TRECVID 2003 - An Overview

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

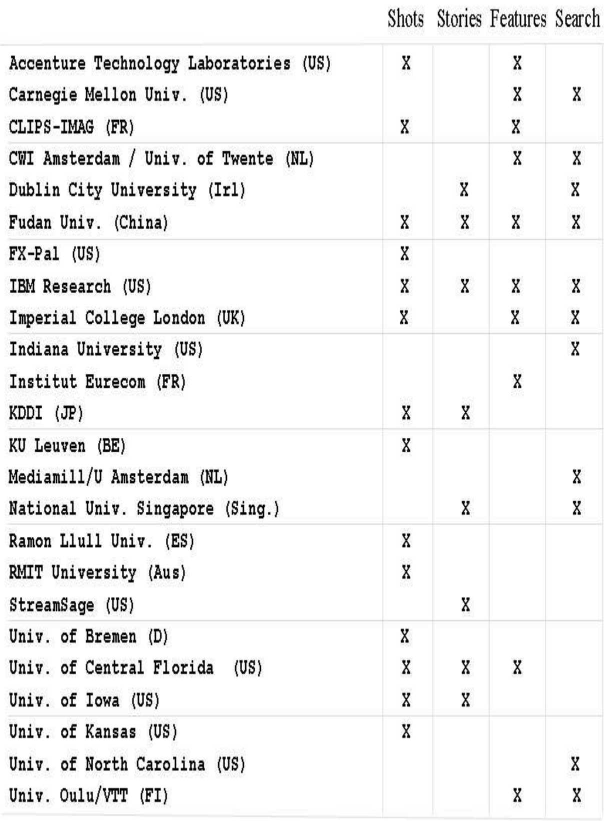
December 4, 2003

1 Introduction

TRECVID 2003 was the third 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 ARDA and NIST.

The evaluation used about 133 hours primarily of US broadcast news video in MPEG-1 format that had been collected for TDT-3 by the Linguistic Data Consortium in 1998. 24 teams representing 5 companies and 19 academic institutions — 4 from Asia/Australia, 10 from Europe, and 10 from the US — participated in one or more of four tasks: shot boundary determination, story segmentation/typing, 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



	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	
Carnegie Mellon Univ. (US)			X	X
CLIPS-IMAG (FR)	X		X	
CWI Amsterdam / Univ. of Twente (NL)			X	X
Dublin City University (Irl)		X		X
Fudan Univ. (China)	X	X	X	X
FX-Pal (US)	X			
IBM Research (US)	X	X	X	X
Imperial College London (UK)	X		X	X
Indiana University (US)				X
Institut Eurecom (FR)			X	
KDDI (JP)	X	X		
KU Leuven (BE)	X			
Mediamill/U Amsterdam (NL)				X
National Univ. Singapore (Sing.)		X		X
Ramon Llull Univ. (ES)	X			
RMIT University (Aus)	X			
StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

of, the evaluation framework — the tasks, data, and measures —, the results, and the approaches taken by the participating groups. For detailed information about the approaches and results, the reader should see the various site reports and the results pages at the back of the workshop notebook.

1.1 New in TRECVID 2003

At the TREC 2002 video track workshop in November 2002, the track set a number of goals for improvement (Smeaton, Over, & Taban, 2002) and in the subsequent months through cooperative effort met almost all of them. As a result the 2003 evaluation differs or extends the previous year's in a number of important ways itemized here:

- There was an increase in the number of participants who completed at least one task - up to 24 from last year's 17.
- The data changed significantly in quality and quantity. We moved from 73 hours of Prelinger Archive videos from the 1930s - 60s to 133 hours of broadcast news from 1998 with commercials, weather, sports, and graphics galore.
- The amount of data and contractual prohibitions against electronic distribution forced us to distribute the data on harddrives. 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.(Gauvain, Lamel, & Adda, 2002)
- 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 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.

2 Data

2.1 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 this year's 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.

2.2 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.

2.3 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

2.4 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.

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 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.

3.1 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 Boreczky and Rowe (1996) and by Ford (1999). This is due to the nature and genre of the video collection we used.

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.

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.

3.3 Approaches in brief

Accenture Technology Laboratories

Extract I-frames from encoded stream; Compute 3 Chi-square values across 3 separate histograms: global intensity, row intensity and column intensity and apply threshold, then combine; This gives indicator location and is followed by frame decoding and fine-grained examination;

CLIPS-IMAG

Based on image differences with motion compensation which uses optical flow as a pre-process and di-

[1] The VirtualDub (Lee, 2001) 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.

rect detection of dissolves; Same as used in TV2001 and TV2002 with little modification; Also includes direct detection of camera flashes;

Fudan University

Reused TV2002 SBD approach based on frame-frame comparison using luminance difference and colour histogram similarity; Adaptive thresholding Detection of camera flashes; GTs are searched seeking a black frame to determine whether they are fades, else dissolves;

FX-Pal

For each frame compute self-similarity against all in a window of past and future frames, as well as cross-similarity between past and future frames; Generates a similarity matrix and examine characteristics of this matrix to indicate cuts and GTs; Includes a clever way to reduce computation costs;

IBM Research

Used SBD from CueVideo system

Imperial College London

Colour histogram similarity of adjacent frames with a constant similarity threshold; Same as TV2002 and showing tradeoff of P vs. R as threshold varies; Good performance for simple approach;

