TRECVID 2004 - An Overview

Wessel Kraaij {wessel.kraaij@tno.nl}
TNO Science & Industry
2600 AD Delft, the Netherlands

Alan F. Smeaton {asmeaton@computing.dcu.ie}
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland

Paul Over {over@nist.gov}
and Joaquim Arlandis {jarlandi@nist.gov}
Retrieval Group
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

February 23, 2005

1 Introduction

TRECVID 2004 was the fourth running of a TREC-style video retrieval evaluation, the goal of which remains to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. Over time this effort should yield a better understanding of how systems can effectively accomplish such retrieval and how one can reliably benchmark their performance. TRECVID is funded by the Advanced Research and Development Activity (ARDA) and the National Institute of Standards and Technology (NIST).

The evaluation used as test data about 70 hours of US broadcast news video in MPEG-1 format that had been collected for TDT-2 by the Linguistic Data Consortium in 1998. 33 teams from various research organizations — 7 from Asia/Australia, 17 from Europe, and 9 from the Americas — participated in one or more of four tasks: shot boundary determination, story segmentation, feature extraction, and search (manual or interactive). Results were scored by NIST using manually created truth data for shot boundary determination and story segmentation. Feature extraction and search submissions were evaluated based on partial manual judgments of the pooled submissions.

This paper is an introduction to, and an overview of, the evaluation framework (the tasks, data, and measures), the results, and the approaches taken by the participating groups[1]. For detailed information about the approaches and results, the reader should see the online proceedings on the TRECVID website (www-nlpir.nist.gov/projects/trecvid).

1.1 New in TRECVID 2004

TRECVID 2004 was the second part of a 2-year cycle using the same tasks and data sources as in 2003 - this to minimize the start-up work for continuing participants and effect of using new test data each year. There was an increase in the number of participants who completed at least one task - up to 33 from last year's 24. See table 1.

The story typing task, which was a subtask of story segmentation in 2003 was dropped for 2004, since the 2003 evaluation had shown that the task was not challenging enough. At the suggestion of the IBM team, a "fully automatic search" task was included late in the development cycle.

[1]The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

Table 1: Participants and tasks

Participants	Country	Task			
AIIA Laboratory	Greece	SB	–	–	–
Bilkent University	Turkey	–	–	–	SE
Carnegie Mellon University	US	–	–	FE	SE
Center for Research & Technology Hellas/ITI	Greece	–	–	–	SE
CLIPS-LSR-LIS	France	SB	SS	FE	SE
CWI / University of Twente	the Netherlands	–	–	–	SE
Dalle Molle Inst. for Perceptual Artificial Intelligence (IDIAP)	Switzerland	–	–	FE	–
Dublin City University	Ireland	–	–	–	SE
Eurecom	France	–	–	FE	–
Fraunhofer (Heinrich Hertz) Institute	Germany	SB	–	–	–
FX Palo Alto Laboratory	US	SB	–	–	SE
IBM Research	US	SB	SS	FE	SE
Imperial College, London	UK	SB	SS	FE	SE
Indiana University	US	–	–	–	SE
KDDI R&D Laboratories	Japan	SB	SS	–	–
Mediamill/University of Amsterdam	the Netherlands	–	–	FE	SE
National Cheng Kung University ELITE Center	Taiwan	–	SS	–	–
National Institute of Informatics	Japan	–	–	FE	–
National Taiwan University	Taiwan	–	–	–	SE
National University of Singapore	Singapore	–	–	FE	SE
RMIT University	Australia	SB	SS	–	–
SAMOVA/IRIT/UPS	France	SB	–	–	–
Tsinghua National Laboratory for Information and Technology	China	SB	–	FE	–
Univeristy of Bremen/TZI	Germany	SB	–	–	–
University of Bordeaux	France	SB	–	–	–
University of Central Florida	US	–	SS	FE	–
University of Iowa	US	SB	SS	FE	–
Queen Mary, University of London	UK	SB	–	–	SE
University of Maryland	US	SB	–	–	–
University of North Carolina at Chapel Hill	US	–	–	–	SE
University of Oulu	Finland	–	–	–	SE
University of Sao Paolo/IME	Brazil	SB	–	–	–
University Rey Juan Carlos	Spain	SB	–	–	–

Task legend. SB: Shot boundary; SS: Story segmentation; FE: Feature extraction; SE: Search

Figure 1: Location of participating groups

More effort was devoted to promoting good experimental designs for the interactive search experiments and strengthening the basis for comparison of systems. As part of this, the Dublin City University team led an effort to define and collect a common set of user demographics and satisfaction data in interactive experiments.

NIST assessors judged twice as large a fraction of the pooled shots submitted in the feature extraction task as last year (20% versus 10%).

2 Data

2.1 Video

All of the 2003 data (CNN Headline News and ABC World News Tonight from January through June of 1998 and a small amount of C-SPAN), common annotations, shared feature results, and truth data were available for system development. Approximately 70 additional hours of CNN Headline News and ABC World News Tonight from October through December of 1998, in MPEG-1, were available for system testing in the four tasks. This data was divided as follows:

A shot boundary test collection for this year's evaluation, comprising about 6 hours, was drawn from the total test collection. It included 12 videos for a total size of about 4.23 gigabytes. The characteristics of this test collection are discussed below. The shot boundary determination test data were distributed by NIST on DVDs just prior to the test period start.

The total test collection exclusive of the shot boundary test set was used for evaluating systems on the story segmentation, feature extraction, and search tasks. This part of the collection was distributed on hard disk drives by the Linguistic Data Consortium (LDC).

2.2 Common shot reference, keyframes, text from speech

The entire story/feature/search collection was automatically divided into shots by Georges Quénot at CLIPS-IMAG. These shots served as the predefined units of evaluation for the feature extraction and search tasks. The story/feature/search test collection contained 128 files/videos and 33,367 reference shots - as compared to 113 files and 35,067 reference shots in the 2003 test data set.

The CLIPS-IMAG group also extracted a keyframe for each reference shot and these were made available to participating groups along with automatic speech recognition (ASR) system output provided by Jean-Luc Gauvain at LIMSI (Gauvain, Lamel, & Adda, 2002).

2.3 Common feature annotation

In 2003 Ching-Yung Lin of IBM headed up a collaborative effort (Lin, Tseng, & Smith, 2003) in which 23 groups used IBM software to manually annotate the development collection of over 60 hours of video content with respect to 133 semantic labels. This data was then available for subsequent use such as training in feature extraction and search. In order to help isolate system development as a factor in system performance each feature extraction task submission, search task submission, or donation of extracted features declared its type:

A - system trained only on common development collection and the common annotation of it

B - system trained only on common development collection but not on (just) common annotation of it

C - system is not of type A or B

3 Shot boundary detection

Movies on film stock are composed of a series of still pictures (frames) which, when projected together rapidly, the human brain smears together so we get the illusion of motion or change. Digital video is also organized into frames - usually 25 or 30 per second. Above the frame, the next largest unit of video both syntactically and semantically is called the shot. A half hour of video, in a TV program for example, can contain several hundred shots. A shot was originally the film produced during a single run of a camera from the time it was turned on until it was turned off or a subsequence thereof as selected by a film editor. The new possibilities offered by digital video have blurred this definition somewhat, but shots, as perceived by a human, remain a basic unit of video, useful in a variety of ways.

Work on algorithms for automatically recognizing and characterizing shot boundaries has been going on for some time with good results for many sorts of data and especially for abrupt transitions between shots. Software has been developed and evaluations of various methods against the same test collection have been published e.g., using 33 minutes total from five feature films (Aigrain & Joly, 1994); 3.8 hours total from television entertainment programming, news, feature movies, commercials, and miscellaneous (Boreczky & Rowe, 1996); 21 minutes total from a variety of action, animation, comedy, commercial, drama, news, and sports video drawn from the Internet (Ford, 1999); an 8-hour collection of mixed TV broadcasts from an Irish TV station recorded in June, 1998 (Browne et al., 2000).

An open evaluation of shot boundary determination systems was designed by the OT10.3 Thematic Operation (Evaluation and Comparison of Video Shot Segmentation Methods) of the GT10 Working

Group (Multimedia Indexing) of the ISIS Coordinated Research Project in 1999 using 2.9 hours total from eight television news, advertising, and series videos (Ruiloba, Joly, Marchand-Maillet, & Quénot, 1999).

The shot boundary task is included in TRECVID both as an introductory problem, the output of which is needed for most higher-level tasks such as searching, and also because it is a difficult problem with which to achieve very high accuracy. Groups can participate for their first time in TRECVID on this task, develop their infrastructure, and move on to more complicated tasks the next year, or they can take on the more complicated tasks in their first year, as some do. Information on the effectiveness of particular shot boundary detection systems is useful in selecting donated segmentations used for scoring other tasks.

The task was to find each shot boundary in the test collection and identify it as an abrupt (cut) or gradual transition, where any transition, which is not abrupt is considered gradual.

3.1 Data

The test videos contained 618,409 total frames (4% more than last year) and 4,806 shot transitions (29% more than last year).

The reference data was created by a student at NIST whose task was to identify all transitions and assign each to one of the following categories: to cuts, which are abrupt, or to one of the other categories containing various kinds of gradual transitions.

cut - no transition, i.e., last frame of one shot followed immediately by the first frame of the next shot, with no fade or other transition;

dissolve - shot transition takes place as the first shot fades out *while* the second shot fades in

fadeout/in - shot transition takes place as the first shot fades out and *then* the second fades in

other - everything not in the previous categories e.g., diagonal wipes.

Software was developed and used to sanity check the manual results for consistency and some corrections were made. Borderline cases were discussed before the judgment was recorded.

The freely available software tool VirtualDub was used to view the videos and frame numbers. The distribution of transition types was as follows:

- 2,774 — hard cuts (57.7%, down from 70.7% in 2003)
- 1,525 — dissolves (31.7%, up from 20.2%)
- 230 — fades to black and back (4.8%, up from 3.1%)
- 276 — other (5.7%, down from 5.9%)

The percentage of gradual transitions increased noticeably. At this point we have not determined why video from the second half of 1998 should be this different from video from the first half of the same year. Gradual transitions are generally harder to recognize than abrupt ones.

