TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking

George Awad {gawad@nist.gov} Jonathan Fiscus {jfiscus@nist.gov}
David Joy {david.joy@nist.gov} Martial Michel {martial.michel@nist.gov}

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

Alan F. Smeaton {alan.smeaton@dcu.ie}
Insight Centre for Data Analytics, Dublin City University

Wessel Kraaij {w.kraaij@liacs.leidenuniv.nl}
Leiden University; TNO, Netherlands

Georges Quénot {Georges.Quenot@imag.fr}
Laboratoire d'Informatique de Grenoble

Maria Eskevich {M.Eskevich@let.ru.nl}
Radboud University, Netherlands

Robin Aly {r.aly@utwente.nl} Roeland Ordelman {roeland.ordelman@utwente.nl}
University of Twente, Netherlands

Marc Ritter {ritter@hs-mittweida.de}
Technische Universitat Chemnitz

Gareth J. F. Jones {gareth.jones@computing.dcu.ie}
ADAPT Centre, Dublin City University, Ireland

Benoit Huet {benoit.huet@eurecom.fr}
EURECOM, Sophia Antipolis, France

Martha Larson {m.a.larson@tudelft.nl}
Radboud University; Delft University of Technology, Netherlands

August 8, 2017

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2016 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 fourteen 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 NIST (National Institute of Standards and Technology) and other US government agencies. In addition, many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2016 represented a continuation of five tasks from 2015, the replacement of the semantic indexing task by a new ad-hoc video search task, and a new pilot video to text description task. 39 teams (see Table 1) from various research organizations worldwide completed one or more of the following seven tasks:

1. Ad-hoc Video Search (AVS)
2. Instance Search (INS)
3. Multimedia Event Detection (MED)
4. Surveillance Event Detection (SED)
5. Video Hyperlinking (LNK)
6. Concept Localization (LOC)
7. Video to Text Description (pilot task) (VTT)

Table 2 represent organizations that registered but did not submit any runs. About 600 new hours of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC.3) were used for ad-hoc Video Search. Unlike previously used professionally edited broadcast news and educational programming, the IACC videos reflect a wide variety of content, style, and source device determined only by the self-selected donors. The instance search task used about 464 hours of the BBC (British Broadcasting Corporation) EastEnders video. A total of almost 4 738 hours from the Heterogeneous Audio Visual Internet (HAVIC) collection of Internet videos in addition to a subset of Yahoo YFC100M videos were used in the multimedia event detection task. For the surveillance event detection task, 11 hours of airport surveillance video was used, while 3,288 hours of blip.tv videos were used for the video Hyperlinking task. The concept localization task used approximately 2.2 million I-frame images for testing. Finally, a new video to text pilot task was proposed this year. The task used about 2000 Twitter vine videos collected through the online API public stream.

Ad-hoc search, 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 chosen by media professionals. Surveillance event detection was scored by NIST using ground truth created by NIST through manual adjudication of test system output. Finally, the new pilot task was annotated by collaboration with Technische Universitat Chemnnitz (TUC) group of Dr. Marc Ritter.

This paper is an overview to the evaluation framework — the tasks, data, measures used in the workshop and high-level results analysis. For detailed information about the approaches and results, the reader should see the various site reports and the results pages available at the workshop proceeding online page [TV16Pubs, 2016].

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 Video Data

2.1 BBC EastEnders video

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

2.2 Internet Archive Creative Commons (IACC.3) video

The IACC.3 dataset consists of 4 593 Internet Archive videos (144 GB, 600 h) with Creative Commons licenses in MPEG-4/H.264 format with duration ranging from 6.5 to 9.5 min and a mean duration of ≈7.8

Table 1: Participants and tasks

Task						Location	TeamID	Participants
−−	*HL*	−−	*MD*	*SD*	AV	NAm+Asia	INF	Beijing U. of Posts and Tele.;U. Autonoma de Madrid; Shandong U.; Xian JiaoTong U.
−−	−−	−−	*MD*	∗∗	–	Asia	BIT_MCIS	Beijing Inst. of Tech., Media Computing and Intelligent System Lab.
−−	∗∗	−−	*MD*	−−	AV	Asia	VIREO	City U. of Hong Kong
−−	*HL*	−−	−−	−−	–	Eur	IRISA	CNRS, IRISA, INSA, Universite de Rennes 1
−−	−−	∗∗	−−	−−	AV	Asia	UEC	U. of Electro-Communications, Tokyo
IN	−−	−−	−−	−−	–	Asia	U_TK	U. of Tokushima
IN	−−	−−	−−	−−	–	Aus	UQMG	U. of Queensland - DKE Group of ITEE
IN	−−	−−	−−	∗∗	∗∗	Eur	insightdcu	Dublin City U.; Polytechnic U. of Catalonia
−−	−−	−−	*MD*	−−	–	NAm	Etter	Etter Solutions
−−	*HL*	−−	−−	−−	AV	Eur	EURECOM	EURECOM
−−	−−	−−	−−	−−	AV	NAm	FIU_UM	Florida International U.; U. of Miami
−−	*HL*	−−	−−	−−	–	NAm	FXPAL	FX PALO ALTO LABORATORY, INC
−−	−−	−−	−−	*SD*	–	Asia	HRI	Hikvision Research Institute
IN	−−	−−	*MD*	*SD*	AV	Eur	ITI_CERTH	Centre for Research and Tech. Hellas
IN	−−	−−	−−	−−	∗∗	Eur	IRIM	EURECOM;LABRI;LIG;LIP6;LISTIC
IN	−−	−−	−−	−−	∗∗	Eur	JRS	JOANNEUM RESEARCH
−−	−−	−−	−−	−−	AV	Eur	ITEC_UNIKLU	Klagenfurt University
−−	−−	−−	−−	−−	AV	Eur+Asia	kobe_nict_siegen	Kobe U.; Natl. Inst. of Inf. and Comm. Tech.;U. of Siegen
−−	−−	−−	*MD*	−−	–	Asia	KoreaUnivISPL	Korea U.
IN	−−	−−	*MD*	−−	–	NAm+Asia	PKU_MI	Peking U.; Rutgers U.
IN	−−	∗∗	*MD*	*SD*	∗∗	Asia	BUPT_MCPRL	Beijing U. of Posts and Telecommunications
IN	∗∗	*LO*	*MD*	*SD*	AV	Asia	NII_Hitachi_UIT	Natl. Inst. of Inf.;Hitachi; U. of Inf. Tech.
IN	−−	−−	−−	−−	–	Asia	WHU_NERCMS	Natl. Eng. Research Center for Multimedia Software, Wuhan U.
−−	−−	−−	*MD*	∗∗	–	Asia	nttfudan	NTT Media Intelligence Laboratories; Fudan U.
IN	∗∗	∗∗	∗∗	∗∗	∗∗	NAm+Asia	PKU_ICST	Peking U.
−−	*HL*	−−	−−	−−	–	Eur	EURECOM_POLITO	Politecnico di Torino Eurecom
−−	−−	−−	−−	*SD*	–	Aus	WARD	U. of Queensland
IN	−−	−−	−−	−−	–	Asia	SIAT_MMLAB	Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
−−	−−	−−	−−	*SD*	–	Asia	SeuGraph	Southeast U. Computer Graphics Lab
IN	−−	−−	−−	−−	–	Asia	TRIMPS_SARI	Third Research Inst., Ministry of Public Security; Chinese Academy of Sciences
−−	−−	*LO*	*MD*	−−	∗∗	Asia	TokyoTech	Tokyo Inst. of Tech.
IN	−−	−−	−−	−−	–	Eur	TUC	TU Chemnitz - Junior Professorship Media Computing - Chair Media Informatics
−−	−−	−−	−−	−−	AV	Eur	IMOTION	U. of Basel; U. of Mons; Koc U.
∗∗	−−	∗∗	*MD*	−−	AV	Eur	MediaMill	U. of Amsterdam
−−	−−	*LO*	−−	−−	–	Aus	UTS_CMU_D2DCRC	U. of Technology, Sydney D2DCRC
−−	−−	−−	−−	−−	AV	Eur	vitrivr	U. of Basel
−−	∗∗	∗∗	∗∗	−−	AV	Asia	Waseda	Waseda U.
−−	∗∗	−−	−−	*SD*	–	Asia	IIP_WHU	Wuhan U.

Task legend. IN:Instance search; MD:Multimedia event detection; HL:Hyperlinking; LO:Localization; SD:Surveillance event detection; AV:Ad-hoc; −−:no run planned; ∗∗:planned but not submitted

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

Approximately 1 200 h of IACC.1 and IACC.2 videos used between 2010 to 2015 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.3 videos.

2.3 iLIDS Multiple Camera Tracking Data

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

The training videos consisted of the ≈100 h of data used for SED 2008 evaluation. The evaluation videos consisted of the same additional ≈50 h of data from the Imagery Library for Intelligent Detection System's (iLIDS) multiple camera tracking scenario data used for the 2009 to 2013 evaluations

Table 2: Participants who did not submit any runs

Task						Location	TeamID	Participants
IN	*HL*	*LO*	*MD*	*SD*	*AV*			
−−	−−	**	−−	−−	**	Eur	PicSOM	Aalto U.
−−	−−	−−	−−	−−	**	Asia	ABZOOBA	Abzooba Inc. India
−−	−−	−−	**	−−	−−	NAm	fork	Arizona state U.
−−	−−	**	**	−−	**	Asia	SamHMS	Beijing Samsung Telecom R&D Center
−−	−−	**	−−	**	−−	NAm	BCTS	Brain Corporation Technical Services
−−	−−	−−	−−	−−	**	NAm	CCNY	City U. of New York; Graduate Center, City U. of New York; NVIDIA Research
**	−−	−−	−−	−−	−−	Asia	CVARL_WU	Computing Center of Computer School at Wuhan U.
**	−−	**	**	**	**	Eur	ADVICE	BASKENT U.
−−	−−	−−	**	−−	−−	Eur	HEU008	Harbin Engineering U.
−−	−−	**	**	−−	−−	Asia	hulustar	HULU LLC
−−	−−	−−	−−	**	−−	Asia	NP	IIT Hyderabad
−−	−−	**	**	−−	−−	Eur	INRIA_STARS	INRIA
−−	−−	−−	−−	−−	**	Asia	TAM	Intel
**	−−	−−	**	−−	**	Asia	Ravi	JNTUK
−−	−−	**	−−	−−	**	Eur	LIG	Laboratoire d'Informatique de Grenoble
**	−−	**	−−	−−	−−	Eur	MetuMedia	Middle East Technical U. Department of Electrical/Electronics Engineering
**	−−	−−	−−	**	−−	Asia	Mitsubishi_Electric	Mitsubishi Electric Corporation
−−	−−	−−	**	**	−−	Asia	MI_TJU	Multimedia Institute, Tianjin U.
−−	−−	−−	**	−−	−−	Asia	nus_action	National U. of Singapore
−−	−−	−−	**	−−	−−	NAm	NEU_MITLL	Northeastern U. and MIT Lincoln Laboratory
**	−−	−−	−−	−−	−−	Asia	NTT	NTT Communication Science Laboratories; NTT Media Intelligence Laboratories
**	**	−−	**	−−	−−	SAm	ORAND	ORAND S.A. Chile
−−	−−	−−	−−	**	−−	NAm	QUPROR	Private Research
**	−−	**	**	**	**	Asia	QUT	Qatar U.
−−	**	**	−−	−−	**	Asia	REGIMVID	REGIM; U. of Sfax
**	−−	−−	**	−−	−−	Asia	saricas	Shanghai Advanced Research Institute, Chinese Academy of Sciences
−−	−−	−−	−−	**	−−	Asia	sjtu_icl	Shanghai Jiao Tong U.
−−	−−	−−	−−	**	−−	Asia	zy_scu	Sichuan U.
**	**	**	**	**	**	Asia	Trimps	The Third Research Institute of the Ministry of Public Security
**	**	**	**	**	**	Asia	HAWKEYE	Tsinghua U.
**	−−	−−	−−	−−	−−	Asia	THSS_IMMG	Tsinghua U. School of Software
**	**	**	**	−−	−−	Eur	TUZ	TUBITAK UZAY
**	−−	−−	−−	−−	−−	Asia	BMC_UESTC	U. of Electronic Science and Technology of China
**	−−	−−	−−	−−	−−	Eur+Asia	Sheffield_UETLahore	U. of Sheffield; U. of Engineering & Technology
−−	−−	−−	**	−−	−−	Eur+Asia	trento_tokyo_univ	U. of Trento
−−	−−	−−	−−	**	−−	Eur+Asia	UniKent	U. of Kent
−−	−−	−−	**	−−	**	Asia	zjgsucvg	Zhejiang Gongshang U.

