HAL
open science

TRECVID 2011 - An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics

Paul Over, George Awad, Jon Fiscus, Brian Antonishek, Martial Michel, Alan F. Smeaton, Wessel Kraaij, Georges Quénot

To cite this version:

Paul Over, George Awad, Jon Fiscus, Brian Antonishek, Martial Michel, et al.. TRECVID 2011 - An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. TRECVID 2011 - TREC Video Retrieval Evaluation Online, Dec 2011, Gaithersburg, MD, United States. hal-00763912

HAL Id: hal-00763912

https://hal.science/hal-00763912v1

Submitted on 11 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics

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

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

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

Wessel Kraaij {wessel.kraaij@tno.nl}
TNO
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

May 8, 2012

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2011 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 National Institute of Standards and Technology (NIST) and other US government agencies. Many organizations and individuals worldwide contribute significant time and effort.

In 2010, TRECVID turned to new and different data and to some new tasks. TRECVID 2011 represented a continuation of the six tasks from 2010 with some new data. 60 teams (see Table 1) from various research organizations — 25 from Asia, 18 from Europe, 12 from North America, 2 from South America, and 3 from Australia — completed one or more of six

tasks:

- content-based copy detection
- instance search
- known-item search
- semantic indexing
- surveillance event detection
- multimedia event detection

200 hours of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC), were used for semantic indexing, known-item search, and copy detection. Unlike previously used professionally edited broadcast news and educational programming, the IACC videos reflect a wide variety of content, style, and source device - determined only by the self-selected donors. About 81 hours of BBC rushes video was reused for the instance search pilot. 45 hours of airport surveillance video was reused for the surveillance event detection task. About 1400 hours from a new collection of Internet videos - Heterogeneous Audio Visual Internet (HAVIC) - was used for development and testing in the multimedia event detection task.

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

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

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

2 Data

2.1 Video

BBC rushes video

Eighty-one hours of unedited (rushes) video provided by the BBC Archive - mostly footage for travel programming - was used for the instance search task.

Internet Archive Creative Commons (IACC) video

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

As in 2010, LIMSI and VecSys research provided automatic speech recognition for the English speech in the IACC video (Gauvain, Lamel, & Adda, 2002).

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 & Quenot, 2008).

Sound and Vision data

In 2006 the Netherlands Institute for Sound and Vision generously provided 400 hours of Dutch TV news magazine, science news, news reports, documentaries, educational programming, and archival video in MPEG-1 format for use within TRECVID. About 180 hours of Sound and Vision video, previously used for testing feature extraction and ad hoc search, were reused in 2010 for testing instance search. This data was available in 2011 for system development

The video had already been automatically divided into shots by Christian Petersohn at the Fraunhofer

(Heinrich Hertz) Institute in Berlin. These shots served as the predefined units of evaluation.

Roeland Ordelman and Marijn Huijbregts at the University of Twente had provided the output of an automatic speech recognition system run on the Sound and Vision data. Christof Monz of Queen Mary, University London had contributed machine translation (Dutch to English) for the Sound and Vision video based on the University of Twente's automatic speech recognition (ASR). The LIMSI Spoken Language Processing Group had produced a speech transcription for the TRECVID 2007-2009 Sound and Vision data using its recently developed Dutch recognizer.

iLIDS Multiple Camera Tracking Data

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

The training video consisted of the ≈100 hours of data used for SED 2008 evaluation. The evaluation video consisted of the same additional ≈50 hours of data from Imagery Library for Intelligent Detection System's (iLIDS) multiple camera tracking scenario data used for the 2009 and 2010 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 (HAVIC) Corpus

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-4's Advanced Audio Coding (ACC) (ACC, 2010) encoded audio.

The training material consisted of: the ≈438 hours of HAVIC material (114 hours used for the MED 2010 pilot evaluation and 324 additional hours of data) and fifteen new event kits - ten of which were used for the formal evaluation.

The evaluation corpus consisted of an additional ≈991 hours of HAVIC video : 17 times more material than MED '10.

3 Semantic indexing

A potentially important asset to help video search/navigation is the ability to automatically identify the occurrence of various semantic features/concepts such as "Indoor/Outdoor", "People", "Speech" etc., which occur frequently in video information. The ability to detect features is an interesting challenge by itself but takes on added importance to the extent it can serve as a reusable, extensible basis for query formation and search. The semantic indexing task was a follow-on to the feature extraction task. It was coordinated by NIST and by Georges Quénot under the Quaero program and had the following additional, new objectives:

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

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

346 concepts were 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.

Two types of submissions are considered: full submissions in which participants submit results for all 346 concepts and lite submissions in which participants submit results for only 50 concepts. TRECVID evaluated 50 concepts - 20 based on judgments done at NIST and 30 done under the Quaero program in France. The 50 concepts evaluated for 2011 were as follows with their feature numbers in brackets. Those marked with an asterisk formed a "lite" subset to which some participants restricted their experiments.

[2] *Adult, [5] Anchorperson, [10] Beach, [21] *Car, [26] Charts, [27] *Cheering, [38] *Dancing, [41] *Demonstration_Or_Protest, [44] *Doorway, [49] *Explosion_Fire, [50] Face, [51] *Female_Person, [52] *Female-Human-Face-Closeup, [53] *Flowers, [59] *Hand, [67] *Indoor, [75] *Male_Person, [81] *Mountain, [83] *News_Studio, [84] *Nighttime, [86] *Old_People, [88] Overlaid_Text, [89] People_Marching, [97] Reporters, [100] *Running, [101] *Scene_Text, [105] *Singing, [107] *Sitting_Down, [108] Sky, [111] Sports, [113] Streets, [123] Two_People, [127] *Walking, [128] *Walking_Running, [227] Door_Opening, [241] Event, [251] Female_Human_Face, [261] Flags, [292] Head_And_Shoulder, [332] Male_Human_Face, [354] News, [392] Quadruped, [431] Skating, [442] Speaking, [443] Speaking_To_Camera, [454] Studio_With_Anchorperson, [464] Table, [470] Text, [478] Traffic, [484] Urban_Scenes

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 Guidelines webpage: `www-nlpir.nist.gov/projects/tv2011/tv2011.html\#sin`.

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, which permits sampling density to vary (Yilmaz, Kanoulas, & Aslam, 2008). This allowed the evaluation to be more sensitive to shots returned below the lowest rank ($\approx$ 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.B collection was used for testing. It contained 137 327 shots.

3.2 Evaluation

Each group was allowed to submit up to 4 runs and in fact 28 groups submitted 68 full runs and 34 lite runs.

For each concept, pools were created and randomly sampled as follows. The top pool sampled 100 % of shots ranked 1 to 100 across all submissions. The bottom pool sampled 8.3 % of ranked 101 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, 268 156 were judged. 1 519 419 shots fell into the unjudged part of the overall samples. All full runs were also treated as lite runs by looking at their performance on just the lite concept subset.

3.3 Measures

The *sample_eval* software, a tool implementing xinfAP, was used to calculate inferred recall, inferred precision, inferred average precision, etc., for each result, given the sampling plan and a submitted run. Since all runs provided results for all evaluated concepts, runs can be compared in terms of the mean inferred average precision across all 50 (or 23 lite) evaluated concepts. The results also provide some information about "within concept" performance.

3.4 Results

Readers should see the results section at the back of the notebook for details about the performance of each run.

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 5 % TPs from the total tested shots percentage. Features "Adult", "indoor", "male-person" and "text" had TPs in over 10 % of the

test shots. On the other hand, features that had the fewest TPs (less than 0.3 %) were "charts", "people-marching", "sitting-down", and "door-opening". It is worth mentioning that the 15 common features from 2010 did not perform well and most of them are far below the 5 % TPs line. The top performing features were more generic by definition than the bottom performing ones which are very specific like "sitting-down" or "people-marching".

Figures 2, 3, and 4 show the results of category A,B and D for full runs. Category A runs used only IACC training data. Catetory B runs used only non-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 random baseline result (as described below) for category A only. All of the runs performed better than the random run which is an improvement compared to TV10 where a small number of runs performed worse than the randomly generated result. Still category A runs are the most popular type and achieve top recorded performances. The random baseline run was done by choosing 10 000 random permutations of all the shot ids and the top 2 000 were selected. The sample_eval program was then applied with the reference qrels file and the obtained xinfAPs were averaged on the 10 000 generated submissions.

Figures 5, 6, and 7 show the results of category A,B, and D for the lite runs respectively together with their median values. As in full runs, category A of lite runs were the best performing in general. Only 1 run from Category C was submitted and achieved a score of 0.01 in full evaluation while 0.017 in lite evaluation.

Figure 8 shows the performance of the top 10 teams across the 50 features. The behavior varied generally across features. For example, some features reflected a large spread between the scores of the top 10 such as feature "Beach"(3), "Charts"(5), "News"(41), "Singing"(27) and "Skating"(43). This indicates that there is still room for further improvement within used techniques, while other features had a tight spread of scores among the top 10 such as feature "Adult"(1),"Face"(11), "indoor"(16),"sitting-down"(28),"two_people"(32), "walking"(33),"walking-running"(34),"door-opening"(35), "head & shoulders"(39), "speaking"(44),"speaking-to-camera"(45), and "text"(48) which may indicate a small variation in used techniques performance. In general, the median scores ranged between 0.002 (feature "Sitting_down") and 0.441 (feature "studio-with-anchorperson"). As a general observation, feature "Sitting_down" had the minimum median score at TRECVID 2010 as well which demonstrates how difficult this feature is for the systems to detect.

The analogous graph for the 23 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 "Explosion", "Female-face-closeup", "Mountain", "Running", "Demonstration_or_protest", and "singing". While the features with tight spread was "Adult", "indoor", "news-studio", and "male-person". As a general observation, still there is an overlap between TV10 largest spread features performance and this year such as "Singing" and "Demonstration_or_protest" which indicates that the same features are probably still hard for most systems to detect.

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 through 12 for full runs and Figures 13 through 15 for lite runs. The left half indicates the sorted top 10 runs, while the right half indicates the order by which the runs are significant according to the randomization test. Different levels of indentation signifies a significant difference according to the test. Runs at the same level of indentation are indistinguishable in terms of the test. In all tests the top ranked run was significantly better than other runs.

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, improved kernel methods, sophisticated fusion strategies, and combination of low and intermediate/high features. In addition, efficiency improvements (e.g., GPU implementations), analysis of more than one keyframe per shot, audio analysis, and using temporal context information. While not so much use of motion information, metadata or ASR. Some systems made use of external (ImageNet 1000-concept) data. Still not many experiments are using external training data (main focus on category A) and no improvements are shown using external training data from the limited experiments reported.

4 Known-item search

The known-item search (KIS) 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.

4.1 Task

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

- automatically return a list of up to 100 video IDs ranked by 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 minutes could elapse from the time the topic is presented to the system/searcher until the search result for the topic was frozen as complete. Interactive systems were able to query a web-based service to find out if a given video file was the known-item sought - this to simulate the fact that searchers looking for their own known-item would recognize it if they found it and stop the search. Each such query was logged and all logs published with the TRECVID workshop results.

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

4.2 Data

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

4.3 Topics

391 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 17 automatic and 12 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 mean inverted rank at which the known item is found or zero if not found. 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 depict for automatic and interactive runs, respectively, the main measure of effectiveness (mean inverted rank of the known-item) and the mean elapsed time across all topics. The highest scoring automatic runs achieved a mean inverted rank of about 0.45. While most runs used the textual metadata associated with the test video, two SCUC did not and yet scored among the top runs for effectiveness. This contradicts the findings from 2010 which, though based on very little data, indicated use of textual metadata was essential to good performance. Using more time does not result in better effectiveness. Adding a human to the search loop significantly improves performance, as expected. Since interactive runs returned only one known-item per topic, mean inverted rank is just the percentage of topics for which the run found the known-item - about 56 % for the highest scoring runs.

For interactive runs, the system could query an "oracle" to find out if a particular item was the known-

item for a given topic. This was intended to simulate a real searcher's recognition of the forgotten known item once it was retrieved. Figure 18 shows the number of calls to the oracle by topic for each of the four teams that used the oracle. There seems to be no clear relation between number of calls and system effectiveness.

Figure 19 shows for each topic, how many runs found the known-item for that topic. The topics are presented in order of how many runs found the known-item for that topic. As in 2010, for a significant number of topics (35 % in 2011 and 22 %in 2010) no run found the known-item. Why this is the case, remains an unanswered question.

In the following, the approaches taken by the participating groups are sketched out. The *Aalto University (PICSOM)* team built a system based on combining simple text search with automatically matched semantic concepts using concept detectors from the semantic indexing task. They tried to improve search by augmenting the metadata and ASR text with the output of optical character recognition. Their automatic runs used text search with a single video-level index containing all the ASR text plus the title, description and subjects from the meta data. They also included text detected by optical character recognition (OCR), lemmatization and used automatic selection of concepts based on matching keywords in the query text. Neither the concept detectors nor the lemmatiztion managed to improve over their baseline.

The *AXES* team comprised researchers from Dublin City University, Ersamus University, the Netherlands Institute for Sound and Vision (NISV), Oxford University, the Indian Institute of Information Technology, and the Fraunhofer Institute. Their system was an initial implementation for a multi-year participation in interactive versions of the known-item and instance search tasks. It incorporated text search based on ASR and metadata using the Lucene search engine, concepts based on a Pyramid Histogram of Visual Word, and a very efficient support vector machine (SVM) classifier. Included as well was similarity search not based on the usual low-level features (color, texture, etc) but on an object search and an elliptical region. The sources were then fused together. A desktop user interface was tested with 14 media professionals as users from NISV in Amsterdam

The *Chinese University of Hong Kong (VIREO)* built on their 2010 participation. They set out to observe the effectiveness of different modalities (metadata, ASR, and concepts) using the same approach as in 2010. Consistent with previous year's results, the evaluation showed that concept-based search is not helpful in known-item search. Textual-based modalities continued to deliver reliable performance especially by use of the metadata. Supplementing the metadata with the ASR feature was no longer able to boost the performance as was the case in 2010.

The *iAD-DCU-CLARITY* team was a collaboration between Dublin City University and several Norwegian Universities and groups, funded by the Norwegian Research Council and it built on participation in interactive known-item search in 2010 with novice users. The team implemented an iPad interface to a KIS video search tool to evaluate different display methodologies for KIS interaction. The system incorporated keyframe clustering based on MPEG-7 features using k-means as well as concept detection for search and for choosing most representative keyframes. The researchers compared baseline non-clustering to a clustering system on a topic by topic basis with 6 interactive users in Oslo and in Dublin.

The *KB Video Retrieval* team (KBVR) fielded a baseline text-only run plus pseudo-relevance feedback and semantic concept re-ranking. They used the Terrier system on ASR and metadata. They incoporated semantic concept re-ranking assuming the known item was retrieved but needs to be "bubbled up" the ranking. The query and initial documents were mapped into a semantic space defined using 130 LSCOM concepts with the description of each concept enhanced using a Wikipedia knowledge base.

The *KSLab at Nagaoka University of Technology* developed an iPad interface for interactive known-item search in their first TRECVID participation. The system searched the metadata using Lucene, refining salient words integrated into retrieval. It used video length as a cue for the user.

Sichuan University of China participated but did not submit a paper.

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 2011 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 Task

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

- 5 or so example frame images drawn at intervals from a video containing the item of interest. For each frame image:
  - a binary mask of an inner region of interest within the rectangle
- an indication of the target type taken from this set of strings (PERSON, CHARACTER, LOCATION, OBJECT)

Figure 20 is an example of the masks provided with each topic.

5.2 Data

Test data: BBC rushes. The rushes were automatically divided into short, roughly equal-length clips (10 s or 20 s) and renamed so the clip name did not indicate the original video. Each clip was to be processed as if no others existed. To see if this seems to have happened, we calculated for each topic the ratio of the following values:

- clips for which both the original and the transform were submitted
- total submitted clips

If systems treated each original and its transform identically then the ratio would be 1. Figure 21 shows this is almost never the case - with the exception of 3 topic results from AXES-DCU runs, each with only 2 of 4 items found by the human searcher and submitted.

Recurring objects, people, and locations were included as needed to create a set of approximately 25 topics. Some video transformations were applied at random to the original test clips - approximating differences you might see if a clip came from a different camera or was taken under different lighting conditions, etc - to create additional test clips.

5.3 Topics

Topics were created to emphasize objects. Topic targets included 17 objects, 6 persons, 2 locations. This represented significant change from 2010 in which topic targets emphasized people and included 8 objects, 8 persons, 5 characters, and 1 location. Figures 22, 23, and 24, show the object targets. Figure 25 shows the person targets. Figures 26, shows the location targets.

5.4 Evaluation, Measures

Each group was allowed to submit up to 4 runs and in fact 13 groups submitted 37 automatic and 4 interactive runs.

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

5.5 Results

Figure 27 is a boxplot showing the distribution of effectiveness scores (average precision) by topic and topic type, as achieved by fully automatic systems. Figure 28 provides the corresponding information for interactive runs. In both cases the number of data points does not allow for strong conclusions but in general systems score best on locations, where they can use the entire frame, next best on objects, and less well on people. Variation across topics is enormous as Figure 29 shows for the top-scoring runs. Surprizingly, some fully automatic runs achieved better effectiveness than most of the interactive runs. Figure 30 shows the results of a randomization test to find significant differences among automatic and among interactive runs.

Figures 31 and 32 explore the question of whether effectiveness is predictable simply from the number of targets available in the test collection. That does not seem to be the case.

Runs that used more elapsed processing time did not generally achieve greater effectiveness - see Figure

33.

In the following, the approaches taken by the participating groups are sketched out.

TNO submitted 3 runs. One used an exhaustive keypoint search, one a bag-of-visual-words approach, and one open-source face recognition software. In terms of effectiveness, they found that the keypoint search significantly outperformed the bag-of-visual-words approach and that face-recognition software can contribute if the queries contain large frontal faces.

The *NII* team explored three different approaches: a) large vocabulary quantization by hierarchical k-means and a weighted histogram intersection based ranking metric, b) combination of similarities based on Glocal quantization of two sets of scale-invariant feature transforms (SIFTs) and color histograms from the full frames, and c) keypoint matching used to compute the similarity between images of the query and images of all videos.

AXES-DCU experimented with an interactive INS search system. 30 media students participated in the study. Their system used a pyramid histogram of visual words based on a dense grid of SIFT features at multiple resolutions. Ranking was achieved using a non-linear chi-square SVM.

AT&T Labs Research based their instance search system on their content-based copy detection work. A baseline run included speeded up robust features (SURF). A normalization technique promoted matches from each query sample image to near the top. They performed outlier analysis, finding weak performance for homogeneous visual characteristics (low contrast, few edges). They experimented with and identified the use of visual content features as a major challenge.

Beijing University of Posts and Telecommunications-MCPRL used features such as hue-staturation-value (HSV) histograms, red-green-blue (RGB) moment, SIFT, SURF, CSIFT, Gabor Wavelet, Edge histograms, local binary patterns (LBP), and histograms of oriented gradients (HoG). Higher weight was given for reranking close-up shots. Specific normalization techniques were developed for each modality. Runs were constructed to compare three (non-specified) score merging strategies.

The *VIREO: City University of Hong Kong* team looked at key differences with search task and content-based copy detection(CCD): region of interest specification, wider definition of relevance than visual copies (e.g., person), and multiple examples with varying conditions (unlike CCD). Their approach incorporated SIFT (Lowe), BoW, and one keyframe per shot. Their four runs contrasted the following: full matching (vireo b) versus partial matching (vireo m), use of weak geometric information (vireo b) versus stronger spatial configuration (vireo s), and use of face matching (vireo f). There was no clearly winning approach. Performance depended on aspects such as size, context uniformity, etc.

Florida International University / University of Miami, in their first participation in the instance search task, employed texture features plus SIFT, Multiple Correspondence Analysis (MCA), and variants enhanced by k-nearest neighbors (KNN) reranking, MCA reranking, SIFT, and 261 extra training images. No significant differences between the runs were found.

The *Instituto de Matematica e Estatistica, University of Sao Paulo* used pyramid histograms of visual words (PHOW) a variant of Dense SIFT (5 pixels distance), and 600 000 descriptors clustered into 300 visual words. Frames were represented as word frequency vectors. The similarity computation was based on chi-square. Only one run was submitted; it scored above median for location topics (where texture was important).

The researchers at *JOANNEUM RESEARCH and Vienna University of Technology* fused four different techniques: face detection (Viola Jones) followed by face matching (Gabor wavelets), BoF (bag of features) with codebook size 100, mean shift segments (color segmentation), and SIFT (Lowe). Fusion took the best result across all topic sample images for all four methods. SIFT-only run performed best, especially well for location type.

