TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics

Paul Over {over@nist.gov}
Jon Fiscus {jfiscus@nist.gov}
Greg Sanders {gregory.sanders@nist.gov}
Barbara Shaw {barbara.shaw@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}

George Awad
Dakota Consulting, Inc.
1110 Bonifant Street, Suite 310
Silver Spring, MD 20910
{gawad@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
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 1, 2013

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2012 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 ten 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 NIST and other US government agencies. Many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2012 represented a continuation of five tasks from 2011 with some variations and significantly new data, along with the initiation of the multimedia event recounting (MER) task. 57 teams (see Tables 1 and 2) from various research organizations — 21 from Europe, 19 from Asia, 15 from North America, 1 from South America, and 1 from Australia — completed one or more of six tasks:

1. Semantic indexing (SIN)
2. Known-item search (KIS)
3. Instance search (INS)
4. Multimedia event detection (MED)
5. Multimedia event recounting (MER)
6. Surveillance event detection (SER)

291 h of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC), were used for semantic indexing and known-item search. 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 91 h of Flickr video was used for the instance search pilot. 45 h of airport surveillance video (iLIDS) was reused for the surveillance event detection task. Almost 4000 h from a new collection of Internet videos – the Heterogeneous Audio Visual Internet Corpus (HAVIC) – was used for development and testing in the multimedia event detection task.

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 by NIST using ground truth created manually by the Linguistic Data Consortium under contract to NIST. The multimedia event recounting task was judged by humans experts in an evaluation designed by NIST.

This paper is an overview of 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 online workshop notebook (TV12Notebook, 2012).

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

Flickr video

Robin Aly at the University of Twente worked in consultation with NIST to create several classes of queries and run them against Flickr video available under a Creative Commons license for research. The videos were then divided into segments of about 10s in duration. A set of 91 videos divided into 74 958 files was chosen independently by NIST, where 21 test topics appropriate for the test videos were created. Each topic contained a very short textual description and example images from Flickr videos not included in the test set

Internet Archive Creative Commons (IACC) video

For 2012, approximately 291 additional hours of Internet Archive videos with Creative Commons licenses in MPEG-4/H.264 and with durations between 10 s and 3.5 min were used as new test data. This dataset is called IACC.1.C. Most videos had some donor-supplied metadata available e.g., title, keywords, and description. 200 h of 2010 IACC test data (IACC.1.A), 200 h of 2010 IACC training data (IACC.1.training), and 287 h of 2011 IACC test data (IACC.1.B) were available for system development.

As in 2010 and 2011, LIMSI and VecSys research (Gauvain, Lamel, & Adda, 2002) 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 346 features against the IACC videos, using an active learning scheme designed to improve the efficiency of the process (Ayache & Quénot, 2008).

iLIDS Multiple Camera Tracking Data

The iLIDS Multiple Camera Tracking data consisted of ≈ 150 h of indoor airport surveillance video collected in a busy airport environment by the United Kingdom (UK) Center for Applied Science and Technology (CAST). The dataset utilized 5, frame-synchronized cameras.

The training video consisted of the ≈100 h of data used for SED 2008 evaluation. The evaluation video consisted of the same additional ≈50 h of data from

Imagery Library for Intelligent Detection System's (iLIDS) multiple camera tracking scenario data used for the 2009, 2010, and 2011 evaluations(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 2011 evaluation.

Heterogeneous Audio Visual Internet Corpus (HAVIC)

The HAVIC Corpus is a large corpus of Internet multimedia files collected by the Linguistic Data Consortium and distributed as MPEG-4 (MPEG-4, 2010) formatted files containing H.264 (H.264, 2010) encoded video and MPEG-4s Advanced Audio Coding (ACC) (ACC, 2010) encoded audio.

The training material consisted of: the 1429 h of HAVIC material (114 h used for the MED 2010 pilot evaluation and 1315 additional hours of data from MED '11) and 30 events: 15 MED '11 events, 10 new events for the Pre-Specified event detection task, and 5 new events for the pilot Ad-Hoc event detection task.

The evaluation corpus was the 3722 h MED Progress Collection which is 3.1 times larger than the MED '11 test collection.

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 terms "features" and "concepts" are used interchangeably through the rest of this document. The ability to detect features is an interesting challenge by itself but takes on added importance 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 2000 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 standard video 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 affords in pooling of results and approximating the basis for calculating recall.

346 concepts had been selected for the TRECVID 2011 semantic indexing task, including 130 concepts tested in 2010. The 346 concepts are those for which there exist at least 4 positive samples in the final community annotation. The goal is to promote research on methods for indexing many concepts and using ontology relations between them. Also it is expected that these concepts will be useful for the content-based (known item) search task. Including TRECVID 2005 to 2010 features favors the reuse of already available annotations and judgments and encourages cross-domain evaluations.

The task remained the same as in 2010 and 2011 but, considering the feedback from a poll about the 2011 issues of the task, it was decided not to increase the number of concepts to be processed. Slight adjustments (clarifications) were made to the concept definitions but the counts remained comparable.

Also considering some feedback from the poll that pointed to a lack of novelty and considering suggestions in this direction, two novelties were introduced as pilot extensions to the participants in 2012:

- A "paired concept" version of the task was added to the light and full ones. It consisted in the detection of pairs of unrelated concepts instead of the detection of simple concepts. The idea was to promote the development of methods for retrieving shots containing a combination of concepts that do better than just combining the output of individual concept detectors. Here are the pairs:

  901 Beach+Mountain

902 Old_People+Flags

903 Animal+Snow

904 Bird+Waterscape_Waterfront

905 Dog+Indoor

906 Driver+Female_Human_Face

907 Person+Underwater

908 Table+Telephone

909 Two_People+Vegetation

910 Car+bicycle

- A "no annotation" version of the tasks: the idea was to promote the development of methods that permit the indexing of concepts in video shots using only data from the web or archives without the need of additional annotations. The training data could for instance consist of images retrieved by a general purpose search engine (e.g., Google) using only the concept name and/or definition with only automatic processing of the returned images. This was not implemented as a new variant of the task like the light, full or pair ones but using additional categories ('E' for training data collected automatically using only the concepts' name and definition and 'F' for training data collected automatically using a query built manually from the concepts' name and definition) for the training types besides the A to D ones (see below). By "no annotation", we meant here that no annotation should be manually done on the retrieved samples (either images or videos). Any annotation done by somebody else prior to the general search did not count. Methods developed in this context could be used for building indexing tools for any concept starting only from a name and a definition for it or from a simple query defined for it.

Three types of submissions were considered: full (F) submissions in which participants submit results for all 346 concepts, lite (L)submissions in which participants submitted results for only 50 concepts, and the paired (P) concept submissions. TRECVID evaluated 46 single concepts - 20 based on judgments done at NIST and 26 done under the Quaero program in France - and 10 paired concepts listed above. The single concepts evaluated for 2012 were as follows. Those marked with an asterisk formed a "lite" subset to which some participants restricted their experiments.

[003] Airplane, [004] *Airplane_Flying, [009] Basketball, [013] *Bicycling, [015] *Boat_Ship, [016] Boy, [017] Bridges, [025] Chair, [031] *Computers, [051] *Female_Person, [054] Girl, [056] Government_Leader, [057] Greeting, [063] Highway, [071] *Instrumental_Musician, [072] Kitchen, [074] *Landscape, [075] *Male_Person, [077] Meeting, [080] Motorcycle, [084] *Nighttime, [085] Office, [095] Press_Conference, [099] Roadway_Junction, [101] *Scene_Text, [105] *Singing, [107] *Sitting_Down, [112] *Stadium, [116] Teenagers, [120] *Throwing, [128] *Walking_Running, [155] Apartments, [163] Baby, [198] Civilian_Person, [199] Clearing, [254] Fields, [267] Forest, [274] George_Bush, [276] Glasses, [297] Hill, [321] Lakes, [338] Man_Wearing_A_Suit, [342] Military_Airplane, [359] Oceans, [434] Skier, [440] Soldiers.

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 which really helped in the search task and contributed to the collaborative nature of TRECVID.

The fuller concept definitions provided to system developers and NIST assessors are listed on the webpage: `http://www-nlpir.nist.gov/projects/tv2012/tv11.sin.500.concepts_ann_v2.xls`

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 average precision (Over, Ianeva, Kraaij, & Smeaton, 2006). This year mean extended inferred average precision (mean xinfAP) was used. It permits 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 that contribute more average precision than those ranked lower.

3.1 Data

The IACC.1.C collection was used for testing. It contained 145 634 shots.

3.2 Evaluation

Each group was allowed to submit up to 4 single concept runs and two additional if they are "no annotation" runs. Each group could also submit up to two paired-concept runs. In fact 25 groups submitted a total of 51 full runs, 40 lite runs, and 16 paired-concept runs. Each full run was also treated as a lite run by looking at their performance on just the lite concept subset. The MediaMill team from the University of Amsterdam provided four baseline runs for the paired-concept subtask, using their single-concept F_A_UvA.Sheldon_1 run as the basis:

- SINPair-baseline-firstconcept.xml: this one contains a pair-run based on a ranking of the first concept only.
- SINPair-baseline-secondconcept.xml: this one contains a pair-run based on a ranking of the second concept only.
- SINPair-baseline-combine-sum.xml: this one contains a pair-run based on a sum of the scores of concept 1 and concept 2.
- SINPair-baseline-combine-mul.xml: this one contains a pair-run based on a product of the scores of concept 1 and concept 2.

For each concept, pools were created and randomly sampled as follows. The top pool sampled 100 % of shots ranked 1 to 200 across all submissions. The bottom pool sampled 10 % of those ranked 201 to 2000 and not already included in a pool. Human judges (assessors) were presented with the pools - one assessor per concept - and they judged each shot by watching the associated video and listening to the audio. In all, 282 949 shot-concept combinations were judged. 1 058 743 shots fell into the unjudged part of the overall samples.

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 46 (or 19 lite) evaluated single concepts. The results also provide some information about "within concept" performance although this is less reliable. This year xinfAP was updated to adjust the average precision (AP) score if the inferred number of true positives for a given concept was greater than the maximum result set size (2000) - so that an AP of 1.0 was possible. Lack of this adjustment, incorporated long ago in the earlier evaluation program trec_eval_video, imposed an artificial limit on the best possible AP score for concepts with more than 2000 true positives, a limit which affected, to various degrees, 10 of 30 concepts evaluated in 2010 and 26 of 50 in 2011.

3.4 Results

Performance varied greatly by feature. Figure 1 shows how many unique instances were found for each tested feature. The inferred true positives (TPs) of 13 features exceeded 1 % TPs from the total tested shots percentage. Features "Female-person", "civilian-person", and "male-person" had TPs in over 5 % of the test shots. On the other hand, features that had the fewest TPs (less than 0.5 %) were "Airplane", "Airplane-flying", "Basketball", "Bicycling", "Bridges", "Kitchen", "Motorcycle", "Office", "sitting-down", "stadium", "Throwing", "Baby", "Military-airplane", "Skier", and "Soldier". The top performing features were more generic by definition than the bottom performing ones which are more specific in category, location or action such as "sitting-down", "stadium", and "Baby". In addition, many of the low performing features are easily confusable by another visually similar features such as "Airplane","Airplane-flying", and "Military-Airplane". Figure 2 shows the number of TPs and false positives (FPs) per feature calculated from the assessors judgments. Some observations from this figure include: only two features "male-person" and "civilian-person" achieved higher TPs compared to FPs while "civilian-person" received the lowest FP. In addition, the feature "Bicycling" received the highest FPs and followed by features "Throwing", "Sitting-down", "Stadium" and "Computers". This may indicate that detecting persons in general was an easier task for participants compared to concepts that include objects, locations or actions.

Figures 3 and 4 show the results of category A, and D for full runs. Category A runs used only IACC training data. Category D runs used both IACC and non-IACC non-TRECVID training data. The graphs show the median values in each category together with a new median baseline run generated by NIST. In the baseline run, for each feature the median rank of each submitted shot is calculated across all sub-

mitted runs in that run type and training category. The final shot median rank value is weighted by the ratio of all submitted runs to number of runs that submitted that shot. One baseline run was generated for each run type and training category. The baseline run can be interpreted as a run that includes mainly the shots that most of the systems agreed to submit and filters out shots that didn't get enough votes from system's runs. Still category A runs were the most popular type and achieve top recorded performances. Only 1 run from category F was submitted and achieved a score of 0.048.

Figures 5, 6, and 7 show the results of category A,D, and F 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. Only 1 run from Category E was submitted and achieved a score of 0.044.

Figure 8 shows the performance of the top 10 teams across the 46 features. Few features reflected a large spread between the scores of the top 10 such as feature "Female-person", and "walking-running", while features such as "Instrumental-Musician", "Motorcycle", "Night-time","Singing","Baby","George-Bush","Glasses", and "Man-Wearing-Suit" had medium spread. The spread in scores may indicate that there is still room for further improvement within used techniques. The majority of the rest of the features had a tight spread of scores among the top 10 which may indicate a small variation in used techniques performance. In general, the median scores ranged between 0.003 (feature "Sitting_down") and 0.825 (feature "Civilian-Person") which is much higher than TRECVID 2011 top median score (0.441). As a general observation, feature "Sitting_down" had the minimum median score at TRECVID 2010 and TRECVID 2011 as well which demonstrates how difficult this feature is for the systems to detect.

The analogous graph for the 15 common features is Figure 9, 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 "Walking-Running" and "Scene-Text", While the features with tight spread was "Bicycling", "Landscape", "Male-person", "Sitting-down", "Stadium" and "Throwing".

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 10 for full runs and Figures 11 through 12 for lite runs. The figures indicate 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 the top ranked run was significantly better than other runs. Also we found that for full runs and category D there was no any significant differences among submitted runs as well as among lite runs and category F.

Figure 13 shows the hits performance for the new subtask of concept-pairs. In general none of the 10 pairs achieved high hits except for the pair "Two-people + vegetation". The performance of all runs can be seen in Figure 14. The top run achieved score 0.076 while the median score was 0.041. Four baseline runs were submitted by Mediamill team based on only the first concept occurrence, the second concept occurrence, the sum of both concepts scores and the product of both concept scores. Surprisingly, the first concept baseline run ranked as the third top score. The results of this first year subtask indicates that systems still have much work to do to find ways and visual evidence for the occurrence of both concepts compared to only depending on detecting one of the concepts and assuming the presence of the other. Figure 15 shows the randomization test on concept-pair runs.

A failure analysis experiment was done at NIST to examine the failure cases of different concepts and see if there was a semantic relationship between confused concepts. For each pair of concepts we found the common shots that were judged as TP for one concept and at the same time as FP for the second concept. Sorting those number of common shots per concept gives an indication of whether this concept was highly confused by other concepts. To mention a few examples, we found that 960 shots were TPs as male-person but FP as sitting-down, 848 shots were TP as walking-running but FP as bicycling, 755 shots were TP as male-person but FP as Glasses, 743 shots were TP as male-person but FP as female-person, 588 shots were TP as landscape but FP as beach+mountain, 560 shots were TP as female-person but FP as girl, 538 shots were TP as walking-running but FP as throwing, 475 shots were TP as landscape but FP as hill, 438 shots were TP as male-person but FP as girl, 415 shots were TP as male-person but FP as George-Bush, and 328 shots were TP as oceans but FP as Boat-ship. From this

experiment it seems that there is some semantic relationship between confused concepts such as it is hard to detect the gender of a person but is easy to just detect a person or it is hard to detect the exact scene in a landscape and even some motion features of different concepts can be very similar such as walking-running vs bicycling or an airplane-flying vs a boat-ship in sea water.

Based on site reports, some general observations on approaches can be made. Systems in general focused on robustness, merging many different representations, use of spatial pyramids, improved bag of word approaches, utilizing Fisher/super-vectors, VLADs (Vector of Locally Aggregated Descriptors), VLATs (Vector of Locally Aggregated Tensors), sophisticated fusion strategies, and combination of low and intermediate/high features. In addition, analysis of more than one keyframe per shot, audio analysis, and using temporal context information was tried. This year some sites focused on metadata or automatic speech recognition (ASR), automatic evaluation of modeling strategies, and consideration of scalability issues. Some participation in the concept-pair task with low performance indicates the need for more research into combining multiple concept detections. Finally, still no improvement using external training data has been observed.

For more detailed results see the on-line workshop notebook (TV12Notebook, 2012) and the workshop papers accessible from the publications webpage (TV12Pubs, 2012).

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 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.

In TRECVID 2010, 78 % of the known-items were found by at least one run; in 2011 65 % were found. Participants were encouraged to focus on why 22 % to 35 % of known-items were not found by current approaches in 2010 and 2011 and what more successful approaches can be developed to reduce that percentage for the new topics of 2012.

4.1 System 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 probability to be 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 min 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.C) was 291 h drawn from the IACC.1 collection using videos with durations between 10 s and 3.5 min.

4.3 Topics

361 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

Each group was allowed to submit up to 4 runs and in fact 9 groups submitted 18 automatic and 15 interactive runs. 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 the measure mean inverted rank at which the known item is found or zero if not found. For example, if a known-item is found first at position 5 in the result, the score for that search is 1/5. 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 were also measured.

4.6 Results

Figures 16 and 17 present the system-level results for effectiveness (mean inverted rank) versus mean elapsed time, for the automatic and interactive runs, respectively. As expected, a human in the loop had a large positive influence on effectiveness at the cost of elapsed time. Calls to the interactive oracle are depicted in Figure 18; more calls did not seem to be correlated with improved effectiveness. Finally, the percentage of topics for which no system found the known item is shown in Figure 19. This situation is not very different from that seen in the 2010 and 2011 results, as shown in Figure 20. In order to find out why those videos were not found we checked randomly 30 % of the not found videos in TRECVID 2012 and found some problems that could conceivably have contributed to why no system found the correct video. Some of the video descriptions we checked were not completely accurate, 1 video had very low quality, 1 query was not realistic (used a name of a child as a visual clue), 1 query had multiple videos that could be a correct answer, 1 query description didn't exist in the ground truth video, and finally, there is very hard queries that can have multiple videos with high similarity to the ground truth video. In those cases systems usually return videos that include some of the concepts mentioned in the query description but not exactly all.

The *AXES* team - a multi-team European Union FP7 project - built on previous participation in 2011. They implemented on-the-fly, query-time training of concept classifiers using external examples (Google Images) based on searchers text input. Their system also used text metadata and incorporated face processing. They made 2.9 M face detections in the KIS data. Their score-based fusion built on their 2011 submission with a focus on integrating multiple search services

The *Beijing University of Posts and Telecommunications (BUPT-MCPRL)* group experimented with two approaches. The first was a traditional text-based technique with a focus on colors, language, places, sound, synonym terms and correlations in an ontology. This yielded the 2nd highest effectiveness. The second was a biologically-inspired method that improved on their 2011 submission using a bottom-up attention model for salient regions in the example images. This approach applied to only 37 of the 361 topics but when used improved performance. Future work will focus on how to automatically determine when to use the technique. They had some submission format issues so some results were depressed.

Building on previous participation in 2011 and 2010, the researchers at *Dublin City University (DCU-iAD-CLARITY)* created an iPad application to be used in lean-back interaction. There were two versions using one keyframe representation and using multiple keyframes per video. Eight novice users participated in a Latin square experimental design. Results suggested multiple keyframes out-perform single keyframe by 1 minute in elapsed time, and also in effectiveness

The *Klagenfurt University (ITEC)* team submitted automatic and interactive runs. They used concepts from the semantic indexing task and employed heuristic voting. Their system relied completely on text-based retrieval with rule-based query expansion and query reduction. The interactive system was based on applying filters (e.g., colors, language, music, etc.) to narrow down results of automatic output; no relevance feedback or iterations (2 users) were included.

Greece's *Centre for Research and Technology (ITI-CERTH)* put their focus was on interface interaction with the VERGE system which integrates visual similarity search, transcription (ASR) search, metadata search, aspect models and semantic relatedness of metadata, along with semantic concepts (from the semantic indexing task). More interestingly they compared shot-based and video-based representations of content, finding video-based substantially better in effectiveness and speed.

The *KB Video (KBVR)* team submitted 3 automatic runs. One used BM25 on ASR and metadata. A second was like the first but with concept expansion using the Large-Scale Concept Ontology for Multimedia (LSCOM). A third was like the first but with concept expansion from Wikipedia. Neither the second nor the third found any improvement, because too many concepts were drawn in, too much noise and semantic drift.

Researchers at Japan's *National Institute of Informatics (NII)* submitted automatic and interactive runs. The automatic ones used metadata, plus Google Translate (automatic) for language-specific topics. Results showed translation worsened effectiveness but this could have been due to the overly aggressive pre-processing. In the interactive system, each video was represented as 5 key frames.

Aalto University (PicSOM) in Finland participated with automatic runs. Their baseline was text search of metadata on which they layered on optical character recognition (OCR) of all keyframes in collection, giving a small improvement. They tried layering on ASR with GNU Aspell spelling correction but found that not beneficial. Using the Google Image Search API to locate images visually similar to visual cues from search, also reduced performance.

At *Peking University* the group submitted automatic and interactive runs, which were top-ranked for effectiveness. The topic text was processed by spelling correction (Aspell), part of speech tagging (Stanford parser) to weight the parts of speech differently, and OCR on video frames, followed by topic term weighting and inflectional normalization from a dictionary. Black and white detection was also included, as was detection and filtering of the video language (French, German, etc.).

