TRECVID 2007 - Overview

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

Wessel Kraaij {wessel.kraaij@.tno.nl}
TNO Information and Communication Technology
Delft, the Netherlands

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

April 16, 2008

1 Introduction

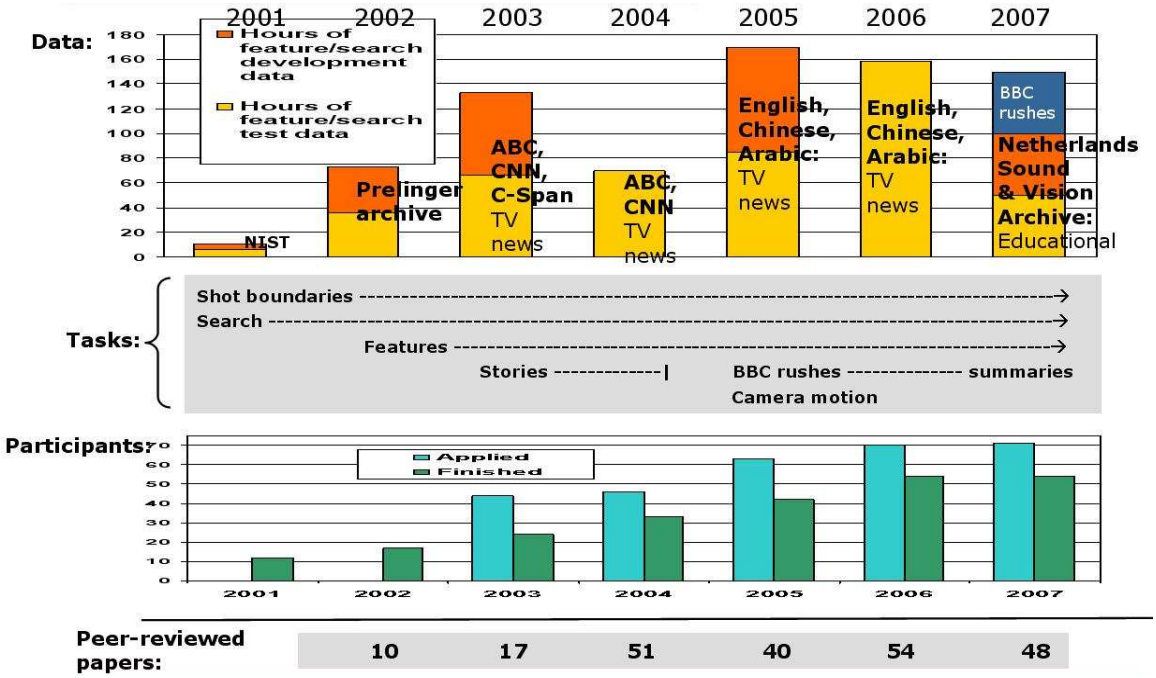
The TREC Video Retrieval Evaluation (TRECVID) 2007 represents the seventh running of a TREC-style (trec.nist.gov) video retrieval evaluation, the goal of which remains to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. Over time this effort should yield a better understanding of how systems can effectively accomplish such retrieval and how one can reliably benchmark their performance. TRECVID 2007 was funded by the US National Institute of Standards and Technology (NIST) and the Intelligence Advanced Research Projects Activity (IARPA).

54 teams (see Table 1 at the end of the paper) from various research organizations — 17 from Asia, 23 from Europe, 12 from the Americas, and 2 from Australia — 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 (rushes) summarization. See Figure 1 for an overview of TRECVID's evolution.

In 2007 TRECVID began what sets out to be a 3-year cycle using new data sources, related to the broadcast news used in 2003-2006 but significantly different. Data for the search and feature tasks was about 100 hours of (MPEG-1) news magazine, science news, news reports, documentaries, educational programming, and archival video almost entirely in Dutch from the Netherlands Institute for Sound and Vision. About 6 additional hours of Sound and Vision data was used for the shot boundary task. The BBC Archive provided about 50 hours of "rushes" - pre-production video material with natural sound, errors, etc. - from several BBC dramatic series for use in the summarization task.

Results were scored by NIST against human judgments. Complete manual annotation of the test set, created by NIST, was used to evaluate shot boundary determination. Feature and search submissions were evaluated based on partial manual judgments of the pooled submissions. The output of summarization systems was manually evaluated at NIST using ground truth created at Dublin City University. Full results for the summarization task were presented and discussed as the TRECVID Video Summarization Workshop at the ACM Multimedia Conference in Augsburg, Germany on September 28, 2007 (Over, Smeaton, & Kelly, 2007).

Figure 1: Evolution of TRECVID



This paper is an introduction to the evaluation framework — the tasks, data, and measures. The results as well as the approaches taken by the participating groups were presented at the TRECVID workshop in November 2007. For detailed information about the approaches and results, the reader should see the various site reports and slides from the workshop available from the publications page of the TRECVID website: www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

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.

1.1 New in TRECVID 2007

The new kinds of data for the feature, search, and shot boundary tasks presented new challenges and made it possible to test how well the broadcast news training data generalized to a related but significantly different sort of video data.

The amount of development and test data for the feature and search tasks was smaller than in previous years and seemed more diverse in content.

No keyframes were provided by NIST. This was to encourage participants to look afresh at how best to train their systems, reconsidering tradeoffs between processing speed, effectiveness, amount of the video processed.

While automatic speech recognition (ASR) and then machine translation (MT) (Dutch to English) was applied to the Sound and Vision videos, TRECVID 2007 required search and feature task participants to submit at least one run based on visual information only - to simulate a situation in which no ASR and MT for the language of a video might be available.

The rushes summarization task was a first attempt at large-scale evaluation of such systems and tested the feasibility of the evaluation framework.

For the first time, all development and test data were distributed via the Internet. Each participating group downloaded up to 100 GB of data from one of four servers at City University Hong Kong, NIST, University of Iowa, or University of Modena.

Table 2: Characteristics of 2007 Sound and Vision test data

	TV2006	TV2007
Dataset length (hours)	~158	~100
Number of shots	79,484	18,142
Average shot length	7 sec	20 sec
Number of unique program titles	11	47

2 Data

2.1 Video

Sound and Vision data

The Netherlands Institute for Sound and Vision generously provided 400 hours of news magazine, science news, news reports, documentaries, educational programming, and archival video in MPEG-1 format for use within TRECVID. TRECVID 2007 used approximately 100 hours of this data in 2007. The amount was kept small because for the first time all the data had to be downloaded and because the data represented a new genre and potential new problems for systems. Table 2 highlights some of the easily quantifiable differences between the 2006 broadcast news video and the Sound and Vision data.

The data was divided as follows:

- 6 hours for the shot boundary task
- 50 hours for development of search/feature detection
- 50 hours for test of search/feature detection

A shot boundary test collection for 2007 was drawn at random from the total collection. It comprised 17 videos for a total size of about 4.08 gigabytes. The characteristics of this test collection are discussed below.

The collections for the search and feature tasks were drawn randomly so as to be balanced across the various program sources. The development data comprised 110 files and 30.6 GB, the test data 109 files and 29.2 GB.

A technical problem that prevented display of shots from one file (BG_37940.mpg, file ID: 200) in the test data was discovered during feature task assessment. As a result all shots from the file were removed from the feature pools and submissions. Search task participants were warned to remove these shots before submission.

BBC Archive data

The BBC Archive provided about 100 hours of rushes data for use in the video summarization task. About half was used for development data and half reserved for testing. The data consisted of raw (i.e., unedited) video footage, shot mainly for five series of BBC drama programs. The drama series included a historical drama set in London in the early 1900's, a series on ancient Greece, a contemporary detective program, a program on emergency services, a police drama, as well as miscellaneous scenes from other programs.

2.2 Common shot reference, ASR, MT

The entire feature/search collection was automatically divided into shots by Christian Petersohn 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 18,142 reference shots (40% of the number used in 2005).

Roeland Ordelman and Marijn Huijbregts at the University of Twente provided the output of an automatic speech recognition system run on the Sound and Vision data. Christof Monz of Queen Mary, University London contributed machine translation (Dutch to English) for the Sound and Vision video based on the University of Twente ASR.

2.3 Common feature annotation

Georges Quénot and Stéphane Ayache of LIG (Laboratoire d'Informatique de Grenoble, formerly CLIPS-IMAG) organized a collaborative annotation for

TRECVID 2007 using an active learning scheme designed to improve the efficiency of the annotation process. About 27 groups participated and shared the resulting ground truth among themselves.

The Multimedia Computing Group at the Chinese Academy of Sciences together with the National University of Singapore provided full annotation of the 2007 training data (using one keyframe per shot).

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

In 2007 there was special interest in how well systems trained on one sort of data generalize to another related, but different type of data with little or no new training data. The available training data contained some that is specific to the Sound and Vision video and some that was not. Therefore three additional training categories were introduced:

a - same as A but no training data (shared or private) specific to any Sound and Vision data has been used in the construction or running of the system.

b - same as B but no training data (shared or private) specific to any Sound and Vision data has been used in the construction or running of the system.

c - same as C but no training data (shared or private) specific to any Sound and Vision data has been used in the construction or running of the system.