IRIM team was a large collaboration of European research groups. They used two representations: bag of visual words (BoVW) (using SURF descriptors) 16 000-word codebook and bag of regions (with HSV histogram as descriptor) 2000-word codebook. For measuring similarity they used BoVW (complement of histogram intersection) and bag-of-regions (BOR) (L1-distance). They made limited use of the mask (only over 8 points for BoVW). The best results came from the merged BOVW/BOR and complete frame approaches.

6 Multimedia event detection

The 2011 Multimedia Event Detection (MED) evaluation was the first, non-pilot evaluation of technologies that search multimedia video clips for complex events of interest to a user. The 2011 effort was scaled up from the 2010 MED Pilot evaluation in four areas:

- Phases of MED system operation. Three phases were defined: metadata generation(video ingest and metadata store creation), event agent generation(event definition ingest and event agent creation), and event agent execution (search). The phases facilitate both glass-box and black-box evaluations.
- Indexing collections. The test collection was 17 times larger than the pilot collection.
- Events tested. The number of events was increased to 10 from 3 for the pilot.
- Evaluation conditions. Two evaluation conditions were introduced to differentiate event agent generation styles and systems that processed all 10 testing events.

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 an expression of some event domain-specific knowledge needed by humans to understand the event definition.
- 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.

6.1 Data

A development and evaluation collection of Internet multimedia (i.e., video clips containing both audio and video streams) clips 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 ≈114 hours of videos.
  - The MED '11 development collection consisting of 10 404 clips totaling ≈ 324 hours of video.
  - Five "training" event kits which were used for internal testing.
  - Ten "testing" events kits which were used for the formal evaluation.
- Evaluation data consisting of 32,062 total clips totaling ≈ 991 hours of video. The evaluation data set included nominally between 80 and 187 positive instances per event.

The MED '11 event names are listed in Table 3 and Table 4.

6.2 Evaluation

Sites submitted system outputs for either all 10 events (referred to as a MEDFull submission) or any

9 or fewer events (referred to as a MEDPart submission). Developers reported how their event agents were constructed: either with human intervention (SemiAutoEAG) or without human intervention (AutoEAG).

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 computer 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.

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 is 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 based on the Detection Threshold. There are two primary measures for computational speed. The first is Metadata Generation Processing Speed (MGPS) which is the real-time factor to complete all steps necessary to build the metadata store. Real-time factor is the total processing time divided by the number of hours of video in the test collection. The second is Event Agent Execution Processing Speed (EAEPS) which is the real-time factor for each event processed during the event agent execution phase.

Participants were provided a Decision Error Tradeoff (DET) curve for each event their system processed and a variety of diagnostic measures using the Normalized Detection Cost (NDC) model. NDC is a weighted linear combination of the system's missed detection probability and false alarm probability. Several variations of NDC were used including: Actual NDC (the NDC based on the Detection Threshold), Minimum NDC (the NDC at the optimum threshold), and NDC at the Target Error Ratio (TER) (the NDC where the DET curve crosses the optimum balance of miss and false alarm probabilities given the NDC weights). Consult the evaluation plan for specifics of the DET curves and these measures.

6.4 Results

Tables 5 and 6 show the P_{Miss} and P_{FA} rates at the systems actual decision points for the primary MED submissions of all 19 participants. All but one team submitted runs for all ten test events. There was a surprising amount of inter-event variability. E014 (Repair an appliance) E011 (Making a sandwich) had the lowest and highest median P_{FA} rates of 0.0095 and 0.0260 respectively - a factor of 2.7. Overall, E008 (Flash Mob) and E010 (Grooming an animal) had the lowest and highest median P_{Miss} rates of 0.0260 and 0.6437 - a factor of 24.7. The majority of the variability is evident in the P_{Miss} rates which varies by almost an order of magnitude more than P_{FA}.

While Tables 5 and 6 show the primary metric, Figure 34 shows a box plot of the NDC at the Target Error Ratio for all primary systems. The graph shows the relative difficulty of events when weighted P_{Miss} and P_{FA} rates are taken into account since the measure combines both the error rates and removes threshold estimation step from the scores. E008 and E015 (Working on a sewing project) have the lowest and highest median NDCs of 0.636 and 1.064 respectively – a difference factor of 1.67.

Focusing on the E008, the event with the lowest median NDC@TER, Figure 35 shows the DET curves for all primary systems. In contrast, Figure 36 shows the DET curves for E015. Both show a wide range of system performance across systems however several systems more than halved their error rates for E008.

Both accuracy and processing speed are critical

components to understanding the efficacy of a technology. Two measures of processing speed were reported by participants, Metadata Generation Processing Speed (MGPS) generation speed and Event Agent Execution Processing Speed (EAEPS). Figure 37 plots the site-reported, Single-Core Adjusted MGPS as a function of accuracy - in this case NDC@TER. The expectation is to see a trend of reduced cost as a function of the Real Time Factor. There is a weak positive correlation for MGPS and a weak negative correlation for the presented Single-Core Adjusted MGPS with an R^2 value of 0.15. MGPS is presumably uncorrelated with performance because it does not account for the size of the cluster.

Participants self-reported their EAEPS. The median runtime across events and systems was 0.0015 x real time (RT) with a mean of 0.077 x RT.

6.5 Summary

In summary, 19 sites participated in the MED '11 evaluation. All but one site participated in all 10 test events. There was a large variation across both events and systems. Systems achieved the lowest error rates on the Flash Mob event and the highest error rates on the Making a sewing project event.

TRECVID '12 evaluation will include the MED Track. There will be a new, larger test corpus and new test events. Given the surprising low error rates, a pilot Ad Hoc MED condition (where the metadata store constructed before the event is known) and a reduced training exemplar tests will be included in the evaluation.

7 Copy detection

As used here, a copy is a segment of video derived from another video, usually by means of various transformations such as addition, deletion, modification (of aspect, color, contrast, encoding, ...), camcording, etc. Detecting copies is important for copyright control, business intelligence and advertisement tracking, law enforcement investigations, etc. Content-based copy detection offers an alternative to watermarking.

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

The required system task was as follows: given a test collection of videos and a set of 11256 queries, determine for each query the place, if any, that some part of the query occurs, with possible transformations, in the test collection. Two thirds of the queries contained copies.

A set of 8 possible video transformations was selected to reflect actually occurring video transformations:

- T1: Simulated camcording (automatic transformation of the query in some ways similar to those resulting from manual camcording)
- T2: Picture in picture Type 1 (The original video is inserted in front of a background video)
- T3: Insertions of pattern
- T4: Compression
- T5: Change of gamma
- T6: Decrease in quality – This includes choosing randomly 3 transformations from the following: Blur, change of gamma (T5), frame dropping, contrast, compression (T4), ratio, white noise
- T8: Post production – This includes choosing randomly 3 transformations from the following: Crop, Shift, Contrast, caption (text insertion), flip (vertical mirroring), Insertion of pattern (T3), Picture in Picture type 2 (the original video is in the background)
- T10: Randomly choose 1 transformation from each of the 3 main categories.

Each transformation was applied to each of 201 untransformed (base) queries using tools developed by IMEDIA to include some randomization at various decision points in the construction of the query set. In total 1608 video-only queries were constructed. For each query, the tools took a segment from the test collection, optionally transformed it, embedded it in some video segment which did not occur in the test collection, and then finally applied one or more transformations to the entire query segment. Some

queries contained no test segment; others were composed entirely of the test segment. Video transformations included camcording simulation, picture-in-picture, insertion of patterns, reencoding, change of gamma, decreasing the quality, and post production alterations. Video transformations used were documented in detail as part of the TRECVID Guidelines.

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

A script to construct 11256 audio + video queries was provided by NIST. These queries comprised all the combinations of transformed audio(7) and transformed video (8) from a given base audio+video query (201).

7.1 Data

The task used the same test data as in 2010. Queries of 2011 are randomly selected and are disjoint to 2010 queries. The Internet Archive video collection was used as a source for reference and non-reference videos. The testing and development videos (11200 files) of 400 hours and duration less than 4.1 min were used as a source from which the test query generation tools chose reference video. While the non-reference video collection was selected from a set of 12480 videos with total duration of 4000 hours and average duration between 10-30 min.

7.2 Evaluation

In total, 22 participant teams submitted 73 runs for evaluation. 41 runs were submitted as balanced runs and 32 as no false alarms. Copy detection submissions were evaluated separately for each transformation, according to:

- How many queries they either find the reference data for or correctly report that there is no reference data (copy) included
- When a copy is detected, how accurately the run locates the reference data in the test data.
- How much elapsed time is required for query processing

After creating the query set, it was found that there exist 8 queries with duplicate reference videos with different file names. For those cases we asked systems to submit all found result items and then before evaluation we removed from the submitted runs all duplicate result items that don't match the ground truth thus we keep the correct reference video and at the same time avoid dropping the whole query from evaluation.

7.3 Measures (per transformation)

- Minimal Normalized Detection Cost Rate: a cost-weighted combination of the probability of missing a true copy and the false alarm rate. For TRECVID 2011 the cost model was identical to that used in 2010 and assumed copies are very rare (e.g. 0.005/hr). Two application profiles were required: the "Balanced" profile in which misses and false alarms are assigned a cost of 1, and the "Nofa" profile in which a false alarm is assigned a cost of 1000 times the cost of a miss. Other realistic scenarios were of course possible. Normalized minimal detection cost rate (minNDCR) reduced in 2010 to two terms involving two variables: probability of a miss (Pmiss) and the number of false alarms (FA). In 2011, the total length of queries per transformation in hours was 4.6 which is similar to 2010. For the "Nofa" profile:

  $$minNDCR = Pmiss + 108.7 * FA$$

  For the same queries under the "Balanced" profile:

  $$minNDCR = Pmiss + 0.1 * FA$$

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

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

7.4 Results

The detection performance among best runs for both profiles across all transformations is shown in Figures

38 to 41. In general this year it could be seen that the detection performance was better than in TRECVID 2010 (lower NDCR values). Note that comparing TV11 to TV10 can be tricky as the queries are not the same, although drawn from the same pool, thus the gain in performance can be either due to real system enhancements or due to easiness of queries. For the balanced profile the actual and optimal results seem to be very close while for "no false alarms" (NOFA) profile there is more difference which indicates that still the NOFA profile is harder for systems. Finally, transformations 3,4 and 5 achieved the best performance which was likely due to the fact that those transformations (insertion of patterns, re-encoding, and change of gamma) are simpler than the others (2,6,8, and 10) including Pict-in-Pict and combined transformations.

A comparison among the detection of the top 10 runs only for both profiles can be shown in the slides (`http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.ccd.slides.pdf`) and shows a gap between the actual median line and optimal median line which indicates that there is still space for more system improvements. This gap in the Nofa profile was much bigger compared to the balanced profile. In general, both profiles achieved actual median scores better than TRECVID 2010. A similar comparison based on the localization shows that both profiles achieved very high performance and better than TRECVID 2010. We also notice that the optimal and actual scores are almost aligned with very small gaps. In terms of efficiency, Figure 42 shows the top 10 processing time performance. Even though the best runs could detect copies in seconds, the majority of other systems were still far from real-time detection. Generally, systems are slower with higher median scores compared to last year. We may conclude that this year systems focused more on enhancing the detection performance but with the price of less efficiency.

The audio transformations 5, 6, and 7 are harder than the other transformations as they include mixing with external speech. This effect is obvious in Figure 43 which compares only the best runs in both profiles based on actual and optimal values. The red circles on the graph show the video transformations that were mixed with audio transformations 5, 6, and 7. In order to study the relation between the three main measures we plotted each two for all transformations (refer to the slides for detailed figures). In general, few systems achieved high localization in short processing time, and most systems which increased the processing time did not gain much in both localization and detection. On the other hand, systems that were good in detection were also good in localization. These observations are true for both application profiles.

Finally, we can draw some general observations from this year's task: First, the task community seems to be stable and even attracting new participants. As for the results, top actual balanced detection scores are very near to top optimal balanced detection scores. At the same time, in general top balanced detection results are better than no false alarm. Comparing TV11 to previous years we find that top balanced detection results are better than 2009 and 2010. Median scores for 2011 (less than or equal to 1) are lower than 2010. Still there is a bigger gap between actual and optimal medians for no false alarm detection results. However, TV11 looks better than TV10. TV11 top localization scores for both profiles are better than TV10 and on average, TV11 systems are slower than TV10. Good detecting systems are also good in localization as in previous years. Audio transformations 5,6 & 7 still seem to be the hardest, while video transformations 3,4,5 & 6 seem to be the easiest. Finally, NDCR and F1 seems to be approaching a ceiling. Readers should see the notebook papers posted on the TRECVID website (`trecvid.nist.gov`) for details about each participant's experiments and results.

8 Surveillance event detection

The 2011 Surveillance Event Detection (SED) evaluation was the fourth in a series of evaluations that focused on event detection in the surveillance video domain. The first such evaluation was conducted as part of the 2008 TRECVID conference series (Over et al., 2008; Rose, Fiscus, Over, Garofolo, & Michel, 2009) followed the next three years as part of the 2009 (Over et al., 2009), 2010 (Over et al., 2010), and 2011 TRECVID. The goal of the evaluation track is to support the development of technologies to detect visual events (people engaged in particular activities) in a large collection of surveillance video data. It was designed to move computer vision technology towards robustness and scalability while increasing core competency in detecting human activities within video. The approach used was to employ real surveillance data, orders of magnitude larger than previous computer vision tests, and multiple, synchronized camera

views. Figure 44 shows the coverage and views from the different cameras included in the data collection used for SED.

The 2009, 2010, and 2011 evaluations supported the same two evaluation tasks as the 2008 evaluation; retrospective event detection and freestyle analysis.

While freestyle analysis was offered in 2011, no site participated in the task.

Retrospective event detection 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).

In 2011, nine teams participated in the retrospective task. Figure 45 presents the list of participants and the number of system runs they provided for each event.

The 2011 evaluation tasks used the same development and evaluation data as the 2009 and 2010 evaluations. While it is unusual to reuse a test set three years in a row, it was done because error rates remain high despite improvements. The reuse allows direct inter year comparison but increases the chance of over-tuning to the data which is of less concern given the high error rates.

8.1 Test Events

Figure 46 lists the seven events used for SED11. The same events were used for the 2009, 2010, and 2011 SED evaluations. There are three classes of events: single person, single person with an object, and multiple person events. The events were chosen to have a range of difficulty for the surveillance domain.

8.2 Event Annotation

For SED11, we define an event to be an observable state change, either in the movement or interaction of people with other people or objects. As such, the evidence for an event depends directly on what can be seen in the video and does not require higher level inference. Annotation guidelines were developed to express the requirements for each event. To determine if the observed action is a taggable event, a reasonable interpretation rule was used. The rule: "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). Videos selected for the evaluation were annotated using the Video Performance Evaluation Resource (ViPER) (Doerman & Mihalcik, 2000) tool by the Linguistic Data Consortium (LDC). Events were represented in ViPER format using an annotation schema that specified each event observation's time interval.

8.3 Data

The development data consisted of the full 100 hours data set used for the 2008 Event Detection evaluation. The video for the evaluation corpus came from the 45-hour Home Office Scientific Development Branch (HOSDB)'s Image Library for Intelligent Detection Systems (iLIDS) (UKHO-CPNI, 2007 (accessed June 30, 2009)) Multi Camera Tracking Training (MCTTR) data set. The evaluation systems processed the full data set however systems were scored on a four-day subset of recordings consisting of approximately fifteen-hours of video data. Both data sets were collected in the same busy airport environment with the same video cameras. The entire video corpus was distributed as MPEG-2 in de-interlaced, Phase Alternating Line (PAL) format (resolution 720 x 576), 25 frames/sec, either via hard drive or internet download.

8.4 Evaluation

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

8.5 Measures of Performance

Since detection system performance is a tradeoff between probability of miss vs. rate of false alarms, this task used the Normalized Detection Cost Rate (NDCR) measure for evaluating system performance as described in the evaluation plan (Fiscus, Rose, & Michel, 2010).

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

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

where:

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

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

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

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

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

The inclusion of decision scores in the system output permits the computation of Detection Error Tradeoff (DET) curves (Martin, Doddington, Kamm, Ordowski, & Przybocki, 1997). DET curves plot P_{miss} vs. R_{FA} for all thresholds applied to the systems decision scores. These plots graphically show the tradeoff between the two error types for the system.

8.6 Results

The NDCRs for the submitted event runs can be found in Figure 47 and contains three NDCR values for each submission: the Actual NDCR which is the NDCR based on the binary decisions produced by the system, the Minimum NDCR which is the lowest NDCR possible based on the decision scores produced by the system and the NDCR at Target Operating Error Ratio (TOER)[1], here β

The difference between the Actual and Minimum NDCRs indicates how well the system-identified decision score threshold (via the binary decisions) was tuned to the NDCR function.

[1]found by searching the DET curve for the point where it crosses the theoretically balancing point where two error types (Miss Detection and False Alarm) contribute equally to the measured NDCR. The Target Operating Error Ratio point is specified by the ratio of the coefficient applied to the False Alarm rate to the coefficient applied to the Miss Probability

Figure 48 contains a single DET curve for each event run with the lowest Actual NDCR. The event with the lowest NDCR was PersonRuns. The highest NDCR events are CellToEar and ObjectPut which are the two Person+Object events.

Figure 49 contains a single DET curve for each event run with the lowest NDCR at TOER. The same two events, CellToEar and ObjectPut, had the highest NDCRs but the lowest NDCR event was PeopleSplitUp.

Figure 50 and Figure 51 present historical views across event, camera, and SED years. Unlike many evaluations, these results are legitimately comparable because the identical test data was used for all years. Figure 50 shows runs with the lowest Actual NDCR per event and Figure 51 shows the runs with the lowest Minimum NDCR. In general, 2010 results were lower compared to 2011 results - both for the Actual and Minimum NDCs. For open, voluntary evaluations such as SED, participants change from year-to-year introducing noise into cross-team analyses. However, it is evident that performance is being improved on the task for several events and camera views. However, it is not known if the improvements are due to over tuning to the training data.

8.7 Summary

This overview of TRECVID 2011 SED 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 paper in the TRECVID publications webpage: `http://www-nlpir.nist.gov/projects/tv2011/tv2011.html`

9 Summing up and moving on

This introduction to TRECVID 2011 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 notebook.

10 Authors' note

TRECVID would not have happened in 2011 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 130 features, and provided judgments for 10 features under the Quaero program.
- Georges Quénot provided the master shot reference for the IACC.1 videos.
- At DCU Kevin McGuinness ran the oracle for interative instance search experiments.
- The LIMSI Spoken Language Processing Group and VexSys Research provided ASR for the IACC.1 videos.