For more detailed results see the on-line workshop notebook (TV12Notebook, 2012) and the workshop papers accessible from the publications webpage (TV12Pubs, 2012).

5 Instance search pilot

An important need in many situations involving video collections (archive video search/reuse, personal 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 2012 this continued as 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 System 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

- several example frame images drawn at intervals from a video containing the item of interest. For each frame image:
  - a binary mask of an inner region of interest within the rectangle, see Figure 21 for an example.
- an indication of the target type taken from this set of strings (person, location, object)

5.2 Data

Test data: a set of queries expected to return many instances of objects, locations, etc. in a set of classes were created by Robin Aly at the University of Twente in consultation with NIST. The queries were then run against Flickr video available under Creative Commons licenses for research. NIST examined the results, defined the test collection, and created the test queries with examples from outside the test collection. The test collection videos were automatically divided into some 74 958 short, roughly equal-length clips at the University of Twente and renamed so the clip name did not indicate the original video. Each clip was to be processed as if no others existed.

5.3 Topics

Topics were created to emphasize objects. Topic targets included 15 objects, 1 person, and 5 locations. Figures 22-24, 25, and 26 show examples of images provided with the object, person, and location topics, respectively.

5.4 Evaluation, Measures

Each group was allowed to submit up to 4 runs and in fact 24 groups submitted 79 automatic and 6 interactive runs.

The submissions were pooled and then divided into strata based on the rank of the result items. For a given topic, the submissions for that topic were judged by a NIST assessor who played each submitted shot and determined if the topic target was present. The assessor started with the highest ranked stratum and worked his/her way down until too few relevant shots were being found or time ran out. Table 3 presents information about the pooling and judging.

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 (MAP) over all queries. While speed and location accuracy were also definitely of interest here, of these two, only speed was measured in the pilot.

5.5 Results

Figures 27 and 28 are boxplots showing the distribution of per-topic average precision scores across all automatic and interactive runs, respectively. Each graph is divided into 3 sections; the leftmost for object topics, the middle for location topics, and the rightmost for people topics. The test collection size is too small to draw strong conclusions about the differences due to topic type. In general, there is, as expected, great variation from topic to topic with most scores being low, though some systems noticeably exceed the median.

Comparing the best performance by topic in interactive versus automatic runs, Figure 29 shows surprisingly that interactive runs outperformed automatic ones on only 8 or the 21 topics.

Partial randomization tests (Figures 30 and 31) reveal that many of the top systems as measured by mean average precision cannot be distinguished from each other but some statistically significant differences do exist as one looks lower in the ranking.

Figure 32 suggests some correlation between the difficulty of a topic as measured by mean average precision and the number of image examples included in the topic. One can also speculate about other possible factors in topic difficulty. Easier topics seem to include examples using the whole frame, interior shots with constant illumination, while more difficult topics have a smaller region of interest sometimes combined with a complex background.

Mean elapsed processing time per topic ranged from 6 s to 87 h. Figure 33 indicates that better effectiveness is not tied to longer processing times. Two runs with sub-minute processing times and MAP $>$ 0.15 were those from Beijing University of Posts and Telecommunications (BUPT) and the City University of Hong Kong (Vireo). BUPT used a very rich combination of local scale-invariant feature transforms (SIFT), regional (e.g., PHOG: capturing spatial layout) and global features as well as linear fusion, and pseudo feedback. Vireo employed SIFT, a bag-of-visual-words (BOVW) (100K), and spatial consistency postfiltering with an inverted file containing all the information necessary for efficient postfiltering.

All participants used local descriptors, most BOVW. A large variety of exploratory experiments with different objectives were carried out. The main team experiments can be grouped by a number of themes. Systems reused techniques from information retrieval such as dimension reduction using visual words, inverted files for fast lookup, feature weighting (e.g., BM25), pseudo-relevance feedback. In dealing with the masked image examples, participants found that fusion of a whole frame run and a masked region of interest run increased performance. Another approach for diminishing the influence of the visual context of a target of interest was to apply blurring, leading to a better INS performance as shown by the Vireo run. Several systems added extra sample images from Internet sources, yielding mixed results. Experiments on finding an optimal query representation looked at how to fuse features, how to exploit spatial constraints - by dropping spacial information regarding local descriptors or via postfiltering techniques - mostly with encouraging results.

For more detailed results see the on-line workshop notebook (TV12Notebook, 2012) and the workshop papers accessible from the publications webpage (TV12Pubs, 2012).

6 Multimedia event detection

The 2012 Multimedia Event Detection (MED) evaluation was the second evaluation of technologies that search multimedia video clips for complex events of interest to a user. The 2012 included three important changes:

- Events tested: 10 new events were added to the Pre-Specified event evaluation for a total of 20

events,

- Evaluation conditions: a pilot Ad-Hoc event evaluation task was supported which tested systems on an additional 5 events, and a new 10-video exemplar event training condition was introduced.
- Indexing collections: a new test collection, the MED Progress Collection, which is 3722 h in duration, was introduced. The Progress set will be used as a test collection until MED 2015.

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 event explication which is a textual listing of some attributes that are often 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.
- 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 containing either an instance of the event or content "related" to 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.

In 2010 and 2011, developers built Pre-Specified event systems where knowledge of the event(s) was taken into account during generation of the metadata store for the test collection. In 2012, the same Pre-Specified (PS) Event task was supported as well as a new pilot Ad-Hoc event task where the metadata store generation was completed before the events where revealed. The pilot Ad-Hoc (AH) Event task was added a year ahead of schedule because systems performed better than expected during MED '11.

6.1 Data

A development and evaluation collection of Internet multimedia (i.e., video clips containing both audio and video streams) clips (HAVIC) 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.

MED participants were provided the following:

- Development data consisting of:
  - The MED '10 data sets consisting of 3,488 clips totaling $\approx$ 114 h of videos.
  - The MED '11 development and evaluation collections consisting of 42,466 clips totaling 1,315 h of video.
  - Fifteen events from MED '11 - ten were included as testing events this year.

- Ten new Pre-Specified "testing" events kits. The new PS events were released in March 2012.
- Five Ad-Hoc "testing" event kits which were provided to researchers 14 days prior to results being due at NIST.

- Evaluation data consisting of MED Progress Test Collection which contained 98 118 videos (3.1 times as many as MED '11 test collection clips) totaling 3,722 h of video (3.8 times as many MED '11 test collection hours of video).

The MED '12 Pre-Specified event names are listed in Table 5 and Table 6 lists the MED '12 Ad-Hoc Events.

6.2 Evaluation

Sites submitted system outputs for either all 20 Pre-Specified events or all 5 Ad-Hoc events (referred to as a MEDFull submissions) or any fraction thereof (referred to as a MEDPart submissions). Developers reported two aspects of how their systems were constructed. First, their event agents were constructed either with human intervention (SemiAutoEAG) or without human intervention (AutoEAG). Second, agents could use all videos supplied with the event kits for training (EKFull) or with a 10-positive and 10-related video subset of the full event kits (EK10Ex).

For each event search a system generates:

- A Score for each search collection clip: A probability value between 0 (low) and 1 (high) representing the system's confidence that the event is present in the clip.
- A Detection Threshold for the event: A probability value between 0 and 1 - an estimation of the detection score at or above which the system will assert that the event is detected in the clip.
- The event agent execution time: The number of seconds used to search for the event in the metadata store.

System developers also reported the compute hardware used to perform indexing and search and the compute time for indexing.

Submission performance was computed using the Framework for Detection Evaluation (F4DE) toolkit. Groups were required to submit a primary run, which is the run they expect to be their best performing system and optionally allowed to submit multiple runs as contrastive conditions. Each team was allowed to submit up to 4 runs plus an additional 2 runs if they participated in the Ad-Hoc Event task.

6.3 Measures

MED system performance was evaluated as a binary classification system by measuring performance of two error types: Missed Detection (MD) errors and False Alarm (FA) errors. NIST reported the primary performance measures for accuracy and processing speed, and a suite of diagnostic measures that may provide a deeper analysis of system performance.

The primary measure for accuracy was the probability of missed detection (the number of missed detection divided by the number of clips containing an event) and false alarms (the number of false alarms divided by the number of clips not containing the event) for the event (P_{Miss} and P_{FA} respectively) based on the Detection Threshold.

There were two primary measures for computational speed expressed as real-time factors. Real-time factor is the total processing time divided by the number of hours of video in the test collection. Two versions of real-time factors were computed: total real-time and single core adjusted real-time. The first speed measurement is Metadata Generation Processing Speed (MGPS) which is the real-time factor to complete all steps necessary to build the metadata store. The second speed measurements is Event Agent Execution Processing Speed (EAEPS) which is the real-time factor for each event processed during the event agent execution phase.

6.4 Results

17 teams participated in the MED '12 evaluation, 6 teams were new. All teams participated in the Pre-Specified, full event kit test processing all 20 events. 13 teams participated in the Pilot Ad-Hoc task.

The MED12 evaluation was the first use of the MED Progress set. Since the Progress set will be used for 4 successive MED evaluations, protecting the statistic of the Progress set is of the utmost importance. NIST reported only P_{Miss} and P_{FA} values at the system's actual decision to prevent revealing statistics of the Progress test rather than decision error tradeoff (DET) curves for each run.

Table 12 presents the P_{Miss} and P_{FA} for the actual decision threshold, averaged over events for the

primary Pre-Specified event task submissions. The P_{Miss} scores for the top 10 performing systems range from 0.211 to 0.357 and the P_{FA}s range from 0.009 to 0.034. Figure 34 presents the same information but displayed in DET Curve space as a single point per site. The top 10 performing systems form two clusters (ECNU, AXES, and TokyoTechCanon) vs. the other 7. The clusters highlight the importance of specifying a common threshold selection criteria. The ECNU, AXES, and TokyoTech thresholds appear to be selected preferring a lower false alarm rate whereas the other 7 targeted a False Alarm rate of 4 %.

Table 8 and Figure 35 show similar depictions of the MED AdHoc Pilot submissions. For the top 9 systems, P_{Miss} scores ranged from 0.200 to 0.325 and the P_{FA} scores ranged from 0.015 to 0.12. Although the Pre-Specified and AdHoc events are different sets of events, encouragingly, the performance range of AdHoc systems overlaps substantially with Pre-Specified event systems.

A new feature to MED '12 was the introduction of the 10 exemplar event training condition. Figure 36 is a DET curve displaying both the participating site's EKFull system (as a solid circle) and the sites' corresponding EK10EX system (as an open circle). The relative average degradation (excluding IBMCU) was 72 % and 58 % for P_{Miss} and P_{FA} respectively. For the most systems, the exemplar reduction had a larger impact on miss rates which is presumably due to a smaller population of examples. In future years, it's expected systems will be able to leverage more content in the event kit text to reduce the degradation.

Participants were asked to report metadata generation computational hardware and Metadata Generation Processing Speed (MGPS) and Event Agent Execution Processing Speed (EAEPS). The MGPS realtime factors of reporting systems ranged from 0.011 to 0.659. The EAEPS realtime factors of reporting systems ranged from 0.443 to 8.402. Participants were also asked to report number of processor cores used for each of their processing steps so that core normalized MGPS and EAEPS could be computed. After reviewing the submissions and talking with participants, it was decided that computing core-normalized measures had no consistent meaning because of the vagaries of modern CPUs and queuing systems. Therefore, the values will not be reported.

6.5 Summary

In summary, 17 teams participated in the MED '12 evaluation. All teams participated in the Pre-Specified event tasks and 13 teams participated on the Pilot Ad-Hoc event task. Ten teams achieved average P_{Miss} error rates below 0.357 and P_{FA} rates below 0.034. The pilot AdHoc task was a success in that system performance did not appear to degrade drastically compared to Pre-Specified systems therefore a scale up AdHoc test will occur next year. The reduced exemplar condition showed that P_{Miss} was affected more by the reduction of event training exemplars.

For more detailed results see the on-line workshop notebook (TV12Notebook, 2012) and the workshop papers accessible from the publications webpage (TV12Pubs, 2012).

7 Multimedia event recounting

The 2012 Multimedia Event Recounting (MER) evaluation was the first, pilot evaluation of technologies that recount multimedia video events detected by MED systems. The evaluation also included MER output generated for known-positive clips.

The purpose, of the 2012 Multimedia Event Recounting (MER) track, was to stimulate the development of technologies that state the evidence that led a Multimedia Event Detection (MED) system to decide that a multimedia clip contains an instance of a specific event.

The initial pilot evaluation of MER consisted of two metrics. The first was to evaluate whether the MER outputs by themselves allow human judges to identify which of five events is represented by a recounting. The second goal was to evaluate whether the MER outputs are sufficiently expressive that judges can match each recounting to the clip from which it was derived.

The key goal was to focus on **content**. Each event kit explicitly defines an *event*. A clip that is *positive* for an event contains an *instance* of that event.

Each event in this evaluation

- 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 orga-

nized and that have significant temporal and semantic relationships to the over-arching activity; and

- is directly observable.

Participation in MER was open to all current TRECVID participants.

7.1 System task

Given an event kit and a test video clip that contains an instance of the event, the MER system was to produce a recounting summarizing the key evidence for the event in the clip. Evidence means observations of scene/context, persons, animals, objects, activities, text, non-linguistic audio, and other evidence supporting the detection of the event. Each observation was associated with an indication of the systems confidence that the observation was correct or accurate.

Systems were asked to produce an XML element for each *observation*, and that element included attributes that gave the following information. Note: the "id" and the "description" were required; the rest of the information was strongly encouraged but optional.

id a unique identifier that can be used in other XML elements to associate elements, e.g., to associate an object or person with an activity

type a list of possible values for the type attribute appears below

description a textual statement of the observation (For example, if the *type* is *object*, the *description* might be *red Toyota Camry.*) The description may be used to state only what is observable (e.g., red Camry) or may also include semantic inferences (e.g., the getaway vehicle).

semantics an optional attribute which can be used if a system restricts its *description* attribute to only what is observable.

startTime an offset into the clip, either as time or as a frame number

endTime an offset into the clip, like the startTime

boundingBox pixel coordinates relative to the upper left corner of the frame for the upper left (row, column) and lower right (row, column) corners of a box that would surround a visible piece of evidence at the *reference time*. No boundingBox should be stated for purely acoustic evidence.

referenceTime a time when the visible piece of evidence is at the location indicated by the *boundingBox*. The reference time should be between startTime and endTime, inclusive.

confidence in the range 0.0 through 1.0, with 1.0 indicating highest confidence

importance in the range 0.0 through 1.0, with 1.0 indicating highest importance

In the XML element for an observation, the *type* attribute will have one of the following values.

scene A "scene or context" is a descriptive set of information flowing from a physical environment. It could include things such as a cityscape, an agricultural farm, a natural setting, a park containing children's swings, or a broad activity such as a soccer game. Also included are unresolved groupings such as a crowd, a clump of trees, or a bunch of houses; or a sub-event for example, lightning striking, a vehicle exploding, or a rock slowly tumbling down a hill.

object An "object" is something inanimate that is visible in the clip. Examples include a tent, suitcase, building, or tree. It is possible for an object to be in motion.

person "Person" means one human being.

animal "Animal" means an animal, not a human.

activity An "activity" is a person or animal doing something. Examples include a person running, putting up a tent, throwing a ball, playing basketball, talking, or hiding. Examples of an activity involving an animal include a dog fetching a stick or a cat chasing a mouse. Note that an activity involves a *living* actor.

text "Text" is

- any text visible in a clip (often referred to as "scene text"), typically captured via OCR
- text overlaid on the clip (titles, closed-captions, etc.), typically captured via OCR, or
- understandable speech (the idea here is the output of an ASR or speech-to-text component)

non-linguistic audio "Non-linguistic audio" (also known as an acoustic event), is sound other than understandable speech. Examples include crash, gunshot, honk, laugh, sneeze, bark, or babble (as of a crowd).

videography Motion of the camera taking the video, or editing done to the video, that is relevant to evidence of the event. Possible examples are "the camera tracks the person" or "the camera zooms in on her hands"

other "Other" is a place for system-defined additional useful information to understand the event, and is intended as an opportunity for MER developers to include evidence that does not fit into the categories of observations described in the preceding list of possibilities. Examples include video quality

In addition to the *observation* elements, the MER output can include *relationship* elements that capture a relationship among observations. For example, if there is an observation element for a man and an observation element for a hammer, there can be a relationship element that says the man is holding the hammer. Relationship elements have a required *semantic_inference* attribute and optional attributes *startTime*, *endTime*, *confidence*, and *importance*. The events used in the TRECVID MER evaluation all inherently involve relationships.

The MER output can optionally include *sequence_of_activities* elements, which are intended for use by MER systems that choose to break down a clip into phases, sub-events, or groups of activities.

Because the purpose of MER systems is to state the evidence for the occurrence of the event, there is an *evidence* element, which groups all the *observation*, *relationship*, and *sequence_of_activities* elements that pertain to the event.

For each clip, participants were to track and report separately: (1) the time required for evidence identification and exaction (including all preprocessing time required to ingest the clip), and (2) the time required for MER output generation.

All participants in the MER Evaluation track were given five event kits (textual description and multimedia clip exemplars), with six evaluation video clips, per event kit, that contained the event defined by the event kit, and tasked to produce a recounting that summarized the key evidence of the event.

The events, chosen from the MED pre-specified events list, were

- **E022** – Cleaning an appliance,
- **E026** – Renovating a home,
- **E027** – Rock climbing,
- **E028** – Town hall meeting, and
- **E030** – Working on a metal crafts project.

For this first TRECVID MER evaluation, the recounting was text-only. Participants were evaluated on 30 multimedia clips that each contained one of five MER events.

MER participants who also participated in the MED (pre-specified) evaluation were required to produce a recounting for each clip that their MED system declared as containing one of the five MER evaluation events. From these outputs, for each of the five MER evaluation events, NIST selected six clips which all the systems had correctly identified as positive. MER outputs for those 30 clips were also evaluated. One system team submitted MER outputs for the MER Evaluation Test Set but not for that system's MED positives, and that system's MER submission could therefore not be evaluated on a fair, equal basis with the other submissions. Results for that system are not included in this paper.

7.2 Data

Three data sets were provided for the 2012 MER evaluation track each containing clips from the MER event set listed above. These three MER data sets are as follows.

1. **MER Development Test Set** – This dataset was limited to 6 video clips from each of the five events in the MER event set, and was provided to support research and a dry run of the evaluation pipeline. There were exactly 30 video clips in this dataset.

2. **MER Evaluation Test Set** – This dataset was limited to 6 video clips from each of the five events in the MER event set, and was provided to support the evaluation specified below. There were exactly 30 video clips in this data set.

3. **MER Progress Test Set** – This dataset was defined for MER participants who also participated in MED. NIST selected exactly 30 positive video clips for evaluation (6 video clips from each of the 5 MER events). MER participants were to generate a recounting for each of the 30 clips in the MER Evaluation Test SET.

All MER participants also participated in MED (pre-specified) and were therefore additionally required (for all five events in the MER event set) to generate a recounting for all MED (pre-specified event) clips that their MED system identified as being above their MED systems decision threshold for being positive for the event of interest. One submission omitted these MER outputs, and its MER outputs therefore could not be judged on an equal (apples to apples) basis with the other submissions. For that reason, results from that submission are omitted from this paper.

All participants were required to participate in a dry run exercise using the **MER Development Test Set** to ensure that both the system outputs were being generated as expected and were parsable by the evaluation pipeline. This exercise also provided insight into how the recounting could be rendered for the judges in the formal evaluation.

Input data formats were as in existing HAVIC data. MER output data formats used ASCII XML text. NIST provided a rendering tool and a MER document type definition (DTD) to be used to specify and validate system output.

7.3 Evaluation

The system's MER outputs for the MER Evaluation Test Set and for the MER Progress Test Set (two corpora) were evaluated by a panel of judges (experienced video analysts and Linguistic Data Consortium staff). The two corpora, and each system, were judged separately. The judges performed two tasks: first, without seeing the clips, the judges attempted to identify which of the five events were represented by each MER output. Secondly, for each MER event and each system separately, the judges were provided with six positive clips along with the output from a system, and attempted to match each recounting with the clip from which it was derived.

NIST assessed the MER outputs by analyzing how accurately the judges were able to perform the two tasks.

7.4 Measures

Several metrics were used in this evaluation.

Metrics for distinguishing one event from another, using only MER output

The system performance metric for this subtask is the fraction of the judgments that correctly identified which of the five events was represented by each MER output, averaged across the events and judges (that set of results is shown in Figure 37. The event difficulty (or confusability) metric for this subtask is the fraction of these judgments that were correct, averaged across the systems and judges (that set of results is shown in Figure 38). In addition, NIST computed the fraction of the judgments that were correct for each combination of system and event, averaged across only the judges (that set of results is shown in Figure 39).

Metrics for distinguishing which clip is described, using MER output plus the clips

The system performance metric for this subtask is the fraction of the matches, of recountings to the clips from which they were derived, that were correct, averaged across the events and judges (that set of results is shown in Figure 40). The event difficulty metric for this subtask is the fraction of the matches (of recountings to clips) that were correct, averaged across the systems and judges (results shown in Figure 41). This event difficulty metric reflects the difficulty or confusability of the clips that were chosen for the event. In addition, NIST computed the fraction of the matches that were correct for each combination of system and event, averaged across only the judges (results shown in Figure 42).