Groups were encouraged to submit at least one pair of runs from their allowable total that helps the community understand how well systems trained on non-Sound-and-Vision data generalize to Sound-and-Vision data.

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 637,805 frames and 2317 shot transitions. This means the 2007 shots are much longer (275.3 frames/shot) on average than in the broadcast news video from 2006 (157.7 frames/shot).

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;

Table 3: Transition types

Search type	2003	2004	2005	2006	2007
% Abrupt	70.7	57.5	60.8	48.7	89.5
% Dissolve	20.2	31.7	30.5	39.9	6
% Fade in/out	3.1	4.8	1.8	1.3	0
% Other	5.9	5.7	6.9	10.1	4.5

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.

The same student has created the shot boundary ground truth for TRECVID since 2001 using the same guideline. 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 significantly different from earlier years (see Table 3) in that the percentage of cuts almost doubled and there were relatively few gradual transitions:

- 2,236 — hard cuts (90.8%)
- 134 — dissolves (5.4%)
- 2 — fades to black and back (1%)
- 91 — other (3.7%)

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 were 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.

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

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

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

Gradual transitions could only match gradual transitions and cuts match only cuts, except in the case of very short gradual transitions (5 frames or less), which, whether in the reference set or in a submission, were treated as cuts. We also expanded each abrupt reference transition by 5 frames in each direction before matching against submitted transitions to accommodate differences in frame numbering by different decoders.

Accuracy for reference gradual transitions successfully detected was measured using the one-to-one matching list output by the detection evaluation. The accuracy measures were frame-based precision and recall. 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 Results

Figures 2, 3 and 4 depict the main results. Note that the small absolute number of gradual transitions in the 2007 data make results for these transitions less generally informative than usual. The results for cuts are strikingly good. The runs submitted by LIG (formerly CLIPS-IMAG) come from essentially the same system as was run in 2006. Figure 5 make clear that the cuts in the 2007 data are easier than in 2006. Good results are possible in times well under that needed to view the videos at normal speed (see Figure 6) and spending more time processing is not correlated with better effectiveness (see Figure 7).

Figure 6: Mean SB runtimes

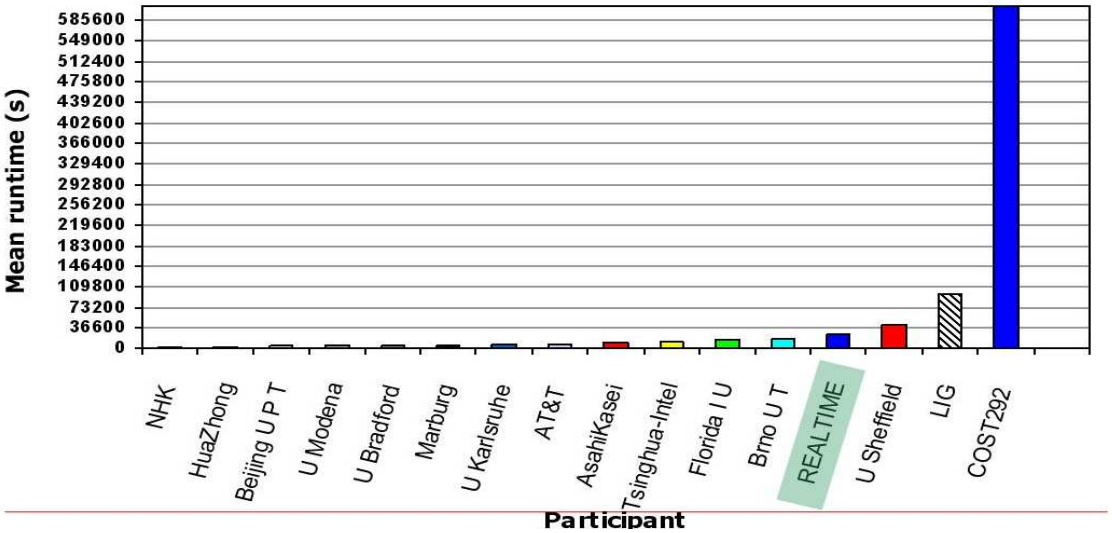


Figure 2: Precision and recall for cuts

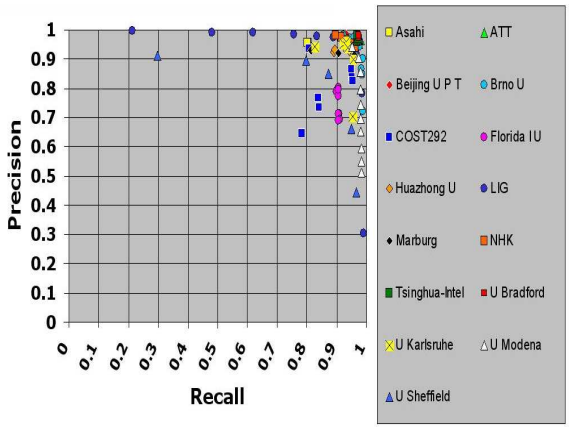


Figure 3: Precision and recall for gradual transitions

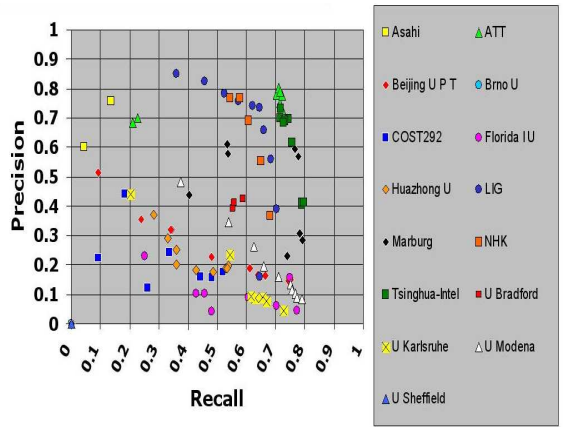


Figure 4: Frame-precision and -recall

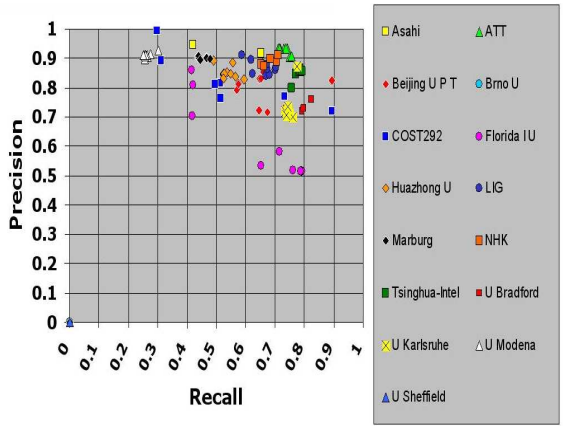


Figure 5: Precision and recall for cuts from same system (CLIPS-IMAG/LIG) run on 2006 and 2007 data

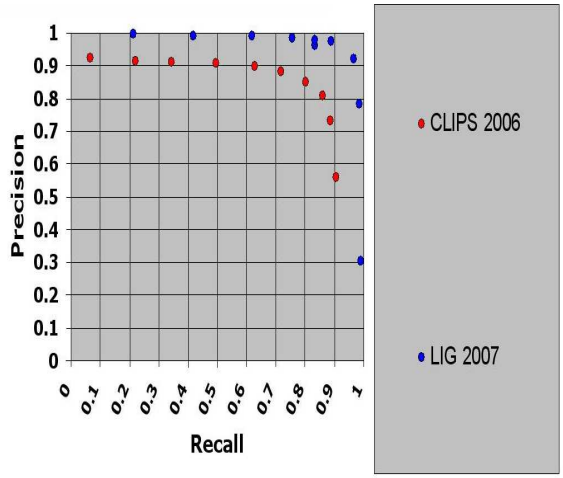
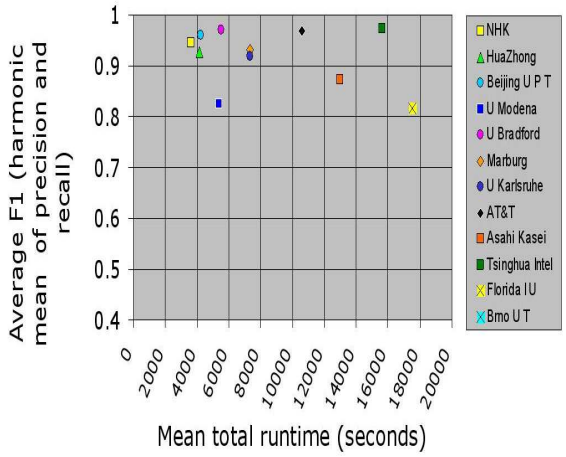


Figure 7: Mean SB runtimes (faster than realtime) versus effectiveness (mean F1 (harmonic mean of precision and recall) for cuts



For detailed information about the approaches and results, the reader should see the various site reports and slides from the workshop available from the publications page of the TRECVID website: www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

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 does take on added importance since it can improve 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 highest possibility of detecting the presence of the feature. The presence of each feature was assumed to be binary, i.e., it was either present or absent in the given standard video shot. If the feature was true for some frame (sequence) within the shot, then it was true for the shot. This is a simplification adopted for the benefits it afforded in pooling of results and approximating the basis for calculating recall.

