TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics

Paul Over, George Awad, Jon Fiscus, Martial Michel, David Joy, Alan F Smeaton, Wessel Kraaij, Georges Quénot, Roeland Ordelman, Robin Aly

TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics

Paul Over{retired},[1][2]George Awad{gawad@nist.gov}
[2]Jon Fiscus{jfiscus@nist.gov},[2]Martial Michel{martial.michel@nist.gov}
[2]David Joy{david.joy@nist.gov},[3]Alan F. Smeaton{Alan.Smeaton@dcu.ie}
[4]Wessel Kraaij{wessel.kraaij@tno.nl},[5]Georges Quénot{Georges.Quenot@imag.fr}
[6]Roeland Ordelman{roeland.ordelman@utwente.nl},[6]Robin Aly{r.aly@utwente.nl}

[1]Dakota Consulting Inc., Silver Spring, MD 20910

[2]Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

[3]Insight Centre for Data Analytics
Dublin City University,Glasnevin, Dublin 9, Ireland

[4]TNO,Delft, the Netherlands
Radboud University Nijmegen, Nijmegen, the Netherlands

[5]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP /
CNRS, LIG UMR 5217, Grenoble, F-38041 France

[6]University of Twente,7500 AE Enschede The Netherlands

August 6, 2016

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2015 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 and other US government agencies. Many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2015 represented a continuation of five tasks from 2014 and the addition of a new task from the Mediaeval workshop series: video hyperlinking. 46 teams (see Table 1) from various research organizations worldwide completed one or more of six tasks:

1. Semantic indexing (SIN)
2. Instance search (INS)
3. Multimedia event detection (MED)
4. Surveillance event detection (SED)
5. Video hyperlinking (LNK)
6. Concept localization (LOC)

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 (British Broadcasting Corporation) EastEnders video was reused for the instance search task. 96k I-frame images were used for testing in the localization task. 11 h of airport surveillance video was used for the surveillance event detection task, and almost 5 200 hours from the HAVIC (Heterogeneous Audio Visual Internet Corpus) collection of Internet videos was used for development and testing in the multimedia event detection task.

Semantic indexing, instance search, multimedia event detection, and localization results were judged by NIST assessors. The video hyperlinking results were assessed by Amazon Mechanical turk (Mturk) workers after initial manual check for sanity while the anchors were choosen by media professional at BBC and Netherlands Institute for Sound and Vision and journalism students. Surveillance event detection was scored by NIST using ground truth created by NIST through manual adjudication of test system output.

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

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

2 Data

2.1 Video

BBC EastEnders video

The BBC in collaboration the European Union's AXES (Access To AudioVisual Archives) 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 471 526 shots), transcripts, and a small amount of additional metadata.

BBC Hyperlinking video

The BBC in collaboration the Mediaeval workshop series provided about 3000 h of mixed BBC programming for research in hyperlinking.

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 will have some metadata provided by the donor available e.g., title, keywords, and description

For 2013, approximately 600 additional 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 2015. Most videos had some donor-supplied metadata available e.g., title, keywords, and description. Approximately 600 h of IACC.1 videos were available for system development.

As in the past, the Computer Science Laboratory for Mechanics and Engineering Sciences (LIMSI) and Vocapia Research provided automatic speech recognition for the English speech in the IACC.2 video.

iLIDS Multiple Camera Tracking Data

The Imagery Library for Intelligent Detection System's (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 iLIDS multiple camera tracking scenario data used for the 2009 - 2013 evaluations [UKHO-CPNI, 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

Table 1: Participants and tasks

Task						Location	TeamID	Participants
−−	∗∗	*LO*	∗∗	−−	SI	Eur	PicSOM	Aalto U.
IN	∗∗	∗∗	*MD*	*SD*	∗∗	Asia	BUPT_MCPRL	Beijing U. of Posts and Telecommunications
−−	−−	−−	*MD*	*SD*	–	Asia	Mcis	Beijing Institute of Technology
IN	∗∗	−−	*MD*	*SD*	SI	Eur	ITI_CERTH	Centre for Research and Tech. Hellas
−−	*HL*	−−	−−	−−	–	Eur	CUNI	Charles U. in Prague
−−	−−	*LO*	−−	−−	–	NAm	CCNY	City College of New York
∗∗	*HL*	−−	*MD*	−−	∗∗	Asia	VIREO	City U. of Hong Kong
−−	*HL*	∗∗	*MD*	*SD*	SI	NAm+Asia	CMU	CMU, CMU-Affiliates
−−	*HL*	−−	−−	−−	–	Eur	DCU_ADAPT_LNK	Dublin City U.
IN	−−	*LO*	−−	−−	SI	Eur	insightdcu	Dublin City U.; U. Polytechnica Barcelona
−−	−−	−−	*MD*	−−	–	NAm	etter	Etter Solutions LLC
−−	*HL*	−−	−−	−−	SI	Eur	EURECOM	EURECOM
−−	−−	−−	−−	−−	SI	NAm	FIU_UM	Florida International U.; U. of Miami
−−	−−	−−	*MD*	−−	–	Asia	BigVid	Fudan U.
−−	−−	−−	−−	*SD*	–	NAm	ibm	IBM
−−	*HL*	−−	−−	−−	–	Eur	IRISA	IRISA Inria Rennes - Bretagne Atlantique
−−	−−	−−	*MD*	−−	–	Asia	KoreaUnivISPL	Korea U.
∗∗	−−	−−	−−	−−	SI	Eur	IRIM	IRIM consortium
−−	−−	∗∗	∗∗	−−	SI	Eur	LIG	Laboratoire d'Informatique de Grenoble
−−	*HL*	−−	∗∗	−−	–	Eur	METU_EE	Middle East Technical U.
IN	∗∗	∗∗	*MD*	∗∗	SI	Asia	NII_Hitachi_UIT	Natl. Inst. of Inf.;Hitachi; U. of Inf. Tech.
−−	−−	−−	*MD*	−−	–	NAm	NEU_MITLL	Northeastern U.; MIT Lincoln Laboratory
−−	−−	−−	*MD*	*SD*	–	Asia	nttfudan	NTT Media Intelligence Laboratories; Fudan U.
IN	−−	−−	−−	−−	–	Asia	NTT	NTT Comm. Science Lab.; NTT Media Intel. Lab.
IN	*HL*	−−	∗∗	−−	–	SAm	ORAND	ORAND S.A. Chile
IN	∗∗	∗∗	∗∗	∗∗	∗∗	Asia	PKU_ICST	Peking University
−−	−−	−−	*MD*	*SD*	–	Asia	BCMI.SJTU	Shanghai Jiao Tong U.
−−	−−	−−	∗∗	*SD*	–	Asia	SeuGraph	Southeast U. Jiulonghu Campus
IN	−−	−−	−−	−−	–	Eur	TUC	Technische Universitaet Chemnitz
IN	−−	*LO*	−−	∗∗	–	Asia	Trimps	Third Research Inst., Ministry of Public Security
∗∗	∗∗	∗∗	∗∗	*SD*	∗∗	Asia	TJU_TJUT	Tianjin U.; Tianjin U. of Technology
∗∗	∗∗	*LO*	*MD*	−−	SI	Asia	TokyoTech	Tokyo Inst. of Tech.
IN	−−	∗∗	∗∗	∗∗	∗∗	Asia	Tsinghua_IMMG	Tsinghua U.
−−	*HL*	−−	−−	−−	–	Asia	TUZ	TUBITAK UZAY
∗∗	∗∗	*LO*	*MD*	−−	SI	NAm+Eur	MediaMill	U. of Amsterdam Qualcomm
IN	∗∗	−−	−−	−−	–	Eur+Asia	Sheffield_UETLahor	U. of Sheffield, UK; U. of Eng. and Tech.,Pakistan
−−	−−	−−	*MD*	−−	SI	Eur+Asia	siegen_kobe_nict	U. of Siegen; Kobe U.; Natl. Inst. of Inf. and Comm. Tech.
−−	−−	−−	∗∗	−−	SI	NAm	UCF_CRCV	U. of Central Florida
−−	∗∗	∗∗	*MD*	−−	SI	Asia	UEC	U. of Electro-Communications, Tokyo
IN	−−	−−	∗∗	∗∗	∗∗	Aus	UQMG	U. of Queensland - DKE Group of ITEE
−−	−−	−−	−−	*SD*	–	Aus	WARD	U. of Queensland
−−	−−	−−	−−	−−	SI	Asia	Waseda	Waseda U.
−−	*HL*	−−	−−	*SD*	–	Asia	IIP_WHU	Wuhan U.
IN	−−	∗∗	−−	−−	–	Asia	U_TK	U. of Tokushima
IN	−−	−−	−−	∗∗	∗∗	Asia	NERCMS	Wuhan U.;Natl. Eng. Research Center for Multimedia Software

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

years and 3 h reused. The overlap allowed some comparison of earlier versus new ground truthing. 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 (LDC) and distributed as MPEG-4 (MPEG-4, 2010) formatted files containing H.264 (H.264, 2010) encoded video and MPEG-4 Advanced Audio Coding (AAC) (AAC, 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.

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 ba-

Table 2: Participants who did not submit any runs

Task						Location	TeamID	Participants
IN	*MD*	*HL*	*LO*	*SD*	*SI*			
**	−−	−−	−−	−−	−−	Asia	DML_BUAA	Beihang U. - Beijing Key Laboratory of Digital Media
−−	**	−−	−−	−−	**	Eur	brno	Brno U. of Tech.
−−	−−	−−	**	−−	−−	Asia	TFV_CASIA	Chinese Academy of Sciences
−−	−−	−−	**	−−	−−	Eur+Asia	Vireo_TNO	City U. of Hong Kong; TNO
−−	**	−−	−−	−−	−−	Eur	TUD	Delft U. of Tech.
**	**	−−	**	**	−−	Asia	FZUCV	Fuzhou U.
−−	−−	−−	**	−−	−−	Asia	HEU	Harbin Engineering U.
−−	−−	−−	**	**	−−	Eur	INRIA_STARS	INRIA
**	**	−−	−−	−−	−−	Eur	JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH
−−	−−	−−	−−	**	−−	Asia	PRLab	Korea U. - Pattern Recognition Lab
**	−−	−−	−−	**	−−	SAm	MindLAB	Mind LAB Research Group
−−	−−	−−	−−	**	−−	Asia	MELCO_ATC	Mitsubishi Electric Corporation
−−	−−	−−	**	−−	−−	Asia	NLPR_13	Natl. Lab. of Pattern Recognition, CAS
−−	−−	−−	−−	**	**	NAm	DFE	NGA/IID
−−	−−	−−	**	−−	−−	Eur	VisQMUL	Queen Mary, U. of London
−−	**	−−	−−	−−	−−	Afr	REGIMVID	REGIM LAB (ENIS) Sfax Tunisia
**	**	**	**	**	**	Asia	SNUMadInfo	Seoul National U.
**	−−	−−	**	**	**	Asia	sjtuis	Shanghai Jiao Tong U.
**	−−	−−	−−	−−	**	Asia	TjuMMTeam	Tianjin U. MM Team
**	−−	−−	**	−−	−−	Asia	MIC_TJU	Tongji U.
**	−−	−−	−−	−−	−−	NAm+Asia	THU_UTSA	Tsinghua U.; U. of Texas, San Antonio
−−	−−	**	**	−−	**	Eur	DAG_CVC	Universitat Autonoma de Barcelona
**	−−	−−	−−	−−	−−	NAm	CVIS	U. of North Texas
**	−−	**	**	**	**	NAm	UCR_Vislab	U. of California, Riverside
**	−−	**	−−	−−	−−	Asia	U_TK	U. of Tokushima

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

sis 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.

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.

Measurement of system progress for a fixed set of concepts and independent of the test data, across 3 years (2013-2015) was concluded this year. Evaluation measures should be able to show how much progress systems achieved in those 3 years by freezing the tested dataset and evaluated concepts while only changing the applied system approaches.

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 Columbia University and VIREO (Video Retrieval Group) team for which there exist annotations on TRECVID 2005 data, and some from the Large Scale Concept Ontology for Multimedia (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 2014 the same list of 60 single concepts were used this year for which participants must submit results in the main task.

In 2015, the task again supported experiments us-

ing the "no annotation" version of the tasks. The idea is 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 of "E" and "F" for the training types of submitted run files besides the A to D ones:

- 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
- E - used only training data collected automatically using only the concepts' name and definition
- F - used only training data collected automatically using a query built manually from the concepts' name and definition

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 year only 1 "main" type of submissions were considered in which participants submitted results for 60 single concepts.

TRECVID evaluated 30 of the 60 submitted single concept results The 60 single concepts are listed in Appendix A. Those that were evaluated are marked with an asterisk.

Concepts were defined in terms a human judge could understand. The fuller concept definitions provided to system developers and NIST assessors are listed on the webpage: `http://www-nlpir.nist.gov/projects/tv2012/tv11.sin.500.concepts_ann_v2.xls`

Work at Northeastern University [Yilmaz and 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 et al., 2006]. This year mean extended inferred average precision (mean xinfAP) was used which permits sampling density to vary [Yilmaz et al., 2008]. This allowed the evaluation to be more sensitive to shots returned below the lowest rank (≈100) previously pooled and judged. It also allowed adjustment of the sampling density to be greater among the highest ranked items that contribute more average precision than those ranked lower.

3.1 Data

The IACC.2.C collection was used for testing. It contained 113 046 shots.

3.2 Evaluation

Each group was allowed to submit up to 4 prioritized main runs and two additional if they are "no annotation" runs. In fact 15 groups submitted a total of 56 runs. In addition to the 56 runs submitted against the IACC.2.C dataset this year, there was 20 runs submitted in TRECVID 2013 and 9 runs submitted in 2014 as part of the progress subtask as well which are all evaluated this year.

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 were evaluated in 2015 only and 15 decided to be common in both years. Randomization tests experiments on the chosen 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, 195 500 shots were judged while 661 519 shots fell into the unjudged part of the overall samples.

3.3 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, given the sampling plan and a submitted run. Since all runs provided results for all evaluated concepts, runs can be compared in terms of the mean inferred average precision across all evaluated single concepts. The results also provide some information about "within concept" performance.

3.4 Results

Performance varied greatly by concept. Figure 1 shows how many unique instances were found for each tested concept. The inferred true positives (TPs) of 6 concepts exceeded 1 % from the total tested shots. Top performing concepts were "Traffic", "Government_Leaders", "Computers", "Old_People", "Instrumental_Musician", "Studio_with_AnchorPerson", "AnchorPerson".

On the other hand, concepts that had the fewest TPs were "Car_Racing", "Basketball", "Motorcycle", "Hill", "Bicycling", "Telephones", "throwing".

Figure 2 shows the results of all the main run submissions (color coded). Category A runs used only IACC training data, Category C used IACC data as well as non-IACC but TRECVID data while Category D runs used IACC and non-trecvid data as well.

The median score across all runs was 0.239 compared to 0.217 in 2014 while maximum score reached 0.362 compared to 0.332 in 2014.

Category D runs were the most popular type and achieve top recorded performances.

Figure 3 shows the distribution of the run scores including the scores of progress runs which were submitted in 2013 and 2014 against the 2015 testing dataset. Most of the progress teams achieved better performance in 2015 compared to their 2013 and 2014 submissions.

Figure 4 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 "Anchorperson", "Bicycling", "Bridges", "cheering", "Dancing", "Kitchen","Telephones", and "Soldiers". While others had more bigger spread such as "Flags", "throwing", "Motorcycle", "Instrumental_Musician", "Government_Leaders", "Computers", and "Bicycling". The spread in scores may indicate the variation in used techniques performance and there is still room for further improvement.

On the other hand few other concepts had very tight spread (performance almost similar) across the top 10 such as "Airplane", "Boat_Ship", "Car_Racing", "Demonstration_or_Protest", "Office", "Press_Conference", and "Hill".

In general, the most common concepts between 2014 and 2015 achieved higher max scores compared to 2014 and the median scores ranged between minimum 0.003 ("Car_Racing") and maximum 0.740 ("Anchorperson").

To test if there were significant differences between the systems' performance, we applied a randomization test [Manly, 1997] on the top 10 runs (Figure 5) as shown in Figure 6. 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 3 ranked runs was significantly better than all other runs while there is no significant difference between the three of them.

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 7 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 "Computers" (1031) and "Telephones"(1117), "Dancing" (1038) and "Instrumental_musician" (1071), "Bus" (1019) and "Traffic" (1478), "Flags"(1261) and "Government_Leader" (1056), and "Motorcycle" (1080) and "Traffic" (1478).

Another experiment to measure how diverse is the submitted runs we measured the percentage of com-

mon shots across the same concepts between each pair of runs. We found that on average about 23% (minimum 14%) 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.C dataset to compare their 2013 and 2014 systems with the 2015 system and measure how much progress they made. Figure 8 shows the best run score by team in both years. 4 out of 6 teams achieved better scores in 2015 compared to 2013 and 2014. One team had 2015 system better than only the 2013 while another team had the 2015 system better than only the 2014 system.

Randomization tests show that the 2015 systems are better than the corresponding 2013 and 2014 systems except for one team (UEC) where their 2014 system where better. The maximum improvement reached 0.212 mean InfAP for the team of EURECOM between their 2015 and 2013 system while the team IRIM improved their 2015 system by 0.070 compared to their 2014 system.

We also measured the performance per concept for each team to find how many concepts were improved in 2015. It can be seen in Figure 9 that most concepts were improved in 2015 compared to 2013 and 2014.

2015 Observations

Finally, to summarize our observations about the overall task performance and general ideas or techniques used by participating teams we found that 2015 task was harder than 2014 that was itself harder than 2013 main task (different data and different set of target concepts). Raw system scores have higher Max and Median compared to 2014 and 2013. Although still relatively low they are regularly improving. Most common concepts between 2014 and 2015 have higher median scores. Most Progress systems improved significantly from 2014 to 2015 as this was also the case from 2013 to 2014. Participation (15 teams) between 2014 and 2015 seems to be stable. In terms of applied methods and techniques we see many more systems based on deep learning where some approaches used trained ImageNet networks. Some reported the usage of parallel deep networks. Data augmentation for training has been applied by multiple teams as well as the usage of multiple frames per shot for prediction. Some teams used gradient and motion features for DCNNs (Deep Convolutional Neural Networks). Some hybrid approaches have been reported where DCNN-based learning has been combined with classical learning. Engineered features are still used mainly as a complementary features (mostly Fisher Vectors, SuperVectors, improved BoW, and similar) but with no new developments. For detailed information about the approaches and results, the reader should see the various site reports [TV15Pubs, 2015] and the results pages in the online workshop notebook [TV15Notebook, 2015].

4 Concept Localization

The localization task challenges systems to make their concept detection more precise in time and space. Currently SIN systems are accurate to the level of the shot. In the localization task, systems are asked to determine the presence of the concept temporally within the shot, i.e., with respect to a subset of the frames comprised by the shot, and, spatially, for each such frame that contains the concept, to a bounding rectangle.

The localization is restricted to 10 concepts from those chosen and used in the semantic Indexing task. However, systems can participate in the localization task without submitting runs in the Semantic Indexing task as both tasks this year are run independently.

For each concept from the list of 10 designated for localization, NIST distributed, about 5 weeks before the localization submissions were due at NIST for evaluation, a subset list of up to 300 shots which were judged as true positives by the human assessors in the Semantic Indexing main task and so contain the concept that needs to be localized. Figure 10 shows the evaluation framework used starting from judging Semantic Indexing results till producing the localization ground truth.

