The TREC-2001 Video Track Report

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

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

April 18, 2002

1 Introduction

New in TREC-2001 was 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 built on publicly available video provided by the Open Video Project of the University of North Carolina at Chapel Hill under Gary Marchionini (Marchionini, 2001), the NIST Digital Video Library (Over, 2001), and stock shot video provided for TREC-2001 by the British Broadcasting Corporation (Richard Wright et al). The track used very nice work on shot boundary evaluation done as part of the ISIS Coordinated Research Project (AIM, 2001).

This paper is an introduction to the track framework — the tasks, data, and measures. For information about results, see the tables associated with the conference proceedings.

TREC research has remained true to its late twentieth century origins, concentrating on retrieval of text documents with only occasional excursions into other media: spoken documents and images of documents. Using TREC as an incubator, the Video Track has pushed into true multimedia territory with respect to formulation of search requests, analysis of multimedia material to be searched (video, audio, transcripts, text in video, music, natural sound, etc), combination of search strategies, and in some cases presentation of results to a human searcher.

The TREC video track had 12 participating groups, 5 from US, 2 from Asia and 5 from Europe. 11 hours of MPEG-1 data was collected and distributed as well as 74 topics or queries. What made these queries particularly interesting and challenging was that they were true multimedia queries as they all had video clips, images, or audio clips as part of the query, in addition to a text description. Participating groups used a variety of techniques to match these multimedia queries against the video dataset, some running fully automated techniques and others involving users in interactive search experiments.

As might be expected for the first running of such a track, the framework was a bit unorthodox by the standards of mature TREC tracks. Participating groups contributed significant amounts of work toward the creation of the track infrastructure. Search systems were called upon to handle a very wide variety of topic types. We hoped exploring more of the possible territory, though it decreased the likelihood of definitive outcomes in any one area this year, would still generate some interesting results and more importantly provide a good foundation for a more focused track in TREC-2002.

In TREC-2001, participating groups were invited to test their systems one or more of the following three tasks/evaluations.

- Shot boundary detection
- Search (fully automatic or interactive)
  - Using known-item topics or queries
  - Using general topics or queries

See the "Approaches" section for a list of the 12 participating groups and information on their systems. Details about each task follow here.

2 Shot boundary detection

Movies on film stock are composed of a series of still pictures (frames) which, when projected, the human brain smears together so we see 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. Software has been developed and evaluations of various methods against the same test collection have been published e.g., using 33 minutes total from 5 feature films (Aigrain & Joly, 1994); 3.8 hrs 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 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 8 television news, advertising, and series videos (Ruiloba, Joly, Marchand-Maillet, & Quénot, 1999).

2.1 Data

The shot boundary test collection for this year's TREC task comprises about half the videos in the overall collection so that each series is represented. The videos are mostly of a documentary nature but vary in their age, production style, and quality. There are 42 videos encoded in MPEG-1 with a total runtime of about 5.8 hours and a total size of 3.34 gigabytes.

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 first of next shot, no fade or combination

dissolve - the first shot fades out *while* the second fades in

fadeout/in - the first shot fades out, *then* the second fades in

other - everything not in the previous categories

The VirtualDub software (Lee, 2001) was used in the Microsoft Windows environment to view the videos and frame numbers. The VirtualDub website contains information about VirtualDub and the MPEG decoder it uses. Twenty of the videos (from the BBC stock shot collection) had no internal transitions and thus no shot boundaries. The collection used for evaluation of shot boundary determination contains 594179 frames and 3176 transitions with the following breakdown as to type (using the post-conference corrected reference data):

- 2066 — hard cuts (65%)
- 975 — dissolves (30.7%)
- 54 — fades to black and back (1.7%)
- 81 — other (2.6%)

The proportion of gradual transitions is about twice that reported by Boreczky and Rowe (1996) and Ford (1999). Gradual transitions are generally harder to recognize than abrupt ones. Table 1 lists the videos with title, source collection, file name, size in megabytes, and run time (mm:ss). Note that the reference data for the video "A new Horizon" (bor10) turned out to have been inadvertently truncated. Consequently, no results for it were ready until immediately after the TREC-2001 workshop.