The feature set was nearly the entire preliminary set of 39 LSCOM-lite features, chosen to cover a variety of target types. Participants were required to build detectors for all 36 features. Requiring this number of detectors was designed to promote the use of generic methods for detector development.

Recent work at Northeastern University (Yilmaz & Aslam, 2006) has resulted in methods for estimating standard system performance measures using relatively small samples of the usual judgment sets so that larger numbers of features can be evaluated using the same amount of judging effort. Tests on past data showed the new measure (inferred average precision) to be a good estimator of mean average precision (Over, Ianeva, Kraaij, & Smeaton, 2006). 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.

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 in 2007 were as follows and are numbered 1-39. The same list was used for 2006 except that features 2 (entertainment), 21 (government leader), and 22 (corporate-leader) were dropped from the list for 2007 since they had proved very difficult to judge. Those evaluated are marked by an asterisk: [1*]Sports, [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,[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 with the detailed feature runs at the back of the notebook and in Appendix B in this paper.

4.1 Data

As mentioned above, the feature test collection contained 109 files/videos and 18,142 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 32 groups submitted a total of 163 runs. Among the 163 runs, 146 used training type A, 7 training type B, 6 training type C and 4 training type a.

TRECVID 2007 required a feature run (among the 6) treating the new video as if no automatic speech recognition (ASR) or machine translation (MT) for the languages of the videos (mostly Dutch) existed - as might occur in the case of video in other less well known languages.

For each feature, all submissions down to a depth of at least 100 (average 154, maximum 240) 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 at the end of the paper. In all, 66,293 shots were judged.

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 (infAP) across all 20 evaluated features as well as "within feature".

Figure 8: infAP by run - top half

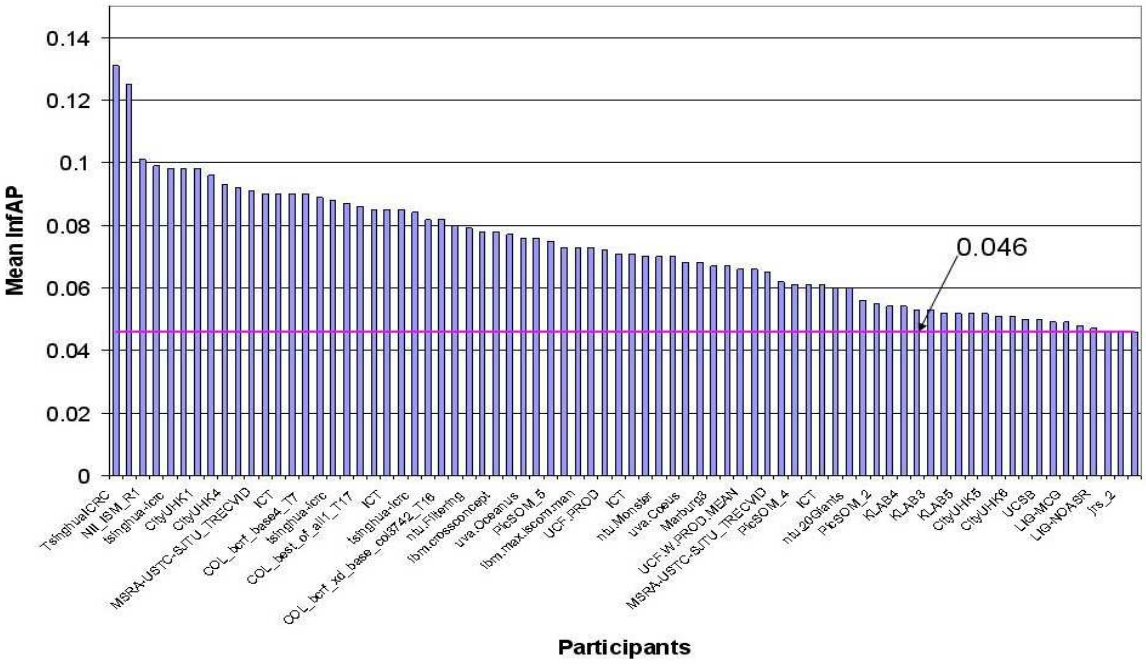


Figure 9: infAP by run - bottom half

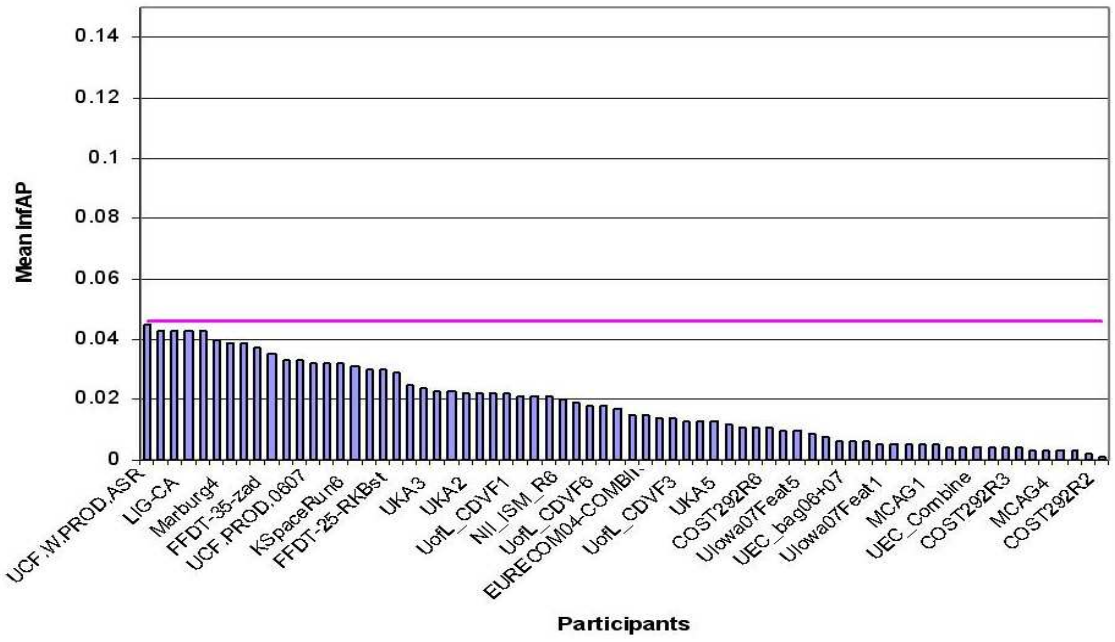


Figure 10: infAP by number of true shots in the feature test data

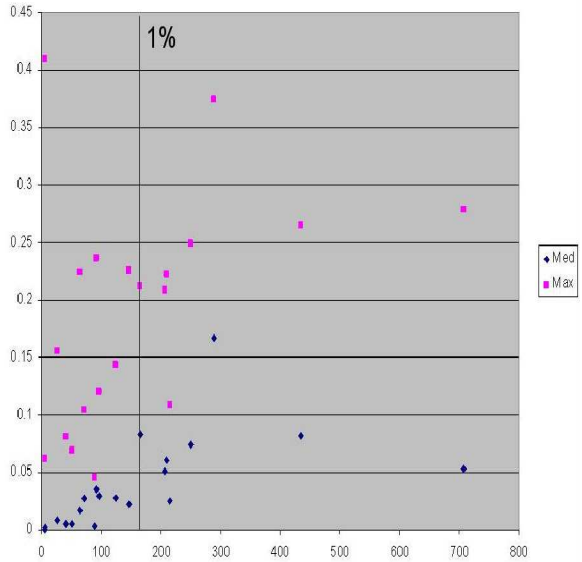
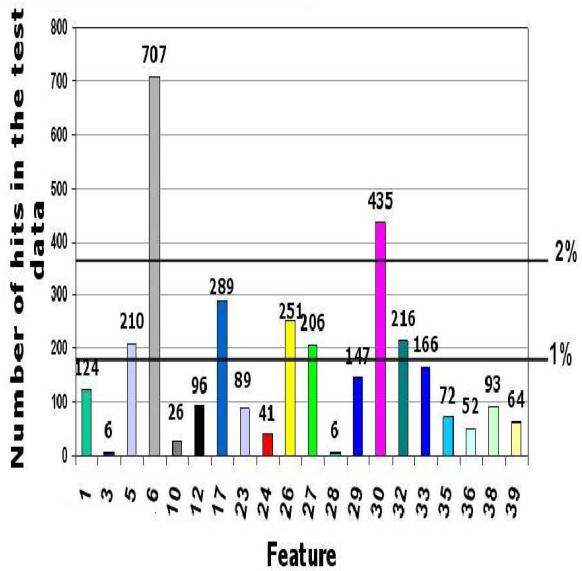


Figure 11: True shots by feature



4.4 Results

Figures 8 and 9 present an overview of the results from runs of type A. Performance varies greatly by feature. Figure 10 shows a weak positive correlation between number of hits possible for a feature and the median or maximum score for that feature.