For each I-Frame within each shot in the list that contains the target, systems were asked to return the x,y coordinates of the upper left and lower right vertices of a bounding rectangle which contains all of the target concept and as little more as possible. Systems may find more than one instance of a concept per I-Frame and then may include more than one bounding box for that I-Frame, but only one will be used in the judging since the ground truth will contain only 1 per judged I-Frame, one chosen by the NIST assessor that is most prominent.

Table 3 describes for each of the 10 localization concepts the number of shots NIST distributed to

Concept ID	*Name*	*shots*	*I-Frames*
3	Airplane	300	7047
5	Anchorperson	300	14119
15	Boat_Ship	300	5874
17	Bridges	300	6054
19	Bus	190	4774
31	Computers	300	15814
80	Motorcycle	99	4165
117	Telephones	145	5851
261	Flags	271	19092
392	Quadruped	300	13949

Table 3: Evaluated localization concepts

systems and the number of I-Frames comprised by those shots:

The larger numbers of I-Frames to be judged for concepts 5,31,261 and 392 within the time allotted caused us to assign some of those images to assessors who had not done the original shot judgments in the semantic indexing task. 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.

4.1 Data

In total, 1 581 537 jpeg I-frames were extracted from the IACC.2.C collection. 2 505 total shots were distributed and included total of 96 739 I-frames.

4.2 Evaluation

This year total of 6 teams finished the task submitting total of 21 runs. For each shot that contains a concept and selected and distributed by NIST, a systematic sampling was employed to select I-frames at regular intervals from the shot. This year an interval value of 2 (every other I-frame) was applied to fit total of 200 hours of human assessors work given that each assessor can judge an average of about 6 000 images. Selected I-frames were displayed to the assessors and 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. In total, 48 136 I-frames were judged.

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 has 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.

4.3 Measures

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 contains 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.

4.4 Results

In this section we present the results based on the temporal and spatial submissions across all submitted runs as well as by results per concepts. Figure 11 shows the mean precision, recall and F-score of the returned I-frames by all runs across all 10 concepts. In general systems reported much higher F-score values compared to the last two years (max F-score reached 0.2 in 2014) as 9 out of 21 runs scored above 0.7 and 8 runs scored above 0.6 f-score.

We believe these high scores are side effect of only localizing true positive shots (output of the semantic indexing task) compared to localizing just raw shots which may include true positive as well as true negative concepts. On the other hand Figure 12 shows the same measure by run for spatial localization (correctly returning a bounding box around the concept). Here the F-scores range were less than the temporal F-score range but still higher than the past two years. Overall 8 out of the 21 runs scored above 0.5 and another 8 runs exceeded 0.4 compared to maximum of about 0.3 score in the last 2 years.

The F-score performance by concept for the top 10 runs is shown in Figures 13 and 14 for temporal and spatial respectively across all runs. In general, most concepts achieved higher temporal scores (between about 0.5 for the concept "bus" and 0.9 for the concept "Flags") compared to spatial localization. A big variation (ranged between 0.3 - 0.4) is noticed in

the performance of about 5 concepts in the spatial domain across the top 10 runs compared to the temporal domain where the performance of the top 10 runs seem to be very close. Finally, both measures are better compared to the past two years given that 8 out of the 10 concepts were fixed.

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 15 and 16 for temporal and spatial localization respectively. We can see in Figure 15 that most concepts achieved very high the precision and recall values (above 0.5).

An interesting observation in Figure 16 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 17 and 18 show visual examples of good and weaker 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. We note here that those examples were chosen to demonstrate the accuracy of some systems in localizing some hard (small,occluded,low-illumination, etc) frames (17) while other systems struggled in localizing very obvious concepts (big, clear,centered, etc) (18).

Summary of observations

It is clear that for the past 3 years the temporal localization was easier than the spatial localization. This current year the scores were significantly high mainly because we aimed to make systems just focus on the localization task bypassing any prediction steps to decide if a video shot include the concept or no as were done in the past two years in the semantic indexing task. This may have caused the task to be relatively easy compared to a real world use case where a localization system would have no way to beforehand know if the video shot already include the concept or no. Thus, next year we plan to give systems raw shots (may include true positive or true negative concepts) simulating a semantic indexing predicted shot list for a given concept. We also plan to test systems on a new set of concepts which may include some actions which span much more frames temporally compared to only objects that may not include much motion.

For detailed information about the approaches and results, the reader should consult the various site reports [TV15Pubs, 2015] and the results pages in the online workshop notebook [TV15Notebook, 2015].

5 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.

5.1 Data

The task was run for three years starting in 2010 to explore task definition and evaluation issues using data of three sorts: Sound and Vision (2010), BBC rushes (2011), and Flicker (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 471 526 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).

5.2 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

Table 4: 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
9129	43684	23079	52.8	240	6336	27.5	265	4.2
9130	43853	16782	38.3	460	8720	52	1735	19.9
9131	43784	20220	46.2	360	7855	38.8	402	5.1
9132	43691	23675	54.2	360	9678	40.9	68	0.7
9133	43647	22386	51.3	200	4940	22.1	112	2.3
9134	43878	20094	45.8	460	9851	49	472	4.8
9135	43625	20948	48	180	4618	22	60	1.3
9136	43648	17801	40.8	320	6884	38.7	83	1.2
9137	43663	18057	41.4	440	8901	49.3	134	1.5
9138	43743	18036	41.2	360	7255	40.2	448	6.2
9139	43712	25520	58.4	100	3185	12.5	33	1
9140	43646	18322	42	280	6047	33	95	1.6
9141	43620	21810	50	200	5079	23.3	52	1
9142	43717	21691	49.6	300	7551	34.8	44	0.6
9143	43724	20436	46.7	280	6903	33.8	105	1.5
9144	43753	22209	50.8	340	7644	34.4	256	3.3
9145	43621	20594	47.2	360	8376	40.7	397	4.7
9146	43881	17891	40.8	280	5390	30.1	528	9.8
9147	39667	22356	56.4	260	6783	30.3	19	0.3
9148	39580	13074	33	460	6464	49.4	1308	20.2
9149	39599	14925	37.7	420	6300	42.2	286	4.5
9150	39591	14542	36.7	360	5348	36.8	1103	20.6
9151	39602	17252	43.6	460	8873	51.4	94	1.1
9152	39578	15727	39.7	460	7032	44.7	638	9.1
9153	44237	13295	30.1	460	5787	43.5	874	15.1
9154	43927	18374	41.8	460	7625	41.5	747	9.8
9155	43722	16238	37.1	460	7348	45.3	127	1.7
9156	44006	17526	39.8	380	6768	38.6	661	9.8
9157	44288	11826	26.7	340	4545	38.4	682	15
9158	43804	19034	43.5	440	7441	39.1	437	5.9

frame image:

  - a binary mask covering one instance of the target
  - the ID of 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 one or more provided images - no video used

E video examples (+ optionally image examples)

5.3 Topics

NIST viewed every 10th test video and developed a list of recurring objects, people, and locations. 30 test queries (topics) were then created (Appendix B). As in 2013 and 2014, the topic targets included mostly small and large rigid objects, logos, and people/animals.

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.

5.4 Evaluation, Measures

Each group was allowed to submit up to 4 runs (8 if submitting pairs that differ only in the sorts of examples used) and in fact 13 groups submitted 43 automatic and 7 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 4 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 (MAP) over all queries. While speed and location accuracy were also definitely of interest here, of these two, only speed was reported.

5.5 Results

Discussion

Figure 19 shows the distribution of automatic run scores (average precision) by topic as a box plot. The topics are sorted by the maximum score with the best performing topic on the left. Median scores vary from nearly 0.5 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 10 of the 15 with the best scores, while such topics make up only 7 of the bottom 15 topics. Figure 20 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 $p < 0.05$;

In Figure 21, a box plot 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, the stationary, rigid targets make up 9 of 12 in the top halve of the topic ranking while such topics make up only 2 of the bottom 12 topics.

Figure 22 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).

The relationship between the two main measures - effectiveness (mean average precision) and elapsed processing time is depicted in Figure 23 for the automatic runs with elapsed times less than or equal to 10 s.

Although some of 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 24. 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 25 shows the relationship between the two category of runs (images only for training OR video and images) and the effectiveness of the runs. The results show that the few runs that took advantage of the video examples are the ones that achieved the highest scores. On the other hand the majority of the rest of the runs used only the image examples. This was the second year video for the images examples was made available and we hope more systems will use those examples in future years for better training data.

In summary, the effectiveness of systems has increased this year compared to the past two years working on the same data and type of queries, the persons category of queries are still the most difficult, the E condition (using video examples) was used by just a few (top runs) teams, and the interactive search task helps improving the MAP of instances with varying backgrounds (mobile).

Approaches

In general, nearly all systems use some form of SIFT local descriptors where large variety of experiments are addressing representation, fusion or efficiency challenges. Most systems also include a CNN (Convolutional Neural Networks) component. There is a better understanding of when CNN can help. Many experiments include post-processing (spatial verifica-

tion, feedback). Few teams started to explore closed captions and fan resources for additional evidence (using topic descriptive text).

A summary of team efforts in order to find an optimal representation includes: Wuhan team reported improvement from processing more frames, the BUPT and PKU-ICST teams combined different feature types (local/global) and fusion of CNN, SIFT BOW (Bag Of Words) and text captions. LAHORE and SHEFFIELD compared 4 different combinations of 4 different local features and 4 matching methods. Trimps team compared BOW based on SIFT with RCNN global features, selective search and CNN with LSH (Locality-Sensitive Hashing) and HOGgles with local features. TU_Chemnitz team explored the classification of the audio track as in 2014. UMQG team presented a new approach based on object detection and indexing where CNN was used to describe extracted objects from video decomposition and then matching the query image with nearest object in a codebook and quantization framework.

In regard to exploiting the query images/videos the Wuhan team manually selected ROI (region of interest) on different query images which helped significantly their system while exploiting the full query video was applied by PKU_ICST, NERCMS, Wuhan and Chemnitz teams.

Different matching experiments are reported by systems. Typically inverted files for fast lookup in sparse BovW space and pseudo relevance feedback for query expansion are mentioned in several reports. Other teams experimented with similarity functions. For example BUPT team used query adaptive late fusion while Wuhan team applied Asymmetric query adaptive matching.

Postprocessing the ranked list results also has been investigated by InsightDCU team where weak geometry consistency check for spatial filtering helped to refine results. The NII-HITACHI team applied DPM (deformable part models) and Fast RCNN in their postprocessing experiments. The Wuhan team applied face and color filters with adjacent shot matching and query text expansion. The NTT team used spatial verification methods such as Ensemble of weak geometric relations and Angle free hough voting in 3D camera motion space. Finally the TU_Chemnitz team used indoor/outdoor detectors based on audio analysis for removal of false matches in addition to clustering similar shot sequences.

In regard to the interactive runs' experiments the TU_Chemnitz improved their system usability to fast review 3500 instances which improved on their automatic results. The PKU_ICST team used 2 rounds of relevance feedback on the initial run and fused the results with the original run results.

For more detailed information about the approaches and results, the reader should see the various site reports [TV15Pubs, 2015] and the results pages in the online workshop notebook [TV15Notebook, 2015].

5.6 Summary and Future Plans

The past 3 years has been successful exercise for participating teams to learn new methods and techniques to detect different set of objects, persons and location instances (about 90 total instance queries between 2013-2015). In the next year we are planning to experiment with a new sort of topics on the same test data. The new topics will ask participants to search for shots containing a specific target person in a specific target location given a set of training videos for named locations and set of ad hoc target persons each with 4 image and video examples.

6 Multimedia Event Detection

The 2015 Multimedia Event Detection (MED) evaluation was the fifth evaluation of technologies that search multimedia video clips for complex events of interest to a user.

The focus of the MED 15 evaluation was to modify the structure of the evaluation to make MED less costly to both participate in and administer. This focus resulted in several simplifications including:

- The Semantic Query evaluation condition was not supported. The MED '14 zero examplar and semantic query conditions are very similar not warranting the extra condition.
- The 100-Exemplar Ad-Hoc evaluation condition was not supported. The motivation to include the 100 exemplar condition was to make an easy, initial test condition five years ago. Given the current performance, 10-exemplars is sufficient. In addition, building a 100-exemplar event kit would be a considerable burden on the the user.
- Multimedia Event Recounting (MER) is no longer a supported evaluation task. The MER evaluations require considerable human annotation effort. In the interest of a lean MED 16, the track was discontinued.

- Reporting of computational resources was greatly simplified from reporting hardware description, computation times at the subcomponent level, and disk usage to participants labelling submissions as the closest match to small, medium, and large cluster configurations.
- Mean Inferred Average Precision [Yilmaz et al., 2008] became the primary metric and references for the (Ad-Hoc) events were generated through pooled assessment.

Despite the several scope reductions, there were two expansions this year including:

- 10 new events were used to evaluate the Ad-Hoc systems.
- A new Interactive Ad-Hoc Event condition was defined.

6.1 Task

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.
- 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.

Within the general area of finding instances of events, the evaluation included three styles of system operation. The first is for Pre-Specified event systems where knowledge of the event(s) was taken into account during generation of the metadata store for the test collection. This style of system has been tested in MED since 2010. The second style is the Ad-Hoc event task where the metadata store generation was completed before the events where revealed. This style of system was introduced in MED 2012. The third style, interactive Ad-Hoc event task, is a variation of Ad-Hoc event detection with 15 minutes of human interaction to search the evaluation collection in order to build a better query. No one participated in this task.

6.2 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 AAC standard.

MED participants were provided the data as specified in the HAVIC data section of this paper. The MED '15 Pre-Specified event names are listed in Table 5 and Table 6 lists the MED '15 Ad-Hoc Events.

6.3 Evaluation

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

- Events: either all 20 Pre-Specified events (PS15) and/or all 10 Ad-Hoc events (AH15).
- Interactivity: Human interaction with query refinement using the search collection.

Table 5: MED '15 Pre-Specified Events

— MED'12 event re-test
Attempting a 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 6: MED '15 Ad-Hoc Events

E061 - Gardeners harvest food
E062 - Land vehicle accident
E063 - Person jumps into natural water
E064 - Cooking on an outdoor grill
E065 - Moving through a flooded street
E066 - Skyscraper window cleaning
E067 - Firefighters battle a fire
E068 - Climbing a tree
E069 - Lecture to an audience
E070 - Team scores a touchdown

- Test collection: either the MED15 Full Evaluation collection (MED15-EvalFull) or a 1 238 hour subset (MED15-EvalSub) collection. The search collections were identical to last year.
- Query Conditions: 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 50 miss clips per event. Only for the PS condition).
- Hardware Definition: Teams self-reported the size of their computation cluster as the closest match to the following three standards:
  - SML - Small cluster consisting of 100 CPU cores and 1 000 GPU cores
  - MED - Medium cluster consisting of 1 000 CPU cores and 10 000 GPU cores
  - LRG - Large cluster consisting of 3 000 CPU cores and 30 000 GPU cores

Full participation requires teams to submit both 10Ex, PS and AH systems.

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.

Rather than submitting detailed runtime measurements to document the computational resources, this year participants labeled their systems as the closest match to a three cluster sizes: small, medium and large. (See above.)

Submission performance was computed using the Framework for Detection Evaluation (F4DE) toolkit. InfAP scores were computed using the procedure described by Yilmaz et al., A simple and efficient sampling method for estimating AP and NDCG [Yilmaz et al., 2008].

6.4 Measures

System output was evaluated by how well the system retrieves and detected MED events in evaluation search video metadata. 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 was Mean Inferred Average Precision.

6.5 Results

16 teams participated in the MED '15 evaluation; 6 teams were new. All teams participated in the Pre-Specified (PS) Event, 10 Exemplar (10Ex) test, processing all 20 events. 6 teams chose to participate in the Ad-Hoc (AH) event, 10 Exemplar (10Ex) test, processing all 10 events. 8 Teams chose to process the MED15-EvalSub set.

For the MED15 evaluation, for the Pre-Specified events we reported both the MAP and Mean Inferred Average Precision (MInfAP) as well as the correlation between the two measures. For the Ad-Hoc events, we reported the Mean Inferred Average Precision. This year, we also reported Average Precision and Inferred Average Precision for both Pre-Specified and Ad-Hoc events by event.

Figures 26 and 27 show the MAP scores per team for both the MED15-EvalFull and MED15-EvalSub sets respectively. Results are broken down by hardware classification and exemplar training condition. As with last years results, MAP scores for MED15-EvalSub are inflated when compared to scores on MED15-EvalFull due to the higher density of positives in the MED15-EvalSub set. That said, MAP scores between MED15-EvalFull and MED15-EvalSub sets are highly correlated; with an R^2 of 0.993.

Figures 28 and 29 show the Pre-Specified Average Precision scores for MED15EvalSub on the 10Ex exemplar training condition broken down by event and team respectively.

In Figure 30, an event effect can be observed on the spread of MAP scores aggregated over teams for Pre-Specified 10Ex. Also note that the MAP scores across "Medium" (MED) sized systems tend to have a tighter range, and are noticably higher than the collection of "Small" (SML) systems. Teams that have participated several years processed the full collection and built medium systems. The notable exceptions are NTT and Fudan, for which this is the second year. That said, it's worth mentioning that 12 of 16 system submissions for the Pre-Specified 10Ex condition were "Small" systems.

For the Mean Inferred Average Precision (MInfAP), we follow Aslam et. al. procedure, Statistical Method for System Evaluation Using Incomplete Judgments [Yilmaz and Aslam, 2006], whereby we use a stratified, variable density, pooled assessment procedure to approximate MAP. In this years evaluation we scored Pre-Specified submissions with both MAP and simulated MInfAP using the reference annotation. Ad-Hoc event references were generated using MInfAP procedures using strata sizes and sampling rates optimized using 2014 Pre-Specified data. Specifically, we define two strata 1-60 with a sampling rate of 100% and 61-200 at 20%. The structure of the strata was determined empirically by generating several strata designs for the MED '14 Pre-Specified submissions using the projected judgment capacity as the limiting factor. The strata design that yielded a higher R^2 correlation coefficient was used. Though initially the strata design was 1-10 sampled at 100%, 11-50 at 60%, and 51-200 at 20%, the judges completed the judgments in half the expected time, so we continued judging additional videos to exhaustively judge to a depth of 60.

Figure 31 shows MInfAP scores for Pre-Specified events. MAP and MInfAP scores were observed to be highly correlated with an R^2 of 0.989, suggesting that MInfAP is indeed a stable metric for follow-on MED evaluations.

For Ad-Hoc, we introduced 10 new events with exemplars defined using the existing training resources. While previous years events have had their event richness controlled, the 10 new Ad-Hoc events created for this years evaluation have not, and as a result were cheaper to create. Only the 10Ex exemplar training condition was supported for Ad-Hoc. Teams processing the Ad-Hoc events were required to run on the MED15EvalFull dataset. Reference generation for these new Ad-Hoc events was done using the aforementioned MInfAP stratified sampling procedure with two strata defined at 1-60 with a sampling rate of 100% and 61-200 at 20%. Results are shown in Figure 32. Figures 33 & 34 show MInfAP scores for the Ad-Hoc events broken down by event and team, respectively.

