Video Manga:
Generating Semantically Meaningful Video Summaries

Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, John Boreczky
FX Palo Alto Laboratory
3400 Hillview Avenue, Palo Alto, CA 94304, USA
+1 650 813 6907
{shingo, foote, andreasg, johnb}@pal.xerox.com

1. ABSTRACT

This paper presents methods for automatically creating pictorial video summaries that resemble comic books. The relative importance of video segments is computed from their length and novelty. Image and audio analysis is used to automatically detect and emphasize meaningful events. Based on this importance measure, we choose relevant keyframes. Selected keyframes are sized by importance, and then efficiently packed into a pictorial summary. We present a quantitative measure of how well a summary captures the salient events in a video, and show how it can be used to improve our summaries. The result is a compact and visually pleasing summary that captures semantically important events, and is suitable for printing or Web access. Such a summary can be further enhanced by including text captions derived from OCR or other methods. We describe how the automatically generated summaries are used to simplify access to a large collection of videos.

1.1 Keywords

Video summarization and analysis, keyframe selection and layout.

2. INTRODUCTION

Video is an information-intensive medium. To quickly get an overview of a video document, users must either view a large portion of the video or consult some sort of summary. In this paper, we describe techniques for automatically creating pictorial summaries of videos using automatic content analysis. While any collection of video keyframes can be considered a summary, our goal is to produce a meaningful and concise representation of the video. We do this by automatically choosing only the most salient images and efficiently packing them into a pictorial summary.

While existing summarization techniques rely chiefly on collecting one or more keyframes from each shot, our approach goes significantly beyond this in a number of ways. Because video summarization is essentially a data reduction process, the degree to which a summary retains important events is a measure of how good it is. The heart of our approach is a measure of importance that we use for summarization. Using the importance measure, keyframes are selected and resized to reflect their importance scores, such that the most important are largest. The differently-sized keyframes are then efficiently packed into a compact summary reminiscent of a comic book or Japanese *manga*.

Our importance measure easily incorporates multiple sources of information. Using low-level automatic analysis, we can successfully find video shots of titles, slides or other visual aids, close-ups of humans, and long shots of audiences [8]. These methods could be used equally well to find anchor shots, reporter “talking heads,” and graphics in a news broadcast. This lets us emphasize important images such as human figures and de-emphasize less important images such as long shots. We also perform optical character recognition on image text that appears in the source video or in synchronized presentation graphics. This gives us text for automatically captioning our summaries (though equivalent text could be found by other means such as closed captions or speech recognition).

A key aspect of our approach is a quantitative assessment of how good a summary actually is. Assessing the quality of our summaries lets us improve them by both removing redundant information and including more relevant data. Though our techniques are applicable to general audio-visual media, we present experimental results in the domain of informal video-taped staff meetings and presentations. Because we have the meeting minutes as ground truth, we can determine what fraction of the events recorded in the meeting minutes are actually depicted in the summary.

At FX Palo Alto Laboratory, regular staff meetings and other presentations take place in a conference room outfitted with several video cameras. All formal meetings and most presentations are videotaped, MPEG-encoded, and made available to staff via the laboratory intranet. These videos

amount to about three hours per week and currently we have more than 160 hours of video in our database. We have built a system that automatically creates an interactive summary for each video on demand. The system is used in daily work at FXPAL. We have also applied our summarization techniques to other video genres such as commercials, movies, and conference videos with good results.

The rest of the paper is organized as follows. In Section 3, we discuss related work. Section 4 describes our methods of creating video summaries. A video segmentation technique using a hierarchical clustering is presented. We also introduce a measure of segment importance. Results of its quantitative evaluation are shown in Section 5. Section 6 describes techniques to enhance video summaries by integrating semantic information. Applications are presented in Section 7. We conclude with directions for future work.

3. RELATED WORK

Many tools have been built for browsing video content [1][2][17][21]. The tools use keyframes so that the contents can be represented statically. These do not attempt to summarize video but rather present video content "as is." Therefore, keyframes are typically extracted from every shot resulting in some redundancy. Some systems select more than one keyframe per shot to visualize camera or object motion. Keyframe layout is typically linear, although some approaches occasionally use other structures [2][17]. Our approach ranks the importance of the video shots and eliminates shots that are of lesser importance or are too similar to other shots.

Shahraray *et al.* at ATT Research have worked on using keyframes for an HTML presentation of video [13]. One keyframe was selected for each shot; uniformly sized keyframes laid out in a column along closed-caption text. This approach has an advantage over watching the entire video of a broadcast news program because the content is randomly accessible. However, it still requires scrolling through a large number of keyframes for videos with many shots.

Taniguchi *et al*. have summarized video using a 2-D packing of "panoramas" which are large images formed by compositing video pans [15]. A "panorama" enables a single keyframe to represent all images included in a shot with camera motion. In this work, keyframes are extracted from every shot and used for a 2-D representation of the video content. Because frame sizes were not adjusted for better packing, much white space can be seen in the summary results.

Yeung *et al.* have made pictorial summaries of video using a "dominance score" for each shot [19]. Though they work towards a goal similar to ours, their implementation and results are substantially different. The sizes and the positions of the still frames are determined only by the dominance scores, and are not time-ordered. While their summaries gave some sense of video content, the lack of sequential organization can make the summaries difficult to interpret.

