TRECVID 2008 –Goals, Tasks, Data, Evaluation Mechanisms and Metrics

Paul Over {over@nist.gov}
George Awad {gawad@nist.gov}
Travis Rose {travis.rose@nist.gov}
Jon Fiscus {jfiscus@nist.gov}
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
Radboud University Nijmegen
Nijmegen, the Netherlands

Alan F. Smeaton {Alan.Smeaton@dcu.ie}
CLARITY: Centre for Sensor Web Technologies / Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland

April 24, 2009

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2008 is a Text REtrieval Conference (TREC)-style video analysis and retrieval evaluation, the goal of which remains to promote progress in content-based exploitation of digital video via open, metrics-based evaluation. Over the last 7 years this effort has yielded a better understanding of how systems can effectively accomplish such processing and how one can reliably benchmark their performance. TRECVID is funded by the Intelligence Advanced Research Projects Activity (IARPA), the US Department of Homeland Security (DHS), and the US National Institute of Standards and Technology (NIST).

77 teams (see Table 1) from various research organizations — 24 from Asia, 39 from Europe, 13 from North America, and 1 from Australia — participated in one or more of five tasks: high-level feature (HLF) extraction, search (fully automatic, manually assisted, or interactive), pre-production video (rushes) summarization, copy detection, or surveillance event detection. The copy detection and surveillance event detection tasks are being run for the first time in TRECVID. Figure 1 depicts the evolution of data, tasks, and participation in TRECVID.

2008 was the second year in what may be a 3-year cycle using new data sources for feature extraction and search, data which is related to the broadcast television news used in 2003-2006 but significantly different. Test data for the search and feature tasks was about 100 hours of (MPEG-1 - a standard developed by the Motion Picture Experts Group) television news magazine, science news, news reports, documentaries, educational programming, and archival video almost entirely in Dutch from the Netherlands Institute for Sound and Vision. An equal amount of video was available for search/feature system development. The combined 200 hours were used in the copy detection task. The British Broadcasting Cor-

poration (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 and in part for copy detection. About 100 hours of surveillance video from the London Gatwick International Airport was provided by the United Kingdom (UK) Home Office for use in the event detection task.

Results were scored by NIST mostly against human judgments. Feature and search submissions were evaluated based on partial manual judgments of the pooled submissions. The output of summarization systems was manually evaluated at Dublin City University using ground truth manually created at NIST. Full results for the summarization task were presented and discussed as the TRECVID Video Summarization Workshop at the Association for Computer Machinery (ACM) Multimedia Conference in Vancouver, BC, Canada on October 31, 2008. Copy detection submissions were evaluated at NIST based on ground truth created automatically using tools donated by the INRIA-IMEDIA group. NIST evaluated the surveillance event detection results using ground truth created manually under contract by the Linguistic Data Consortium

This paper is an introduction to the evaluation framework — the tasks, data, and measures. It also provides an overview of the results. For the details of the approaches taken by the participating groups, their hypotheses, and conclusions please see the notebook papers available on 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.

2 Data

2.1 Video

Sound and Vision data

The Netherlands Institute for Sound and Vision generously provided 400 hours of television 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 — half for development and half for evaluation of feature extraction and search systems. All the 2007 data was available for system development in 2008. An additional 100 hours were used for evaluation in TRECVID 2008

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

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 35,766 reference shots.

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's automatic speech recognition (ASR).

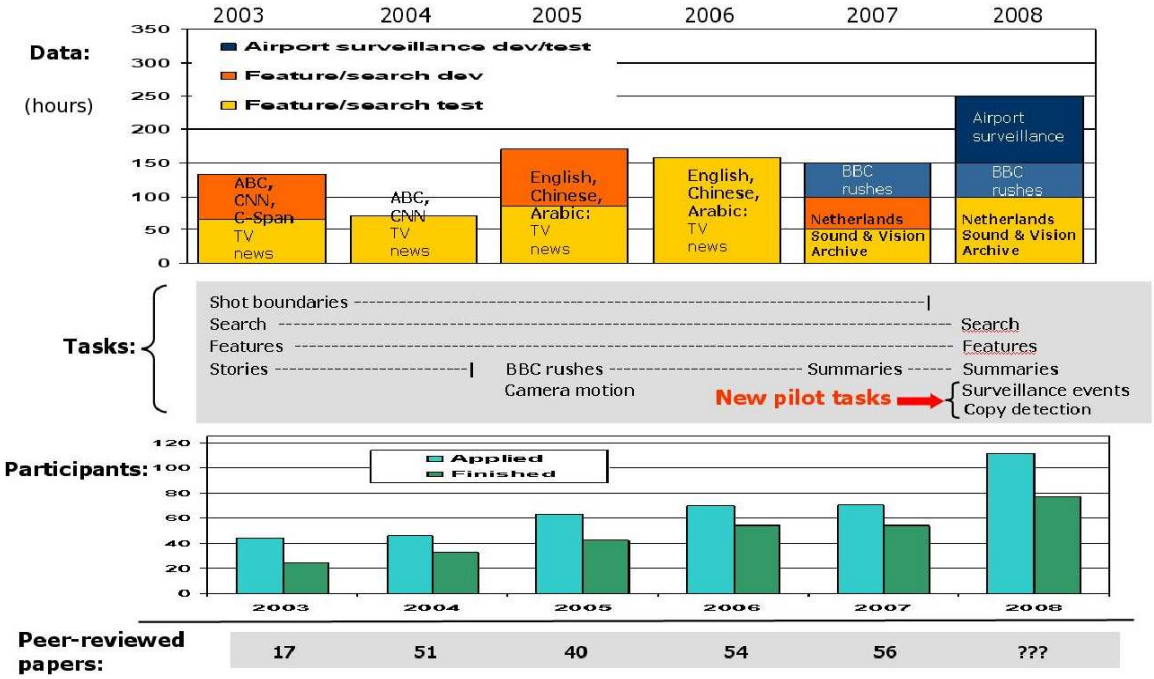
BBC Archive data - rushes

The BBC Archive provided rushes video for use in the video summarization task. The material 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. About 35 hours (57 clips), with associated ground truth and automatic summaries for half of that, were available for system development. About 18 hours (40 clips) were reserved for system evaluation.

Gatwick Airport surveillance video

The UK Home Office provided about 100 hours (10 days $\times$ 2 hours/day $\times$ 5 cameras) of surveillance video from London's Gatwick International Airport. The video was annotated for a set of 10 events. About half was distributed as development data/annotation and half reserved for evaluation.

Figure 1: Evolution of TRECVID



2.2 Common feature annotation

Georges Quénot and Stéphane Ayache of LIG (Laboratoire d'Informatique de Grenoble, formerly CLIPS-IMAG) organized a collaborative annotation of 20 features in the TRECVID 2008 search/feature development data using an active learning scheme designed to improve the efficiency of the process. About 40 groups created 1.2 million image × concept annotations and shared the resulting ground truth among themselves (Ayache & Quénot, 2008).

The Multimedia Computing Group at the Chinese Academy of Sciences together with the National University of Singapore provided full annotation for 20 features of the 2008 training data.

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.

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

There continued to be special interest in how well feature/search 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 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 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 20 features were drawn from the Large Scale Ontology for Multimedia (LSCOM) feature set so as to be appropriate to the Sound and Vision data. Some feature definitions were enhanced for greater clarity, so it is important that the TRECVID feature descriptions be used and not the LSCOM descriptions.

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. All 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 2008 were as follows and are numbered 1-20. All were evaluated. [1] Classrooms, [2] Bridge, [3] Emergency Vehicle, [4] Dog, [5] Kitchen, [6] Airplane flying, [7] Two people, [8] Bus, [9] Driver, [10] Cityscape, [11] Harbor, [12] Telephone, [13] Street, [14] Demonstration Or Protest, [15] Hand, [16] Mountain, [17] Nighttime, [18] Boat Ship, [19] Flower, [20] Singing.

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

3.1 Data

As mentioned earlier, the feature test collection contained 219 files/videos and 35,766 reference shots, but four test files were ignored in the testing due to problems displaying shots from these long files (BG_36684, BG_37970, BG_38162, BG_8887) in the assessment system. Removing these files left 215 files and 33,726 shots. Testing feature extraction and search on the same data offered the opportunity to assess the quality of features being used in search.

3.2 Evaluation

Each group was allowed to submit up to 6 runs and in fact 43 groups submitted a total of 200 runs.

For each feature, all submissions down to a depth of at least 90 (average 129, maximum 220) result items

Figure 4: Effectiveness of category a runs

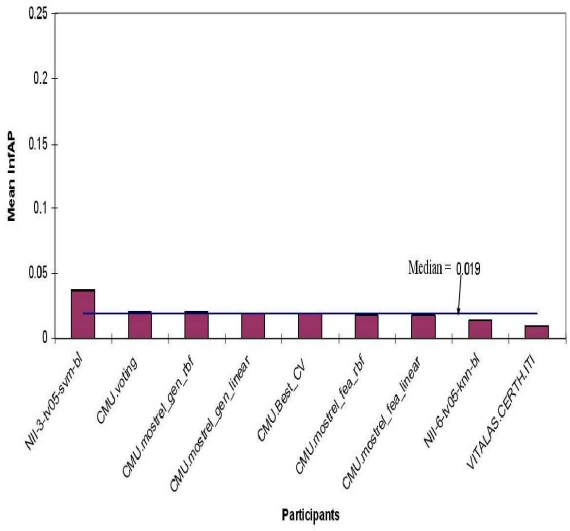
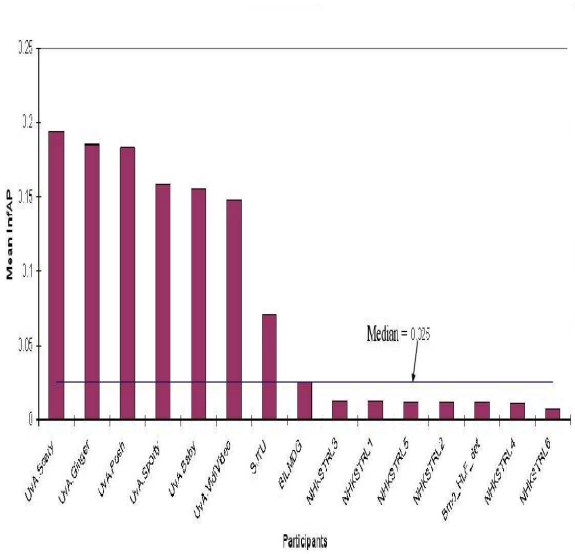


Figure 5: Effectiveness of category B runs



(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 3. In all, 67774 feature-shot pairs were judged.

3.3 Measures

The *trec_eval* software, a tool used in the main TREC activity since it started in 1992, was used to calculate recall, precision, inferred average precision, etc., for each result.(See http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools.) 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".

3.4 Results

Figures 2 and 3 present an overview of the results from runs of category A. While systems that submitted runs for category A training are the most popular, more interest is noticed for special training data of category B and C and unrelated data of category