As with MAP, MInfAP scores are sensitive to the event richness of the test collection as demonstrated in Figures 35 & 36.

For detailed information about the approaches and results, the reader should see the various site reports [TV15Pubs, 2015] and the results pages in the online workshop notebook [TV15Notebook, 2015].

6.6 Summary

In summary, all 16 teams participated in the Pre-Specified (PS), 10 Exemplar (10Ex) test, processing all 20 events, with MAP scores ranging from 0.07675 to 0.2132 (median of 0.1818) for primary systems that ran over MED15-EvalFull, and MAP scores ranging from 0.00515 to 0.2776 (median of 0.1371) for primary systems that ran over MED15-EvalSub (includes MED15-EvalFull systems scored on the MED15-EvalSub subset). 6 teams chose to participate in the Ad-Hoc (AH) event, 10 Exemplar (10Ex) test, processing all 10 events, with MInfAP scores ranging from 0.1619 to 0.4248 (median of 0.3552) for primary systems over the MED15-EvalFull set. 8 of 16 teams chose to process the

MED15-EvalSub set, and most teams built "Small" hardware systems.

Participation in the Ad-Hoc condition was limited this year, as only teams able to process the full search collection (MED15-EvalFull) could participate. This year we limitted testing to the 10 Exemplar (10Ex) condition, and this is the first year we have reported event-specific scores. For this year, no teams participated in the Interactive Event Query test.

The transition from MAP to MInfAP produced correlated results, which will allow us to test with new collections without exhaustive annotation. This is particularly relevant to next years MED evaluation, as a subset of Yahoo!'s YFCC100M set will be used as an additional search set.

For MED '16, it's clear that Ad-Hoc events are a key aspect of MED, and current capabilities indicate that testing on the 10 Exemplar (10Ex) training condition is feasible. Also, Ad-Hoc events are inexpensive to create and dessiminate, and don't rely on constructed data. MED '16 participants can expect 10 new Ad-Hoc events, and a subset of the YFCC100M data set as an additional search set.

7 Interactive Surveillance Event Detection

The 2015 Surveillance Event Detection (SED) evaluation was the eigth 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 et al., 2009] and again in 2009, 2010, 2011, 2012, 2013 and 2014. 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 2015, the evaluation test data used a 9-hour subset (EVAL15) from the total 45 hours available of the test data from the Imagery Library for Intelligent Detection System's (iLIDS)[UKHO-CPNI, 2009] Multiple Camera Tracking Scenario Training (MCTTR) data set collected by the UK Home Office Centre for Applied Science and Technology (CAST) (formerly Home Office Scientific Development Branch's (HOSDB)). This 9-hours is a subset of the 11-hour SED14 Evaluation set that was generated following a crowdsourcing effort in order to generate the reference data. The difference is due to the removal of "camera4" from this set as it had little events of interest (for camera coverage, see Figure 37).

In 2008, NIST collaborated with 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, 2014 and 2015 evaluations.

Those events are:

- CellToEar: Someone puts a cell phone to his/her head or ear
- Embrace: Someone puts one or both arms at least part way around another person
- ObjectPut: Someone drops or puts down an object
- 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
- PersonRuns: Someone runs
- Pointing: Someone points

New for 2015 is a 2-hour "Group Dynamic Subset" subset (SUB15), taken from the 9-hour EVAL15, limited to three specific events: Embrace, PeopleMeet and PeopleSplitUp.

In 2015, 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 multiple 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. 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).

7.1 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.

System performance was assessed on EVAL15 and/or SUB15. 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.

In 2014, event annotation was performed by requesting past participants to run their algorithms against the entire subset of data. A confidence score obtained from the participant's systems was created. A tool developed at NIST was then used to review event candidates. A first level bootstrap data was created out of this process and refined as actual test data evaluation systems from participants were received to generate a second level bootstrap reference which was then used to score the final SED results. The 2015 data uses subsets of this data.

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

7.2 Evaluation

For EVAL15, sites submitted system outputs for the detection of any 3 of 7 possible events (PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, and Pointing). Outputs included the temporal extent as well as a confidence score and detection decision (yes/no) for each event observation. Developers were advised to target a low miss, high false alarm scenario, in order to maximize the number of event observations.

SUB15 followed the same concept, but only requesting 1 out of 3 possible events (Embrace, PeopleMeet and PeopleSplitUp).

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

7.3 Measures

Since detection system performance is a tradeoff between probability of miss vs. rate of false alarms, this task used the Normalized Detection Cost Rate (NDCR) measure for evaluating system performance. NDCR is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per 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.

7.4 Results

SED15 saw 10 sites partipates (see Figure 38), six from China, two from the USA, and one each for Japan, Australia and Greece.

Presented here are the comparative results for PersonRuns (Figure 39) and Embrace (Figure 40) presenting the 2014 and 2015 Retrospective DET curves. For individual participants' results, please refer to the results pages on the TRECVID website.

8 Video Hyperlinking

8.1 System Task

The focus of the Video Hyperlinking task is the automatic generation of video hyperlink *targets* given manually generated *anchors* in source videos. Both targets and anchors are video segments with a start time and an end time (jump-in/out points) and are derived from a substantial collection of videos. The goal of the task is to return a ranked list of target video segments in decreasing likelihood of being *about* an anchor video segment. Target video segments should not be derived from the same video as the anchor video segment. Furthermore, hyperlinking targets may not overlap with previously returned segments for this anchor. Finally, in order to facilitate ground truth annotation, the length of returned target segments is restricted to be between 10 seconds and 120 seconds.

A typical use case of video hyperlinking would be the exploration of large quantities of locally archived or distributed video content via a link structure created at the level of video segments. The video hyperlinking use case is distinguished from other use cases (e.g., recommendation) by its focus on "give me more information *about* an anchor" instead of "give me more segments *similar* to this anchor" [Ordelman et al., 2015].

8.2 Data

The anchors and link targets are taken from a collection of 2,686 hours of English language broadcast TV video from the BBC, including human generated textual metadata and available subtitles. The video collection for both training and test anchor sets consists of content originally broadcast between 12.05.2008 and 31.07.2008. The average length of each video is roughly 30 minutes.

Along with the video data and metadata provided by the BBC, the output of several content analysis methods were provided to the participants. Based on the audio channel, automatic speech recognition, speaker diarization[1], and prosodic feature extraction[2] [Eyben et al., 2013] were calculated. The computer vision groups at University of Leuven and University of Oxford provided the output of concept detectors for 1,537 concepts from ImageNet using different training approaches [Tommasi et al., 2014, Chatfield and Zisserman, 2013].

8.3 Topics

In the video hyperlinking task, the search topics used in standard video retrieval tasks were replaced with video segments that represent anchors of video hyperlinks. We define an anchor to be the triple of: video (v), start time (s) and end time (e). Anchors were selected manually by media professionals from the BBC and the Netherlands Institute for Sound and Vision, students majoring in journalism, and task organisers, as we expected these groups to be the most capable of understanding the novel concept of video hyperlinking. They were instructed about the anchor generation task and provided with an example "second screen" type of scenario that put the participant in a role of a producer of a video programme that s/he wants to enrich with video hyperlinks. They were also provided with guidelines previously used for wikification: hyperlinks may help to understand the anchor better, hyperlinks may contain relevant information about the anchor given what you are currently looking for, hyperlinks may contain information about occurring objects, places, people and events that appear in the video.

In total, the media professionals generated 135 anchors. From those anchors, we selected a subset of 100 anchors for evaluation when those were correctly defined in terms of start and end time, description of potential hyperlinks that we could use at the evaluation stage, and when those anchors were targeting video segments with content available within our collection. On average, the selected anchors were 71 seconds long.

[1] Speech recognition and speaker diarization were created by the LIMSI-CNRS/Vocapia VoxSigma system, the LIUM CMU Sphinx based system, and NST/Sheffield system

[2] Prosodic features were extracted using the OpenSMILE tool version 2.0

8.4 Evaluation, Measures

The ground truth was generated by pooling the top 10 results of participants runs and asking MTurk [3] workers to annotate pairs of an anchor and a result target segment as relevant or non-relevant to the anchor. We provided workers with both anchor and target video segments, as well as a textual description of potential hyperlink target video segments that the users had in mind when defining anchors manually. The AMT workers were asked to support their decision on relevance with a textual explanation of their decision, and with the choice of predefined options, e.g. 'Video 2 fits given description', 'Video 2 is connected to Video 1', 'Same location', 'Same objects', 'Same persons', 'Same topic being discussed', 'Other'. This metadata of the workers decisions was provided to the task participants for follow up analysis of their submissions. A small subset of about 200-300 out of a total of 19742 assessments were manually checked by the task organisers in order to confirm whether AMT workers had understood the task correctly. AMT workers submissions were accepted automatically, when all the required decision metadata fields had been filled in, and the answer to the test questions were correct. The evaluation metrics used were standard Mean Average Precision (MAP) and an adaptation of MAP called Mean Average interpolated Segment Precision (MAiSP) which is based on previously proposed adaptations of MAP for this task [Liu and Oard, 2006, Kamps et al., 2006, Eskevich et al., 2012]. For MAP computation, we assume that a result segment is relevant if it overlaps with a segment that was judged relevant (see also [Aly et al., 2013]).

Standard MAP assumes that the cost of finding relevant information within a suggested relevant segment is negligible. By contrast, MAiSP takes into account the number of relevant seconds that can be watched from the start of the segment to compute the precision with which relevant content has been retrieved and reflect expected user effort to find and extract the relevant information (see [Racca and Jones, 2015] for an extensive exploration of adapted MAP metrics for this type of tasks). In MAiSP, user effort is measured as the number of seconds of video that the user watches, and user satisfaction as the number of seconds of new relevant content that the user can watch starting from the start time of the segment. The user is assumed to stop watching at the segment's end if no relevant content continuous after this point, or at the end of the continuing relevant content otherwise. Precision is then calculated as the ratio of relevant seconds watched to total seconds watched and recall as the ratio of relevant seconds watched to total seconds that are known relevant in the collection. As systems may return segments of varying length, precision is computed at 101 fixed-recall points rather than at fixed-positions in the ranks.

[3] http://www.mturk.com

8.5 Results

For more detailed information about the approaches and results, the reader should consult the various site reports [TV15Pubs, 2015] and the results pages in the online workshop notebook [TV15Notebook, 2015].

In general, the results of the participants show that runs perform similarly in terms of MAP (Figure 42) and MAiSP (Figure 43). Using the MAP measure we see the runs from CMU/SMU having the strongest performance and also we see that results quickly decline. Using the MAiSP measure, the decline is less extreme and the the team of DCU has a slightly stronger performance than CMU/SMU. For the best submitted runs the correlation between the two measures is low (Figure 44). Possible reasons for this is that some runs were over fitted to a particular measure. Besides mean performance over all anchors, we also looked at the performance of particular anchors over all runs. To study how well accurate targets can be provided in the ideal case, we study the distribution of the maximum performance for an anchor over all runs (Figure 45). Roughly one third of the anchors have a maximum performance lower than 0.3 MAiSP.

9 Appendix A: Semantic Indexing Concepts

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

10 Appendix B: Instance search topics

9129 OBJECT - "this silver necklace"

9130 OBJECT - "a chrome napkin holder"

9131 OBJECT - "a green and white iron"

9132 OBJECT - "this brass piano lamp with green shade"

9133 OBJECT - "this lava lamp"

9134 OBJECT - "this cylindrical spice rack"

9135 OBJECT - "this turquoise stroller"

9136 OBJECT - "this yellow VW beetle with roofrack"

9137 OBJECT - "a Ford script logo"

9138 PERSON - "this man with moustache"

9139 OBJECT - "this shaggy dog (Genghis)"

9140 OBJECT - "a Walford Gazette banner"

9141 OBJECT - "this guinea pig"

9142 OBJECT - "this chihuahua (Prince)"

9143 PERSON - "this bald man"

9144 OBJECT - "this doorknocker on #27"

9145 OBJECT - "this jukebox wall unit"

9146 OBJECT - "this change machine"

9147 OBJECT - "this table lamp with crooked body"

9148 OBJECT - "this cash register (at the cafe)"

9149 LOCATION - "this Walford Community Center entrance from street"

9150 OBJECT - "this IMPULSE game"

9151 LOCATION - "this Walford Police Station entrance from street"

9152 OBJECT - "this PIZZA game"

9153 OBJECT - "this starburst wall clock"

9154 OBJECT - "this neon Kathy's sign"

9155 OBJECT - "this dart board"

9156 OBJECT - "a 'DEVLIN' lager logo"

9157 OBJECT - "this picture of flowers"

9158 OBJECT - "this flat wire 'vase with flowers"'

References

[Aly et al., 2013] Aly, R., Eskevich, M., Ordelman, R., and Jones, G. J. F. (2013). Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval Tasks. *CoRR*, abs/1312.1913.

[Chatfield and Zisserman, 2013] Chatfield, K. and Zisserman, A. (2013). Visor: Towards on-the-fly large-scale object category retrieval. In *Computer Vision–ACCV 2012*, pages 432–446. Springer.

[Eskevich et al., 2012] Eskevich, M., Magdy, W., and Jones, G. J. F. (2012). New Metrics for Meaningful Evaluation of Informally Structured Speech Retrieval. In *Proceedings of ECIR 2012*, pages 170–181, Barcelona, Spain.

[Eyben et al., 2013] Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of ACM Multimedia 2013*, pages 835–838, Barcelona, Spain. ACM.

[Kamps et al., 2006] Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., and Robertson, S. (2006). INEX 2007 evaluation measures. In *Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) Focused access to XML Documents*, pages 24–33. Springer.

[Liu and Oard, 2006] Liu, B. and Oard, D. W. (2006). One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 673–674. ACM.

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

[Ordelman et al., 2015] Ordelman, R. J., Eskevich, M., Aly, R., Huet, B., and Jones, G. J. (2015). Defining and evaluating video hyperlinking for navigating multimedia archives. In *LIME 2015, 3rd International Workshop on Linked Media, co-located with WWW 2015, 18 May 2015, Florence, Italy*, Florence, ITALIE.

[Over et al., 2006] Over, P., Ianeva, T., Kraaij, W., and Smeaton, A. F. (2006). TRECVID 2006 Overview. `www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6overview.pdf`.

[Racca and Jones, 2015] Racca, D. N. and Jones, G. J. F. (2015). Evaluating Search and Hyperlinking: An Example of the Design, Test, Refine Cycle for Metric Development. In *Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany.

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

[Strassel et al., 2012] Strassel, S., Morris, A., Fiscus, J., Caruso, C., Lee, H., Over, P., Fiumara, J., Shaw, B., Antonishek, B., and Michel, M. (2012). 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).

[Tommasi et al., 2014] Tommasi, T., Tuytelaars, T., and Caputo, B. (2014). A Testbed for Cross-Dataset Analysis. *CoRR*, abs/1402.5923.

[TV15Notebook, 2015] TV15Notebook (2015). Trecvid 2015 workshop notebook. `http://www-nlpir.nist.gov/projects/tv2015/active/tv15.workshop.notebook`.

[TV15Pubs, 2015] TV15Pubs (2015). Trecvid 2015 workshop proceeding. `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.15.org.html`.

[UKHO-CPNI, 2009] 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 and Aslam, 2006] Yilmaz, E. and Aslam, J. A. (2006). Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management (CIKM)*, Arlington, VA, USA.

[Yilmaz et al., 2008] Yilmaz, E., Kanoulas, E., and 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, pages 603–610, New York, NY, USA. ACM.

