The TRECVID 2007 BBC Rushes Summarization Evaluation Pilot

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.
& Adaptive Information Cluster
Dublin City University
Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

Philip Kelly
Centre for Digital Video Proc.
& Adaptive Information Cluster
Dublin City University
Glasnevin, Dublin 9, Ireland
kellyp@eeng.dcu.ie

ABSTRACT

This paper provides an overview of a pilot evaluation of video summaries using rushes from several BBC dramatic series. It was carried out under the auspices of TRECVID. Twenty-two research teams submitted video summaries of up to 4% duration, of 42 individual rushes video files aimed at compressing out redundant and insignificant material. The output of two baseline systems built on straightforward content reduction techniques was contributed by Carnegie Mellon University as a control. Procedures for developing ground truth lists of important segments from each video were developed at Dublin City University and applied to the BBC video. At NIST each summary was judged by three humans with respect to how much of the ground truth was included, how easy the summary was to understand, and how much repeated material the summary contained. 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 78% and results indicate that while it is difficult to exceed the performance of baselines, a few systems did. [1]

Categories and Subject Descriptors

H.5.1 [**Information Systems Applications**]: Information Interfaces & Presentation—*Multimedia Information Systems*

[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.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

Keywords

Video summarization, Evaluation, Benchmarking

1. INTRODUCTION

For several years, the TRECVID evaluation campaigns (Smeaton and Over (2003), Smeaton, Over, and Kraaij (2004, 2006)) have mainly explored the evaluation of video information retrieval systems using a variation of the Cranfield-TREC methodologies. In 2007, TRECVID struck off in a related but significantly different direction — a first, or pilot, attempt at a large-scale evaluation of video summarization systems.

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 (Truong & Venkatesh, 2006; Taskiran, Pizlo, Amir, Ponceleon, & Delp, 2006). 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 this paper we present an overview of the TRECVID 2007 video summarization evaluation campaign including a description of the goals of the evaluation, the video data used, the task set for the participating groups, the evaluation approach used, including the procedure used for creating the ground truth. We also include an overview of the results of the 22 groups (see Table 1) who completed the summarization activity, though the details of each group's activities 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. PREVIOUS WORK ON EVALUATING VIDEO SUMMARIES

There have been a number of earlier studies of video summarization, some of which include evaluation of the approaches taken. These tend to have looked at related but different situations to what we address in the TRECVID 2007 summarization task and several are specialized to a specific genre. Some are extrinsic, i.e., in terms of how a summary helps in some task, rather than intrinsic i.e., direct evaluations and most do not compare summaries to the full video being summarized.

Ding, Marchionini, and Tse (1997) carried out an extrinsic evaluation of slideshow summaries. They looked at the effect of presentation rates and participant characteristics on whether participants could tell that images (taken from a slideshow summary) were from the summary and whether participants could choose textual descriptions of objects (that occurred in the summary) from a list. In each case, half the images or object descriptions were distractors.

Ferman and Tekalp (2003) report an intrinsic evaluation in which a "neutral observer" evaluated four summaries (two methods applied to two MPEG-7 test videos), apparently with knowledge of the video to be summarized, to determine the number of redundant or missing frames based on how well the frames aided identification of objects and events.

Komlodi and Marchionini (1998) studied the usability of three user interface designs for presentation of key frames extracted from videos: 4 key frame static, 12 key frame static, and 12 key frame dynamic. The study attempted to test the designs in a setting meant to resemble a real-life information-seeking situation. Comparison was in terms of object/action identification, gist comprehension, selection precision/recall, examination time, and user satisfaction. Participants watched the summary and then worked with object identification lists, wrote a sentence to capture the gist, and answered a multiple choice question about the best one-sentence description of the summary. Static storyboards were shown to be significantly better than dynamic ones in supporting object identification and memory: "The animation of the key frames made it more difficult for subjects to make out, recognize, and remember details of the pictures".

Christel, Smith, Taylor, and Winkler (1998) carried out two extrinsic evaluations of various kinds of video skims as measured by comprehension, navigation, and user satisfaction in two tasks: fact-finding and gisting. The video came from 3 public television series. The gisting task is relevant to the summarization task under discussion here. After viewing a skim, the participants in the experiment were shown lists of text-phrases and thumbnail images and were asked to indicate which phrases and images best represented the material covered by the skim. The lists were "populated with independently validated text phrases and representative images". The text was not lifted verbatim from the original video but composed by Informedia staff "as video descriptors for library use". The images on the other hand came from the original video and in their second experiment the images all appeared in each skim as well. Participants in the second experiment were asked if "the image was part of the video [summary] they had just seen".

He, Sanocki, Gupta, and Grudin (1999) directly evaluated three techniques for automatic summarization of four online presentations (slides, audio, video (mainly talking head)) and author-generated summaries in a user study. Participants completed quizzes before and after viewing each summary to gauge any change in their understanding of the topic due to having watched the summary. While watching the summary they were allowed to pause, jump backwards, or jump forwards. The summaries were about 20% the size of the original presentations.

Ekin, Tekalp, and Mehrotra (2003) performed an intrinsic evaluation of three types of summaries tailored very specifically to soccer games: all slow-motion segments in a game, all goals in a game, and slow-motion segments classified using object-based features (referee, penalty box). Compression rates were on average 12.78% for all slow-motion summaries and 4.68% for all goal summaries. Precision and recall were the measures used.

More recent work by Taskiran et al. (2006) tested the informativeness of video summaries in two ways. In an extrinsic test, 48 participants who had seen only the summary (1 play, straight through) answered 10 multiple choice questions (1 of 4 choices was correct) composed based on the closed-caption transcripts of the full video by two authors. In an intrinsic test, each summary was evaluated based on how many of the above questions were answered in the summary and on user satisfaction. The emphasis in the algorithm's test and the ground truth questions was strictly on the text-from-speech. The number of summaries evaluated was small.

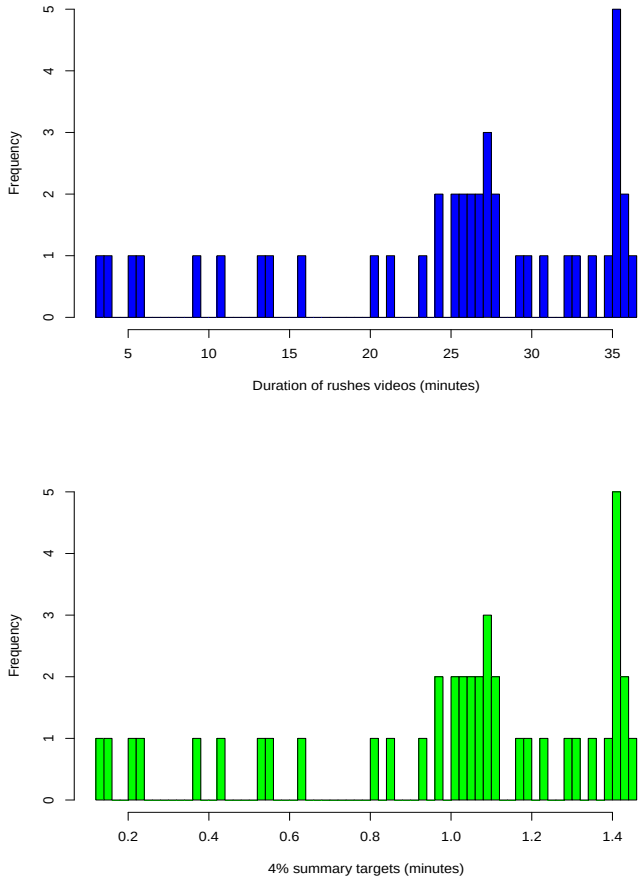
Finally, Marchionini (2006) presented a theoretical discussion of two sets of human cognitive performance measures for object and action recognition and inference from gists, as well as user satisfaction measures — all with respect to visual surrogates. The object and action recognition measures made no reference to the objects and actions of the full video. The inference measures used ground truth based on full knowledge of the video to be summarized. Given a summary and list of words describing objects, users were asked to select the words that describe objects they saw in the summary. For example, given a video summary and list of statements describing the video to be summarized, choose the statement that best describes the video to be summarized (based only on the summary), or given a summary and a list of keyframes, choose the keyframes that belong to the video to be summarized (based only on summary).