Looking at general statistics from the results we can find that the top two features with maximum hits (see Figure 11) are "meeting" (707 hits) and "car" (435 hits) - perhaps because there is a lot of research concerning detecting popular objects such as people and cars within the object recognition community. On the other hand, a feature such as "weather", which had good performance in 2006, has only 6 hits against the 2007 data. These results show that the high level feature extraction task can promote the research efforts concerning detecting new, unfamiliar, and unpopular kind of features which can be of more benefit eventually and can encourage new application domains in the future.

Regarding category "A" results, the maximum mean infAP reached 0.131 while for category "a" it reached 0.049. Category "B" reached maximum 0.132 while category "C" achieved 0.073. Figure 12 shows the performance of the top 10 runs across the features when compared to the median. To test if there is a real significant differences between the systems per-

Figure 13: Randomization test for significant differences in top 10 feature runs

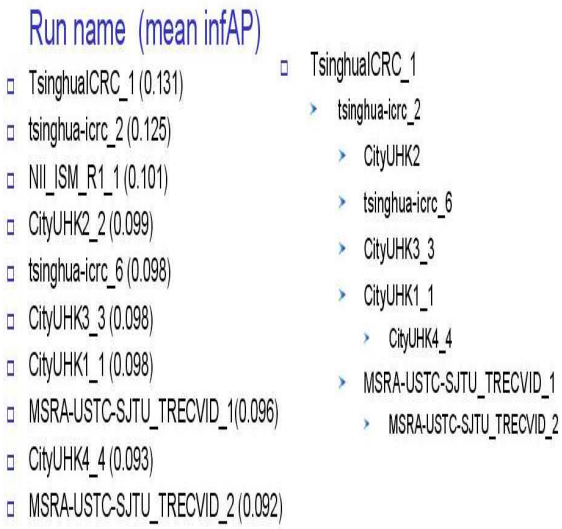
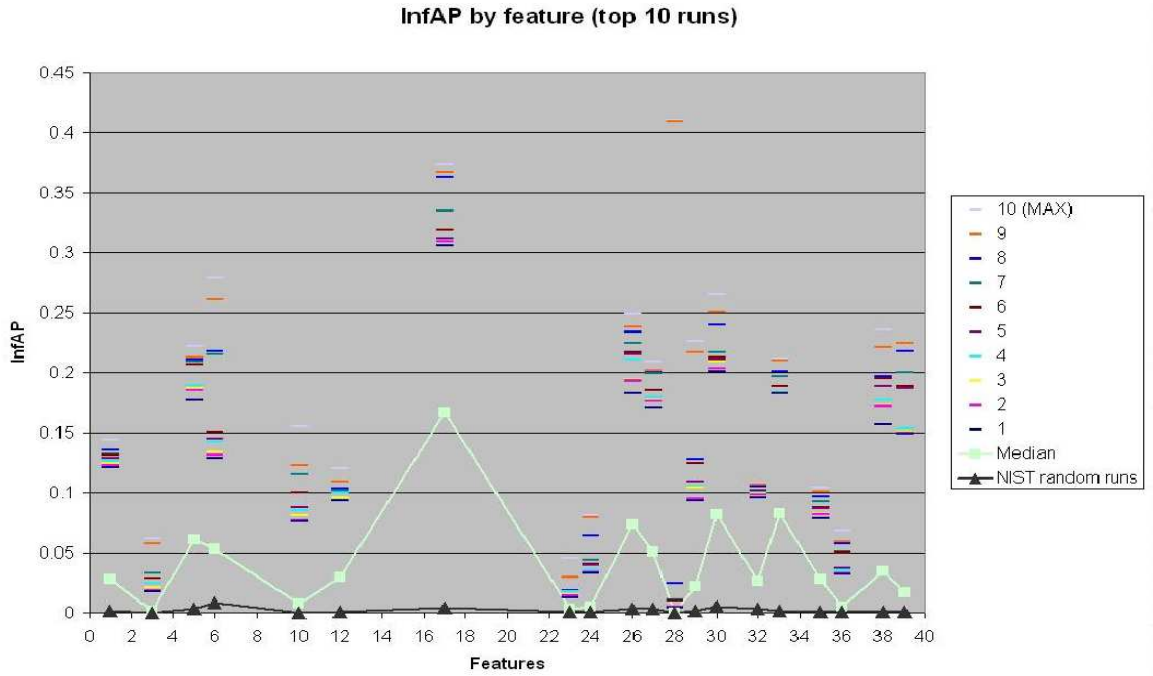


Figure 12: infAP by feature - top 10 runs



formance, we applied a randomization test (Manly, 1997) on the top 10 runs as shown in Figure 13. The left half indicates the sorted top 10 runs, while the right half indicates the order by which the runs are significant according to the randomization test.

We asked the participants to fill out some metadata information to help us draw some general observations about the systems. From this metadata, we found that many groups depended only on visual information while using audio information as well yielded only mixed results across sites. Few groups experimented with alternative keyframe extraction methods. A good observation is that temporal analysis gained more attention, which is healthy for more sophisticated features. Some grey-scale-specific approaches and learning from unlabeled data has been explored by some groups as well. Regarding the classifiers architecture, most groups, including the top 10 runs, tried to build a generic architecture, while few built specific feature classifiers.

The number of classifiers used for fusion varied greatly from 1 to 231. Figure 14 plots the performance vs. the number of classifiers used for fusion. The general trend seems to be that increasing the number of classifiers helps to improve the perfor-

Figure 14: Number of classifiers vs. MAP

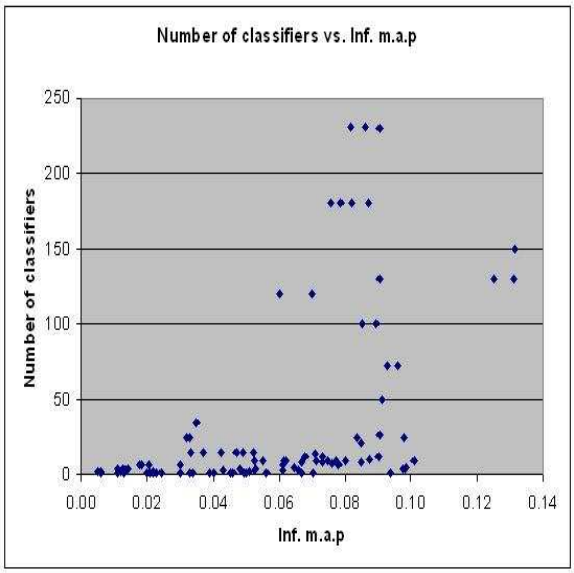


Figure 15: People class vs. MAP

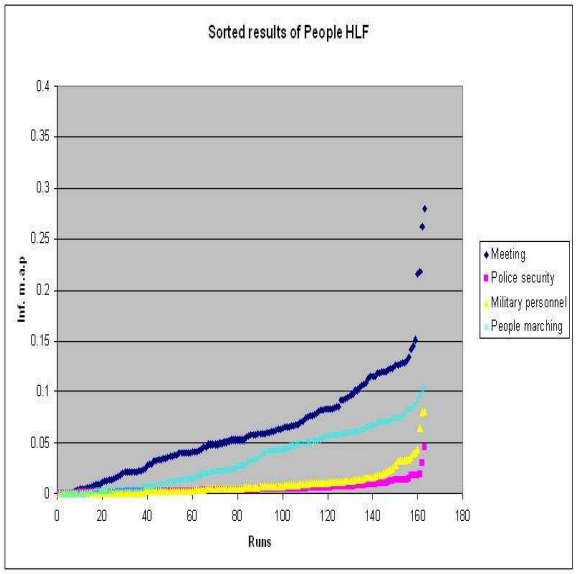
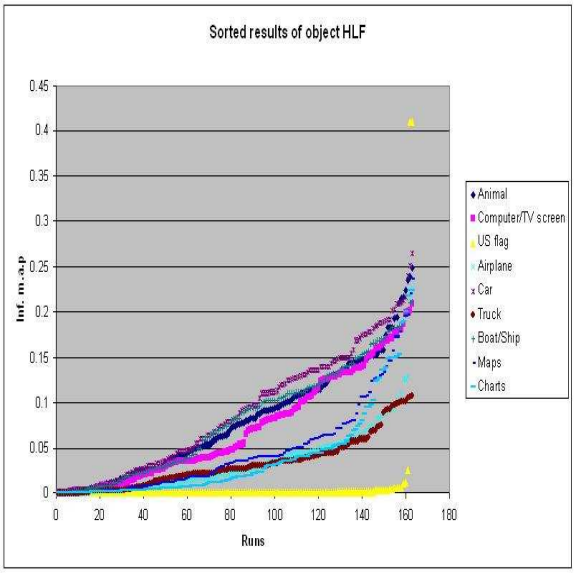


Figure 16: Object class vs. MAP



mance but we can also see that some groups achieved similar results using smaller numbers of classifiers. Also, the increase of classifiers comes at the cost of heavy computations and long running time. Regarding the hardware platform among the participants, we found that most of them used single CPUs while some groups used medium (2 to 16) to large clusters (60 to 284). Finally, the metadata indicated that training time varied between 25 minutes to 25 hours across the different systems while testing time varied between 1 minute to 3 hours.