3.2 Evaluation and measures

Participating groups in this task were allowed up to 10 submissions and these were compared automatically to the shot boundary reference data. Each group determined the different parameter settings for each run they submitted. Seventeen groups submitted runs.

Detection performance for cuts and for gradual transitions was measured by precision and recall where the detection criteria required only a single frame overlap between the submitted transitions and the reference transition. This was to make the detection independent of the accuracy of the detected boundaries. For the purposes of detection, we considered a submitted abrupt transition to include the last pre-transition and first post-transition frames so that it has an effective length of two frames (rather than zero).

Analysis of performance individually for the many sorts of gradual transitions was left to the participants since the motivation for this varies greatly by application and system.

Gradual transitions could only match gradual transitions and cuts match only cuts, except in the case of very short gradual transitions (5 frames or less), which, whether in the reference set or in a submission, were treated as cuts. We also expanded each abrupt reference transition by 5 frames in each direction before matching against submitted transitions to accommodate differences in frame numbering by different decoders.

Accuracy for reference gradual transitions successfully detected was measured using the one-to-one matching list output by the detection evaluation. The accuracy measures were frame-based precision and recall. Note that a system could be very good in detection and have poor accuracy, or it might miss a lot of transitions but still be very accurate on the ones it finds.

Measuring complexity

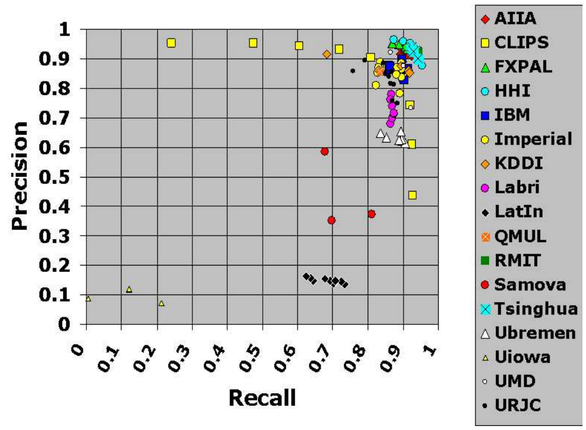
After the guidelines were complete, a requirement for complexity data was added at the request of participants. Because not all systems were designed with speed in mind, there was some difficulty defining a set of usable measures that every system could provide. Nevertheless, most groups were able to supply time spent in decoding and in segmentation. This information makes an interesting addition to the existing effectiveness measures.

3.3 Approaches in brief

The runs from *CLIPS* used the same approach as in previous years, detecting cuts by image comparisons after motion compensation and gradual transitions (GTs) by comparing norms of first and second temporal derivatives of images. The *FX Palo Alto Laboratory* combined pairwise similarities between images in the locality and supervised classification. They calculated multiple pairwise similarities in 2003 and used a new classification scheme for 2004. The *Fraunhofer / Heinrich Hertz Institute* detected cuts using pixel and edge differences with adaptive thresholding, photo and slo-mo detection. For gradual transitions, they used edge energy statistics, pixel and histogram differences, and a wipe detector. *IBM Research* fielded the same CueVideo System as used previously but the test data encoding (different than that of the development data) caused some problems for their decoder. *Imperal College London* used distances between color histograms of frames over a range of timescales. Their system was largely unchanged from previous years.

KDDI Labs extended their 2003 approach by adding edge features from DC image, color layout, and support vector machine (SVM) learning. They noted that including edge features added a lot of improvement for gradual transitions. *RMIT University* submitted 10 runs using ASR as well as 10 regular runs. A new post-processing step was added to their 2003 GT detection scheme. Post-processing improved detection of gradual transitions. ASR was used in post-processing to remove GTs that coincided with spoken words but was ineffective as it dramatically reduced recall while precision remains. *SAMOVA-IRIT Toulouse's* approach was based on detection and tracking of changes in illumination. Frame size was reduced to 44x30 pixels. *Tsinghua University* used a fade in/out detector based on detecting monochrome frames. Their cut detector used 2nd order derivatives of color histogram and pixelwise comparisons, flash detection, and GT elimination.

Figure 2: Precision and recall for cuts



Their GT detector used the same features as their cut detector plus motion vectors. The *University of Bordeaux - LaBRI* estimated robust global camera motion, determined P-frame peaks to compute motion and frames statistics, then measured similarity between compensated adjacent I-frames.

The *University of Bremen / TZI* used RGB histogram differences within a 5-frame window, edge change ratio between consecutive frames and in a 10-frame window, followed by block-based motion analysis. GT detection used a similar approach to cut detection but with a finite state automaton. The *University of Iowa* employed a combination of color histogram similarity, aggregate color distance for equivalent pixel pairs, and edge distances, but there were technical problems in the submission. The *University of Maryland* didn't use the usual color histogram or edge differences but rather a 2max ratio. For dissolves, they used the skipping image distance, a function of similarity between "distant" frames. The *University Rey Juan Carlos* looked at color histogram differences and different bin sizes. Redistribution of boundary values yielded multi-resolution histograms.

Details from AIIA Laboratory, Queen Mary University of London, and the University of Sao Paolo were not available for this overview.

3.4 Results discussion

As illustrated in Figure 2 and Figure 3, performance on gradual transitions lags, as expected, behind that on abrupt transitions, where for some uses the problem may be considered a solved one. While progress in detection of gradual transitions may be possible, it is not clear what user/application would require such

Figure 3: Precision and recall for gradual transitions

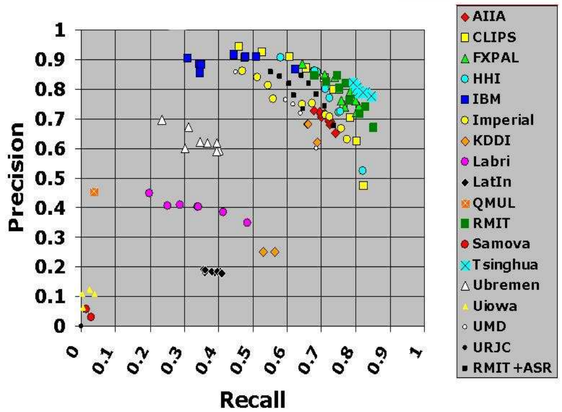
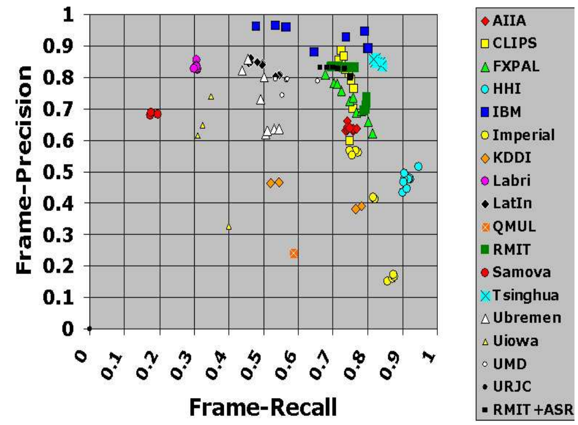


Figure 4: Frame-precision and frame-recall for gradual transitions



improvement.

The collection of information on processing time was incomplete because it was included late and some systems were not designed to provide such timings. It should also be noted that some systems may not have been designed for speed. Where available, this information did illuminate systems from a new angle - one that may be critical to some applications but not others.

We observed that there was a large range in the computational cost associated with performing the task which seemed to average around real-time, and going up to as much as three times real-time. One can also see that performing among the top systems in terms of precision and recall is compatible with top speed. This is demonstrated, e.g., by the runs submitted by the Fraunhofer Institute or KDDI, whose systems processed video in about 5% and 3% of real time respectively.

There continues to be strong interest in the shot boundary detection task and even after 4 annual iterations of the cycle novel approaches continue to emerge.

4 Story segmentation

A different way to decompose digital video and in particular news shows is to segment at the story level. News shows consist of a series of news items and publicity items. The story segmentation task was defined as follows: given the story boundary test collection, identify the story boundaries with their location (time) in the given video clip(s).

The definition of the story segmentation task is based on manual story boundary annotations made by LDC for the TDT-2 project and thus LDC's definition of a story was used in the task. A news story was defined as a segment of a news broadcast with a coherent news focus which contains at least two independent, declarative clauses. Other coherent non-news segments were labeled as "miscellaneous", merged together when adjacent, and annotated as one single story.

Story boundaries do not necessarily coincide with shot boundaries as an anchor person can present several stories during one shot. Stories often span multiple shots, e.g., when the anchor introduces a reporter at a different location.

Unlike TRECVID 2003, the 2004 edition of TRECVID did not include the story classification subtask. Results for the story classification subtask from 2003 were very good and the general conclusion was that the task was too easy. A more difficult task (a refined classification scheme including "sports" , "finance" , "health" , "politics" etc.) as proposed by some participants was not considered as a suitable task for TRECVID 2004, since classification would be dominated by textual features and ground truth for such a task was not available.

The TRECVID story segmentation task differs from the TDT-2 story segmentation task in a number of important ways:

- TRECVID uses a subset of TDT2 dataset and only uses video sources.
- The video stream is available to enhance story segmentation.
- The task is modeled as a retrospective action, so it is allowed to use global data.

With the TRECVID 2003/2004 story segmentation task, the goal was to show how video information can enhance or completely replace existing story segmentation algorithms based on text.

In order to concentrate on this goal there were several required runs from participants in this task:

- Video + Audio (no ASR/CC)
- Video + Audio + LIMSI ASR
- LIMSI ASR (no Video + Audio)

Additional optional runs using other ASR and/or closed-captions-based transcripts were also allowed to be submitted.

4.1 Data

The story test collection used for evaluation contained 3,105 story boundaries from 118 videos. Ten videos from the test set were not evaluated because the TDT truth data (based on timing in an analogue version of the video) could not be automatically aligned with the ASR from the MPEG-1. The number of stories found per video varied between a minimum of 14 and a maximum of 42.