2.2 Evaluation

Submissions were compared to the shot boundary reference data using a modified version of the protocol proposed for 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. The version used in TREC has the following features:

- A short gradual transition (less than 6 frames) was treated as a cut

Table 1: Shot Boundary Determination Test Collection

Shot Boundary Test Videos				
Title	*Source*	*File*	*Size (MB)*	*Run time (mm:ss)*
Challenge at Glen Canyon	OV	bor03	240.5	26:56
The Great Web of Water	OV	bor08	251.0	28:07
A new Horizon	OV	bor10	149.4	16:44
The Rio Grande - Ribbon of Life	OV	bor12	121.9	13:39
Lake Powell - Jewel of the Colorado	OV	bor17	247.2	27:41
NASA 25th Anniversary Show - Seg. 5	OV	anni005	66.9	6:19
NASA 25th Anniverary Show - Seg. 9	OV	anni009	72.4	6:50
Spaceworks - Episode 3	OV	nad28	262.7	29:26
Spaceworks - Episode 6	OV	nad31	260.1	29.08
Spaveworks - Episode 8	OV	nad33	247.1	27:40
A&S Reports Tape 4 - Report 260	OV	nad53	128.0	14:20
A&S Reports Tape 5 - Report 264	OV	nad57	63.4	7:06
Senses and Sensitivity - Lecture 3	OV	senses111	484.1	48:16
Aircraft Hangar Fires...	NIST	ahf1	90.2	9:00
Enhanced Aerial Lift Controller	NIST	eal1	92.3	9:00
Portsmouth Flexible Manufacturing Workstation	NIST	pfm1	84.1	8:15
25 BBC stock shot videos between 00:19 and 4:27 in length	BBC	---	353	31:43
Totals --->			3.342 GB	5.8 hrs.

- A submitted cut matched a reference cut if the latter fell entirely within the boundaries of the former after the former has been extended 5 frames on each end.
- Gradual transitions matched if the intersection was at least 0.333 of the longer and 0.499 of the shorter transition — the default values from the earlier ISIS evaluation scheme.

For the purposes of evaluation, the categories were divided into two:

- cuts - cuts
- graduals - dissolves, fades to black and back, and other

2.3 Measures

For continuity with earlier work, the following measures were calculated by NIST: inserted transition count, deleted transition count, correction rate, deletion rate, insertion rate, error rate, quality index, correction probability, recall, and precision. See Ruiloba et al. (1999) for details on the definitions of these measures.

2.4 Issues/Lessons

There were several unexpected issues that cropped up during the running and subsequent evaluation of the shot boundary determination task.

Varying frame numbering

Different MPEG-1 decoders produced slightly different frame numbering from the same video source file. This caused problems for evaluation of cuts since, initially, exact matches were required. A fixed shift of plus or minus 2 and then plus or minus 5 for an entire file was used until evidence was found that in some cases the shift of frame numbers varied within a file. The solution to this problem was eventually the algorithm described above, immediately under "Evaluation". The TREC video mailing list was quite active on this point and contributed to addressing the problem. The applicability of the 11-frame window to new data, is unknown and as an alternative for the future, a standard decoder or set of decoders could be mandated for determining frame numbers in the submission. Workshop participants generally felt this would be impractical for them.

Test collection available in advance

Although they did not know specifcially which files would be used, the shot boundary test collection was available to the participating groups long before the test began. Groups were reminded that systems to be tested could not have been trained on any of the test collection files — standard research practice anyway. It would however be preferable in future to use test video not generally available before the test.