In a trial to observe the difference in performance among the different feature classes, we divided the features into four main classes namely people, events, location/scene, and objects. Figures 15, 16, 17, and 18 show the sorted results across the 163 runs for each class. Looking at these plots, we can see that the object class got the top performance followed by location then people and finally event. These results suggest that still static features are more easily detectable such as objects and locations, while more dynamic features like people activities and events are more challenging - which is not surprising. More effort is needed toward detecting dynamic features and events which in real life applications can be very important. In brief, there are a lot of factors that affect the final systems performance such as the features used (color, texture, edge, audio,

Figure 17: Location class vs. MAP

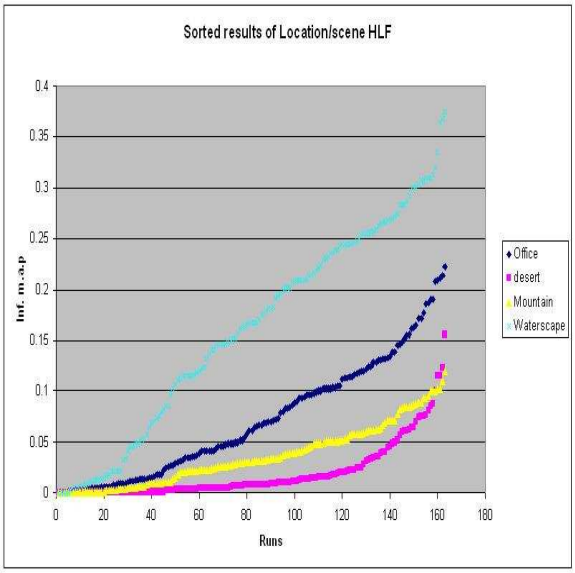
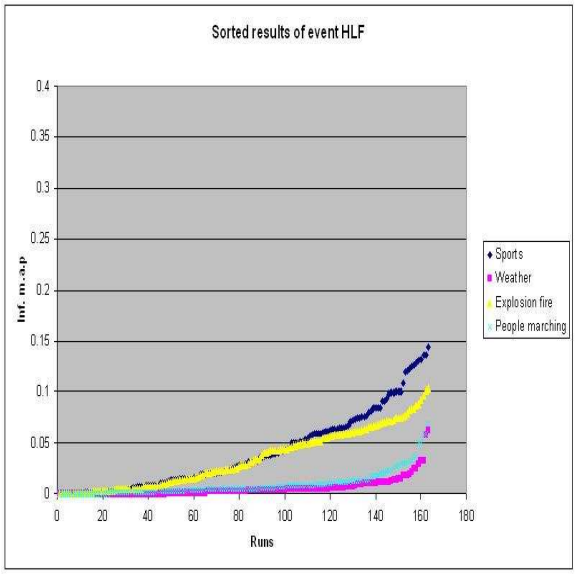


Figure 18: Event class vs. MAP



... etc), the classifiers adopted (SVM, Adaboost, NN, ...etc) each with it's various parameters, fusion methods (late, early,...etc), quality of training data (different sources, frequency of positive examples, annotation methods,...etc), running time, scalability to handle more feature detectors, so on. For detailed information about the approaches and results, the reader should see the various site reports and slides from the workshop available from the publications page of the TRECVID website: www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

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 a need for video — 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. In interactive searchs, the searcher can reformulate the query as many times as time allows, reacting to the results of each search. In automatic runs, the topic is input to the system, which produces a result without any human involvement. Socalled "manual" searches were allowed to provide a simple middle ground between fully automatic and full interactive.

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 allows a human expert in the search system interface, to interpret the topic and create an optimal query in an attempt to make the problem more tractable. 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 own 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.).

Two baseline runs were required of every manual and 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 and one using not text from ASR/MT, as though we were dealing with video in a language for which ASR/MT was not available. The goal of the baseline submissions is to help provide a basis for answering questions such as how much (if any) using visual information (or ASR/MT) helps over just using text in searching.

One participant, FX Palo Alto Laboratory, carried out a new variant of the interactive task, collaborative search, in which the focus was on 2 or more people working synchronously on a query, playing different search roles ("prospector" and "miner") sharing search terms, results, etc.

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 pre-supposed the existence of the sort of very effective video search tool which participants are working to develop.

What was left was to work backwards from the test collection with a number of goals in mind. Rather than attempt to create a representative sample, NIST has in the past tried to get an approximately equal number of each of the basic types (generic/specific and person/thing/event), though in 2006 generic topics dominated over specific ones. The 2007 topics were almost all generic due to the diversity of the collection and the resulting difficulty finding enough examples of named people, objects, events, or places. Generic topics may be more dependent on the visual information than the specific which usually score high on text based (baseline) search performance. Also, the 2007 topics reflect a deliberate emphasis on events.

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 2007 based on Armitage & Enser, 1996, is provided in Table 5. In 2007 all topics are generic and there was a deliberate emphasis on event topics. Figure 19 shows the reasonably varied distribution of number of hits per topic. Topic 199 (" Find shots of a person walking or riding a bicycle") stands out due to an ambiguity that was introduced when "walking or" was inserted. The intended meaning was "Find shots of a person walking a bicycle or riding a bicycle" but the final formulation allowed shots of bicycle-riding people to be included and there are lots of those in the test collection.

Table 5: 2007 Topic types

	Named			Generic		
Topic	Person, thing	Event	Place	Person, thing	Event	Place
197				X	X	
198				X	X	
199				X	X	
200				X	X	
201				X		X
202				X	X	
203				X		X
204				X	X	
205				X	X	
206				X		X
207				X		X
208				X		X
209				X		
210				X	X	
211				X		
212				X	X	
213				X	X	
214				X		
215				X		X
216				X		
217				X	X	
218				X	X	
219	X					
220				X		X

Figure 19: Hits in the test set by topic

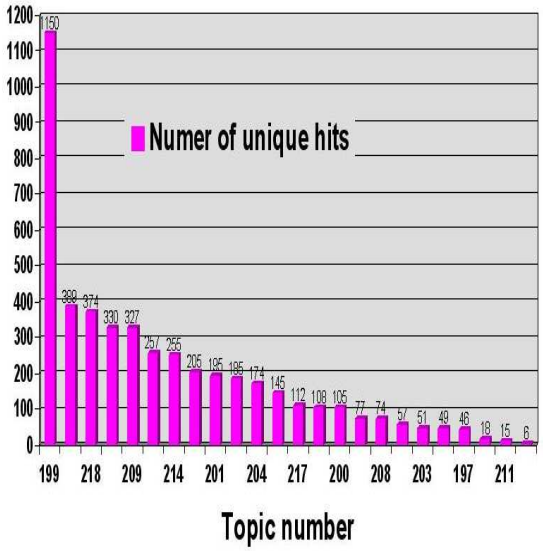


Figure 20: Runs by type

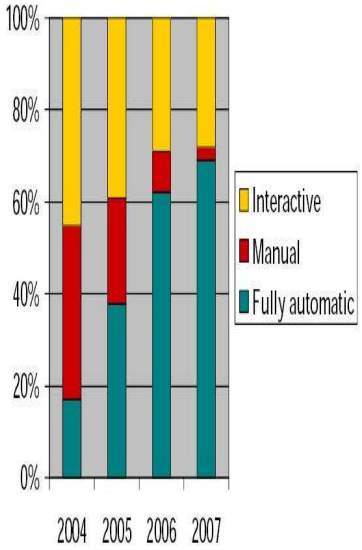


Table 6: Search type statistics

Search type	2004	2005	2006	2007
Fully automatic	17 %	38 %	62 %	69 %
Manually assisted	38 %	23 %	9 %	3 %
Interactive	45 %	39 %	29 %	28 %

5.3 Evaluation

Groups were allowed to submit a total of up to 6 runs of any types in the search task. In fact 24 groups submitted a total of 118 runs - 33 interactive runs, 4 manual ones, and 81 fully automatic ones. The trends seen in 2005 and 2006 leveled off in 2007 as shown in Table 6 and Figure 20.

All submitted runs from each participating group contributed to the evaluation pools. For each topic, all submissions down to a depth of at least 30 (average 84, maximum 160) 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 listening to the audio. The maximum result set depth judged and pooling and judging information for each topic are listed in Table 7 at the end of this paper.