4.2 Evaluation

Each participating group could submit up to 10 runs. In fact, eight groups submitted a total of 50 runs.

Since story boundaries are rather abrupt changes of focus, story boundary evaluation was modeled on the evaluation of shot boundaries (the cuts, not the gradual boundaries). A story boundary was expressed as a time offset with respect to the start of the video file in seconds, accurate to nearest hundredth of a second. Each reference boundary was expanded with a fuzziness factor of five seconds in each direction, resulting in an evaluation interval of 10 seconds. A reference boundary was detected when one or more computed story boundaries lay within its evaluation interval. If a computed boundary did not fall in the evaluation interval of a reference boundary, it was considered a false alarm.

4.3 Measures

Performance on the story segmentation task was measured in terms of precision and recall. Story boundary recall was defined as the number of reference boundaries detected divided by the total number of reference boundaries. Story boundary precision was defined as the (total number of submitted boundaries minus the total amount of false alarms) divided by total number of submitted boundaries. In addition, the F-measure ($\beta = 1$)was used to compare performance across conditions and across systems.

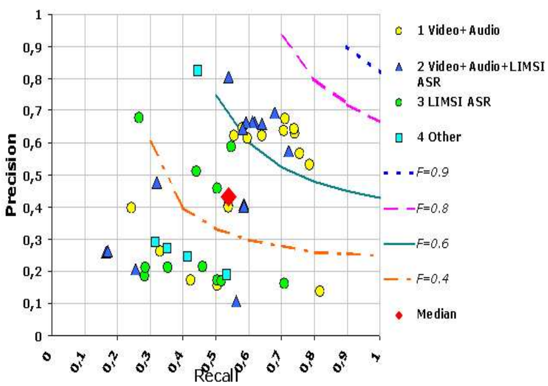
4.4 Approaches in brief

The *CLIPS-LSR-LIS* team used an approach which logically combined single feature detectors. Candidates included shot boundaries plus long pauses. They used audio change based on peak detection on BIC curve, speaker segmentation (LIMSI), example-based jingle detection, and cue phrases in the ASR. It appears that audio features boosted precision of SB baseline at the cost of some recall and adding ASR improved both recall and precision. *IBM Research* called their approach visual clue cluster construction (based on the information bottleneck principle). They used features from text, video, and rich prosody. *Imperial College London* Observed that an anchor shot starts a new story and finishes the previous story. Based on ASR, they merged segments when similarity exceeded threshold. Using a window size of 5 tokens, they calculated a score determined by number of similar words, weighted by type and distance. Their system included anchor detection using K-NN classifier trained on 2003 data.

KDDI R&D Laboratories combined features using SVM, did section-specialized segmentation (top stories, headline sports, etc.) and anchor shot segmentation using pauses. Their system made use of a shot segmenter based on shot duration and shot density, average RMS, avg RMS first n frames, freq of audio class (silence, speech, music, noise), motion (horizontal, vertical, total, intensity), as well as color layout and distance of various frames. *RMIT University* detected story boundaries in terms of anchor shots or speech pauses fro condition 1. For condition 2 they segmented the ASR using shot boundaries. Story boundaries were then defined by detecting minima of a weighted vocabulary overlap score in a window consisting of several shots. In condition 3 they proceeded as for condition 2 but candidate boundaries were based on speech pauses only. The *University of Central Florida* system used a news show grammar. It gave special treatment to weather and sports stories (no anchors) and merged adjacent similar stories (to reduce false alarm rate). For condition 1 the *University of Iowa* submitted shot boundaries as story boundaries. For condition 3 they used an extended set of trigger phrases.

Details about the approaches taken by the National Cheng Kung University ELITE Center were not avail-

Figure 5: Precision and recall by condition



able for this overview.

4.5 Results discussion

Figure 5 shows the ASR-only runs performs mostly worse than the other conditions, which use video and audio. By averaging, systems from condition 2 were more conservative than condition 1 given their lower recall and higher precision. Figure 6 allows one to compare conditions and participants. For most systems the conditions allowing the use of audio and video yield better results. This tendency is clearer in Figure 7, which displays the best run for each condition and team using the F-measure to combine precision and recall.

An error analysis including boundary categorization was made over a selection of the 6 most competitive runs in each condition. False positives were found to be mostly unique regardless of their condition. Categorization of the most popular missing boundaries turned out to be boundaries linking a regular news and a miscelaneous segment and were specially difficult for condition 1, while brief segments (sports, headlines, or money) were difficult for conditions using ASR (2,3). Boundaries between regular news were mostly harder for conditions using ASR than for condition 1.

4.6 Comparability with TDT-2 results

Results of the TRECVID 2004 story segmentation task, as in TRECVID 2003, cannot be directly compared to TDT-2 results because the evaluation datasets differ and different evaluation measures are used. TRECVID 2003 participants showed a

Figure 6: Precision and recall by condition and system

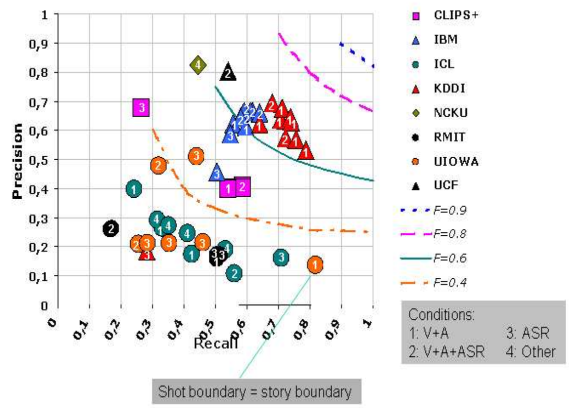


Figure 7: F-Measure by condition

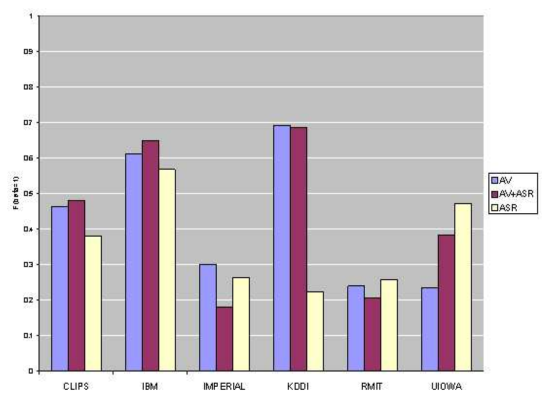
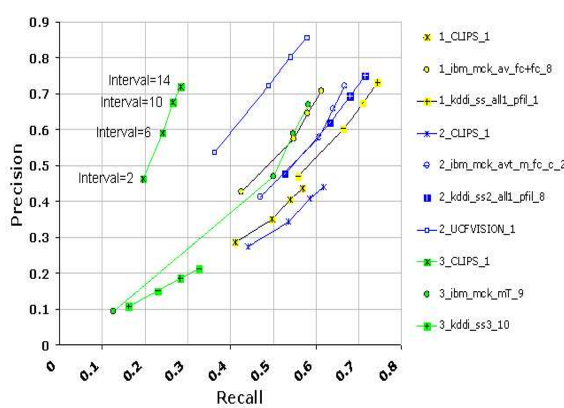


Figure 8: Effect of evaluation interval



preference for a precision/recall-oriented evaluation, whereas TDT used (and is still using) normalized detection cost. Finally, TDT was modeled as an online task, whereas TRECVID examines story segmentation in an archival setting, permitting the use of global information. However, the TRECVID story segmentation task provides an interesting testbed for cross-resource experiments. In principle, a TDT system can be used to produce an ASR+CC or ASR+CC+Audio run as demonstrated by IBM during TRECVID 2003.

4.7 Issues

There are several issues which remain outstanding with regard to this task and these include the relatively small size of the test collection used in TRECVID compared to that used in TDT. There is not a lot we can do about this since we are constrained by the availability of news data in video format which has story boundary ground truth available to us.

The procedure to align ASR transcripts with the manual story boundaries was automatic in TRECVID 2004, unlike TRECVID 2003 when it was manual. Each video offset used for alignment was computed as an average of a number of candidate values. Videos with an offset having a standard deviation above 1 were rejected from evaluation. The average of the standard deviations was 0.2032 seconds.

The evaluation interval of 10 seconds was chosen during the preparation of TRECVID 2003. This is a smaller interval than used at TDT (TDT is using 15 seconds) but made deliberately large in order to make the evaluation insensitive to the somewhat peculiar definition of TDT2 annotation standards (which have become more intuitive in later TDT corpora). This year additional results (see Figure 8) were generated for smaller and larger evaluation intervals to get an idea how precisely story boundary determination can be done. Too small values of the evaluation interval are not meaningful, since the ground truth ASR file was aligned automatically to the digital video files. From this point of view the evaluation interval should be well beyond twice the standard deviation of the estimated offset.

5 Feature extraction

A potentially important asset to help video search/navigation is the ability to automatically identify the occurrence of various semantic features such as "Indoor/Outdoor","People", "Speech" etc., which occur frequently in video information. The ability to detect features is an interesting challenge by itself but it would take on added importance if it could serve as an extensible basis for query formation and search. The high-level feature extraction task was first tried in TRECVID in 2002 and many of the issues which that threw up were tackled and overcome in TRECVID 2003. The feature extraction task has the following objectives:

- to continue work on a benchmark for evaluating the effectiveness of detection methods for various semantic concepts
- to allow exchange of feature detection output for use in the TRECVID search test set prior to the search task results submission date, so that a greater number of participants could explore innovative ways of leveraging those detectors in answering the search task queries in their own systems.

The feature extraction task was as follows. Given a standard set of shot boundaries for the feature extraction test collection and a list of feature definitions, participants were asked to return for each feature that they chose, at most the top 2,000 video shots from the standard set, ranked according to the highest possibility of detecting the presence of the feature. The presence of each feature was assumed to be binary, i.e., it was either present or absent in the given standard video shot. If the feature was true for some frame (sequence) within the shot, then it was true for the shot. This is a simplification adopted for the benefits it afforded in pooling of results and approximating the basis for calculating recall.