Task legend. IN:instance search; MD:multimedia event detection; HL:Hyperlinking; LO:Localization; SD:surveillance event detection; AV:Ad-hoc; −−:no run planned; ∗∗:planned but not submitted

[UKHO-CPNI, 2009] .

2.4 Heterogeneous Audio Visual Internet (HAVIC) Corpus

The HAVIC Corpus [Strassel et al., 2012] is a large corpus of Internet multimedia files collected by the Linguistic Data Consortium and distributed as MPEG-4 (MPEG-4, 2010) formatted files containing H.264 (H.264, 2010) encoded video and MPEG-4 Advanced Audio Coding (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. Approximately 98 003 clips with a total duration of 3 712.89 h and total size of 1300 GB were reused from the MED15 task as an evaluation collection.

2.5 Yahoo Flickr Creative Commons 100M dataset (YFCC100M)

The YFCC100M dataset [Thomee et al., 2016] is a large collection of images and videos available on Yahoo Flickr. All photos and videos listed in the collection are licensed under one of the Creative Commons copyright licenses. The YFCC100M dataset is comprised of 99.3 million images and 0.7 million videos. Only a subset of the YFCC100M videos (100 000 Clips with a total duration of 1 025.06 h and total size of 352 GB) are used for evaluation.

2.6 Blip10000 Hyperlinking video

The Blip10000 data set consists of 14 838 videos for a total of 3 288 h from blip.tv. The videos cover a broad range of topics and genres. It has automatic speech recognition transcripts provided by LIMSI, and user-contributed metadata and shot boundaries provided by TU Berlin. Also, video concepts based on the MediaMill MED Caffe models are provided by EURECOM.

3 Ad-hoc Video Search

The previous Semantic Indexing task which has run from 2010 to 2015 addressed the problem of automatic assignment of predefined semantic tags representing visual or multimodal concepts to video segments. More and more concepts were trained and developed over the course of those six years. However, testing individual visual concepts is not very realistic in a real-world setting as an average user would more likely be interested in searching for those concepts in a particular context or in a combined form. This year a new Ad-hoc search task was introduced to model the end user video search use-case, who is looking for segments of video containing persons, objects, activities, locations, etc. and combinations of the former.

It was coordinated by NIST and by Georges Quénot at the Laboratoire d'Informatique de Grenoble.

The Ad-hoc video search task was as follows. Given a standard set of shot boundaries for the IACC.3 test collection and a list of 30 Ad-hoc queries, participants were asked to return for each query, at most the top 1 000 video clips from the standard set, ranked according to the highest possibility of containing the target query. The presence of each query 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 query 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 query 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 query target, such as photos, paintings, models, or toy versions of the target (e.g picture of Barack Obama vs Barack Obama himself), was NOT grounds for judging the query to be true for the segment. Containing video of the target within video may be grounds for doing so.

Like it's predecessor, in 2016 the task again supported experiments using the "no annotation" version of the tasks: the idea is to promote the development of methods that permit the indexing of concepts in video clips using only data from the web or archives without the need of additional annotations. The training data could for instance consist of images or videos retrieved by a general purpose search engine (e.g. Google) using only the query definition with only automatic processing of the returned images or videos. This was implemented by adding the categories of "E" and "F" for the training types besides A and D:[1]

- A - used only IACC training data
- D - used any other training data
- E - used only training data collected automatically using only the official query textual description
- F - used only training data collected automatically using a query built manually from the given official query textual description

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.

Two main submission types were accepted:

- Fully automatic runs (no human input in the loop): System takes a query as input and produces result without any human intervention.
- Manually-assisted runs: where a human can formulate the initial query based on topic and query interface, not on knowledge of collection or search results. Then system takes the formulated query as input and produces result without further human intervention.

TRECVID evaluated 30 query topics. Some samples are listed in Appendix A.

[1] Types B and C were used in some past TRECVID iterations but are not currently used.

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 clips 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.3 collection was used for testing. It contained 335 944 video clips.

3.2 Evaluation

Each group was allowed to submit up to 4 prioritized main runs and two additional if they were "no annotation" runs. In fact 13 groups submitted a total of 52 runs, from which 22 runs were manually-assisted and 30 were fully automatic runs.

For each query topic, pools were created and randomly sampled as follows. The top pool sampled 100 % of clips ranked 1 to 200 across all submissions after removing duplicates. The bottom pool sampled 11.1 % of ranked 201 to 1000 clips and not already included in a pool. 10 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, 187 918 clips were judged while 371 376 clips fell into the unjudged part of the overall samples.

3.3 Measures

The *sample_eval* software (`http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/sample_eval/`), 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 topics, runs can be compared in terms of the mean inferred average precision across all evaluated query topics. The results also provide some information about "within topic" performance.

3.4 Results

The frequency of correctly retrieved results varied greatly by query. Figure 1 shows how many unique instances were found to be true for each tested query. The inferred true positives (TPs) of only 1 query exceeded 0.5 % from the total tested clips. Top 5 found queries were "a person playing drums indoors", "a person wearing a helmet", "soldiers performing training or other military maneuvers", "military personnel interacting with protesters", and "a woman wearing glasses". On the other hand, the bottom 5 found queries were "a sewing machine", "a person lighting a candle", "a man shake hands with a woman", "a man with beard and wearing white robe speaking and gesturing to camera", and "people shopping". The complexity of the queries or the nature of the dataset may be factors in the different frequency of hits across the 30 tested queries. Figure 2 shows the number of unique clips found by the different participating teams. From this figure and the overall scores it can be shown that top performing automatic runs indeed were among the most unique clips contributors, while on the contrary, top performing manually-assisted runs were among the least unique clips contributors. Given that manually-assisted runs performed in general better than fully automatic runs, we can conclude that humans helped the system in retrieving more common clips but not necessarily unique clips.

Figures 3 and 4 show the results of all the manually-assisted and fully automatic run submissions respectively. Except for the Waseda team runs which achieved the maximum InfAP score of 0.177, all manually-assisted runs almost reached maximum InfAP scores of 0.047. On the other hand, the fully automatic run scores are more smooth and reached a maximum of 0.054 and median score of 0.024. We should also note here that few runs were submitted under the training category of E (6 runs) and F (0 runs) while the majority of runs were of type D. Compared to the previous 6 years of semantic indexing task that was running to detect single concepts (e.g airplane, animal, bridge,...etc) it can be shown from the results that the ad-hoc task is still very hard and

systems still have a lot of room to research methods that can deal with unpredictable queries composed of one or more concepts.

Figures 5 and 6 show the performance of the top 10 teams across the 30 queries. 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 a specific team. A team's scores can rank differently across the 30 queries. As expected there are more queries within the manually-assisted runs that achieved higher scores compared to their corresponding ones in the automatic runs. Top query scores from both run types came from the queries: "a person playing drums indoors", "the 43rd president George W. Bush sitting down talking with people indoors", "a choir or orchestra and conductor performing on stage", "palm trees", "one or more people at train station platform", "any type of fountains outdoors", "a person sitting down with a laptop visible", and "a person wearing a helmet". Bottom query scores from both run types came from the queries: "a person jumping", "a man shake hands with a woman", "a woman wearing glasses", "a person drinking from a cup, mug, bottle, or other container", and "people shopping". A main theme among the top performing queries is their composition of more common visual concepts compared to the bottom ones which require more temporal analysis for some activities (shaking hands, jumping, drinking, shopping, and discriminating between man and woman). In general there is a noticeable spread in score ranges among the top 10 runs which may indicate the variation in the performance of the used techniques and that there is still room for further improvement.

To test if there were significant differences between the systems' performance, we applied a randomization test [Manly, 1997] on the top 10 runs for manually-assisted submissions as shown in Figures 7 and 8 using significance threshold of p¡0.05. The figures indicate the order by which the runs are significant according to the randomization test. Different levels of indentation signify a significant difference according to the test. Runs at the same level of indentation are indistinguishable in terms of the test. In this test the top 2 ranked runs were significantly better than all other runs while there is no significant difference between the two of them. On the other hand for automatic runs, the randomization test did not reveal that there is any run that is significantly better than the rest among the top 10 submissions.

Among the submission requirements, we asked teams to submit the processing time that was consumed to return the result sets for each query. Figure 9 plots the reported processing time vs the InfAP scores among all run queries for automatic runs. It can be shown that spending more time did not necessarily help in many cases and few queries achieved high scores in less time. There is more work to be done to make systems efficient and effective at the same time.

In order to measure how were the submitted runs diverse we measured the percentage of common clips across the same queries between each pair of runs. We found that on average only about 7.5 % (minimum 0 %) of submitted clips are common between any pair of runs. These results show the diversity of the used approaches and their output. In comparison, the average was about 23 % and the minimum was 14 % for the previous year of the semantic indexing task.

2016 Observations

A summary of general observations can be drawn to show that most teams relied on intensive visual concept indexing, leveraging on past semantic indexing tasks and used popular datasets for training such as ImageNet. Deep learning approaches dominated teams' methods and used pretrained models. Different methods applied manual or automatic query transformation approaches. Fusion of concept scores (e.g. Waseda team) was investigated by most teams to combine useful results that satisfy the queries. Ad-hoc search is more difficult than simple concept-based tagging as shown by the big gap between past semantic indexing best performance and the new Ad-hoc search task. Manually-assisted runs performed better than fully-automatic suggesting more work needs to be done for query understanding and knowledge transfer between the human experience in formulating the query and the automatic system. Most systems did not provide real-time response for an average system user. In addition, the slowest systems were not necessarily the most effective. Finally the E and F runs are still rare compared to A and D. For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV16Pubs, 2016] in the online workshop notebook proceedings.

4 Instance search

An important need in many situations involving video collections (archive video search/reuse, personal video organization/search, surveillance, law enforcement, protection of brand/logo use) is to find more video segments of a certain specific person, object, or place, given one or more visual examples of the specific item. Building on work from previous years in the concept detection task [Awad et al., 2016] the instance search task seeks to address some of these needs. For the past six years (2010-2015) the instance search task has tested systems on retrieving specific instances of individual objects, persons and locations. This year systems were tested on a new query type, to retrieve specific persons in specific locations.

4.1 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 Flickr (2012). Finding realistic test data, which contains sufficient recurrences of various specific objects/persons/locations under varying conditions has been difficult.

In 2013 the task embarked on a multi-year effort using 464 h of the BBC soap opera EastEnders. 244 weekly "omnibus" files were divided by the BBC into 471 523 video clips to be used as the unit of retrieval. The videos present a "small world" with a slowly changing set of recurring people (several dozen), locales (homes, workplaces, pubs, cafes, restaurants, open-air market, clubs, etc.), objects (clothes, cars, household goods, personal possessions, pets, etc.), and views (various camera positions, times of year, times of day).