KDDI

For cuts, preprocess the encoded MPEG-1 stream to locate high inter-frame differences using motion vectors then decode likely frames and test for luminance and chrominance differences; For dissolves, detect gradual changing over time using DCT activity data; Specific detection looking for wipes, and for camera flashes; Because it processes encoded stream, 24x real time on PC;

KU Leuven

Adaptive thresholding on the average intensity differences between adjacent frames; Includes motion compensation which computes an affine transformation between consecutive frames;

Ramon Llull University

Global colour histogram differences as a measure of discontinuity is used to detect cuts; For GTs, a method to account for linear colour variation of images across the duration of the GT, with specific treatment of moving objects during the GT which can distort this

RMIT University

Target GTs; Using a moving window of (200) frames, use current frame as a QBE against all in the window with a 6-frame DMZ around current frame; Based on frame-frame similarity and adaptive thresholding; A refinement on TV2002

TZI/University of Bremen

Combination of 3 approaches: changes in image luminance; gray level histogram differences; FFT feature extraction; Combined, with adaptive thresholding;

University of Central Florida

Colour histogram intersection of frames with subsampling of video at 5fps; This gives approximate location of shot bounds, followed by fine-grained frame-frame comparison using 24-bin colour histogram; Post-processing to detect abrupt changes in illumination (camera flashes); Also determined transition types;

University of Iowa

Comparison of adjacent frames based on: 512-bin global colour histogram, 60x60 pixel thumbnail vs. thumbnail based on pixel/pixel, Sobel filtering and detected edge differences.Then Boolean and arithmetic product combinations of these;

University of Kansas

No details available at this time

3.4 Results

See the results pages at the back of notebook for detailed information about the performance of each submitted run.

Most techniques are based on frame-frame comparisons, some with sliding windows. Comparisons are based on colour and on luminance, mostly. Some use adaptive thresholding, some dont. Most operate on decoded video strea.; Some have special treatment of motion during GTs, of flashes, of camera wipes. Performances are getting better.

As illustrated in Figures 1,2, and 3, performance on gradual transitions lags, as expected, behind that

Figure 1: Precision and recall for cuts

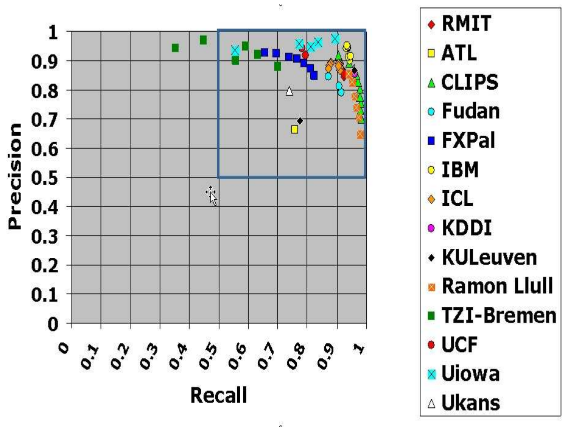


Figure 2: Precision and recall for cuts (zoom)

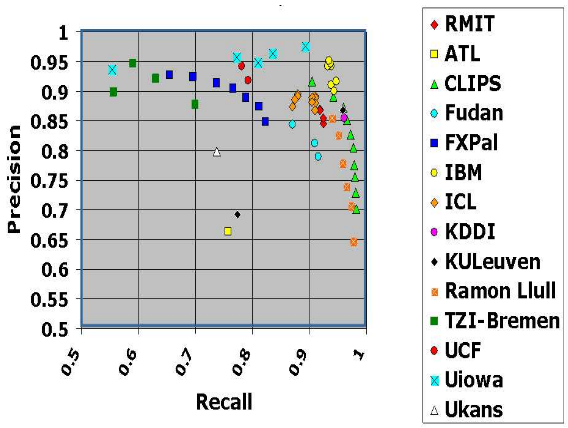


Figure 3: Precision and recall for gradual transitions

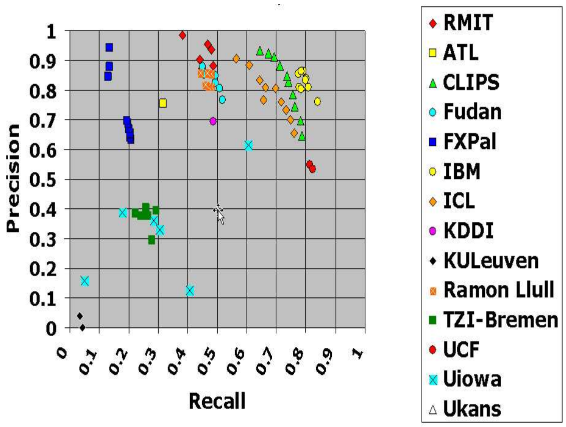
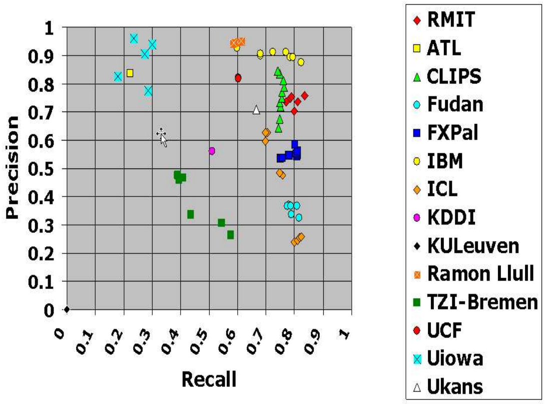


Figure 4: Frame-precision and frame-recall for gradual transitions



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 4 indicates that ???

