TREC Video Retrieval Evaluation:
A Case Study and Status Report

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

Wessel Kraaij {kraaij@tpd.tno.nl}
Department of Data Interpretation
Information Systems Division
TNO TPD
2600 AD Delft, the Netherlands

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

June 1, 2004

1 Abstract

The TREC Video Retrieval Evaluation is a multi-year, international effort, funded by the US Advanced Research and Development Agency (ARDA) and the National Institute of Standards and Technology (NIST) to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. Now beginning its fourth year, it aims over time to develop both a better understanding of how systems can effectively accomplish such retrieval and how one can reliably benchmark their performance. This paper can be seen as a case study in the development of video retrieval systems and their evaluation as well as a report on their status to-date.

After an introduction to the evolution of the evaluation over the past three years, the paper reports on the most recent evaluation TRECVID 2003: the evaluation framework — the 4 tasks (shot boundary determination, high-level feature extraction, story segmentation and typing, search), 133 hours of US television news data, and measures —, the results, and the approaches taken by the 24 participating groups.

Table 1: 2001 Evaluation framework

2 Origins and evolution

The impetus for the TREC video retrieval evaluation effort came from the observation that while quantities of digital video were growing at a rapidly increasing rate and interesting research was being done, there was no widely available basis for scientific comparison of approaches: common training and test data, benchmark tasks, established evaluation procedures and measures. The TREC video retrieval evaluation was established in 2001 to explore the feasibility of addressing this need.

2.1 2001 - Feasibility

In 2001, 12 research groups used 11 hours of publically available educational/informational MPEG-1 video from NIST and the Open-Video Project [?] in 2 tasks: shot boundary determination and search (automatic and interactive). Shot boundary detection called for the identification of all transitions between shots and their categorization as abrupt (i.e., a "cut") or gradual. In the search task, systems were given the search test collection and a set of topics. Each topic was a multimedia statement of information need comprising a text description of the video needed along with optional example clips, images, and audio files. NIST provided guidelines as to the types of needs desired: video of generic or named people, objects, events, locations, and combinations of the foregoing. For each topic they were asked to return a list of up to 100 shots which met the need described in the topic.

Not sure of what sorts of topics the various participating groups could likely handle, we asked each group to contribute a few topics for a total 74. Some of the topics were designed to have a known, small number of relevant clips in the test collection, so search results could be evaluated automatically. Submissions for the remaining topics were judged manually for relevance at NIST.

The initial running of the evaluation showed it to be feasible. Participants found it useful. Assessors were able to judge a little over 100 clips per hour with inter-assessor agreement equal to or better than for judgments of text document relevance in TREC-2 and TREC-4.

One problem for the search evaluation was the lack of predefined units of retrieval. Such units would haved allowed us to pool results and judge a clip only once even though it was submitted by several systems. We could also have estimated recall for the more than the known-item searches. Even for the known item searches, a fuzzy match between the known items and the submitted clips was necessary.

The worst case situation in which the topics devised by a given group were too easy for them while the other topics were too hard did not occur. Several groups found their own topics quite challenging and most groups had some success with topics other than their own. But the topics needed to be regularized.

The relationship of the non-text examples to the information need as a whole was usually a complex one and varied greatly. The meaning of an image or clip is famously hard to determine [?]. What aspects, spacio-temporal extents of a video clip or image exemplified the information need and to what degree? This posed some problems in assessment but much more in the automatic translation of the topic to a system-specific query.

Table 2: 2002 Evaluation framework

We needed much more search test data in order to draw conclusions of any interesting scope. Comparison of systems was limited by confounding factors such as different training data. Small numbers of runs per group limited the number of approaches that could be compared.

With respect to shot boundary determination, differences in frame numbering from different decoders meant we needed to incorporate a "fuzzy" match (plus or minus 5 frames) in the shot boundary evaluation. Requiring the use of a particular decoder was deemed impractical.

Although we profited by the experience of the OT10.3 Thematic Operation of the GT10 Working Group of the ISIS Coordinated Research Project [?] in designing the evaulation, we decided on a simpler set of measures for 2002: precision and recall. In addition there was a desire to separate the measurement of whether each gradual transition was detected from the measurement of how accurately each detected one was located. For more information see the track report [?].

2.2 2002 - Expanding and settling down

In 2002, 17 research groups used 73 hours of video mainly from the Prelinger Archive [?] for system development and testing. In addition we used video from the Open-Video Project and some stock shot videos provided by the BBC Archive. The Prelinger material comprised digitized versions of educational, advertising, educational, industrial, and amateur films made in the 1930s-70s. Although the material contained various encoding anomalies it represented a true archive one could imagine someone wanting to search for historical footage to be incorporated in a new production.