Figure 1: SIN: Histogram of shot frequencies by concept number

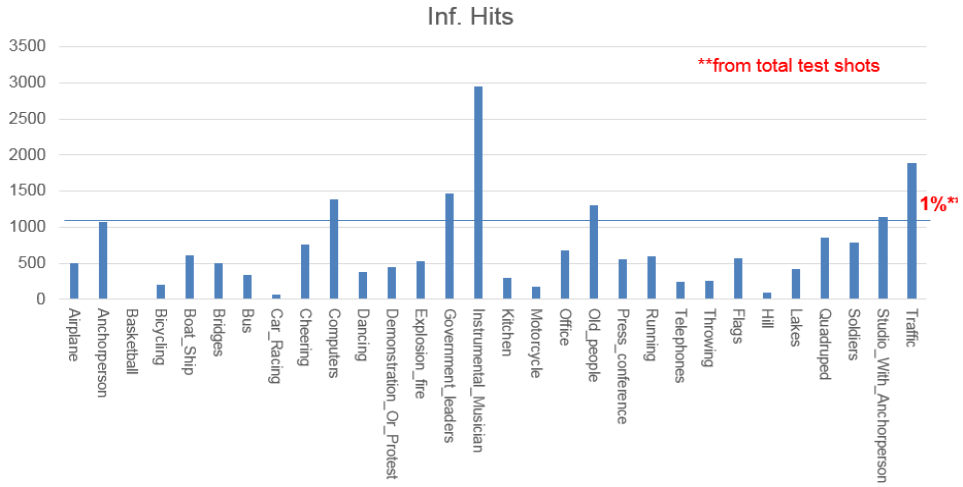


Figure 2: SIN: xinfAP by run - 2015 submissions

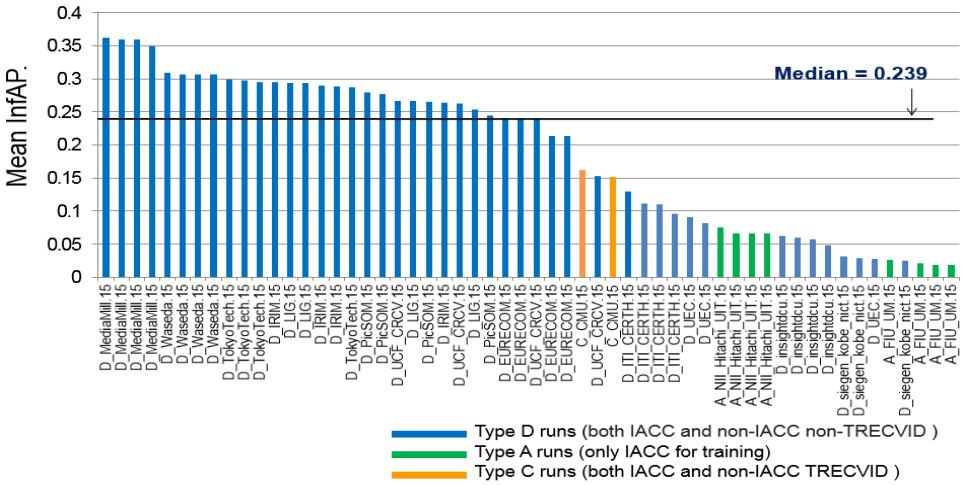


Figure 3: SIN: xinfAP by run - 2015 submissions including Progress runs

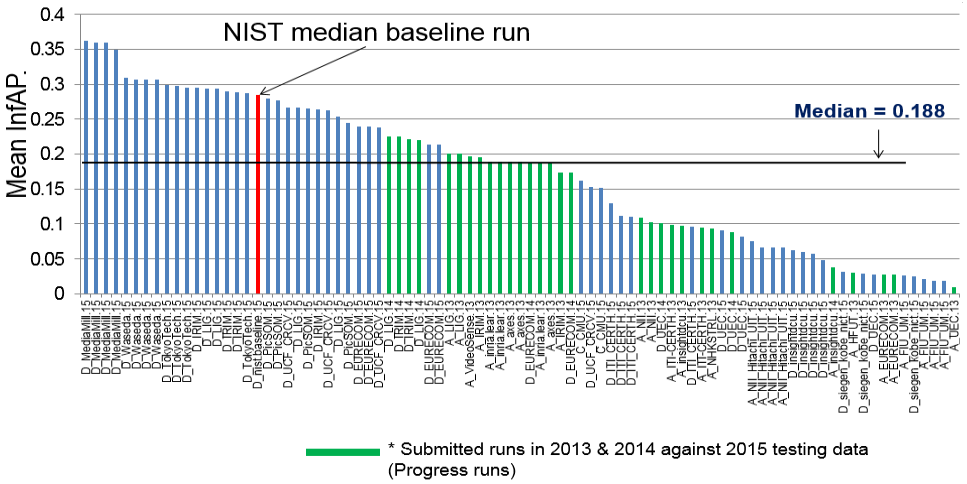


Figure 4: SIN: top 10 runs (xinfAP) by concept number

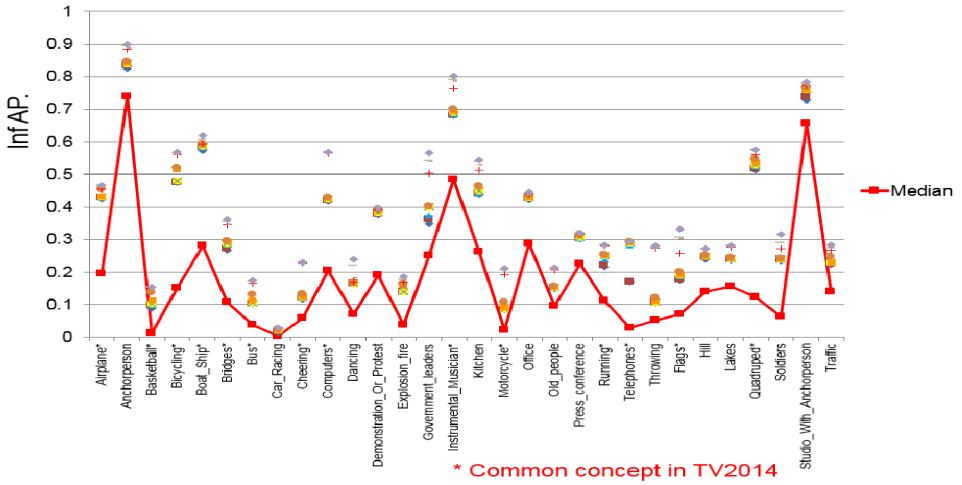


Figure 5: SIN: top 10 main runs

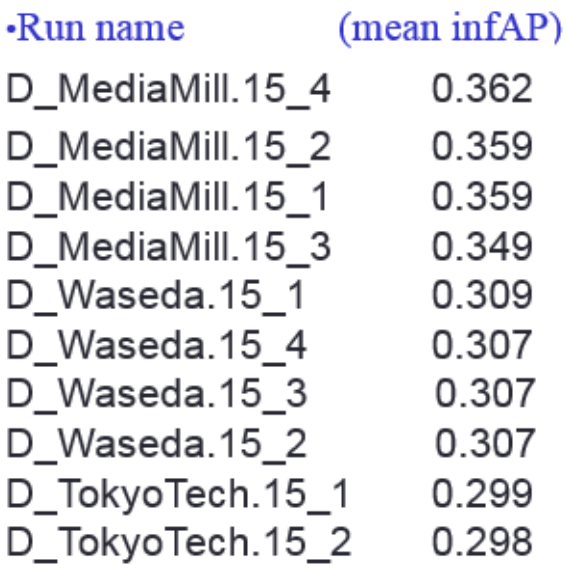


Figure 6: SIN: Significant differences among top 10 main runs

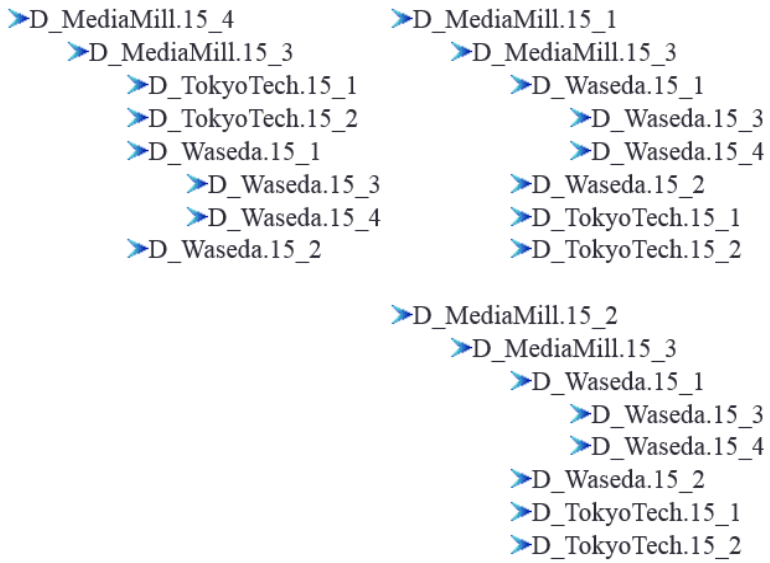


Figure 7: SIN: Confusion analysis across concepts

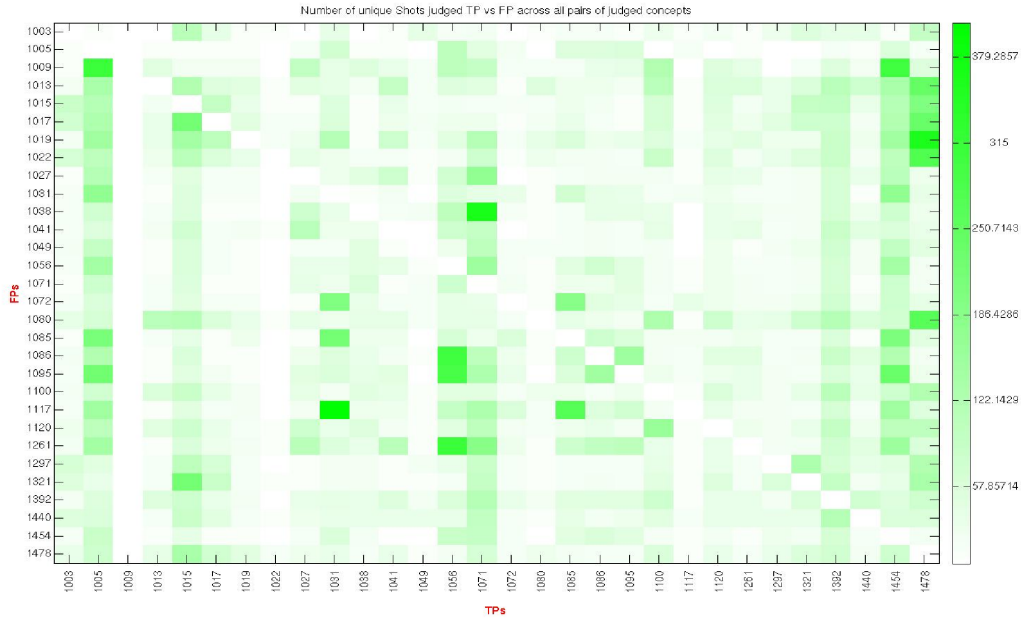


Figure 8: SIN: Progress subtask - Comparing best runs in 2013, 2014 and 2015 by team

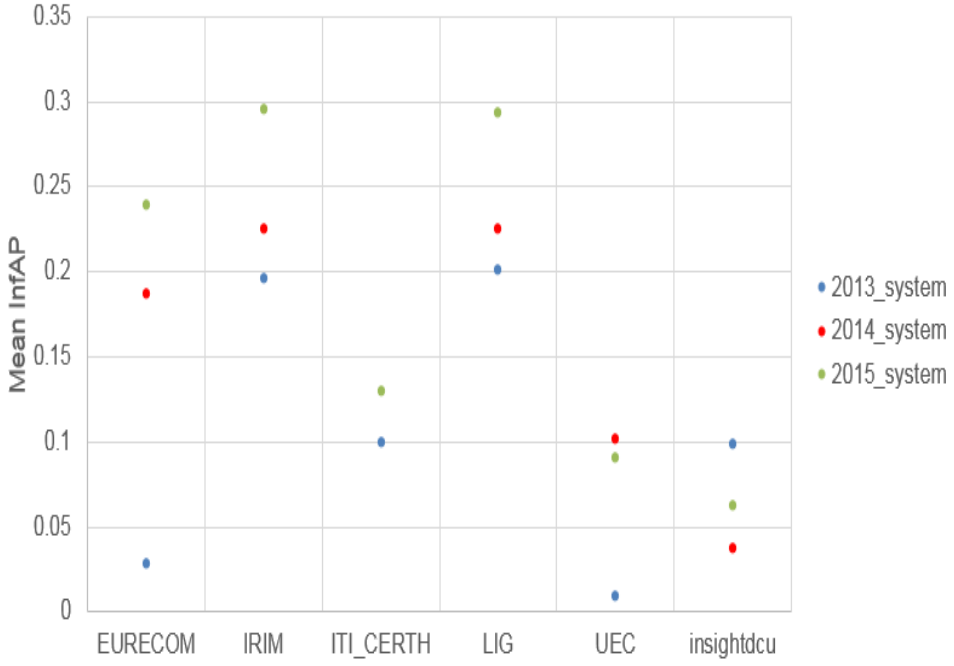
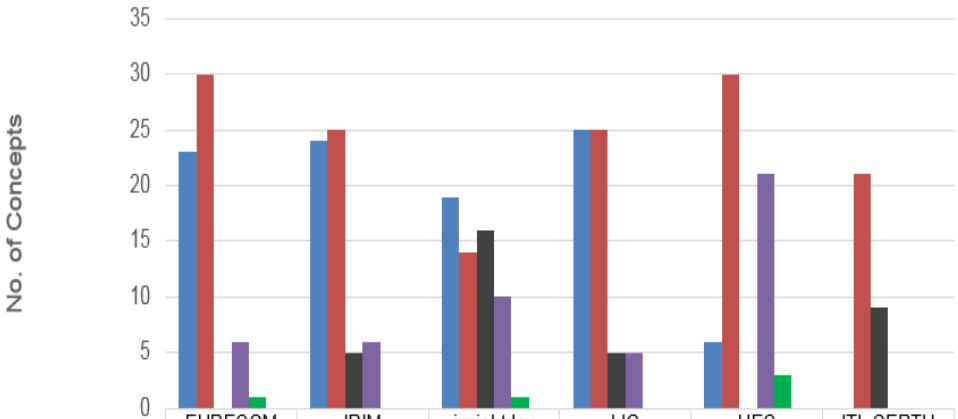


Figure 9: SIN: Progress subtask - Concepts improved vs weakened by team



	EURECOM	IRIM	insightdcu	LIG	UEC	ITI_CERTH
better than 2014	23	24	19	25	6	
better than 2013	30	25	14	25	30	21
worse than 2013	0	5	16	5	0	9
worse than 2014	6	6	10	5	21	
same as 2014	1	0	1	0	3	
same as 2013	0	0	0	0	0	0