Figure 6: Effectiveness of category C runs

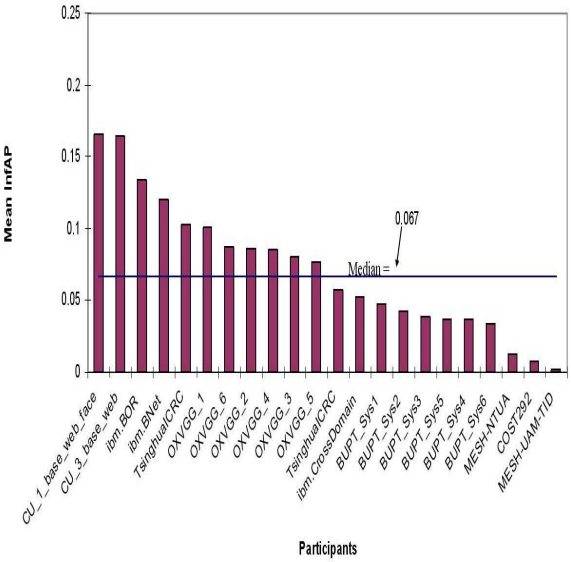


Figure 2: infAP by run - top half

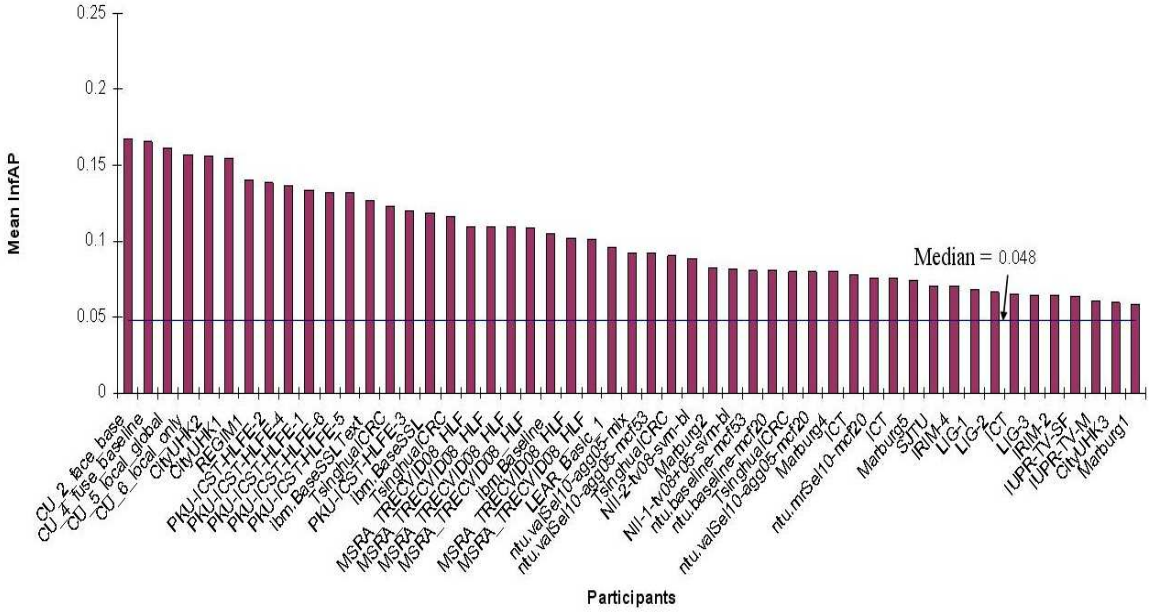


Figure 3: infAP by run - bottom half

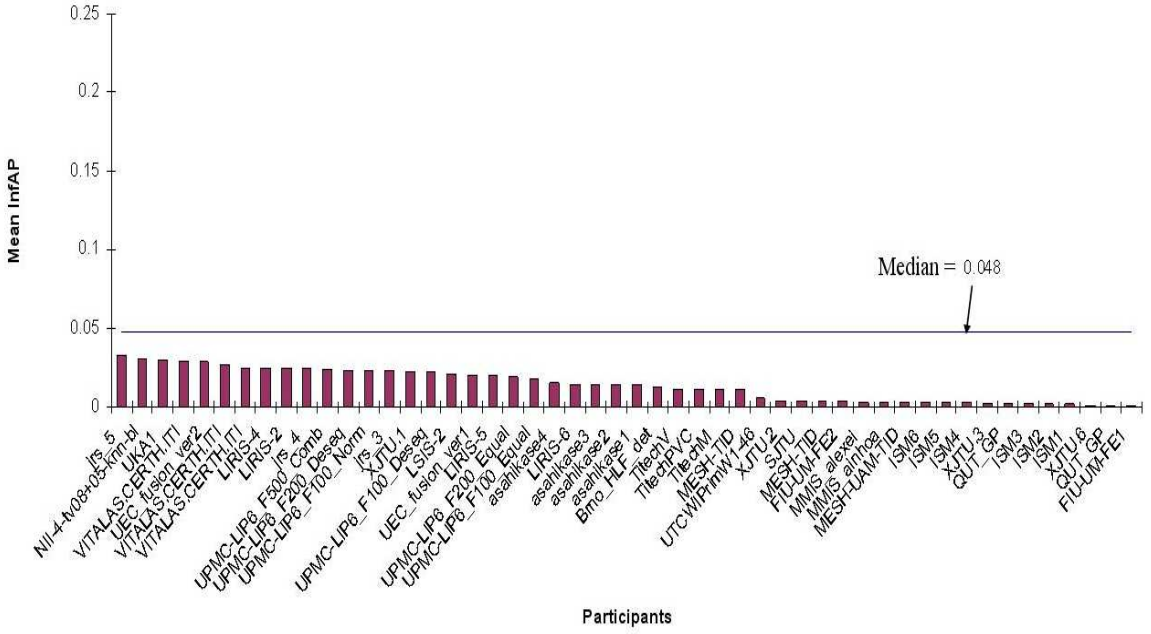


Figure 7: Effectiveness of category c runs

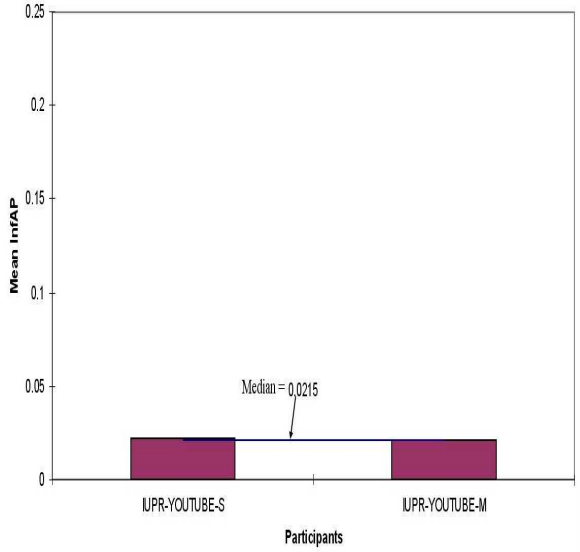


Figure 8: Frequencies of shots with each feature

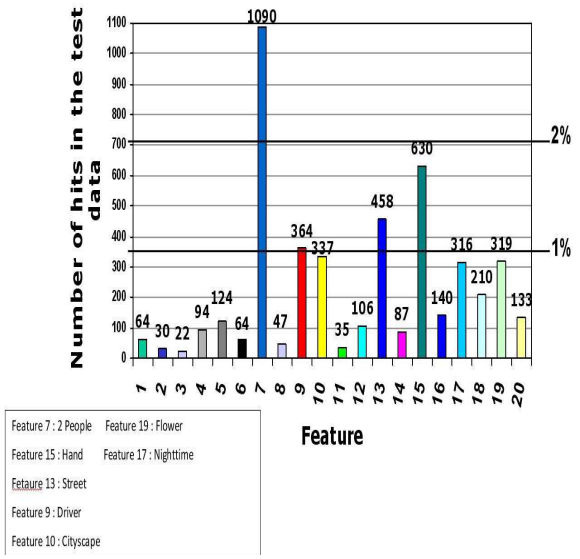


Figure 9: Top 10 runs (infAP) by feature

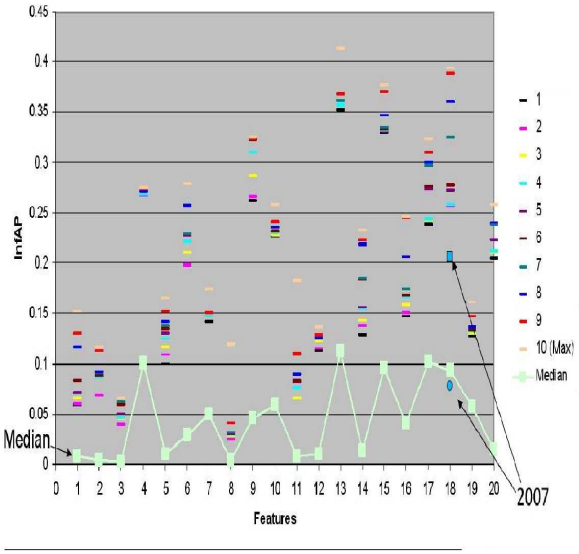


Figure 10: Effectiveness versus number of hits

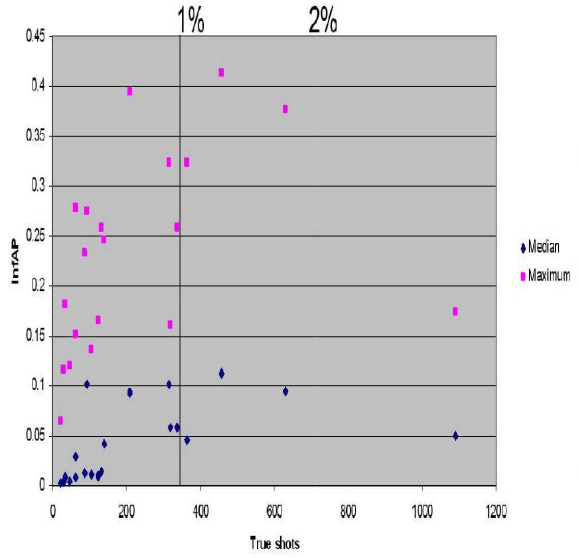


Figure 11: Significant differences among top 10 A-category runs

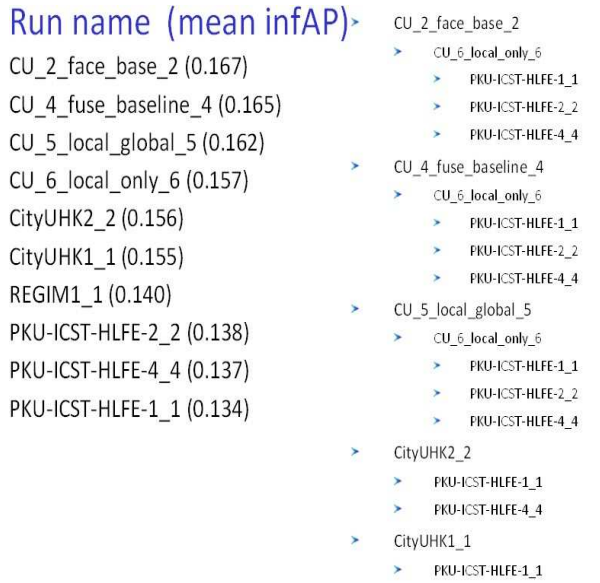


Figure 12: Significant differences among top 10 a-category runs

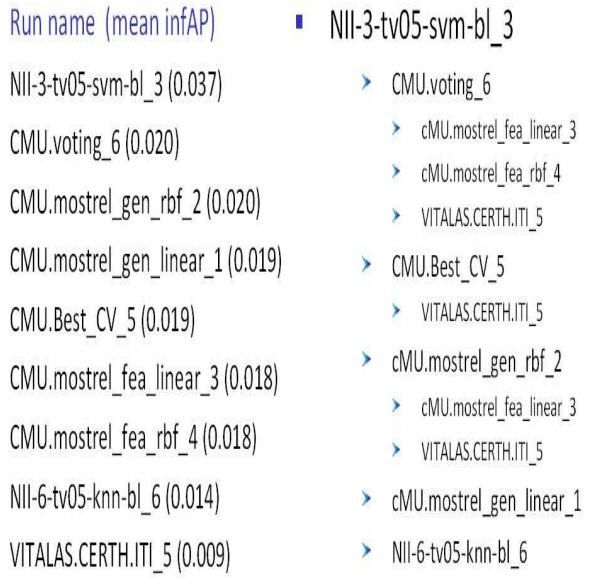


Figure 13: Significant differences among top 10 B-category runs

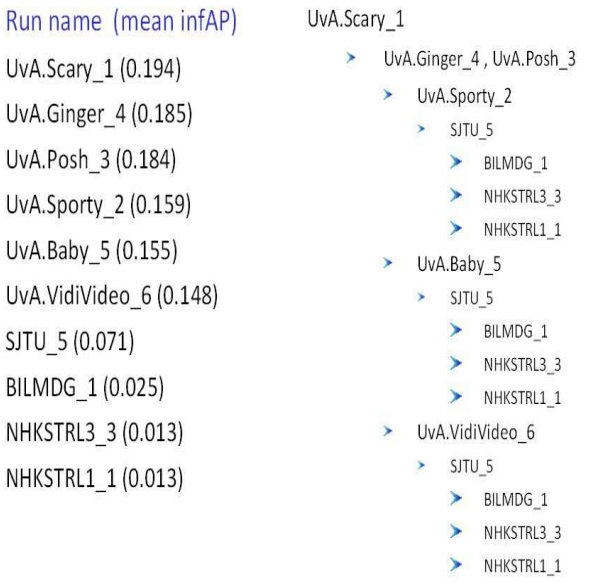


Figure 14: Significant differences among top 10 C-category runs

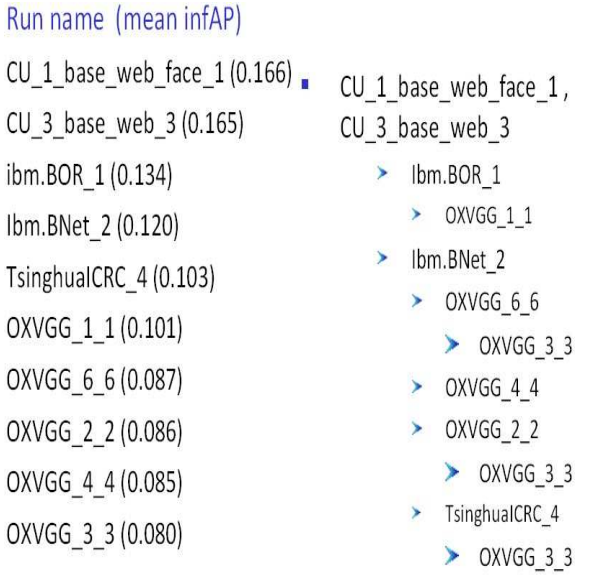


Figure 15: Significant differences among top 10 A/a-category runs by group

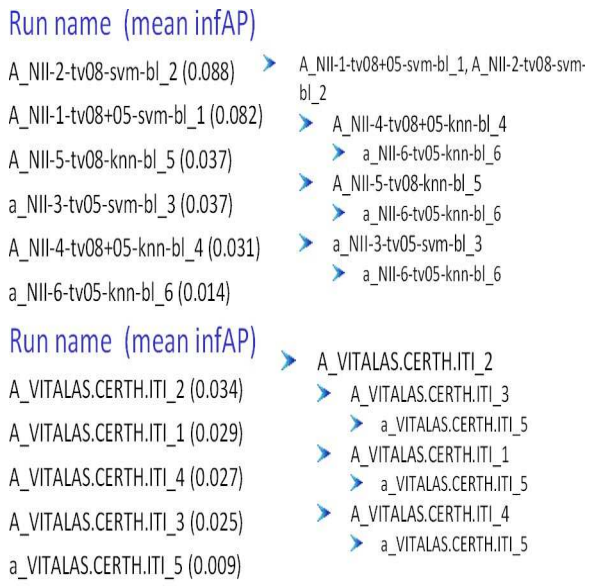
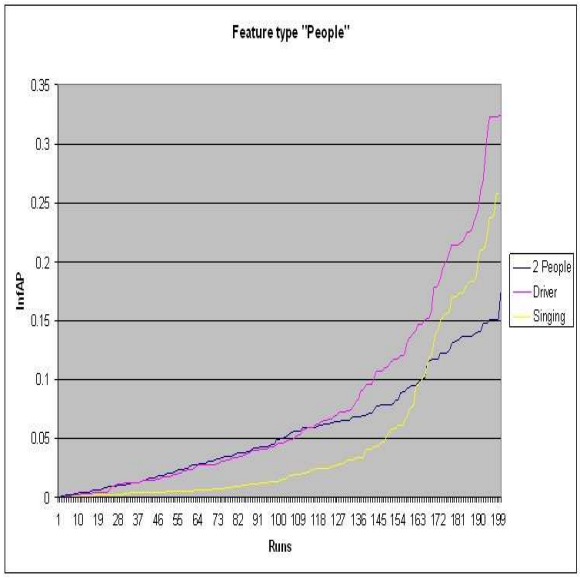


Figure 16: Features of category "People"



a and c. Figures 4 through 7 present the other training categories performance. For the first year, category B runs achieve higher performance than category A runs. Also, category C runs (using data from Flickr, Youtube and Peekaboom) are on par with A runs. Performance varies greatly by feature. Figure 8 shows how unique instances were found for each tested feature. Four features (2 people, driver, street, and hand) exceeded 1% hits from the total tested shots percentage, while only the "people" feature exceeded 3%. On the other hand, features that had lowest hits were "Emergency-vehicle", "bridge", "bus", and "harbor". It can also be shown that other features such as "Cityscape", "Flower" and "Nighttime" received hits very near to the 1%. Two features "Mountain" and "boat-ship" in TRECVID 2007 were in common with TRECVID 2006 HLF task. We realized that the number of hits have increased for Mountain from 96 in last year to 140 this year and for boat-ship increased from 166 last year to 210 this year. The increase of hits for the same features across successive years indicates that systems are becoming more mature and familiar with how to handle those features. Figure 9 shows the performance of the top 10 teams across the 20 features. The behavior varies generally across features. For example some features reflect big spread between the scores of the top 10 such as feature "boat-ship", "demonstration-or-protest","classroom", and "mountain" indicating that there is still room for further improvement, while other features had tight spread of scores among the top 10 such as feature "dog" followed by less tight spread as in features "emergency-vehicle" and "telephone". In general, the median scores ranged between 0.003 (feature "Emergency-Vehicle") and 0.113 (feature street). Figure 10 shows a weak positive correlation between number of hits possible for a feature and the median or maximum score for that feature. To test if there are significant differences between the systems performance, we applied a randomization test (Manly, 1997) on the top 10 runs for each run category as shown in Figures 11 through 15. 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. Figure 15 applies the randomization test on runs that used sound and vision data vs runs that did not use sound and vision data for training across same teams.

Figures 16 through 19 show the performance of the submitted runs for each of 4 main features categories. We divided the 20 tested features into features that represent people, objects, events, and locations. It seems that for each of the categories we can find a set of easy and hard features. For example, in the "object" category we can see that features like Bus

Figure 17: Features of category "Location"

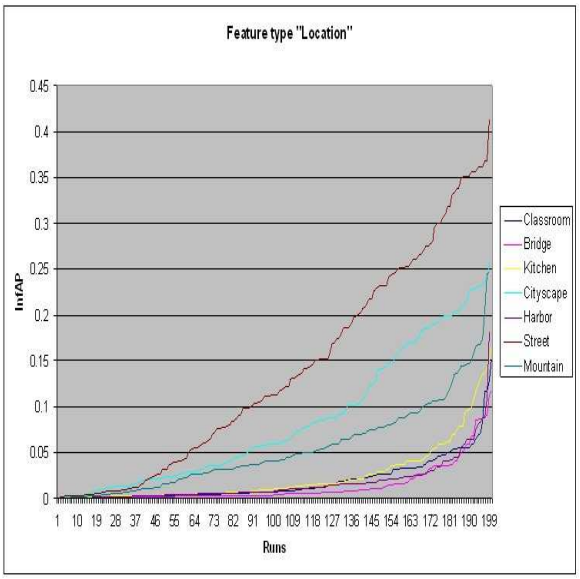


Figure 18: Features of category "Event"

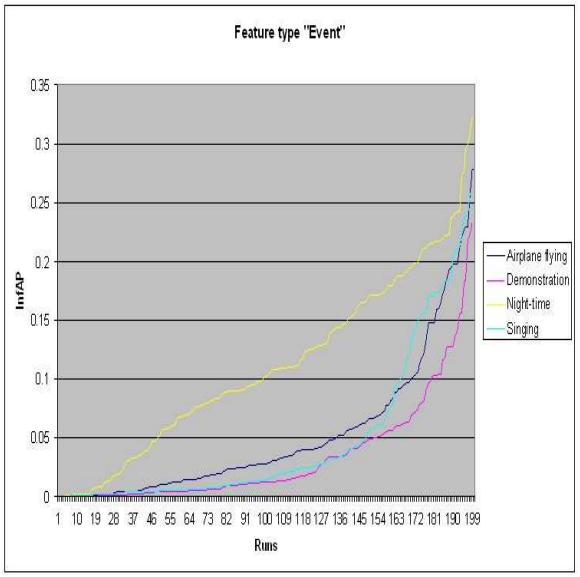


Figure 19: Features of category "Object"

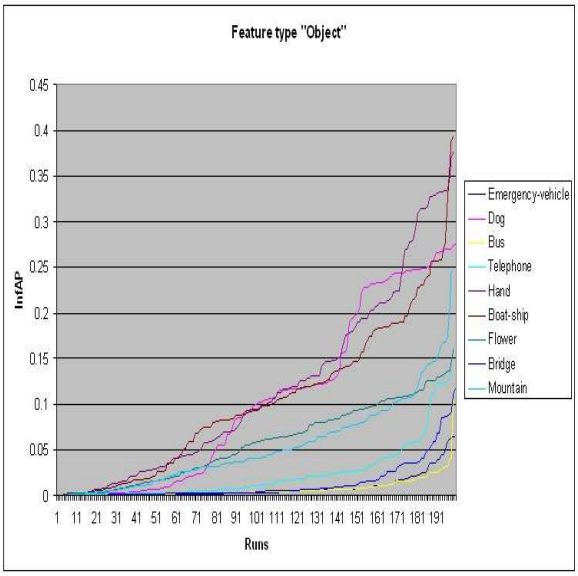
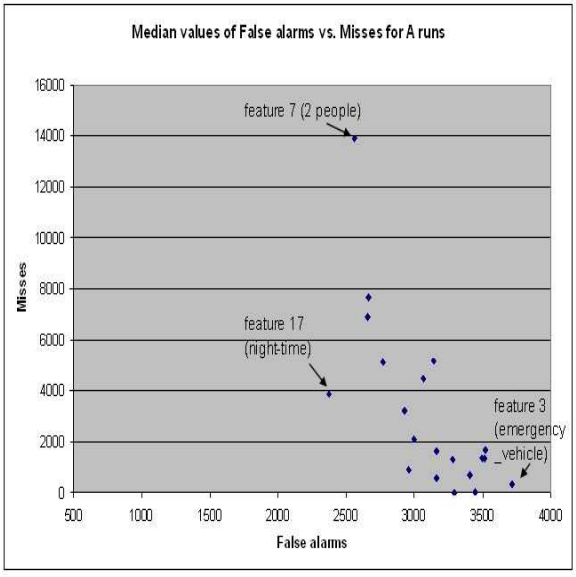


Figure 20: False alarms vs. misses for category A runs



