

7-2008

Evaluating Audio Skimming and Frame Rate Acceleration for Summarizing BBC Rushes

Michael G. Christel
Carnegie Mellon University

Wei-Hao Lin
Carnegie Mellon University

Bryan Maher
Carnegie Mellon University

Follow this and additional works at: <http://repository.cmu.edu/compsci>

Published In

Proceedings of the 2008 international Conference on Content-Based Image and Video Retrieval. CIVR '08. , 407-416.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Computer Science Department by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Evaluating Audio Skimming and Frame Rate Acceleration for Summarizing BBC Rushes*

Michael G. Christel
CS Dept. and HCI Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
1-412-268-7799

christel@cs.cmu.edu

Wei-Hao Lin
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
1-412-268-6591

whlin@cs.cmu.edu

Bryan Maher
Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
1-412-268-8970

bsm@cs.cmu.edu

ABSTRACT

For the first time in 2007, TRECVID considered structured evaluation of automated video summarization, utilizing BBC rushes video. In 2007, we conducted user evaluations with the published TRECVID summary assessment procedure to rate a *cluster* method for producing summaries, a *25x* (sampling every 25th frame), and *pz* (emphasizing pans and zooms). Data from 4 human assessors shows significant differences between the *cluster*, *pz*, and *25x* approaches. The best coverage (text inclusion performance) is obtained by *25x*, but at the expense of *25x* taking the most time to evaluate and judged as being the most redundant. Method *pz* was easier to use than *cluster* and rated best on redundancy. A question following the TRECVID workshop was whether simple speed-ups would still work at *50x* or *100x*, leading to a study with 15 human assessors looking at *pzA* (*pz* but with better audio), *25x*, *50x*, and *100x* summaries (these latter 3 with an unsynchronized more comprehensive audio track as well). *100x* gives the fastest time on task but with poor usability and performance. *PzA* gives the best usability measures but poor time on task and performance. *25x* does well on performance as before, with *50x* doing just as well but with much less time on task and better ease of use and redundancy scores. Based on these results, *50x* with its audio skimming is recommended as the best way to summarize video rushes materials.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – *evaluation, video*

General Terms

Experimentation, Human Factors

Keywords

TRECVID, video summarization, video skim, video surrogate, video abstract, user studies, benchmarking, evaluation

*© ACM, 2008. This is the authors' version of the work. It is posted here by permission of the ACM for your personal use. Not for redistribution. The definitive version was published in:

Proc. of the 2008 International Conference on Image and Video Retrieval CIVR'08, July 7–9, 2008, Niagara Falls, Canada.
<http://doi.acm.org/10.1145/1386352.1386405>

1. INTRODUCTION

Video as an information type can take a great deal of time to locate, download, and view. Video summaries can help direct viewers to relevant content, saving effort, network resources, and increasing end user satisfaction. Video summaries can take many forms, for example as text labels, single thumbnails, storyboards of thumbnails, or dynamic slide shows [8, 9]. This paper concentrates on playable video summaries, experimenting with summaries that have durations of one-twenty-fifth (4%) or smaller compared to the target video.

Song and Marchionini note that in the information science literature, a *surrogate* is a condensed representation constructed to stand for a complete information object, and report that *video surrogates* are meant to help people quickly make sense of the content of a video before downloading or seeking more detailed information [7]. Truong and Venkatesh define *video abstracting* as a mechanism allowing the user “to gain certain perspectives of a video document without watching/addressing the video in its entirety” [9]. Christel et al. similarly define *multimedia abstraction* as preserving and communicating “in a compact representation the essential content of a source video” and a *video skim* as a “temporal, multimedia abstraction that incorporates both video and audio information from a longer source” [3]. These terms all describe the summaries studied here, but as the work builds from the TRECVID 2007 BBC rushes evaluation pilot, the paper will use the term “video summary” as that was used most frequently in pilot task reports (e.g., [6]). The TRECVID pilot organizers define a summary as presenting “a condensed version of some information, such that various judgments about the full information can be made using only the summary and taking less time and effort than would be required using the full information source” [6]. This type of video summary is meant to serve both an indicative and informative function as defined in [8], giving the video summary all of the important information contained in the full information source. In a world of information overload, summaries have widespread application as compact surrogates returned by searches as previews, or used to give someone an efficient overview of a vast or unfamiliar video collection [6].

This paper details two evaluation experiments for the TRECVID 2007 BBC Rushes Summarization track. Our Carnegie Mellon University (CMU) Informedia research group has investigated the utility of automated video summarizations for news and documentaries, i.e., for produced materials, since the mid-1990s [3]. However, most of the Informedia summaries were based on produced broadcast news and documentaries, with redundancies edited out, and with good automatic speech recognition transcripts

available. In contrast, the BBC rushes are video takes from before the editing process, with much redundancy and mixed quality audio.

Section 2 overviews the TRECVID assessment framework for evaluating BBC rushes summaries. Section 3 reports on the different fully automated summarization techniques used for the first experiment. Section 4 discusses the first experiment, with Section 5 discussing the motivation for and development of additional techniques. These techniques are tested in a second experiment, reported in Section 6, with ending sections offering discussion and presenting conclusions based on these two empirical studies with human subjects.

2. SUMMARY ASSESSMENT FRAMEWORK

In the TRECVID 2007 task of BBC rushes summarization, there were 42 individual rushes videos in the test set, and a maximum size of 4% duration (1/25 of the target) for the video summary. Twenty-two research teams participated, each submitting a single video summary for every test video. Further details on the tasks and results are presented in the overview paper [6], with this section summarizing from that work so that the procedure and terminology in follow-up experiments reported here can be fully understood.

The 2007 TRECVID evaluation pilot provides a reasonably large video collection to be summarized, a uniform method of creating ground truth, and a uniform scoring mechanism. The video data consisted of raw (i.e., unedited) video footage, shot mainly for five series of BBC drama programs. The data was provided to TRECVID for research purposes by the BBC Archive. The drama series included a historical drama set in London in the early 1900s, a series on ancient Greece, a contemporary detective program, a program on emergency services, a police drama, as well as miscellaneous scenes from other programs. About 50 videos were provided to participating groups as development data and 42 were withheld for use in testing the systems once developed. Each set of videos represented a random sample balanced with respect to the number of videos from each series. The test videos had a minimum duration of 3.3 minutes and a maximum duration just under 36.4 minutes, with the mean duration being 25 minutes. Sample ground truth was provided for about half of the development videos and ground truth was also created for the test videos.