4 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)

4.1 Data

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

4.2 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.

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 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.

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.

4.4 Approaches in brief

Fudan University

Segmentation: Anchor detection based on clustering and heuristics, Commercial detection based on ?, and ASR segmentation using a variant of Text-tiling Rule based and Maxent classifiers. News classification: GMM/Maxent using music, commercial and speech proportion as features.

IBM Research

KDDI

Segmentation: All shots are classified as ANCHOR, REPORT or COMMERCIAL, using audio and motion intensity, color into SVM. Subsequently rule based segmentation. Direct classification of boundaries, using the features of two shots before and after the boundary candidate. SVM Classification: SVM for NEWS-NEWS, NEWS-MISC and MISC NEWS

National University of Singapore

StreamSage/Dublin City University

ASR only segmentation runs. Three methods: lexical chaining to define topically coherent segments, Variant of text-tiling, Use methods 1 and 2 for compiling a list of cue-phrases that announce topic introduction or closure

University of Central Florida

Combined segmentation and classification: Story boundaries are marked by blank frames. Long stories imply news; short stories imply non-news. Merge adjacent non-news stories. Conclusion: story length is a strong feature for news classification

Figure 5: Story Segmentation: Recall & Precision by Condition

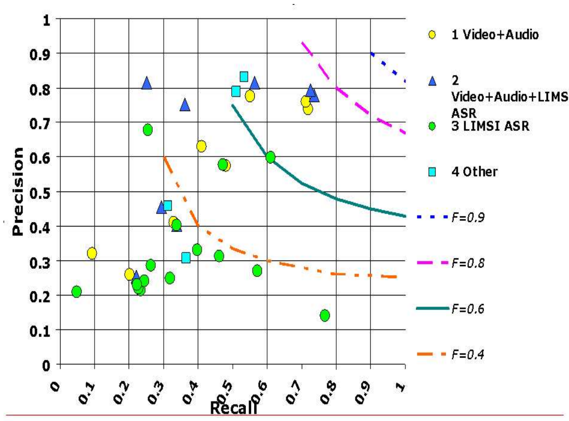
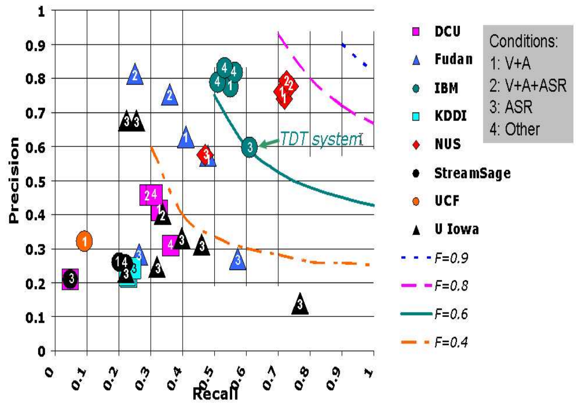


Figure 6: Story Segmentation: Recall & Precision by System and Condition



University of Iowa

4.5 Results

See the table in the results section of the notebook for details.

Figures 5,6,7, 8, 9,10,11,12 show ???

Video provides strong clues for story segmentation and even more for classification, best runs are either type 1 or 2. AV runs generally have a higher precision. Combination of AV and ASR gives a small gain for segmentation. Most approaches are generic. Are the combination methods optimal? Are the ASR segmentation runs state of the art?

Figure 7: Story Segmentation: F-measure by System

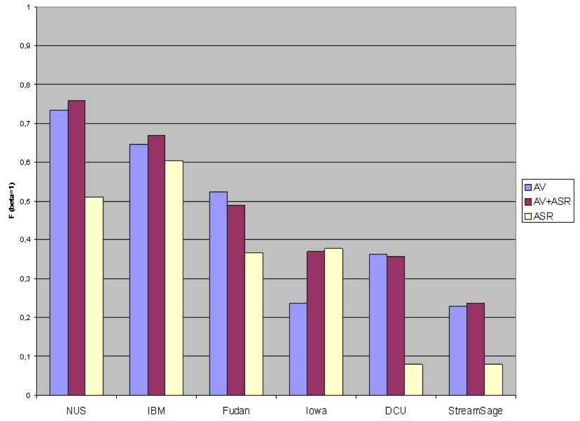


Figure 8: Story typing: Recall & Precision by Condition

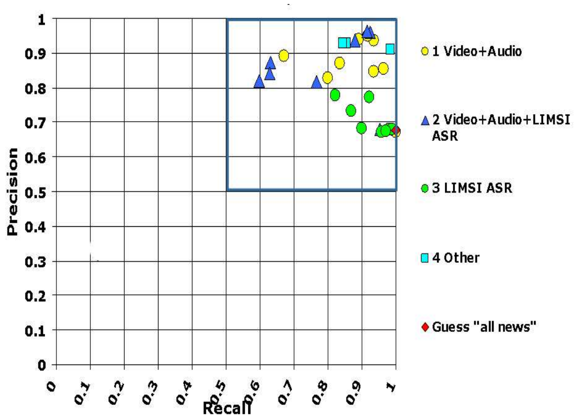


Figure 9: Story typing: Recall & Precision by Condition (zoomed)