Figure 10: Concept Localization Evaluation Framework

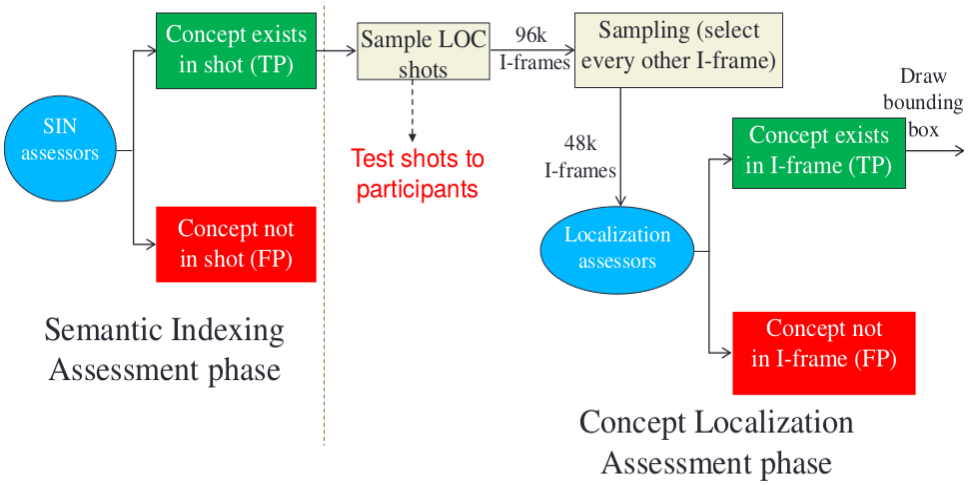


Figure 11: LOC: Temporal localization results by run

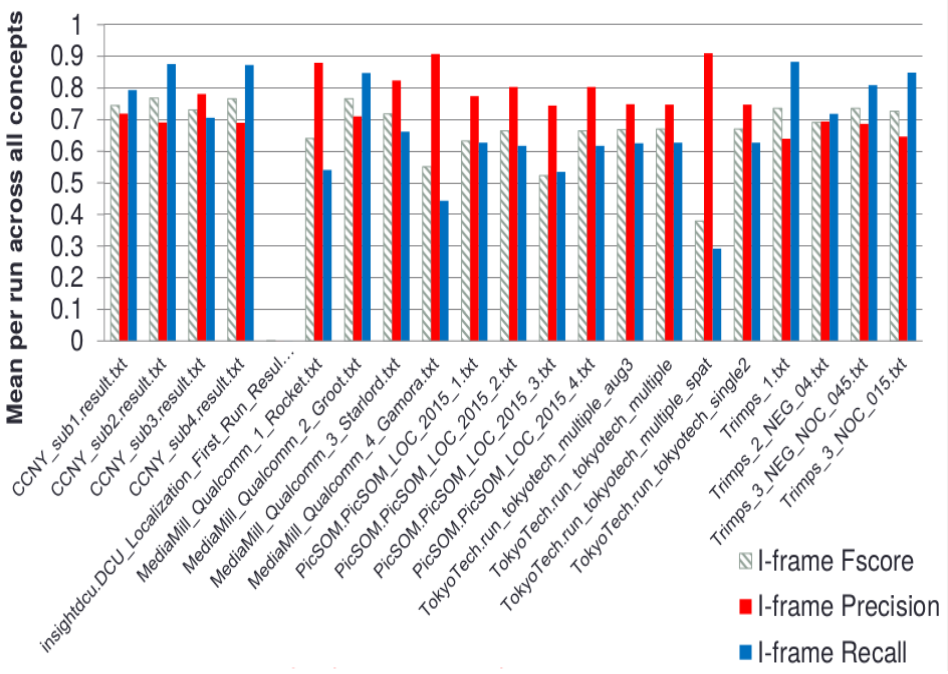


Figure 12: LOC: Spatial localization results by run

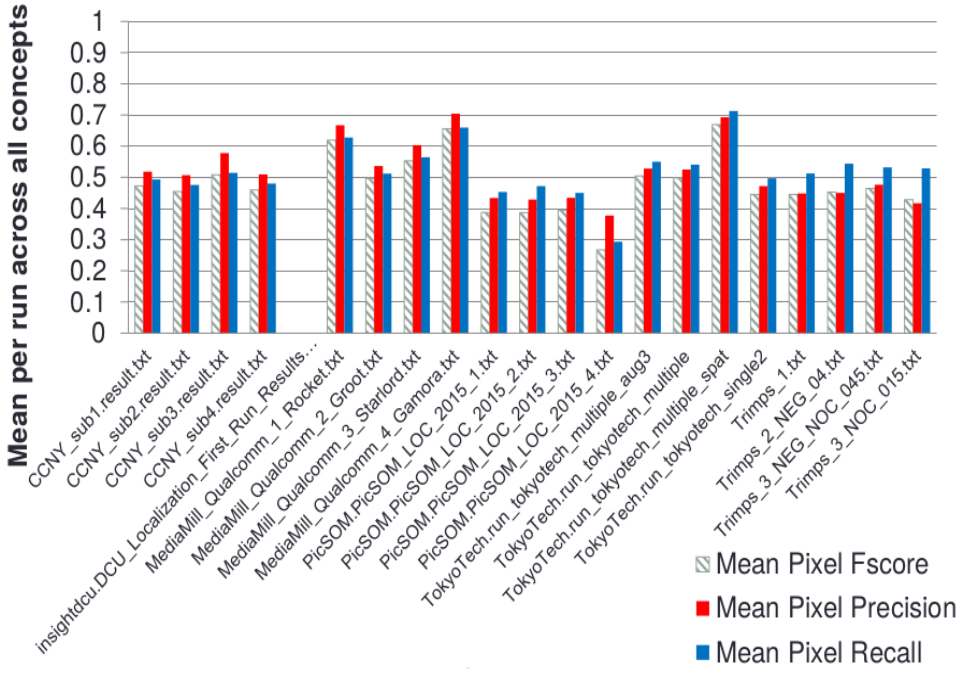


Figure 13: LOC: Temporal localization by concept

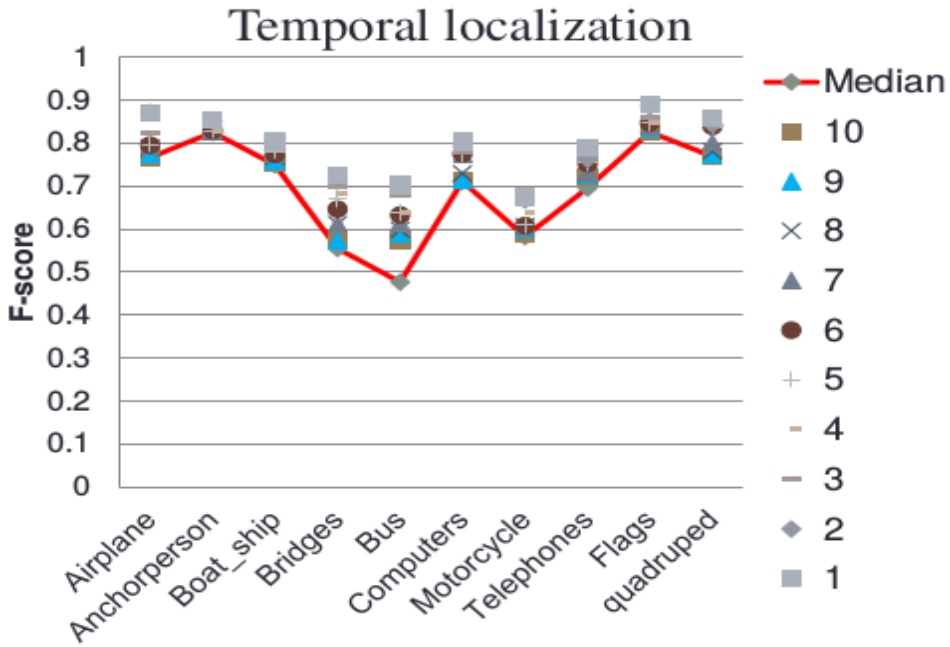


Figure 14: LOC: Spatial localization by concept

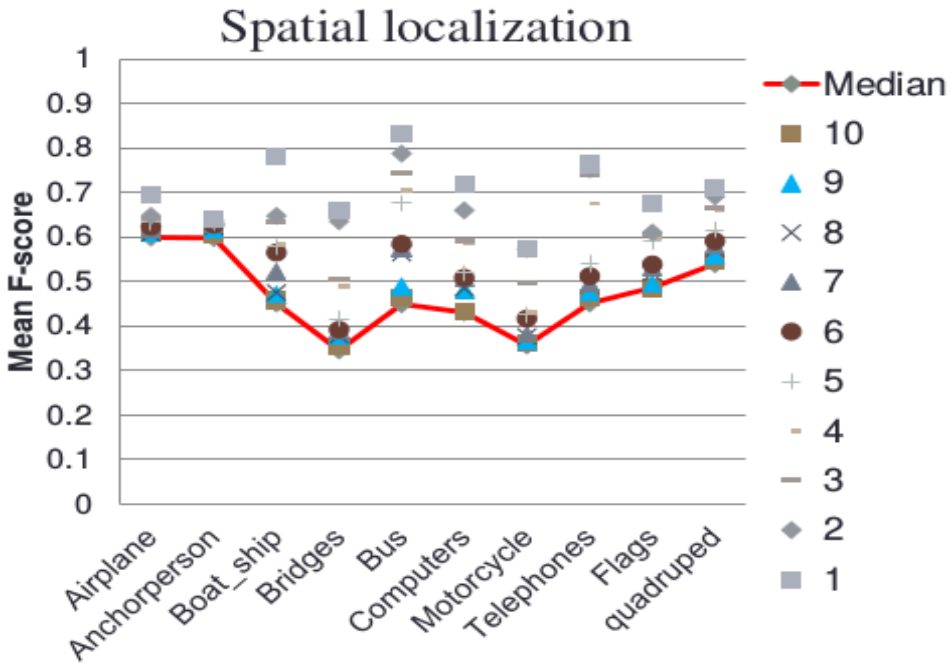


Figure 15: LOC: temporal precision and recall per concept for all teams

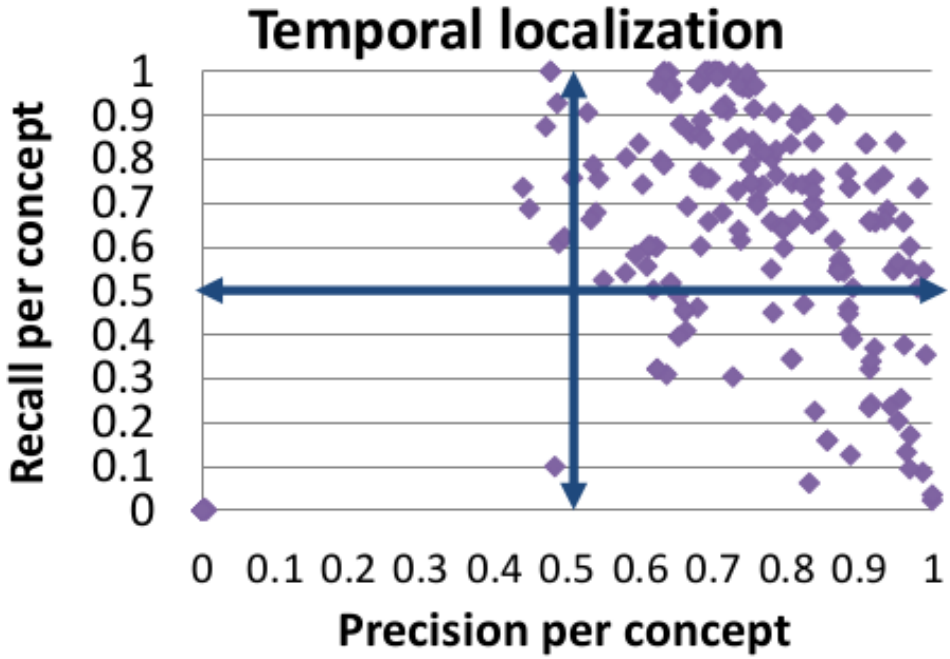


Figure 16: LOC: spatial precision and recall per concept for all teams

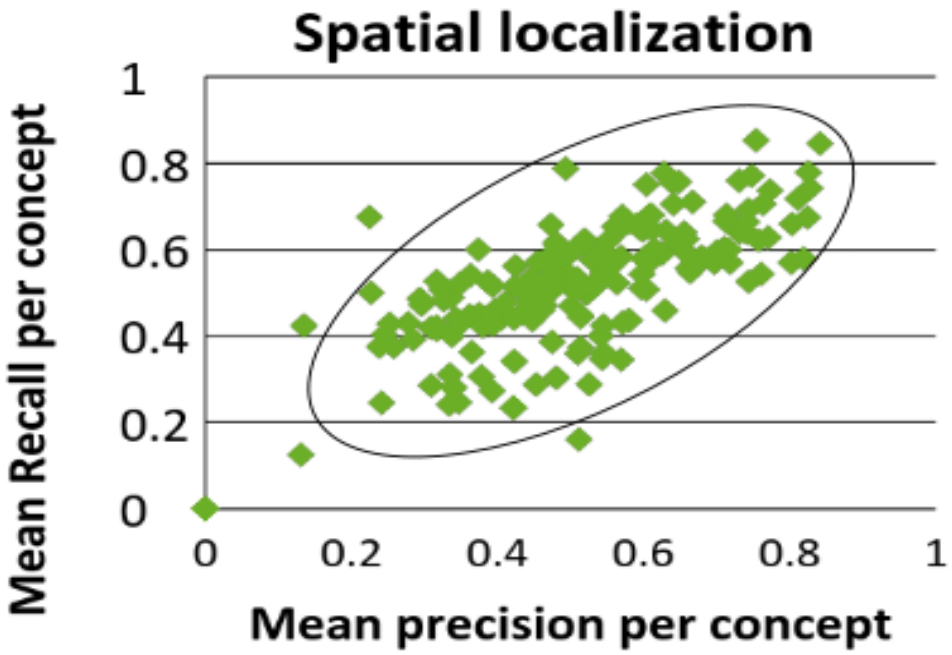


Figure 17: LOC: Samples of good spatial localization

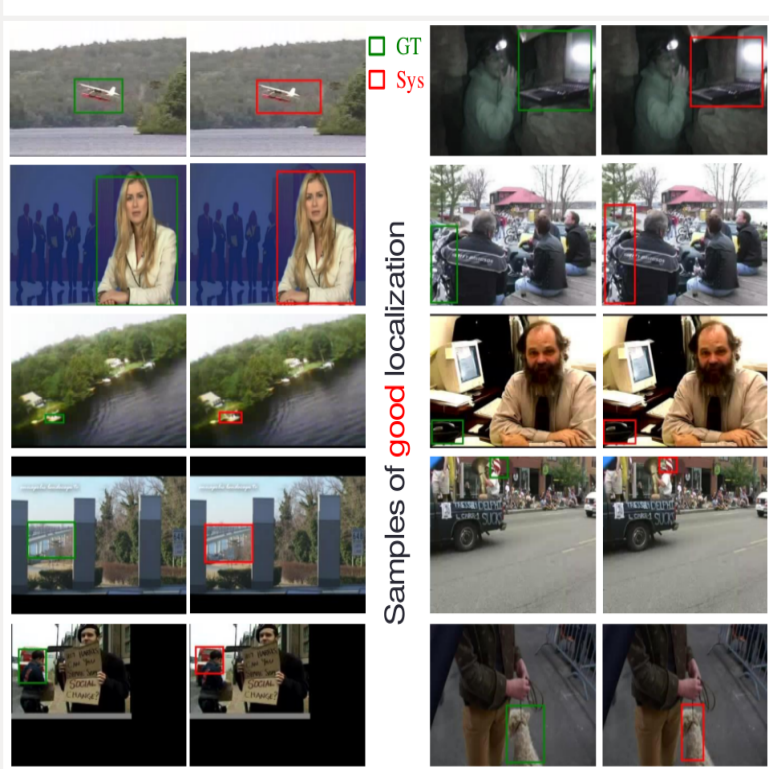


Figure 18: LOC: Samples of weaker spatial localization

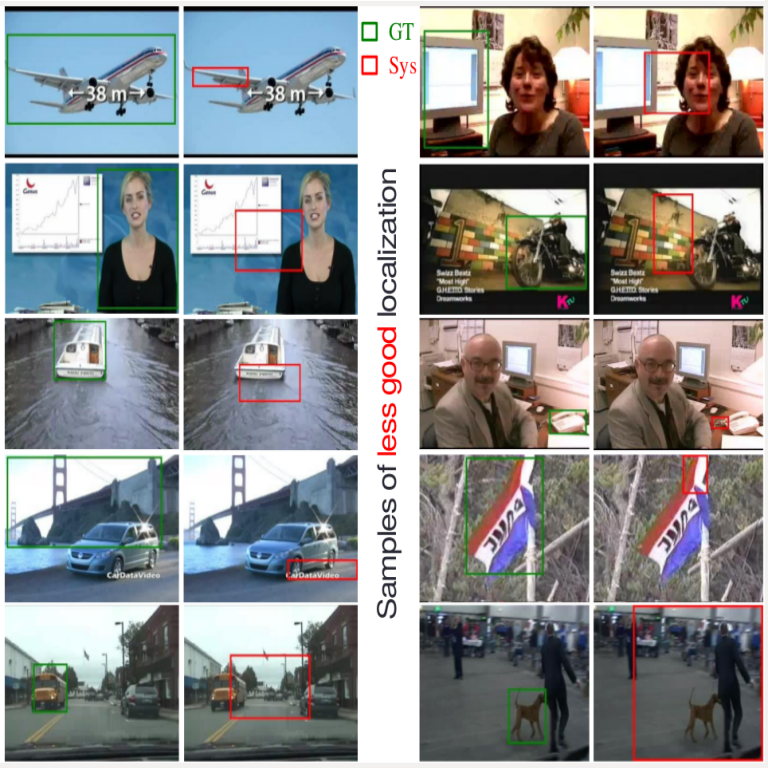


Figure 19: INS: Boxplot of average precision by topic for automatic runs

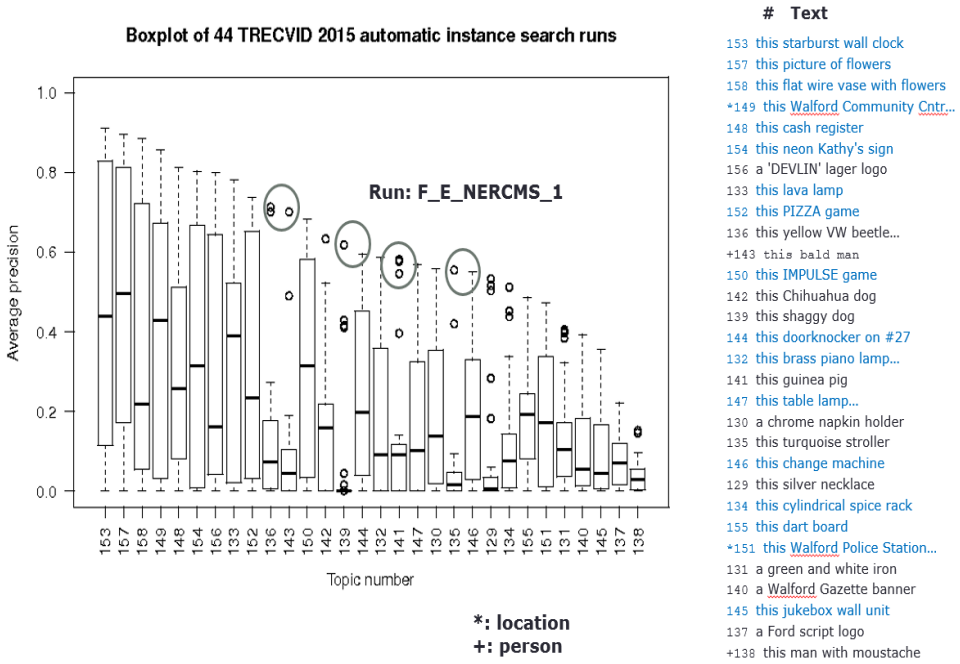