4.2 System task

The instance search task for the systems was as follows. Given a collection of test videos, a master shot reference, a set of known location/scene example videos, and a collection of topics (queries) that delimit a person in some example videos, locate for each topic up to the 1000 clips most likely to contain a recognizable instance of the person in one of the known locations.

Each query consisted of a set of

- The name of the target person
- The name of the target location
- 4 example frame images drawn at intervals from videos containing the person of interest. For each frame image:
  - a binary mask covering one instance of the target person
  - the ID of the shot from which the image was taken

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)

4.3 Topics

NIST viewed a sample of test videos and developed a list of recurring people, locations and the appearance of people at certain locations. In order to test the effect of persons or locations on the performance of a given query, the topics tested target persons across the same locations. In total this year we asked systems to find 7 target persons across 5 target locations. 30 test queries (topics) were then created (Appendix B).

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

4.4 Evaluation

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 41 automatic and 7 interactive runs (using only the first 20 topics). Each interactive search was limited to 5 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 clips were being found or time ran out. Table 3[2] presents information about the pooling and judging.

[2]Please refer to Appendix B for query descriptions.

Table 3: Instance search pooling and judging statistics

Topic number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number relevant	% judged that were relevant
9159	45016	14915	33.1	280	5226	35	92	1.8
9160	44880	15726	35	140	2580	16.4	68	2.6
9161	44919	14866	33.1	120	2318	15.6	13	0.6
9162	44557	16898	37.9	200	4029	23.8	31	0.8
9163	44899	13997	31.2	180	2867	20.5	305	10.6
9164	45354	15231	33.6	460	8647	56.8	890	10.3
9165	45352	11726	25.9	360	4337	37	1169	27
9166	45377	13275	29.3	280	3841	28.9	849	22.1
9167	45420	10905	24	520	5669	52	1614	28.5
9168	45825	14533	31.7	200	3691	25.4	763	20.7
9169	45796	13118	28.6	440	5708	43.5	715	12.5
9170	45833	12968	28.3	200	2902	22.4	247	8.5
9171	45809	14152	30.9	520	7272	51.4	546	7.5
9172	45843	12827	28	340	4123	32.1	785	19
9173	45818	15837	34.6	360	6792	42.9	1135	16.7
9174	45817	15147	33.1	380	6213	41	428	6.9
9175	45787	14446	31.6	220	4097	28.4	99	2.4
9176	45835	16249	35.5	200	3581	22	231	6.5
9177	45732	15322	33.5	280	4786	31.2	321	6.7
9178	45887	14243	31	460	7217	50.7	896	12.4
9179	39734	13280	33.4	180	3047	22.9	49	1.6
9180	39733	12201	30.7	220	3462	28.4	144	4.2
9181	39256	14320	36.5	520	8504	59.4	574	6.7
9182	39221	11973	30.5	200	3152	26.3	134	4.3
9183	39207	13000	33.2	220	3507	27	116	3.3
9184	39786	13438	33.8	420	6379	47.5	1243	19.5
9185	39741	14009	35.3	220	3655	26.1	88	2.4
9186	39751	12827	32.3	180	3139	24.5	81	2.6
9187	39784	14885	37.4	140	2677	18	38	1.4
9188	39743	13271	33.4	220	3326	25.1	136	4.1

4.5 Measures

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

4.6 Results

Figure 10 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 0.4 down to almost 0.0. Per-topic variance varies as well with the largest values being associated with topics that had the best performance. Two main factors might be expected to affect topic difficulty this year, the target person or the location. From the analysis of the performance of topics, it can be shown that for example the persons "Dot" and "Brad" were easier to find as 3 "Dot" topics were among the top 15 topics compared to only 1 in the bottom 15 topics. Similarly, 4 "Brad" topics were among the top 15 topics compared to only 1 in the bottom 15 topics. In addition, it seems that the public location "Pub" made it harder to find the target persons at as 5 out

of the bottom 15 topics were at the location "Pub" compared to only 2 in the top 15 topics.

Figure 11 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 $p < 0.05$.

Figure 12 shows the box plot of the interactive runs performance. For the majority of the topics, they seem to be equally difficult when compared to the automatic runs. A common observation is that the location "Pub" is still among the majority of the bottom (most difficult) 10 topics (topics 168, 178, 159, 173, and 164). In general interactive results boosted some of the hard topics compared to automatic runs performance.

Figure 13 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 14 for the automatic runs with elapsed times less than or equal to 10 s. Only 3 teams reported processing time below 10 s. Among those, the UQMG team that used more time, did not achieve higher score compared to a much faster team (TUC).

Figure 15 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 runs that took advantage of the video examples achieved the highest scores compared to using only image examples. These results are consistent to previous years. However, the majority of the rest of the runs used only the image examples. This was the third year where the video for the images examples was made available and we hope more systems will use those examples in future years for better training data.

4.7 Summary of observations

The new task using the same Eastenders dataset followed a stable number of finishing rate from participants. One team submitted manual runs or special types of interactive runs with very high average precision scores. However, after researching the approach, the team and the organizers agreed that although the run is inspiring and innovative, it does not fall under neither the automatic, nor the interactive categories that the task planned for in the guidelines. Thus this led to more discussions about if the task needs an additional third category (Manual) run type and specifically how teams can handle their prior knowledge about the closed world of the Eastenders dataset and the previous ground truth data. Few teams submitted interactive runs where they mainly focused on relevance feedback and cleaning up result lists. Specific methods for detecting and recognizing faces mainly based on CNN (Convolutional Neural Networks) helped significantly, while learning locations was difficult since they are usually occluded by people. The best location strategies combined CNN and BOVW (Bag of Visual Words) using traditional SIFT (Scale-Invariant Feature Transform) features. In general almost all systems had dedicated pipelines for persons and locations. More work is observed dealing with scene threading or linking related clips. In addition, some teams experimented with exploring external data such as closed captions and fan resources.

A summary of the different approaches used by participating teams in order to find an optimal representation includes using SIFT, VGG19 (Visual Geometry Group), VGG-places-205, SIFT BOW, CNN ImageNet for locations and using DLIB (Dynamic Library) detection, VGG-face, VGG16-faces, Openface using CNN, person reidentification based on tracking clothes, ASR (Automatic Speech Recognition) search, face tracking, and SADR (Scale-Adaptive Deconvolutional Regression) network for persons.

In regard to exploiting the query images/videos the Wuhan team manually selected ROI (region of interest) on different query images which helped their system significantly especially for finding correct locations. The JRS team blurred the area outside the region of interest mask for persons while InsightDCU only used the face part of the masked target person. The PKU team applied different transformations on the sample query images for CNNs while the WUHAN team used extra images from the web for characters and locations. The full query video clips were exploited for query expansion by the teams of IRIM, PKU and SIAT.

Different matching and ranking experiments are reported by systems. Typically systems fuse the locations and character search results. The BUPT team applied query adaptive late fusion method similarly to previous year, the WUHAN team applied Asymmetrical query adaptive matching, the SIAT and WHU teams used Hamming embedding, while

the TUC teams used a linear weighted fusion between person and location giving more weight to person results. A semi-supervised learning for discarding noisy videos was applied by the PKU team.

Postprocessing the ranked list results also has been investigated by the IRIM team where they filtered credits, ads, opening and ending credit segments. The NII-HITACHI team applied geometric verification and CNN filtering. The SIAT team used spatial verification for locations, while the TU Chemnitz team used improved version of semantic sequence clustering.

Readers should see the online proceedings for individual teams' performance and runs.

5 Multimedia event detection

The 2016 Multimedia Event Detection (MED) evaluation was the sixth evaluation of technologies that search multimedia video clips for complex events of interest to a user.

The focus of MED 15 was to make MED less costly to both participate in and administer. MED 16 continues that trend by replacing a portion of the test set with an equal number of videos from the Yahoo Flickr Creative Commons 100M dataset (YFCC100M), which is new to MED this year. The YFCC100M dataset is more readily accessible and contains shorter duration videos than the HAVIC dataset.

The MED 16 evaluation protocol is identical to MED 15, with the following modifications:

- Replaced roughly half of the test set with a subset of the YFCC100M dataset videos.
- Introduced 10 new Ad-Hoc (AH) events.
- Scored both Pre-Specified (PS) and AH event sets using Inferred Mean Average Precision [Yilmaz et al., 2008], reference generated through pooled assessment.

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 a mnemonic title for the event.
- An event definition which was a textual definition of the event.
- An event explication which was an expression of some event domain-specific knowledge needed by humans to understand the event definition.
- An evidential description which was a textual listing of the attributes that are indicative of an event instance. The evidential description provides a notion of some potential types of visual and acoustic evidence indicating the event's existence but it was not an exhaustive list nor was it to be interpreted as required evidence.
- A set of illustrative video examples containing either an instance of the event or content related to the event. The examples were illustrative in the sense they helped form the definition of the event but they did not demonstrate all the inherent variability or potential realizations.

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 were revealed. This style of system was introduced in MED 2012. The third style 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. As with MED 15, no one participated in this task.

5.1 Data

A development and evaluation collection of Internet multimedia (i.e., video clips containing both audio and video streams) clips was provided to MED participants.

The HAVIC data, which was collected and distributed by the Linguistic Data Consortium, consists

Table 4: MED '16 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

of publicly available, user-generated content posted to the various Internet video hosting sites. Instances of the events were collected by specifically searching for target events using text-based Internet search engines. All video data was reviewed to protect privacy, remove offensive material, etc., prior to inclusion in the corpus. Video clips were provided in MPEG-4 formatted files. The video was encoded to the H.264 standard. The audio was encoded using MPEG-4's Advanced Audio Coding (AAC) standard.

The YFCC100M data, collected and distributed by Yahoo!, consists of photos and videos licensed under one of the Creative Commons copyright licenses. While the entire YFCC100M dataset consists of 99.3 million images and 0.7 million videos, only a subset of 100 000 randomly selected[3] videos were chosen for this years evaluation.

MED participants were provided the data as specified in the HAVIC and YFCC100M data sections of this paper. The MED '16 Pre-Specified event names are listed in Table 4, and Table 5 lists the MED '16 Ad-Hoc Events.

[3]Clips included in the YLI-MED Corpus, [Bernd et al., 2015] were excluded from selection. Clips not hosted on the multimedia-commons public S3 bucket were also excluded, see http://mmcommons.org/

Table 5: MED '16 Ad-Hoc Events

E051 - Camping
E052 - Crossing a Barrier
E053 - Opening a Package
E054 - Making a Sand Sculpture
E055 - Missing a Shot on a Net
E056 - Operating a Remote Controlled Vehicle
E057 - Playing a Board Game
E058 - Making a Snow Sculpture
E059 - Making a Beverage
E060 - Cheerleading

5.2 Evaluation

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

- Events: all 20 Pre-Specified events (PS16) and/or all 10 Ad-Hoc events (AH16).
- Interactivity: Human interaction with query refinement using the search collection.
- Test collection: either the MED16 Full Evaluation collection (MED16-EvalFull) or a 783 h subset (MED16-EvalSub) collection.
- 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 100 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 10 Ex, 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, participants labeled their systems as the closest match to one of three cluster sizes: small, medium and large. (See above.)

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

5.3 Measures

System output was evaluated by how well the system retrieved and detected MED events in the 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 measure for performance was Inferred Mean Average Precision. For the Pre-Specified events, we report both Mean Average Precision (MAP) and Mean Inferred Average Precision (MInfAP) for the HAVIC Progress subset of the evaluation data, as well as the correlation between them. MInfAP is the sole metric reported for the YFCC100M subset and the complete MED16-EvalFull and MED16-EvalSub evaluation sets.

