TRECVID 2010 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics

Paul Over {over@nist.gov}
George Awad {gawad@nist.gov}
Jon Fiscus {jfiscus@nist.gov}
Brian Antonishek {brian.antonishek@nist.gov}
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

Martial Michel
Systems Plus
One Research Court, Suite 360
Rockville, MD 20850
{martial.michel@nist.gov}

Alan F. Smeaton {Alan.Smeaton@dcu.ie}
CLARITY: Centre for Sensor Web Technologies
School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland

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

Georges Quénot {Georges.Quenot@imag.fr}
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP /
CNRS, LIG UMR 5217, Grenoble, F-38041 France

April 15, 2011

1 Introduction

The Text Retrieval Conference's (TREC's) Video Retrieval Evaluation (TRECVID) 2010 was a 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 10 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 National Institute of Standards and Technology (NIST) and other US government agencies. Many organizations and individuals worldwide also contribute significant time and effort.

In 2010, TRECVID turned to new and different data and to some new tasks. 73 teams (see Table 1) from various research organizations — 27 from Europe, 32 from Asia, 12 from North America, 1 from Africa, and 1 from South America — completed one or more of six tasks:

1. content-based copy detection (CCD)
2. instance search (INS)
3. known-item search (KIS)
4. semantic indexing (SIN)
5. surveillance event detection (SED)
6. multimedia event detection (MED)

400 hours of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC), were used for semantic indexing, known-item search, and copy detection. Unlike previously used professionally edited broadcast news and educational programming, the IACC videos reflect a wide variety of content, style, and source device — determined only by the self-selected donors. About 180 hours of Sound and Vision video was reused for the instance search pilot. 45 hours of airport surveillance video was reused for the surveillance event detection task. About 100 hours from a new test collection of Internet videos (HAVIC) was used for the multimedia event detection pilot.

Copy detection submissions were evaluated at NIST based on ground truth created automatically with tools donated by the INRIA-IMEDIA group. Instance search results were judged by NIST assessors — similarly for the semantic indexing task with additional assessments done in France under the European Quaero program (QUAERO, 2010). Known-item search topics and associated ground truth were created by NIST assessors, so submissions could be scored automatically. Multimedia and surveillance event detection were scored using ground truth created manually by the Linguistic Data Consortium under contract to NIST.

This paper is an introduction to the evaluation framework — the tasks, data, and measures for the workshop. For detailed information about the approaches and results, the reader should see the various site reports and the results pages available at the back of the workshop notebook and on the TRECVID website.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, 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

Internet Archive Creative Commons (IACC) video

Approximately 8 000 Internet Archive videos (50 GB, 200 h) with Creative Commons licenses in MPEG-4/H.264 and with durations between 10 seconds and 3.5 minutes were used as test data. Most videos had some donor-supplied metadata available e.g., title, keywords, and description. Another 3 200 IACC videos (50 GB, 200 h) with durations between (3.6 and 4.1) min were designated for use in system development.

LIMSI and VecSys research provided automatic speech recognition for the English speech in the IACC video.

Georges Quénot and Stéphane Ayache of LIG (Laboratoire d'Informatique de Grenoble) again organized a collaborative annotation by TRECVID participants of 130 features against the IACC videos. using an active learning scheme designed to improve the efficiency of the process (Ayache & Quénot, 2008).

Sound and Vision data

In 2006 the Netherlands Institute for Sound and Vision generously provided 400 hours of Dutch television news magazine, science news, news reports, documentaries, educational programming, and archival video in MPEG-1 format for use within TRECVID. About 180 hours of Sound and Vision video, previously used for testing feature extraction and ad hoc search, were reused in 2010 for testing instance search.

The video had already been automatically divided into shots by Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin. These shots served as predefined units of evaluation.

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

speech recognition (ASR). The LIMSI Spoken Language Processing Group had produced a speech transcription for the TRECVID 2007-2009 Sound and Vision data using its recently developed Dutch recognizer.

iLIDS Multiple Camera Tracking Data

The iLIDS Multiple Camera Tracking data consisted of ≈150 hours of indoor airport surveillance video collected in a busy airport environment by the United Kingdom (UK) Home Office Scientific Development Branch (HOSDB). The dataset utilized 5 frame-synchronized cameras.

The training video consisted of the ≈100 hours of data used for SED 2008 evaluation. The evaluation video consisted of an additional ≈50 hours of data from Imagery Library for Intelligent Detection System's (iLIDS) multiple camera tracking scenario (UKHO-CPNI, 2007 (accessed June 30, 2009)).

One third of the evaluation video was annotated by the Linguistic Data Consortium using a triple-pass annotation procedure. Seven of the ten annotated events were used for the 2010 evaluation.

Heterogeneous Audio Visual Internet (HAVIC) Corpus

The Heterogeneous Audio Visual Internet (HAVIC) Corpus is a new, large corpus of Internet multimedia files collected by the Linguistic Data Consortium. The corpus contained ≈3 400 video clips consists of ≈114 hours of MPEG-4 (MPEG-4, 2010) formatted files containing H.264 (H.264, 2010) encoded video and MPEG-4's Advanced Audio Coding (ACC) (ACC, 2010) encoded audio. The data was collected to specifically contain 100 instances of three events: "Assembling a Shelter", "Batting in a run", "Making a Cake". The data was evenly divided up into a ≈57 hour development set and a ≈57 hour evaluation set – each set containing ≈50 instances per event.

3 Semantic indexing

A potentially important asset to help video search/navigation is the ability to automatically identify the occurrence of various semantic features/concepts 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 takes on added importance

Figure 1: xinfAP by run (cat. C) - Full

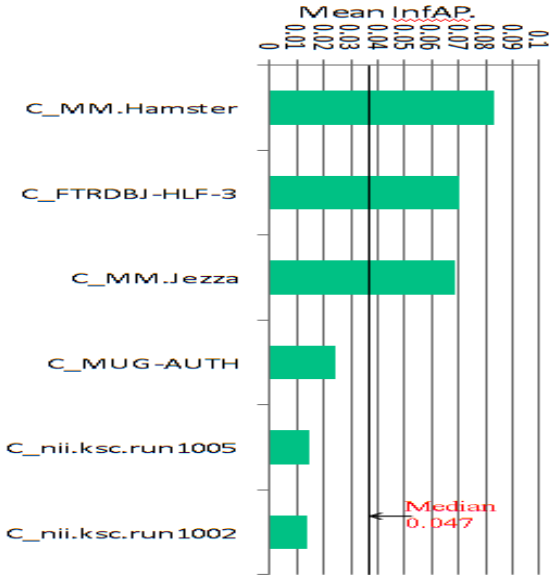


Figure 2: xinfAP by run (cat. D) - Full

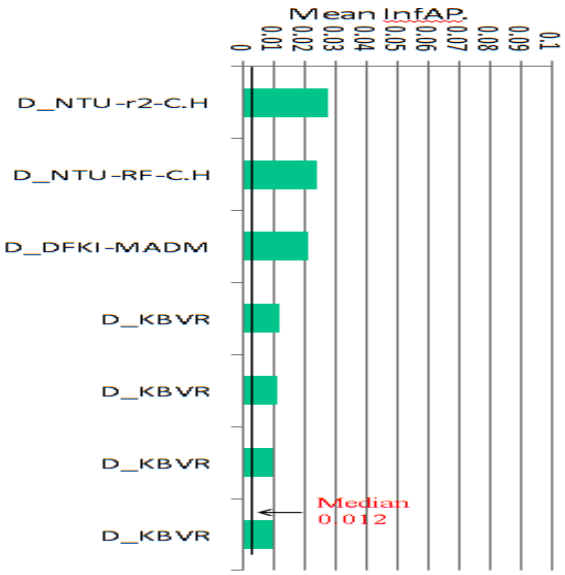
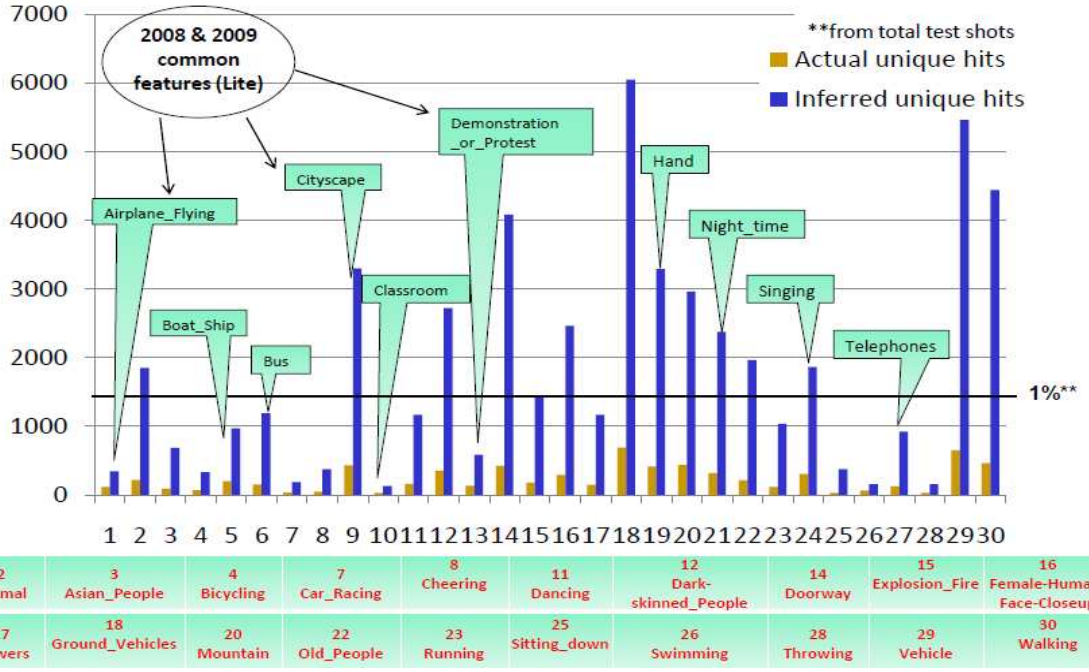


Figure 3: Frequencies of shots with each feature



to the extent it can serve as a reusable, extensible basis for query formation and search. The semantic indexing task was a follow-on to the feature extraction task. It was coordinated by NIST and by Georges Quénot under the Quaero program and had the following additional, new objectives:

- to increase the number of semantic concepts most systems can extract and the number evaluated
- to support experiments using relations in a simple ontology among the concepts to be detected
- to offer a "lite" version of the task to encourage new participation

The semantic indexing task was as follows. Given a standard set of shot boundaries for the semantic indexing test collection and a list of concept definitions, participants were asked to return for each concept in the full set of concepts, at most the top 2 000 video shots from the standard set, ranked according to the highest possibility of detecting the presence of the concept. The presence of each concept was assumed to be binary, i.e., it was either present or absent in the given shot. If the concept 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.

130 concepts were selected for the TRECVID 2010 semantic indexing task. These included all the TRECVID "high level features" from 2005 to 2009 plus a selection of Large Scale Concept Onotology for Multimedia (LSCOM: `www.lscom.org`) concepts so that we ended up with a number of generic-specific relations among them. The goal was to promote research on methods for indexing many concepts and using ontology relations between them. Also it was expected that these concepts would be useful for the content-based known-item search task. Including TRECVID 2005 to TRECVID 2009 features favored the reuse of already available annotations and judgments and encouraged cross-domain evaluations.

Two types of submissions were considered: full submissions in which participants submit results for all 130 concepts and lite submissions in which participants submit results for only 10 concepts. TRECVID only evaluated 30 concepts — 20 based on judgments done at NIST and 10 done under the Quaero program in France. The 10 features from the lite set

