The TREC-2002 Video Track Report

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

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

March 5, 2003

1 Introduction

TREC-2002 saw the second running of the Video Track, the goal of which was to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. The track used 73.3 hours of publicly available digital video (in MPEG-1/VCD format) downloaded by the participants directly from the Internet Archive (Prelinger Archives) (internetarchive, 2002) and some from the Open Video Project (Marchionini, 2001). The material comprised advertising, educational, industrial, and amateur films produced between the 1930's and the 1970's by corporations, nonprofit organizations, trade associations, community and interest groups, educational institutions, and individuals. 17 teams representing 5 companies and 12 universities — 4 from Asia, 9 from Europe, and 4 from the US — participated in one or more of three tasks in the 2001 video track: shot boundary determination, feature extraction, and search (manual or interactive). Results were scored by NIST using manually created truth data for shot boundary determination and manual assessment of feature extraction and search results.

This paper is an introduction to, and an overview of, the track framework — the tasks, data, and measures — the approaches taken by the participating groups, the results, and issues regrading the evaluation. For detailed information about the approaches and results, the reader should see the various site reports in the final workshop proceedings.

1.1 New in TREC 2002

At the TREC 2001 video track workshop in November 2001, 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 2002 track differs from the first running in 2001 in a number of important ways itemized here:

- There was an increase in the number of participants, up to 17 from last year's 12, and an increase in the data where a total of overview 73 hours of VCD/MPEG-1 data were identified for use in development and testing — up from 11 hours last year.
- A semantic feature extraction task was added. 10 features (e.g., cityscape, face, instrumental sound, monologue speech) were defined by a group of interested track participants and systems attempted with some success to find shots containing a given feature.
- Several groups volunteered to extract sets of these features from the test video and share their results with other groups allowing those other groups to use that feature detection in the search task. These feature detections were distributed in an MPEG-7 format developed by IBM.
- This year the track used a common set of shot definitions, donated by the CLIPS-IMAG group and formatted by Dublin City University whereas previously each group had defined their

own shot boundaries. Results for the feature detection and search tasks were reported in terms of these predefined units — allowing for pooling of results.

- The 25 topics for the search task were developed by NIST rather than by the participants and were released 4 weeks before the search results were due. These were again true multimedia queries as they all had video clips, images, or audio clips as part of the query, in addition to a text description. They 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). Unlike last year, where the topics were either known item or general, this year's topics were all general.

- The very difficult task of fully automatic topic-to-query translation was set aside for a future TREC video track. Searching in this year's track could be interactive with full human access to multiple interim search results, or "manual". In manual searches a human with knowledge of the query interface but no direct or indirect knowledge of the search test set or search results was given one chance to translate each topic to what he or she believed to be the most effective query for the system being tested.

- The shot boundary detection test set was not announced until 3 weeks before the submissions were due at NIST for evaluation. New and revised measures were used to separate a system's ability to detect shot transitions by identifying at least one of the frames in the transition from the accuracy with which a system locates the entire transition (frame-recall and frame-precision).

- Elapsed search time was added as measure of effort for the interactive search task and groups were encouraged to gather and report data on searcher characteristics and satisfaction.

Details about each of the three tasks follow.

2 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 the video track as an introductory problem, the output of which is needed for higher-level tasks such as search. Groups can participate for the first time on this task, develop their infrastructure, and move on to more complicated tasks the next year. Information on the effectiveness of particular systems is useful in selecting donated segmentations used for scoring other tasks.

2.1 Data

The shot boundary test collection for this year's TREC task comprises 4 hours and 51 minutes of video, slightly smaller than last year. The videos are mostly of a documentary/educational nature but were varied in their age, production style, and quality. There were 18 videos encoded in MPEG-1 with a total size of 2.88 gigabytes. The videos contained 545,068 total frames and 2,090 shot transitions (according to the manually created reference data.)

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.

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 2,090 transitions with the following breakdown as to type:

- 1466 — hard cuts (70.1%)
- 511 — dissolves (24.4%)
- 63 — fades to black and back (3.0%)
- 50 — other (2.4%)

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.

2.2 Evaluation

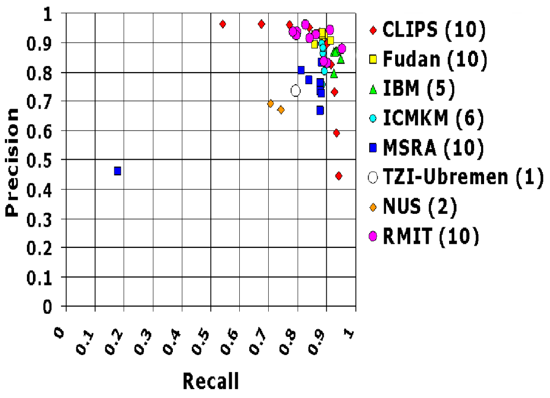
Participating groups in this task were allowed up to 10 submissions and these were compared automatically to the shot boundary reference data. Each group determined the different parameter settings for each run they submitted. Detection performance for cuts and for gradual transitions was measured by 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.

As last year, 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 1: Precision and recall for cuts



[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.

Figure 2: Precision and recall for gradual transitions

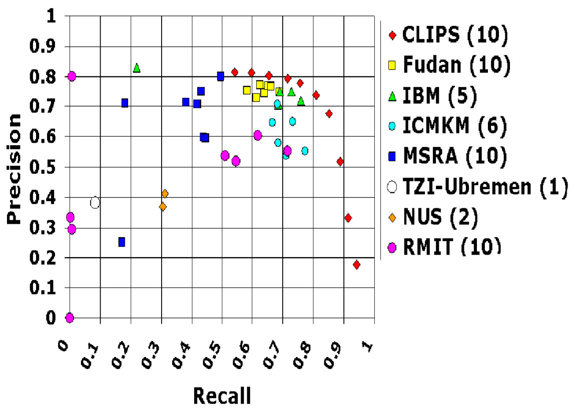
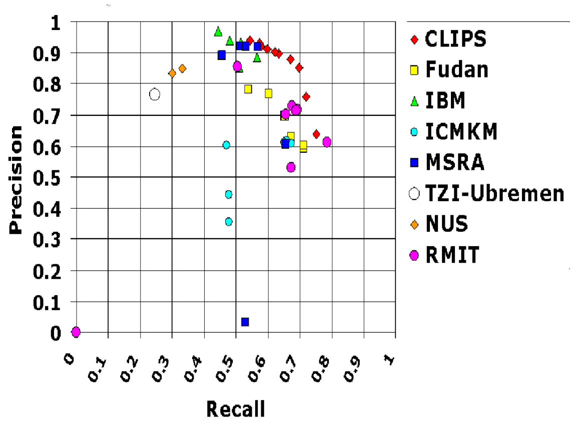


