TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics

Paul Over {over@nist.gov}
Jon Fiscus {jfiscus@nist.gov}
Gregory Sanders {gregory.sanders@nist.gov}
David Joy {david.joy@nist.gov}
Martial Michel {martial.michel@nist.gov}
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

George Awad
Dakota Consulting, Inc.
1110 Bonifant Street, Suite 310
Silver Spring, MD 20910
{gawad@nist.gov}

Alan F. Smeaton {Alan.Smeaton@dcu.ie}
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland

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

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

April 11, 2016

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2014 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 dozen years this effort has yielded a better understanding of how systems can effectively accomplish such processing and how one can reliably benchmark their performance. TRECVID is funded by the NIST with support from other US government agencies. Many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2014 represented a continuation of five tasks from 2013. 40 teams (see Tables 1 and 2) from various research organizations worldwide completed one or more of five tasks:

1. Semantic indexing
2. Instance search
3. Multimedia event detection
4. Multimedia event recounting
5. Surveillance event detection

Some 200 hours of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC.2), were used for semantic indexing. 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 464 h of BBC EastEnders video was reused for the instance search task. 45 h of airport surveillance video was reused for the surveillance event detection task. Almost 5 200 hours from the HAVIC collection of Internet videos was used for

development and testing in the multimedia event detection task.

Semantic Indexing and Instance search results were judged by NIST assessors. Multimedia event detection was scored by NIST using ground truth created manually by the Linguistic Data Consortium under contract to NIST. The multimedia event recounting task was judged by humans experts in an evaluation designed by NIST. Surveillance event detection was scored by NIST using ground truth created by NIST through manual adjudication of test system output.

This paper is an overview of the evaluation framework — the tasks, data, and measures for the workshop. For detailed information about the approaches and results, the reader should see the various site reports (TV14Pubs, 2014) and the results pages in the online workshop notebook (TV14Notebook, 2014) (part of the Guidelines) on the TRECVID website (trecvid.nist.gov).

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

2 Data

2.1 Video

BBC EastEnders video

The BBC in collaboration the European Union's AXES project made 464 h of the popular and long-running soap opera EastEnders available to TRECVID for research. The data comprise 244 weekly "omnibus" broadcast files (divided into 471 527 shots), transcripts, and a small amount of additional metadata.

Internet Archive Creative Commons (IACC.2) video

7300 Internet Archive videos (144 GB, 600 h) with Creative Commons licenses in MPEG-4/H.264 format with duration ranging from 10 s to 6.4 min and a mean duration of almost 5 min. Most videos have some metadata provided by the donor available e.g., title, keywords, and description.

For 2013 - 2015, approximately 600 hours of Internet Archive videos with Creative Commons licenses in MPEG-4/H.264 and with durations between 10 seconds and 6.4 min were used as new test data. This data was randomly divided into 3 datasets: IACC.2.A, IACC.2.B, and IACC.2.C. IACC.2.B was the test dataset for semantic indexing in 2014. Most videos had some donor-supplied metadata available e.g., title, keywords, and description. Approximately another 600 h of IACC.1 videos were available for system development.

As in the past, LIMSI and Vocapia Research provided automatic speech recognition (Gauvain, Lamel, & Adda, 2002) for the English speech in the IACC.2 video.

iLIDS Multiple Camera Tracking Data

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

The training video consisted of the ≈100 h of data used for SED 2008 evaluation. The evaluation video consisted of the same additional ≈50 h of data from Imagery Library for Intelligent Detection System's (iLIDS) multiple camera tracking scenario data used for the 2009 - 2013 evaluations (UKHO-CPNI, 2007 (accessed June 30, 2009)).

In 2014, system performance was assessed on an 11-hour subset of the evaluation corpus. The subset contained 8 h different from the subset used in previous years and 3 h reused. The overlap allowed some comparison of earlier versus new groundtruthing. The same set of seven events used since 2011 were evaluated.

Heterogeneous Audio Visual Internet (HAVIC) Corpus

The HAVIC Corpus (Strassel et al., 2012) 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 Advanced Audio Coding (ACC) (ACC, 2010) encoded audio.

The HAVIC systems used the same, LDC-provided development materials as in 2013 but teams were also

able to use site-internal resources. The LDC-provided data included:

- Event kits [290 h] (event training material for 40 events),
- Research Resources [314 h] (development resources composed of MED11 Development data and a portion of the MED11 Test data that may be altered, amended or annotated in any way participants need to facilitate their research),
- MEDTest [837 h] (a site-internal testing data set composed of a subset of the MED11 Test data that is structured as fixed background [non-event] clip set and additional positive examples for test events),
- KindredTest [675 h] (an internal testing data structured as a fixed set of background [non-event] clips that contain a 'city building exterior' and the same event positives as used in the MEDTest collection).

The evaluation corpus was doubled this year to be 7 580 hours of video call MED14-EvalFull. The data set consisted of the 3 722 hour MED Progress Collection and a new, 3 858 hour data set called HAVIC Novel1. Teams could choose to process either the full evaluation collection or a 1 238 hour subset called MED14-EvalSub.

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 at the Laboratoire d'Informatique de Grenoble.

3.1 System task

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.

Judges at NIST followed several rules in evaluating system output. 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. In concept definitions, "contains x" or words to that effect are short for "contains x to a degree sufficient for x to be recognizable as x to a human" . This means among other things that unless explicitly stated, partial visibility or audibility may suffice. The fact that a segment contains video of a physical object representing the concept target, such as photos, paintings, models, or toy versions of the target, was NOT grounds for judging the concept to be true for the segment. Containing video of the target within video may be grounds for doing so.

Two novelties were introduced as pilot extensions to the participants in 2013 and were continued in 2014:

- measurement of system progress for a fixed set of concepts and independent of the test data, across 3 years (2013-2015)
- a new optional "localization" subtask with the goal of spatially and temporally localizing 10 detected concepts inside the I-Frames of the video shots

500 concepts were selected for the TRECVID 2011 semantic indexing task. In making this selection, the organizers drew from the 130 used in TRECVID 2010, the 374 selected by CU/Vireo for which there exist annotations on TRECVID 2005 data, and some from the LSCOM ontology. From these 500 concepts, 346 concepts were selected for the full task in 2011 as those for which there exist at least 4 positive samples in the final annotation. Similarly to 2013 the same list of 60 single concepts were used this year for which participants must submit results in the main task. Also, the same 10 concepts for localization used in 2013 were again chosen as a subset of the main task concepts.

In 2014 the task again supported experiments using the "no annotation" version of the tasks: the

idea was to promote the development of methods that permit the indexing of concepts in video shots using only data from the web or archives without the need of additional annotations. The training data could for instance consist of images retrieved by a general purpose search engine (e.g., Google) using only the concept name and/or definition with only automatic processing of the returned images. This was again be implemented by using the additional categories if "E" and "F" for the training types besides the A to D ones.

There was a change to a stricter interpretation of the system training categories A through D - namely, *all* data used for training at any level of any system component should be considered.

- A - used only IACC training data
- B - used only non-IACC training data
- C - used both IACC and non-IACC TRECVID (S&V and/or Broadcast news) training data
- D - used both IACC and non-IACC non-TRECVID training data

This means that even just the use of something like a face detector that was trained on non-IACC training data would disqualify the run as type A. This implied that some systems accepted in category A in the previous years were placed in categories B, C or D with the new rules.

Three types of submissions were considered: "main" in which participants submitted results for 60 single concepts, "loc" in which main task participants submitted localization results for 10 concepts drawn from the 60 main concepts, and finally "progress" in which participants submitted independent results for all and only the 60 main task concepts but against the IACC.2.B, and IACC.2.C data.

TRECVID evaluated 30 of the 60 submitted single concept results and all of the 10 submitted concept localization results. The 60 single concepts are listed below. Those that were evaluated in the main task are marked with an asterisk. The subset evaluated for localization are marked with ">".

3 * > Airplane
5 Anchorperson
6 Animal
9 * Basketball
10 * Beach
13 * Bicycling
15 * > Boat_Ship
16 Boy
17 * > Bridges
19 * > Bus
22 Car_Racing
25 * > Chair
27 * Cheering
29 * Classroom
31 * Computers
38 Dancing
41 * Demonstration_Or_Protest
49 Explosion_Fire
52 Female-Human-Face-Closeup
53 Flowers
54 Girl
56 Government-Leader
57 Greeting
59 * > Hand
63 * Highway
71 * Instrumental_Musician
72 Kitchen
77 Meeting
80 * > Motorcycle
83 * News_Studio
84 * Nighttime
85 Office
86 Old_People
89 People_Marching
95 Press_Conference
97 Reporters
99 Roadway_Junction
100 * Running
105 * Singing
107 Sitting_Down
112 * Stadium
115 Swimming
117 * > Telephones
120 Throwing
163 * Baby
227 Door_Opening
254 * Fields
261 * > Flags
267 Forest
274 * George_Bush
297 Hill
321 * Lakes
342 Military_Airplane
359 * Oceans
392 * > Quadruped
431 Skating
434 * Skier

440 Soldiers
454 Studio_With_Anchorperson
478 Traffic

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 with the detailed semantic indexing runs at the back of the workshop notebook and on the webpage: `http://www-nlpir.nist.gov/projects/tv2012/tv11.sin.500.concepts_ann_v2.xls`

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

3.2 Data

The IACC.2.B collection was used for testing. It contained 106 913 shots while the IACC.2.C collection used in the "Progress" task contained 113 161 shots. In the localization subtask, 1 573 832 jpeg I-frames were used for testing. Automatic Speech Recognition (ASR) output on IACC.2 videos was provided by LIMSI (Gauvain et al., 2002) and a past community annotation of concepts was organized by LIG and LIF groups (Ayache & Quénot, 2008) on sound and vision as well as Internet Archive videos from 2007-2013 were available for use in system development.

3.3 Evaluation

Each group was allowed to submit up to 4 prioritized main runs and two additional if they are "no annotation" runs, one localization run was allowed with each main submission, and up to 2 progress runs was allowed on the progress dataset IACC.2.C. In total 15 groups submitted a total of 54 main runs, 4 localization runs, and 9 progress runs against IACC.2.C dataset. In addition to the 54 main runs submitted against the IACC.2.B dataset this year, there were 21 runs submitted in TRECVID 2013 as part of the progress subtask and which were evaluated this year as well.

Main concepts

The 30 evaluated single concepts were chosen after examining TRECVID 2013 60 evaluated concept scores across all runs and choosing the top 45 concepts with maximum score variation such that 15 concepts were evaluated in 2014 only, 15 will be evaluated in 2015 only and 15 will be common in both years including the subset of 10 concepts for localization. Randomization tests experiments on the choosen concepts revealed consistent performance of system ranks when compared with trecvid 2013 results.

For each concept in the main task, pools were created and randomly sampled as follows. The top pool sampled 100 % of shots ranked 1-200 across all submissions. The bottom pool sampled 11.1 % of ranked 201-2000 shots 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. Once the assessor completed judging for a topic, he or she was asked to rejudge all clips submitted by at least 10 runs at ranks 1 to 200. In all, 191 717 were judged. 664 023 shots fell into the unjudged part of the overall samples.

Localization

For the localization subtask judging proceeded as follows. For each shot found to contain a concept in the main task, a systematic sampling was employed to select I-frames at regular intervals from the shot. This year an interval value of 3 was applied to fit 200 hours of human assessors work given that each assessor can judge about 6000 images. Selected I-frames were displayed to the assessors. For each image the assessor was asked to decide first if the frame contained the

concept or not, and, if so, to draw a rectangle on the image such that all of the visible concept was included and as little else as possible. Figure 1 shows the evaluation framework. In accordance with the guidelines, if more than one instance of the concept appeared in the image, the assessor was told to pick just the most prominent one and box it in and stick with selecting it unless its prominence changed and another target concept had to be selected.

Assessors were told that in the case of occluded concepts, they should include invisible but implied parts only as a side effect of boxing all the visible parts.

The following table describes for each of the 10 localization concepts the number of shots judged to contain the concept and the number of I-Frames comprised by those shots:

Concept	*Name...*	*True shots*	*I-Frames*
3	Airplane	194	2885
15	Boat_Ship	517	15880
17	Bridges	222	6101
19	Bus	223	6158
25	Chair	1077	73142
59	Hand...	120	1976
80	Motorcycle	196	6075
117	Telephones	211	11964
261	Flags	468	16814
392	Quadruped	485	29858

The larger numbers of I-Frames to be judged for concepts 25 and 392 within the time allotted caused us to assign some of those images to assessors who had not done the original shot judgments. Such additional assessors were given the rules that the original assessors used to judge if the concept exists or not in the video and told to make use of these rules as a guide for their judgments and localization.

3.4 Measures

Main concepts

The *sample_eval* software, a tool implementing xinfAP, was used to calculate inferred recall, inferred precision, inferred average precision, etc., for each result. Since all runs provided results for all evaluated concepts, runs can be compared in terms of the mean inferred average precision across all evaluated single concepts. The results also provide some information about "within concept" performance.

Localization

Temporal and spatial localization were evaluated using precision and recall based on the judged items at two levels - the frame and the pixel, respectively. NIST then calculated an average for each of these values for each concept and for each run. For each shot that was judged to contain a concept, a subset of the shot's I-Frames was sampled, viewed and annotated to locate the pixels representing the concept. The set of annotated I-Frames was then used to evaluate the localization for the I-Frames submitted by the systems.

3.5 Results

Single Concepts

Performance varied greatly by concept. Figure 2 shows how many unique instances were found for each tested concept. The inferred true positives (TPs) of 7 concepts exceeded 1 % from the total tested shots. Top performing concepts were "Chair", "Forest", "Singing", "Nighttime", "Instrumental_Musician", "Running", "News_Studio", "Boat_Ship".

On the other hand, concepts that had the fewest TPs were "Baby", "Basketball", "Skier", "Airplane", "George_Bush", "stadium", "lakes".

Figure 3 shows the results of all the main run submissions (color coded). Category A runs used only IACC training data, while Category D runs used IACC and non-trecvid data as well. The median score across all runs was 0.217 while maximum score reached 0.332. Also, the median baseline run score automatically generated by NIST is plotted on the graph with score 0.273.

Category D runs were the most popular type and achieve top recorded performances. Only 4 runs from category E (no annotation) were submitted and achieved a maximum score of 0.080.

Figure 4 shows the distribution of the run scores including the scores of progress runs which were submitted in 2013 against the 2014 testing dataset. Most of the progress teams achieved better performance in 2014 compared to their 2013 submissions. The max and median scores (0.206) across all runs were better than 2013 scores as well. However, we can not conclude that in general systems performance are getting significantly better because in 2013 the number of submitted runs were much more than in 2014 and several runs had very low scores affecting the median score to go down.

Figure 5 shows the performance of the top 10 teams across the 30 concepts. Note that each series in this plot just represents a rank (from 1 to 10) of the scores, but not necessary that all scores at given rank belong to specific team. Team's scores rank differently across the 30 concepts.

Some concepts reflected a medium spread (approx. 0.1) between scores of the top 10 such as feature "Basketball", "Beach", "Chair", "Motorcycle", "Running", "Baby","Flags", "George_Bush", and "Quadruped". While others had more bigger spread such as "Demonstration_Protest", "Computers", and "Bicycling". The spread in scores may indicate the variation in used techniques performance and there is still room for further improvement. The majority of the rest of the concepts had a tight spread of scores among the top 10.

In general, the median scores for common concepts were higher in 2014 than in 2013 and scores ranged between minimum 0.015 ("Telephones") and maximum 0.66 ("News-Studio"). As a general observation, both concepts had the minimum and maximum median scores last year as well which demonstrates that probably systems performance in general didn't change too much and more research into new techniques are needed to tackle the most diffecult concepts.

To test if there were significant differences between the systems' performance, we applied a randomization test (Manly, 1997) on the top 10 runs (Figure 6) as shown in Figure 7. The figure indicates the order by which the runs are significant according to the randomization test. Different levels of indentation signify a significant difference according to the test. Runs at the same level of indentation are indistinguishable in terms of the test. In this test the top 2 ranked runs was significantly better than all other runs.

To further perform failure analysis on the submitted results we ran an experiment to count number of shots submitted for each pair of concepts that were judged as a TP in one concept and as a FP in the other concept. This experiment essentially can help in identifying confused concepts due to high visual similarity or due to overlapping context or background information. Figure 8 shows the matrix across all pairs. Dark green slots refers to high number of shots while light green refers to low number of shots. From this figure we can notice high confusion between different pairs such as "Chair" (1025) and "Classroom" (1029), "Chair" and Telephones (1117), "Computers" (1031) and "Telephones", "Bridges" (1017) and "Highway" (1063), "Instrumental_Musician" (1071) and "Singing" (1105), "Forest" (1267) and "Lakes" (1321), and "Lakes" and "Oceans" (1359).

Another experiment to measure how diverse is the submitted runs we measured the percentage of common shots across the same concepts between each pair of runs. We found that on average about 30% (minimum 22%) of submitted shots are common between any pair of runs. These results show the diversity of the used approaches and their output.

Progress

A total of 6 teams submitted progress runs against the IACC.2.B dataset to compare their 2013 system with the 2014 system and measure how much progress they made. Figure 9 shows the best run score by team in both years. 5 out of 6 teams achieved better scores in 2014 compared to 2013 and randomization tests show that the 2014 runs are better than the corresponding 2013 runs. The maximum improvement reached about 0.2 mean InfAP.

We also measured the performance per concept for each team to find how many concepts were improved in 2014. It can be seen in Figure 10 that most concepts were improved in 2014 compared to 2013 with 3 teams reaching 100% improvement (all 30 concepts improved).

Concept Localization

Figure 11 shows the mean precision, recall and F-score of the returned I-frames by all runs across all 10 concepts.