Huang *et al*. have created summaries of news broadcasts, as reported in [9]. Story boundaries were pre-determined based on audio and visual characteristics. For each news story, a keyframe was extracted from a portion of video where keywords were detected the most. Their method nicely integrated information available for news materials, but relies heavily on the structured nature of broadcast news and would not apply to general videos.

Some approaches to summarization produce video skims [4][11][14]. A video skim is a very short video that attempts to capture the essence of a longer video sequence. Although video skims may be useful for some purposes, the amount of time required for viewing suggests that skimmed video is not appropriate for a quick overview. Christal *et al.* [4] require each segment in a skimmed video to have a minimum duration which limits the lower boundary for compactness. Producing a single image allows our video summaries to be viewed at-a-glance on a web page or printed on paper.

4. SELECTING AND PACKING KEYFRAMES

A typical keyframe extraction algorithm is described in [22]. The video is first segmented into shots and then the frames of each shot are clustered to find one or more keyframes for that shot. In contrast, our approach does not rely on an initial segmentation; we cluster all the frames of the video (or a sub-sampled representation). This approach yields clusters of similar frames regardless of their temporal continuity. We use smoothed three-dimensional color histograms in the YUV color space to compare video frames [7].

4.1 Clustering Similar Frames

To find groups of similar frames, we use a bottom-up method that starts with assigning each frame to a unique cluster. Similar frames are clustered by iteratively merging the two closest clusters at each step. This hierarchical clustering results in a tree-structured representation with individual frames at the leaves. At the root node of the tree is the maximal cluster consisting of all the frames. The children of each node are the sub-clusters that were merged to form the node, and so forth down to the leaves. We record the distance of the merged clusters with each node so that it can be used to select a desired number of clusters by thresholding. We select an optimal threshold by finding the knee in the curve of the cluster diameter. Once the distance gets large enough, irrelevant frames start being incorporated into the same cluster, and the cluster diameter begin to increase rapidly. A typical case of the distance transition is shown in Figure 1.

Once clusters are determined, we can segment the video by determining to which cluster the frames of a contiguous segment belong. This avoids all the pitfalls of on-the-fly shot detection, as it does not generate spurious frames due to motion, pans, or fades. In particular, it does not rely on manually-tuned thresholds for good performance, and thus works well across a variety of video genres.

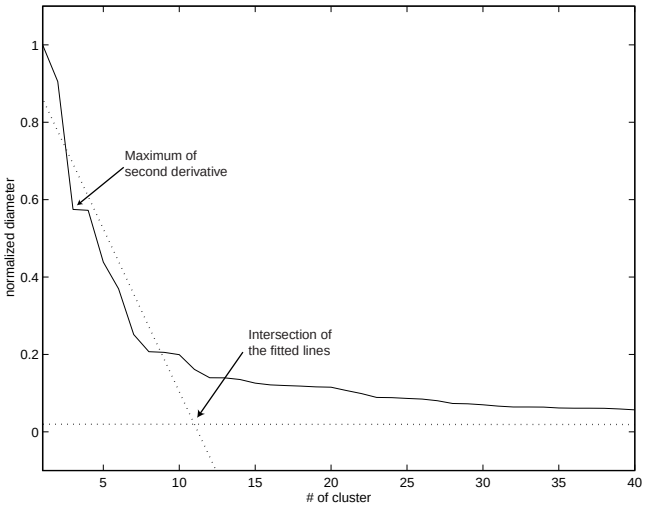

Figure 1. Finding a Threshold for Segmentation

4.2 Importance Score

For a good summarization, we must still discard or de-emphasize many segments. To select appropriate keyframes for a compact pictorial summary, we use the importance measure introduced in [16]. This calculates an importance score for each segment based on its rarity and duration, such that a segment is deemed less important if it is short or very similar to other segments. This penalizes both repetitive shots and shots of insignificant duration. Given C clusters in the video, a measure of normalized weight W_i for cluster i is computed as

$$W_i = \frac{S_i}{\sum_{j=1}^{C} S_j} \quad (1)$$

where S_i is the total length of all segments in cluster i, found by summing the length of all segments in the cluster. W_i is the proportion of segments from the whole video that are in cluster i.

A segment is important if it is both long and rare, that is, it does not resemble most other segments. Thus weighting the segment length with the inverse of the cluster weight yields a measure of segment importance. Thus the importance I of segment j (from cluster k) is

$$I_j = L_j \log \frac{1}{W_k} \quad (2)$$

where L_j is the length of the segment j.

The importance measure becomes larger if the segment is long, and smaller if the cluster weight is large (meaning the segment is common). The contribution from the length and the cluster weight can be balanced by weighting the reciprocal of W_i by a factor other than unity.

A particular advantage of the importance score is that it easily incorporates evidence derived from other sources. We use this in several ways, for example in Section 6 we weight the importance score with automatic analysis that can detect shots of humans. Thus human images are more important and are favored in the summary.

4.3 Selecting and Preparing Keyframes

Segments with an importance score higher than a threshold are selected to generate a pictorial summary. The appropriate number of keyframes will vary depending on the video type and duration. We found that one eighth of the maximum importance score as a threshold value resulted in a good selection of keyframes [16]. By using the maximum score as a normalizing factor, the threshold is automatically adjusted to the contents. This method has shown to work on many video types such as movies and commercials. The algorithm can also be altered to select a variable number of segments so the summary length can be precisely controlled. We use this feature in the experiments of Section 5.