Since simple rankings provide no information about which differences are significant, partial randomization tests were applied to the top 8 automatic and interactive runs to see which differences are not likely to be due to chance ($p < 0.05$). The results of these tests are depicted in Figures 21 and 22.

5.4 Measures

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

5.5 Results

Various tables and figures provide different views of the search results. Figure 23 presents the mean, median, and max results by topic for automatic, manual, and interactive search runs. Figure 24 presents the median MAP scores across all runs by topic. The results for the top 10 runs (sorted by MAP) are presented individually for each type of run in Figures 25, 26, and 27, respectively. The number of uniquely

Figure 23: MAP by topic

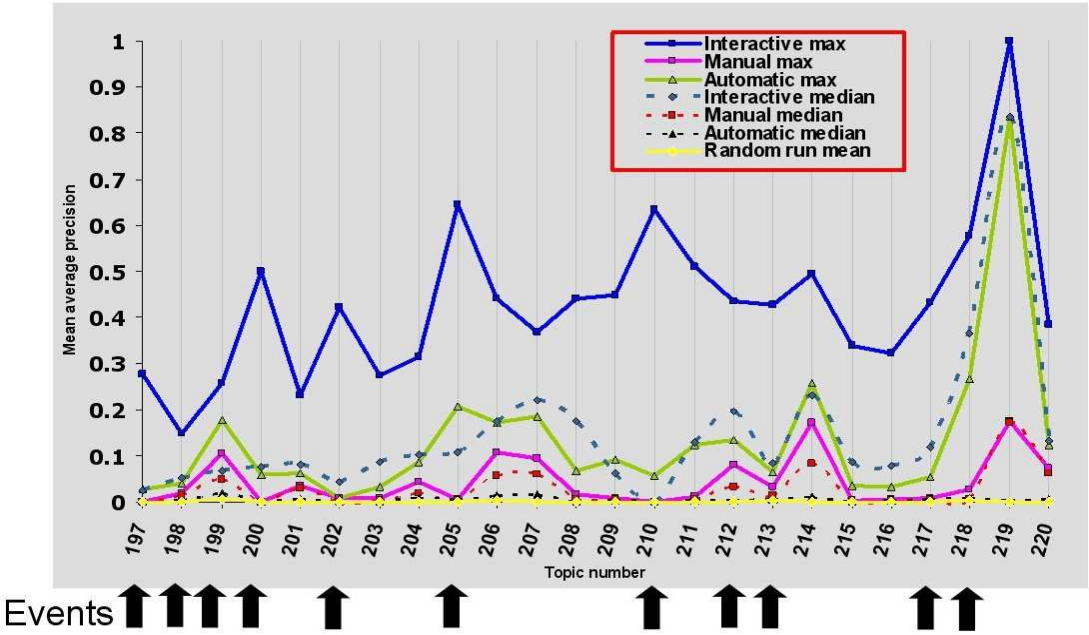


Figure 21: Randomization test on top 8 automatic search runs

Figure 22: Randomization test on top 8 interactive search runs

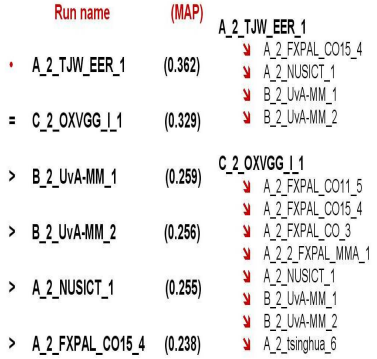


Figure 24: Topics sorted by median MAP

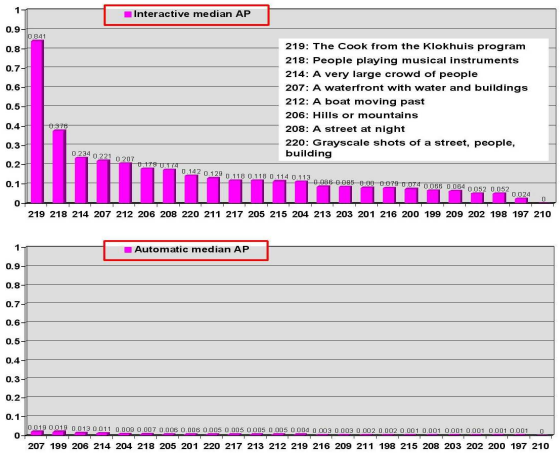


Figure 25: Top 10 automatic search runs

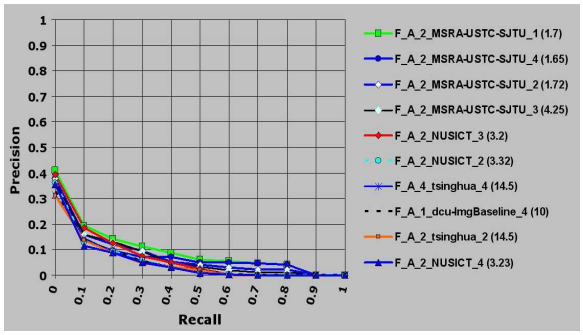


Figure 26: Top 10 manual search runs

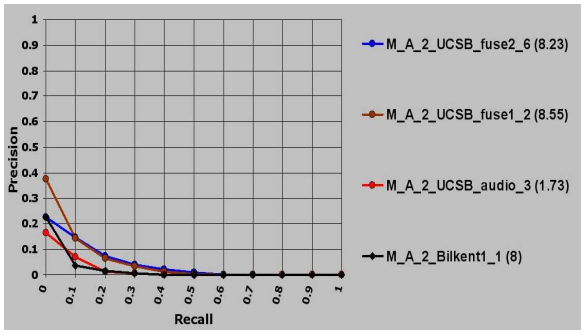


Figure 27: Top 10 interactive search runs

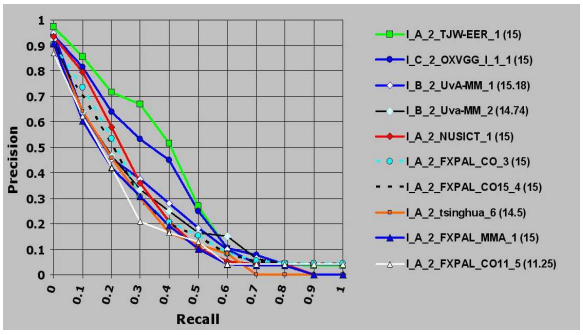
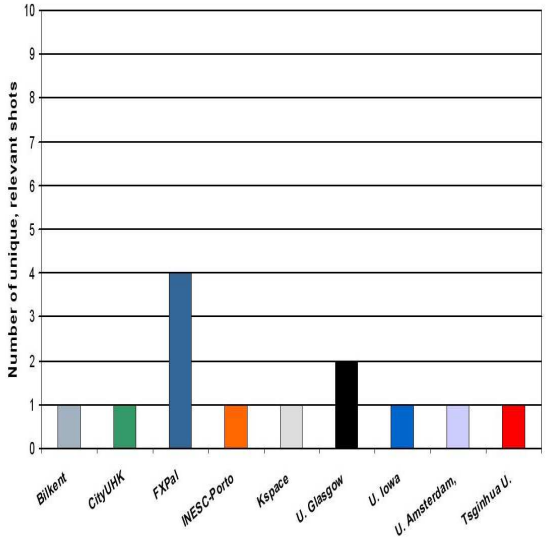


Figure 28: Unique relevant by team



relevant submitted by any given team is very small as shown in Figure 28.

The following section discusses results from the top 10 interactive and automatic systems. Results for automatic runs were in general quite low, due in part perhaps to the quality of the ASR and MT output as well as cross-domain training/testing effects. Another likely cause for low scores was the fact that the 2007 topics emphasized events and deemphasized named entities. Text-only runs were often outperformed by runs using visual or multimodal information.

IBM's automatic search runs combined text-, visual-, and concept-based retrieval. Randomization testing indicated only the run TJW-TMSV-C.qind, which used 50 additional concept detectors trained on external data and a large sample of webpages used to improve the use of WordNet, seems to have produced significant improvement over the text run (TJW-Text). IBM's interactive run achieved relatively good results using a new interactive system, which switched between a tagging and a browsing mode. It should be noted that while the average time spent on a topic was no more than 15 minutes, individual topic search times may have exceeded the limit set in the guidelines.

Oxford University submitted just one search run from a system which used the output of high-level feature analysis for quick object search and provided the user with various tools for expanding the set of desired shots - in the temporal neighborhood, to near duplicates, based on color, texture, etc. Access to external image sources (Google Image) was also provided.

The University of Amsterdam (MediaMill) - The MediaMill team submitted two interactive search runs featuring two different user interfaces: the CrossBrowser (UvA-MM1) and the newer ForkBrowser (UvA-MM2). A partial randomization test confirmed the observation in the team paper that there was no significant difference in the effectiveness of the two interfaces, measured in terms of mean average precision. Of more interest is the lower amount of user interaction required when using the ForkBrowser.