Figure 20: INS: Randomization test results for top automatic runs

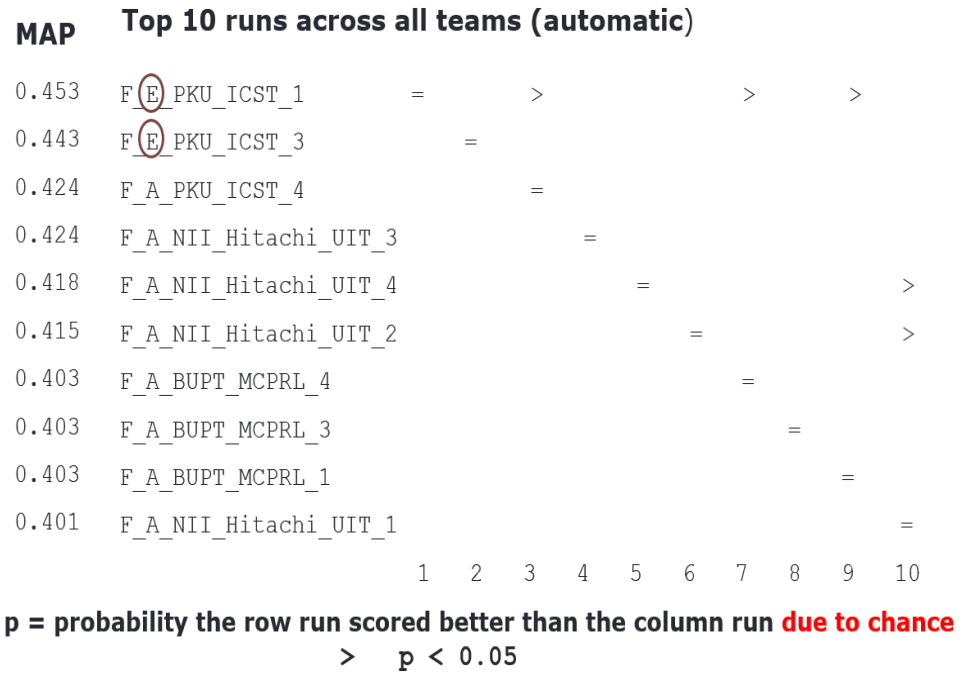


Figure 21: INS: Boxplot of average precision by topic for interactive runs

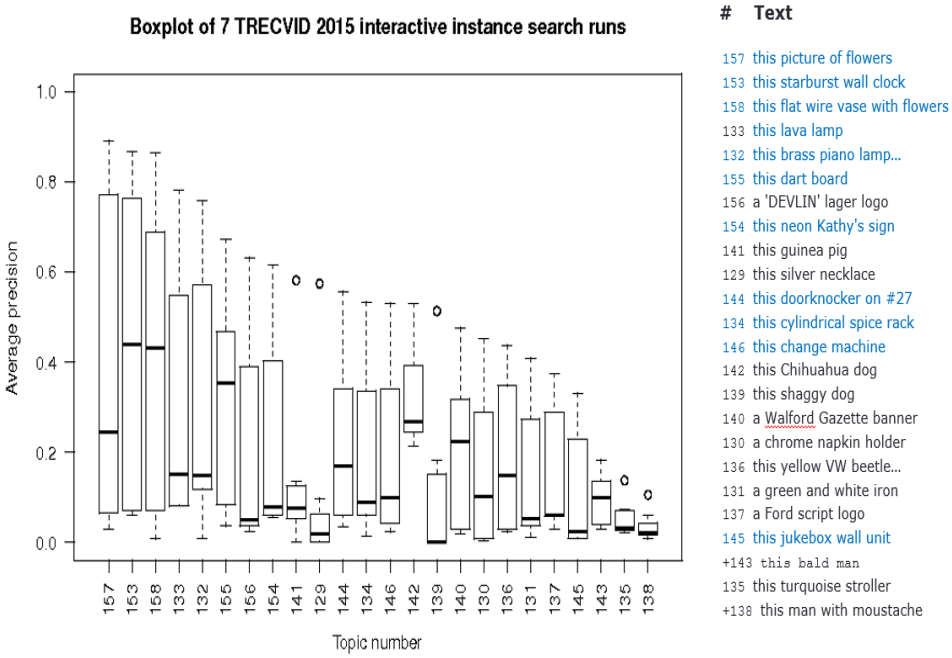


Figure 22: INS: Randomization test results for top interactive runs

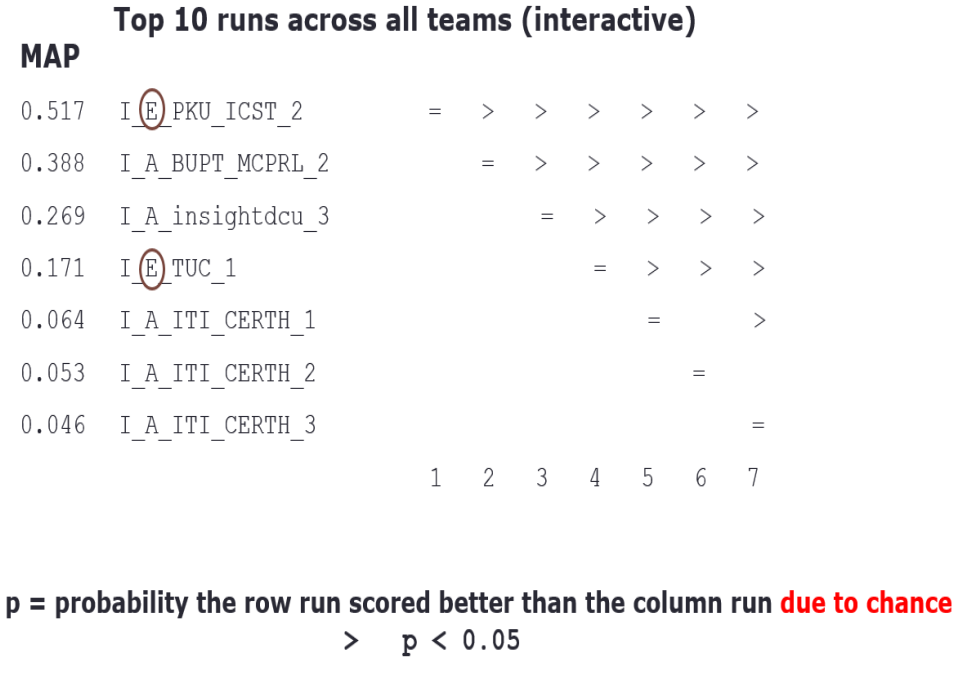


Figure 23: INS: Mean average precision versus time for fastest runs

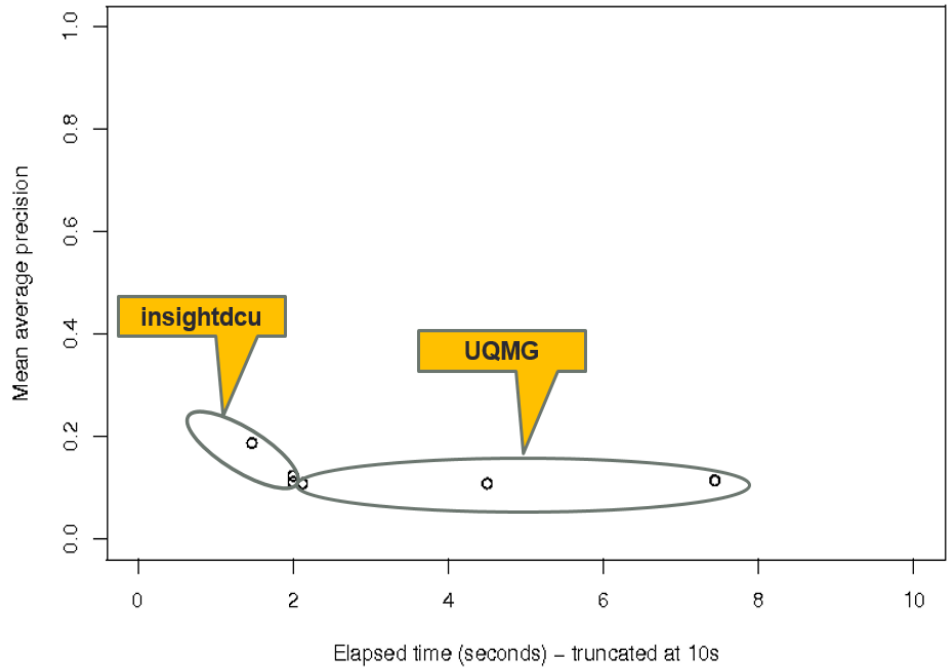


Figure 24: INS: Number of true positives versus average precision

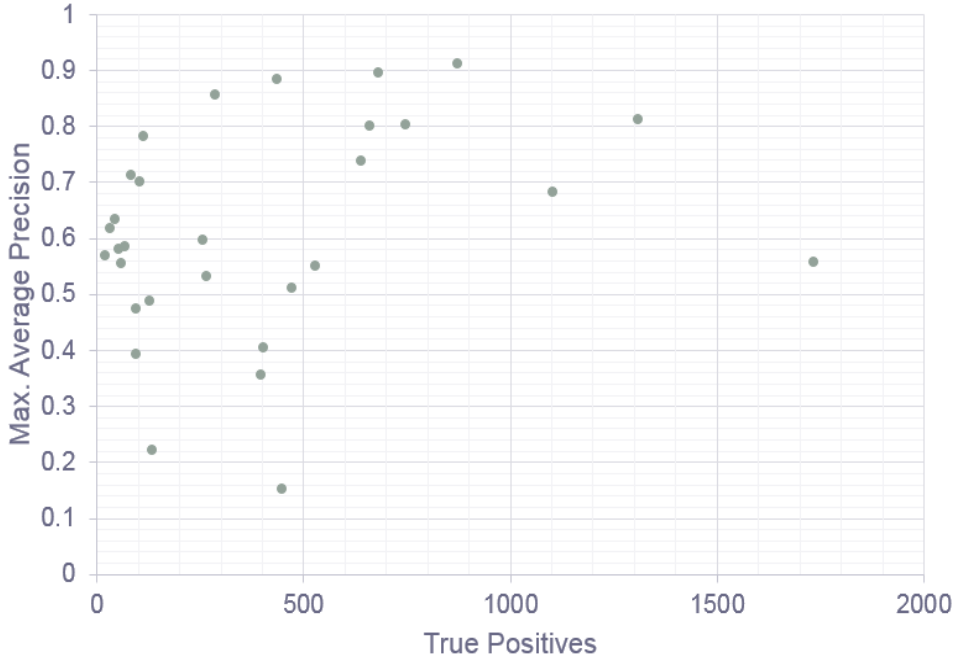


Figure 25: INS: Effect of number of topic example images used

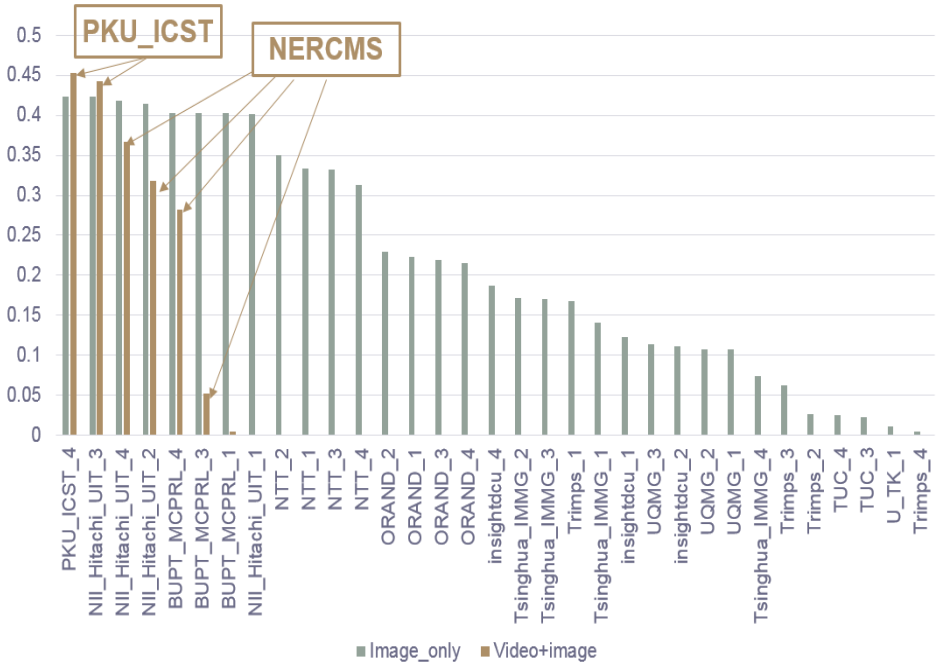


Figure 26: MED: Pre-Specified MAP scores per team for MED15-EvalFull

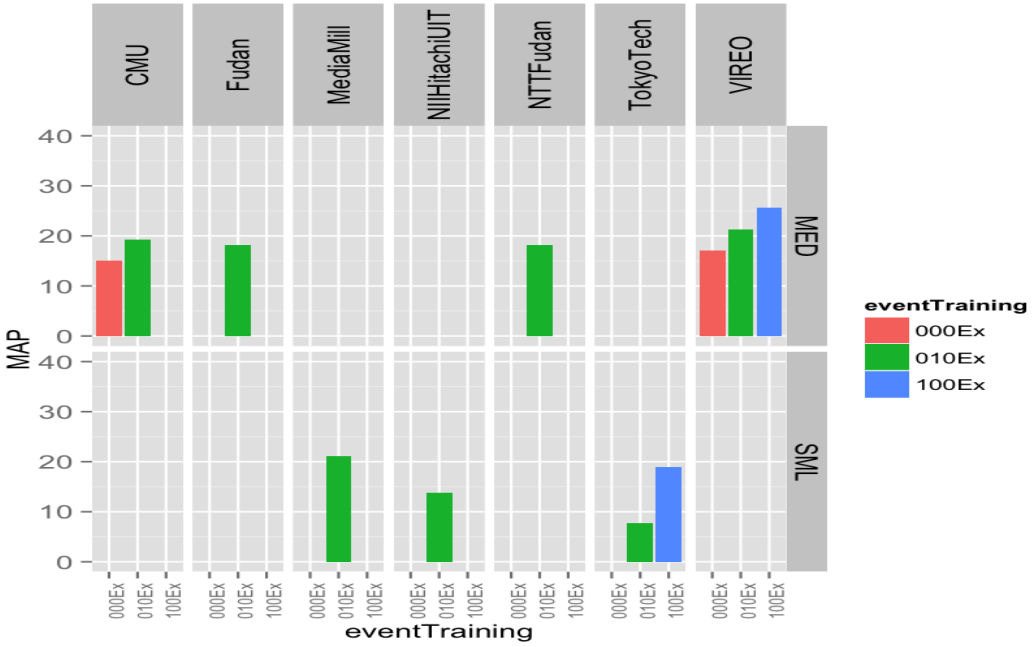


Figure 27: MED: Pre-Specified MAP scores per team for MED15-EvalSub

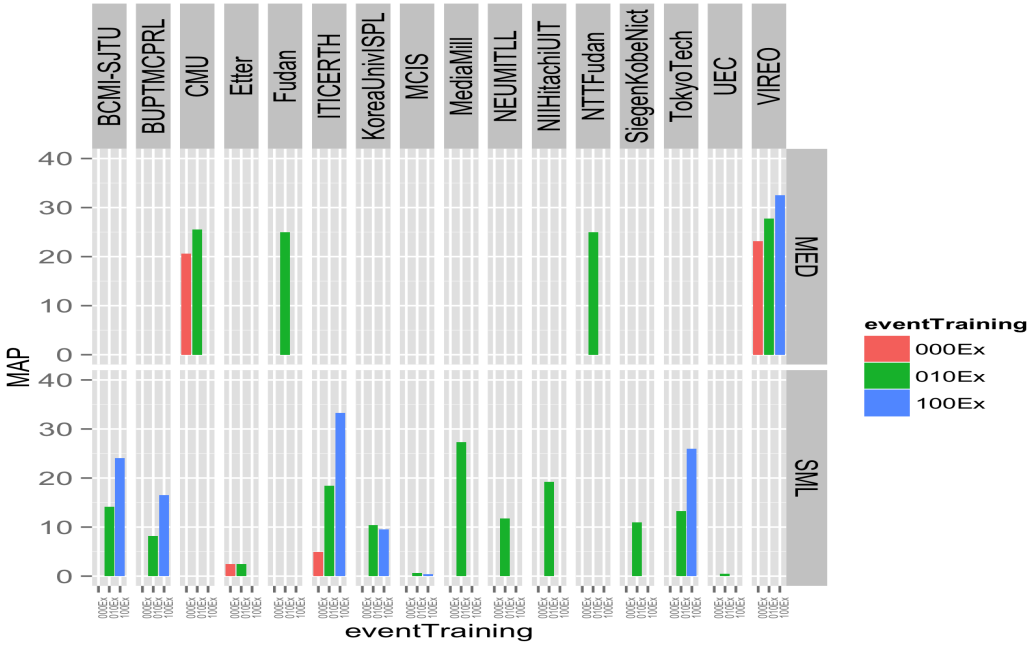


Figure 28: MED: Events vs Systems (MED15-EvalSub 10Ex exemplar)

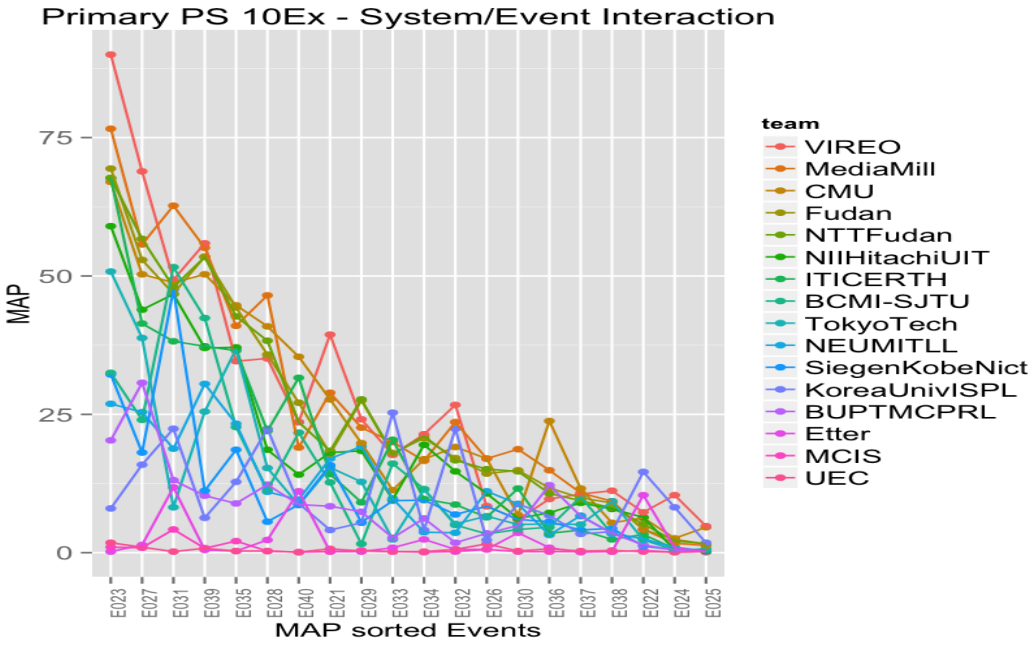


Figure 29: MED: Systems vs Events (MED15-EvalSub 10Ex exemplar)

