The TRECVid 2008 BBC Rushes Summarization Evaluation

Paul Over
Information Access Division
Information Technology Lab.
National Institute of Standards and Technology
Gaithersburg, MD. 20899, USA
over@nist.gov

Alan F. Smeaton
Centre for Digital Video Proc.
& CLARITY: Centre for Sensor Web Tech.
Dublin City University Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

George Awad
Information Access Division
Information Technology Lab.
National Institute of Standards and Technology
Gaithersburg, MD. 20899, USA
gawad@nist.gov

October 31, 2008

Abstract

This paper describes an evaluation of automatic video summarization systems run on rushes from several BBC dramatic series. It was carried out under the auspices of the TREC Video Retrieval Evaluation (TRECVid) as a followup to the 2007 video summarization workshop held at ACM Multimedia 2007. 31 research teams submitted video summaries of 40 individual rushes video files, aiming to compress out redundant and insignificant material. Each summary had a duration of at most 2% of the original. The output of a baseline system, which simply presented each full video at 50 times normal speed was contributed by Carnegie Mellon University (CMU) as a control. The 2007 procedures for developing ground truth lists of important segments from each video were applied at the National Institute of Standards and Technology (NIST) to the BBC videos. At Dublin City University (DCU) each summary was judged by 3 humans with respect to how much of the ground truth was included and how well-formed the summary was. Additional objective measures included: how long it took the system to create the summary, how long it took the assessor to judge it against the ground truth, and what the summary's duration was. Assessor agreement on finding desired segments averaged 81%. Results indicated that while it was still difficult to exceed the performance of the baseline on in-

cluding ground truth, the baseline was outperformed by most other systems with respect to avoiding redundancy/junk and presenting the summary with a pleasant tempo/rhythm. [1]

ISBN: 978-1-60558-303-7

Categories and Subject Descriptors: H.5.1 [**Information Interfaces & Presentation**]: Multimedia Information Systems - video

General Terms: Experimentation

Keywords: Benchmarking, Evaluation, TRECVid, Video summarization

1 Introduction

For several years, the TRECVid evaluation campaigns ([26, 27, 28]) have mainly explored the evaluation of video information retrieval system components such as shot boundary detection, feature detection and search, using a variation of the Cranfield-TREC methodologies. In 2007, TRECVid introduced a new track as a first attempt at a large-scale evaluation of video summarization systems. Twenty-two research groups participated and the results of that effort were presented and discussed at a workshop at the ACM Multimedia Conference in 2007 (TVS07) [20].

A summary presents 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. A video summary can take various forms: e.g., keyframes (simple, static storyboards, dynamic slideshows), video skims (at fixed or variable speeds, etc.) or more complicated multidimensional browsers [31, 29]. A video summary can exploit the human visual system's native strengths in quickly scanning large numbers of images and facilitating recognition of objects and events. 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 an unfamiliar video collection. Video summarization is thus a key video content service, along with browsing and searching.

In the overview of the 2007 TRECVid rushes summarization task, [20], several earlier studies of video summarization were discussed, some of which included evaluation of the approaches taken. These tend to have looked at related, but different, situations to what was addressed in TVS07 and several were specialized to a specific genre. Some were extrinsic, i.e., in terms of how a summary helps in some tasks, rather than intrinsic i.e. direct evaluations, and most did not compare summaries to the full video being summarized.

These several examples of previous work in evaluating video summaries, show that there is definite interest in somehow quantifying the effectiveness of an automatically-generated video summary. However, the datasets used have been small and based on the efforts of just single groups. In TVS07/08 TRECVid provided a reasonably large video collection to be summarized, a uniform method of creating ground truth and a uniform scoring mechanism.

In this paper we present an overview of the TRECVid 2008 Video Summarization evaluation (TVS08) which built on TVS07 but used new test data, a larger set of participating research groups, and improvements to the evaluation measures based on lessons learned. What follows includes a description of the goals of the evaluation, the video data used, the task set for the participating groups, and the evaluation approach used, including the procedure used for creating the ground truth. We also include an overview of the results of the 31 groups who completed the summarization activity and a very high level overview of the different approaches taken by the groups. The details of each group's activities

[1] Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

can be found in their own individual papers. In the next section we present a brief overview of previous related work in video summarization.

2 Video data

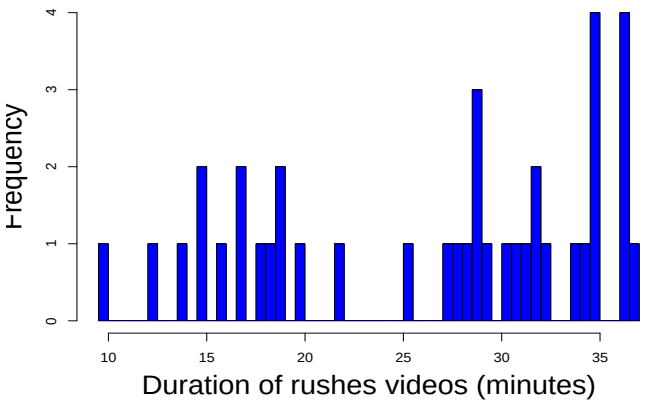
The video to be summarized in the TRECVid 2007/2008 summarization evaluation was of a particular sort that presents special problems and opportunities. It consisted of unedited video footage, shot mainly for five series of BBC drama programs, and was provided to TRECVid for research purposes by the BBC Archive. The drama series included a historical drama set in London in the early 1900's, 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 42 videos were provided to participants as development data and 40 were withheld for testing the systems once developed. One video (MRS336774) presented problems in creating the ground truth and summaries for it were not evaluated. Each set of videos represented a random sample balanced with respect to the number of videos from each original TV series. The test videos had a minimum duration of 9.8 minutes and a maximum duration 36.9 minutes, with the mean duration being 26.6 minutes and Figure 1 presents the distribution of the 39 video durations for those used in testing.

Sample ground truth was available for all of the development videos and ground truth was also created for the test videos as described later.

The rushes contained scenes of people in various everyday situations, both indoor and outdoor. Some actors appear repeatedly in the same and in different settings, sometimes with different clothing, etc. Other people may be seen only once. There was scripted dialog as well as natural sounds of the director, crew, the shooting environment, etc. There was a great deal of redundancy of various sorts as scenes were shot and then re-shot, with the camera runs leading up to/between/after scenes, etc. Crew appeared now and then as well as video of clapboards at scene and at "take" boundaries.

Rushes are potentially very valuable as re-usable video content but are largely unexploited because only the original production team knows what the rushes contain and metadata is generally very limited, e.g., indexing by program, department, name, date. Twenty to forty hours of rushes may be shot for each hour of finished programing produced [34]. It is hoped that the ability to summarize such rushes might contribute significantly to an overall rushes management and exploitation solution.

Figure 1: Distribution of test video durations (minutes)



3 System task

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. It was to be no longer than 2% of the duration of the

video being summarized. This meant that the average video would have a summary lasting at most 32 seconds.

The choice of 2% was based on a consensus of participants in the 2007 evaluation that significantly more redundancy could be wrung out of the rushes than was required in TVS07 which had been set at 4%. Both targets are somewhat arbitrary, as no complete, detailed information about redundancy in each of the test videos was available. The motivation for choosing these compression factors included the following considerations. The rushes are highly redundant and a couple of manual experiments indicated all the unique content might fit in a 10% summary. It was hoped the requirement for greater compression would encourage researchers to explore more than just selection of frames from the full video as the means of compression. While 32 seconds may be a relatively long summary from the point of view of a recreational searcher wanting a preview of a video, it seemed within reason for a professional working with a rushes database.

Ideally one would not restrict the types of summary created (skims, interactive storyboards, etc.) but this would have complicated the evaluation. So to simplify things, each summary was limited to a single MPEG-1 file of up to a given maximum duration which would be displayed during evaluation using the original video's frame rate/size. In its simplest form it could have been just a subset of frames from the video to be summarized in the original sequence. However, it could also have been more creative — presenting the viewer with multiple smaller frames at once, adjusting their sizes, changing the sequence of original frames, etc., and while the restriction of allowing submissions only as MPEG-1 video did constrain interactive engagement with the summary, it did not limit participants' creativity in summary presentation.

4 Evaluation

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, how much redundant and useless video the summary contained, and whether the summary had a pleasant tempo/rhythm.

At NIST, 5 retired adults with computer skills were hired, trained, and then spent a total of 110 person-hours watching eight assigned test videos each. They created for each video a list of items identifying the video segments they felt should be included in a good summary. Each item was identified as a person, thing, or event which occurred in the segment and distinguished it from other segments. Each list was reviewed at NIST against the full video and revised to normalize any extremes in level of detail, correct any ambiguities, and maximize the economy of expression. For a detailed description of ground truth creation see the instructions in Appendix A.

At Dublin City University, the submitted and baseline summaries were then evaluated by 10 hired assessors, some of whom were graduate students, using software written by NIST for that purpose in TVS07. Each submitted summary and each baseline summary of each of the 39 test videos was judged by three different assessors. Unless explicitly noted otherwise, scores presented in the following are means of the three judgments for any summary and measure.

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 21-item) ground truth list from the original video content. Each ground truth element uniquely identified an important segment from the full video by noting included objects/events, sometimes with camera motion specified. The assessor viewed the summary only once in a 125 mm x 102 mm mplayer [17] window at 25 frames per second using only the "play" and "pause" controls and then determined which of the designated segments appeared in the summary. The process of trying to find the listed segments was timed to yield a measure of assessor effort.

The evaluation also collected usability/satisfaction information from the assessors with reference to each system's summary style. Based on the results in TVS07, the question about redundancy was kept but the two other questions were new and based on the observation that the TVS07 baselines seemed worse than the better automatic summaries but in ways the

TVS07 usability measures failed to capture.

In all three cases, a statement was made about each summary and the assessor indicated on a 5-point Likert scale the degree to which he or she (dis)agreed with the statement:

1. "This summary contains many color bars, clapboards, all black or all white frames."

2. "This summary contains many nearly identical segments."

3. "This summary is presented in a pleasant tempo and rhythm."