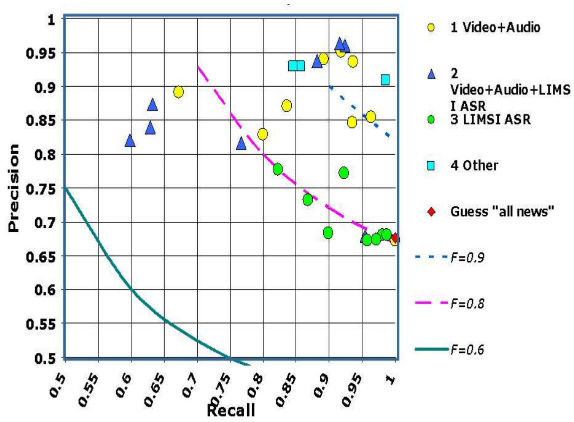


Figure 10: Story typing: Recall & Precision by System

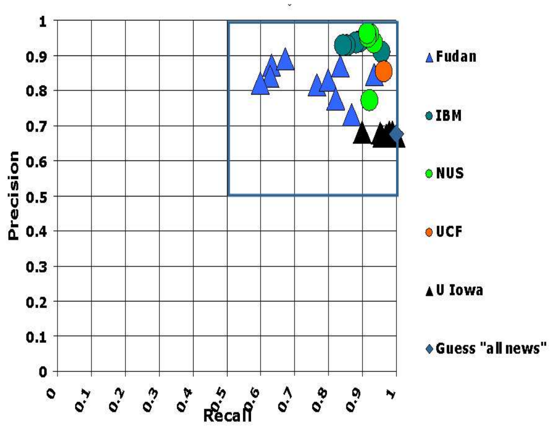


Figure 11: Story typing: Recall & Precision by Condition and System (zoomed)

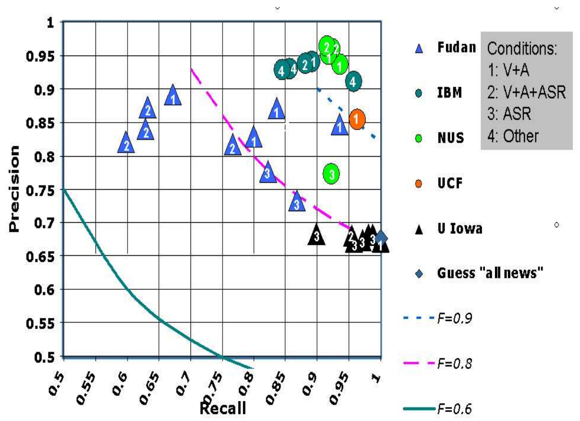
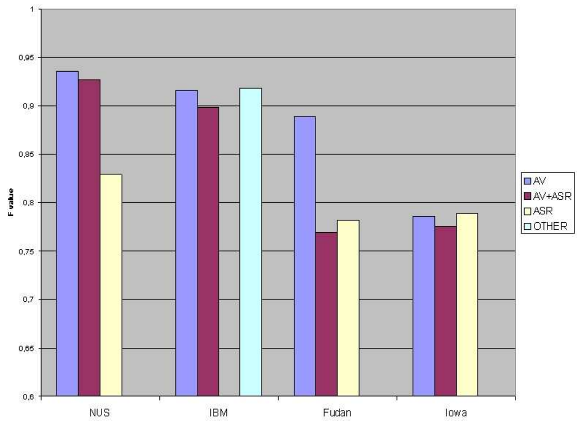


Figure 12: Story typing: F-measure by System



4.6 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 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.

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 2003 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. Other issues associated with the particulars of the TRECVID2003 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. Should this task be repeated in 2004?

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 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] sporting event, [24] weather news, [25] zoom in, [26] physical violence, [27] Madeleine Albright. The full definitions are listed with the detailed feature runs at the back of the notebook.

5.1 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.

Table 2: Feature pooling and judging statistics

Feature number	Total submitted	Unique submitted	% total that were unique	Max result depth pooled	Num judged	% unique that were judged	Num true	% judged that were true
11	70000	21142	30.2	100	2130	10.1	1045	49.1
12	52000	18700	36.0	100	1615	8.6	854	52.9
13	53452	20180	37.8	200	2820	14.0	1493	52.9
14	62000	21300	34.4	150	2515	11.8	923	36.7
15	58000	19351	33.4	100	1850	9.6	367	19.8
16	68000	18847	27.7	150	2170	11.5	1055	48.6
17	66296	20917	31.6	100	1936	9.3	235	12.1
18	66491	18025	27.1	150	1921	10.7	893	46.5
19	68436	21980	32.1	150	3150	14.3	717	22.8
20	62122	16229	26.1	150	1900	11.7	258	13.6
21	52000	10435	20.1	100	1020	9.8	266	26.1
22	64000	23040	36.0	350	2765	12.0	2429	87.8
23	70655	22264	31.5	150	2382	10.7	585	24.6
24	68519	21156	30.9	100	1051	5.0	166	15.8
25	36000	5267	14.6	350	1405	26.7	1175	83.6
26	60000	20323	33.9	150	1283	6.3	340	26.5
27	51376	17907	34.9	100	1035	5.8	35	3.4

5.2 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 2 for details.