In the shot boundary detection task, a gradual transition was considered detected if the submitted transition overlapped by at least one frame with the transition as defined in the ground truth. Then two new measures for gradual transitions - frame-recall and frame-precision were defined to measure how accurately systems found the gradual transitions. The number of runs allowed was increased to 10 to allow for more experiments. Several groups chose to use the runs to explore their control of precision-recall tradeoffs.

One of the participating groups, CLIPS-IMAG, defined standard shots for use as pre-defined units of retrieval.

A high-level feature extraction task was added to shot determination and search. 10 features (e.g., outdoors, face, cityscape, text overlay, instrumental sound, monologue) were defined jointly by the participants and deemed to be true of a shot if they were present in at least one frame. For each feature the participants were asked to return a ranked list of up to 1000 standard shots from the feature test collection for which the feature was true. While the ability to detect such features may have a variety of applications, the main question to be answered was whether and how they could be used in executing searches on as yet unseen topics.

Several groups volunteered to share the output of some of their feature detectors with other groups. Dublin City University contributed a standard set of keyframes, one per standard shot. Microsoft Research Asia and Spoken Language Processing Group at LIMSI provided the output of their automatic speech recognition systems. These donations allowed groups without feature detectors, ASR engines, keyframe extracters, etc. to nevertheless join the experiments in use of features in search. To the extent that multiple systems used the same features, keyframes, etc. comparability and isolation of the system effect was increased.

The fully automatic, extremely difficult translation of topics to queries was replaced with a "manual" search, which allowed a human system expert one chance at formulating an optimal query from each topic. Elapsed time was added as a measure of effort for interactive searches. Average precision was added as a measure for search and feature extraction results, i.e., the requirement that systems rank their output by confidence was added even though some methods (e.g. support vector machines) may yield only binary decisions.

NIST created 25 topics that reflected the kinds of queries posed by real video archive searchers [?] — requests for video of named or generic people things, events, places, and combinations of foregoing. This was done by viewing most of the test videos with the audio on, making notes about possible topic targets, reviewing these for commonalities across videos, choosing some candidates, and fleshing them out with examples from the Internet or if need be the test collection. The diversity of the test collection made this process difficult. The size made it time-consuming.

Table 3: 2003 Evaluation framework

Table 4: Participants and Tasks in 2003

Creating training data and adapting to very different sorts of video material each year increased the proportion of participant time spent on start-up. It also made it more difficult to draw conclusions about whether improvement was due to changes in systems or in the data.

Differences in the availability and quality of training data from group to group obscured differences in systems and added to the cost of participation. Comparability within a group's systems was reduced when some groups used different humans for each manual run. Interactive experiments were not taking advantage of best practices in experimental design to block for topic and searcher effects.

Topic development needed to be isolated more from knowledge of the test collection to avoid biasing the topics artificially toward one medium or approach, e.g., to avoid the possibility that words from the test video audio were used as topic text thus emphasizing the importance of text/ASR. The latter could distort results bearing on the question of whether/when queries and systems using more than text achieve better results than text-only queries. While there were exceptions, text-only runs more frequently than not were achieving better results than text+audiovisual runs.

2.3 2003 - Start of a 2-year cycle

In 2003, 24 research groups (see Table **??** used 133 hours of US newscasts from 1998 and some data from C-SPAN from roughly the same period. The amount of data and contractual prohibitions against electronic distribution forced us to distribute the data on IDE hard drives. This was managed by LDC and worked surprisingly well. A little over 30 drives were shipped; all arrived in good working order. The number of features to be automatically extracted grew from 10 to 17 with some feature definitions re-used from last year. A news story segmentation and typing task was added to examine the effectiveness of using full audio and/or visual cues over just text from ASR.

Ching-Yung Lin of IBM headed up a collaborative effort to annotate the development data. Jean-

Luc Gauvain of the Spoken Language Processing Group at LIMSI provided automatic speech recognition (ASR) output for the entire collection.[?]. Georges Quenot of the CLIPS-IMAG group once again provided a common set of shot boundary definitions and this year added keyframes to this and provided this, and the LIMSI ASR output, in MPEG-7 format.