and emergency-vehicle received the worst scores (one of the reasons might be that they were confused with each other as they have high similarity). On the other hand, features like "hand", "boat", and "dog" received top scores in this category. Those features might be easier to be recognize because the hand has characteristic color and features, the boat will be highly correlated with the existance of water and dog is the only animal in that category of features. Regarding the other feature categories, it can be shown that for the location category, the street feature received the highest score while features like "bridge", harbor" and "classroom" were at the bottom. The features "night-time" and "driver" achieved the highest score in the event and people categories respectively while the "Demonstration" and "singing" features achieved the lowest scores. In general the street feature achieved the highest score across all categories followed by "boat", "hand", "driver" and "night-time" while the "emergency-vehicle" achieved the lowest score among all categories followed by "bridge" and "bus" features. Figure 20 plots the false alarms vs. misses for each of the 20 features for runs of type A. The numbers in that plot were calculated based on the median values of the confusion matrix of the 20 features. We also did the same experiment for all other run types and found that they all almost show the same pattern. In general as the false alarms increases the misses decreases and vice versa. An interesting observation was found in those plots concerning a set of features having almost the same relative locations in all run types. Those features are "2 people" which has the highest miss rate across all run types, the feature "night-time" which has low miss and low false alarms thus better detection in general, and finally features "classroom","bridge", and "emergency-vehicle" which were confused highly with all other features giving low miss rate and high false alarms. We think that systems tried to achieve high accuracy for the "2 people" feature (maybe because it can be an easy feature to detect) so they reduced the false alarms but this came with the cost of high misses especially because this feature occurs very frequently in the test data. Also, for the feature "night-time" we think systems achieved good results as expected because the color feature can easily discriminate between those type of videos and other normal day-time videos. We can summarize some general observations from this year's task in the following points. Participation is still increasing and more interest are noticed for categories B and C submissions. Submissions in category B achieved best performance while category C is on a par with category A. There are hardly any feature specific approaches. Approximately, 50% of the runs use salient or scale-invariant feature transform (SIFT) points, while approx. 30% of the runs do some form of temporal analysis. The number of classifiers used for fusion ranges between 1 and more than 1160 and there is large variety in classifier architecture and choice of feature representations. The hardware used is usually a single central processing unit, however several medium and large clusters exist. Testing times vary between 10m and 150h per feature. Readers should see the results on the TRECVID website for details about the performance of each run.

4 Search

The search task in TRECVID was a multimedia extension of its text-only analogue. Video search systems were presented with topics — formatted multimedia 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.

4.1 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 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.).

One baseline run was required of every manual system — a run based only on the text from the provided English ASR/machine translation (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.

4.2 Data

As mentioned above, the search test collection (identical to the that for the feature task) contained 219 files/videos and 35766 reference shots, but four test files were ignored in the testing due to problems displaying shots from these long files (BG_36684, BG_37970, BG_38162, BG_8887) in the assessment system. Removing these files left 215 files and 33726 shots.

4.3 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 2008 topics are 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 2008 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.

NIST developed 48 topics for use in testing fully automatic search systems. Half of that set were used to test manual and interactive systems. The 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 2008 based on Armitage & Enser, 1996, is provided in Table 5. In 2008 all topics were generic and there was a deliberate emphasis on event topics.

