TRECVID 2006 - An Overview

Paul Over[◇], Tzveta Ianeva[◇†], Wessel Kraaij[‡], and Alan F. Smeaton[○]

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

[○]Adaptive Information Cluster /
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland
asmeaton@computing.dcu.ie

[‡]TNO Information and Communication Technology
Delft, the Netherlands
wessel.kraaij@.tno.nl

[†]Departament d' Informàtica
Universitat de València
València, Spain
tzveta.ianeva@uv.es

March 21, 2007

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2006 represents the sixth running of a TREC-style video retrieval evaluation, the goal of which remains to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. Over time this effort should yield a better understanding of how systems can effectively accomplish such retrieval and how one can reliably benchmark their performance. TRECVID is funded by the Disruptive Technology Office (DTO) and the National Institute of Standards and Technology (NIST) in the United States.

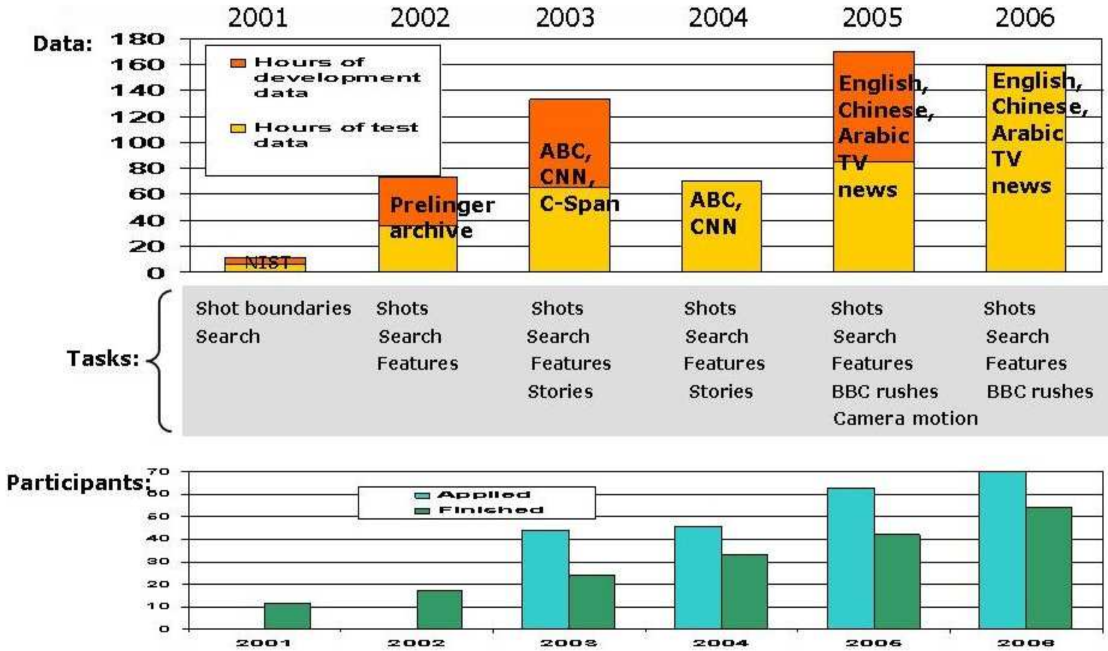
Fifty-four teams (twelve more than last year) from various research organizations — 19 from Asia, 19 from Europe, 13 from the Americas, 2 from Australia and 1 Asia/EU team — participated in one or more of four tasks: shot boundary determination, high-level feature extraction, search (fully automatic, manually assisted, or interactive) or pre-production video management. Results for the first 3 tasks were scored by NIST using manually created truth data. Complete manual annotation of the test set was used for shot boundary determination. Feature and search submissions were evaluated based on partial manual judgments of the pooled submissions. For the fourth exploratory task participants evaluated their own systems.

Test data for the search and feature tasks was about 150 hours (almost twice as large as last year) of broadcast news video in MPEG-1 format from US (NBC, CNN, MSNBC), Chinese (CCTV4, PHOENIX, NTDTV), and Arabic (LBC, HURRA) sources that had been collected in November 2004. The BBC Archive also provided 50 hours of "rushes" - pre-production travel video material with natural sound, errors, etc. - against which participants could experiment and try to demonstrate functionality useful in managing and mining such material.

This paper is an overview of the evaluation framework — the tasks, data, and measures — as well as to the results and the approaches taken by the participating groups. For detailed information about the approaches and results, the reader should see the various site reports in the Publications area of the TRECVID website.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Figure 1: Evolution of TRECVID



1.1 New in TRECVID 2006

While TRECVID 2006 continued to work primarily with broadcast news in Arabic, English, and Chinese, a significant portion of the test data came from programs not represented in the development data. This presents a test of how well feature detectors generalize and how searching broadcast TV news works on material from broadcasters other than those on which a search system has been trained.

Participants in the high-level feature task were required to submit results for 39 individual features defined by the DTO workshop on Large Scale Ontology for Multimedia (LSCOM) as the "LSCOM-lite" feature set, rather than some self-selected subset thereof. This was intended to promote the use of generic means for the training of feature detectors.

NIST planned to evaluate only 10 of the submitted features but by using a new measure of average precision based on sampling, was able to evaluate 20 of the 39 feature results submitted by each group.

The size of the feature and search test collection was nearly doubled over that used in 2005.

Participants were given access to two new sets of auxiliary data:

- the MediaMill Challenge data, which included 101 low-level features, estimated 101 MediaMill high-level concepts, and resulting rankings for the 2005 and 2006 test data

- the manual LSCOM annotations of the development data for 449 features

These were provided to participants in time for them to be used as part of their feature and/or search submissions.

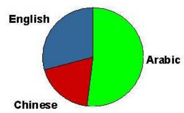
The BBC rushes presented special challenges (e.g., video material with mostly only natural sound, errors, lots of redundancy) and a special opportunity since such material is potentially valuable but currently inaccessible. The rushes differed in content from those use in 2005 - e.g., by containing more interviews.

There was an increase in the number of participants who completed at least one task - up to 54 from last year's 42. See Table 1 for a list of participants and the tasks they undertook. This represents another steady increase in the evolution of TRECVID in this 6th year of the annual cycle.

Figure 2: Hours of data by language

(Some test data programs not represented in the training data)

TRECVID Use		Lang	Source	Program	Hours	
tv5	tv6	Eng	NBC	NIGHTLYNEWS	9.0	
tv5	tv6	Eng	CNN	LIVEFROM	14.5	
	tv6	Eng	CNN	COOPER	8.3	
	tv6	Eng	MSN	NEWSLIVE	14.5	
						46.3
tv5	tv6	Chi	CCTV4	DAILY_NEWS	9.2	
	tv6	Chi	PHOENIX	GOODMORNCN	7.5	
	tv6	Chi	NTDTV	ECONFRNT	7.8	
	tv6	Chi	NTDTV	FOCUSINT	5.2	
						29.7
tv5	tv6	Ara	LBC	LBCNAHAR	35.3	
tv5	tv6	Ara	LBC	LBCNEWS	39.5	
	tv6	Ara	ALH	HURRA_NEWS	7.8	
						82.6
						158.6 hours



More Arabic than in 2005

2 Data

2.1 Video

The 2005 development *and* test data were made available to participants as development data for 2006. The total amount of news video available as test data in 2006 for the evaluated tasks was about 159 hours of video: 83 in Arabic, 30 in Chinese, 46 in English. The data were collected by the Linguistic Data Consortium (LDC) during November and December of 2005, digitized, and transcoded to MPEG-1.

A shot boundary test collection for 2006, comprising about 7.5 hours, was drawn at random from the total collection. It comprised 13 videos for a total size of about 4.64 gigabytes. The characteristics of this test collection are discussed below. The shot boundary determination test data were distributed by NIST on DVDs just prior to the test period start.

The total news collection minus the shot boundary test set was used as the test data for the high-level feature task as well as the search task. Both the development and test data were distributed on hard disk drives by the LDC.

2.2 Common shot reference, keyframes, ASR

The entire feature/search collection was automatically divided into shots at the Fraunhofer (Heinrich Hertz) Institute in Berlin. These shots served as the predefined units of evaluation for the feature extraction and search tasks. The feature/search test collection contained 259 files/videos and 79,484 reference shots (up from 45,765 in 2005).

A team at Dublin City University's Centre for Digital Video Processing extracted a keyframe for each reference shot and these were made available to participating groups.

BBN provided ASR/MT output for the Chinese and Arabic videos using the then current version of their latest MT research system, which is believed to reflect the state of the art at the time. The LDC provided ASR for the English videos.

2.3 Common feature annotation

In 2005 each of about 100 researchers from some two dozen participating groups annotated a subset of some 39 features in the development data using a tool developed by CMU or a new one from IBM. The total set of annotations was made available to all TRECVID 2006 participants — for use in training feature detectors and search systems.

In order to help isolate system development as a factor in system performance each feature extraction task submission, search task submission, or donation of extracted features declared its type as one of the following:

A - system trained only on common TRECVID development collection data, the common annotation of such data, and any truth data created at NIST for earlier topics and test data, which is publicly available. For example, common annotation of 2005 training data and NIST's manually created truth data for 2005 could in theory be used to train type A systems in 2006.

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

C - system is not of type A or B

Since by design there were multiple annotators for most of the common training data features but it was not at all clear how best to combine those sources of evidence, it seemed advisable to allow groups using the common annotation to choose a subset and still qualify as using type A training. This was the equivalent of adding new negative judgments. However, no new positive judgments could be added.

3 Shot boundary detection

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

The shot boundary task is included in TRECVID as an introductory problem, the output of which is needed for most higher-level tasks. Groups can work for their first time in TRECVID on this task, develop their infrastructure, and move on to more complicated tasks the next year, or they can take on the more complicated tasks in their first year, as some do. Information on the effectiveness of particular shot boundary detection systems is useful in selecting donated segmentations used for scoring other tasks.

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

3.1 Data

The shot boundary test videos contained a total of 597,043 frames and 3,785 shot transitions.

The reference data was created by a student at NIST whose task was to identify all transitions and assign each to one of the following categories:

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

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

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

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

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

The freely available software tool [1] VirtualDub was used to view the videos and frame numbers. The distribution of transition types was as follows:

- 1,844 — hard cuts (48.7%)
- 1,509 — dissolves (39.9%)
- 51 — fades to black and back (1.3%)
- 381 — other (10.1%)

This distribution has shifted toward more gradual transitions as Table 2 shows. In addition, short graduals — those with lengths of 1 to 5 frames, have increased as well (see Table 3). These are judged very strictly by the evaluation measures since they are cuts but without the 5-frame extension of boundaries to cover differences in decoders.

3.2 Evaluation and measures

Participating groups in this task were allowed up to 10 submissions and these were compared automatically to the shot boundary reference data. Each group determined different parameter settings for each run they submitted. Twenty-one groups submitted runs. The runs are evaluated in terms of how well they find all and only the true shot boundaries and how much clock time is required for their systems to do this.

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

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

Gradual transitions could only match gradual transitions and cuts match only cuts, except in the case of very short gradual transitions (5 frames or less), which, whether in the reference set or in a submission, were treated as cuts. We also expanded each

[1] The VirtualDub (Lee, 2001) website contains information about VirtualDub tool and the MPEG decoder it uses.

Table 2: Transition types

Search type	2003	2004	2005	2006
% Abrupt	70.7	57.5	60.8	48.7
% Dissolve	20.2	31.7	30.5	39.9
% Fade in/out	3.1	4.8	1.8	1.3
% Other	5.9	5.7	6.9	10.1

Table 3: Short graduals (1-5 frames)

	2003	2004	2005	2006
% of all transitions	2	10	14	24
% of all graduals	7	24	35	47
% of SG's = 1 frame	41	88	83	82

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. These measures evaluate the performance of gradual shot transitions in terms of the numbers of frames overlapping in the identified and the submitted gradual transitions and thus higher performance using these is more difficult to achieve than for non-frame precision and recall. Note that a system could be very good in detection and have poor accuracy, or it might miss a lot of transitions but still be very accurate on the ones it finds.