All runs reported much higher precision (reaching a maximum above 40 %) than recall or F-score (maximum 20%). On the other hand Figure 12 shows the same measure by run for spatial localization (correctly returning a bounding box around the concept). Here all three scores were almost close to each other for all runs reaching maximum little above 30% precision. Contrary to 2013 results, this year spatial F-score results are better than the temporal F-scores but, as all runs come from only 1 team, it is hard to draw any conclusions.

The average true positive I-frames vs average false positive I-frames for each run is shown in Figure 13. For all runs the average False positive I-frames are almost double the average true positive I-frames even for top run. Runs that tried to be more conservative

in reporting I-frames didn't gain much in terms of F-score measure. There is a big challenge for systems to try to balance the accuracy of the returned I-frames while still achieving high F-score measure.

The F-score performance by concept is shown in Figures 14 and 15 for temporal and spatial respectively across all runs. In general, most concepts achieved higher spatial scores compared to temporal localization with the concept "Flags" reporting maximum score of more than 70% in spatial and more than 50% in temporal. We notice very low maximum scores for the concept "Hand" in both localization types. Finally, all 4 run's scores are very near to each other in both localization types across all concepts except the concept "Telephones" which varied in spatial scores among the 4 runs.

To visualize the distribution of recall vs precision for both localization types we plotted the results of recall and precision for each submitted concept and run in Figures 16 and 17 for temporal and spatial localization respectively. We can see in Figure 16 that most concepts missed a lot of true positve I-frames achieving low recall scores while some concepts achieved good precision scores but at the cost of low recall.

An interesting observation in Figure 17 shows that systems are good in submitting an accurate approximate bounding box size which overlaps with the ground truth bounding box coordinates. This is indicated by the cloud of points in the direction of positive correlation between the precision and recall for spatial localization.

Figures 18 and 19 show some samples of good and less good spatial localization results based on F-scores. The green boxes on the left column display the ground truth bounding box as decided by the human assessors while the red box on the right column displays the submitted result from a run.

2014 Observations

Finally, to summarize our observations about the overall task performance and general ideas or techniques used by participating teams we found that the main task was little harder than 2013 because of the new dataset used and different target concepts tested and evaluated. The raw system scores have higher Max and Median compared to 2013, but still relatively low, and most common concepts with 2013 have higher median scores. In regard to the progress task most systems improved significantly from 2013 to 2014. In the localization subtask, runs missed a lot of true positive I-frames but submitted boxes that approximate the true bounding box in size and with some overlap in location.

Systems approaches were similar to 2013 ones with many innovations. Bag of visual words were still very common, used in combination with many different variations in feature extraction and image representation options. Many used dense and pyramidal feature extraction, spatial information encoding with fisher vectors, MFCC audio features and trajectory-based features, multiple keyframes per shot, semantic features, hard negative mining and pseudo-relevance feedback. More teams this year took up deep learning approaches to train their classifiers. Some teams used trained ImageNet networks and made use of the hidden layers in deep convolutional networks. A new approach based on fast local area independent representaiton was used in the localization subtask.

Finally, we anticipate more research innovations in the coming years to explore the promising directions such as deep learning and new fast image and feature representations specially with the usage of parallel computing and GPUs.

For detailed information about the approaches and results, the reader should see the various site reports (TV14Pubs, 2014) and the results pages in the online workshop notebook (TV14Notebook, 2014).

4 Instance search

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. The instance search task seeks to address some of these needs.

4.1 Data

The task was run for three years, starting in 2010, to explore task definition and evaluation issues. It used data of three sorts: Sound and Vision (2010), BBC rushes (2011), and Flickr (2012). Finding realistic test data which contains sufficient recurrences of various specific objects/persons/locations under varying conditions has been difficult.

In 2013 the task embarked on a multi-year effort using 464 h of the BBC soap opera EastEnders. 244 weekly "omnibus" files were divided by the BBC into 471 523 shots to be used as the unit of retrieval. The

videos present a "small world" with a slowly changing set of recurring people (several dozen), locales (homes, workplaces, pubs, cafes, restaurants, open-air market, clubs, etc.), objects (clothes, cars, household goods, personal possessions, pets, etc.), and views (various camera positions, times of year, times of day).

4.2 System task

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

- a brief phrase identifying the target of the search
- 4 example frame images drawn at intervals from videos containing the item of interest. For each frame image:
  - a binary mask of one or more regions of interest (ROI) covering one instance of the target, each region bounded by a single polygon
  - the shot from which the image was taken
- an indication of the target type taken from this set of strings (OBJECT, PERSON)

Information about the use of the examples was reported by participants with each submission. The possible categories for use of examples were as follows:

A example 1 only
B examples 1 and 2 only
C examples 1, 2, and 3 only
D all four examples only
E video examples (+ optionally image examples)

4.3 Topics

NIST viewed every 10th test video and developed a list of recurring objects, people, and locations. 30 test queries (topics) were then created. As in 2013, the topic targets included mostly small and large rigid objects, logos, and people/animals.

Based on information that emerged during the evaluation, three topics were dropped from the scoring:

- 9100: a SLUPSK vodka bottle - had only 2 true positives
- 9113: a sanitation worker's yellow-green vest - the topic text was too restrictive
- 9117: pay phone - there was a late revision from "a" to "this"

The guidelines for the task allowed the use of metadata assembled by the EastEnders fan community as long as this use was documented by participants and shared with other teams.

4.4 Evaluation, Measures

Each group was allowed to submit up to 4 runs and in fact 23 groups submitted 107 automatic and 12 interactive runs (using only the first 24 topics). Each interactive search was limited to 15 minutes.

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

This 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 reported.

4.5 Results

Discussion

Figure 20 shows the distribution of automatic run scores (average precision) by topic as a boxplot. The topics are sorted by the maximum score with the best performing topic on the left. Median scores vary from nearly 0.8 down to almost 0.0. Per-topic variance varies as well with the largest values being associated with topics that had the best performance. Many factors might be expected to affect topic difficulty. All things being equal, one might expect targets with less variability to be easier to find. Rigid, static objects would fall into that category. In fact for the automatic runs, topics with targets that are stationary, rigid objects make up 9 of the 14 with the best scores, while such topics make up only 2 of the bottom 13 topics. Figure 21 documents the raw scores

of the top 10 automatic runs and the results of a partial randomization test (Manly,1997) and sheds some light on which differences in ranking are likely to be statistically significant. One angled bracket indicates $p < 0.05$; two indicate $p < 0.01$.

In Figure 22, a boxplot of the interactive runs performance, the relative difficulty of several topics varies from that in the automatic runs but in the majority of cases is the same. Here, unlike the case with the automatic runs, stationary, rigid targets are equally represented (5 of 11) in the top and bottom halves of the topic ranking. Figure 23 shows the results of a partial randomization test. Again, one angled bracket indicates $p < 0.05$ (the probability the result could have been achieved under the null hypothesis, i.e., could be due to chance); two indicate $p < 0.01$. The relationship between the two main measures - effectiveness (mean average precision) and elapsed processing time is depicted in Figure 24 for the automatic runs with elapsed times less than or equal to 10 s. Although the highest effectiveness is correlated with the longest elapsed times, at levels below that, the same effectiveness was achieved across the full range of elapsed times. The relationship between the number of true positive and the maximum effectiveness on a topic is shown in Figure 25. For topics with less than 500 true positives there seems to be little correlation; for those with more than 500 true positives, maximum effectiveness seems to rise with the number of true positives.

Figure 26 shows the relationship between the number of topic example images used and the effectiveness of the runs. (Scores for multiple runs from a team with the same number of image examples used were averaged.) With few exceptions, using more image examples resulted in better effectiveness. However, using the video associated with each image example did not produce any improvement in effectiveness over using just all four image examples. This was the first year video for the images examples was made available.

Approaches

Nearly all systems used some form of SIFT local descriptors, but there was a large variety of experiments addressing representation, fusion, or efficiency challenges. The trend was moving to larger bag of visual words (BoVW) vocabularies, larger numbers of keyframes (in the case of Nagoya University: all). New in 2014 were several experiments with convolutional neural networks (CNN) for intermediate features. There was increased focus on post-processing (e.g., spatial verification, feedback). The effectiveness of new methods was not consistent across teams so further research is needed.

A typical INS system comprised the following processing:

1. Processing clips
    - Keyframe choice (1 per shot - 5fps-all frames)
    - Keyframe downsizing?
2. Representation
    - Global (HSV, LBP, CNN, etc.)
    - Local (Detection methods, Choice of descriptors)
    - Cluster to BoVW (1M words, hard/soft, etc.)
3. Matching
    - Similarity function(idf weighting,
    - Weighting ROI vs. background
4. Postprocessing
    - spatial verification
    - Face/color filtering
5. Feedback
6. Fusion of scores
    - Average pooling

System developers addressed the issue of dealing with topic information. Teams considered how to exploit the masks (focus versus background). MediaMill compared mask, full, and fused. BUPT assumed the boundary region of mask contained relevant local points. VIREO experimented with background context modelling using a "stare" model and found it helps. Teams experimented with combining sample images. Several teams used joint average querying to combine samples into a single query. Some teams tried exploiting the full video clip for query expansion. NII tracked interest points in ROI and found it helped some, but interlaced video raised issues. OrangeBJ found no gains. Tokyotech tried tracking and warping the mask with a small gain. VIREO found tracking objects in query video helped if video quality is good (often not the case).

Participating researchers worked on finding an optimal representation for the videos. Teams tried processing more frames (IRIM, Nagoya), combining different feature types (local/global), reviewed techniques and their results (IRIM), combined BoVW and CNN (BUPT). Some groups combined multiple keypoint detectors and multiple descriptors. Nagoya found a single descriptor (Hessian Affine ROOTSift) was almost as good as a combination of 6, yet was more efficient. ORAND used no quantization codebook, kept raw keypoints, and faced a scaling issue. Sheffield compared SIFT, HOG, global features. Experiments with MPEG-7 features were carried out by TU Chemnitz and TelecomItalia; they seemed OK for mid-sized rigid objects. INSIGHTDCU explored the potential of convolutional neural networks (CNN) in promising experiments with a small-scale dataset. The approach seemed to be useful as a representation that could help improve BOVW, but not sufficiently discriminative for primary search keys.

Several teams experimented with how best to match topics to videos. Typically inverted files were used for fast lookup in sparse BovW space (Lucene). NII used an asymmetric similarity function (2013); it was tested by IRIM to no effect, but Nagoya found it helped. VIREO found a new normalization term in the cosine similarity function helped to increase recall. Collection statistics were used by some teams - BM25 enhancements for weighting (NTT-NII) helped, as did IDF adjusted for burstiness (INSIGHTDCU). Pseudo relevance feedback and query expansion were explored by NTT-CSL, who used ROI features for reranking and found it promising.

In studies involving post-filtering, NII tested an improved spatial verification method; Nagoya found that spatial verification helped; OrangeBJ used a face detector for filtering hits for topics involving faces but got no improvement; Wuhan University applied a face filter and color filter; TU Chemnitz employed an indoor/outdoor detector based on audio analysis for removing false matches.

In the matter of system architecture and efficiency JRS experimented with compact VLAT signatures; but a particular signature was not sufficiently discriminative; TU Chemnitz tried PostgreSQL on grid platform; MIC_TJ (Tongjing Univ) tried hybrid parallelization using CPU's, GPU's and map/reduce; ORAND approximated K nearest neighbors (KNN) on unquantized local descriptors; Nagoya worked on efficient re-ranking methods (involving spatial verification); and CERTH built a complete index in RAM.

Several teams built interactive systems. OrangeBJ (BUPT and Orangelabs) had strong performance using a "relative rerank method". BUPT_MCPRL used an automatic system without Convolutional Neural Networks for a small gain. ORAND propagated labels to similar shots in same scene using a similarity shot graph. INSIGHTDCU found a system using positive images for new queries outperformed one using them for training an SVM. AXES implemented pseudo relevance feedback and an interactive check. TUC_MI (Chemnitz) found MPEG-7 color descriptors were not sufficiently discriminative. ITI_CERTH tested shots vs scene presentation and found that shot-based presentation yielded better results.

No information was available from the following teams: ATTLABS, PKU_ICST, U_TK, Tsinghua_IMMG. For details on the other teams' work please see the online workshop notebook. In addition, slides from the National Institute of Informatics, Japan (NII), Nagoya University (NU), NTT Communication Science Laboratories (NTT_CSL), Beijing University of Posts and Telecommunications (BUPT), ORAND S.A. Chile (ORAND) can be found on the TRECVID publications webpage.

For more detailed information about the approaches and results, the reader should see the various site reports (TV14Pubs, 2014) and the results pages in the online workshop notebook (TV14Notebook, 2014).

5 Multimedia event detection

The 2014 Multimedia Event Detection (MED) evaluation was the fourth evaluation of technologies that search multimedia video clips for complex events of interest to a user. There were four major changes in 2014.

- 10 new events were used to evaluate the Ad-Hoc systems.
- NIST built an evaluation Input/Output server that controlled the release of data, specified the order of modules: Metadata Generation (MG), Semantic Query Generation (SQG), Event Query Generation (EQG), and Event Search (ES) to run on the team's servers, and collected time stamps for all module calls.
- Teams reported hardware computing resources for all modules specified in the evaluation plan.

- The HAVIC Novel 1 collection was added to last year's PROGRESS collection, doubling the size of the evaluation collection to 7 580 hours.

A user searching for events, complex activities occurring at a specific place and time involving people interacting with other people and/or objects, 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 was an mnemonic title for the event.
- An event definition which was a textual definition of the event.
- An event explication which was a textual listing of some attributes that are often indicative of an event instance. The evidential description provided a notion of some potential types of visual and acoustic evidence indicating the event's existence but it was not an exhaustive list nor was it to be interpreted as required evidence.
- An evidential description which was 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 was not an exhaustive list nor was 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 were illustrative in the sense they helped form the definition of the event but they did not demonstrate all the inherent variability or potential realizations.

Developers built Pre-Specified event systems where knowledge of the event(s) was taken into account during generation of the metadata store for the test collection. In 2014, the second Ad-Hoc event task was conducted where the metadata store generation was completed before the events where revealed.

5.1 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 data as specified in the HAVIC data section of this paper. The MED '14 Pre-Specified event names are listed in Table 4 and Table 5 lists the MED '14 Ad-Hoc Events.

5.2 Evaluation

Sites submitted MED system outputs testing their systems on the following dimensions:

- Events: either all 20 Pre-Specified events (PS14) and/or all 10 Ad-Hoc events (AH14).
- Subsystems: with or without pseudo relevance feedback.
- Test collection: either the MED14 Full Evaluation collection (MED14-EvalFull) or a 1 238 hour subset (MED14-EvalSub) collection.
- Query Conditions: Semantic Query (the event text), 0 Ex (the event text and the 5 000-clip Event Background collection "EventBG"), 10 Ex (the event text, EventBG, and 10 positive and 10 miss clips per event), 100 Ex (the event text, EventBG, and 10 positive and 10 miss clips per event).

Full participation would mean teams would submit 8 runs, (PS and AH events * 4 queries).

For each event search a system generated:

- A rank for each search clip in the evaluation collection: A value from 1 (best rank) to N representing the best ordering of clips for the event.

- 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 Rank Threshold for the event: A threshold on the ranks optimizing the R_o for the system.

- 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.

System developers also reported the hardware components used and computation times of the metadata generation, event query generation, and event search modules as well as the metadata store size.

Submission performance was computed using the Framework for Detection Evaluation (F4DE) toolkit.

5.3 Measures

System output was evaluated by how well the system retrieved and detected MED events in evaluation search video metadata and by the computing resources used to do so. The determination of correct detection was at the clip level, i.e. systems provided a response for each clip in the evaluation search video set. Participants had to process each event independently in order to ensure each event could be tested independently.

The primary evaluation measures for performance were Mean Average Precision (MAP) and Minimal Acceptable Recall (Ro).

There were three primary measures for computational speed expressed as real-time factors. Real-time factor was the total processing time divided by the number of hours of video in the test collection. Three aspects speed factors were computed: (1) Metadata Generation Processing Speed, (2) Event Query Generation Processing Speed, and (3) Event Search Processing Speed.

5.4 Results

17 teams participated in the MED '14 evaluation; 6 teams were new. All teams participated in the Pre-Specified (PS) Event, 10 Exemplar (10Ex) test processing all 20 events as well as the Ad-Hoc (AH) event, 10 Exemplar (10Ex) test processing all 10 events. 5 teams chose to process the MED14Eval-Subset.

The MED14 evaluation re-used the MED Progress Evaluation collection and added the Novel 1 Evaluation collection. Since the evaluation data will be used through 2015 MED evaluations, protecting the statistic of the evaluation data is of the utmost importance, NIST reported only Mean Average Precision for each run.

Table 6 presents the MAP (averaged over events) for the PS and AH task submissions for all training exemplar conditions and for non-Psuedo Relevance Feedback (noPRF) and Pseudo Relevance Feedback (PRF) systems.

First, the MAPs for MED14Eval-Full and MED14Eval-Sub differ by a large amount; an average of 43% and 45% for the PS and AH 10EX systems respectively. While the MAP discrepancy is large, if one accounts for the change in positive richness in the subset, the MAPs are similar as expected from a randomly selected subset.

The SQ and 0EX conditions remain difficult for most teams with the exception of CMU who achieved a MAP of 14.9 and 15.5 for their SQ and 0Ex condition, PS event system on the MED14Eval-Full. This is 180% higher than the nearest team. Similarly, CMU achieved a MAP of 11.7 for their SQ and 0Ex, Ad-Hoc event system which is 185% higher than the nearest team.

The MED evaluation reuses the PROGESS collection and events so that yearly changes in performance can be measured. Figure 27 shows the change on MAP scores over time for the same 10 events processing the 10Ex event kits. As can be seen in the graph, all teams that participated in the condition improved their MAP scores this year.