For each segment chosen, the frame nearest the center of the segment is extracted as a representative keyframe. Frames are sized according to the importance measure of their originating segments, so that higher importance segments are represented with larger keyframes.

Larger keyframes help guide users' attention to important segments. In the current implementation, if the importance of a given frame is between 1/8 and 1/4 of the maximum, it is assigned the smallest frame size. Frames scoring more than 1/4 but less than 1/2 of the maximum are sized twice the smallest size, and frames scoring higher than 1/2 are sized three times larger than the smallest. Table 1 shows distributions of different sized frames. Note that the keyframe sizes distributions are similar for three out of four categories despite the very different source genres.

Category	omitted	Size 1	Size 2	Size 3
meetings	0.744	0.125	0.079	0.051
commercials	0.360	0.253	0.240	0.147
movies	0.747	0.130	0.079	0.044
produced	0.739	0.140	0.087	0.033

Table 1: Distribution of Assigned Frame Sizes

4.4 Frame-Packing Algorithm

Once an appropriate set of frames has been selected, they may be laid out to form a pictorial abstract of the video sequence. A two-dimensional layout is most appropriate for a printed synopsis, such as a "comic book" format. Given that the selected frames have multiple sizes, a sequence of frames must be found that both fills space efficiently and represents the original video sequence well.

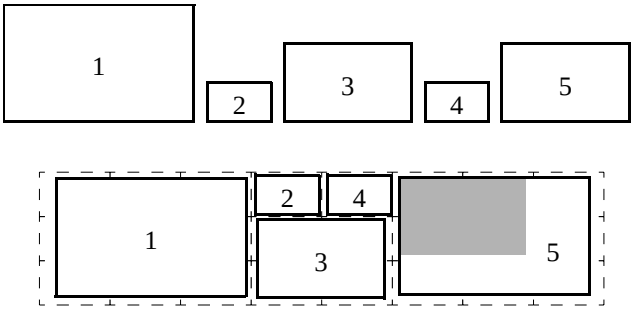

Figure 2. Packing Keyframes into a Row Block

We use a frame-packing algorithm to find the frame sequence for a particular "block" or sub-region of the entire area. The block to be packed is divided into a grid, such that one unit of the grid will hold the smallest frame. One "row block," or row of columns across the grid, is packed at a time. Once a row block has been packed with frames, further row blocks are considered iteratively until all frames have been packed. To pack one row block, a "block exhaustive" algorithm is used; all possible packing are considered and the one with the best packing and least frame resizing is chosen. Though this is an exhaustive algorithm, the number of possible combinations in one block is small enough to make this approach very rapid. More details of the algorithm are described in [16].

An example of this row block packing procedure is depicted in Figure 2. The rectangles at the top are the original frame sequence, sized by importance. The bottom picture illustrates the frames packed into a row block of height 3 and width 8. Note that frame 5 has been resized from its original size (indicated as a gray rectangle) for a better packing with minimal white space. This approach will work for many genres of video. For example, Figure 3 is an example of a summary from an MPEG7 reference pop music video [10].

5. QUANTITATIVE EVALUATION

5.1 Preliminary Experiments

Any collection of keyframes can be considered a "video summary," however our goal is to produce a meaningful display that optimally presents the events in the summarized video. To this end, we evaluated how well the importance score reflects the actual importance of the video.

Meeting minutes produced by a human secretary give a list of important events in our weekly staff meetings, which are also videotaped. The minutes and videos were created independently from each other. To evaluate our summaries, we measured the proportion of important events that are recognizable in our meeting video summaries.

We used 23 meeting videos with a total length of 943 minutes for our experiments. Those videos were recorded from various signal sources including multiple cameras, PC video output, and a VCR. The videos were not post-produced; all editing was done during the recording by a human operator. The operator could remotely pan and zoom the cameras as

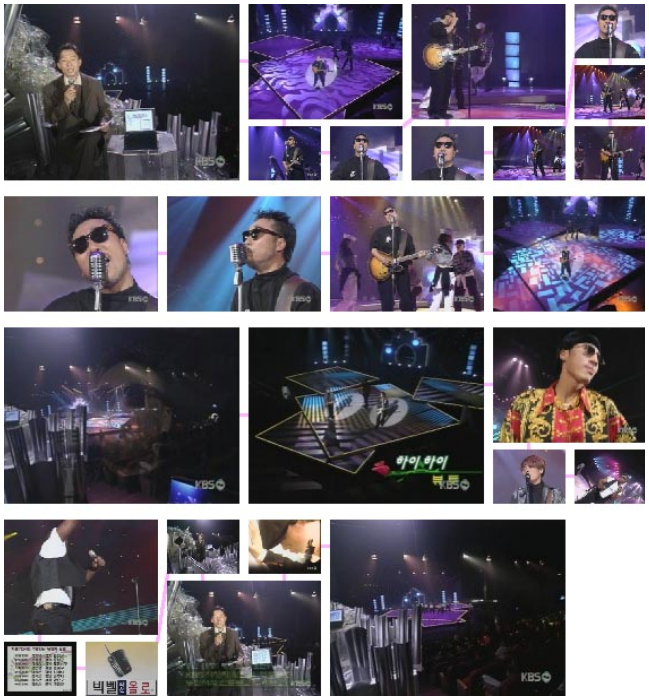

Figure 3. An Example of the Comic Book Style Summary

well as switch between cameras and other sources such as the VCR and the computer display.