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

Accenture Technology Laboratories

People: Skin tone detection, count faces. Weather: 200¡length¡1000 + color distribution + position of overlay text. Female Speech: Audio based gender detection + face + moving lips

Carnegie Mellon University

CLIPS-IMAG

1 feature: Madeleine Albright. How would a blind person locate a shot containing Madeleine Albright? Speaker detection (acoustic model) M.A. is probably mentioned in one of the preceding shots

CWI Amsterdam, University of Twente

14 features. Working hypothesis: Feature extraction == query by sample Generative probabilistic retrieval model (same as used for search task), divide frame in pixel blocks Take a sample of the annotated frames, rank the keyframes based on the likelihood that they generate the query sample

Fudan University

Scene features: grid, color histogram, edge direction, texture, KNN, AdaBoost. Vegetation, Weather: texture+color, SVM, GMM, MaxEnt Objects: - Car: Schneiderman - Animal: vegetation with KNN - Aircraft: detect context of aircraft. Audio: female speech : 12-MFCC, Pitch, 10-LPC

IBM Research

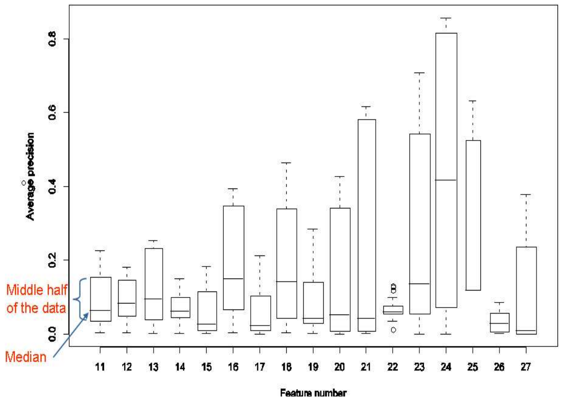
Imperial College London

Feature 16: Vegetation - Based on grass detector using a colour feature and KNN

Institut Eurecom

Apply LSI. 15 features. Keyframes are segmented into regions. Regions are clustered using K-means. Cluster X frame matrix is reduced by LSI. Use new feature space for GMM and KNN detectors.

Figure 13: Feature extraction: Average Precision by Feature



University of Central Florida

2 features. Weather news: Color histogram similarity Non-studio setting: Taken as all non anchor shots

University of Oulu/VTT

Extracted 15 features using: Motion, Temporal color correlogram, Edge gradients, Several low level audio features (used for outdoors, vehicle noise, sport, monologue. Feature fusion based on Borda count voting

5.5 Results

See the results section at the back of the notebook for details about the performance of each run.

Figures 13, 14, 15, 16, 17, 18, 19, 20, and 21 show ???

Some feature detectors had quite good results. Are features well chosen for search? Is detection quality good enough? Which combination methods work well? Which dont?

5.6 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.

Figure 14: Feature extraction: Average Precision by Feature for Top 10 Runs

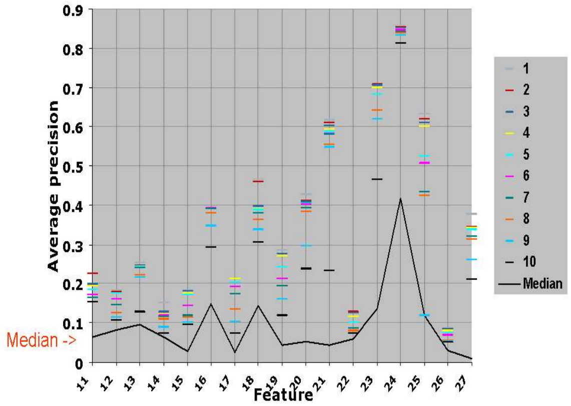


Figure 15: Feature extraction: Average Precision by Feature for Top 5 Runs

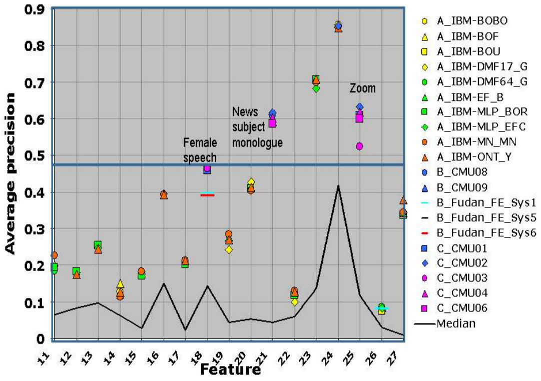


Figure 16: Feature extraction: Average Precision by Feature for Top 5 Runs (easier features?)

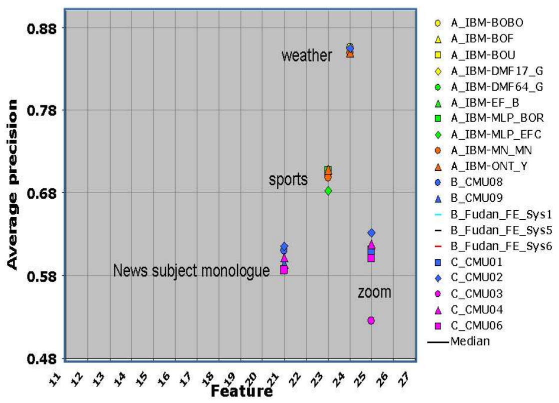


Figure 17: Feature extraction: Average Precision by Feature for Top 5 Runs (harder features?)