Minimal Acceptable Recall measures the system's ability to set a retrieval threshold based on minimizing the tradeoff between improving recall at the expense of additional retrieved videos. Figure 28 shows the stacked bars of the AdHoc, 10Ex, NoPRF systems. The full height of the bar indicates the Ro at the system's threshold and the lower bar indicates the lowest Ro achievable with an oracle-set threshold. The difference between the two heights indicates how well the threshold was set. There were three cliques of systems: three teams missed the optimum Ro by 6-7%, 5 teams missed the optimum by 11-39%, and three teams over 140%.

Teams reported a range of statistics describing the computational resources used during the evaluation. We present a few of the salient statistics here. Figure 29 shows the number of CPU and GPU cores

used to process the evaluation collection. There was a wide range hardware systems used to process the MED14Eval-Full set. BBNVISER used the most CPU cores at 2,432 cores which was slightly larger than CMU who used 2,400 cores but added 30,000 GPU cores in 12 GPU units. MediaMill used the fewest CPUs at 16. From a CPU/MAP tradeoff perspective, the MediaMill's MAP score of 15.1 is a modest degradation from BBNVISER's was 18.0 MAP despite the 99.3% reduction in cores.

The size of the metadata generated for a search collection is an important deployability factor for MED systems. The MED evaluation did not require developers to engineer their systems attempting to minimize metadata size; however they were asked to report the disk size of their metadata. Figure 30 shows the Real Size Factor (RS) (the metadata size/the video size) for metadata derived from the signal, Automatic Speech Recognition(ASR)/Optical Character Recongition (OCR), and semantic (actions/objects/etc.) data. A real size factor of 1 means the metadata size equals the video size. The consistent pattern is metadata for the signal < semantic data < ASR/OCR as one would expect.

5.5 Summary

In summary, all 17 teams participated in the Pre-Specified (PS), 10 Exemplar (10Ex) test processing all 20 events as well as the Ad-Hoc (AH), 10 Exemplar (10Ex) test processing all 10 events. 5 of 17 teams chose to process the MED14Eval-Subset collection. Performance has steadily improved for the 10Ex condition since 2012. The Semantic Query condition remains a challenge for most teams however CMU's new techniques are closing the gap between exemplar-based and semantic-based queries.

TRECVID '15 evaluation will include the MED Track. Proposed changes include the introduction of 10 new Ad-Hoc events selected randomly from existing HAVIC data.

For more detailed information about the approaches and results, the reader should see the various site reports (TV14Pubs, 2014) and the results pages in the online workshop notebook (TV14Notebook, 2014).

6 Multimedia event recounting

The 2014 Multimedia Event Recounting (MER) evaluation was the third evaluation of technologies that recount the multimedia video events detected by MED systems.

In more detail, the purpose, of the 2014 Multimedia Event Recounting (MER) track, was to stimulate the development of technologies that state the *key evidence* that led a Multimedia Event Detection (MED) system to decide that a multimedia clip contains an instance of a specific event and to allow human users to rapidly and accurately find clips of interest via the recountings. The 2014 TRECVID MER evaluation assesses just the recounting of the evidence.

The 2014 evaluation of MER consisted of six metrics, described briefly here and in more detail later. Note that for each of the following metrics, ratings are pooled for all MER events. And each of the first five metrics are rated on a Likert-type scale of "Strongly Agree" to "Strongly Disagree". *Event Query Quality*, is the judge-provided rating of whether a given event query is concise and logical. *Tag Quality*, is the judge-provided rating of each piece of key evidence regarding how well the tag name captures the contents of the evidence. *Spatial Localization*, is the judge-provided rating of how well a piece of key evidence is localized in space, scored only if the piece of evidence has a visual component, and bounding boxes are provided. *Temporal Localization*, is the judge-provided rating of how well a piece of key evidence is temporally localized, assuming the piece of evidence is not a keyframe. *Evidence Quality*, is the judge-provided rating of how convincing the evidence for a given recounting was as a whole, that the recounted video contains an instance of the event. An Evidence Quality rating is first requested for whether the judge thinks the set of key evidence convinces the judge that the clip contains an instance of the event of interest (Key Evidence Quality), if the key evidence (alone) is not convincing, all evidence is shown and judges are again asked to rate the Evidence Quality (All Evidence Quality). *Recounted Percent*, is the sum of evidence durations for all recountings, divided by the sum duration of all of the corresponding videos, but that ratio was also computed for each recounting (there is one recounting for each clip for each system), and the distribution of those ratios is also provided for each system.

Each *event* was explicitly defined by an *Event Kit*. A clip that is *positive* for an event contains an *instance* of that event.

Each event in this evaluation

- is a complex activity occurring at a specific place and time;

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

Participation in MER 2014 was open to all 2014 TRECVID MED participants whose system always produced a recounting for each clip that their MED system deemed to be positive (that is, identified as being above their MED system's decision threshold for being positive) for the MED 10-Ex condition.

Input data formats were as in the existing HAVIC data. MER output data formats used ASCII XML text. NIST provided a MER XSD schema to be used to specify and validate system output.

The systems recountings were evaluated by a panel of judges. NIST created a MER Workstation to view and judge the recountings, and NIST provided it to the MER participants and the judges.

We are interested in queries that a human would consider to be logical and concise, and recountings that state the evidence in a way that human readers find easily understandable.

6.1 System task

Given an event kit and a test video clip that the team's MED system deems to contain an instance of the event, the MER system was to produce a recounting containing the evidence used to support that decision. Evidence means observations such as scene/context, persons, animals, objects, activities, text, non-linguistic audio, and other evidence supporting the detection of the event. Each piece of evidence was associated with an indication of the system's confidence that the evidence is correct. For each piece of evidence in the recounting, the system indicated whether or not that piece of evidence was key, in the sense that it needs to be viewed by the MER judge to convince them that the event occured in the video.

For each piece of evidence, the recounting was to include pointers to the evidence in the clip, indicating

- temporally, where in the clip the piece of evidence occurs, and
- spatially, where in the frame the evidence occurs (if visible evidence, optional).

In addition, each piece of evidence was assigned a multimedia type, drawn from the following list:

- *visual*: (not involving Audio, or OCR)
- *audio-visual*: (not involving OCR)
- *ocr*: (text via OCR)
- *audio*: (without ASR textual transcription)
- *asr*: (transcribed via ASR)

Systems specified XML tag elements in their queries for the retrieval of evidence from the videos. Each tag included the following attributes:

id a unique identifier that can be used in other XML elements to associate elements, e.g., in the equation of a parent node.

name a textual label for the tag (e.g., *pickup truck*).

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

For each tag, systems produced an XML element for each piece of evidence retrieved by that tag. These elements included attributes with the following information.

key indicating whether or not the piece of evidence is key

start begin time of the piece of evidence

end end time of the piece of evidence. For keyframes this value would be set to the start time

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

start and end bounding boxes using the convention of upper left and lower right points, specified the spatial localization. These attributes were optional, and only permitted for evidence with a visual component

text transcribed from ASR or OCR. Only permitted for ASR and OCR evidence

A piece of evidence is a spatio-temporal pointer to an excerpt from the video. It contains a *start* and an *end* time, given in seconds. If the piece of evidence is not purely auditory, an initial and final bounding

boxes within the frame, consisting of pixel coordinates of the upper-left and lower-right corners of the bounding box, relative to the upper-left corner of the frame may be included. If the piece of evidence is ASR or OCR evidence, the transcribed *text* is also included. For implementation reasons, each piece of evidence was required to be of a certain type (visual, audio-visual, ocr, audio, asr).

The MER Evaluation was performed on the MED 10-Ex condition. NIST chose, for evaluation, ten events and up to fifteen videos where all participating MER teams made a positive MED decision. Only a few events did not provide a full set of fifteen videos.

The ten 2014 MER evaluation events, chosen from both the MED pre-specified and ad-hoc events, the chosen pre-specified events were:

- E027 Rock climbing,
- E031 Key keeping,
- E032 Wedding shower,
- E036 Felling a tree,
- E037 Parking a vehicle, and
- E039 Tailgating.

The chosen ad-hoc events were:

- E043 Busking,
- E044 Decorating for a celebration,
- E045 Extinguishing a fire, and
- E050 Teaching dance choreography.

6.2 Evaluation Procedures

Using the MER workstation, the judge reviewed the event kit name and description, read over the query, and then assessed the query according to their agreement with the following statement: "This seems like a concise and logical query that would be created for the event"

Following that, the judge then assessed the recounting by:

1. Assessing each piece of key evidence by:
   (a) Reading the tag name of the enclosing tag element
   (b) Viewing the piece of evidence in it's entirety
   (c) Rating their agreement with the following statements:
      - "[*tag name*] correctly captures the contents of the snippet."
      - "The system chose the right window of time to present the evidence." (Only for non-keyframe snippets)
      - "The system chose the right bounding box(es) to isolate the evidence." (Only when bounding boxes are included)
2. After assessing each piece of key evidence, rating their agreement with the following statement:
   - "The evidence presented convinces me that the video contains the "*Event Name*" event."
3. If the judge was not convinced (rated either "Neutral", "Disagree", or "Strongly Disagree"), they proceeded by:
   - Reviewing all of the evidence
   - Then rating their agreement with the following statement:
     – "Now the evidence presented convinces me that the video contains the "*Event Name*" event."

the E031 Key evidence was revealed alongside the key evidence

Each of the statements posed to judges for MER assessment have responses on a Likert-scale with the following five levels of agreement:

- "Strongly Agree"
- "Agree"
- "Neutral"
- "Disagree"
- "Strongly Disagree"

The MER Workstation shows the structure and contents of the query during query judgement. Following that, the query is fleshed out with the recounting elements for evidence judgement. Once each piece of key evidence has been judged, judges may freely navigate through the key evidence.

6.3 Metrics

NIST measured the following characteristics of the recountings for each system.

Qualitative Measures:

For each of the judge-provided ratings for *Event Query Quality*, *Tag Quality*, *Spatial Localization*, *Temporal Localization*, and *Evidence Quality*, NIST computed the percentage breakdown of responses for each individual measure, i.e.:

- (*Total number of "Stongly Agree"*) / (*Total number of responses*)
- (*Total number of "Agree"*) / (*Total number of responses*)
- (*Total number of "Neutral"*) / (*Total number of responses*)
- (*Total number of "Disagree"*) / (*Total number of responses*)
- (*Total number of "Stongly Disagree"*) / (*Total number of responses*)
- (*Total number of null responses*) / (*Total number of responses*)

When it is not appropriate to request a response from the judges, the response is considered to be null or "Not Available". As in the case of *Temporal Localization* for keyframe evidence, or *Spatial Localization* for evidence where bounding boxes have been omitted.

Recounted Percent:

The total time of all key pieces of evidence across recountings as a percentage of total video duration.

(*Total duration of key pieces of evidence*) / (*Total duration of videos to be assessed*)

6.4 Results

For more detailed information about the approaches and results, the reader should see the various site reports (TV14Pubs, 2014) and the results pages in the online workshop notebook (TV14Notebook, 2014). The results pages include a graph showing that, for all teams, the human judges found the evidence recounted by the systems to be distinctly more convincing for clips that actually contained an instance of the event of interest (the target or positive clips) than for the clips that did not (the non-target or negative clips).

7 Interactive surveillance event detection

The 2014 Surveillance Event Detection (SED) evaluation was the seventh evaluation focused on event detection in the surveillance video domain. The first such evaluation was conducted as part of the 2008 TRECVID conference series (Rose, Fiscus, Over, Garofolo, & Michel, 2009) and again in 2009, 2010, 2011, 2012 and 2013. It was designed to move computer vision technology towards robustness and scalability while increasing core competency in detecting human activities within video. The approach used was to employ real surveillance data, orders of magnitude larger than previous computer vision tests, and consisting of multiple, synchronized camera views.

For 2014, the evaluation test data used a new 11-hour subset from the total 45 hours available of the test data from the Imagery Library for Intelligent Detection System's (iLIDS) Multiple Camera Tracking Scenario Training (MCTTR) data set (UKHO-CPNI, 2007 (accessed June 30, 2009)) collected by the United Kingdom's Home Office Science and Development Branch. Given that this test data was never annotated, a crowdsourcing effort was conducted in order to generate the reference data.

In 2008, NIST collaborated with the Linguistics Data Consortium (LDC) and the research community to select a set of naturally occurring events with varying occurrence frequencies and expected difficulty. For this evaluation, we define an event to be an observable state change, either in the movement or interaction of people with other people or objects. As such, the evidence for an event depends directly on what can be seen in the video and does not require higher level inference. The same set of seven 2010 events were used for the 2011, 2012, 2013 and 2014 evaluations.

7.1 System task

In 2014, the retrospective event detection (rSED) and interactive event detection (iSED) tasks were supported.

- The retrospective task is defined as follows: given a set of video sequences, detect as many event observations as possible in each sequence. For this evaluation, a single-camera condition was used as the required condition (multiple-camera input was allowed as a contrastive condition). Furthermore, systems could perform mul-

tiple passes over the video prior to outputting a list of putative events observations (i.e., the task was retrospective).

- The interactive task is defined as follows: given a collection of surveillance video data files (e.g., that from an airport, or commercial establishment) for preprocessing, at test time detect observations of events based on the event definition and for each return the elapsed search time and a list of video segments within the surveillance data files, ranked by likelihood of meeting the need described in the topic. Each search for an event by a searcher can take no more than 25 elapsed minutes, measured from the time the searcher is given the event to look for until the time the result set is considered final. Note that iSED is not a short latency task. Systems can make multiple passes over the data prior to presentation to the user.

The annotation guidelines were developed to express the requirements for each event. To determine if the observed action is a taggable event, a *reasonable interpretation rule* was used. The rule was, "if according to a reasonable interpretation of the video, the event must have occurred, then it is a taggable event". Importantly, the annotation guidelines were designed to capture events that can be detected by human observers, such that the ground truth would contain observations that would be relevant to an operator/analyst. In what follows we distinguish between event types (e.g., parcel passed from one person to another), event instance (an example of an event type that takes place at a specific time and place), and an event observation (event instance captured by a specific camera).

7.2 Data

The development data consisted of the full 100 hours data set used for the 2008 Event Detection (Rose et al., 2009) evaluation. The video for the evaluation corpus came from the approximate 50 hour iLIDS MCTTR data set. Both data sets were collected in the same busy airport environment. The entire video corpus was distributed as MPEG-2 in Phase Alternating Line (PAL) format (resolution 720 x 576), 25 frames/sec, either via hard drive or Internet download. Figure 31 shows the coverage and views from the different cameras used for data collection.

System performance was assessed on a new 11-hour subset of the evaluation corpus, which is different from past Evaluations. Like SED 2012 and after, systems were provided the identity of the evaluated subset so that searcher time for the interactive task was not expended on non-evaluated material. This 11-hour subset is composed of about 3 hours taken from the SED13 dataset as well as 8 hours that were not annotated. The new data was generated using a *System Mediated Crowdsourcing* effort where a few selected past participants were asked to run their latest algorithm on the entire 45 hours of data. We then performed an event instance confidence analysis, generating a percentage confidence that a given number of systems acknowledge an event true. This was followed by a set of human reviewing each event occurrence above a certain threshold. Out of 3300 event occurrence reviewed, about 600 (close to 20 %) were confirmed as true occurrences. Each true event instance found made the reference used for scoring system inputs.

After the test results were received, a second event instance confidence analysis was performed from the actual system inputs received, which was also followed by another set of human reviewing all new event occurrence found. Out of 2600 event occurrence reviewed, about 300 more (12%) new event occurrence were added.

This extended reference was then used to score the final SED results.

7.3 Evaluation

Sites submitted system outputs for the detection of any 3 of 7 possible events (PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, and Pointing). Additional details for the list of event used can be found in Figure 32. For each instance observation, sites are asked to identify each detected event observation by:

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

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

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

Events observations were represented in the Video Performance Evaluation Resource (ViPER) format using an annotation schema that specified each event observation's time interval.

7.4 Measures

Since detection system performance is a tradeoff between probability of miss vs. rate of false alarms, this task used the Normalized Detection Cost Rate (NDCR) measure for evaluating system performance. NDCR is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per time unit). At the end of the evaluation cycle, participants were provided a graph of the Decision Error Tradeoff (DET) curve for each event their system detected; the DET curves were plotted over all events (i.e., all days and cameras) in the evaluation set. The videos were annotated using the Video Performance Evaluation Resource (ViPER) tool. Events were represented in ViPER format using an annotation schema that specified each event observation's time interval.

7.5 Results

There were 4 participants in 2014 (see figure 33), for a total of 38 Interactive Event Runs and 52 Retrospective Event Runs.

Since this is a new dataset, there are no comparables plots available yet for the different events of interest; therefore what follows are simply the primary Retrospective and Introspective submissions per site for the events: Embrace (see figure 34), PeopleMeet (see figure 35), PeopleSplitUp (see figure 36), PersonRuns (see figure 37), and Pointing (see figure 38).

For more detailed information about the approaches and results, the reader should see the various site reports (TV14Pubs, 2014) and the results pages in the online workshop notebook (TV14Notebook, 2014).

8 Summing up and moving on

Looking back on the development of TRECVid since 2001 and of TREC since a decade before then, it is interesting to see how the idea of a semi-competitive benchmarking or evaluation campaign has really become an embedded feature within our discipline. An earlier study of the scholarly impact of TRECVid (Thornley, Johnson, Smeaton, & Lee, 2011) shows how widespread the use of data, principally data resources and scoring mechanisms, has become and even a cursory examination of the content of our major conferences like ACM MULTIMEDIA and the ACM International Conference on Multimedia Retrieval (ICMR) shows the impact that we have. However another less visible impact is how TREC and then TRECVid have led the evolution of coordinated research efforts from across the world, right across the disciplines.

Of course we can point at IR-related benchmarking like CLEF, INEX, FIRE and others, and similar benchmarking in the vision community like PASCAL, but then we saw the emergence of coordinated research focusing on narrow and specific tasks in association with multimedia IR conferences. The ACM MULTIMEDIA Grand Challenge series is one example, the VideoBrowser Showdown at the MMM conference is another. In other disciplines like the semantic web, we also saw coordinated challenges emerge run by a grassroots organisation but sponsored by a company, Elsevier in this case[1]. Companies then started to take a more active role in sponsoring these challenge events, mostly because they are the gatekeepers of the data that is used to drive these challenges. As such we have now seen companies like Yahoo!, Microsoft and Google sponsor some of these and all the time these help to push out the barriers and even define what makes up our discipline.