A fixed number of segments with the highest importance score were selected for each summary. Each segment was represented by one keyframe extracted from the middle. (The actual number of selected keyframes varies because of ties in the importance score.)

We reformatted the meeting minutes so that so that all events are expressed as a short headline such as "Jim, announcement," "Peter, introduced" or "Joe, trip report." 126 total events were described in the meeting minutes.

The videos were reviewed by the authors, and the video sequences were subjectively compared with the events written in the minutes. We found that 88.5 percent of the events could be recognized in the complete video sequences. Because a summary cannot include information not in the source video, this is the upper limit of summarization performance and is represented as the dotted line in the figures. We used reasonably strict criteria for an event to be "recognizable." For example, if someone presents a slide show, a video or summary must contain a good close-up shot of the person and one of the slide images for the event to be judged as recognized. If an event is shown in a video only partially, we count it as half coverage. Some events were not properly captured in the video. This was often observed when a person was standing outside camera range, or the person could not be recognized because of poor lighting, misaimed camera, or other factors.

Figure 4 shows results of this experiment. Event coverage (circles) are the fraction of events that are recognizable in

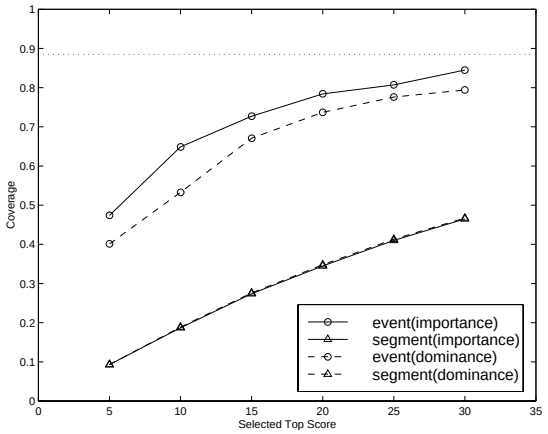

Figure 4. A Comparison of Event and Segment Coverage

the summary keyframes. Segment coverage (triangles) show the ratio of the number of segments used to the total number of segments. We also compared our importance score with the "dominance score" (dashed lines) described in [19], which uses segment length as its importance measure. Our importance score suppresses repetitive shots to be selected yielding more informative set of keyframes. The experiment clearly shows that our method produces better summaries, because they represent more recognizable events with fewer keyframes.

5.2 Eliminating Further Redundancy

Using our measure of summarization coverage lets us further improve our summaries. We can now eliminate more redundant information, resulting in a more compact summary, while experimentally verifying that we have not lost significant event coverage.

Some redundancy is caused by the thresholding of importance scores. For example, two or more consecutive segments may be from the same cluster if all segments between them fall below the threshold. This kind of redundancy can be eliminated using the following two heuristics, applied repeatedly until all redundant keyframes are eliminated.

5.2.1 Removing duplicates

If two consecutive frames are from the same cluster, the smaller frame is eliminated. If both frames have the same size, the latter one is eliminated while the earlier one may be enlarged.

5.2.2 Removing stepping-stones

In a dialogue, it is common for two scenes to alternate. This results in redundant smaller frames using the simple thresholds as described. If two frames from the same cluster are only separated by a single frame, the smaller one is removed. If both frames have the same size, the latter one is eliminated while the earlier one may be enlarged. This approach prevents over-selections of repeated pairs of shots, which are common in long dialogues.

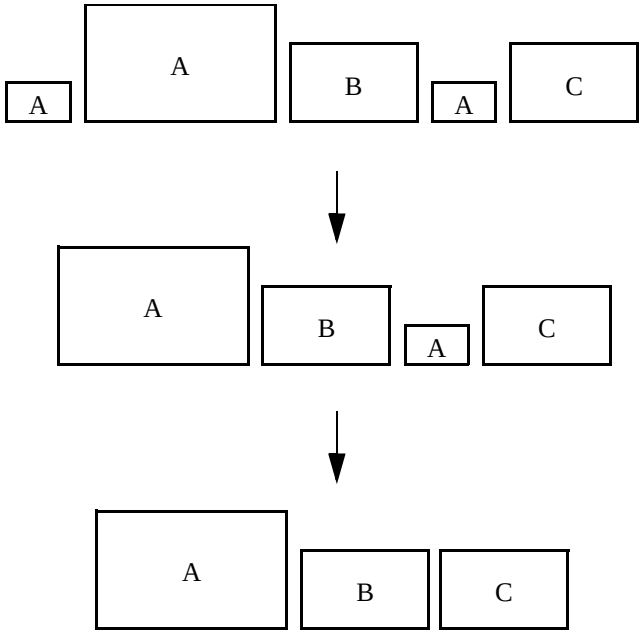

Figure 5. Redundancy Elimination Procedures

Figure 5 shows how these heuristics can reduce the summary size without compromising its effectiveness. Five frames of different sizes are shown, from clusters A, A, B, A, C respectively. The first frame is removed because the next frame is both bigger and from the same cluster (A). The fourth frame is also from cluster A and has the same size as the first, but it is not removed at this point since neither of the adjacent frames is from the same cluster. During the second step, the third frame is eliminated because its one-frame-away neighbor is both bigger and from the same cluster.