The summaries were presented to the assessors grouped by the full video being summarized. Such groups were not split across multiple assessors, so any assessor differences are spread evenly across all systems. When working with a new group of summaries (i.e., with a new video to be summarized) the assessor was also learning a new list of ground truth items to look for. The order of presentation of summaries within a group was therefore randomized with respect to systems to randomly assign any bias due to learning effects. In addition, the first five summaries of each group were judged again at the end of the session to mitigate the presumed start-up bias and provide some input on assessment reliability. The scores from the initial judging were not used in the final averages. Before beginning to judge summaries in a group, the assessor was instructed to play the full video (at about 5 × realtime) as many times as desired while studying the list of groundtruth segments.

Objective measures included system effort as measured by elapsed time to create the summary (as reported by the participants), size of the summary as determined by mplayer, and ease of understanding the summary content as reflected in assessor time-on-task in judging which of the ground truth segments were included in the summary.

To recap, the measures used for each summary were:

- percentage of desired segments found as judged by assessor
- presence of junk (color bars, clapboards, empty frames), as judged by the assessor
- amount of near redundancy, as judged by assessor
- satisfaction with tempo and rhythm of presentation, as judged by the assessor
- assessor time taken to determine presence/absence of desired segments
- duration of summary relative to the 2% duration target
- elapsed time for summary creation

In TVS07 there was some debate in designing the evaluation about how much time and control the assessor should have while viewing each summary. On the one hand, allowing unlimited (re)play and pausing could have allowed evaluation of summaries under conditions no real user would tolerate. This would have yielded unrealistic results. On the other hand the assessment situation is not a realistic one in so far as assessors not only watched the summaries but also had to record their judgments. Allowing only one play-through of each summary at normal speed (25 fps) seemed to place too great a weight on the visual acuity and memory capacity of the assessors. The compromise reached was to allow only one play-through at normal speed but to allow unlimited pausing. The time spent in pause as well as the number of pauses was recorded by the assessment software.

5 2008 Participants and their Approaches

Thirty-one groups completed submission of summaries for the test videos and these are listed in Table 1, along with a code used to refer to them through the remainder of this paper. We now present a thumbnail overview of most of the participants' approaches. Twenty-six of the participants have summary papers describing their approaches in more detail in the proceedings of this workshop and further details beyond these overviews can be had in those

papers. The other five participants are *Asahikasei Co.* from Japan, *NTT Cyber Solutions Laboratories* also from Japan, *Helsinki University of Technology* from Finland, *University of Sheffield* from the UK, and *City University of Hong Kong* from Hong Kong. These groups are expected to describe their approaches in the proceedings of the TRECVid conference in Gaithersburg in November 2008.

The team from *AT&T Labs* in New Jersey, USA [15] used standard clustering of visual characteristics of the original video in order to detect redundancy and re-takes and they did this using shots and sub-shots as their logical segments. The team also incorporated junk frame removal and applied saliency detection to detect the (visually) most important segments to include in the generated summary. Their generated summaries consisted of full frame summaries with variable speed playback, also showing the positional offset in terms of the original source video, overlaid on the screen during playback.

A similar approach was taken by the team from *Brno University of Technology* in the Czech Republic [2] who extracted low level visual features from each frame in the original source video, including low level features from regions within each frame. The source video was divided into 1-second segments rather than shots and the visual features were then used as inputs to clustering, in order to detect redundancy and re-takes. Junk shots (vertical color bars, blank screens and clapper boards) were explicitly removed so as not to appear in the generated summaries. The summary was rendered with a variable speed during playback. This speed changed depending on characteristics of the segment in the original video. The layout of the playback also included a visual indication of the position of the summary playback within the original video source.

The *University of Bradford* in the UK, working with the *Fraunhofer Institute* in Germany [22] sought to model rushes as an hierarchical structure and to exploit this structure in deciding what to include in the summary. A k-NN clustering approach was used based on visual similarity between shot keyframes. Face detection, audio and motion characteristics were also used, and junk shots were explicitly removed. The unit of information in this team's approach was the shot, and the generated summary consisted of frame playback with frame number/offset and 1-second dissolves between frames during the summary playback, in order to indicate shot bounds.

The team at *Carnegie Mellon University* created one baseline video summarization system and submitted its output for evaluation along with other group submissions. The baseline simply presented the entire video at $50\times$ normal speed - a strategy arrived at after study of various alternatives [6]. The baseline was mute with no audio whatsoever. Unlike the 2007 baselines, the 2008 baseline made no attempt to remove redundancy or junk frames. CMU's second submitted run was based on enhancing the baseline with junk frame removal using color and SIFT features, generating a comprehensible audio track, and emphasizing pans and zooms as camera motion. The team re-assessed $25\times, 50\times$ and $100\times$ summaries, and found $50\times$ to be the best performer.

The *COST Action 292 Group* is a large consortium of European research partners from the Netherlands, UK, France, Italy and Spain and extended their 2007 summarization participation by developing new approaches to detecting repetition [18]. As with most other groups, junk frames were explicitly detected and removed and the unit of manipulation was the scene rather than the shot or frame. This team, like some others, used face detection and camera motion, and extracted MPEG-7 color layout descriptors for each frame in the original video as input to their clustering approach. For their generated summary, this team did not use any fast forward and their summary segments tended to be longer in playback duration than others.

Like many others, the team from *Dublin City University* in Ireland [3] also worked at the shot level, and removed junk shots from their processing. This group made two submissions using two techniques for shot selection to be included in the final summary, one technique based on linear discriminant analysis and the other based on principal components analysis. Once shots had been selected for inclusion in the generated summary, sub-shots of 2 to 3 seconds were selected and some of them were played back at an accelerated rate. A storyboard of shot keyframes was generated and included at both the start, and the

Table 1: 2008 Participating teams

Code	**Team**	Ref.
asahikasei	Asahikasei Co.	
ATTLabs	AT&T Labs	[15]
Brno	Brno University of Technology	[2]
BU_FHG	University of Bradford and Fraunhofer Institute	[22]
CMU	Carnegie Mellon University	[6]
COST292	COST Action 292 group	[18]
DCU	Dublin City University	[3]
ETIS	ETIS Laboratory	[11]
EURECOM	Institut EURECOM	[8]
FXPAL	FX Palo Alto Laboratory Inc.	[5]
GMRV-URJC	Universidad Rey Juan Carlos	[30]
GTI-UAM	Universidad Autonoma de Madrid	[32]
ipan_uoi	University of Ioannina, Greece	[4]
IRIM	GDR ISIS - IRIM consortium	[21]
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH	[1]
K-Space	K-Space EU FP6 Network of Excellence	[9]
NHKSTRL	NHK Science and Technical Research Laboratories	[24]
NII	National Institute of Informatics	[13]
nttlab	NTT Cyber Solutions Laboratories	
PicSOM	Helsinki University of Technology	
PolyU	The Hong Kong Polytechnic University	[14]
QUT_GP	Queensland University of Technology	[25]
REGIM	École Nationale d'Ingénieurs de Sfax ENIS	[10]
Sheffield	University of Sheffield	
thu-intel	Intelligent Multimedia Group at Tsinghua U., Intel China Research Center	[33]
TokyoTech	Tokyo Institute of Technology	[35]
UEC	University of Electro-Communications	[19]
UG	University of Glasgow	[23]
UPMC-LIP6	Universite Pierre et Marie Curie - LIP6	[7]
VIREO	City University of Hong Kong	
VIVA-LISTIC	University of Ottawa - SITE	[12]

end, of the generated summary. A smooth zoom from the opening storyboard to the playback window (occupying 80% of the screen) took place at the start of the summary, and as the summary transitioned from shot to shot this was tracked on the storyboard.

The *ETIS Laboratory* in France [11] set out to detect "semantic" shot boundaries and to compare nearby shots in order to detect re-takes. They used 1 out of every 4 frames to keep computation costs down and based primarily on hue-saturation-value (HSV) color. Junk frames were also removed here, again using color histograms. Once these were eliminated, the amount of motion for each remaining shot was computed as an indicator of the amount of action in a shot. Shots of greater than 1 second duration were then candidates for inclusion in the summary.

Institut Eurécom based in Sophia Antipolis, France [8], extracted HSV color features from each frame in the source video and performed sequence alignment. This was inspired by its application in bioinformatics, and was used here to address the variable times taken during occurrences of re-takes and redundant segments. Following this they also did clustering to detect redundancy and removed junk frames explicitly. The generated summary consisted of a series of keyframes, occupying about 80% of the frame size, with icons and time offset indicators. The generated summary ended with a keyframe storyboard to provide a re-cap of what was included.

FX Palo Alto Laboratory Inc. in California, USA [5] used the metadata donated by the NHK Science and Technical Laboratory for junk frame removal and segmented the video using a combination of motion, audio and color features, and then clustered based on these. Two runs were submitted which vary in the methods for clip similarity and selection of clips for inclusion. The generated summary had a 0.25 second overlap fade transition between clips and an overlay of a transparent timeline and visual cues to indicate the amount of duplication from the original video.

The team from the *Universidad Rey Juan Carlos* in Spain [30] aimed to exploit low-level features only, built around their extraction from keyframes. Candidate segments for the generated summary were selected based on shot bound detection with n keyframes per shot based on activity or motion within the shot. A filtering stage removed junk frames and detects duplicates based on keyframe similarity. The final summary was a concatenation of keyframes.

The *Universidad Autonoma de Madrid* in Spain [32] extended their 2007 system for on-the-fly summarization. Last year their summarization technique was not able to predict or control the duration of the generated summary but this year the team used dynamic generation of binary trees, allowing realtime, on-the-fly summaries to be generated. These allowed progressive summary generation as the original video was either captured or processed. Unfortunately the structure of the TVS evaluation did not reward such progressive summary generation but nonetheless the resulting system generated impressive output.

A group from the *University of Ioannina* in Greece [4], first-time participants in the TRECVid summarization task, also segmented the source video into shots and extracted visual features, specifically HSV color histograms for every 5th frame. Keyframes for each shot were selected and shot-shot similarity was based on using the keyframes in order to detect repeating shots or re-takes. Junk frames were removed and the shot-shot similarities reduced the set of shots from the original video into the subset to be incorporated into the summary.