Figure 4: xinfAP by run (cat. A) - Full

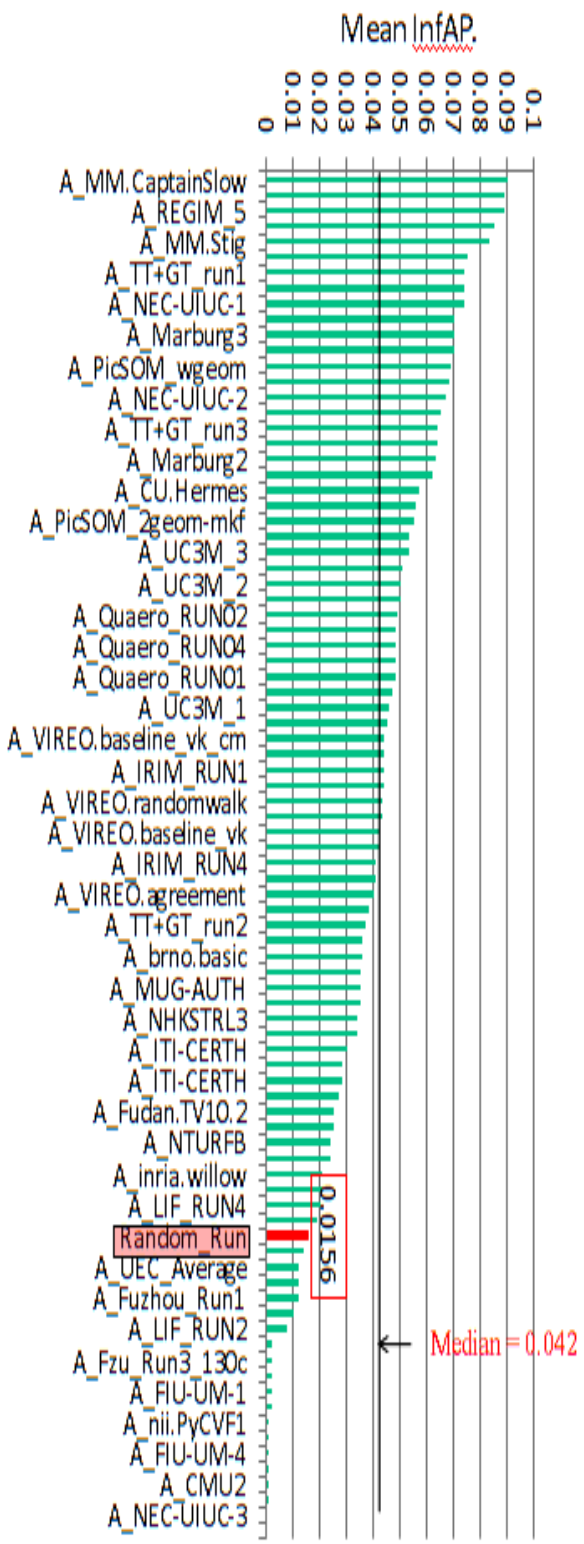


Figure 5: xinfAP by run (cat. A) - Lite

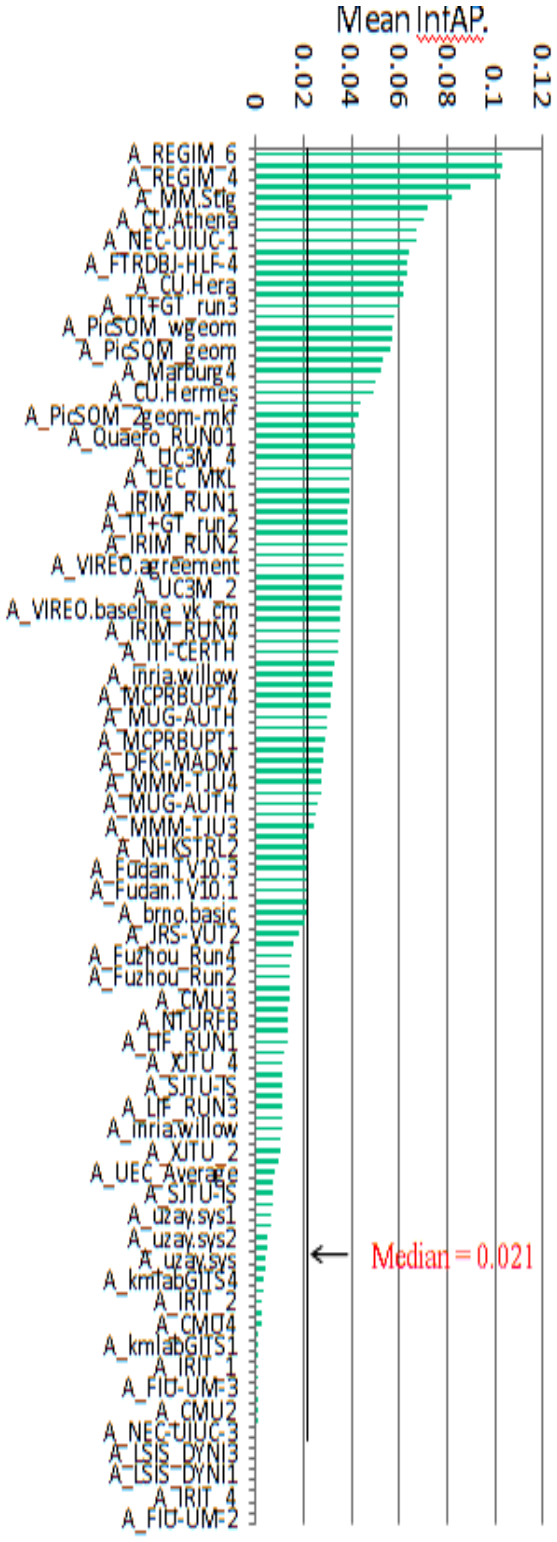


Figure 6: xinfAP by run (cat. B) - Lite

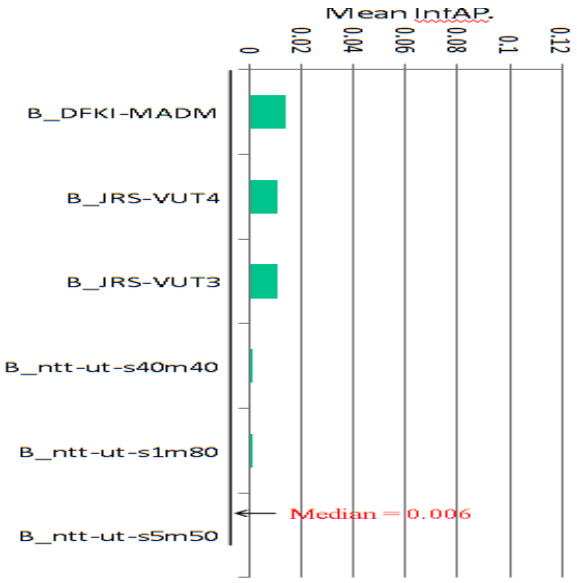


Figure 7: xinfAP by run (cat. C) - Lite

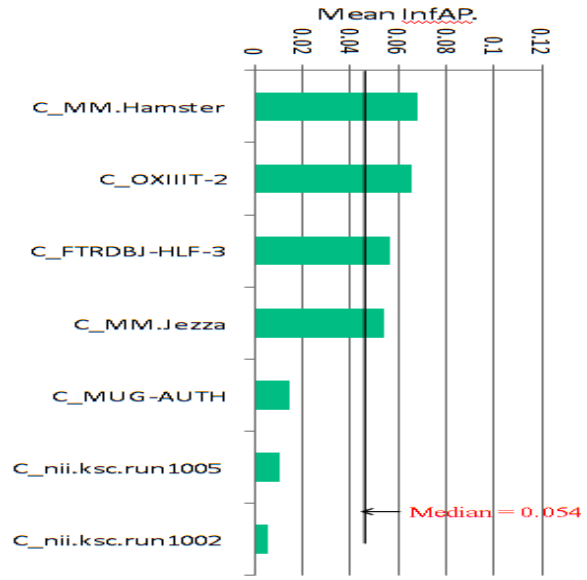
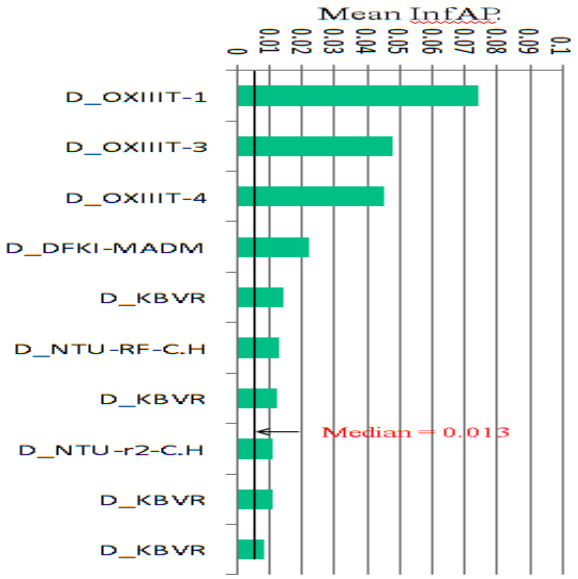


Figure 8: xinfAP by run (cat. D) - Lite



were included in the 20 judged at NIST. The 10 light concepts overlap with 2008 and 2009 high-level feature task. The 2010 30 evaluated concepts were as follows. Those marked with an asterisk are the light concepts:

[4] * Airplane-flying, [6] Animal, [7] Asian-People, [13] Bicycling, [15] * Boat-Ship, [19] * Bus, [22] Car-Racing,[27] Cheering, [28] * Cityscape, [29] * Classroom, [38] Dancing, [39] Dark-skinned-People, [41] * Demonstration-Or-Protest, [44] Doorway, [49] Explosion-Fire, [52] Female-Human-Face-Closeup, [53] Flowers, [58] Ground-Vehicles, [59] * Hand, [81] Mountain, [84] * Nighttime, [86] Old-People, [100] Running, [105] * Singing, [107] Sitting-down, [115] Swimming, [117] * Telephone, [120] Throwing, [126] Vehicle, and [127] Walking.

Concepts were defined in terms a human judge could understand. Some participating groups made their feature detection output available to participants in the search task.

The fuller concept definitions provided to system developers and NIST assessors are listed with the detailed semantic indexing runs in Appendix B in this paper.

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

Figure 9: Top 10 runs (xinfAP) by feature - Full

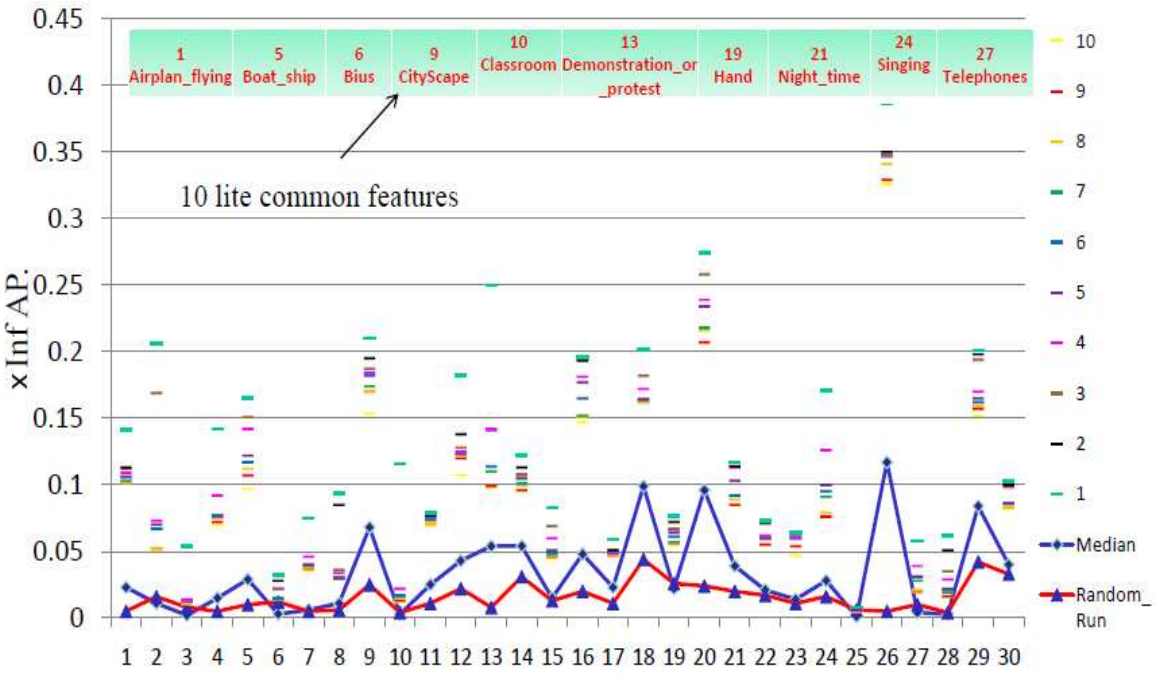
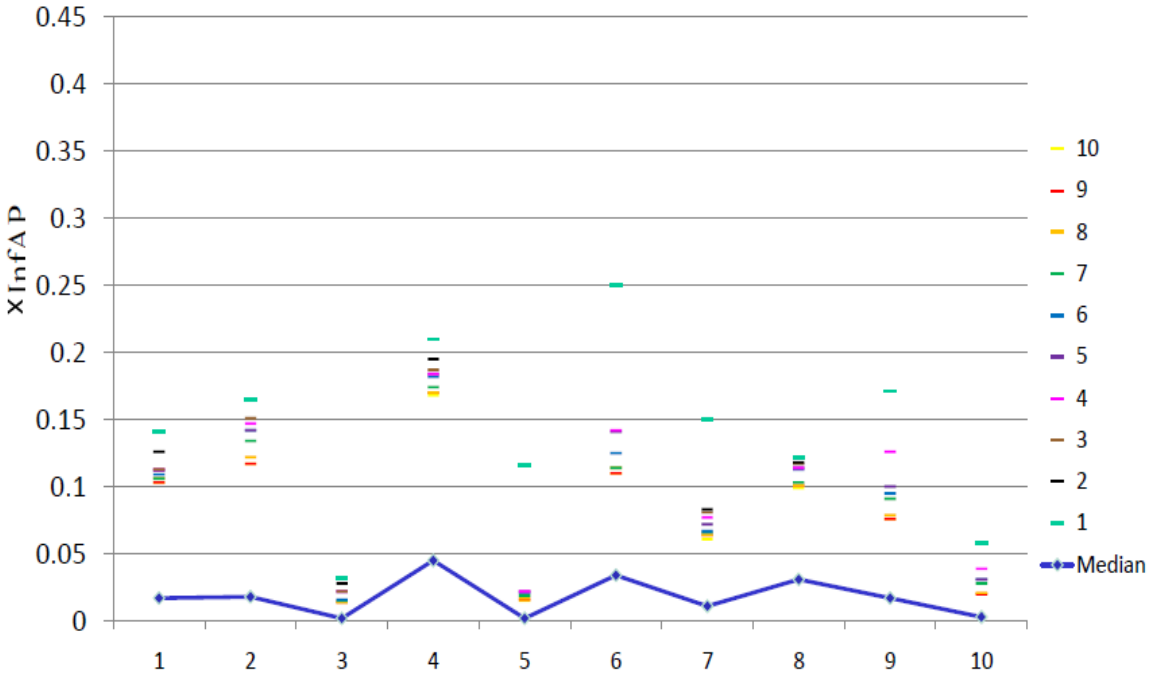


Figure 10: Top 10 runs (xinfAP) by feature - Full + Lite



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 average precision (Over, Ianeva, Kraaij, & Smeaton, 2006). This year mean extended inferred average precision (mean xinfAP) was used, which permited sampling density to vary (Yilmaz, Kanoulas, & Aslam, 2008). This allowed the evaluation to be more sensitive to shots returned below the lowest rank (≈100) previously pooled and judged. It also allowed adjustment of the sampling density to be greater among the highest ranked items,which contribute more to average precision than those ranked lower.

3.1 Data

As mentioned earlier, the IACC test collection contained approximately 8 000 files/videos in MPEG-4/H.264 format and 146 788 shots. Development data contained 3200 files/videos and approx. 119 685 shots. Testing concept detection and known item search on the same data offered the opportunity to assess the quality of concepts being used in search.

3.2 Evaluation

Each group was allowed to submit up to 4 runs and in fact 39 groups submitted a total of 150 runs including 100 full runs and 50 "lite" runs.

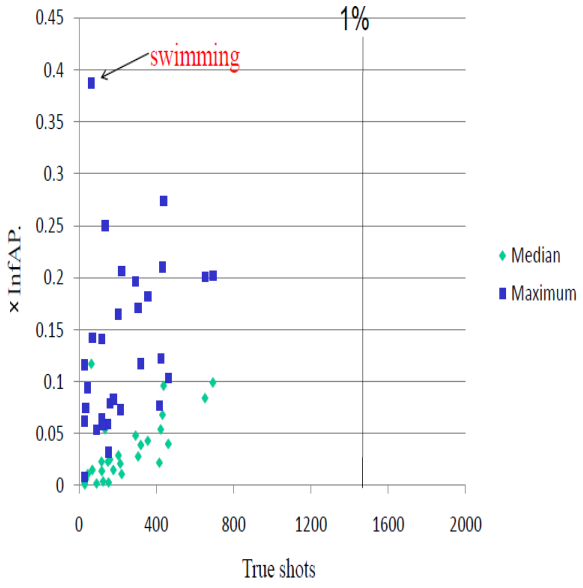
For each concept, 3 pools of shots were created as follows for use with sample_eval. The sample_eval evaluation tool is discussed below under "Measures".

1. The top pool comprised all unique shots ranked 1-10 in any submission. The medium pool included all unique shots 11-100 in any submission and not in the top pool The bottom pool held all unique shots ranked 101-2000 in any submission and not in the top or medium pools.

2. Each pool was then partitioned into a judged and an unjudged part using random sampling. The following fractions of each pool were marked for judgment: 100 % of the top pool, 20 % of middle pool, and 5 % of the bottom pool.

3. The part to be judged, (top, middle and bottom combined and randomly ordered) was presented to humans for judgment. The shots in the unjudged part were marked unjudged. After the human judgments were recorded, the union of the judged and unjudged parts made up the ground truth (qrels) used by sample_eval to score each submission.

Figure 11: Effectiveness versus number of true positives



4. Human judges (assessors) – one assessor per concept – judged each shot by watching the associated video and listening to the audio. In all, 117 058 shots were judged. 1 537 314 shots fell into the unjudged part of the overall samples. All full runs were also treated as lite runs by looking at their performance on just the 10-feature lite subset.

3.3 Measures

The sample_eval software, a tool implementing xinfAP, was used to calculate inferred recall, inferred precision, inferred average precision, etc., for each result, given the sampling plan and a submitted run. Since all runs provided results for all evaluated concepts, runs can be compared in terms of the mean inferred average precision across all 30 (or 10 lite) evaluated concepts. The results also provide some information about "within concept" performance.