Figure 3: Frame-precision and frame-recall for gradual transitions



2.3 Results

As illustrated in Figure 1 and Figure 2, performance on gradual transitions lags, as expected, behind that on abrupt transitions, where for some uses the problem may be considered a solved one. The numbers in parentheses give the number of runs submitted by each group. Some groups (e.g., CLIPS and RMIT) used their runs to explore a number of precision-recall settings and seem to have good control of this tradeoff. Figure 3 indicates that at the level of frames in gradual transitions, the best systems have better precision than they do in detecting those transitions but their frame-level recall scores tend to be lower than for simple detection.

3 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 had the following objectives:

- to begin work on a benchmark for evaluating the effectiveness of detection methods for various semantic concepts
- to allow exchange of feature detection output based on the TREC Video Track 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.

The 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 the list, at most the top 1000 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 so as to be manageable in this first implementation 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.

Much to the appreciation of the track as a whole, some participating groups made their feature detection output available to participants in the search task and this will be discussed in the section describing the search task.

The features to be detected were defined as follows for the system developers and for the NIST assessors:

Outdoors segment contains a recognizably outdoor location, i.e., one outside of buildings. Should exclude all scenes that are indoors or are close-ups of objects (even if the objects are outdoors)

Indoors segment contains a recognizably indoor location, i.e., inside a building. Should exclude all scenes that are outdoors or are close-ups of objects (even if the objects are indoors).

Face segment contains at least one human face with the nose, mouth, and both eyes visible. Pictures of a face meeting the above conditions count.

People segment contains a group of two or more humans, each of which is at least partially visible and is recognizable as a human.

Cityscape segment contains a recognizably city/urban/suburban setting.

Landscape segment contains a predominantly natural inland setting, i.e., one with little or no evidence of development by humans. For example, scenes consisting mostly of plowed/planted fields, pastures, orchards would be excluded. Some buildings, if small features on the overall landscape, should be OK. Scenes with bodies of water that are clearly inland may be included.

Text Overlay segment contains superimposed text large enough to be read.

Speech a human voice uttering words is recognizable as such in this segment

Instrumental Sound sound produced by one or more musical instruments is recognizable as such in this segment. Included are percussion instruments.

Monologue segment contains an event in which a single person is at least partially visible and speaks for a long time without interruption by another speaker. Pauses are OK if short.

3.1 Data

23.26 hours (96 videos containing 7,891 standard shots) were randomly chosen from the total available data, to be used solely for the development of feature extractors. 5.02 hours (23 videos containing 1,848 standard shots) were randomly chosen from the remaining material for use as a feature extraction test set.

Table 1: Features and total hits

Feature name	Feature number	Shots submitted	Shots judged (pooled)	Total hits
Outdoors	1	12353	1821	962
Indoors	2	9143	1801	351
Face	3	7181	1688	415
People	4	4440	1233	486
Cityscape	5	9346	1656	521
Landscape	6	7208	1524	127
Text overlay	7	8120	1699	110
Speech	8	15800	1599	1382
Instrumental sound	9	11388	1846	1221
Monologue	10	5092	1319	38

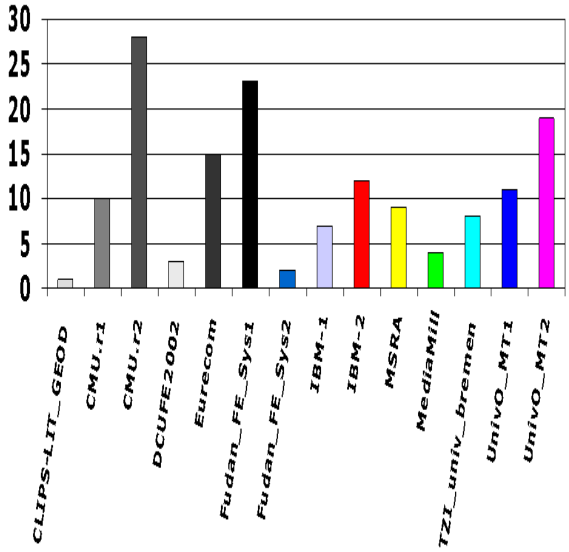
3.2 Evaluation

This year all result sets from all runs were fully assessed manually to create reference data. Basically, the feature extraction definitions were treated like topics of the form: “I want shots for which this feature is true.”

3.3 Measures

The trec_eval software, a tool available via trec.nist.gov, 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”.

Figure 4: The number of true shots contributed uniquely by run



3.4 Issues

It should be noted that in the case of some features (speech, instrumental sound) the number of shots in the feature extraction test set containing the feature approached or exceeded the maximum size of the submitted result set (1,000) and represented a large portion of the entire feature test collection size (1,848 shots) — see Table 1. While the performance of a random baseline was high in these cases, the median performance was still well above it. Where more hits exist than a result can hold, an artificial upper bound on possible average precision scores exists — namely for feature 8 (speech) 0.724 and for feature 9 (instrumental sound) 0.819.

3.5 Results

Figure 5 summarizes the results by feature for all of the runs at the median or above. Included as a dotted line in this figure is the baseline - the average for 100,000 randomly created result sets for each feature. The artificial upper limit on average precision mentioned above is indicated by a white triangle for features 8 and 9.

Results vary in their dispersion among features as well as in their mean. While the random baseline is high, almost all of the runs are well above it. While there was a lot of overlap in the shots submitted for a given feature, Figure 4 shows the relatively small number of true shots contributed uniquely by a given system – summed over all features. Not all systems submitted results for all features. The large overlap is no doubt due in part to the relatively small size of the test set (1,848 shots) in comparison to the size of the result (1,000 shots).