3.3 Approaches and Results

Participants continue to experiment with new variations on existing work but the best performances, being already very good, are hard to improve upon in terms of effectiveness or speed. This in spite of evidence that the 2006 data was harder than in 2005. Very good effectiveness continues to be combined with high speed. Accuracy results are depicted in Figures 3 - 5. Mean runtimes and mean runtimes versus accuracy are shown in Figures 6 - 9. See the individual group papers on the TRECVID website for details about various approaches.

Figure 3: Precision and recall for cuts

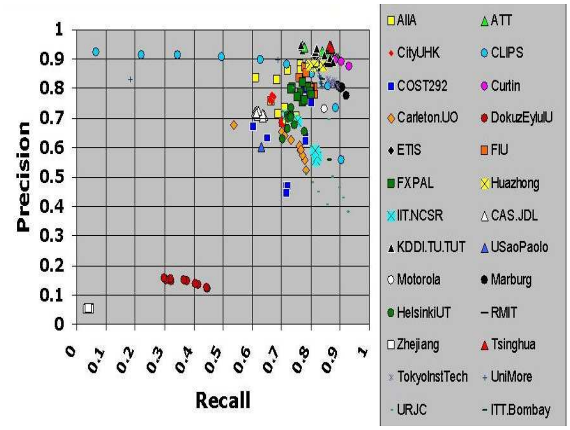


Figure 4: Precision and recall for gradual transitions

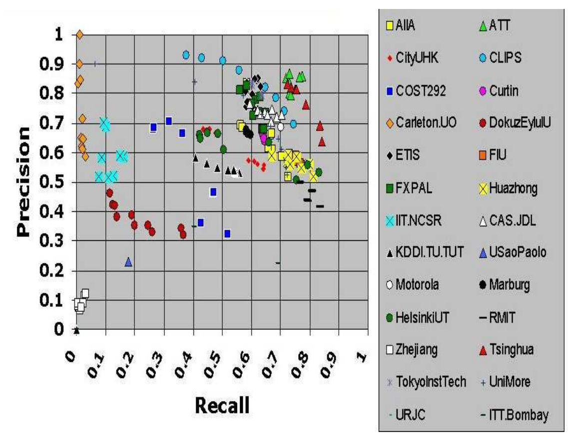


Figure 5: Frame-precision and -recall

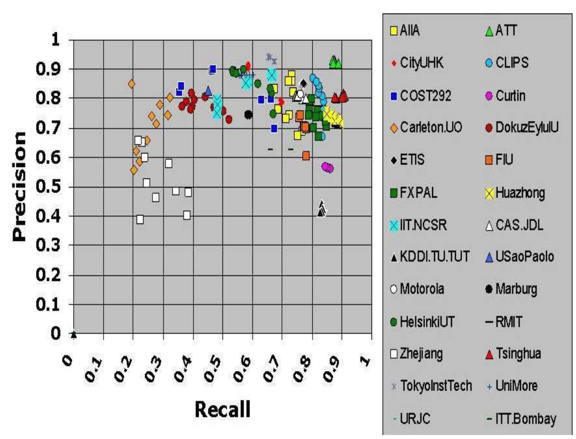


Figure 6: Mean SB runtimes

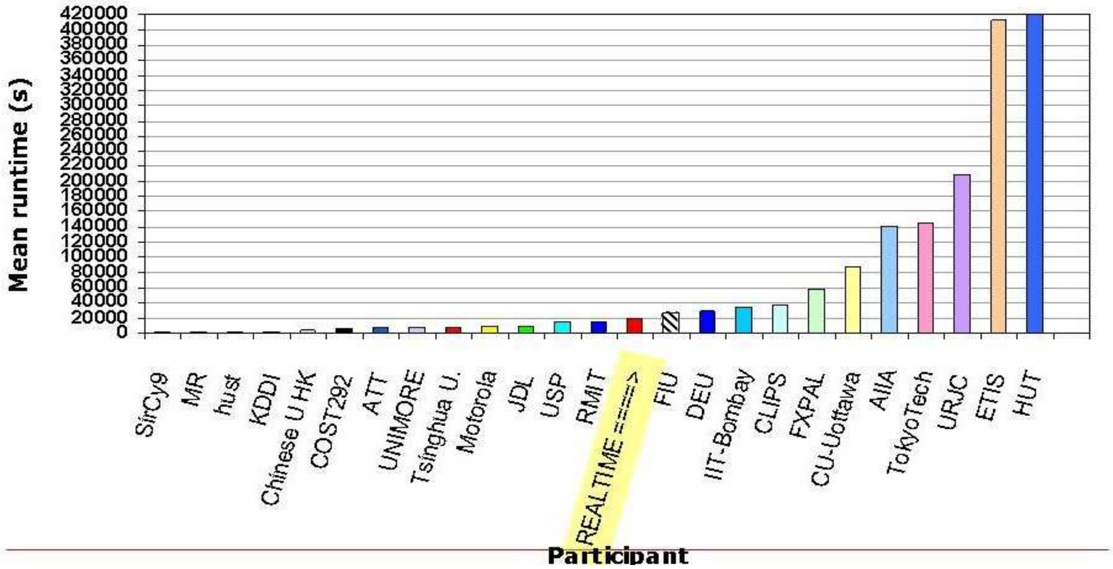


Figure 7: Mean SB runtimes better than realtime

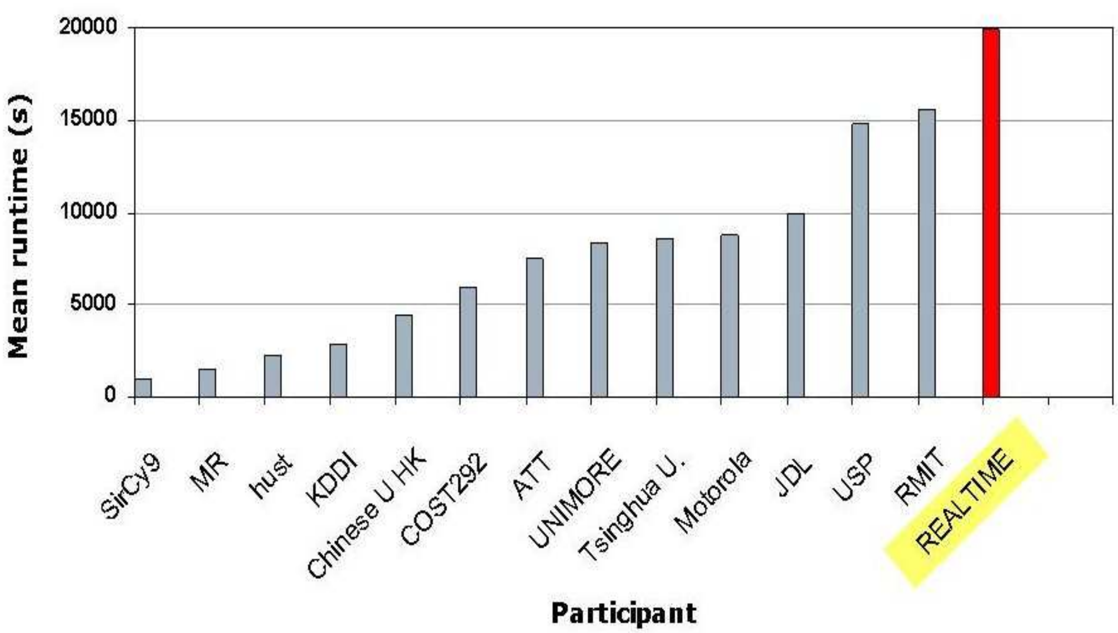


Figure 8: Mean SB runtimes (faster than realtime) versus effectiveness (mean F1 (harmonic mean of precision and recall) for cuts

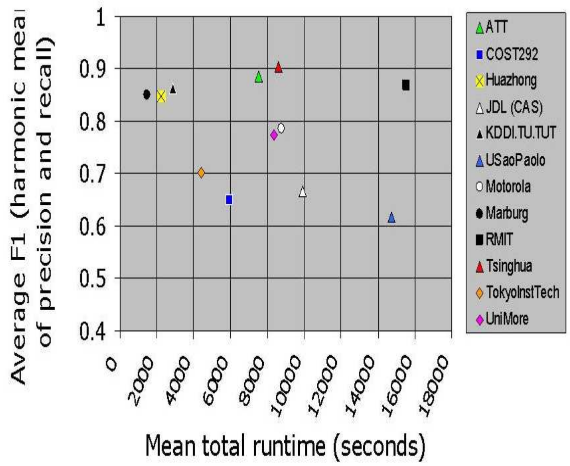
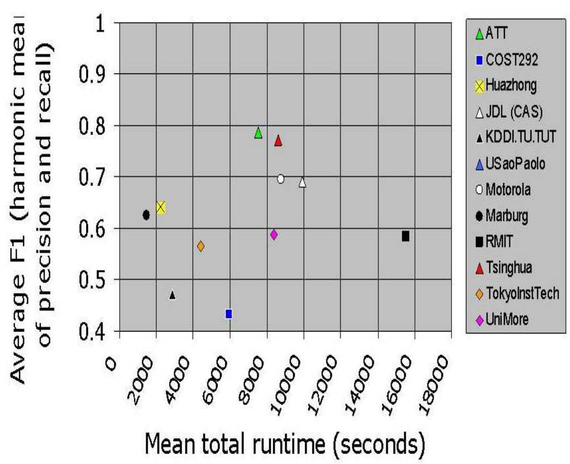


Figure 9: Mean SB runtimes (faster than realtime) versus effectiveness (mean F1 (harmonic mean of precision and recall) for graduals



4 High-level 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 would take on added importance if it could serve as a reusable, extensible basis for query formation and search. The feature extraction task has the following objectives:

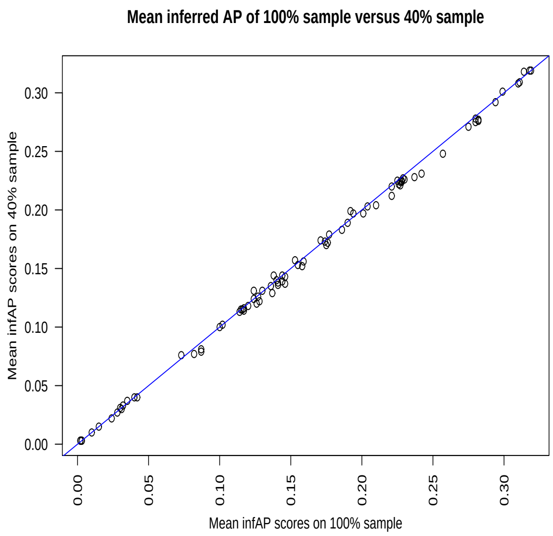
- to continue work on a benchmark for evaluating the effectiveness of detection methods for various semantic concepts
- to allow exchange of feature detection output for use in the TRECVID search test set prior to the search task results submission date, so that a greater number of participants could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

The feature extraction task was as follows. Given a standard set of shot boundaries for the feature extraction test collection and a list of feature definitions, participants were asked to return for each feature in the full set of features, at most the top 2,000 video shots from the standard set, ranked according to the probability 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 the entire preliminary set of 39 LSCOM-lite features, chosen to cover a variety of target types. In the past groups were allowed to choose from a subset of 10 features those they wished to develop detectors for. By increasing the number of detectors required, the aim was to promote generic methods for detector development.