The feature set was taken largely from those in the common annotation from TRECVID 2003. It was modified in on-line discussions by track participants. The number of features to be detected was kept small (10) so as to be manageable in this iteration of TRECVID and the features were ones for which more than a few groups could create detectors. Another consideration was whether the features could, in theory at least, be used in executing searches on the video data as part of the search task, though the topics did not exist at the time the features were defined. Finally, feature definitions were to be in terms a human judge could understand. Some participating groups made their feature detection output available to participants in the search task which really helped and contributed to the collaborative nature of TRECVID.

Table 2: Feature pooling and judging statistics

Feature number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number true	% judged that were true
28	106000	24795	23.4	300	5971	24.1	441	7.4
29	91892	21161	23.0	175	3153	14.9	19	0.6
30	134764	21183	15.7	300	5215	24.6	409	7.8
31	96000	25509	26.6	175	3557	13.9	43	1.2
32	117183	26226	22.4	250	6175	23.5	374	6.1
33	116612	23790	20.4	175	3175	13.3	103	3.2
34	99999	22044	22.0	200	3442	15.6	62	1.8
35	98000	23554	24.0	300	5614	23.8	1695	30.2
36	96000	24598	25.6	275	6256	25.4	292	4.7
37	96000	21854	22.8	300	5312	24.3	938	17.7

The features to be detected were defined (briefly) as follows for the system developers and for the NIST assessors. This year features are numbered 28-37: [28] Boat/ship, [29] Madeleine Albright, [30] Bill Clinton, [31] Train, [32] Beach, [33] Basket scored, [34] Airplane takeoff, [35] People walking/running, [36] Physical violence, and [37] road. Three of them were the same as 2003 (29, 36, and 37) and two were similar but more restrictive (34 was just "Aircraft" and 35 was "more than two people"). The full definitions are listed in the guidelines on the TRECVID website.

5.1 Data

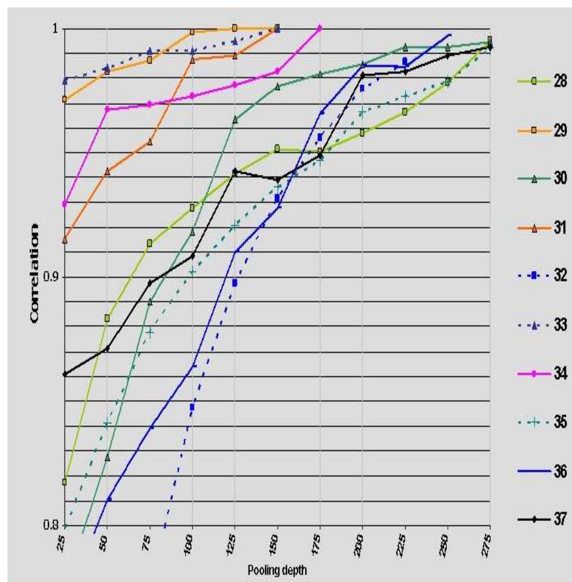
As mentioned above, the feature test collection contained 128 files/videos and 33,367 reference shots. Testing feature extraction and search on the same data offered the opportunity to assess the quality of features being used in search.

5.2 Evaluation

Each group was allowed to submit up to 10 runs. In fact 12 groups submitted a total of 83 runs.

Pooling was carried out differently than in 2003. All submissions were divided into strata of depth 25. So, for example, stratum A contained result set items 1-25 (those most likely to be true), stratum B items 26-50, etc. A subpool for each stratum was formed from the unique items from that stratum in all submissions and then randomized. To even out further the rate at which assessors could be expected to find true shots, the first several subpools were re-merged, re-randomized, and re-divided into subpools. Assessors were presented with the subpools in "alphabetical" order until they had judged the redivided set and then ran out of time or stopped finding true shots.

Figure 10: Correlations (Kendall's tau) between runs scores (avgP) based on full judgments vs. subsets



At least the top 4 sub-pools were judged completely for each feature. Beyond this, in some cases, the last subpool assessed may not have been completely judged. The maximum result set depth judged and pooling and judging information for each feature is listed in Table 2. Figure 10 shows that scoring based on somewhat shallower pools results in relatively small number of changes in pairwise ranking among the runs for a given feature.

After the evaluation, a study of the population of false positive shots found was made. We focused on

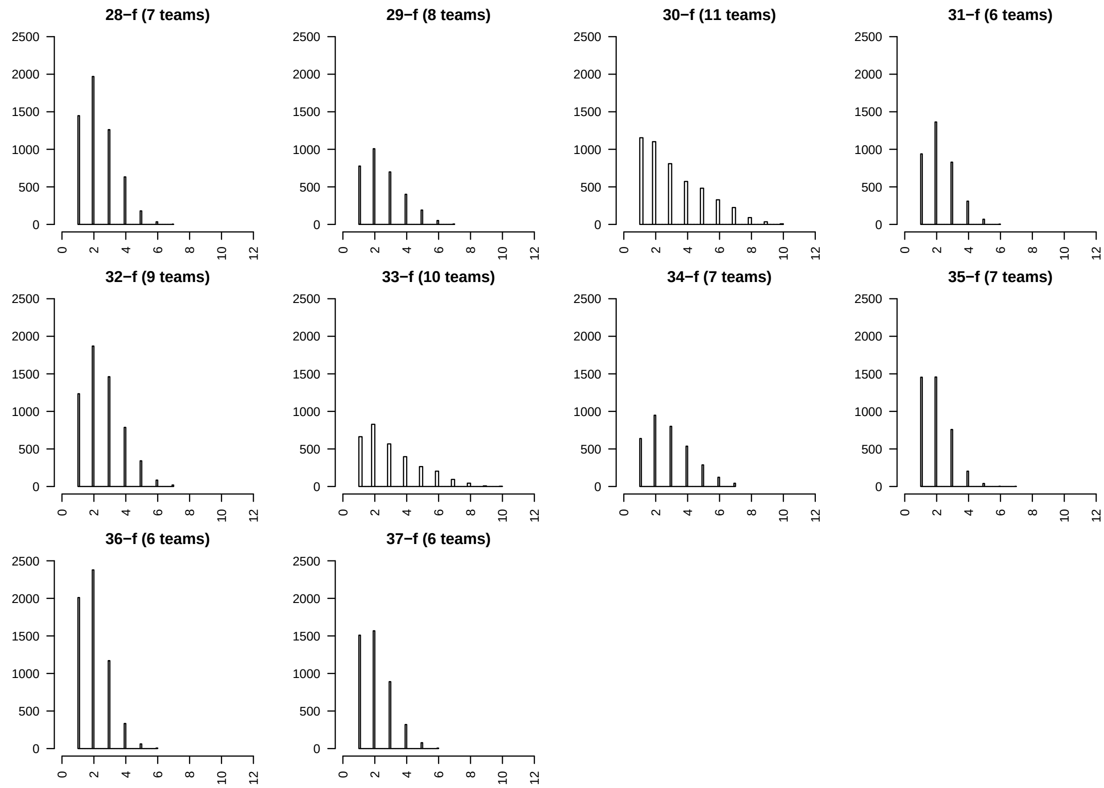


Figure 9: Number of unique false positive shots (Y axis) submitted by a given number of teams (X axis).

false positive coincident between most of the groups, trying to find out reasons for that. Figure 9 shows the number of false positive coincident between a number of systems from different groups. For some of the features (30, 33, and 34), shots with higher number of coincidences were selected and reviewed. For these shots, the most frequent reasons for misclassification were: similar but no matching features, audio referencing the feature but no image, and frozen images matching features.

5.3 Measures

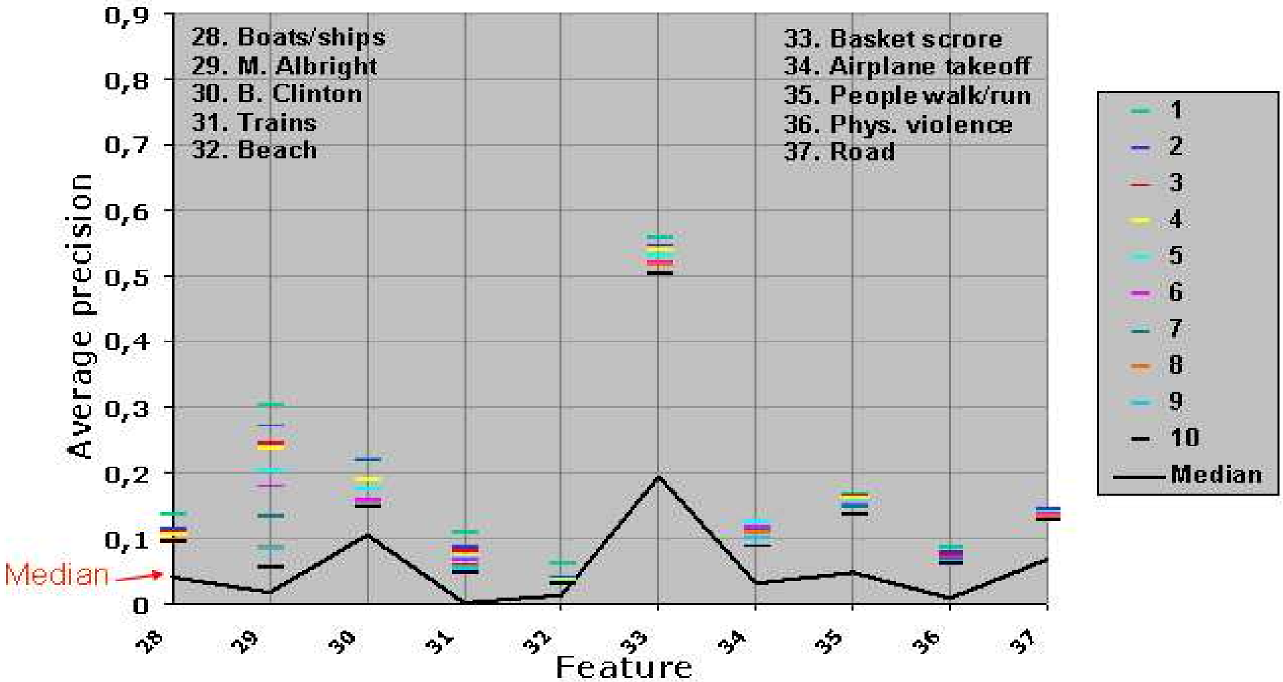
The trec_eval software, a tool used in the main TREC activity since it started in 1991, was used to calculate recall, precision, average precision, etc., for each result. In experimental terms the features represent fixed rather than random factors, i.e., we were interested at this point in each feature rather than in the set of features as a random sample of some population of features. For this reason and because different groups worked on very different numbers of features, we did not aggregate measures at the run-level in the results pages at the back of the notebook. Comparison of systems should thus be "within feature". Note, that if the total number of shots found for which a feature was true (across all submissions) exceeded the maximum result size (2,000), average precision was calculated by dividing the summed precisions by 2,000 rather than by the the total number of true shots.