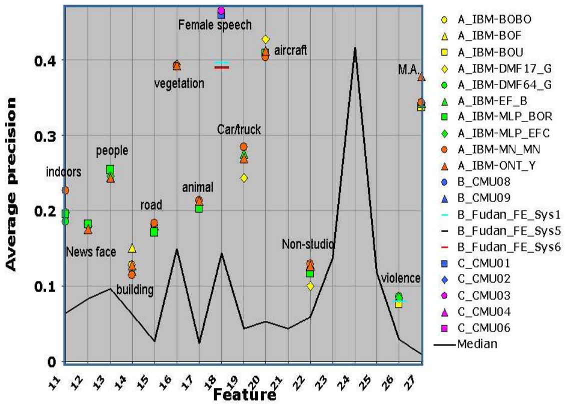


Figure 18: Feature extraction: Average Precision for Best and Median Runs by True Reference Shots

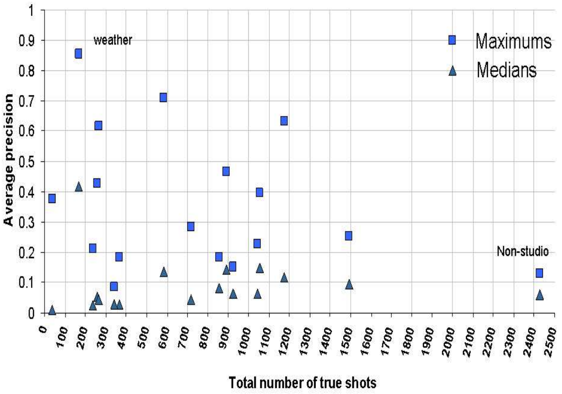


Figure 19: Feature extraction: True Shots Contributed Uniquely by Run

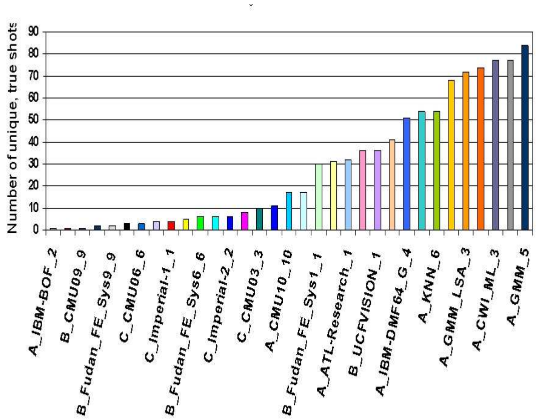


Figure 20: Feature extraction: True Shots Contributed Uniquely by Feature and Run

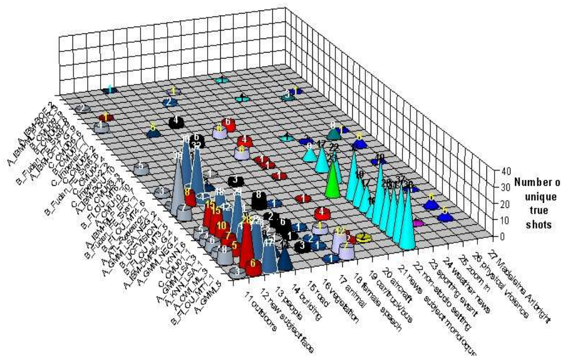
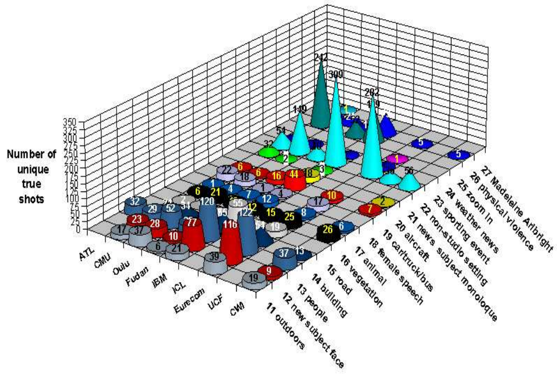


Figure 21: Feature extraction: True Shots Contributed Uniquely by Feature and Group



6 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.

6.1 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 (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 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 (Enser & Sandom, 2002).

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 with the search run results at the back of the notebook.

Table 3: Search pooling and judging statistics

Topic number	Total submitted	Unique submitted	% total that were unique	Max result depth pooled	Num judged	% unique that were judged	Num relevant	% judged that were relevant
100	53321	16150	30.3	50	1435	8.9	87	6.1
101	48425	16119	33.3	100	2111	13.1	104	4.9
102	48784	13276	27.2	50	932	7.0	183	19.6
103	45622	16938	37.1	50	1017	6.0	33	3.2
104	51136	15698	30.7	50	1355	8.6	44	3.2
105	49793	14930	30.0	50	1249	8.4	52	4.2
106	49180	16142	32.8	50	1268	7.9	31	2.4
107	48111	15101	31.4	50	1265	8.4	62	4.9
108	47508	17871	37.6	100	2211	12.4	34	1.5
109	47653	16287	34.2	50	1362	8.4	16	1.2
110	45362	18041	39.8	50	1328	7.4	13	1.0
111	49255	16939	34.4	50	1499	8.8	13	0.9
112	50369	16888	33.5	100	1987	11.8	228	11.5
113	49913	16280	32.6	50	1354	8.3	62	4.6
114	48691	16705	34.3	100	2520	15.1	26	1.0
115	50683	15709	31.0	100	2478	15.8	106	4.3
116	47492	16473	34.7	50	1291	7.8	12	0.9
117	49968	17612	35.2	100	3169	18.0	665	21.0
118	46689	16943	36.3	50	1328	7.8	6	0.5
119	41971	16869	40.2	50	1372	8.1	18	1.3
120	31291	9976	31.9	150	1610	16.1	47	2.9
121	47787	17381	36.4	100	1200	6.9	95	7.9
122	47462	16712	35.2	50	1328	7.9	122	9.2
123	49087	16792	34.2	50	1000	6.0	45	4.5
124	49397	14706	29.8	50	1408	9.6	10	0.7

6.3 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 3 for details.