The number of features to be evaluated was at first kept small (10) so as to be manageable in this iteration of TRECVID. However, recent work at Northeastern University (Yilmaz & Aslam, 2006) had resulted in methods for estimating standard system performance measures using relatively small samples of the usual judgment sets so that larger numbers of

Figure 10: Comparing MAP and mean infAP using 40% sample on 2005 data



features can be evaluated using the same amount of judging effort.

Using TRECVID 2005 high-level feature task results, an analysis of the new estimate for average precision - inferred average precision (infAP) - at various levels of judgment sampling (80%, 60%, 40%, and 20%) showed very good estimation of average precision in terms the of actual values of the measures. By design, infAP using a 100% sample is equal to average precision.

System rankings as measured by Kendall's tau (normalized number of pairwise swaps) vary little for better samples:

- 80% sample 0.986
- 60% sample 0.987
- 40% sample 0.970
- 20% sample 0.951

Furthermore, results of a randomization test showed no swaps in 2,053 significant pairwise differences ($p < 0.05$) found when measured using mean infAP versus mean average precision (MAP).

As a result, it was decided to use a 50% sample of the usual feature task judgment set, calculate inferred average precision instead of average precision, and evaluate 20 features from each group rather than the initially planned 10. Systems were compared in terms of the mean inferred average precision scores across the 20 features.

Features were defined in terms a human judge could understand. Some participating groups made their feature detection output available to participants in the search task which really helped in the search task and contributed to the collaborative nature of TRECVID.

The features to be detected were as follows and are numbered 1-39. Those evaluated are marked by an asterisk: [1*]Sports, [2]Entertainment, [3*]Weather, [4]Court, [5*]Office, [6*]Meeting, [7]Studio, [8]Outdoor, [9]Building, [10*]Desert, [11]Vegetation, [12*]Mountain, [13]Road, [14]Sky, [15]Snow, [16]Urban, [17*]Waterscape-Waterfront, [18]Crowd, [19]Face, [20]Person, [21]Government-Leader, [22*]Corporate-Leader, [23*]Police-Security, [24*]Military, [25]Prisoner, [26*]Animal, [27*]Computer-TV-screen, [28*]Flag-US, [29*]Airplane, [30*]Car, [31]Bus, [32*]Truck, [33]Boat-Ship, [34]Walking-Running, [35*]People-Marching, [36*]Explosion-Fire, [37]Natural-Disaster, [38*]Maps, [39*]Charts.

The full definitions provided to system developers and NIST assessors are listed in Appendix B.

4.1 Data

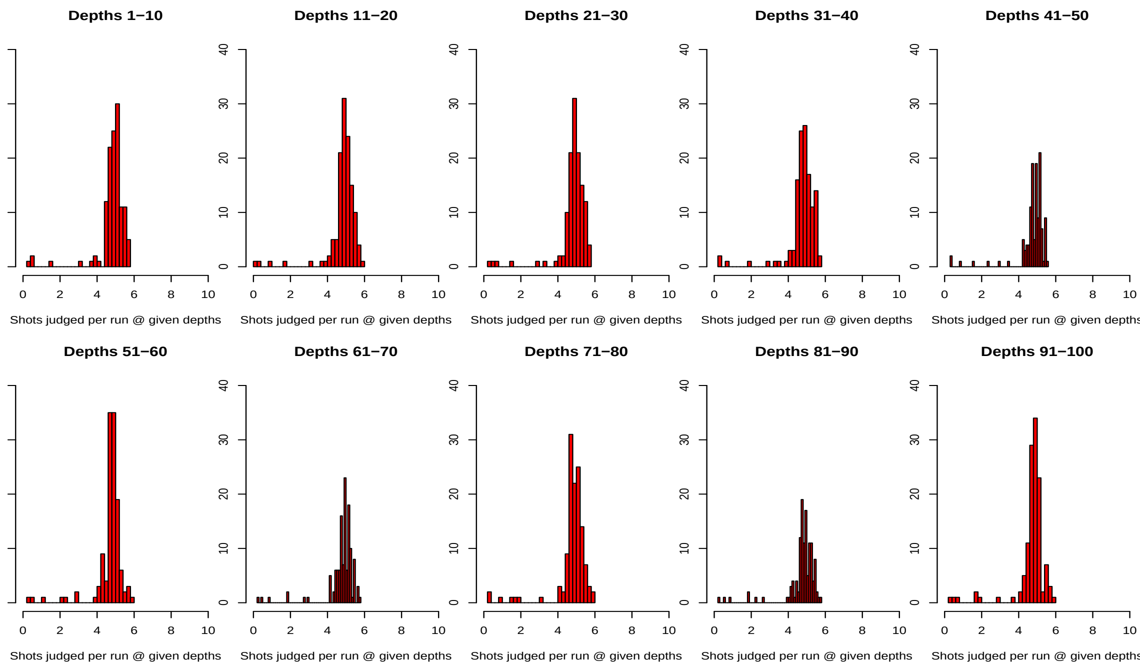
As mentioned above, the feature test collection contained 259 files/videos and 79,484 reference shots. Testing feature extraction and search on the same data offered the opportunity to assess the quality of features being used in search.

4.2 Evaluation

Each group was allowed to submit up to 6 runs and in fact 30 groups submitted a total of 125 runs.

For each feature, all submissions down to a depth of at least 100 (average 145, maximum 230) result items (shots) were pooled, removing duplicate shots, randomized and then sampled to yield a random 50% subset of shots to judge. Human judges (assessors) were presented with the pools - one assessor per feature - and they judged each shot by watching the associated video and listening to the audio. The maximum result set depth judged and pooling and judging information for each feature is listed in Table 4. In all, 66,769 shots were judged.

Figure 11: Distribution of shots judged per submission at various depths



A post-workshop analysis of the density of judgments for each submission indicates that runs received approximately equal amounts of judging as indicated in Figure 11. The few runs with significantly less than half of the top shots judged are those that contained only one or a couple shots due to processing problems or violation of the rule requiring results be submitted for all features.

4.3 Measures

The *trec_eval* software, a tool used in the main TREC activity since it started in 1991, was used to calculate recall, precision, inferred average precision, etc., for each result. Since all runs provided results for all evaluated features, runs can be compared in terms of the mean inferred average precision across all 20 evaluated features as well as "within feature".

4.4 Approaches in brief

The requirement that participants build detectors for all of the LSCOM-lite features did not reduce the number of participating groups; it grew over 2005. Support vector machines are still the dominant classifier with robust results. Good systems combined representations at multiple granularities (local, regional, global) with use of salient point representations gaining ground. Good systems combined various types of features (color, texture, shape, edges, acoustic, face, text). About a quarter of systems looked at more than just the keyframe for each shot. Many interesting multimodal and concept fusion experiments were carried out. Multi-concept fusion still seems of limited use, perhaps because there is not enough concepts that support each other in the relatively small set.

Figures 12 and 13 show the results of a detailed comparative analysis of feature extraction systems. Abbreviations for features in the tables are as follows: c: color, t: texture, s: shape, e: edges, a: acoustic, f: face, T: text. Blue cells mark the focus of work in a particular group. Yellow cells indicate groups that gave oral presentations at the workshop.

4.5 Results

Figure 14 presents a general picture of how scores are distributed for each feature. Median scores are with few exceptions quite low but scores vary widely

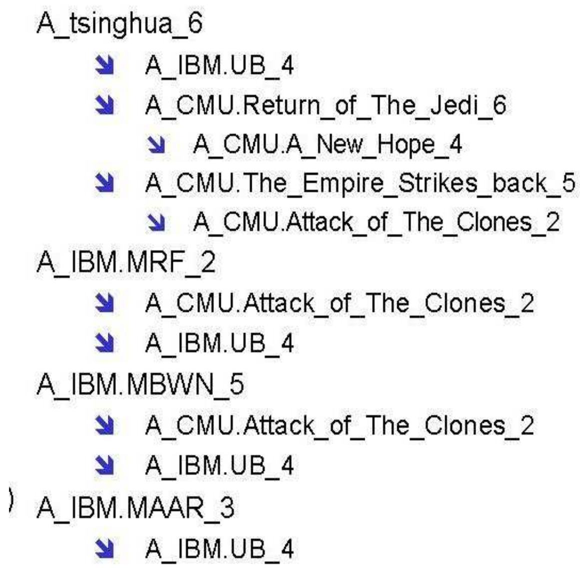
Figure 12: Approaches to high-level feature extraction

Cat.	run tag best run	best	repr. granularity	features	temporal analysis	classifier	multimodal fusion	multiconcept fusion	generic?
A	tsinghua	0.192	global,grid, segm. point	c,t,T,f	camera motion, motion act.	svm	weight-select, rankboost, stacked SVM	stackedSVM, rules	
A	IBM.MAAR	0.170	?	?		svm,?	?	?	
A	CMU.A_New_Hope	0.148	grid (5x5)+points	c,t,T		svm	logistic regression, early, late, borda	multi discr RF (chi2 selection)	
A	COL1	0.142	SIFT points/grid	c,t,T	EMD	svm	average fusion	boosting CRF (PMI selection)	
A	ucb_1best	0.122	points	a,e,T	shot context	svm	svm	svm	
A	UCF.CE.PROB	0.119		c,e		svm	average/product/KDE		
B	MM.bottom	0.117	global, grid, point			svm/ log reg / LD	early/ late fusion	svm	
A	KSpace-base	0.110	grid	c,t,e,T	camera motion	svm	bayesian (DS)		gen + specific
A	CityUHK1	0.106	points+grid	c,t	EMD	svm	average fusion		
A	MSRA_TRECVID	0.086	global, grid	c,t,s,f,T		SVM, KDE, manifold ranking, t-graph	weighted fusion, also looked at unlabeled data		
A	NTU	0.073							
B	PicSOM_F7	0.064	grid	c,t,T	motion act. average c,t, for shot	SOM	linear combination	handpicked negative concepts	
B	FXPAL-06Beta	0.059	MM	MM		svm		DRF / chi2	
B	OXVGG_A	0.053	points (sparse/dense)	c,e,f		SVM	Borda Count		

Figure 13: Approaches to high-level feature extraction (continued)

Cat.	run tag best run	best	repr. granularity	features	temporal analysis	classifier	multimodal fusion	multiconcept fusion	generic?
A	i2Rnus	0.040	grid	c,t,T	frame clustering, bigrams	SVM,LDF, GMM		cond prob	
A	NII_ISM_R1	0.033	overlapping grid	loc. bin. pat.		SVM			
B	clips.local-reuters-kernel-prod	0.031	local + global	c,t,T		SVM			
A	TokyoTech1	0.030							
A	ZJU	0.029	global	c,t,e,T,a		SVM	multimodal subspace correlation propag		
C	kddi.SiriusCy3	0.026	grid + points	s		Haar/KNN			not all
A	Bilkent1	0.021	grid	c,t,e,T		KNN			
B	TZI_Avg	0.021		c,T,e,f,a	every 20th frame	SVM	weighted average, prob. relax. labelling	cond prob	+specific
A	UEC_Common	0.006							
A	Glasgow.Sheffield01	0.005		T		tfidf			
A	LIP6.FuzzyDT	0.004	grid	p,c		fuzzy decision trees			
A	EUR01-SVM	0.002	points	c,t		SVM		NN	
A	FD_SCM_BN	0.001	points	c,t		GMM/SVM		cond. P	
A	icl.jhu_4	0.001	grid	c,t,T		likelihood ratio (HMM)	source adaptation		
C	UIowa06FE01	0.001							
A	COST292R1	0.000	points/grid/LSA	c,t / T		NN/Bayes			not all

Figure 16: Randomization test for significant differences in top 10 feature runs



for many features. Explanations of per-feature differences even for a single system require extensive data analysis. Figure 15 shows how close together the top ten runs for each feature tend to be. Figure 16 details the set of significant differences among the top 10 runs using a randomization test.