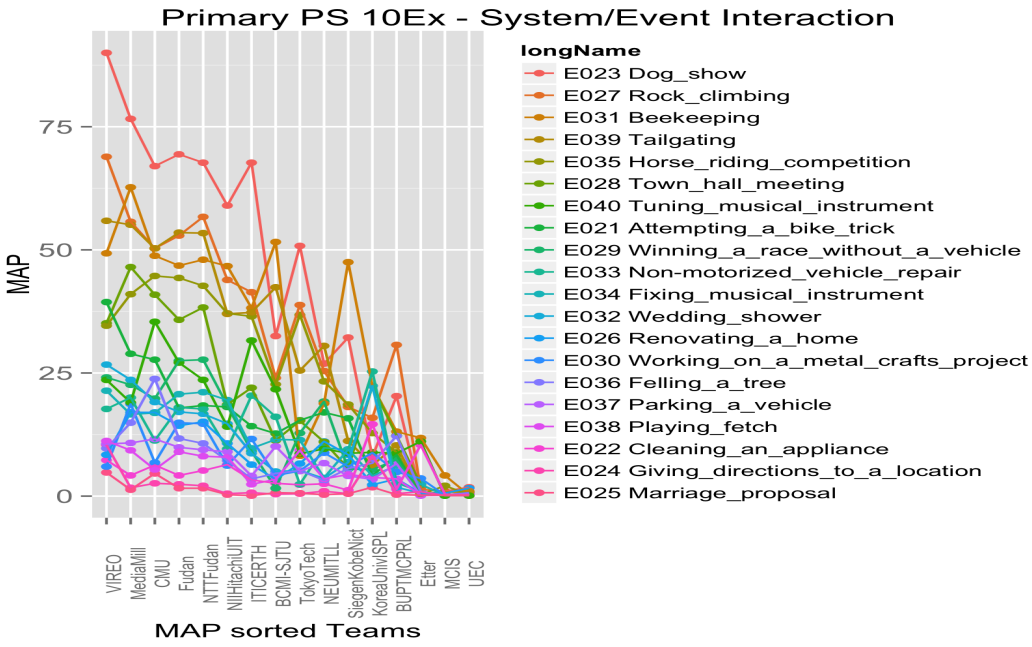


Figure 30: MED: Event effect on systems performance grouped by hardware size

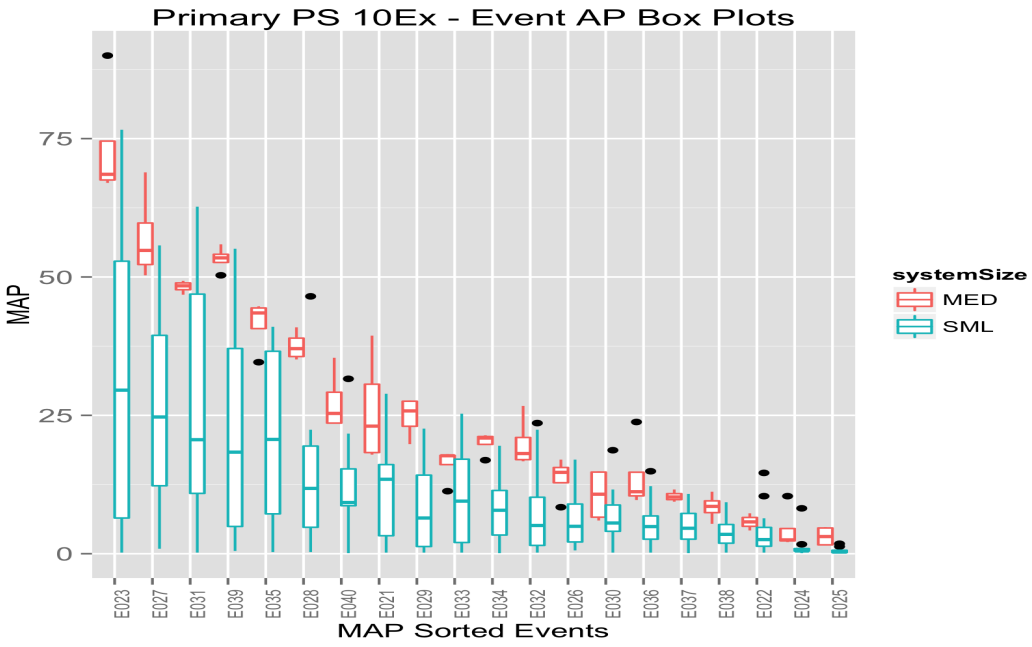


Figure 31: MED: InfMAP scores for Pre-Specified events

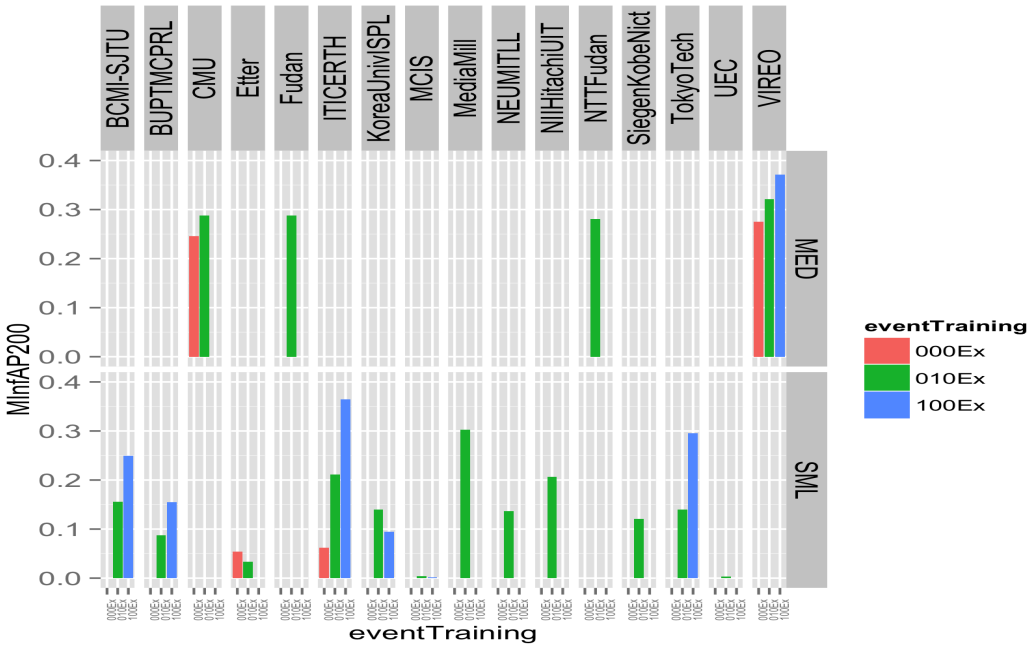


Figure 32: MED: MInfAP scores for Ad-Hoc events

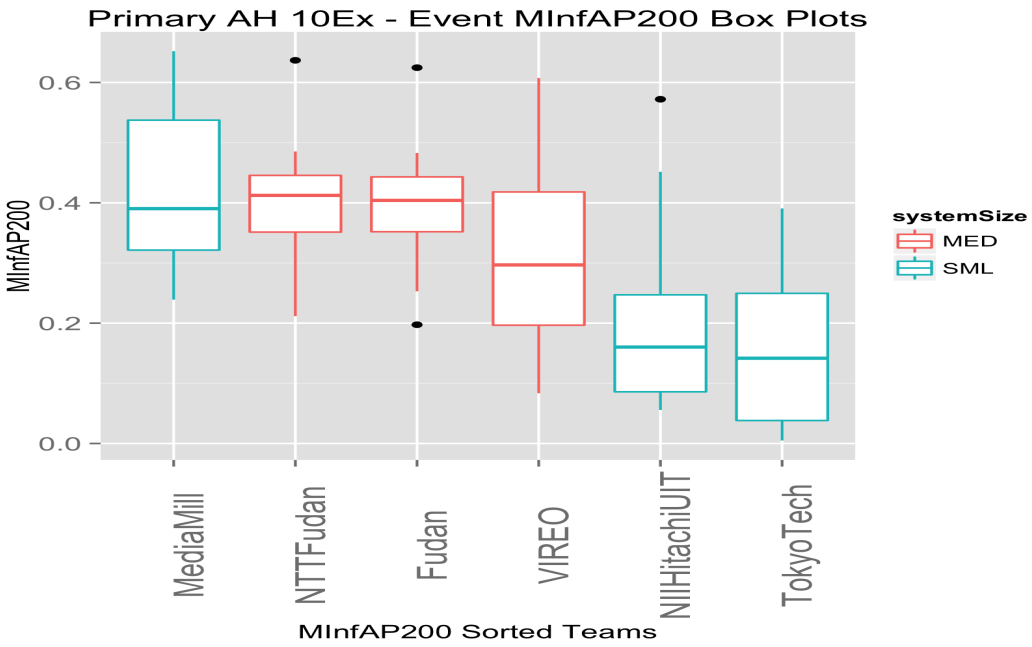


Figure 33: MED: Events vs Systems (MED15-EvalFull 10Ex exemplar) Ad-Hoc events

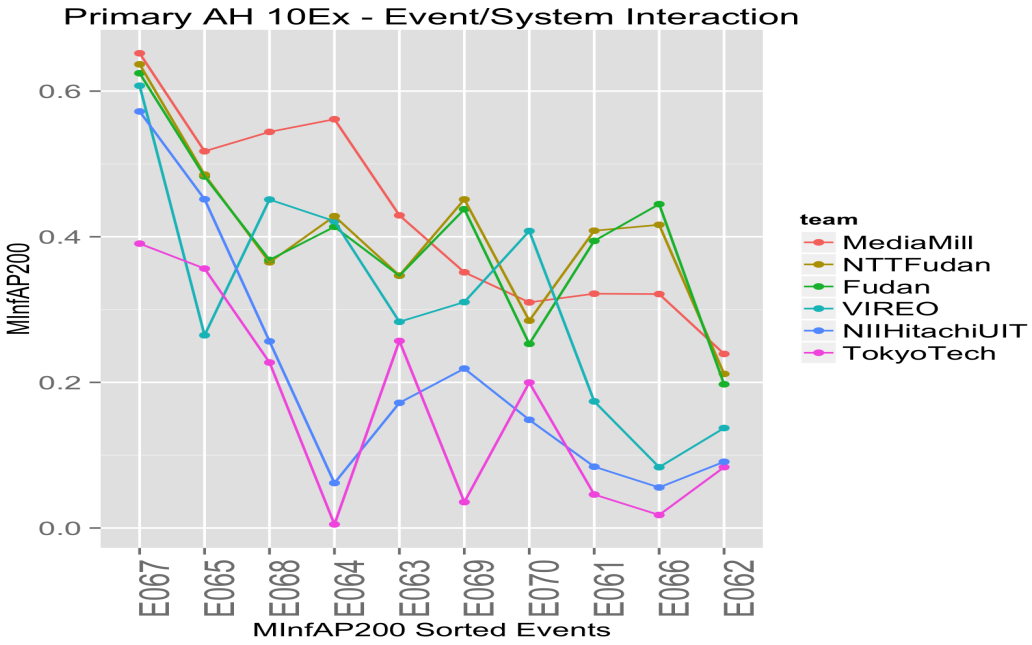


Figure 34: MED: Systems vs Events (MED15-EvalFull 10Ex exemplar) Ad-Hoc events

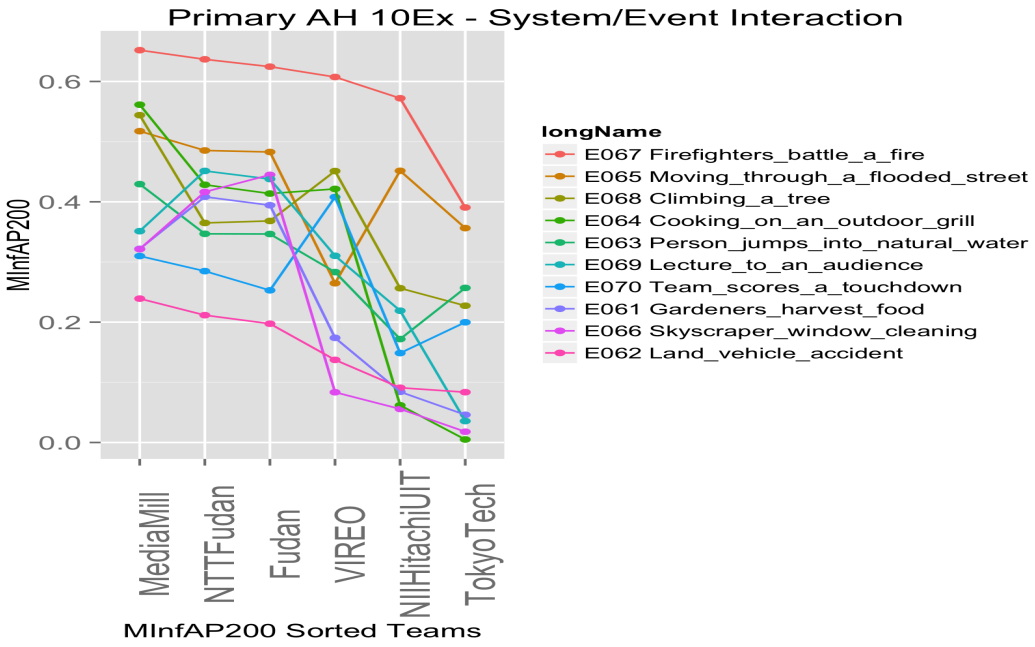


Figure 35: MED: Event Richness

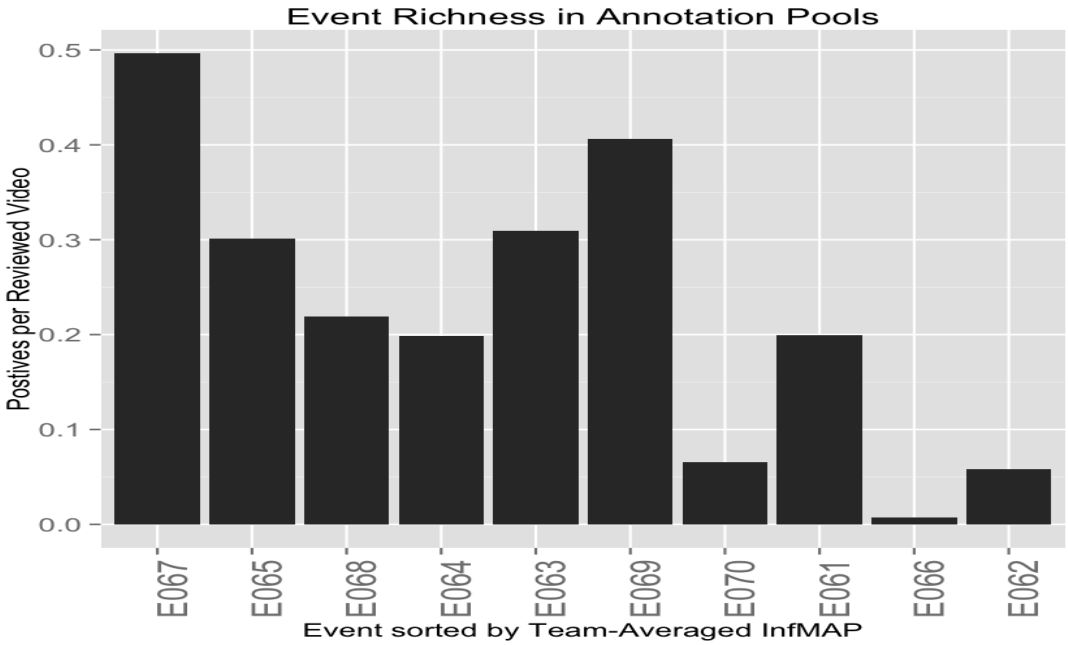
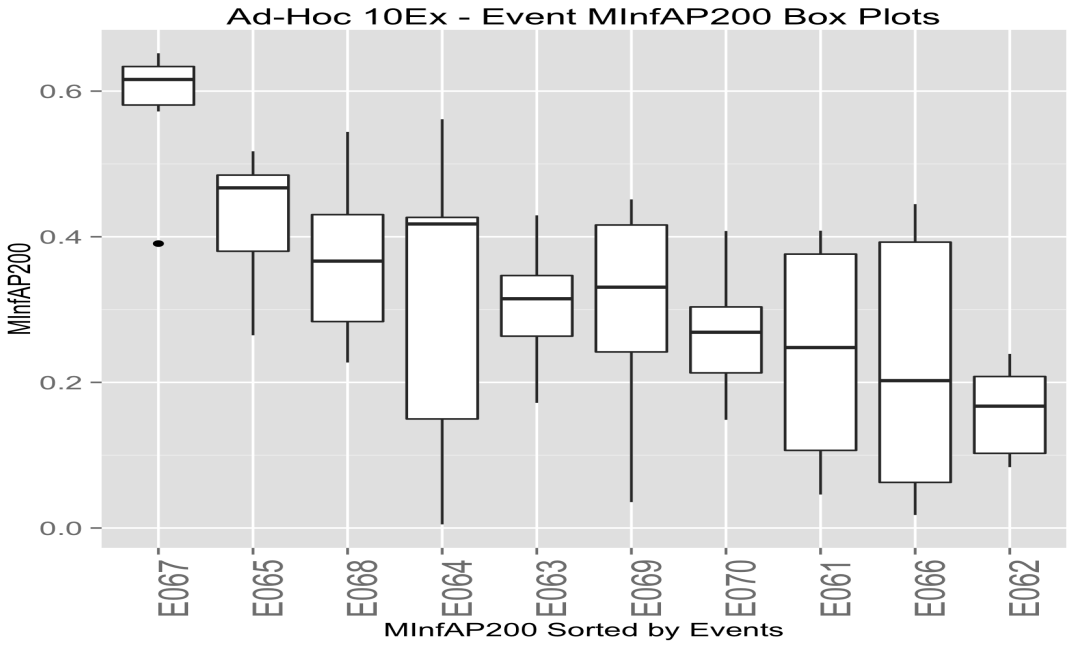