5.4 Approaches in brief

CLIPS-LSR-LIS used ASR, audio plus video, and their combination. They used fusion by linear combination. Their system employed lexical context, color histograms, texture, and face detection/recognition. *Carnegie Mellon University* made uni-modal features (e.g., color histogram, texture, edge, fast Fourier transforms,mel-frequency cepstral coefficients, kinetic energy, optical flow, face detection, VOCR detection) for the test and training data available to all TRECVID participants. CMU used SVMs to combine the uni-model features into multimodal ones, which in turn were used to create extractors for the TRECVID test features.

Eurecom tried 3 fusion techniques (SVM, KNN, GA). Their system employed visual features (color histograms, LSI), text (2000 word vocabulary fre-

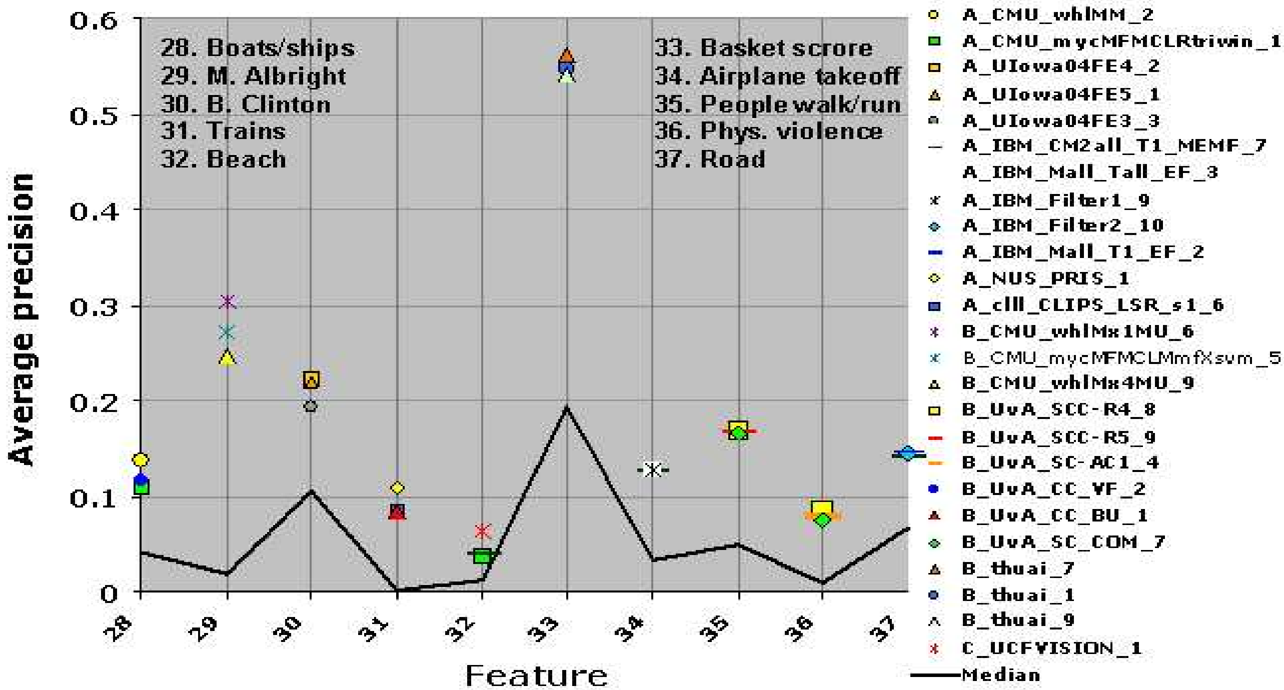
Figure 11: Average precision for top 10 runs



quency vector), and motion (camera motion & motion histogram extracted from MPEG motion vectors). The *IBM Research* team used cross feature ensemble learning: feature extraction, SVM and MaxEnt classification, ensemble fusion or MaxEnt fusion, removal of anchor shots. They used visual features (correlograms, histograms, edges, color moments, wavelet texture, co-occurrence texture) and textual ones. The *Imperial College London* system used a nearest neighbor approach. Features included global HSV, thumbnails, convolution, word "basketball", and linear combination of inverse normalized distances. The *University of Amsterdam / MediaMill* project developed a generic approach for all features — the "Semantic Value Chain". A lexicon of 32 semantic concepts was trained on 2003 data and common annotation. Video analysis was seen as inverted authoring. They employed SVM & a parallel architecture. Features included words associated with concept, visual colorspace corrected for intensity and object size variation), layout (shot length, overlayed text, silence, voice over), and content (faces, face location, cars, object motion, frequent speaker, overlayed text length, named entity), and capture (camera distance, camera work, camera motion).

The *National University of Singapore* used visual auto-concept annotation, fusion of text and specialized detectors; 4X4 blocks, block clustering, SVM training. Their approach was generic except for the use of a face detector for person X. Features considered included colorhistogram, texture, edge histogram, informative keywords selected using IDF and MI using Google. The *National Institute of Informatics* team worked on face detection , face alignment and face recognition using positive training examples MA/BC and negative training examples from 1210 individuals in FERET database for training SVM classifier, and a Gaussian radial basis function kernel. The *Tsinghua National Laboratory for Information and Technology* used an SVM classifier for several features, then voting, interpolation with text based score, and finally filtering of known false positives. Features considered were visual and text features, clusters of positive shots, and position of shot in show. The *University of Central Florida* developed specific detectors for 5 features using fusion by means of SVM or Adaboost features based on text, face, colorhistogram, correlogram, skin, motion, and sports shot. The *University of Iowa* combined general low level features and fusion by sum or prod-

Figure 12: Average precision for top 3 runs by feature



uct. Features used were color histograms, edge histograms, feature-specific cropping; text (manually selected trigger words), and anchor.

No details about the approaches taken by IDIAP were available for this overview.

5.5 Results discussion

Figure 11 summarizes the results by feature for the top 10 runs by feature. The line at the bottom of the graph indicates the location of the median performance. Results vary greatly in their mean and dispersion from feature to feature. "Madeleine Albright" in particular shows a very wide variability in detection results. "Basketball score" stands out with its high scores. Why that is we are not sure.

Figure 12 shows the top 3 runs by name for each feature on a truncated scale. It is necessary to draw on runs from 8 groups to account for the top 3 for each feature. Figures 13 and 14 show to what extent a given run or a group's runs contributed true shots not contributed by any other run or group, respectively.

The features for 2004 were more difficult than those for 2003. The average rate of true shots per 1000 test shots was 13 in 2004 versus 21 in 2003. The maximum rate of true shots per 1000 test shots was 51 in 2004 versus 69 in 2003. Hard features are sometimes easy (basketball hoop). Most features were rare enough for reliable pooling. The average of the top 10 runs for "Madaleine Albright" dropped but the corresponding results for "road" and "phyiscal violence" were unchanged from 2003. Many groups chose training type A, increasing comparability.

5.6 Issues

The choice of the features and the characteristics of the test collection can cause problems for the evaluation framework. One feature (35. People walking/running) turned out to be very frequent in its occurrence in the collection. This affects the pooling and judging in ways we have yet to measure.

The repetition of video material in commercials and in repeated news segments can increase the frequency of true shots for a feature and reduce the usefulness of the recall measure. Finally, the issue of interaction between the feature extraction and the search tasks still needs to be examined so that search can benefit more from feature extraction.

Finally, the encoder used by LDC to encode the

Figure 13: True shots contributed uniquely by run and feature

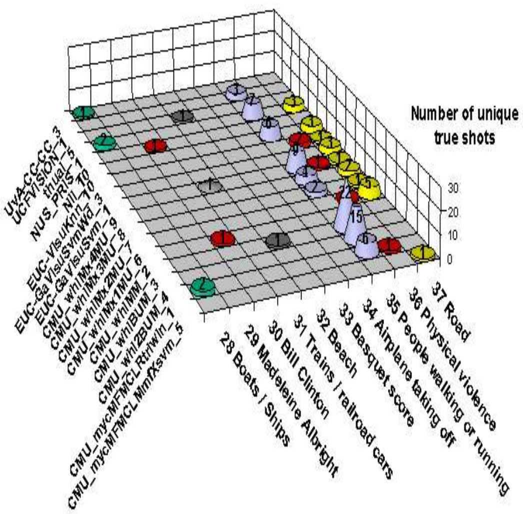
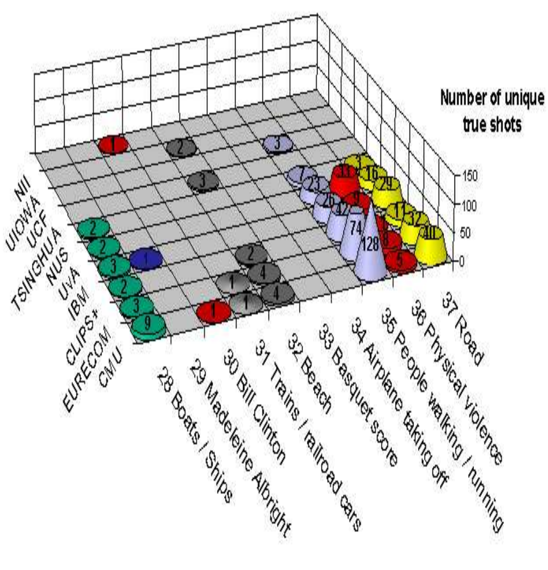


Figure 14: True shots contributed uniquely by group and feature



development data was different from that used to encode the test data. This caused problems for machine learning approaches which learned decoder-dependent characteristics of the development data not shared by the test data.