Do the systems retrieve results only or mainly from the English videos? No, as Figure 18 shows, the proportion of true shots for each language usually mirrors the proportion of video for the language. Did the systems tend to find true shots only or mainly in the video from programs with examples in the training data? No, as Figure 17 illustrates, the systems find true shots in both and the distribution seems to be proportional to the amount of data with and without examples in the training set.

While it is important to average over variation due to feature in order to describe system performance, it can be instructive to look at which groups found true shots for each feature that no other group found. Figure 19 displays this information.

4.6 Issues

There remain a number of issues for discussion concerning the feature detection task, as follows:

Figure 17: Feature results for known/new programs

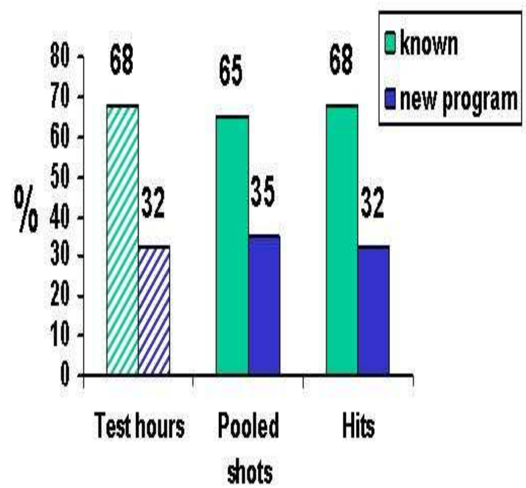


Figure 19: Unique true feature shots by team

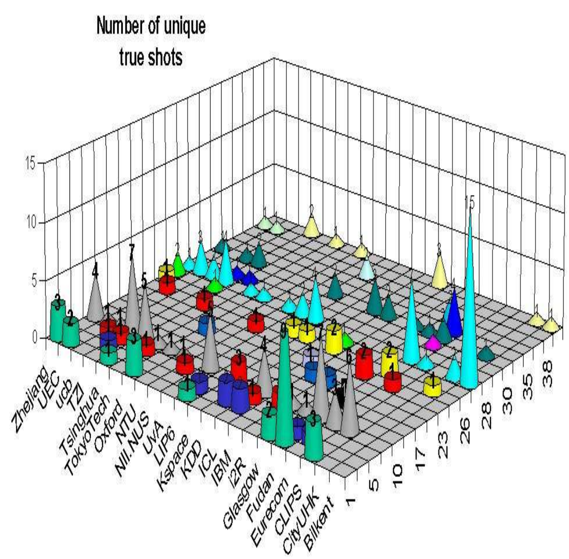


Figure 14: InfAP by feature

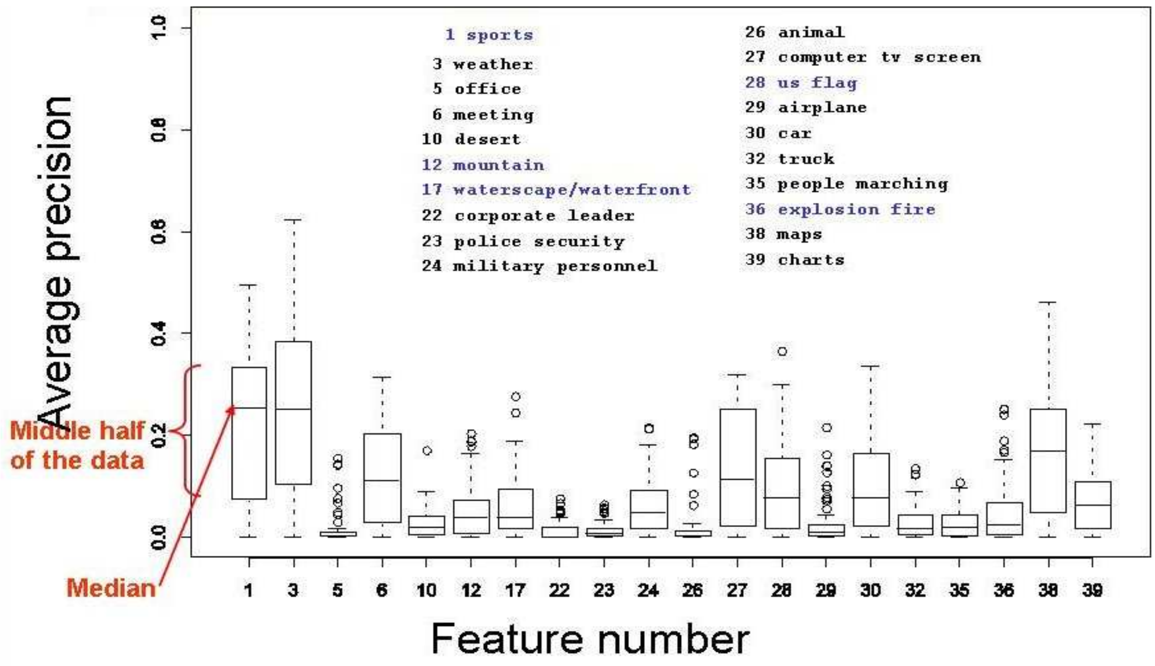


Figure 15: infAP by feature - top 10 runs

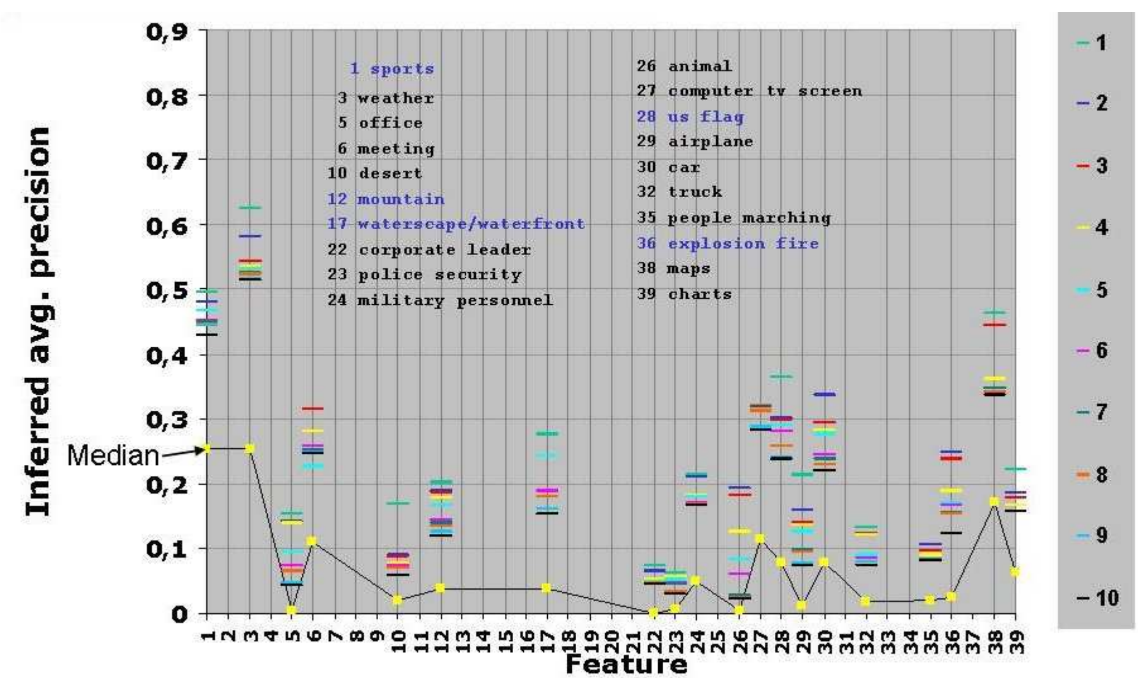
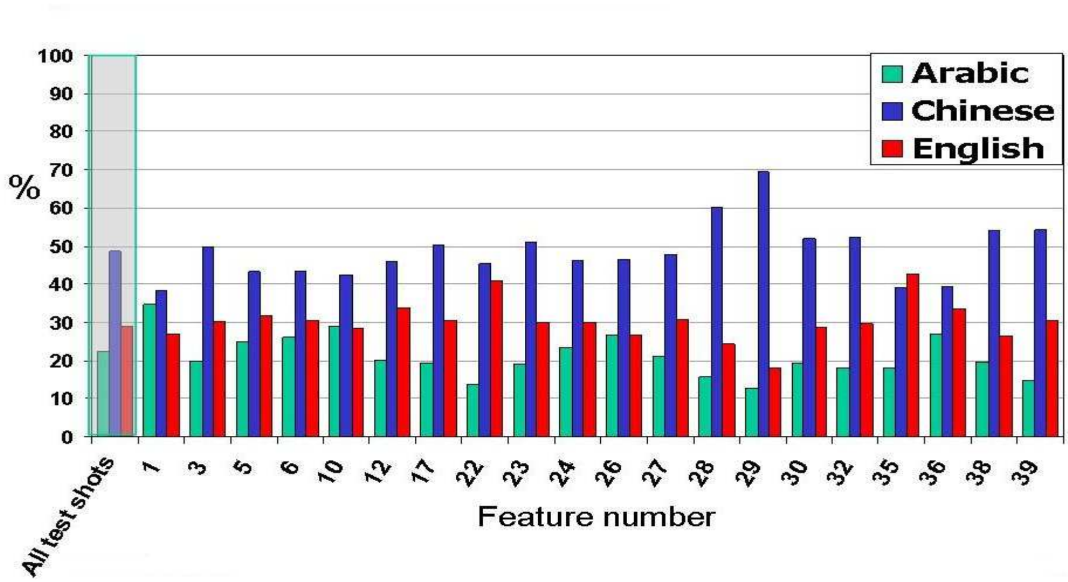


Figure 18: True features by language



1. The costs and benefits of sampling on top of pooling need further discussion and study. This year we decided to introduce a new sampling method for choosing submitted shots to be manually assessed in order to expand the number of features that could be judged. This is an example of yet another trade-off we make in benchmark evaluation campaigns.

2. The repetition of advertisement clips in the development and test data, which occurred in 2005 when the development and test data all came from the month of November 2004, was not the case in 2006 where the development data came from 2004 and the test data from 2005. In general the repetition of video material in commercials and in repeated news segments can increase the frequency of true shots for a feature and reduce the usefulness of the recall measure. The extent of this redundancy and its effect on the evaluation have yet to be examined systematically.

3. Finally, the issue of the interaction between the feature extraction and the search tasks still needs to be explored so that search can benefit even more from feature extraction.

5 Search

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

5.1 Interactive, manually assisted, and automatic search

As was mentioned earlier, three search modes were allowed, fully interactive, manually assisted, and fully automatic. A big problem in video searching is that topics are complex and designating the intended meaning and interrelationships between the various pieces — text, images, video clips, and audio clips — is a complex one and the examples of video, audio, etc. do not always represent the information need exclusively and exhaustively. Understanding what an image is of/about is famously complicated (Shatford, 1986).

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

One baseline run was required of every manual system — a run based only on the text from the provided English ASR/MT output and on the text of the topics. A baseline run was also required of every automatic system — a run based only on the text from the provided English ASR/MT output and on the text of the topics. The reason for the requirement for the baseline submissions is to help provide a basis for answering the question of how much (if any) using visual information helps over just using text in searching.

5.2 Topics

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

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

What was left was to work backward from the test collection with a number of goals in mind. Rather than attempt to create a representative sample, NIST has tried to get an approximately equal number of each of the basic types (generic/specific and person/thing/event), but in 2006 generic topics dominated over specific ones. Generic topics are more dependent from the visual information than the specific which usually score high on text based (baseline) search performance. 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 24 multimedia topics developed by NIST for the search task express the need for video (not just information) concerning people, things, events, etc. and combinations of the former. The topics were designed to reflect many of the various sorts of queries real users pose: requests for video with specific people or types of people, specific objects or instances of object types, specific activities or instances of activity (Enser & Sandom, 2002).