The system task given to participants was an abstraction of a real world video summarization task: given a video, automatically create a generic video summary by compressing the original video to remove redundant and unclear footage. The summary was to be constructed to maximize a viewer's efficiency in recognizing the main (primarily visual) objects and events from the original video as quickly as possible. To simplify evaluation, each summary was limited to a single MPEG-1 video file of a maximum duration of 4% of the target video, which would be displayed during evaluation using the original video's frame rate and size.

The quality of each summary was evaluated directly by objective and subjective means. Subjective measures included the fraction of important segments from the full video included (IN), how easy it was to find the desired content (EA), and how much redundant video the summary contained (RE). An objective measure was the ease of understanding the summary content as reflected in assessor time-on-task (TT) judging which ground truth segments were included in the summary. Time on task was recorded as the time

spent watching the video summary, including time spent in pause. The human assessor could only play the video summary once at normal speed but could pause the playback an unlimited number of times.

At NIST, 7 retired adults with computer skills spent a total of 221 hours judging the 1008 submitted summaries (22 research groups plus 2 baseline systems), using software written by NIST for that purpose. Each submitted summary and each baseline summary of each of the 42 test videos were judged by three different assessors. The assessment interface is the same one used for the experiments in this paper, with the same phrase sets used for each of the test videos.

Procedures for developing ground truth lists of important segments from each BBC rushes video were developed at Dublin City University. Full details are in the overview paper and its appendices [6], with the result being a text phrase describing an important object, possibly modified with event (e.g., "red hot air balloon ascending"), camera pan/zoom event (e.g., "pan across room"), or both. Each human judge (assessor) was given the summary for a video and a chronological list of up to 12 phrases randomly sampled from a longer (on average 24-item) ground truth list from the original video content. The assessor viewed the summary once in a 125 mm x 102 mm playback area with only pause/play control, determining which of the designated segments appeared in the summary.

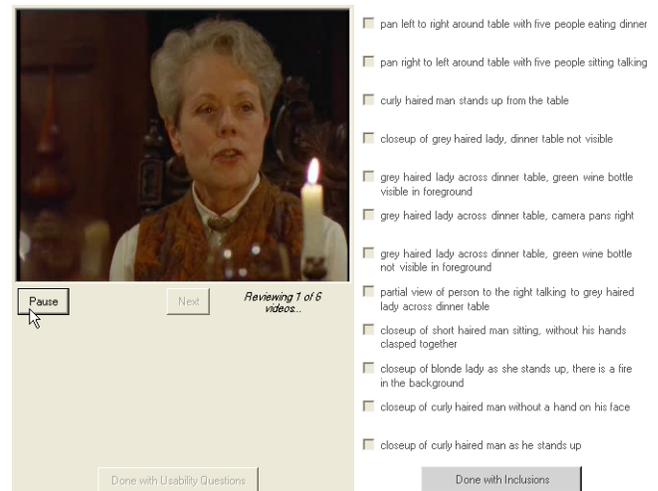


Figure 1. Assessment interface, step 1: review video and text inclusion list.

For each target test video, the assessor began with the interface as shown in Figure 1. The assessor was instructed via a paper form to play the target video, a five times real-time overview of the full video being summarized with no audio, as many times as desired while studying the list of segments for judging. The assessor then graded one summary at a time, shown as in Figure 2, with the ability to check on/off the text phrases (ground truth) judged as being in the summary. The assessor could pause as much as desired, but not reseek with the summary being played at the target video's frame rate (25 fps) only once.

The timer stopped when the assessor clicked "Done with Inclusions" to mark the end of the ground truth judgment task. The assessor then graded the summary on ease of use (EA) and redundancy (RE) using 2 questions: "It is easy to see and understand what is in this

summary” and “This summary contains more video of the listed inclusions than what is needed” as shown in Figure 2 (left labeled strongly agree, right strongly disagree). For follow-up analysis, NIST decided to have 1 indicate poor and 5 excellent on these scales, so the assessor selection was shifted from [1, 5] on the EA question to [5, 1], so that an EA reporting of 5 meant strongly agreed easy to use and RE of 5 meant strongly agree no redundancy. The assessor graded summaries grouped by their target video. The order of presentation of the summaries for a target video was randomized to randomly assign any bias due to learning effects.

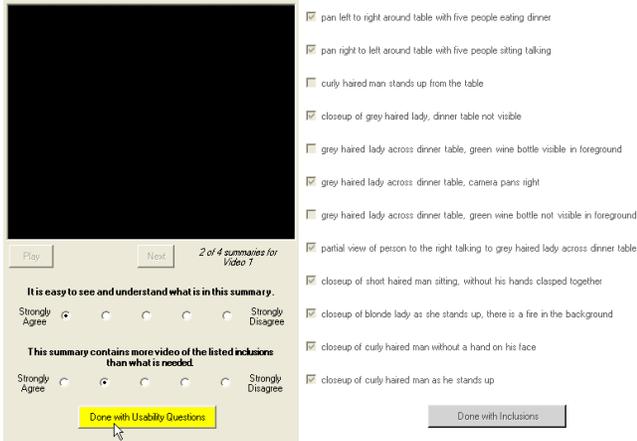


Figure 2. Assessor interface, steps 2 and 3: play summary which shows at upper left, pausing at will; then fill out 2 5-point ratings (with completed playback area now blanked) before continuing on to next summary for the target video.

3. AUTOMATED SUMMARIZATION TECHNIQUES

As a control to help in gauging success for the TRECVID summarization task across all participants, our Informedia group at Carnegie Mellon University developed a very simple baseline approach to 4% summary generation. Building blocks of one second each were chosen based on that duration being close to the lower limit that humans can comfortably recognize non-trivial visual content on the screen, e.g., text overlays on the screen are always shown for at least that long. Research on automatic shot detection makes use of empirical observation and also chooses one second as minimal shot duration [4]. The one-second building blocks were then trivially assembled as follows: divide the target video (i.e., the video to be summarized) into segments of 25 seconds each and then include the middle second from each segment into the summary. This baseline was labeled CMUBASE1, the uniform sampling baseline. Despite not taking into consideration any sort of noise-shot filtering or skipping over the often noisy lead-in to rush videos such as lengthy color bar shots, this CMUBASE1 still proved quite difficult to beat in the NIST-conducted TRECVID summary evaluations [6].