5.4 Results

12 teams participated in the MED '16 evaluation; 1 team was new. All teams participated in the Pre-Specified (PS) Event condition, processing the 20 PS events. All but one team completed the required 10 Exemplar (10Ex) PS evaluation condition. 4 teams chose to participate in the Ad-Hoc (AH) portion of the evaluation, which was optional, processing the 10 AH events. 7 teams chose to process the MED16-EvalSub set. This year, only one team submitted runs for a "Medium" (MED) sized system.

Figures 16 and 17 show the MAP scores per team on the Progress subset for both the MED16-EvalFull and MED16-EvalSub respectively. Results are broken down by hardware classification and exemplar training condition. Teams who processed MED16-EvalFull but not MED16-EvalSub were scored on the MED16-EvalSub subset in figure 17 and subsequent figures featuring MED16-EvalSub. As with previous years' results, MAP scores for MED16-EvalSub are inflated when compared to scores on MED16-EvalFull due to the higher density of positives in the MED16-EvalSub set. That said, MAP scores between the MED16-EvalFull and MED16-EvalSub sets are highly correlated; with an R^2 of 0.996.

Figures 18 and 19 show the Pre-Specified Average Precision scores for the Progress subset of MED16EvalSub on the 10Ex exemplar training condition broken down by event and team respectively. An event effect can be observed in figure 19, with most events showing a tight range of scores excluding low-scoring systems.

For the Mean Inferred Average Precision (MInfAP), we follow Yilmaz et al.'s procedure, Statistical Method for System Evaluation Using Incomplete Judgements [Yilmaz and Aslam, 2006], whereby we use a stratified, variable density, pooled assessment procedure to approximate MAP. Event references for both PS and AH event sets for the YFCC100M subset were generated using InfAP procedures using strata sizes and sampling rates used during last years evaluation, see [Over et al., 2015]. Specifically, we define two strata 1-60 with a sampling rate of 100 % and 61-200 at 20 %. We refer to Inferred Average Precision measures using these parameters as InfAP200 in subsequent figures. As with last year, we found that this method produces MInfAP scores which are highly correlated with MAP scores for corresponding sets; R^2 of 0.99 on the MED16-EvalSub Progress subset using simulated[4] MInfAP.

Figures 20 and 21 show the MInfAP scores for the PS event condition on MED16-EvalFull and MED16-EvalSub respectively. Note that both HAVIC Progress and YFCC100M portions of the evaluation set are included here. As an aside, figure 23 shows the top 200 ranked clips, by dataset, for a sample of systems (high, medium and medium-low scoring systems as Team1, Team2, and Team3 respectively) and events; demonstrating the heterogeneous nature of the datasets.

This year, we introduced 10 new Ad-Hoc events. We used the stratified sampling method detailed above to select clips for scoring AH. While these AH events were already fully annotated for the HAVIC Progress portion of the evaluation set by the LDC, selected clips from the YFCC100M subset were annotated by NIST annotators. Figure 22 shows the AH event condition MInfAP scores on MED16-EvalFull. Note that MED16-EvalFull, 10Ex were the only evaluation conditions supported for AH.

Figures 24 and 25 show the Inferred Average Precision scores for the Ad-Hoc event set broken down

[4]Clips were selected using the InfAP procedure, but used the original HAVIC reference annotation

by event and team respectively. Note the overall low performance of the "Crossing a barrier" (E052) event, during annotation we noticed that many of the high-ranking clips reported by teams were clips of children escaping from cribs, which is explicitly mentioned in the event kit text as not constituting a positive instance of the event.

As with MAP, MInfAP scores are sensitive to the event richness (true positives) of the test collection as demonstrated in Figures 26 and 27.

For detailed information about the approaches and results, the reader should see the various site reports in the online workshop notebook [TV16Pubs, 2016].

5.5 Summary

In summary, 11 of 12 teams participated in the Pre-Specified (PS), 10 Exemplar (10Ex) test, processing all 20 events, with MAP scores on the Progress subset of MED16-EvalFull ranging from 21.35 to 28.96 (median of 27.01), and MAP scores on the Progress subset of MED16-EvalSub (including MED16-EvalFull submissions scored on the subset for teams who did not make a MED16-EvalSub submission) ranging from 0.42 to 35.45 (median of 29.79). The MInfAP scores for these same MED16-EvalFull submissions over the entire evaluation set (i.e. Progress + YFCC100M) ranged from 27.43 to 39.40 (median of 36.62), while MED16-EvalSub submissions (again including MED16-EvalFull submissions scored on MED16-EvalSub) ranged from 0.26 to 38.52 (median of 33.87).

For the Ad-Hoc, 10 Exemplar evaluation condition, in which teams are required to process the MED16-EvalFull set, only 4 of 12 teams participated, with MAP scores on the Progress subset of MED16-EvalFull ranging from 15.43 to 25.38 (median of 24.62). MInfAP scores on the MED16-EvalFull set, including both Progress and YFCC100M, ranged from 30.70 to 46.28 (median of 44.77).

As with last year, no teams participated in the Interactive Event Query test. Consequently, we will not support the Interactive Event Query test for MED '17.

For MED '17 we will continue using the Inferred Average Precision procedure given the strong correlation between MAP and MInfAP scores, and the low cost of annotation. We also intend to release the Progress annotations, and introduce a new set of Ad-Hoc events. MED '17 participants can also expect to see both the HAVIC and YFCC100M data again, though the exact makeup of the evaluation set is yet to be determined.

6 Surveillance event detection

The 2016 Surveillance Event Detection (SED) evaluation was the ninth 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 continued from 2009 till 2015. 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 2016, the evaluation test data used a 10-hour subset (EVAL16) from the total 45 h available of the test data from the Imagery Library for Intelligent Detection System's (iLIDS)[UKHO-CPNI, 2009] Multiple Camera Tracking Scenario Training (MCTTR) dataset. This dataset was collected by the UK Home Office Centre for Applied Science and Technology (CAST) (formerly Home Office Scientific Development Branch's (HOSDB)). EVAL16 added 1 h to the EVAL15 set.

This 10 h dataset contains a subset of the 11-hour SED14 Evaluation set that was generated following a crowdsourcing effort in order to generate the reference data. Since 2015, "camera4" is not used, as it had few events of interest.

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

Introduced in 2015 was a 2-hour "Group Dynamic Subset" (SUB15) limited to three specific events: Embrace, PeopleMeet and PeopleSplitUp. This dataset was reused in 2016 as SUB16.

In 2016, only the retrospective event detection was 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 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).

6.1 Data

The development data consisted of the full 100 h data set used for the 2008 Event Detection [Rose et al., 2009] evaluation. The video for the evaluation corpus came from the approximate 50 h iLIDS MCTTR dataset. Both datasets 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 EVAL16 and/or SUB16. Like SED 2012 and after, systems were provided the identity of the evaluated subset.

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 and 2016 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.

6.2 Evaluation

For EVAL16, sites submitted system outputs for the detection of any 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.

SUB16 followed the same concept, but only using 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.

6.3 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 Detection 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.

SED16 results will be presented using three metrics:

1. **A**ctual NDCR (Primary Metric), computed by restricting the putative observations to those with true actual decisions.

2. **M**inimum NDCR (Secondary Metric), a diagnostic metric found by searching the DET curve for its minimum cost. The difference between the value of Minimum NDCR and Actual NDCR indicates the benefit a system could have gained by selecting a better threshold.

3. **N**DCR at Target Operating Error Ratio (NDCR@TOER, Secondary Metric), is another diagnostic metric. It is found by searching the DET curve for the point where it crosses the theoretical balancing point where two error types (Miss Detection and False Alarm) contribute equally to the measured NDCR. The Target Operating Error Ratio point is specified by the ratio of the coefficient applied to the False Alarm rate to the coefficient applied to the Miss Probability.

More details on result generation and submission process can be found within the TRECVID SED16 Evaluation Plan [5].

6.4 Results

SED16 saw 8 sites participate (see Figure 28), four from China, and one each for Australia, Greece, Japan, USA and Vietnam.

Figure 29 shows, per Event and per Metric the systems with the lowest NDCR for the 2016 SED Evaluation (only on primary submissions).

Figure 30, 31 and 32 present the SED16 Results for the PeopleSplitUp, Embrace and PersonRuns events. For additional individual results, please see the TRECVID SED proceedings.

Despite the introduction of the "Group Dynamic Subset" (SUB16), the task is still very difficult. From those results we see that performance improvement has slowed, and related to this problem, only two of the eight participants (down from ten last evaluation cycle) participated in SED for over two years, six are teams that joined in the past two years.

We propose to continue SED17 using the same data, metrics and evaluation protocols as EVAL16 and SUB16.

[5] ftp://jaguar.ncsl.nist.gov/pub/SED16/TRECVid-SED16-EvaluationPlan.pdf

7 Video hyperlinking

7.1 System task

The high-level definition of the Video Hyperlinking (LNK) task in 2016 is the same as that of the 2015 edition of the task [Over et al., 2015]. The task requires the automatic generation of hyperlinks between given manually defined *anchors* within source videos and *target* videos from within a substantial collection of videos. Both targets and anchors are video segments with a start time and an end time. The result of the task for each anchor is a ranked list of target videos in decreasing likelihood of being *about* the content of the given anchor. Targets have to fulfill the following requirements: i) they must be from different videos than the anchor, ii) they may not overlap with other targets in the same anchor, finally iii), in order to facilitate ground truth annotation, the targets must be between 10 and 120 seconds in length.

The 2016 edition of the LNK task has the following main differences from the 2015 edition:

- The task switched from the BBC broadcast collection (professionally generated content) to a subset of the Blip10000 collection [Schmiedeke et al., 2013] crawled from blip.tv, a website that hosted semi-professional user-generated content.

- The 2016 anchors were multimodal, i.e., the information about suitable targets, or the information request, is a combination of both audio and visual streams. We focus on the multimodality in video interpretations as these are likely to be shared between multiple viewers. We apply the simple reasoning that the videomaker and the viewer are likely to interpret the video in the same way: in other words, the information that the videomaker intends to convey with the video is the information that the viewer receives by watching the video. Because we cannot know the intent of the videomaker directly, we look for

a reliable way to judge his or her intent. Specifically, we make the assumption that intent to convey information is obviously present when the videomaker uses both the visual and the audio channel to get a single message across. In the future, other reasoning may also be interesting to explore.

The multimodality of anchors was ensured within a two step anchor creation framework: the verbal-visual anchors were first defined by two media professionals, further these anchors were verified via crowdsourcing [Eskevich et al., 2017].

- The relevance assessment framework was split into two steps carried out using the Amazon MTurk platform: general vetting of the submitted target video segments, and collection of detailed relevance descriptions.

7.2 Data

The Blip10000 dataset used for the 2016 task consists of 14,838 semi-professionally created videos [Schmiedeke et al., 2013]. As part of the task release, automatically detected shot boundaries were provided [Kelm et al., 2009], together with automatic speech recognition (ASR) transcripts [Lamel, 2012] originally provided with this dataset.

Additionally, new versions of ASR transcripts and visual features were made available for the task. The new set of ASR transcripts were created by LIMSI using the 2016 version of their neural network acoustic models in their ASR system. The visual concepts were obtained using the BLVC CaffeNet implementation of the so-called AlexNet, which was trained by Jeff Donahue (@jeffdonahue) with minor variation from the version described in [Krizhevsky et al., 2012]. The model is available with the Caffe distribution [6]. In total, detection scores for 1000 visual concepts were extracted, with the five most likely concepts for each keyframe being released along with their associated confidence scores.