Figure 36: MED: InfAP vs Events



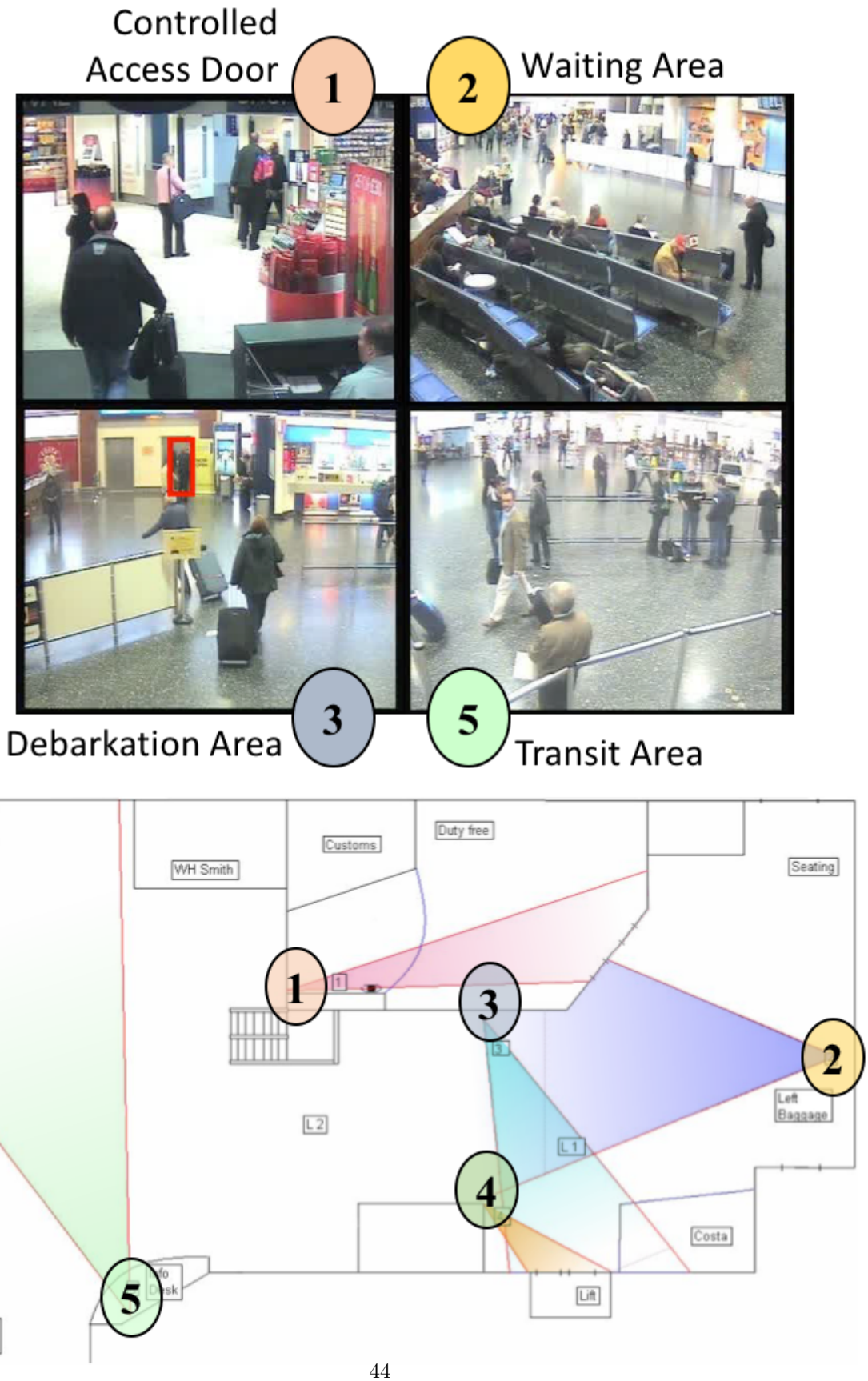


Figure 37: Camera views and coverage

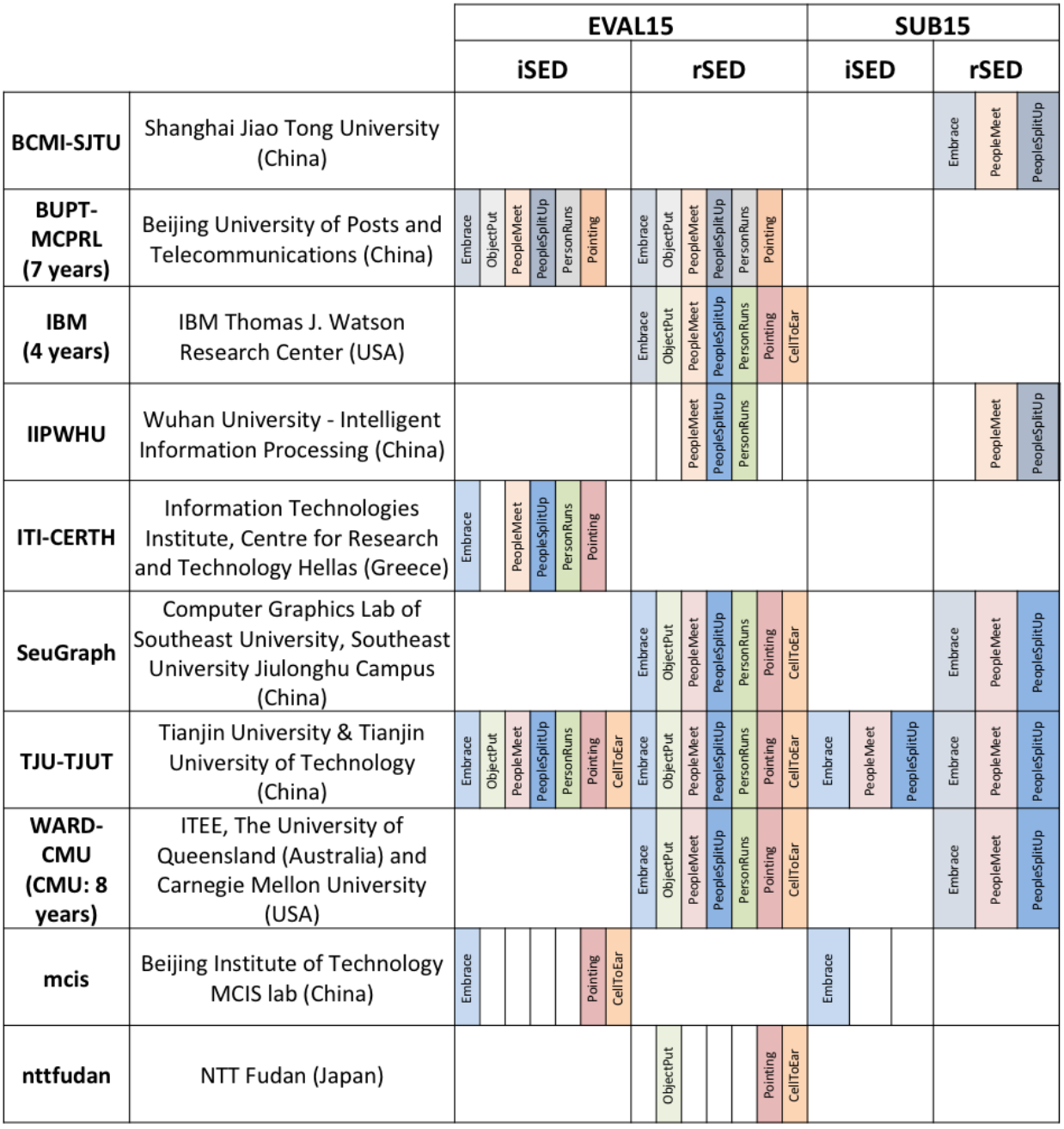


Figure 38: SED15 Participants

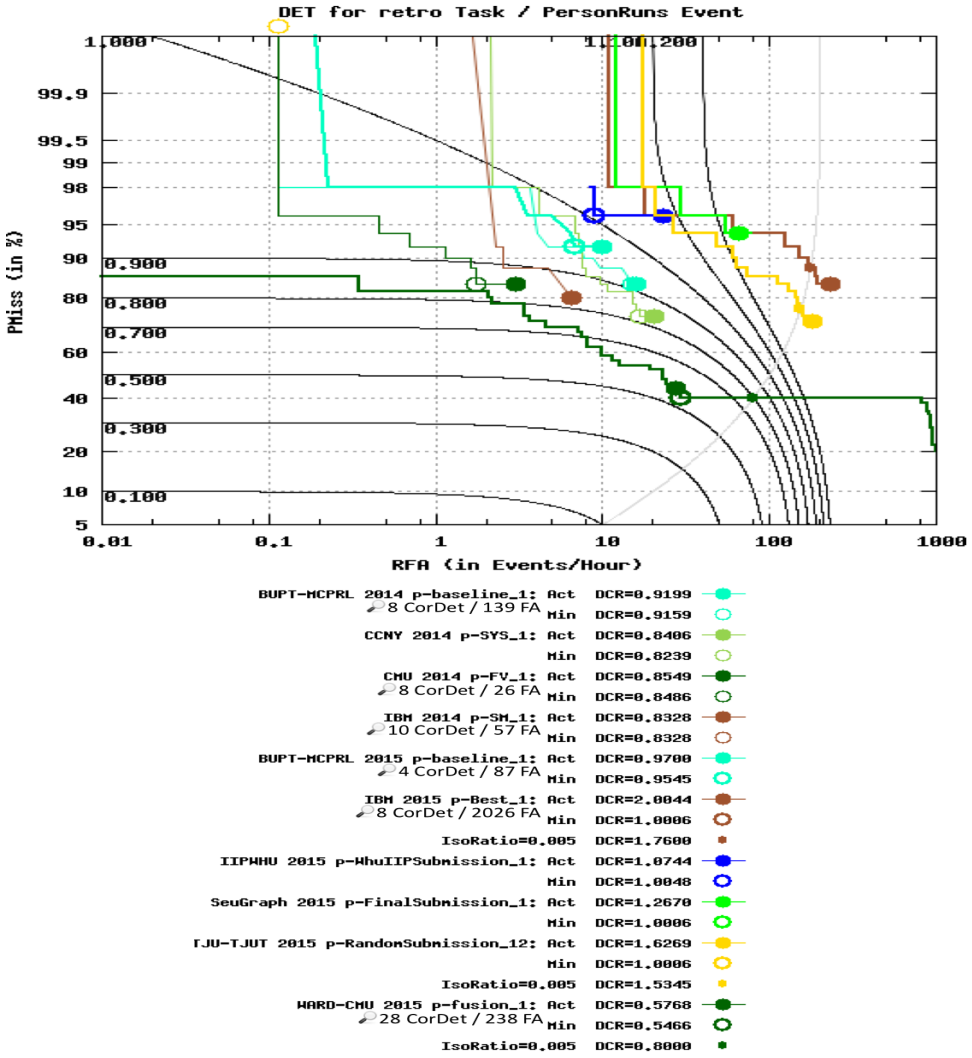


Figure 39: PersonRuns: 2014 and 2015 rSED

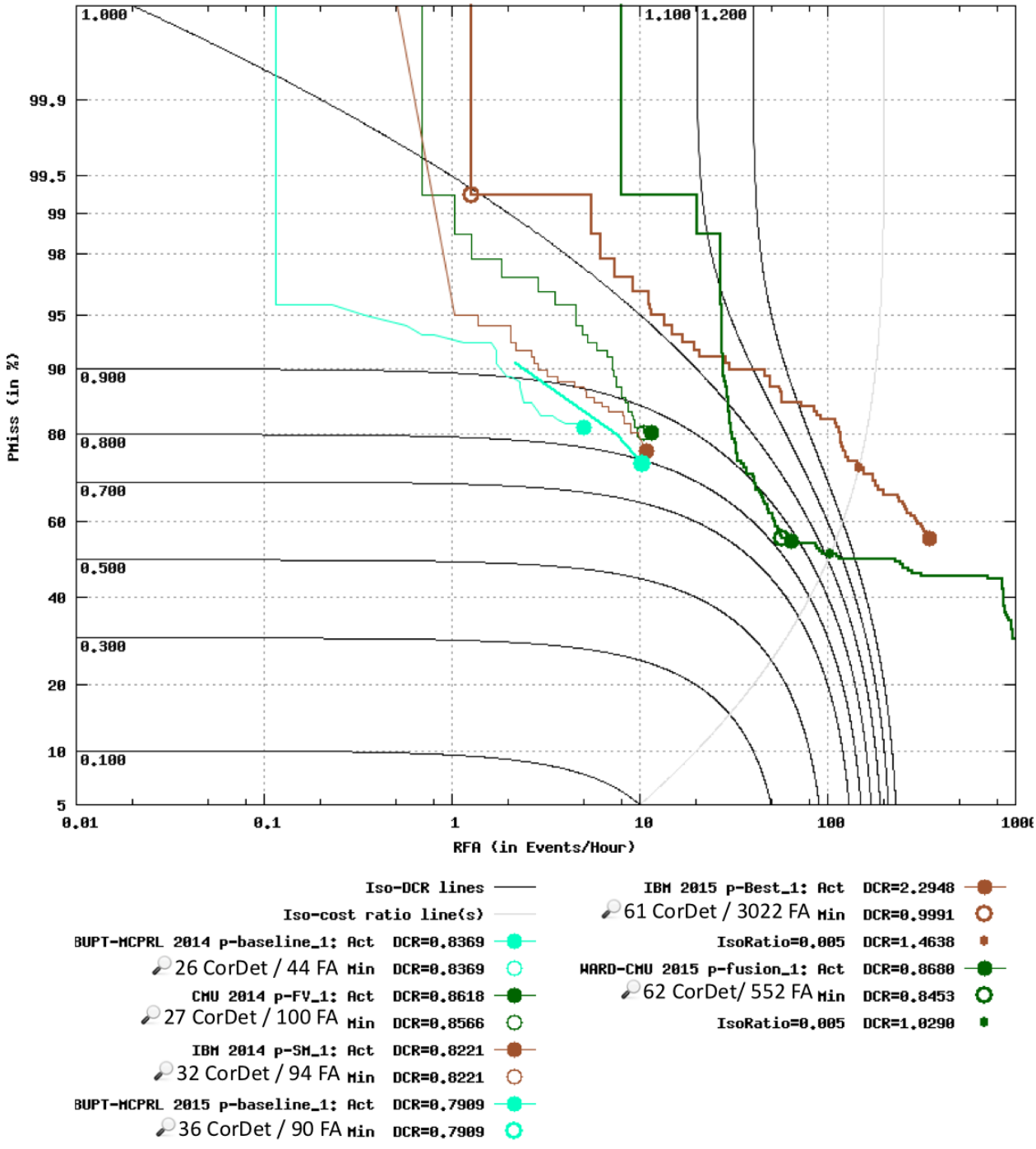


Figure 40: Embrace: 2014 and 2015 rSED (only repeat site)

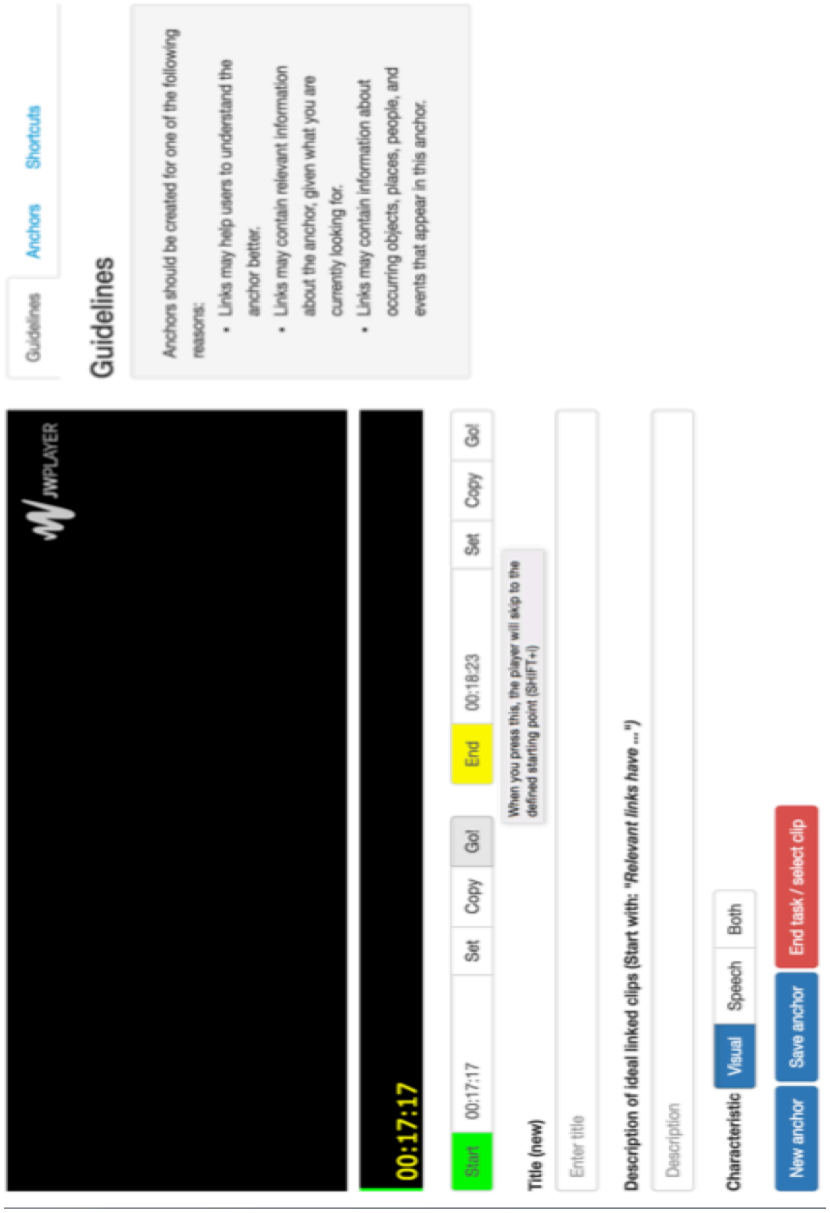


Figure 41: Screenshot of the user interface for manual video hyperlink generation

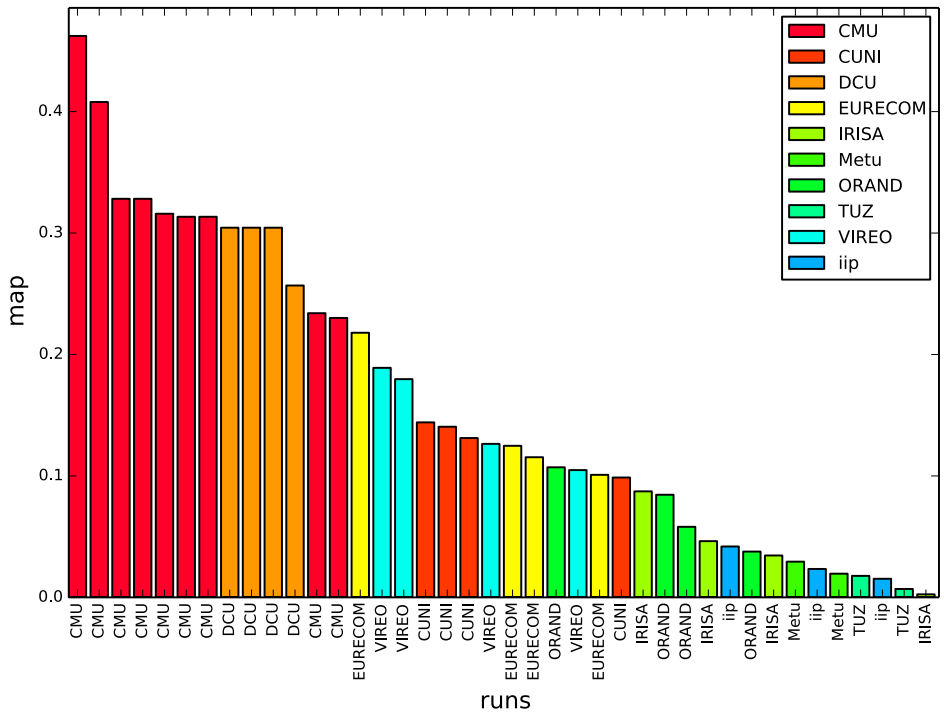


Figure 42: The results of the Video Hyperlinking evaluation based on Mean Average Precision (MAP).

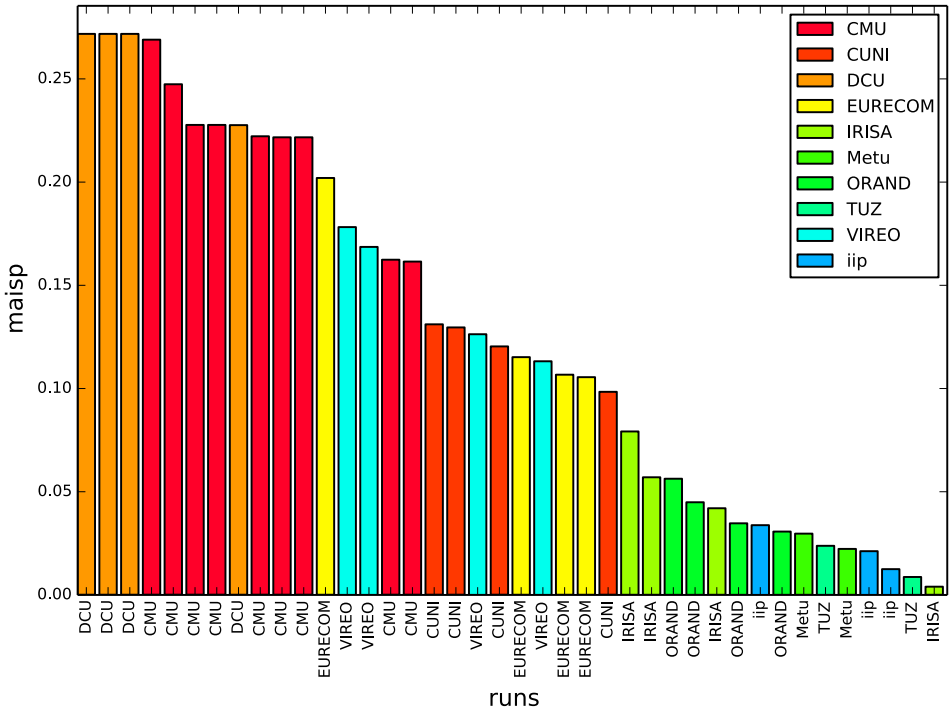


Figure 43: The results of the Video Hyperlinking evaluation based on Mean Average interpolated Segment Precision (MAiSP).

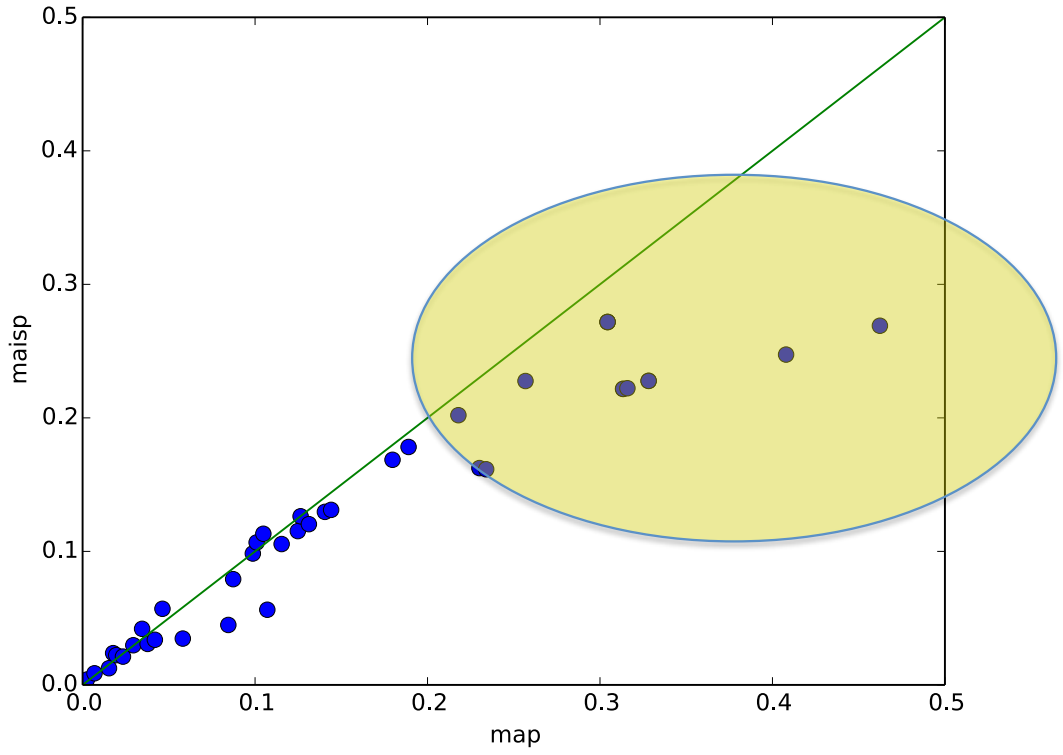


Figure 44: Plot of the correlation of the two measures MAP and MAiSP.

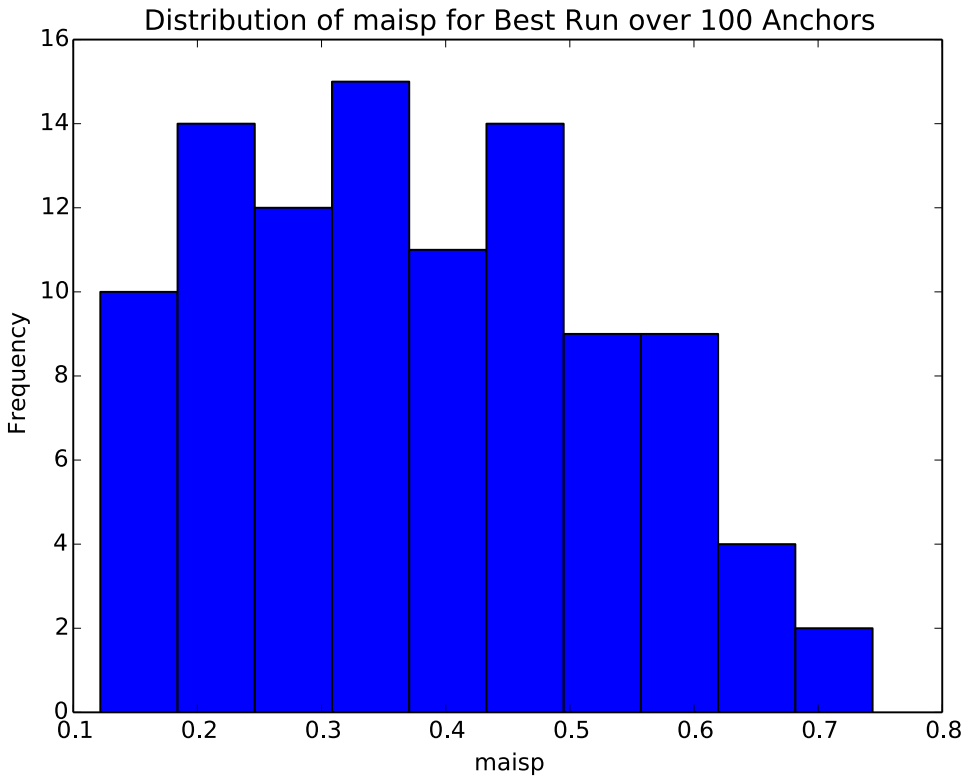


Figure 45: Distribution of the maximum performance of an anchor.