Figure 6 shows coverage results of applying the above procedures to the summaries of Figure 4. The number of keyframes is reduced by up to 25 with less than 5 percent degradation in coverage. Thus the summary size has been

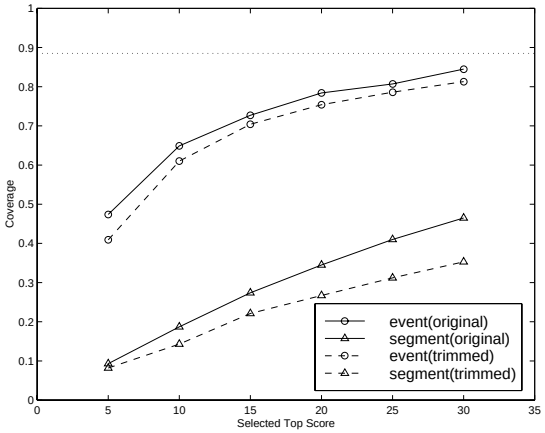

Figure 6. Event and Segment Coverage for Trimmed Selection

significantly reduced without seriously degrading the coverage.

6. ENHANCEMENTS FOR MEETING SUMMARIZATION

6.1 Weighted Importance Score

Although the results of the previous section show that we are generating reasonable summaries, they can be improved even more by considering additional information. A particular advantage of the importance measure introduced in Section 4.2 is that it easily incorporates additional evidence. This section shows how automatically detecting human close-ups and other image types can further improve the summaries.

A particular advantage of the importance measure formulated in (2) is that it can be easily weighted to incorporate other sources of information as shown in (3). A_t is a predetermined amplification factor for category t. $P_t(s_j)$ is an estimate of the probability that shot or segment s_j belongs to the category t.

$$I_j = L_j \log \frac{\sum_t A_t p_t(s_j)}{W_k} \quad (3)$$

The performance evaluation was done using the same video in the previous section. To analyze the video images into various classes, we used the transform approach of [8]. Diagonal-covariance Gaussian models of DCT coefficients were trained on examples of 6 video types such as human close-ups, long shots, crowds, and slides. For each video, frames are extracted every half second and classified using the maximum-likelihood class model. The confidence of correct classification is estimated by normalizing the correct model likelihood by the mean of the other class models likelihood. Because many frames do not match any model, this number gives the confidence that a given frame is a member of the maximum-likelihood class. These confidence values can be incorporated in the importance score so that particular shot classes are favored in the summary. Furthermore, the class confidence score can be weighted to vary its influence on the importance score (and thus its predominance in the summary). It can even be weighted so that it is less likely to appear in the summary by using an amplification factor A of greater than one; this is appropriate for less informative segments such as long shots.

Figure 7 shows how the event coverage is improved by including class confidences in the importance score; close-ups were emphasized using an amplifier A of 0.5, while slides and long shots were de-emphasized by using an amplifier of 2.0. This improves the event coverage ratio, as slides and long shots are less critical to event recognition, as shown in Figure 7. By selecting the top 20 scores, 92 percent of visually recognizable events were covered.

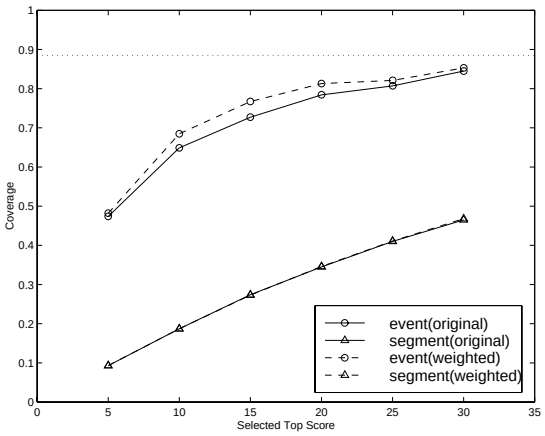

Figure 7. Event and Segment Coverage for Weighted Selection

Figures 8 and 9 show how incorporating the confidence scores improves the summary: the long shot of Figure 8 has been replaced with the close-up in Figure 9. Any other source of information can be similarly integrated into the importance score. For example, laughter and applause are easy to identify in the audio soundtrack, and indicate the conclusion of a significant or amusing segment. We have used image classification in combination with speaker identification to automatically find speaker boundaries in a video, that is, the time extent in which a person is speaking [6]. Other appropriate information might be keywords extracted from time-aligned text or speech recognition. If it could be detected, the audiences' mean heart rate could be used to identify important sections of a video!

6.2 Enhancing Summaries with Text Captions

Many of the videos in our video database depict meetings. Figure 10 shows an example of a summary with meeting minutes displayed as captions. Textual annotations enhance the quality of the video summary. The pictorial layout captioned with text from minutes is a better summary than either of the individual parts. If the meeting minutes were taken on a computer, they can be automatically time-stamped and aligned with the video.