Data inconsistencies

Two issues were identified in the distributed version of the collection.

- For one video the wrong ASR file was provided. Here, we blacklisted the video, totally excluding it from the results and evaluation.
- With regard to the metadata creation history, not all types of metadata were created using the original files, rather some made use of intermediate extracted content in the form of extracted audio for the ASR transcripts. This led to the misalignment issue between ASR transcripts and keyframe timecodes, i.e. for some video files, the length of the provided *'.ogv'* encoding was shorter than the encoding for which the shot cut detection and keyframe extraction was performed. In these cases, it was possible for a run that used visual data only to return segments that did not exist in the ASR transcripts, which were derived from the *'.ogv'* video files. For 416 video files, circa 3 % of all the data, the keyframes extended more than five minutes over the supplied *'.ogv'* video, which corresponds to 138 h of extension. To make the evaluation comparable, we ignored all results after the end time of the *'.ogv'* video files across the collection.

7.3 Anchors

Anchors in the video hyperlinking task are essentially comparable to the search topics used in a standard video retrieval tasks. As in the 2015 edition of the task, we define an anchor to be the triple of: video (v), start time (s) and end time (e).

In 2016, we focused on multimodal anchors. Specifically, we selected anchors in which the videomaker, i.e., the person who created the video, is using both the audio and video modalities in order to convey a message. Our definition of multimodality was applied using the following test. If the message of the video segment could be completely understood using only the visual or only the audio channel, the segment was not considered multimodal. In other words, for a segment to be multimodal, the viewer needs both channels in order to appreciate its intended message in full. In order to easily locate candidate segments that are potentially multimodal according to this test, we compiled a list of the following speech cues, which we take to be associated with situations in which people are showing something while talking about it: 'can see', 'seeing here', 'this looks', 'looks like', 'showing', and 'want to show'. For practical reasons, we also limited anchors to be between 10 and 60 seconds long. The anchors were selected by anchor creators, who reviewed the speech segments in which one of the cues was recognized. Anchor creators watched each can-

[6] See http://caffe.berkeleyvision.org/ for details.

didate segment and assessed whether or not it passed the test for multimodality. In total, two creators generated 94 anchors and corresponding descriptions of potentially relevant targets, i.e., information request descriptions that were further used in the evaluation process. Three of these 94 anchors were later discarded from the evaluation because the crowdsourcing anchor verification step did not confirm them as truly multimodal, while one additional anchor was used as example at this stage [Eskevich et al., 2017]. This resulted in the final list of the 90 multimodal anchors that were used in the 2016 task edition as test set.

7.4 Evaluation

Ground truth

The ground truth was generated by pooling the top 5 results of all formally submitted participant runs (20), and running the assessment tasks on the Amazon Mechanical Turk (AMT)[7] platform[8]. Overall, the ground truth creation proceeded in two stages:

- 'Target Vetting': The top 5 targets for each anchor from the participants' runs were assessed using a so-called forced choice approach, which constrains the crowdworkers' responses to a finite set of options. Concretely, the crowdworkers were given a target video segment and five textual targets descriptions (one of them being taken from the actual anchor that the target in question has been retrieved for). The task for the workers was to choose a definition that they felt was best suited to a given video segment. In case they chose the target description of the original anchor, this was considered to be a judgment of relevance. In case the target was unsuitable for any of the anchors, i.e., it was considered non-relevant, the crowdworkers were expected not to be comfortable making the choice among the five given options. For each top-5 anchor–target pair we collected three crowdworkers' judgments. The final relevance decision was made based on the majority of the relevance judgments.

- 'Video-to-Video Relevance Analysis': the crowdworkers were shown both the anchor and target video segments, and were asked to give a textual description (2-3 natural language sentences) of the relevance relationship, i.e., what made the target relevant to the anchor.

The Target Vetting stage for all the participants' submissions involves large-scale crowdsourcing submissions processing, which is not feasible to carry out manually. For this reason, we ran a manual check of a small subset of crowdworkers submissions to the Target Vetting stage in order to confirm that the task was understood correctly. Beyond this subset, the submissions were accepted or rejected automatically, according to a procedure that checked whether all the required decision metadata fields had been filled in, and whether the answers to the test questions were correct.

Initially, we aimed at providing ground truth from the top 10 results of the 20 submitted runs. However, the top-10 rank positions contained a total of 12 758 non-overlapping segments. Due to limited assessment resources, we focused on the top-5 rank positions from each run, comprising in total 7 216 targets. Of these targets, 2526 were identified as relevant and 4690 non-relevant.

7.5 Measures

The evaluation metrics were chosen to reflect diverse aspects of system performance. Specifically, the metrics were Precision at rank 5 (Precision@5), and an adaptation of Mean Average Precision called Mean Average interpolated Segment Precision (MAiSP), which is based on previously proposed adaptations of MAP for this task [Racca and Jones, 2015]. Precision at rank 5 was chosen as the ground truth judgments were collected for the top 5 rank positions of all submitted runs, which means this metric reflects the quality of all of the top-ranked results that were assessed. The MAiSP metric takes into account whether the relevant content is retrieved up to rank-position 1000 in the list. This metric enables a comparison between the runs below rank position 5 in terms of user effort measured in the amount of time that needs to be spent to access relevant content.

7.6 Results

Five groups submitted four runs each, resulting in 20 run submissions, which were used for ground truth creation and assessment using the metrics described above.

The Readers should see the online proceedings for individual teams' performance and runs. An overall

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

[8] For all HITs details, see: `https://github.com/meskevich/Crowdsourcing4Video2VideoHyperlinking/`

comparison of the systems' performance according to Precision at rank 5 and MAiSP are given in Figures 33-34.

In terms of Precision@5, 3 systems (IRISA and two INF runs) achieved scores above 0.5, with FXPAL, EURECOM and EURECOM.POLITO following. The order of the teams changes when results were evaluated with respect to MAiSP (INF achieves the highest score, IRISA the third, while more of the EURECOM and FXPAL runs achieve similar scores between 0.10 and 0.12).

The systems that combined multiple modalities in their approaches achieved higher scores according to both metrics. This finding is consistent with the fact that the anchors were defined to be multimodal, suggesting that the targets would also be multimodal, and that both audio and visual modalities would contribute to finding them.

8 Concept localization

The localization task challenges systems to make their concept detection more precise in time and space. Currently other video search tasks such as Adhoc and instance search 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 a subset of 10 concepts from those chosen and used in the semantic Indexing task between 2012 and 2015 and building on the work done in previous years [Awad et al., 2016]. This year a different set of concepts was tested than those tested in the past 3 years. In addition, most of the concepts were dynamic in nature compared to the object concepts used in previous years.

For each concept from the list of 10 designated for localization, NIST distributed[9] a subset list of up to 1000 clips where each video shot may or may not contain the concept.

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, the one chosen by the NIST assessor as the most prominent.

[9]The data was available to the teams about 5 weeks before the localization submissions were due at NIST for evaluation

Concept	*Name*	*clips*	*IFrames*
6	Animal	997	31330
13	Bicycling	998	21912
16	Boy	998	34230
38	Dancing	983	31584
49	Explosion_fire	983	20816
71	Instrument_Musician	1000	30374
100	Running	1000	24842
107	Sitting_Down	1000	52779
434	Skier	1000	32900
163	Baby	1000	17298

Table 6: Evaluated localization concepts

Table 6 describes for each of the 10 localization concepts the number of clips NIST distributed to systems and the number of I-Frames comprised by those clips.

8.1 Data

In total, 2 205 140 jpeg I-frames were extracted from the IACC.2 collection. 9 959 total clips were distributed and included a total of 298 065 I-frames.

8.2 Evaluation

For each shot that contains a concept and selected and distributed by NIST, all I-frames were selected and 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, 55 789 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 the most prominent one and continue selecting it unless its prominence changed and another target concept became more prominent.

Assessors were instructed 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.

In total, 11 runs were submitted this year by 3 teams.

8.3 Measures

Temporal and spatial localization were evaluated using precision and recall based on the judged items at two levels - the frame as the basis for temporal localization and the pixel bounding box for spatial localization. NIST then calculated an average for each of these values for each concept and for each run.

The set of annotated I-Frames was then used to evaluate the localization for the I-Frames submitted by the systems.

8.4 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 35 shows the mean precision, recall and F-score of the returned I-frames by all runs across all 10 concepts. In general systems' performance almost doubled the maximum F-score values compared to the years of 2013 and 2014 as the max F-score this year reached about 0.45. We should note here that we can not compare the performance to last year as only true positive clips were given to systems to localize in 2015. In addition, this year concepts are mainly action oriented and so are more difficult to detect and localize compared to 2013 and 2014 when concepts were mainly objects.

On the other hand Figure 36 shows the same measure by run for spatial localization (correctly returning a bounding box around the concept). Here the F-scores range was less than the temporal F-score range but still higher (reached 0.27) than maximum 2013 spatial F-scores and almost near 2014 maximum F-score.

The F-score performance by concept for the top 10 runs is shown in Figures 37 and 38 for temporal and spatial respectively across all runs. In general, most concepts achieved higher temporal scores compared to spatial localization. Also a noticeable resemblance between the performance of the concepts across the two measures is clear. In both measures, the top performed concepts were Animal, Bicycling, Instrumental_Musician, and Baby, while the weak performed concepts were Boy and sitting_down.

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 39 and 40 for temporal and spatial localization respectively. We can see in Figure 39 some concepts submitted a lot of non-target I-frames which resulted in low precision and high recall. Hard concepts achieved recall and precision values of less than 0.5, while very few concept results achieved good recall and precision above 0.5.

Figure 40 shows an interesting observation. Systems are good at 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.

8.5 Summary of observations

It is clear that for the past 4 years temporal localization was easier than the spatial localization and systems could approximate the ground truth box sizes. Performance of the action/dynamic concepts achieved higher scores than object concepts tested in 2013-2014 which is a good sign that systems are getting better and more sophisticated. During the human assessment at NIST, it was proven that asking human judges to localize dynamic concepts is very hard task as they had to watch each video clip several times to verify the concept. As the finishing rate of teams this year was very low (3 teams finished out of 21 signed up), it was decided to discontinue the task while keeping the past data and resources available to the community as a benchmark to evaluate their new systems. Finally, readers should see the online proceedings for individual teams' performance and runs.

9 Video to Text Description

Automatic annotation of videos using natural language text descriptions has been a long-standing goal of computer vision. The task involves understanding of many concepts such as objects, actions, scenes, person-object relations, the temporal order of events throughout the video and many others. In recent years there have been major advances in computer vision techniques which enabled researchers to start practical work on solving the challenges posed in video captioning.

There are many use case application scenarios which can greatly benefit from technology such as video summarization in the form of natural language,

including facilitating the search and browsing of video archives using such descriptions, describing videos to the blind, etc. In addition, learning video interpretation and temporal relations among events in a video will likely contribute to other computer vision tasks, such as prediction of future events from the video.

This year a new showcase/pilot task was proposed and launched within TRECVid which we refer to as "Video to Text Description" (VTT) and describe in this section.