3.4 Results

Performance varied greatly by feature. Figure 3 shows how many unique instances were found for

Figure 12: Significant differences among top A-category full runs

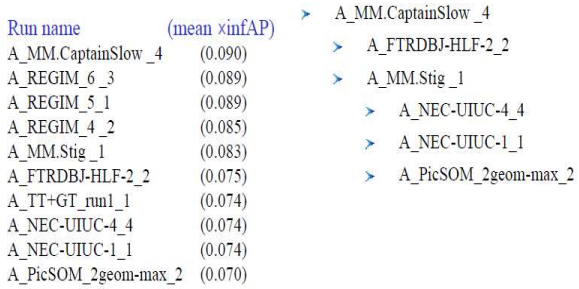


Figure 13: Significant differences among top C-category full runs

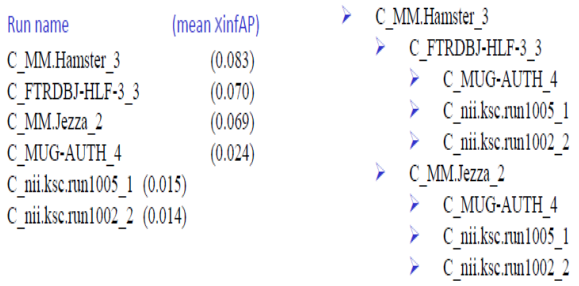


Figure 14: Significant differences among top D-category full runs

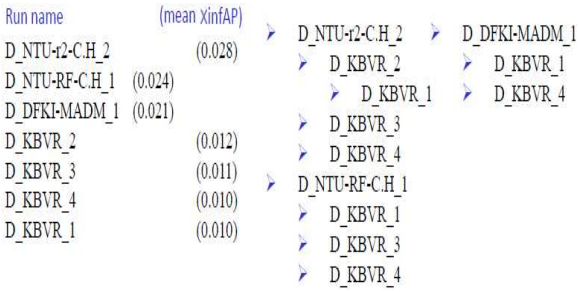


Figure 15: Significant differences among top A-category lite runs

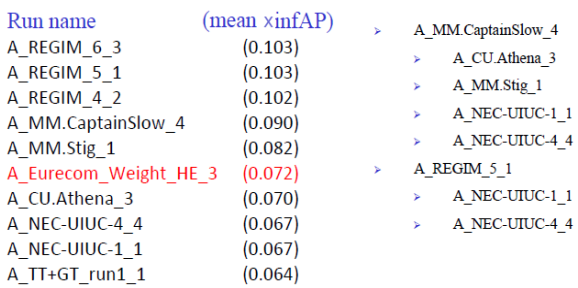


Figure 16: Significant differences among top B-category lite runs

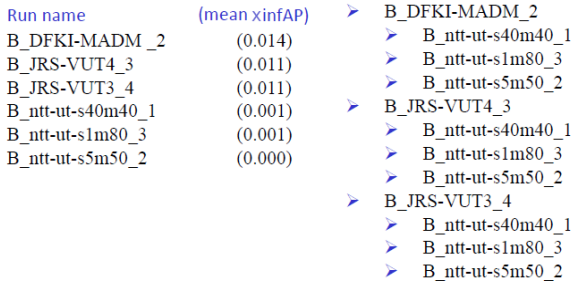


Figure 17: Significant differences among top C-category lite runs

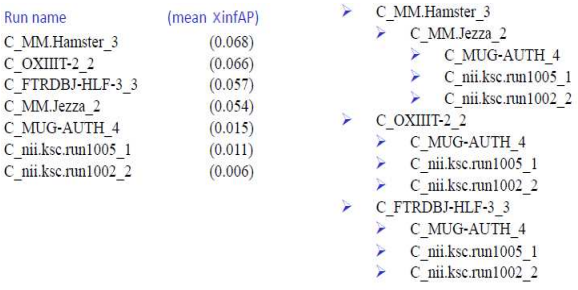
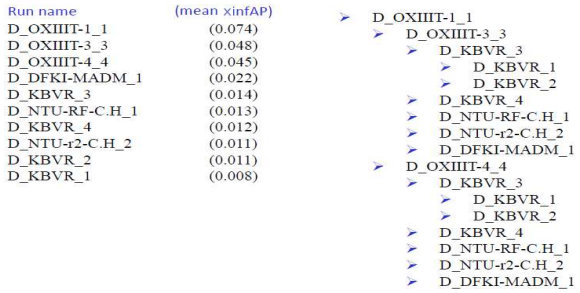


Figure 18: Significant differences among top D-category lite runs



each tested feature. The inferred true positives (TPs) of 13 features exceeded 1 % TPs from the total tested shots percentage. Features "Vehicle" and "Ground_vehicle" had TPs in over 3 % of the test shots. On the other hand, features that had the fewest TPs were "Classroom", "Swimming", "Throwing", and "Car_racing". It can also be shown that features such as "Explosion_fire" received TPs very near to the 1 %. It is worth mentioning that 4 features namely "Hand", "City_scape", "Night", and "Singing" of the 10 common features from 2008 and 2009 were among the top performing features.

Figures 4, 1, and 2 show the results of category A,C and D for full runs. The graphs show the median values in each category together with a random baseline result (as described below) for category A. A small number of runs are below the randomly generated result. Still category A runs are the most popular type and achieve top recorded performances.

For the random baseline, two estimations were made. The first one relies on the idea that the performance of a random run theoretically depends only upon the concept frequency or the total number of found TPs. For each concept, 10 000 result sets were randomly constructed with the density of TPs estimated from the actual density of TPs for the concept in the judged pools. The trec_eval program was then applied on each of them and the obtained MAPs were averaged on the 10 000 produced sets. The obtained value for the random run with this method was of 0.0156 (this is the value selected for inclusion in Figure 4.

For the second random baseline estimation, 10 000 random permutations of all the shots ids were generated and the top 2 000 were selected. The sample_eval program was then applied with the reference qrels file and the obtained xinfAPs were averaged on the 10,000 generated submissions. The obtained value for the random run with this method was of 0.000265 $\pm$ 0.000147.

The value obtained with the second method is much lower than the value obtained with the first method. It is also much smaller than the estimated average concept frequency that is of 0.0123. The source of this difference is under investigation.

How reusable is a set of judgments based on pooling? In particular how much of a difference would a system see between results based on pools it contributed to versus results based on pools it did not contribute to? This is always a question about TRECVID results using pooling. The use of stratified sampling (sample_eval) seemed possibly to further complicate the question of reusability. So we ran a "what if" (hold-one-out) test - what if my run were evaluated (e.g. after TRECVID 2010) so it couldn't contribute to the pools? Would its score be significantly different from the official score it got when it could be represented in the pooling?

Out of 150 SIN runs, 91 contributed at least one unique feature-shot before pooling and sampling. Each of those runs was evaluated against the official ground truth they had contributed to and against a new temporary ground truth from which all their uniquely contributed feature-shots had been removed. For 28 of the 91 pairs a randomization test (10 000 iterations, $p < 0.05$) found a significant difference between the official and the hold-one-out results, but the largest difference was 0.0015 and most differences were less than 0.001 and so not likely to be important for practical purposes. We conclude there is good evidence for the reusability of TV2010 SIN judgments using sample_eval.

Figures 5, 6, 7 and 8 show the results of category A,B,C and D for the lite runs respectively together with their median values. As in full runs, category A of lite runs were the best performing in general. Category A runs used only IACC training data. Catetory B runs used only non-IACC training data. Category C runs used both IACC and non-IACC TRECVID training data. Category D runs used both IACC and non-IACC non-TRECVID training data.

Figure 9 shows the performance of the top 10 teams across the 30 features. The behavior varied generally across features. For example some features reflected a large spread between the scores of the top 10 such as feature "Animal", "Bicycling", "Singing", and "Demonstration_or_protest". This indicates that

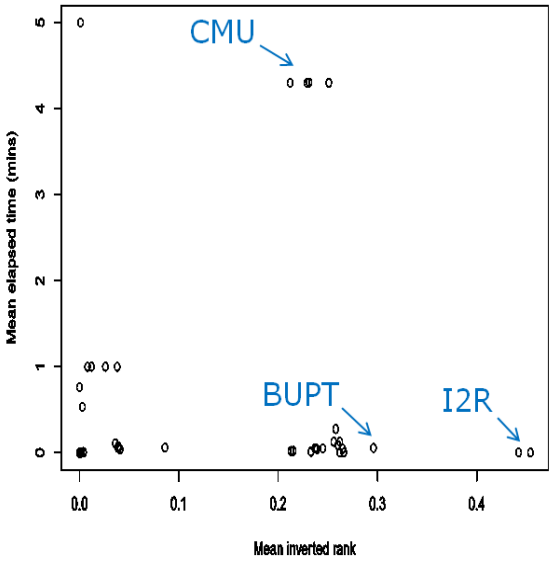
there is still room for further improvement, while other features had a tight spread of scores among the top 10 such as feature "Sitting_down","Dancing", "Flowers", and "Running". In general, the median scores ranged between 0.001 (feature "Sitting_down") and 0.117 (feature Swimming). As a general observation, feature "Sitting_down" had the minimum spread across the top 10 and at the same time the minimum median score across all systems, which demonstrates how difficult this feature is for the systems to detect. Also, it can be shown on the graph that the median curve was above the random baseline run generated by NIST except for 8 features, for which the random and median values were very close and in general were from the low performance features.

A similar graph for the 10 common features is Figure 10 which shows the performance of the top 10 teams for both the lite and full runs. Features that reflected a large spread between the scores of the top 10 are "Hand", "Classroom", "Demonstration_or_protest", and "singing". While the lowest performing feature was "Bus". As a general observation, the top 10 performance for the majority of the common features were less than the top 10 scores for 2009. This was probably due to the high variation between this year's data and the last 3 years'. More research is needed toward developing systems that generalize well among different datasets.

Figure 11 shows the relation between the xinfAP and number of true shots detected by systems in terms of their median and maximum values for the 30 features. A positive correlation between number of TPs and accuracy can be in general concluded. There are a few features (e.g "Swimming") where relatively small number of TPs produced high xinfAP scores. It can also be shown that the maximum TPs for a feature didn't exceed $\approx 0.5\,\%$ of the number of test shots.

To test if there were significant differences between the systems performance, we applied a randomization test (Manly, 1997) on the top 10 runs for each run type and training category as shown in Figures 12 through 14 for full runs and Figures 15 through 18 for lite runs. 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. Different levels of indentation signifies a significant difference according to the test. Runs at the same level of indentation are indistinguishable in terms of the test. In all tests except one (Figure 15) the top ranked run was significantly better than other

Figure 19: Mean inverted rank versus mean elapsed time for automatic runs



runs.

Based on site reports, some general observations on approaches can be made. Experiments involved focusing on robustness, merging many different representations, use of spatial pyramids, sophisticated fusion strategies, efficiency improvements (e.g. use of graphics processing units), analysis of multi keyframe per shot, audio analysis, using temporal context information and less highlighting on motion information, metadata or ASR. Some usage of training data from YouTube were utilized. As in previous years, most runs were in training category A (i.e less external data). The most common features used were Scale-invariant feature transforms (SIFT), color, and edge histograms and their variations. Audio features were mainly based on mel-frequency cepstral coefficients (MFCC). Still the most common classifier used is support vector machines (SVM). Readers should see the notebook papers posted on the TRECVID website (trecvid.nist.gov) for details about each participant's experiments and results.

4 Known-item search

The known-item search task models the situation in which someone knows of a video, has seen it before, believes it is contained in a collection, but doesn't

Figure 20: Mean inverted rank versus mean elapsed time for interactive runs

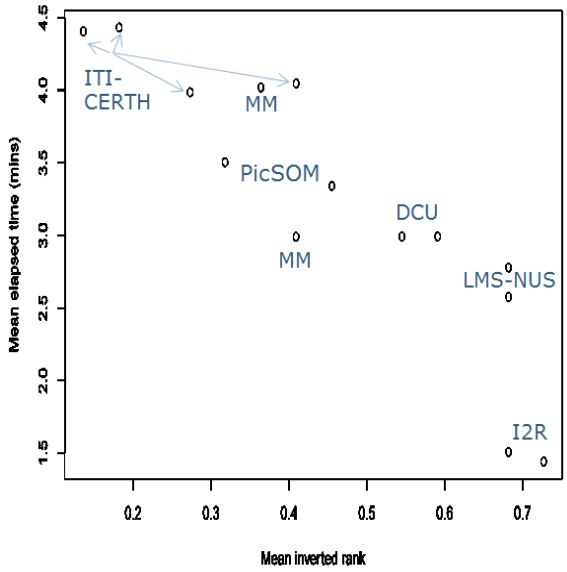


Figure 21: Topic variation (1)

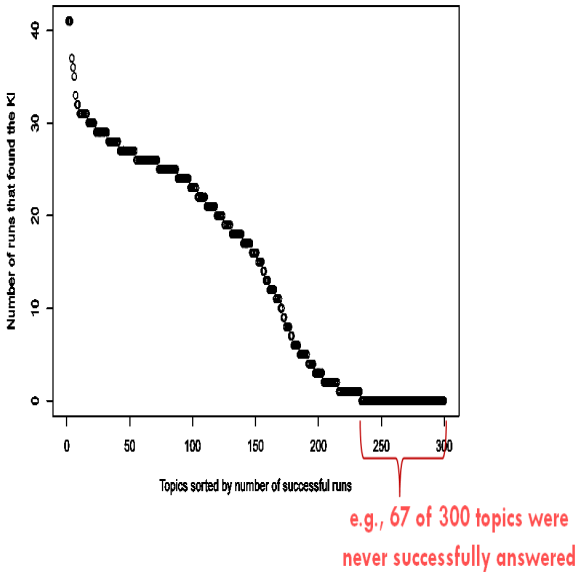


Figure 22: Topic variation (2)

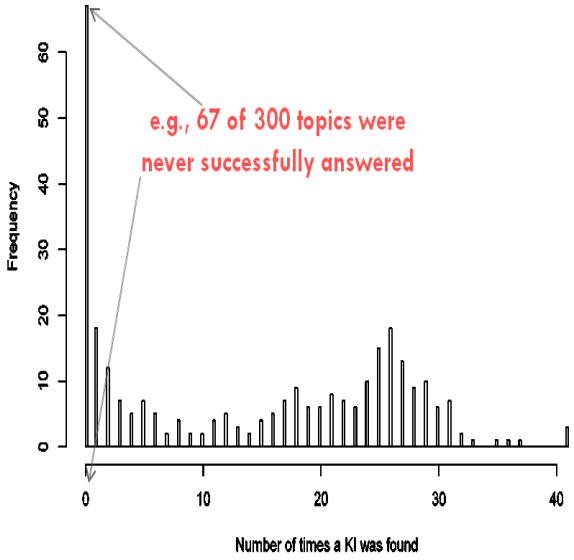
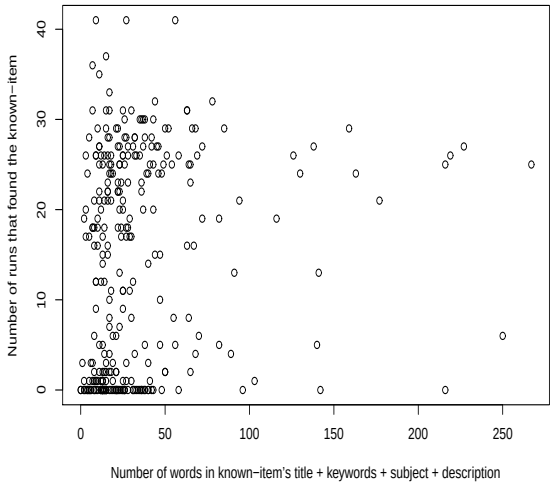


Figure 23: Topic difficulty versus amount of metadata



know where to look. To begin the search process, the searcher formulates a text-only description, which captures what the searcher remembers about the target video. This task is very different from the TRECVID ad hoc search task, in which the systems began with a textual description of the need together with several image and video examples of what was being looked for.

4.1 Task

Given a text-only description of the video desired (i.e. a topic) and a test collection of video with associated metadata:

- automatically return a list of up to 100 video IDs ranked by confidence of being the one sought. There was no time limit on automatic searches but the elapsed time for each search - from the time the topic is presented to the system until the search result for the topic is frozen as complete - had to be submitted with the system output; or
- interactively return the ID of the sought video and elapsed time to find it. No more than 5 minutes could elapse from the time the topic is presented to the system/searcher until the search result for the topic was frozen as complete. Interactive systems were able to query a web-based service to find out if a given video file was the known-item sought – this to simulate the fact that searchers looking for their own known-item would recognize it if they found it and stop the search. Each such query was logged and all logs published with the TRECVID workshop results.

The topic also contained a list of 1 to 5 words or short phrases, each identifying an object/person/location that should be visible in the target video.

4.2 Data

The test data set (IACC.1.A) was 200 hours drawn from the IACC.1 collection using videos with durations between 10 seconds and 3.5 minutes.