4 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 100 shots from the videos in the search test collection which met the need. The list was to be prioritized based on likelihood of relevance.

4.1 Data to be searched

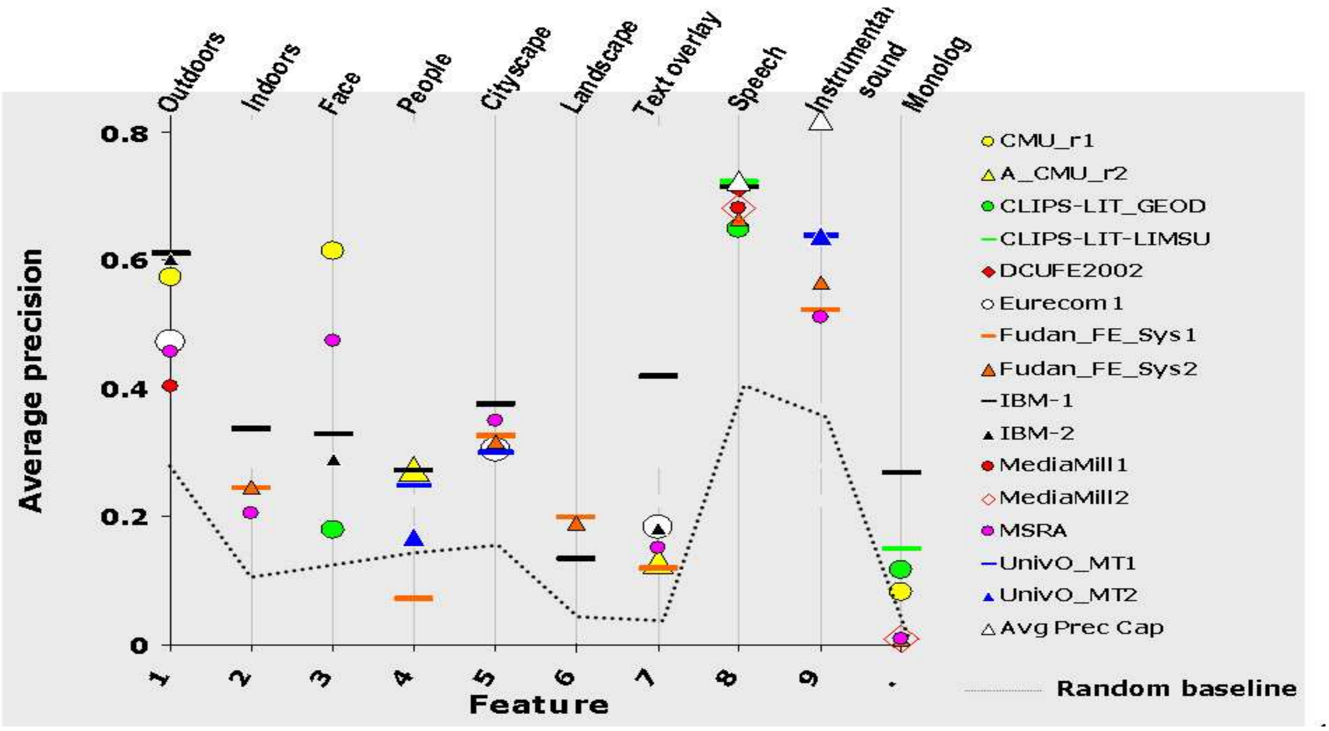
40.12 hours (176 videos containing 14,524 master shots) were randomly chosen from the identified collection to be used as the search test collection.

The video data was chosen because it represented an established archive of publicly available material that one can easily imagine being searched for information as well as historically interesting material that could be included in new video products. Publicly available video collections of any significant size are extremely hard to find. While we are not aware of any systematic study of the characteristics of the Internet Archive movie material, some details can be found in individual site papers. Collection characteristics will affect the scope of any conclusions drawn here.

Groups were allowed to develop their systems with knowledge of the search test collection — the topics being the surprise element. This was designated training pattern A. Other groups preferred to develop their systems without knowledge of the search test set. This training pattern was designated B. Results were labeled with these designations as were the feature extractions donated by some of the groups.

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 exhaus-

Figure 5: Average precision by feature and run



tively. 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 is 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.).

4.2 Topics

The topics were designed as multimedia descriptions of an information need, such as someone searching a large archive of video might have in the course of collecting material to include in a larger video or to answer questions. Today this may be done largely by searching descriptive text created by a human when the video material was added to the archive. The track's search scenario envisioned allowing the searcher to use a combination of other media in describing his or her need. How one might do this naturally and effectively is an open question. This year 25 topics were created by NIST, who had intended to create 50, but due to time pressures, this was not possible. Each topic contained a text description of the user information need. Examples in other media, e.g., one more video clips, still images, audio files illustrating the information need, were optional. Table 2 presents an overview of the topics, their types, and the number of relevant shots found for each topic.

Comparing the TREC video topic types to distributions of actual queries against video archives is nearly impossible due to lack of published studies, differences in archive content and searcher characteristics, amount of mediation, etc. However, Armitage and Enser (1996) provide some real world reference points which may be of interest. Comparing the distribution of TREC video track topics types to a sample of 370 submitted to the BBC Natural History Unit and 388 submitted to the British Film Institute's Na-