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 but that to date, the datasets used have been small and based on the efforts of just one group. In the 2007 TRECVID evaluation pilot we provided a reasonably large video collection to be summarized, a uniform method of creating ground truth and a uniform scoring mechanism. The next section describes the video data we used.

3. VIDEO DATA

The video to be summarized in the TRECVID 2007 evaluation campaign was of a particular sort that presents special problems and opportunities. It consisted of raw (i.e., 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,

Figure 1: Distribution of test video durations (top) and the 4% target summary durations (bottom) (minutes)



a program on emergency services, a police drama, as well as miscellaneous scenes from other programs. About 50 videos were provided to participating groups as development data and 42 were withheld for use in testing the systems once developed. Each set of videos represented a random sample balanced with respect to the number of videos from each series. The test videos had a minimum duration of 3.3 minutes and a maximum duration just under 36.4 minutes, with the mean duration being 25 minutes. Figure 1 presents the distribution of the 42 test video durations and the duration of the 4% target summaries. Sample ground truth was provided for about half of the development videos and ground truth was also created for the test videos.

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 "take" boundaries.

Rushes are potentially very valuable 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. (Wright, 2005). It is hoped that the ability to summarize such rushes might contribute significantly to an overall rushes management and exploitation solution.

4. 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 4% of the duration of the video being summarized. This meant that the average video (25 minutes in length) would have a summary lasting at most 1 minute.

The choice of 4% was somewhat arbitrary, as no complete, detailed information about redundancy in each of the test videos was available. The motivation for choosing this compression factor 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 60 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 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. But 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.

5. 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 easy it was to find the desired content, and how much redundant video the summary contained.

At NIST, 7 retired adults with computer skills spent a total of 221 hours (about 4 hours per day over about 9 days) judging the summaries using software written by NIST for that purpose. Each submitted summary and each baseline summary of each of the 42 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 24-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 construction of this ground truth is described in Section 6 below. The assessor viewed the summary once in a 125 mm x 102 mm mplayer (mplayer, 2007) 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 effort.

The evaluation also collected usability/satisfaction information from the assessor with reference to each system's summary style. 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 following: "It is easy to see and understand what is in this summary."

A second question with the same answer scale as the first was added to provide an estimate of how much redundant material was included in the summary: "This summary contains many nearly identical segments"

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 group to mitigate the presumed start-up bias and provide some input on assessment reliability. Before beginning to judge summaries in a group, the assessor was instructed to play the full video (at c.5 times realtime) as many times as desired while studying the list of desired 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
- ease of finding desired content as judged by assessor
- amount of near redundancy as judged by assessor
- assessor time taken to determine presence/absence of desired segments
- size of summary (number of frames) relative to the 4% duration target
- elapsed time for summary creation

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.

6. CREATING GROUND TRUTH DATA

In order for a video summary to be evaluated, a ground truth list of the important video segments occurring in the original video was required. This involved manual annotation of around 25 development data videos and 50 test data videos, however, as outlined in section 6.4, a number of test and development videos were dropped from the final sets. This resulted in a final ground truth set of 21 development and 42 test videos.

6.1 Preliminary work

The initial conception of ground truth was a list of objects and events, but the application of this view to even a small sample of the data quickly made it clear that there would be too many such items - even using some guidelines from library cataloging to limit the detail. As a result the working notion of ground truth was changed to be a list of important video segments, each identified by means of a distinctive object or event occurring in the segment.

However, the structure of the rushes is such that the same action is often filmed from multiple viewpoints and/or at multiple distances and these differences are significant and should not be ignored in creating the ground truth. As a result, items in the ground truth list often had to be lengthened with qualifications concerning camera angle, distance, or some other information to make each item description unique. This had potential consequences for the assessment procedure since it could make it harder for the assessor to keep the items in mind and separate. Also, the number of ground truth items for each video was on average some two dozen.

In response to this potential problem the guidelines for ground truth creation emphasized minimal, unique item descriptions and the evaluation, as discussed in Section 5, incorporated sampling the full ground truth, study of the full video, and double judging of the first 5 summaries in each group of summaries.

6.2 Creating ground truth for development data

The initial stage in the ground truth creation process involved the formalization of the ground truth process as a set of guidelines, (see Appendix A), which could then be issued to each person involved in the ground truth creation process to ensure a consistent ground truth format. The creation of the guidelines was an iterative process. Initially, a first draft was created which was issued to a group of four people to test the guidelines on sample BBC video data. The resultant ground truth data was examined by the group as a whole to identify any issues, such as unexpected scenarios where no guidelines exist, ambiguous guidelines, and ground truth format issues. Using this input the guidelines were updated and reissued. This cycle of updating guidelines, testing and reviewing, was iterated until the set of guidelines became coherent and stable. The final guidelines, as issued to people involved in the ground truth creation process, consisted of an overview, the guidelines for creating and formatting ground truth data, and guidelines on the approach of how to apply these, and they are presented in Appendix A.

Using the guidelines, the creation of the ground truth for the development data was undertaken involving a new group of eleven people. Each person was presented with a number of videos from the development dataset and resultant ground truths highlighted a number of issues including ambiguity, inclusion of multiple or unimportant events, inclusion of identical events, or events contextually linked to other events in the ground truth. To filter out these errors, each person was issued with a brief checklist, included here in Appendix B, adding an extra check used after a ground truth had been created to ensure that it adhered to the guidelines.

One issue that was left outstanding was to ensure that the granularity of the ground truth data was consistent throughout the entire development dataset. The guidelines state that only *significant* objects or events should be listed in the ground truth data, as insignificant objects should not be part of the generated video summaries. The inclusion of this rule, however, makes the ground truth generation inherently subjective. To address this, a ground truth granularity consistency test was undertaken. When all the development ground truth data was obtained, a final granularity test was undertaken by two people, working in close contact, to thoroughly filter each ground truth to obtain consistency throughout the entire dataset.

6.3 Creating ground truth for test data

A different approach was undertaken during the generation of the ground truth for the test data because of the larger number of videos. When allocating test videos for each person to ground truth, it was ensured that each person's allocation overlapped by a single video with one other person, and overlapped by another video with a second person. Though this increased the workload, it simplified the granularity consistency test for the ground truth for the test data.

