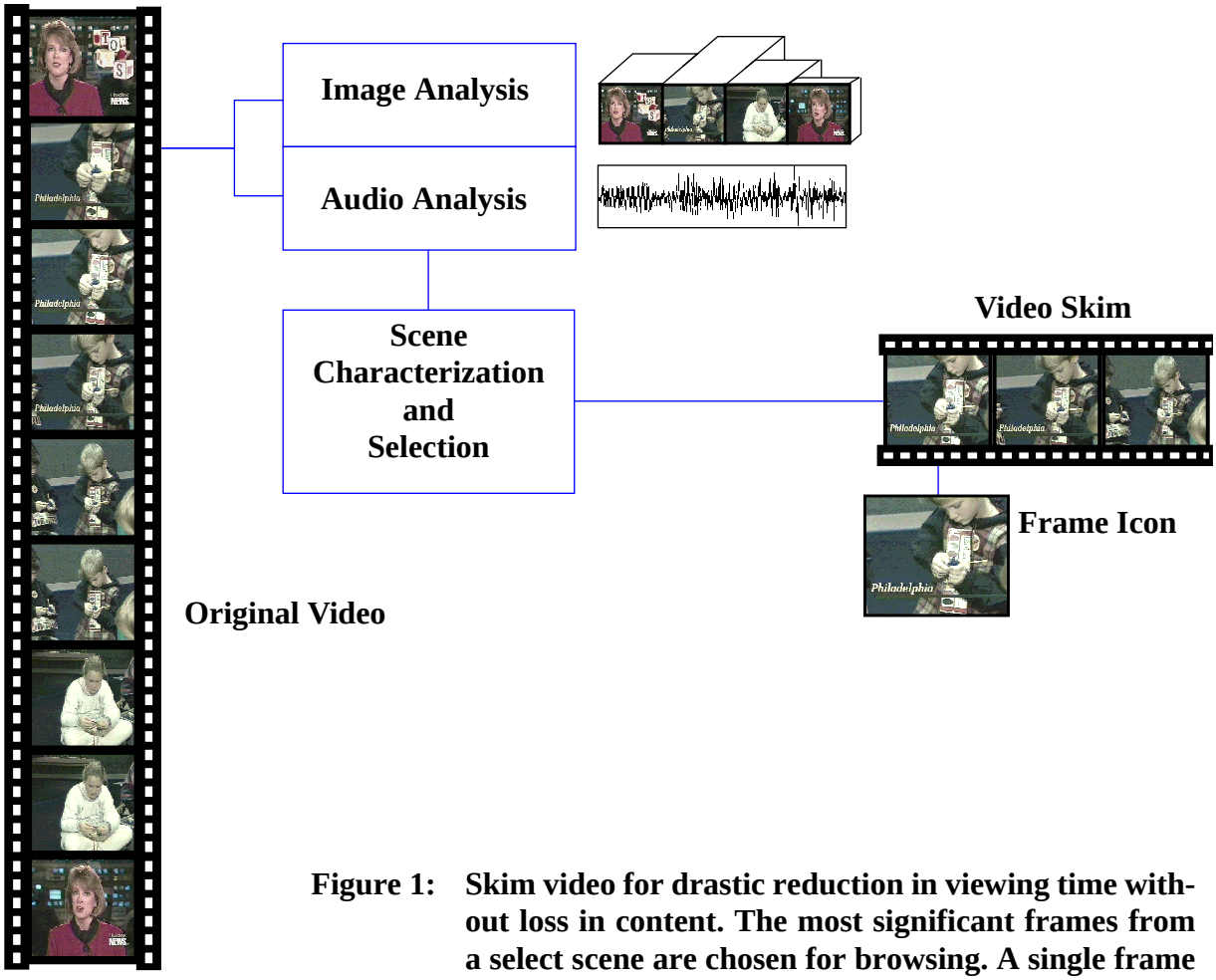


Figure 1: Skim video for drastic reduction in viewing time without loss in content. The most significant frames from a select scene are chosen for browsing. A single frame is selected from the skim for iconic representation.

The critical aspect of compressing 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 can 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 the objects that appear in it, the associated words and the structure of the video scene. The skim is the smallest comprehensible video representation of the original segment. The lowest level of compaction is a single icon which could naturally be extracted from the skim video frames since they contain the most significant information.

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. The results from this system will show the utility of the video skim as an effective means of browsing.

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 University [9], will consist of thousands of hours of video made available to a user upon request. To access the library the information embedded within the digital video must be easy to locate, manage and display. Even with intelligent content-based search algorithms being developed [6], [15], multiple segments will be returned to insure retrieval of pertinent information and the users will often need to view them to obtain final selections.

For many users, a query of interest is not always a full-length film. Unlike video-on-demand, video libraries will provide informational access in the form of brief, content-specific segments as well as full-featured videos. These segments will act as "video paragraphs" for the entire broadcast, allowing the user to view the complete video by moving from one segment to the next. In video libraries, the user will want to "skim" the relevant portions of video for the segments that are related in content to their query. To avoid time consuming searches, there must exist technology to organize these collections so users can effectively retrieve and browse the video data for specific content.

Browsing Digital Video

For the purpose of browsing, techniques such as increasing the video playback speed and displaying video at fixed intervals offer little to convey content. Speeding up the video rate eliminates the majority of the audio information and distorts much of the image information[3], while showing separate video sections at fixed intervals merely gives a random estimate of the overall content. Recently, techniques have proposed browsing representations based on information within the video [12], [13], [14], [16]. These systems are primarily based on the motion of the video, placement of individual scenes changes, and image statistics such as color and shape. Presently, no system automatically utilizes the specific contents of video, such as audio information, specific types of objects in video, or areas of significance from camera structure. Browsing must entail not only decreased viewing time, but also must preserve the essential message of the video.

An ideal browser would display only the video pertaining to a scene's content, suppressing irrelevant data. A separate video, containing only the images pertinent to content, would be considerably smaller than the original source and could be used to skim the video in browsing. To extract the significant images from a video would result in a smaller, content-specific version of the original. The audio portion of this video should also consist of the significant audio or spoken words, instead of simply using the synchronized portion of the audio corresponding to the selected images.

The compacted video of the original could be used to view several segments or an entire broadcast in much less time without losing the content and could be called during playback. The user could sample many segments without actually viewing each in it's entirety. The level of compaction should be adjustable so a user could view sections with as much or as little content as needed. To view a full-length feature, the user could watch at the lowest level of compression to maintain content, while still reducing the viewing time from hours to minutes.

Video Skims

We describe a method to create a short synopsis of a video segment, a skimmed video. Using various techniques in image and language understanding, we can extract the significant

Keywords: video library, browsing, integrated technology, video paragraph, skimming

Video Skimming for Quick Browsing based on Audio and Image Characterization

Michael A. Smith and Takeo Kanade
July 30, 1995
CMU-CS-95-186

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Digital video is rapidly becoming an important source for information, entertainment and a host of multimedia applications. With the size of these 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 short synopsis of the original. The extraction of significant information, such as specific objects, audio keywords and relevant video structure, is made possible through the integration of techniques in image and language understanding. The resulting skim is much smaller, and 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. 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.