Table 2: Overview of topics

Topic #	Abbreviated text description of needed information/shot	Topic Types Panofsky-Shatford mode/facet categories (minus abstract) after Armitage & Enser 1996								Number of examples in the topic		Shots sub-mitted	Shots judged (pooling top 50 from each result)	Shots judged relevant
		Specifc				Generic								
		S1	S2	S3	S4	G1	G2	G3	G4					
		person; group; tning	event; action	location	linear time; date; period	kind of person; thing	kind of event; action; condition	kind of place: geograph-ical; architec-tural	cyclical time: season; time of day	Video	Image			
75	Eddie Rickenbacker	x								2	2	2668	850	15
76	Raymond H. Chandler	x								3	0	3036	625	47
77	pictures of George Washington	x								1	1	2521	931	3
78	depictions of Abraham Lincoln	x								1	1	2637	1014	6
79	people spending leisure time at the beach					x	x	x		4	0	3109	1055	55
80	one or more musicians					x	x			2	0	2829	860	63
81	football players					x				4	0	2311	890	15
82	women standing in long dresses					x	x			3	0	2696	1058	170
83	Golden Gate Bridge	x		x						0	5	2529	936	33
84	Price Tower in Bartlesville, OK	x		x						0	1	2409	816	4
85	Washington Square Park's arch in NYC	x		x						1	0	2708	909	7
86	overhead views of cities					x		x		4	0	3041	1112	105
87	oil fields, rigs, derricks					x		x		1	0	2721	1002	40
88	map of the continental US			x		x				4	0	2569	969	72
89	a living butterfly					x	x			0	2	2325	979	10
90	snow-capped mountain peaks or ridges					x		x		3	0	2785	926	75
91	one or more parrots					x				1	1	2228	880	17
92	sailboats, clipper ships, etc. with sails unfurled					x				4	2	2860	921	47
93	live beef or dairy cattle					x				5	0	3622	1003	161
94	groups of people walking in an urban environment					x	x	x		3	0	3168	1175	303
95	a nuclear explosion with a mushroom cloud					x	x			3	0	2658	951	17
96	one or more US flags flapping	x				x	x			2	0	2458	1055	31
97	microscopic views of living cells					x	x			2	0	2968	859	82
98	a locomotive approaching the viewer					x	x			5	0	2729	998	56
99	a rocket or missile taking off					x	x			2	0	2438	907	11

tional Film and Television Archive one sees the same predominance of non-abstract types and roughly the same percentage of type overlap (i.e., multi-category queries). However, the TREC queries have about half as many requests for specific persons and things and two to five times as many requests for generic persons and things. Whether this may be due to any degree to librarian/archivist mediation (e.g, substitution of a request for a known example for a generic request) is unknown.

4.3 Evaluation

The top 50 items (half) of each submitted result set was judged by a NIST assessor. Double judging in TREC 2001 indicated a high degree of assessor agreement for both relevant and non-relevant shots, so NIST did not do double judgments for TREC 2002.

4.4 Measures

The trec_eval program was used to calculate recall, precision, average precision, etc. The interested reader should see the back of the proceeding results pages for details on the performance of individual runs.

It should be noted that in the case of topics 82, 86, 93, and 94, as with evaluation in the feature extraction task, the number of relevant shots exceeded the maximum size of the submitted result set (100) — see Table 2. Where more relevant shots exist than a result can hold, an artificial upper bound on possible average precision scores exists — namely for topic 82 - 0.588, 86 - 0.952, 93 - 0.621, and 94 - 0.330.

4.5 Issues

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 avenue was also not open to us for two main reasons. First, the collection used is so diverse that creating a question that has answers in several videos is next to impossible without

Figure 6: Top 10 manual search runs

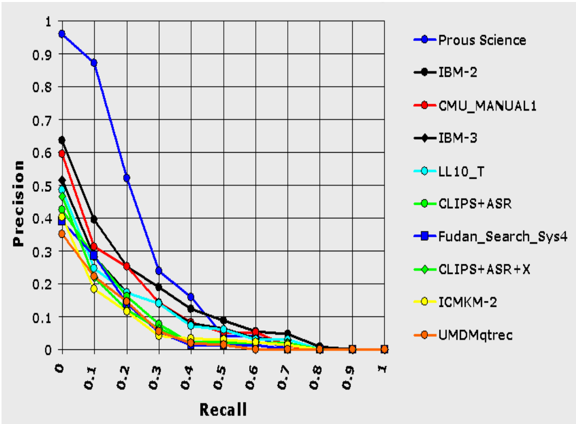
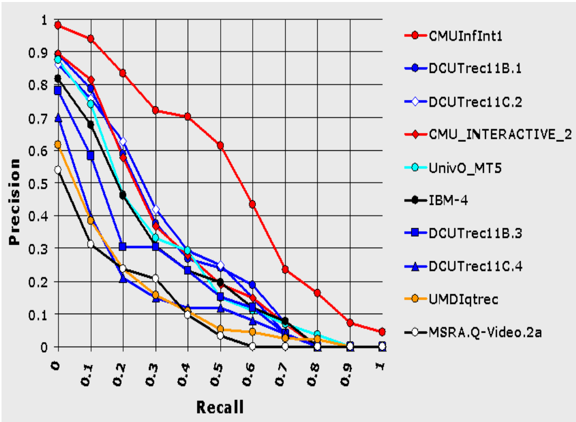


Figure 7: Top 10 interactive search runs



detailed knowledge of the collection. Second, NIST had no video search system in place which could be used.

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.

In general NIST tried to be sure there were relevant shots with relatively large images of the target person, thing, or event. When choosing examples for the topics, NIST tried to find at least some that seemed to resemble the target shot in shape, color, and/or texture. This was often not possible, nor is it likely the estimate of similarity corresponded in any meaningful way with that of the automatic systems.

Sometimes words from the audio were incorporated into the wording of the topic. This leaves open the possibility that some topics were in fact generally biased toward approaches using automatic speech recognition. On the other hand some information needs make demands unlikely to be supported by text from the audio e.g., requests for specific relative object/camera motion (98: locomotive approaching the viewer), some events/activities (96: US flags flapping), etc. A full analysis on the presence or absence of topic keywords in the audio track for relevant shots would be required to determine whether this is the case and has yet to be done.

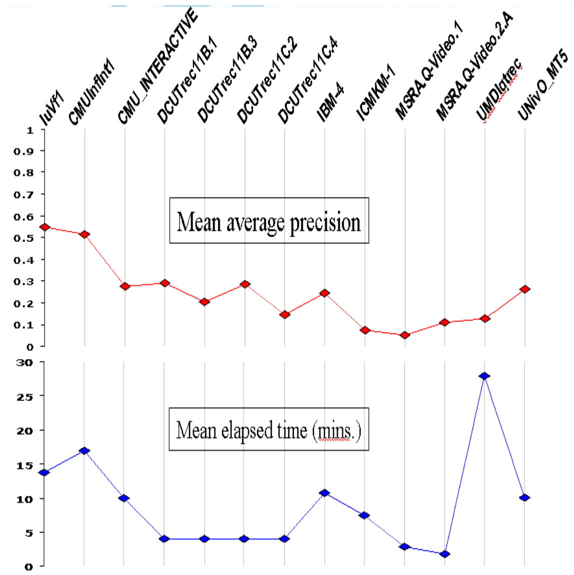
The nature of the test collection for 2003 and the possible use of a search tool to validate minimal numbers of relevant shots (even if a related system is likely to participate in the evaluation), should allow the creation of topics uncontaminated by the details of the test collection.