Encouraged by the TRECVID summary task organizers, we also tried a more sophisticated baseline using simple color clustering. Using our own shot boundary detector, we lowered the threshold of sufficient differences between adjacent frames to detect a shot compared to broadcast news, allowing any dramatic motion to create a shot change. Hence there were more shots than normally seen in edited broadcast video, with 25423 shots in the test set of 42 videos

(average shot length with such oversampling: 2.64 seconds). From the start of each shot (near the dramatic change) we extracted a keyframe, and partitioned this into a five by five grid. In each grid cell, we extracted the mean and standard deviation of hue, saturation and value (HSV color space). One keyframe from each shot was used in per-video k-means clustering, with the number of clusters set to the number of seconds (rounded down) in the 4% summary. For example, with a 10 minute video (600 seconds), we would have a target summary length of 24 seconds (4%), and therefore cluster the data into 24 clusters. From each cluster, one second from the middle of the shot closest to the centroid was included in the summary. We did not consider merely displaying the keyframe for one second, as events frequently involve actor and/or camera motions, which would be lost in any static representation. This second baseline was labeled CMUBASE2, the simple clustering baseline, and as with CMUBASE1, it did not incorporate any noise shot filtering.

Based on the reports and demonstrations from the TRECVID Video Summarization Workshop [1], most participants in this task did attempt noise shot reduction, eliminating irrelevant shots as shown in Figure 3.

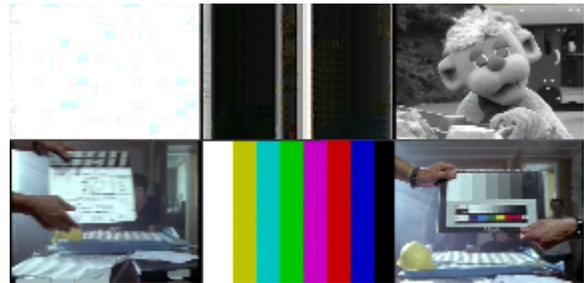


Figure 3. Examples of 6 types of irrelevant shots seen in BBC rushes video: white frame, black frame, grayscale image, clapper, color bar, and color calibration chart.

Such attempts at noise reduction, even when coupled with additional work to automatically eliminate redundancy and emphasize important sequences in the target video, did not often result in video summaries that were markedly better than CMUBASE1 or CMUBASE2. Based on the inclusion (IN) metric, the fraction of ground truth judged to be in the summary, the two baseline systems performed in the upper quartile of all systems. A partial randomization test found no significant difference between the baselines at the 0.05 level of significance. The same test found 3 of 22 systems significantly better than both baselines but indistinguishable from each other [6]. For EA, with one or two exceptions, most systems scored nearly the same, clustering around neither strongly agree nor strongly disagree that the summary was easy to understand and use. The randomization test found the baselines indistinguishable from each other and only one system significantly better than the baselines with respect to ease of understanding and use. The randomization test found no significant difference between the baselines on the redundancy (RE) measure, but most systems were significantly better than the baselines.

For our CMU NIST-judged run, we decided to focus on a few specific summarization features:

- (1) How would our own noise-filtering improve on the baselines, removing clearly irrelevant material as illustrated in Figure 3?

- (2) What if we improve the audio component of the summary?
- (3) What if we forfeit redundancy and try to emphasize inclusion/recall by simple video frame rate acceleration rather than extracting one second windows?
- (4) What if we forfeit full inclusion by emphasizing sequences of importance – pan/zoom sequences – and fold in variable rate playback?

3.1 The *cluster* Video Summary

Our official submission into the NIST-run evaluation, *cluster*, was created based on iterative color clustering with noise filtering, backfilling of unused space and audio coherence, and is described elsewhere in more detail [5]. We had noticed that the summaries we created and iterated with on the development set of rush data included a number of shots that were clearly irrelevant, as illustrated in Figure 3. For each of these types we built automatic detectors that tried to identify the classes of frames for future exclusion [5].

The next step after k-means clustering was to eliminate all clusters which were predominantly composed of a clearly irrelevant class. We were now left with fewer clusters, and so we clustered the data again to end up with the targeted number of clusters, with one cluster for each second of video in the target summary. From these clusters, we selected the first second of the shot whose keyframe was closest to the cluster centroid.

Since some shots were shorter than one second, we were again left with a little extra room in the summary (below our target of 4%). To maximize the IN score, we wanted to include as much and as diverse information as possible. The leftover space was ‘backfilled’ by selecting one second from a shot that was furthest from a cluster centroid, effectively an outlier. The procedure was repeated until the resulting summary was just below 4%.

We did not mute the summaries since we felt that understanding the acoustic context would help to more quickly understand the visual events. We ran automatic speech recognition on the audio track to identify speech and non-speech regions. Given an edit list of segments based on visual characteristics, we selected the corresponding time boundaries in the ASR transcripts, and determined which edits contain speech and where silences separated the speech transcripts using Signal-to-Noise Ratio calculation. Earlier research on skims [3] has shown that choppy audio is very distracting, and in that research we had successfully used the SNR segmentation to obtain reasonable acoustic phrases in news skims. For *cluster*, we initialized an audio edit list with the mid-point of each visual edit instruction, found the nearest SNR boundaries to each audio edit segment, and extended the currently shorted audio edit segment to this boundary. The process stopped when the total duration of the summary (4%) was reached. This simple approach favors playing coherent, recognizable audio segments, related to the visual segments, but loses full audio/video synchronization. Keeping some audio representation in a multimodal video summary was a recommendation from an earlier empirical study [7], which also advised that tight audio-visual synchronization may not be necessary in a video surrogate.

3.2 A Simple Speed-Up Summary: *25x*

Our research group debated intensely over which one of our automated methods should be submitted to NIST for evaluation.