4.3 Topics

300 text-only topics were created by NIST assessors. For each of the random sample of IACC videos assigned to them, they were told to watch the video at least once, pause, and then formulate a brief textual query that would likely be satisfied only by the video they just watched. Finally they were asked to choose from the topic 1 to 5 objects, people, or events and list those as part of the topic.

4.4 Evaluation

Since the target video was determined for each topic as during topic creation, evaluation could be automatic.

4.5 Measures

Automatic runs were scored against the ground truth using mean inverted rank at which the known item is found or zero if not found. Note: "mean inverted rank" means the same thing as the older term "mean reciprocal rank". In TRECVID 2011 and beyond we will drop "mean inverted rank" and use "mean reciprocal rank" instead. For interactive runs, which returned either one or no known items per topic, mean inverted rank measures the fraction of all topics for which the known item was found. For interactive runs elapsed time and user satisfaction (Likert scale 1-7 (most satisfied) were also measured.

4.6 Results

Fifteen runs were submitted and evaluated. Of those, 5 were interactive and 15 automatic. The highest mean average inverted rank for interactive runs was 0.727 (I_A_YES_I2R_INTERACTIVE_KIS_2_1) and for automatic runs it was 0.454 (F_A_I2R_AUTOMATIC_KIS_2_1). As shown in Figure 19, most automatic runs required about the same mean elapsed time but in that time achieved a wide range of mean inverted rank scores. For interactive runs, as seen in Figure 20, mean inverted rank improved as mean elapsed time decreased.

The topics varied in how many systems/runs were able to find the known-item. 67 of 300 topics were not found by any run. Figures 21 and 22 present two views of the same topic distribution. In general results suggested that use of topic text and video metadata was the best approach with automatic speech recognition adding some benefit. Did the 67 topics all systems failed on simply lack metadata? Figure 23, using word count in the title, description, keywords and subject fields as a rough measurement of "amount of metadata", suggests this is not the case. Most of the 67 topics have between 0 and 50 words

Figure 24: Average precision for automatic runs by topic/type

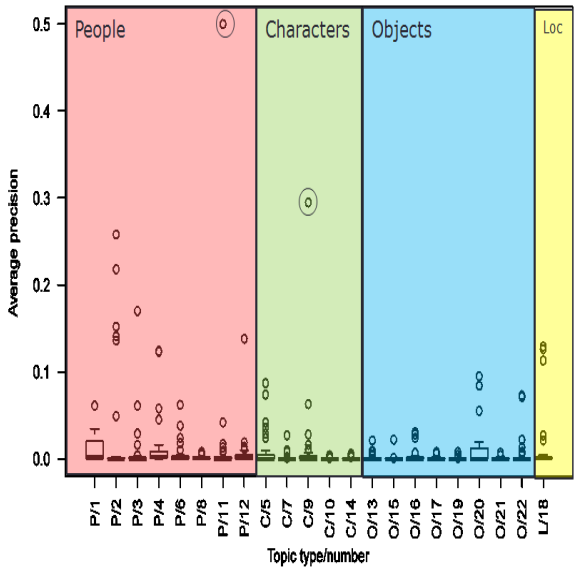
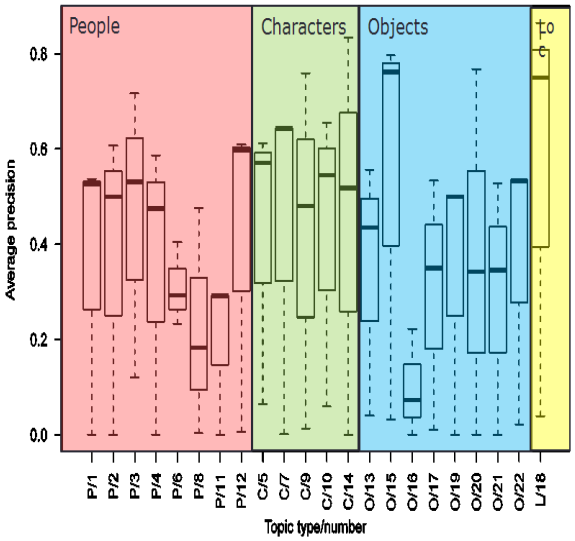


Figure 25: Average precision for interactive runs by topic/topic



of metadata, but so do many topics which multiple systems handled successfully.

To investigate if a topic had actually two duplicate true answers we conducted a small experiment were we counted all the submitted pairs of [topic, video] by all runs. Then we checked manually the sorted list of counts from the highest count for the pairs that does not match the ground truth. In theory, if many runs agree on a specific video for a topic then there is a big chance that it could be a duplicate for the true known item video in the ground truth. Our results revealed a duplicate video for topic 92 and 250. Also, for topic 250 (John Kerry and text about him) many runs returned videos very similar to the true known item video which may indicate that either this topic was not a good candidate or the language used to describe the topic needed to be more specific and less general.

For details about approaches and results the reader is referred to the notebook papers on the TRECVID website: `www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html`.

Figure 26: Example character targets

5 Instance search pilot

An important need in many situations involving video collections (archive video search/reuse, per-

Figure 27: Example location targets

Figure 28: Example object targets

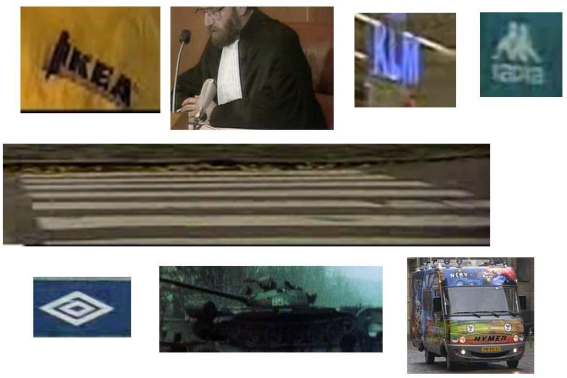
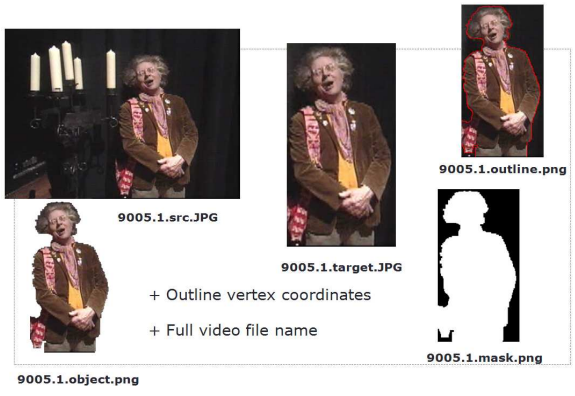


Figure 29: Example people targets

Figure 30: Example segmentations



sonal video organization/search, surveillance, law enforcement, protection of brand/logo use) is to find more video segments of a certain specific person, object, or place, given one or more visual examples of the specific item.

In 2010 this was a pilot task — evaluated by NIST but intended mainly to explore task definition and evaluation issues using data and an evaluation framework in hand. The task was a first approximation to the desired full task using a smaller number of topics, a simpler identification of the target entity, and less accuracy in locating the instance than would be desirable in a full evaluation of the task.

5.1 Task

The instance search task for the systems was as follows. Given a collection of test videos, a master shot reference, and a collection of queries that delimit a person, object, or place entity in some example video, locate for each query the 1000 shots most likely to contain a recognizable instance of the entity. Each query consisted of a set of...

- 5 or so example frame images drawn at intervals from a video containing the item of interest. For each frame image:
  - the rectangular region within the frame image, containing the item of interest
  - a binary mask of an inner region of interest within the rectangle
  - the inner region against a gray background
  - the frame image with the inner region region outlined in red
  - a list of vertices for the inner region region
- the video from which the images were selected
- an indication of the target type taken from this set of strings (PERSON, CHARACTER, LOCATION, OBJECT)

5.2 Data

Test data: Sound and Vision data from TRECVID 2007-2009 (tv9.sv.test).

5.3 Topics

In a first approximation to the full task, most queries were created by NIST and targeted actors that appeared as themselves or as characters in Sound and Vision programs – in different clothes, costumes, settings, etc. In a few cases objects (including logos) and locations were targeted. Figures 26-29 show images of all search targets from topics. Figure 30 shows the various segmentations of an example target image provided to systems as part of the topic.

As this was a pilot task, participants were encouraged to help by examining the test data and contributing up to 5 topics per team with non-person/character targets. Several teams did so. See Appendix A for a listing of the topics.

5.4 Evaluation, Measures

This pilot version of the task was treated as a form of search and evaluated accordingly with average precision for each query in each run and per-run mean average precision over all queries. While speed and location accuracy were also definitely of interest here, of these two, only speed was measured in the pilot.

For each topic, the runs were pooled and presented to the human judges in the order the shots were ranked by the systems. The depth to which each topic's pool was judged depended in part on how many true positives were found as the depth of the shots being judged increased. See Table 3 for details on the pooling and judging.

5.5 Results

15 research teams submitted runs. Although not part of the official task design, NIST allowed the ITI-CERTH team to run an interactive version of the task to provide additional context for the results. The top half of the automatic runs had mean average precision (MAP) scores ranging from 0.01 to 0.033. The two ITI-CERTH interactive runs achieved much higher MAP scores: 0.524 and 0.534.

There was considerable difference in performance from topic to topic. Figure 24 depicts the distribution of scores by topic for the people, character, and object types. Figure 25 does the same for the much smaller number of interactive runs.

During the TRECVID 2010 Workshop there was a panel discussion out of which came the suggestion that if we continue to use small targets, then we should use better quality video.

Results this year were of a very preliminary nature. For details about approaches and results the reader is referred to the notebook papers on the TRECVID website: `www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html`.

6 Multimedia event detection pilot

The 2010 Multimedia Event Detection (MED) pilot evaluation was the first evaluation of technologies that search multimedia video clips for events of interest to a user. An event for MED:

- is a complex activity occurring at a specific place and time;
- involves people interacting with other people and/or objects;
- consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and semantic relationships to the overarching activity;
- is directly observable.

A user searching for events in multimedia material may be interested in a wide variety of potential events. Since it is an intractable task to build special purpose detectors for each event a priori, a technology is needed that can take as input a human-centric definition of an event that developers (and eventually systems) can use to build a search query. The events for MED were defined via an event kit which consisted of:

- An event name which is an mnemonic title for the event.
- An event definition which is a textual definition of the event.
- An evidential description which is a textual listing of the attributes that are indicative of an event instance. The evidential description provides a notion of some potential types of visual and acoustic evidence indicating the event's existence but it is not an exhaustive list nor is it to be interpreted as required evidence.
- A set of illustrative video examples each containing an instance of the event. The examples are illustrative in the sense they help form the definition of the event but they do not demonstrate all the inherent variability or potential realizations.

The 2010 MED evaluation was a pilot for two main reasons. First, only three events, Assembling a Shelter, Batting in a run, and Making a cake, were used for the evaluation. Future evaluations will involve ten events and new events will be tested each year. Second, the data resources were small on the order of 100 hours as opposed to 1000s of hours for future evaluations.

Figure 31: TRECVID 2010 MED Participants Chart

2010 Participants 7 Sites, 45 Submission Runs		Number of Submissions		
		assembling_shelter	batting_in_run	making_cake
Center for Research and Technology, Hellas - Informatics and Telematics Institute	CERTH-ITI	9	9	9
Carnegie Mellon University	CMU	8	8	8
Columbia University / University of Central Florida	Columbia-UCF	6	6	6
IBM T. J. Watson Research Center / Columbia University	IBM-Columbia	10	10	10
KB Video Retrieval (Etter Solutions LLC)	KBVR	1	1	1
Mayachitra, Inc.	Mayachitra	2	2	2
Nikon Corporation	NIKON	9	9	9
Total Submissions per Event		45	45	45

For this pilot MED evaluation, there were seven teams that submitted results. Though a team only needed to submit scores for a single event to be able to participate, it was encouraging to see that each team submitted results for all three events. Most teams also submitted multiple contrastive systems in addition to their primary system bringing the total number of submission runs processed to 45 for the seven participating teams. Shown in Figure ≈31 are the participating teams and their submissions for the 2010 MED events.

6.1 Data

A new collection of Internet multimedia (i.e., video clips containing both audio and video streams) was provided to MED participants. The data, which was collected and distributed by the Linguistic Data Consortium, consists of publicly available, user-generated content posted to the various Internet video hosting sites. Instances of the events were collected by specifically searching for target events using text-based Internet search engines. All video data was reviewed to protect privacy, remove offensive material, etc., prior to inclusion in the corpus.

Video clips were provided in MPEG-4 formatted files. The video was encoded to the H.264 standard.

The audio was encoded using MPEG-4's Advanced Audio Coding (AAC) standard.

The video data collection was divided into two data sets:

- Development data consisted of 1746 total clips (c. 56 hours). The development data set included nominally 50 instances of each of the three MED events and the rest of the clips were not on any of the three MED events.
- Evaluation data consisted of 1742 total clips (c. 59 hours). The evaluation data set included normally 50 instances per event.

6.2 Evaluation

Sites submitted system outputs for any combination of the three events. Outputs included a detection score which expresses the strength of evidence supporting the existence of the event and detection decision (yes/no) for each event observation.

Submission performance was computed using the Framework for Detection Evaluation (F4DE) toolkit. Groups were required to submit a primary run, which was the run they expect to be their best performing system and optionally allowed to submit multiple runs with contrastive conditions.

6.3 Measures

Since detection system performance is a tradeoff between probability of miss and false alarms, this task used the Normalized Detection Cost (NDC) measure for evaluating system performance. NDC is a weighted linear combination of the system's Missed Detection Probability and False Alarm Probability. NDC is defined in terms of the system (S) and a particular event (E) as follows:

$$P_{Miss} = \frac{N_{miss}(S,E)}{N_{Target}(E)}$$

$$P_{FalseAlarm} = \frac{N_{FalseAlarm}(S,E)}{N_{NonTarget}(E)}$$

$$NDC(S,E) = \frac{Cost_{Miss} \times N_{miss}(S,E) + Cost_{FalseAlarm} \times N_{FalseAlarm}(S,E)}{MINIMUM(Cost_{Miss} * P_{Target}, Cost_{FalseAlarm} * (1 - P_{Target}))}$$

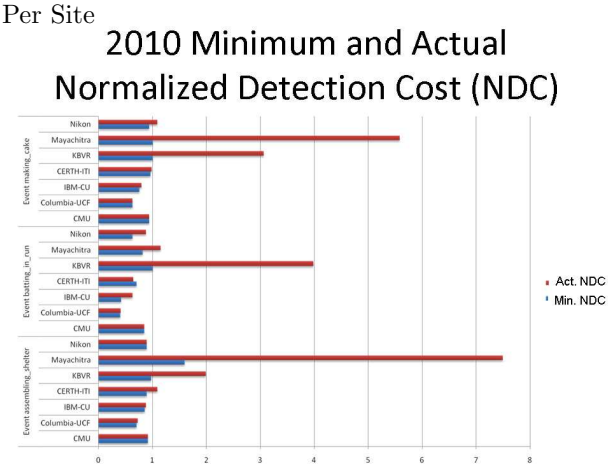
Where the event detection constants were assigned these values: $Cost_{Miss} = 80$, $Cost_{FA} = 1$ and $P_{Target} = 0.001$.

A perfect NDC score is 0. NDC is scaled so that an NDC of 1.0 would be the cost of a system with no output (no false alarms and all misses). NDC may exceed 1.0.

Using the submitted decision scores allowed us to compute Decision Error Tradeoff (DET) curves for the systems. Participants were provided with a graph of the DET curve plotting $P_{Miss} vs. P_{FalseAlarm}$ for each event their system participated in.

6.4 Results

Figure 32: TRECVID 2010 MED - NDC Per Event, Per Site



Shown in Figure 31 are the participating teams' computed NDC values for each of the events. Both the actual NDC of the submitted system and the minimal (optimal) NDC of the system based on the submitted decisions scores is shown.