This model of proposing and then part-funding grand challenge ideas to see what the research community can come up with is not new and certainly not restricted to IR tasks, examples being the DARPA grand challenge for the development of autonomous cars or the U.S. Agency for International Development (USAID) sponsorship of the Fighting Ebola Grand Challenge for Development. While these other grand challenges and benchmarking activities have grand longterm ambitions, back to our own discipline we must continue to ensure that the benchmarking campaigns that we support remain true to the Cranfield model with replicable results and easy access to data including document, queries, ontologies, or whatever other resources are needed to complete the task.

This overview of TRECVID 2014 has provided ba-

[1] http://challenge.semanticweb.org/

sic 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 online workshop notebook linked from the Guidelines document.

9 Authors' note

TRECVID would not have happened in 2014 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:

- Koichi Shinoda of the TokyoTechCanon team agreed to host a copy of IACC.2 data
- Georges Quénot provided the master shot reference for the IACC.2 videos.
- The LIMSI Spoken Language Processing Group and VexSys Research provided ASR for the IACC.2 videos.
- Noel O'Connor and Kevin McGuinness at Dublin City University along with Robin Aly at the University of Twente worked with NIST and Andy O'Dwyer plus William Hayes at the BBC to make the BBC EastEnders video available for use in TRECVID

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

10 Appendix A: Instance search topics

9099 OBJECT - a checkerboard band on a police cap

9100 OBJECT - a SLUPSK vodka bottle

9101 OBJECT - a Primus washing machine

9102 OBJECT - this large vase with artificial flowers

9103 OBJECT - a red, curved, plastic ketchup container

9104 PERSON - this woman

9105 OBJECT - this dog, Wellard

9106 OBJECT - a London Underground logo

9107 LOCATION - this Walford East Station entrance

9108 OBJECT - these 2 ceramic heads

9109 OBJECT - a Mercedes star logo

9110 OBJECT - these etched glass doors

9111 OBJECT - this dartboard

9112 OBJECT - this HOLMES lager logo on a pump handle

9113 OBJECT - a yellow-green sanitation worker vest

9114 OBJECT - a red public mailbox

9115 PERSON - this man

9116 PERSON - this man

9117 OBJECT - this pay phone

9118 OBJECT - a Ford Mustang grill logo

9119 PERSON - this man

9120 OBJECT - a wooden park bench, straight-backed, with flat arm rests

9121 OBJECT - a Royal Mail red vest

9122 OBJECT - this round watch with black face and black leather band

9123 OBJECT - a white plastic kettle with vertical blue window

9124 PERSON - this woman

9125 OBJECT - this wheelchair with armrests

9126 OBJECT - a Peugeot logo

9127 OBJECT - this multicolored bust of Queen Victoria

9128 OBJECT - this F pendant

References

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

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

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

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

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

Strassel, S., Morris, A., Fiscus, J., Caruso, C., Lee, H., Over, P., et al. (2012, may). Creating havic: Heterogeneous audio visual internet collection. In *Proceedings of the eight international conference on language resources and evaluation (lrec'12).* Istanbul, Turkey: European Language Resources Association (ELRA).

Thornley, C. V., Johnson, A. C., Smeaton, A. F., & Lee, H. (2011, April). The Scholarly Impact of TRECVid (2003-2009). *J. Am. Soc. Inf. Sci. Technol.*, *62*(4), 613–627. Available from http://dx.doi.org/10.1002/asi.21494

TV14Notebook. (2014). http://www-nlpir.nist.gov/projects/tv2014/active/tv14.workshop.notebook.

TV14Pubs. (2014). http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.14.org.html.

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

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

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

Table 1: Participants and tasks

Task					Location	TeamID	Participants
−−	**	−−	−−	SI	Eur	PicSOM	Aalto U.
IS	MD	−−	−−	−−	Eur	AXES	Access to Audiovisual Archives
IS	MD	−−	**	**	NAm	ATTlabs	AT&T Labs Research
IS	−−	−−	SD	−−	Asia	BUPT_MCPRL	Beijing U. of Posts and Telecommunications
−−	MD	−−	−−	−−	Asia	MCIS	Beijing Inst. of Tech., China Inst. for Infocomm Research
−−	MD	MR	SD	SI	NAm	CMU	Carnegie Mellon U.
IS	MD	MR	−−	SI	Eur	ITI_CERTH	Centre for Research and Technology Hellas
−−	−−	−−	SD	−−	NAm	CCNY	City College of New York
IS	MD	MR	−−	SI	Asia,Eur	VIREO-TNO	City U. of Hong Kong, TNO
−−	**	−−	−−	SI	Eur	EURECOM	EURECOM
−−	−−	−−	−−	SI	NAm	FIU_UM	Florida International U., U. of Miami
−−	MD	−−	−−	**	Asia	Fudan	Fudan U.
**	**	−−	SD	**	NAm	IBM	IBM Research
−−	MD	−−	−−	**	Eur	INRIA_LIM_VocR	INRIA LEAR LIMSI Vocapia Research
IS	**	−−	−−	SI	Eur	insightdcu	Insight Centre for Data Analytics
IS	**	−−	−−	SI	Eur	IRIM	IRIM consortium
IS	**	−−	−−	**	Eur	JRS	JOANNEUM RESEARCH
−−	MD	−−	−−	−−	Asia	KU_ISPL	Korea U.
−−	**	−−	−−	SI	Eur	LIG	LIG consortium
IS	−−	−−	−−	−−	Asia	NU	Nagoya U.
IS	MD	−−	**	**	Asia	NII	National Institute of Informatics
IS	**	−−	−−	−−	Asia	NTT_CSL	NTT Communication Science Laboratories
IS	MD	−−	**	−−	SAm	ORAND	ORAND S.A. Chile
IS	**	−−	−−	SI	Asia	OrangeBJ	Orange Labs International Center Beijing
IS	**	−−	**	**	Asia	PKU-ICST	Peking U. ICST
**	**	−−	−−	SI	Asia	PKUSZ_ELMT	Peking U. ELMT
−−	MD	MR	−−	−−	NAm	BBNVISER	Raytheon,UMD,ColumbiaU,CMU,intuVision,Polar Rain
−−	MD	MR	−−	−−	NAm,Eur	SRI_SESAME	SRI International, U. Amsterdam, USCa
−−	MD	MR	−−	**	NAm	SRI_AURORA	SRI International, Sarnoff, UCF, UMass, Cyc
IS	−−	−−	−−	−−	Eur	TUC_MI	Technische Universität Chemnitz
IS	−−	−−	−−	−−	Eur	TelecomItalia	Telecom Italia
IS	MD	−−	−−	SI	Asia	TokyoTech-Waseda	Tokyo Institute of Technology, Waseda U.
IS	MD	−−	−−	−−	Asia	MIC_TJ	Tongji U.
IS	−−	−−	**	−−	Asia	Tsinghua_IMMG	Tsinghua U.
**	**	−−	**	SI	Asia	UEC	U. of Electro-Communiacations, Tokyo
IS	−−	−−	−−	−−	Asia	U_TK	U. of Tokushima
IS	MD	MR	−−	SI	Eur	MediaMill	U. of Amsterdam
−−	−−	−−	−−	SI	NAm	CRCV_UCF	U. of Central Florida
IS	−−	−−	−−	−−	Eur,Asia	Sheffield_UETLahore	U. of Sheffield, U. of Engineering & Technology (PK)
IS	−−	−−	**	−−	Asia	NERCMS	Wuhan U.

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

Table 2: Participants who did not submit any runs

Task					Location	TeamID	Participants
IN	MD	MR	SD	SI			
−−	**	−−	**	**	Asia	BIT	Beijing Institute of Technology - BIT
**	−−	−−	−−	**	Eur	CEALIST	CEA
−−	**	−−	−−	−−	Asia	djcaisa	Chinese Academy of Science (graduate student)
−−	**	−−	−−	**	NAm	ClipMine	ClipMine
**	−−	−−	−−	**	Asia	DUT	Dalian University of Technology
**	**	−−	**	**	Asia	VSAG_IITH	Indian Institute of Technology Hyderabad
−−	**	−−	−−	**	Asia	ECNU	Institute of Computer Applications
−−	**	−−	−−	−−	Asia	JBF	Institute of Automation, Chinese Academy of Science
−−	−−	−−	**	−−	NAm	Kitware	Kitware Inc.
−−	−−	−−	**	−−	Asia	MELCO_ATC	Mitsubishi Electric Corporation
**	−−	−−	−−	−−	SAm	SSIG_NPDI_VIPLAB	Pontifical Catholic U. MG, Federal U. MG
**	−−	−−	**	−−	NAm,Asia	srad	Samsung Research America, Samsung Electronics Korea
−−	**	−−	−−	**	Asia	SRC_Beijing	Samsung Research Center Beijing
−−	−−	−−	**	−−	Asia	SeSaMe_NUS	SeSaMe Centre, IDMI (NUS)
−−	**	−−	−−	−−	Asia	MMLab	Shenzhen Institutes of Advanced Technology (CAS)
−−	**	−−	−−	−−	Asia	SEU	Southeast university
**	**	−−	**	**	Asia	MMM_TJU	Tianjin University
**	**	−−	**	**	Asia	img_thu	Tsinghua University - Intelligent Multimedia Group
−−	**	−−	−−	−−	NAm	UCSD.Triton	University of California, San Diego
**	**	−−	**	**	NAm	UofTML	University of Toronto - Machine Learning
**	**	−−	**	**	Aus	UQMG	University of Queensland

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

Table 3: Instance search pooling and judging statistics

Topic number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number relevant	% judged that were relevant
9099	110315	29864	27.1	340	9448	31.6	494	5.2
9100	111809	36485	32.6	240	11121	30.5	2	0.0
9101	109543	17537	16.0	460	7061	40.3	1568	22.2
9102	111583	24491	21.9	320	8551	34.9	398	4.7
9103	109640	29558	27.0	460	14558	49.3	1818	12.5
9104	110824	40892	36.9	200	10246	25.1	342	3.3
9105	109730	44266	40.3	160	8363	18.9	97	1.2
9106	110217	34113	31.0	220	6369	18.7	243	3.8
9107	110345	29046	26.3	260	8371	28.8	229	2.7
9108	109318	28738	26.3	200	6373	22.2	121	1.9
9109	110374	35219	31.9	200	8104	23.0	104	1.3
9110	109488	20413	18.6	200	3793	18.6	444	11.7
9111	109828	26794	24.4	320	6844	25.5	416	6.1
9112	108912	14410	13.2	460	4720	32.8	846	17.9
9113	110777	36691	33.1	340	13230	36.1	359	2.7
9114	109594	39319	35.9	200	8860	22.5	387	4.4
9115	111200	35393	31.8	400	15911	45.0	277	1.7
9116	110082	37416	34.0	180	7360	19.7	238	3.2
9117	110083	22138	20.1	260	5731	25.9	1738	30.3
9118	110062	35135	31.9	140	6322	18.0	4	0.1
9119	110436	32131	29.1	140	5528	17.2	180	3.3
9120	109669	28389	25.9	180	6489	22.9	189	2.9
9121	109886	41551	37.8	240	10621	25.6	730	6.9
9122	109837	41183	37.5	420	18471	44.9	211	1.1
9123	101871	29936	29.4	460	13667	45.7	831	6.1
9124	102727	39436	38.4	120	4969	12.6	29	0.6
9125	102881	35693	34.7	280	11514	32.3	57	0.5
9126	101886	32837	32.2	120	4617	14.1	36	0.8
9127	102595	29103	28.4	160	3843	13.2	582	15.1
9128	102853	39325	38.2	280	11577	29.4	278	2.4

Table 4: MED '14 Pre-Specified Events

Testing Events
MED'12 event re-test
Bike trick
Cleaning an appliance
Dog show
Giving directions
Marriage proposal
Renovating a home
Rock climbing
Town hall meeting
Winning a race without a vehicle
Working on a metal crafts project
MED'13 event re-test
Beekeeping
Wedding shower
Non-motorized vehicle repair
Fixing a musical instrument
Horse riding competition
Felling a tree
Parking a vehicle
Playing fetch
Tailgating
Tuning a musical instrument

Table 5: MED '14 Ad-Hoc Events

Testing Events
Baby shower
Building a fire
Busking
Decorating for a celebration
Extinguishing a fire
Making a purchase
Modeling
Doing a magic trick
Putting on apparel
Teaching dance choreography

Table 6: MED '14 Mean Average Precisions for Pre-Specified Event and Ad-Hoc Event Systems. '*'=Debugged submission. '+'=Late Submission

			Scores on MED14Eval Full (200K videos)							Scores on MED14Eval Sub (32K videos)						
			MED MAP							MED MAP						
			1_noPRF				2_PRF			1_noPRF				2_PRF		
			SQ	0Ex	10Ex	100Ex	SQ	0Ex	10Ex	SQ	0Ex	10Ex	100Ex	SQ	0Ex	10Ex
AH	MED14Full	ATTLabs			1.3	9.7						4.3	19.7			
		AXES			17.8	29.8						24.8	35.4			
		Aurora	2.2	* 2.2	* 12.7	* 25.6	0.9	* 1.9		4.3	* 4.3	* 17.9	* 30.7	2.2	* 3.8	
		BBNVISER	4.1	5.4	16.7	33.6				7.3	8.6	23.2	37.9			
		CMU	11.7	11.7	18.2	35.3		17.7	20.7	17.4	17.4	23.9	40.5		24.3	26.3
		Fudan			11.3	24.1						15.9	29.5			
		INRIA-LIM-VocR			18.4	31.0						25.3	36.6			
		MediaMill	2.4		12.3	26.6				5.1		18.2	33.4			
		NII			7.4	22.0						12.7	26.5			
		Sesame	2.4		* 16.9	* 32.8				4.9		* 24.1	* 40.6			
		TokyoTech			10.0	25.6						14.2	29.6			
		VIREO	2.7	3.5	10.3					4.4	5.8	16.8				
	MED14Sub	ITI_CERTH										18.3	33.1			
		KU-ISPL										+ 2.1	+ 2.8			
		MCIS										16.1				
		MIC									0.4	0.9	3.2			
		ORAND										5.1	11.6			
PS	MED14Full	ATTLabs			5.3	11.1						7.2	14.5			
		AXES			12.7	28.5						18.9	36.6			
		Aurora	3.5	3.5	* 13.5	* 25.7				6.7	6.7	* 19.6	* 32.5			
		BBNVISER	5.3	5.7	18.0	29.8				8.8	10.0	24.8	36.9			
		CMU	14.9	15.5	19.4	32.3		18.1	20.3	20.0	21.2	25.7	39.6		24.1	26.2
		Fudan			+ 10.7	+ 22.1						+ 15.0	+ 29.2			
		INRIA-LIM-VocR			14.0	29.7						20.0	37.9			
		MediaMill	3.6		15.1	24.3				7.3		20.6	29.8			
		NII			8.0	21.6						11.6	28.1			
		Sesame	5.1		18.3	29.9				8.6		23.7	38.1			
		TokyoTech			8.0	21.9						13.4	29.2			
		VIREO	4.0	5.2	12.4	15.8				6.0	7.7	18.5	23.1			
	MED14Sub	ITI_CERTH										15.1	30.3			
		KU-ISPL										+ 2.4	+ 4.7			
		MCIS										15.6				
		MIC									* 0.2	* 0.6	* 2.7			
		ORAND										1.2	5.0			