Should we emphasize aesthetics over INclusion, how much time does a viewer need to identify a pan/zoom, should detected faces or people be given a priority, is there a role for audio, does the audio need to be synchronized as earlier work showed that news summaries with asynchronous audio were jarring – all were questions we considered. Among the most heated discussion was whether a simple *25x* summary, which merely speeds up the playback by selecting every 25th frame, was too simple and therefore embarrassing to submit to evaluation, even though our informal tests revealed it would likely score very high on the INclusion metric, but also required much effort to watch.

By simply sampling every 25th frame, you create a 4% video summary, which we label *25x*. We will use this labeling convention throughout, that sampling every Nth frame produces a *Nx* summary which appears to play back at N times normal speed. The audio is incomprehensible at *25x* playback, but some of the BBC rushes dialogue seemed to hold value based on casual inspection of the development data. So, we wanted to augment the *25x* video with a regular speed narration. We chose 4% audio content based on the algorithm used create the audio associated with the *cluster* summaries. For the visual component of the *25x* summary, select every 25th frame with no consideration given for noise-filtering.

3.3 A Domain-Specific Summary: *pz*

We noted in the instructions to the task that camera events, i.e., pans and zooms, were emphasized as being important. This serves as a form of domain expertise: for future users skimming through summaries of BBC rushes, they will likely want to identify pans and zooms. Rather than hope that our cluster method somehow captures pans and zooms well enough, we created a *pz* method as follows that still makes use of the clusters discussed in Section 3.1:

1. All pans and zooms longer than 1 second are automatically tagged. All clusters are identified as in Section 3.1.
2. Each cluster is represented in time order in the summary. If a cluster has a pan or zoom, the longest one is used to represent the cluster. Otherwise, the representation is chosen based on having video with faces (we assumed faces to be important to humans) and not noise video, where noise video includes color bars, white shots, and clapper shots.
3. If no face video and no pan/zoom exist for the cluster, the cluster representation is as done for Section 3.1.
4. Pans/zooms are kept in up to 6-second runs, using the central 6 seconds if the identified run was longer. To save time in the summary, however, pan/zoom sequences longer than 2 seconds had their durations cut in half by sampling the video at twice normal rate but using the first half of the audio (so audio playback is normal rate).
5. If the resulting summary is too long, pans/zooms are shortened down to 1 second in length as needed until we reach 4%.

The *cluster*, *pz*, and *25x* summaries were all less than the upper bound of 4% of the original video’s duration for each of the 42 test set videos. The *cluster* summary did not distinguish itself from the two baselines in NIST assessment, as shown in Table 1. We ran the same assessment with CMU judges to gauge any relative differences between the *cluster*, *25x*, and *pz* summaries.

Table 1. NIST TRECVID official results for *cluster* and two baselines.

	CMUBASE1	CMUBASE2	<i>cluster</i>
TT (secs.)	105.66	100.48	101.83
IN	0.59	0.58	0.60
EA (5 best)	3.44	3.41	3.37
RE (5 best)	3.52	3.50	3.62

4. USER EVALUATION: *cluster*, *25x*, *pz*

Four CMU students and staff (3 male, average age 28) conducted two passes through the 42 test videos, resulting in 84 summary assessments using the NIST protocol presented in Section 2. Assessors played the same 5-times speed overview as in Figure 1, and then judged all the summaries for the target video as shown in part in Figure 2. The assessment order of the 3 summary types was counterbalanced to remove any bias due to learning and reinforcement effects.

The announced pairwise agreement in NIST-judging which of the (up to 12) desired items from the full video were included in the summary was on average 78% [6]. The agreement between our CMU assessors was 80.6%. We tested our *cluster* again to see how well CMU assessors agree with NIST assessment, and the numbers correlate well for IN and EA, correlation coefficient $r=0.8$ and 0.86 , NIST IN means for IN and EA 0.6 and 3.37, CMU assessors' means 0.61 and 3.06 respectively. For TT and RE ($r=0.43$ and 0.24), CMU assessors took a bit more time (likely because they only had 3 summaries per video to grade) and were more lenient on redundancy: NIST TT and RE means 101.8 and 3.67; CMU assessors 109.9 and 4.17 respectively. These are for the same exact *cluster* summaries on the 42 test videos graded at NIST and then later at CMU.

The point of the comparison between CMU and NIST grading is not to check NIST's grading accuracy, but to note that IN and EA numbers are comparable so that in later discussion these numbers for the CMUBASE1 and CMUBASE2 NIST-judged runs can be contrasted with the summary forms directly assessed at CMU. The main point of the 4-judge user evaluation was to see relative differences between *cluster*, *25x*, and *pz*. Figure 4 overviews the differences on the TT, IN, EA, and RE measures.

Significant differences were found using ANOVA, 2 degrees of freedom, $p < 0.002$ across all four measures: $F=8.14$ for TT, $F=82.83$ for IN, $F=6.66$ for EA, and $F=119.51$ for RE. The Tukey HSD test confirms the following significant differences at $p < 0.01$: for TT, *25x* is slower than the others; for IN, *25x* produces better performance; for EA, *cluster* is worse than *pz*; for RE, *25x* is worse than the others.

If the main objective of the summary is to maximize recall of text inclusions, i.e., produce the highest IN score, then *25x* is an excellent method, with its 0.87 mean (0.92 median) far outstripping these other two runs and all other NIST submitted runs whose IN means ranged from 0.25 to 0.68 as graded at NIST. Such excellent performance comes at a cost: the TT metric for *25x* was higher (but still exceeded by some of the NIST graded runs), and the acknowledged redundancy in the *25x* summary was quite high (the RE measure). RE and EA were included as metrics to help with assessing utility and end-user satisfaction, but while *25x* was

acknowledged as redundant, its ease of use measure (EA) was actually better than that for *cluster*. Such conflicts in assessing video summaries are discussed further in [8]: optimizing for one parameter like ease of use often comes at the expense of another like redundancy. The NIST overview report notes another such conflict with worse RE often leading to better IN: "redundancy does seem to make it more likely the ground truth items will be included and found...perhaps because it makes the assessor's job easier" [6]. We believe the inclusion of an audio narrative made the *25x* video summary more playable by end users, improving its EA score despite its high redundancy.