9.1 Data

A dataset of more than 30k Twitter Vine videos has been collected automatically. Each video has a total duration of about 6 s. In this showcase/pilot task a subset of 2 000 Vine videos was randomly selected and annotated manually twice by two different annotators. In total, 4 non-overlapping sets of 500 videos each were given to 8 annotators to generate a total of 4 000 text descriptions. Those 4 000 text descriptions were split into 2 sets corresponding to the original 2 000 videos. Annotators were asked to include and combine into 1 sentence, if appropriate and available, four facets of the video they are describing:

- Who is the video describing (e.g. concrete objects and beings, kinds of persons, animals, or things)
- What are the objects and beings doing? (generic actions, conditions/state or events)
- Where is the video taken (e.g. locale, site, place, geographic location, architectural)
- When is the video taken (e.g. time of day, season)

9.2 System task

The task set for participants was as follows: given a set of about 2 000 URLs of Vine videos and two sets (A and B) of text descriptions (each composed of 2 000 sentences), systems were asked to work and submit results for at least one of two subtasks:

- Matching and Ranking: Return for each video URL a ranked list of the most likely text description that corresponds (was annotated) to the video from each of the sets A and B.
- Description Generation: Automatically generate for each video URL a text description (1 sentence) independently and without taking into consideration the existence of sets A and B.

9.3 Evaluation

The matching and ranking subtask scoring was done automatically against the ground truth using mean inverted rank at which the annotated item is found. The Description generation subtask scoring was done automatically using standard metrics from machine translation (MT) including METEOR [Banerjee and Lavie, 2005] and BLEU [Papineni et al., 2002]. BLEU (bilingual evaluation understudy) is a metric used in MT and was one of the first metrics to achieve a high correlation with human judgements of quality. It is known to perform more poorly if it is used to evaluate the quality of individual sentence variations rather than sentence variations at a corpus level. In the VTT task the videos are independent thus there is no corpus to work from, so our expectations are lowered when it comes to evaluation by BLEU. METEOR (Metric for Evaluation of Translation with explicit ORdering) is based on the harmonic mean of unigram or n-gram precision and recall, in terms of overlap between two input sentences. It redresses some of the shortfalls of BLEU such as better matching synonyms and stemming, though the two measures seem to be used together in evaluating MT.

Systems taking part in VTT were encouraged to take into consideration and use the four facets that annotators used as a guideline to generate their automated descriptions. In addition to using MT metrics, an experimental semantic similarity metric (STS) [Han et al., 2013] was applied. This metric measures how semantically similar the submitted description is to the ground truth descriptions (both A and B). In total, 11 teams signed up for the pilot/showcase task and 7 of those finished, submitting 46 individual runs to the matching and ranking subtask and 16 runs to the description generation subtask.

9.4 Results

Readers should see the online proceedings for individual teams' performance and runs but here we present a high-level overview.

The first set of results for the caption-ranking subtask are shown in Figures 41 and 42 and shows the mean inverted rank scores for all submitted runs,

color-coded by participating group. From Figure 41 we can see that submitted runs from different groups cluster together indicating each group submitted runs which were slight variants of each other. To interpret these results we can say that with a mean inverted rank value of greater than 0.1, on average systems can find the correct caption from a set of 2 000 captions, at rank position 9 or 10.

Figure 42 shows the same results but differentiating the A runs from B runs and from this we can see that systems seem to perform better on the B runs than on the A runs, though what that means is not very clear.

If we explore these results a bit deeper, then Figure 43 shows the rank positions at which the top 3 videos from 2 000, in terms of how "easy" they were to rank, were ranked by the submitting groups.

Video number 324, for example, was found within the top-1p rank positions by 5 of the runs, while for some of the runs it was not located until almost rank position 1 000. This demonstrates the variable difficulty in this task, some systems perform well on some videos but badly on others.

The videos shown in Figure 44, for example, are among the easiest to rank highly, right across the submitted runs with their captions being *"a woman and a man are kissing each other"* and *"a dog imitating a baby crawling across the floor in a living room"*. On the other hand, the videos in Figure 45 are among the most challenging with their captions being *"3 balls hover in front of a man"* and *"a person wearing a costume and carrying a chainsaw"*. What makes these latter two videos most challenging is the unusual nature of the actions that take place. In the first case the 3 basketballs do hover in front of the subject as it is a video of a magic trick, and in the second case there is a man with a chainsaw who chases another man, so both represent atypical behaviour.

Moving on to the results from the caption generation sub-task, Figures 46 and 47 show the performance of runs submitted by 5 participating groups using the BLEU and METEOR metrics respectively.

The BLEU results in Figure 46 are difficult to interpret because, for example, multiple results from single sites are scattered throughout the results list whereas one would expect results from a single site to cluster as each site submits only minor variations of its own system for generating captions. This may be due to the issues associated with using BLEU for this tak, as mentioned earlier. The METEOR results in Figure 47 show results for each site are indeed clustered and thus may be more reliable. However, for both BLEU and METEOR, trying to interpret the values of the system scores is impossible so their real value is in comparison only.

However, in order to give the reader some insight into the captions actually generated, Figure 48 shows a keyframe from one of the videos where a baby crawls forward across what appears to be a livingroom carpet, the camera zooms out to reveal a dog behind the baby and the dog does indeed mimic the way the baby crawls with its hind legs trailing behind it. Below are the submitted captions for this video from across the groups (there are duplicate captions among the submissions).

- a girl is playing with a baby
- a little girl is playing with a dog
- a man is playing with a woman in a room
- a woman is playing with a baby
- a man is playing a video game and singing
- a man is talking to a car
- A toddler and a dog

What this shows is that there are good systems that do well, and others that do not do well in terms of the captions that they generate, just as there are videos which are easier to caption than others, and each approach does well on some videos and badly on others, but not consistently so. For detailed information about the approaches and results, the reader should see the various site reports in the online workshop notebook [TV16Pubs, 2016].

9.5 Conclusions and Observations

The first observation to make about the VTT pilot sub-tasks is that there was good participation from among TRECVid groups and that there are submitted captions and caption rankings from across the groups, which generate impressive results. Not all generated captions or caption rankings are correct or impressive, but there are enough good ones in the results of this pilot task to be encouraged. One of the quirks of the results was that B runs did better than A in matching and ranking while A did better than B in the semantic similarity, but that may be just an artefact of the annotation.

In terms of metrics used, METEOR scores are higher than BLEU, and in retrospect we should have

used the CIDEr metric (Consensus-based Image Description Evaluation) [Vedantam et al., 2015] and as can be seen in their TRECVid papers, some participants did include this metric in their write-ups anyway. The CIDEr metric has been reported to show high agreement with consensus as assessed by humans. Using sentence similarity, the notions of grammaticality, saliency, importance and accuracy (precision and recall) are inherently captured by the metric. We plan to use again the three MT metrics including CIDEr in future evaluations so that a comparison can be made to assess which one can be more reliable as a metric.

STS, the semantics-based metric which we also used as a metric has some issues based on our anecdotal of its coverage for the language. This makes us ask what makes more sense, MT or semantic similarity, and which metric measures real system performance in a realistic application. To illustrate this, Figure 49 shows the values of the STS metric where captions A and B are measured against each other, for each of the 2 000 videos. One would expect that the A and B captions would be semantically similar, perhaps even identical and so we would hope for a large number of the 2 000 similarity measures to be at, or close to, a value of 1,0. Instead, as Figure 49 illustates, the median similarity is only 0.545 with a disappointing tailoff of similarities close to a value of 0. That tells us that either the A and B annotators got things terribly wrong, which is unlikely given the quality assurance we used in producing the manual annotations, or the STS measure is having difficulty measuring similarity across just a short video caption, or the vocabulary used in the captioning creates difficulties for the STS computation. Whatever the reason, STS similarity values would not add much value to interpreting the real performance of submitted runs.

One thing that became apparent as we looked at the approaches taken by different participants is that there are lots of available training sets for this task, including MSR-VTT, MS-COCO, Place2, ImageNet, YouTube2Text, and MS-VD. Some of these even have manual ground truth captions generated with Mechanical Turk such as the MSR-VTT-10k dataset [Xu et al., 2016] which has 10 000 videos, is 41.2 h in duration and has 20 annotations for each video.

This provides a rich landscape for those wishing to use machine learning in all its various forms within the VTT task and particpants in VTT have used all of these at some point.

Finally, while this was a pilot task in TRECVid in 2016, there are other video-to-caption challenges like the ACM MULTIMEDIA 2016 Grand Challenge where images from YFCC100M with captions were used in a caption-matching/prediction task for 36,884 test images. For participants in this Grand Challenge task, the majority of participants used CNNs and RNNs just as in VTT at TRECVid 2016, but unlike TRECVid, the ACM MM Grand Challenge setup does not give an opportunity for participants' results to be aggregated and disseminated at the ACM MM Conference, so it is difficult to gauge its overall impact in terms of comparison across participating systems, unlike in TRECvid where this happens at the workshop. We plan to continue working and improving the task next year with a similar setup and possible more human annotations.

10 Summing up and moving on

This overview to TRECVID 2016 has provided basic information on the goals, data, evaluation mechanisms, metrics used and high-level results analysis. Further details about each particular group's approach and performance for each task can be found in that group's site report. The raw results for each submitted run can be found at the online proceeding of the workshop [TV16Pubs, 2016].

11 Authors' note

TRECVID would not have happened in 2016 without support from the National Institute of Standards and Technology (NIST). The research community is very grateful for this. Beyond that, various individuals and groups deserve special thanks:

- Koichi Shinoda of the TokyoTech team agreed to host a copy of IACC.2 data.
- Georges Quénot provided the master shot reference for the IACC.3 videos.
- The LIMSI Spoken Language Processing Group and Vocapia Research provided ASR for the IACC.3 videos.
- Noel O'Connor and Kevin McGuinness at Dublin City University along with Robin Aly at the University of Twente worked with NIST and Andy O'Dwyer plus William Hayes at the BBC to make the BBC EastEnders video available for use in TRECVID.

- Maria Eskevich, Roeland Ordelman, Robin Aly, Gareth Jones, Benoit Huet, and Martha Larson at University of Twente, Radboud University, Dublin City University, EURECOM and Delft University of Technology for coordinating the Video hyperlinking task.
- Marc Ritter at TUC Chemnitz for supporting the Video to Text task annotations.

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

12 Acknowledgments

The video hyperlinking work has been partially supported by: ESF Research Networking Programme ELIAS (travel grants for Serwah Sabetghadam and Maria Eskevich); BpiFrance within the NexGenTV project, grant no. F1504054U; Science Foundation Ireland (SFI) as a part of the ADAPT Centre at DCU (13/RC/2106), the Dutch National Research Programme COMMIT/ and the CLARIAH (www.clariah.nl) project. The Video-to-Text work has been partially supported by Science Foundation Ireland (SFI) as a part of the Insight Centre at DCU (12/RC/2289). We would like to thank Ayana Yaegashi, and Serwah Sabetghadam for their work on anchor creation and crowdsourcing anchor verification and target vetting assessments.

References

[Awad et al., 2016] Awad, G., Snoek, C. G., Smeaton, A. F., and Quénot, G. (2016). Trecvid Semantic Indexing of Video: A 6-year retrospective. *ITE Transactions on Media Technology and Applications*, 4(3):187–208.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.

[Bernd et al., 2015] Bernd, J., Borth, D., Elizalde, B., Friedland, G., Gallagher, H., Gottlieb, L. R., Janin, A., Karabashlieva, S., Takahashi, J., and Won, J. (2015). The YLI-MED Corpus: Characteristics, Procedures, and Plans. *CoRR*, abs/1503.04250.

[Eskevich et al., 2017] Eskevich, M., Larson, M., Aly, R., Sabetghadam, S., Jones, G. J. F., Ordelman, R., and Huet, B. (2017). Multimodal Video-to-Video Linking: Turning to the Crowd for Insight and Evaluation. In *23rd International Conference on Multimedia Modeling(MMM)*, Reykjavik, Iceland.