7.5 Results

For detailed results on each run's performance, see the on-line workshop notebook (TV12Notebook, 2012) and the workshop papers accessible from the publications webpage (TV12Pubs, 2012). That level of voluminous detail is omitted from this paper.

8 Interactive surveillance event detection

The 2012 Surveillance Event Detection (SED) evaluation was the fifth evaluation focused on event detection in the surveillance video domain. The first such evaluation was conducted as part of the 2008 TRECVID conference series (Rose, Fiscus, Over,

Garofolo, & Michel, 2009) and again in 2009, 2010, and 2011. 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 consisting of multiple, synchronized camera views.

For 2012, the evaluation re-used the 2009 test corpus and 2010 events. The major change for 2012 was the introduction of the Interactive SED Task.

In 2008, NIST collaborated with the Linguistics Data Consortium (LDC) and the research community to select a set of naturally occurring events with varying occurrence frequencies and expected difficulty. 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 same set of seven 2010 events were used for the 2011 and 2012 evaluations.

For 2012, the evaluation re-used the 2009 test corpus. The test data was the Imagery Library for Intelligent Detection System's (iLIDS) (UKHO-CPNI, 2007 (accessed June 30, 2009)) Multiple Camera Tracking Scenario Training (MCTTR) data set collected by the United Kingdom's Home Office Science and Development Branch.

In 2012, the Retrospective Surveillance Event Detection (rSED) and Interactive Surveillance Event Detection (iSED) tasks were supported.

- The retrospective task is 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). The retrospective task addresses the need for automatic detection of events in large amounts of surveillance video. It requires application of several Computer Vision techniques, involves subtleties that are readily understood by humans, yet difficult to encode for machine learning approaches, and can be complicated due to clutter in the environment, lighting, camera placement, traffic, etc.
- The interactive task is defined as follows: given a collection of surveillance video data files (e.g., that from an airport, or commercial establishment) for preprocessing, at test time detect observations of events based on the event definition and for each return the elapsed search time and a list of video segments within the surveillance data files, ranked by likelihood of meeting the need described in the topic. Each search for an event by a searcher can take no more than 25 elapsed minutes, measured from the time the searcher is given the event to look for until the time the result set is considered final. Note that iSED is not a short latency task. Systems can make multiple passes over the data prior to presentation to the user.

  The Motivation for an interactive task is that SED remains a difficult task for humans and systems. Also, Interactivity and relevance feedback have been effectively employed in other tasks.

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).

8.1 Data

The development data consisted of the full 100 h data set used for the 2008 Event Detection (Rose et al., 2009) evaluation. The video for the evaluation corpus came from the approximate 50 h iLIDS MCTTR data set. Both data sets were collected in the same busy airport environment. 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 Internet download. Figure 43 shows the coverage and views from the different cameras used for data collection.

System performance was assessed on the same 15-h subset of the evaluation corpus as the 2009 Evaluation. Unlike previous SED evaluations, systems were provided the identify of the evaluated subset so that searcher time for the interactive task was not expended on non-evaluated material. Event annotation was performed be the LDC using a three-pass annotation scheme. The multi-pass process improves the human annotation recall rates.

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.

8.2 Evaluation

Sites submitted system outputs for the detection of any 3 of 7 possible events (PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, and Pointing). Additional details for the list of event used can be found in Figure 44. For each instance observation, sites are asked to identify each detected event observation by:

- the temporal extent (beginning and end frames)
- a decision score: a numeric score indicating how likely the event observation exists with more positive values indicating more likely observations (normalized)
- an actual decision: a boolean value indicating whether or not the event observation should be counted for the primary metric computation

Developers were advised to target a low miss, high false alarm scenario, in order to maximize the number of event observations.

Groups were allowed to submit multiple runs with contrastive conditions. System submissions were aligned to the reference annotations scored for missed detections / false alarms.

8.3 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 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.

8.4 Results

There were 12 participants in 2012 (see Figure 45), for a total of 108 Interactive Event Runs and 95 Retrospective Event Runs.

Figure 46 presents the event-averaged lowest NDCR by site's iSED vs rSED for the 8 sites that submitted both types of runs. Out of those 8 sites, 5 show some reduction in their NDCR, with two large reductions (BrnoUT by 19 % and BUPT-MCPRL by 29 %).

Comparable results since 2009 for rSED, and adding the 2012 iSED results to the plots are present in Figures 47 to 53. In those plots, one can see that Single-person (PersonRuns, PeopleSplitUp, Pointing) and Multi-Person (PeopleMeet, Embrace) events show evidence of yearly improvements, still not approaching human performance. Person+Object (ObjectPut, CellToEar) events remain difficult.

For detailed results see the on-line workshop notebook (TV12Notebook, 2012) and the workshop papers accessible from the publications webpage (TV12Pubs, 2012).

9 Summing up and moving on

This overview of TRECVID 2012 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. The raw results for each submitted run can be found in the results section at the back of the online notebook (TV12Notebook, 2012).

10 Authors' note

TRECVID would not have happened in 2012 without support from the National Institute of Standards and Technology (NIST) and the Intelligence Advanced Research Projects Activity (IARPA). The research community is very grateful for this. Beyond that, various individuals and groups deserve special thanks:

- Alan Smeaton and Brian Boyle at DCU arranged for the mirroring of the video data.

- Georges Quénot with Franck Thollard, Andy Tseng, Bahjat Safadi from LIG and Stéphane Ayache from LIF shared coordination of the semantic indexing task, organized the community annotation of concepts, and provided judgments for 31 concepts under the Quaero program.
- Michal Hradiš and his team at the Brno University of Technology provided 300 h of additional annotations for the IACC.1.C video.
- Georges Quénot provided the master shot reference for the IACC.1.C videos.
- The LIMSI Spoken Language Processing Group and VexSys Research provided ASR for the IACC.1.C videos.
- Cees Snoek helped choose the SIN concept-pairs and provided 4 baseline concept-pair runs
- Robin Aly at the University of Twente worked with NIST to develop various queries and ran them against Flickr to form the basis of the INS test data for 2012, available under Creative Commons licensing
- Kevin McGuinness at Dublin City University ran the oracle there for interactive systems in the known-item search task.

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

9048 OBJECT - Mercedes star

9049 OBJECT - Brooklyn bridge tower

9050 OBJECT - Eiffel tower

9051 OBJECT - Golden Gate Bridge

9052 OBJECT - London Underground log

9053 OBJECT - Coca-cola logo - letters

9054 LOCATION - Stonehenge

9055 OBJECT - Sears/Willis Tower

9056 LOCATION - Pantheon interior

9057 OBJECT - Leshan Giant Buddha

9058 OBJECT - US Capitol exterior

9059 OBJECT - baldachin in Saint Peter's Basilica

9060 PERSON - Stephen Colbert

9061 OBJECT - Pepsi logo - circle

9062 OBJECT - One World Trade Center building

9063 LOCATION - Prague Castle

9064 OBJECT - Empire State Building

9065 LOCATION - Hagia Sophia interior

9066 LOCATION - Hoover Dam exterior

9067 OBJECT - MacDonald's arches

9068 OBJECT - PUMA logo animal

References

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

Gauvain, J., Lamel, L., & Adda, G. (2002). The LIMSI Broadcast News Transcription System. *Speech Communication*, *37*(1-2), 89—108.

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.* `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, December). The TRECVid 2008 Event Detection Evaluation. In *IEEE Workshop on Applications of Computer Vision (WACV).* IEEE.

TV12Notebook. (2012). `http://www-nlpir.nist.gov/projects/tv2012/active/workshop.notebook`.

TV12Pubs. (2012). `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.12.org.html`.

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, November). 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.

12 Tables and Figures

Table 1: Participants and tasks

Task						Location	TeamID	Participants
IN	KI	−−	−−	−−	SI	Europe	PicSOM	Aalto U.
IN	−−	−−	−−	−−	**	Europe	Bilkent	Bilkent U.
−−	−−	**	**	SD	SI	NorthAm	INF	Carnegie Mellon U.
IN	−−	**	−−	−−	SI	Europe	CEALIST	CEA
IN	**	MD	MR	−−	SI	Asia	VIREO	City U. of Hong Kong
**	−−	MD	MR	−−	−−	NorthAm	CU	Columbia U.
−−	−−	−−	−−	SD	−−	Asia	SJTU_BCMI	Shanghai Jiaotong U.
IN	−−	−−	−−	−−	−−	SouthAm	PRISMA-Orand	U. of Chile
IN	−−	−−	−−	−−	**	Asia	U_Tokushima	U. of Tokushima
IN	KI	MD	MR	−−	−−	Europe	DCU_IAD	Dublin City U., IAD
IN	KI	MD	−−	−−	−−	Europe	AXES	Acess to Audiovisual Archives
−−	−−	−−	−−	SD	−−	Europe	dcu_savasa	Dublin City U., CLARITY
**	**	**	−−	−−	SI	Europe	ECL_Liris	Ecole Centrale de Lyon, Universit de Lyon
−−	−−	**	−−	−−	SI	Europe	EURECOM	EURECOM - Multimedia Communications
−−	−−	−−	−−	−−	SI	Europe	VideoSense	EURECOM VideoSense Consortium
**	−−	−−	−−	SD	**	Europe	Brno	Brno U. of Technology
−−	−−	−−	−−	−−	SI	NorthAm	FIU_UM	Florida International U. U. of Miami
IN	−−	−−	−−	−−	SI	Asia	FTRDBJ	France Telecom Orange Labs (Beijing)
IN	−−	−−	−−	−−	−−	Europe	MADM	German Research Center for AI
−−	−−	**	−−	−−	SI	Asia	kobe_muroran	Kobe U.; Muroran Institute of Technology
**	**	MD	MR	SD	SI	NorthAm	IBM	IBM T. J. Watson Research Center
**	KI	MD	MR	−−	SI	Europe	ITI_CERTH	Informatics and Telematics Institute
**	−−	**	−−	−−	SI	Europe	Quaero	INRIA, IRIT, LIG, U. Karlsruhe
IN	−−	−−	−−	−−	−−	Europe	ARTEMIS.Ubi..	Inst. TELECOM; TELECOM SudParis; Bell Labs, Fr.
−−	−−	−−	−−	SD	−−	Asia	BJTU_SED	Beijing Jiaotong U.
−−	−−	MD	−−	−−	SI	Asia	ECNU	East China Normal U.
IN	KI	**	−−	−−	**	Asia	PKU_ICST	Peking U.
−−	−−	−−	−−	SD	−−	Asia	PKU_OS	Peking U. (OS)
IN	−−	**	−−	−−	SI	Europe	JRS.VUT	JOANNEUM RESEARCH; Vienna U. of Technology
−−	KI	**	−−	−−	−−	NorthAm	KBVR	KB Video Retrieval
−−	−−	MD	MR	−−	−−	NorthAm	GENIE	Kitware Inc.
−−	KI	−−	−−	−−	−−	Europe	ITEC_KLU	Klagenfurt U.
IN	−−	**	−−	−−	SI	Europe	IRIM	Indexation et Recherche d'Inform. MM GDR-ISIS
IN	KI	**	−−	SD	**	Asia	BUPT.MCPRL	Beijing U. of Posts and Telecommunications
IN	KI	MD	−−	**	SI	Asia	NII	National Institute of Informatics
**	**	**	−−	**	SI	Asia	NHKSTRL	NHK Science and Technical Research Labs
IN	−−	MD	**	−−	−−	Asia	NTT_NII	NTT Comm. Sci. Labs, Natl. Inst. of Informatics
−−	−−	−−	−−	−−	SI	Asia	ntt	Dalian U. of Technology
IN	−−	MD	−−	−−	−−	Asia	IMP	Osaka Prefecture U.
−−	−−	−−	−−	SD	−−	Asia	PKU_NEC	Peking U. and NEC Labs China
−−	−−	−−	−−	SD	−−	Austral	SAIVT	Queensland U. of Technology
−−	−−	MD	MR	−−	−−	NorthAm	BBNVISER	Raytheon BBN Technologies
IN	−−	−−	−−	−−	−−	Austral	RMIT	RMIT U. School of CS&IT
−−	**	−−	−−	−−	SI	Asia	IRC_Fuzhou	Fuzhou U.
−−	−−	MD	MR	−−	−−	NorthAm	SESAME	SRI International SESAME
−−	−−	MD	MR	−−	−−	NorthAm	Aurora	SRI International Sarnoff Aurora
**	−−	−−	−−	−−	SI	NorthAm	stanford	Stanford U.
−−	−−	−−	−−	SD	−−	NorthAm	MediaCCNY	The City College of New York Media Team
IN	−−	−−	−−	−−	−−	Europe	TNOM3	TNO
−−	−−	MD	**	−−	SI	Asia	TokyoTechCanon	Tokyo Institute of Technology and Canon

Task legend. IN:instance search; KI:known-item search; MD:multimedia event detection; MR:multimedia event recounting; SD: surveillance event detection; SI:semantic indexing; −−:no run planned; **:planned but not submitted

Table 2: Participants and tasks (continued)

Task						Location	TeamID	Participants
IN	−−	MD	MR	−−	SI	Europe	MediaMill	U. of Amsterdam
IN	−−	−−	−−	SD	−−	NorthAm	UCSB_UCR_VCG	U. of California, Santa Barbara
**	**	MD	−−	−−	SI	Asia	UEC	U. of Electro-Communications
−−	−−	−−	−−	−−	SI	Europe	GIM	U. of Extremadura
−−	−−	−−	−−	SD	−−	NorthAm	VIVA_uOttawa	U. of Ottawa
IN	−−	−−	−−	−−	−−	Europe	sheffield_harbin	U. of Sheffield
IN	−−	−−	−−	−−	**	NorthAm	ATTLabs	AT&T Labs Research

Task legend. IN:instance search; KI:known-item search; MD:multimedia event detection; MR:multimedia event recounting; SD: surveillance event detection; SI: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
9048	73379	29603	40.3	160	7344	24.8	50	0.7
9049	74937	29044	38.8	160	6827	23.5	57	0.8
9050	74368	30892	41.5	140	5637	18.2	27	0.5
9051	75524	26618	35.2	140	4538	17.0	29	0.6
9052	73347	31769	43.3	160	7061	22.2	59	0.8
9053	75850	31498	41.5	160	6879	21.8	61	0.9
9054	74329	31173	41.9	300	12585	40.4	40	0.3
9055	70829	29531	41.7	300	11937	40.4	36	0.3
9056	75585	25616	33.9	300	9740	38.0	43	0.4
9057	74042	26457	35.7	300	10848	41.0	21	0.2
9058	73017	27583	37.8	240	9549	34.6	79	0.8
9059	74308	30410	40.9	240	10005	32.9	30	0.3
9060	74384	31171	41.9	180	7362	23.6	45	0.6
9061	74389	32866	44.2	200	9369	28.5	13	0.1
9062	74016	30266	40.9	300	12282	40.6	9	0.1
9063	74367	27394	36.8	300	11257	41.1	25	0.2
9064	72406	30779	42.5	300	12521	40.7	51	0.4
9065	75752	29303	38.7	300	11634	39.7	10	0.1
9066	74843	29489	39.4	180	7722	26.2	12	0.2
9067	75268	29103	38.7	180	7427	25.5	5	0.1
9068	73929	29447	39.8	180	7237	24.6	15	0.2

Table 4: 2011 Teams not submitting any runs

IN	KI	MD	MR	SD	SI	Location	TeamID	Participants
−−	−−	**	−−	**	−−	Europe	AIT_MKWT	Athens Information Technology
−−	−−	**	**	−−	−−	Austral	ANU	Australian National U.
**	−−	**	−−	**	−−	NorthAm	TrackingResearch	BAE Systems, Inc.
−−	−−	−−	−−	−−	**	Asia	BJTU_SIN	Beijing Jiaotong U. Semantic Indexing
−−	**	−−	−−	**	−−	Europe	F4K	Catania U., CWI Amsterdam, U. of Edinburgh
**	−−	−−	−−	−−	**	NorthAm	RITLCS	Rochester Institute of Technology
−−	−−	**	−−	**	**	Asia	BIT	Beijing Institute of Technology
−−	−−	**	−−	−−	−−	Asia	TheBundVideo	Fudan U.
**	−−	−−	−−	−−	−−	Asia	MCVL	Huazhong U. of Science & Technology
−−	−−	−−	−−	−−	**	Asia	NUDTISEL	Information System Engineering Lab
−−	−−	−−	−−	**	−−	Europe	WILLOW	INRIA - WILLOW
**	−−	−−	−−	**	**	Europe	inria_texmex	INRIA - Texmex
**	−−	**	**	−−	−−	Europe	INRIA_LEAR	INRIA's Lear group
**	**	**	−−	**	**	Asia	THU_FRDC_NWPU	Tsinghua U.; Fujitsu R&D, NW Polytech. U.
**	**	**	**	−−	−−	Asia	IVS	Korea Advanced Inst. of Science & Technology
**	**	**	−−	**	**	Europe	METU_EEE	Middle East Technical U.
−−	**	**	−−	−−	**	Eur.+Asia	METU_TODAI	Middle East Technical U.; U. of Tokyo
**	**	**	**	**	**	Asia	MMM_TJU	Multimedia Institute of Tianjin U.
−−	−−	−−	−−	**	−−	Asia	CAS_Team	Institute of Automation
**	**	**	−−	**	**	Europe	NDRC	National Digital Research Center
−−	−−	−−	−−	−−	**	Asia	lixuan	National Laboratory of Pattern Recognition
−−	−−	−−	−−	**	**	Asia	PostechCVlab	Postech
**	−−	−−	−−	−−	−−	Europe	RGU	Robert Gordon U.
**	**	**	−−	−−	**	Europe	Lincoln	Brayford Pool U. of Lincoln
**	**	−−	−−	−−	**	Asia	IMMG	School of Software, Tsinghua U.
−−	−−	−−	−−	**	−−	Asia	SJTU_IS2012	SJTU
**	−−	−−	−−	−−	−−	Asia	GTIL.Sysu	Sun Yat-sen U.
−−	−−	−−	−−	**	**	Asia	CVS_TJUT	Tianjin U. of Technology
**	**	−−	−−	−−	−−	NorthAm	VISLab	UC Riverside, VISLab
**	**	−−	−−	−−	**	Europe	marburg	U. of Marburg
−−	−−	−−	−−	**	−−	NorthAm	UCR_VCG	U. of California, Riverside
**	−−	−−	−−	**	−−	NorthAm	VRL_UCSB_TEAM	U. of California, Santa Barbara
**	**	**	−−	−−	**	SouthAm	RECOD	U. of Campinas (UNICAMP)
−−	−−	**	−−	**	−−	NorthAm	UCFCVL	U. of Central Florida
−−	**	−−	−−	−−	**	Europe	Glasgow_IR	U. of Glasgow Information Retrieval
−−	−−	−−	−−	−−	**	Asia	MONASH_MULTI..	U. Sunway Campus Malaysia
**	**	**	−−	**	**	Africa	REGIM_VIDEO	Universit de Sfax
−−	−−	**	−−	−−	−−	NorthAm	USC_TRECVID	U. of Southern California
−−	**	−−	−−	−−	**	NorthAm	yorku	York U.