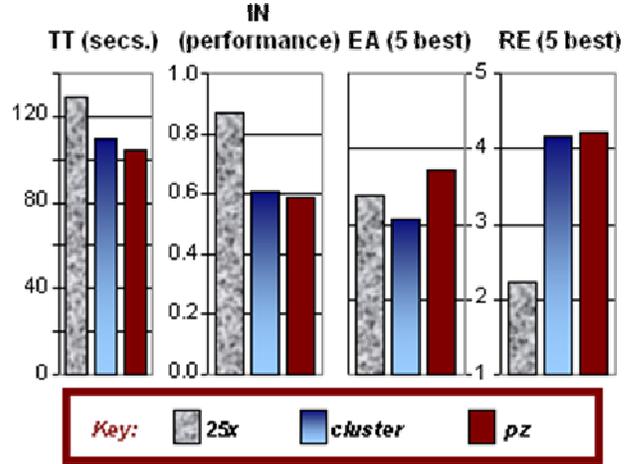


Figure 4. Mean TT, IN, EA, and RE collected from 84 evaluations for each of *cluster*, *25x*, and *pz* summaries using NIST protocol (conducted twice across 42 test set videos).

If the main objective is to produce a video summary type that users would not mind playing over and over, then of course additional satisfaction metrics and longitudinal studies could be employed to better address that objective. Even with just EA and RE, though, *pz* shows itself to be an improved summary type than *cluster* by bringing in some domain knowledge. Namely, for video like BBC rushes where color bars, all white shots, and clapper bars are noise, people are important, and pans and zooms are likely to be looked for later, an emphasis on pans and zooms first, then faces, and dropping out noise works well for EA and RE as a strategy. The EA measure for *pz* was significantly better than that for the *cluster* method which did not emphasize pans or zooms, and its RE mean was the highest as well for the 3 tested methods. One reason for little separation on TT and IN between *pz* and *cluster* is the large overlap in the automated methods to produce each, and especially the steps 2 and 5 (Section 3.3) for *pz* where pans/zooms are dropped rather than clusters being dropped when the assembled edit list to produce the summary is too long in duration. Future work includes testing more aggressive *pz* methods that preserve pans and zooms at the expense of clusters and anticipated coverage, i.e., rather than shorten pans/zooms, drop clusters. One immediate concern we had, though, was that the audio of *pz* did not make use of the audio composition strategy discussed in Section 3.1. Instead, it kept the audio synchronized to video, so that when small clips of video are composed together via the process of Section 3.3, small audio clips were joined together as well, even if they broke in mid-word or expressed something inaudible or without any speech. We improved the audio of the *pz* method by using the same audio as in

25x and *cluster*, relabeling this strategy *pzA* and testing it in a second empirical evaluation discussed below.

5. MOVING TO MORE ACCELERATION

CMU developed an automated inclusion score (IN score) metric against the development data for use in post-hoc analyses into various aspects of summary methods like *cluster* [5]. One such analysis looked at the effects of different compression rates using *cluster*, CMUBASE1, and CMUBASE2. Figure 5 shows that based on automatic evaluation of INclusion, the iterative clustering slightly outperforms the baseline uniform result (CMUBASE1) as well as the baseline clustering result (CMUBASE2) at 4%. This difference between approaches shrinks at lower summary compression rates, but increases as the target summaries become shorter. This data hints that for the BBC rushes, a 4% or longer summary may not show much relative difference in IN score, regardless of its construction, and in fact the NIST-judged summaries only rarely differentiated themselves from the baselines [6]. However, at 2% there are vast differences in CMU’s iterative clustering (*cluster*) and the baselines, with iterative clustering still producing a good automated IN score.

It could be that CMU clustering algorithms can find unique shots up to a certain level with this set of BBC rushes video data. After that, the clustering performance reaches a plateau and uniform sampling/speed-up summaries start to dominate the inclusion scores. From Figure 5, iterative clustering can find much more unique shots than baselines before 2%, but after 2% iterative clustering fails to find more unique shots (note the leveling off in Figure 5 for *cluster*), and baselines start to dominate the inclusion scores.

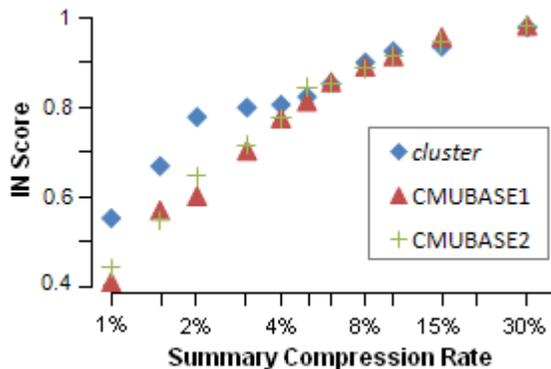


Figure 5. Comparison of different summary compression rates vs. automatically determined INclusion scores of labeled events from the training (development) data, with x axis in log scale.

25x speed-up summaries are shown in Section 4 to be superior in including important events in the BBC rushes, but how fast can we speed up summaries? Clearly we cannot keep speeding up a video and expect speed-up summaries of Nx rates with N growing infinitely large to still include most of the events in the original video. However, we do want to optimize summary space, too. If 25x and 50x speed-up summaries achieve similar performance in terms of inclusion scores, we prefer 50x because it is shorter. Hence, there appears to be a trade-off between summary length and the number of events included in a summary.

We suspect that the key factor to the length-inclusion score tradeoff in speed-up summaries is an event’s redundancy. An event repeated only few times is very likely to be missed in speed-up summaries. On the other hand, an event repeated many times is very unlikely to be missed. We illustrate the idea in the following thought experiment. Consider a video with length of 100 units. There is only one event of length one unit in the video. If the event is repeated n times, how likely is it that a k -x speed-up summary includes the event? The probability that the event is included in a speed-up summary is

$$\Pr(\text{hit}) = 1 - \left(1 - \frac{n}{100}\right)^{\frac{100}{k}}$$

We plot the probabilities that 5x, 25x, 50x, and 100x speed-up summaries include an event that repeats from $n = 10$ times to $n = 100$ times in Figure 6. The likelihood that an event is included in a summary is positively related to the summary’s inclusion score, and thus we can regard the probability as a surrogate inclusion score. When an event is repeated as few as 20 times, there is a significant difference in the probability between the speed-up summaries: 5x is 99% likely to include the event, but 25x is only 60% likely to include the event. 5x summaries obtain high inclusion scores by paying in greater summary length to obtain more samples. On the other hand, when an event is repeated as many as 80 times, there is little difference between 5x, 25x, and even 50x summaries. 5x and 25x speed-up summaries then become unwise choices for highly repeated events, because the 50x summary is significantly shorter and still very likely to include the event.