The process for creating a consistent granularity across the test ground truth data was initially begun in a manner similar to before, whereby by two people, working in close contact, filtered ground truth from a single individual, creating a final, possibly altered, *template* ground truth. Where a second person in the group had created a ground truth for the same video, then this was compared to the first template, giving an indication of the granularity of the second person's ground truths. If the granularity between the two were similar, it was assumed that all the second person's ground truths were at a similar granularity, and therefore the rest of the second person's ground truths were not tested against the relevant videos, instead they were just checked against the rules in the appendix. If, however, there were large discrepancies between a template ground truth and the second person's ground truth, each of the second person's ground truths were inspected more thoroughly.

6.4 Ground truth data discussion

The length of the ground truth data created by our users for a given video tends to be proportional to both the amount of activity occurring and the structure of the video. For this summarization task the structure of the videos we used tends to be fairly static, with fixed cameras and repeated scripted scenes. Video *MRS146241*, for example, is both structured and has relatively little foreground object activity, resulting in just four elements in the ground truth;

- **"side view of woman sitting on couch with garden in background"**
- **"woman sitting on couch faces camera with garden in background"**
- **"camera zooms in on woman crying"**
- **"close up of woman, head and shoulders in view"**

Video *MRS205522*, however, is less structured as it has few repeating shots. This results in a lot of original shots in that video, thereby increasing the number of elements in the ground truth. This video results in a total of 77 elements of ground truth data including;

- **"close up of blonde woman putting lipstick on woman with brown hair"**
- **"camera tilts down to brown haired woman waving"**
- **"camera pans as woman with brown hair walks through airport"**
- **"empire state building"**
- . . .

The current ground truth process, as is, lends itself well to this sort of structured video. However, videos with a lot of foreground activity or videos that have a relatively free structure, whereby the cameras either tend not to be static and have few repeating scenes, tend to result in large ground truths. It is for this reason that a few videos, such as *MRS114850*, *MRS020502*, *MRS020503*, and *MRS205531*, were dropped from the original test set of videos as a consensus was not determined on how to deal with such videos. Evaluating against such video ground truths would create a measure that was more of an indication of the person doing the ground truth rather than the performance of the summarization system.

In future systems it may be possible to automatically separate strictly scripted videos and unstructured material, so systems could handle both differently. As regards the ground truth of such videos, perhaps the line in the ground

truth guidelines, *"You can include segments from the unscripted portion of the video if they are substantial enough and seem as though they might be reusable."*, could be used to alter the granularity of the ground truth, so that the granularity could be proportional to the level of how structured the video is.

7. PARTICIPANTS AND THEIR APPROACHES

Twenty-two 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 each participant's approach. 17 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 *AT&T Labs* system adopted a multimodal approach to rushes summarization. Their system relied on speech and face detection to create a human-centric video summary. The video was first segmented into shots and three keyframes were selected for each shot. Based on the dissimilarities among corresponding keyframes, AT&T Labs then computed a shot distance matrix, and applied an hierarchical agglomerative clustering algorithm to remove redundancy. For each cluster, the longest shot was kept, and the total budget (less than 4% of the original duration) was assigned to all chosen shots based on their durations. Within each shot, they then picked one continuous segment that contained the most speech and the greatest number of face occurrences. The final video summary was then the concatenation of all selected segments in their original time order.

Brno University of Technology in the Czeck Republic took an approach to summarization based on shot boundary detection, detection of "junk" frames which should be excluded from a summary, and the remaining shots are clustered using principal component analysis (PCA). The generated summary is not just a sequence of keyframes or video clips from the original video but is composed of thumbnails and extra textual information such as shot duration, etc.

Carnegie Mellon University provided the evaluation with two baseline summarization approaches, described in § 8 of this paper. In addition, their own submission was created based on iterative color clustering with noise filtering, backfilling of unused space and audio coherence. Noise filtering included removing unusable shots, such as color bars and clapper shots, and the clustering of the remaining frames used k-means clustering in order to detect clusters of repeated shots from which only one would be included in the summary. Audio coherence was achieved by applying automatic speech recognition and composing an audio track for the summary based on the shots selected from visual analyses, above, but breaking the audio only when there was a pause in speech. This yields a more coherent overall summary.

The *City University of Hong Kong (CityU)* developed a summarization approach based on a complex and detailed analysis of the original video. This included shot boundary detection, object detection, camera motion estimation, keypoint matching and tracking, audio classification and speech recognition. Their system filtered undesirable shots in a one-pass operation and then minimized the inter-shot redundancy by repetitive shot detection, which is particularly well-suited to rushes content. A representability measure was then used to model the presence of objects and four audio-visual events in a video clip, namely motion activity of objects, camera motion, scene changes, and speech content. The video clips with the highest representability scores were selected for inclusion in the generated summary.

Columbia University exploited the highly-similar and repetitive nature of the rushes content, and leveraged their near-duplicate image detector, used in previous TRECVID campaigns, and weighted concept similarity scores to produce an ordered summary with montage-capable rendering. They used shot segmentation and ASR systems provided by AT&T Labs and performed dynamic programming to better guarantee audio continuity in the final output. Founded on image clustering, this approach was found to give a fair allocation of time to both scenes with human interaction and environment, establishing shots intentionally framed during direction.

The *COST-292* group is a collaboration among 25 academic and industry partners from the European Union.Their summarization approach was performed in 3 steps. Firstly, based on a frame clustering algorithm, segments in the video with similar visual content were identified and extracted. Then, using a combination of face detection, camera motion detection and audio excitement analysis algorithms, they calculated an "importance function" for segments along the whole run of the video. In the final step they picked the maximum importance point of each and every cluster and used this to create the summary. The length of the sequence which each cluster contributes to the overall summary was proportional to the mean importance of that corresponding cluster.

Curtin University adopted an approach based on selecting summary clips from a tight clustering among shots/keyframes produced via SIFT matching. Scale-invariant feature transforms (or SIFT) are based on extracting distinctive features from images or video keyframes, and are used to match different views or perspectives of the same object or scene. Normally, SIFT implementations are computationally expensive but in this work the authors examined an approach without complex implementation in terms of concept detection or excerpt assembly (i.e, no picture- in-picture, split screen and special transitions). Although the approach dis not perform very well in terms of concept inclusion, the assessors ranked the resulting summary as easy to understand and with a low level of redundancy. The authors performed an analysis of their results from TRECVID that revealed several promising directions for resolving shortcomings of their technique.