The topic creation process at NIST was revised to eliminate or reduce tuning of the topic text or examples to the test collection. More effort was devoted to promoting good experimental designs for the interactive search experiments. In an effort to support more analysis of various approaches, the maximum number of runs each group could submit was increased to 10 for most tasks. The size of result sets were similarly increased to accommodate the results of extraction for frequently occurring features and topics with many relevant shots to 1000 for search and 2000 for feature extraction. To handle this more effectively despite shortened judgment time, NIST attempted to pool to different depths for different topics based on number of true/relevant shots found. Details follow.

3 Data

3.1 Video

Approximately 133 hours of video in MPEG-1 were available for system development and testing in the four tasks. This data was divided as follows.

A shot boundary test collection for the 2003 evaluation, comprising about 6 hours, was drawn from the total collection. It comprised 13 videos for a total size of about 4.9 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 collection exclusive of the shot boundary test set was ordered by date. The first half was used for system development, while the second half was used for testing — for story segmentation, feature extraction, and search. Eight files were withdrawn from the originally planned test collection due to poor quality. This part of the collection was distributed on harddrives by LDC.

3.2 Common shot reference, keyframes, ASR

The entire story/feature/search collection was automatically divided into shots by Georges Quenot at CLIPS-IMAG. These shots served as the predefined units of evaluation for the feature extraction and search tasks. The development collection contained 133 files/videos and 35067 shots as defined by the common shot reference. The test collection contained 113 files/videos and 32318 shots.

The CLIPS-IMAG group also extracted a keyframe for each reference shot and these were made available to participating groups along with ASR output provided by Jean-Luc Gauvain at LIMSI.

3.3 Common feature annotation

Ching-Yung Lin of IBM headed up a collaborative effort 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 other tasks. 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

3.4 Additional data

In addition to the MPEG-1 video data there was data created for the TDT task which was made available to TRECVID. This included the output of an automatic speech recognition system (*.as1) and a closed-captions-based transcript. The transcript was available in two forms, firstly as simple tokens (*.tkn) with no other information for the development and test data and secondly as tokens grouped into stories (*.src_sgm) with story start times and type for the development collection. The times in the TDT ASR and transcript data were based on the analogue version of the video and so were offset from the MPEG-1 digital version. LDC provided alignment tables so that the old times could be used with the new video.

Details about each of the four tasks follow.

4 Shot boundary detection

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 [?]; 3.8 hours total from television entertainment programming, news, feature movies, commercials, and miscellaneous [?]; 21 minutes total from a variety of action, animation, comedy, commercial, drama, news, and sports video drawn from the Internet [?]; an 8-hour collection of mixed TV broadcasts from an Irish TV station recorded in June, 1998 [?].

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 [?].

The shot boundary task was 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 to try 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 identify each shot boundary in the test collection and identify it as an abrupt or gradual transition.

4.1 Data

The test videos contained 596,054 total frames (10% more than last year) and 3,734 shot transitions (78% 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:

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 combination;

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 [1] was used to view the videos and frame numbers. The collection used for evaluation of shot boundary determination contains 3,734:

- 2,644 — hard cuts (70.7%)
- 753 — dissolves (20.2%)
- 116 — fades to black and back (3.1%)
- 221 — other (5.9%)

The percentage of gradual transitions remained about the same as in last year's antique videos, but among the gradual transitions there was a shift away from dissolves and toward more exotic wipes, fades, etc. Gradual transitions are generally harder to recognize than abrupt ones. The proportion of gradual transitions to hard cuts in this collection is about twice that reported by boreczky96 and by ford99. This is due to the nature and genre of the video collection we used.

4.2 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.

Detection performance for cuts and for gradual transitions was measured by precision and recall

[1] The VirtualDub [?] website contains information about VirtualDub tool and the MPEG decoder it uses. The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

Figure 1: Precision and recall for cuts (zoom)

Figure 2: Precision and recall for gradual transitions

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.

Figure 3: Frame-precision and frame-recall for gradual transitions

4.3 Results discussion

Most techniques were based on frame-frame comparisons, some with sliding windows. Comparisons were based on colour and on luminance, mostly. Some used adaptive thresholding.. Most operated on decoded video stream. Some had special treatment of motion during gradual transitions, of flashes, of camera wipes. Performance was getting better.

For cuts, the results of all but two runs lay within the upper right quadrant shown in Figure **??**. As illustrated in Figure **??**, performance on gradual transitions lagged, as expected, behind that on abrupt transitions, where for some uses the problem may be considered a solved one. Some groups (e.g., CLIPS, Ramon Llull University, FX-Pal) used their runs to explore a number of precision-recall settings and seem to have good control of this trade-off. Figure **??** indicates that ???