Task legend. IN:instance search; KI:known-item search; MD:multimedia event detection; MR:multimedia event recounting; SD: surveillance event detection; SI:semantic indexing; −−:no run planned; ∗∗:planned but not submitted

Table 5: MED '12 Pre-Specified Events

Testing Events
MED'11 event re-test
Birthday Party
Changing a vehicle tire
Flash mob gathering
Getting a vehicle unstuck
Grooming an animal
Making a sandwich
Parade
Parkour
Repairing an appliance
Working on a sewing project
New for MED'12
Attempting a bike trick
Cleaning an appliance
Dog show
Giving directions to a location
Marriage proposal
Renovating a home
Rock climbing
Town hall meeting
Winning a race without a vehicle
Working on a metal crafts project

Table 6: MED '12 Ad-Hoc Events

Testing Events
Doing homework or studying
Hide and seek
Hiking
Installing flooring
Writing text

Table 7: MED '12 Event-Averaged, P_{Miss} and P_{FA} for Primary Pre-Specified Event Systems

		EvAvg-PFA	EvAvg-PMiss
AXES	p-LFdnbig_2	0.009	0.357
BBNVISER	p-Baseline_2	0.026	0.256
CERTH-ITI	p-visual_1	0.001	0.898
CMU	p-ensembleKRSVM_1	0.034	0.211
DCU-iAD-CLARITY	p-MultiModels_1	0.125	0.579
ECNU	p-baseline_1	0.012	0.303
Genie	p-MixAndMatch_2	0.026	0.335
IBMCU	p-IFAYL-Fusion_1	0.027	0.355
MediaMill	p-FusionAll-lateDBG_1	0.032	0.274
NII	p-FusionGlobalFeatures_1	0.005	0.816
NTT-NII	p-baseline_1	0.115	0.872
OPU	p-fusion_1	0.071	0.663
SRIAURORA	p-LLFeatHLFeatAsrOcrLFGM_1	0.030	0.261
Sesame	p-fusionWMroot-lateDBG_1	0.029	0.224
TokyoTechCanon	p-GSSVM7PyramidCcScv-r1_1	0.014	0.354
UEC	p-Sys_1	0.156	0.615
VIREO	p-FUSIONALLREG_1	0.011	0.682

Table 8: MED '12 Event-Averaged, P_{Miss} and P_{FA} for Primary AdHoc Event Systems

		EvAvg-PFA	EvAvg-PMiss
AXES	p-LFdnbig_1	0.015	0.325
BBNVISER	p-Baseline_4	0.033	0.222
CMU	p-SVM_1	0.035	0.208
DCU-iAD-CLARITY	p-MultiModels_1	0.395	0.400
Genie	c-MixAndMatchAdHoc_1	0.029	0.313
IBMCU	p-Fusion_1	0.024	0.410
MediaMill	p-FusionAll_1	0.031	0.263
NTT-NII	p-baseline_1	0.670	0.372
OPU	p-fusion_1	0.073	0.643
SRIAURORA	p-LLFeatHLFeatAsrOcrLFGM_1	0.027	0.299
TokyoTechCanon	p-GSSVM7PyramidCcScv-r5_1	0.120	0.245
UEC	p-Sys_1	0.118	0.552

Figure 1: SIN: Frequencies of shots with each feature

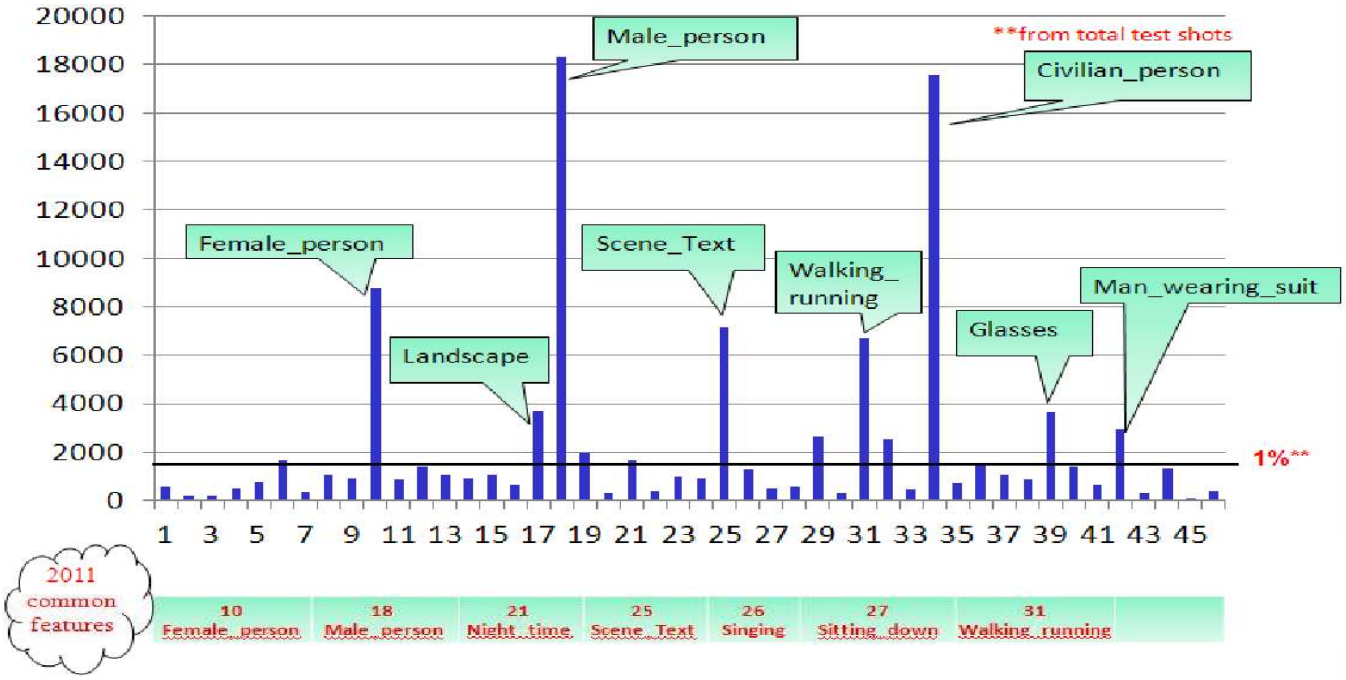


Figure 2: SIN: True positives and False positives per evaluated feature

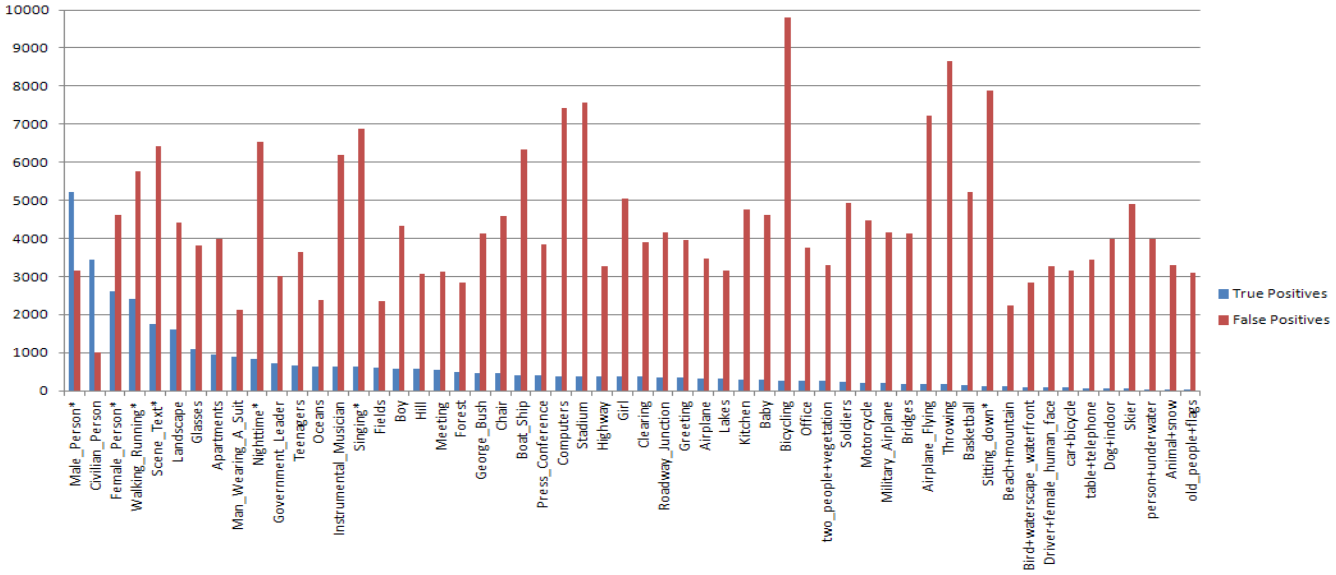


Figure 3: xinfAP by run (cat. A) - Full

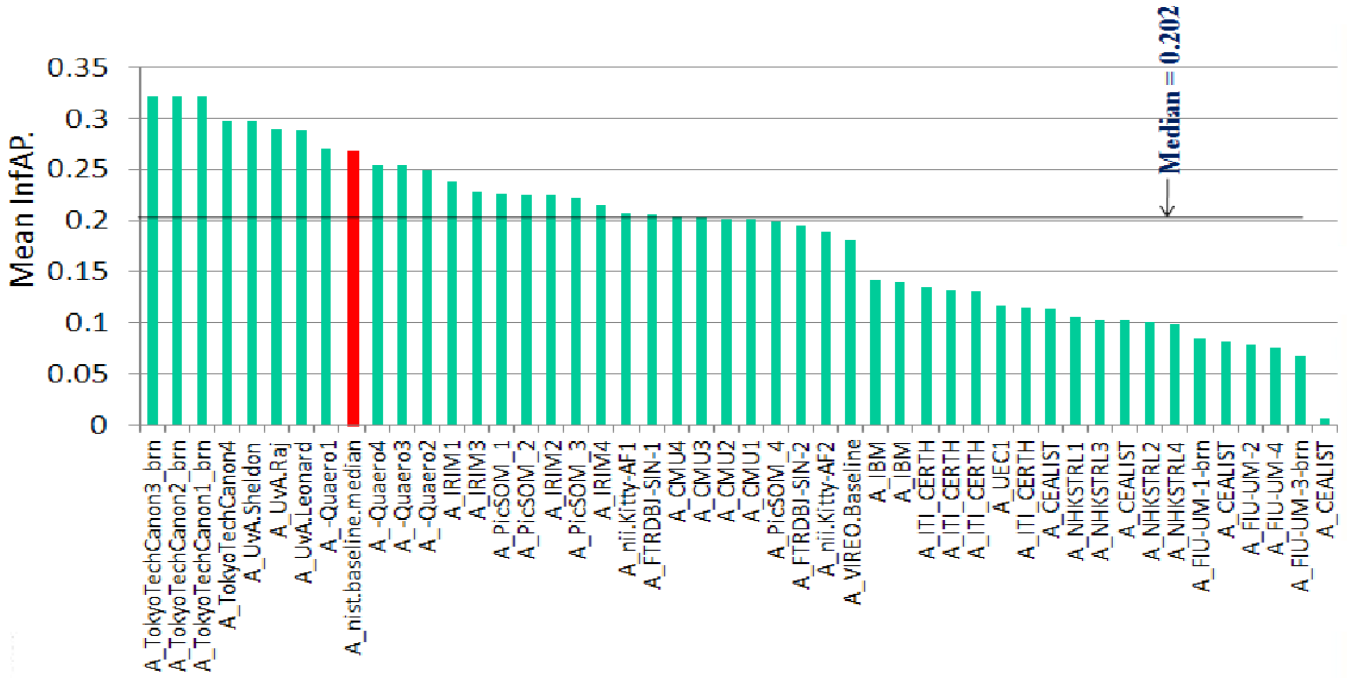


Figure 4: xinfAP by run (cat. D) - Full

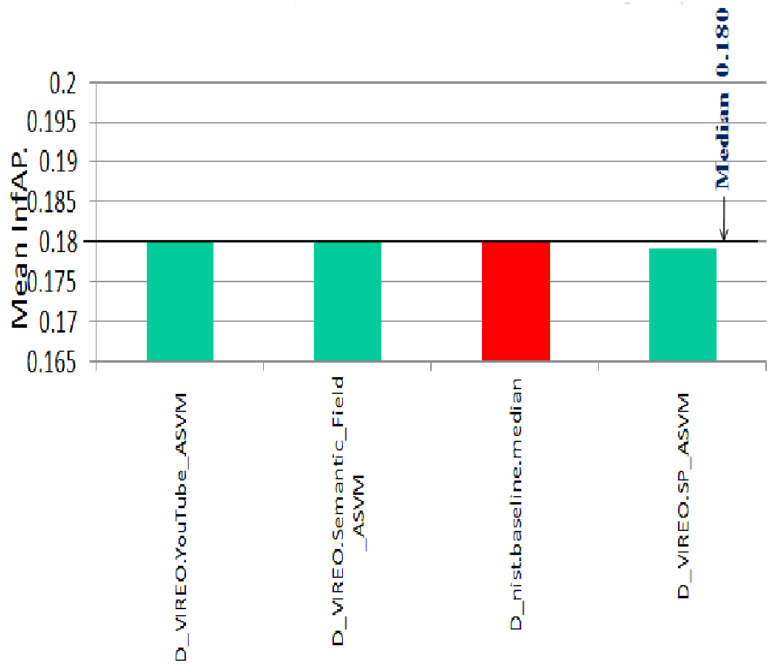


Figure 5: xinfAP by run (cat. A) - Lite

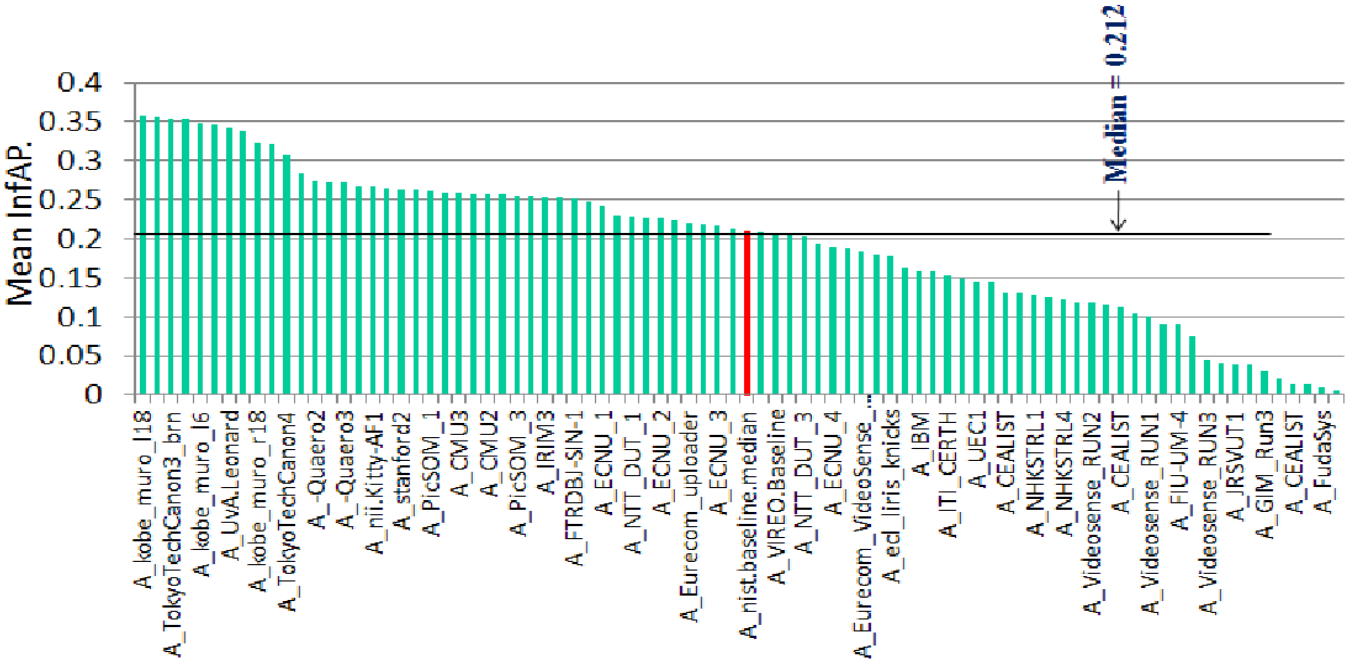


Figure 6: xinfAP by run (cat. D) - Lite

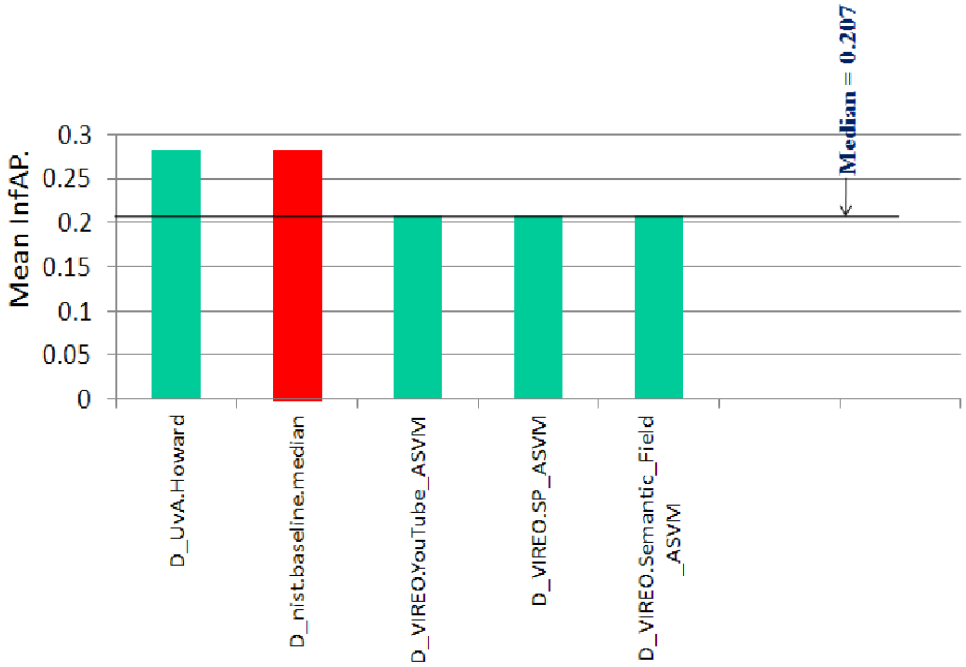


Figure 7: xinfAP by run (cat. F) - Lite

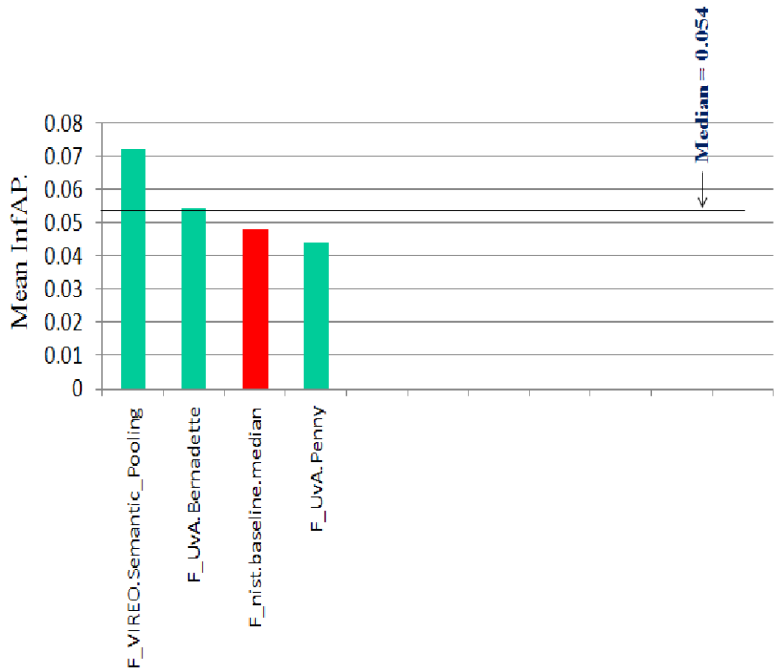


Figure 8: Top 10 runs (xinfAP) by feature - Full

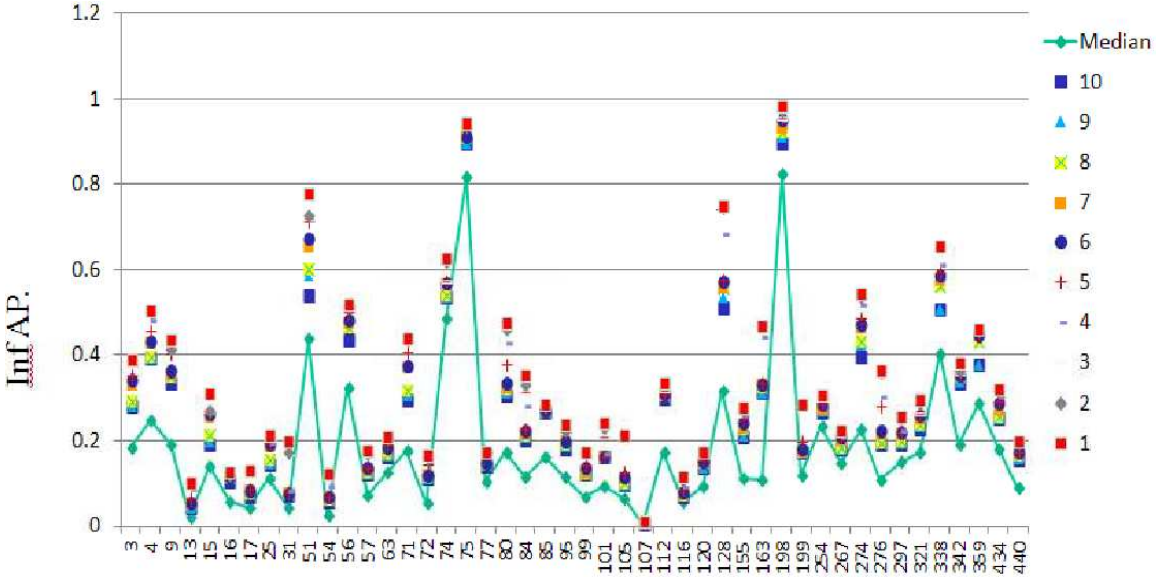


Figure 9: Top 10 runs (xinfAP) by feature - Full + Lite

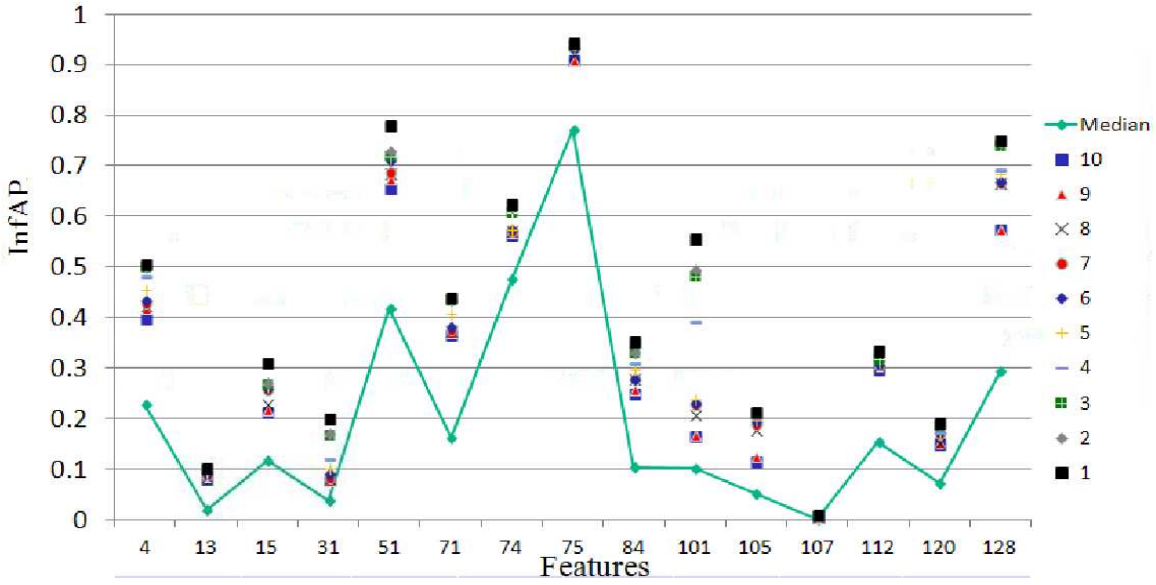


Figure 10: Significant differences among top A-category full runs

- A_TokyoTechCanon3_brn_3
  - A_UvA.Raj_2
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Sheldon_1
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Leonard_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_TokyoTechCanon4_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4