The results from FX Palo Alto provide evidence that the amount of time alloted to the collaborative had a significant effect. A randomization test with 10000 repetitions found the 15 minute collaborative run significantly better ($p < 0.05$) than the 11 minute one, which in turn is significantly better than the 7 minutes one. No significant difference was found by the above test between the best single-user and collaborative runs. The suggestion by the FX Pal team that collaborative searching may be especially suited to difficult searches deserves further study.

Tsinghua submitted automatic runs using various combinations of text-based, example-based, and concept-based search. A randomization test found each of the runs using a combination of approaches was significantly better than the text-based run, but no significant differences ($p < 0.05$) when comparing the combination runs to each other. Randomization testing did not find a significant difference between the interactive run with default options and the one with manually adjusted options.

MSRA-USTC-SJTU's system includes components for query pre-processing, query analysis, unimodal search, multimodal fusion, re-ranking, and result refinement. Randomization testing showed their best run, using fusion of text-based and concept mapping methods performed significantly ($p < 0.05$) better than other non-baseline runs and the baseline.

NUS-ICT focused on query analysis and retrieval using high-level, motion, and visual features. Their query analysis included determination of query class, extraction of high-level features from the query terms and example images, as well as extraction of motion patterns from the query's example video. Randomization tests on NUS-ICT's automatic search runs indicated that fusion using the multimedia query performed significantly ($p < 0.05$) better than fusion using only the text query (with or without motion pattern extraction). All runs using fusion outperform the visual baseline, which outperformed the text baseline.

DCU submitted 6 automatic search runs using only low-level features and ASR/MT. Randomization testing shows finds the run using the baseline visual expert achieved significantly ($p < 0.05$) better results than any of the other runs. All runs performed better than the run using only ASR/MT.

For detailed information about the approaches and results, the reader should see the various site reports and slides from the workshop available from the publications page of the TRECVID website: www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

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 prod-

uct. 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).

In 2005 and 2006 TRECVID sponsored exploratory tasks aimed at investigating rushes management with a focus on how to eliminate redundancy and how to organize rushes in terms of some useful features. For 2007 a pilot evaluation was carried out in which systems created simple video summaries of BBC rushes from several dramatic series compressed to at most 4% of the full video's duration and designed to minimize the number of frames used and present the information in ways that maximized the usability of the summary and speed of objects/event recognition. Summaries of largely scripted video can take advantage of the associated structure and redundancy, which seem to be different for other sorts of rushes, e.g., the travel rushes experimented with in 2005/6.

Such a summary could be returned with each video found by a video search engine which is similar to text search engines when they return short lists of keywords (in context) for each document found - to help the searcher decide whether to explore a given item further without viewing the whole item. Alternatively it might be input to a larger system for filtering, exploring and managing rushes data.

Although in this pilot task the notion of visual summary was limited to a single clip to be evaluated using simple play and pause controls, there was still room for creativity in generating the summary. Summaries need not have been series of frames taken directly from the video to be summarized and presented in the same order. Summaries could contain picture-in-picture, split screens, and results of other techniques for organizing the summary. Such approaches raised interesting questions of usability.

For practical reasons in planning the assessment an upper limit on the size of the summaries was needed. Different use scenarios could motivate different limits. One might involve passing the summary to downstream applications that support, clustering, filtering, sophisticated browsing for rushes exploration, management, reuse. There was minimal emphasis on compression.

Assuming the summary should be directly usable by a human, then at least it should be usable by a professional, looking for reusable material, and willing to watch a summary longer than someone with more recreational goals.

Therefore longer summaries than a recreational user would tolerate were allowed but results were scored so that systems that could meet a higher goal (much shorter summary) could be identified, Each submitted summary had a duration which was at most 4% of the video to be summarized. That gave a mean maximum summary duration of 60 seconds with a range from 7 - 87 seconds).

6.1 Data

The BBC Archive provided about 300 Beta-SP tapes, which NIST had read in and converted to MPEG-2. NIST then transcoded the MPEG-2 files to MPEG-1. Ground truth was created by Dublin City University for about half of the development clips and all the test data.

6.2 Evaluation

At NIST, all the summary clips for a given video were viewed using mplayer on Linux in a window 125mm x 102mm @ 25 fps in a randomized order by a single human judge. In a timed process, the judge played and/or paused the video as needed to determine as quickly as possible which of the segments listed in the ground truth for the video to be summarized are present in the summary.

The judge was also asked to assess the usability/quality of the summary. This included answering the following two questions with 5 possible answers for each - where only the extremes are labeled: "Strongly agree" and "strongly disagree".

1. It is easy to see and understand what is in this summary.

2. This summary contains more video of the desired segments than was needed.

This process was repeated for each test video. Each summary was evaluated by three judges.

The output of two baseline systems was provided by the Carnegie Mellon University team. One was

a uniform sample baseline within the 4% maximum. The other was based on a sample within the 4% maximum from clusters built on the basis of a simple color histogram.

6.3 Measures

Per-summary measures were:

- fraction of the ground truth segments found in the summary
- time (in seconds) needed to check summary against ground truth
- number of frames in the summary
- system time (in seconds) to generate the summary
- usability scores

Per-system measures were the means of the per-summary measures over all test videos.

6.4 Results

A detailed discussion of the results is available in the workshop papers as part of the ACM Digital Library. See (Over et al., 2007) for an introduction and overview. Slides from the workshop are available from the TRECVID video summarization workshop page at www-nlpir.nist.gov/projects/tv7.acmmm.

7 Summing up and moving on

This introduction to TRECVID 2007 has provided basic information on the goals, data, evaluation mechanisms and metrics used. Further details about each particular group's approach and performance can be found in that group's site report - available from the publications page of the TRECVID website: www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

8 Authors' note

TRECVID would not happen without support from IARPA and NIST and 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, to Peter Wilkins at the Centre for Digital Video Processing at Dublin City University (DCU) for formating the master shot reference definition and to Phil Kelly also at Dublin City University (DCU) for co-ordinating the creation of the summarization ground truth.

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

Roeland Ordelman and Marijn Huijbregts at the University of Twente provided the output of an automatic speech recognition system (Huijbregts, Ordelman, & Jong, 2007) run on the Sound and Vision data.

Christof Monz of Queen Mary, University London contributed machine translation (Dutch to English) for the Sound and Vision video.

Georges Quénot and Stéphane Ayache of LIG (Laboratoire d'Informatique de Grenoble, formerly CLIPS-IMAG) organized a collaborative annotation and more than two dozen groups contributed to that effort (Ayache & Quénot, 2007).

The Multimedia Content Group at the Chinese Academy of Sciences together with the National University of Singapore provided full annotation of the 2007 training data (using one keyframe per shot).

Carnegie Mellon University created two baseline summarization runs to help put the summarization results in context.

Shih-Fu Chang at Columbia University made available the models and features they used in detecting 374 LSCOM concepts.

Yu-Gang Jiang at City University Hong Kong donated 374 LSCOM concept detectors (SVM detectors of local feature, color and texture separately).

Once again we appreciate Jonathan Lasko's careful creation of the shot boundary truth data once again - his seventh and final year doing this work.

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.

0197 Find shots of one or more people walking up stairs (I/2, V/6, R/46)

0198 Find shots of a door being opened (I/0, V/7, R/185)

0199 Find shots of a person walking or riding a bicycle (I/2, V/4, R/1150)

0200 Find shots of hands at a keyboard typing or using a mouse (I/3, V/7, R/105)

0201 Find shots of a canal, river, or stream with some of both banks visible (I/4, V/6, R/195)

0202 Find shots of a person talking on a telephone (I/3, V/5, R/49)

0203 Find shots of a street market scene (I/3, V/4, R/51)

0204 Find shots of a street protest or parade (I/4, V/4, R/174)

0205 Find shots of a train in motion (I/3, V/7, R/108)

0206 Find shots with hills or mountains visible (I/4, V/9, R/330)

0207 Find shots of waterfront with water and buildings (I/4, V/3, R/257)

0208 Find shots of a street at night (I/4, V/7, R/74)

0209 Find shots with 3 or more people sitting at a table (I/4, V/4, R/327)

0210 Find shots with one or more people walking with one or more dogs (I/4, V/5, R/18)

0211 Find shots with sheep or goats (I/4, V/4, R/15)

0212 Find shots in which a boat moves past (I/4, V/4, R/77)

0213 Find shots of a woman talking toward the camera in an interview - no other people visible (I/0, V/6, R/389)

0214 Find shots of a very large crowd of people (fills more than half of field of view) (I/4, V/4, R/255)

0215 Find shots of a classroom scene with one or more students (I/4, V/6, R/145)

0216 Find shots of a bridge (I/5, V/5, R/57)

0217 Find shots of a road taken from a moving vehicle through the front windshield (I/0, V/5, R/112)

0218 Find shots of one or more people playing musical instruments such as drums, guitar, flute, keyboard, piano, etc. (I/3, V/10, R/374)

0219 Find shots that contain the Cook character in the Klokhuis series (I/1, V/4, R/6)

0220 Find grayscale shots of a street with one or more buildings and one or more people (I/4, V/6, R/205)

10 Appendix B: Features

1 Sports: Shots depicting any sport in action

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

22 DROPPED - 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.

Ayache, S., & Quénot, G. (2007). *TRECVID 2007 Collaborative Annotation using Active Learning.* URL: http://clips.imag.fr/mrim/georges.quenot/articles/trec07.pdf.