6.4 Measures

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

6.5 Approaches in brief

Carnegie Mellon University

Interactive: same system as TV2002. Split topics among 5 individuals, text search across ASR, CC, OCR with storyboarding of keyframes, layout under user control, filtering based on features; another run used improved version with more effective visualisation and browsing;

Manual: multiple retrieval agents across colour, texture, ASR, OCR and some features, combined in different ways, incl. Negative pseudo-RF and "co-retrieval".

CWI Amsterdam, University of Twente

merging information from multiple modalities: run separate Qs for each topic example, combine different models of Qs,- combine sims from system / user judgments; to build a language model for each shot. Precomputing NNs for each keyframe in data; Interactive better than manual and combination of text/visual better than text solo

Dublin City University

Variation of Fischlar in interactive setting with 16 users, 7 mins each, doing 12 topics. Two system variations were ASR search only and ASR plus query image vs. shot keyframe. Both had shot-level browsing, user controlled ASR/image search balance, RF allowed by expanding text and/or image. Aim was to see if users used and benefited from text and image.

Fudan University

Manual search using 4 different approaches and then combinations:- ASR, - colour histogram, - multiple feature (colour hist, edge, coocurrence texture). - "pecial search" where user selects most appropriate for topic, from 1. human face recog, 2. general shot features, 3. multiple features, 4. motion (camera and object), 5. colour/texture, 6. colour regions.;

IBM Research

Examined Spoken Document Retrieval and content based techniques in manual runs. SDR used automatic and phonetic techniques and SDR fusion across multiple match functions, re-ranking shots based on color blobs.; Also did fully automatic multiple example content-based (which is beyond "manual") and fusion of content-based and SDR-based via linear weighting.;

Imperial College London

Used ASR and 11 low-level colour/texture, disregarding image footer likely to contain news ticker. Features include global colour, colour from frame centre, colour structure descriptors, RGB colour moments,44x27 pixel gray thumbnails, convolution filters, variance, image smoothness and uniformity, ASR.. Retrieval of kNNs, thumbnails on 2D display, RF by user movement of thumbnails. 2x manual, 4x interactive runs, results good.

Indiana University

Used ASR and built a system around interactive text search and query expansion plus video shot browsing; Interactive search with 1 subject doing all topics, 15 mins max but used only 10 mins; Future work is to include search based on visual features;

Mediamill - University of Amsterdam

Interactive search with 22 groups of 2 users (in pairs?), using a combination of:- CMU donated features, - derived "concepts" from LSI over ASR, - keywords from ASR, to yield an active set of 2,000 shots then a snazzy shot browser to select examples. Only 1 of 11 complete runs submitted. Used 1 system so no local variant to compare against, and selectively combined sets of users outputs per topic to generate submission; "Best" (per topic) objectively selected by submitting the result where the most shots were selected by the users

National University of Singapore

1. News story retrieval based on ASR and using WordNet and web to expand the original query, POS tagging of query; 2. Filter shots from story based on shot features; 3. Use image and video matching to re-rank remaining shots; In interactive runs user views top 100 shots and marks relevant ones. Results show marked impact of manual vs. interactive, I.e. user RF;

University of North Carolina

Compare ASR-only, features-only, ASR+features, in interactive search task; Features: aggregated results of 10 groups from 17 features used in extraction task;ASR was LIMSI, combination was 2xASR; 36 searchers, each doing 12 topics over systems in 15 mins per topic; Shot browser had annotated storyboard of keyframe + ASR, lots of pre- and post-questionnaire analysis

Results: no statistical difference in Precision, but statistical difference in recall where features-only was less than the other two poor feature recognition accuracy ? Large variability in time taken per search, avg 4 to 6 minutes; Much evaluation of users perception and satisfaction; Some helpful pointers on future assessment of interactive search;

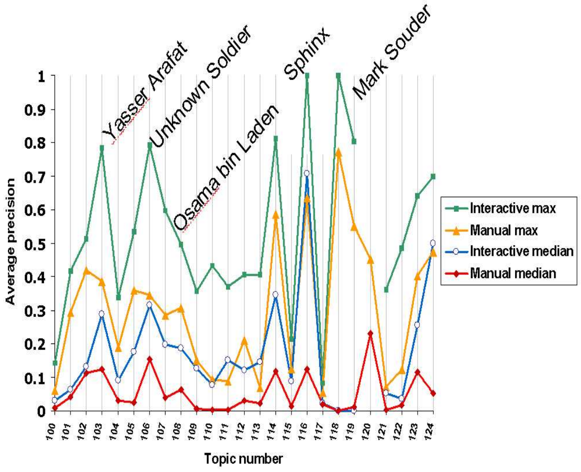
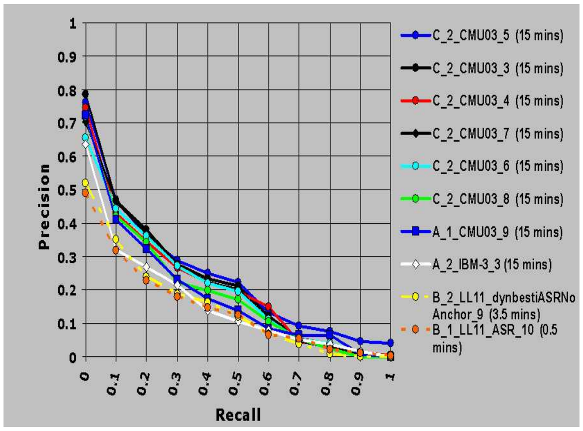


Figure 23: Search: Precision & Recall For Top 10 Manual Runs (with mean manual elapsed time)



University of Oulu/VTT

VIRE has interactive cluster/temporal shot browsing and shot similarity based on visual (colour, edge structure, motion), conceptual (15x features from feature set) and lexical (from ASR) similarity; Manual runs .. Pre-select combinations of features and images from topic; Interactive runs ...8 people, 2 systems, 9.5 mins per topic, (a) browse by visual features only and (b) browse by visual features plus ASR ...result indicates no significant difference;

6.6 Results

See the results pages at the back of the notebook for information about each search run's performance.