The *GDR ISIS - IRIM consortium* [21] consists of several research labs from France, which combined their resources to make a submission to the summarization workshop. The approach taken here was to generate low-level features for detected shots, including both an audio level indicator and a motion activity level. They also used mid-level features including face detection, explicit detection and removal of junk frames, and camera motion. All these features were used to select video segments for inclusion in the generated summary based on a k-NN clustering which also used color features on every 4th frame from of the original video.

JOANNEUM Forschungsgesellschaft from Graz, Austria [1], implemented two different approaches to summary generation, one based on hidden Markov models and one using a rule-based approach to selecting segments to include in the summary, and the group submitted two runs, one for each technique.

They also used clustering in order to detect re-takes and redundancy, and factored in a face detection module to help indicate which segments are more important for inclusion. They also incorporated junk frame removal, and they used shots from their shot boundary detection module, as their unit of information.

The *K-Space EU FP6 Network of Excellence* [9] is a large consortium from which 6 partners from the UK, Ireland, Germany, Austria and two from France, combined in this summarization task. The team used 3 independent techniques for segmenting the video which were then fused, followed by two independent techniques for redundancy detection which included face detection and hierarchical agglomerative clustering of 1 second video segments. The final summary was 1.5× fast forward in one of the two submitted runs, and 4.0× fast forward with non accelerated audio and a transparent timeline overlaid in the other.

The *NHK Science and Technical Research Laboratories* in Japan [24] performed shot boundary detection and also manipulated sub-shots which were detected based on motion in the video. Junk frames were detected and removed and duplicate scenes or re-takes were detected using keyframes which were detected, in turn, from sub-shots. These were then used for shot-shot similarity, which was ultimately based on color. The generated summary was a concatenation of sub-shots.

The *National Institute of Informatics* in Japan, working with Chulalongkorn University in Thailand [13] developed and tested two approaches to generating rushes summaries. The first used shot boundary detection, also detected sub-shots and extracts keyframes. This was followed by junk frame elimination and redundant or repeated shot elimination based on using the keyframes as the shot (and sub-shot) representatives. In the second approach, they used all frames of each sub-shot fragment in order to detect and eliminate redundancy. The second approach was more computationally expensive than the first and appeared to perform better.

The Hong Kong Polytechnic University worked with Nanjing University in China to use both audio and visual information in their summarization submissions [14]. Like most groups, they did shot detection followed by shot removal or pruning. Shot bounds were determined using color histograms taken from regions within each frame and then junk frames were removed explicitly. Shots and sub-shots of short duration were discarded and keyframes by clustering frames within shots using color histograms and choosing the maximum stability. Re-takes and redundancy were detected by clustering sub-shots and the final summary generated was based on keyframes from remaining shots.

Queensland University of Technology in Australia [25] took an approach to summarization that was based on trying to make summaries as pleasant to watch as possible. This group followed the regular approach of shot boundary detection based on color histograms, then shot clustering from which a minimum spanning tree of the cluster graph is constructed. From each cluster, the longest shot is selected. Junk shots and frames are then removed and the number of faces in each shot, the amount of motion and the size of the cluster are all used to score and rank shots for inclusion in the summary. Summaries are then generated and include a speedup of up to 2× for some shots.

The *École Nationale d'Ingénieurs de Sfax ENIS* in Tunisia [10] is a first-time participant and used shot boundary detection to segment the video and then automatically filter shots which are less then 2 seconds in length. Junk shots were detected and removed, and sub-shots with little movement were determined as likely to be camera setup and so were not included in the summary. A genetic algorithm was then used for selecting the final sub-shots for selection in the summary.

A team composed of researchers from the *Intelligent Multimedia Group at Tsinghua U.* and *Intel China Research Center* in China [33] used hierarchical clustering to select representative keyframes and used dynamic programming to remove redundant retakes. They also did junk frame detection and used color histograms and color layout as low level feature representations, both for the whole frame and for regions. For this team, the unit of video being manipulated were 1 second clips, not shots.

The *Tokyo Institute of Technology* team, from Japan [35], focused on the number of scenes from the

original video to be included in the final generated summary. Their units of processing were shots and sub-shots, from which they extracted color features, as well as optical flow characteristics, and used these as the basis for their clustering. From this they then selected segments for inclusion in the final summary. No explicit searching for junk frames was done and some crept into the final summaries, which affected the performance figures.

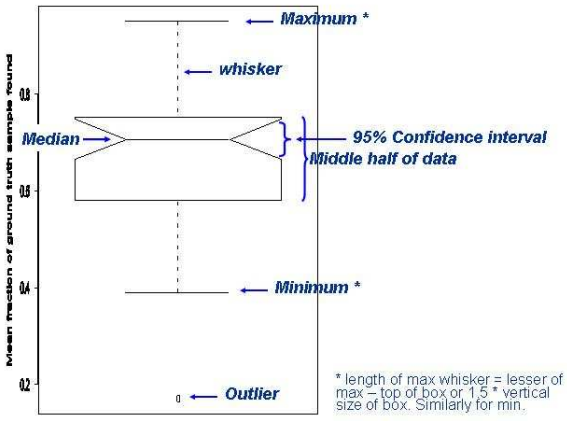
The *University of Electro-Communications* in Tokyo, Japan [19], took an approach that segmented the original video into shots by comparing adjacent frames using color histograms. This was followed by a k-means clustering using color histograms to compute shot-shot similarity. Junk shots were then removed, specifically searching for the sound of a clapper board. To select shots for inclusion in the final summary the approach used face detection as well as the output of shot clustering.

The *University of Glasgow* [23] divided original rushes video into shots. They used, as their unit of information, what is referred to as *sub-sequences*. Their approach used multiple keyframes, taken from each shot, as the unit for computing shot-shot similarity which is used in clustering. Color histograms were extracted from each of 3×3 regions in each frame, with extra weighting given to the region in the middle and at the corners of the frame. Junk frames were explicitly removed as well as "meaningless views", corresponding to over/under exposure.

A team from the *Universite Pierre et Marie Curie - LIP6* in Paris, France [7], began processing the original video by doing shot boundary detection. Shot-shot similarity, excluding shots of less than 2 seconds, was then computed based on color histograms of regions in so-called characteristic frames, and similar shots were then *stacked*. This correspondeds to retakes of shots from the original video. For generating the summary, an adaptive acceleration technique was used, changing playback speed based on the (visual) similarity of frames adjacent in the generated summary.

Finally, the *University of Ottawa - SITE* group in Canada combined with the *Université de Savoie* in France and a group from LAPI, University of Bucharest in Romania to study the spatio-temporal activity levels of input videos [12]. This was done by generating a spatio-temporal matrix of interest points, with explicit removal of junk frames. This matrix was then used to detect repeated clips and segments for elimination and the remaining clips with the highest activity levels were used to generate the summary.

Figure 2: Example of Tukey-style boxplot



6 2008 Results

In this section we present an initial, largely graphic, exploratory analysis of the evaluation results. As mentioned earlier, details of each group's techniques and an exploration of each individual group's approaches and performance appear in the individual group papers in these proceedings. The overall results are of individual measures and are presented as boxplots. Figure 2 gives an explanation of the conventions used in the Tukey-style boxplots. Unless explicitly noted otherwise, scores presented in the following are means of the three judgments for any summary and measure.