Figure 1: Concept Localization Evaluation Framework

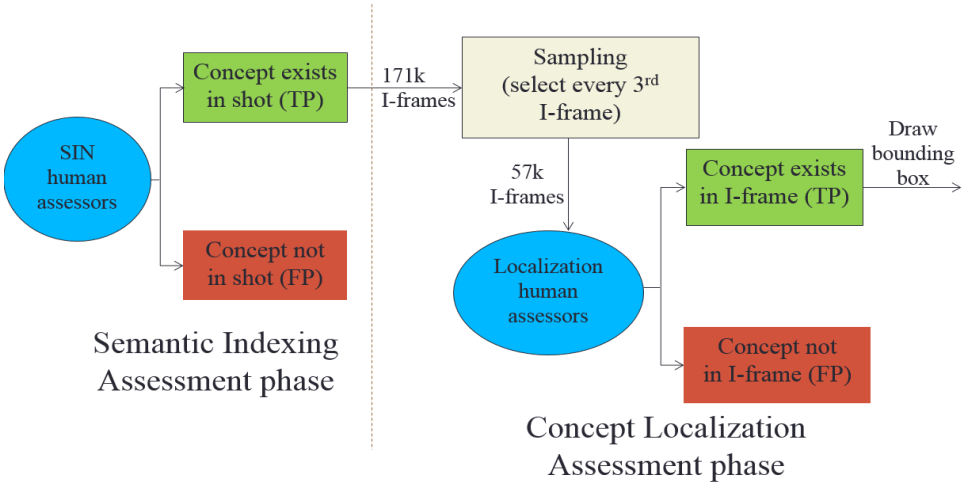

Figure 2: SIN: Histogram of shot frequencies by concept number

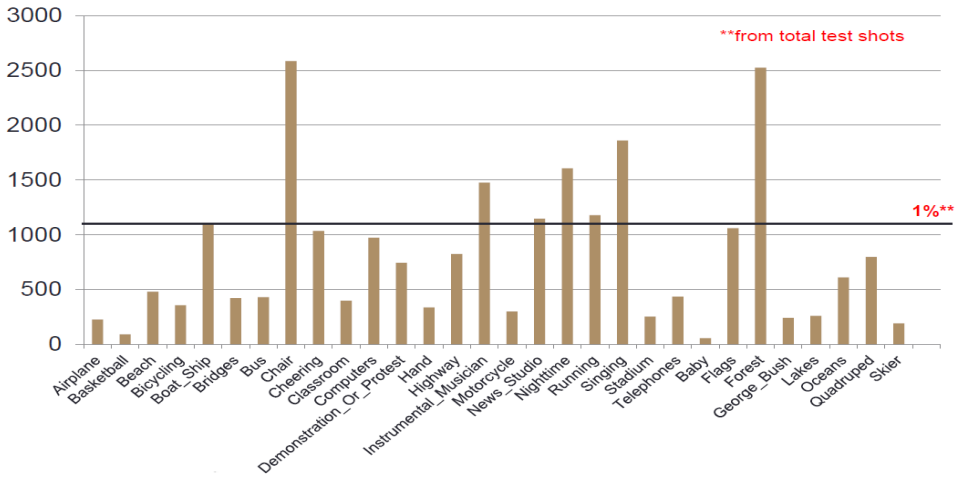

Figure 3: SIN: xinfAP by run - 2014 submissions

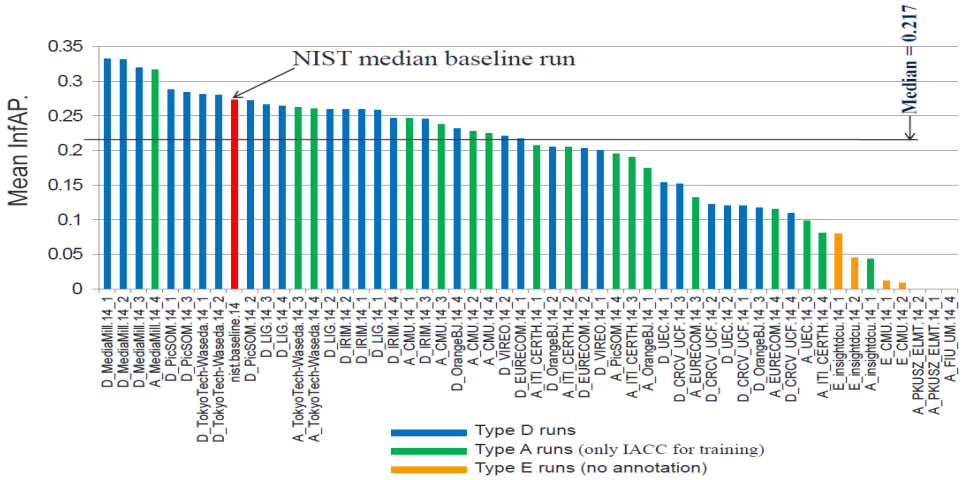

Figure 4: SIN: xinfAP by run - 2014 submissions including Progress runs

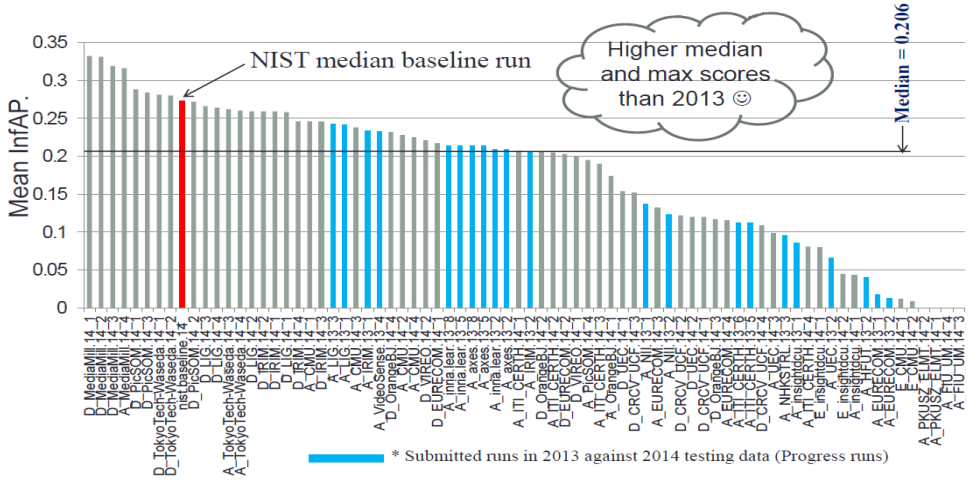

Figure 5: SIN: top 10 runs (xinfAP) by concept number

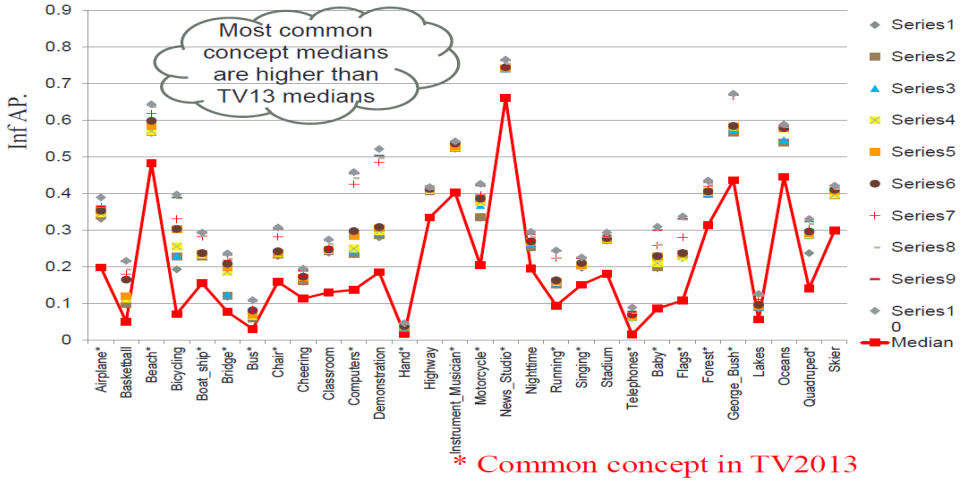

Figure 6: SIN: top 10 main runs

- D_MediaMill.14_1
  - D_MediaMill.14_3
    - D_TokyoTech-Waseda.14_2
    - D_TokyoTech-Waseda.14_1
      - D_LIG.14_3
    - D_PicSOM.14_1
      - D_PicSOM.14_3
        - D_PicSOM.14_2
        - D_LIG.14_3
  - D_MediaMill.14_4
    - D_TokyoTech-Waseda.14_2
    - D_TokyoTech-Waseda.14_1
      - D_LIG.14_3
    - D_PicSOM.14_1
      - D_PicSOM.14_3
        - D_PicSOM.14_2
        - D_LIG.14_3
- D_MediaMill.14_2
  - D_MediaMill.14_3
    - D_TokyoTech-Waseda.14_2
    - D_TokyoTech-Waseda.14_1
      - D_LIG.14_3
    - D_PicSOM.14_1
      - D_PicSOM.14_3
        - D_PicSOM.14_2
        - D_LIG.14_3
  - D_MediaMill.14_4
    - D_TokyoTech-Waseda.14_2
    - D_TokyoTech-Waseda.14_1
      - D_LIG.14_3
    - D_PicSOM.14_1
      - D_PicSOM.14_3
        - D_PicSOM.14_2
        - D_LIG.14_3

Figure 7: SIN: Significant differences among top 10 main runs

Run name	(mean infAP)
D_MediaMill.14_1	0.332
D_MediaMill.14_2	0.331
D_MediaMill.14_3	0.319
A_MediaMill.14_4	0.316
D_PicSOM.14_1	0.288
D_PicSOM.14_3	0.284
D_TokyoTech-Waseda.14_1	0.281
D_TokyoTech-Waseda.14_2	0.280
D_PicSOM.14_2	0.272
D_LIG.14_3	0.266