4.4 Evaluation

Groups were allowed to submit a total of up to 6 runs of any types in the search task. In fact 27 groups submitted a total of 124 runs — 34 interactive runs, 8 manual ones, and 82 fully automatic ones. The trends seen in 2005 and 2006 in terms of groups migrating away from interactive search and towards fully automatic, with a dwindling participation in manual

Figure 22: Hits in the test set by topic

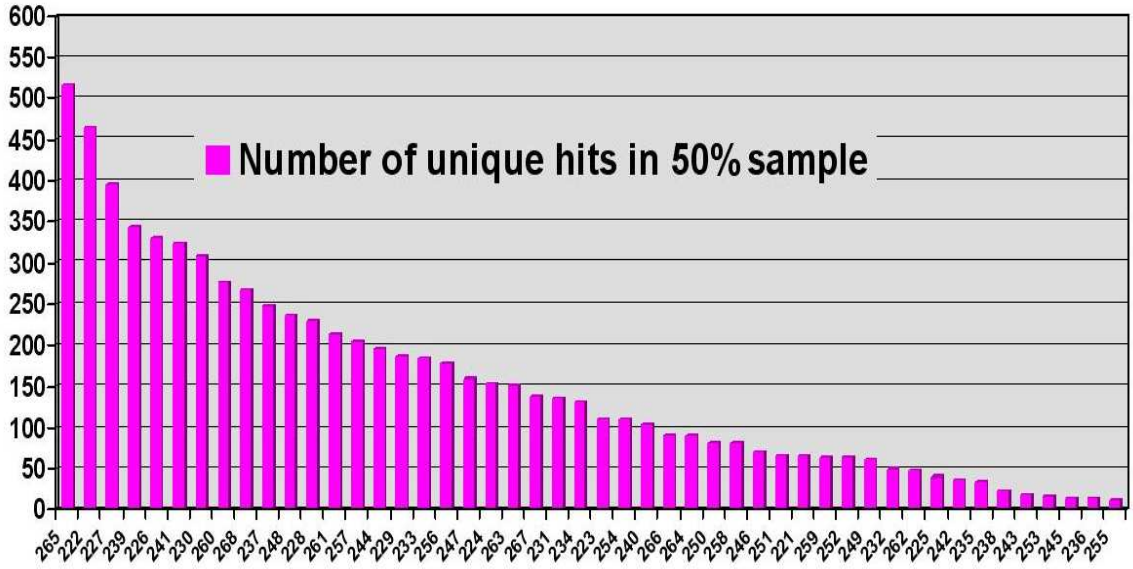
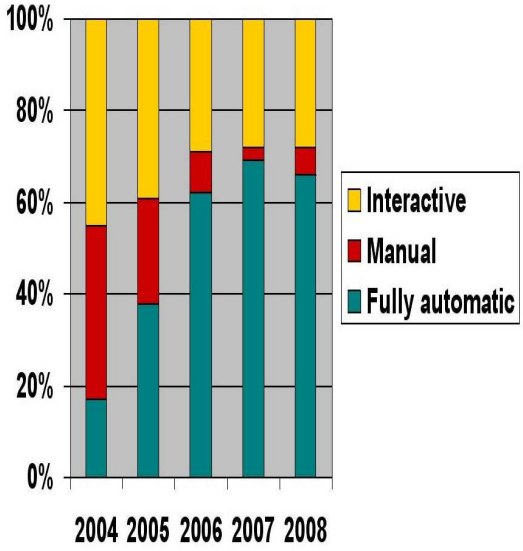


Figure 21: Search runs by type



search, leveled off in 2007 and 2008 as shown in Figure 21.

All submitted runs from each participating group contributed to the evaluation pools. For each topic, all submissions down to a depth of at least 40 (average 67, maximum 100) result items (shots) were pooled, duplicate shots were removed and randomized. Human judges (assessors) were presented with a 50% random sample of 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 is listed in Table 4 for details. Figure 22 shows the number of relevant shots found for each topic in the 50% judged sample.

4.5 Measures

The *trec_eval* program was used to calculate estimated recall, estimated precision, and inferred average precision (infAP) based on a 50% sample of the judgement pools.

4.6 Results

Figures 23, 24, and 25 show the estimated precision/recall curves for the top automatic, manual, and interactive search runs, respectively. Performance rises significantly with added human contribution.

Figure 23: Top 10 automatic search runs

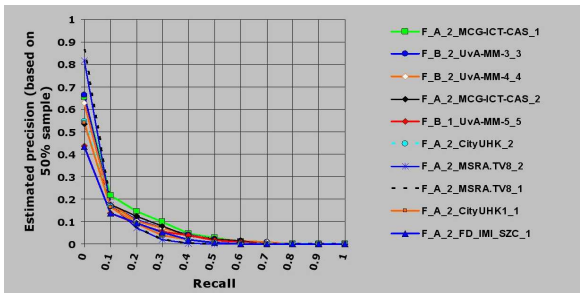


Figure 24: All manual search runs

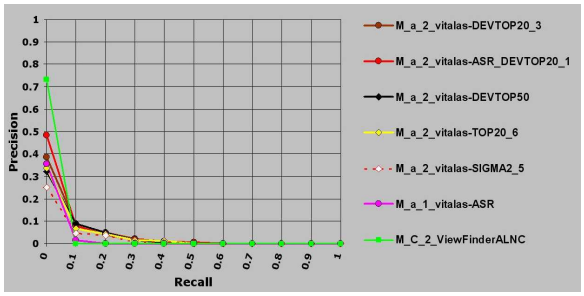


Figure 25: Top 10 interactive search runs

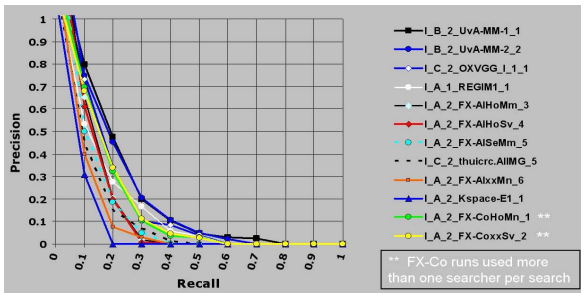


Figure 26: Randomization test on top 10 automatic search runs

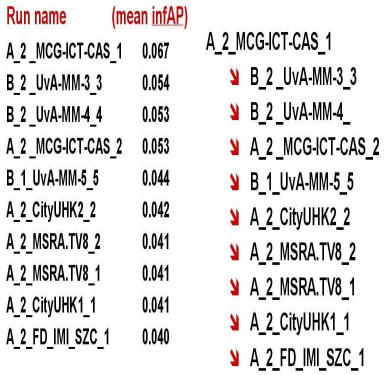
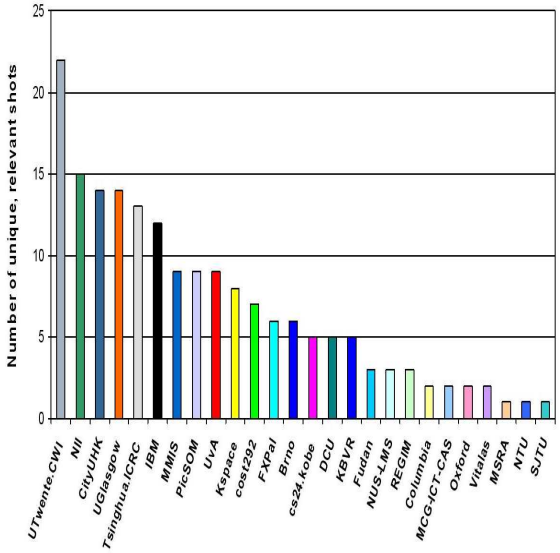


Figure 27: Randomization test on top 10 interactive search runs

Figure 28: Unique relevant by team



A partial randomization test (Manly, 1997) on the top runs indicates there are significant ($p < 0.05$) differences as shown in Figures 26 and 27.

Another interesting difference in systems is how many responsive shots were returned only by a given team's runs as shown in Figure 28. If this number is low, that suggests that the pooled assessments would still be useful in judging a system even if it had not contributed to the pools that were judged. For example, if the 22 unique hits found by UTwente-CWI all came from their 5 automatic runs, they would represent 0.3 % of all the hits found. The number of hits found uniquely by a team's runs may point to opportunities for other systems to improve their performance. Interestingly, the two teams with the highest number of unique hits (UTwente-CWI and NII) both trained their systems on video not taken from the Sound and Vision source.

Underneath the averages across topics, performance varies widely as shown in Figure 29. Figure 30 shows the text of the topics on which system performed best, something not easily predicted based on a single factor. Figure 31 shows the performance of runs using text only (speech via machine translation from the video and the text description from the topic) versus runs that (also) use visually encoded information from the video to be searched and the topic.

Figure 30: Topics sorted by median infAP

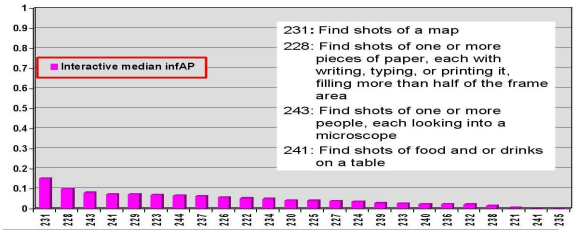


Figure 31: Text-only versus text-plus runs

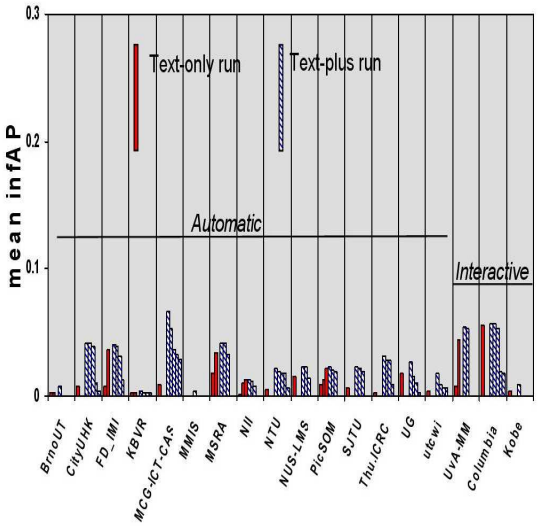
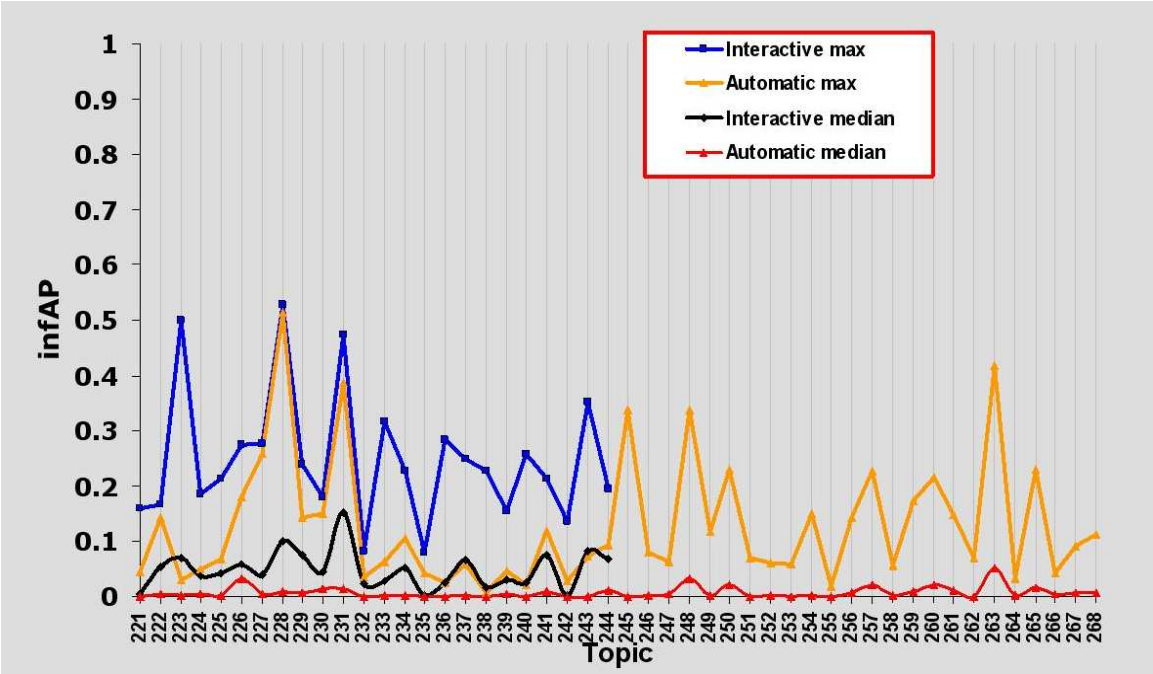


Figure 29: infAP by topic



By design, TRECVID sets a high-level search task and applies summative measures for effectiveness, speed, and usability for systems. This allows participants to focus on the specific components and research questions of interest to them and a very wide variety of issues are addressed each year. The particular hypotheses tested and the conclusions drawn are best understood in the context of the each participant's experiments as presented in their notebook papers on the TRECVID publications page (www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html).

