10-2008

Exploring the Utility of Fast-Forward Surrogates for BBC Rushes

Michael G. Christel
Carnegie Mellon University, christel@cs.cmu.edu

Alexander Hauptmann
Carnegie Mellon University, alex@cs.cmu.edu

Wei-Hao Lin
Carnegie Mellon University

Ming-Yu Chen
Carnegie Mellon University

Jun Yang
Carnegie Mellon University

See next page for additional authors

Published In
Proceedings of the 2nd ACM Trecvid Video Summarization Workshop . TVS '08. , 35- 39.

Exploring the Utility of Fast-Forward Surrogates for BBC Rushes*

Michael G. Christel, Alexander G. Hauptmann, Wei-Hao Lin,
Ming-Yu Chen, Jun Yang, Bryan Maher, and Robert V. Baron
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
1-412-268-7799

{christel, hauptmann, whlin, mychen, juny, bsm, rvb}@cs.cmu.edu

ABSTRACT
This paper discusses in detail our approaches for producing the video summaries submitted to the TRECVID 2008 BBC rushes summarization task, including the baseline method. Empirical work produced during and after the TRECVID 2007 rushes summarization task gave strong evidence that a simple 50x method (sampling every 50th frame) provides excellent coverage (text inclusion performance). Our submissions for TRECVID 2008 investigated the effects of junk frame removal, including a comprehensible audio track, and emphasizing pans and zooms when backfilling to reclaim the space removed with the noise shots from the original 50x set. Results show that 50x based methods provide excellent coverage as expected. There were limited effects for the other strategies to improve user satisfaction, with the discussion providing some insights for future video summary development and evaluation work.

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

General Terms
Experimentation

Keywords
TRECVID, video summarization, video skim, video surrogate, video abstract, benchmarking, evaluation

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. This paper concentrates on playable video summaries, experimenting with summaries that have durations of one-fiftieth (2%) or smaller compared to the target video.

Proc. TRECVID Summarization Workshop (in association with the ACM Multimedia Conference), October 31, 2008, Vancouver, British Columbia, Canada.

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 [8]. Christel et al. 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, 7]). The TRECVID summarization task 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 [9], giving the video all of the important information contained in the video.

This paper details the Carnegie Mellon University (CMU) Informedia research group's participation in the BBC rushes summarization task for TRECVID 2008. Our work builds from conclusions first published in the seminal work "How Fast Is Too Fast? Evaluating Fast Forward Surrogates for Digital Video" by Wildemuth et al. [10], which won the best paper award for JCDL 2003. As an acknowledgment to their work, we describe our submissions as "fast-forward surrogates" in the title of this paper. Section 2 discusses this type of summary further, Section 3 presents the CMU strategy for producing rush video summaries, Section 4 overviews results, and Section 5 presents conclusions.

2. FAST-FORWARD Nx SURROGATES
Wildemuth and her co-authors found that for produced documentary materials without the redundancy of rushes, fast-forward surrogates with accelerated playback can be effective [10]. They 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. 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]. Our Informedia research group has

also investigated the utility of automated video summarizations for news and documentaries, i.e., for produced materials, since the mid-1990s [3]. The TRECVID 2008 BBC rushes video summarization task provided a perfect opportunity to test whether conclusions reached with produced materials would transfer to rushes materials as well. The cited work ([3, 10]) conducted user studies based on produced broadcast news and documentaries, with redundancies edited out, and (for [3]) 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 [6].

Before any work was begun on TRECVID 2008, we conducted an empirical study with 15 human assessors on the TRECVID 2007 rushes summarization task [2], set up specifically to look at the utility of 25x, 50x, and 100x fast-forward surrogates for the BBC rushes. By simply sampling every 25th frame, you create a 4% video summary, which we label 25x in agreement with [10]. 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 augmented the 25x, 50x, and 100x video with regular speed narration to produce the tested summaries [2]. For the visual component of the Nx summaries, every Nth frame was selected with no consideration given for noise-filtering. The user study was conducted using the TRECVID 2007 assessment framework.

The results of the study showed that a move to greater acceleration, from 25x to 50x, has 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 satisfaction metrics. For the tested material, 50x is recommended, with 100x dropping off significantly in both performance and rated usability [2]. When the invitation to participate came out for the TRECVID 2008 rushes summarization task with the target summary size reduced from 4% (used in 2007) to 2%, this allowed for the 50x summary to be submitted as a video surrogate having an established empirical record of success [2, 10]. The next section discusses how we tweaked attributes of the 50x summary to investigate the relative contributions of other such attributes, using the common benchmark metrics discussed in [6].