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

Huijbregts, M., Ordelman, R., & Jong, F. de. (2007). Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the second international conference on semantics and digital media technologies (SAMT).* Berlin: Springer Verlag.

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

Manly, B. F. J. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology* (2nd ed.). London, UK: Chapman & Hall.

Over, P., Ianeva, T., Kraaij, W., & Smeaton, A. F. (2006). *TRECVID 2006 Overview.* URL: http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6overview.pdf.

Over, P., Smeaton, A. F., & Kelly, P. (2007). The TRECVID 2007 BBC rushes summarization evaluation pilot. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization* (pp. 1–15). New York, NY, USA: ACM Press.

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			
Asahi Kasei Corporation	Japan	SB	**	–	–
AT&T Labs	USA	SB	–	–	SU
Beijing Jiaotong University (Northern Jiaotong Univ.)	China	–	–	SE	–
Beijing University of Posts and Telecommunications	China	SB	–	–	–
Bilkent University	Turkey	**	FE	SE	**
Brno University of Technology	Czech Republic	SB	FE	**	SU
Carnegie Mellon University	USA	–	**	**	SU
City University of Hong Kong (CityU)	China	–	FE	SE	SU
Columbia University	USA	–	FE	**	SU
COST292 Team	EU	SB	FE	SE	SU
Curtin University	Australia	**	–	–	SU
CWI-CTIT-UTwente team	Netherlands	–	**	SE	–
Dublin City University	Ireland	–	–	SE	SU
École Nationale Supérieure des Télécommunications / TSI	France	–	FE	–	–
Etter Solutions Research Group	USA	–	–	SE	–
Florida International University, FIU-UM	USA	SB	**	–	–
Fraunhofer Institute IAIS and University of Bradford	EU	SB	**	–	–
Fudan University	China	–	FE	SE	–
FX Palo Alto Laboratory Inc.	USA	**	**	SE	SU
Helsinki University of Technology	Finland	**	FE	SE	SU
Huazhong University of Science and Technology	China	SB	**	**	**
IBM T. J. Watson Research Center	USA	**	FE	SE	**
Institute for Systems and Computer Engineering of Porto	Portugal	–	**	SE	–
Institut EURECOM	France	–	FE	–	SU
JOANNEUM RESEARCH Forschungsgesellschaft mbH	Austria	**	FE	–	SU
KDDI R&D Labs, Inc., Tokushima U., Tokyo U	Japan	**	FE	–	SU
K-Space	EU	–	FE	SE	–
LIG (Laboratoire d'Informatique de Grenoble)	France	SB	FE	**	**
LIP6 - Laboratoire d'Informatique de Paris 6	France	–	FE	–	SU
MSRA-USTC-SJTU Team (Microsoft Research Asia- ...)	China	–	FE	SE	**
Multimedia Content Analysis Group (CAS)	China	–	FE	–	–
Multimedia Computing Group (CAS) / National University of Singapore	China,Singapore	–	FE	SE	**
National Institute of Informatics	Japan	–	FE	–	SU
National Taiwan University	Taiwan	–	FE	**	SU
NHK Science and Technical Research Laboratories	Japan	SB	**	–	–
Oxford University	UK	–	FE	SE	–
Philipps University Marburg	Germany	SB	FE	**	**
The Hong Kong Polytechnic University	China	–	–	–	SU
Tokyo Institute of Technology	Japan	**	FE	**	**
Tsinghua University / Intel China Research Center	China	SB	FE	SE	SU
Universidad Autónoma de Madrid	Spain	–	**	–	SU
Universidy of Jaén (SINAI)	Spain	–	–	SE	–
University of Karlsruhe (TH)	Germany	SB	FE	–	–
University of Amsterdam (MediaMill Team)	Netherlands	–	FE	SE	–
University of California, Berkeley	USA	–	FE	**	–
University of California, Santa Barbara	USA	–	FE	SE	SU
University of Central Florida	USA	–	FE	SE	**
University of Electro-Communications	Japan	–	FE	**	–
University of Glasgow	UK	–	–	SE	SU
University of Iowa	USA	**	FE	SE	–
University of Louisville	USA	–	FE	–	–
University of Modena and Reggio Emilia (Italy)	Italy	SB	**	–	**
University of Queensland	Australia	–	–	SE	–
University of Sheffield	UK	SB	–	–	SU

Task legend. SB: Shot boundary; FE: High-level features; SE: Search; SU: Rushes summarization; **: no runs

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	293764	17453	5.9	150	3296	18.9	124	3.8
3	284814	17296	6.1	170	3360	19.4	6	0.2
5	289509	17173	5.9	150	3289	19.2	210	6.4
6	291522	17324	5.9	120	3319	19.2	707	21.3
10	290028	17476	6.0	140	3298	18.9	26	0.8
12	295266	17377	5.9	180	3311	19.1	96	2.9
17	299160	16900	5.6	240	3249	19.2	289	8.9
23	288896	17547	6.1	100	3239	18.5	89	2.7
24	292336	17507	6.0	120	3373	19.3	41	1.2
26	298252	17410	5.8	160	3235	18.6	251	7.8
27	290991	17387	6.0	140	3282	18.9	206	6.3
28	281010	17503	6.2	130	3370	19.3	6	0.2
29	287745	17487	6.1	150	3287	18.8	147	4.5
30	295604	17393	5.9	140	3283	18.9	435	13.3
32	289844	17408	6.0	140	3409	19.6	216	6.3
33	289285	17185	5.9	190	3318	19.3	166	5.0
35	292668	17210	5.9	180	3328	19.3	72	2.2
36	288378	17484	6.1	120	3359	19.2	52	1.5
38	284727	17434	6.1	170	3354	19.2	93	2.8
39	281735	17386	6.2	190	3334	19.2	64	1.9

Table 7: 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
197	117593	17815	15.1	40	2324	13.0	46	2.0
198	114535	17709	15.5	80	3992	22.5	185	4.6
199	112646	17360	15.4	100	4606	26.5	1150	25.0
200	112500	17721	15.8	70	3847	21.7	105	2.7
201	113076	16733	14.8	90	3836	22.9	195	5.1
202	113519	17432	15.4	30	1887	10.8	49	2.6
203	114586	17308	15.1	50	2454	14.2	51	2.1
204	113660	16902	14.9	100	4020	23.8	174	4.3
205	112851	16935	15.0	120	4834	28.5	108	2.2
206	110890	16613	15.0	160	5406	32.5	330	6.1
207	114965	15536	13.5	80	2991	19.3	257	8.6
208	114017	16733	14.7	60	2926	17.5	74	2.5
209	117016	17393	14.9	100	5044	29.0	327	6.5
210	116346	17624	15.1	60	3095	17.6	18	0.6
211	110253	16810	15.2	70	3115	18.5	15	0.5
212	113930	16771	14.7	100	3600	21.5	77	2.1
213	116373	17129	14.7	70	3485	20.3	389	11.2
214	118236	16798	14.2	70	3050	18.2	255	8.4
215	111850	17492	15.6	130	5976	34.2	145	2.4
216	111714	16930	15.2	70	3265	19.3	57	1.7
217	114875	17606	15.3	100	4755	27.0	112	2.4
218	117674	17517	14.9	80	4129	23.6	374	9.1
219	111948	17688	15.8	30	1768	10.0	6	0.3
220	118279	16132	13.6	150	5147	31.9	205	4.0

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

Participants	Country	SB	FE	SE	RU
AIIA Laboratory	Greece	**	–	–	–
Artificialife	Canada	–	–	–	**
Chinese University of Hong Kong	China	**	**	**	**
ETIS Laboratory	France	**	**	**	**
INRIA	France	–	**	–	–
IRISA/INRIA Rennes - TEXMEX team F218	France	**	–	**	–
Johns Hopkins University	USA	–	**	–	–
Massachusetts Institute of Technology	USA	**	**	**	**
RMIT University School of CS&IT	Australia	**	–	**	**
RWTH Aachen University	Germany	**	**	–	–
Technical University Berlin	Germany	**	–	–	–
The Open University	UK	**	**	**	–
University Rey Juan Carlos	Spain	**	–	**	**
University of California, San Diego	US	–	**	**	**
University of Kocaeli	Turkey	**	–	–	–
U. of North Carolina at Chapel Hill	USA	–	**	–	**
University of Trieste	Italy	–	**	**	–

Task legend. SB: Shot boundary; FE: High-level features; SE: Search; RU: BBC rushes summarization; **: Group applied but didn't submit any runs