Figure 8: SIN: Confusion analysis across concepts

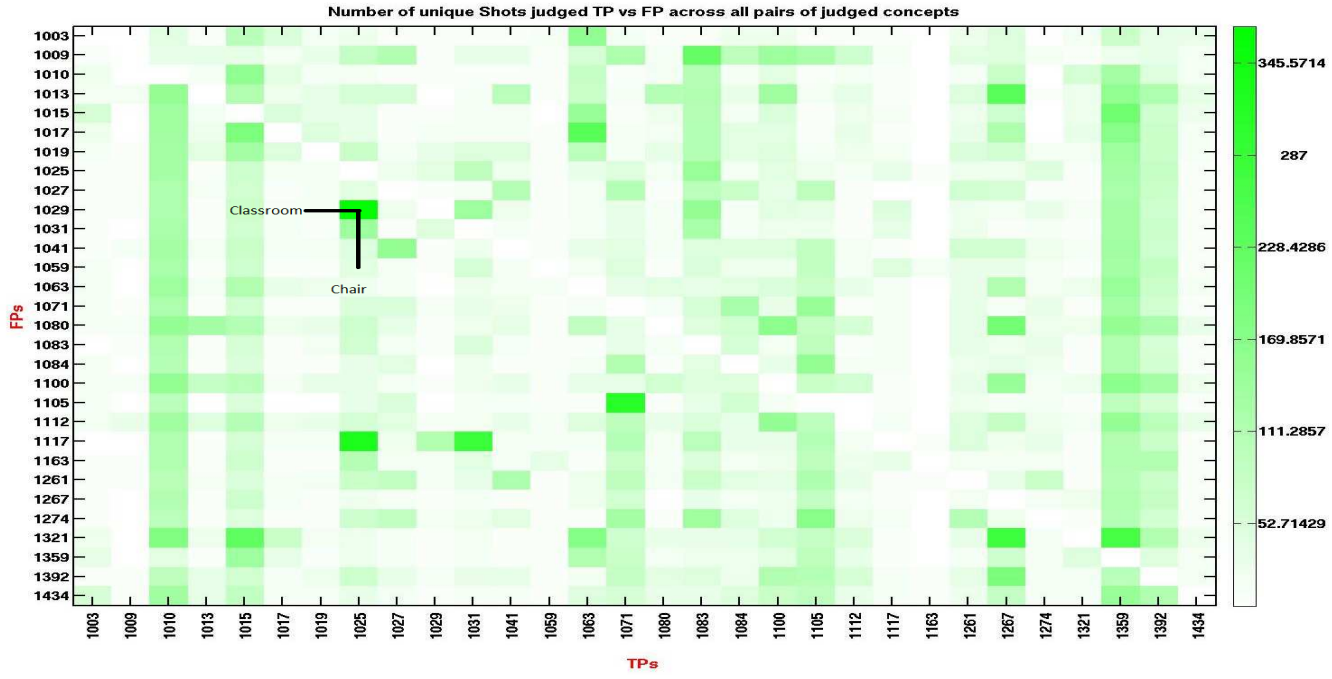

Figure 9: SIN: Progress subtask - Comparing best runs in 2013 vs 2014 by team

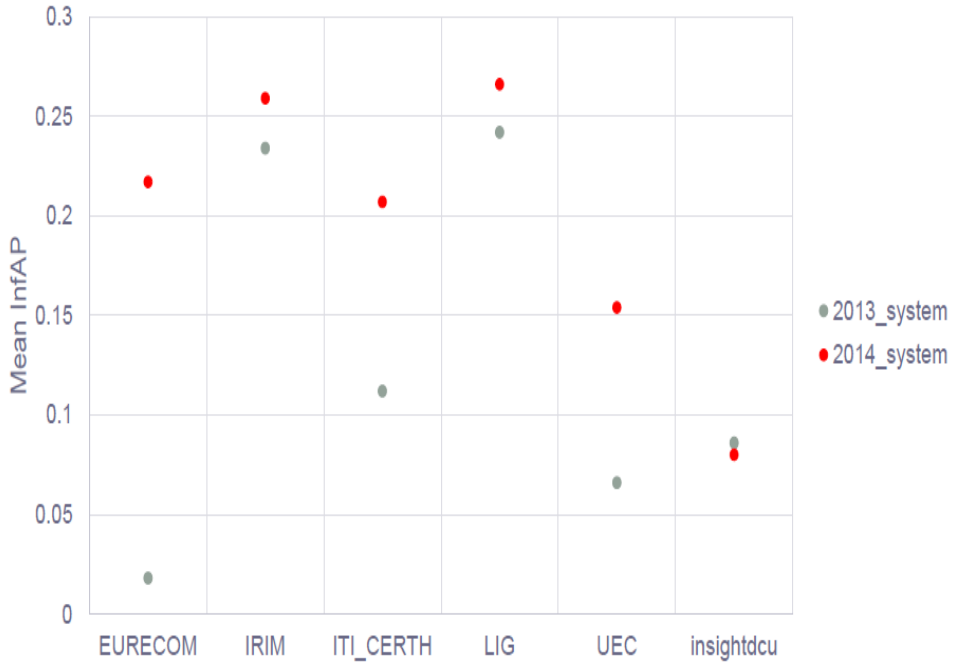

Figure 10: SIN: Progress subtask - Concepts improved vs weakened by team

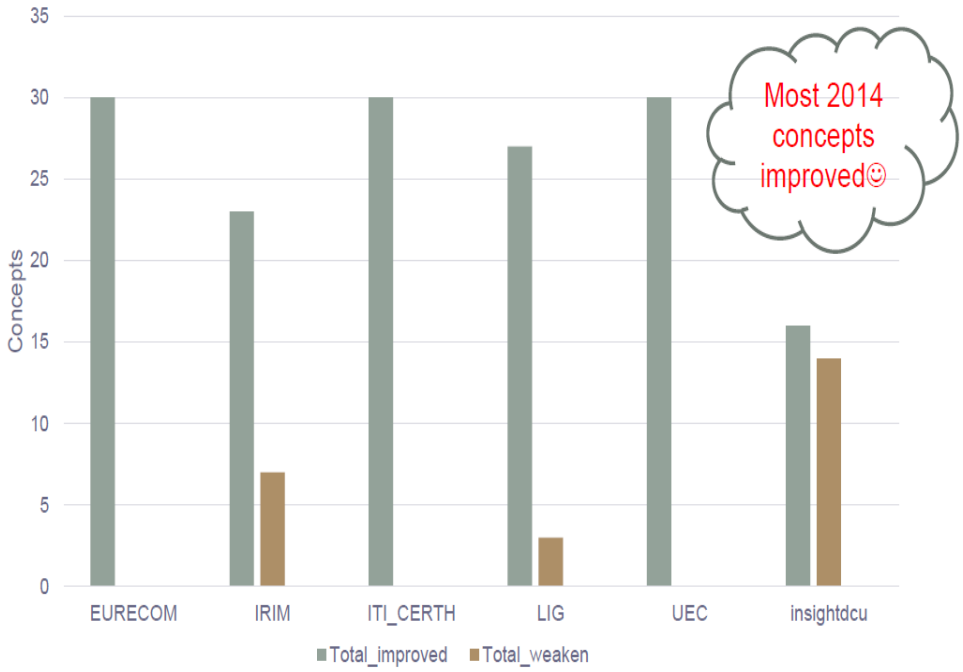

Figure 11: LOC: Temporal localization results by run

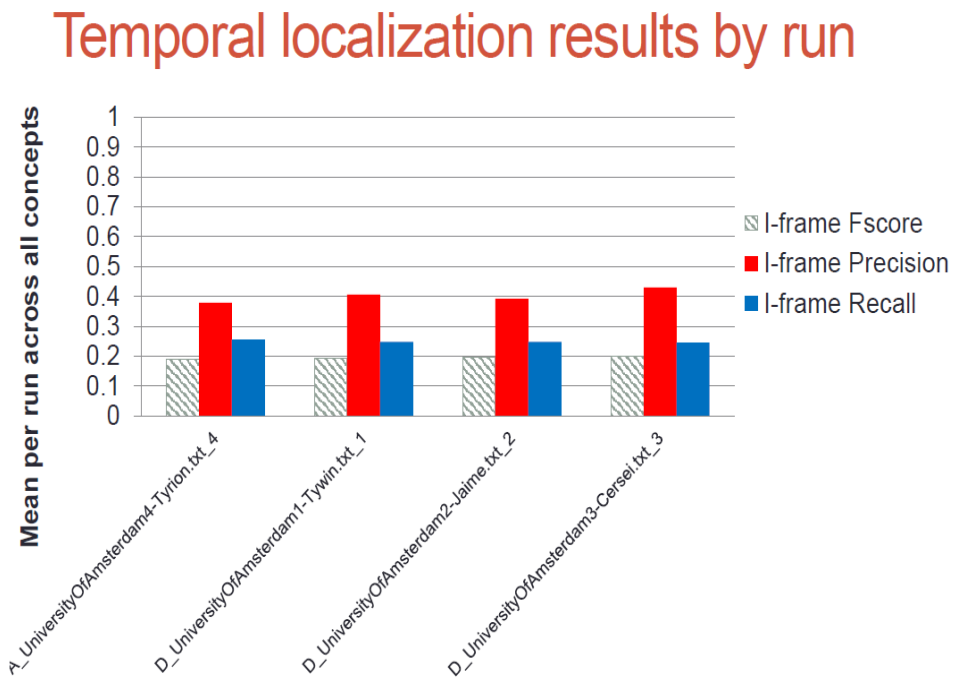

Figure 12: LOC: Spatial localization results by run

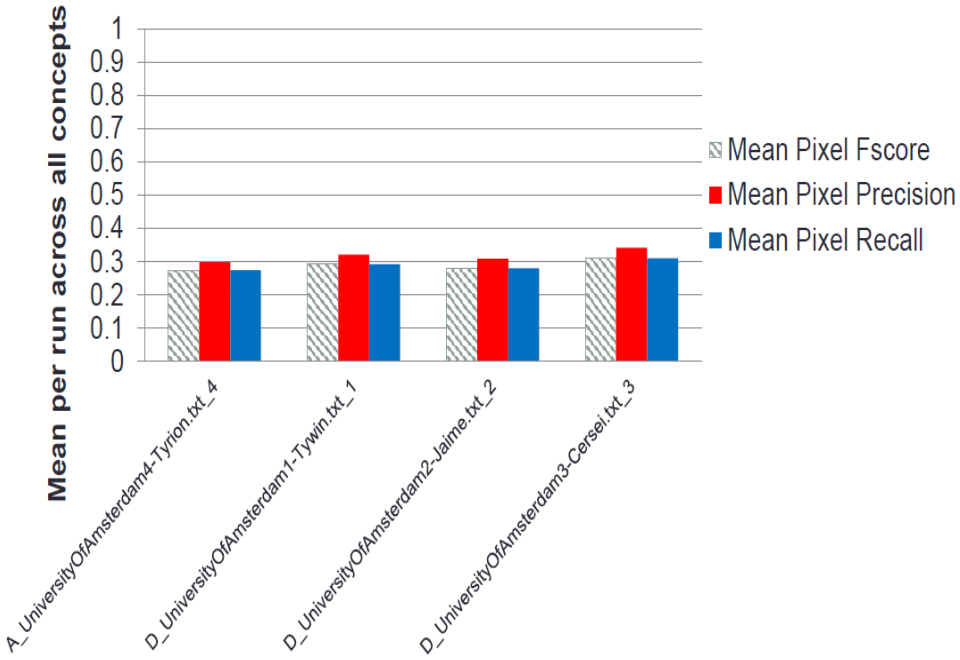

Figure 13: LOC: TP vs FP I-frames by run

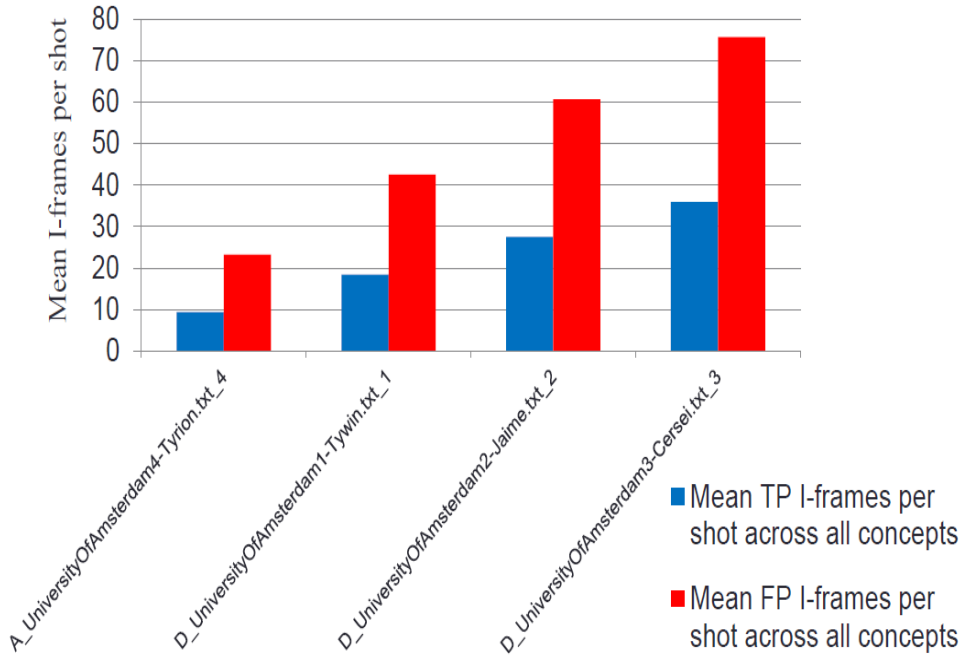

Figure 14: LOC: Temporal localization by concept

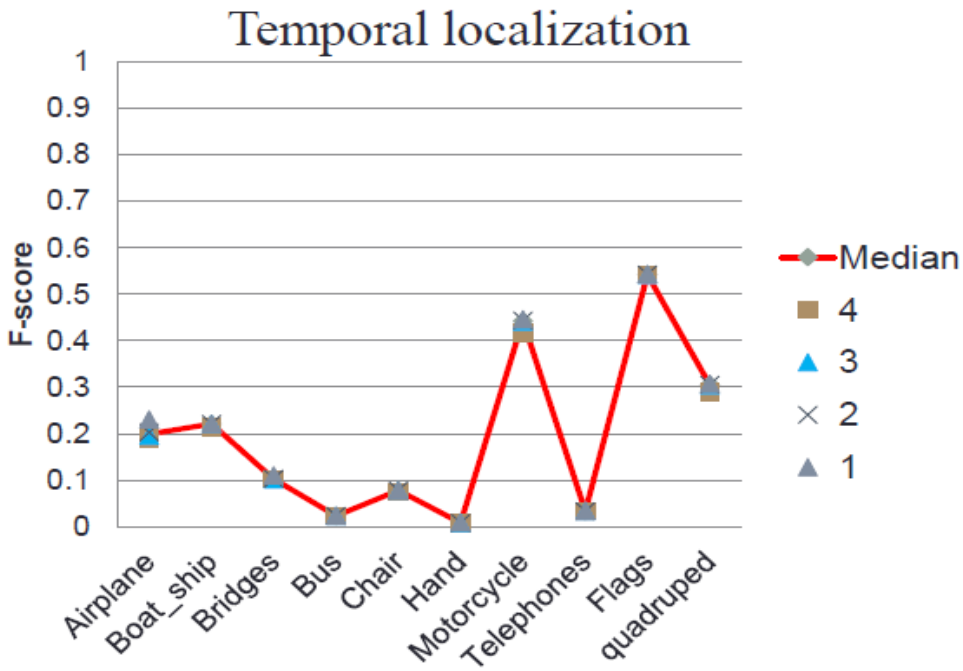

Figure 15: LOC: Spatial localization by concept

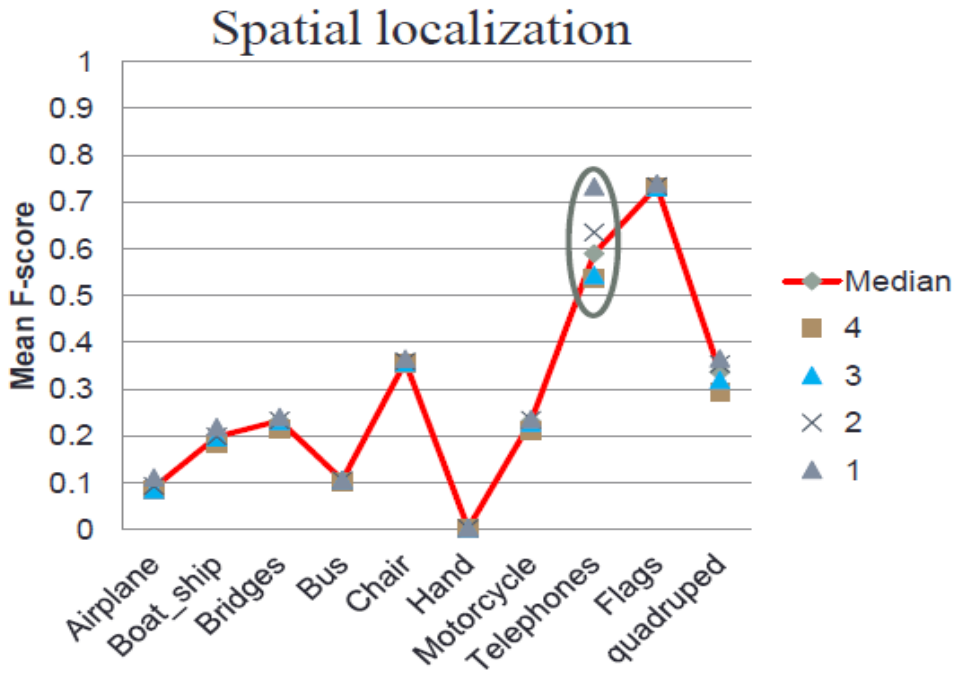

Figure 16: LOC: temporal precision and recall per concept for all teams

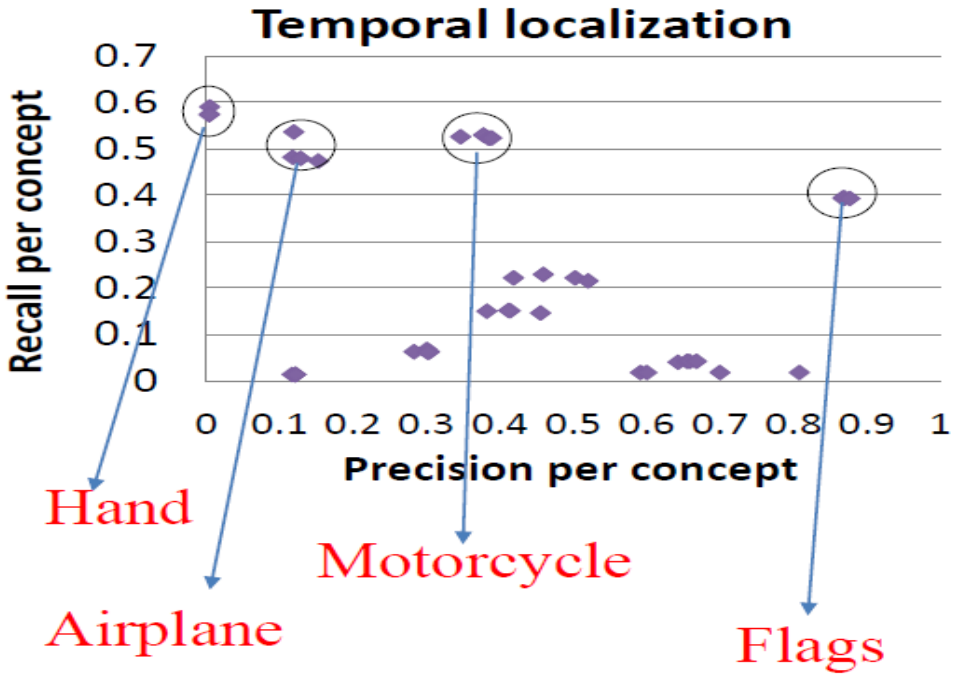

Figure 17: LOC: spatial precision and recall per concept for all teams

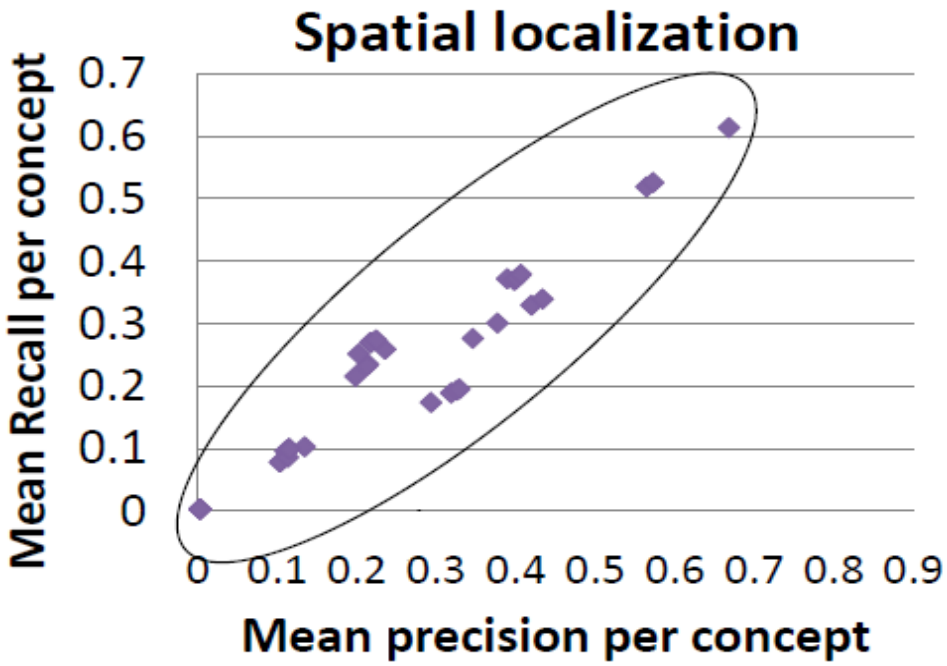

Figure 18: LOC: Samples of good spatial localization

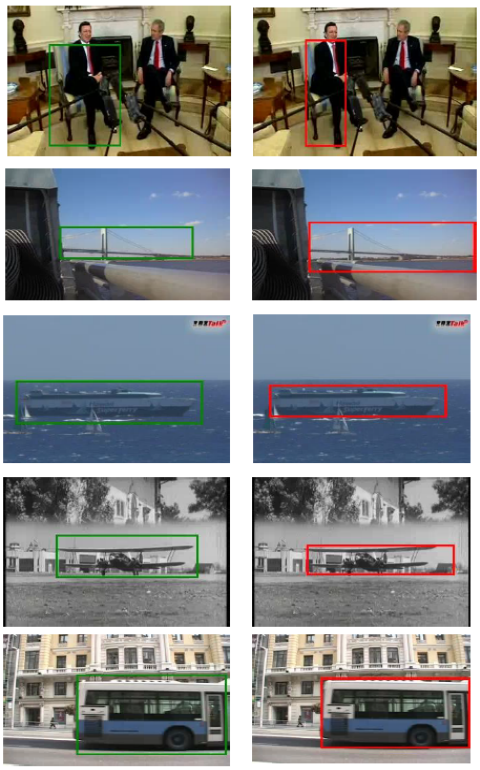

Figure 19: LOC: Samples of less good spatial localization

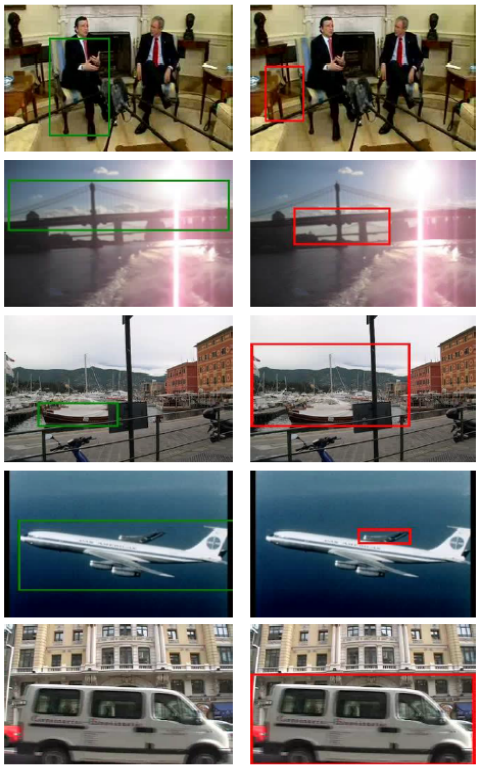

Figure 20: INS: Boxplot of average precision by topic for automatic runs

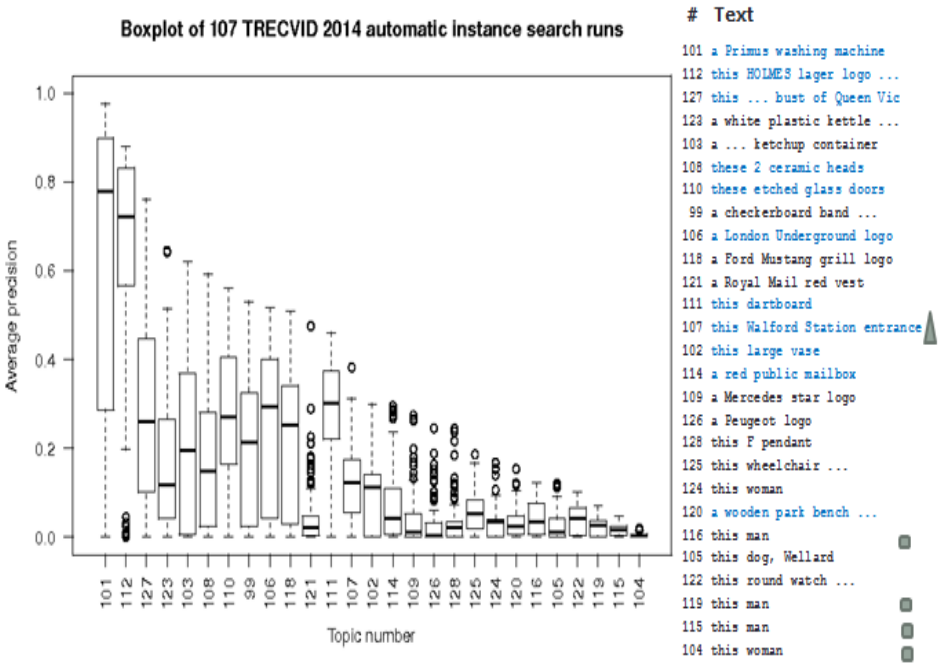

Figure 21: INS: Randomization test results for top automatic runs

MAP

0.325	F_D_NII_2	1	=		>>	>>	>>	>>	>>	>>	>>	>>
0.304	F_D_NU_1	2		=	>>	>>	>>	>>	>>	>>	>>	>>
0.234	F_D_NTT_CSL_1	3			=						>	>>
0.232	F_D_PKU-ICST_2	4				=				>	>	>>
0.227	F_D_MediaMill_1	5					=					>
0.227	F_D_BUPT_MCPRL_1	6						=				>>
0.213	F_D_IRIM_1	7							=			>>
0.197	F_D_VIREO_3	8								=		>
0.183	F_D_ORAND_4	9									=	
0.167	F_D_OrangeBJ_2	10										=
			1	2	3	4	5	6	7	8	9	10