3. AUTOMATED SUMMARIZATION TECHNIQUES

As a control to help gauge success for the TRECVID summarization task across all participants, our Informedia group at Carnegie Mellon University produced a very simple baseline approach to 2% summary generation: sample every 50th frame. No consideration is given to the audio track at all, and the produced baseline has no audio component. No junk shot filtering is applied, so the percentage of junk frames in the baseline will be roughly equivalent to the percentage of junk frames in the source BBC rush video. The visual persistence of the junk video in 50x may not register with the viewer as much if all the junk shots in the source are brief. Consider a clapper shot of just under two seconds that produces exactly one frame for viewing in the 50x summary: a viewer may not notice this junk frame that appears for 1/25 of a second. Alternatively, a 50-second sequence of color bars will be represented in the 50x summary as a one second color bar steady shot that will be noticeable. Hence, the simplicity of the baseline algorithm was expected to result in lower subjective ratings, but we purposefully kept the 50x baseline simple so as to measure the contributions of folding in junk frame removal and other automated processing techniques. Based on the reports and demonstrations from the TRECVID 2007 Video Summarization Workshop [7], most participants in this task did attempt junk frame removal in 2007, eliminating irrelevant shots as shown in Figure 1.

Figure 1. Examples of 4 types of irrelevant (junk) shots seen in BBC rushes video: white frame, black frame, clapper to mark scene takes, and color bar.

Such noise reduction attempts did not often result in video summaries that were markedly better than the baselines in 2007. Perhaps the source rushes material did not contain the volume of junk frames needed for such reduction to matter for the inclusion (IN) metric, or for the subjective measures used in 2007. For 2008, in order to more directly measure effects of junk frame removal perhaps, the published guidelines (http://www-nlpir.nist.gov/projects/tv2008/tv2008.html) stated that human evaluators would directly rate whether the summary "contains color bars, clapboards, and/or totally black or totally white frames" – the JU metric. Given this emphasis on junk frame removal, we obviously chose to include such noise reduction attempts in both of our submitted runs CMU.1 and CMU.2.

We did not mute the summaries since we felt that understanding the acoustic context would help to more quickly understand the visual events. Hence, for both CMU.1 and CMU.2 we included audio, in fact the same audio. The audio edit list was not a 50x sampling, which would have been incomprehensible, but rather a collection of 1x (normal playback) audio phrases, of course no longer time-aligned to every point of the video summary visuals. Our 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 [8], which also advised that tight audio-visual synchronization may not be necessary in a video summary.

For our CMU NIST-judged runs, we decided to focus on a few specific summarization features. First, how would our own junk frame removal improve on the baseline, removing clearly irrelevant material as illustrated in Figure 1? Second, does the audio component improve satisfaction with the summary or time on task? Finally, if we aggressively remove junk frames to have more space available in the 2% target size for other frames, and then backfill those frames by emphasizing sequences of importance – pan/zoom sequences – and fold in variable rate playback, is there a difference on any of the collected metrics?

3.1 Visual Processing

Our NIST runs labeled CMU.1 and CMU.2 in the evaluation utilized automatic junk frame detectors. There are four different kinds of junk frames we want to remove from BBC

summarizations, as shown in Figure 1. We extract three different features to construct our junk frame detectors: HSV color histogram features, SIFT features and speech recognition features. Color histogram features provide solid performance to detect black frames, white frames, and color bar frames, as witnessed in the TRECVID 2007 summarization runs [2, 5]. However, the main challenge is to detect clapper frames. SIFT features and speech recognition features are extracted to provide different information other than global color appearance.

Scale-invariant feature transform (SIFT) is an algorithm in computer vision to detect and describe local features in images. The detection and description of local image features can help in object recognition. The SIFT features are based on the appearance of the object at particular interest points, are invariant to image scale and rotation, and are also robust to changes in illumination, noise, occlusion and minor changes in viewpoint. In clapper detection, a key difficulty is that clappers vary in appearance. Hence, global features by themselves may not detect clappers accurately. Our SIFT clapper detection has three major steps: (1) SIFT feature detection and description based on [5]; (2) bag-of-word quantification; and (3) classification. Each interest point is described by a 128-dimension vector. For each key frame, the number of extracted SIFT features is different. Therefore, we try to use bag-of-word approach to quantify feature distribution of each key frame. The idea of the bag-of-word approach is to quantify SIFT features into a fixed number of types. We use k-means clustering to find the conceptual meaningful types and each cluster (or type) is treated as a visual word in bag-of-word approach. In the end, each frame is presented by a visual word histogram feature and this is the bag-of-word feature of the image. We train our clapper classification based on bag-of-word features.