5 BBC rushes summarization

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).

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 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 presentation. 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 2% of the video to be summarized. That gave a mean maximum summary duration of about 32 seconds.

5.1 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 at NIST for all the test data.

5.2 Evaluation

At Dublin City University all the summary clips for a given source video were viewed using mplayer on Linux in a window 125mm x 102mm @ 25 fps in a randomized order. A single human judge judged all summary clips from the same source video and several judges took part in the evaluation[1]. 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 questions with 5 possible answers for each - where only the extremes are labeled: "Strongly agree" and "strongly disagree".

1. "This summary contains many color bars, clapboards, all black or all white frames."

2. "This summary contains many nearly identical segments."

3. "This summary is presented in a pleasant tempo and rhythm."

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 2% maximum. The other was based on a sample within the 2% maximum from clusters built on the basis of a simple color histogram.

5.3 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

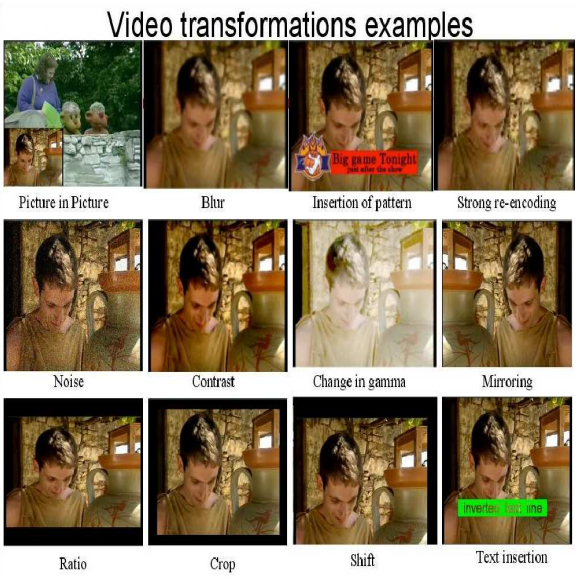
Per-system measures were the means of the per-summary measures over all test videos.

[1]This part of the evaluation was sponsored by the European Commission under contract FP6-027026 (K-Space)

5.4 Results

A detailed discussion of the results is available in the workshop papers and slides available from the TRECVID Video Summarization Workshop webpage at www-nlpir.nist.gov/projects/tv8.acmmm and in the ACM Digital Library (e.g. in the overview paper - http://portal.acm.org/citation.cfm?doid=1463563.14635

Figure 32: Examples of video transformations



6 Copy detection

As used here, a copy is a segment of video derived from another video, usually by means of various transformations such as addition, deletion, modification (of aspect, color, contrast, encoding, ...), camcording, etc. Detecting copies is important for copyright control, business intelligence and advertisement tracking, law enforcement investigations, etc. Content-based copy detection offers an alternative to watermarking. The TRECVID copy detection task was carried out in collaboration with members of the IMEDIA team at INRIA and built on the Video Copy Detection Evaluation Showcase at CIVR 2007

The required system task was as follows: given a test collection of videos and a set of 2010 queries (video-only segments), determine for each query the place, if any, that some part of the query occurs, with possible transformations, in the test collection. Two thirds of the queries contained copies.

A set of 10 possible transformations was selected to reflect actually occurring transformations and applied to each of 201 untransformed (base) queries using tools developed by IMEDIA to include some randomization at various decision points in the construction of the query set. For each query, the tools took a segment from the test collection, optionally transformed it, embedded it in some video segment which did not occur in the test collection, and then finally applied one or more transformations to the entire query segment. Some queries contained no test segment; others were composed entirely of the test segment. Video transformations included camcording, picture-in-picture, insertion of patterns, reencoding, change of gamma, decreasing the quality, and post production alterations. Video transformations used were documented in detail as part of the TRECVID Guidelines and examples frames are depicted in Figure 32.

Since detection of untransformed audio copies is relatively easy, and the primary interest of the TRECVID community is in video analysis, it was decided to model the required copy detection task with video-only queries. However, since audio is of importance for practical applications, there were two additional optional tasks: a task using transformed audio-only queries and one using transformed audiovideo queries.

1407 audio-only queries were generated by Dan Ellis at Columbia University along the same lines as the video-only queries: an audio-only version of the set of 201 base queries was transformed by seven techniques that were intended to be typical of those that would occur in real reuse scenarios: (1) bandwidth limitation (2) other coding-related distortion (e.g. subband quantization noise) (3) variable mixing with unrelated audio content.

A script to construct 14070 audio + video queries was provided by NIST. These queries comprised all the combinations of transformed audio(7) and transformed video (10) from a given base audiovideo query (201). In this way participants could study the effectiveness of their systems for individual audio and video transformations and their combinations.

6.1 Data

All of the 2007 and 2008 Sound and Vision data were used as a source (200 hours) from which the test query generation tools chose reference video. The

2007 BBC rushes video was used as a source for non-reference video.

6.2 Evaluation

In total in 2008, 22 participant teams submitted 55 runs for evaluation. 48 runs were submitted for video-only evaluation, 1 run for audio-only and 6 runs for mixed (audiovideo). Copy detection submissions were evaluated separately for each transformation, according to:

- How many queries they find the reference data for or correctly tell us there is none to find
- When a copy is detected, how accurately the run locates the reference data in the test data.
- How much elapsed time is required for query processing

6.3 Measures (per transformation)

- Minimal Normalized Detection Cost Rate: a cost-weighted combination of the probability of missing a true copy and the false alarm rate. For TRECVID 2008 the cost model assumed a scenario in which copies are very rare (e.g. 0.5/hr) and assigned misses a cost 10 times that of a false alarm. Other realistic scenarios were of course possible. Minimal normalized detection cost rate (minNDCR) reduced in 2008 to two terms involving two variables: probability of a miss (Pmiss) and the number of false alarms (FA).

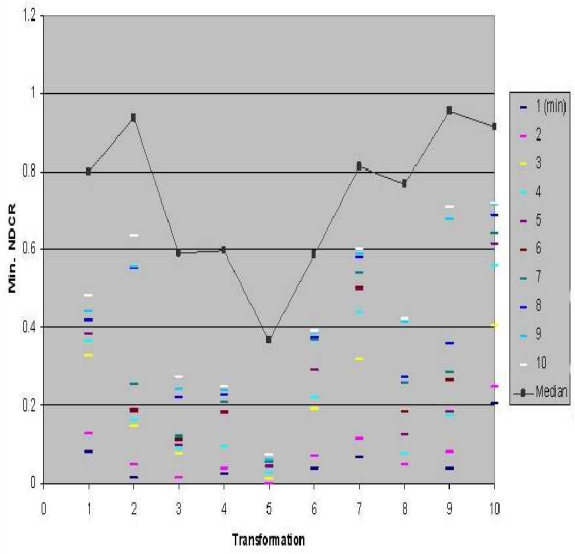
$$minNDCR = Pmiss + FA/24.9$$

- Copy location accuracy: mean F1 (harmonic mean) score combining the precision and recall of the asserted copy location relative to the ground truth location
- Copy detection processing time: mean processing time (s)

6.4 Results

Figures 33, 34, and 35 present the best results for the three main measures - minimum NDCR, F1, and processing time, respectively. From the figures of the F1 and cost measures, it can be shown that there is still room for participants to improve as there is a noticed

Figure 33: Video transformations vs. Min NDCR (Top 10)



spread among the top 10 performance for almost all of the transformations. Only one noticed tight spread exist in transformation 5 (Change of gamma) which also achieved the minimum cost among all transformations. Regarding the processing time, the top 10 achieved maximum about 20 second which is reasonable near a real-time performance for a copy detection system. Figure 36 shows the percentage of submitted items that were false alarms for the top runs. Some systems achieved very low false alarm rates (reaching 0) which is also very good performance for practical systems.

Figure 37 plots the relationship between minimum NDCR and F1 for each video transformation.There appears to be little correlation between systems that are good in separating copies from non-copies (low NDCR) and those also good in localization. Also we noticed that transformation 10 probably makes it hardest to detect copies. This can be justified by the fact that transformation 10 is a combination of 5 transformations. Similarly, Figure 38 graphs F1 versus processing time. From that graph we can conclude that increasing processing time did not enhance localization. Only few systems achieved high localization in small processing time. Figure 39 compares minimum NDCR against processing time. We can see that increasing the processing time did not reduce the

Figure 34: Video transformations vs. F1 (Top 10)

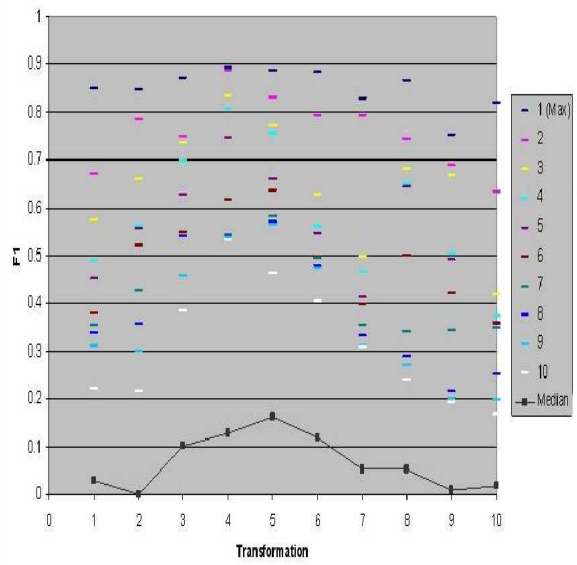


Figure 35: Video transformations vs. Processing time (Top 10)

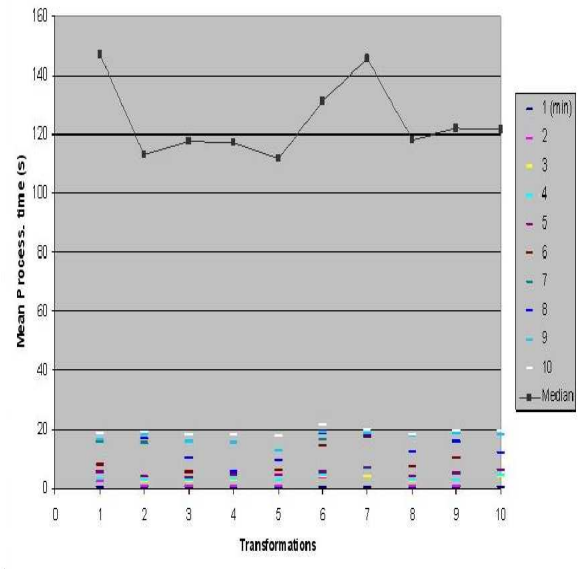
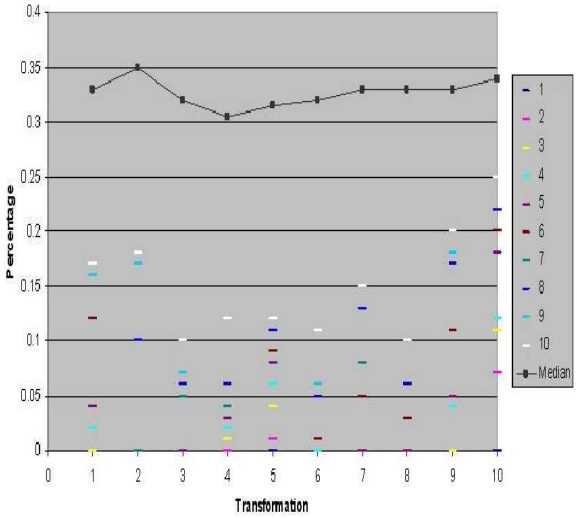


Figure 36: Video transformations vs. False alarms (Top 10)



cost or make the systems stronger. Few good systems are fast with low cost.

Figure 40 presents the best minimum NDCR scores for the audio + video queries for each combination of the audio and video transformation. For purposes of rough comparison, it also shows the scores for the best video-only queries. As the number of submitted audio + video runs are too limited, we can not make general conclusions. However, the relative effect of audio transformations seems similar across video transformations; it seems that using audio (when no speech is mixed in) helped to decrease the cost across transformations compared to using only video (except in video transformation 5).