Figure 3: Variation in included ground truth by video file

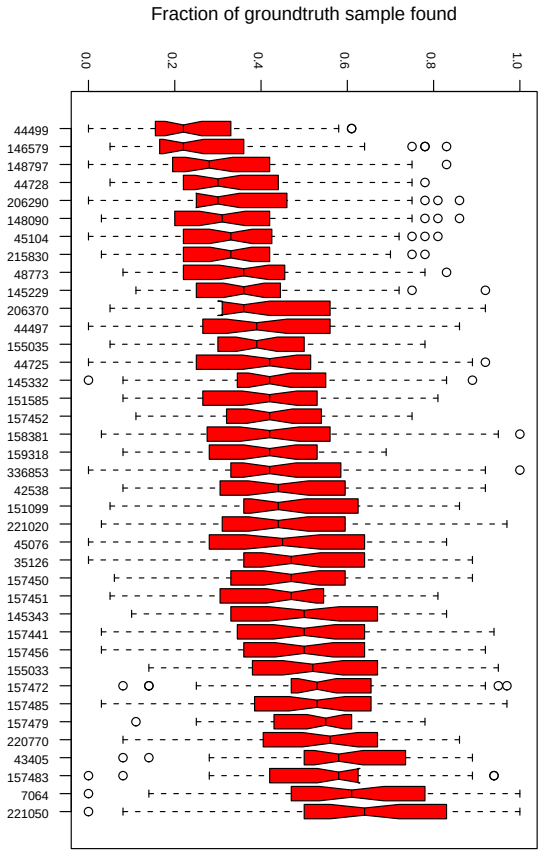


Figure 4: Distribution of included ground truth scores

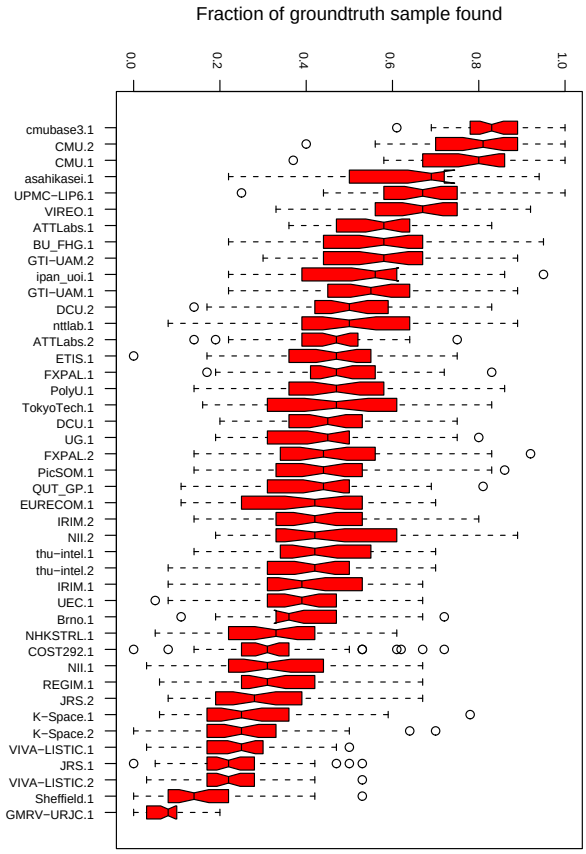
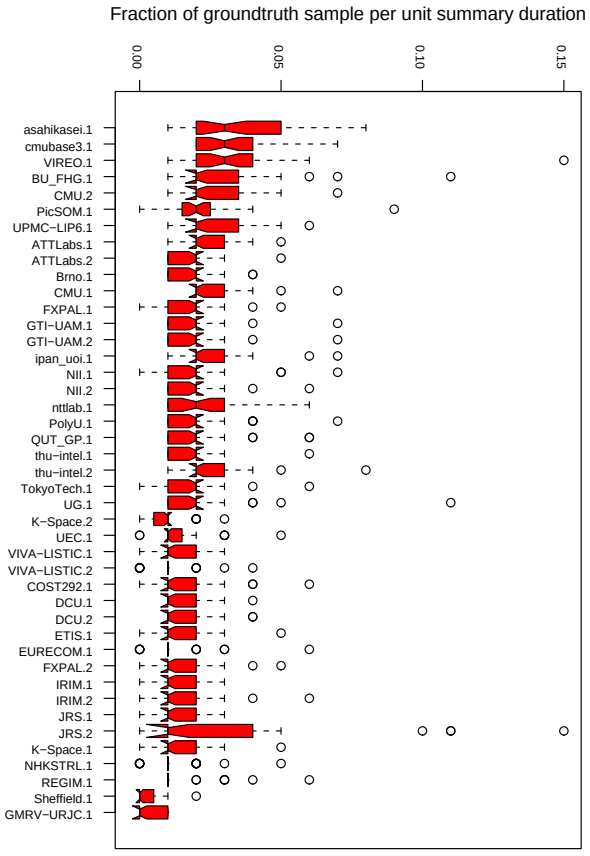


Figure 5: Distribution of included ground truth per unit of summary duration



6.1 Inclusion of ground truth content

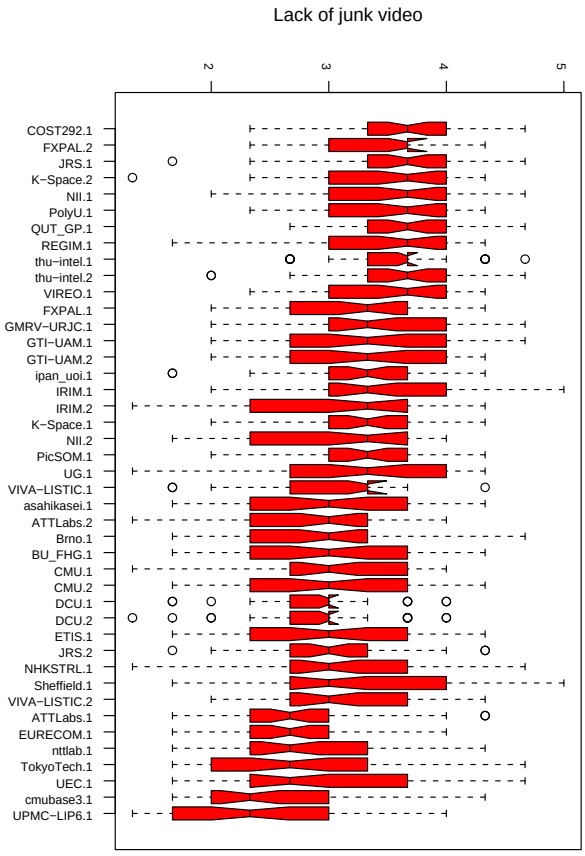
The fraction of ground truth included in a summary could range from 0 to 1 with a granularity of 0.08 (= 1/12). Figure 3 shows the variation in system performance by video file. Note that the effect of any video file on results is confounded with that of the three assessors assigned to judge the summaries of that video file.

In general, it is very difficult to conclude why systems scored high for specific videos and scored low on others. This could be due to the effect of the nature of the ground truth, the assessors' judgments or the video content itself. After checking the set of ground truth of these videos and the set of assessors who judged them, we found that there is no clear evidence of the effect of the ground truth or assessors on the summary scores. However, one plausible effect is the video content itself. Videos that tend to have different scenes in different locations, actors, background, etc., seem to be good candidates for high summary scores, while videos with limited scenes, actors and locations can be more confusing to the systems especially when the systems are looking to remove the high redundancy in the video content. However, all systems were tested on all the video files so any effect due to the videos is distributed equally across all systems and will cancel out.

Figure 4 shows the results by system. The median fraction of included ground truth for all summaries from each participant ranged from 0.08 to 0.83. The baseline system performed at the top of all systems. Since it contained the entire video anyway, it might be expected to do a good job of including the expected ground truth – barring problems due to the speed at which it was presented. A partial randomization test [16] using 10,000 repetitions found the baseline performed significantly better ($p<0.05$) than all other systems in including ground truth.

Figure 5 plots the fraction of ground truth included per unit of summary duration. The view of included ground truth, rewards conciseness. A partial randomization test [16] using 10,000 repetitions found one run (asahikasei.1) performed significantly better ($p<0.05$) than the baseline in terms of included ground truth per unit of summary duration.

Figure 6: Distribution of "lack of junk" scores



6.2 Subjective measures of well-formedness

Measures of three aspects of summary well-formedness were included to reflect usability concerns. These were chosen to highlight three specific characteristics of good summaries that were not well caught by the 2007 evaluation. Scores on all three occupy a narrow range, a difference of only approximately one choice on the Likert scale apart, when one disregards the two or three lowest scoring systems. In this regard the well-formedness scores are less useful than hoped for in distinguishing systems not at the extremes. It is reassuring that the baseline was rated high in redundancy and high in junk video content as would be expected from summaries which presented the entire video at high speed. We take these as sanity checks on the evaluation process. The baseline's low score on pleasant tempo/rhythm raises serious doubts about its user acceptance within the evaluation framework.

6.2.1 Lack of junk video

The lack of junk score was an integer ranging from 1 (worst) to 5 (best). "Junk" was defined as color bars, clapper boards, completely black or completely white frames. Figure 6 shows the results for this measure. The scores ranged from 2.33 to 3.67 with the baseline, which did not attempt to remove junk frames, scoring second-worst.

6.2.2 Redundancy

The lack of redundancy score was an integer ranging from 1 (worst) to 5 (best). The scores for lack of redundancy (Figure 7), ranged between 2 and 4, where 5 signifies that the assessor "strongly disagreed" that the summary contained many repeated segments. Again this year, greater redundancy is correlated with better scores on inclusion of ground truth (see Figure 8) - perhaps because repetition makes the included content easier to see. The baseline performed worse than all 31 submitted runs, as expected since it did not attempt to remove redundant footage.

Figure 7: Distribution of "lack of redundancy" scores

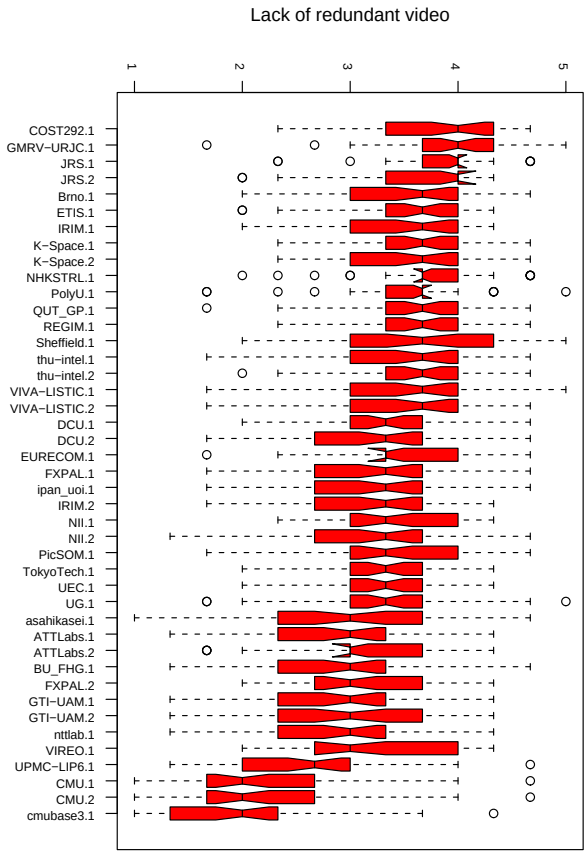
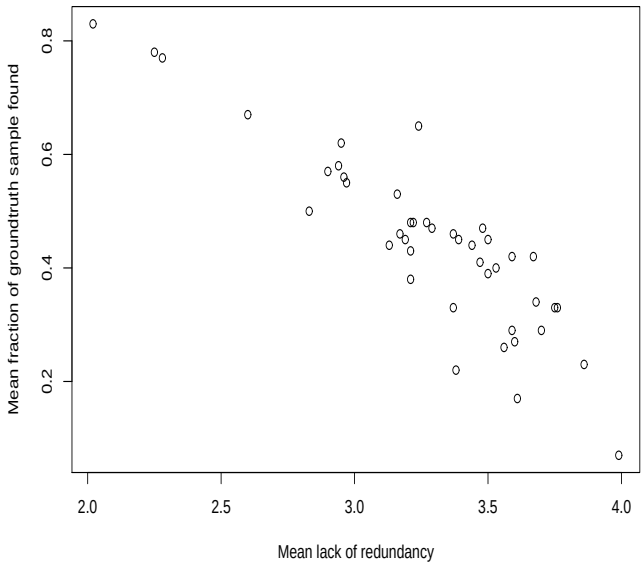


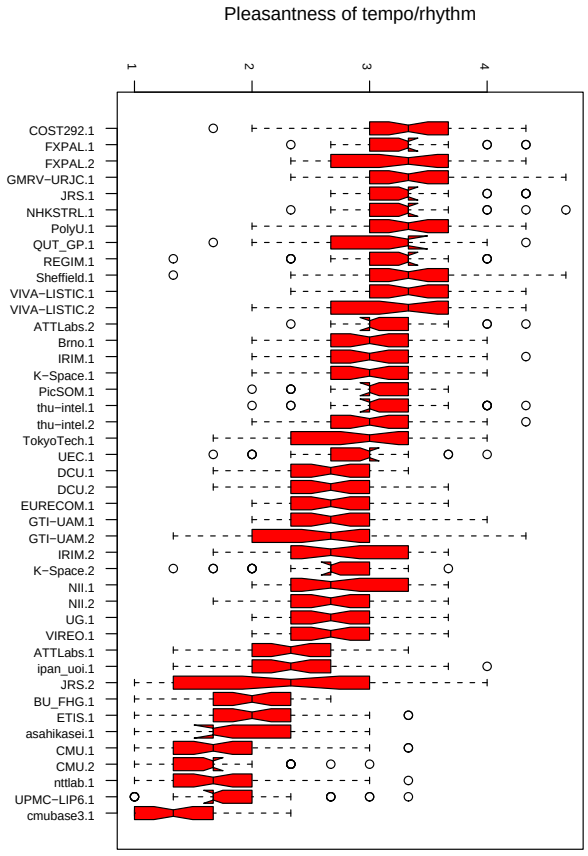
Figure 8: Lack of redundancy vs. ground truth included



6.2.3 Pleasant tempo and rhythm

The pleasant tempo/rhythm score was an integer ranging from 1 (worst) to 5 (best). Figure 9 shows the scores for systems. Scores ranged from 1.33 to 3.33.