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

9023 OBJECT - setting sun

9024 LOCATION - upstairs, inside the windmill

9025 OBJECT - fork

9026 OBJECT - trailer

9027 OBJECT - SUV

9028 OBJECT - plane flying

9029 LOCATION - downstairs, inside the windmill

9030 OBJECT - the Parthenon

9031 OBJECT - yellow dome with clock

9032 OBJECT - spiral staircase

9033 OBJECT - newsprint balloon

9034 OBJECT - tall, cylindrical building

9035 OBJECT - tortoise

9036 OBJECT - all yellow balloon

9037 OBJECT - windmill seen from outside

9038 PERSON - female presenter X

9039 PERSON - Carol Smilie

9040 PERSON - Linda Robson

9041 OBJECT - monkey

9042 PERSON - male presenter Y

9043 PERSON - Tony Clark's wife

9044 OBJECT - American flag

9045 OBJECT - lantern

9046 PERSON - grey-haired lady

9047 OBJECT - airplane-shaped balloon

12 Tables

Table 1: Participants and tasks

Task						Location	TeamID	Participants
–	–	KIS	–	–	SIN	Europe	PicSOM	Aalto U.
–	INS	KIS	* * *	–	* * *	Europe	AXES-DCU	Access to Audiovisual Archives
CCD	INS	–	–	–	–	NorthAm	ATTLabs	AT&T Labs Research
–	–	–	MED	–	–	NorthAm	BBN-VISER	BBN, UMD, Columbia, UCF team
–	–	–	–	SED	–	Asia	BJTU_SED	Beijing Jiaotong U.
–	–	–	–	–	SIN	Asia	BJTU_SIN	Beijing Jiaotong U.
CCD	INS	KIS	–	SED	SIN	Asia	BUPT-MCPRL	Beijing U. of Posts & Telecom. - MCPRL
CCD	–	–	* * *	* * *	SIN	Europe	brno	Brno U. of Technology
–	* * *	* * *	MED	SED	SIN	NorthAm	CMU-Informedia	Carnegie Mellon U.
–	–	KIS	MED	–	SIN	Europe	ITI-CERTH	Centre for Research and Technology Hellas
–	–	–	MED	–	–	NorthAm	ADDLIV21CM	Charles Stark Draper Laboratory, Inc.
–	–	–	–	SED	–	Asia	IRDS-CASIA	Chinese Academy of Sciences - IRDS-CASIA
–	INS	KIS	MED	–	SIN	Asia	VIREO	City U. of Hong Kong
CCD	–	–	–	SED	–	NorthAm	CRIM-VISI	Computer Research Inst. of Montreal
–	–	KIS	MED	–	SIN	Europe	DCU-iAD-CLARITY	Dublin City U.
–	–	–	* * *	–	SIN	Asia	ECNU	East China Normal U.
–	–	–	–	–	SIN	Europe	Liris-Imagine	Ecole Centrale de Lyon, U. de Lyon
–	–	* * *	* * *	–	SIN	Europe	EURECOM	EURECOM
–	INS	–	–	–	SIN	NorthAm	FIU-UM	Florida International U.
CCD	–	–	–	–	SIN	Asia	FTRDBJ	France Telecom Orange Labs (Beijing)
* * *	–	–	MED	–	* * *	NorthAm	IBM	IBM T. J. Watson Research Center
CCD	–	* * *	MED	* * *	* * *	Europe	INRIA-LEAR	INRIA-LEAR
CCD	* * *	–	–	–	–	Europe	INRIA-TEXMEX	INRIA/IRISA
–	INS	–	–	–	–	Europe	ARTEMIS-Ubimedia	Inst. Telecom (SudParis), Bell Labs France
–	–	–	–	–	SIN	Europe	VideoSense	Institut Eurecom
–	INS	* * *	–	–	–	SouthAm	CAUVIS-IME-USP	Instituto de Matematica e Estatistica - USP
* * *	* * *	* * *	* * *	* * *	SIN	Asia	IMG@...	Tsinghua U., Fujitsu
CCD	–	–	–	–	–	Europe	ITU_MSPR	Istanbul Technical U.
–	INS	–	* * *	–	SIN	Europe	JRS-VUT	Joanneum Research, Vienna U. of Tech.
–	–	KIS	–	–	–	NorthAm	KBVR	KB Video Retrieval
CCD	* * *	–	–	–	–	Asia	KDDILabs	KDDILabs
–	–	–	MED	–	–	NorthAm	GENIE	Kitware Inc
. –	–	* * *	MED	–	SIN	Asia	cs24_kobe	Kobe U.
* * *	INS	* * *	* * *	* * *	SIN	Europe	IRIM	Laboratoire d'Informatique de Grenoble
–	* * *	KIS	–	–	–	Asia	KSLab-NUT	Nagaoka U. of Technology in Japan
* * *	INS	* * *	MED	* * *	SIN	Asia	NII	National Inst. of Informatics
–	INS	–	–	–	–	Europe	TNO	Netherlands Org. for Applied Sci. Research
* * *	* * *	* * *	* * *	SED	SIN	Asia	NHKSTRL	NHK Science and Technical Research Labs
–	–	–	MED	–	* * *	Asia	Nikon	Nikon Corporation
CCD	–	–	–	–	–	Asia	NTT-CSL	NTT Communication Science Labs-CSL
–	INS	–	–	–	–	Asia	NTT-NII	NTT Communication Science Labs-NII
–	–	–	–	–	SIN	Asia	NTT-SL-ZJU	NTT Cyber Solutions Lab.
CCD	* * *	–	–	–	–	Asia	IMP	Osaka Prefecture U.
CCD	–	–	–	SED	–	Asia	PKU-IDM	Peking U.-IDM
CCD	–	–	–	–	–	SouthAm	PRISMA	PRISMA-U. of Chile
–	* * *	–	MED	–	SIN	Europe	Quaero	Quaero consortium
CCD	–	–	–	–	–	Australia	RMIT	RMIT U. School of CS&IT
–	* * *	KIS	–	–	–	Asia	SCUC	Sichuan U. of China
* * *	* * *	* * *	MED	* * *	* * *	NorthAm	Aurora	SRI International Sarnoff Aurora
–	–	–	MED	–	–	NorthAm	SESAME	SRI International - SESAME
CCD	–	–	–	* * *	–	Asia	SYSU-GITL	Sun Yat-sen U. - GITL

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

Table 2: Participants and tasks (continued)

Task						Location	TeamID	Participants
CCD	–	–	–	–	–	Europe	Telefonica.research	Telefonica Research
–	* * *	* * *	MED	–	* * *	Australia	ANU	The Australian National U.
* * *	* * *	* * *	* * *	SED	* * *	Asia	TJUT	Tianjin U. of Technology
CCD	INS	–	–	–	–	Asia	tokushima_U	Tokushima U.
–	–	–	MED	SED	SIN	Asia	TokyoTech+Canon	Tokyo Inst. of Technology, Canon Corp.
CCD	–	–	–	–	SIN	Europe	iupr-dfki	U. of Kaiserslautern
* * *	* * *	–	* * *	–	SIN	Europe	marburg	U. of Marburg
–	* * *	* * *	MED	–	SIN	Europe	MediaMill	U. of Amsterdam
–	* * *	* * *	MED	–	SIN	Asia	UEC	U. of Electro-Communications
CCD	–	–	–	–	SIN	Australia	UQMSG	U. of Queensland
CCD	* * *	–	–	–	* * *	NorthAm	USC-UTSA	USC Viterbi School of Engineering
CCD	* * *	* * *	* * *	* * *	* * *	Asia	XJTU	Xi'an Jiaotong U.
CCD	* * *	–	–	* * *	–	Asia	ZJU_CS_IV	Zhejiang U.

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

Table 3: MED '11 Training Events

Training Events
E001 - Attempting a board trick
E002 - Feeding an animal
E003 - Landing a fish
E004 - Wedding ceremony
E005 - Working on a woodworking project

Table 4: MED '11 Test Events

Testing Events
E006- Birthday Party
E007 - Changing a vehicle tire
E008 - Flash mob gathering
E009 - Getting a vehicle unstuck
E010 - Grooming an animal
E011 - Making a sandwich
E012 - Parade
E013 - Parkour
E014 - Repairing an appliance
E015 - Working on a sewing project

Table 5: TRECVID 2011 MED: Primary Systems P_{Miss} and P_{FA}s for Actual Decisions (a)

	E006		E007		E008		E009		E010	
	P_{FA}	P_{Miss}	P_{FA}	P_{Miss}	P_{FA}	P_{Miss}	P_{FA}	P_{Miss}	P_{FA}	P_{Miss}
ADDLIV21CM	0.0320	0.5161	0.0506	0.7748	0.0433	0.5227	0.0885	0.6526	0.0724	0.7471
ANU	0.0102	0.4624	0.0043	0.5495	0.0020	0.4242	0.0027	0.5474	0.0031	0.7931
BBNVISER	0.0429	0.1398	0.0487	0.1622	0.0451	0.0227	0.0416	0.1368	0.0505	0.2644
CERTH-ITI	0.0133	0.9946	0.0538	0.8829	0.0353	0.8712	0.0533	0.7368	0.0409	0.8276
CMU	0.0092	0.4785	0.0104	0.4144	0.0066	0.2727	0.0126	0.2737	0.0069	0.6552
DCU	0.0223	0.6129	0.0239	0.6036	0.0256	0.8864	0.0334	0.6632	0.0246	0.5402
GENIE	0.0056	0.6667	0.0084	0.6396	0.0082	0.4318	0.0090	0.5053	0.0077	0.7586
IBM	0.0162	0.4516	0.0197	0.3694	0.0137	0.1894	0.0220	0.2737	0.0335	0.4598
INRIA-LEAR	0.0144	0.5484	0.0004	0.8739	0.0034	0.4848	0.0026	0.6526	0.0056	0.7471
MediaMill	0.0161	0.7043	0.0158	0.6126	0.0152	0.4545	0.0169	0.5368	0.0170	0.6437
NII	0.1354	0.3548	0.0591	0.5676	0.1588	0.0833	0.0662	0.3579	0.1546	0.4598
Nikon	0.0001	1.0000	0.0002	1.0000	0.0107	1.0000	0.0002	1.0000	0.0007	1.0000
Quaero	0.0008	0.9946	0.0006	0.9820	0.0207	0.4167	0.0115	0.6526	0.0058	0.8736
SRI-AURORA	0.0495	0.2634	0.0593	0.2342	0.0795	0.0530	0.0363	0.2105	0.0720	0.3103
Sesame	0.0161	0.6935	0.0158	0.6306	0.0152	0.4545	0.0167	0.5368	0.0170	0.6437
TokyoTech-Canon	0.0131	0.4946	0.0086	0.4505	0.0120	0.1894	0.0116	0.2842	0.0089	0.6437
UEC	0.4661	0.0591	0.5896	0.1171	0.6078	0.0076	0.4626	0.0526	0.6616	0.0230
VIREO	0.1797	0.3763	0.0781	0.8198	0.0106	0.6515	0.0720	0.3053	0.1407	0.5747
cs24-kobe			0.0301	0.7748						
mean	0.0579	0.5451	0.0567	0.6031	0.0619	0.4120	0.0533	0.4655	0.0735	0.6092
median	0.0161	0.5053	0.0197	0.6126	0.0152	0.4280	0.0194	0.5211	0.0208	0.6437

Table 6: TRECVID 2011 MED: Primary Systems P_{Miss} and P_{FA}s for Actual Decisions (b)

	E011		E012		E013		E014		E015	
	P_{FA}	P_{Miss}	P_{FA}	P_{Miss}	P_{FA}	P_{Miss}	P_{FA}	P_{Miss}	P_{FA}	P_{Miss}
ADDLIV21CM	0.0567	0.7071	0.0386	0.5931	0.0475	0.6442	0.0087	0.4872	0.0578	0.6914
ANU	0.0158	0.5786	0.0004	0.8874	0.0030	0.6635	0.0033	0.4615	0.0003	0.8642
BBNVISER	0.0326	0.2786	0.0305	0.1342	0.0380	0.1058	0.0329	0.1795	0.0548	0.3086
CERTH-ITI	0.0253	0.9786	0.0252	0.8918	0.0794	0.9423	0.0032	0.9744	0.0585	0.8642
CMU	0.0098	0.5429	0.0131	0.3463	0.0070	0.3173	0.0071	0.2692	0.0048	0.5556
DCU	0.0267	0.7143	0.0264	0.6623	0.0231	0.6923	0.0301	0.6026	0.0204	0.7284
GENIE	0.0062	0.7571	0.0071	0.6277	0.0098	0.4615	0.0056	0.4487	0.0080	0.7037
IBM	0.0350	0.4071	0.0211	0.3117	0.0275	0.1827	0.0104	0.3718	0.0274	0.4568
INRIA-LEAR	0.0271	0.6714	0.0035	0.6494	0.0013	0.5288	0.0011	0.6667	0.0211	0.6049
MediaMill	0.0164	0.6357	0.0148	0.5801	0.0164	0.5385	0.0164	0.5513	0.0165	0.6420
NII	0.1094	0.4357	0.0916	0.3074	0.1414	0.2981	0.0641	0.3462	0.0335	0.6914
Nikon	0.0079	0.9929	0.0001	1.0000	0.0045	0.9904	0.0001	1.0000	0.0024	1.0000
Quaero	0.0008	0.9357	0.0045	0.9264	0.0134	0.7788	0.0020	0.6282	0.0098	0.8272
SRI-AURORA	0.0586	0.3071	0.0614	0.1039	0.0505	0.0769	0.0336	0.2949	0.0499	0.3333
Sesame	0.0163	0.6214	0.0147	0.5671	0.0165	0.5385	0.0164	0.5513	0.0164	0.6420
TokyoTech-Canon	0.0118	0.6071	0.0134	0.4286	0.0084	0.3750	0.0036	0.5128	0.0109	0.5802
UEC	0.7044	0.0143	0.5848	0.0216	0.6254	0.0385	0.3418	0.1282	0.3560	0.2222
VIREO	0.1136	0.5714	0.0749	0.4069	0.0526	0.4327	0.0444	0.5000	0.0584	0.5432
mean	0.0708	0.5976	0.0570	0.5248	0.0648	0.4781	0.0347	0.4986	0.0448	0.6255
median	0.0260	0.6142	0.0180	0.5736	0.0198	0.4951	0.0095	0.4936	0.0208	0.6420

Table 7: 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
9023	35987	12264	34.1	320	5838	47.6	86	1.5
9024	35883	10546	29.4	200	3301	31.3	109	3.3
9025	36568	11934	32.6	400	6751	56.6	105	1.6
9026	36701	11848	32.3	100	2304	19.4	22	1.0
9027	36187	10389	28.7	240	3950	38.0	32	0.8
9028	36633	10609	29.0	240	3903	36.8	64	1.6
9029	36421	11452	31.4	140	2641	23.1	69	2.6
9030	36154	12232	33.8	200	3937	32.2	31	0.8
9031	36719	12722	34.6	200	3780	29.7	177	4.7
9032	36727	11570	31.5	340	5978	51.7	49	0.8
9033	36327	12001	33.0	220	4120	34.3	45	1.1
9034	36137	11101	30.7	160	3131	28.2	27	0.9
9035	36822	11776	32.0	300	5415	46.0	57	1.1
9036	35412	12179	34.4	260	4723	38.8	48	1.0
9037	36706	11948	32.6	260	5025	42.1	70	1.4
9038	36861	11995	32.5	220	4461	37.2	21	0.5
9039	37050	13119	35.4	260	5297	40.4	34	0.6
9040	36827	12513	34.0	160	3275	26.2	43	1.3
9041	36877	12548	34.0	320	6140	48.9	108	1.8
9042	36998	12657	34.2	240	4858	38.4	84	1.7
9043	36928	11761	31.8	520	7860	66.8	287	3.7
9044	35769	11127	31.1	200	3594	32.3	25	0.7
9045	36724	12055	32.8	280	5499	45.6	28	0.5
9046	37088	12139	32.7	280	4936	40.7	139	2.8
9047	35218	11443	32.5	240	4374	38.2	70	1.6

Table 8: 2011 Teams not submitting any runs

CCD	INS	KIS	MED	SED	SIN	Location	TeamID	Participants
−	−	−	∗∗∗	∗∗∗	∗∗∗	Europe	AIT-MKWT	Athens Information Technology
−	−	−	∗∗∗	−	−	Asia	BJTU_MED	Beijing Jiaotong U.
∗∗∗	−	−	−	−	−	Asia	Kerwise	Beijing U. of Posts & Telecom
∗∗∗	−	∗∗∗	−	∗∗∗	−	Asia	AICS	Chinese Academy of Sciences - AICS
−	−	∗∗∗	∗∗∗	−	−	NorthAm	Columbia	Columbia U.
∗∗∗	∗∗∗	−	−	∗∗∗	−	Asia	MCG-DUT	Dalian U. of Technology
−	−	∗∗∗	−	−	−	Europe	DMIR	Delft U. of Technology
−	∗∗∗	∗∗∗	∗∗∗	∗∗∗	−	Europe	duth	Democritus U. of Thrace - EECE
−	−	−	−	−	∗∗∗	Asia	NISTFudan7	Fudan U.
∗∗∗	−	−	−	−	−	Asia	HUNAN.NISL	Hunan U.
−	−	∗∗∗	−	−	−	Asia	I2R	I2R, Singapore
−	−	−	∗∗∗	∗∗∗	−	Europe	PULSAR	INRIA Sophia-Antipolis
−	−	−	−	∗∗∗	∗∗∗	Europe	willow	INRIA-willow
−	−	−	−	−	∗∗∗	Europe	LIF	LIF de Marseille
∗∗∗	∗∗∗	∗∗∗	−	−	−	Asia	NJUSE	Nanjing U.
∗∗∗	∗∗∗	−	∗∗∗	∗∗∗	−	Asia	NTU-EEE	Nanyang Technological U.
∗∗∗	−	−	∗∗∗	−	−	Asia	CCU	National Chung Cheng U.
−	∗∗∗	∗∗∗	−	−	−	Asia	HFUT-NUS	National U. of Singapore
−	−	−	∗∗∗	−	−	Australia	NICTA	National ICT Australia
−	∗∗∗	−	−	−	−	Asia	NUS-LV	National U. of Singapore
−	−	−	∗∗∗	∗∗∗	∗∗∗	NorthAm	NECLA	NEC Labs America
−	−	−	−	−	∗∗∗	Asia	SmcNPU	Northwest Polytechnical U.
∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	NorthAm	PURDUE_RVL_TRECVID	Purdue U.
∗∗∗	−	∗∗∗	−	−	∗∗∗	Asia	IDARE	Shandong U.
∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	Asia	SJTU-ICIP	Shanghai Jiao Tong U. - ICIP
−	∗∗∗	−	−	∗∗∗	∗∗∗	Asia	SJTU-IS	Shanghai Jiao tong U. - IS
−	−	−	∗∗∗	∗∗∗	−	NorthAm	AntsMiner	Stevens Inst. of Technology
∗∗∗	−	−	−	−	−	Asia	SYSU_SEC_LAB	Sun Yat-Sen U. - SECLAB
∗∗∗	∗∗∗	−	−	∗∗∗	∗∗∗	Europe	TUBITAK_UZAY	TUBITAK
−	−	∗∗∗	−	−	∗∗∗	Europe	SINAI	Universidad de Jan - SINAI
−	−	−	−	∗∗∗	−	SouthAm	UNS	Universidad Nacional del Sur
−	−	∗∗∗	−	−	−	Europe	ITEC-UNIKLU	Universitaet Klagenfurt
∗∗∗	∗∗∗	−	−	−	∗∗∗	NorthAm	BU-CAS-UTSA	U. of Texas at San Antonio
−	−	−	∗∗∗	−	−	NorthAm	CVLAB_UCF	U. of Central Florida
∗∗∗	−	−	−	−	−	Europe	CVSSP_Surrey	U. of Surrey
−	−	−	∗∗∗	−	∗∗∗	Europe	DCAPI	U. of Lincoln
−	−	−	−	∗∗∗	−	Europe	mr00034	U. Of Surrey
∗∗∗	−	∗∗∗	−	−	∗∗∗	SouthAm	RECOD	U. of Campinas
∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	Africa	REGIM	U. of Sfax
∗∗∗	∗∗∗	−	−	−	∗∗∗	NorthAm	USTC-UTSA-TSU	U. of Texas at San Antonio
−	∗∗∗	−	−	∗∗∗	−	Europe	VIP-UMA	U. of Malaga
∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	SouthAm	We.Find.Video	U. of Sao Paulo