5 Story segmentation and typing

The new story segmentation and classification task was as follows: given the story boundary test collection, identify the story boundaries with their location (time) and type (miscellaneous or news) in the given video clip(s)

A story can be composed of multiple shots, e.g. an anchorperson introduces a reporter and the story is finished back in the studio-setting. On the other hand, a single shot can contain story boundaries, e.g. an anchorperson switching to the next news topic.

The definition of the story segmentation task was 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 segments were labeled as "miscellaneous".

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

- TRECVID 2003 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.
- TRECVID 2003 has a story classification task (which is optional).

With TRECVID 2003's story segmentation task, the goal was to show how video information can enhance or completely replace existing story segmentation algorithms.

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)

5.1 Data

The story test collection contained 2,929 story boundaries. About 67.6% of the material was classified as "news" in the ground truth.

5.2 Evaluation

Each group could submit up to 10 runs. In fact eight groups submitted a total of 41 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.

5.3 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 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.

Figure 4: Story Segmentation: Recall & Precision by System and Condition

Figure 5: Story Segmentation: F-measure by System

The evaluation of story classification was defined as follows: for each reference news segment, we checked in the submission file how many seconds of this timespan were marked as news. This yielded the total amount of correctly identified news subsegments in seconds. News segment precision was defined as the total time of correctly identified news subsegments divided by total time of news segments in the submission. News segment recall was defined as the total time of correctly identified news subsegments divided by the total time of reference news segments.

5.4 Results discussion

Video provided strong clues for story segmentation and even more for classification. For segmentation, Figures **??** and **??** show that for most systems, type 1 and or type 2 runs had higher precision than the type 3 (ASR only) run. In classification most systems exceeded the precision of an "always guess news" run. Figures **??** and **??** show type 1 and type 2 results always had better precision than those of type 3. Results for recall varied as reflected in the F-scores shown in Figure **??**. Most approaches were generic. But were the combination methods optimal and were the ASR segmentation runs state of the art?

5.5 Comparability with TDT-2 results

Results of the TRECVID 2003 story segmentation task cannot be directly compared to TDT-2 results because the evaluation datasets differ and different evaluation measures are used. TRECVID

Figure 6: Story typing: Recall & Precision by Condition (zoomed)

Figure 7: Story typing: Recall & Precision by Condition and System (zoomed)

Figure 8: Story typing: F-measure by System

2003 participants have shown a preference for a precision/recall-oriented evaluation, whereas TDT used (and is still using) normalized detection cost. Finally, TDT was modeled as an on-line task, whereas TRECVID examines story segmentation in an archival setting, permitting the use of global information. However, the TRECVID 2003 story segmentation task provides an interesting testbed for cross-resource experiments. In principle, a TDT system could be used to produce an ASR+CC or ASR+CC+Audio run. In fact, as indicated in Figure ??, IBM's ASR-only run was used their TDT software.

5.6 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 2003 compared to that used in TDT. There was not a lot we could do about this since we were constrained by the availability of news data in video format which has story boundary ground truth available to us. Other issues associated with the particulars of the TRECVID 2003 experiment include the alignment of audio/video, closed captions and ASR transcripts with the manual story bounds, the correct use of clipping points, and the definition of a news story as used in the TDT task.

6 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 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 task 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 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 suggested in on-line discussions by track participants. The number of features to be detected was kept small (17) 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 using the topics. The topics did not exist yet at the time the features were defined. The 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.

The features to be detected were defined as follows for the system developers and for the NIST assessors. Last year's were 1-10; this year's are numbered 11-27: [11] outdoors, [12] news subject face, [13] people, [14] building, [15] road, [16] vegetation, [17] animal, [18] female speech, [19] car/truck/bus, [20] aircraft, [21] news subject monologue, [22] non-studio setting, [23]

Table 5: Feature pooling and judging statistics

Figure 9: Feature extraction: Average Precision by Feature

sporting event, [24] weather news, [25] zoom in, [26] physical violence, [27] Madeleine Albright.

6.1 Data

As mentioned above, the test collection contained 113 files/videos and 32318 shots. For feature extraction this represented an dramatic increase from last year's 1848 shots. Testing feature extraction and search on the same data offered the opportunity to assess the quality of features being used in search.

6.2 Evaluation

Each group was allowed to submit up to 10 runs. In fact 10 groups submitted a total of 60 runs.