From a brief survey regarding the used approaches among participants, we can find that generally techniques used can be divided into transformation-specific or more generic techniques. The most used features are SIFT descriptors, block-based features and edge histograms. There is a major trade-off between localization, effectiveness and speed. Some groups achieved very good results while others found the task very diffecult. In the future, we need to invistigate more realistic transformations that are actually used in copied videos in real life situations. This might be done by dropping very complicated transformations that are a combination of other transfor-

Figure 37: Relationship between F1 and cost across video transformations

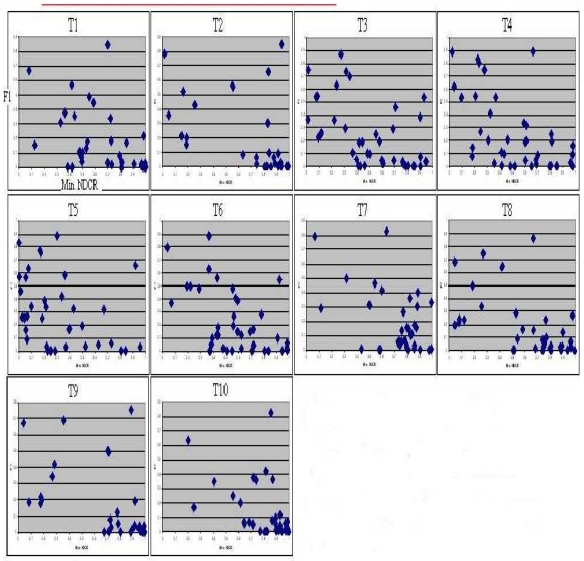
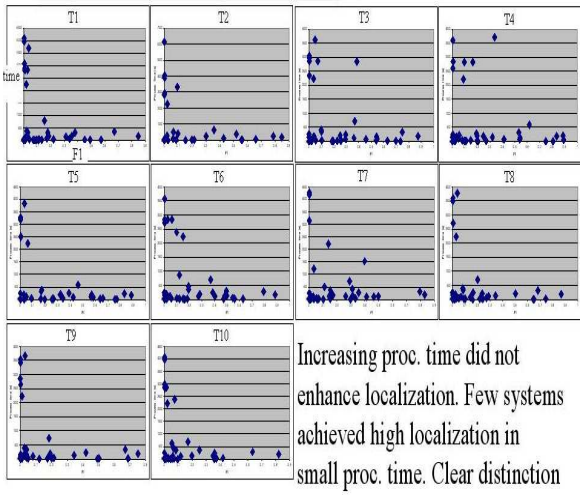
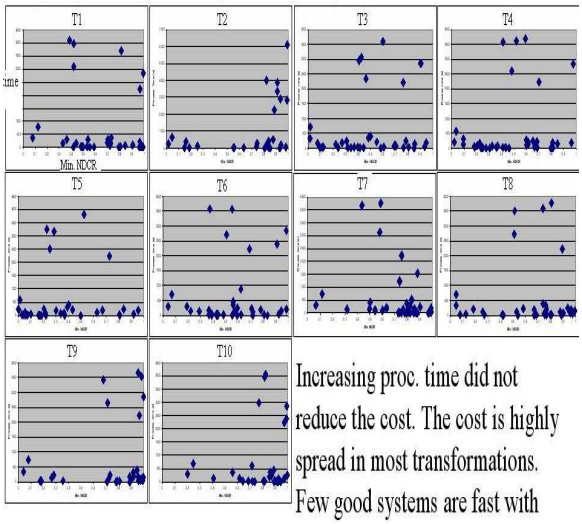


Figure 38: Relationship between F1 and processing time across video transformations



Increasing proc. time did not enhance localization. Few systems achieved high localization in small proc. time. Clear distinction between good systems and immature systems.

Figure 39: Relationship between processing time and cost across video transformations



Increasing proc. time did not reduce the cost. The cost is highly spread in most transformations. Few good systems are fast with low cost.

Figure 40: audiovideo runs vs. video only

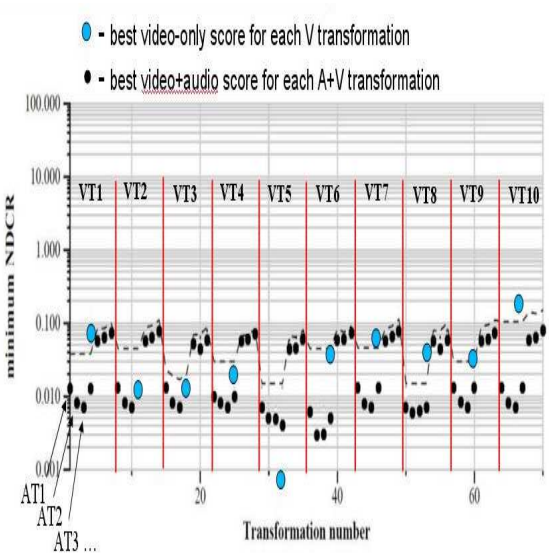
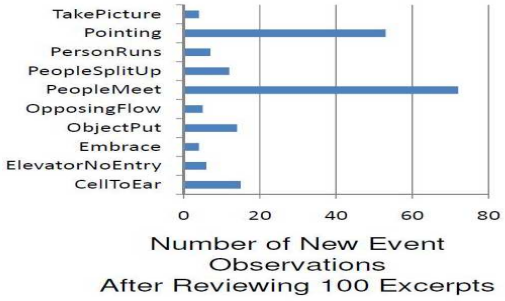


Figure 41: Camera views

Figure 43: Effect of adjudications on annotations



mations and found by this year's result to be very difficult to detect. Also, more encouragement for participants to submit audiovideo runs will be very important as the audiovideo runs seems to enhance the detection performance. In general, the pilot task for this year has achieved its goals in terms of getting all pieces together such as query composing and transforming, and attracting participants from the computer vision community. Readers are asked to see the results pages and workshop paper from each participating group on the TRECVID website for detailed information about each system's performance.

7 Surveillance event detection pilot

To help promote the development of computer vision techniques for event understanding, NIST proposed a formal evaluation that addresses video event detection from a large corpus of naturally collected video (starting with 5 cameras × 20 hours = 100 hours of

Figure 44: Effect of adjudications on scoring

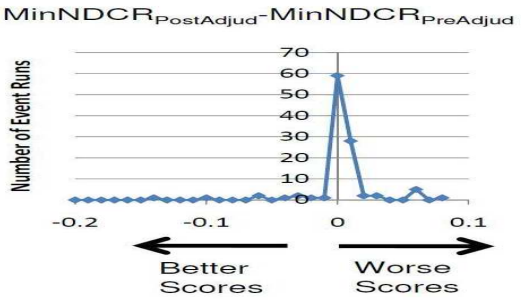


Figure 45: Distributions underlying detection scoring

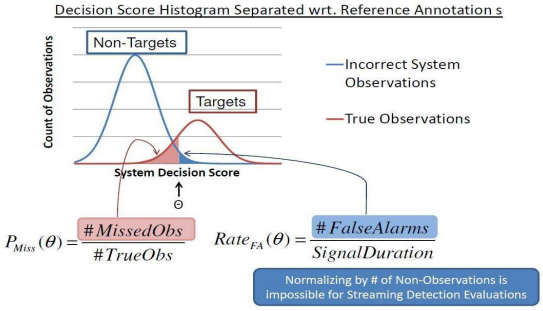


Figure 46: Plotting Detection Error Tradeoff (DET) curves

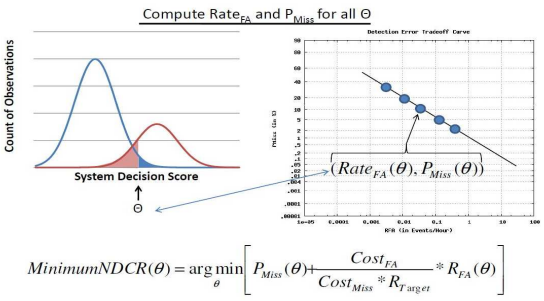
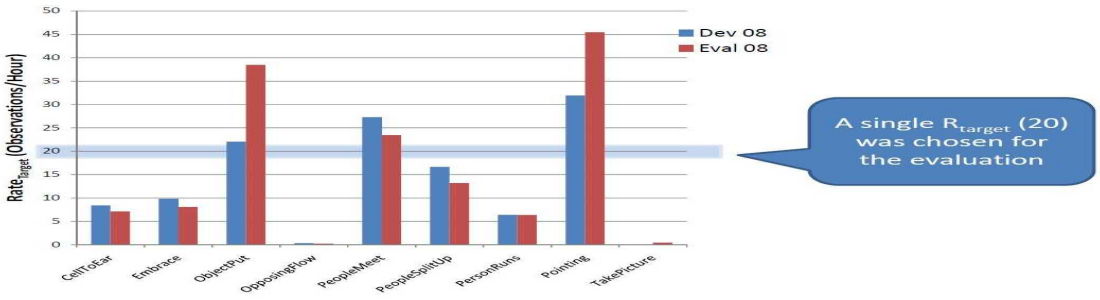


Figure 42: Event rates



surveillance, collected by the United Kingdom Home Office) as illustrated in Figure 41. While previous event detection efforts have been smaller in scope, the use of a large video corpus collected "in the wild" enabled the discovery of a set of naturally occurring events and allowed their frequencies (Figure 42) to be characterized.

The goal of this pilot evaluation was to move computer vision technology towards robustness and scalability while increasing core competency. The approach was to employ real surveillance data that is orders of magnitude larger than previous computer vision tests, and that consists of multiple, synchronized camera views. Further, NIST collaborated with the Linguistics Data Consortium (LDC) and the research community to select a variety of naturally occurring events. These events were of varying frequency and difficulty.

The evaluation supported two tasks: (a) retrospective event detection, (b) freestyle analysis. The first event detection task was defined as follows: given a set of video sequences, detect as many event observations as possible in each sequence. For this evaluation, a single-camera condition was used as the required condition (multiple-camera input was allowed as a contrastive condition). Further, systems could perform multiple passes over the video prior to outputting a list of putative event observations (i.e., the task was retrospective). For freestyle analysis, participants were asked to define tasks pertinent to the airport video surveillance domain and that could be implemented on the data set. Freestyle submissions were to include rationale, clear definitions of the task, performance measures, reference annotations, and a baseline system implementation.

Planning telecons were held with researchers and

Figure 47: Minimum versus actual NDCR

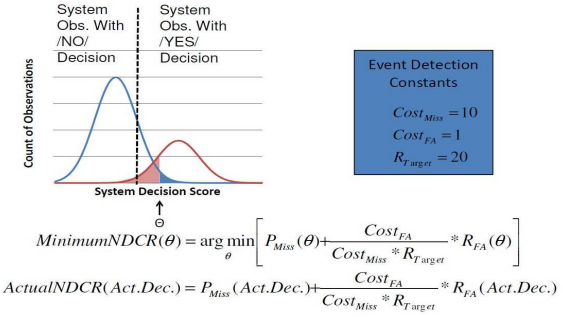


Figure 48: Participants by event

	Cell To Ear	Elevator NoEntry	Embrace	ObjectPut	Opposing Flow	People Meet	People Split Up	Person Runs	Pointing	Take Picture
AIT		×			×			×		
BUT		×		×	×			×		
CMU	×	×	×	×	×	×	×	×	×	×
DCU		×	×		×	×		×		
FD					×			×		×
IFP-UIUC-NEC	×	×	×	×	×	×	×	×	×	×
Intuvision		×			×					×
MCG-ICT-CAS		×	×		×	×	×	×		×
NHKSTRL		×			×			×		
QMUL-ACTIVA		×			×			×		
SJTU		×			×	×		×	×	
THU-MNL	×				×			×		
TokyoTech						×	×	×		
Toshiba		×			×			×		
UAM				×	×			×		
UCF				×	×			×		×
Total	**3**	**11**	**4**	**5**	**15**	**6**	**4**	**15**	**3**	**6**

the LDC to discuss the data, develop the task, discuss the annotation guidelines, etc. For this evaluation, we define an event to be an observable state change, either in the movement or interaction of people with other people or objects. As such, the evidence for an event depends directly on what can be seen in the video and does not require higher level inference. The annotation guidelines were developed to express the requirements for each event. To determine if the observed action is a taggable event, a "reasonable interpretation rule" was used. The rule was, "if according to a reasonable interpretation of the video, the event must have occurred, then it is a taggable event". Importantly, the annotation guidelines were designed to capture events that can be detected by human observers, such that the ground truth would contain observations that would be relevant to an operator/analyst. In what follows we distinguish between event types (e.g., parcel passed from one person to another), event instance (an example of an event type that takes place at a specific time and place), and an event observation (event instance captured by a specific camera).

7.1 Data

As noted above, the video data consisted of 100 hours of indoor airport surveillance from London Gatwick Airport (denoted by the airport code "LGW"). A portion of the video data was released as an online microcorpus (5 cameras $\times$ 4 minutes) to facilitate discussion about the naturally occurring events with the research community.