- A_TokyoTechCanon2_brn_2
  - A_UvA.Raj_2
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Sheldon_1
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Leonard_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_TokyoTechCanon4_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4

- A_TokyoTechCanon1_brn_1
  - A_UvA.Raj_2
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Sheldon_1
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Leonard_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_TokyoTechCanon4_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4

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

Run name	(mean infAP)
L_A_kobe_muro_l18_3	0.358
L_A_TokyoTechCanon1_brn_1	0.355
L_A_TokyoTechCanon3_brn_3	0.354
L_A_TokyoTechCanon2_brn_2	0.353
L_A_kobe_muro_l6_1	0.348
L_A_UvA.Sheldon_1	0.346
L_A_UvA.Leonard_4	0.342
L_A_UvA.Raj_2	0.338
L_A_kobe_muro_r18_2	0.323
L_A_kobe_muro_l5_4	0.320

A_kobe_muro_l18_3
- L_A_kobe_muro_l6_1
  - L_A_kobe_muro_l5_4
- L_A_kobe_muro_r18_2

Figure 12: Significant differences among top D-category lite runs

Run name	(mean infAP)
L_D_UvA.Howard_3	0.282
L_D_VIREO.Semantic_Field_ASVM_5	0.207
L_D_VIREO.SP_ASVM_4	0.207
L_D_VIREO.YouTube_ASVM_3	0.207

L_D_UvA.Howard_3
- L_D_VIREO.Semantic_Field_ASVM_5
- L_D_VIREO.SP_ASVM_4
- L_D_VIREO.YouTube_ASVM_3

Figure 13: Frequencies of shots with each feature for concept-pairs

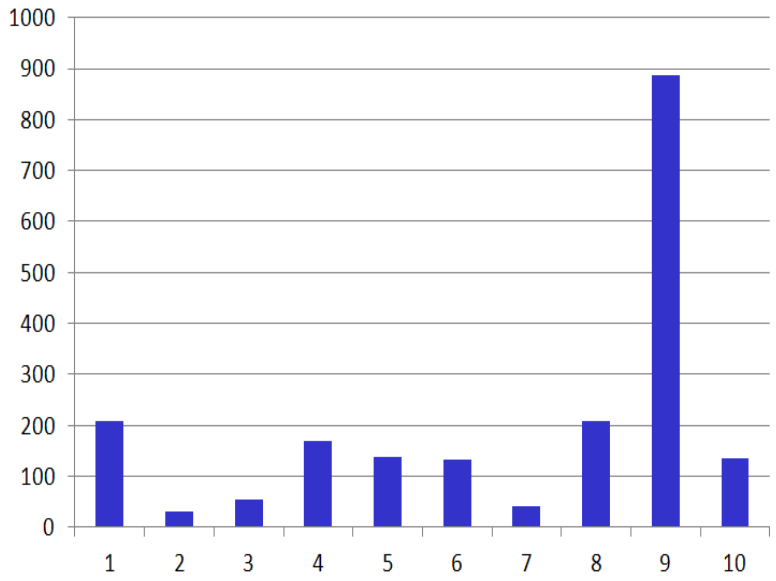


Figure 14: xinfAP by run (cat. A) - Full concept-pairs

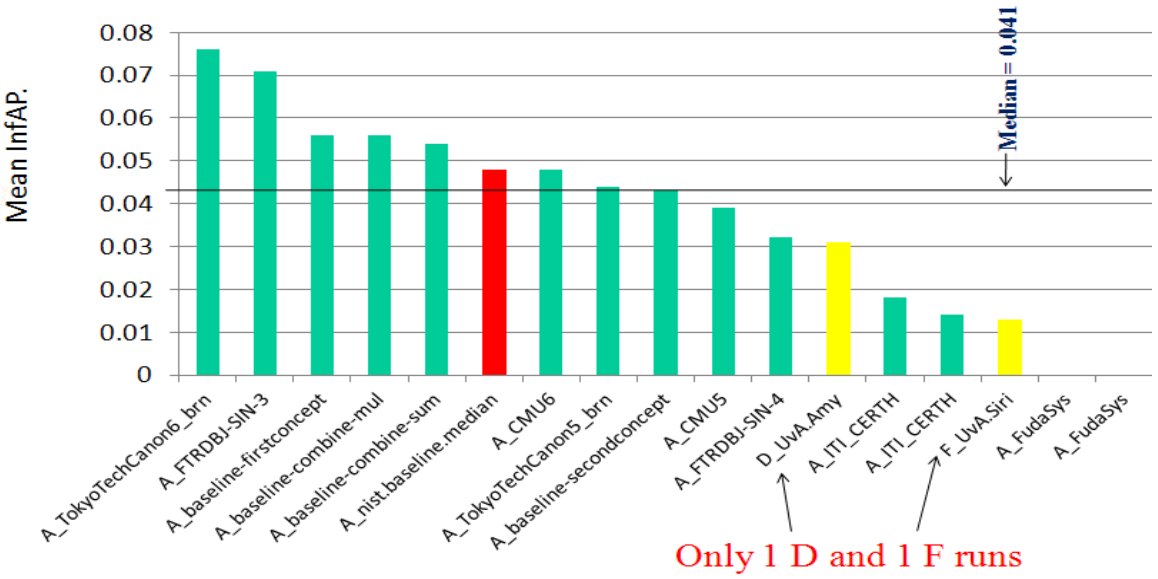


Figure 15: Significant differences among top A-category full concept-pairs runs

Run name	(mean infAP)
P_A_TokyoTechCanon6_brn_6	0.076
P_A_FTRDBJ-SIN-3_3	0.071
P_A_baseline-firstconcept_3	0.056
P_A_baseline-combine-mul_1	0.056
P_A_baseline-combine-sum_2	0.054
P_A_CMU6_1	0.048
P_A_TokyoTechCanon5_brn_5	0.044
P_A_baseline-secondconcept_4	0.043
P_A_CMU5_2	0.039
P_A_FTRDBJ-SIN-4_4	0.032

- A_TokyoTechCanon6_brn_6
  - A_CMU5_2
  - A_FTRDBJ-SIN-4_4
- A_TokyoTechCanon5_brn_5
  - A_FTRDBJ-SIN-4_4
- A_FTRDBJ-SIN-3_3
  - A_baseline-secondconcept_4

Figure 16: KIS: Mean inverted rank versus mean elapsed time for automatic runs

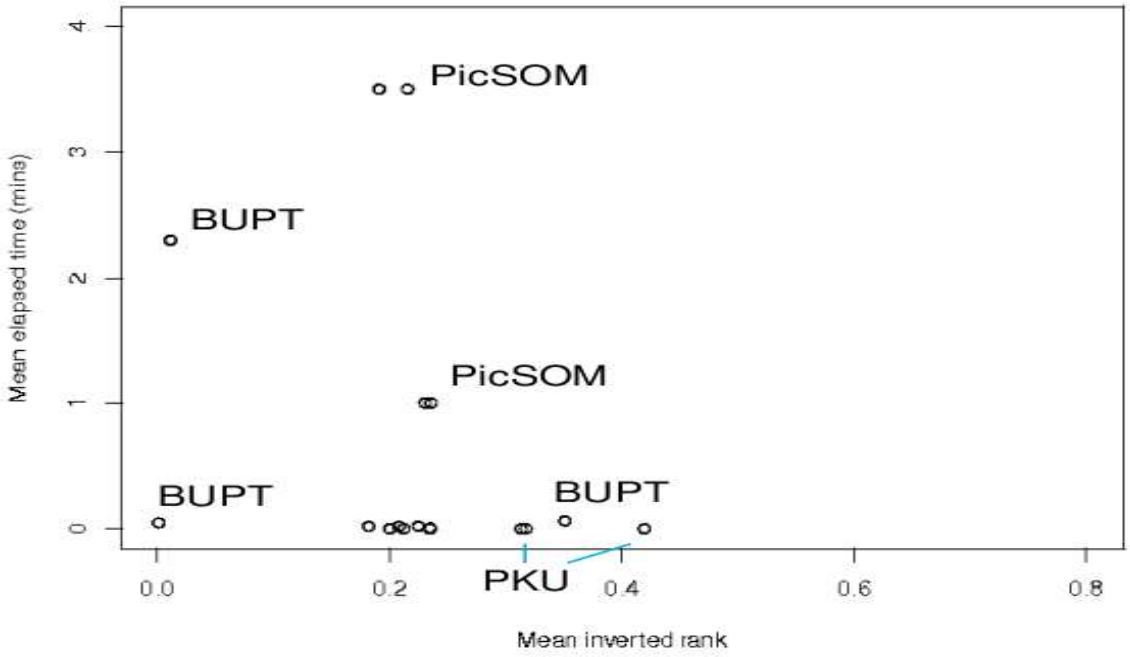


Figure 17: KIS: Mean inverted rank versus mean elapsed time for interactive runs

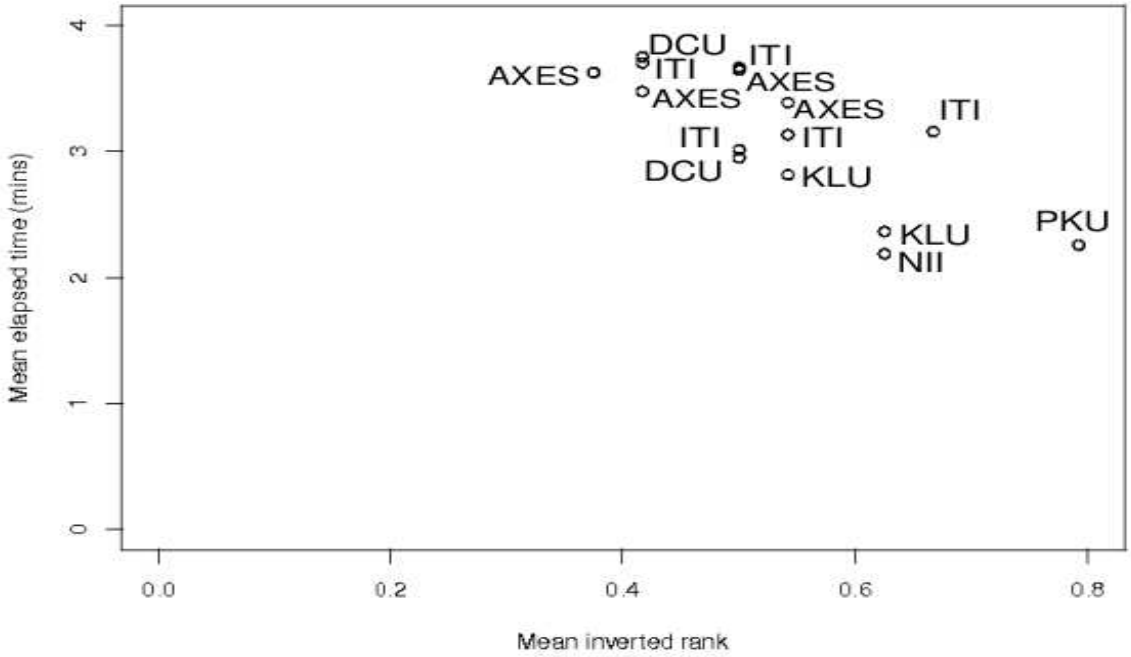


Figure 18: KIS: Oracle calls by topic and team

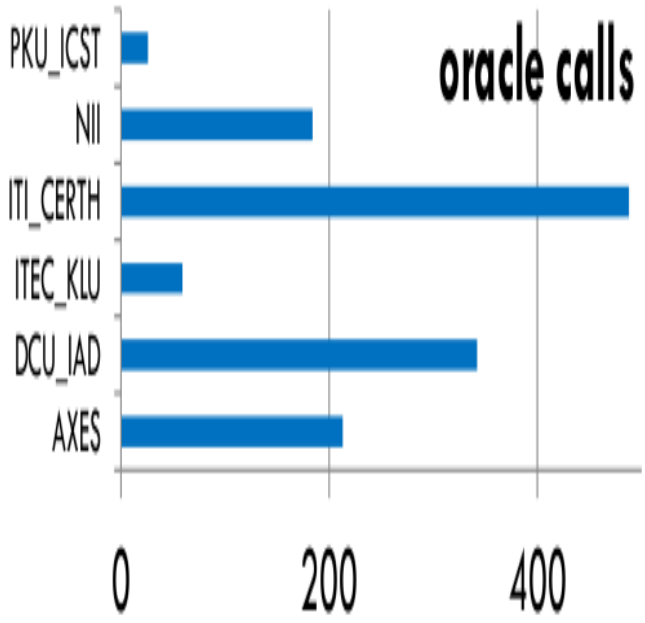


Figure 19: KIS: Runs finding known items

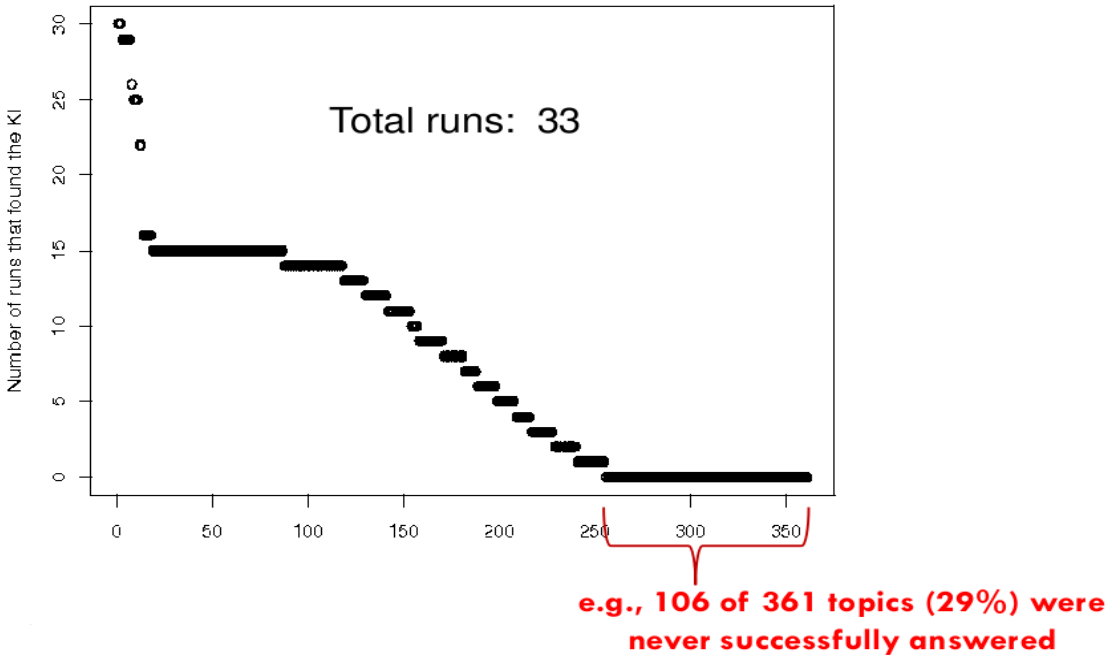
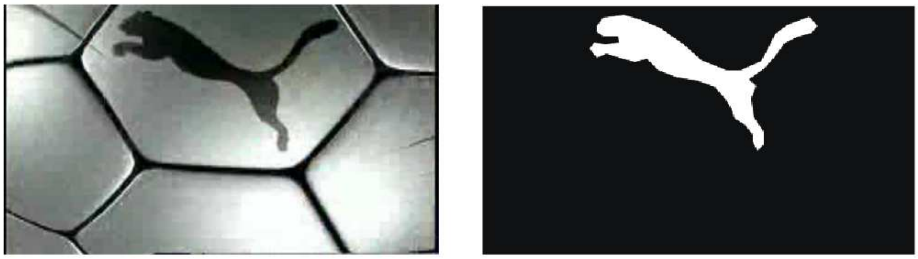


Figure 20: KIS: Known items never found

	Interactive		Automatic	
2012	2/24	17%	108/361	29%
2011	6/25	24%	142/391	36%
2010	5/24	21%	69/300	22%