All submissions were pooled but in stages and to varying depths depending on the number of shots with the feature found. See Table **??** for details.

6.3 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.

Figure 10: Feature extraction: Average Precision by Feature for Top 10 Runs

Figure 11: Feature extraction: Average Precision by Feature for Top 5 Runs

Figure 12: Feature extraction: Average Precision by Feature for Top 5 Runs (easier features?)

Figure 13: Feature extraction: Average Precision by Feature for Top 5 Runs (harder features?)

Figure 14: Feature extraction: Average Precision for Best and Median Runs by True Reference Shots

Figure 15: Feature extraction: True Shots Contributed Uniquely by Feature and Run

Figure 16: Feature extraction: True Shots Contributed Uniquely by Feature and Group

6.4 Results discussion

NEED TO INCORPORATE SOME DISCUSSION THAT HIGHLIGHTS/SUMMARIZES APPROACHES IN RELATION TO RESULTS

As indicated in the boxplots of Figure **??**, the range of performance varied greatly from feature to feature. While median performance was generally low, best performance was in several cases quite good. Figure **??** shows the performance of the top ten runs for each feature. More than half were well above the median and again the spread varied from feature to feature with many results very close to eachother. The top five runs per feature were even closer to eachother as one can see in Figures **??**, **??**, and **??**.

One simple predictor of success might be the total number of true shots found — the more shots the easier the feature? Figure **??** suggests there may be such a relationship for the median performance but not for the best.

Systems can distinguish themselves through superior average precision but their ability to find shots other systems haven't can also be valuable and suggest ways in which systems can improve their performance. Figures **??**, and **??** break down the counts for each feature of true shots uniquely returned by each run or site.

6.5 Issues

The choice of the features and the characteristics of the test collection cause problems for the evaluation framework. Some features turned out to be very frequent. 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. A compressed schedule meant less time to manual judgment of submission than was desirable. More true shots are likely to exist. The question of how the inclusion of these would affect the absolute and relative performances remains for the time being an open research question.

7 Search

The search task in the Video Track 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.

7.1 Interactive vs manual search

As was mentioned earlier, two search modes were allowed, fully interactive and manual, though no fully automatic mode was included, a choice which has advantages as well as disadvantages. A big problem in TREC video searching is that topics were complex and designating the intended meaning and interrelationships between the various pieces — text, images, video clips, and audio clips — is a complex one and 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 [**?**].

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.

7.2 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 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; 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 don't want to make the topics too difficult.

The videos in the test collection were viewed and notes made about their content in terms of people, things, and events, named or unnamed. Those that occurred in more than one video became candidates for topics. This process provided a rough idea of a minimum number of relevant shots for each candidate topic. The third goal was the most difficult since there is no reliable way to predict the hardness of a topic.

The 25 multimedia topics developed by NIST for the search task expressed 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 [?].

The topics were constructed based on a review of the test collection for relevant shots, but this year the topic creation process was designed to eliminate or reduce tuning of the topic text or examples to the test collection. Potential topic targets were identified 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.

Table 6: Search pooling and judging statistics

Figure 17: Search: Average Precision by Topic

7.3 Evaluation

Groups were allowed to submit up to 10 runs. In fact 11 groups submitted a total of 37 interactive runs and 38 manual ones. In addition, 4 supplemental interactive runs were submitted and evaluated though they did not contribute to the pools.

All submissions were pooled but in stages and to varying depths depending on the number of relevant shots found. See Table ?? for details.

7.4 Measures

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

7.5 Results discussion

NEED TO INCORPORATE SOME DISCUSSION THAT HIGHLIGHTS/SUMMARIZES APPROACHES IN RELATION TO RESULTS

Performance of the top 10 manual search runs is depicted in Figure ?? with little spread or difference in manual effort. Time spent did not seem to be a determining factor in the top interactive system performance shown in Figure ??. Informedia, an establish system, stood out especially at high recall but a number of other newer, quite different systems from other research groups performed similarly overall. The CMU02 run was essentially the 2002 system run on 2003 data, while CMU01 contained enhancements which appear to have worked. IS THIS NOT POSSIBLE TO SAY BECAUSE OF THE SEARCHER EFFECT CONFOUNDING EACH?

Figure 18: Search: Precision & Recall For Top 10 Manual Runs (with mean manual elapsed time)

Figure 19: Search: Precision & Recall For Top 10 Interactive Runs (with mean total elapsed time)

Figure 20: Search: Average Precision For Best Interactive by Total Number Relevant

As expected, when one looks below the averages as one must in doing success/failure analysis, results varied greatly from topic to topic as shown in Figure ??. Again, number of relevant shots was not a dominant factor in predicting best performance ??. And systems and groups could still learn from each other in finding all the relevant shots — as shown in Figures ??, and ??.