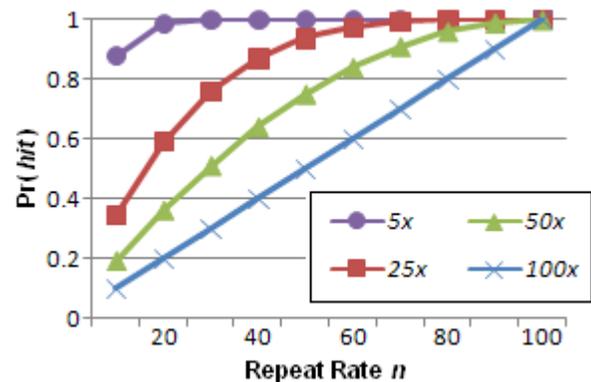


Figure 6. The probability of an event detected by 5x, 25x, 50x, and 100x speed-up summaries (y axis) vs. the number of times the event is repeated in the original video (x axis).

We know that the BBC rushes contain highly redundant material based on 25x summaries’ high inclusion scores in the previous study, but it is still not clear how redundant the BBC rushes actually are. 25x or 4% summaries, after all, are arbitrary, and may not be an optimal summary length for the BBC rushes. As argued in the above thought experiment, we may significantly shorten summary length without losing important events in the original video. Therefore, we decided to empirically study 50x and 100x speed-up summaries on the BBC rushes.

From a review on video abstraction comes the caution that “in order to ensure that humans do not perceive any discontinuity in the video

stream, a frame rate of at least 25 fps is required” [9]. The BBC rushes have a high degree of visual redundancy, however, often repeating the same scene in multiple “takes.” There is the likelihood that the summary of 25x worked well precisely because of this redundancy: the first time through, some events may have been missed, but as the 25x summary represented every repeated take, the takes are shown over and over with the viewer filling in and grasping what otherwise might have been missed if every take were represented exactly once. The follow-up study looks to see whether for such highly redundant material as the BBC rushes, further acceleration to 50x or 100x will work well.

There is some evidence that even for produced documentary materials without the redundancy of rushes, fast-forward surrogates with accelerated playback can be effective. Wildemuth et al. tested 32x, 64x, 128x, and 256x video surrogates (video that samples every 32, 64, 128, or 256 frames, with no audio component) with four target documentary video segments of lengths 9:19, 14:00, 14:09, and 19:48. They conclude from an empirical study with 45 participants and six measures of human performance that 64x is the recommended speed for the fast forward surrogate, supporting good performance and user satisfaction [10]. They note that for videos less than 10 minutes, 64x “does not produce enough frames to create a fast forward surrogate of useful length” and so plan to use 32x for shorter videos.

This study indicates that 50x and perhaps 100x will achieve success using the NIST rushes summary evaluation protocol. Based on the earlier success with 25x keeping an audio track, we added audio to 50x and 100x using the same procedure as described in Section 3.1, except that the target size was set at 2% and 1%, respectively.

6. SECOND USER STUDY: 25x, pz, 50x, 100x

Study participants were recruited through the “Experiment Scheduling Site” web page provided by the Center for Behavioral Decision Research at Carnegie Mellon University. This page attracts subjects from the Pittsburgh community within walking distance of the University of Pittsburgh and Carnegie Mellon University, predominantly but not exclusively college students. The 15 subjects (8 female, 7 male; age range [21, 35] with average age 25.7) who participated in this study had no prior experience with the interface or data under study and no connection with the research group. These subjects were given the same instructions and interface used in the NIST assessment. They were asked to complete up to 6 groups, with each group starting with the interface of Figure 1 for a target video followed by 4 summaries for that video: 25x, 50x, 100x, and pzA. The order of presentation of the summaries for a target video was randomized to randomly assign any bias due to learning effects. 13 of the subjects completed 6 groups in 50-70 minutes. One subject stopped after 55 minutes and an assessment of 5 groups. The fifteenth subject was also given 6 groups of video, but only the one leftover video from the subject assessing only 5 was kept for the final analysis, which consisted of 2 judgments each across the 42 test videos.

Figure 7 overviews the differences on the TT, IN, EA, and RE measures. Significant differences were found using ANOVA, 3 degrees of freedom, $p < 0.0001$ across three measures: $F=91.06$ for TT, $F=24.31$ for IN, and $F=11.59$ for EA. Significant differences were found with the RE measure as well, $F=3.94$, $p = 0.009$. The

Tukey HSD test confirms the following significant differences at $p < 0.01$: for TT, 25x and pzA are both slower than 50x and 100x, and 50x is slower than 100x; for IN, 25x and 50x both produce better performance than 100x and pzA; for EA, 100x is worse than the others; for RE, 25x is worse than pzA (and, at $p < 0.05$, 25x is worse than 100x).

If the main objective of the summary is to maximize recall of text inclusions, i.e., produce the highest IN score, then 25x confirms itself to be an excellent method, with its 0.73 mean superior to all other NIST submitted runs as graded at NIST. Notably, however, the IN performance of 50x is also excellent with mean 0.68, not statistically different from 25x in this experiment. Such excellent performance for 25x comes at a cost not incurred by 50x: the TT time on task metric for 25x was the highest (along with pzA) of those evaluated, with 50x taking only two-thirds of the time.

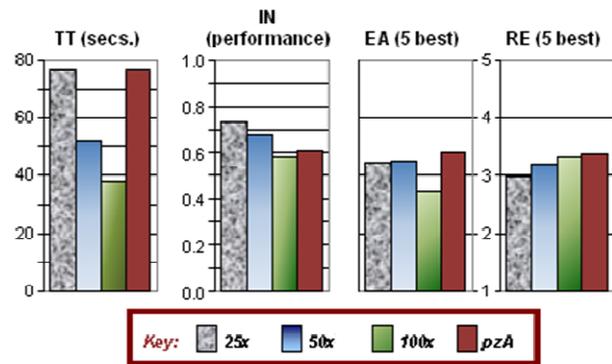


Figure 7. Mean TT, IN, EA, and RE collected from 84 evaluations for each of 25x, 50x, 100x, and pzA summaries using NIST protocol (conducted twice across 42 test videos).