The topics were constructed based on a review of the test collection for relevant shots. The topic creation process was the same as in 2003 – designed to eliminate or reduce tuning of the topic text or examples to the test collection. Potential topic targets were identified while watching the test videos with the sound off. Non-text examples were chosen without reference to the relevant shots found. When more examples were found than were to be used, the subset used was chosen at random. The topics are listed in Appendix A. A rough classification of topic types for TRECVID 2006 based on Armitage & Enser, 1996, is provided in Table 7.

5.3 Evaluation

Groups were allowed to submit a total of up to 6 runs of any types in the search task. In fact 26 groups (up from 20 in 2005) submitted a total of 123 runs (up from 112) - 36 interactive runs, 11 manual ones, and 76 fully automatic ones. The trends seen in 2005 continue in 2006 with strong growth in the proportion of automatic runs, and at the same time a strong reduction in the proportion of manual, and a decrease in the proportion interactive runs, as shown in Table 6.

All submitted runs from each participating group contributed to the evaluation pools. For each topic, all submissions down to a depth of at least 70 (average 83, maximum 130) result items (shots) were pooled, duplicate shots were removed and randomized. Human judges (assessors) were presented with the pools — one assessor per topic — and they judged each shot by watching the associated video and listing to the audio. The maximum result set depth judged

Table 6: Search type statistics

Search type	2004	2005	2006
Fully automatic	17 %	38 %	62 %
Manually assisted	38 %	23 %	9 %
Interactive	45 %	39 %	29 %

and pooling and judging information for each feature is listed in Table 5 for details.

5.4 Measures

Once again, the *trec_eval* program was used to calculate recall, precision, average precision, etc.

5.5 Results

As in the past, scores vary greatly by topic. Figures 26 and 27 depict this variation. Figures 28 and 29 highlight the ability of some groups to find shots that meet the topic's need and were not found by any other group.

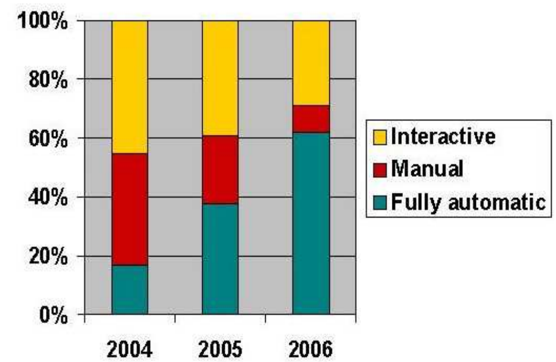
Precision/recall curves for the top 10 runs appear in Figure 21 (interactive), Figure 23 (manual) and Figure 24 (automatic). The top runs are difficult to distinguish in the graphs. Statistical analysis finds some differences not likely due to chance but also confirm the impression that many runs are performing at about the same level. Statistically significant differences among the top eight interactive runs are listed in Figure 22 according to a partial randomization test. The same test was run on the top eight automatic runs with the results as indicated in Figure 25. The symbols used in these Figures have the following interpretation:

- "*" indicates the run being compared against
- "=" means no difference in terms of MAP
- ">" means significantly better in terms of MAP

Automatic search

The top eight automatic runs were submitted by the IBM (TJW), CMU, NUS/I2R and Columbia University teams. The ranking in terms of MAP can be seen on Figure 25, but how significantly different they are? The answer, as depicted in the second column of Figure 25 is:

Figure 20: Runs by type



- Top two runs A_2_TJW_Qclass_4 and A_2_TJW_Qcomp_2 are significantly better than last three runs but are not significantly different from each other.
- Top third and fifth runs A_2_CMU_Taste_5 and B_2_i2Runs_1 are significantly better than last two runs.
- Top five runs can't be distinguished.

Interactive search

The top eight interactive runs were submitted by the CMU, MediaMill, University of Central Florida and FXPAL teams. A partial randomization test of the hypothesis that these search runs, whose effectiveness is measured by (mean) average precision, are significantly different - against the null hypothesis that the differences are due to chance was performed. The used significance level was fixed to $p < 0.05$. The ranking in terms of MAP can be seen on Figure 22, with the following interpretation of significant differences, as depicted in the second column of Figure 22:

- Top run A_2_CMU_See_1 is significantly better than all the other runs.
- Second top run B_2_UvA-MM_1 is significantly better than last five runs but can't be distinguished from third top run A_2_CMU_Hear_2.
- No significant difference in MAP among the remaining six runs A_2_CMU_Hear_2, ..., B_1_FXPAL4UNC_4.

Figure 26: MAP by topic

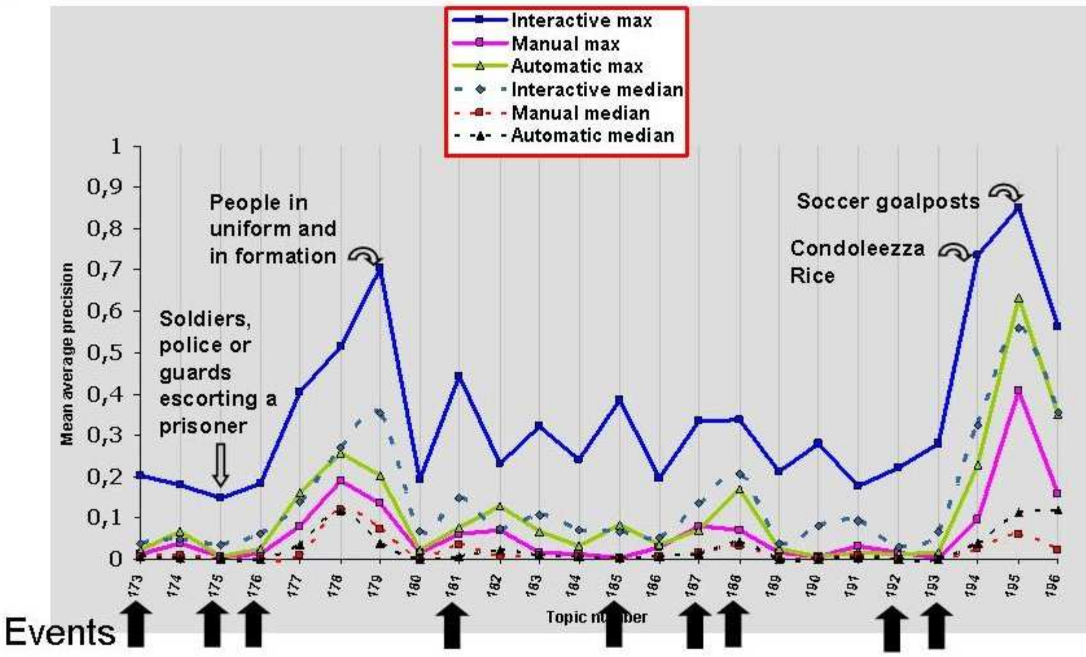


Figure 21: Top 10 interactive search runs

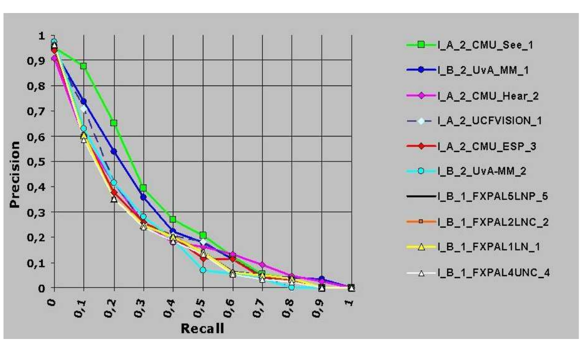


Figure 22: Randomization test on top 8 interactive search runs

	Run name	(MAP)
*	A_2_CMU_See_1	(0.303)
>	B_2_UvA-MM_1	(0.267)
>	A_2_CMU_Hear_2	(0.226)
>	A_2_UCFVISION_1	(0.225)
>	A_2_CMU_ESP_3	(0.216)
>	B_2_UvA-MM_2	(0.212)
>	B_1_FXPAL5LNP_5	(0.210)
>	B_1_FXPAL4UNC_4	(0.210)

A_2_CMU_See_1
- B_2_UvA-MM_1
  - A_2_UCFVISION_1
  - A_2_CMU_ESP_3
  - B_2_UvA-MM_2
  - B_1_FXPAL5LNP
  - B_1_FXPAL4UNC
- A_2_CMU_Hear_2