Shown in Figure 32 are the participating teams' computed NDC values for each of the events. Both the actual NDC of the submitted system and the minimal (optimal) NDC of the system based on the submitted decisions scores is shown. Figure 36 shows the highest performing overall system (including all primary and contrastive systems) for the three events.

Figures 33, 34, and 35 show the primary system's values for each of the three individual events: Assembling a Shelter, Batting in a Run, and Making a Cake.

6.5 Summary

This overview of the TRECVID 2010 MED task has provided an introduction to the goals, data, evaluation methods and metrics, and results. For more

Figure 33: TRECVID 2010 MED - Each Team's Primary System for "Assembling a Shelter" Event

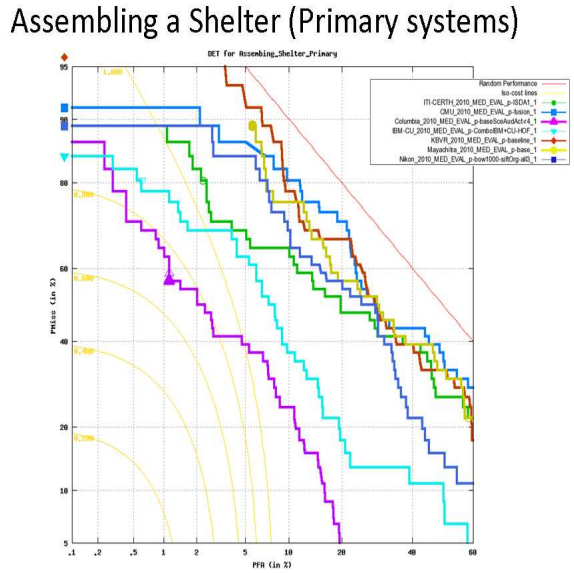


Figure 34: TRECVID 2010 MED - Each Team's Primary System for "Batting in a Run" Event

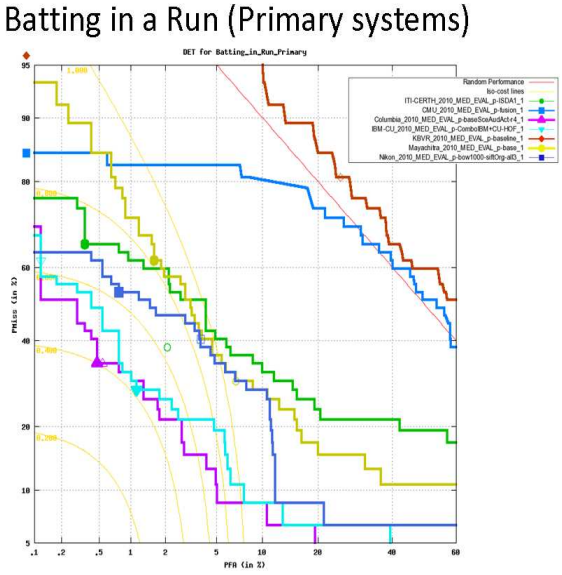


Figure 35: TRECVID 2010 MED - Each Team's Primary System for "Making a Cake" Event

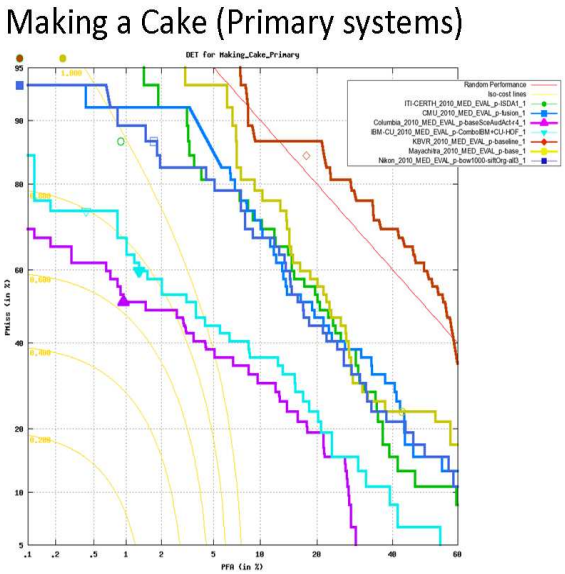
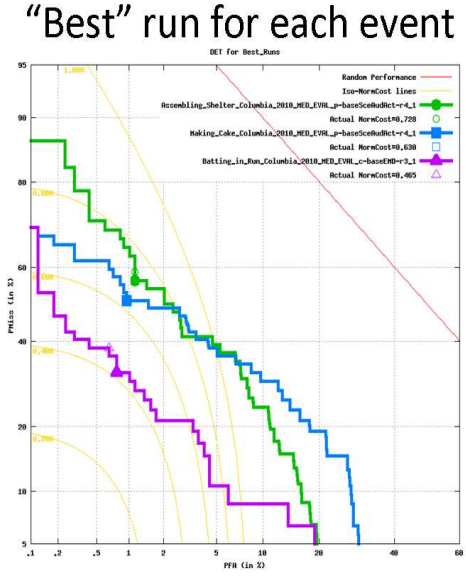


Figure 36: TRECVID 2010 MED - Best Systems for each Event



in-depth information, or detailed approaches and results of individual participants, the reader is referred to the notebook papers on the TRECVID website: `www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html`.

7 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, advertisement tracking, law enforcement investigations, etc. Content-based copy detection offers an alternative to watermarking.

As the audio plays an important role in detecting copied videos, this year systems were required to submit runs for only one required query type task (video + audio queries). Systems had the option to individually evaluate video-only and audio-only query types. Two application profiles were required to be simulated. One that required a balanced cost for misses and false alarms and one that required no false alarms (thus very high cost for false alarms). Systems were required to submit a decision score threshold believed to correspond to the best performance for the run.

The required system task was as follows: given a test collection of videos and a set of 11 256 queries, 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 8 possible video transformations was selected to reflect actually occurring video 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. In total 1608 video-only queries were constructed. 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 simulation, 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.

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 11 256 audio + video queries was provided by NIST. These queries comprised all the combinations of transformed audio(7) and transformed video (8) from a given base audio+video query (201).

7.1 Data

The new Internet Archive video collection was used as a source for reference and non-reference videos. This year's testing and development videos (11 200 files) of 400 hours and duration less than 4.1 min were used as a source from which the test query generation tools chose reference video. While the non-reference video collection was selected from a set of 12 480 videos with total duration of 4000 hours and duration between 10-30 min.

7.2 Evaluation

In total in 2010, 22 participant teams submitted 78 runs for evaluation. 41 runs were submitted as balanced runs and 37 as no false alarms. 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

After creating the query set, it was found that 5 base queries have to be dropped from evaluation because there exist multiple answers for them in the reference set or because some were taken from original corrupted videos.

7.3 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 2010 the cost model assumed copies are very rare (e.g. 0.005/h) then two application profiles were required. The Balanced profile in which misses and false alarms are assigned a cost of 1, and the NOFA profile in which a false alarm is assigned a cost of 1000 times the cost of a miss. Other realistic scenarios were of course possible. Normalized minimal detection cost rate (minNDCR) reduced in 2010 to two terms involving two variables: probability of a miss (Pmiss) and the number of false alarms (FA). The total length of queries per transformation in hours was 4.6. For the "Nofa" profile:

  $$minNDCR = Pmiss + 108.7 * FA$$

  For the same queries under the "Balanced" profile:

  $$minNDCR = Pmiss + 0.1 * FA$$

- Copy location accuracy: average harmonic mean (F1) score combining the precision and recall of the asserted copy location versus the ground truth location

- Copy detection processing time: mean processing time (s)

Finally, the submitted run threshold was used to calculate the actual Normalized Detection Cost Rate (NDCR) and F1 and those results were compared to the minNDCR and F1 using the optimal threshold calculated by the DET curve.

7.4 Results

The detection performance among best runs for both profiles across all transformations is shown in Figures 37 to 40. In general this year it could be seen that the detection performance was better than in TRECVID 2009 (lower NDCR values). For the balanced profile a noticeable difference can be seen between the actual and optimal results while for "no false alarms" (NOFA) profile the difference was very small. Finally, transformations 3,4 and 5 achieved the best performance which was likely due to the fact that those transformations (insertion of patterns, re-encoding,