Figure 21: INS: Example segmentations



Source **Mask**

Figure 22: INS: Example object targets 1/3

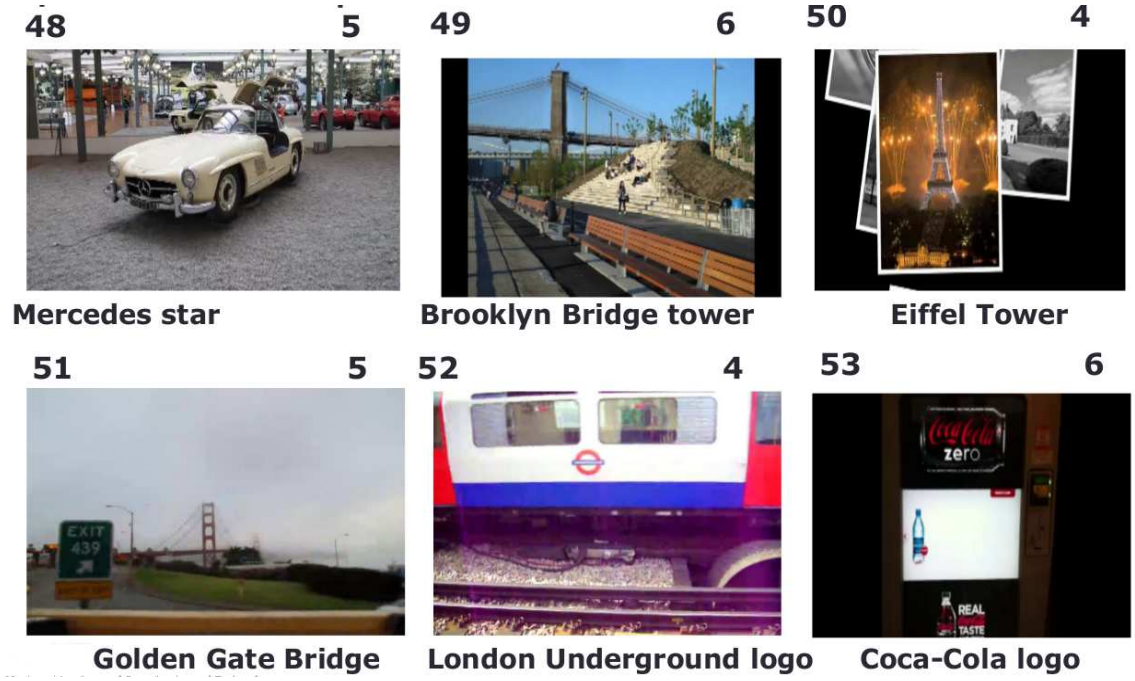


Figure 23: INS: Example object targets 2/3

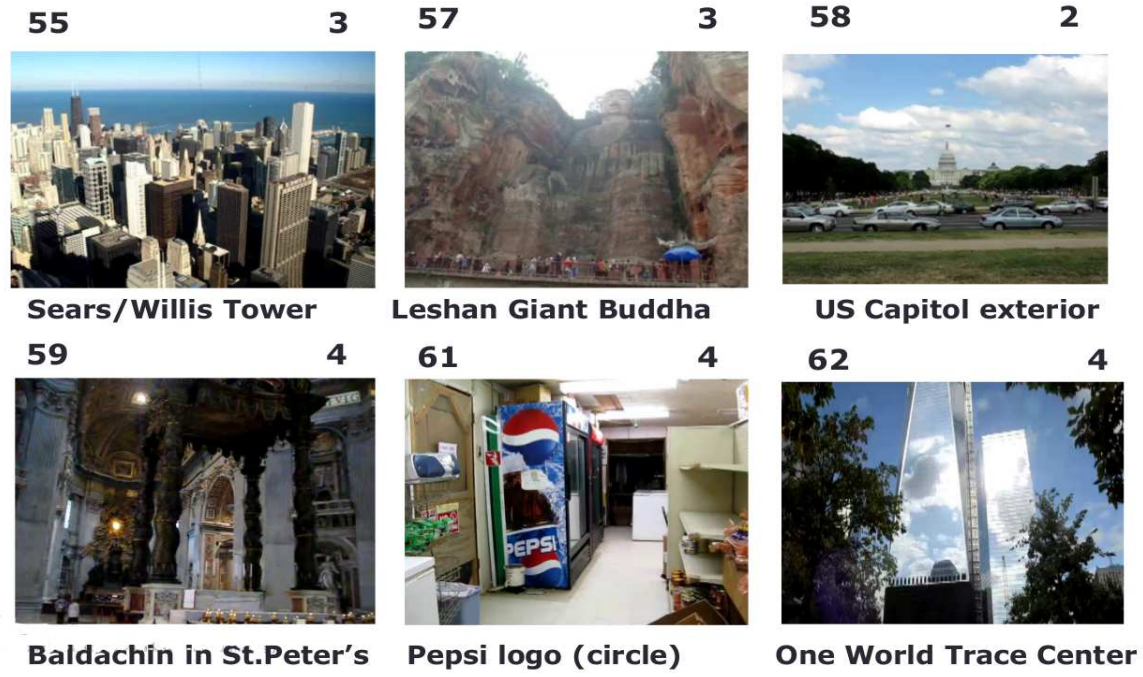


Figure 24: INS: Example object targets 3/3

Figure 25: INS: Example people target

Figure 26: INS: Example location targets

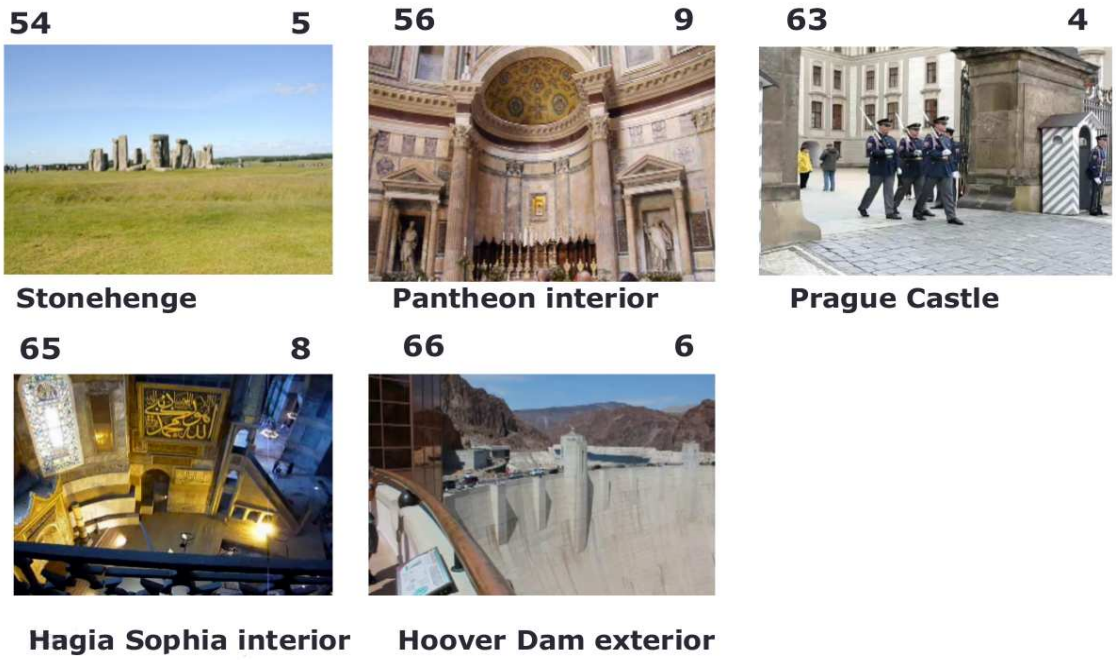


Figure 27: INS: Average precision for automatic runs by topic/type

.

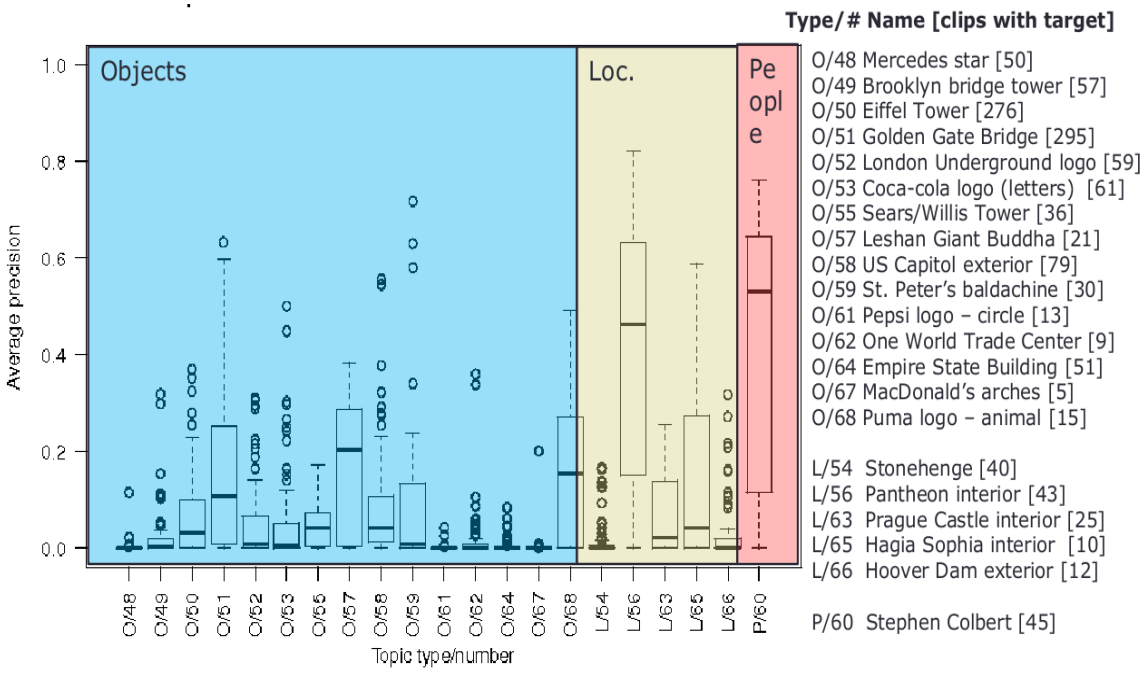


Figure 28: INS: Average precision for interactive runs by topic/type

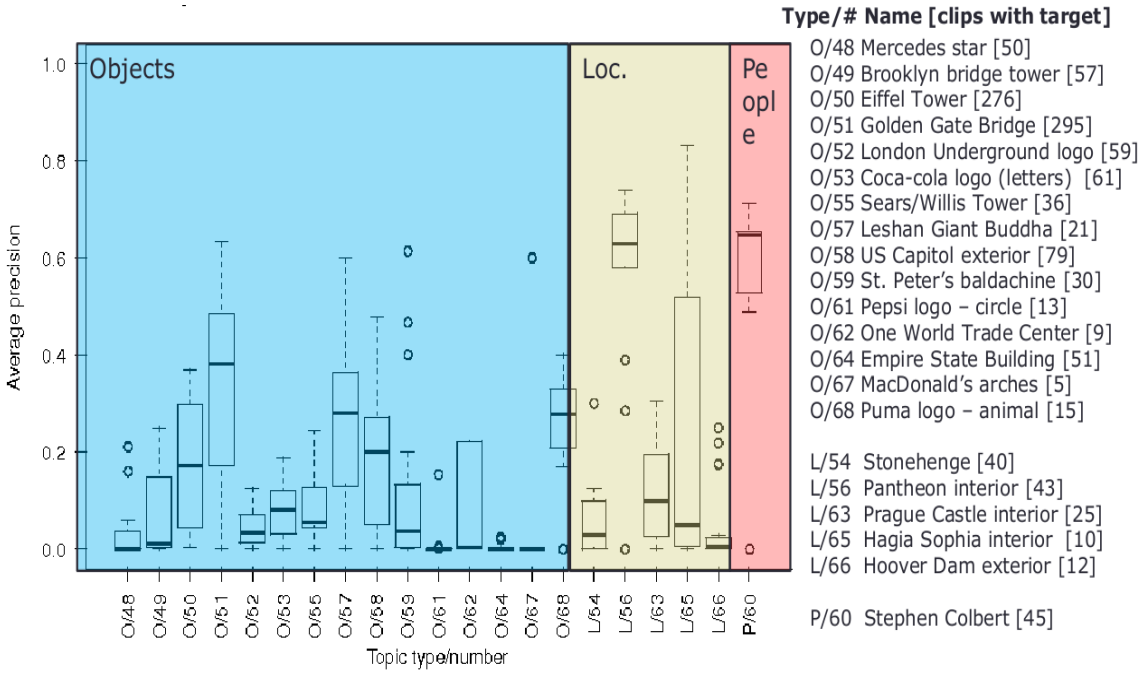


Figure 29: INS: AP by topic for top runs

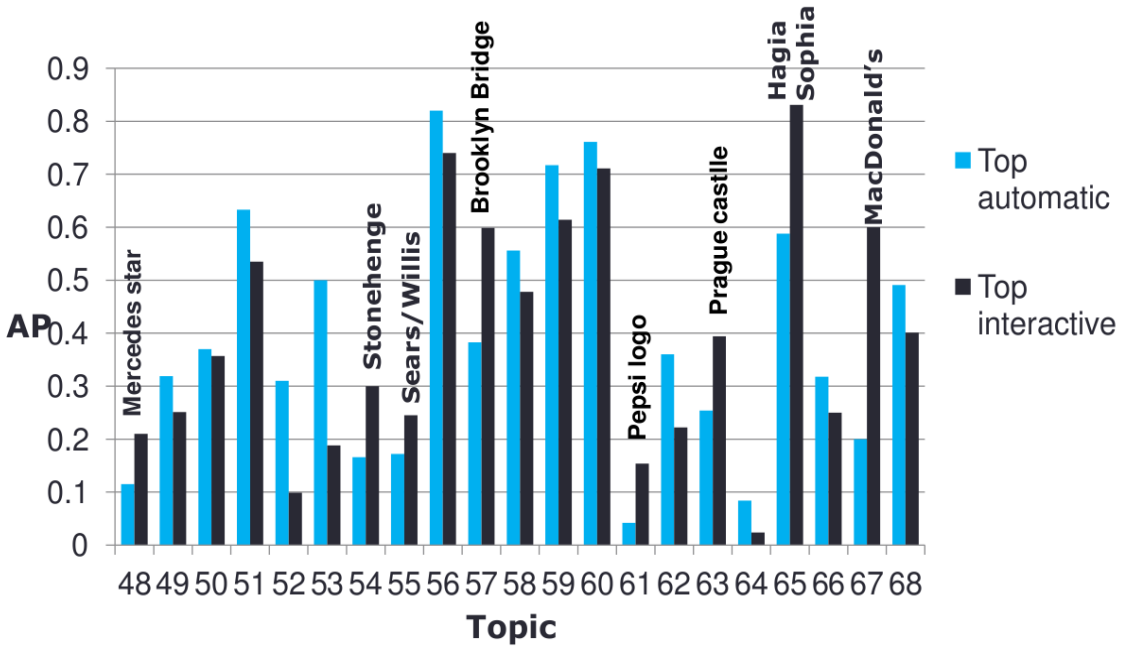


Figure 30: INS: Randomization testing for significant differences (interactive runs)

```
I X N PKU-ICST-MIPL 2
  I X N AXES_3 3

I X N AXES_4 4
I X N AXES_2 2
I X N AXES_1 1
I X N FTRDBJ 4
```

Figure 31: INS: Randomization testing for significant differences (automatic runs)

```
F X N BUPT.MCPRL 2
  F X N PKU-ICST-MIPL 4

F X N PKU-ICST-MIPL 1
  F X N PKU-ICST-MIPL 3
  F X N PKU-ICST-MIPL 4

F X N BUPT.MCPRL 3
  F X N PKU-ICST-MIPL 3
  F X N vireo_bl 4
  F X N vireo_dto 1
  F X N PKU-ICST-MIPL 4
  F X N JRSVUT2 1

F X N vireo_dtc 2
F X N vireo_dtcv 3
```

Figure 32: INS: MAP vs. number examples

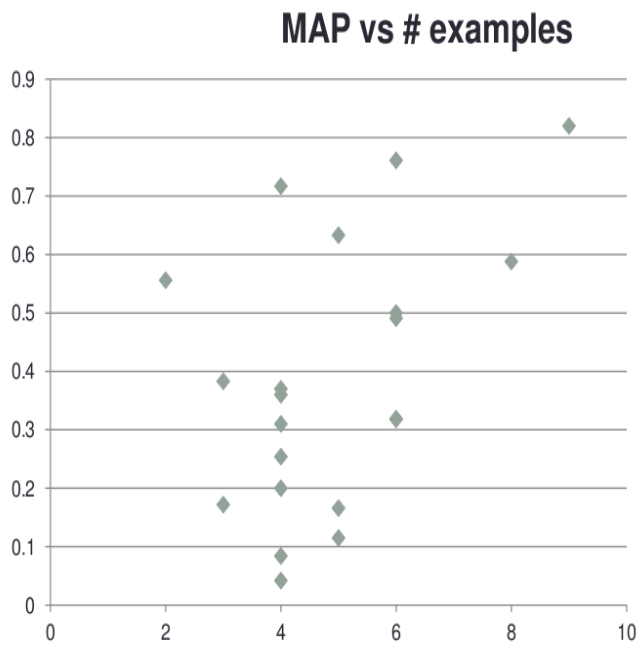


Figure 33: INS: MAP vs. elapsed time

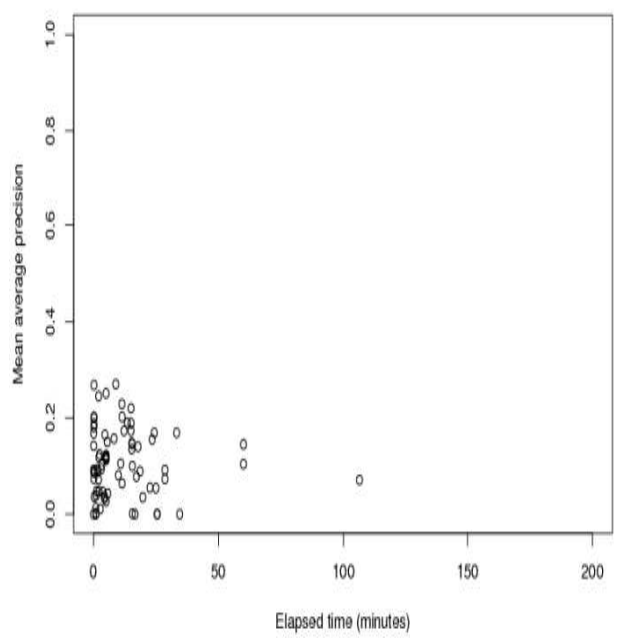


Figure 34: DET Curve visualization of Actual Decision P_{Miss} and P_{FA}s for Primary Pre-Specified Event Systems

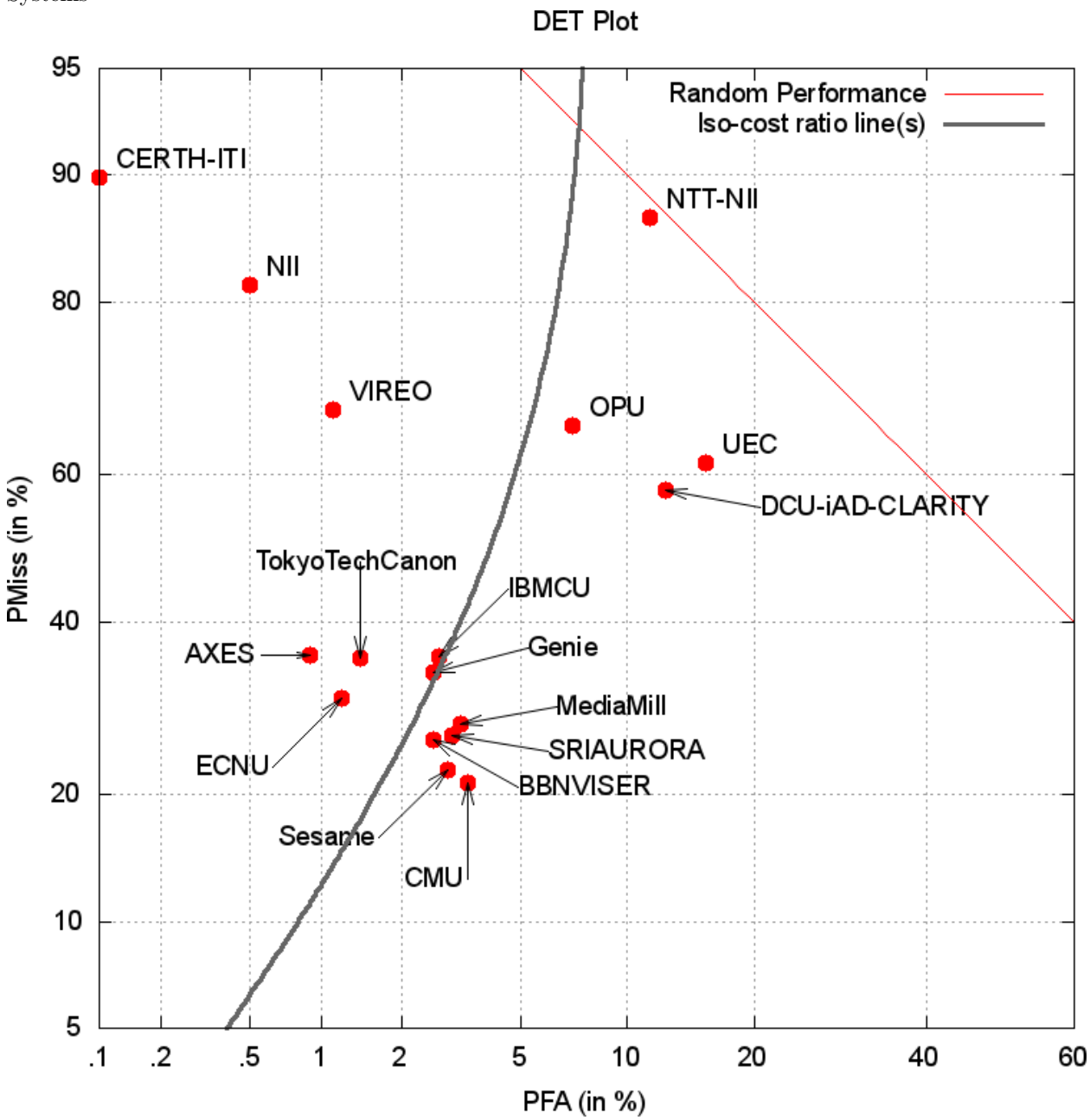


Figure 35: DET Curve visualization of Actual Decision P_{Miss} and P_{FA}s for Primary Ad-Hoc Event Systems