Figure 23: Top 10 manual search runs

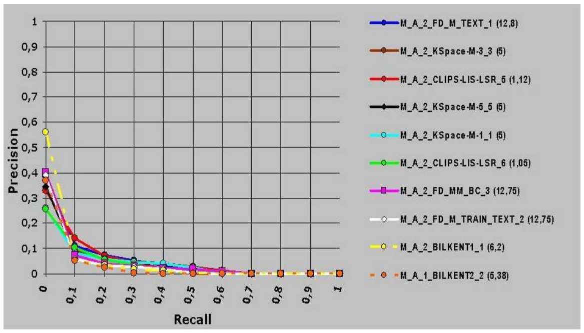


Figure 24: Top 10 automatic search runs

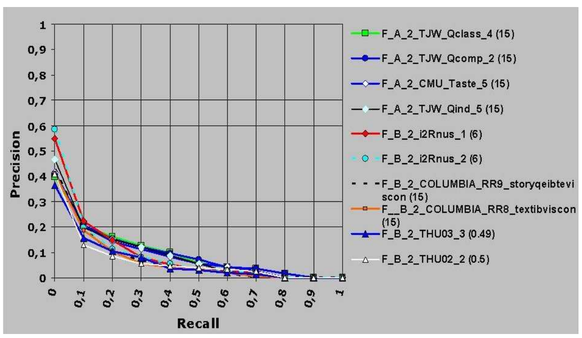


Figure 25: Randomization test on top 8 automatic search runs

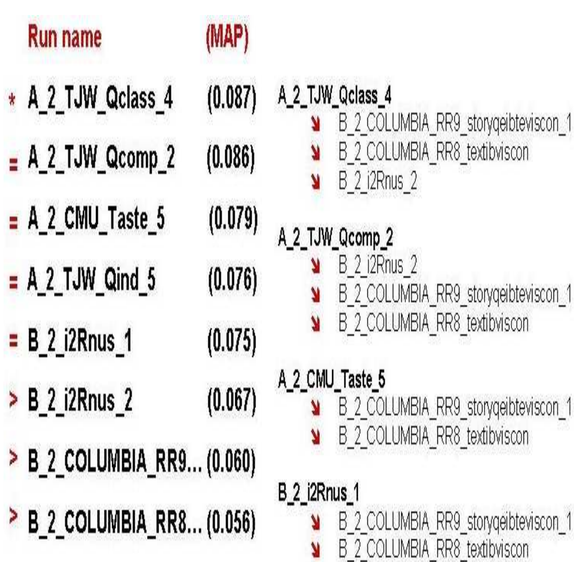


Figure 27: Topics sorted by median MAP

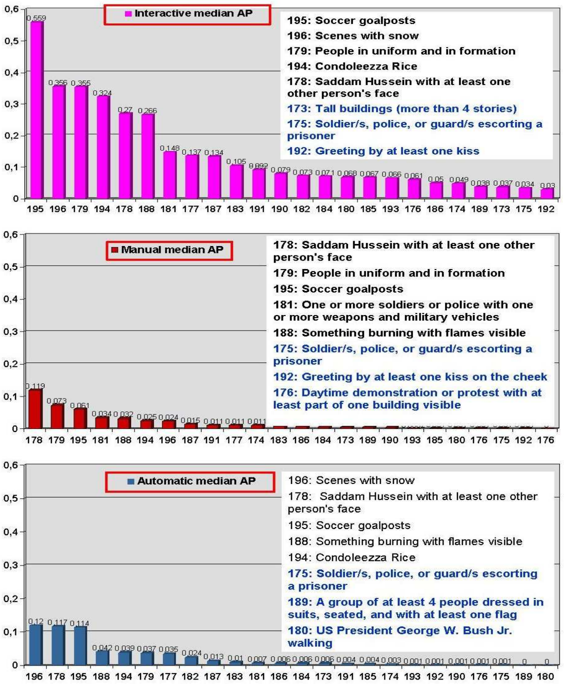


Figure 28: Unique relevant by team

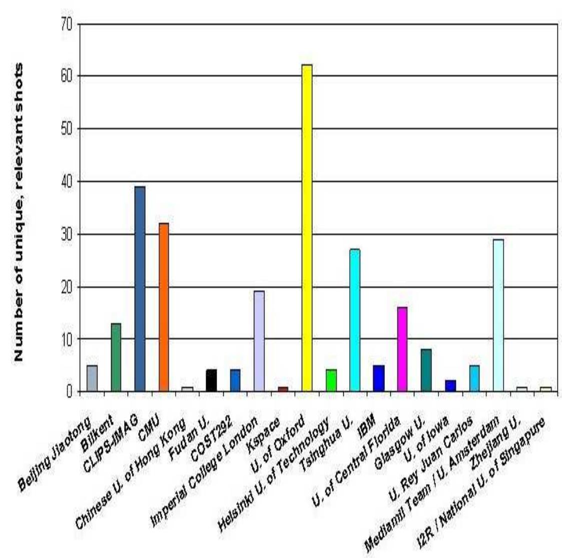
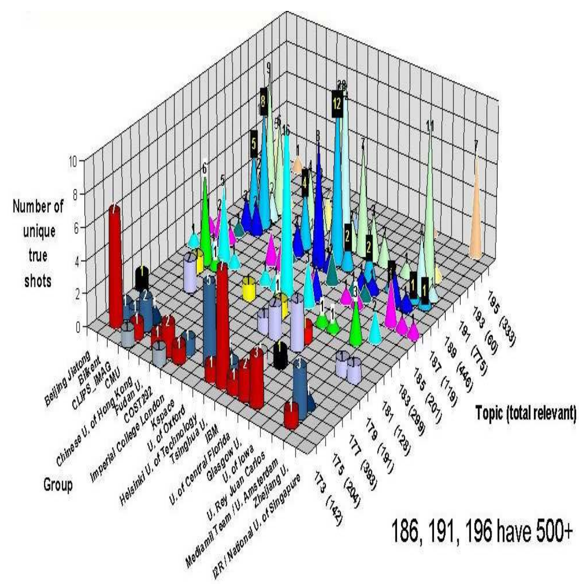


Figure 29: Unique relevant by team and topic



5.6 Approaches in brief

Given the large variation in approaches, browsing interfaces, mixture of features/concepts and learning algorithms we describe in the next sections only the systems that produced best runs, as shown in Figure 25 and Figure 22.

Automatic search

Figure 25 as stated earlier presents the top eight automatic runs based on their MAP score together with the significant differences among them.

The best IBM runs (first, second and fourth in Figure 25) were produced by ranked lists generated by speech-based, visual-based and semantic-based runs, and re-ranked using a model-based approach. The top two used query-dependent fusion, while the fourth used independent fusion. Improved speech-based retrieval includes automatic query refinement at the story level, with story boundaries automatically extracted and provided by Columbia University. Improved visual retrieval is based on a combination of two light-weight learning algorithms – modified k-Nearest Neighbor classifier and SVM with pseudo-negative sampling and bagging over five features (global color correlogram, color moments grid, global co-occurrence texture, wavelet texture grid and semantic model vectors). Model-based retrieval and re-ranking uses detectors for the 39 LSCOM-lite concepts. The query-dependent fusion approach is based on query-class categorization. In other words, the input query text is analyzed and the generated query features are assigned to query classes as Sports, Named-Person, Unnamed-Person, etc. The 24 TRECVID 2005 queries are used for training in order to learn a set of combination weights. Note that although the difference in MAP between the top 3 IBM runs is not statistically significant, the top two runs based on query-dependent fusion are significantly different from the last three runs, highlighting the performance of query-class-dependent fusion approaches using query class assignments.

The best CMU run (third in Figure 25) is produced by a relevance-based probabilistic retrieval model exploiting all modalities. The retrieval is based on the "ranking logistic regression" algorithm. Query information is incorporated into the probabilistic retrieval model by assigning the query to one out of five defined query types: Named person, Named object, General object, Sports, and Scene. Note that statistically this run has similar performance as the top two but is significantly better than the last two runs.

The best NUS/I2R runs (fifth and sixth in Figure 25) are produced by retrieval carried out based on the query-content. The query analysis is enhanced by including *query-HLF* and *query-event* contents in addition to the previously used *keywords, query-type* and *query-class*. The query-HLF measures the importance of a HLF with respect to a query, while the query-event links the query to possible event groups. The retrieval framework includes the detection of near duplicate keyframes (NDK). Similar keyframes are grouped together and Scale-invariant feature transform (SIFT)-based image matching is applied within the cluster for NDK detection. During retrieval relevant segments are obtained in 3 stages: 1) pseudo story-level retrieval, 2) multimodal shot-level re-ranking and 3) pseudo relevance feedback based on the top results returned. Pseudo story segmentation is obtained by the use of the story boundaries provided by Columbia University, followed by a second level segmentation by the use of anchor-person shots. Shots within the pseudo story segments are re-ranked based on the induced query-class, query-HLF, and the NDK.

The Columbia University team explored the potential of using a large set of automatically detected concepts for improving automatic search. Their runs rely on the combination of concept search results with a number of other search components. Results are query-class-dependently fused, with weights varying depending upon the type of query. The largest improvement (30% in comparison to the text story baseline) of any individual method is produced by the use of concept detectors in concept-based search. The best Columbia University runs (last two in Figure 25) are produced by a query class dependent combination of concept-based search, visual-example search and: information bottleneck (IB) re-ranked story-based search (in eighth top run), IB re-ranked story-based search with query expansion fused with example-based text search (in seventh top run). The text example-based search uses the video examples for a given topic as positive examples and negative examples are pseudo-sampled. SVM classifiers are trained using the examples with associated tf-idf features. The concept-based search uses SVM automatic detectors for 374 of the 449 concepts included in the LSCOM annotations.SVMs are built over three visual features: color moments on a $5X5$ grid, Gabore textures over the entire image, and an edge direction histogram. The same features are used in the representation of query videos and images in the visual example-based search, where test images are scored based on their Euclidean distance from the query images. The query-class dependent fusion uses five query classes: "named entity", "sports", "concept", "named entity with concept", and "general".

All automatic runs discussed above yielded relative improvement of 85% over text-only baseline runs.

Interactive search

Figure 22 presents the top eight interactive runs based on their MAP score, together with the significant differences among them.

The CMU interactive search system uses only the LSCOM-Lite concepts and their efforts were toward improving interface efficiency and exploring the effects of varying interfaces utilizing output of the automated search runs. An impressive number of shots were reviewed within the allowed 15 minute due to the improved efficiency of the Informedia interface. The CMU runs (first, third and fifth in Figure 22) all use as a starting point for interactive search the output of the ranked shots from the CMU fully automatic search runs. The fifth top run (... _ESP_...) implements video retrieval using manual browsing with variable, resizable pages on one set of images derived automatically, which gives the user more control over the display. It also takes into account the trade-offs between serial and parallel presentation of thumbnail imagery to assess shot relevance for a given topic. This run also uses the following strategy; when a shot in the ranked list of queries is marked as relevant by the user, neighbors of this shot are inserted at the top of the shots to be presented to the user. The third top run (... _Hear_ ...) uses the Informedia storyboard interface, working only with the ranked shot output from the automatic search and no query functionality. The benefits of additional query capabilities ("query-by-text", "query-by-image-example", "query-by-LSCOM-lite-concept") used in the top run (... _See_...) are confirmed by its statistically out-performance from all other runs.

The MediaMill interactive search system implements a *query by object matching* algorithm that uses difference of Gaussians detector and SIFT descriptor. Using a lexicon of 491 learned concept detectors, the MediaMill engine finds the most appropriate concept detector given the topic using an ontology. Two ways for displaying video threads (query result, visual, semantic, top-rank, textual and time threads) are supported. The CrossBrowser display mode shows two fixed directions, the query result and time threads, while the second multi-dimensional RotorBrowser shows a variable number of directions. The best MediaMill runs (third and sixth in Figure 22) are produced with expert users using the MediaMill search engine with CrossBrowser and RotorBrowser respectively. The out-performance of the CrossBrowser run over the RotorBrowser run is partly due to the fact that the largest portion of results is generated from the initial query results and the time thread, and these threads are the only ones that are available.

The University of Central Florida search system uses an improved on-line video retrieval system called PEGAS. Two ways of search are supported: text search using any known words, and search by known image/video example. Given the query, the search engine retrieves the relevant shots from the feature index system. Relevance feedback using visual features is implemented by global matching (via color correlogram, $5X5$ grid color moments and edge histogram and L1 distance measure for image-to-image similarity) and region-based refinement (via mean shift segmentation algorithm and EMD distance measure for image-to-image similarity). The Lucene full-text index is adopted to index the ASR information. The best UCF run (fifth in Figure 22) involves interactive search using text and visual information. A temporal

"K-neares neighbor" method is used as final step in the relevant feedback.

The FXPAL team used six different methods for interactive search: two text-only, three combination of text and concept-based ranking, and one combination of text and concept-based ranking using only positive concept examples. Concept-enhanced systems outperformed the text-only systems. The interactive system interface displays query results as a list of story thumbnails, sized in proportion to their query relevance. Basic units of retrieval during queries are the story segments built by text-based latent semantic analysis (LSA) of the ASR transcripts. Data pre-processing includes the production of text indices for both shot-level and story-level segmentation. Color correlograms are pre-computed for each shot thumbnail image. The search engine supports exact keywords text search (using a Lucene back end and story ranking based on the tfidf keywords values), LSA-based text search (using the cosine similarity function), and a combination of the two with averaged combined score and image search. An overall score formed by the combination of the documents scores from the text search and the document scores from the image similarity (based on the color correlograms) is used to sort the results. Additional concept-based search is supported using the 101 concepts provided by the MediaMill team. The best FXPAL runs (last two in Figure 22) are produced by combined Lucene/LSA text query and concept-based ranking with bracketing. Bracketing implies the inclusion in the results of the shots immediately adjacent to all shots marked relevant by the user, right after the user-selected shots and before the highest ranked unjudged shots.

5.7 Issues and Observations

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

During TRECVID 2006 top performances on all search types are down in comparison with TRECVID 2005. It is hard to tell whether this is because the the TRECVID 2006 test collection is twice as big, there are half as many relevant shots, topics were harder, etc.

Manual runs no longer outperform automatic ones. Is this because there were so few manual submissions or becauase of the higly improved automatic retrieval carried out based on the query-content?

Many of the best performing retrieval approaches use story-level retrieval, story ranking based on tf-idf keywords values, query-class dependent fusion by query-class categorization and automatically detected concepts. Color correlogram and color moments grid were were the most used visual features. Interactive search systems use impressively efficient user interfaces and very often bracketing (including in the results shots neigboring the relevant ones.)

6 BBC rushes management

Rushes are the raw video material used to produce a video. Twenty to forty times as much material may be shot as actually becomes part of the finished product. Rushes usually have only natural sound. Actors are only sometimes present. Rushes contain many frames or sequences of frames that are highly repetitive, e.g., many takes of the same scene re-done due to errors (e.g. an actor gets his lines wrong, a plane flies over, etc.), long segments in which the camera is fixed on a given scene or barely moving, etc. A significant part of the material might qualify as stock footage - reusable shots of people, objects, events, locations. Rushes are potentially very valuable but are largely unexploited because only the original production team knows what the rushes contain and access is generally very limited, e.g., indexing by program, department, name, date (Wright, 2005).