Single reference

A second reference set was started but could not be completed in time. Finishing it would allow one to gauge the variability in system evaluation due to inter-annotator disagreements. For the final results we did check the shot boundary reference in cases where more than a couple systems told us there was a transition we did not have. This resulted in the addition of 20 transitions. We also completed the reference for the bor10.mpg file which had been inadvertently truncated.

3 The Search Tasks

The search tasks in the Video Track were extensions of their text-only analogues. The systems, some 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 shots

from the videos in the test collection which met the need.

In the case of the Video Track, the topics contained not only text but possibly examples (including video, audio, images) of what is needed. The topics expressed a very wide variety of needs for video clips: of a particular object or class of objects, of an activity/event or class of activities/events, of a particular person, of a kind of landscape, on a particular subject, using a particular camera technique, answering a factual question, etc. See Table 3 for an overview of the topics and their makeup.

The boundaries for units of retrieval to be identified - shots - were not predefined for all systems and each system made its own independent judgment of what frame sequences constituted a relevant shot. This had important consequences for evaluation.

The evaluation of video retrieval, whether for known-items or general searching, presents a larger, if not harder, set of problems than evaluations of text-only retrieval and we are not aware of any other large, open evaluation of content-based retrieval from digital video. Wide-spread use of video data, when it exists, is often limited by cost and intellectual property rights. Details about each of the tasks follow.

Although the track decided early on that it should work with more than text from audio, systems were allowed to use transcripts created by automatic speech recognition (ASR). Any group which did this had to submit a run without the ASR or one using only ASR — as a baseline. At least two groups used ASR.

Table 2: Additional video to be searched

Additional test videos				
Title	*Source*	*File*	*Size (MB)*	*Run time (mm:ss)*
The Colorado	OV	bor02	178.3	19:58
The Story of Hoover Dam	OV	bor07	246.1	27:24
Wetlands Regained	OV	bor09	126.5	14:01
Giant on the Bighorn	OV	bor11	125.4	14:03
Take Pride in America	OV	bor14	103.0	11:32
How Water Won the West	OV	bor19	100.8	11:17
NASA 25th Anniverary Show - Seg. 6	OV	anni006	97.6	9:13
NASA 25th Anniverary Show - Seg. 10	OV	anni010	184.8	17:27
Spaceworks - Episode 5	OV	nad30	266.1	29:48
Spaceworks - Episode 7a	OV	nad32	259.3	29.03
A&S Reports Tape 4 - Report 259	OV	nad52	129.7	14:31
A&S Reports Tape 4 - Report 262	OV	nad55	131.2	14:41
A&S Reports Tape 5 - Report 265	OV	nad58	68.8	7:42
Senses and Sensitivity - Lecture 4	OV	senses114	486.4	48:30
Telepresence Miscoscopy	NIST	dbe1	94.3	12:30
NIST in 5 Minutes and 41 Seconds	NIST	n5m1	65.9	5:41
A Decade of Business Excellence for America	NIST	ure1	85.1	8:50
A Uniquely Rewarding Experience	NIST	ydh1	128.1	12:23
25 BBC stock shot videos between 00:11 and 3:40 in length	BBC	---	301.8	27:08
Totals --->			2.96 GB	5.4 hrs.

3.1 Data to be searched

The test collection for the search task consisted of the collection used for the shot boundary determination task plus another six or so hours of similar video as listed in Table 2. The only manually created information that search systems were allowed to use was that which was already as part of the test collection, namely: the existing transcripts associated with the NIST files and the existing descriptions associated with the BBC material.

3.2 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 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.

For a number of practical reasons, the topics were created by the participants. This was not an easy or quick process. Each group was asked to formulate five or more topics they could imagine being used by someone searching a large video archive. Twelve sets of topics were submitted. NIST submitted topics as well, did some selection, and negotiated revisions. All the topics were pooled and all systems were expected to run on the union, if at all possible. The worst-case scenario in which each group found it's topics too easy and everyone else's topics too hard to learn something did not occur. Several groups found their own topics quite challenging and most groups had some success with topics other than their own.