The entire video corpus was distributed as MPEG-2 in Phase Alternating Line (PAL) format (resolution 720 x 576), 25 frames/sec, either via hard drive or downloaded from several internet mirrors. Both the development and evaluation video data were released at once to allow the most compute time for feature extraction, tracking algorithms, etc. The development set (devset) annotations were released incrementally as they became available. The evaluation set (evalset) annotations were released after final scores were provided to participants.

The videos were annotated using the Video Performance Evaluation Resource (ViPER) tool. Events were represented in ViPER format using an annotation schema that specified each event observation's time interval. For system outputs, in addition to temporal extent, DetectionDecision and DetectionScore values were required.

7.2 Evaluation

Sites submitted system outputs for the detection of any 3 of 10 required events (PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, Pointing, ElevatorNoEntry, OpposingFlow, and TakePicture). Outputs included the temporal extent as well as a confidence score and detection decision (yes/no) for each event observation. Developers were advised to target a low miss, high false alarm scenario, in order to maximize the number of event observations.

A dry run was carried out for one day of collection (10 camera hours) from the devset to test the evaluation infrastructure. A formal evaluation was carried out for five days of collection (approx. 50 camera hours). Groups were allowed to submit multiple runs with contrastive conditions. System submissions were aligned to the reference annotations and initially scored for missed detections / false alarms.

Although the LDC performed exhaustive annotations over the entire video corpus, analysis of dual annotations indicated there would likely be missed event observations in the reference data. In order to develop a more complete reference annotation, NIST and LDC collaborated to review the most likely missed annotations based on system outputs. This "adjudication" process was limited by time and budget, so a prioritized interval list was created based on the agreement across systems or the strength of the decision scores. Figure 43 shows the effect of adjudication on the annotation. Following adjudication and annotation enrichment, system submissions were re-scored. Figure 44 shows how it affected the scoring. The post-adjudication scores were provided to participants at the TRECVid workshop.

7.3 Measures

Since detection system performance is a tradeoff between probability of miss vs. rate of false alarms, this task used the Normalized Detection Cost Rate (NDCR) measure for evaluating system performance. NDCR is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per unit time). Participants were provided with a graph of the Decision Error Tradeoff (DET) curve for each event their system detected; the DET curves were plotted over all events (i.e., all days and cameras) in the evaluation set. Figure 45 shows the distributions on which the scoring is based. Figure 46 shows the relationship between the basic dis-

tributions and the DET curve. Figure 47 depicts the difference between minimum versus actual NDCR.

7.4 Results

Seventeen research groups completed the required task. Table 48 shows which groups worked on which events. Readers are asked to see workshop notebook papers on the TRECVID website for details about each participating group's work.

8 Summing up and moving on

This introduction to TRECVID 2008 has provided basic information on the goals, data, evaluation mechanisms and metrics used. Further details about each particular group's approach and performance for each task can be found in that group's site report on the TRECVID website.

9 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.

City University of Hong Kong, the Laboratoire d'Informatique de Grenoble, and the University of Iowa helped out in the distribution of video data by mirroring them online.

We are grateful to Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin for providing the master shot reference, to Roeland Ordelman and Marijn Huijbregts at the University of Twente for donating the output of their automatic speech recognition system run on the Sound and Vision data, and to Christof Monz of Queen Mary, University London, who contributed machine translation (Dutch to English) for the Sound and Vision video.

INRIA's Nozha Boujemaa, Alexis Joly, and Julien Law-to led the design of the copy detection task, in particular regarding the definitions of the video transformations. They provided an independent person, Laurent Joyeux, who created the original 201 queries and applied the 10 video transformations in a process blind to the ground truth. Dan Ellis at Columbia University devised and applied the audio transformations to produce the audio-only queries for copy detection.

Georges Quénot and Stéphane Ayache of LIG (Laboratoire d'Informatique de Grenoble organized a collaborative annotation and 40 groups contributed 1.2 million concept × image judgments. The Multimedia Content Group at the Chinese Academy of Sciences provided full annotation of test features for 2008 training data including location rectangles for object features. Columbia University and the City University of Hong Kong contributed detection scores for the 2008 data CU-VIREO374. The University of Amsterdam provided 2 benchmarks for assessing mappings of topics to concepts for video retrieval.

Phil Kelly at Dublin City University (DCU) assisted with the assessment of the rushes summaries and this assessment was sponsored by the European Union under contract FP6-027026 (K-SPace). Carnegie Mellon University created a baseline summarization run to help put the summarization results in context.

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

10 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.

221 Find shots of a person opening a door (I/0, V/5, R/65)

222 Find shots of 3 or fewer people sitting at a table (I/2, V/5, R/465)

223 Find shots of one or more people with one or more horses (I/2, V/3, R/111)

224 Find shots of a road taken from a moving vehicle, looking to the side (I/0, V/5, R/153)

225 Find shots of a bridge (I/2, V/4, R/40)

226 Find shots of one or more people with mostly trees and plants in the background; no road or building can be seen (I/1, V/3, R/330)

227 Find shots of a person's face filling more than half of the frame area (I/2, V/5, R/395)

228 Find shots of one or more pieces of paper, each with writing, typing, or printing it, filling more than half of the frame area (I/2, V/4, R/230)

229 Find shots of one or more people where a body of water can be seen (I/2, V/5, R/187)

230 Find shots of one or more vehicles passing the camera (I/1, V/4, R/307)

231 Find shots of a map (I/2, V/1, R/136)

232 Find shots of one or more people, each walking into a building (I/0, V/5, R/49)

233 Find shots of one or more black and white photographs, filling more than half of the frame area (I/0, V/5, R/184)

234 Find shots of a vehicle moving away from the camera (I/2, V/5, R/130)

235 Find shots of a person on the street, talking to the camera (I/2, V/4, R/35)

236 Find shots of waves breaking onto rocks (I/2, V/2, R/14)

237 Find shots of a woman talking to the camera in an interview located indoors – no other people visible (I/2, V/5, R/248)

238 Find shots of a person pushing a child in a stroller or baby carriage (I/1, V/6, R/23)

239 Find shots of one or more people standing, walking, or playing with one or more children (I/2, V/5, R/343)

240 Find shots of one or more people with one or more books (I/2, V/5, R/104)

241 Find shots of food and or drinks on a table (I/2, V/4, R/323)

242 Find shots of one or more people, each in the process of sitting down in a chair (I/0, V/5, R/37)

243 Find shots of one or more people, each looking into a microscope (I/2, V/2, R/18)

244 Find shots of a vehicle approaching the camera (I/2, V/11, R/195)

245 Find shots of a person watching a television screen – no keyboard visible (I/2, V/2, R/15)

246 Find shots of one or more people in a kitchen (I/2, V/5, R/70)

247 Find shots of one or more people with one or more animals (I/3, V/4, R/159)

248 Find shots of a crowd of people, outdoors, filling more than half of the frame area (I/2, V/5, R/237)

249 Find shots of a classroom scene (I/2, V/3, R/62)

250 Find shots of an airplane exterior (I/2, V/5, R/82)

251 Find shots of a person talking on a telephone (I/2, V/7, R/66)

252 Find shots of one or more people, each riding a bicycle (I/2, V/5, R/63)

253 Find shots of one or more people, each walking up one or more steps (I/2, V/6, R/17)

254 Find shots of a person talking behind a microphone (I/2, V/5, R/110)

255 Find shots of just one person getting out of or getting into a vehicle (I/2, V/5, R/12)

256 Find shots of one or more people, singing and/or playing a musical instrument (I/2, V/5, R/177)

257 Find shots of a plant that is the main object inside the frame area (I/2, V/8, R/204)

258 Find shots of one or more people sitting outdoors (I/2, V/4, R/81)

259 Find shots of a street scene at night (I/2, V/5, R/64)

260 Find shots of one or more animals – no people visible (I/2, V/5, R/276)

261 Find shots of one or more people at a table or desk, with a computer visible (I/2, V/5, R/213)

262 Find shots of one or more people in white lab coats (I/2, V/4, R/48)

263 Find shots of one or more ships or boats, in the water (I/0, V/5, R/151)

264 Find shots of one or more colored photographs, filling more than half of the frame area (I/0, V/5, R/91)

265 Find shots of a man talking to the camera in an interview located indoors – no other people visible (I/2, V/5, R/516)

266 Find shots of more than 3 people sitting at a table (I/2, V/5, R/91)

267 Find shots with the camera zooming in on a person's face (I/0, V/5, R/138)

268 Find shots of one or more signs with lettering (I/2, V/5, R/268)

11 Appendix B: Features

1 Classroom: a school– or university-style classroom scene. One or more students must be visible. A teacher and teaching aids (e.g. blackboard) may or may not be visible.

2 Bridge: a structure carrying a pathway or roadway over a depression or obstacle. Such structures over non-water bodies such as a highway overpass or a catwalk (e.g., as found over a factory or warehouse floor) are included.

3 Emergency Vehicle: external view of, for example, a police car or van, fire truck or ambulance. There may be other sorts of emergency vehicles. Included may be UN vehicles, but NOT military vehicles

4 Dog: any kind of dog, but not wolves

5 Kitchen: a room where food is prepared, dishes washed, etc.

6 Airplane flying: external view of a heavier than air, fixed-wing aircraft in flight - gliders included. NOT balloons, helicopters, missiles, and rockets

7 Two people: a view of exactly two people (not as part of a larger visible group)

8 Bus: external view of a large motor vehicle on tires used to carry many passengers on streets, usually along a fixed route. NOT vans and SUVs

9 Driver: a person operating a motor vehicle or at least in the driver's seat of such a vehicle

10 Cityscape: a view of a large urban setting, showing skylines and building tops. NOT just street-level views of urban life

11 Harbor: a body of water with docking facilities for boats and/or ships such as a harbor or marina, including shots of docks. NOT shots of offshore oil rigs, piers that do not look like they belong to a harbor or boat dock

12 Telephone: any kinds of telephone, but more than just a headset must be visible.

13 Street: a regular paved street NOT a highway, dirt road, or special type of road or path

14 Demonstration Or Protest: an outdoor, public exhibition of disapproval carried out by multiple people, who may or may not be walking, holding banners or signs

15 Hand: a close-up view of one or more human hands, where the hand is the primary focus of the shot.

16 Mountain: a landmass noticably higher than the surrounding land, higher than a hill, with the slopes visible

17 Nighttime: a shot that takes place outdoors at night. NOT sporting events under lights

18 Boat Ship: exterior view of a boat or ship in the water, e.g. canoe, rowboat, kayak, hydrofoil, hovercraft, aircraft carrier, submarine, etc.

19 Flower: a plant with flowers in bloom; may just be the flower

20 Singing: one or more people singing PageRanksinger(s) visible and audible, solo or accompanied, amateur or professional

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. (2008). Video Corpus Annotation Using Active Learning,. In *Proceedings of the 30th european conference on information retrieval (ecir'08)* (pp. 187–198). Glasgow, UK.

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.

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.