6 Search

The search task in TRECVID was an extension of its text-only analogue. Video search systems, all of which included a human in the loop, were presented with topics — formatted descriptions of an information need — and were asked to return a list of up to 1,000 shots from the videos in the search test collection which met the need. The list was to be prioritized based on likelihood of relevance.

6.1 Interactive vs manual search

As was mentioned earlier, two search modes were allowed, fully interactive and manual, and although no fully automatic mode was formally included, we did facilitate a late pilot of fully automatic submissions. A big problem in TREC video searching is that topics are complex and designating the intended meaning and interrelationships between the various pieces — text, images, video clips, and audio clips — is a complex one. The examples of video, audio, etc. do not always represent the information need exclusively and exhaustively. Understanding what an image is of/about is famously complicated (Shatford, 1986).

The definition of the manual mode allowed a human, expert in the search system interface, to interpret the topic and create an optimal query in an attempt to make the problem less intractable. The cost of the manual mode in terms of allowing comparative evaluation is the conflation of searcher and system effects. However if a single searcher is used for all manual searches within a given research group, comparison of searches within that group is still possible. At this stage in the research, the ability of a team to compare variants of their system is arguably more important than the ability to compare across teams, where results are more likely to be confounded by other factors hard to control (e.g. different training resources, different low-level research emphases, etc.).

One baseline run was required of every manual system — run based only on the text from the LIMSI ASR output and on the text of the topics.

6.2 Topics

Because the topics have a huge effect on the results, the topic creation process deserves special attention here. Ideally the topics would have been created by real users against the same collection used to test the systems, but such queries were not available.

Alternatively, interested parties familiar in a general way with the content covered by a test collection could have formulated questions which were then checked against the test collection to see that they were indeed relevant. This is not practical either because it presupposed the existence of the sort of very effective video search tool which participants are working to develop.

What was left was to work backward from the test collection with a number of goals in mind. Rather than attempt to create a representative sample, NIST tried to get an equal number of each of the basic types: generic/specific and person/thing/event, though in no way do we wish to suggest these types are equal as measured by difficulty to systems. Another important consideration was the estimated number of relevant shots and their distribution across the videos. The goals here were as follows:

- For almost all topics, there should be multiple shots that meet the need.
- If possible, relevant shots for a topic should come from more than one video.
- As the search task is already very difficult, we didn't want to make the topics too difficult.

The 24 multimedia topics developed by NIST for the search task express the need for video (not just information) concerning people, things, events, locations, etc. and combinations of the former. The topics were designed to reflect many of the various sorts of queries real users pose: requests for video with specific people or types of people, specific objects or instances of object types, specific activities or locations or instances of activity or location types (Enser & Sandom, 2002).

The topics were constructed based on a review of the test collection for relevant shots. The topic creation process was the same as in 2003 – designed to eliminate or reduce tuning of the topic text or examples to the test collection. Potential topic targets were identified by watching the test videos with the sound off. Non-text examples were chosen without reference to the relevant shots found. When more examples were found than were to be used, the subset used was chosen at random. The topics are listed in Appendix A. A rough classification of topic types for TRECVID 2003 and 2004 based on Armitage & Enser, 1996 is provided in Tables 4 and 5. At the request of participants, the fraction of topic involving action was increased in 2004.

6.3 Evaluation

Groups were allowed to submit up to 10 runs. In fact 16 groups (up from 11 in 2003) submitted a total of 67 interactive runs (up from 37), 52 manual ones (up from 38), and 23 fully automatic ones. Automatic runs did not contribute to the evaluation pools. In addition, 10 supplemental runs were submitted and evaluated though they also did not contribute to the evaluation pools.

Pooling was carried out differently than in 2003. All submissions were divided into strata of depth 10. So, for example, stratum A contained result set items 1-10 (those most likely to be true), stratum B items 11-20, etc. A sub-pool for each stratum was formed from the unique items from that stratum in all submissions and then randomized. To even out further the rate at which assessors could be expected to find true shots, the first several sub-pools were re-merged, re-randomized, and re-divided into subpools. Assessors were presented with the subpools in "alphabetical" order until they had judged the re-divided set and then ran out of time or stopped finding true shots. At least the top 5 sub-pools were judged completely for each topic. Beyond this, in some cases, the last sub-pool assessed may not have been completely judged. The average depth judged was 85. The maximum result set depth judged and pooling and judging information for each feature is listed in Table 3 for details. No relevant shots were found for topic 146 (slalom skiing) so it was not included in the evaluation.

After the workshop all the runs were re-evaluated using several sets of partial truth judgments based on shallower pooling than was used officially – in order to gather information about the effect of doing less judging. Figures 15, 16, and 17 show the mean average precision for the top interactive, manual, and automatic runs respectively. In most cases, while the actual mean average precision score changes significantly used using the shallowest pools, the relative ranking of systems in mostly the same.

6.4 Measures

The trec_eval program was used to calculate recall, precision, average precision, etc.

Table 3: Search pooling and judging statistics

Topic number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number relevant	% judged that were relevant
125	79074	21184	26.8	70	3061	14.4	154	5.0
126	80618	17844	22.1	100	2772	15.5	118	4.3
127	81765	20621	25.2	80	2743	13.3	64	2.3
128	77068	18559	24.1	80	2278	12.3	115	5.0
129	77705	19488	25.1	90	2581	13.2	16	0.6
130	79381	17447	22.0	140	3096	17.7	162	5.2
131	77229	19600	25.4	100	3227	16.5	86	2.7
132	82248	20270	24.6	60	2679	13.2	41	1.5
133	77112	16712	21.7	60	1216	7.3	46	3.8
134	75655	18769	24.8	60	1468	7.8	22	1.5
135	74713	16942	22.7	110	2685	15.8	54	2.0
136	76251	17671	23.2	80	2218	12.5	19	0.9
137	76851	17762	23.1	80	1568	8.8	106	6.8
138	81438	22862	28.1	90	4063	17.8	97	2.4
139	78970	19806	25.1	60	2074	10.5	55	2.6
140	73659	22130	30.0	70	2524	11.4	69	2.7
141	73898	20516	27.8	70	2728	13.3	54	2.0
142	73341	20697	28.2	50	1810	8.7	41	2.3
143	76931	22129	28.8	110	4608	20.8	39	0.8
144	81356	17557	21.6	70	2487	14.2	96	3.9
145	74838	21431	28.6	70	2638	12.3	67	2.5
146	72067	20372	28.3	90	2953	14.5	0	0
147	79597	20441	25.7	110	3708	18.1	85	2.3
148	79143	20152	25.5	140	4478	22.2	194	4.3

Figure 15: Effect of pool depth on evaluation of top 10 interactive search runs

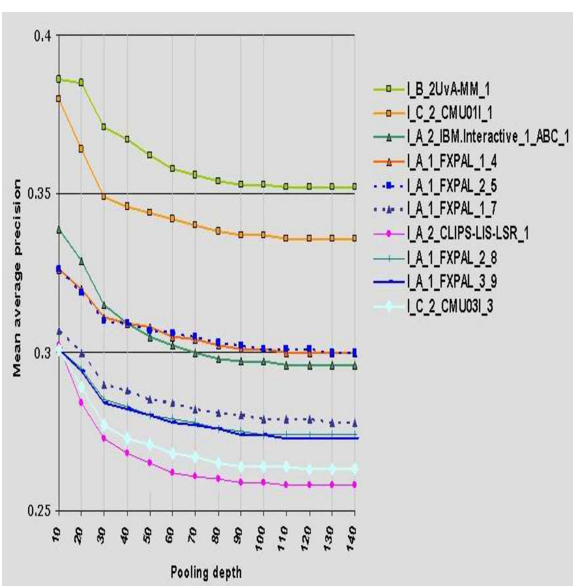


Figure 16: Effect of pool depth on evaluation of top 10 manual search runs

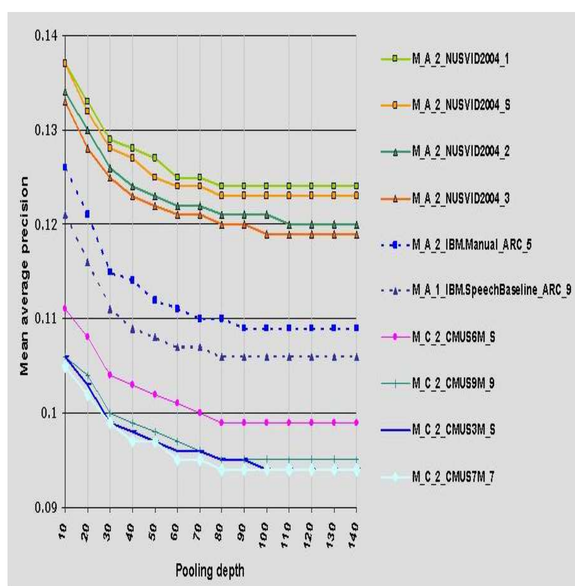
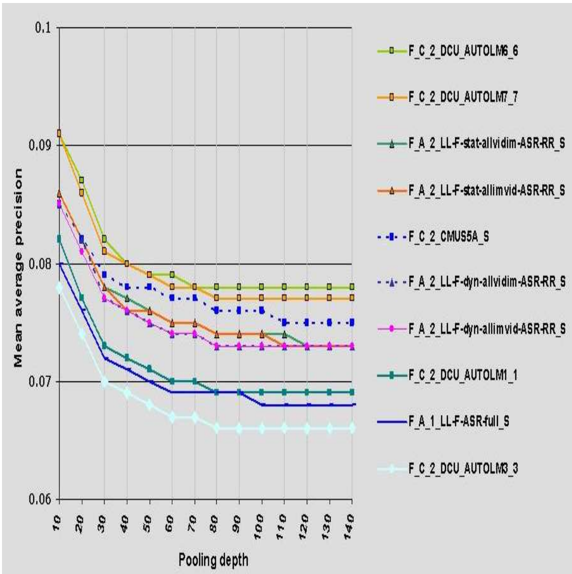


Figure 17: Effect of pool depth on evaluation of top 10 automatic search runs



6.5 Approaches in brief

Bilkent University used keyword-based ASR and color-based keyframe image matching along with CMU's donated features. *Carnegie Mellon University* looked at novice versus expert searchers and visual-only searching versus a full system that uses ASR and closed captioning. They used questionnaires and transaction logs to analyze their system's performance. They also submitted manual and automatic runs which used text retrieval to find candidate shots and the re-ranked the candidates by linearly combining scores from multimodal features or re-ranked weights trained by logistic regression. The *ITI/SCHEMA* team, a collaboration among ITI Thessaloniki, Munich University of Technology, and Dublin City University modified the MPEG-7 experimentation model (XM) so it could be used for low-level feature extraction and descriptor matching for search and retrieval.