p = probability the row run scored better than the column run due to chance

>> p < 0.01

> p < 0.05

Figure 22: INS: Boxplot of average precision by topic for interactive runs

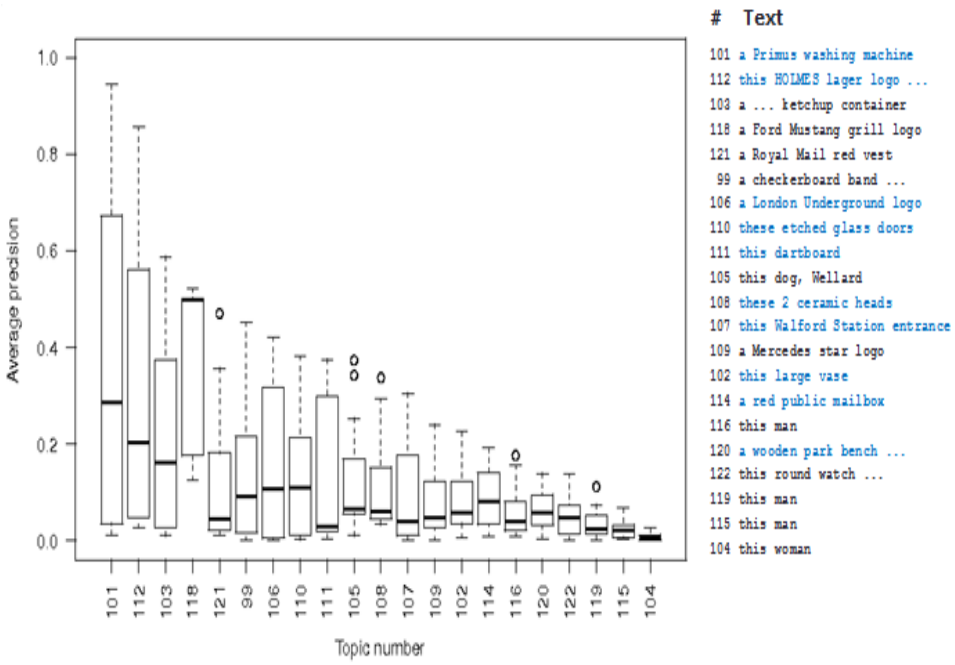

Figure 23: INS: Randomization test results for top interactive runs

MAP

0.317	I_D_PKU-ICST_3	1	=	>>	>>	>>	>>	>>	>>	>>
0.249	I_D_OrangeBJ_3	2		=		>	>	>>	>>	>>
0.237	I_D_BUPT_MCPRL_2	3			=		>	>>	>>	>>
0.174	I_D_ORAND_3	4				=		>>	>>	>>
0.135	I_D_insightdcu_2	5					=		>>	>>
0.108	I_D_AXES_1	6						=	>	>>
0.037	I_E_TUC_MI_1	7							=	
0.032	I_D_ITI_CERTH_1	8								=
			1	2	3	4	5	6	7	8

p = probability the row run scored better than the column run due to chance

>> p < 0.01

> p < 0.05

Figure 24: INS: Mean average precision versus time for fastest runs

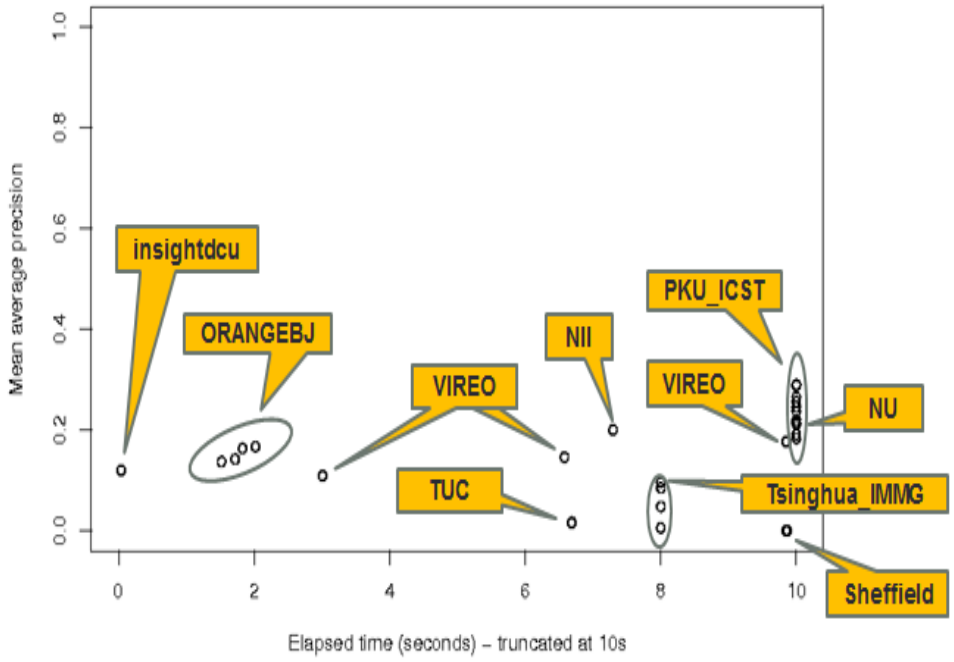

Figure 25: INS: Number of true positives versus average precision

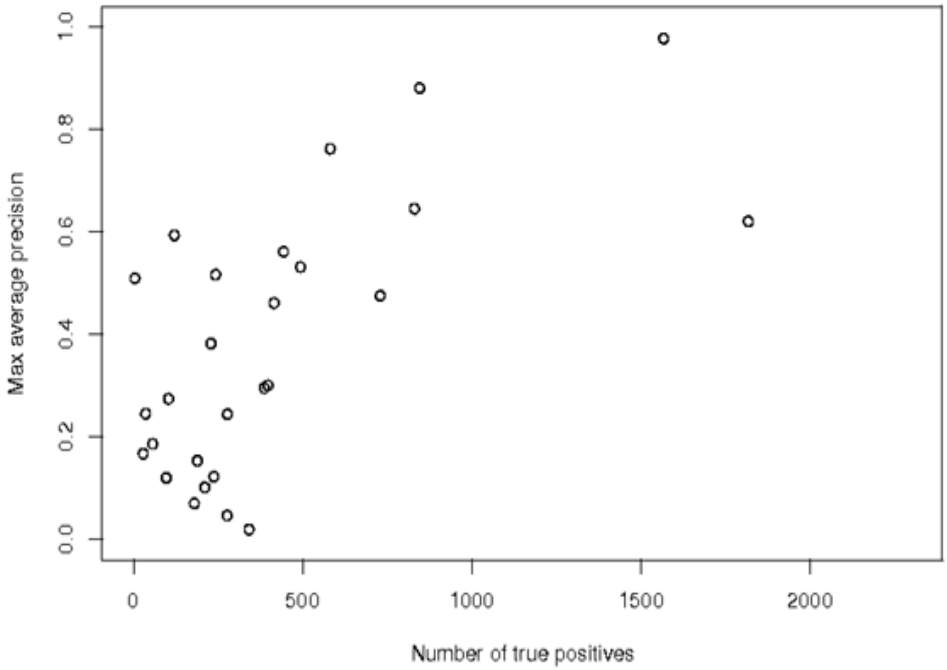

Figure 26: INS: Effect of number of topic example images used

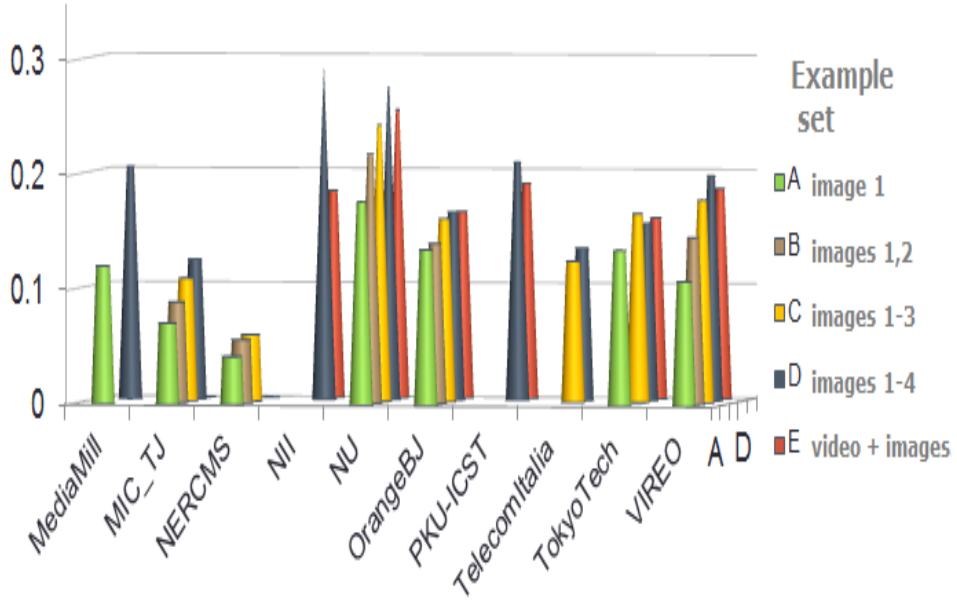

Figure 27: MED: Historical MAP scores for Pre-Specified event, 10Ex systems common events '12-'14

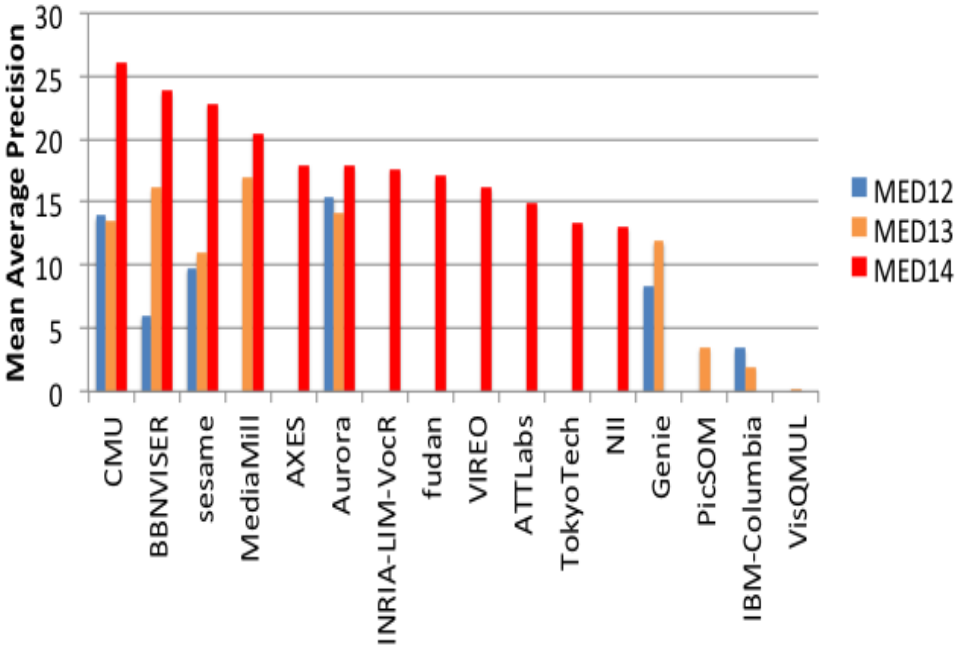

Figure 28: MED: Ro scores for Ad-Hoc, 10Ex systems

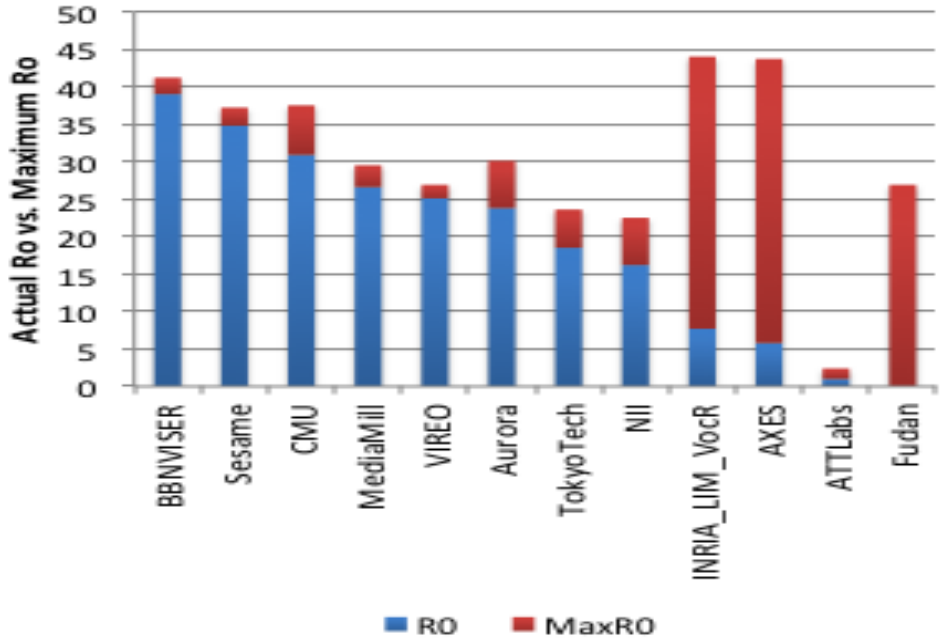

Figure 29: MED: CPU and GPU core counts for metadata generation on the MED14Eval-Full and MED14Eval-Sub collections

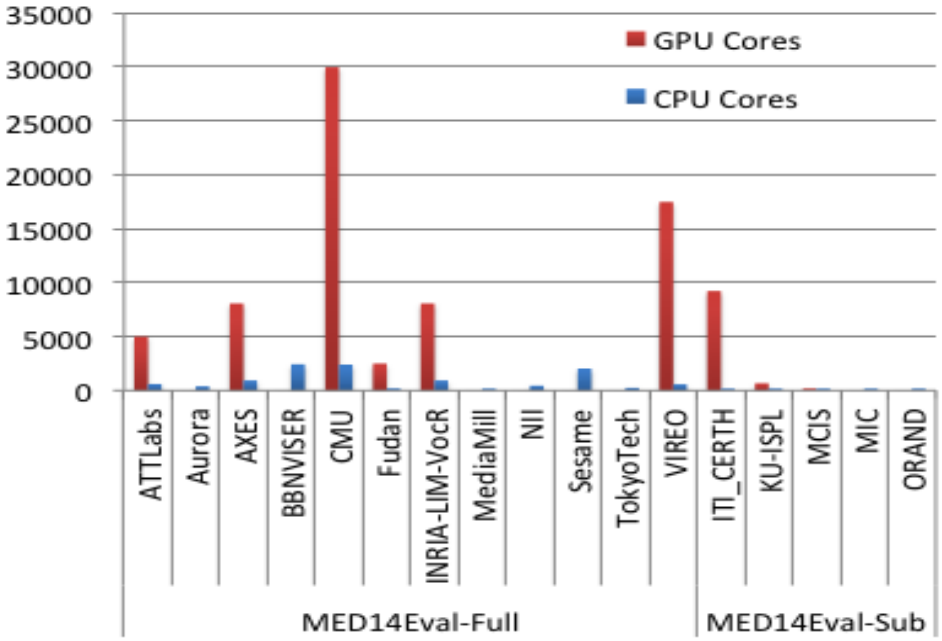

Figure 30: MED: Metadata sizes in terms of Real Size (metadata size/video size) for the MED14Eval-Full and MED14Eval-Sub collections broken down by the data type: signal (video and audio features), metadata (tags, actions, objects, etc.) and ASR/OCR

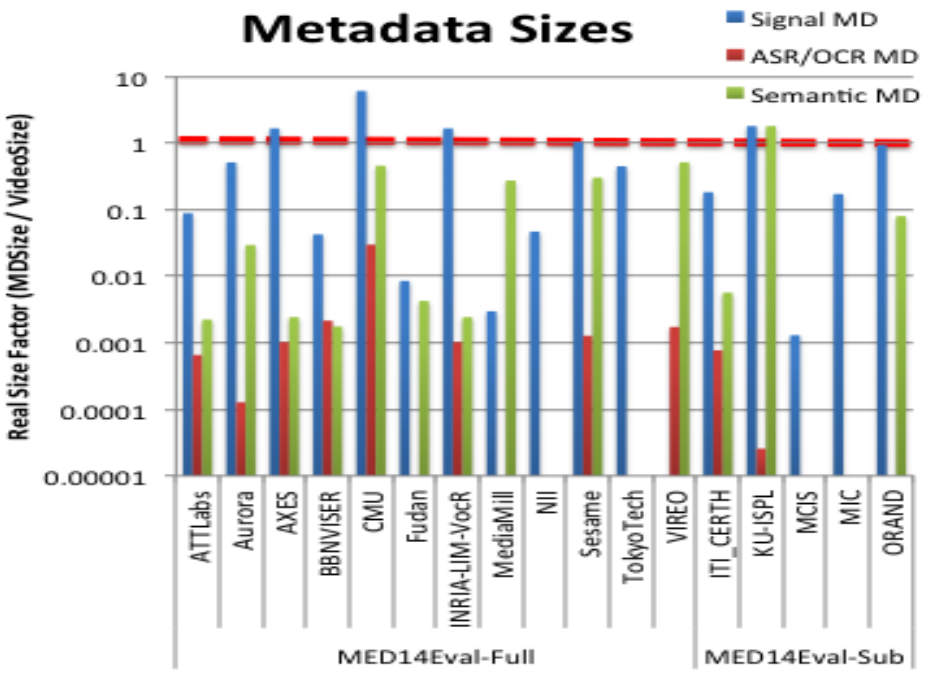

Figure 31: SED: Camera views and coverage

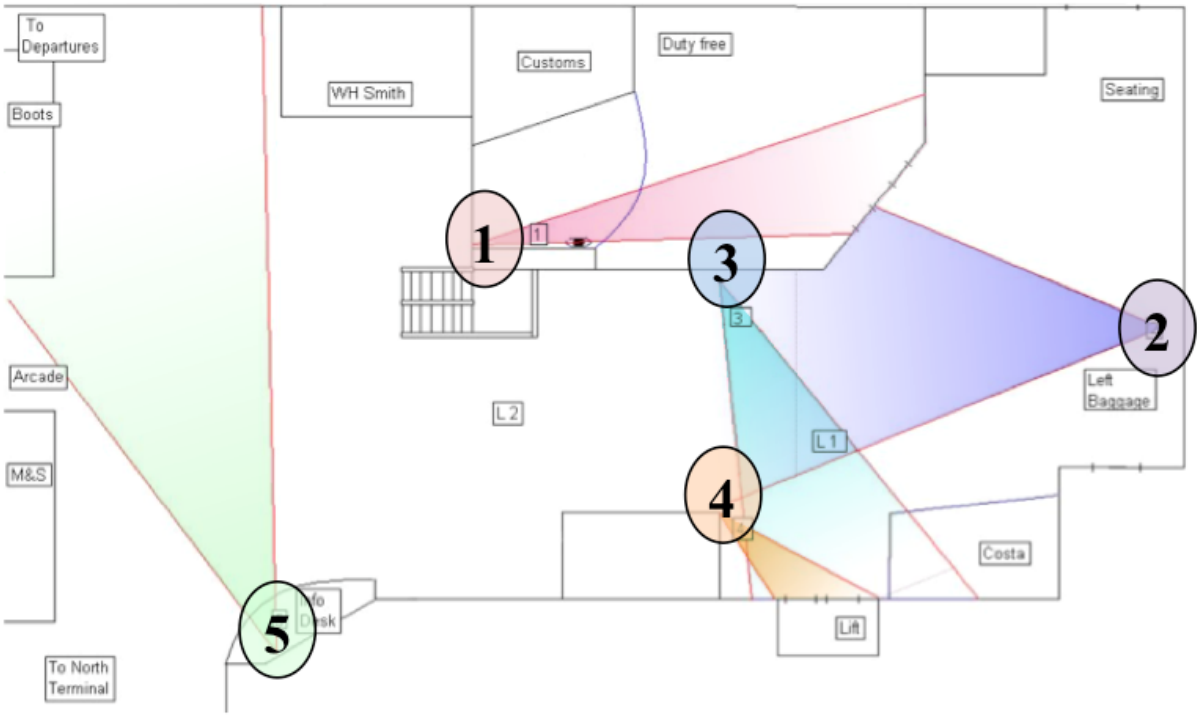

Figure 32: SED: Event name and definition

Single Person events	
PersonRuns	Someone runs
Pointing	Someone points
Single Person + Object events	
CellToEar	Someone puts a cell phone to his/her head or ear
ObjectPut	Someone drops or puts down an object
Multiple People events	
Embrace	Someone puts one or both arms at least part way around another person
PeopleMeet	One or more people walk up to one or more other people, stop, and some communication occurs
PeopleSplitUp	From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame

Figure 33: TRECVID 2014 SED Participants Chart

4 SED 2014 Participants

(with number of systems per event)

		Single Person				Person + object				Multiple People					
		PersonRuns		Pointing		CellToEar		ObjectPut		Embrace		PeopleMeet		PeopleSplitUp	
		iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED
7 years in a row	Carnegie Mellon University [CMU]	3	3	3	3	3	3	3	3	3	3	3	3	3	3
6 years in a row	Multimedia Communication and Pattern Recognition Labs, Beijing University of Posts and Telecommunications [BUPT-MCPRL]	2	2	2	2					2	2	2	2	2	2
3 years in a row	IBM Thomas J. Watson Research Center [IBM]	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	The City College of New York Media Lab [CCNY]		2		2		2		2		2		2		2
		6	8	6	8	4	6	4	6	6	8	6	8	6	8

Total iSED Runs	38
Total rSED Runs	52

Figure 34: SED'14: rSED and iSED - Embrace

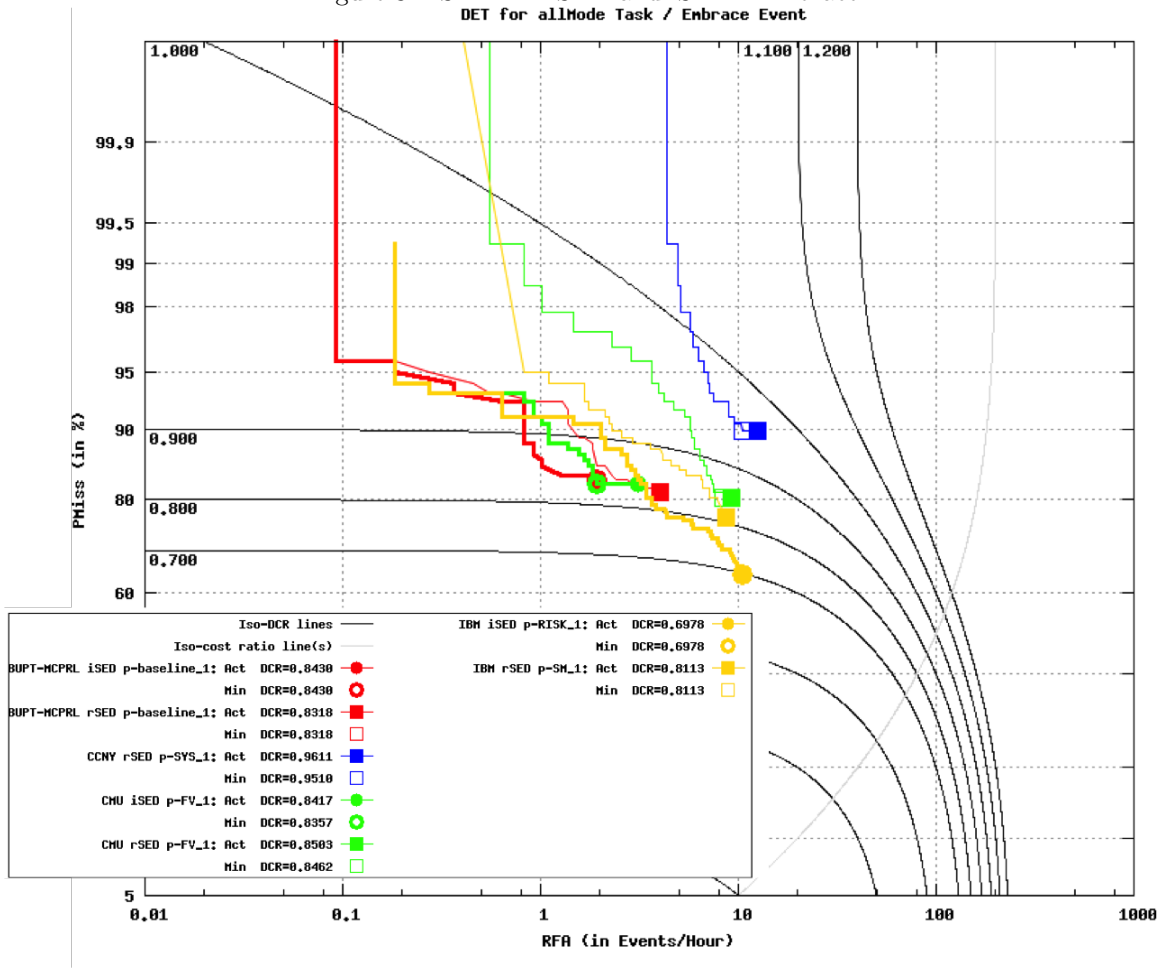

Figure 35: SED'14: rSED and iSED - PeopleMeet

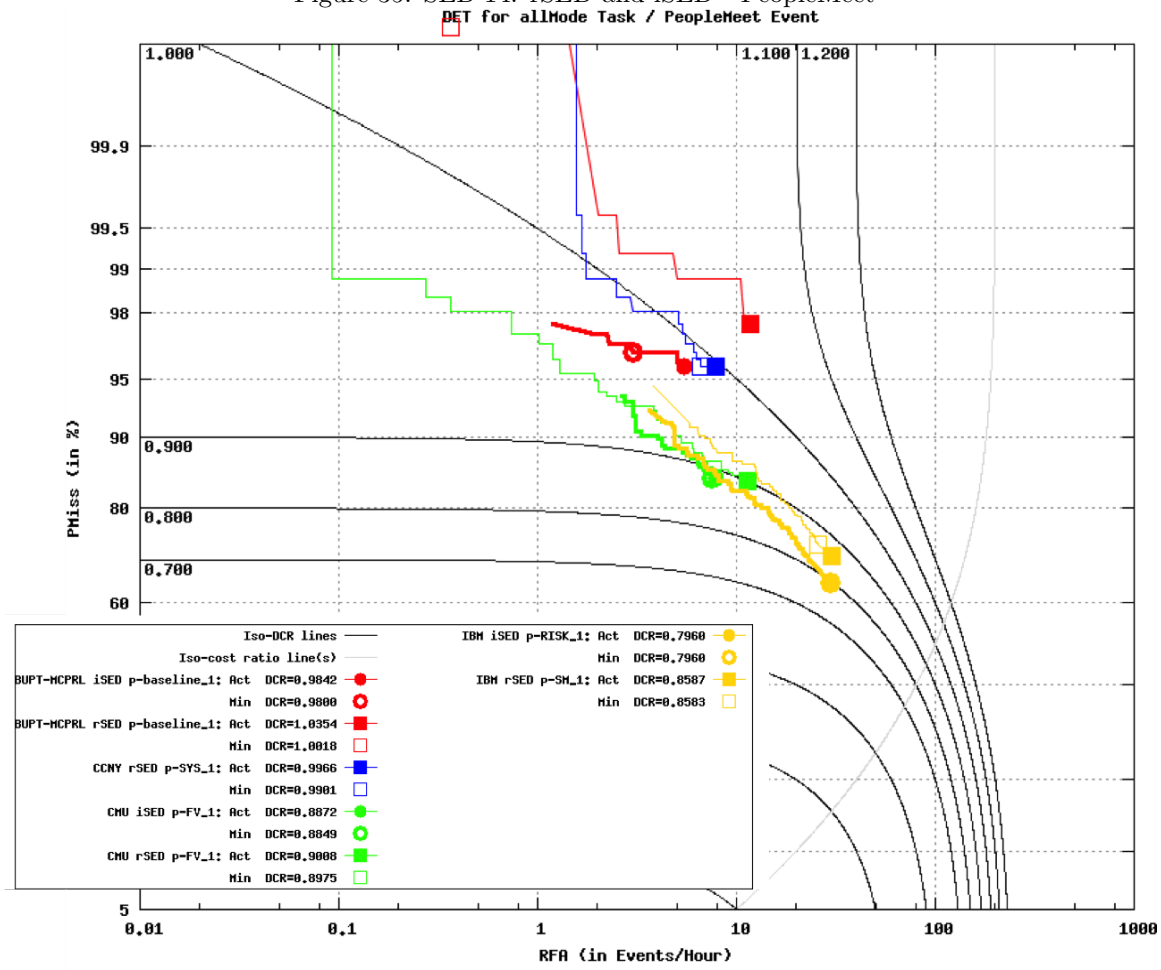

Figure 36: SED'14: rSED and iSED - PeopleSplitUp

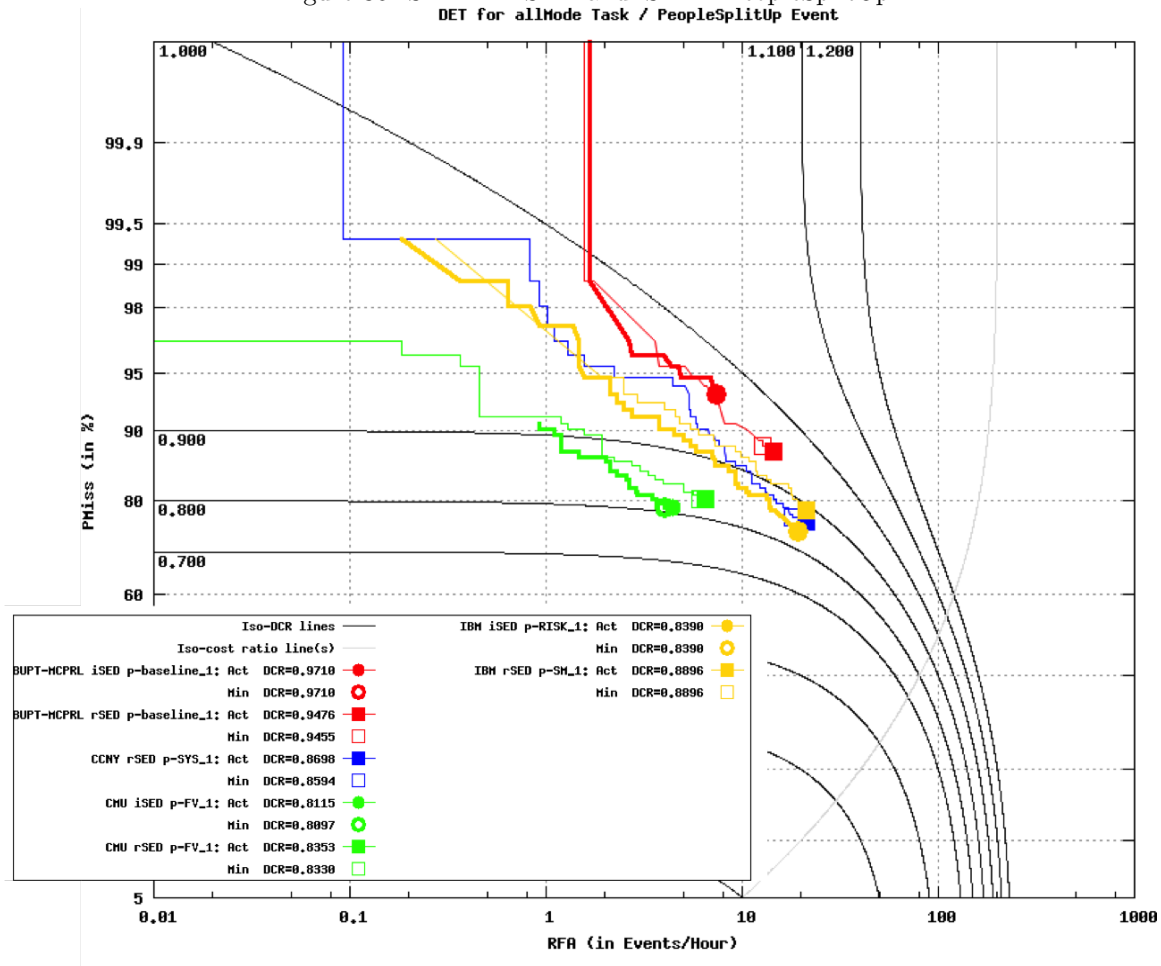

Figure 37: SED'14: rSED and iSED - PersonRuns

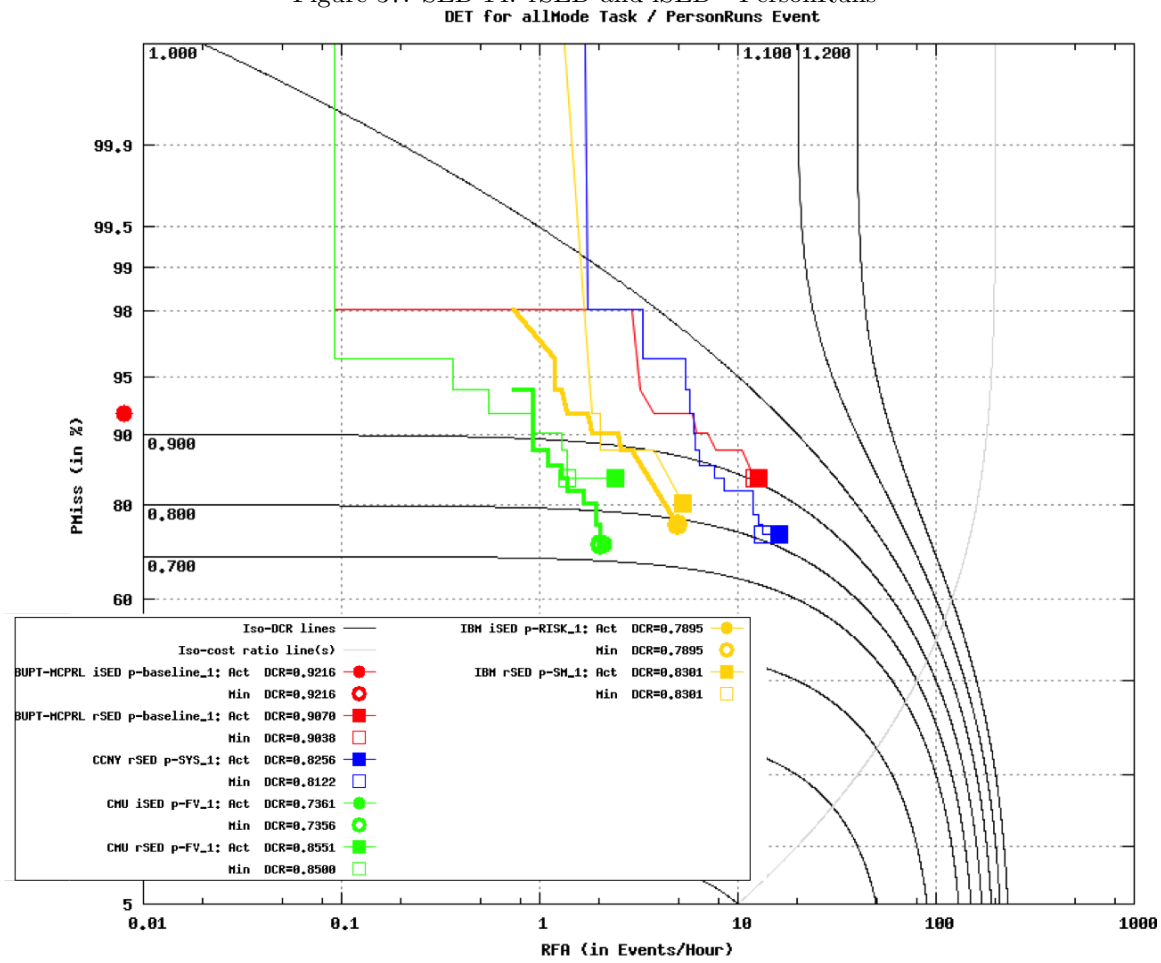

Figure 38: SED'14: rSED and iSED - Pointing