Other video summarization methods use text, typically from closed-caption subtitles or manual transcriptions. However, most informal videos in our domain are not closed-cap-

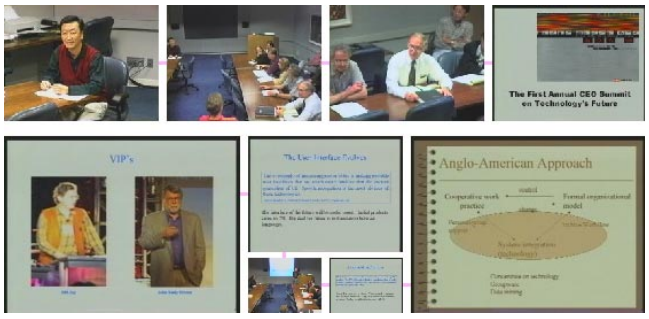

Figure 8. Unweighted Summary

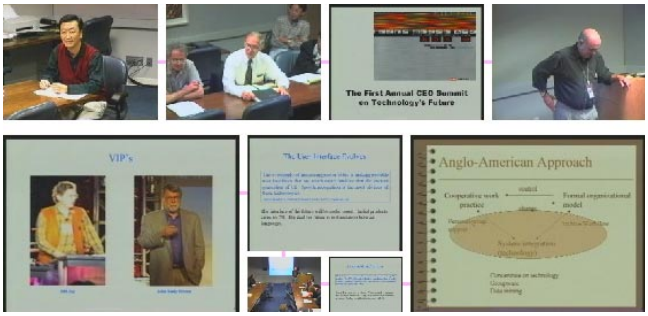

Figure 9. Summary Weighted to Include Human Shots

tioned and transcriptions are not likely to be provided. In our domain of informal meeting recordings using far-field microphones, the current state-of-the-art of speech recognition is not robust enough to be practical at present [20].

In our meeting room, we have instrumented the display system such that a high-resolution screenshot is automatically captured every 5 seconds. This is automatically analyzed using optical character recognition to extract the text of presentation slides, Web pages, and even conventional overhead transparencies (using a rostrum camera instead of an overhead projector). This automatically gives us time-stamped text for summary labeling and searching. We use the text output for labeling summary images, as well as indexing and retrieval by keyword search. The text is processed to remove stop words and formal names are shortened to produce a text summary suitable for inclusion in the video summary.

To display this text as captions, it may be aligned with segments shown with small keyframes or no keyframes at all. These segments need to be represented by larger keyframes so there is room for the captions. This can be done by forcing the corresponding segments to have large keyframes, or by increasing the segment importance scores so that the size of their keyframes increase. The importance score calculation is again easily extended so that the presence of a caption can increase a segment's importance.

6.3 Other Representation

Keyframes need not be closely packed; indeed good graphic design rules stress the use of white space to set off important information. To this end, we have developed a free-form

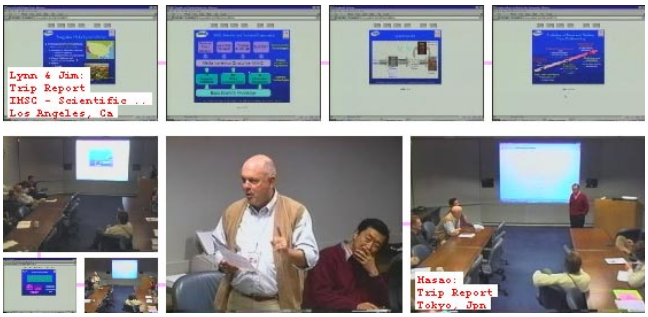

Figure 10. Pictorial Summary with Captions

representation in which keyframes are shown in temporal order along a path; the contents of a video can be obtained by following the path. This layout makes it easy to add captions without covering parts of the images. To create the example shown in Figure 11, automatic speaker boundary detection is applied to a video [6]. Then a frame for the first slide of each presentation is extracted using a frame classification technique [8]. Text is retrieved from the screen image corresponding to the extracted frame for the summary. In the video, there were two announcements followed by two presentations using slides.

7. APPLICATIONS

The video summarization technique shown in this paper is useful for many applications. Any video library, or other collection of videos could use this technique to abstract each video in the collection. Armed with such an abstract, a viewer could quickly find the desired video in even a large collection. Our summaries also help viewers to quickly locate interesting passages within a longer video, using active interfaces (see Figure 12).

We have implemented an interactive version of the pictorial summary as an abstract-level browsing tool for videos. Moving the mouse over the displayed frames highlights the frame and the corresponding segment in the timeline. This display allows users to explore the temporal properties of a video. At a glance, they can see both the visual representation of an important segment and its corresponding time interval.

Once an interesting segment has been identified, clicking on its keyframe starts video playback from the beginning of that segment. Our automatic clustering approach combined with the importance scoring provides good enough segment boundaries to aid the exploration of a video. This interface makes it easy to check promising passages of a video. If a passage turns out to be uninteresting after all, other segments can be easily reviewed just by clicking on their keyframes.

While the initial display provides a good summary of the video, we also implemented a way to explore further. A keyframe shown in a summary is representing a segment, and there might be several neighbors that have importance scores low enough that they are not represented by a keyframe. In our system we have implemented a feature to show the keyframes for such segments. This provides additional context for the one being explored.

The interaction technique for exploring video segments in greater detail has been met with enthusiasm by our pilot users. It promises to be an effective means for rapidly exploring a video without having to wait for playback.

8. CONCLUSIONS AND FUTURE WORK

In conclusion, we have presented methods for automatically generating concise and semantically significant summaries of general videos. The validity of our techniques have been evaluated and proven quantitatively through experimental results. Most prior systems are tuned for specific materials

Figure 11. A "Free-form" Pictorial Summary with Captions

such as broadcast news or documentary films. Although we used meeting videos to evaluate our methods, we stress that our methods will work for nearly any genre of video, as Figure 3 illustrates. Even video sources containing long, few, or relatively unchanging shots can be summarized, though the result will contain many similar keyframes. This accurately captures the monotonous content of the source.