Figure 9: Distribution of "pleasant tempo/rhythm" scores



6.3 Assessment time

The median times for judging summaries against ground truth varied, as shown in Figure 10. Per-system medians range from 21.67 to 61.67 seconds. Figure 11 suggests more time spent judging inclusions correlated with higher scores on included ground truth, but the evaluation provides no insight into which was cause and which was effect, if either.

It may be that the assessment time might have some impact on the rate of inclusion of ground truth. There is a case for examining whether judgement time has a correlation with either the duration of the summary, or with the rate of inclusion of ground truth, but initial examination of this did not reveal anything major. This remains a topic for more detailed investigation which we hope to do at a later stage.

6.4 Duration of summary

Most summaries were at or under the 2% limit on duration, as can be seen in the boxplots in Figure 16 where negative values indicate the summary was larger than the target. There was no penalty in the scoring for this violation of the guidelines, but neither did excess duration correlate with including more of the ground truth material as shown in Figure 17.

6.5 Summary creation time

Summary creation times ranged widely from 8 seconds to 16 hours. The median summary creation time was about 32 minutes. Some systems were not optimized for speed in this initial pilot. Longer summary creation times do not correlate well with better results on any of the subjective quality measures, as can be seen in Figures 12, 13, 14, and 15.

Figure 10: Distribution of total inclusion assessment time (seconds)

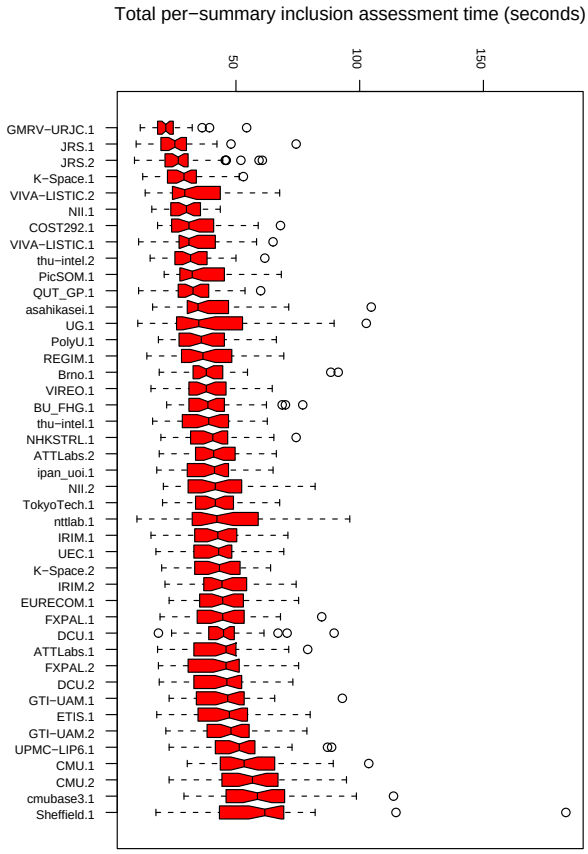
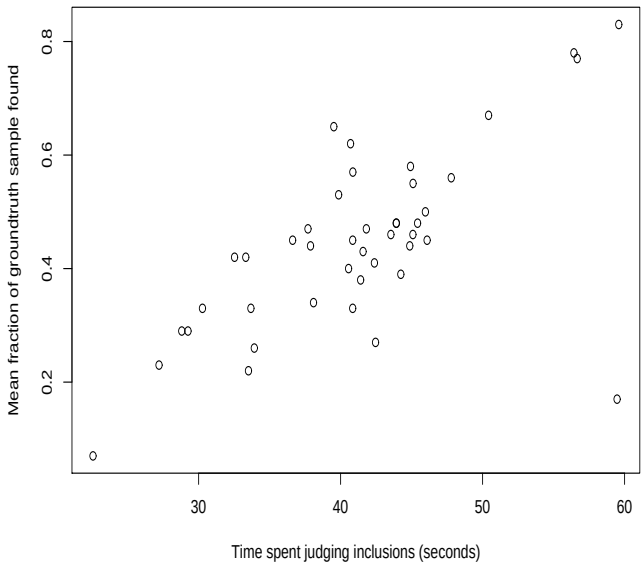


Figure 11: Time spent judging inclusions (seconds) vs. ground truth included



6.6 Summary of results

The following table presents the medians for the major measures for each system, sorted by fraction of ground truth found. Sorting by other measures would yield very different rankings. All times are in seconds. Each of the four scores in the rightmost four columns is a median of the means of the three assessor judgments for each summary and measure. The fraction of inclusions found ranges from 0 to 1. The other three scores range from 1 through 5, where 5 is best.

System.priority	Summary duration	Target summary size - actual	Time judging inclusions	- SUBJECTIVE SCORES --	Fraction of ground truth inclusions found	Lack of junk	Lack of redundancy	Tempo, rhythm
cmubase3.1	33.9	0.40	58.67	0.83	2.33	2.00	1.33
CMU.2	33.9	0.40	56.67	0.81	3.00	2.00	1.67
CMU.1	33.9	0.40	53.33	0.80	3.00	2.00	1.67
asahikasei.1	19.5	9.64	34.67	0.69	3.00	3.00	1.67
VIREO.1	23.6	7.63	38.00	0.67	3.67	3.00	2.67
UPMC-LIP6.1	33.6	0.82	51.33	0.67	2.33	2.67	1.67
GTI-UAM.2	34.1	0.20	48.00	0.58	3.33	3.00	2.67
BU_FHG.1	22.9	7.94	38.67	0.58	3.00	3.00	2.00
ATTLabs.1	29.7	4.82	46.00	0.58	2.67	3.00	2.33
ipan_uoi.1	28.0	5.17	41.33	0.56	3.33	3.33	2.33
GTI-UAM.1	34.3	0.12	46.67	0.55	3.33	3.00	2.67
nttlab.1	25.0	1.05	42.33	0.50	2.67	3.00	1.67
DCU.2	33.3	1.43	46.33	0.50	3.00	3.33	2.67
TokyoTech.1	32.4	1.58	41.67	0.47	2.67	3.33	3.00
PolyU.1	26.0	3.07	36.00	0.47	3.67	3.67	3.33
FXPAL.1	34.4	0.23	44.67	0.47	3.33	3.33	3.33
ETIS.1	33.1	0.92	47.33	0.47	3.00	3.67	2.00
ATTLabs.2	30.9	3.45	41.00	0.47	3.00	3.00	3.00
UG.1	23.8	2.37	35.00	0.45	3.33	3.33	2.67
DCU.1	33.1	1.30	45.00	0.45	3.00	3.33	2.67
QUT_GP.1	21.5	7.17	32.67	0.44	3.67	3.67	3.33
PicSOM.1	22.1	4.05	32.33	0.44	3.33	3.33	3.00
FXPAL.2	34.4	0.23	46.00	0.44	3.67	3.00	3.33
thu-intel.2	19.6	12.32	31.67	0.42	3.67	3.67	3.00
thu-intel.1	28.1	4.09	39.00	0.42	3.67	3.67	3.00
NII.2	32.6	0.75	41.67	0.42	3.33	3.33	2.67
IRIM.2	34.4	-0.10	44.33	0.42	3.33	3.33	2.67
EURECOM.1	34.3	-0.01	44.67	0.42	2.67	3.33	2.67
UEC.1	32.3	2.06	43.00	0.39	2.67	3.33	3.00
IRIM.1	34.4	-0.08	42.67	0.39	3.33	3.67	3.00
Brno.1	30.0	4.42	38.00	0.36	3.00	3.67	3.00
NHKSTRL.1	32.3	0.90	40.67	0.33	3.00	3.67	3.33

Figure 12: Summary creation time vs. ground truth included

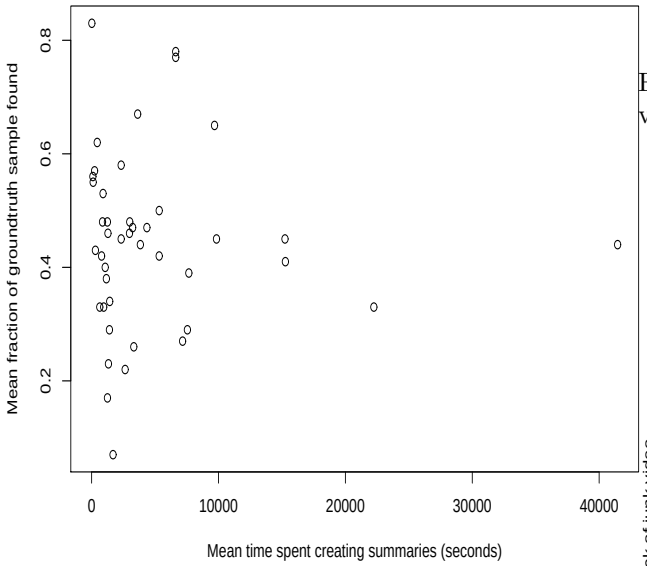
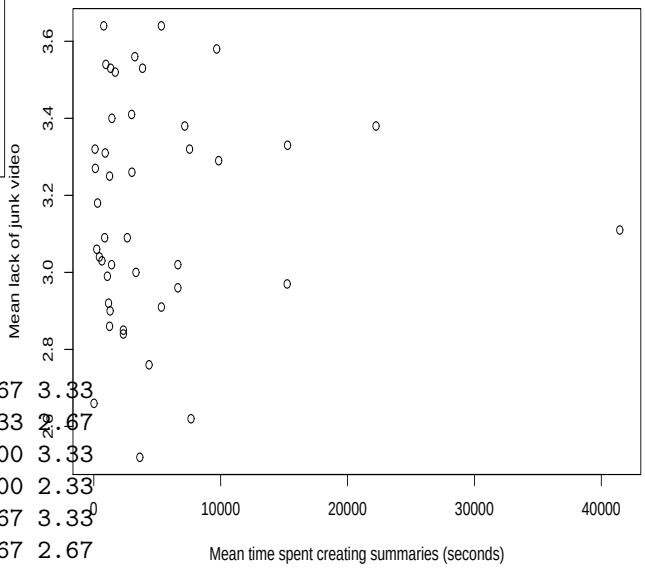