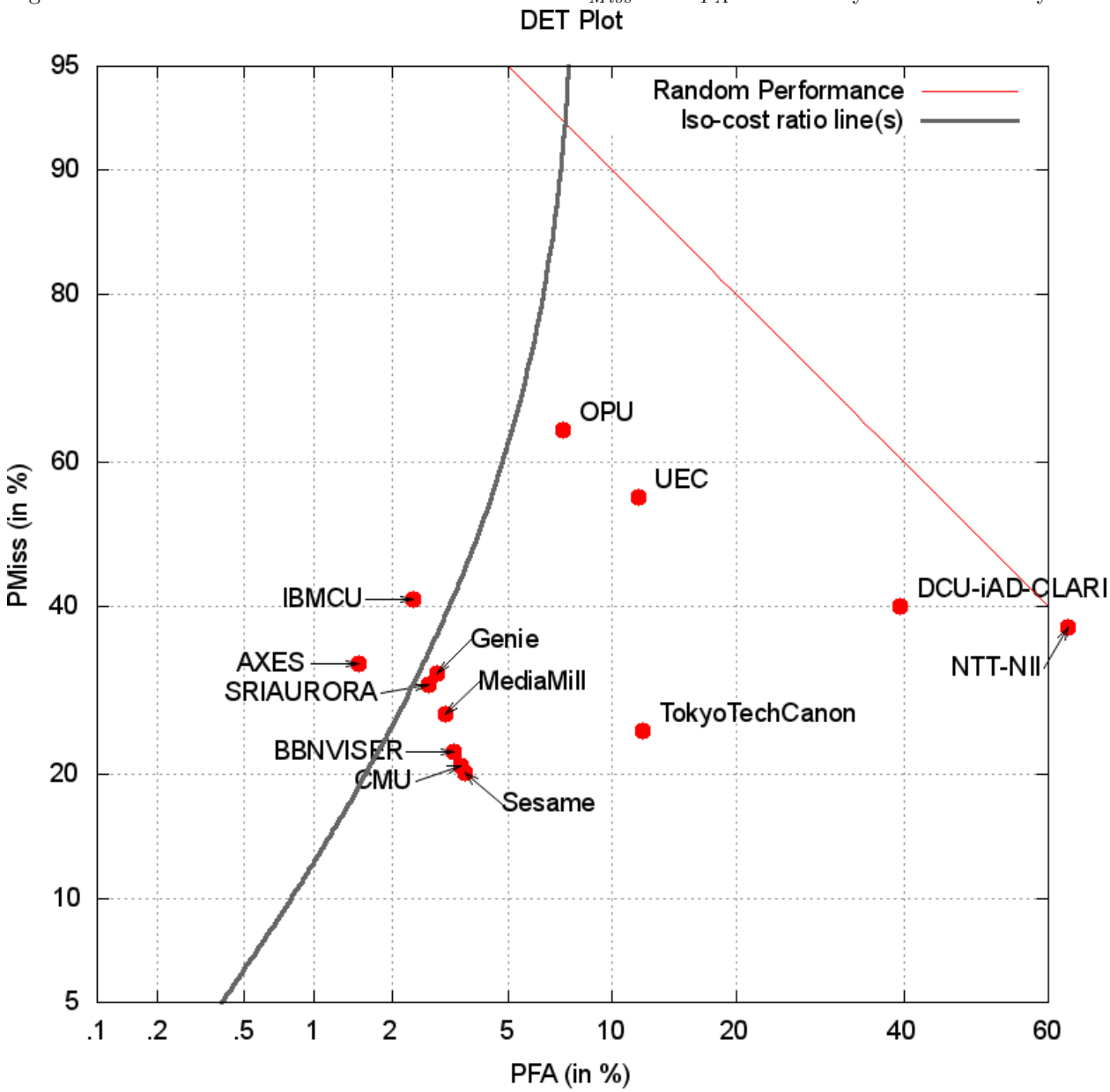


Figure 36: DET Curve visualization of Actual Decision P_{Miss} and P_{FA}s contrasting Full Event Kit and 10 Exemplar Pre-Specified Event Systems