Figure 37: Top runs based on Actual DET score in balanced profile

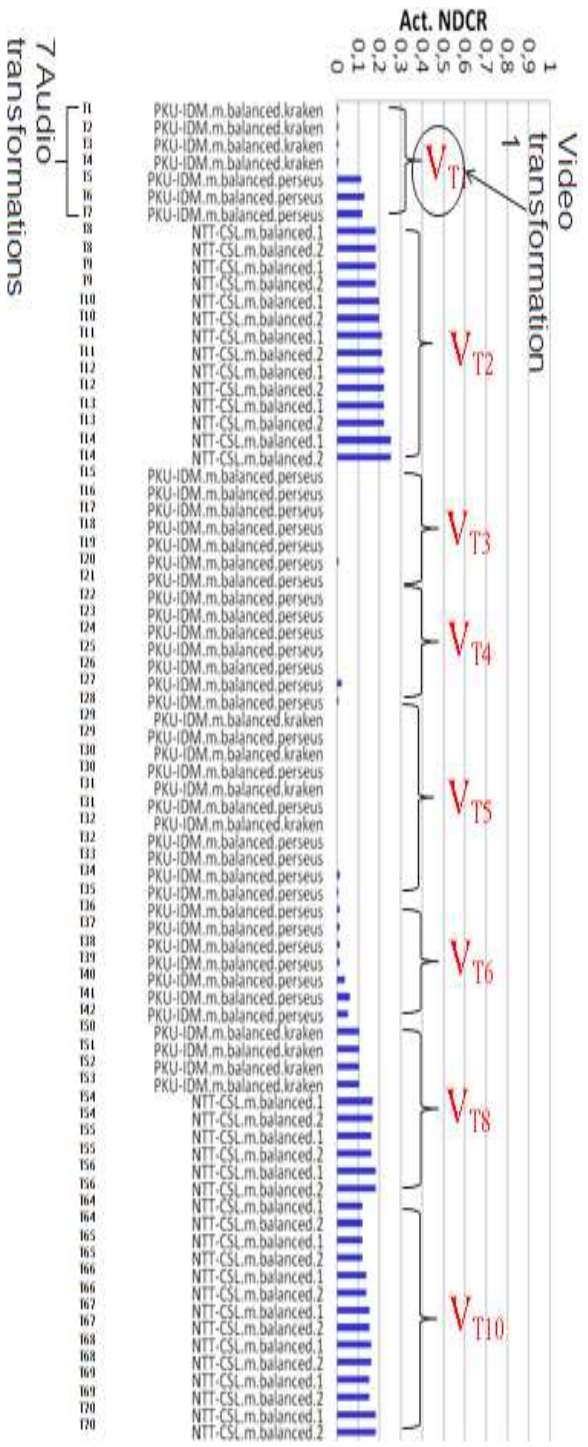


Figure 38: Top runs based on Optimal DET score in balanced profile

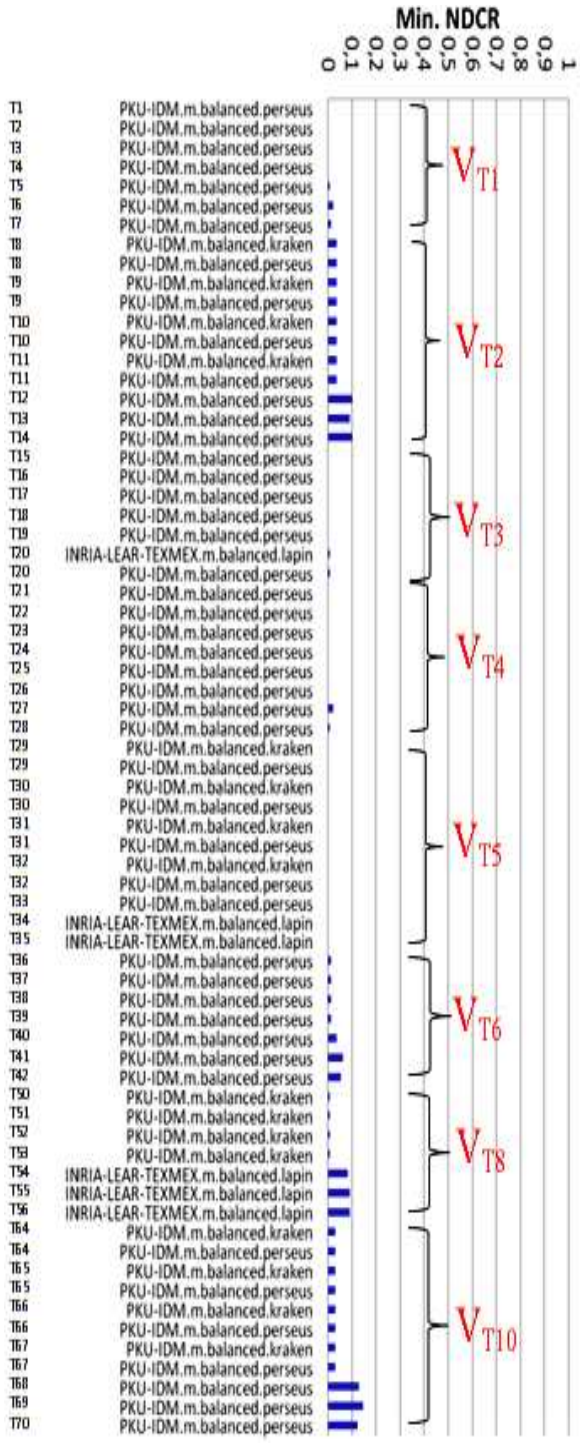


Figure 39: Top runs based on Actual DET score in Nofa profile

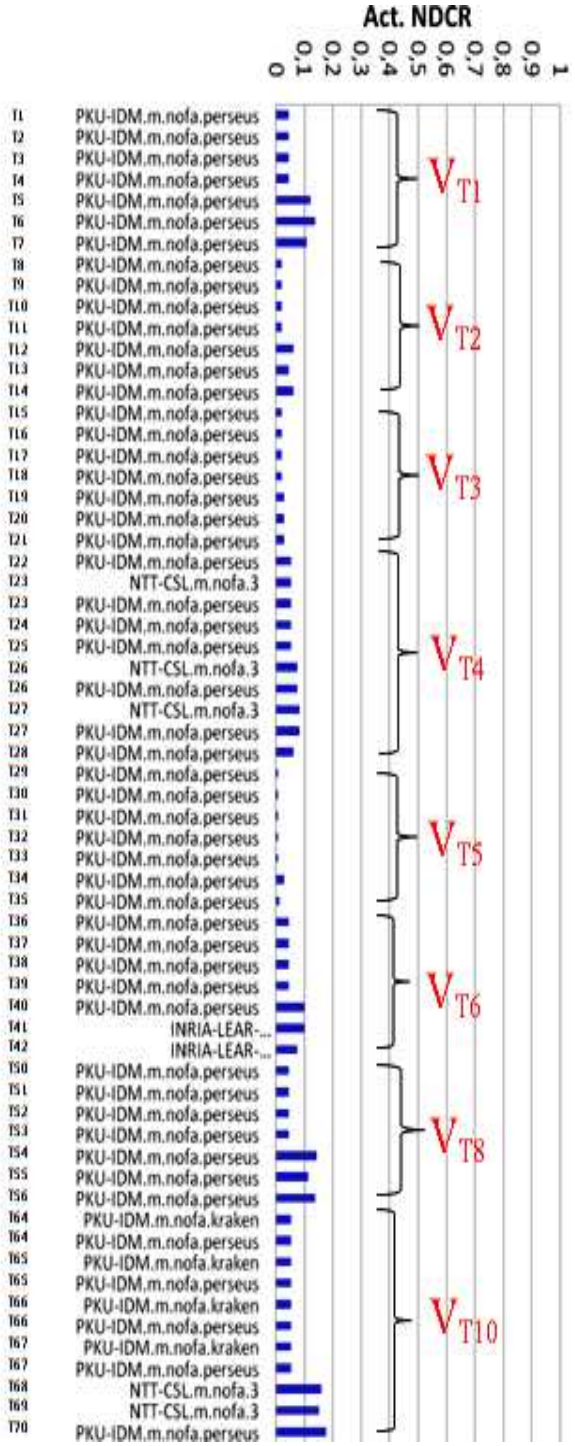


Figure 40: Top runs based on Optimal DET score in Nofa profile

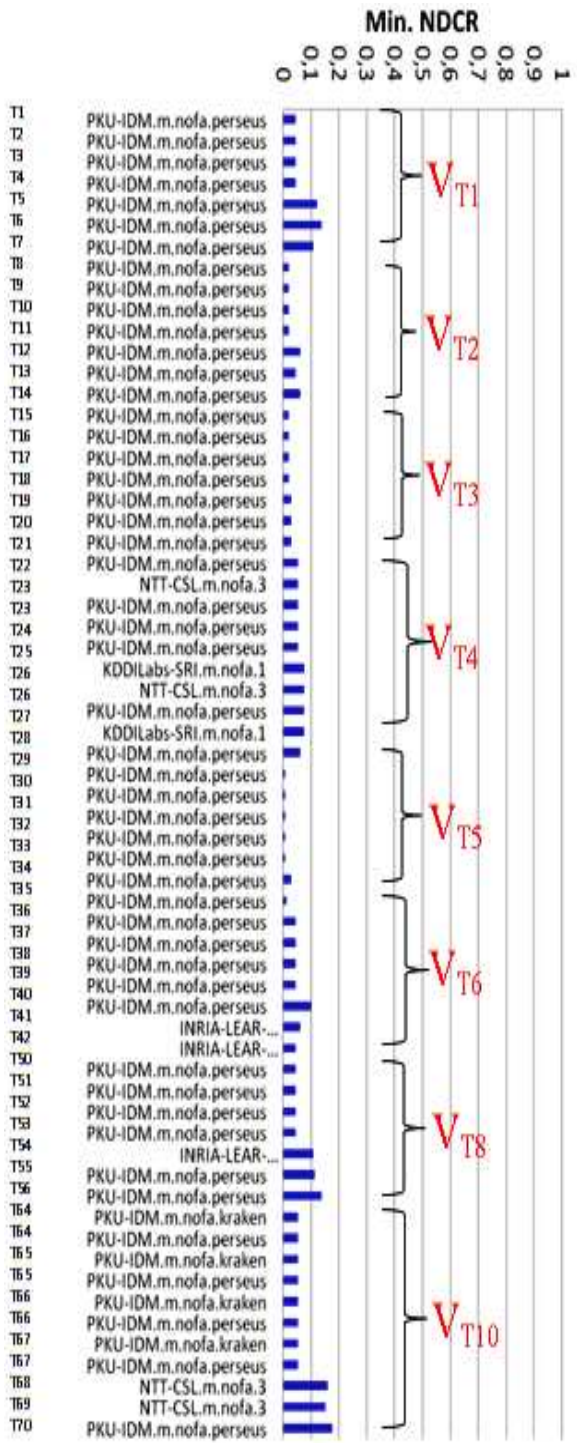


Figure 41: Top 10 runs DET score in balanced profile

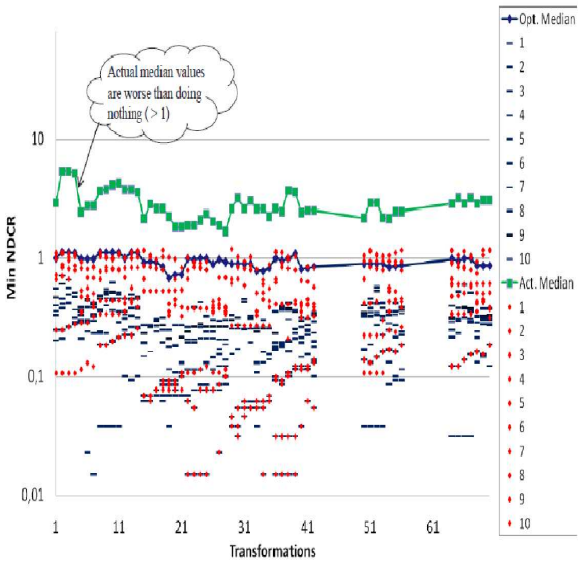


Figure 42: Top 10 runs DET score in Nofa profile

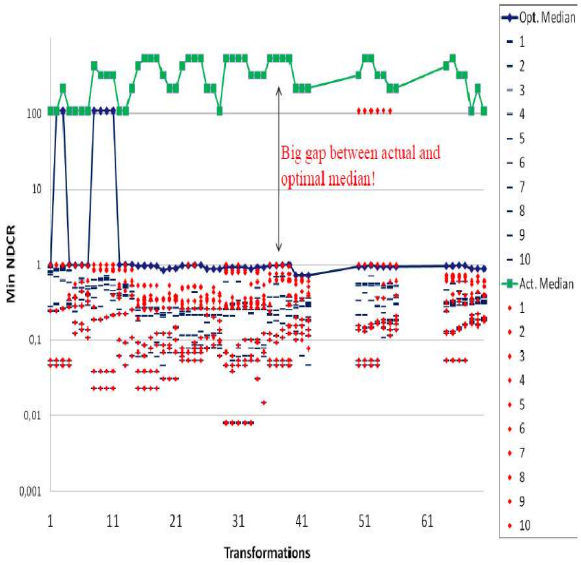


Figure 43: Top 10 runs localization in balanced profile

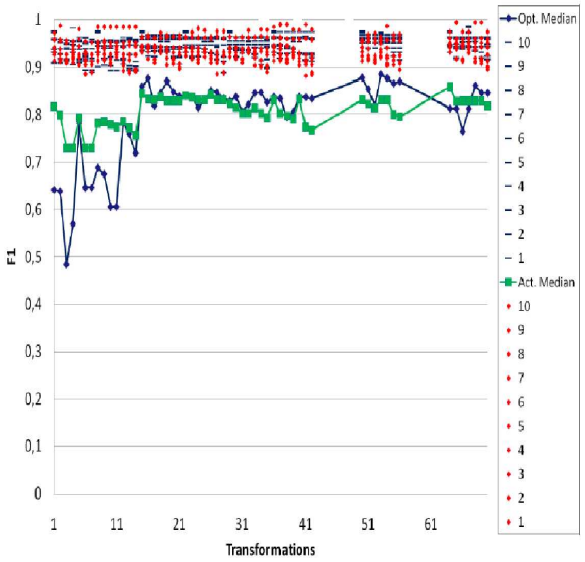


Figure 44: Top 10 runs localization in Nofa profile

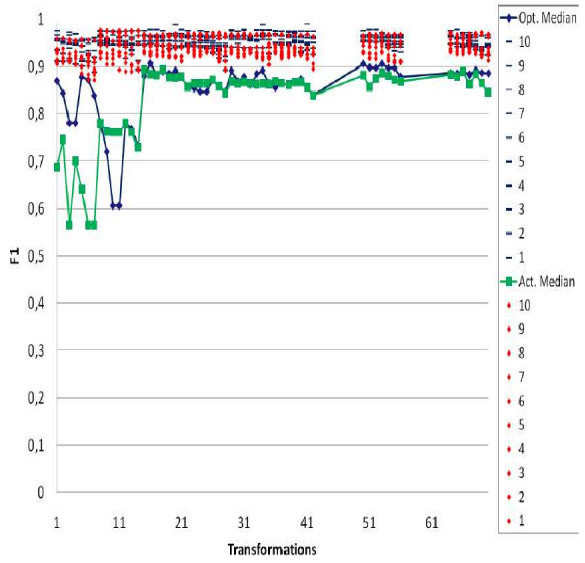
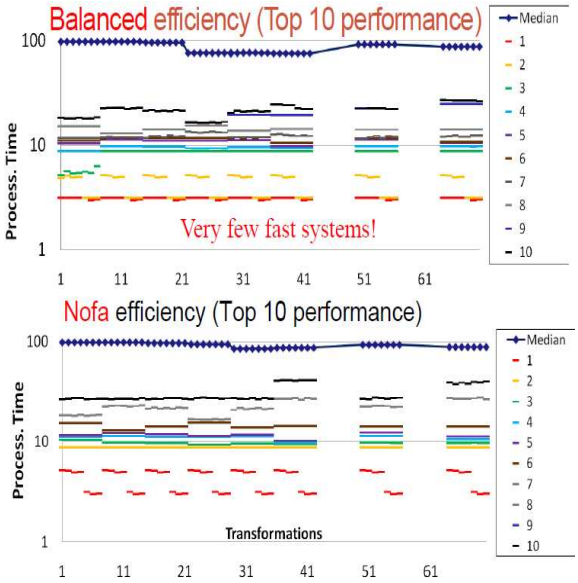


Figure 45: Top 10 runs efficiency in both profiles



and change of gamma) are simpler than the others with combined transformations.

A comparison among the detection of the top 10 runs only for both profiles is shown in Figures 41 and 42. The gap between the actual median line and optimal median line shows that there is still space for more system improvements. This gap in the Nofa profile was much bigger compared to the balanced profile. Although the top 10 runs achieved better results than in TRECVID 2009, the actual and optimal medians were worse than in 2009. A similar comparison based on the localization is shown in Figures 43 and 44. The top 10 runs performed almost equally (and very good) in localization and better compared to 2009 results. We notice that the optimal median was better than the actual median for most of the transformations except few ones. In terms of efficiency, Figure 45 shows the top 10 processing time performance. Even though the best runs could detect copies in seconds, the majority of other systems were still far from real-time detection.

The audio transformations 5, 6, and 7 are harder than the other transformations as they include mixing with external speech. This effect is obvious in Figure 46 which compares only the best runs in both profiles based on actual and optimal values. The red circles on the graph show the video transformations that were mixed with audio transformations 5, 6, and

Figure 46: Comparing best runs

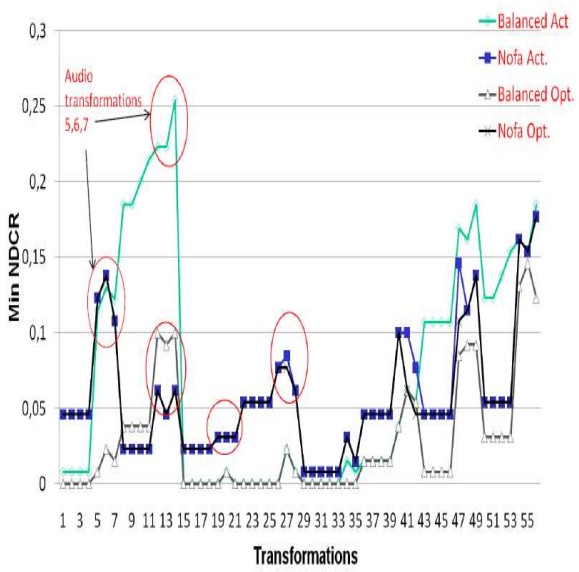


Figure 47: Localization vs Processing time

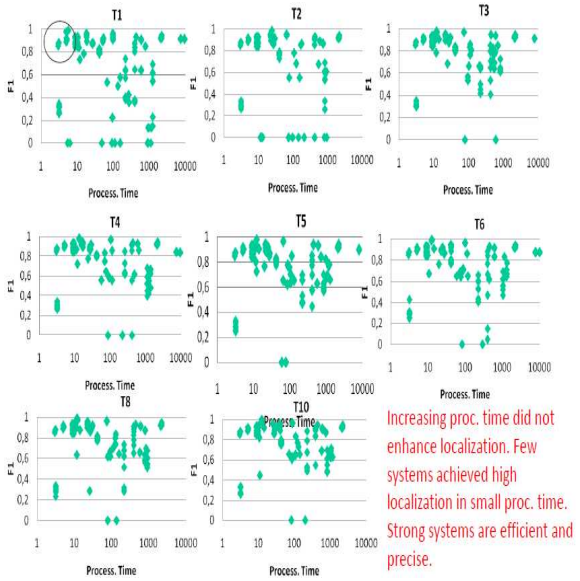


Figure 48: Detection vs Processing time

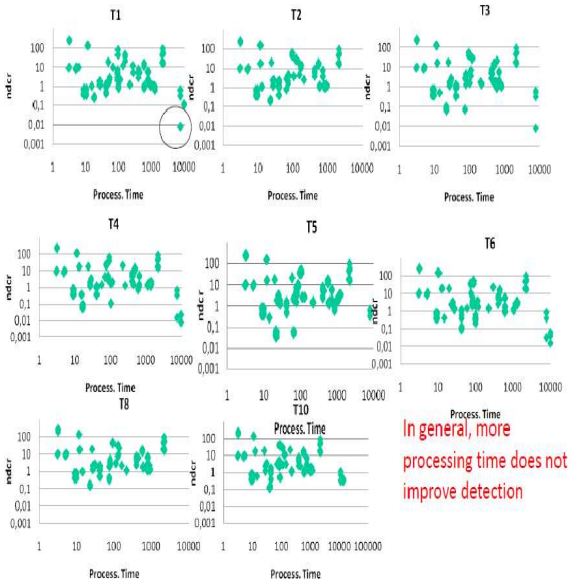
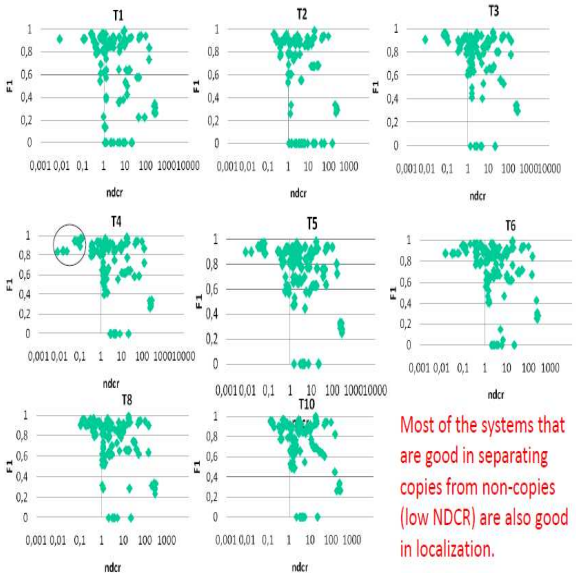


Figure 49: Detection vs Localization



7. In order to study the relation between the three main measures we plotted each two for all transformations in Figures 47, 48, and 49. In general, few systems achieved high localization in short processing time, and most systems which increased the processing time did not gain much in both localization and detection. On the other hand, systems that were good in detection were also good in localization. These observations are true for both application profiles.

Finally, we can draw some general observations from this year's task: Some systems (including first-timers) achieved very good results, while the task was difficult for many others. There was substantial room for improvement available for the balanced profile indicated by difference between actual and optimal results and difference across top runs. Determining the optimal threshold was still a major hurdle. Some systems achieved better NDCR scores compared to 2009. However the median values were higher as the 2010 dataset is very different. Most of the systems were still far from real-time detection while good detecting systems were also good in localization. Complex transformations (audio or video) were indeed more difficult. Camcording was a difficult transformation for some systems. Some submissions were using only the video modality (e.g. IBM, Nanjing University, National Taiwan Normal University, Univ. of Chile, City University of Hong Kong) while audio modality helped to reduce the false alarm rate for picture-in-picture video transformations. Most teams fused audio and video at the decision level. Queries with short copied segments tend to be missed. In regard to the used techniques, the most popular features used were SIFT, speeded up robust features (SURF), direction-adaptive residual transforms (DART), color, texture, and edge histograms for video and MFCC and weighted advanced stability feature (WASF) for audio features. Bag of visual words based techniques are the most popular approaches reported. Readers should see the notebook papers posted on the TRECVID website (`trecvid.nist.gov`) for details about each participant's experiments and results.