We have also presented several methods to enhance the visual summaries. Weighting keyframe selection based on knowledge of the video contents improved the summaries by including more relevant information. Text information either from manual transcripts or OCR also enriched the results. Our video *manga* captures semantically important events with a compact arrangement of small images, and is thus suitable for Web-based access or other low-bandwidth applications.

The methods we have presented are flexible enough to allow considerable room for improvement. We discuss three possibilities below.

Because our summaries are especially suitable for printing, we have developed paper-based user interfaces for video. Paper summaries are enhanced using barcodes or glyph technology to encode hot links directly on the paper. For example, scanning the code associated with a keyframe calls up the proper video from an encoded URL and starts playback at the appropriate time. We envision a system where meeting participants could be handed printed summaries as soon as they leave the meeting, so that they can review the meeting as desired.

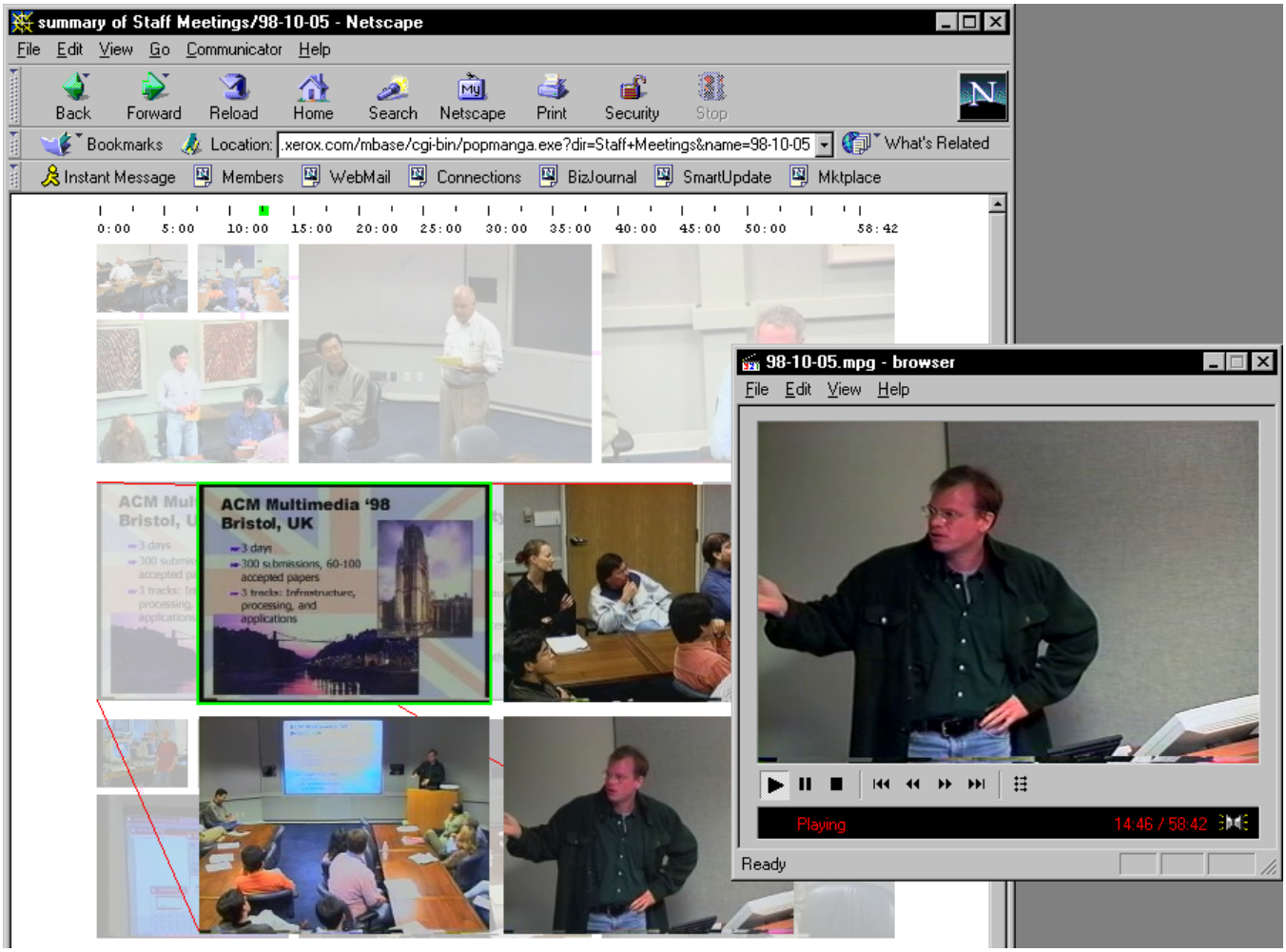

Figure 12. Interactive *Manga* Summary of a Meeting Video

Another enhancement can be made based on person tracking. We can already identify segments that contain human images. We are working on using motion and color information to locate human images precisely in the keyframes. This would allow us to add captions directly on the keyframes without obscuring faces. Ultimately, these could include "speech balloons" to complete the comic-book metaphor.