4.6 Results

The results in terms of mean average precision for the top ten manual runs are presented in Figure 6 and those for the top ten interactive runs in Figure 7, each list sorted by mean average precision. Another measure for interactive runs which was gathered was total elapsed time for each topic search. Figure 8 contrasts the two measures. Time spend varied widely from an average of just over 1 minute to just under 30 minutes per topic. No simple relationship between elapsed search time and effectiveness as measured by mean average precision is apparent.

The number of relevant shots contributed uniquely

Figure 8: MAP vs mean elapsed time



by each run is presented in Figure 9. As expected, interactive runs contribute the most.

The search task results in this report are based on manual relevance judgments for the top (most relevant) half (50 shots) in each submitted result set. The bottom half of each result has also been judged manually and this yielded few additional relevant shots except in the case of a couple topics which already had more than the average number of relevant shots. Fourteen of the twenty-five topics had no change in the number of relevant shots. For 8 the number of relevant shots grew 11% or less, for 3 it grew 20 - 24% (topics 82 - 20%; 94 - 21%; 96 - 24%). Figure 10 illustrates the distribution of relevant shots in the top versus the bottom half of the result sets.

Looking underneath the averages at the performance by topic, one can see that considerable variability exists across the set of topics and that some topics were harder than others. Figure 11 and Figure 12 show these together with two covariates: number of relevant shots and relevant videos. Manual results for topics 76, 84, 90, and 97 stand out. Why are they better? No single, simple explanation suffices. Topics with more relevant shots/videos or topics containing video examples from the search test collection (see small vertical arrows in Figure 12) are not necessarily easier.

The jury is still out with respect to two important search issues. The reliable usefulness of features in search generally or in specific situations has yet to be

Figure 9: Relevant shots contributed uniquely by run

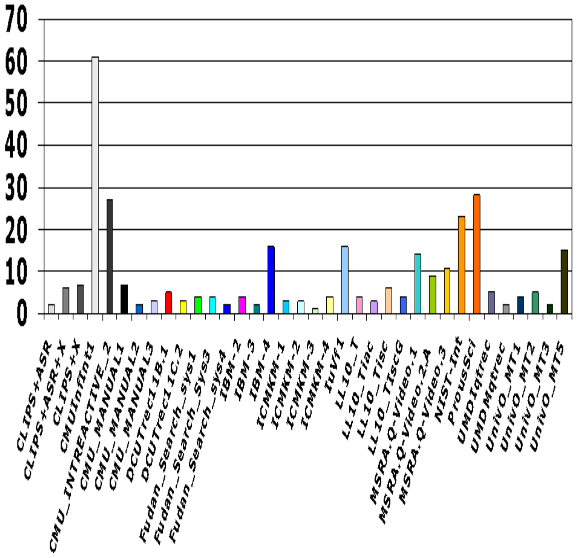


Figure 10: Distribution of relevant shots in result sets

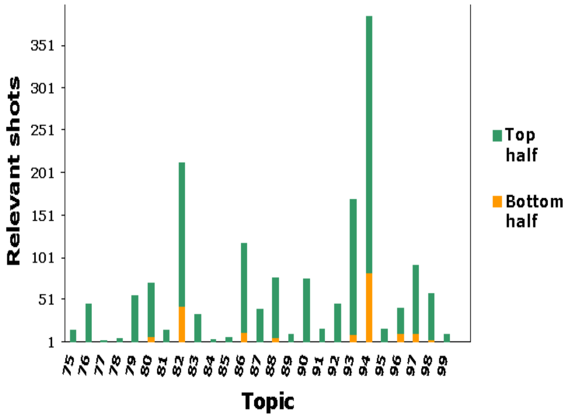


Figure 11: Interactive search: average precision by topic

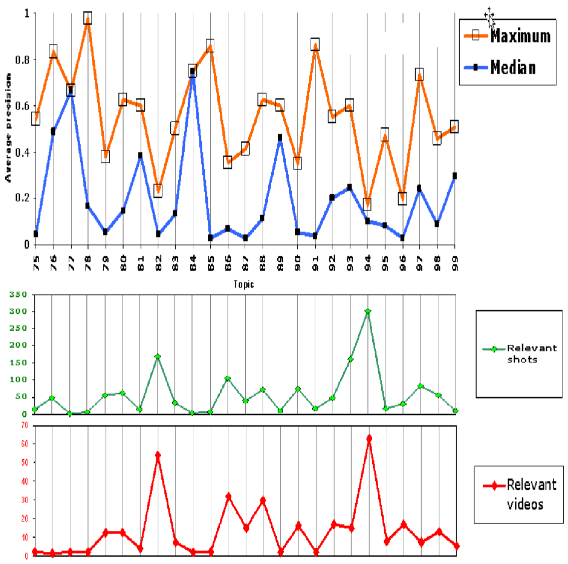
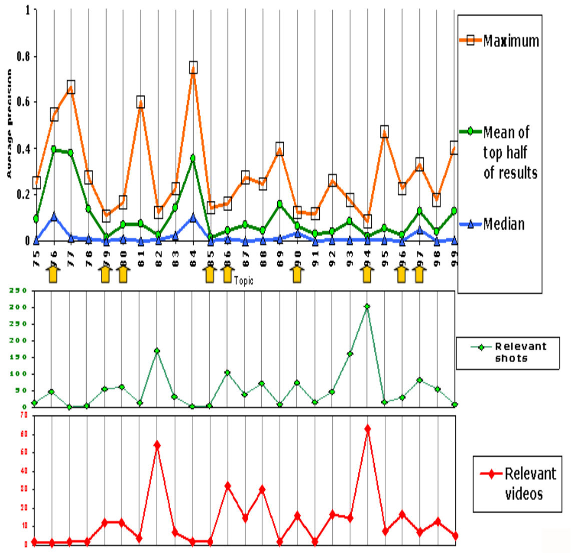


Figure 12: Manual search: average precision by topic



demonstrated. Similarly, the proper role and usefulness of the non-text topic elements is not yet clear. Matching the text of the topic against the text derived by automatic speech recognition on the video's audio track usually delivered better overall results than searches based on just the visual elements in the topic or combinations of the text and other elements. It is too early to draw convincing conclusions about either issue, but see the participants' papers for some interesting observations.