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

13 Figures

Figure 1: Frequencies of shots with each feature

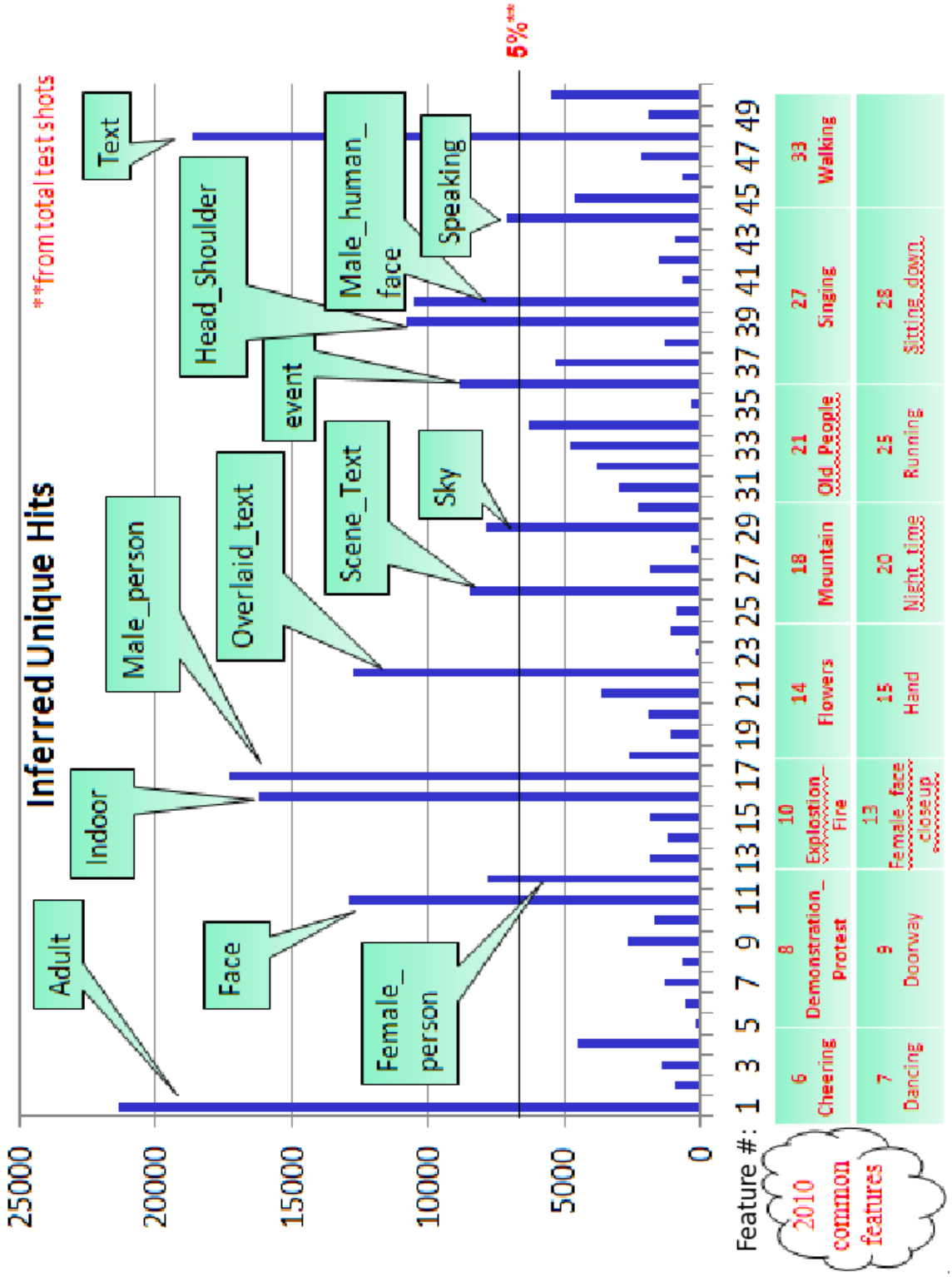
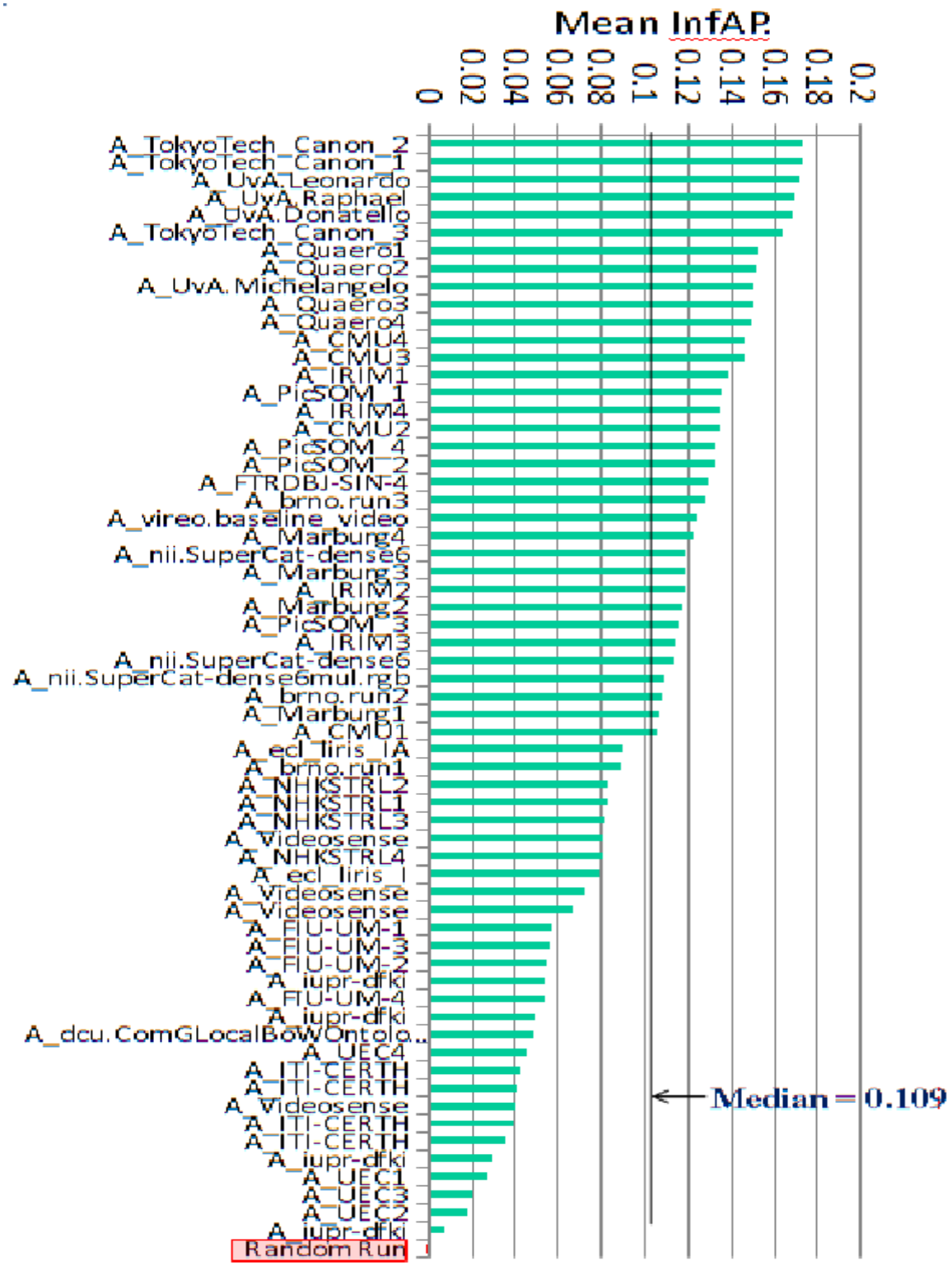


Figure 2: xinfAP by run (cat. A) - Full



[H]

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

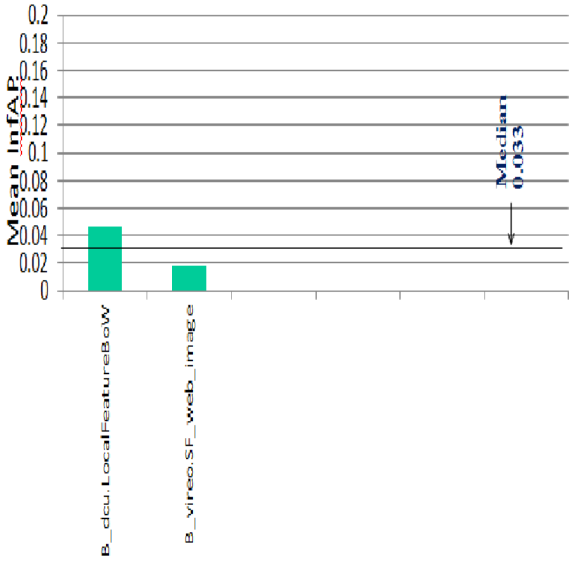


Figure 4: xinfAP by run (cat. D) - Full

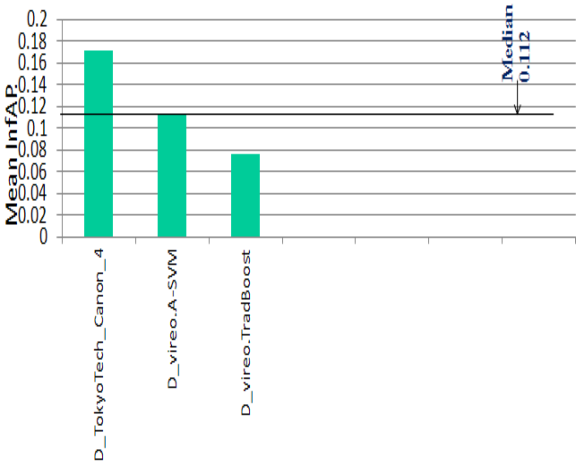


Figure 5: xinfAP by run (cat. A) - Lite

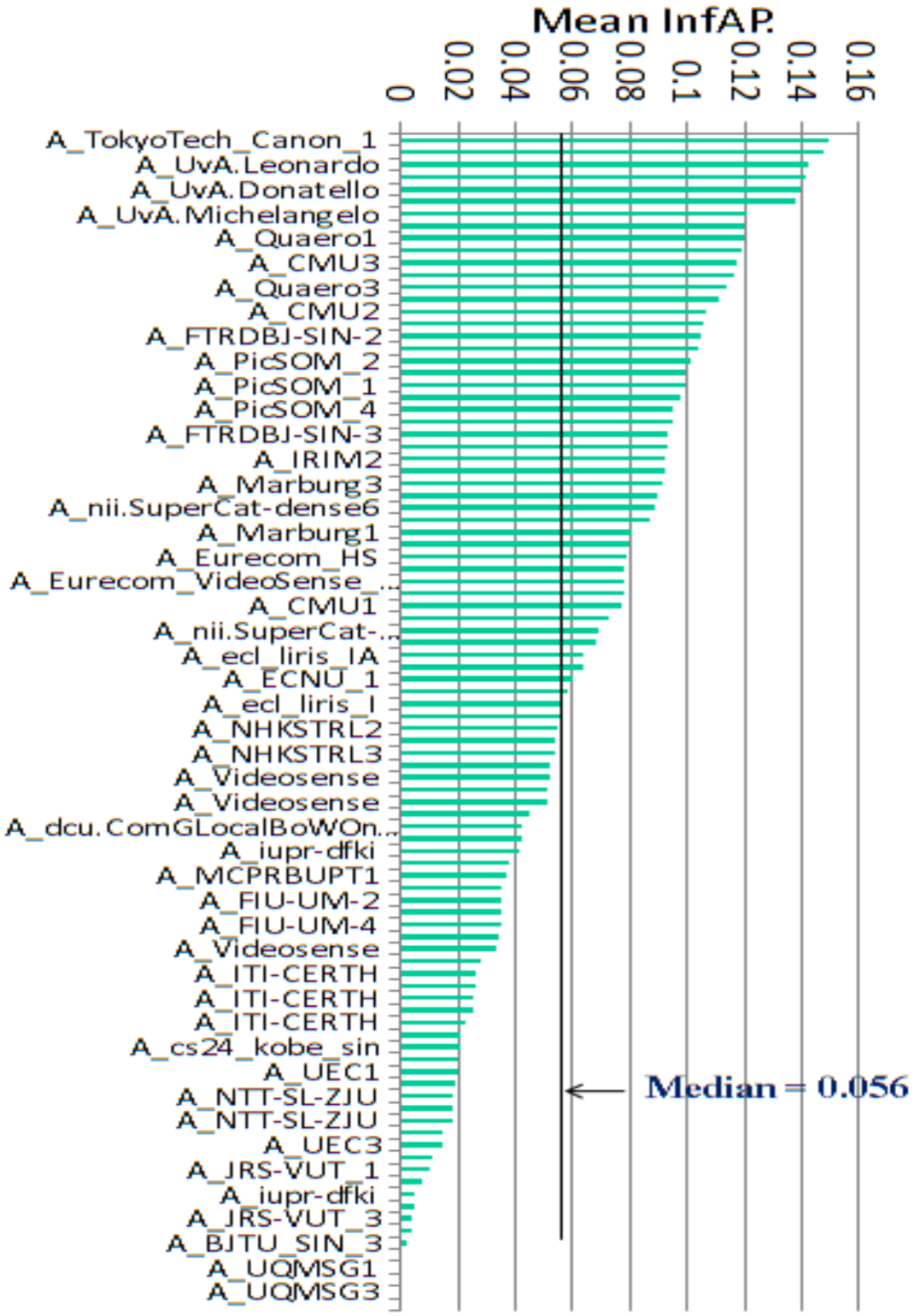


Figure 6: xinfAP by run (cat. B) - Lite

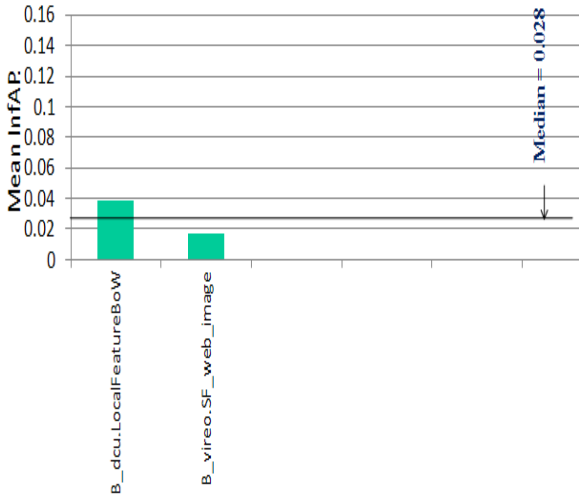


Figure 7: xinfAP by run (cat. D) - Lite

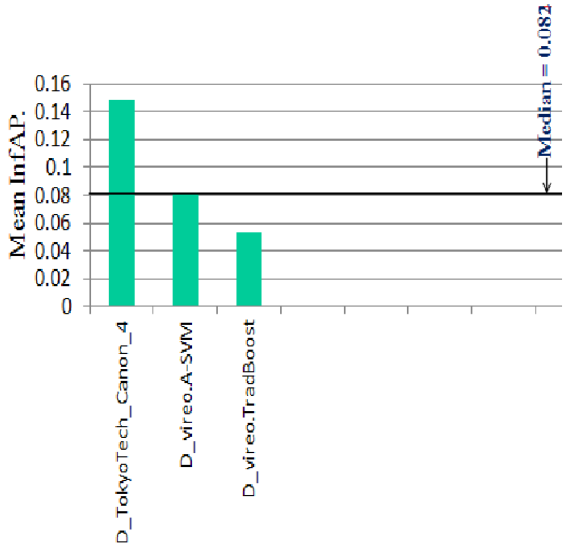


Figure 8: Top 10 runs (xinfAP) by feature - Full

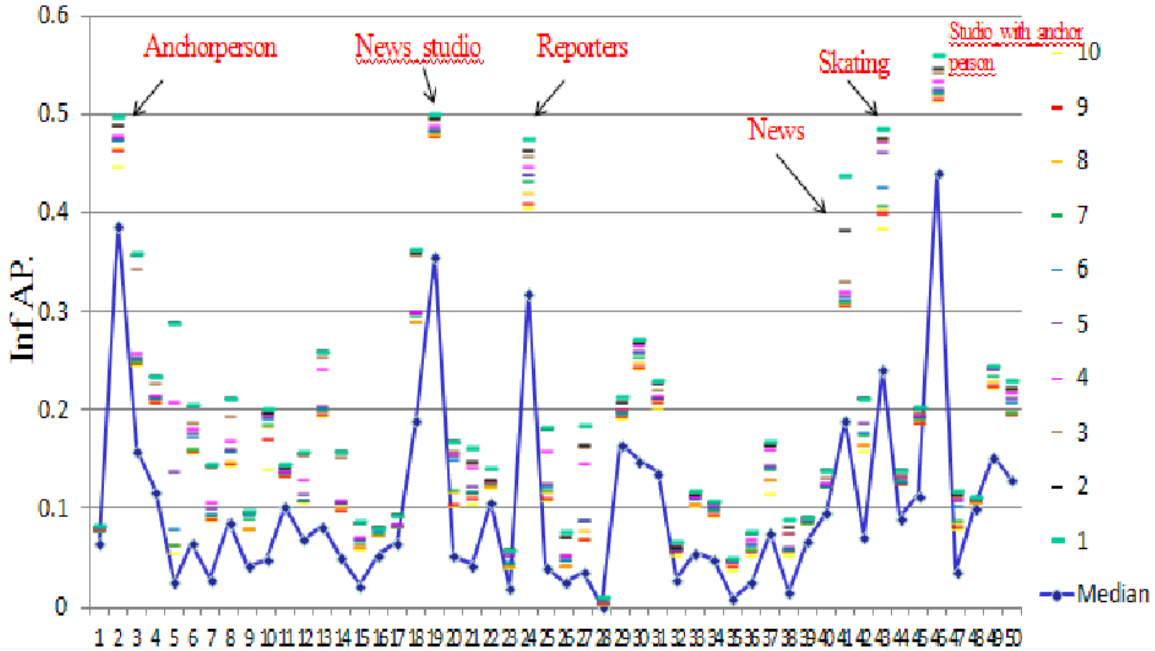


Figure 9: Top 10 runs (xinfAP) by feature - Full + Lite

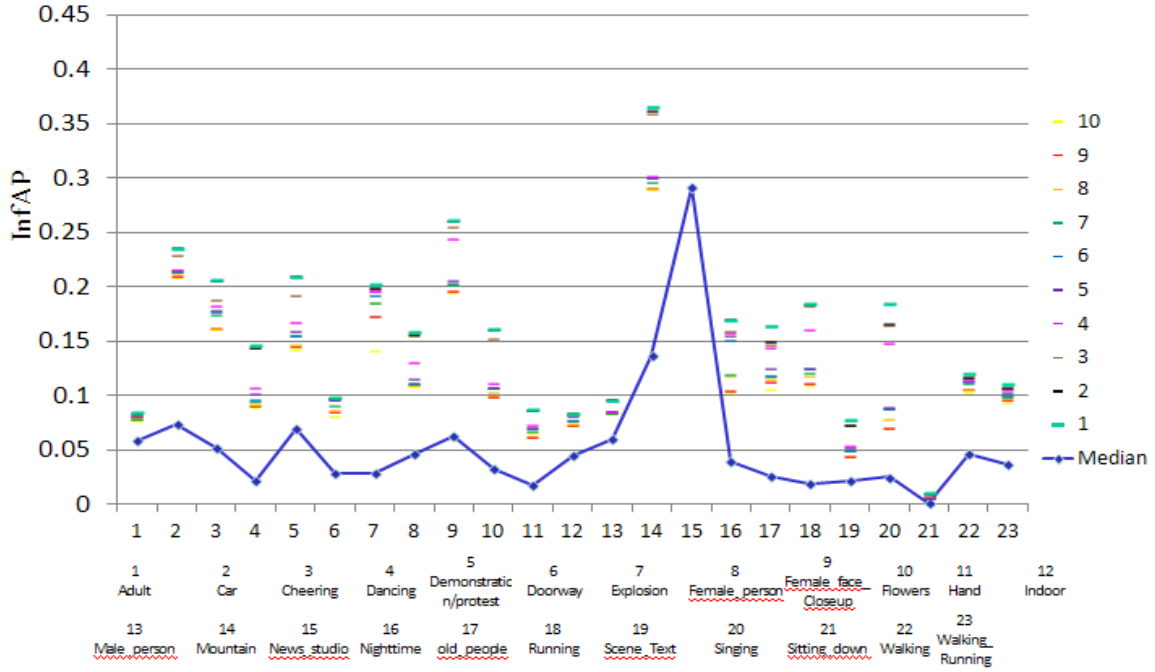


Figure 10: Significant differences among top A-category full runs

Run name	(mean infAP)
A_TokyoTech_Canon_2	0.173
A_TokyoTech_Canon_1	0.173
A_UvA.Leonardo_1	0.172
A_UvA.Raphael_3	0.170
A_UvA.Donatello_2	0.168
A_TokyoTech_Canon_3	0.164
A_Quaero1	0.153
A_Quaero2	0.151
A_UvA.Michelangelo_4	0.150
A_Quaero3	0.150

- A_UvA.Leonardo_1
  - A_UvA.Raphael_3
    - A_Quaero1
      - A_Quaero2
    - A_Quaero3
    - A_UvA.Michelangelo_4
  - A_UvA.Donatello_2
    - A_Quaero1
      - A_Quaero2
    - A_Quaero3
    - A_UvA.Michelangelo_4
- A_TokyoTech_Canon_2
  - A_TokyoTech_Canon_3
    - A_UvA.Michelangelo_4
    - A_Quaero1
      - A_Quaero2
    - A_Quaero3
- A_TokyoTech_Canon_1
  - A_TokyoTech_Canon_3
    - A_UvA.Michelangelo_4
    - A_Quaero1
      - A_Quaero2
    - A_Quaero3

Figure 11: Significant differences among top B-category full runs

Run name	(mean infAP)
B_dcu.LocalFeatureBoW_2	0.046
B_vireo.SF_web_image_4	0.019

- B_dcu.LocalFeatureBoW_2
  - B_vireo.SF_web_image_4

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

Run name	(mean infAP)
D_TokyoTech_Canon_4	0.172
D_vireo.A-SVM_3	0.112
D_vireo.TradBoost_2	0.076

- D_TokyoTech_Canon_4
  - D_vireo.A-SVM_3
    - D_vireo.TradBoost_2

Figure 13: Significant differences among top A-category lite runs

Run name	(mean infAP)
A_TokyoTech_Canon_1	0.149
A_TokyoTech_Canon_2	0.148
A_UvA.Leonardo_1	0.142
A_UvA.Raphael_3	0.141
A_UvA.Donatello_2	0.140
A_TokyoTech_Canon_3	0.138
A_UvA.Michelangelo_4	0.121
A_Quaero1	0.120
A_CMU4	0.120
A_Quaero2	0.119