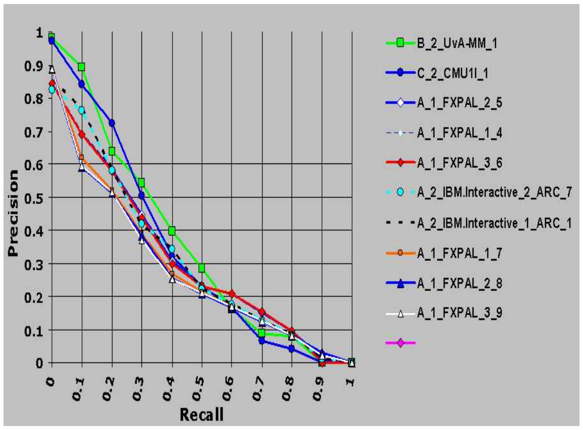
The *CLIPS-IMAG* system included a user-controlled combination of 5 modalities: keyword search on ASR, visual similarity to example images based on color histograms, use of 15 semantic categories (features), visual similarity to positive images (relevance feedback), and temporal closeness to positive images. CLIPS's submissions included manual and interactive runs with 4 users; The *Lowlands (CWI Amsterdam & University of Twente)* team continued their investigation of generative probabilistic models for video retrieval. They built dynamic (shot), static (keyframe), and language (ASR) models of each shot, created queries from topics automatically as well as by manual construction and selection of visual examples. Combining visual and text scored assuming independence gave better results than each modality on its own.

Dublin City University investigated interactive search based on text plus image versus text alone and found the combination better. In manual search using face filtering didn't seem to help. Their automatic runs were based on language models over ASR and visual features. They used multiple sources of text evidence (ASR, CC, OCR) and query expansion. The *FX Palo Alto Laboratory* emphasized interface elements facilitating rich visualization and quick and easy exploration of search results. Their search engine provided text search over the ASR and keyframe search by image similarity. *Imperial College London* explored content-based search plus k-nearest neighbor (kNN) browsing versus kNN browsing alone. They integrated relevance feedback, temporal browsing, historical browsing, and content-based search into a tight user-oriented interface. They redesigned their 2003 GUI image browser and placed strong emphasis on supporting the user task.

Indiana University submitted a manual baseline run searching ASR output. One user searched all topics with a five minute time limit for each. An interactive run searched ASR output. One user searched all topics with a ten minute time limit for each. Each topic began with a query generated during manual run. The ASR search was supported by tf/idf term weighting. They built on the previous years's ViewFinder image browsing interface. The *University of Amsterdam / MediaMill* team used a set of 32 semantic concepts - closely linked to their feature extraction task submissions. Global color histograms for image querying were also used. Topics close to their 32 concepts worked well, e.g. ice hockey, bicycle, Clinton. They used expert users. The *National Taiwan University* submitted manual search runs only. They aligning ASR tokens and high level features (concepts) using WordNet for word-word distances based on distance to common ancestor in WN. They considered the time delay in ASR vs image.

The *National University of Singapore* implemented a generic query analysis module, used 6 query-specific models, and the fusion of multi-modality features like text, OCR, visual concepts, etc. They borrowed ideas from text-based definition question-answering approaches. The *University of North Carolina at Chapel Hill* built 3 different in-

Figure 18: Top 10 interactive search runs



teractive search systems: text only (using ASR and MySQL full text search), features-based (using 10 IBM donated features), and semantic clusters of stories (using U.Iowa story segmentation and LSI on the ASR yielding semantic clusters). They created an extensive user questionnaire and carried out a detailed user study; The *University of Oulu / VTT* carried out manual and interactive search experiments. Manual runs investigated combinations of different search engines on visual (low level image features), concept (feature) and text (ASR and CC). Interactive runs used ASR only versus ASR plus visual and found them equivalent. They made improvements to their 2003 browser with extra support for user tasks;

No details were available from Queen Mary University of London and IBM Research at the time this overview was written.

6.6 Results

The results in terms of recall and precision for the top ten interactive, manual, and automatic runs (sorted by mean average precision (MAP)), are presented in Figures 18, 19, and 20 respectively. Mean elapsed time for each interactive run was approximately 15 minutes. The times for each depicted manual run is noted after its name. There is no clear relationship between manual effort as measured by elapsed time and precision/recall.

The median of the top 10 interactive runs is higher for 2004 than for 2003 and the runs are generally closer together. Manual runs are as expected considerably lower in performance than fully interactive runs and the median of the top 10 manual runs is lower than in 2003. The top 10 fully automatic runs

Figure 19: Top 10 manual search runs

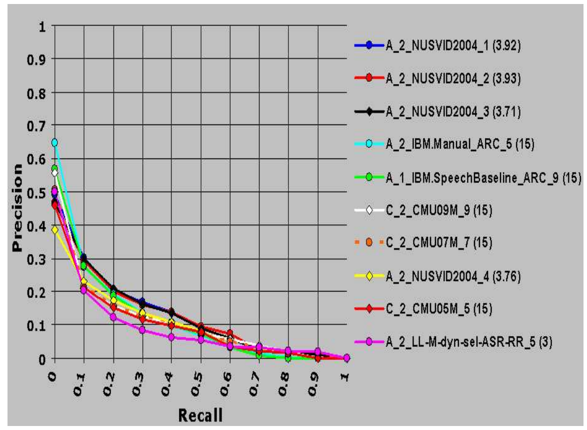


Figure 20: Top 10 automatic search runs

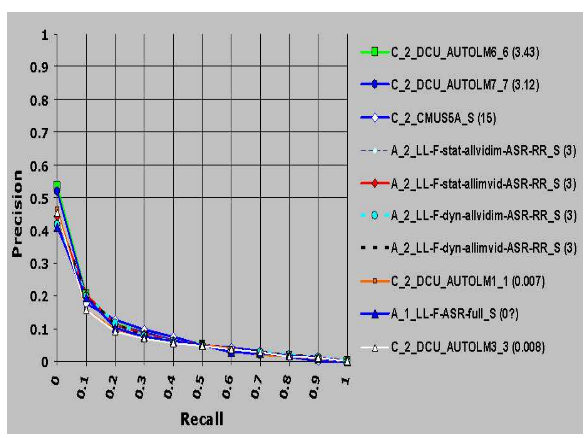


Figure 21: Relevant shots contributed uniquely by run

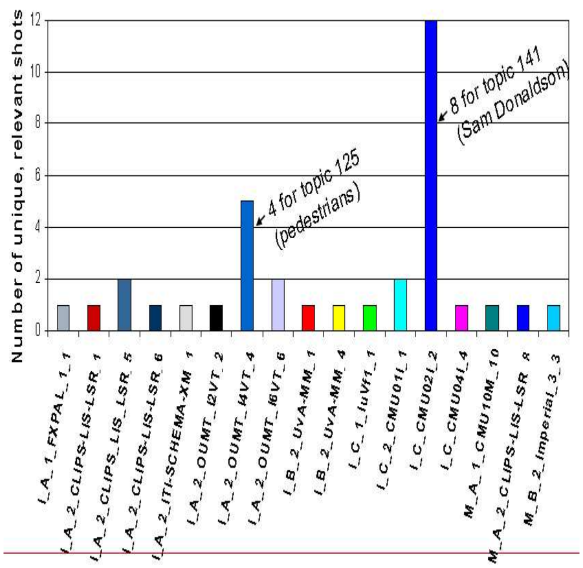
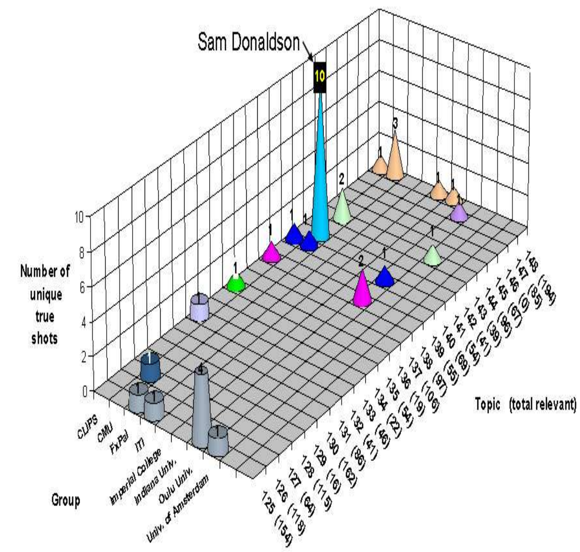


Figure 22: Relevant shots contributed uniquely by group and topic



are still closer together and, while performance drops, it is surprisingly still within the lower range top ten manual runs.

Looking underneath the averages at the variability due to topic, Figure 23 shows the best average precision by topic. A better understanding of relative performance per topic requires detailed knowledge of the system, the topic, and the test data. See the participants' papers for some interesting observations and tentative conclusions.