5 Approaches in brief

The following is a list of the groups that took part in one or more of the video track tasks and very short self-descriptions of the approaches taken by each participating research group. For detailed information the reader should consult the relevant system-specific paper in the proceedings.

5.1 Carnegie Mellon University (US)

The Informedia Project participated in the feature extraction task and both the manual and interactive search tasks. For the feature classification tasks, their standard approach was to hand label the feature training data using a labeling efficient interface, which allowed undergraduates to label one hour of video in 10 minutes for the presence/absence of one classification type (indoor, outdoor, etc.) They then extracted a set of standard low level image features such as HSV color histogram values, textures, EDH edge features, aggregrated line features, MPEG motion vectors and derived camera motion. These features were combined in a Support Vector Machine training process to produce a classification model for each category. Exceptions to this 'generic' image classifier approach were a custom developed face detector, a heuristic text detector and a decision-tree based people detector which used the face class as an input feature. Audio features were derived for the audio-based classes using a GMM model, and the monologue classifier combined both face output and audio features.

For the interactive track CMU used a modified version of the Informedia Digital Video Library System client, which was expanded to incorporate the classifier features and made more efficient to enable rapid display and exploration of large video data sets. It also incorporated an interface to multiple image search engines based on RGB or Munsell color, Texture, with different 3x3, 5x5, 7x7 blocks or QBIC-style image matching. An expert Informedia user, who did not have knowledge of the current TREC video collection, obtained the answers attempting to achieve high recall rather than speedy results. For the manual track, CMU submitted three systems: the first system was quite similar to last year's video track submission, combining speech recognition transcripts and OCR and image information in a linear fashion, while the second and best system extended the first system by incorporating the movie title and description information as text. This second system also added pseudo-relevance feedback for image retrieval as an additional combination module. Finally CMU submitted a third run using only the speech transcripts for text-only queries, without any relevance feedback or query expansion.

5.2 CLIPS-IMAG Grenoble (France)

This group used almost the same system for shot boundary detection as the one used for the TREC-2001 evaluation. This system detects "cut" transitions by direct image comparison after motion compensation and "dissolve" transitions by comparing the norms of the first and second temporal derivatives of the images. It also has a special module for detecting photographic flashes and filtering them as erroneous "cuts". Some parameters controlling the existing modules have been tuned using the TREC-2001 SBD corpus and reference segmentation, and a global parameter for the tuning of the recall versus precision compromise has been inserted.

The CLIPS group extracted only features 3 (faces), 4 (people), 8 (speech) and 10 (monologue). Face and people detection were based on a face detection tool publicly available from CMU run on one keyframe automatically extracted for each shot. The results were ranked according to the presence of a face and its size for feature 3 and according to the presence of at least two faces and the total size for feature 4. For features 8 and 10, they used the output of two different speech recognition systems, one from CLIPS-IMAG (GEOD team) and the other from LIMSI-CNRS, the same output as used by the group from Dublin. For feature 8, the length of detected speech segment within shots was used for ranking the results. For feature 10, the results were ranked using a combination of the length of a speech segment and the presence of a face.

Finally, CLIPS submitted three manual runs for the search task. One based only on speech transcription, on based only on a combination of donated features, and one based on a combination of both.

5.3 Dublin City University (Ireland)

DCU submitted results for three of the features from the feature set, namely speech, instrumental sound (music) and faces. Each technique worked directly on the encoded MPEG-1 bitstream. Speech extraction was based on measuring the duration of the rate of energy peaks of the audio signal. The same technique was extended to include rhythm and harmonicity for music detection while skin masks were used to detect the presence of faces. For the Search Task this group developed an interactive video retrieval system which used all 10 features identified earlier, three of which were the result of their own extractions, and the rest were donations from other groups. Twelve test users each ran the full 25 topics by formulating queries, browsing results and submitting results. The group ran two variations of their system, one which used the features plus the ASR transcript provided by LIMSI, and the other which used just the ASR transcript. All topic searches were limited to 4 minutes in total elapsed time.

5.4 Eurecom (France)

This group submitted runs under the feature extraction task. Their approach avoided complete decoding of the MPEG stream, basing decisions instead on the classification of the DCR macro-blocks — at some cost to the precision of the analysis. The work can be seen as an exploration of a "low-cost" baseline.

5.5 Fudan University (China)

Fudan University participated in the shot segmentation, feature extraction, and search tasks.

In the shot segmentation task, Fudan used most parts of their TREC-2001 shot segmentation system. The parameters used in the system were trained and adjusted based on the TREC-2001 video collection. According to the performance on TREC-2001 video collection, they selected the system parameters to generate the submissions. They added fade in/out detection to the system this year although the shot segmentation task did not include it. Evaluation showed that the system had a good balance between precision and recall. Comparing F-Value, the rank of the best result for all the changes, cut changes and gradual changes was 3, 3 and 9 (out of 54 systems). On gradual accuracy, frame-recall of the system was better than frame-precision. Compared with other submitted systems, their system was located at the middle in gradual accuracy.

In the feature extraction task, they developed a new video feature extraction system. It consisted of five sub-systems: outdoor / indoor detection, cityscape / landscape detection, face / people detection, text detection and speech / music / monologue detection. In each sub-system, a value calculated by whatever methods and features were used for ranking. Evaluation showed that the system worked well on these features: Cityscape, Landscape, Indoor and Music.

In the search task, Fudan submitted four runs. Considering the difficulty of search topics, they did not process all of the topics in each run. The whole architecture of the search system was almost the same as last year. However, there were some improvements in face recognition and object search. Fudan tried a fast manifold-based approach to face recognition in the TREC-2002 Search Task. This can be used when there are only few different images of a specific person and this process runs fast.