We are also working on integrating audio information into our pictorial summaries. For an example, the summaries could be enhance by including "sound bites" with each frame. This would be especially valuable for the Web or other environments lacking the bandwidth for full-motion video. A Manga summary with hot links to important segments of the audio could deliver most of the relevant information in the video in a highly compact form. However there are several issues to be solved. Finding and segmenting the relevant audio is a difficult problem. Assumptions can be hardly made to simplify analysis of audio characteristics since speech and music are not always clear or do not even exist in general videos. We also need to consider ways to visualize audio information in the visual summaries. Further research needs to be done on both the signal processing and user interface.

9. ACKNOWLEDGMENTS

Thanks to John Doherty for producing the meeting videos in our corpus. Peter Hodgson helped us tremendously with graphic design issues.

10.REFERENCES

[1] Aigrain, P., Joly, P. and Longueville, V., "Medium Knowledge-Based Macro-Segmentation of Video into Sequences," *Intelligent Multimedia Information Retrieval*, AAAI Press/The MIT Press, pp. 159-173, 1997.

[2] Arman, F., Depommier, R., Hsu, A. and Chiu, M.-Y., "Content-based Browsing of Video Sequences," in *Proc. ACM Multimedia 94*, San Francisco, October 1994, pp. 97-103.

[3] Boreczky, J. and Rowe, L., "Comparison of Video Shot Boundary Detection Techniques," in *Proc.SPIE Conference on Storage and Retrieval for Still Image*

and Video Databases IV, San Jose, CA, February, 1996, pp. 170-179.

[4] Christal, M., Smith, M., Taylor, C. and Winkler, D., "Evolving Video Skims into Useful Multimedia Abstractions," in *Human Factors in Computing Systems, CHI 98 Conference Proceedings* (Los Angeles, CA), New York: ACM, pp. 171-178, 1998.

[5] Foote, J., Boreczky, J., Girgensohn, A. and Wilcox, L., "An Intelligent Media Browser using Automatic Multimodal Analysis," in *Proc. ACM Multimedia '98*, Bristol, England, pp. 375-380, 1988.

[6] Foote, J., Boreczky, J., and Wilcox, L., "Finding Presentations in Recorded Meetings Using Audio and Video Features," in *Proc. ICASSP '99*, Vol. 6, pp. 3045-3048, 1999.

[7] Girgensohn, A. and Boreczky, J., "Time-Constrained Keyframe Selection Technique," in *IEEE Multimedia Systems '99*, IEEE Computer Society, Vol. 1, pp. 756-761, 1999.

[8] Girgensohn, A. and Foote, J., "Video Frame Classification Using Transform Coefficients," in *Proc. ICASSP '99*, Vol. 6, pp. 3045-3048, 1999.

[9] Huang, Q., Liu, Z. and Rosenberg, A., "Automated Semantic Structure Reconstruction and Representation Generation for Broadcast News," in *Proc. IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII*, Vol. 3656, pp. 50-62, 1999.

[10] ISO MPEG 7 Content Set, Item V20, "Korea's Pop Singers' Live Music Show", Korean Broadcasting System, 1998.

[11] Pfeiffer, S., Lienhart, R., Fischer, S. and Effelsberg, W., "Abstracting digital movies automatically," in *Journal of Visual Communication and Image Representation*, 7(4), pp. 345-353, December 1996.

[12] Rasmussen, E., Clustering Algorithms, In W. B. Frakes & R. Baeza-Yates (Eds.), *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, pp. 419-442, 1992.

[13] Shahraray, B. and Gibbon, D. C., "Automated Authoring of Hypermedia Documents of Video Programs," in *Proc. ACM Multimedia 95*, San Francisco, November, pp. 401-409, 1995.

[14] Smith, M. and Kanade, T., "Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques," in *Proc. Computer Vision and Pattern Recognition*, pp. 775-781, 1997.

[15] Taniguchi, Y., Akutsu, A. and Tonomura, Y., "PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing," in *Proc. ACM Multimedia 97*, pp. 427-436, 1997.

[16] Uchihashi, S. and Foote, J., "Summarizing Video Using a Shot Importance Measure and a Frame-Packing Algorithm," in *Proc. ICASSP '99*, Vol. 6, pp. 3041-3044, 1999.

[17] Yeo, B-L. and Yeung, M., "Classification, Simplification and Dynamic Visualization of Scene Transition Graphs for Video Browsing," in *Proc. IS&T/SPIE Electronic Imaging '98: Storage and Retrieval for Image and Video Databases VI*.

[18] Yeung, M. M., Yeo, B. L., Wolf, W. and Liu, B., "Video Browsing using Clustering and Scene Transitions on Compressed Sequences," in *SPIE Vol. 2417 Multimedia Computing and Networking 1995*, pp. 399-413, Feb. 1995.

[19] Yeung, M. and Yeo, B-L., "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," in *IEEE Trans. Circuits and Sys. for Video Technology*, Vol. 7, No. 5, pp. 771-785, Oct. 1997.

[20] Yu, H., Clark, C., Malkin, R. and Waibel, A., "Experiments In Automatic Meeting Transcription Using JRTK," in *Proc. ICASSP 98*, pp. 921-924, 1998.

[21] Zhang, H. J., Low, C. Y., Smoliar, S. W. and Wu, J. H., "Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution," in *Proc. ACM Multimedia 95*, San Francisco, November 1995, pp. 15-24

[22] Zhuang, Y., Rui, Y., Huang, T.S. and Mehrotra, S., "Adaptive Key Frame Extraction Using Unsupervised Clustering," in *Proc. ICIP '98*, Vol. I, pp. 866-870, 1998.