Looking closer at the TT results, note that the duration of 25x is 4% of the target video, with 50x only 2% of the target video and 100x only 1% of the target video. If the summaries were never paused in the assessment interface of Figure 2 and judgments of inclusions occurred on the fly as the summary video played, then the TT results would show 100x being a quarter of 25x and 50x being twice that of 100x and half that of 25x. This linear pattern did not hold because the assessors did pause the summaries. Before the study, we were not sure whether 100x would be paused so much more frequently than 50x, and 50x more than 25x, to the level that the summary duration differences would not translate into TT differences. Figure 7 shows that such increased pausing as sampling goes from 25x to 100x did not flatten out the TT results. While the TT differences are not mapped exactly to duration, there is the significant separation with 100x taking on average only 38.5 seconds, 50x taking 52.8 seconds, and 25x and the other 4% summary, pzA, taking the longest, at 77.5 seconds and 77.4 seconds respectively.

The fast time on task for 100x comes at a cost. It is a significantly worse performer on IN vs. 25x and 50x, and it also is the worst summary in terms of ease of use. The subjects clearly felt that 100x was not a satisfying summary.

As for the RE measure, the acknowledged redundancy in the 25x summary was quite high, significantly different from both 100x and pzA. RE and EA were included as metrics to help with assessing utility and end-user satisfaction, and for both measures, 50x was not scored significantly differently from the top-rated summaries in this

experiment. We believe the inclusion of an audio narrative along with the sped-up video made the 25x video summary more playable by end users, and that this audio narrative helped ease of use even when shortened to half its duration at 50x. Only at 100x, with the audio duration now at 1% of the target video length, did usability suffer because at this extreme skimming rate the audio segments put into the summary became choppy and brief, at the word level rather than phrase level, exactly the characteristics of audio found to be problematic in prior video skim evaluations [3].

These 15 participants, with no connection to the research group or familiarity with the rushes videos or summarization task, did not have the same level of care as the assessors recruited at NIST or for the study reported in Section 4. These participants spent much less time with each summary assessment task. Ignore the 1% and 2% compression rate summaries for now and consider only the 4% ones for 25x and *pzA*: on average, the participants spent 77.5 seconds assessing a summary, whereas in Section 4 (with 4% summaries) an average of 129.1 seconds were spent (109.9 with *cluster*), and NIST assessors spent an average of 101.8 seconds with *cluster*. They were more centrist in their ratings on EA and RE as well (i.e., more likely to choose 3 on the 1-to-5 scale). The result was a depressed level of inclusion performance for 25x, which was the same exact summary tested by both assessors: Figure 4 shows the mean IN score at 0.87 with Figure 7 showing the 15 participants produced an IN score mean for 25x of 0.73. The faster TT had another consequence: the pairwise agreement in judging which of the (up to 12) desired items from the full video were included in the summary was on average 65.4%. Broken down by summary type, the pairwise agreement for 25x, 50x, 100x, and *pzA* was 68%, 67%, 62%, and 64%, respectively.

The dramatically lower time on task showed a compelling motivation for these subjects – many of whom were students with busy class schedules – to be done with the study as quickly as possible with TT for them being more important than maximizing accuracy on IN. This may be insightful for temporal video summary work in general: end users may be most interested in the time savings as reflected by TT, and as long as ease of use is not too onerous (as with 100x and its EA score under 3), faster summaries are best. Hence, 50x is even more compelling as the option of choice over 25x based on its significantly faster TT.

7. DISCUSSION

Taken collectively, a few strong patterns emerge from these two experiments. Simple speedup, at least when accompanied by discernable spoken narrative, at 25x and 50x work extremely well for the IN performance metric, covering the target video better than all other submitted NIST runs and other methods tested here. As concluded by Song and Marchionini [7], multimodal surrogates have value, even if the audio and video are not fully synchronized. In an empirical study [7], participants were able to easily use the combined aural and visual multimodal surrogates even though they were not synchronized, suggesting that synchronization of different media channels may not be necessary in surrogates as it is in full video.

The experiments here began with a premise, that audio has value, and we included audio in all tested summaries. Follow-up experiments could test the premise directly by running similar within-subjects experiments using the NIST BBC rushes summary assessment protocol as described here against 25x with and without

audio, and 50x with and without audio. Perhaps the baseline summary for the TRECVID 2008 BBC Rushes summarization task should be 50x without any audio, with at least one group submitting 50x with audio to isolate any benefits that extra channel offers.

Prior work has already established that audio, if not properly composed, can severely detract from video summaries. If the audio track for a summary is composed of very brief snippets that crop words, it leads to decreased satisfaction [3]. If audio playback is accelerated its pitch changes and comprehensibility drops, and if natural speech boundaries are not respected listeners are negatively affected; Arons overviews the issues well in his seminal audio-only summary work with SpeechSkimmer [2]. Enough NIST summarization task research participants submitted summaries without much if any concern for the accompanying audio that in the TRECVID 2007 workshop it was noted that NIST assessors, becoming frustrated with incomprehensible or annoying audio, turned off the speakers or refused to wear headphones. In the overview report the organizers comment that one assessor “noted that listening to the audio was unnecessary and distracting” [6]. This may be true for even a majority of submissions, but what if a research group tries to carefully craft the audio to provide a multimodal video summary of greater value? Logistically, how can review of the video summary audio be encouraged, or even required? Perhaps future video summarization evaluations should force the listening of the audio as well, with groups encouraged to mute the audio completely (as was done in the roughly five times playback of the full target video that led off assessment as shown in Figure 1) instead of leaving it in an incomprehensible state.

The move to greater acceleration, from 25x to 50x, had significant benefits. The accelerated 2% summary provided excellent performance equivalent to 25x, but with dramatically faster time on task, and no significant drop in the ease of use or redundancy metrics from the top-rated systems tested here.