Figures 22,23,24,25,26, and 27 show ???

Figure 24: Search: Precision & Recall For Top 10 Interactive Runs (with mean total elapsed time)

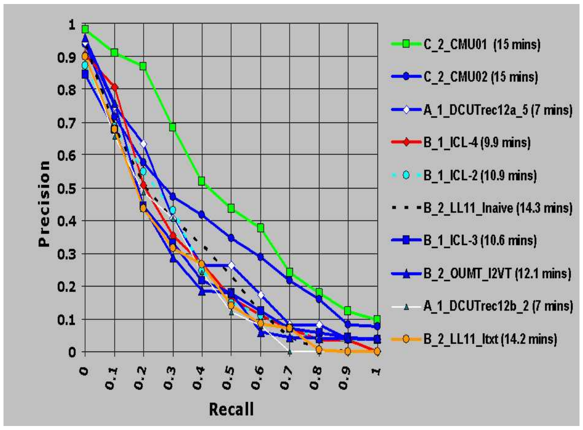


Figure 25: Search: Average Precision For Best Interactive by Total Number Relevant

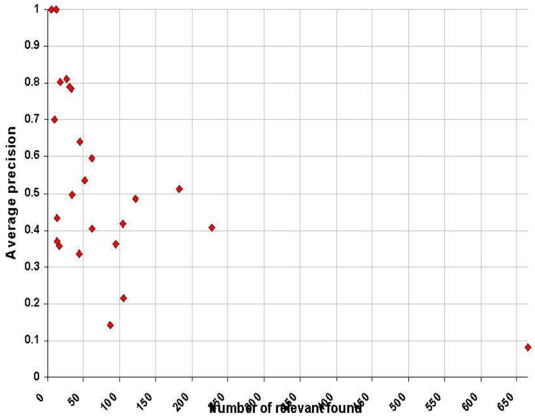


Figure 26: Search: Relevant Shots Contributed Uniquely by Run

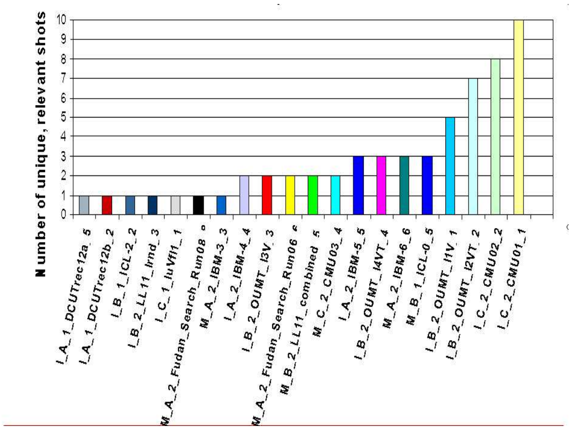
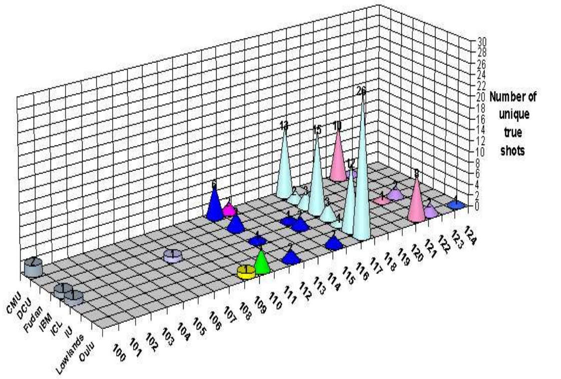


Figure 27: Search: Relevant Shots Contributed Uniquely by Topic and Group



Lots of variation, interesting shot browsing interfaces, mixture of interactive and manual. Approximately as much use of donated features as TV2002. A lot more participation, more runs, better at the upper end quite respectable curves! Nearly a dozen groups can now complete the search task and the demos are impressive.

6.7 Issues

The implications of the variable depth pooling have yet to be investigated.

7 Summing up and moving on

This overview of the TREC-2003 Video Track has provided basic information on the goals, data, evaluation mechanisms and metrics used. Further details about a 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.

In 2004 the track is likely to repeat the same tasks on data from the same sources but using data taken from later in 1998. This should reduce the startup time for continuing participants and make it easier to isolate the effect of system modifications on results. The development data for 2004 will comprise both the 2003 development and test data. We are already working with ARDA and LDC to make an additional 80 hours of CNN/ABC news video from 1998 available as test data in 2004. Distribution will again be by disk drive. We hope this will be available much earlier than was the case in 2003. CLIPS-IMAG and LIMSI generously have agreed to provide the common shot definition, keyframes, and ASR one more time.

8 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.

9 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]

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.

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.

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.

Smeaton, A., Over, P., & Taban, R. (2002). The trec-2001 video track report. In E. M. Voorhees & D. K. Harman (Eds.), *The Tenth Text REtrieval Conference (TREC-2001).* Gaithersburg, MD, USA.