All topics contained a text description of the user information need. Examples in other media were optional. There were indicators of the appropriate processing. And finally, if the need was conceived as a hunt for one or more known-items, then the list of known-items was included. Here is a summary of the topic layout:

- Text description of the information need
- Examples of what is needed
  - video clip illustrating what is needed

Table 3: Overview of topics

Topic #	Inter-active	Auto-matic	Text description of needed information/shot	Number of examples			Known items
				Video	Image	Audio	
1	Y		number of spikes on Statue of Liberty's crown		1		10
2		Y	liftoff of the Space Shuttle	4			
3		Y	vehicle traveling on the moon	1			2
4		Y	mountains as prominent scenery	1			8
5		Y	water skiing	1			5
6		Y	scenes with a yellow boat	1			4
7		Y	pink flower		1		1
8	Y	Y	the planet Jupiter		2		6
9		Y	people who are water skiing	1			
10		Y	swimming pools	1			
11		Y	people on the beach	1			
12		Y	surface of Mars	1		1	4
13		Y	speaker talking in front of the US flag	2		2	2
14	Y	Y	astronaut driving lunar rover over lunar surface	2			5
15	Y	Y	corn on the cob		1		4
16	Y	Y	deer with its antlers	1	1		2
17	Y	Y	airliner landing		1		3
18	Y	Y	John Deere tractor		2		2
19	Y	Y	lunar rover from Apollo missions		2		5
20	Y	Y	pictures of Ron Vaughn, President of Vaughncraft				1
21	Y	Y	pictures of Ronald Reagan speaking		3	3	1
22	Y	Y	pictures of Harry Hertz		2		5
23	Y	Y	images of Lou Gossett, Jr.		3		2
24	Y	Y	all other pictures of R. Lynn Bonderant	1			
25	Y	Y	scene from Star-Wars with R2D2 and 3CPO		2		1
26	Y	Y	given summary, find the full scene sequence	1			1
27	Y	Y	biplane flying over a field	1	1		4
28	Y	Y	sailing boat on a beach		1		2
29	Y	Y	hot air balloon in the sky		1		5
30	Y	Y	governmental buildings looking like Capitol	1			4
31	Y	Y	waterskier behind a speed boat		2		7
32	Y	Y	chopper landing			3	1
33	Y	Y	additional shots of white fort	1			1
34	Y	Y	Ronald Reagan reading speech about Space Shuttle		1		1
35	Y	Y	Where else does this person appear?	1			11
36	Y	Y	Where else does this person appear?	1			7
37	Y		other examples of rocket and shuttle launches	7		7	
38	Y		other examples of fires	4			
39	Y		other examples of airplanes taking off	3		3	
40		Y	all monologue shots	2			
41		Y	all shots with at least 8 people	2			
42		Y	all shots with David J. Nash	1			
43		Y	all shots with a specific landscape: grassland	1			
44		Y	all shots with specific camera technique: pan & tilt	1			
45		Y	other shots of cityscapes	1			
46		Y	other shots of sailing boats	1			
47		Y	clips that deal with floods	1			
48		Y	overhead zooming-in views of canyons...	8			
49		Y	other clips from the lecture showing/explaining example graphic	9			
50		Y	other examples of natural outdoors scenes with birds	8		10	
51		Y	other examples of splashing water in natural outdoors environment	7		10	
52	Y	Y	space shuttle on launch pad	6	2		
53	Y	Y	pictures of the Perseus high altitude plane		3		
54	Y	Y	clips showing Glen Canyon dam	1			
55	Y	Y	pictures of Hoover Dam	1			
56	Y	Y	clips of rockets taking off	2			
57	Y	Y	footage of explosions, blasting of hillsides	1			
58	Y	Y	additional shots of Lynn Bonderant	1			
59	Y	Y	launch of the Space Shuttle	3	1		
60	Y	Y	explosions in progress		1		60
61	Y	Y	environmental degradation	3	1	1	
62	Y	Y	how long has Baldrige Award existed				3
63		Y	clips of different interviewees	7			
64		Y	clips of different male interviewees	4		3	
65		Y	gradual shot changes	1			
66	Y	Y	clips talking about water projects	1			
67	Y	Y	segments of aircraft X-29	2	5		10
68	Y	Y	segment with a(n expert) person showing the X-29	2	5		1
69	Y	Y	logo of Northwest Airlines		5		2
70	Y	Y	identify the producer of each item				3
71	Y	Y	scenes with street traffic (cars, trucks, maybe people)		1		18
72	Y	Y	other similar clips containing a rocket launch	2			
73		Y	all shots with a specific landscape: lake	2			
74		Y	all shots with specific camera technique: zoom	1			