8 Surveillance event detection

The 2010 Surveillance Event Detection (SED) evaluation was the third evaluation focused on event detection in the surveillance video domain. The first such evaluation was conducted as part of the 2008 TRECVID conference series (Over et al., 2008; Rose, Fiscus, Over, Garofolo, & Michel, 2009) followed the next year as part of the 2009 TRECVID (Over et al., 2009). The goal of the evaluation track was to support the development of technologies to detect visual events (people engaged in particular activities) in a large collection of video data. It was designed to move computer vision technology towards robustness and scalability while increasing core competency in detecting human activities within video. The approach used was to employ real surveillance data, orders of magnitude larger than previous computer vision tests, and multiple, synchronized camera views.

Figure 50: TRECVID 2010 SED Participants Chart

2010 SED Participants (11 Sites)		Single Person		Single Person + Object		Multiple People		
		PersonRuns	Pointing	CellToEar	ObjectPut	Embrace	PeopleMeet	PeopleSplitUp
3 years in a row	Carnegie Mellon University [CMU]	11	11	11	11	11	11	11
	NHK Science and Technical Research Laboratories [NHKSTRL]		1	1	1			
2 years in a row	Beijing University of Posts and Telecommunications, MCPRL [BUPT-MCPRL]	2	2		2	2		
	Peking University, IDM [PKU-IDM]	4				4	4	4
	Simon Fraser University [SFU]	1						
	Tokyo Institute of Technology and Georgia Institute of Technology [TTandGT]	1					1	1
new	Centre de Recherche Informatique de Montréal [CRIM]	1	1		1			
	Institut National de Recherche en Informatique et en Automatique, WILLOW [INRIA-WILLOW]	6	6	6	6	6	6	6
	Intelligent Perception Group of Beijing JiaoTong University [IPG-BJTU]		1	1	1	1		
	Queen Mary University of London [QMUL-ACTIVA]	1						
	Tianjin University [TJU]	8	8	8	8	8	8	8
	Total Participants per Event	9	7	5	7	6	5	5

The 2009 evaluation supported the same two evaluation tasks as the 2008 evaluation, retrospective event detection and freestyle analysis, and the same set of ten events were used. In the 2010 evaluation, the list of events was reduced to seven, as three events were removed from the list of evaluated upon events: ElevatorNoEntry, OpposingFlow, and TakePicture. While freestyle analysis was supported in 2010, no site participated in the task.

Retrospective event detection 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). Furthermore, systems could perform multiple passes over the video prior to outputting a list of putative events observations (i.e. the task was retrospective).

In 2010, eleven teams participated in the retrospective task. Figure 50 presents the list of participants and the number of experiments they provided for each

event.

The 2010 evaluation tasks used the same data that was distributed in 2009, used in the following ways:

1. The 2008 Event Detection development and evaluation data sets were both designated as development resources thus expanding the development material to 100 camera-hours.

2. The 2009 evaluation data set was reused for the 2010 evaluation.

8.1 Event Annotation

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. Annotation guidelines were developed to express the requirements for each event. To determine if the observed action is a tagable 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 tagable 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). Videos selected for the evaluation were annotated using the Video Performance Evaluation Resource (ViPER) tool by the Linguistic Data Consortium (LDC). Events were represented in ViPER format using an annotation schema that specified each event observation's time interval.

8.2 Data

The development data consisted of the full 100 hours data set used for the 2008 Event Detection evaluation. The video for the evaluation corpus came from the 45-hour Home Office Scientific Development Branch (HOSDB)'s Image Library for Intelligent Detection Systems (iLIDS) Multi Camera Tracking Training (MCTTR) data set. The evaluation systems processed the full data set however systems were scored on a four-day subset of recordings consisting of approximately fifteen-hours of video data. Both data sets were collected in the same busy airport environment with the same video cameras. The entire video corpus was distributed as MPEG-2 in de-interlaced, Phase Alternating Line (PAL) format (resolution 720 x 576), 25 frames/s, either via hard drive or Internet download.

8.3 Evaluation

Sites submitted system outputs for the detection of possible events. Outputs included the temporal extent as well as a decision score (indicating the strength of evidence supporting the observation's existence) and detection decision (yes/no) for each event observation. Developers were advised to target a low miss, high false alarm scenario via the scoring metrics in order to maximize the number of event observations. A dry run was carried out for one day of collection from the development data in order to test the system's ability to generate compliant system outputs capable of being scored by the evaluation infrastructure.

8.4 Measures of Performance

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 as described in the evaluation plan (Fiscus & Michel, 2010).

NDCR is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per unit time).

$NDCR = P_{miss} + \beta \times R_{FA}$

where:

$P_{Miss} = N_{misses}/N_{Ref}$

$R_{FA} = N_{falseAlarms}/N_{CamHrs}$

$\beta = \frac{Cost_{FA}}{Cost_{Miss} \times R_{Target}}$

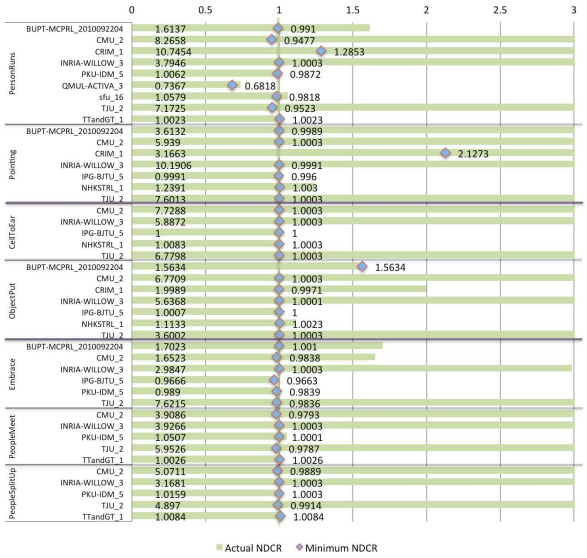
here: $C_{Miss} = 10$, $C_{FA} = 1$ and $R_{Target} = 20/hour$, therefore $\beta = 0.05$.

For this task NDCR is normalized to have the range of $[0, +\infty)$ where 0 would be for perfect performance, 1 would be the cost of a system that provides no output, and $+\infty$ is possible because false alarms are included in the measure.

The inclusion of decision scores in the system output permits the computation of Decision Error Tradeoff (DET) curves. DET curves plot P_{miss} vs. R_{FA} for all thresholds applied to the system's decision scores. These plots graphically show the tradeoff between the two error types for the system.

8.5 Results

Figure 51: TRECVID 2010 SED - NDCR Per Event, Best Run Per Site



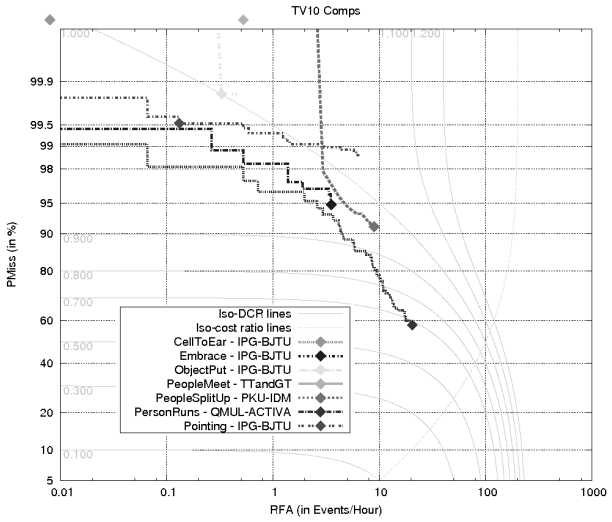
The NDCRs for the submitted event runs can be found in Figure 51 and contains two NDCR values for each submission: the Actual NDCR which is the NDCR based on the binary decisions produced by the system and the Minimum NDCR which is the lowest NDCR possible based on the decision scores produced by the system. The difference between the Actual and Minimum NDCRs indicates how well the system-identified decision score threshold (via the binary decisions) was tuned to the NDCR function.

Figure 52 contains a single DET curve for each event. The curve presents the primary metric selection of the best system per event, selected using the lowest Minimum NDCR.

Figure 53 contains a single DET curve for each event. Also present on the Figure is the iso-cost line for the β value which is computed using R_{Target}, $Cost_{Miss}$, and $Cost_{FA}$ and represents the characteristics of a theoretical application. Developers were asked to tune their systems to that application. In this figure, the best system per event was selected as the system whose curve crosses the iso-cost line and is the lowest on this cost line.

Figure 54 presents the improvements in the Minimum NDCR values between 2009 and 2010 results. In almost all cases, participants did improve results from 2009 to 2010 on the same set of data. This result is the result of an improvement of participants normalizing their system not only per event but also per camera to obtain a better match rate.

Figure 52: TRECVID 2010 SED - Best System Per Event, selected using the minimum Actual NDCR



9 Summing up and moving on

This introduction to TRECVID 2010 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 publications webpage: `www-nlpir.nist.gov/projects/tv2010/tv2010.html`.

10 Authors' note

TRECVID would not have happened in 2010 without support from the National Institute of Standards and Technology (NIST), the Intelligence Advanced Research Projects Activity (IARPA), and the Department of Homeland Security (DHS). The research community is very grateful for this. Beyond that, various individuals and groups deserve special thanks:

- Brewster Kahle (Internet Archive's founder) and R. Manmatha (U. Mass, Amherst) suggested in December of 2008 that TRECVID take another look at the resources of the Archive.

Figure 53: TRECVID 2010 SED - Best System Per Event, selected using the iso-cost line

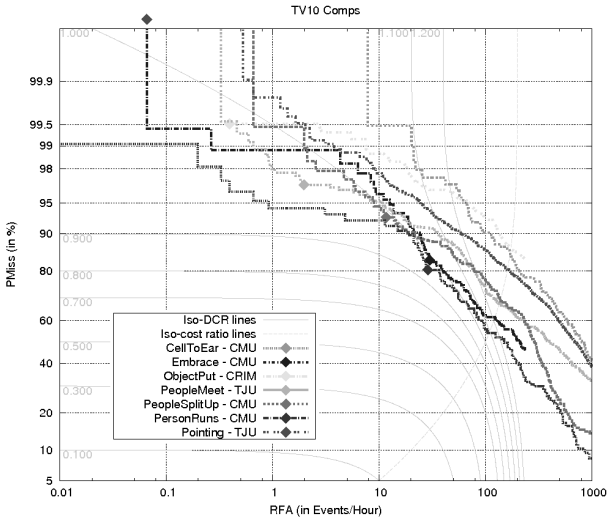
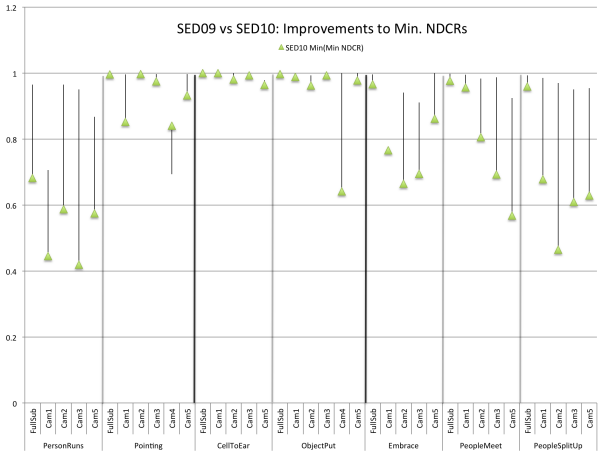


Figure 54: TRECVID 2010 SED - Improvements to MinNDCR between 2009 vs 2010 systems



- Cara Binder and Raj Kumar helped explain how to query and download automatically from the Internet Archive.

- Shin'ichi Satoh at Nantional Institute of Informatics along with Alan Smeaton at Dublin City University and Brian Boyle at HEANet arranged for the mirroring of the video data.

- Georges Quénot with Franck Thollard, Andy Tseng, Bahjat Safadi from LIG and Stéphane Ayache from the Laboratoire D'informatique Fondamentale (LIF) shared coordination of the semantic indexing task, organized the community annotation of 130 features, and provided judgments for 10 features under the Quaero program.

- Georges Quénot provided the master shot reference for the IACC.1 videos.

- At Dublin City University Colum Foley and Kevin McGuinness helped segment the instance search topic examples and built the oracle for interactive instance search.

- The Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI) - Spoken Language Processing Group and VexSys Research provided ASR for the IACC.1 videos.

- Laurent Joyeux (INRIA-Roquencourt) updated the copy detection query generation code.

- Matthijs Douze from INRIA-LEAR volunteered a camcorder simulator to automate the camcording transformation for the copy detection task.

- Emine Yilmaz (Microsoft Research) and Evangelos Kanoulas (U. Sheffield) updated their xinfAP code (sample_eval.pl) to estimate additional values and made it available.

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

11 Appendix A: Instance search topics

9001 PERSON - George W. Bush

9002 PERSON - George H. W. Bush

9003 PERSON - J. P. Balkenende

9004 PERSON - Bart Bosch

9005 CHARACTER - Professor Fetze Alsvanouds from the University of Harderwijk (Aart Staartjes)

9006 PERSON - Prince Bernhard

9007 CHARACTER - The Cook (Alberdinck Thijn: Gijs de Lange)

9008 PERSON - Jeroen Kramer

9009 CHARACTER - Two old ladies, Ta en To

9010 CHARACTER - one of two officeworkers (Kwelder of Benema en Kwelder: Harry van Rijthoven)

9011 PERSON - Colin Powell

9012 PERSON - Midas Dekkers

9013 OBJECT - IKEA logo on clothing

9014 CHARACTER - Boy Zonderman (actor in leopard tights and mesh top: Frank Groothof)

9015 OBJECT - black robes with white bibs worn by Dutch judges and lawyers

9016 OBJECT - zebra stripes on pedestrian crossing

9017 OBJECT - KLM Logo

9018 LOCATION - interior of the Dutch parliament

9019 OBJECT - Kappa logo

9020 OBJECT - Umbro logo

9021 OBJECT - tank

9022 OBJECT - Willem Wever van

12 Appendix B: Concepts/Features

The features labeled with an asterisk comprise the "lite" subset.

004 *Airplane Flying - An airplane flying in the sky

006 Animal - Shots depicting an animal (no humans)

007 Asian People - People of Asian ethnicity

013 Bicycling - A person riding a bicycle

015 *Boat-Ship - Shots of a boat or ship

019 *Bus - Shots of a bus

022 Car Racing - Shot of scenes at car races

027 Cheering - One or more people cheering or applauding

028 *Cityscape - View of a large urban setting, showing skylines and building tops. (Not just street-level views of urban life)

029 *Classroom - Images of school or university style classroom scenes

038 Dancing - one or more, not necessarily with eachother

039 Dark-skinned People - People who are dark skinned due to African or African/American descent (ethnicity)

041 *Demonstration or Protest - One or more people protesting. May or may not have banners or signs

044 Doorway - An opening you can walk through into a room or building

049 Explosion Fire - Shots of an explosion or a fire

052 Female-Human-Face-Closeup - Closeup of a female human's face (face must clearly fill more than 1/2 of height or width of a frame but can be from any angle and need not be completely visible)

053 Flowers - Pictures of flowers

058 Ground vehicles - Vehicles refers to ground vehicles, which includes any of the following: Agricultural vehicle (tractor,combine), Armored vehicle, Automobile, Bicycle, Bus, Construction vehicle, Emergency vehicle, Limousine, Livestock carrier, Motor Scooter, Motorcycle, Truck, Box truck, Pickup truck, RV, bulldozer, quads. Excludes interior of cars, vehicles badly destroyed

059 *Hand - A close-up view of one or more human hands, where the hand is the primary focus of the shot

081 Mountain - Shots depicting a mountain or mountain range with the slopes visible

084 *Nighttime - Shots that take place (outdoors) at night. Included is the continuation of story if ambiguous (if a story starts at night, it probably ends at night). Excluded are sports events under lights

086 Old People - Seniors or elderly people

100 Running - One or more people running

105 *Singing - One or more people singing

107 Sitting Down - Person in the act of sitting down

115 Swimming - One or more people swimming

117 *Telephones - All kinds of phones. If only the headset is visible, it was not included

120 Throwing - A person throwing some object

126 Vehicle - Any thing used for transporting people or goods, such as a car, bus, truck, cart, plane, etc.

127 Walking - One or more people walking

References

ACC. (2010). http://www.iso.org/iso/catalogue_detail.htm?csnumber=42739.