Figure 13: Summary creation time vs. lack of junk video



REGIM.1	28.0	2.65	36.67	0.31	3.67	3.67	3.33
NII.1	20.6	13.32	30.00	0.31	3.67	3.33	2.67
COST292.1	22.8	8.44	31.00	0.31	3.67	4.00	3.33
JRS.2	14.0	14.20	26.67	0.28	3.00	4.00	2.33
VIVA-LISTIC.1	22.1	3.92	31.00	0.25	3.33	3.67	3.33
K-Space.2	34.1	0.02	43.33	0.25	3.67	3.67	2.67
K-Space.1	19.7	11.62	29.00	0.25	3.33	3.67	3.00
VIVA-LISTIC.2	22.1	2.92	29.33	0.22	3.00	3.67	3.33
JRS.1	18.5	13.38	25.33	0.22	3.67	4.00	3.33
Sheffield.1	50.1	-16.83	61.67	0.14	3.00	3.67	3.33
GMRV-URJC.1	13.0	23.02	21.67	0.08	3.33	4.00	3.33

Figure 14: Summary creation time vs. lack of redundancy

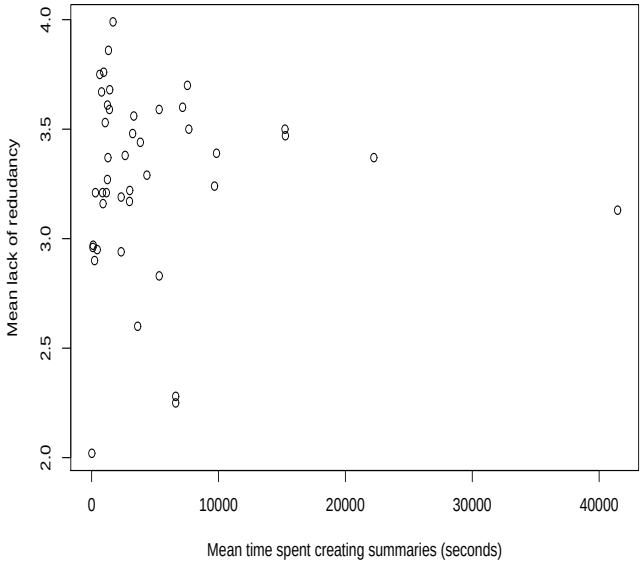


Figure 15: Summary creation time vs. pleasantness of tempo/rhythm

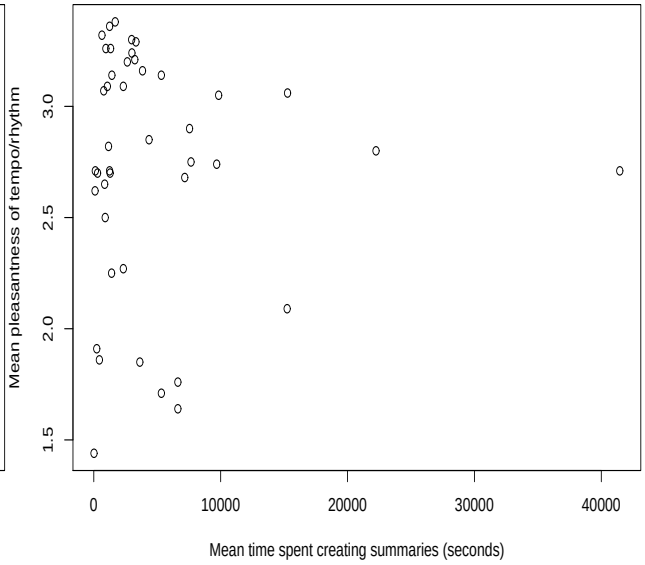


Figure 16: Distribution of excess summary duration (2% duration target - actual summary duration (seconds)

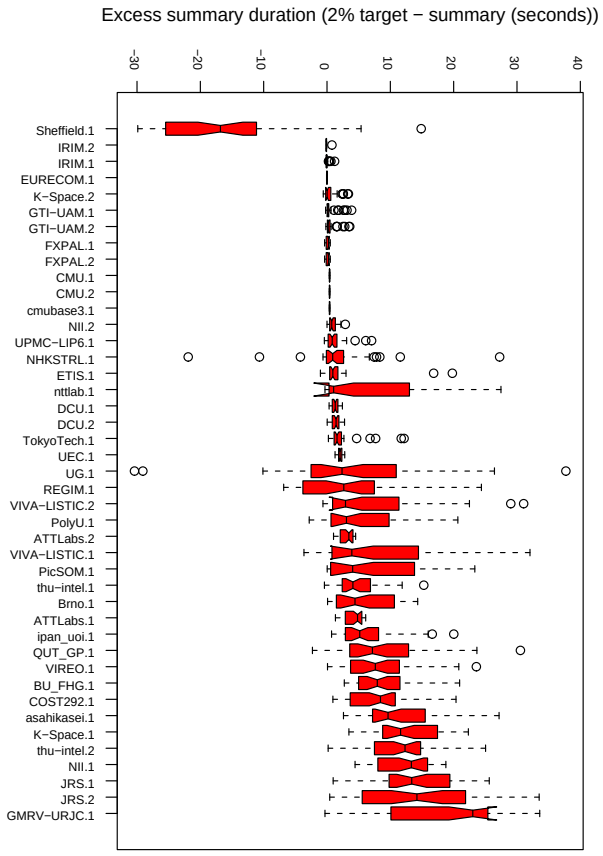
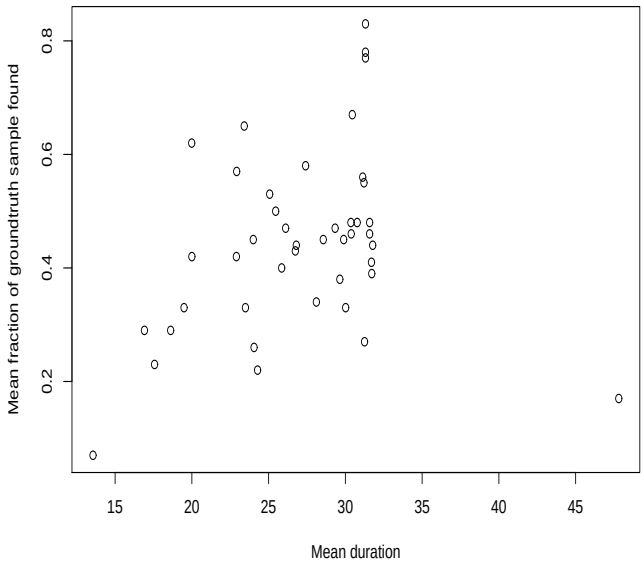


Figure 17: Summary duration vs fraction of ground truth included



7 Evaluating the evaluation

As in 2007, assessor comments indicated they believed they understood the assessment task instructions and were able to carry them out using the software developed for this purpose. Several noted that summaries presented at much faster then original speed made it difficult to see what was included.

The fact that the "fast forward through the whole video" baseline summaries were judged to contain a high proportion of the ground truth but also a high amount of redundancy and junk video provides evidence for the fact that the evaluation was measuring what was intended, and we take encouragement from this.

Triple assessments of each submitted summary provided data in inter-assessor agreement. At the most detailed level of comparison - the binary judgments of the presence or absence of individual ground truth items - mean agreement was 81.7% (median = 83%) compared to 50% agreement that could be expected from chance alone. Agreement exceeded slightly that found in 2007 (78%). The fraction of agreements on a judgment of "no inclusion", which might just be due to inability to see the included material, did not change markedly from 2007 (53.8%) to 2008 (57.2%), although the average summary duration was cut approximately in half.

Pairwise differences in judgments of summary well-formedness showed more consistency than in 2007. The mean and median differences in 2008 are all very close to 1. In 2007 they were 1.443 for ease of understanding and 1.366 for redundancy. Figures 18, 19, and 20 illustrate the 2008 results.

In order to avoid learning start-up effects as each assessor began to judge summaries for a new video and had to get acquainted with a new set of ground truth to look for, the first five summaries were judged again by the same assessor later in the sequence. The scores from the first judgments were not used in the evaluation but can be looked at separately for information on within-assessor consistency. Figures 21, 22, 23, and 24 depict the distribution of differences in the within-assessor score pairs for repeated summaries. The mean differences are 0.07 for included ground truth, 0.6 for each of the 3 quality questions.