  - still image illustrating what is needed
  - audio illustrating what is needed
- Processing recommendations
  - indication of whether topic is for interactive processing
  - indication of whether topic is for automatic processing
- list of known-items, if any defined

If examples to illustrate the information need were included then these were to come from outside the test data. They could be taken from NIST or Open-Video material not part of the test collection or from other public domain sources. If the example came from the test collection, the topic's text description was to be such that using a video quotation from the test collection is plausible, e.g., "I want to find all the OTHER shots dealing with X." A search for a single shot could not be described with example video or images from the target shot.

3.3 Evaluation of known-item searches

The known-item search submissions were evaluated by NIST using a variation of the algorithm used in the shot boundary determination task. Matching a submitted item to a known-item defined with the topic was a function of the length of the known-item, the length of the submitted item, the length of the intersection, and two variables:

- KI coverage: minimum value for the ratio of the length of the intersection to the length of the known-item, i.e., how much of the known-item was captured by the submitted item
- RI coverage: minimum value for the ratio of the length of the intersection to length of the submitted result item, i.e., how much of the submitted result item was on target

The evaluation was run with four different settings of the two variables — as examples. In the absence of an application, a choice of particular settings would be arbitrary. The four settings reported to participants were the four combinations of 0.333 and 0.666. The pages at the back of the TREC-2001 proceedings report results where the length of the intersection must be at least 0.666 of the length of the known-item and at least 0.333 of the submitted item.

The performance of systems/runs can't be compared directly since they attempt different subsets of topics and may or may not include a human in the loop though we are dealing with rather small differences. It may be worth noting that the ranking of the systems/runs based on these values appear to be fairly stable across different match parameter settings as measured by Kendall's tau (see Table 4).

Table 4: Stability of known-item search system rankings as match parameter settings vary

Kendall's tau for recall-ranked systems by matching-parameter settings				
KI,RI settings	0.333, 0.333	0.333, 0.666	0.666, 0.333	0.666, 0.666
0.333, 0.333		0.923	0.881	0.814
0.333, 0.666			0.838	0.876
0.666, 0.333				0.890
0.666, 0.666				

Kendall's tau for precision-ranked systems by matching-parameter settings				
KI,RI settings	0.333, 0.333	0.333, 0.666	0.666, 0.333	0.666, 0.666
0.333, 0.333		0.957	0.914	0.876
0.333, 0.666			0.900	0.900
0.666, 0.333				0.942
0.666, 0.666				

3.4 Known-item measures

The measures calculated for the evaluation of known-item searching were precision and recall. It should be noted that a result set item could match more than one known-item and a known-item could match more than one result set item. In calculating precision, credit was given if a result set item matched at least one known-item. In calculating recall, credit was given for all known-items that a result item matched. The number of known-items varied from 1 to 60 with a mean of 5.63, so the upper bound on precision in a result set of 100 items was quite low.

3.5 Known-item issues/lessons

Evaluation of the known-item searches turned out to be more difficult than we anticipated. Because neither the known-items nor the result items were chosen from a predefined set of shot bounds or other video segments, a parameterized matching procedure was defined as described above. It is not yet clear if/how system performance across a range of parameter settings is most usefully reported and depicted. If retrieval and evaluation could be done in terms of a reasonable set of predefined segments, the matching problem might be avoided.