For each search topic, Fudan combined the similarities coming from different modules such as face recognition, text recognition, color histogram comparison, ASR text etc. In their submission, Sys1 only used the information returned by their own search modules. There was no ASR Text and Feature Extraction results used. However, feature extraction confidence was useful for some topics. So in the runs labeled Sys2 and Sys3, they combined feature extraction confidence into the searching. Sys2 used their own feature extraction results and Sys3 used the reference feature extraction results provided by IBM and MediaMill. In Sys4, they combined the ASR Results provided by LIMSI. NIST's evaluation showed that their searching system was not effective in several topics. In their future work, Fudan plans to pay more attention to image similarity calculation.

5.6 IBM Research, Almaden and T.J. Watson (US)

IBM participated in the shot boundary detection, feature extraction and search tasks. This large group explored several diverse methods for video analysis, indexing, and retrieval, which included automatic descriptor extraction, statistical modeling, and multimodal fusion. In the shot boundary detection task, they explored several methods for making SBD more robust to poor video quality. Some of the methods explored include using localized edge gradient histograms and comparing pairs of frames at greater temporal distances. In the feature detection task the IBM group explored several methods for automatic descriptor extraction and statistical modeling and made significant efforts to manually annotate the Feature Training and Validation collections. First, using the Feature Training collection, they built statistical models of the concepts, exploring a variety of descriptors including color histograms, wavelet texture, edge histograms, color correlograms, motion vectors, audio spectrum features, and so on. They also investigated different discriminant modeling methods (e.g., support vector machines). Once the individual statistical models were constructed, they explored different fusion methods for maximizing retrieval effectiveness on the Feature Validation collection. The resulting fused classifiers were then applied to the Feature Test collection. Overall, feature detection results were submitted for all ten feature classes.

For the search task the IBM group investigated both manual and interactive methods of searching, submitting four runs as follows: (1) Manual searching using content-based retrieval (CBR) without knowledge of the Search Test collection; (2) Manual searching using spoken document retrieval (SDR) based on automatic speech recognition results; (3) A combination of CBR and SDR in manual searching; (4) Interactive use of CBR and SDR;

5.7 Imperial College London (United Kingdom)

Imperial College London used a shot-boundary detection scheme based on a multi-timescale detection algorithm in which colour histogram differences were examined over a range of frames. At each frame they calculated a distance measure for each of a range of timescales, and made decisions on whether a cut or gradual change had occurred according to where coincident peaks occurred in these distance measures. For the search task, they took a representative key frame for each shot and derived a number of low-level features including illumination-invariant colour representations, text from ASR and convolution filters. Query images were tested for similarity to a shot in the test set using the k-nearest neighbours approach. A novel relevance feedback system was then employed to allow the user to modify the query and update the results.

5.8 Indiana University (US)

At Indiana University researchers have developed a system named ViewFinder for the purpose of providing access to video content for a project named the Cultural digital Library Indexing Our Heritage (CLIOH). They took this existing system, made notable modifications, and applied it to the interactive search task, submitting one interactive search run.

5.9 Lowlands Group (the Netherlands)

This group participated in the search task by evaluating a probabilistic model for the retrieval of multimodal documents. The model was based on Bayes decision theory and combined models for text based search with models for visual search. The textual model, applied to the LIMSI transcripts, was based on the language modeling approach to text retrieval. The visual model, a mixture of Gaussian densities, described keyframes selected from shots. Both models had been proven successful on media specific retrieval tasks. Their contribution was the combination of both techniques in a unified model, ranking shots on ASR-data and visual features simultaneously. To further improve the query, they experimented with query expansion by adding additional example images found using Google image search. While the expansion process needed human involvement, they hoped the results would identify potential benefits of automatic expansion techniques for video search.

5.10 The MediaMill Group (the Netherlands)

The MediaMill Group performed feature extraction by evaluating a system aimed at training models for semantic concepts on a specific collection by active learning. The system was geared to feature classification for specific collections, to exploit characteristics of domain and collection, and to allow for user definition of problem-specific semantic concepts. Using the i-Notation system, annotators provided learning examples to the system in an efficient way. For active learning (i.e. classifier feedback during an annotation session) as well as final classification, a Maximum Entropy classifier was used. Binning was applied to provide the mapping of numerical values to binary values necessary for Maximum Entropy. A fixed pool of sixty visual descriptors was used as input for the Maximum Entropy classifier for all eight visual TREC features, so that extension of the approach to any other visual feature is trivial.

5.11 Microsoft Research Asia (China)

This team participated in the shot boundary, features and search tasks. For shot boundary detection, the submission was based on the last year's work but concentrated on improving gradual transition (GT) detection. The main feature for SBD was frame difference, the total difference of the binwise histogram comparison between two consequent frames in the R, G and B channels. Shot boundaries were then determined according to a set of heuristic rules. For feature extraction, multiple key frames were extracted for each shot and feature extraction was performed on these images. For the indoor, outdoor, cityscape and landscape features, trained models were employed based on color moment and edge direction histograms, aggregated over all keyframes from a shot. Face detection from keyframes and text overlay also ran on the multiple keyframes from each shot. The audio feature extraction was based on a support vector machine classifier with inputs based on low-level audio analysis.

This group used the Q-Video video retrieval system in the search task. Manual searching was performed using a combination of Color Moment (CM), Dominant Color (DC), HSV Histogram (HSVH), Color Layout (CL), Edge Histogram (EH), Color Texture Moment (CTM), Kirsh Direction Density (KDD), Wavelet feature (WF) and Motion Texture (MT) with different distance metrics employed for different feature sets. For interactive searching, users browsed retrieved shots and their feedback, both positive and negative, was fed into an SVM-based learning procedure for each topic, making it a kind of learning-based relevance feedback.

5.12 National University of Singapore (Singapore)

This group took part in the shot boundary detection task and used an expanded version of their previous temporal multi-resolution analysis (TMRA) work by introducing a new feature vector based on motion, incorporating functions to detect flash and camera/object motion, and selecting automatic thresholds for noise elimination based on the type of video. The framework can be used to extract meaningful keyframes and provides a unified approach to detection of gradual transitions and cuts.

5.13 Prous Science (Spain)

This company submitted runs under the search task but details have not been provided in writing at the time of writing this summary report. An overview paper describing the approach taken by Prous Science may become available along with other video track site reports, at a later time.