Like many other groups, *Dublin City University* also presented a generated summary as a set of keyframes selected from highly ranked shots and used a gradual transition between these keyframes as well as an audio cue to indicate a transition. This was done to overcome the human phenomenon of change blindness, whereby we tend to take time to register a change of scenery, such as a shot change in a video summary. The DCU approach for keyframe selection was based shot boundary detection and keyframe selection, followed by determining which shots/keyframes to include in the summary based on the amount of motion in the shot and the number of faces appearing. The summary was rendered as a series of keyframes, with a scrollbar within the

Table 1: Participating teams

AT&T Labs	attlabs
Brno University of Technology	brno
Carnegie Mellon University	cmu
City University of Hong Kong (CityU)	cityu
Columbia University	colu
COST292	cost292
Curtin University	curtin
Dublin City University	dcu
FX Palo Alto Laboratory Inc.	fxpal
Helsinki University of Technology	hut
Hong Kong Polytechnic University	hkpu
Institut EURECOM	euecom
JOANNEUM RESEARCH Forschungsgesellschaft mbH	joanneum
KDDI R&D Labs, Inc., Tokushima U., Tokyo U.	kddietal
LIP6 - Laboratoire d'Informatique de Paris 6	lip6
National Institute of Informatics	nii
National Taiwan University	ntu
Tsinghua University - Intel Chinese Research Center	thu-icrc
University of California at Santa Barbara	ucal
University of Glasgow	uglasgow
Universidad Autónoma de Madrid	umadrid
University of Sheffield	usheff

summary video to indicate the offset of the keyframe in the overall video, along with an iconic indication of the amount of movement and the number of faces in that shot. The DCU group also included a detailed failure analysis of their performance.

The *FX Palo Alto Laboratory Inc.* generated summaries based on identifying segments of the original video where the color distribution and camera motion are similar, an approach well-suited to rushes video with its repetitive structure. Unusually among the participating groups, audio features were used, in this case to identify clapboard appearances so they could be excluded. Representative segments from each cluster were then played back within the summary at a higher rate than in the original video based on the detected camera motion in the segment. Pitch-preserving audio processing was also used to capture the sense of the original audio and metadata about each segment was overlaid on the summary to help the viewer understand the context of the summary segments from the original video.

Helsinki University of Technology have been using Self-organizing Maps (SOMs) in various TRECVID tasks for some time and in their approach to video summarization they used SOMs in multiple stages. Their approach involved an initial shot boundary detection followed by shot similarity assessment and pruning to eliminate redundant shots, with both stages implemented using multiple parallel SOMs and within the PicSOM framework. They also pruned out "junk" frames by developing specific detectors for color bar test screens, black frames and white frames (wf). They also applied face detection using Intels OpenCV Library and motion estimation, and all this was used to select 1 second clips to compose the generated summary.

The *Hong Kong Polytechnic University*, like most others, used an approach of shot bound detection to structure the rushes video. This was followed by detection of "noise" shots including clapper boards, color rainbows, and blank shots and this detection was done using a combination of visual, and audio, characteristics. Remaining shots are then examined for visual redundancy, the criterion being color histogram-based similarity indicating repeated shots. One second clips are then taken from non-redundant shots and used to compose the final summary.

Institut EURECOM presented a summarization approach whereby the video was segmented into shots, and the most important and non redundant shots were selected for inclusion in the summary. During presentation in the summary, shots were dynamically accelerated according to their motion activity in order to maximize the content included in the summary per time unit. In addition, shots to appear in the summary were optimally grouped into sets of four and presented simultaneously using a split-screen display, though this did not rate highly with the assessors in terms of ease of use.

JOANNEUM RESEARCH developed an approach based on clustering takes of one scene shot from the same camera position. Their technique uses a variant of the LCSS algorithm to find matching subsequences in sequences of extracted features from the source video, making it well-suited to the repetitive nature of rushes video. The generated summary is a sequence of video clips taken from the original video, and where appropriate there is a text insert in the summary showing the number of alternative takes of the scene.

KDDI R&D Labs, Inc., Tokushima U., and Tokyo U. used a strategy for rushes summarization based on prioritizing rushes segments using low level features and eliminating segments with low priorities, instead of detecting exact objects and camera movement based on relatively high level features. By analyzing audio-visual features in the compressed domain, their proposed method is very fast and can summarize rushes at 1.7% of the original video file duration time.

LIP6 - Laboratoire d'Informatique de Paris 6 took an in-

teresting, and effective, approach to video summarization based on shot boundary detection and detection and elimination of redundant, repeating shots in a technique they call "stacking". For construction of the video summary, they used a technique to vary the playback speed for each shot, depending on the shot duration, which was rated quite well in the assessment and overall performance of this technique was good.

The *National Institute of Informatics* in Japan, developed a summary generation technique which uses video decomposition and feature extraction followed by clustering of the resulting fragments, which are subsequently merged into segments. The segments are then skimmed at a rate which can vary and which is calculated automatically by the system and the summary is constructed.

The *National Taiwan University* had an interesting approach to automatic detection of clapperboards, which feature extensively in some rushes video. First, though, they applied shot boundary detection to structure the video into shots and sub-shots, followed by the detection of "junk" shots, such as clapperboards. Using shot similarity based on color histograms and motion vectors from the compressed domain, they than clustered the remaining shots using hierarchical agglomerative clustering to detect and remove duplicate or repeated shots, and used the remainder to compose the summary.

Tsinghua University and the Intel Chinese Research Center in Beijing, China developed an approach which first constructs story boards for the original video in order to let the user know about the main scenes and main actors in the video. Then they detect and remove useless or "junk" frames, e.g. color bar, near-monochrome/ abrupt/shaking frames, and clapper boards etc. Finally, they construct the video skim by key frame clustering, important factor analysis and repetitive segment detection and removal. The distinguishing feature about this approach is the construction of the storyboard in the early stage of the video analysis.

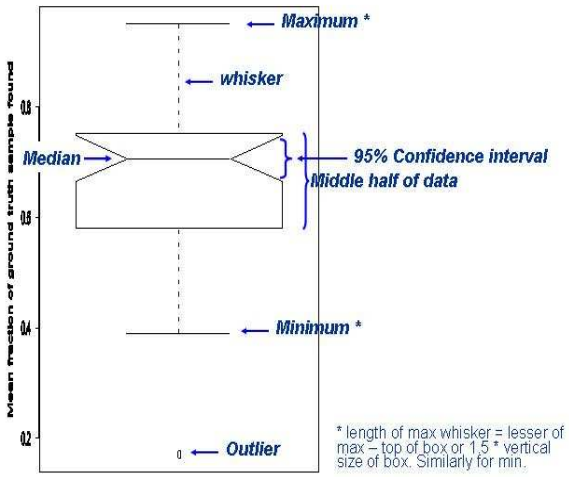
The *University of California at Santa Barbara* used high-level feature fusion to identify segments for inclusion in the video summary. Their approach aims to capture distinct video events using a variety of features including k-means based weighting, speech, camera motion, significant differences in HSV colorspace, and a dynamic time warping (DTW) based feature that suppresses repeated scenes.

The *University of Glasgow* took an approach of formalizing the summarization process as an 0-1 Knapsack optimization problem. Like others, they had a 3-phase process involving analysis (shot segmentation, feature extraction, raw video discrimination and shot clustering), content selection (weighting the importance of video segments by an attention model) and summary generation.

The *Universidad Autónoma de Madrid* took an unusual approach of realtime generation of video summaries. That means that the received video is processed and either written to a summary, or skipped, while it is being received without considering the fragments of video following, and without the possibility of deleting already written sequences. This means that the processing, which is based on color histograms for both shot bound detection and clip selection, has to be fast.

Finally, the *University of Sheffield* generated video summaries by simply concatenating a number of continuous frames (clips), which were extracted from the middle of each detected shot. This is similar to one of the CMU baselines (CMUBASE2, see §8) except that in this case the number of frames taken from each shot was fixed at 4% of the shot length. A color histogram method was employed to detect shot boundaries by calculating the differentiation value between adjacent frames.