Figure 18: Pairwise score differences in lack of junk video

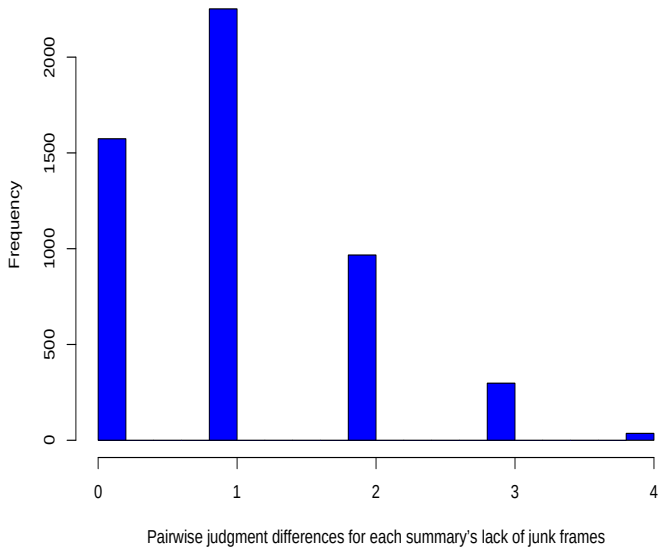


Figure 19: Pairwise score differences in lack of redundant video

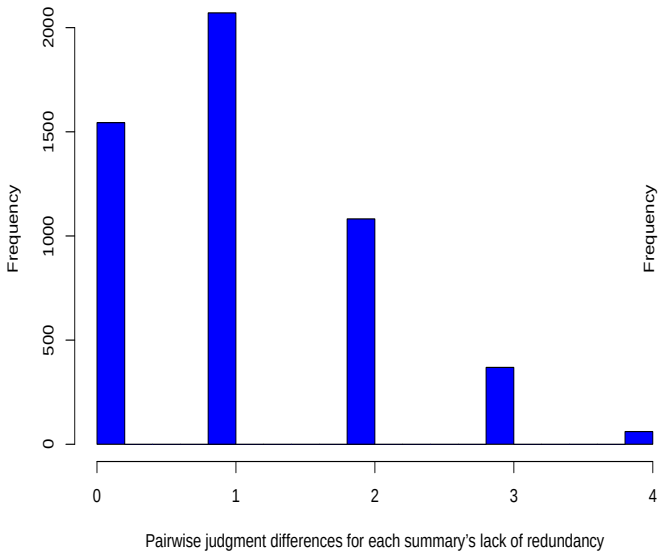


Figure 20: Pairwise score differences in pleasantness of tempo/rhythm

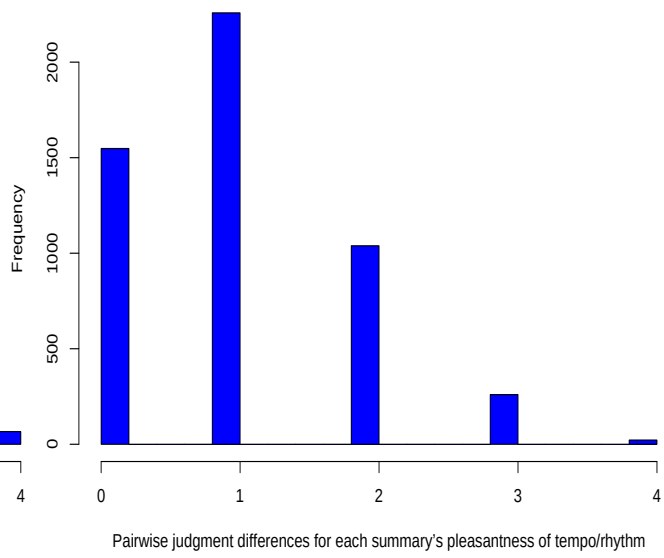


Figure 21: Within-assessor score differences on included ground truth

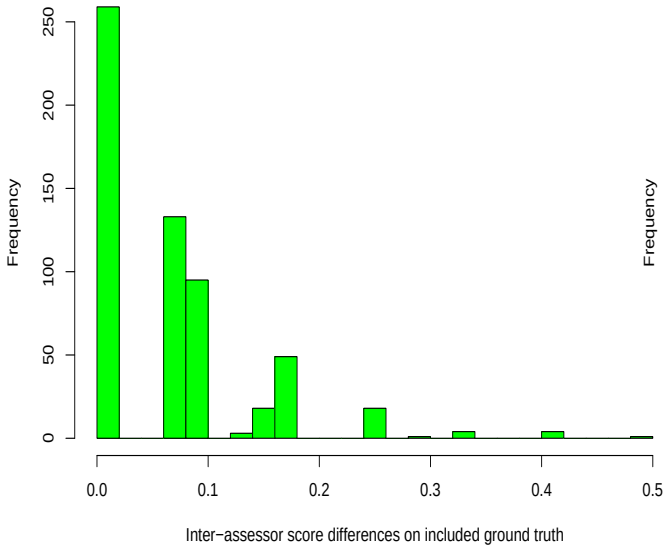


Figure 22: Within-assessor score differences on lack of junk

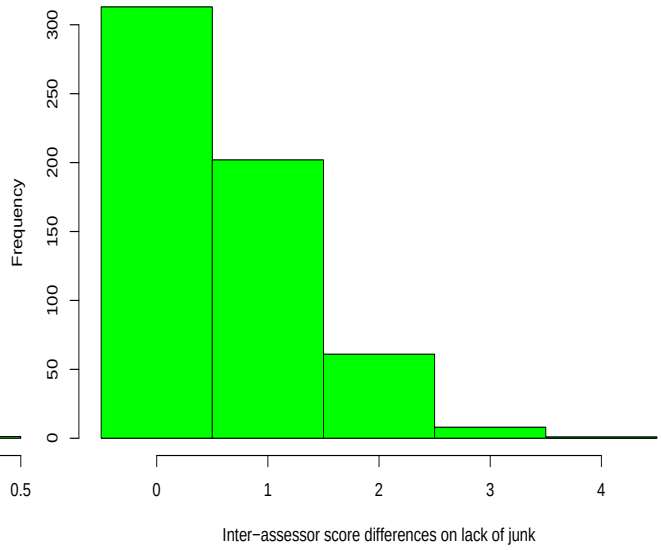


Figure 23: Within-assessor score differences on lack of redundancy

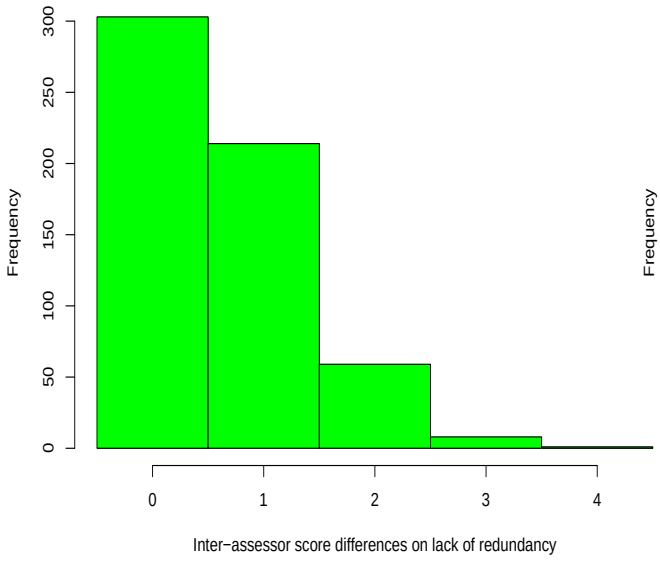
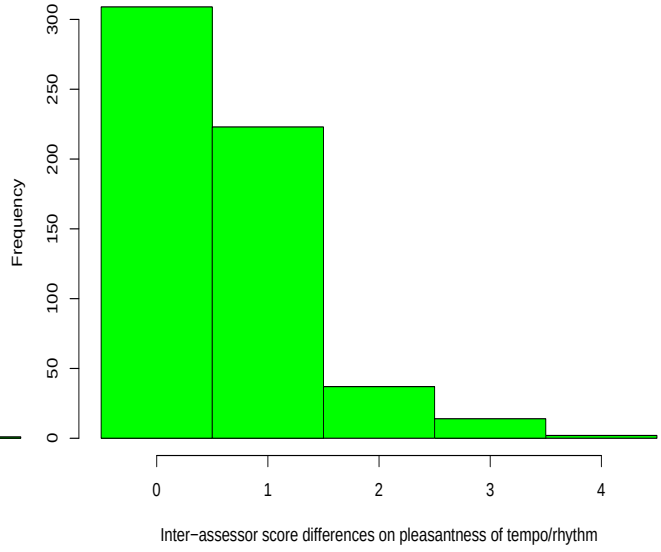


Figure 24: Within-assessor score differences on pleasantness of tempo/rhythm



8 Conclusions

There are many things which we can conclude from this year's TRECVid BBC rushes video summarization evaluation and the first, and most fundamental, is that the evaluation framework seems to have worked again to produce credible results. Clearly, systems stepped up to doubling the compression which had been required in 2007 and scores did not suffer significantly.

It is interesting to examine the techniques used by participants. Almost all groups used some form of shot boundary detection and since there were no gradual transitions in rushes, shot boundary detection for hard cuts worked well. Most techniques used some form of color histogram, and some used motion as well. Almost all participants explicitly looked for junk frames in order to remove them, as well as removing shots of a short duration. Most systems used some form of clustering and these differed in how the shot-shot comparisons were made but using color, with or without regional color, face detection using the OpenCV technique, and motion in the video, were all common. For generating summaries, most groups simply appended selected shots or subshots but many did speedup of video. Use of fast forward seems to be correlated with better scores on included ground truth but worse scores on the other subjective measures of summary quality. Few systems used overlays such as timeliness, on the generated summaries.

It is apparent that the similarities among approaches taken were very strong and a very homogeneous set of approaches were tried. With such homogeneity among approaches one could expect very similar performance results.

While the 3 quality measures that we used in 2008 did detect defects in the baseline not found in 2007, the 2008 baseline was radically different. Perhaps including the 2007 baseline in 2008 would have been ideal but that is not now possible.

In general the narrow range of scores for the well-formedness measures makes them less useful than hoped for distinguishing systems. Using summary duration as a normalizing factor provided a view of the results on including ground truth that showed only one system performing significantly better than the baseline.

Several groups invested a lot of computation time in generating their summaries and one would have expected payback in terms of performance. However, increased time spent in summary creation did not usually yield a better summary on any of the measures. This is similar to the effect we observed in the shot boundary detection task in TRECVid.

Finally, as with all intrinsic evaluations such as this, one is left wondering what real users in a real work environment would think of the summaries produced.