5.14 The University of Oulu (Finland)

The MediaTeam research group participated in collaboration with VTT Technical Research Centre of Finland to do the feature extraction, manual and interactive search tasks. In the feature extraction task they participated in detecting people, cityscapes, landscapes, speech and instrumental sound. The visual features used were based on spatio-temporal correlation of oriented gradient edge directions. Features from the audio signal consisted of various statistical measurements from signal power and energy. Representative shots for each feature class were selected from the feature development set to guide the vision-based feature detection. This group's video browsing and retrieval system contains a multi-modal indexing structure to access video shots. It uses combinations of self-organizing feature maps and semantic filters in content-based topic queries. It also provides a novel way to navigate interactively through vast collection of video shots based on a lattice-shaped browsing view. The view combines temporal coherence with metric shot similarities.

5.15 RMIT University (Australia)

RMIT participated in the shot boundary detection task, where they used the techniques of query by example (QBE) and ranked results, both often used in content-based image retrieval (CBIR). Each frame in turn was considered as an example query on the image collection formed by the other frames within a moving window. Transitions were detected by monitoring the relative ranks of these frames in the results list.

5.16 University of Bremen (Germany)

This group submitted runs under the shot boundary detection and feature detection tasks.

The shot detection approach was based on histogram differences. It was divided into two steps - feature extraction and shot boundary detection. Firstly, the histogram differences were calculated for the entire video in real time. Secondly, shot boundaries were detected. The advantage of this approach was the possibility to set adaptive thresholds for the shot boundary detection considering all extracted features of the complete video sequence. The adaptive threshold was set to a percentage of the maximum of all calculated difference values of the video. In the case of gradual changes, often multiple shot boundaries were detected. Therefore multiple detected shot boundaries that followed each other within a short temporal interval were grouped together and a gradual change was detected beginning with the first and ending with the last shot boundary in the interval.

For the feature extraction task the group examined whether it was possible to classify indoor and outdoor shots by their color distribution. In order to analyze the color distribution, first order statistical features were used, which were extracted from the histograms of the three color channels (RGB) and the grey level histogram. The features calculated from each histogram were average, variance, and amount of peaks, normalized to an interval [0...1]. In order to classify the shots into indoor and outdoor shots, a feed forward neural net with backpropagation learning was trained. At the input layer the 12 statistical features mentioned above were presented. The output layer consisted of two neurons that take on values between 0 and 1 measuring the probability for the features indoors or outdoors to be present in the shot. Two hidden layers each with 20 neurons were initialized with random weights. In order to train the neural net, some videos from the feature development collection were chosen. The shots were classified manually to generate 323 training data sets, 178 for indoors and 145 for outdoors.

In order to classify the shots from the feature extraction test collection, a set of n key frames was extracted from each shot. Every k-th frame of a shot was used as a key frame, but in order to be more independent of inaccuracies during the shot detection and of gradual changes (e.g., wipes, fades, or dissolves) a number of frames around the shot boundaries were skipped. In order to classify a shot, the set of n key frames was presented to the neural net. For each of the two output neurons a list was obtained containing n values, one for each key frame. The median for each list was calculated to obtain the final probabilities for the shot to be indoors or outdoors. In order to measure the accuracy of the classification result, the difference between the median values of the indoors and the outdoors neuron was calculated. If the difference exceeded a threshold the shot was classified to contain the feature with the higher probability. The difference was also used for the ranking.

5.17 University of Maryland (US)

The University of Maryland led a team made up of researchers from INSA Lyon (France) and the Universities of Maryland (US) and Oulu (Finland), and participated in the text feature extraction task and the search task. For search they provided a weighted query mechanism by integrating 1) text (OCR and ASR) content using full text and n-grams through the MG system, 2) color correlogram indexing of shots and images reported last year in TREC, and 3) ranked versions of the extracted binary features. All of the features are normalized, and a variety of distance measures are used to index into the collection. The command line version of the interface allowed users to make various queries, store them and use weighted combinations to generate a compound query.

In their interactive search experiments, most users generated their initial manual queries with the command line interface, and then explored a ranked collection of clips with an interactive interface. The interactive interface treated each video clip as a visual object in a multi-dimensional space, and each "feature" of that clip was mapped to one dimension. The user could visualize in two dimensions by placing any two features on the horizontal and vertical axis. Additional dimensions could be visualized by adding attributes to each object. Color, for example, could be used to represent a third feature dimension, size a fourth and shape a fifth dimension. Dynamic range sliders were provided for all features.

6 Summing up and moving on

This overview of the TREC-2002 Video Track has provided basic information on the track structure, data, evaluation mechanisms and metrics used, and a snapshot of what most of the participants did in their experiments. 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 the final proceedings or under "Publications" on the trec.nist.gov website.

In 2003 the track will become an independent evaluation with a one- or two-day workshop (TRECVID 2003) immediately preceding TREC. The guidelines will be developed during the first quarter of 2003. The following are likely:

- using 120 hours of 1998 news video (MPEG-1) in 2003 and more of the same/similar in 2004
- continuing the three basic tasks: segmentation, feature extraction, search
- perhaps attempting detection of higher-level segments: stories, scenes
- keeping most of the features, but adding some appropriate to news
- striving for better system comparability in the search task
- creating more topics, perhaps 50, unbiased by detailed knowledge of the test collection
- significantly increasing the sizes of the search and especially the feature test collections.

The latest information about the TREC video retrieval evaluation efforts, past and present, is available from the track website at www-nlpir.nist.gov/projects/trecvid.

7 Authors' note

We are particularly grateful to Rick Prelinger and Niall O'Driscoll for their help with the Internet Archive data. We appreciate Jonathan Lasko's painstaking creation of the shot boundary truth data. The track would not have been possible without the software development work and general collaboration of Ramazan Taban, who has returned home to France and the job market. Our thanks to John Garofolo and Jose Joeman for their helpful suggestions on an earlier draft.

Finally, we would like to thank all the track participants and other contributors on the mailing list, and especially those groups who provided shot boundary and feature extraction output for use by others. These combined efforts made this running of the track possible. The spirit of the track was again a very positive one.

References

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

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

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

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

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

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

The Internet Archive Movie Archive home page. (2002). URL: http://www.archive.org/movies/.

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

Marchionini, G. (2001). *The Open Video Project home page.* URL: www.open-video.org.

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.