Figure 2: Example of Tukey-style boxplot



8. BASELINES

The team at Carnegie Mellon University created two baseline video summarization systems and submitted their outputs for evaluation along with other group submissions. They describe their algorithms as follows:

CMUBASE1 is a uniform baseline of 4% summaries in which 1 second was selected for every 25 seconds of original video. This 1 second chunk starts at 12.5 seconds into the current 25 second window and ends at 13.5 seconds. The 1-second chunks were then appended together to generate the overall video summary.

CMUBASE2 used a CMU shot boundary detector. The threshold for detecting sufficient differences between adjacent frames was lowered compared to broadcast news, to detect shot boundaries where there is dramatic motion. Hence there were more shots ('denser') than one would normally see, with 26,268 shots in the development set. From each shot a keyframe was extracted and partitioned into a 5×5 grid. In each grid cell, the mean and standard deviation of hue, saturation and value (in the HSV color space) was extracted. All keyframes for a video were used in a K-means clustering, with the number of clusters set to the number of seconds (rounded down) in the 4% summary. From each cluster, the single shot closest to the centroid was selected, and one second from the middle of this shot is used for inclusion in the summary. (Hauptmann, 2007).

Figure 3: Distribution of "ground truth included" scores

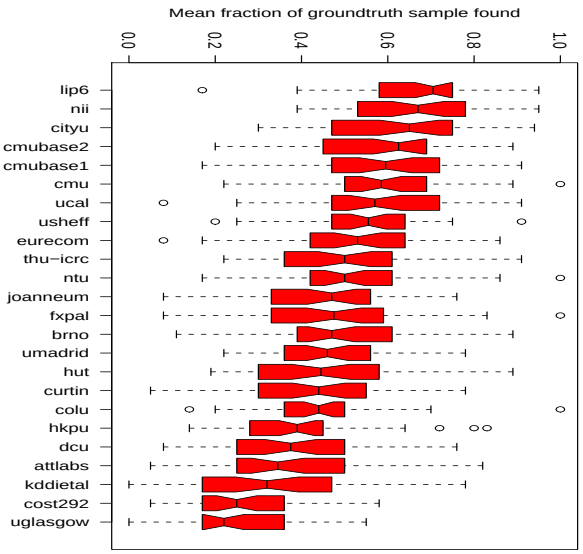
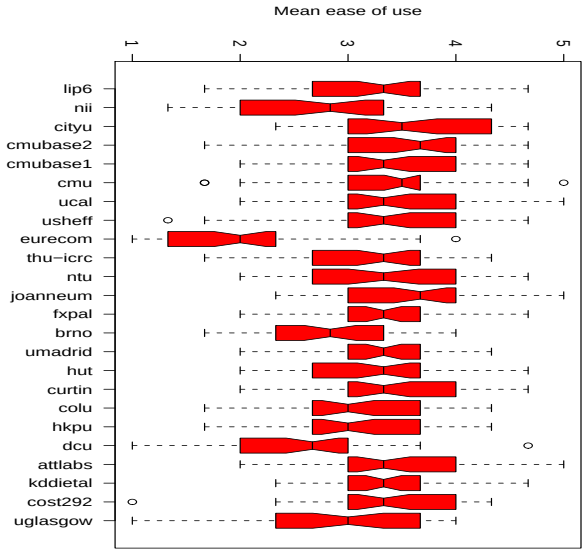


Figure 4: Distribution of "ease of use" scores



9. RESULTS

In this section we present an initial, largely graphic, exploratory analysis of the results produced by evaluating the summaries from the 22 participating groups plus the two baseline systems. 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.

9.1 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 that the median fraction of included ground truth for all summaries from each participant, ranged from 0.22 to 0.70. The two baseline systems performed in the upper quartile of all systems. A partial randomization test (Manly, 1997) found no significant difference between the baselines at the 0.05 level of significance. The same test found three systems (cityu, lip6, and nii) significantly better than both baselines but indistinguishable from each other.

9.2 Ease of understanding

The ease of understanding score was an integer ranging from 1 (worst) to 5 (best). Figure 4 shows the ease of understanding for systems, sorted by the median fraction of included ground truth. With one or two exceptions, most systems scored nearly the same, clustering around "neither strongly agree nor strongly disagree" that the summary was easy to understand and use. The randomization test found the baselines indistinguishable from each other and only the cityu system significantly better than the baselines with respect to ease of understanding and use.

9.3 Redundancy

The lack of redundancy score was an integer ranging from 1 (worst) to 5 (best). The scores for lack of redundancy (Figure 5), again sorted by median fraction of included ground truth, ranged very narrowly between 3 and 4, where 5 signifies that the assessor "strongly disagreed" that the summary contained many repeated segments. Again the randomization test found no significant difference between the baselines on this measure, but most systems were significantly better than one or both baselines, including attlabs, cityu, cost292, curtin, hkpu, hut, joanneum, kddietal, ntu, thu-icrc, umadrid. Redundancy does seem to make it more likely the ground truth items will be included and found (see Figure 6) – perhaps because it makes the assessor's job easier.

9.4 Assessment time

The median times for judging summaries against ground truth varied, as shown in Figure 7. Per-system medians range from 52.5 to 117.67 seconds. Figure 8 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.

9.5 Size of summary

Most summaries were at or under the 4% limit on duration, as can be seen in the boxplots in Figure 9 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 10.

Figure 5: Distribution of "lack of redundancy" scores

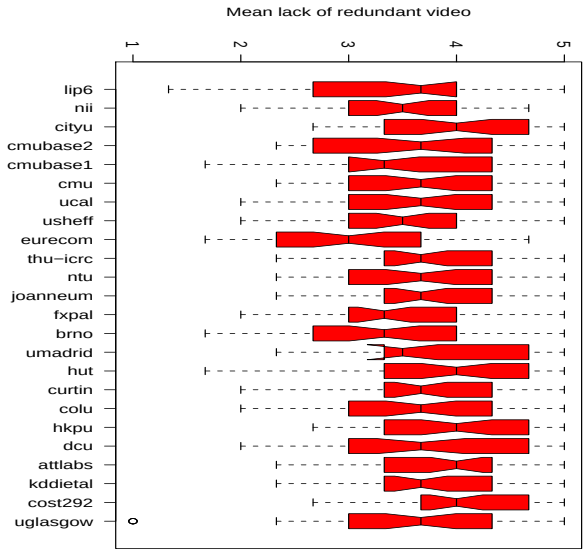


Figure 6: Lack of redundancy vs. ground truth included

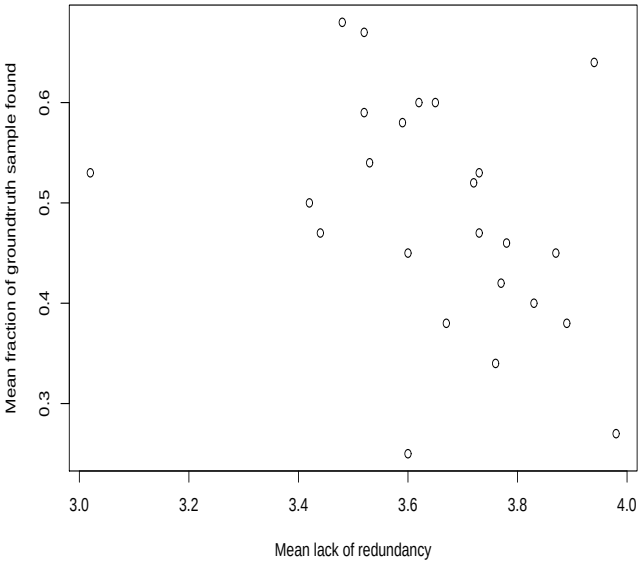


Figure 7: Distribution of total inclusion assessment time (seconds)