6.1 Data and task definition

The BBC Archive provided about 50 hours of rushes shot for BBC programming along with some metadata. The training and test sets are composed of 49 and 48 videos respectively. TRECVID participants were invited to 1) build a system to help a person, unfamiliar with the rushes to browse, search, classify, summarize, etc. the material in the archive, and 2) devise their own way of evaluating such a system's effectiveness and usability.

Twelve groups worked on the rushes task and submitted notebook papers describing their efforts but only a few groups provided actual evaluation. These groups' approaches are briefly described next.

6.2 Approaches in brief

The **Chinese Academy of Sciences** proposed a three step system:

- *Structuralization of rushes video.* Shot-boundary detection and key frame extraction using an unsupervised clustering method that extracts fewer keyframes for low-activity shots. Scene boundary detection using the keyframes by constructing a weighted undirected shot similarity graph (SSG) and transforming the shot clustering problem into a graph partitioning problem.

- *Redundancy detection and semantic feature extraction.* Three types of redundant shots – color-bar shot, black or gray background shot and very-short shot (less than 10 frames) – are detected by extracting their uniform visual features as template. Detection of repetitive shots is based on spatiotemporal slices instead of key frames. Once redundant shots are removed the following concepts are detected: face, interview, person, crowd, waterbody, building and outdoors.

- *Interactive interface.* Folder browsing, media file, playing-back, hierarchical browsing and a storyboard subwindows compose the interface. Concept legend and a color bar under each key frame are used to display whether the corresponding key frame is redundant, repetitive, or contains the concepts.

For evaluation purposes standard precision and recall measures are used. Face detection is achieved with precision of 90.8% (334/368) and recall of 84.8% (334/394). Person detection results are precision of 86.7% (39/45), recall of 95.1% (39/41). Interview detection is seen as a high level semantic concept containing both face and speech information. The difficulty of the task is represented in the results, best precision of 84.2% (223/265) and recall of 77.2% (223/289) are worse than face/person detections and achieved when using an intersection fusion method of face detection and audio classification techniques. The average precision for redundant shot detection is 99.05% and average recall 100.00%. As expected repetitive shots are harder to detect, average precision of 84.0% and average recall of 77.0%.

The **Curtin University of Technology** summarized the rushes data by extracting a set of frames to represent a scene in the original clip. The approach is based on the assumption that a good summary should be - *Concise* (up to 10 frames for scene representation); *Stable* (without frames containing high motion, people moving in front of the camera, intermedia camera transitions); *Help the user identify* characters, events, locations etc.

The implementation is based on Data-Preparation - Shot segmentation by a method of applying an adaptive-threshold on the discontinuity curve. Keyframes extraction based on visual dissimilarity from that of the last keyframe. Scene boundary identification via the use of SIFT features. Shot clusters - formed by SIFT feature matching. Shot/keyframe characterization: Interview shots - explicitly detected; Shots with dominant faces - using a face detector; Unstable Frames - detected by the use of optical flow vectors. The summarization algorithm is a process of selecting which shot and which keyframe from each shot to be included in the summary.

The evaluation is based on the judgment of people, unfamiliar with the field, following given criteria answering questions such as: How many out of the total number of main characters are captured in the summary? Does the order of summary frames correspond to the relative importance of characters/objects and setting in the original video sequence? How many frames are considered redundant? How many frames are considered missing?

Twelve summary examples are presented. According to the testers two were perfect, three needed one shot more, the rest contained redundant shot/s. Some of the results show the importance of producer-expert judgment.

Example: City View. The city is captured with 17 shots at different camera configurations and time of the day. The summary consists of only one shot. While two testers considered that as sufficient, two testers thought that it should include the city shot of the sun-set too.

The testers observed that while it is easy to identify redundant frames, it is harder to identify missing frames from the summary set. The technique has problem of separating different interviewee if the interview location is the same.

The **DFKI** proposed a video retrieval framework based on a combination of features (spatiograms) and shot distance measures (Jensen-Shannon divergence). They performed redundancy detection (by clustering) and interview experiments.

- *Clustering.* Subset of 149 shots of BBC Rushes Test data were manually selected and labeled to obtain 33 clusters. Runs with features based on the color attribute were performed. The proposed approach is tested together with several baseline methods. All runs are tested in two ways: NN classification and Clustering, and error rate is measured. In the case of NN classification each of the 149 shots was removed from the dataset, and the remaining 148 shots were used as training samples. In the case of clustering, distance matrices were used and for each clustering result the manual labels were considered as ground truth. None of the tested approaches has shown the capability to reliably structure shots of a complex dataset into meaningful clusters. Best performance in both cases is achieved by the proposed framework. NN classification yield error rate of 14.4% significantly outperforming the unsupervised clustering, error rate of 42.3%.

- *Interview experiment.* Features and similarity measures applied over (MPEG-4) motion vectors. Subset of 1404 shots of BBC Rushes Test data were manually labeled with "showing an interview" or "not showing an interview". Shots were split up into 702 training and 702 testing disjoint sets. One color-based and several using extracted MPEG motion vectors runs were done For each run, 3-NN classification is performed on the resulting distance matrix. Precision and recall is used for evaluation. Best performance was achieved by the proposed framework with precision of 69% and recall of 84%, indicating that spatiograms make an excellent fast-to-extract, compress domain descriptor. Misclassification's were due mainly when people were present in the image, but it wasn't an interview.

The **Joanneum Research** provided a pictorial summary by the implementation of an extensible and feature-independent framework for content browsing. The feature specific parts are implemented as plug-ins. The framework is based on a concept on which the user starts from the full content set and restricts this set by the following iterative process: 1) clustering (in terms of a selected feature) 2) selecting representative items for each cluster 3) visualization of the clusters using the selected media items 4) The user: selects relevant item or repeat clustering (starting from 1)) or filter (selects relevant clusters and discards the others)

The framework is composed of the following components:

- *Data store.* Keyframes stored in file system, metadata stored in a relational database, or in the file system

- *Indexing service.* Feature specific indexing performed by a set of indexing plug-ins. Each plug-in handles a certain feature by reading the required information from MPEG-7 description and creating the necessary database entries.

- *Summarizer.* Handles clustering, filtering and selection of representative media items.

- *User interface* Visualizes the current content set and cluster structure and allows selecting items as input for filtering or as relevant results. A key frame based visualization (light table) is implemented.

The video browsing tool uses the following features: camera motion, motion activity, audio volume, face occurrence, global color similarity and object similarity. The result of feature extraction is one metadada description per video.

The evaluation method is based on formulation of tasks (textual descriptions of the content to be retrieved). Seven non-expert users using the tool completed four tasks. The measures used for evaluation were: precision, recall, total time for completing the task (max. 10 minutes) and number of selected (correct and false) items, measured every 30 seconds. Ground truth data were produced by the use of a video annotation tool. All Rushes test data shots that are in a village/town/city environment have been annotated. For these segments, a description of the location, salient objects and persons and concepts is made.

The task that yields the worst precision and recall, and at the same time was the one with the shortest average working time was "Find a zoom in on an interviewee". In all results recall is always much lower than precision meaning that there is always a number of relevant segments in the content that have not been discovered. The evaluation has shown that the improvement of recall is the most important task for the future.

The **University of Marburg** presented another approach based on unsupervised clustering of shots to eliminate redundant shots. The system consists of a summarization and a retrieval component. First, video cut detection is applied to the rushes material. Then, rushes shots are segmented further to obtain sub-shots with respect to events like speech, silence, camera motion and face appearances. Summarization is achieved by clustering those sub-shots using high-level audiovisual features. The resulting clusters are visualized and the user can navigate interactively through these sub-shots to search the videos. Two visualization techniques were compared: Sammon mapping and classical multidimensional scaling (MDS). The experiments demonstrated that both methods work equally well. Furthermore, users can employ an automated search using the retrieval component with respect to the following features: number of faces, shot size, pan, tilt, zoom, silence, speech, music, action, background noise, and interview. In addition, a user can select arbitrary frame regions to search for similar regions in the rushes material (query-by-example). In the experiments, the rushes test data were used and the top-50 precision was measured for several features. These experiments showed very good retrieval results for audio and face features, while the work with camera-motion features showed room for improvement, especially in the case of ranking in the retrieval list. Some interesting observations are, for music (top-50 precision), 43% of the false positives included chirping of birds or blowing of a whistle, which can be related to the concept of music. For camera motion features, the movement of a big foreground object along the desired camera-movement axis, which is hard to distinguish even for humans was the reason for the falsely retrieved sub-shots.

6.3 Observations

A large number of groups can build systems to ingest, analyze, and allow user filtering and summarization. Most approaches are based on detecting redundancy through clustering. Surprising emphasis on audio classification. Few groups did actual evaluation. Those that did did classic ad hoc search. Results suggest the importance of producer-like tester.

Readers are invited to see the site papers in the workshop notebook for for further details about their approaches and results.

The experience learned from this exploration will be used for the addition of a workshop on video summarization, to be held as part of the ACM Multimedia Conference 2007, with report out at TRECVID 2007.

7 Summing up and moving on

This overview of TRECVID 2006 has provided information on the goals, data, evaluation mechanisms and metrics used. In addition it contains an overview of the approaches and results. Further details about each particular group's approach and performance can be found in that group's site report in the Publications section of the TRECVID website.

8 Authors' note

TRECVID would not happen without support from DTO and NIST. The research community is very grateful for this. Beyond that, various individuals and groups deserve special thanks.

We are particularly grateful to Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin for providing the master shot reference and to the team at the Centre for Digital Video Processing at Dublin City University (DCU) for formating the master shot reference definition and selecting keyframes.

City University of Hong Kong, the University of Amsterdam, and the University of Iowa helped out in the distribution of rushes data by mirroring the them online.

Bing Xiang, John Makhoul, and Ralph Weischedel at BBN provided the output of their then latest ASR/MT engines on the Arabic and Chinese sources.

Cees Snoek and other from the MediaMill team at the University of Amsterdam donated baseline results for 101 features trained on the 2005 and then the 2006 development data.

Carnegie Mellon University, Columbia University, and IBM provided annotations for 449 LSCOM features on the 2005 development data.

We appreciate Jonathan Lasko's painstaking creation of the shot boundary truth data once again.

Finally, we want to thank all the participants and other contributors on the mailing list for their enthusiasm and diligence.