6.7 Issues

The implications of pooling/judging depth on relevant shots found and on system scoring and ranking have yet to be investigated thoroughly.

7 Summing up and moving on

This overview of the TREC-2004 Video Track has provided basic information on the goals, data, evaluation mechanisms and metrics used. Further details about each particular group's approach and performance can be found in that group's site report. The raw results for each submitted run can be found in the results section of at the back of the notebook.

8 Authors' note

TRECVID would not happen without support from ARDA and NIST and the research community is very grateful for this.

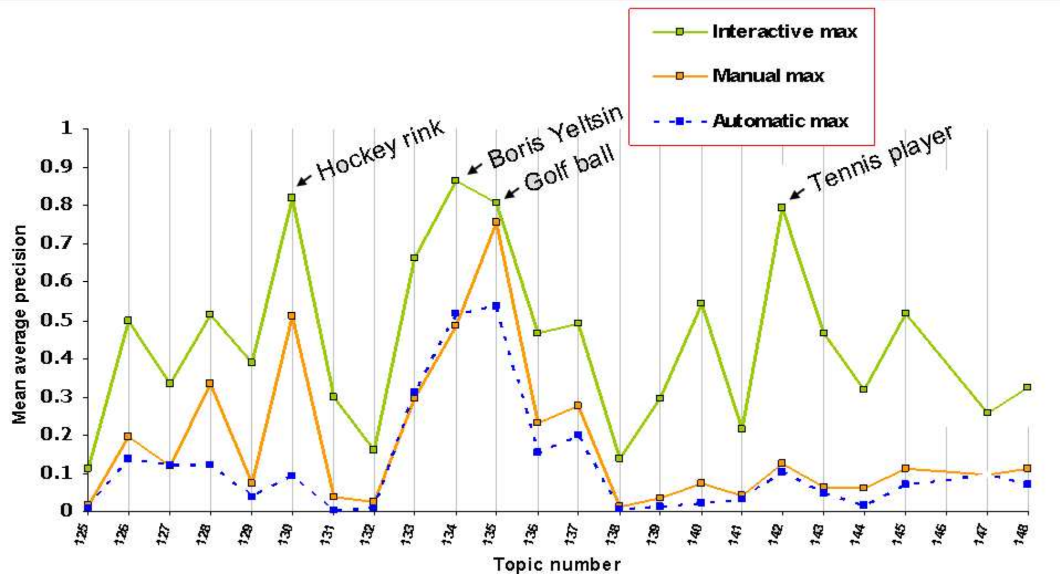
Beyond that, various individuals and groups deserve special thanks. We are particularly grateful once more to Kevin Walker and his management at LDC for making the data available despite administrative problems beyond their control. We appreciate Jonathan Lasko's painstaking creation of the shot boundary truth data oncer again. Special thanks again to Jean-Luc Gauvain at LIMSI for providing the output of their automatic speech recognition system for the entire collection, and to Georges Quénot at CLIPS-IMAG for once more creating the common shot reference, selecting the keyframes, and formatting the ASR output for distribution.

Finally, we would like to thank all the participants and other contributors on the mailing list for their enthusiasm, patience, and sustained hard work.

9 Appendix A: Topics

The text descriptions of the topics are listed below followed in brackets by the associated number of image examples (I), video examples (V), and relevant shots (R) found during manual assessment the pooled runs.

Figure 23: Best average precision by topic



125 Find shots of a street scene with multiple pedestrians in motion and multiple vehicles in motion somewhere in the shot. (I 1, V 2, R 154)

126 Find shots of one or more buildings with flood waters around it/them. (I 2, V 4, R 118)

127 Find shots of one or more people and one or more dogs walking together. (I 0, V 6, R 64)

128 Find shots of US Congressman Henry Hyde's face, whole or part, from any angle. (I 5, V 1, R 115)

129 Find shots zooming in on the US Capitol dome. (I 2, V 3, R 16)

130 Find shots of a hockey rink with at least one of the nets fully visible from some point of view. (I 2, V 3, R 162)

131 Find shots of fingers striking the keys on a keyboard which is at least partially visible.(I 0, V 4, R 86)

132 Find shots of people moving a stretcher. (I 0, V 5, R 41)

133 Find shots of Saddam Hussein.(I 3, V 2, R 46)

134 Find shots of Boris Yeltsin. (I 3, V 4, R 22)

135 Find shots of Sam Donaldson's face - whole or part, from any angle, but including both eyes. No other people visible with him. (I 1, V 4, R 54)

136 Find shots of a person hitting a golf ball that then goes into the hole.(I 0, V 3, R 19)

137 Find shots of Benjamin Netanyahu. (I 4, V 4, R 106)

138 Find shots of one or people going up or down some visible steps or stairs. (I 4, V 4, R 97)

139 Find shots of a handheld weapon firing. (I 4, V 4, R 55)

140 Find shots of one or more bicycles rolling along. (I 3, V 3, R 69)

141 Find shots of one or more umbrellas. (I 5, V 5, R 54)

142 Find more shots of a tennis player contacting the ball with his or her tennis racket. (I 3, V 4, R 41)

143 Find shots of one or more wheelchairs. They may be motorized or not. (I 4, V 4, R 39)

144 Find shots of Bill Clinton speaking with at least part of a US flag visible behind him. (I 2, V 2, R 96)

145 Find shots of one or more horses in motion. (I 2, V 5, R 67)

Table 4: 2003 Topic types

	Named			Generic		
Topic	Person, thing	Event	Place	Person, thing	Event	Place
100				X		
101				X	X	
102				X	X	
103	X					
104				X	X	
105				X	X	
106	X		X			
107				X	X	
108	X					
109				X		
110				X	X	
111				X	X	
112				X		
113				X		X
114	X					
115				X		X
116	X					
117				X	X	X
118	X					
119	X					
120	X					
121				X		
122				X		
123	X					
124				X	X	X

Table 5: 2004 Topic types

	Named			Generic		
Topic	Person, thing	Event	Place	Person, thing	Event	Place
125				X	X	X
126				X		
127				X	X	
128	X					
129	X					
130				X		X
131				X	X	
132				X	X	
133	X					
134	X					
135	X					
136				X	X	
137	X					
138				X	X	
139				X	X	
140				X	X	
141				X		
142				X	X	
143				X		
144	X			X	X	
145				X	X	X
147				X	X	
148				X		

146 Find shots of one or more skiers skiing a slalom course with at least one gate pole visible. (I 1, V 4, R 0 - this topic was dropped from the evaluation)

147 Find shots of one or more buildings on fire, with flames and smoke visible. (I 0, V 4, R 85)

148 Find shots of one or more signs or banners carried by people at a march or protest. (I 5, V 6, R 194)

10 Appendix B: Features

28 Boat/ship: segment contains video of at least one boat, canoe, kayak, or ship of any type.

29 Madeleine Albright: segment contains video of Madeleine Albright

30 Bill Clinton: segment contains video of Bill Clinton

31 Train: segment contains video of one or more trains, or railroad cars which are part of a train

32 Beach: segment contains video of a beach with the water and the shore visible

33 Basket scored: segment contains video of a basketball passing down through the hoop and into the net to score a basket - as part of a game or not

34 Airplane takeoff: segment contains video of an airplane taking off, moving away from the viewer

35 People walking/running: segment contains video of more than one person walking or running

36 Physical violence: segment contains video of violent interaction between people and/or objects

37 Road: segment contains video of part of a road, any size, paved or not

References

Aigrain, P., & Joly, P. (1994). The automatic real-time analysis of film editing and transition effects and its applications. *Computers and Graphics*, *18*(1), 93—103.

Armitage, L. H., & Enser, P. G. B. (1996). *Information Need in the Visual Document Domain: Report on Project RDD/G/235 to the British Library Research and Innovation Centre.* School of Information Management, University of Brighton.

Boreczky, J. S., & Rowe, L. A. (1996). Comparison of video shot boundary detection techniques. In I. K. Sethi & R. C. Jain (Eds.), *Storage and Retrieval for Still Image and Video Databases IV, Proc. SPIE 2670* (pp. 170–179). San Jose, California, USA.

Browne, P., Smeaton, A. F., Murphy, N., O'Connor, N., Marlow, S., & Berrut, C. (2000). Evaluating and Combining Digital Video Shot Boundary Detection Algorithms. In *IMVIP 2000 - Irish Machine Vision and Image Processing Conference.* Belfast, Northern Ireland: URL: www.cdvp.dcu.ie/Papers/IMVIP2000.pdf.

Enser, P. G. B., & Sandom, C. J. (2002). Retrieval of Archival Moving Imagery — CBIR Outside the Frame. In M. S. Lew, N. Sebe, & J. P. Eakins (Eds.), *Image and Video Retrieval, International Conference, CIVR 2002, London, UK, July 18-19, 2002, Proceedings* (Vol. 2383). Springer.

Ford, R. M. (1999). A Quantitative Comparison of Shot Boundary Detection Metrics. In M. M. Yueng, B.-L. Yeo, & C. A. Bouman (Eds.), *Storage and Retrieval for Image and Video Databases VII, Proceedings of SPIE Vol. 3656* (pp. 666–676). San Jose, California, USA.

Gauvain, J., Lamel, L., & Adda, G. (2002). The LIMSI Broadcast News Transcription System. *Speech Communication*, *37*(1-2), 89—108.

Lee, A. (2001). *VirtualDub home page.* URL: www.virtualdub.org/index.

Lin, C.-Y., Tseng, B. L., & Smith, J. R. (2003). *Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets.* URL: http://www-nlpir.nist.gov/projects/tvpubs/papers/ibm.final.paper.pdf.

Ruiloba, R., Joly, P., Marchand-Maillet, S., & Quénot, G. (1999). Towards a Standard Protocol for the Evaluation of Video-to-Shots Segmentation Algorithms. In *European Workshop on Content Based Multimedia Indexing.* Toulouse, France: URL: clips.image.fr/mrim/georges.quenot/articles/cbmi99b.ps.

Shatford, S. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging and Classification Quarterly*, *6*(3), 39—61.