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 5: 2008 Topic types

	Named			Generic		
Topic	Person, thing	Event	Place	Person, thing	Event	Place
221				X	X	
222				X	X	
223				X		
224				X	X	X
225				X		
226				X		
227				X		
228				X		
229				X		X
230				X	X	
231				X		
232				X	X	X
233				X		
234				X	X	
235				X	X	X
236				X	X	
237				X	X	
238				X	X	
239				X	X	
240				X		
241				X		
242				X	X	
243				X	X	
244				X	X	
245				X	X	
246				X		X
247				X		
248				X		X
249						X
250				X		
251				X	X	
252				X	X	
253				X	X	
254				X	X	
255				X	X	
256				X	X	
257				X		
258				X		X
259						X
260				X		
261				X		
262				X		
263				X		
264				X		
265				X	X	
266				X		
267				X	X	
268				X		

Table 1: Participants and tasks

Task					Location	Runprefix	Participants
–	**	FE	RU	–	Asia	asahikase	Asahikasei Co.
–	ED	–	–	–	Europe	–	Athens Information Technology
*	–	–	RU	–	NorthAm	ATTLabs	AT&T Labs - Research
–	ED	–	–	–	NorthAm	–	Beckman Institute
CD	**	**	–	**	Asia	IIS_BJTU	Beijing Jiaotong University
CD	**	FE	–	–	Asia	BeijingUPT	Beijing University of Posts and Telecommunications
CD	–	FE	–	**	Europe	BilkentUMDG	Bilkent University MDG
CD	ED	FE	**	SE	Europe	Brno	Brno University of Technology
–	ED	FE	RU	**	NorthAm	CMU	Carnegie Mellon University
CD	ED	FE	–	SE	Asia	MCG-ICT-CAS	Chinese Academy of Sciences (MCG-ICT-CAS)
CD	**	FE	RU	SE	Asia	VIREO	City University of Hong Kong
–	–	FE	–	–	Europe	LSIS	CNRS LSIS
CD	–	FE	–	SE	NorthAm	CU	Columbia University
CD	–	–	–	–	NorthAm	CRIMontreal	Computer Research Institute of Montreal
CD	**	FE	RU	SE	Europe	COST292	COST292
CD	**	FE	RU	SE	Europe	COST292	Delft University of Technology
–	ED	–	RU	SE	Europe	DCU	Dublin City University
–	**	FE	RU	SE	Europe	REGIM	École Nationale d'Ingénieurs de Sfax ENIS
*	**	**	RU	**	Europe	ETIS	ETIS Laboratory
–	**	FE	–	–	NorthAm	FIU	Florida International University
CD	ED	FE	–	SE	Asia	FD	Fudan University
–	–	–	RU	SE	NorthAm	FX	FX Palo Alto Laboratory
*	**	FE	RU	**	Europe	IRIM	GDR ISIS - IRIM consortium
CD	–	FE	RU	SE	Europe	PicSOM	Helsinki University of Technology TKK
CD	**	FE	**	SE	NorthAm	IBM	IBM Watson Research Center
CD	–	**	–	**	Europe	INRIA-IMEDIA	INRIA-IMEDIA
CD	–	FE	–	–	Europe	INRIA-LEAR	INRIA-LEAR
–	**	**	RU	–	Europe	EURECOM	Institut EURECOM
–	ED	–	–	–	NorthAm	–	intuVision, Inc.
CD	–	–	–	–	Europe	ITU	Istanbul Technical University
*	–	FE	–	–	Europe	IUPR	IUPR-DFKI
*	**	FE	RU	–	Europe	JRS	JOANNEUM RESEARCH
–	–	–	–	SE	NorthAm	KBVR	KB Video Retrieval
–	–	**	–	SE	Asia	cs24_kobe	Kobe University
–	–	**	RU	SE	Europe	KSPACE	K-Space
*	–	FE	–	**	Europe	LIG	Laboratoire d'Informatique de Grenoble
–	**	FE	**	**	Europe	LIRIS	Laboratoire LIRIS (LYON)
*	**	FE	**	SE	Asia	MSRA	Microsoft Research Asia
CD	**	FE	RU	SE	Asia	NII	National Institute of Informatics
*	**	FE	–	SE	Asia	NTU	National Taiwan University
–	–	–	–	SE	Asia	NUSLMS	National Unversity of Singapore
*	ED	FE	RU	**	Asia	NHKSTRL	NHK Science and Technical Research Laboratories
–	–	**	RU	–	Asia	nttlab	NTT Cyber Solutions Laboratories
CD	–	–	–	–	Europe	OrangeLabs	Orange Labs - France Telecom Group
–	**	–	RU	–	Asia	–	Osaka University
*	**	FE	**	**	Asia	PKU	Peking University
–	ED	–	–	–	Europe	–	Queen Mary, University of London (QMUL)
–	ED	FE	–	SE	Asia	SJTU	Shanghai Jiao Tong University
–	–	FE	–	–	Asia	ISM	The Institute of Statistical Mathematics
*	–	FE	–	SE	Europe	MMIS	The Open University
*	–	–	RU	–	Asia	PolyU	The Hong Kong Polytechnic University
CD	**	–	–	–	Europe	TNO	TNO-ICT
–	ED	FE	RU	–	Asia	TiTech	Tokyo Institute of Technology
–	ED	–	–	–	Asia	–	Toshiba Corporation
CD	**	FE	RU	SE	Asia	thuicrc	Tsinghua University and Intel China Research Center
–	ED	–	**	–	Asia	Thu-intel	Tsinghua University-MNL
–	–	FE	–	–	Europe	MESH	UAM-NTUA-Telefonica I+D

Task legend. CD:copy detection; ED:event detection; FE:feature detection; RU:rushes summarization; SE:search; **:no runs submitted

Table 2: Participants and tasks (continued)

Task					Location	Participants
–	ED	–	RU	–	Europe	Universidad Autonoma de Madrid
–	–	FE	–	–	Europe	Universidad Carlos III de Madrid
–	–	–	RU	**	Europe	Universidad Rey Juan Carlos
*	**	FE	RU	–	Europe	Universite Pierre et Marie Curie - LIP6
–	–	FE	RU	–	Australia	Queensland University of Technology
CD	ED	**	–	**	NorthAm	University of Central Florida
CD	**	–	RU	–	Europe	University of Bradford
CD	–	**	RU	SE	Europe	University of Glasgow
–	–	FE	–	–	Europe	University of Karlsruhe (TH)
*	**	FE	**	**	Europe	University of Marburg
*	**	FE	RU	**	Asia	University of Electro-Communications
–	**	FE	–	SE	Europe	University of Amsterdam
*	–	FE	–	SE	Europe	University of Oxford
–	–	FE	–	SE	Europe	University of Twente and CWI
–	**	**	RU	**	Europe	University of Ioannina, Greece
*	–	–	RU	–	Europe	University of Sheffield
–	**	–	RU	–	NorthAm	University of Ottawa - SITE
–	–	–	–	SE	NorthAm	University of Alabama
–	–	FE	–	SE	Europe	VITALAS: CERTH-ITI (GR), CWI (NL), U. Sunderland (UK)
*	–	FE	–	–	Asia	Xi'an Jiaotong University

Task legend. CD:copy detection; ED:event detection; FE:feature detection; RU:rushes summarization; SE:search; **:no runs submitted

Table 3: 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	365148	30926	8.5	100	3348	10.8	64	1.9
2	365827	29584	8.1	110	3497	11.8	30	0.9
3	352645	31843	9.0	90	3369	10.6	22	0.7
4	355055	31226	8.8	110	3454	11.1	94	2.7
5	354894	31051	8.7	90	3299	10.6	124	3.8
6	354336	29253	8.3	150	3317	11.3	64	1.9
7	367576	31805	8.7	90	3394	10.7	1090	32.1
8	363937	30202	8.3	110	3330	11.0	47	1.4
9	362502	30744	8.5	110	3358	10.9	364	10.8
10	360184	26793	7.4	160	3460	12.9	337	9.7
11	361481	27660	7.7	170	3387	12.2	35	1.0
12	355656	31579	8.9	100	3402	10.8	106	3.1
13	361038	28318	7.8	150	3398	12.0	458	13.5
14	362959	30244	8.3	120	3364	11.1	87	2.6
15	359031	30269	8.4	120	3324	11.0	630	19.0
16	359950	26791	7.4	180	3377	12.6	140	4.1
17	358513	24644	6.9	220	3383	13.7	316	9.3
18	368627	27452	7.4	170	3389	12.3	210	6.2
19	366551	29858	8.1	130	3436	11.5	319	9.3
20	367611	30498	8.3	110	3488	11.4	133	3.8

Table 4: 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
221	102445	26398	25.8	60	1890	7.2	65	3.4
222	106747	25056	23.5	40	1424	5.7	465	32.7
223	101553	25995	25.6	70	2020	7.8	111	5.5
224	102813	22537	21.9	70	1787	7.9	153	8.6
225	96721	22908	23.7	80	1949	8.5	40	2.1
226	104633	22140	21.2	70	1851	8.4	330	17.8
227	103089	24953	24.2	60	1900	7.6	395	20.8
228	100329	24243	24.2	70	1845	7.6	230	12.5
229	104066	23443	22.5	70	1959	8.4	187	9.5
230	101482	23421	23.1	70	1815	7.7	307	16.9
231	97431	23470	24.1	90	1886	8.0	136	7.2
232	103579	24816	24.0	40	1291	5.2	49	3.8
233	101743	26868	26.4	50	1716	6.4	184	10.7
234	100193	24151	24.1	60	1696	7.0	130	7.7
235	104788	26180	25.0	50	1833	7.0	35	1.9
236	97028	23344	24.1	50	1284	5.5	14	1.1
237	105530	25544	24.2	60	2046	8.0	248	12.1
238	100917	26694	26.5	50	1644	6.2	23	1.4
239	105561	25454	24.1	50	1775	7.0	343	19.3
240	104087	26891	25.8	50	1671	6.2	104	6.2
241	100757	24289	24.1	70	1878	7.7	323	17.2
242	103593	25230	24.4	50	1619	6.4	37	2.3
243	101309	27394	27.0	60	1921	7.0	18	0.9
244	101435	24191	23.8	70	1831	7.6	195	10.6
245	75157	23254	30.9	40	988	4.2	15	1.5
246	75966	23124	30.4	80	1854	8.0	70	3.8
247	77157	21261	27.6	80	1915	9.0	159	8.3
248	77095	20846	27.0	80	1798	8.6	237	13.2
249	73407	22298	30.4	80	1816	8.1	62	3.4
250	70422	20280	28.8	100	1844	9.1	82	4.4
251	75447	22448	29.8	80	1827	8.1	66	3.6
252	76145	20870	27.4	90	1777	8.5	63	3.5
253	76129	22333	29.3	60	1395	6.2	17	1.2
254	75787	23164	30.6	50	1286	5.6	110	8.6
255	75265	23149	30.8	70	1614	7.0	12	0.7
256	76288	22728	29.8	80	1856	8.2	177	9.5
257	75059	22525	30.0	70	1548	6.9	204	13.2
258	75234	22580	30.0	50	1211	5.4	81	6.7
259	75336	19490	25.9	70	1428	7.3	64	4.5
260	76321	20899	27.4	90	1879	9.0	276	14.7
261	77197	21579	28.0	90	1930	8.9	213	11.0
262	75577	22500	29.8	80	1846	8.2	48	2.6
263	75477	17568	23.3	100	1654	9.4	151	9.1
264	73329	23711	32.3	70	1648	7.0	91	5.5
265	76709	21730	28.3	80	1897	8.7	516	27.2
266	76614	21097	27.5	50	1113	5.3	91	8.2
267	76975	22696	29.5	50	1223	5.4	138	11.3
268	73672	22460	30.5	60	1437	6.4	268	18.6

Table 6: 2008 Participants not submitting runs (or at least papers in the case of optional tasks)

CD	ED	FE	RU	SE	Location	Participants
–	**	–	–	–	NorthAm	Arete Associates
–	**	–	–	–	Asia	Beihang University
*	**	–	–	–	Europe	Bilkent University
*	**	–	–	–	Europe	Chemnitz University of Technology
–	**	–	**	–	Asia	Chubu University
–	**	–	–	–	Europe	Delft University of Technology
*	–	–	–	–	Europe	Digital Systems & Media Computing Laboratory
*	–	–	–	–	Europe	sEff^2 Videntifier
–	**	–	–	–	Asia	Harbin Engineering University
*	–	**	**	–	Asia	KDDI R&D Laboratories, Inc.
*	**	**	–	**	Asia	Nanyang Technological University
–	**	**	–	–	Asia	National Electronics and Computer Technology Center (NECTEC)
*	–	–	–	–	NorthAm	Northrop Grumman
–	**	–	–	–	NorthAm	Objectvideo Inc
*	**	**	–	–	NorthAm	Rensselaer Intelligent Systems Lab
–	**	–	**	**	Australia	RMIT University School of CS&IT
–	**	**	**	**	Australia	Ryerson University
–	**	–	–	–	Europe	SCOVIS consortium
–	**	**	–	–	Europe	TELECOM ParisTech
–	**	**	–	–	NorthAm	TiChen.Net LLC
–	**	–	–	**	Europe	Universidade do Porto/ INESC-Porto
–	**	**	**	–	Europe	Università degli Studi di Modena e Reggio Emilia
–	**	–	–	–	NorthAm	University of California at Los Angeles
–	**	–	–	–	NorthAm	University of Maryland, College Park - CaNVid
–	**	–	–	–	NorthAm	University of Maryland College Park - CVL
–	**	–	–	–	NorthAm	University of Texas at Austin
–	**	–	–	–	Europe	University of Glasgow - MIAUCE
*	**	–	–	–	Asia	University of Mysore
*	–	–	–	–	Australia	University of Queensland, Brisbane
–	–	**	–	–	NorthAm	University of California, Irvine
–	**	–	–	–	NorthAm	University of California, Santa Barbara
–	**	**	–	**	NorthAm	University of Iowa
*	**	**	**	**	NorthAm	University of Memphis
–	**	–	–	–	NorthAm	University of Southern California
*	–	–	–	–	NorthAm	Vercury
–	–	–	–	**	NorthAm	Video Retrieval GMU
*	–	–	–	–	Europe	Vienna University of Technology
–	**	–	–	**	NorthAm	VIKI
–	–	–	–	**	Europe	Yahoo! Research Barcelona

Task legend. CD: Copy detection; ED: event detection; FE: Feature extraction; RU: rushes summarization; SE: Search; **: Group applied but didn't submit any runs