- A_UvA.Leonardo_1
  - A_UvA.Donatello_2
    - A_CMU4
    - A_Quaero1
    - A_Quaero2
    - A_UvA.Michelangelo_4
  - A_UvA.Raphael_3
    - A_CMU4
    - A_Quaero1
    - A_Quaero2
    - A_UvA.Michelangelo_4
- A_TokyoTech_Canon_1
  - A_TokyoTech_Canon_3
    - A_CMU4
    - A_Quaero1
    - A_Quaero2
    - A_UvA.Michelangelo_4
  - A_TokyoTech_Canon_2
    - A_TokyoTech_Canon_3
      - A_CMU4
      - A_Quaero1
      - A_Quaero2
      - A_UvA.Michelangelo_4

Figure 14: Significant differences among top B-category lite runs

Run name	(mean infAP)
B_dcu.LocalFeatureBoW_2	0.039
B_vireo.SF_web_image_4	0.017

- B_dcu.LocalFeatureBoW_2
  - B_vireo.SF_web_image_4

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

Run name	(mean infAP)
D_TokyoTech_Canon_4	0.148
D_vireo.A-SVM_3	0.082
D_vireo.TradBoost_2	0.054

- D_TokyoTech_Canon_4
  - D_vireo.A-SVM_3
    - D_vireo.TradBoost_2

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

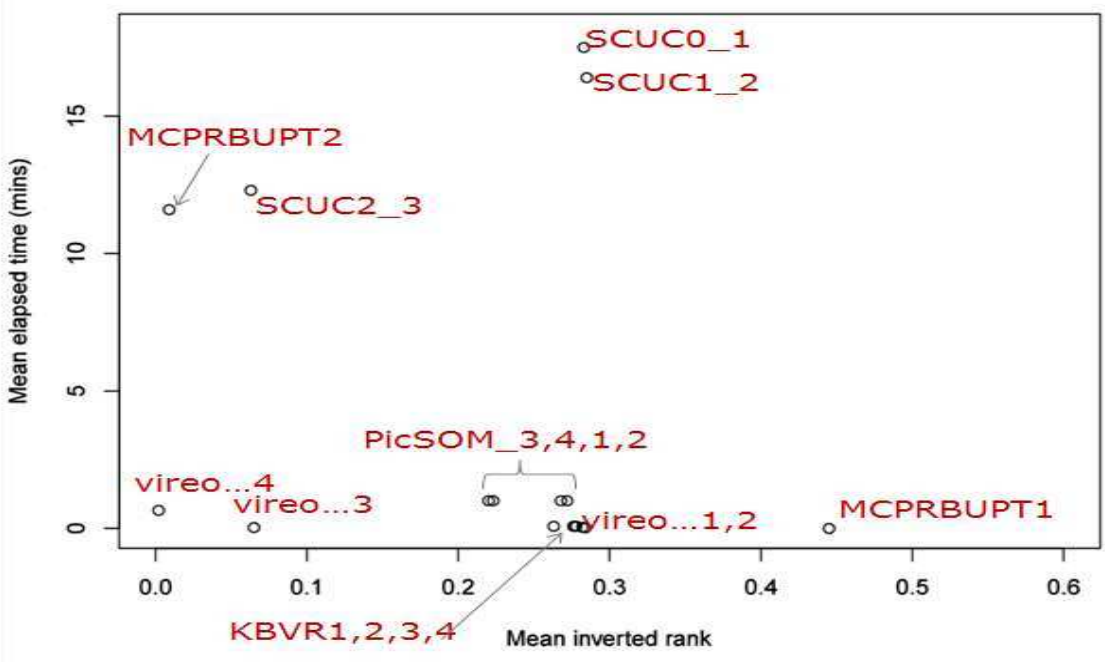


Figure 17: Mean inverted rank versus mean elapsed time for interactive runs

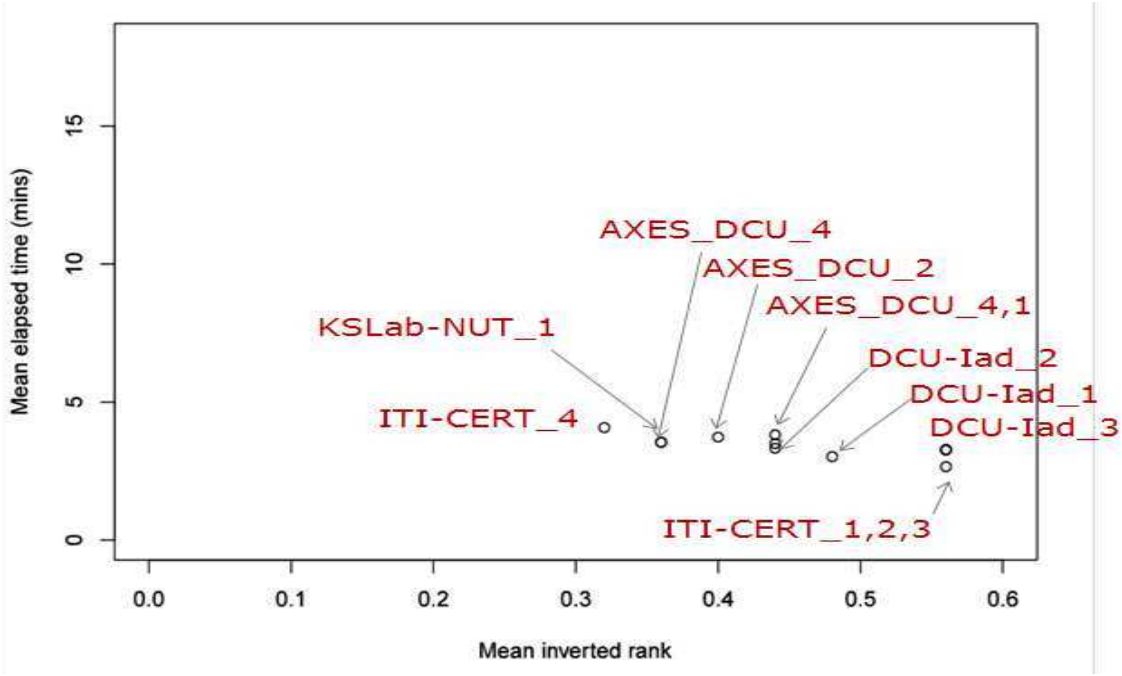


Figure 18: Oracle calls by topic and team

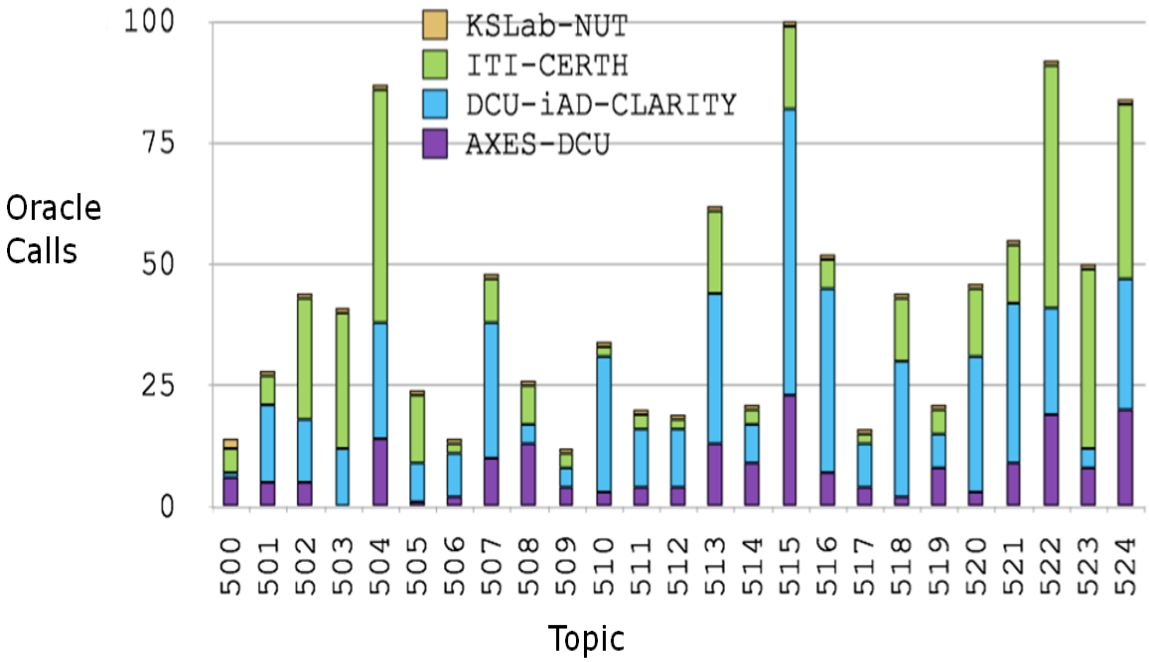


Figure 19: Runs finding known items

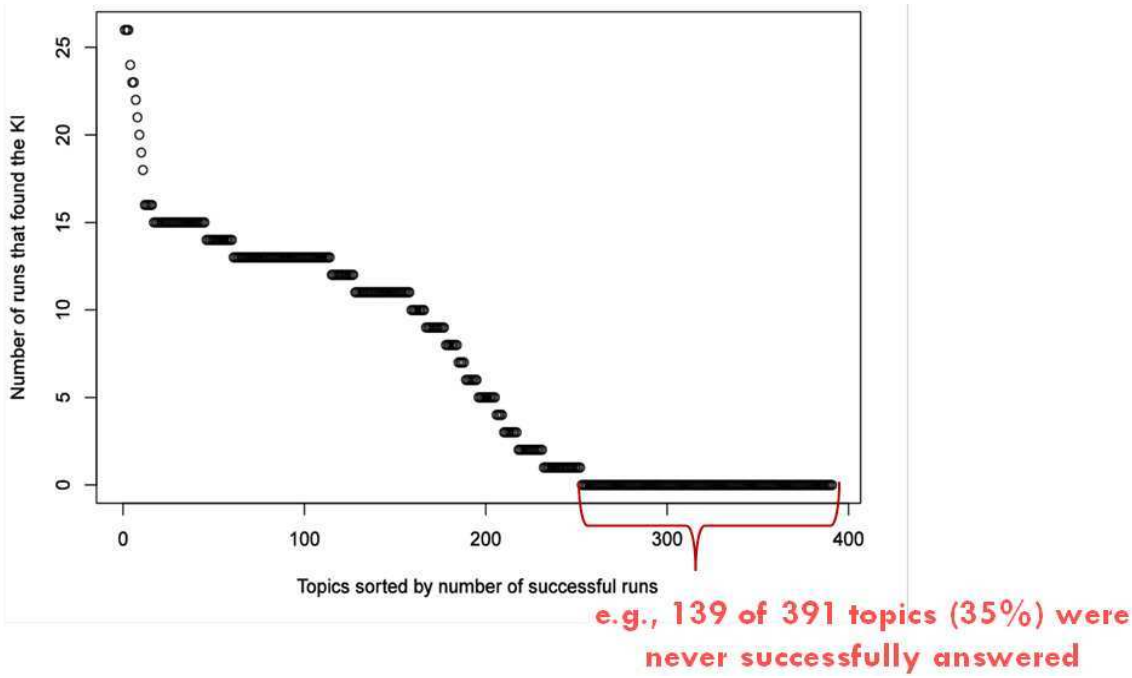


Figure 20: Example segmentations

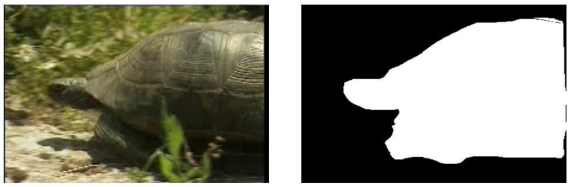


Figure 21: Original versus transformed

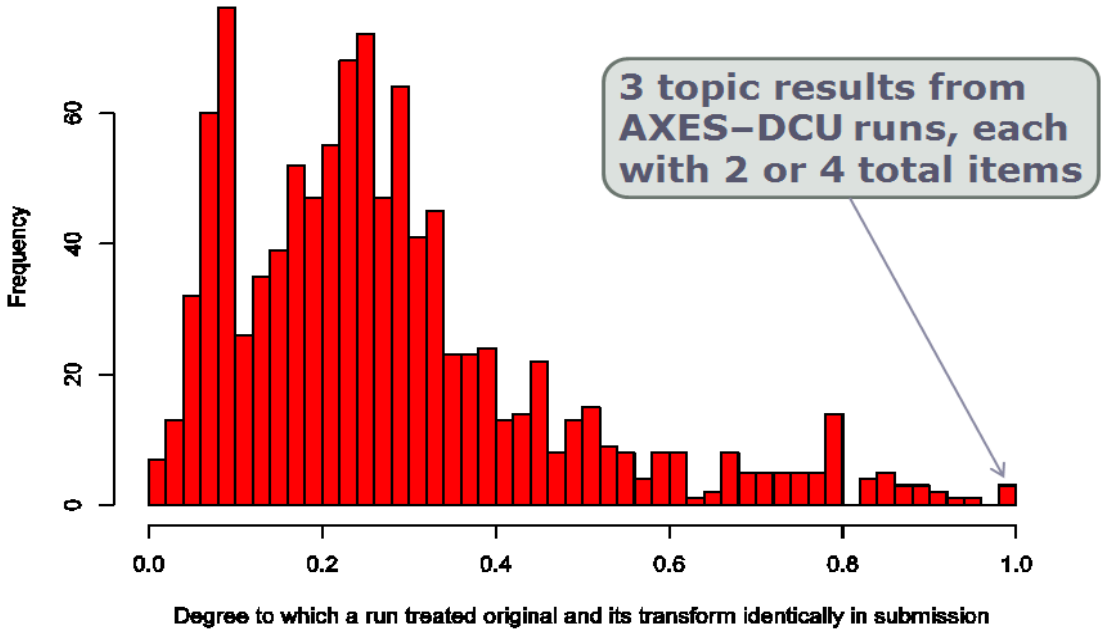


Figure 22: Example object targets 1/3

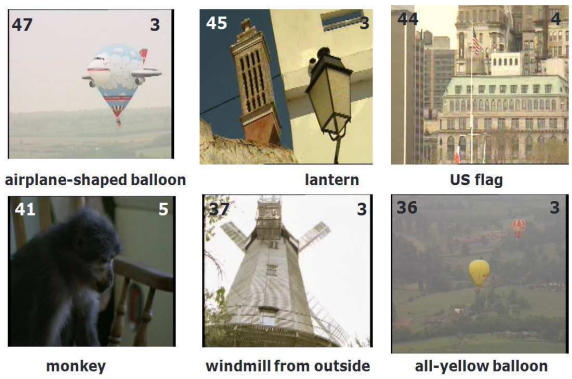


Figure 23: Example object targets 2/3

Figure 24: Example object targets 3/3

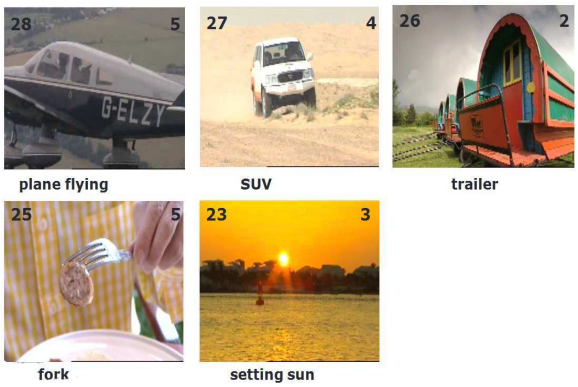


Figure 25: Example people targets

Figure 26: Example location targets

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

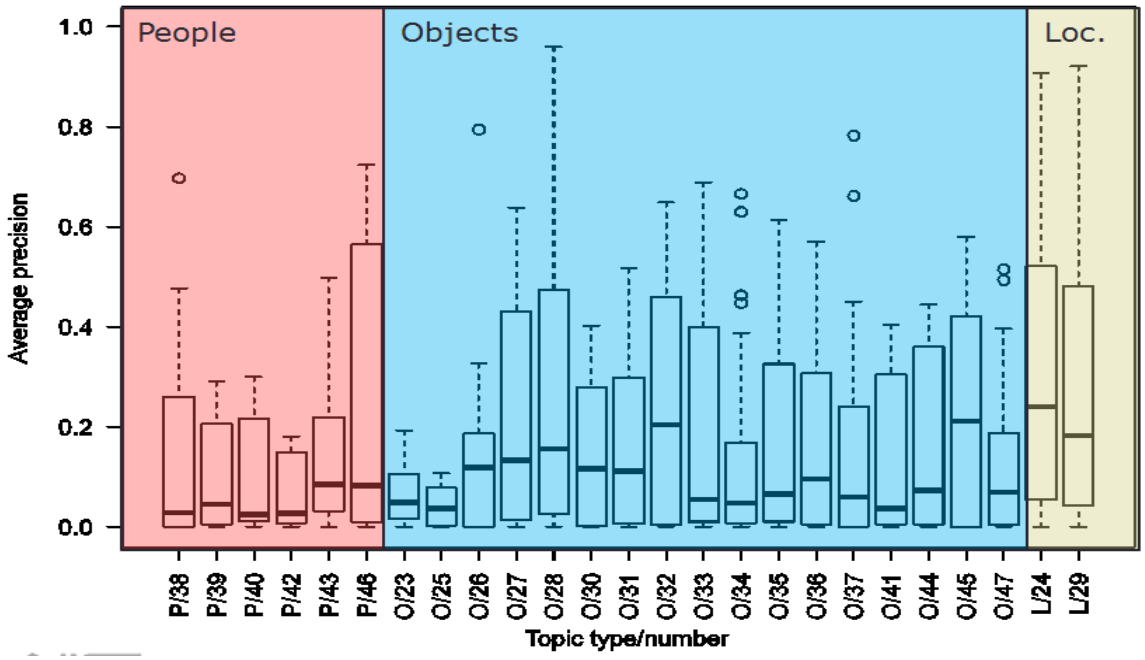


Figure 28: Average precision for interactive runs by topic/type

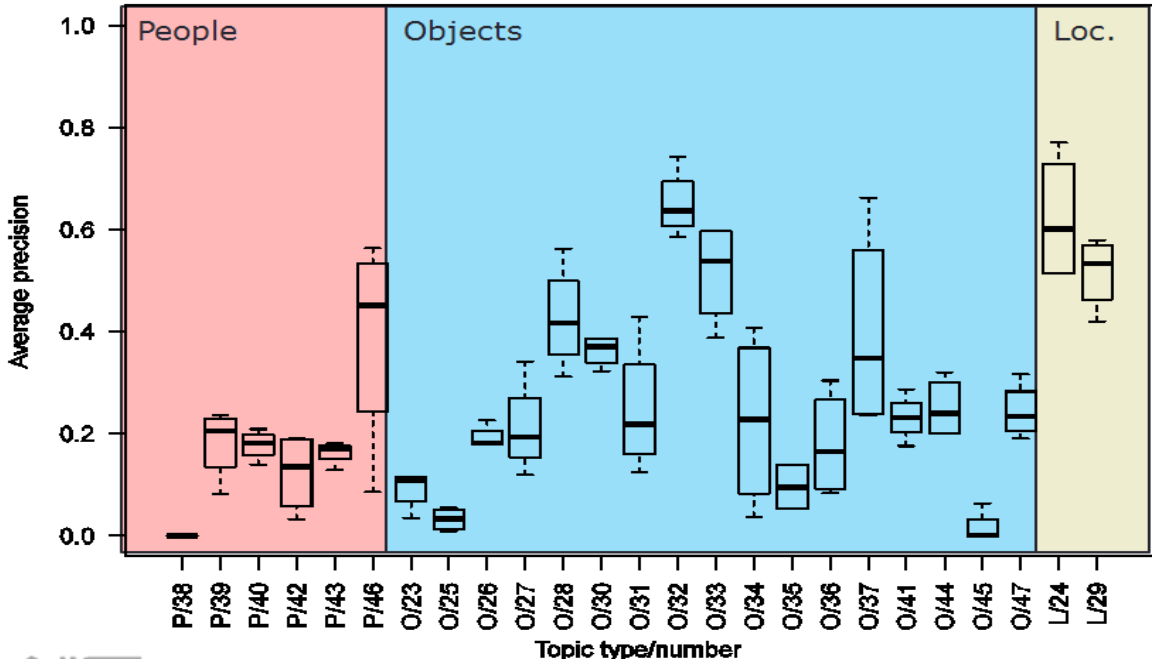


Figure 29: AP by topic for top runs

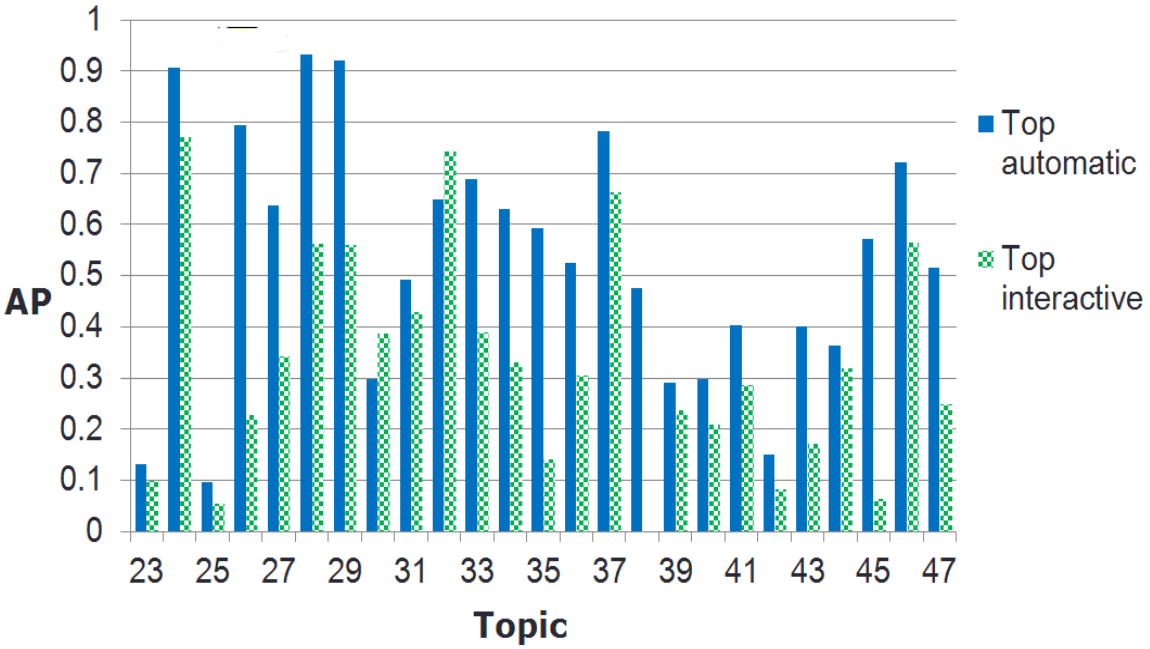


Figure 30: Randomization testing for significant differences

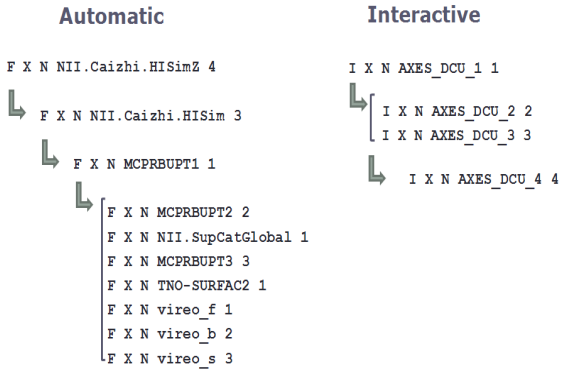


Figure 31: AP vs. number found (automatic)