SOMETHING NEEDS TO BE SAID ABOUT RESULTS THAT SHOW TEXT-ONLY IS NO LONGER KING.

DID PEOPLE FOLLOW UP THE BETTER INTERACTIVE DESIGN WITH THE APPROPRIATE ANALYSIS?

7.6 Issues

The implications of the variable depth pooling have yet to be investigated. It was very time-consuming to manage. A compressed schedule meant less time to manual judgment of submission than was desirable. More relevant shots are likely to exist. The question of how the inclusion of these would affect the absolute and relative performances remains for the time being an open research question.

8 Summing up and moving on

In the TREC Video Retrieval Evaluation, wheels turn within wheels. As system developers devise, test, reject, accept, refine, and/or combine approaches to the various tasks, the evaluation framework, within which they work, evolves.

Figure 21: Search: Relevant Shots Contributed Uniquely by Run

Figure 22: Search: Relevant Shots Contributed Uniquely by Topic and Group

Fundamental evaluation issues require continuing attention: finding available data, modeling interesting tasks appropriate for that data, choosing informative measures, controling the overarching experiment so that the system effect can be isolated from confounding factors such as training data/method and a human in the loop, but allowing for enough diversity that surprising solutions can still emerge.

Fundamental questions specific to video retrieval remain for both the evaluation and the research systems: how does one clearly express an information need using non-textual examples; will pre-defined features provide an effective basis for executing as yet unseen queries much as natural features (words) do for text retrieval; what is the optimal division of labor in video retrieval between the human with highly evolved visual capabilities and a machine ; how should content-based and concept-based approaches be combined; and so on.

In pursuit of answers to these and other questions, the evaluation continues in 2004, completing the 2-yr cycle on the 1998 news video with about 80 hours of new test data and adjustments to some tasks and procedures. Information about the plans for 2004 as well as what happened in the past evaluations, including detailed reports on the participants' experiments, is available from the website: www-nlpir.nist.gov/project/trecvid.

9 Authors' note

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

Beyond that, various individuals and groups deserve special thanks. We are particularly grateful 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. Special thanks are due to Ching-Yung Lin at IBM for heading up and supporting the common feature annotation effort, to Jean-Luc Gauvain at LIMSI for providing the output of their automatic speech recognition system for the entire collection, and to Georges Quenot at CLIPS-IMAG for creating the common shot reference, selecting the keyframes, and formating the ASR output for distribution. CMU provided NIST with a version of Informedia to be used in exploring the test and development collections and we are grateful for this and all their help with the installation process.

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

10 Appendix A: Topics

The text descriptions of the topics are listed below followed in brackets by the total count of relevant submitted shots found.

100 - Find shots with aerial views containing both one or more buildings and one or more roads [87]

101 - Find shots of a basket being made - the basketball passes down through the hoop and net [104]

102 - Find shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at [183]

103 - Find shots of Yasser Arafat [33]

104 - Find shots of an airplane taking off [44]

105 - Find shots of a helicopter in flight or on the ground [52]

106 - Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery [31]

107 - Find shots of a rocket or missile taking off. Simulations are acceptable [62]

108 - Find shots of the Mercedes logo (star) [34]

109 - Find shots of one or more tanks [16]

110 - Find shots of a person diving into some water [13]

111 - Find shots with a locomotive (and attached railroad cars if any) approaching the viewer [13]

112 - Find shots showing flames [228]

113 - Find more shots with one or more snow-covered mountain peaks or ridges. Some sky must be visible behind them. [62]

114 - Find shots of Osama Bin Laden [26]

115 - Find shots of one or more roads with lots of vehicles [106]

116 - Find shots of the Sphinx [12]

117 - Find shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings) [665]

118 - Find shots of Congressman Mark Souder [6]

119 - Find shots of Morgan Freeman [18]

120 - Find shots of a graphic of Dow Jones Industrial Average showing a rise for one day. The number of points risen that day must be visible. (Manual only) [47]

121 - Find shots of a mug or cup of coffee. [95]

122 - Find shots of one or more cats. At least part of both ears, both eyes, and the mouth must be visible. The body can be in any position. [122]

123 - Find shots of Pope John Paul II [45]

124 - Find shots of the front of the White House in the daytime with the fountain running [10]