[Han et al., 2013] Han, L., Kashyap, A., Finin, T., Mayfield, J., and Weese, J. (2013). UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 44–52.

[Kelm et al., 2009] Kelm, P., Schmiedeke, S., and Sikora, T. (2009). Feature-based Video Key Frame Extraction for Low Quality Video Sequences. In *10th Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2009, London, United Kingdom, May 6-8, 2009*, pages 25–28.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114.

[Lamel, 2012] Lamel, L. (2012). Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data. In Tavast, A., Muischnek, K., and Koit, M., editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012*, volume 247 of *Frontiers in Artificial Intelligence and Applications*, pages 1–8. IOS Press.

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

[Over et al., 2015] Over, P., Fiscus, J., Joy, D., Michel, M., Awad, G., Kraaij, W., Smeaton, A. F., Quénot, G., and Ordelman, R. (2015). TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2015*. NIST, USA.

[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.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[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.

[Schmiedeke et al., 2013] Schmiedeke, S., Xu, P., Ferrané, I., Eskevich, M., Kofler, C., Larson, M. A., Estève, Y., Lamel, L., Jones, G. J. F., and Sikora, T. (2013). Blip10000: A Social video Dataset containing SPUG content for Tagging and Retrieval. In *Multimedia Systems Conference 2013 (MMSys '13)*, pages 96–101, Oslo, Norway.

[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).

[Thomee et al., 2016] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). YFCC100M: The New Data in Multimedia Research. *Commun. ACM*, 59(2):64–73.

[TV16Pubs, 2016] TV16Pubs (2016). http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.16.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/.

[Vedantam et al., 2015] Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

[Xu et al., 2016] Xu, J., Mei, T., Yao, T., and Rui, Y. (2016). MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[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.

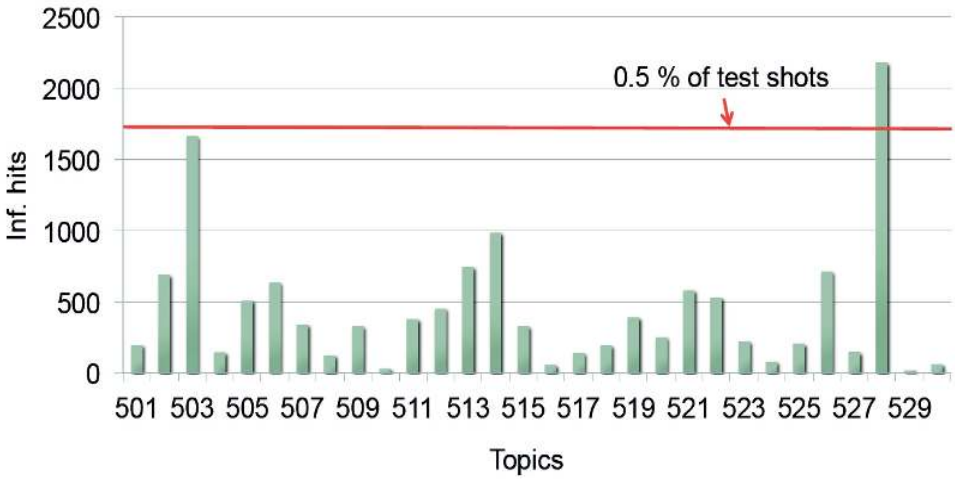

Figure 1: AVS: Histogram of shot frequencies by query number

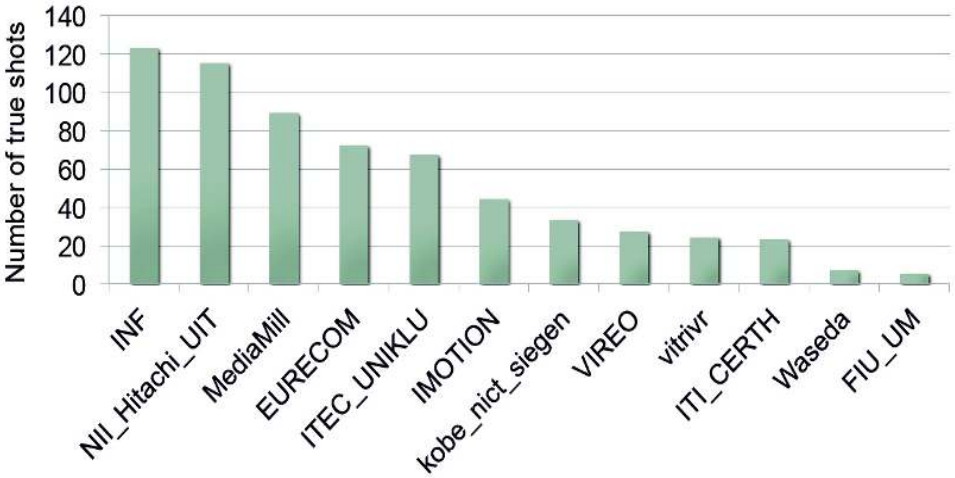

Figure 2: AVS: Unique shots contributed by team

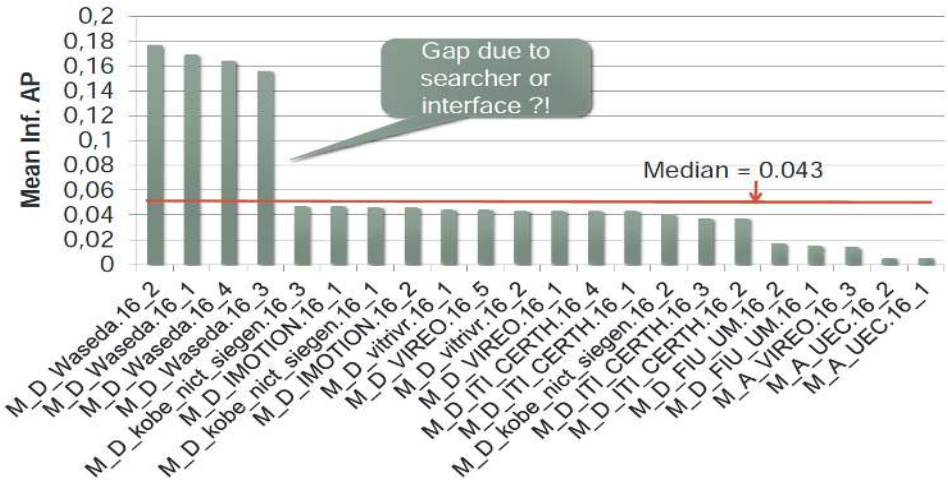

Figure 3: AVS: xinfAP by run (manually assisted)

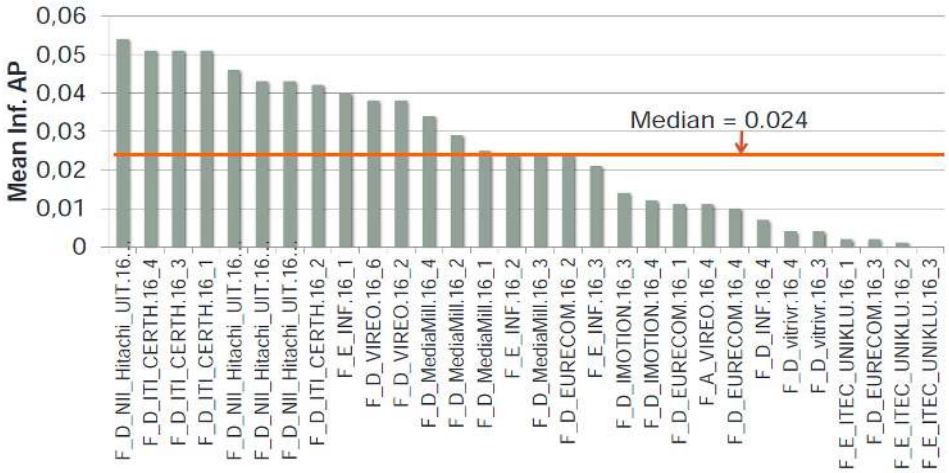

Figure 4: AVS: xinfAP by run (fully automatic)

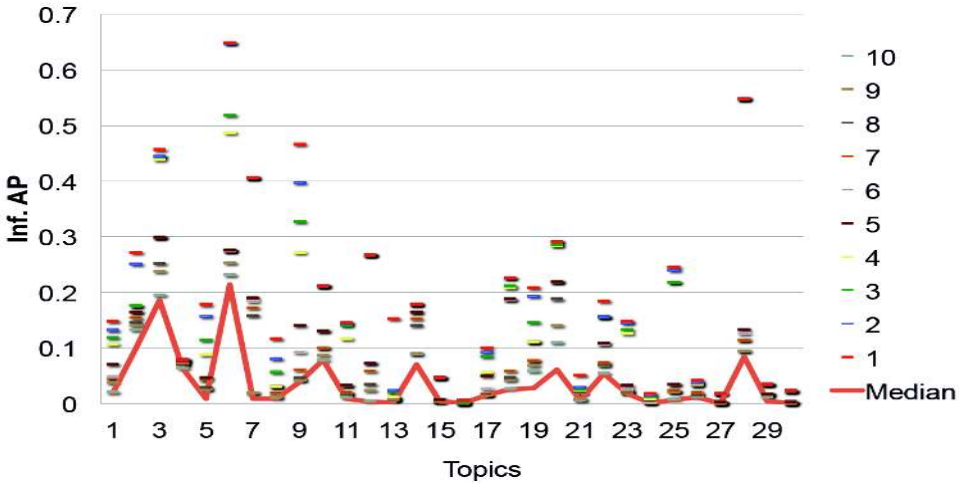

Figure 5: AVS: Top 10 runs (xinfAP) by query number (manually assisted)

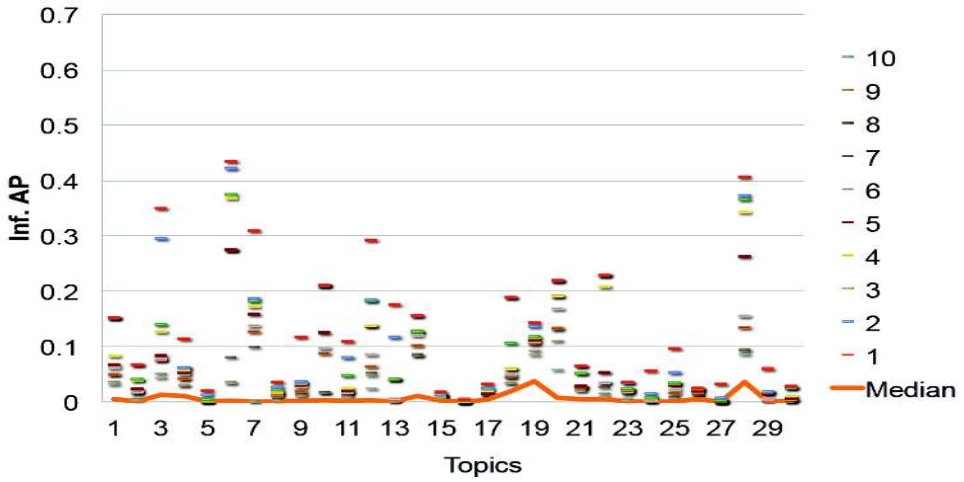

Figure 6: AVS: Top 10 runs (xinfAP) by query number (fully automatic)