We also found speech recognition provides distinguishing features to detect clappers. Prior work by FX Palo Alto signaled the possible benefits of audio for clapper detection [1], but we looked for spoken words rather than the clapper sound as we also wanted to detect clapper sequences where there was no audible "snap" of a board clapping against another board. Just as the visual cues varied for clappers, so did the audio, but based on inspection of the development set we felt that there were some distinguishing key phrases with clappers, like "Take One!", "Take Two!", and "Action!" We set up heuristic rules to retrieve speech recognition to utilize this information to detect clappers by audio appearance. We combined the detection results from global appearance (HSV color histogram), local appearance (SIFT features) and audio appearance (speech recognition). Table 1 shows our cross-validation detection performance on the BBC Rush training set.

Table 1. Performance of different automated clapper detectors against the training set.

	Color	SIFT	ASR	Combination
Precision	0.65	0.7	0.17	0.31
Recall	0.22	0.37	0.2	0.58

Color in the table means HSV color histogram, SIFT is the result from SIFT features, ASR is the result from speech recognition and combination is the result by combining all three features. From the result, we can see ASR has fairly poor precision but overall it helps recall. Since all three features are from three different views, they complement each other to detect more clappers. This results in overall recall that is much higher; however, the precision drops because of the poor ASR result.

For CMU.1, we make use of the SIFT clapper detector, finding fewer clappers but with higher precision. For CMU.2, we attempt to score better on the junk frame existence subjective scale (JU) by more aggressively removing clapper frames according to the Combination clapper detector. We acknowledge that some non-clapper content is thrown out, but we have better recall of clapper frames as well. CMU.2 has more space available after junk frame removal for backfilling. As was done in our 2007 runs [2, 5], we emphasize automatically detected camera pans and zooms, using those frames to replace the junk frames removed by the two methods. The replacement frames come from the longest pans/zooms found for the original video with reasonable confidence. On average, CMU.2 eliminated 6.8% more frames as junk frames based on the more aggressive Combination clapper filtering, producing a video summary with more pan/zoom frames than CMU.1.

The visual processing begins with the frames that are in the 50x sampling, which we will term CMU keyframes. Black, white, and color bar junk frames are eliminated from this set based on the HSV color histogram classifier that CMU also used last year. Clapper shots are detected via SIFT for CMU.1, and more aggressively folding in ASR cues for CMU.2. Temporal cues are used to more aggressively eliminate junk frames: CMU keyframes adjacent to detected junk frames are also considered junk and removed. Also, the temporal neighborhood is broadened for clappers: CMU keyframes adjacent to frames adjacent to clapper frames are also removed. Detected pans/zooms are added back in as space allows at a playback rate of 2x (if space allows) or 4x so that the pan or zoom can be easily interpretable. As first noted in [5], we are leveraging from cinematic principles that pans and zooms are used to emphasize important visuals, but unfortunately for rush videos this is not always the case. A pan or zoom may also be occurring at the setup of a take to get the camera focused appropriately. We opted for slower playback of pans/zooms so that the pan/zoom event could be more easily recognized by the viewer of the video summary.

3.2 Aural Processing

We first ran the SAIL LABS Technology (http://www.sail-technology.com) speech recognizer over the video. The source footage is not ideal for automatic speech recognition (ASR), as in rushes materials the speaker is often not properly microphoned and environmental aural noise not controlled because the sound track is anticipated to be cleaned up later during the production process. Since the ASR error rate was quite high, we did not rely heavily on the content of the spoken text, but exploited other characteristics. The output of the recognizer was split into phrases, based on the duration of silences in the speech and whether a speaker change was detected. We wished to collect audio snippets bounded by silences based on earlier research on skims [3] showing 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.

To achieve a balance in coverage of the video, we divided the video into equal segments, where the number of segments was determined by the target length of the summary video ($1/50^{th}$ of

the original full video) and the average duration of the SAIL-recognized phrases, i.e.

Number of Segments = TargetLength / AvgPhraseDuration

Now we divided the full video into the same number of segments, and the algorithm then selected one phrase of near average duration which occurred in each of the segments in the full video. This allowed us to insert some audio from every portion of the video. Near the end of the target time of the summary video, shorter phrases became acceptable if longer phrases no longer fit into the overall time budget. At the end, if the resulting audio segment was shorter than the target time, the remaining time was padded with silence.