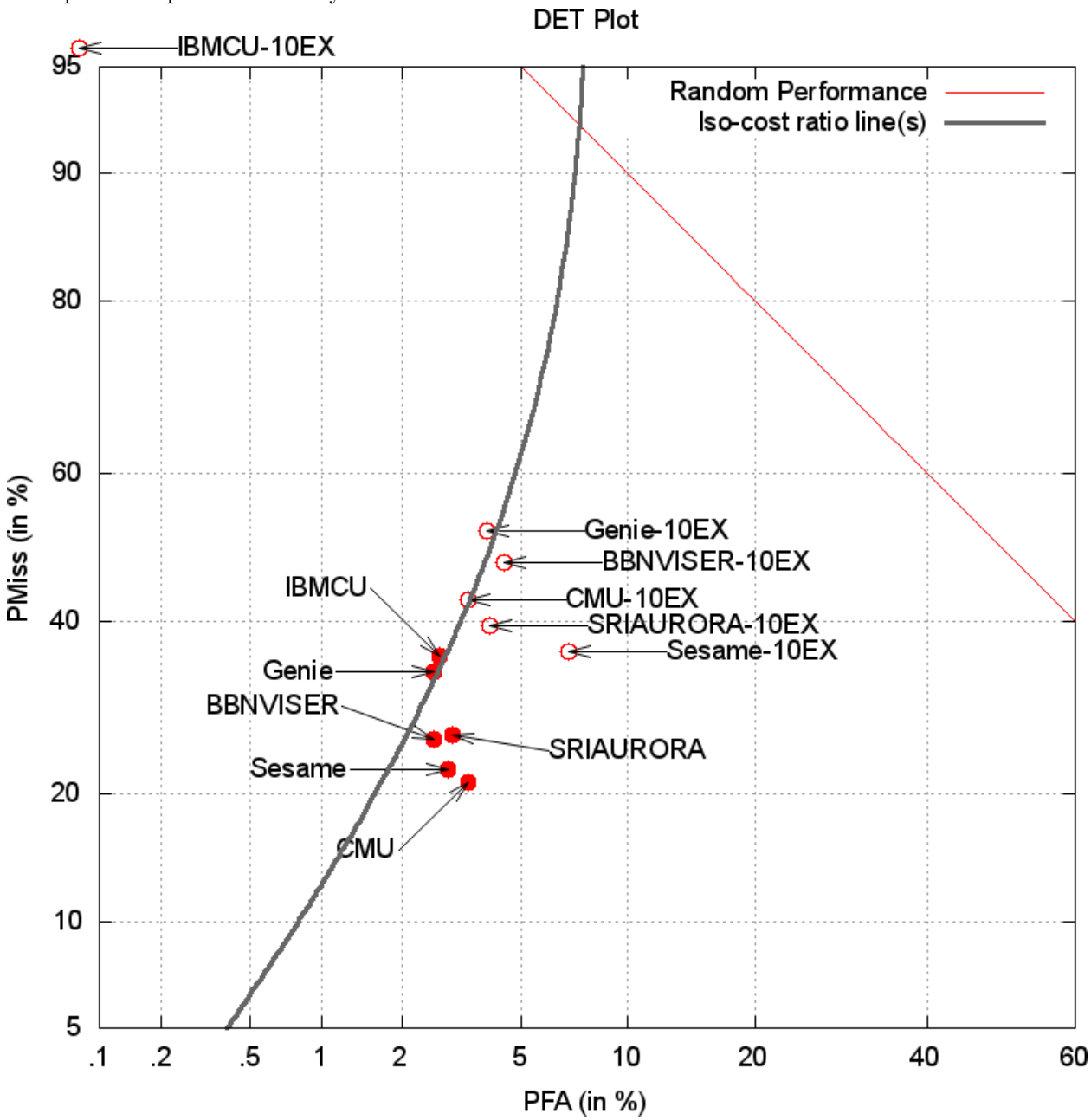


Figure 37: MER-to-Event: results by system

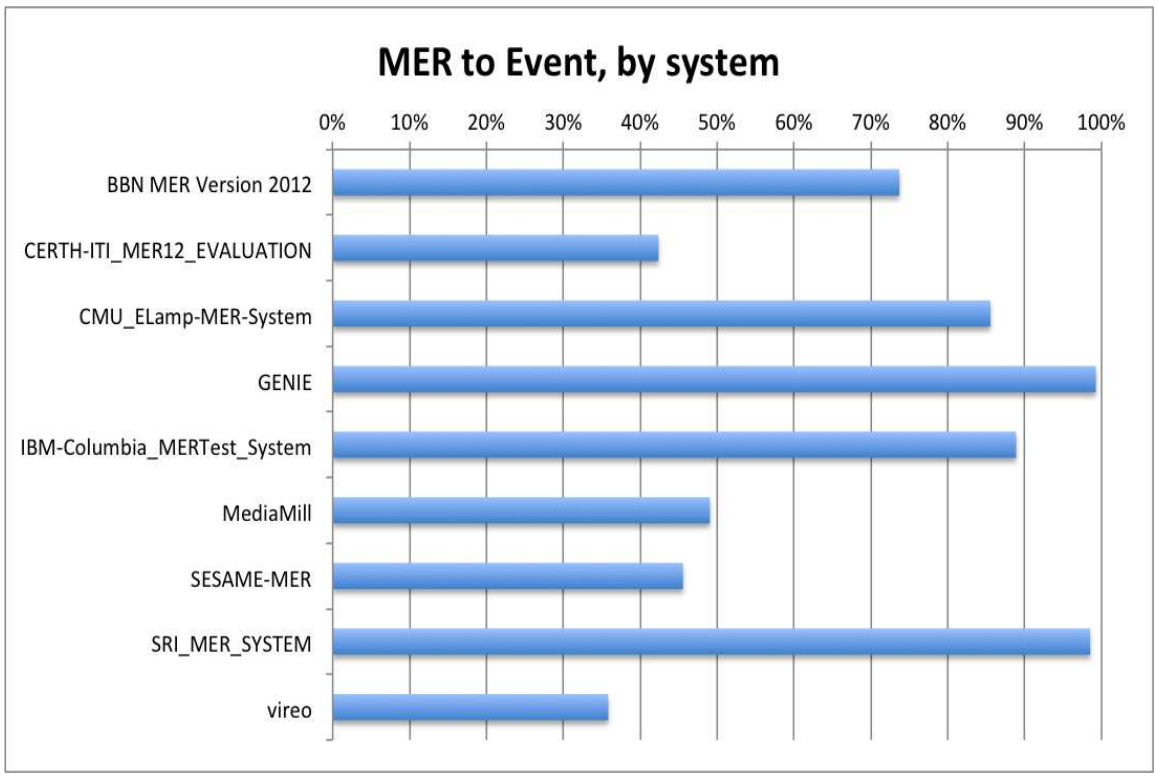


Figure 38: MER-to-Event: results by event

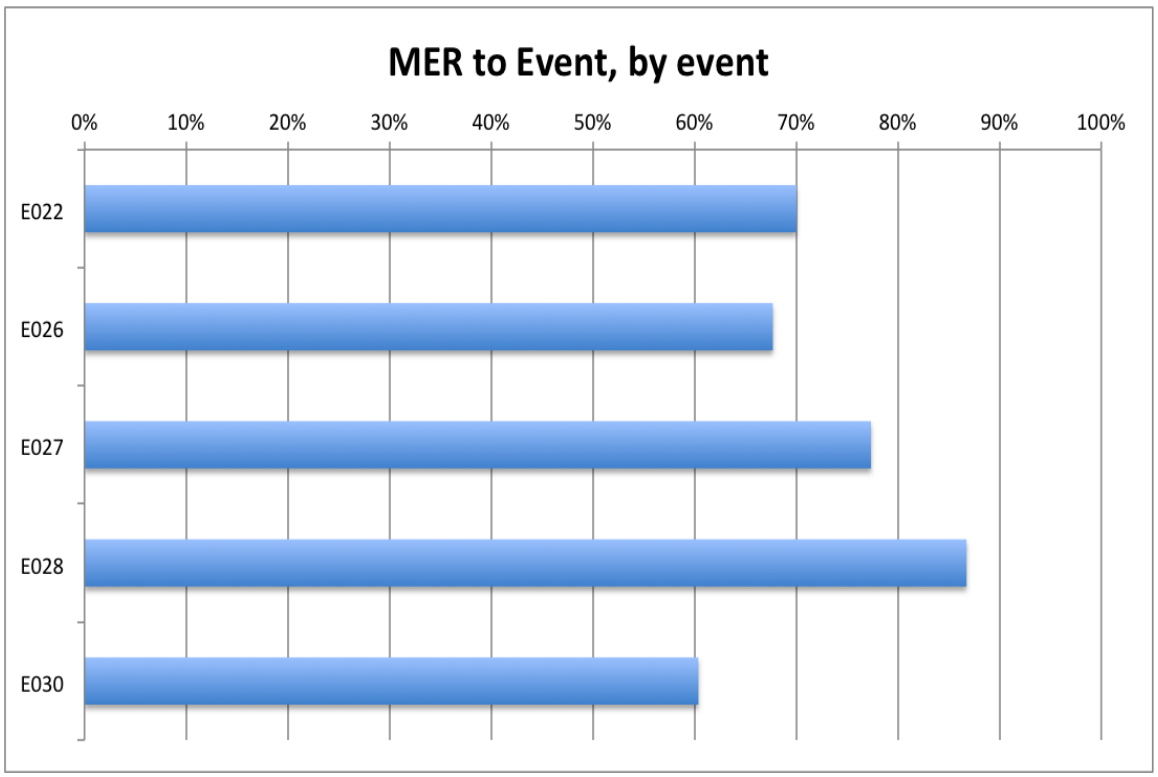


Figure 39: MER-to-Event: results by system, by event

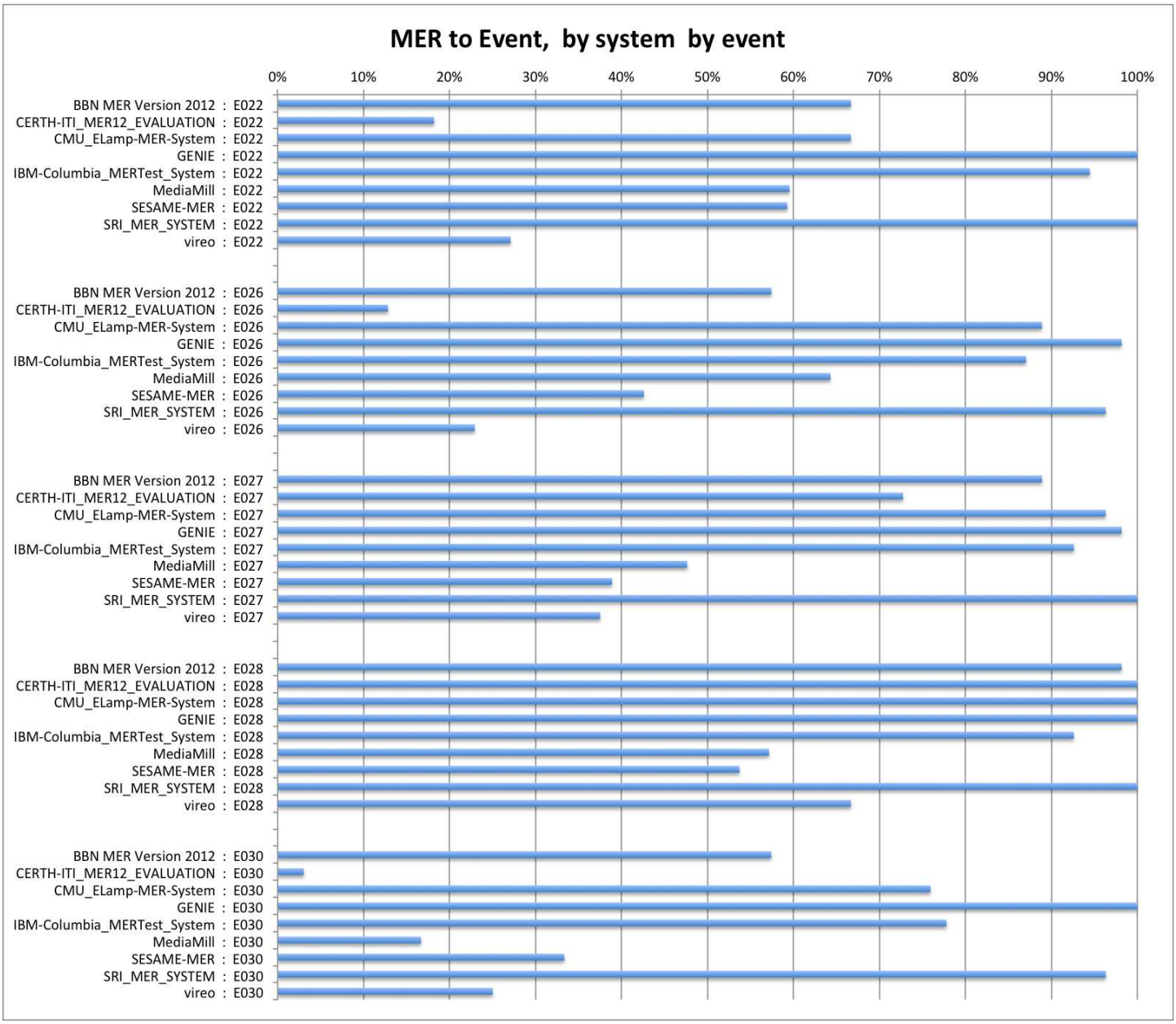


Figure 40: MER-to-Clip: results by system

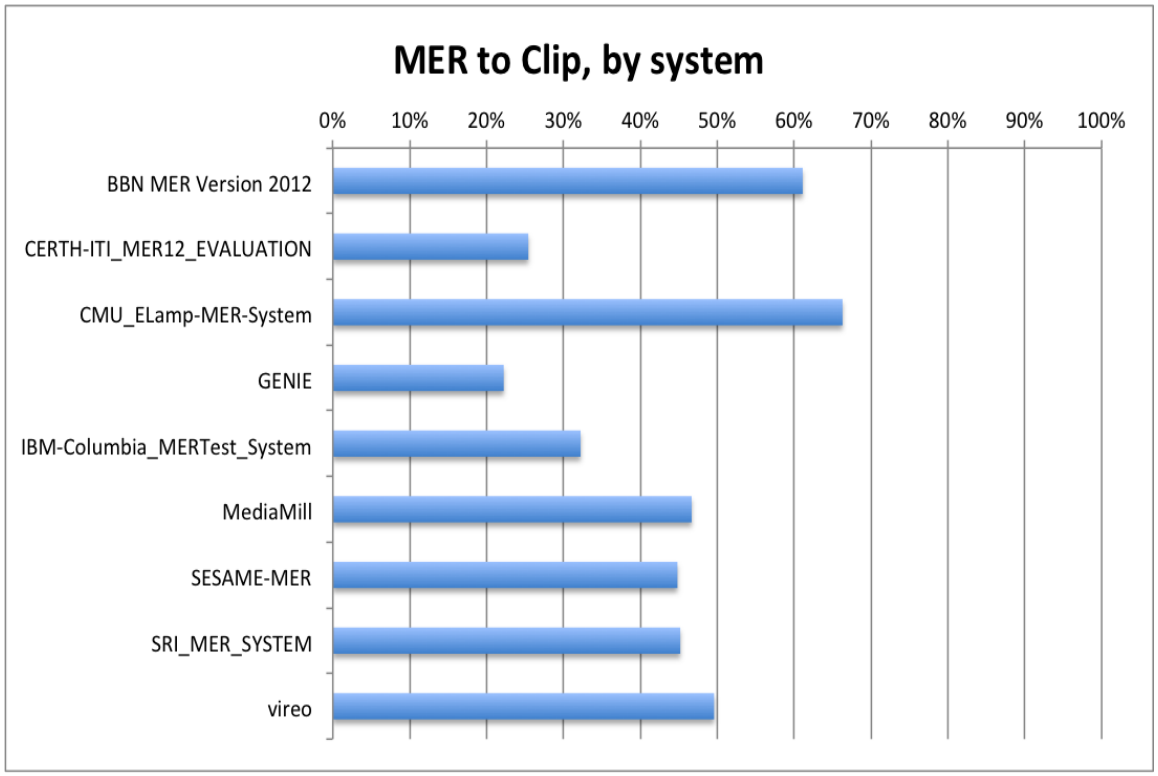


Figure 41: MER-to-Clip: results by event

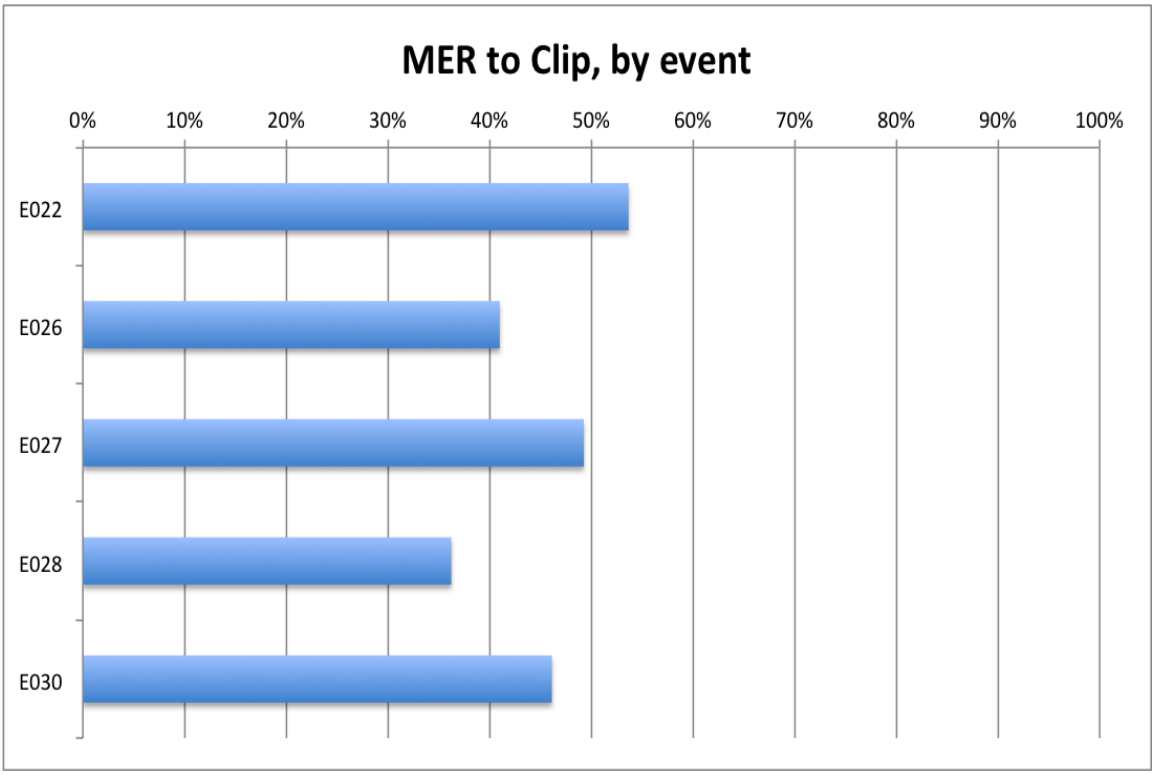


Figure 42: MER-to-Clip: results by system, by event

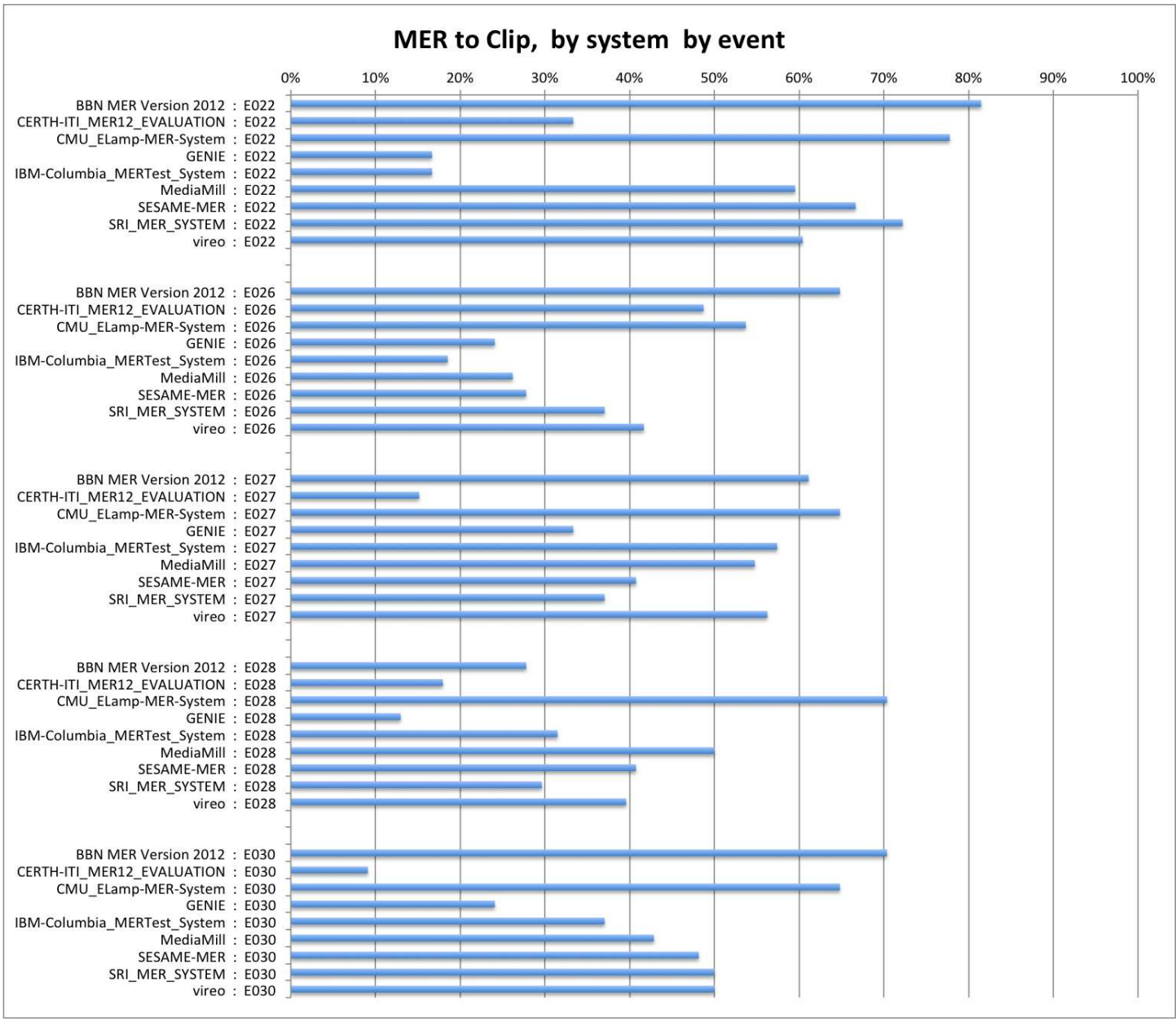


Figure 43: Camera views and coverage

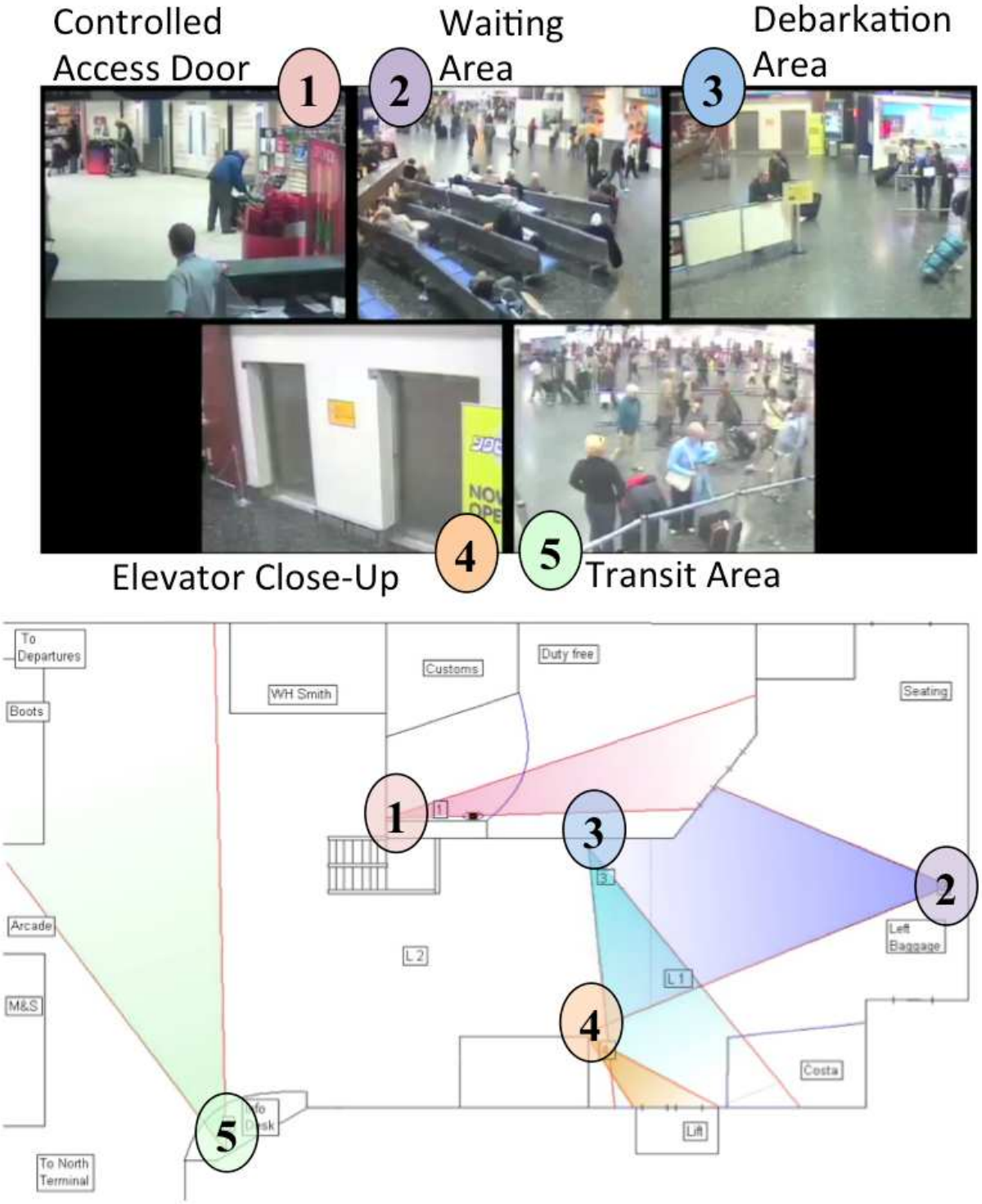


Figure 44: Event name, their rate of occurrences in Instances per Hour (IpH) / their average duration (in seconds) and Definition

Single Person events		
PersonRuns	7.02 / 2.80	Someone runs ← ***Lowest frequency***
Pointing	69.74 / 1.53	Someone points ← ***Highest frequency***
Single Person + Object events		
CellToEar	12.73 / 0.78	Someone puts a cell phone to his/her head or ear
ObjectPut	40.74 / 1.07	Someone drops or puts down an object
Multiple People events		
Embrace	11.48 / 6.13	Someone puts one or both arms at least part way around another person
PeopleMeet	29.46 / 4.89	One or more people walk up to one or more other people, stop, and some communication occurs
PeopleSplitUp	12.27 / 10.36	From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame

Figure 45: TRECVID 2012 SED Participants Chart

2012 SED Participants

(with number of systems per event)

		Single Person				Person + object				Multiple People					
		PersonRuns		Pointing		CellToEar		ObjectPut		Embrace		PeopleMeet		PeopleSplitUp	
		iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED
5 years in a row	Carnegie Mellon University & IBM [CMU-IBM]	5	7	5	7	5	6	5	7	5	7	5	7	5	7
4 years in a row	Multimedia Communication and Pattern Recognition Labs, Beijing University of Posts and Telecommunications [BUPT-MCPRL]	1	3	1	3			1	3	1	3	1	3	1	3
	Peking University, NEC Laboratories [PKUNEC]	3	1	3	1	3	1	3	1	3	1	3	1	3	1
3 years in a row	Beijing Jiaotong University [BJTU-SED]			1	1					1	1				
NEW	Brno University of Technology [BrnoUT]	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Dublin City University [dcu-savasa]	2	2	2	5			2	5						
	The City College of New York Media Team [MediaCCNY]		1		1		1		1		1		1		1
	Institute of Computer Science and Technology, Peking University [PKU-OS]	1	1	1	1			1	1						
	Queensland University of Technology [saivt]									1		1		1	
	Shanghai Jiaotong University, Center for Brain-like Computing and Machine Intelligence [SJTUBCMI]			1	1			1	1			1	1	1	1
	Video Computing Group, University of California Santa Barbara [UcsbUcrVcg]	1		1		1		1		1		1		1	
	University of Ottawa [VIVA-uOttawa]	1													
		15	16	16	21	10	9	15	20	13	14	13	14	13	14

Figure 46: Event-Averaged, Lowest NDCR by Site: iSED vs. rSED

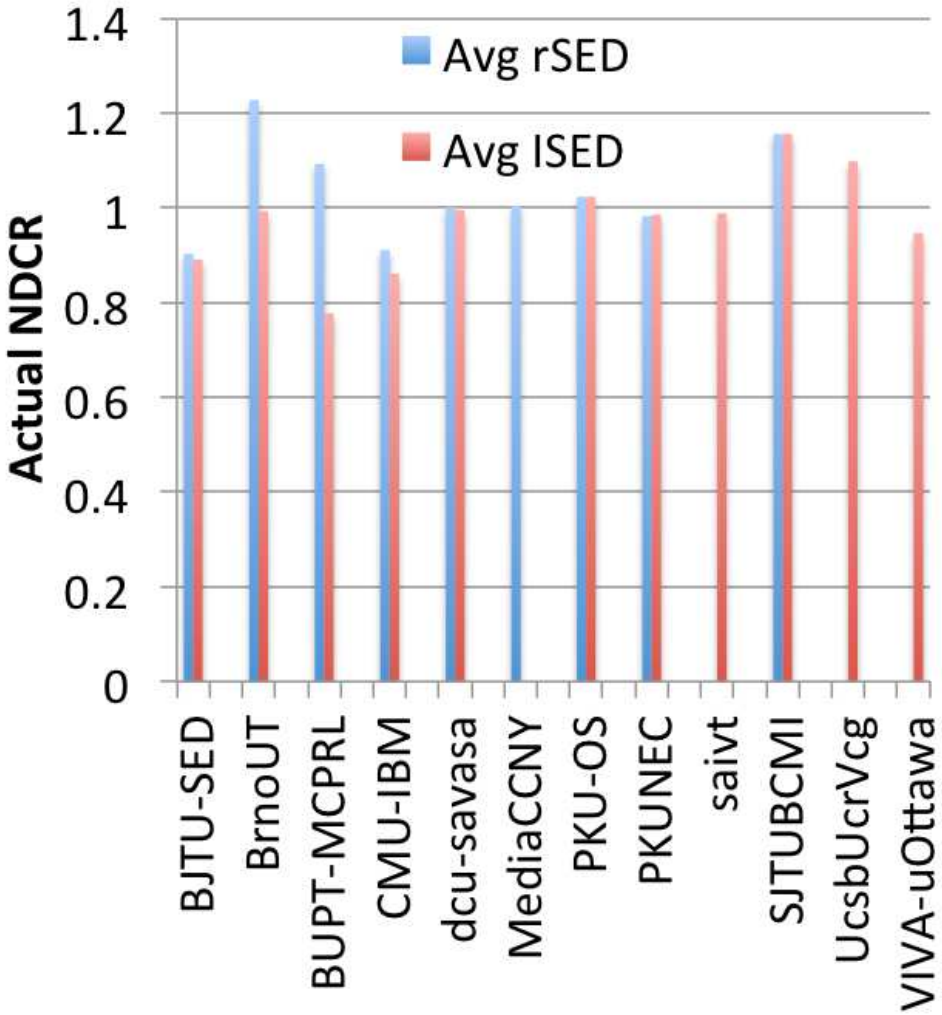


Figure 47: TV09-12 PeopleMeet

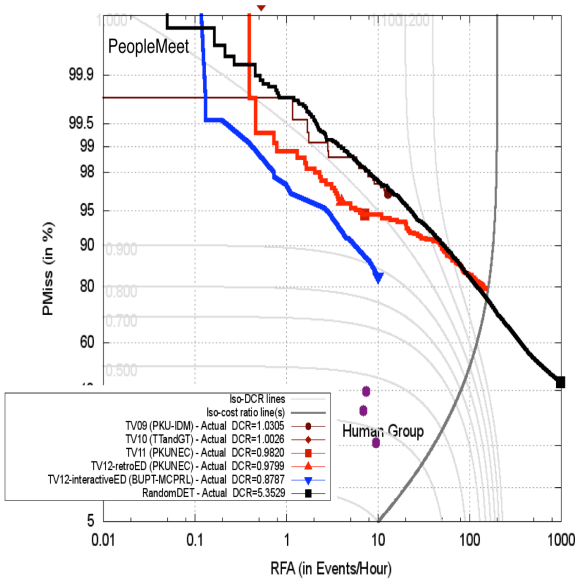


Figure 49: TV09-12 ObjectPut

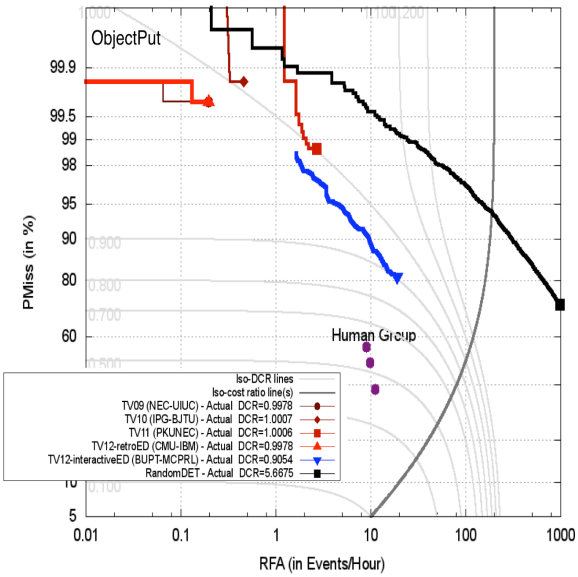


Figure 48: TV09-12 Embrace

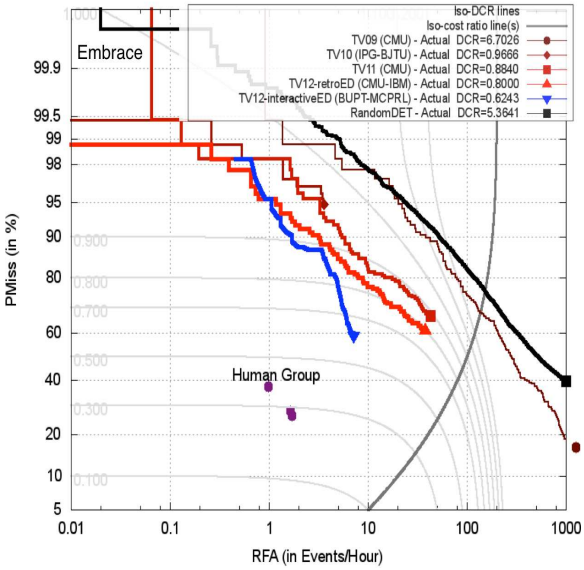


Figure 50: TV09-12 CellToEar

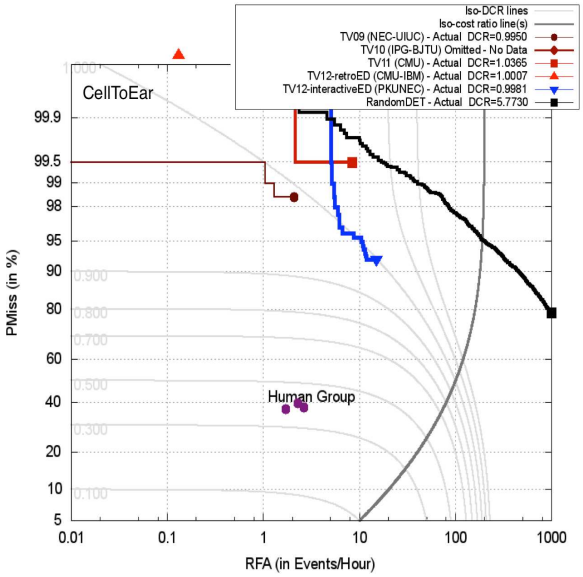


Figure 51: TV09-12 PersonRuns

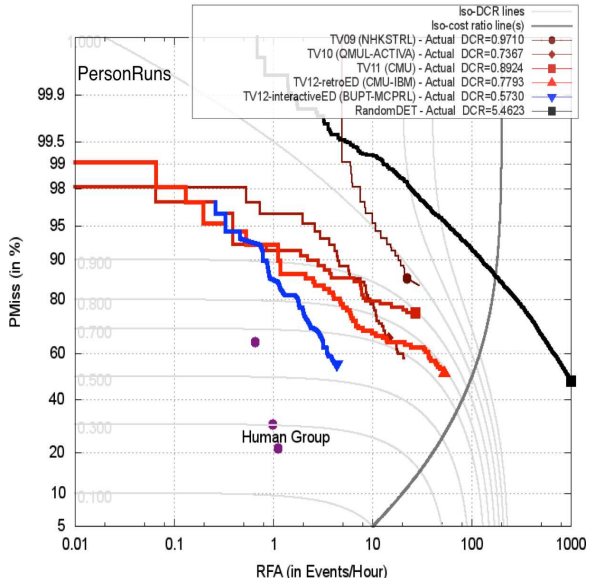


Figure 52: TV09-12 PeopleSplitUp

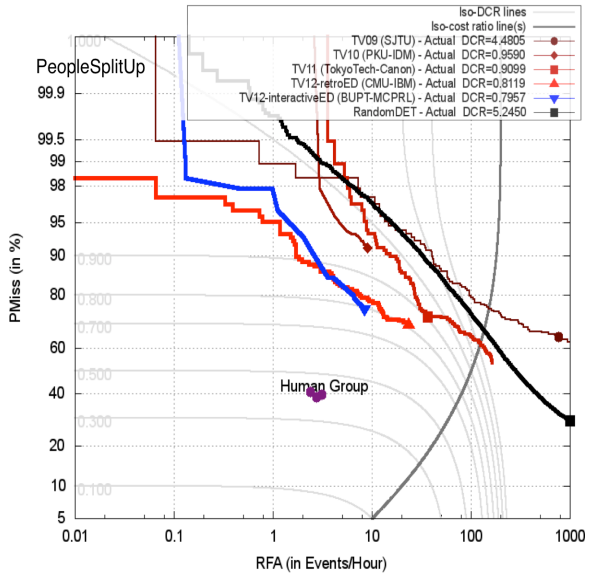
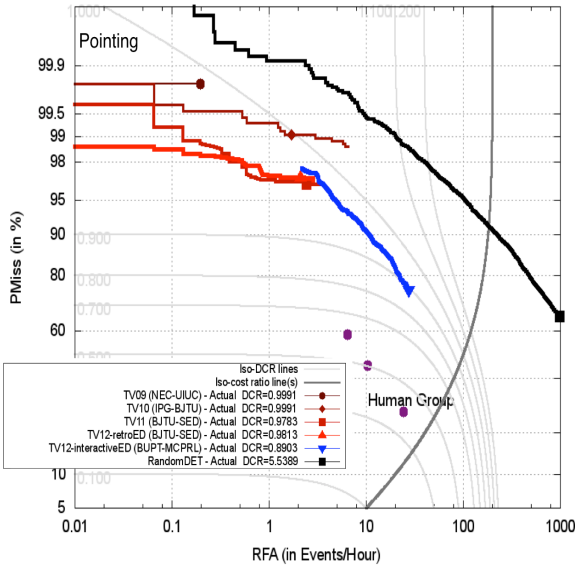



Figure 53: TV09-12 Pointing