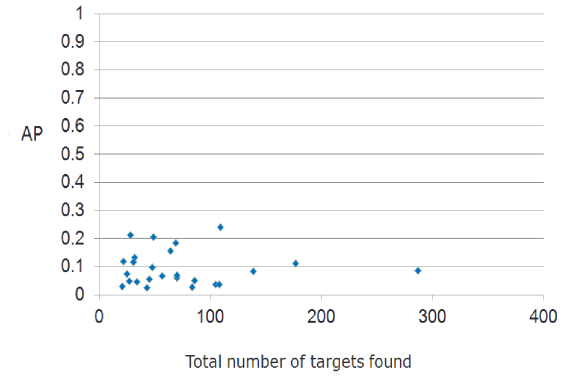


Figure 32: AP vs. number found (interactive)

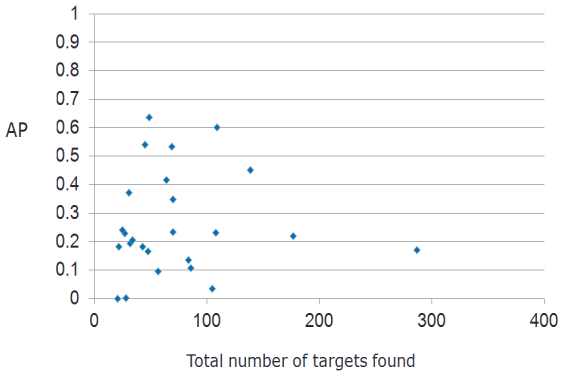


Figure 33: MAP vs. elapsed time

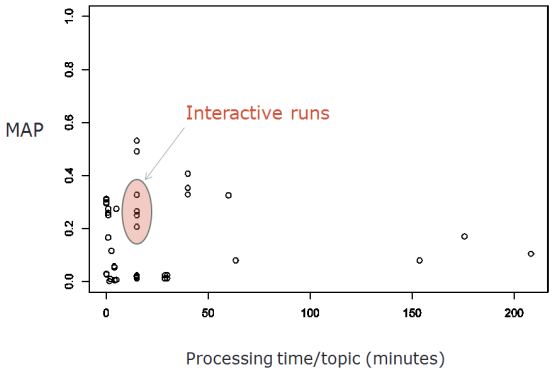


Figure 34: TRECVID 2011 MED: Primary Systems NDCs at the Target Error Ratio

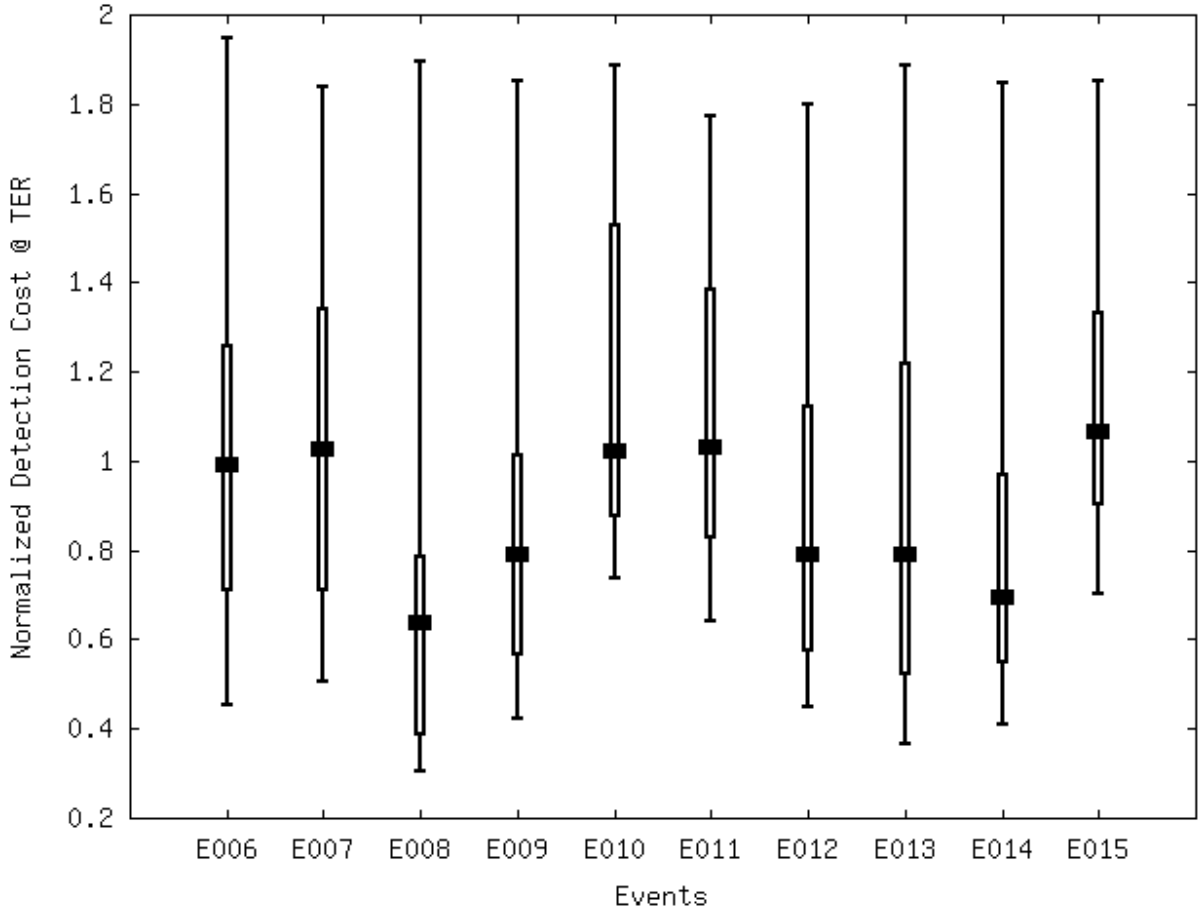


Figure 35: Primary Systems for E008

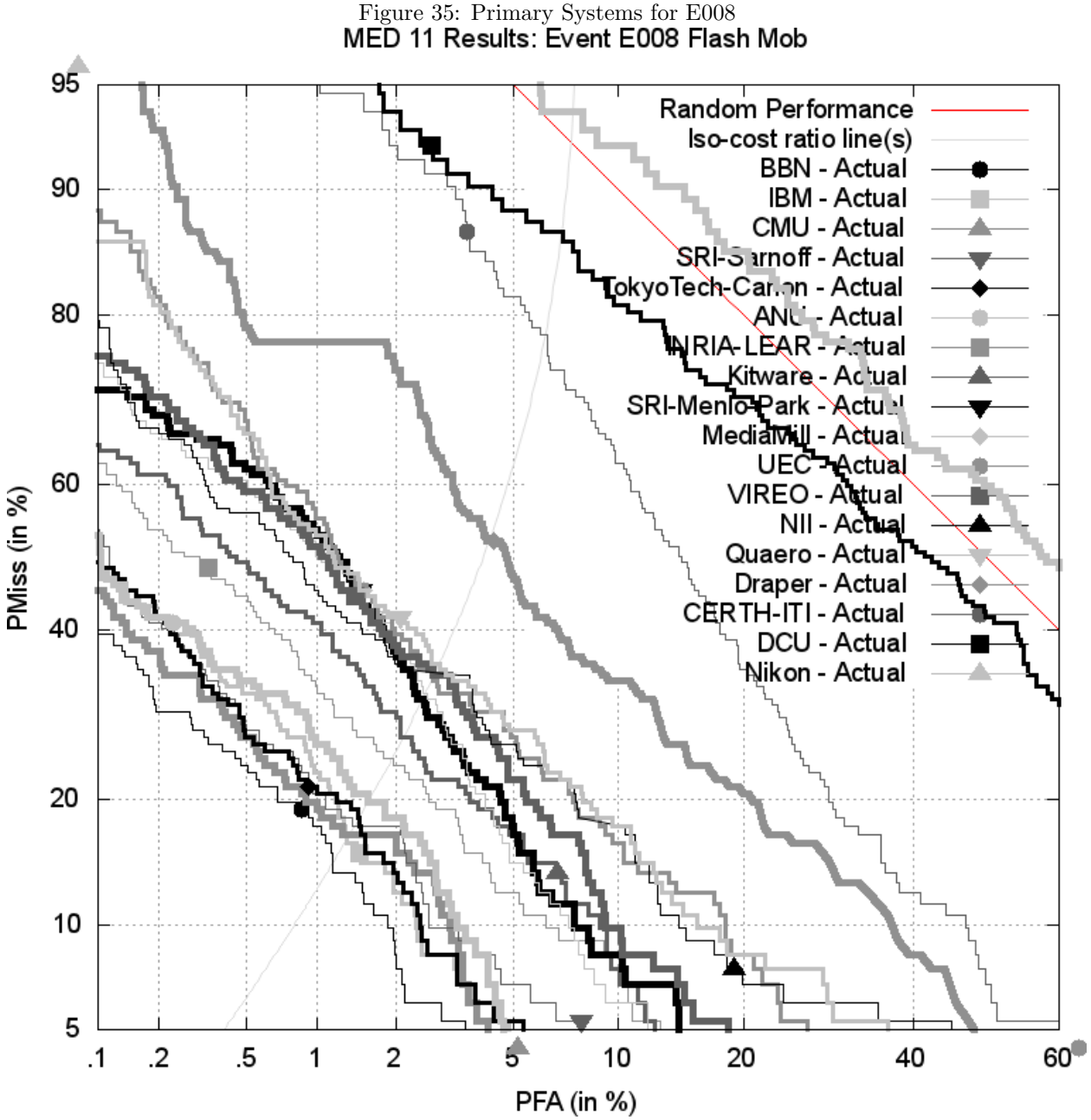


Figure 36: Primary Systems for E015

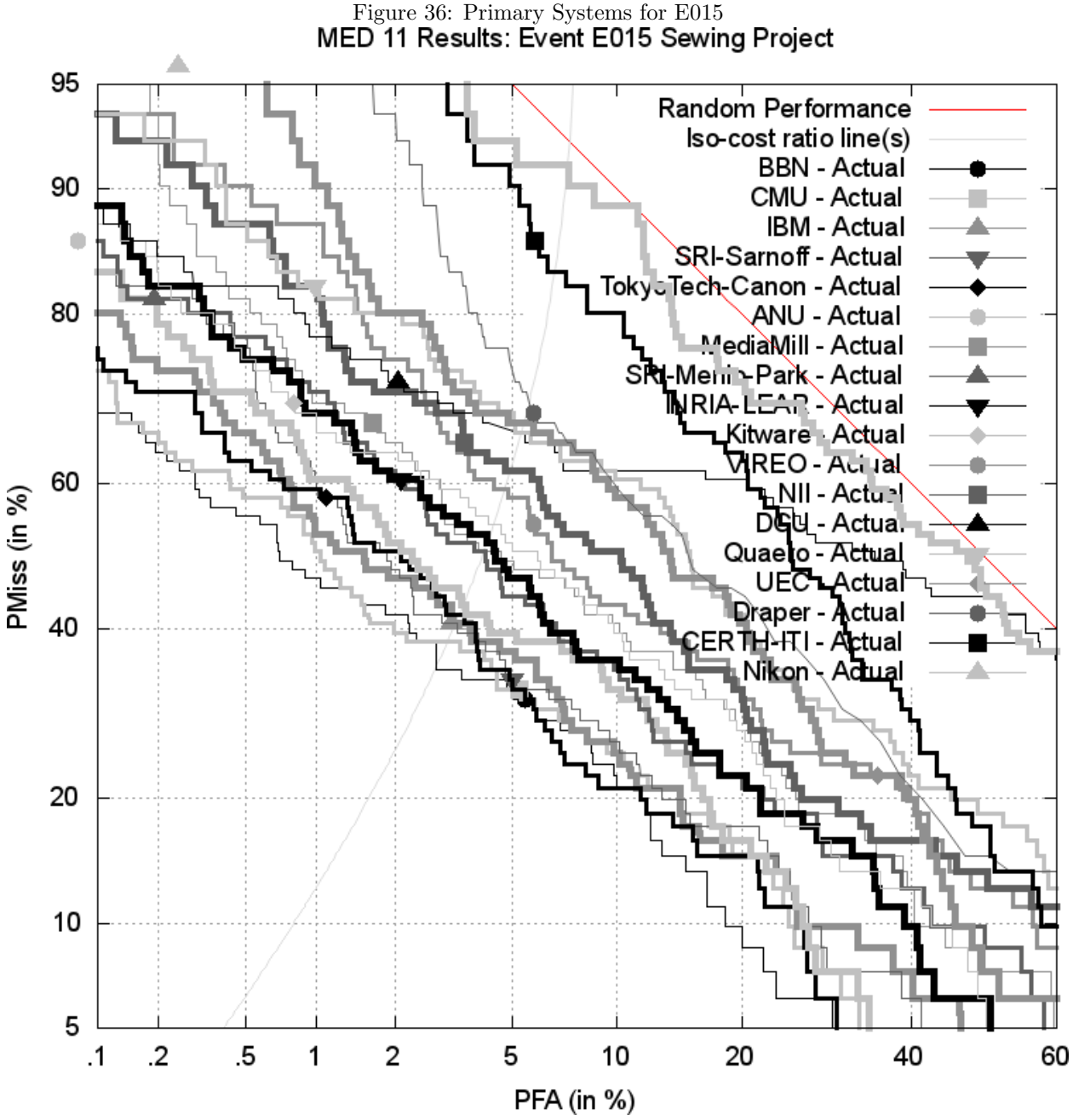


Figure 37: NDDC@TER as a function of Single-Core Normalized CDR Generation Speed

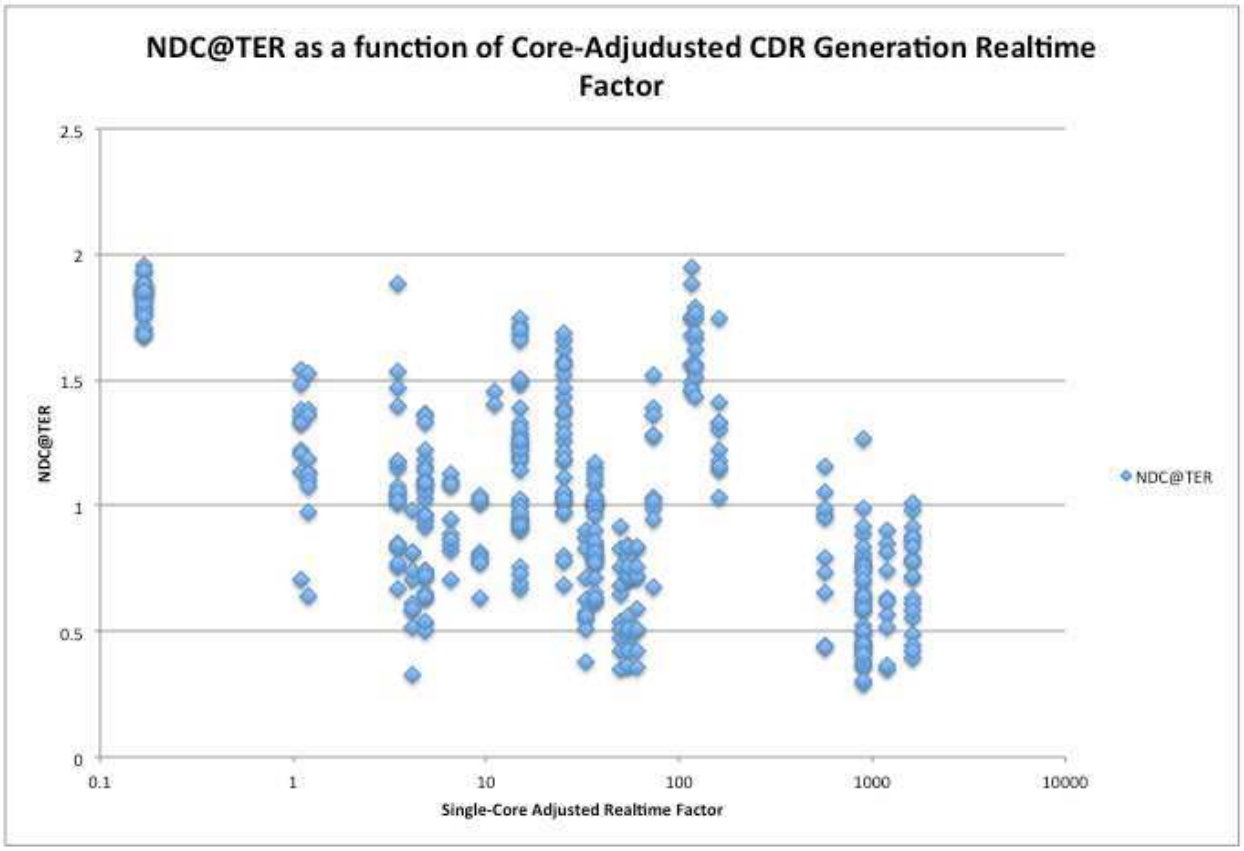


Figure 38: Top runs based on Actual DET score in balanced profile

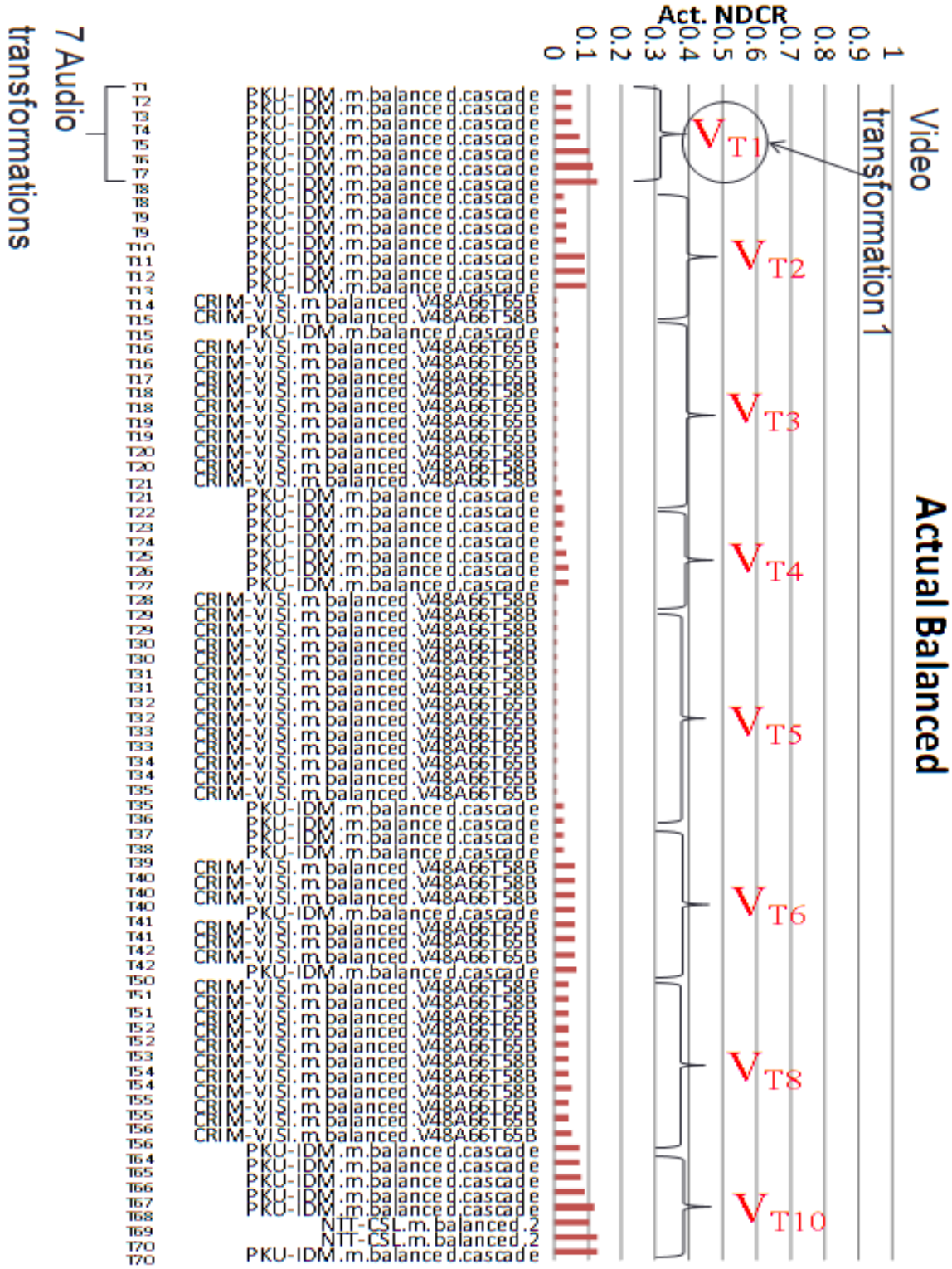


Figure 39: Top runs based on Optimal DET score in balanced profile

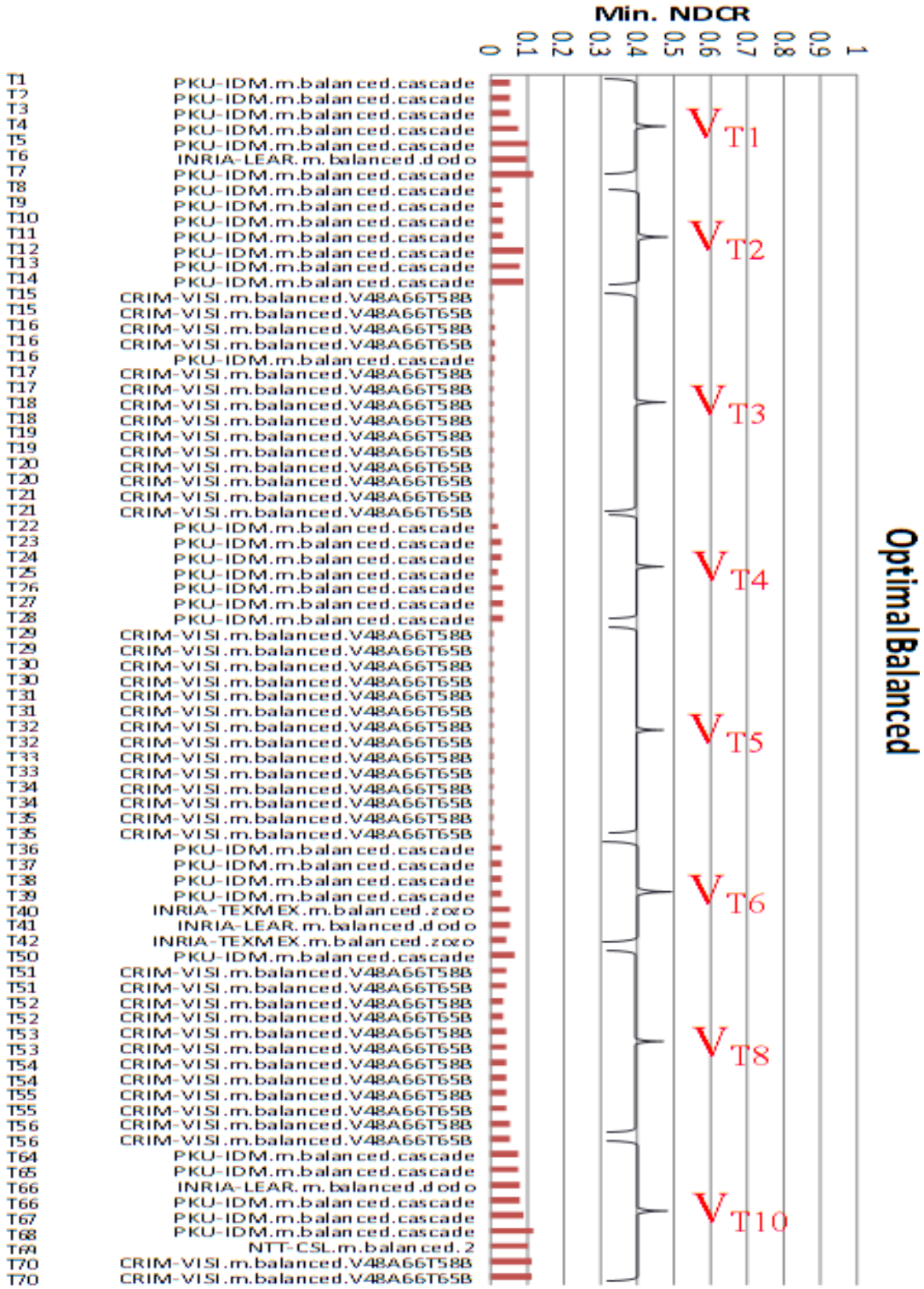


Figure 40: Top runs based on Actual DET score in Nofa profile

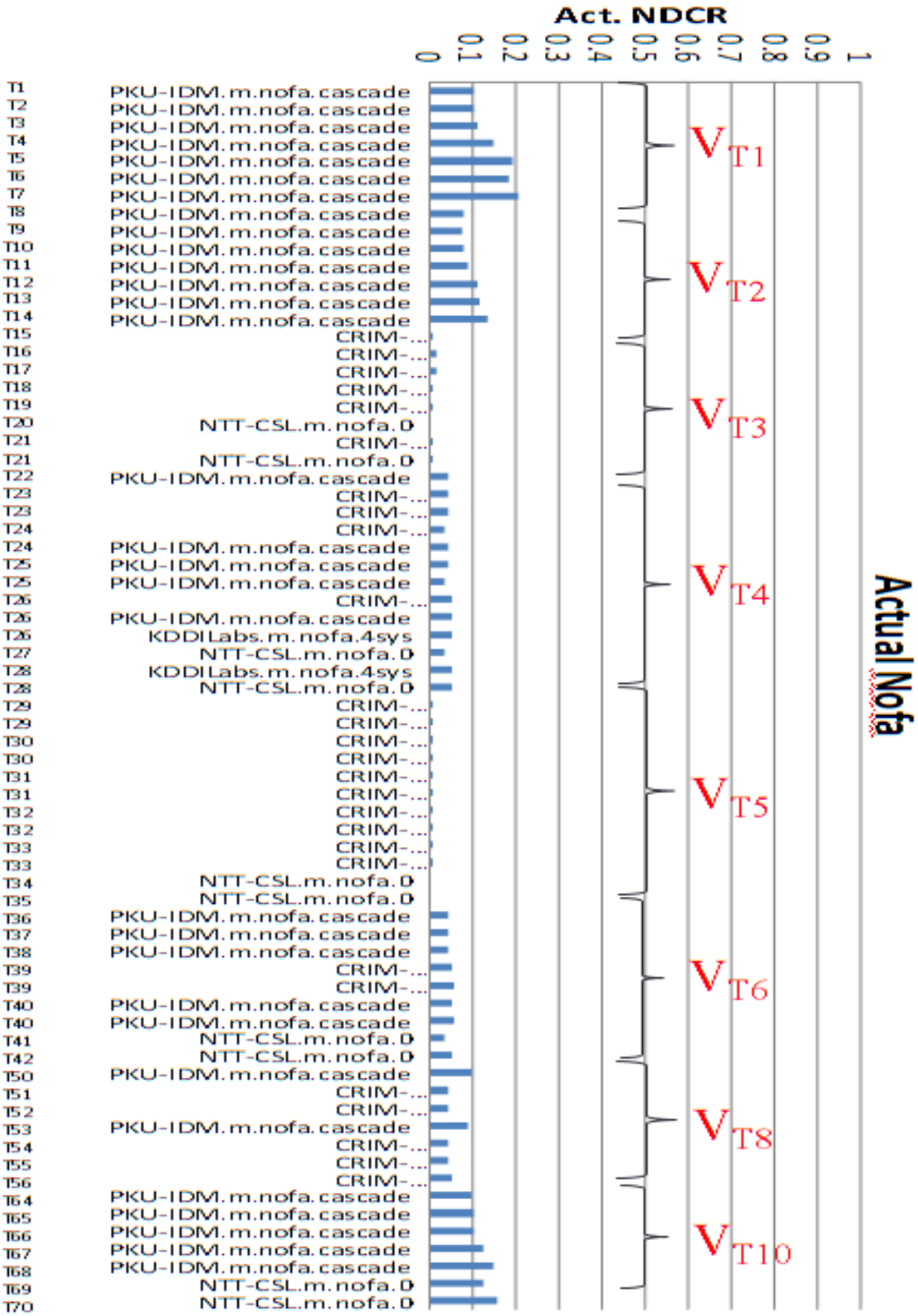


Figure 41: Top runs based on Optimal DET score in Nofa profile

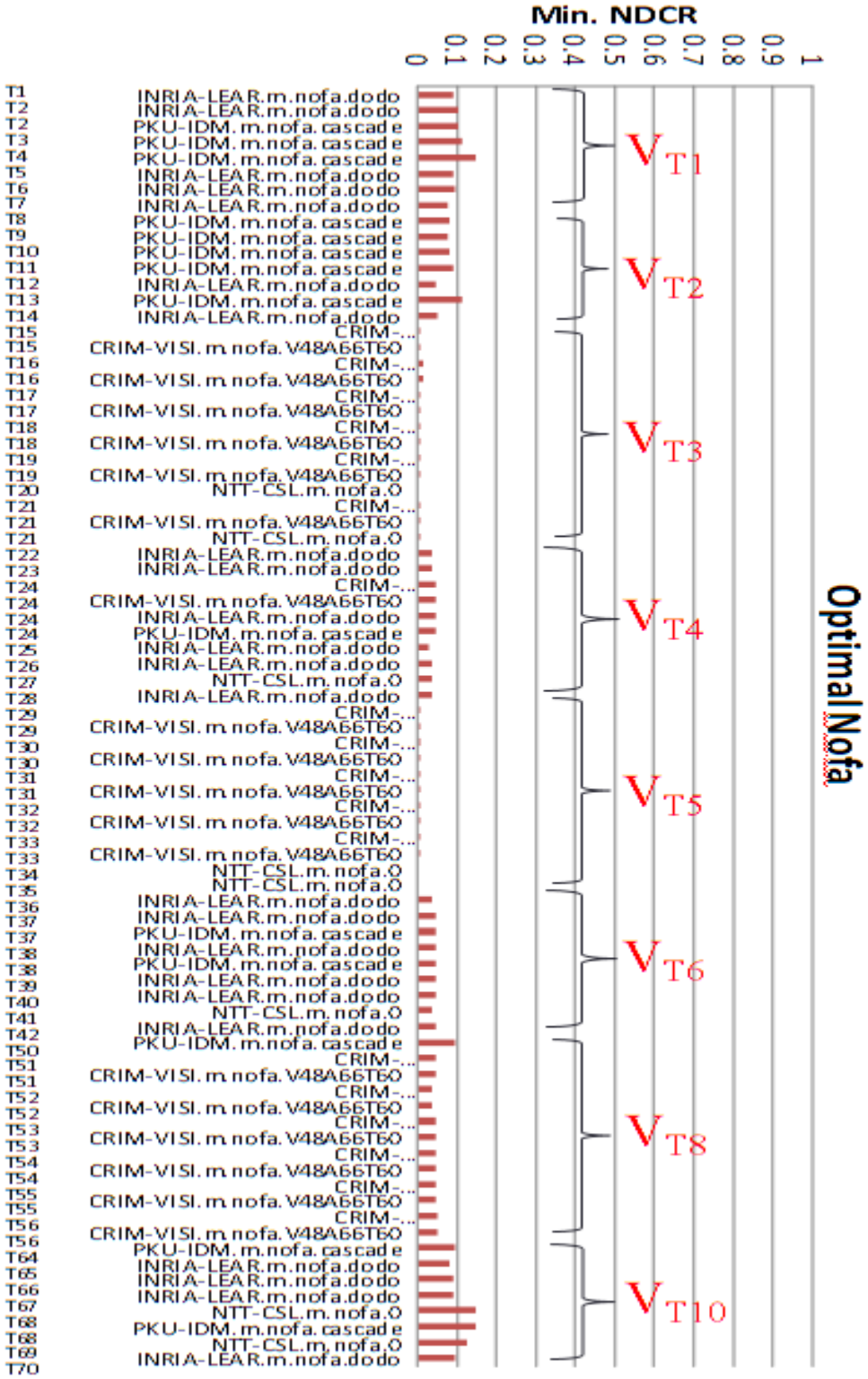


Figure 42: Top 10 runs efficiency in both profiles

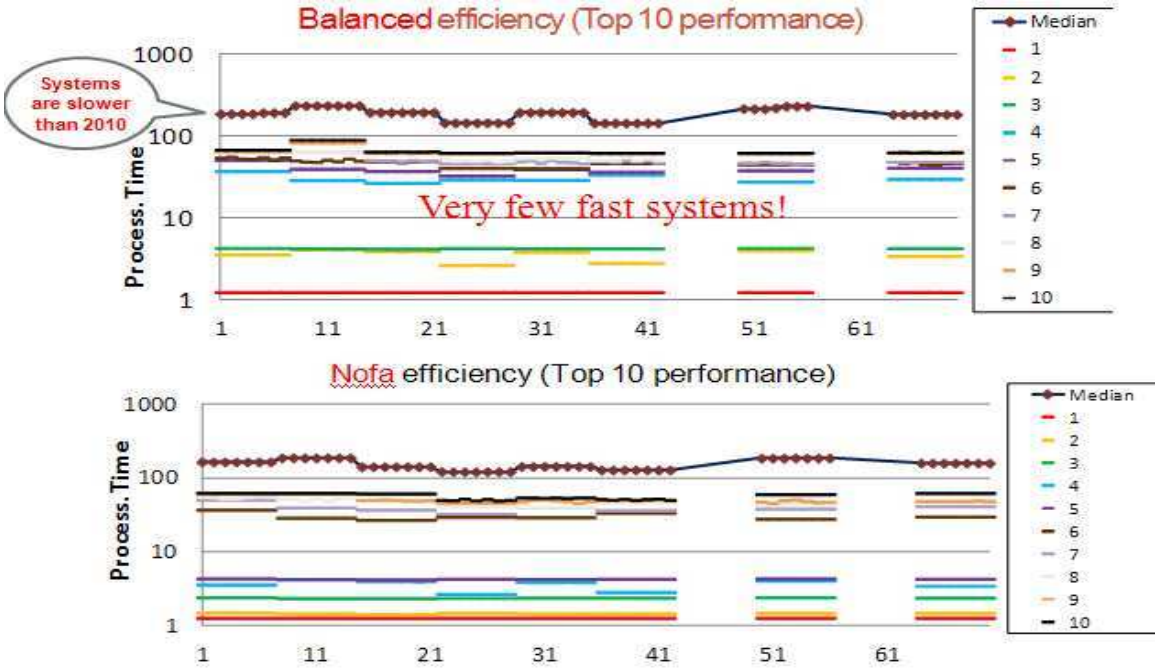


Figure 43: Comparing best runs

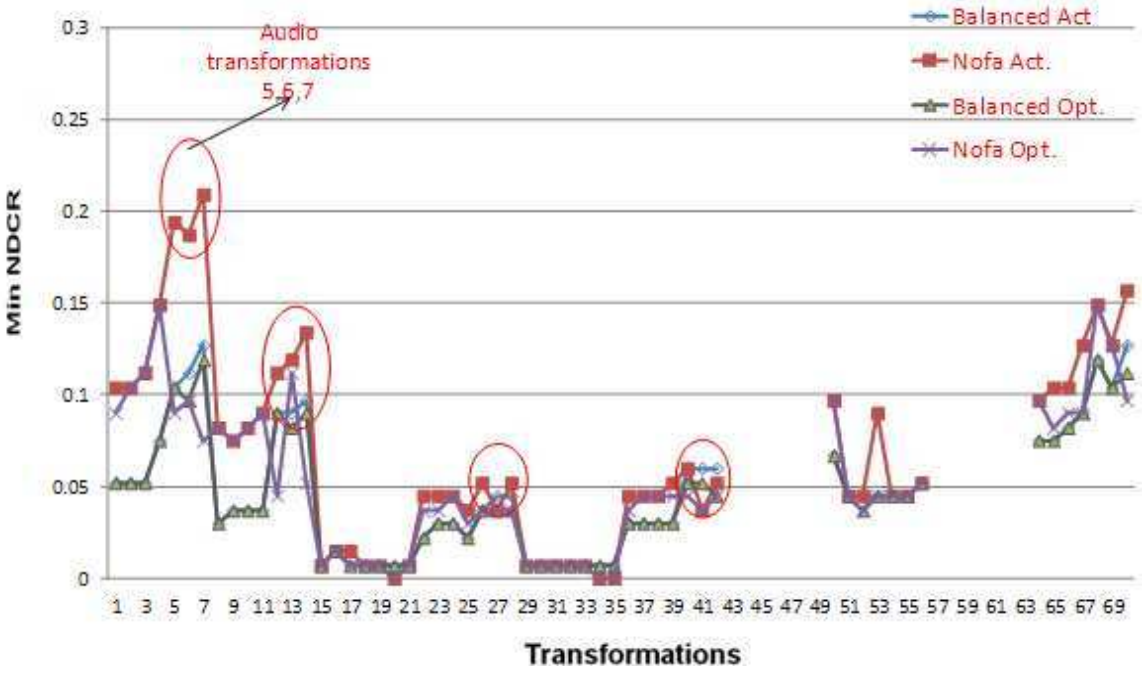


Figure 44: Camera views and coverage

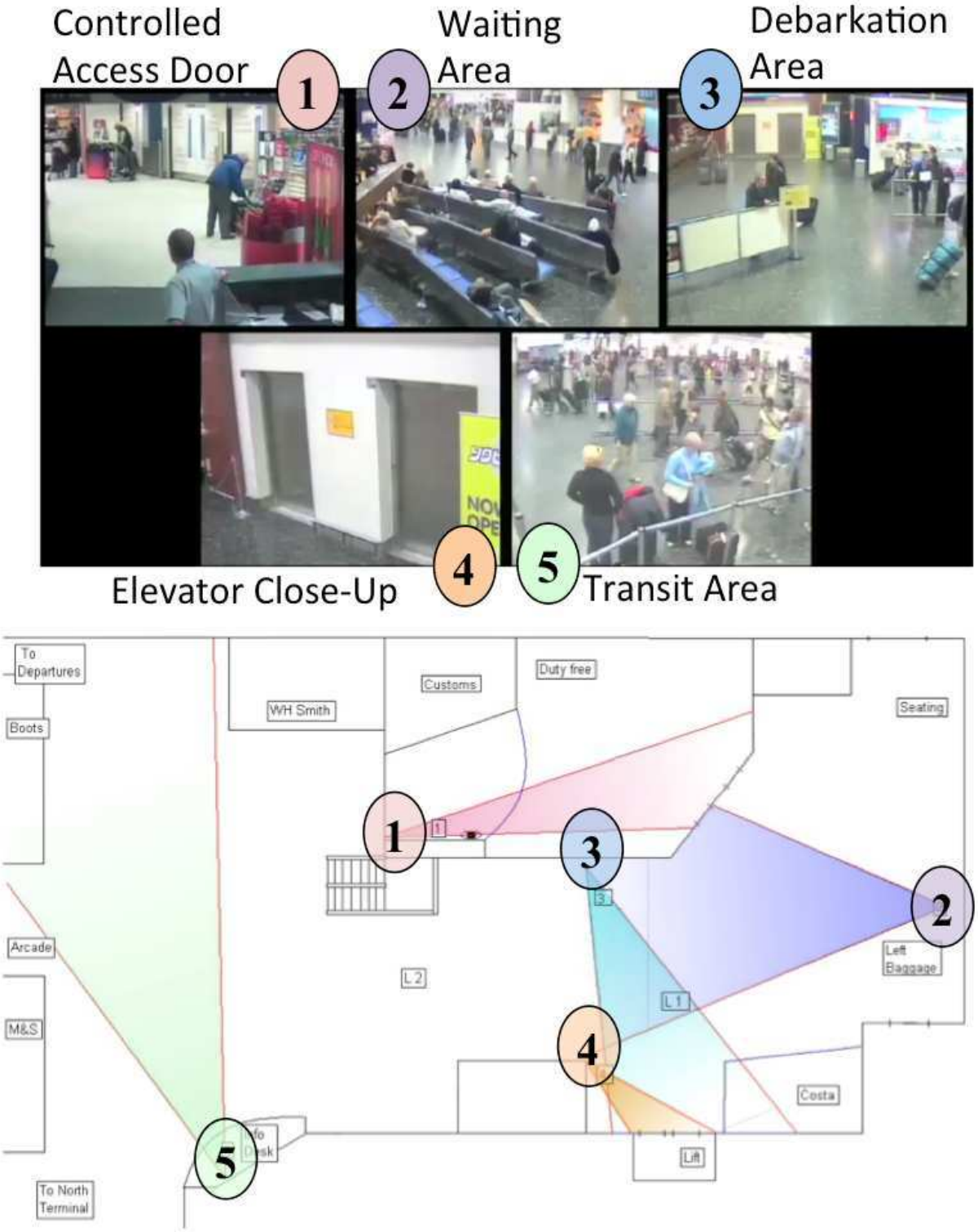


Figure 45: TRECVID 2011 SED Participants Chart

2011 SED Participants (with number of systems per event) 9 participating sites		Single Person		Single Person + Object		Multiple People		
		PersonRuns	Pointing	CellToEar	ObjectPut	Embrace	PeopleMeet	PeopleSplitUp