Table 5: Raw counts of video assessment (dis)agreements

Counts of assessor (dis)agreements by type		
	B: Relevant	B: Not relevant
A: Relevant	1524	587
A: Not relevant	553	4729

3.6 Evaluation of general searches

Submissions for the general search topics were evaluated by retired information analysts at NIST. They were instructed to familiarize themselves with the topic material and then judge each submitted clip relevant if it contained material which met the need expressed in the topic as they understood it, even if there was non-relevant material present. Otherwise they were told to judge the clip as not relevant. They used web-based software developed at NIST to allow them to (re)play the video, audio, and image examples included in the topic as well as the submitted clips.

We had time to get a second set of judgments of the submitted materials. The raw counts of the ways in which the pairs of assessments (dis)agree are as shown in Table 5.

There were 7393 pairs of judgments. Overall, the two assessors agreed 84.6% of the time. On average, if either one of the assessors said the item was relevant, the other agreed 72.8% of the time. On average, if either one of the assessors said the item was not relevant, the other agreed 89.2% of the time. This is as good or better than the agreement among assessors judging text documents as measured in TREC-2 and TREC-4.

3.7 General Search Measures

The measure calculated for the evaluation general searching was precision.

We also made an effort to calculate a partial recall score. Each result item that was judged relevant and came from a file covered by the shot boundary reference was compared to the shots defined by the shot boundary reference. A reference shot was marked as relevant if at least one relevant result item matched it. A result item matched if it overlapped with the reference shot and the overlap was at least one third of the result item and at least two thirds of the reference shot. A result item could match more than one reference shot.

Table 6: Raw counts of intra-assessor assessment (dis)agreements

Intra-assessor (dis)agreements					
Result item types by times judged	Total items of each type	Number of total agreements		Number of disagreements	Disagreements as percent of total items
		Rel	Not Rel		
1	3849	-----	-----	-----	-----
2	1633	564	1054	15	1%
3	91	29	59	3	3%
4	1	1	0	0	0%

Once the relevant reference shots for each topic has been identified, each submission was evaluated against this partial list of relevant shots. The same matching criteria as above were applied in deciding which result items matched relevant reference shots. The table at the back of these proceedings shows the results of this procedure.

3.8 General Search Issues/Lessons

No pooling

Some groups submitted runs from multiple related systems which returned identical shots. No attempt was made to remove these since, lacking predefined retrieval units, we did not expect to be able to pool results and so did not try. This means some shots were assessed more than once by the same assessor. This set could be looked at as a sort of "natural experiment" for information on within-assessor consistency.

Interpretation of topics

Questions from the assessors about how to interpret the topics raised important issues in multimedia topic formulation. Basically the problems had to do with the relationship between the text and non-textual parts of the topic. Often it was not clear that all of the example was exemplary, but there was no way to indicate, even to a human, what aspects of the example to emphasize or ignore.

4 Approaches in brief

The following are very short descriptions of the approaches taken by each participating research group. For detailed information the reader should consult the relevant system- specific paper in these proceedings.

- Carnegie Mellon University

  Search: with, and without, the Sphinx speech recognition system, both automatic and interactive searches; Minor changes to the Informedia system; Used colour histogram matching, texture, video OCR, face detection and speech recognition;

- CLIPS IMAG Grenoble (Fr)

  Shot boundary detection (SBD): where there is significant motion between adjacent frames, uses motion compensation based on optical flow, and a photo flash detector, and a dissolve detector;

- Dublin City University (Irl)

  SBD: some work on macroblock patterns but only on partial dataset;

  Search: interactive, to evaluate the effectiveness of 3 different keyframe browsers (timeline, slideshow, hierarchical), used 30 real users, each doing 12 topics using 3 browsers each;