We also placed a number of constraints on the phrases that were added. A phrase had to contain at least two new words to be included in the summary. This avoided adding repetitions of identical or nearly identical utterances from repeated scenes. A certain set of phrases was always ignored or eliminated. This included the very first phrase in video since this phrase frequently contained the director's instructions or startup talk. We also excluded all phrases containing the words "Action", "Quiet", "Take", "Scene", as well as any phrases containing numbers (as in "take two", "scene five"). Deleting any phrase with one of these keywords allowed us to remove phrases containing director's instructions despite a high speech recognition error rate for anything not spoken directly near the microphone. One word phrases were also eliminated from the process.

4. RESULTS: *50x baseline, CMU.1, CMU.2*

As expected based on the empirical study conducted with BBC rush video from TRECVID 2007 [2], the inclusion score (IN) metric showed that the 50x strategy, and our derivatives CMU.1 and CMU.2, provide excellent coverage of the original video. The IN score, an estimation of recall, is plotted for the submitted runs in Figure 2. The simplistic strategy of Nx sampling, with N=50 to provide a 2% summary, worked as well as the 25x 4% summary did in evaluations of TRECVID 2007 rush summaries [2, 5].

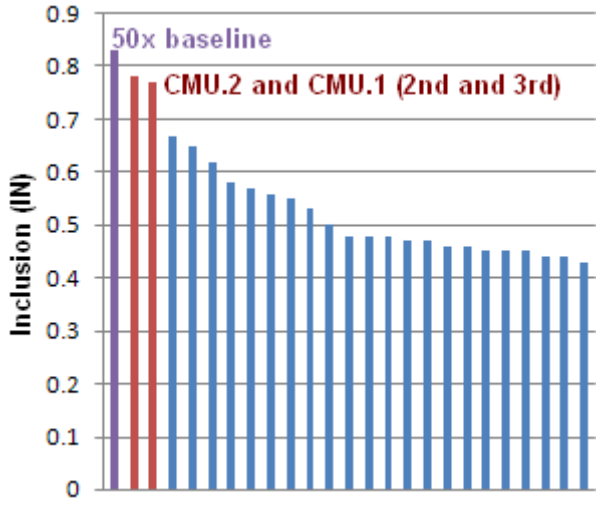

Figure 2. Top 25 IN scores for TRECVID 2008 BBC rush video summarization runs.

From 2007, we reported that "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 …all other NIST submitted runs" [2]. We echo that same conclusion: if the main objective of the summary is to provide excellent coverage of the original video to maximize recall of text inclusions, then 50x is an excellent method. There is a limit on N for Nx speed-ups, and we have gathered some empirical evidence from the TRECVID 2007 rush video test set that 100x does not work as well as 50x [2], but at 50x, we see confirmation in the results of Figure 2 that users can capably perform on the inclusion task armed with 50x summaries or the derivatives CMU.1 and CMU.2 that were based on 50x.

Such excellent performance comes at a cost: the time on task TT metric for 50x, CMU.1, and CMU.2 averaged 59.59, 56.66, and 56.43 seconds, respectively, the slowest, third slowest, and fourth slowest of all of the NIST graded runs. The lengths of the summaries were backfilled to be 2% summaries, on purpose to focus attention on differences in summary make-up rather than summary length. The average length was 31.31 seconds, so that the task time was about double the video summary duration. To accomplish greater than 75% inclusion score, the viewer had to invest 4% of the time that it would take to view the full video, so while these CMU.1, CMU.2, and 50x baseline summaries did slow down time on task, there is still tremendous savings over watching the original full video: 4% vs. 100%.

We anticipated a separation of our two CMU runs and the 50x baseline on the subjective metrics JU (summary contained lots of junk, 1 strongly agree – 5 (best) strongly disagree), RE (summary contained lots of duplicate video: 1 strongly agree – 5 (best) strongly disagree), and TE (summary had a pleasant tempo/rhythm: 1 strongly disagree – 5 (best) strongly agree). Table 2 shows the results. The baseline scored the absolute lowest for all NIST graded runs on these measures, with the CMU runs likewise scoring at the bottom of the scale for RE and TE. The one metric where we attempted to improve based on some cleverness in better clapper recall was in the JU metric, and with JU the CMU.1 and CMU.2 separate themselves from the baseline, but not from each other. Hence, the increased aggressiveness in removing potential clapper shots with CMU.2 over CMU.1 was not distinguishing enough to produce significant differences in any of the collected metrics IN, TT, JU, RE, and TE.