4 years in a row	Carnegie Mellon University [CMU]	9	9	9	9	9	9	9
	NHK Science and Technical Research Laboratories [NHKSTRL]		1	1	1	1		
3 years in a row	Beijing University of Posts and Telecommunications, MCPRL [BUPT-MCPRL]	2	2		2	2	1	1
	Peking University, NEC [PKU-NEC]		3		3	3	3	3
	Tokyo Institute of Technology, Canon [TokyoTech-Canon]	3					3	3
2 years in a row	Centre de Recherche Informatique de Montréal [CRIM]	1	1	1	1	1		
	Beijing Jiaotong University [BJTU-SED]		1	1				
	Tianjin University [TJUT-TJU]	13	13	13	13	13	13	13
New	Chinese Academy of Sciences [IRDS-CASIA]	5	5	5	5	5	5	5
	Total participants per event	6	8	6	7	7	6	6

Figure 46: 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 47: TRECVID 2011 SED: NDCR Per Event, Primary System Per Site

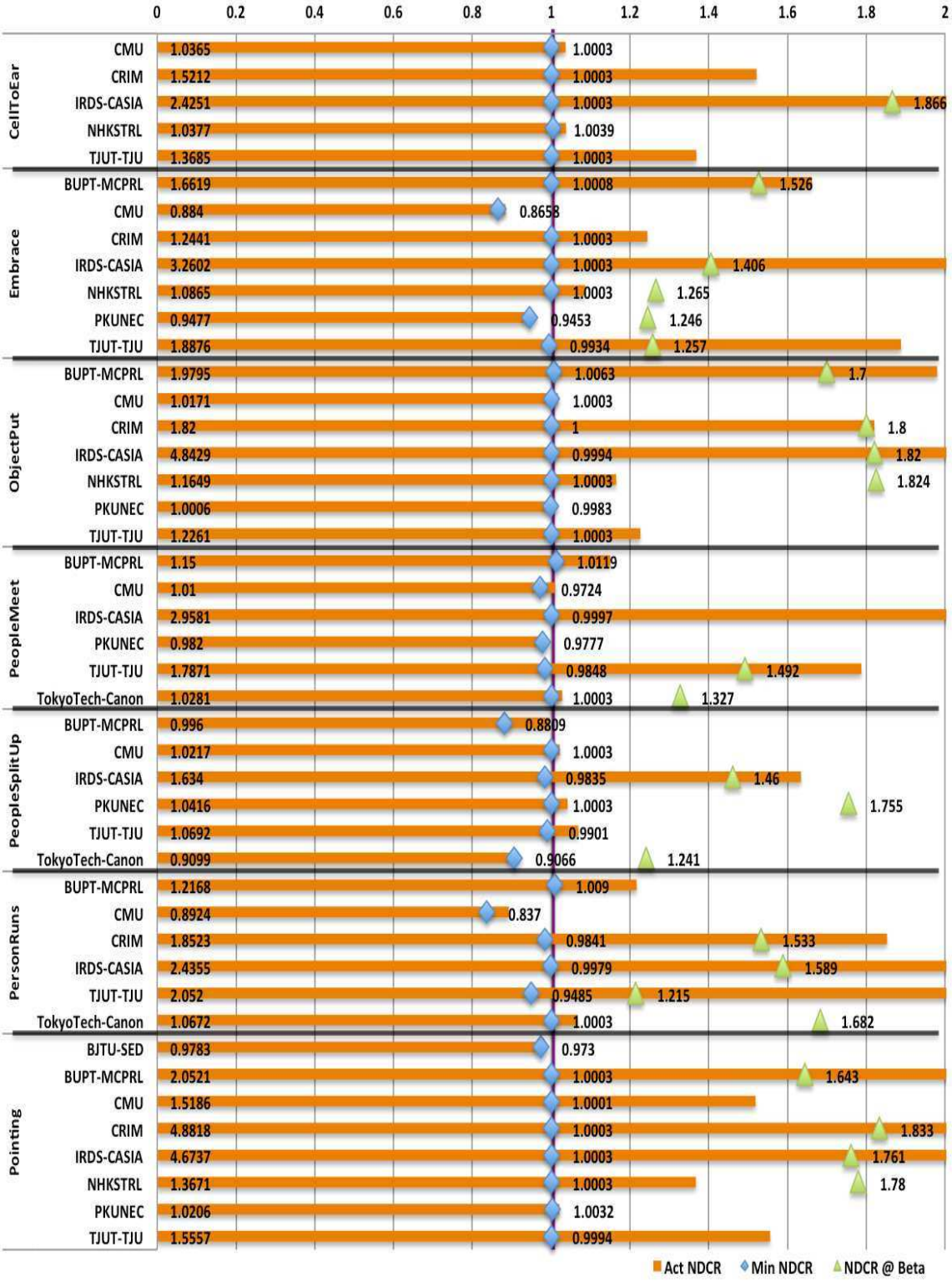


Figure 48: DET Plot Showing SED11 Systems with Lowest Actual NDCR (Per Event, Primary System per Site)

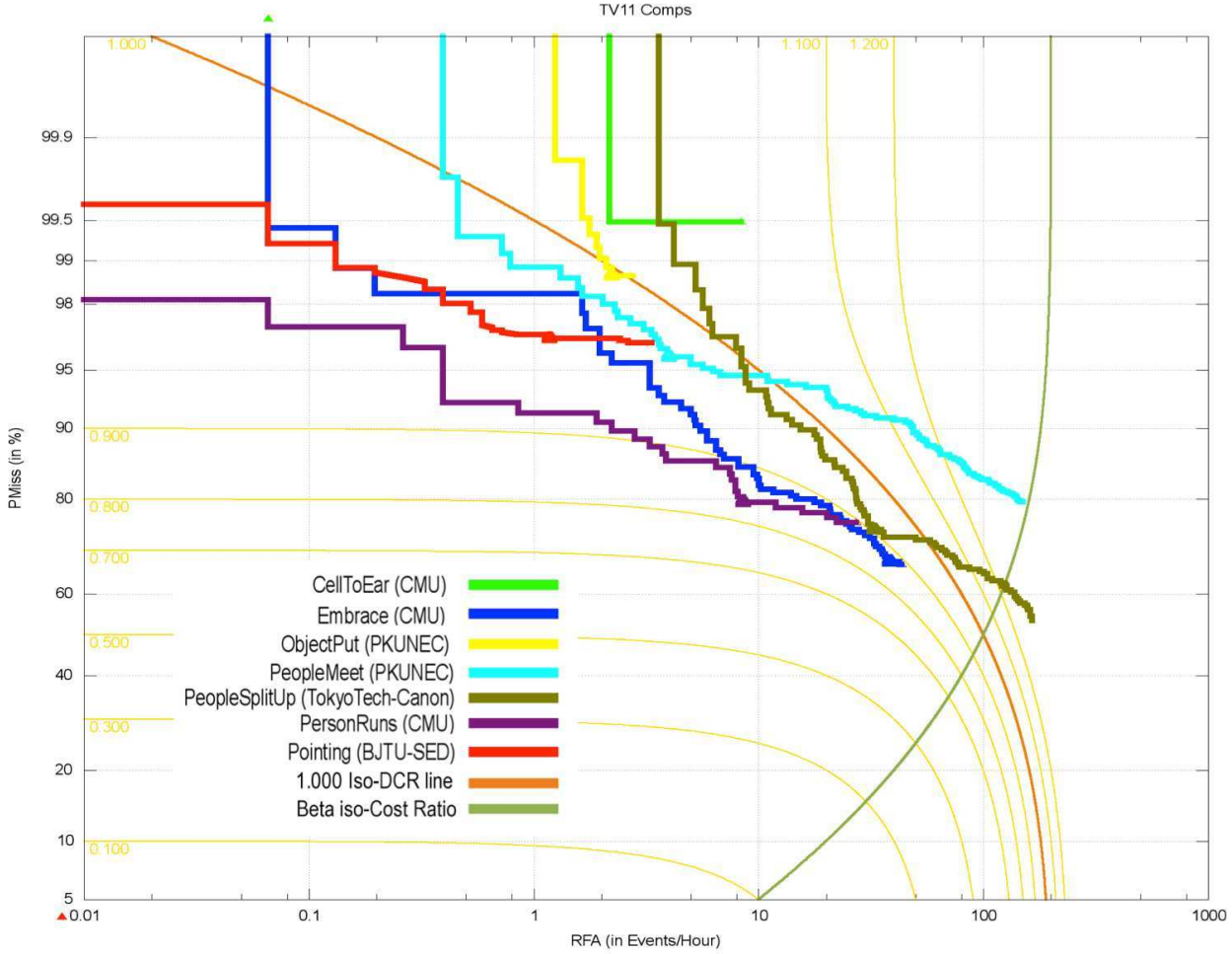


Figure 49: DET Plot Showing SED11 Systems with Lowest NDCR at TOER (Per Event, Primary System per Site)

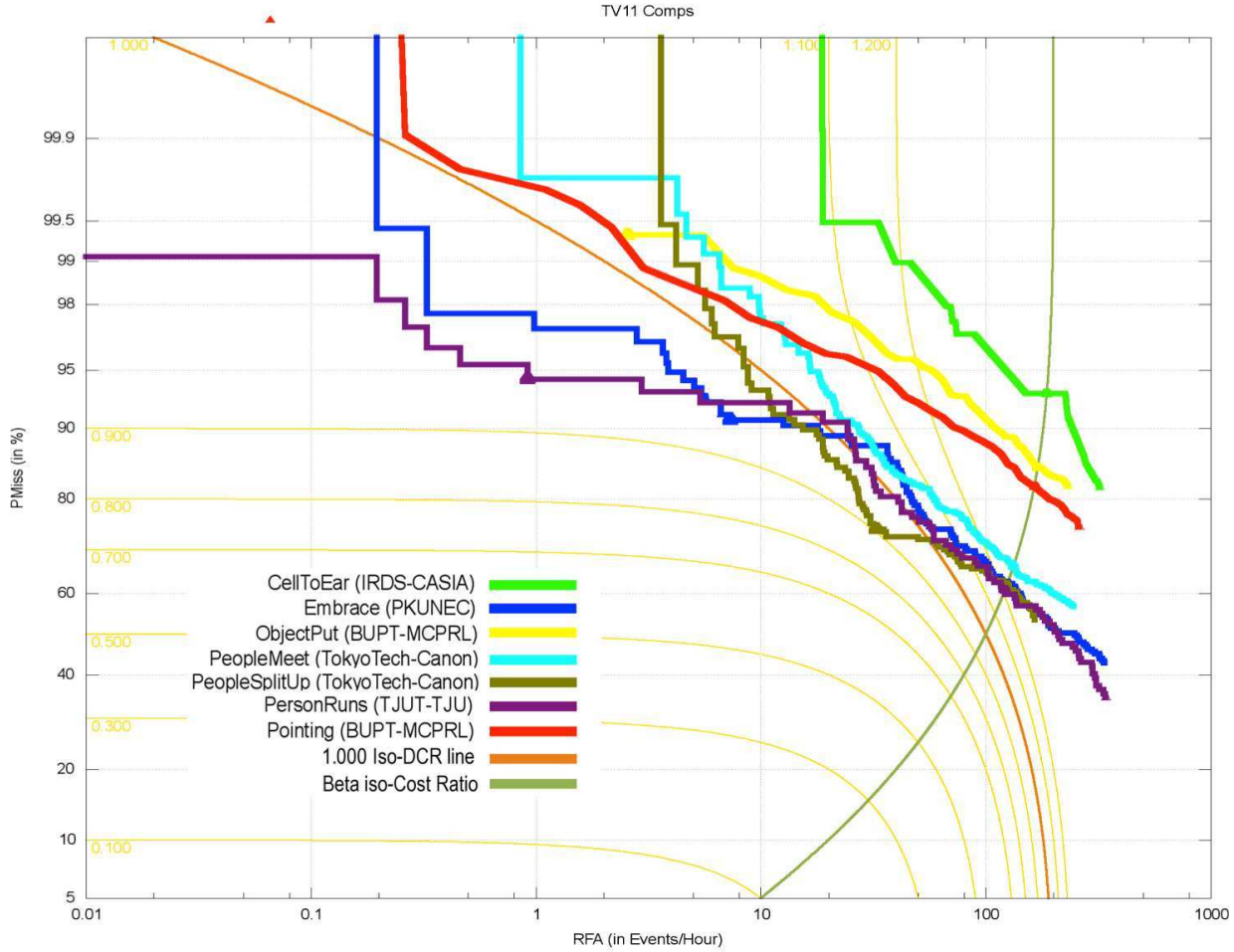
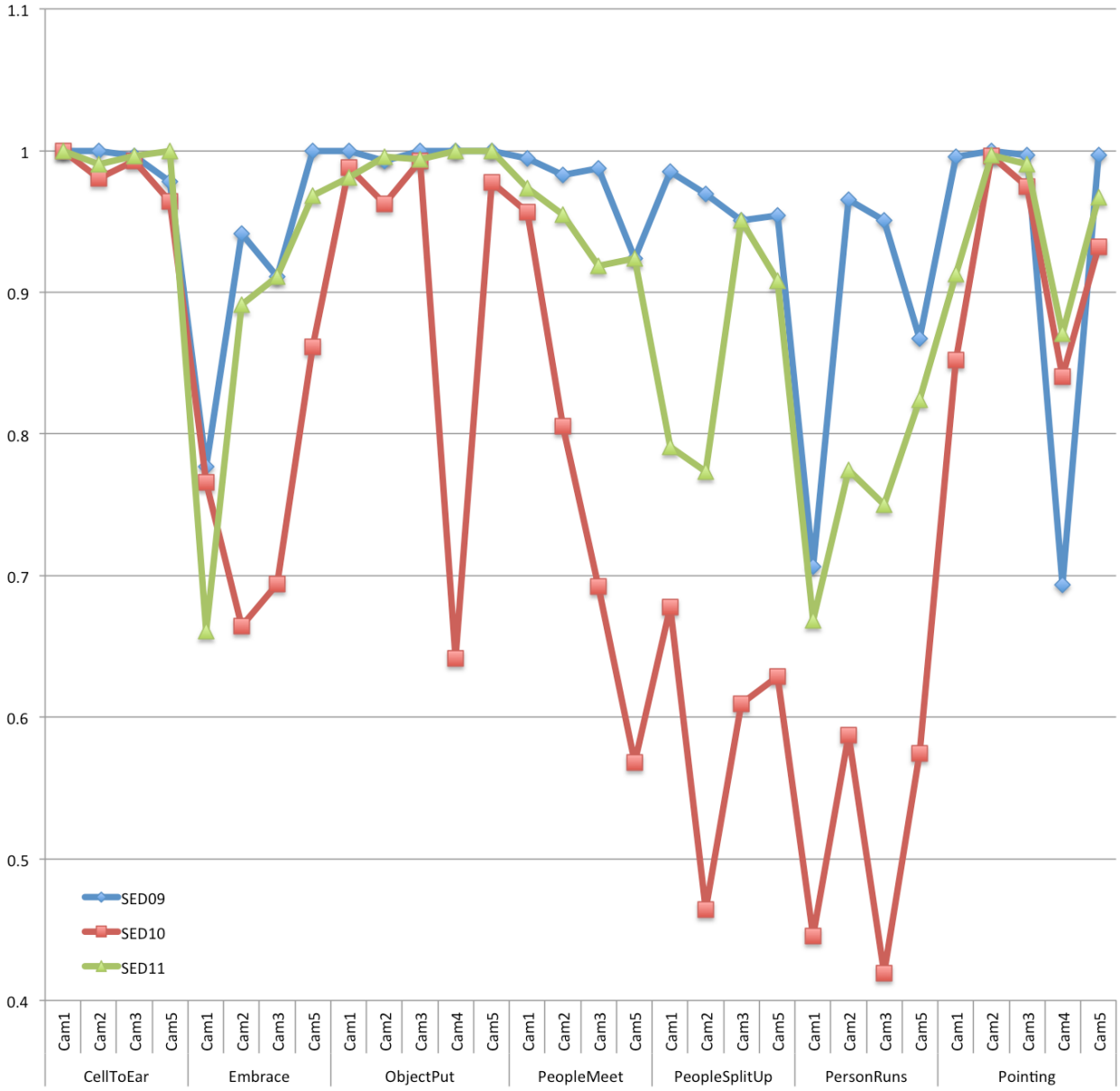


Figure 50: Compared TRECVID SED: Lowest Actual NDCR per camera (Per Event, All Systems)

Figure 51: Compared TRECVID SED: Lowest Minimum NDCR per camera (Per Event, All Systems)



References

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

Ayache, S., & Quenot, G. (2008). Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval (ECIR)* (pp. 187–198). Glasgow, Scotland.

Doerman, D., & Mihalcik, D. (2000). Tools and Techniques for Video Performance Evaluation. In *ICPR* (pp. 167–170).

Fiscus, J., Rose, R., & Michel, M. (2010). *2010 TRECVid Event Detection evaluation plan.* http://www.itl.nist.gov/iad/mig//tests/trecvid/2010/doc/EventDet10-EvalPlan-v01.htm.

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

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

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

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET Curve in Assessment of Detection Task Performance. In *Proceedings of Eurospeech '97* (p. 1899-1903). Rhodes, Greece.

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

Over, P., Awad, G., Fiscus, J., Antonishek, B., Michel, M., Smeaton, A. F., Kraaij, & Quénot, G. W. (2010). *TRECVID 2010 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics.* http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/tv10overview.pdf.

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

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

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

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

Rose, T., Fiscus, J., Over, P., Garofolo, J., & Michel, M. (2009). The TRECVid 2008 Event Detection Evaluation. In *IEEE Workshop on Applications of Computer Vision (WACV).* IEEE.

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

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

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