- Fudan University (China)

  SBD: used frame differences based on luminance and colour histograms;

  Search: for 17 topics, calculated camera motion, face detection and recognition, video text and OCR, speaker recognition and clustering, speech recognition and speaker gender detection;

- Glasgow University (UK)

  SBD: Examining the frequency of occurrence of macroblock types on compressed files, technique not tuned to gradual transitions;

- IBM Groups Almaden and T.J. Watson (US)

  SBD: used the IBM CueVideo toolkit;

  Search: with, and without, speech recognition, automatic and interactive searching tasks; based on the semi-automatic construction of models for different kinds of scenes, events and objects - extensive experiments;

- Imperial College (UK)

  SBD: used colour histograms but by comparisons across a range of frame distances, instead of the usual adjacent frames;

- Johns Hopkins University (US)

  SBD: based on colour histogram and luminance;

  Search: treated video as a sequence of still images and used colour histograms and texture to match query images and topic video keyframes vs. video data keyframes; no processing of text or audio; no previous video experience;

- Lowlands Group (NL)

  Search: both automatic and interactive searching, used output from CMU speech processing plus recognition of video text via OCR, detector for the number of faces on-screen, camera motion (pan, tilt, zoom), scene detectors, and models of lazy, interactive users;

- Microsoft Research Asia (China)

  SBD: working on uncompressed video, 2 techniques for hard and for gradual shots, integrated together; very elaborate SBD technique;

- University of Maryland (US)

  SBD: based on examining macroblock and DCT coefficients;

  Search: temporal colour correlogram (a colour histogram with the spatio-temporal arrangement of colours considered) is used to automatically retrieve from video topic examples;

- University of North Texas (US)

  Search: did 13 of the general search topics; used a keyframe extractor and an image retrieval tool to match topics which had exemplar video or images;

5 Summing up and moving on

The track revealed that there are still a lot of issues to be addressed successfully when it comes to evaluating the performance of retrieval on digital video information and it was encouraging to see so much interest from the community who specialise in evaluation of interactive retrieval, in what was achieved in the video track.

Overall, the track was a great success with more participants than expected and the promise of even more groups next year. However the real impact of

the track was not in the measurement of the effectiveness of one approach to retrieval from digital video archives over another approach but was in the fact that we have now shown that there are several groups working in this area worldwide who have the capability and the systems to support real information retrieval on large volumes of digital video content. This year's TREC video track was a wonderful advertisement for what some current content-based video retrieval systems are capable of and of the potential we have for future development.

For next year it is hoped that we will be able to use a new dataset which will be greater in size, and more challenging in nature - perhaps as much as 100 hours if we can get such data. It is expected that we will repeat the searching task with a more focussed set of topics, though we will still use multimedia topic descriptions. We are also likely to have a variety of detection tasks such as the occurrence of faces, text, camera motion, speech and dialogue properties, etc. to be included in addition to the automatic detection of shot boundaries as was done this year. Finally, some participants may use MPEG-7 as an interchange format. All of the decisions on these, and other, topics will be made over the TREC Video mailing list in the coming months.

6 Authors' note

More information about the track is available from the track website at www-nlpir.nist.gov/projects/trecvid. The interaction (e.g., topics, submissions, and evaluation output) was based on XML for which DTDs are available on the website.

Finally, we would like to thank all the track participants and other contributors on the mailing list whose combined efforts made the first running of the track possible. The spirit of the track was been very positive. Special thanks to everyone who early on did the tedious work of watching the videos and making up candidate topics and more recently to Jan Baan et al at TNO for help in better addressing the varying frame numbering problem as deadlines loomed.

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.

AIM. (2001). *AIM home page in French.* URL: www-asim.lip6.fr/AIM.

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.

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.

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.

Over, P. (2001). *NIST Digital Video Collection home page.* URL: www-nlpir.nist.gov/projects/dv.

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.