9 Appendix A: Topics

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

Table 7: 2006 Topic types

	Named			Generic		
Topic	Person, thing	Event	Place	Person, thing	Event	Place
173				X	X	
174				X		
175				X	X	
176				X	X	
177				X	X	
178	X					
179	X			X		
180				X		
181	X				X	
182				X		
183				X		X
184				X		
185				X	X	
186				X		X
187				X	X	
188				X	X	
189				X		
190				X		
191				X		
192				X	X	
193				X	X	
194	X					
195				X		
196				X		X

0173 Find shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.) (I/0, V/4, R/142)

0174 Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible (I/3, V/4, R/675)

0175 Find shots with one or more people leaving or entering a vehicle (I/0, V/10, R/204)

0176 Find shots with one or more soldiers, police, or guards escorting a prisoner (I/0, V/4, R/111)

0177 Find shots of of a daytime demonstration or protest with at least part of one building visible (I/4, V/4, R/393)

0178 Find shots of US Vice President Dick Cheney (I/3, V/3, R/99)

0179 Find shots of Saddam Hussein with at least one other person's face at least partially visible (I/8, V/0, R/191)

0180 Find shots of multiple people in uniform and in formation (I/3, V/5, R/197)

0181 Find shots of US President George W. Bush, Jr. walking (I 0, V/5, R/128)

0182 Find shots of one or more soldiers or police with one or more weapons and military vehicles (I/2, V/6, R/307)

0183 Find shots of water with one or more boats or ships (I/3, V/5, R/299)

0184 Find shots of one or more people seated at a computer with display visible (I/3, V/4, R/440)

0185 Find shots of one or more people reading a newspaper (I/3, V/4, R/201)

0186 Find shots of a natural scene - with, for example, fields, trees, sky, lake, mountain, rocks, rivers, beach, ocean, grass, sunset, waterfall, animals, or people; but no buildings, no roads, no vehicles (I/2, V/4, R/523)

0187 Find shots of one or more helicopters in flight (I/0, V/6, R/119)

0188 Find shots of something burning with flames visible (I/3, V/5, R/375)

0189 Find shots of a group including least four people dressed in suits, seated, and with at least one flag (I/3, V/5, R/446)

0190 Find shots of at least one person and at least 10 books (I/3, V/5, R/295)

0191 Find shots containing at least one adult person and at least one child (I/3, V/6, R/775)

0192 Find shots of a greeting by at least one kiss on the cheek (I/0, V/5, R/98)

0193 Find shots of one or more smokestacks, chimneys, or cooling towers with smoke or vapor coming out (I/3, V/2, R/60)

0194 Find shots of Condoleeza Rice (I/3, V/7, R/122)

0195 Find shots of one or more soccer goalposts (I/3, V/4, R/333)

0196 Find shots of scenes with snow (I/3, V/6, R/692)

10 Appendix B: Features

1 Sports: Shots depicting any sport in action

2 Entertainment: Shots depicting any entertainment segment in action

3 Weather: Shots depicting any weather related news or bulletin

4 Court: Shots of the interior of a court-room location

5 Office: Shots of the interior of an office setting

6 Meeting: Shots of a Meeting taking place indoors

7 Studio: Shots of the studio setting including anchors, interviews and all events that happen in a news room

8 Outdoor: Shots of Outdoor locations

9 Building: Shots of an exterior of a building

10 Desert: Shots with the desert in the background

11 Vegetation: Shots depicting natural or artificial greenery, vegetation woods, etc.

12 Mountain: Shots depicting a mountain or mountain range with the slopes visible

13 Road: Shots depicting a road

14 Sky: Shots depicting sky

15 Snow: Shots depicting snow

16 Urban: Shots depicting an urban or suburban setting

17 Waterscape,Waterfront: Shots depicting a waterscape or waterfront

18 Crowd: Shots depicting a crowd

19 Face: Shots depicting a face

20 Person: Shots depicting a person (the face may or may not be visible)

21 Government-Leader: Shots of a person who is a governing leader, e.g., president, prime-minister, chancellor of the exchequer, etc.

22 Corporate-Leader: Shots of a person who is a corporate leader, e.g., CEO, CFO, Managing Director, Media Manager, etc.

23 Police,security: Shots depicting law enforcement or private security agency personnel

24 Military: Shots depicting the military personnel

25 Prisoner: Shots depicting a captive person, e.g., imprisoned, behind bars, in jail or in handcuffs, etc.

26 Animal: Shots depicting an animal, not counting a human as an animal

27 Computer,TV-screen:Shots depicting a television or computer screen

28 Flag-US: Shots depicting a US flag

29 Airplane: Shots of an airplane

30 Car: Shots of a car

31 Bus: Shots of a bus

32 Truck: Shots of a truck

33 Boat,Ship: Shots of a boat or ship

34 Walking,Running: Shots depicting a person walking or running

35 People-Marching: Shots depicting many people marching as in a parade or a protest

36 Explosion,Fire: Shots of an explosion or a fire

37 Natural-Disaster: Shots depicting the happening or aftermath of a natural disaster such as earthquake, flood, hurricane, tornado, tsunami

38 Maps: Shots depicting regional territory graphically as a geographical or political map

39 Charts: Shots depicting any graphics that is artificially generated such as bar graphs, line charts, etc. (maps should not be included)

References

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.

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.

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

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

Wright, R. (2005). *Personal communication from Richard Wright, Technology Manager, Projects, BBC Information & Archives.*

Yilmaz, E., & Aslam, J. A. (2006). Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management (CIKM).* Arlington, VA, USA.

Table 1: Participants and tasks

Participants	Country	Task			
Accenture Technology Labs	USA	–	–	–	RU
AIIA Laboratory	Greece	SB	–	–	–
AT&T Labs - Research	USA	SB	–	SE	RU
Beijing Jiaotong U.	China	–	–	SE	–
Bilkent U.	Turkey	–	FE	SE	–
Carnegie Mellon U.	USA	–	FE	SE	–
Chinese Academy of Sciences (CAS/MCG)	China	–	–	–	RU
Chinese Academy of Sciences (CAS/JDL)	China	SB	–	–	–
Chinese U. of Hong Kong	China	–	FE	SE	–
City University of Hong Kong (CityUHK)	China	SB	FE	SE	–
CLIPS-IMAG	France	SB	FE	SE	–
Columbia U.	USA	–	FE	SE	–
COST292 (www.cost292.org)	EU	SB	FE	SE	RU
Curtin U. of Technology	Australia	SB	–	–	RU
DFKI GmbH	Germany	–	–	–	RU
Dokuz Eylul U.	Turkey	SB	–	–	–
Dublin City U.	Ireland	–	–	SE	–
Florida International U.	USA	SB	–	–	–
Fudan U.	China	–	FE	SE	–
FX Palo Alto Laboratory Inc	USA	SB	FE	SE	–
Helsinki U. of Technology	Finland	SB	FE	SE	–
Huazhong U. of Science and Technology	China	SB	–	–	–
IBM T. J. Watson Research Center	USA	–	FE	SE	RU
Imperial College London / Johns Hopkins U.	UK/USA	–	FE	SE	–
Indian Institute of Technology at Bombay	India	SB	–	–	–
NUS / I2R	Singapore	–	FE	SE	–
IIT / NCSR Demokritos	Greece	SB	–	–	–
Institut EURECOM	France	–	FE	–	RU
Joanneum Research Forschungsgesellschaft	Austria	–	–	–	RU
KDDI / Tokushima U. / Tokyo U. of Technology	Japan	SB	FE	–	–
K-Space (kspace.qmul.net)	EU	–	FE	SE	–
Laboratory ETIS	Greece	SB	–	–	–
LIP6 - Laboratoire d'Informatique de Paris 6	France	–	FE	–	–
Mediamill / U. of Amsterdam	The Netherlands	–	FE	SE	–
Microsoft Research Asia	China	–	FE	–	–
Motorola Multimedia Research Laboratory	USA SB	–	–	–	
National Taiwan U.	Taiwan	–	FE	–	–
NII/ISM	Japan	–	FE	–	–
RMIT U. School of CS&IT	Australia	SB	–	SE	–
Tokyo Institute of Technology	Japan	SB	FE	–	–
Tsinghua U.	China	SB	FE	SE	RU
U. of Bremen TZI	Germany	–	FE	–	–
U. of California at Berkeley	USA	–	FE	–	–
U. of Central Florida	USA	–	FE	SE	–
U. of Electro-Communications	Japan	–	FE	–	–
U. of Glasgow / U. of Sheffield	UK	–	FE	SE	–
U. of Iowa	USA	–	FE	SE	–
U. of Marburg	Germany	SB	–	–	RU
U. of Modena and Reggio Emilia	Italy	SB	–	–	–
U. of Ottawa / Carleton U.	Canada	SB	–	–	–
U. of Oxford	UK	–	FE	SE	–
U. of Sao Paolo	Brazil	SB	–	–	–
U. Rey Juan Carlos / Dublin City U.	Spain	SB	–	SE	RU
Zhejiang U.	China	SB	FE	SE	–

Task legend. SB: Shot boundary; FE: High-level features; SE: Search ; RU: BBC rushes

Table 4: Feature pooling and judging statistics

Feature number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number true	% judged that were true
1	233646	47108	20.2	220	3334	7.1	679	20.4
3	232793	47111	20.2	230	3264	6.9	474	14.5
5	236583	56072	23.7	110	3483	6.2	292	8.4
6	234686	46967	20.0	140	3427	7.3	1498	43.7
10	234730	47675	20.3	130	3353	7.0	172	5.1
12	234749	46306	19.7	140	3351	7.2	163	4.9
17	234391	44099	18.8	150	3255	7.4	427	13.1
22	233658	52982	22.7	110	3371	6.4	22	0.7
23	233292	56100	24.0	100	3434	6.1	340	9.9
24	233456	47047	20.2	130	3254	6.9	612	18.8
26	235465	53551	22.7	110	3270	6.1	243	7.4
27	238532	46689	19.6	140	3290	7.0	1556	47.3
28	230852	51552	22.3	130	3254	6.3	231	7.1
29	238438	50829	21.3	140	3262	6.4	166	5.1
30	234328	45793	19.5	140	3361	7.3	750	22.3
32	236076	49727	21.1	120	3390	6.8	238	7.0
35	233127	46895	20.1	140	3250	6.9	150	4.6
36	232393	49268	21.2	130	3384	6.9	221	6.5
38	231767	44126	19.0	210	3375	7.6	511	15.1
39	228361	47485	20.8	190	3407	7.2	329	9.7

Table 5: Search pooling and judging statistics

Topic number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number relevant	% judged that were relevant
173	115248	28312	24.6	80	3669	13.0	142	3.9
174	117517	29734	25.3	70	3743	12.6	675	18.0
175	113024	33293	29.5	70	3919	11.8	204	5.2
176	114012	30063	26.4	70	3916	13.0	111	2.8
177	115904	27297	23.6	90	3542	13.0	393	11.1
178	112852	30145	26.7	100	3287	10.9	99	3.0
179	116894	26503	22.7	110	3497	13.2	191	5.5
180	115272	34038	29.5	70	4408	13.0	197	4.5
181	117850	28141	23.9	80	3457	12.3	128	3.7
182	117135	26353	22.5	80	3484	13.2	307	8.8
183	115522	28584	24.7	90	3763	13.2	299	7.9
184	115214	34229	29.7	70	3516	10.3	440	12.5
185	117167	31236	26.7	70	3436	11.0	201	5.8
186	117836	31430	26.7	70	3611	11.5	523	14.5
187	113495	27800	24.5	100	3697	13.3	119	3.2
188	114389	32715	28.6	90	3577	10.9	375	10.5
189	117763	36079	30.6	70	4138	11.5	446	10.8
190	117687	33855	28.8	70	3706	10.9	295	8.0
191	117858	32807	27.8	70	3559	10.8	775	21.8
192	110356	37242	33.7	70	3936	10.6	98	2.5
193	114040	33806	29.6	70	3738	11.1	60	1.6
194	112202	33741	30.1	100	3786	11.2	122	3.2
195	113326	31201	27.5	130	3348	10.7	333	9.9
196	118109	24117	20.4	110	3375	14.0	692	20.5

Table 8: Participants not submitting runs (or at least papers in the case of rushes task)

Participants	Country	Task			
Cambridge U.	UK	–	–	–	–
Fraunhofer-Institute for Telecommunications	Germany	–	–	–	–
INESC-Porto	Portugal	–	–	–	–
Indian Institute of Technology at Kharagpur	India	–	–	–	–
Language Computer Corporation (LCC)	USA	–	–	–	–
LowLands team (CWI + Twente U.)	the Netherlands	–	–	–	–
Nagoya U.	Japan	–	–	–	–
Northwestern U.	USA	–	–	–	–
Ryerson U.	Australia	–	–	–	–
Tampere U. of Technology	Finland	–	–	–	–
U. of East Anglia	UK	–	–	–	–
U. of Kansas	USA	–	–	–	–
U. of North Carolina at Chapel Hill	USA	–	–	–	–
U. of Washington	USA	–	–	–	–
U. of Wisconsin-Milwaukee	USA	–	–	–	–

Task legend. SB: Shot boundary; HL: High-level features; SE: Search ; RU: BBC rushes