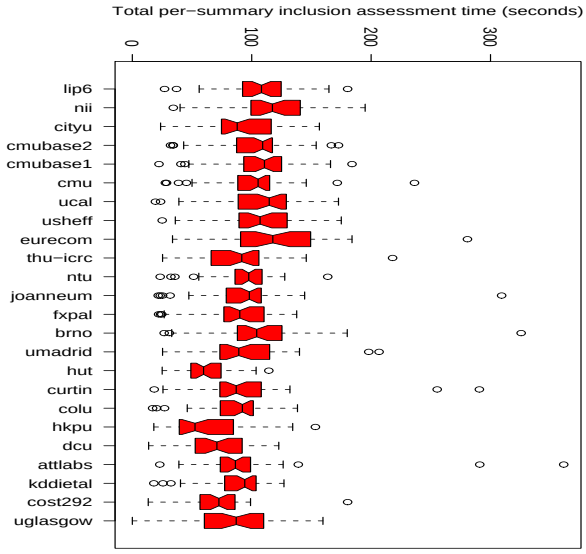
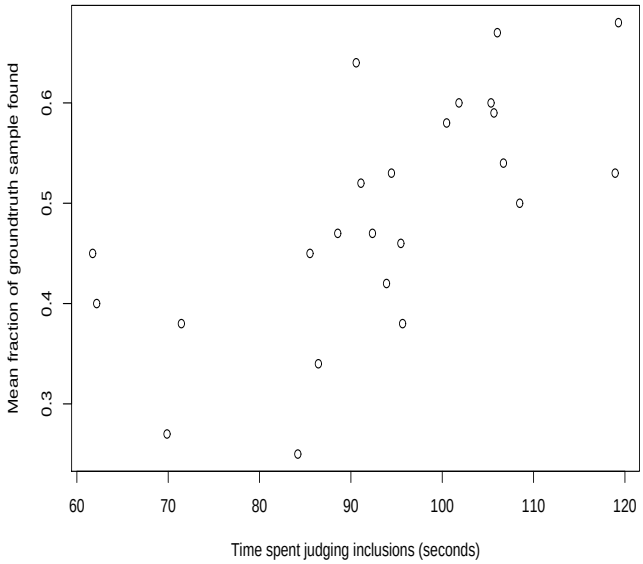


Figure 8: Time spent judging inclusions (seconds) vs. ground truth included



9.6 Summary creation time

With a couple of outliers, summary creation times all lie in approximately the same narrow range as can be seen in Figure 11. The median summary creation time was 19.23 minutes. Some systems were not optimized for speed in this initial pilot.

9.7 Summary of results

The following table presents the medians for the major measures for each system. All times are in seconds. Each score is the mean of the 3 assessments for that summary and measure.

System code	Summary duration	(Target summary size - actual)	Total time judging inclusions	Non-paused time judging	Fraction of inclusions found	Ease of use	Lack of redundancy
CMUBASE1	66.40	-2.28	110.67	66.67	0.60	3.33	3.33
CMUBASE2	64.60	-0.89	109.17	63.83	0.62	3.67	3.67
attlabs	59.15	5.49	86.33	59.50	0.35	3.33	4.00
brno	48.40	16.01	104.17	50.83	0.47	2.83	3.33
cityu	42.15	15.03	87.83	45.33	0.65	3.50	4.00
cmu	61.90	1.20	105.33	61.33	0.59	3.50	3.67
colu	53.65	10.03	92.17	53.50	0.44	3.00	3.67
cost292	47.30	16.06	72.50	49.83	0.25	3.33	4.00
curtin	60.65	1.29	87.00	61.00	0.44	3.33	3.67
dcu	40.90	8.65	70.83	42.67	0.38	2.67	3.67
eurecom	43.20	14.73	117.67	43.33	0.53	2.00	3.00
fxpal	59.50	2.49	90.00	57.33	0.47	3.33	3.33
hkpu	26.20	30.60	52.50	29.00	0.39	3.00	4.00
hut	26.10	37.22	59.67	28.67	0.44	3.33	4.00
joanneum	57.50	5.05	98.00	58.67	0.47	3.67	3.67
kddietal	63.45	0.10	94.17	64.33	0.32	3.33	3.67
lip6	60.40	3.02	108.17	60.17	0.70	3.33	3.67
nii	68.00	-2.84	117.33	72.00	0.67	2.83	3.50
ntu	60.80	2.25	97.50	62.83	0.50	3.33	3.67
thu-icrc	41.10	15.34	91.67	45.33	0.50	3.33	3.67
ucal	63.60	0.37	114.67	68.50	0.57	3.33	3.67
uglasgow	61.70	5.40	87.00	66.17	0.22	3.00	3.67
umadrid	49.65	10.05	89.17	52.17	0.46	3.33	3.50
usheff	66.98	-2.12	107.00	69.33	0.55	3.33	3.50

10. EVALUATING THE EVALUATION

The TRECVID 2007 video summarization effort was a pilot, not only for many if not most of the participating systems, but also for the evaluation framework.

The assessors seemed to grasp the task and in informal conversations indicated they believed they were able to judge the summaries. They were asked for written feedback on the assessment training, process, documentation, system, and experience as a whole. They all reacted positively to the task of evaluating summaries and to the assessment software. One commented on the difficulty of using the 5-point scales and some mentioned not being sure of what they saw when the summaries or summary segments were very short

Figure 9: Distribution of 4% duration target - summary duration (seconds)