Table 2. Results on 5-point scale subjective metrics.

	JU	RE	TE
50x baseline	2.66	2.02	1.44
CMU.1	3.02	2.28	1.76
CMU.2	2.96	2.25	1.64

5. DISCUSSION AND CONCLUSIONS

50x video summaries allow for very high inclusion scores by users willing to spend up to double the video summary duration in pausing, playing, and reviewing the summary. Removing junk frames is noticed by users (as noted by improved JU score in Table 2), as well as better RE and TE scores for CMU.1 and CMU.2 over the baseline. However, the extremely fast apparent

playback of 50x taxes the user such that the subjective metrics of Table 2 rank at the bottom of all the graded runs. Such conflicts in assessing video summaries are discussed further in [9]: optimizing for one parameter like rhythm often comes at the expense of another like coverage or duration. The NIST overview report in 2007 noted 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" [7]. In fact, Nx fast forward surrogates for rush videos containing redundant takes relies on the takes being repeatedly represented in consecutive temporal order in the apparent N-times normal speed playback. For N=50 as studied here, a 50 second take would be represented with exactly 1 second of visual material in the 50x summary. If that take is repeated 3 times, the user sees the first second showing the first take at 50 times speed, then sees it again in another second, and then again in a third. The redundant playback reinforces messages that perhaps would not be recognized if each scene were represented exactly once in the summary, rather than multiple sequential times via multiple takes. Redundancy allows N=50; without such redundancy, there is likely a much smaller limit on N that is reasonable for Nx summaries to still generate excellent recall performance.

We believe the inclusion of an audio narrative made the video summaries more playable by end users, but unlike the 2007 evaluation there was no "EA" metric for ease of use, that being replaced by JU and TE metrics. Since the 50x foundation for CMU submissions has acknowledged shortcomings in rhythm, opting for constant pacing to maximize coverage, the audio addition in CMU.1 and CMU.2 was not enough to move these runs higher in the TE rankings, and audio did not matter of course for JU (and was chosen to not worsen RE). Hence, we feel the contribution of audio was not adequately captured in these sets of metrics, and remains an issue for further study. We also need to run experiments of 50x vs. 50x+junk-removal vs. 50x-with-audio. Given the limit on submissions, here we bundled both junk frame removal and audio addition in our CMU runs, so we can't definitively state whether differences in Table 2 from the baseline are due to the audio, the junk frame removal, or both (although JU is likely due only to the improved junk frame removal).

We also still seem to be at early stages of video summary evaluation, so investigating subtle differences, like the clapper detector difference that distinguished CMU.1 from CMU.2, may be too premature. We need to understand higher order effects first, such as the contributions of an audio track or varying N for Nx summaries, before tweaking more subtle video summary creation filters. We feel the three runs reported here definitely establish the 50x family of summaries as superior for recall performance, requiring users to spend up to double the summary time in interactions.

The assessment framework provided by NIST and the TRECVID organizers for 2007 and 2008 allows the international research community to systematically address video summarization for a given genre of video, with the test genre being BBC rushes materials. The obvious can be stated: a verbatim extraction of a few seconds from the full video will have great tempo (TE), 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 50x fast forward surrogates with audio and junk frame removal in terms of usability and performance for the BBC rushes materials.

6. 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.

7. REFERENCES

[1] Chen, F., Cooper, M., and Adcock, J. Video Summarization Preserving Dynamic Content. In *Proc. ACM Workshop on TRECVID Video Summarization* (Augsburg, Germany, Sept. 2007), 40-44.

[2] Christel, M.G., Lin, W.-H., and Maher, B. Evaluating Audio Skimming and Frame Rate Acceleration for Summarizing BBC Rushes. In *Proc. CIVR* (Niagara Falls, July 2008).

[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] 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.

[5] Lowe, D. G. Object Recognition from Local Scale-Invariant Features. In *Proc. International Conf. on Computer Vision* (Kerkyra, Greece, Sept. 1999), 1150-1157.

[6] Over, P., Smeaton, A.F., and Awad, G. The TRECVID 2008 BBC Rushes Summarization Evaluation. In *TVS '08: Proc. ACM Workshop on TRECVID Video Summarization* (Vancouver, BC, Canada, Oct. 2008), 1-20.

[7] 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.

[8] 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.

[9] 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.

[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.