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.

Fiscus, J., & Michel, M. (2010). *2010 TRECVID Event Detection Evaluation Plan.* http://www.itl.nist.gov/iad/mig//tests/trecvid/2010/doc/EventDet10-EvalPlan-v01.htm.

H.264. (2010). http://www.itu.int/rec/T-REC-H.264-201003-I/en.

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

MPEG-4. (2010). http://mpeg.chiariglione.org/.

Over, P., Awad, G., Fiscus, J., Michel, M., Kraaij, W., & Smeaton, A. a. (2009). *TRECVID 2009 – Goals, Tasks, Data, Evaluation Mechanisms and Metrics.* http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/tv9overview.pdf.

Over, P., Awad, G., Fiscus, J., Rose, R., Kraaij, W., & Smeaton, A. a. (2008). *TRECVID 2008 – Goals, Tasks, Data, Evaluation Mechanisms and Metrics.* http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/tv8overview.pdf.

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

QUAERO. (2010). *QUAERO homepage.* www.quaero.org/modules/movie/scenes/home/.

Rose, T., Fiscus, J., Over, P., Garofolo, J., & Michel, M. (2009). The TRECVid 2008 Event Detection Evaluation. In *Proceedings of the workshop on applications of computer vision (wacv)* (pp. 1–8). Snowbird, Utah, USA.

UKHO-CPNI. (2007 (accessed June 30, 2009)). *Imagery library for intelligent detection systems.* http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids/.

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.

Yilmaz, E., Kanoulas, E., & Aslam, J. A. (2008). A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 603–610). New York, NY, USA: ACM.

Table 1: Participants and tasks

Task						Location	Participants
– – –	* * *	KIS	* * *	– – –	SIN	Europe	Aalto University School of Science & Technology
– – –	– – –	– – –	– – –	– – –	SIN	Europe	Aristotle University of Thessaloniki
CCD	– – –	– – –	– – –	– – –	– – –	Asia	Asahikasei Co.
CCD	INS	* * *	* * *	– – –	* * *	NorthAm	AT&T Labs - Research
– – –	– – –	– – –	* * *	SED	– – –	Asia	Beijing Jiaotong Univ.
CCD	INS	KIS	– – –	SED	SIN	Asia	Beijing Univ. of Posts and Telecom.-MCPRL
CCD	* * *	– – –	* * *	– – –	SIN	Europe	Brno Univ. of Technology
– – –	* * *	KIS	MED	SED	SIN	NorthAm	Carnegie Mellon Univ.
* * *	* * *	KIS	– – –	– – –	* * *	Asia	Chinese Academy of Sciences - MCG
CCD	– – –	KIS	– – –	* * *	SIN	Asia	City Univ. of Hong Kong
– – –	* * *	– – –	MED	– – –	SIN	NorthAm	Columbia Univ.
– – –	– – –	– – –	– – –	SED	– – –	NorthAm	Computer Research Inst. of Montreal
– – –	* * *	– – –	– – –	– – –	SIN	Europe	DFKI-MADM
– – –	INS	KIS	– – –	– – –	– – –	Europe	Dublin City Univ.
– – –	* * *	– – –	* * *	* * *	SIN	Europe	EURECOM
– – –	* * *	– – –	– – –	– – –	SIN	NorthAm	Florida International Univ.
– – –	* * *	– – –	– – –	– – –	SIN	Asia	France Telecom Orange Labs (Beijing)
– – –	– – –	– – –	– – –	– – –	SIN	Asia	Fudan Univ.
* * *	– – –	– – –	– – –	– – –	SIN	Asia	Fuzhou Univ.
* * *	INS	KIS	– – –	– – –	* * *	Europe	Hungarian Academy of Sciences
CCD	* * *	* * *	MED	– – –	* * *	NorthAm	IBM T. J. Watson Research Center
– – –	INS	KIS	MED	– – –	SIN	Europe	Informatics and Telematics Inst.
CCD	* * *	* * *	* * *	* * *	* * *	Europe	INRIA-TEXMEX
– – –	– – –	– – –	* * *	SED	SIN	Europe	INRIA-willow
– – –	* * *	– – –	– – –	– – –	SIN	Europe	IRIT - Equipe SoratoriesAMoVA
– – –	– – –	KIS	– – –	– – –	– – –	Asia	Inst.for Infocomm Research
CCD	– – –	– – –	– – –	– – –	– – –	Europe	Istanbul Technical Univ.
– – –	INS	– – –	– – –	* * *	SIN	Europe	JOANNEUM RESEARCH
– – –	INS	KIS	MED	* * *	SIN	NorthAm	KB Video Retrieval
CCD	– – –	– – –	* * *	* * *	* * *	Asia	KDDI R&D Labs and SRI International
– – –	– – –	– – –	– – –	– – –	SIN	Europe	Lab. d'Informatique Fondamentale de Marseille
– – –	INS	* * *	* * *	– – –	SIN	Europe	Lab. d'Informatique de Grenoble for IRIM
– – –	– – –	– – –	– – –	– – –	SIN	Europe	LSIS / UMR CNRS & USTV
– – –	– – –	– – –	MED	– – –	– – –	NorthAm	Mayachitra, Inc.
CCD	INS	– – –	– – –	* * *	– – –	Asia	Nanjing Univ.
CCD	– – –	– – –	– – –	– – –	– – –	Asia	National Chung Cheng Univ.
CCD	INS	* * *	* * *	* * *	SIN	Asia	National Inst. of Informatics
– – –	* * *	– – –	– – –	– – –	SIN	Asia	National Taiwan Univ.
– – –	* * *	KIS	* * *	* * *	– – –	Asia	National Univ. of Singapore
* * *	* * *	* * *	* * *	SED	SIN	Asia	NHK Science and Technical Research Labs
– – –	– – –	– – –	MED	– – –	– – –	Asia	Nikon Corporation
CCD	– – –	– – –	– – –	– – –	– – –	Asia	NTNU and Academia Sinica
CCD	– – –	– – –	– – –	– – –	– – –	Asia	NTT Communication Science Labs-CSL
– – –	INS	– – –	– – –	– – –	– – –	Asia	NTT Communication Science Labs-NII
– – –	– – –	KIS	– – –	– – –	SIN	Asia	NTT Communication Science Labs-UT
– – –	* * *	* * *	– – –	– – –	SIN	Europe	Oxford/IIIT
CCD	– – –	– – –	– – –	SED	– – –	Asia	Peking Univ.-IDM
– – –	– – –	– – –	* * *	– – –	SIN	Europe	Quaero consortium
– – –	– – –	– – –	– – –	SED	– – –	Europe	Queen Mary, Univ. of London
– – –	– – –	* * *	– – –	– – –	SIN	Asia	Ritsumeikan Univ.
CCD	– – –	* * *	– – –	– – –	* * *	Asia	Shandong Univ.
– – –	– – –	– – –	– – –	– – –	SIN	Asia	SHANGHAI JIAOTONG UNIVERSITY-IS
– – –	– – –	– – –	– – –	SED	– – –	NorthAm	Simon Fraser Univ.
CCD	– – –	– – –	– – –	* * *	* * *	Asia	Sun Yat-sen Univ. - GITL
CCD	– – –	– – –	– – –	– – –	– – –	Europe	Telefonica Research

Task legend. CCD:copy detection; INS:instance search; KIS:known-item search; MED:multimedia event detection; SED: surveillance event detection; SIN:semantic indexing; – – –:no run planned; * * *:planned but not submitted

Table 2: Participants and tasks

Task						Location	Participants
* * *	* * *	* * *	* * *	SED	SIN	Asia	Tianjin Univ.
– – –	INS	– – –	– – –	– – –	– – –	Europe	TNO ICT - Multimedia Technology
* * *	INS	– – –	– – –	– – –	– – –	Asia	Tokushima Univ.
– – –	* * *	– – –	* * *	SED	SIN	Asia/NorthAm	Tokyo & Inst. of Techn.+ Georgia Inst. of Techn.
CCD	* * *	* * *	* * *	* * *	* * *	Asia	Tsinghua Univ.-IMG
CCD	* * *	– – –	– – –	* * *	SIN	Europe	TUBITAK - Space Technologies Research Inst.
– – –	– – –	– – –	– – –	– – –	SIN	Europe	Universidad Carlos III de Madrid
– – –	INS	KIS	* * *	* * *	SIN	Europe	Univ. of Amsterdam
CCD	– – –	– – –	– – –	– – –	– – –	Europe	Univ. of Brescia
CCD	– – –	– – –	– – –	– – –	– – –	SouthAm	Univ. of Chile
– – –	* * *	* * *	* * *	* * *	SIN	Asia	Univ. of Electro-Communications
– – –	– – –	– – –	* * *	* * *	SIN	NorthAm	Univ. of Illinois at U-C & NEC Labs America
* * *	* * *	KIS	– – –	– – –	– – –	Europe	Univ. of Klagenfurt
* * *	* * *	– – –	* * *	– – –	SIN	Europe	Univ. of Marburg
* * *	* * *	* * *	– – –	* * *	SIN	Africa	Univ. of Sfax
– – –	– – –	* * *	– – –	* * *	SIN	Asia	Waseda Univ.
* * *	INS	* * *	* * *	* * *	* * *	Asia	Xi'an Jiaotong Univ.
* * *	– – –	KIS	* * *	* * *	* * *	NorthAm	York Univ.

Task legend. CCD:copy detection; INS:instance search; KIS:known-item search; MED:multimedia event detection; SED: surveillance event detection; SIN:semantic indexing; – – –:no run planned; * * *:planned but not submitted

Table 3: Instance 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
9001	36874	18349	49.8	150	3514	19.2	61	1.7
9002	36779	18153	49.4	100	2514	13.8	28	1.1
9003	36828	18800	51.0	150	3730	19.8	140	3.8
9004	35507	17745	50.0	160	3897	22.0	140	3.6
9005	36836	19423	52.7	100	2706	13.9	36	1.3
9006	37578	18904	50.3	120	3066	16.2	27	0.9
9007	38722	19965	51.6	100	2712	13.6	14	0.5
9008	36850	19544	53.0	260	6397	32.7	135	2.1
9009	37515	18797	50.1	130	3381	18.0	75	2.2
9010	36457	19531	53.6	120	3165	16.2	68	2.1
9011	36256	16508	45.5	100	2333	14.1	4	0.2
9012	36476	19287	52.9	170	4376	22.7	174	4.0
9013	37230	18276	49.1	100	2367	13.0	25	1.1
9014	38082	17974	47.2	100	2583	14.4	15	0.6
9015	38126	18173	47.7	140	3286	18.1	80	2.4
9016	35811	16873	47.1	160	3412	20.2	27	0.8
9017	35540	16249	45.7	110	2335	14.4	20	0.9
9018	37201	16996	45.7	120	2795	16.4	52	1.9
9019	32932	15374	46.7	100	1925	12.5	6	0.3
9020	36035	18134	50.3	100	2254	12.4	38	1.7
9021	35162	16496	46.9	170	3622	22.0	28	0.8
9022	38585	18207	47.2	100	2400	13.2	15	0.6

Table 4: 2010 Participants not submitting any runs

Task						Location	Participants
− − −	− − −	− − −	− − −	− − −	∗ ∗ ∗	NorthAm	AKiiRA Media Systems Inc.
− − −	− − −	− − −	∗ ∗ ∗	∗ ∗ ∗	− − −	Asia	Beijing University of Post-Telecom.-MCU
∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	NorthAm	Binatix Inc.
− − −	− − −	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	− − −	Asia	Chinese Academy of Sciences - AIBT
∗ ∗ ∗	− − −	− − −	− − −	− − −	− − −	NorthAm	CMART Systems
− − −	∗ ∗ ∗	∗ ∗ ∗	− − −	− − −	∗ ∗ ∗	Europe	Commissariat á l'Energie Atomique, LIST
− − −	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	NorthAm	CompuSensor Technology Corporation Address
− − −	− − −	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	Europe	Consorzio Milano Ricerche
∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	− − −	NorthAm	Florida Atlantic University
− − −	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	− − −	− − −	NorthAm	Harvard
∗ ∗ ∗	− − −	− − −	− − −	∗ ∗ ∗	− − −	Asia	Hong Kong Polytechnic University
− − −	∗ ∗ ∗	− − −	− − −	∗ ∗ ∗	− − −	Europe	Imagelab - University of Modena and Reggio Emilia
− − −	∗ ∗ ∗	− − −	− − −	∗ ∗ ∗	∗ ∗ ∗	NorthAm	ITT Geospatial Systems
− − −	∗ ∗ ∗	∗ ∗ ∗	− − −	− − −	∗ ∗ ∗	Asia	Kobe University
∗ ∗ ∗	− − −	− − −	− − −	− − −	− − −	Asia	Korea Advanced Institute of Science and Technology
− − −	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	Asia	Mangalore University
− − −	− − −	− − −	− − −	− − −	∗ ∗ ∗	Asia	National Cheng Kung University
− − −	∗ ∗ ∗	∗ ∗ ∗	− − −	− − −	∗ ∗ ∗	Europe	Open University
∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	Asia	Peking University-ICST
− − −	− − −	∗ ∗ ∗	∗ ∗ ∗	− − −	− − −	Europe	Politecnico Di Milano
− − −	∗ ∗ ∗	− − −	− − −	− − −	− − −	Austrail	RMIT University School of CS&IT
∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	Asia	Shanghai Jiao Tong Univresity-IICIP
− − −	− − −	∗ ∗ ∗	− − −	− − −	− − −	Asia	Sichuan University of China
∗ ∗ ∗	− − −	− − −	− − −	− − −	− − −	Asia	Sun Yat-sen University - IST
− − −	∗ ∗ ∗	− − −	− − −	− − −	− − −	Asia	Tsinghua University-THEEIE
− − −	− − −	− − −	− − −	∗ ∗ ∗	∗ ∗ ∗	Europe	Universidad Autnoma de Madrid
− − −	− − −	− − −	− − −	∗ ∗ ∗	− − −	SouthAm	Universidade Federal do Paran
− − −	− − −	∗ ∗ ∗	− − −	− − −	− − −	NorthAm	University of South Carolina
∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	Europe	University of Ioannina
− − −	− − −	∗ ∗ ∗	− − −	− − −	− − −	Asia	University of Malaya
− − −	− − −	∗ ∗ ∗	− − −	− − −	∗ ∗ ∗	NorthAm	University of Maryland
∗ ∗ ∗	∗ ∗ ∗	− − −	∗ ∗ ∗	− − −	− − −	NorthAm	University of Ottawa
∗ ∗ ∗	∗ ∗ ∗	∗ ∗ ∗	− − −	− − −	∗ ∗ ∗	SouthAm	University of Sao Paulo
− − −	− − −	− − −	∗ ∗ ∗	− − −	∗ ∗ ∗	Europe	University of Sheffield
∗ ∗ ∗	− − −	− − −	− − −	− − −	∗ ∗ ∗	Europe	University of Surrey
∗ ∗ ∗	∗ ∗ ∗	− − −	∗ ∗ ∗	− − −	∗ ∗ ∗	Asia	Zhejiang University

Task legend. CCD:copy detection; INS:instance search; KIS:known-item search; MED:multimedia event detection; SED: surveillance event detection; SIN:semantic indexing; − − −:no run planned; ∗ ∗ ∗:planned but not submitted