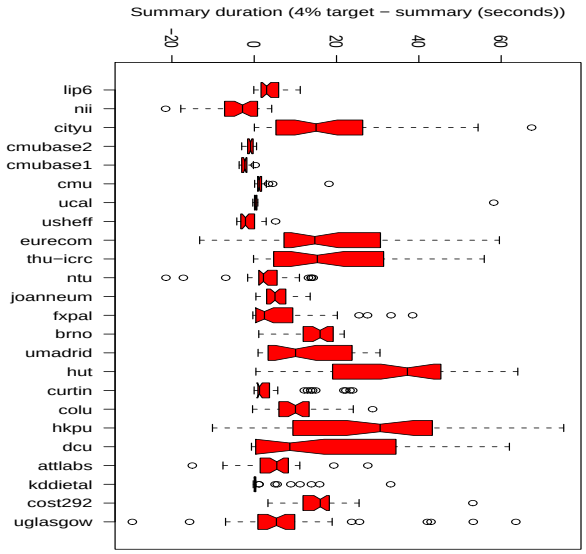


Figure 10: Excess duration (seconds over 4%) vs ground truth included

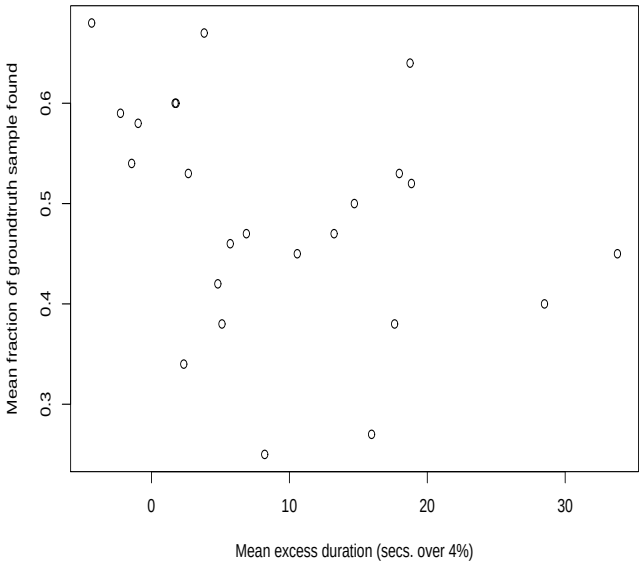
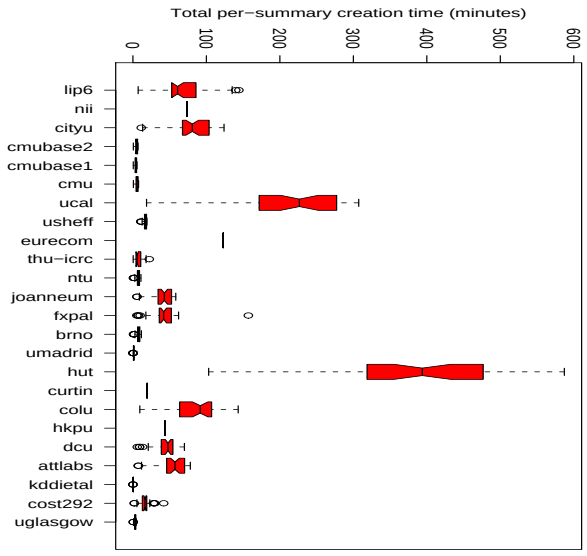


Figure 11: Distribution of system processing time (minutes)



or the videos were dark. One noted that listening to the audio was unnecessary and distracting. About the summaries, two complained that those with multiple simultaneous views (picture in picture) were difficult to follow. Comments indicate a couple of areas for improvement of the evaluation and further investigation of the results.

One assessor suggested the "ease of understanding" question may have been misinterpreted or misapplied. Scores on "ease of understanding" did not seem to correlate as expected with time spent in judging included ground truth (see Figure 12) but other confounding factors may explain that. Do the "ease of understanding" scores seem reasonable when looking at some of the actual summaries? An examination of the summary styles versus their scores on "ease of understanding" suggests assessors may just not have liked summaries that varied from showing selected segments from the full video at the original speed. Table 2 presents evidence for this possible interpretation.

The fact that there were three judgments for each summary allows an examination of assessor agreement. The initial exploration of that data is presented here. First we look at the sizes and numbers of differences in the judgments of included groundtruth, ease of understanding, and amount of repeated video. Figure 14 shows the distribution of pairwise score differences in assessors judging included ground truth. The agreement looks high. Digging deeper, comparing pairs of assessors in their binary judgments of individual included ground truth for a given summary, Figure 13 shows a mean and median agreement of 78 % compared to the 50% agreement that can be expected by chance alone;

Figures 15 and 16 depict the size and number of pairwise differences in the scores for ease of understanding and redundancy. Agreement here is less than for included ground truth. There were clearly more exact agreements (differ-

Table 2: Systems, ease of understanding, enhancements over original video

joanneum	3.67	
CMUBASE2	3.67	
cmu	3.50	
cityu	3.50	
usheff	3.33	
umadrid	3.33	
ucal	3.33	
thu-icrc	3.33	scene/cast list at beginning only
ntu	3.33	
lip6	3.33	some artifacts in the video?
kddietal	3.33	
hut	3.33	
fxpal	3.33	info at bottom of large window
curtin	3.33	
cost292	3.33	
CMUBASE1	3.33	
attlabs	3.33	
uglasgow	3.00	slide show without fade-in/out
hkpu	3.00	
colu	3.00	a few small pics at right of larger
nii	2.83	fast forward
brno	2.83	multiple small images below larger
dcu	2.67	slide show with fade-in/out
eurecom	2.00	4 windows in 1

Figure 12: Time to judge ground truth inclusions versus mean ease of understanding scores

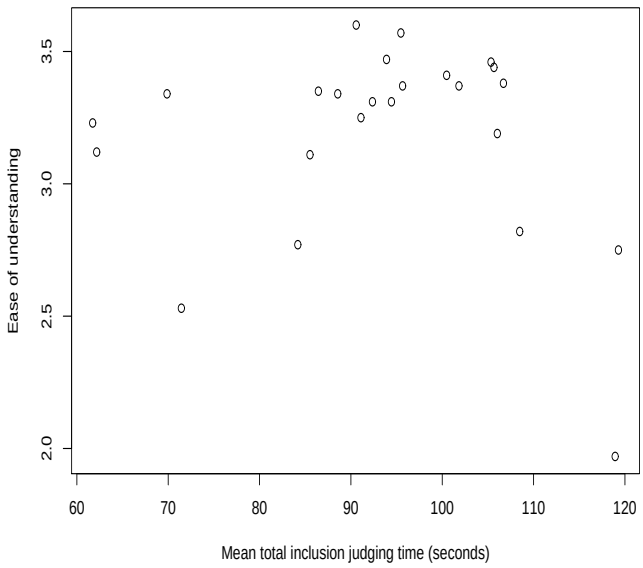


Figure 13: Distribution of pairwise agreement between assessors in binary judgments of included ground truth

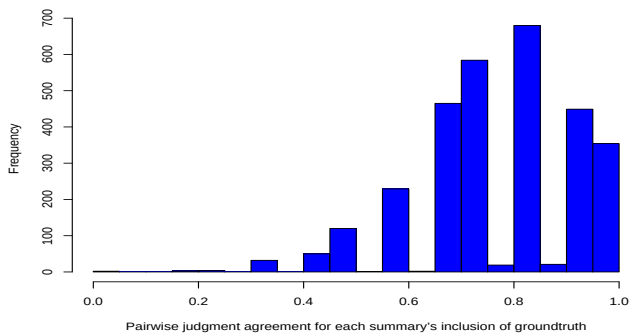
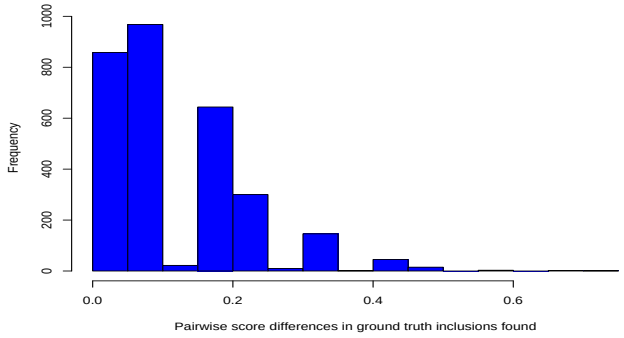


Figure 14: Pairwise score differences between assessors judging included ground truth



ences of size zero) on redundancy (about 800) than on "ease of understanding" (about 600). But the mean difference in scores for ease of understanding (1.433) is not much larger than that for redundancy (1.366). These data need further study and the two questions need to be tested with revised instructions and possible wider scales if they are to be used again.