Acknowledgments

The authors would like to thank several individuals and groups for making this video summarization evaluation possible. We are grateful to the BBC archives and to Richard Wright for providing the video data, to NIST and Intelligence Advanced Research Projects Activity (IARPA) and to the European Commission under contract FP6-027026 (K-Space) for sponsoring the evaluation, to the assessors at NIST who created the ground truth and to the assessors at Dublin City University for performing the evaluation, to Philip Kelly at Dublin City University for helping to organize the summary judging, to Carnegie Mellon University for providing the baseline results once again, to several sites for mirroring the video data to allow distribution to participating groups over the Internet, to the program committee and several others for reviewing papers and finally, to all the participating groups for taking part. AS was partly sponsored by Science Foundation Ireland under grant 03/IN.3/I361 and by the European Commission under contract FP6-027026 (K-Space).

A Ground truth creation guidelines

Here we present the final ground truth guidelines as issued to people involved in the ground truth creation process.

Background

A good video summary shows the viewer segments containing examples of the main objects and events depicted in the video it summarizes, filtering out the *unclear* and the *predictable*. One way to evaluate such a summary is to have a human summarizer create a filtered list of such segments, each identified uniquely in terms of an object or event. Then the summary can be compared to the list to see how many of the desired objects/events (i.e., segments) it contains.

Segments

The task of the ground truth creator is to watch a video, select desirable segments, and then identify each uniquely by noting an object (animate or inanimate) or event (i.e., one or more objects involved in some action) occurring in the segment. The number of segments will vary with the video.

It is the nature of rushes that some scenes and parts of scenes will be shot multiple times. The variations in such re-takes, while important to the director, will likely be below the level that matters to a highly compressed summary. That is, the summary need only include one instance. An exception might be something that goes wrong and might have a separate use from other takes that proceed mostly as expected.

A desirable segment should not cross shot boundaries and the ground truth might identify multiple such segments within a single shot while not including extremely short segments separately unless they seem very interesting. The ground truth can include segments from the unscripted portion of the video if they are substantial enough and seem as though they might be reusable. However, they should *not* include the starting/ending clap boards of scenes and takes or the color bars at the beginning.

Items

The object/event cue for each desired segment should be as simple as possible while still identifying the segment uniquely within the video. Uniqueness is primary. For example, if there are two women in a video then the ground truth should include two segments (a close-up of each) and will specify some distinguishing modifiers, e.g., **"woman with glasses"** vs. **"woman with red hair"**, so the person judging the summary against the list can tell when s/he has seen each of the women designated.

Each item needs to be independent of context and should not refer to another, e.g., **"view of road from different angle"** would not be included. Items should be clear even if the order of entries in the ground truth of items was randomized or only a subset was used.

Many videos contain alternate shots of some object/person at different ranges and this is addressed by mentioning what is visible (shoulder and head vs. head only).

Each item should take one of the following forms. either an object (no event or camera event) such as an "antique car" or an "old woman", or a combination of object(s) + event such as a "red hot air balloon ascending" or "people talking", or a combination of object(s) + camera event such as a "pan across room" or a "zoom in on newspaper page", or a combination of object(s) + event + camera event such as a "zoom in on red hot air balloon ascending" or a"zoom in on blimp's cabin touching the water". The set of allowable camera events is limited to: zoom in, zoom out, or pan, where a zoom or pan is an event and a close-up is a state.

The purpose of each item in the ground truth of objects/events is to identify an important segment from the video to be summarized. The item must do this uniquely in the context of that video and minimally, by means of a key object/event, so someone can tell when they see the designated segment in the summary. It is *not* to describe the video's objects/events as one would in traditional annotation of content.

Procedure

The procedure for the ground truth creation process was to play the video at normal speed through one take of a scene, select the distinct segments and enter them as ground truth elements as described above. The creator then re-watched the scene to supplement/check the elements, fast forwarding through the

other takes of the same scene unless something really different and interesting happens.

B Ground truth data checklist

Is each element in your ground truth *UNIQUE* ? as no two elements should be the same

Is each element in your ground truth *INDEPENDENT* ? as each element should stand on its own, e.g., **"View of road from different angle""** is not independent as it assumes you know what the original angle was before it became **"different"**

Is each element/event you have listed *SIGNIFICANT* ? don't list something unless it is clear and complete enough to be useful once found, except if its presence is surprising enough to trump its obscurity or incompleteness

Is there *ONE OBJECT/EVENT* per element ? as there should be no more than 1 per element

Does any element have any *UNNECESSARY DETAIL* ? only the minimum amount of detail that is needed to uniquely describe an element should be given

Is there any element with only *CAMERA MOVEMENT* ? e.g., **"Camera pans right"** probably needs more substance as it unlikely to be the only time in the video when the camera pans right, something like **"Camera pans right onto an object"** gives a more accurate description

References

[1] W. Bailer and G. Thallinger. Comparison of Content Selection Methods for Skimming Rushes Video. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[2] V. Beran, M. Hradis, P. Zemcika, A. Herout, and I. Reznicek. Video Summarization at Brno University of Technology. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[3] H. Bredin, D. Byrne, H. Lee, N. E. O'Connor, and G. J. Jones. Dublin City University at the TRECVid 2008 BBC Rushes Summarisation Task. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[4] V. Chasanis, A. Likas, and N. Galatsanos. Video Rushes Summarization Using Spectral Clustering and Sequence Alignment. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[5] F. Chen, J. Adcock, and M. Cooper. A Simplified Approach to Rushes Summarization. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[6] M. G. Christel, A. G. Hauptmann, W.-H. Lin, M.-Y. Chen, B. Maher, and R. V. Baron. Exploring the Utility of Fast-Forward Surrogates for BBC Rushes. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[7] M. Detyniecki and C. Marsala. Adaptive Acceleration and Shot Stacking for Video Rushes Summarization. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[8] E. Dumont and B. Mérialdo. Sequence Alignment for Redundancy Removal in Video Rushes Summarization. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[9] E. Dumont, B. Merialdo, S. Essid, W. Bailer, H. Rehatschek, D. Byrne, H. Bredin, N. E. O'Connor, G. J. Jones, A. F. Smeaton, M. Haller, A. Krutz, T. Sikora, and T. PIatrik. Rushes Video Summarization Using a Collaborative Approach. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[10] M. Ellouze, H. Karray, and A. M. Alimi. REGIM, Research Group on Intelligent Machines, Tunisia, at TRECVID 2008, BBC Rushes Summarization. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[11] D. Gorisse, F. Precioso, S. Philipp-Foliguet, and M. Cord. Summarization Scheme based on Near-duplicate Analysis. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[12] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. Ionescu. Video Summarization from Spatio-Temporal Features. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[13] D.-D. Le, N. Putpuek, N. Cooharojananone, C. Lursinsap, and S. Satoh. Rushes Summarization Using Different Redundancy Elimination Approaches. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[14] Y. Liu, Y. Liu, and T. Ren. Rushes Video Summarization using Audio-Visual Information and Sequence Alignment. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[15] Z. Liu, E. Zavesky, B. Shahraray, D. Gibbon, and A. Basso. Brief and High-Interest Video Summary Generation: Evaluating the AT&T Labs Rushes Summarizations. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[16] B. F. J. Manly. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman & Hall, London, UK, 2nd edition, 1997.

[17] mplayer. Mplayer - the Movie Player. URL: www.mplayerhq.hu/design7/news.html, 2007.

[18] S. Naci, U. Damnjanovic, B. Mansencal, J. Benois-Pineau, C. Kaes, and M. Corvaglia. The COST292 Experimental Framework for RUSHES Task in TRECVID 2008. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[19] A. Noguchi and K. Yanai. Rushes Summarization Based on Color, Motion and Face. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[20] P. Over, A. F. Smeaton, and P. Kelly. The TRECVid 2007 BBC rushes summarization evaluation pilot. In *TVS '07: Proceedings of the International Workshop on TRECVid Video Summarization*, pages 1–15, New York, NY, USA, 2007. ACM.

[21] G. Quénot, J. Benois-Pineau, B. Mansencal, E. Rossi, M. Cord, F. Precioso, D. Gorisse, P. Lambert, B. Augereau, L. Granjon, D. Pellerin, M. Rombaut, and S. Ayache. Rushes Summarization by IRIM Consortium: Redundancy Removal and Multi-Feature Fusion. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[22] J. Ren and J. Jiang. Hierarchical Modeling and Adaptive Clustering for Real- time Summariza-

tion of Rush Videos in TRECVID'08. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[23] R. Ren, P. Punitha, and J. Jose. Rushes Redundancy Detection. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[24] M. Sano, Y. Kawai, N. Yagi, and S. Satoh. Video Rushes Summarization utilizing Retake Characteristics. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[25] J. Sasongko, C. Rohr, and D. Tjondronegoro. Efficient Generation of Pleasant Video Summaries. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[26] A. F. Smeaton and P. Over. TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In *Proc. of the International Conference on Image and Video Retrieval (CIVR)*, pages 451–456. Springer, 2003.

[27] A. F. Smeaton, P. Over, and W. Kraaij. TRECVID: evaluating the effectiveness of information retrieval tasks on digital video. In *Proceedings of the 12th Annual ACM international Conference on Multimedia*, pages 652–655. ACM Press New York, NY, USA, 2004.

[28] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, October 2006. ACM Press.

[29] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. J. Delp. Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia*, 8(4):775–791, August 2006.

[30] P. Toharia, O. D. Robles, L. Pastor, and Ángel Rodríguez. Combining Activity and Temporal Coherence with Low-level Information for Summarization of Video Rushes. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[31] B. T. Truong and S. Venkatesh. Video Abstraction: A Systematic Review and Classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1):1–37, 2006.

[32] V. Valdés and J. M. Martínez. Binary Tree Based On-line Video Summarization. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[33] T. Wang, S. Feng, P. P. Wang, W. Hu, S. Zhang, W. Zhang, Y. Du, J. Li, J. Li, and Y. Zhang. THU-Intel at Rush Summarization of TRECVID 2008. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.

[34] R. Wright. Personal communication, Technology Manager, Projects, BBC Information & Archives, 2005.

[35] K. Yamasaki, K. Shinoda, and S. Furui. Automatically Estimating Number of Scenes for Rushes Summarization. In *Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia*, New York, NY, USA, 2008. ACM.