The failure to demonstrate additional improvements through folding in domain knowledge, e.g., emphasizing inclusion of camera effects like pan and zoom into the summary, was disappointing. For the first experiment, *pz* had audio synchronized to its video, suffering from chopped words and poor comprehensibility compared to the cleaner audio track used with 25x and *cluster*. The same visual footage was used in the second experiment for this summary treatment, but using the clean audio of 25x, but this *pzA* summary had poor task performance, even though it had the best average ratings for ease of use and redundancy. While this summary technique knowingly gives up coverage (vs. the speed-up techniques) to increase satisfaction and playability, two immediate corrections to the *pzA* method could be made that may increase its performance effectiveness. First, rather than keeping some attempt at coverage by representing all clusters (step 2 of Section 3.3), instead keep all pans/zooms at some minimum playback length, giving up on 100% cluster representation if necessary. Secondly, attempt to fold in further knowledge of the rushes footage so that only relevant, meaningful pans/zooms are kept. For example, the camera operator frequently moves the camera or frames and reframes a scene at the start of a take, and these “setting up the take” shots carry little meaning, but are recognized as pans or zooms and included into the *pz* and *pzA* summaries. Further processing to identify not just pans/zooms but pans/zooms from within takes (rather than setting up takes) would improve the relevancy of the summary visual footage, and perhaps boost IN, EA, and RE scores

to where they compete well with the frame rate acceleration methods.

As to the future of playable temporal video summaries, for general purpose use it may be impractical or impossible to define which attributes are most important. If coverage matters, then the IN metric is critical. If detail matters, e.g., to be able to identify people in pan effects, then coverage can be sacrificed for detail. As pointed out by Arons, the human in the loop should be leveraged, with his SpeechSkimmer allowing for intelligent filtering of recorded speech [2]: “the intelligence is provided by the interactive control of the human, in combination with the speech segmentation techniques.”

An excellent video summary technique is likely one where the user has interactive control, e.g., using a $50x$ summary until a neighborhood of interest is reached and then a pzA summary to see details within the pan. Wildemuth et al., note that fast-forward surrogates should ideally be controllable by end users who can adjust the speed based on content characteristics and personal preferences [10]. Interactive video summaries have received emphasis by others as well, e.g., in [8] the authors note the following, while acknowledging the difficulties of setting up assessment frameworks for such interactive, dynamic summaries:

Our summarization approach quantifies these two concepts and maximizes a weighted sum of both detail and coverage functions to obtain a tradeoff between the two. This approach enables the user to change the weights and regenerate the video summary of a program with more detail or more coverage, depending on a particular application.

Even assuming interactive adjustment, there remains the question of what a video summary should look like for new users or those unwilling or unable to further tune the summary playback. The empirical investigations conducted here help frame the parameters that can be used for default settings of playable video summaries.

8. CONCLUSIONS

The assessment framework provided by NIST and the TRECVID organizers for 2007 allows the international research community to systematically address video summarization for a given genre of video, with 2007's test genre being BBC rushes materials. By taking the assessment framework and text inclusions listings, one can conduct follow-up investigations as we did here comparing the relative merits of various summarization methods. The duration of the summary and audio content is controlled to be exactly 4% for tested $25x$ and pzA methods, with 2% and 1% durations for $50x$ and $100x$. Without such control, such as with trying to reach conclusions across the broad set of submitted summaries of various durations and audio quality graded by NIST, it is difficult to state what video summary features lead to what sort of utility. The obvious can be stated: a verbatim extraction of a few seconds from the original full video will have very easy playability (EA), little redundancy (RE), very fast playback (TT), but very poor coverage (IN performance). We endeavored in these experiments to move beyond the obvious and explore at what point frame rate acceleration drops off in terms of usability and performance for the

BBC rushes materials. As noted in [6], caution regarding the scope of conclusions is, as always, appropriate because rushes of dramatic series can look quite different from other less dialogue-based rushes. For the tested material, $50x$ is recommended, with $100x$ dropping off significantly in both performance and rated usability.

9. ACKNOWLEDGMENTS

Our thanks to NIST and the TRECVID organizers for enabling this video summarization evaluation. This work is supported by the National Science Foundation under Grant No. IIS-0205219 and Grant No. IIS-0705491.

10. REFERENCES

- [1] *Proc. ACM Int'l Workshop on TRECVID Video Summarization* (Augsburg, Germany, in conjunction with *ACM Multimedia*, Sept. 28, 2007), ISBN: 978-1-59593-780-3.
- [2] Arons, B. SpeechSkimmer: A System for Interactively Skimming Recorded Speech. *ACM TOCHI* 4(1), 1997, 3-38.
- [3] Christel, M.G., Smith, M.A., Taylor, C.R., & Winkler, D.B. Evolving Video Skims into Useful Multimedia Abstractions. In *Proc. ACM CHI '98* (Los Angeles, April 1998), 171-178.
- [4] Hanjalic, A. Shot-Boundary Detection: Unraveled or Resolved? *IEEE Transactions on Circuits and Systems for Video Technology* 12(2), 2002, 90-105.
- [5] Hauptmann, A.G., Christel, M.G., Lin, W.-H., Maher, B., Yang, J., Baron, R.V., and Xiang, G. Clever Clustering vs. Simple Speed-Up for Summarizing BBC Rushes. In *Proc. ACM Workshop on TRECVID Video Summarization* (Augsburg, Germany, Sept. 2007), 20-24.
- [6] Over, P., Smeaton, A.F., and Kelly, P. The TRECVID 2007 BBC Rushes Summarization Evaluation Pilot. In *Proc. ACM Workshop on TRECVID Video Summarization* (Augsburg, Germany, Sept. 2007), 1-15.
- [7] Song, Y., and Marchionini, G. Effects of Audio and Visual Surrogates for Making Sense of Digital Video. In *Proc. ACM CHI '07* (San Jose, CA, April-May 2007), 867-876.
- [8] Taskiran, C.M., Pizlo, Z., Amir, A., Ponceleon, D., and Delp, E. J. Automated Video Program Summarization Using Speech Transcripts. *IEEE Transactions on Multimedia* 8(4), 2006, 775-791.
- [9] Truong, B.T., and Venkatesh, S. Video Abstraction: A Systematic Review and Classification. *ACM Trans. Multimedia Computing, Communications, and Applications (TOMCCAP)* 3(1), 2007, 1-37.
- [10] Wildemuth, B.M., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., and Gruss, R. How Fast Is Too Fast? Evaluating Fast Forward Surrogates for Digital Video. In *Proc. Joint Conf. Digital Libraries* (Houston, TX, May 2003), 221-230.