While there exist various ways of measuring inter-assessor agreement, for practical purposes disagreement that affects significant differences in system rankings is of primary importance. To examine that question in a first pass, 7 sets of results were created - each based on withholding the judgments of a different assessor. A randomization test at the 0.05 level of significance was run on each set of results to produce a list of significantly different systems. No significantly different systems swapped positions due to changing the assessor withheld for the included ground truth or repeated video measures. For the ease of use measure, 6 pairs of significantly different systems swapped positions out of all system pairs in all 21 pairwise comparisons of the 7 sets of results. There were no changes in significantly different rankings on included ground truth, ease of understanding, or repeated video when comparing results with one assessor held out to the official results. The outcome is similar when comparing results produced by 3 random sets of single-judgment assignments to each other or to the official results. Further study of within- and between-assessor agreement is planned.

Figure 15: Pairwise score differences between assessors judging ease of understanding

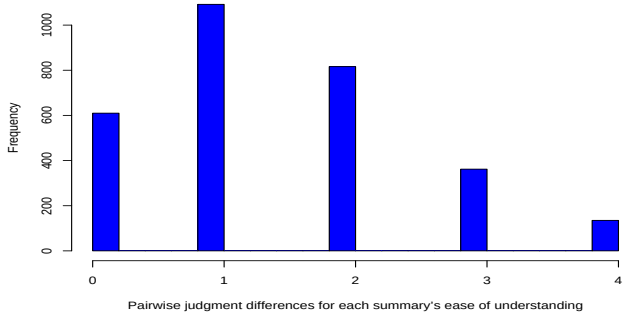
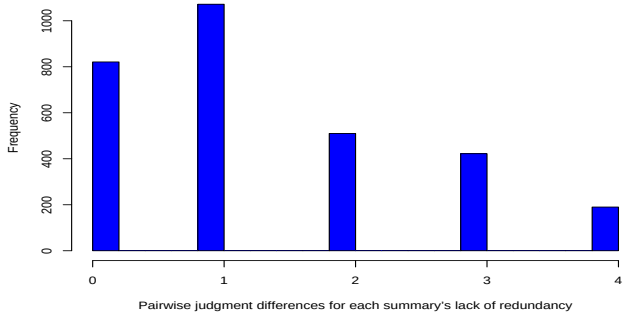


Figure 16: Pairwise score differences between assessors judging redundancy



11. CONCLUSIONS

The completion of the first large-scale multi-participant evaluation of rushes video summarization marks a new phase in the development of video summarization techniques. We now have a common dataset, ground truth and metrics which we can use to evaluate generated summaries. Caution regarding the scope conclusions is, as always, appropriate - for example because rushes of dramatic series can look quite different from other less dialogue-based rushes. But the framework, while not perfect, represents a good starting point on which we can build in the future.

There is informal evidence that results presented here correspond, in specific cases, to expectations we would have had about video summarization and we found surprisingly good agreement on the most detailed part of the evaluation, the inclusion of ground truth in the summaries. Results suggest that systems were able to do something sensible within the guidelines and perhaps that the 4% target could have been even smaller. Summarization is not a computationally fast operation as currently implemented and there is certainly scope for further improvement. Baseline systems using simple techniques were hard to beat but it can be done. Judgment of the ease of use and the amount of redundancy in summaries may be too coarse or insensitive, but it is also possible that there are few large difference among the systems. These measures do need to be refined and improved. Finally, it is comforting to note that the assessors themselves found the evaluation quite doable.

Acknowledgments

The authors would like to thank several individuals and groups for making the video summarization pilot evaluation possible. We are grateful to the BBC archives and to Richard Wright for providing the data, to NIST and DTO for sponsoring the evaluation campaign, to the assessors at NIST who judged the summaries, to Science Foundation Ireland under grant number 03/IN.3/I361 and the team at Dublin City University for creating the ground truth, to several sites for mirroring the video data to allow distribution to participating groups over the Internet, to the program committee for advice throughout and to the program committee and several others for reviewing papers and finally, to all the participating groups for taking part.

APPENDIX

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.

object (no event or camera event) e.g.,

- **"antique car"**
- **"old woman"**

object(s) + event e.g.,

- **"red hot air balloon ascending"**
- **"people talking"**

object(s) + camera event e.g.,

- **"pan across room"**
- **"zoom in on newspaper page"**

object(s) + event + camera event e.g.,

- **"zoom in on red hot air balloon ascending"**
- **"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 angleD""** 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

Christel, M. G., Smith, M. A., Taylor, C. R., & Winkler, D. B. (1998). Evolving Video Skims into Useful Multimedia Abstractions. In *Proceedings of the ACM CHI '98 Conference on Human Factors in Computing Systems* (pp. 171–178). Los Angeles, California, USA.

Ding, W., Marchionini, G., & Tse, T. (1997). Previewing Video Data: Browsing Key Frames at High Rates Using a Video Slide Show. In *Proceedings of the International Symposium on Research, Development and Practice in Digital Libraries* (pp. 151–158). Tsukuba Science City, Japan.

Ekin, A., Tekalp, A. M., & Mehrotra, R. (2003). Automatic Soccer Video Analysis and Summarization. *IEEE Transactions on Image Processing*, *12*(7), 796–807.

Ferman, A. M., & Tekalp, A. M. (2003). Two-Stage Hierarchical Video Summary Extraction to Match Low-Level User Browsing Preferences. *IEEE Transactions on Multimedia*, *5*(2), 244–256.

Hauptmann, A. (2007). *Personal communication, School of Computer Science at Carnegie Mellon University.*

He, L., Sanocki, E., Gupta, A., & Grudin, J. (1999). Auto-summarization of audio-video presentations. In *Proceedings of the 7th ACM International Conference on Multimedia* (pp. 489–498). Orlando, Florida: ACM Press New York, NY, USA.

Komlodi, A., & Marchionini, G. (1998). Key frame preview techniques for video browsing. In *Dl '98: Proceedings of the third acm conference on digital libraries* (pp. 118–125). New York, NY, USA: ACM Press.

Manly, B. F. J. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology* (2nd ed.). London, UK: Chapman & Hall.

Marchionini, G. (2006). Human Performance Measures for Video Retrieval. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval* (pp. 307–311). New York, NY, USA: ACM Press.

mplayer. (2007). *Mplayer - the Movie Player.* URL: www.mplayerhq.hu/design7/news.html.

Smeaton, A. F., & Over, P. (2003). TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. *Proc. of the International Conference on Image and Video Retrieval (CIVR).*

Smeaton, A. F., Over, P., & Kraaij, W. (2004). TRECVID: evaluating the effectiveness of information retrieval tasks on digital video. *Proceedings of the 12th Annual ACM international Conference on Multimedia*, 652–655.

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

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

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

Wright, R. (2005). *Personal communication, Technology Manager, Projects, BBC Information & Archives.*