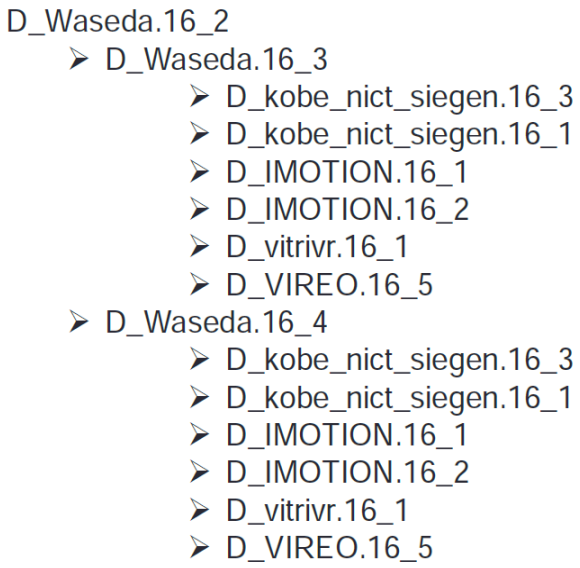

Figure 7: AVS: Statistical significant differences (top 10 manually-assisted runs)

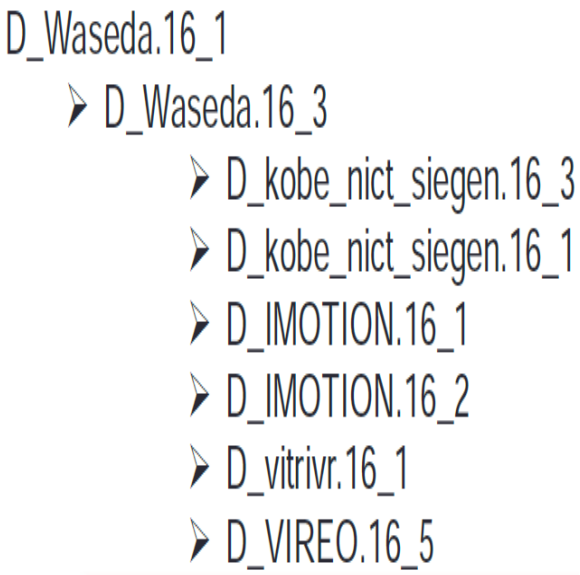

Figure 8: AVS: Statistical significant differences (top 10 manually-assisted runs)

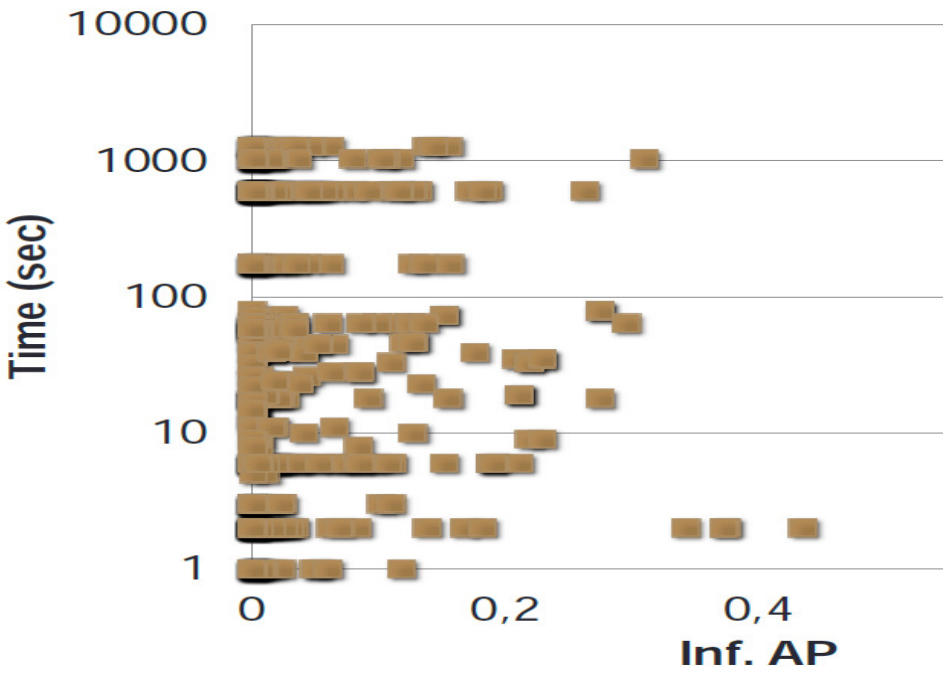

Figure 9: AVS: Processing time vs Scores (fully automatic)

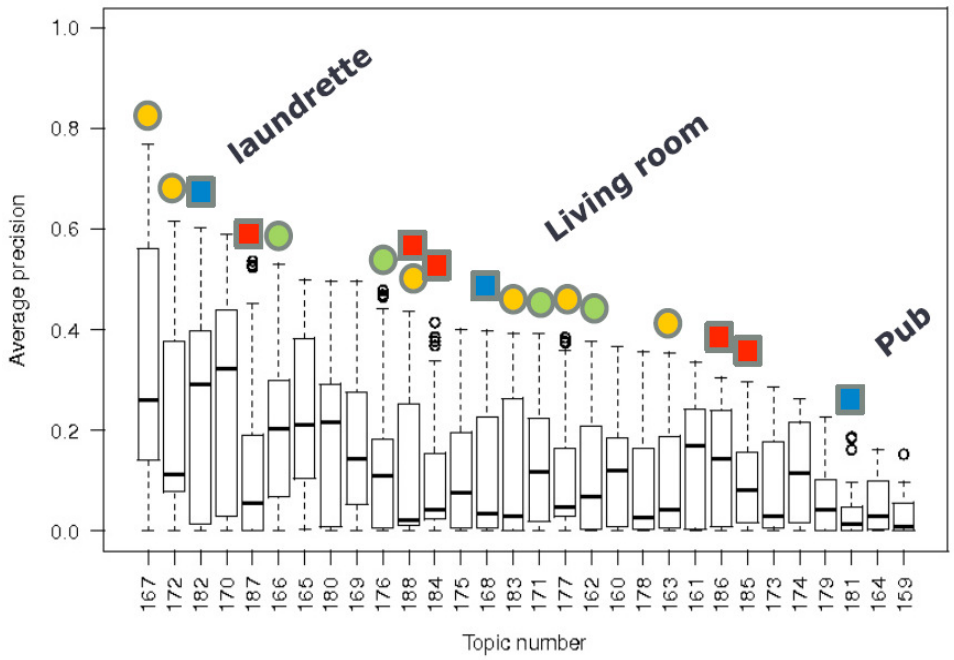

Figure 10: INS: Boxplot of average precision by topic for automatic runs. Yellow circle: living room, red square: Pat, green circle: foyer, blue square: pub

```
0.370  F_E_PKU_ICST_1          =  >  >  >  >  >  >  >  >  >
0.364  F_E_PKU_ICST_3             =     >  >  >  >  >  >  >
0.349  F_E_PKU_ICST_5                =     >  >  >  >  >  >
0.335  F_A_PKU_ICST_4                   =  >  >  >  >  >  >
0.328  F_A_PKU_ICST_6                      =     >  >  >  >
0.317  F_A_PKU_ICST_7                            =  >  >  >  >
0.244  F_A_NII_Hitachi_UIT_1                        =        >
0.230  F_A_NII_Hitachi_UIT_4                           =
0.230  F_A_BUPT_MCPRL_3                                    =
0.229  F_A_NII_Hitachi_UIT_2                                    =
                              1  2  3  4  5  6  7  8  9  10
```

Figure 11: INS: Randomization test results for top automatic runs. "E":runs used video examples. "A":runs used image examples only.

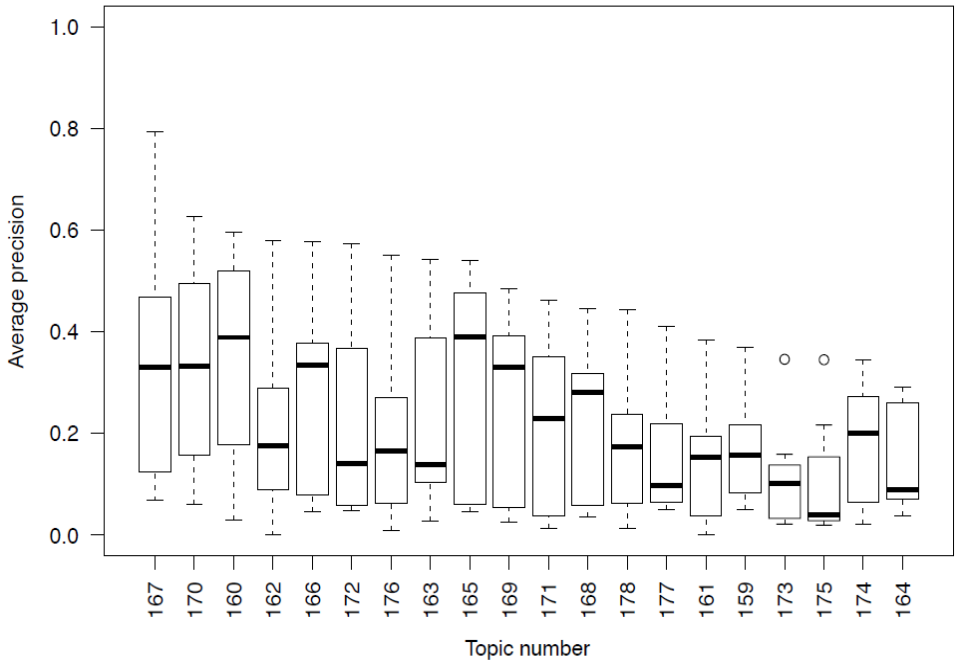

Figure 12: INS: Boxplot of average precision by topic for interactive runs

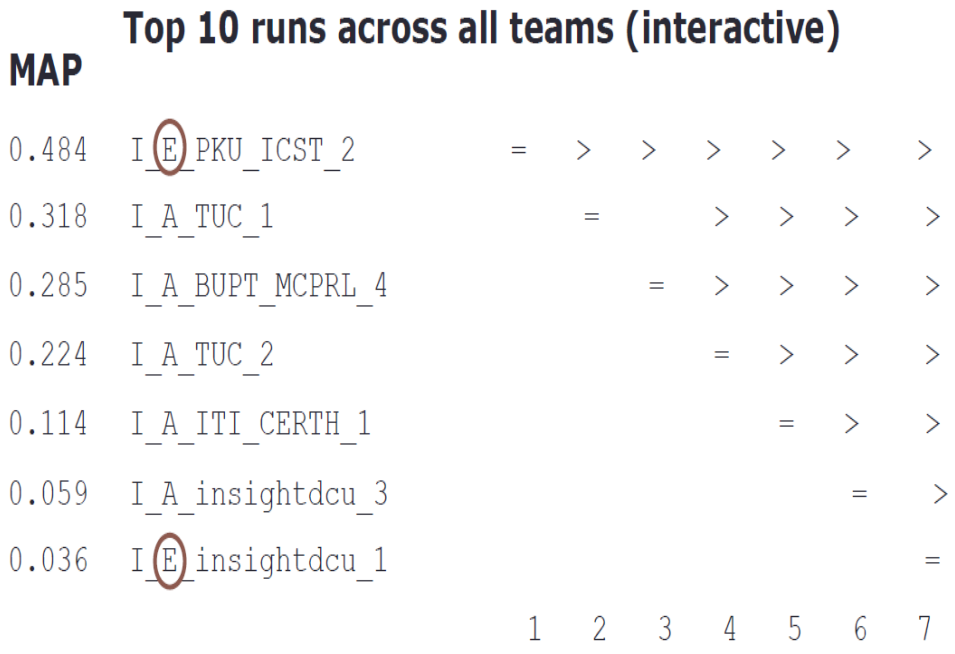

Figure 13: INS: Randomization test results for top interactive runs. "E":runs used video examples. "A":runs used image examples only.

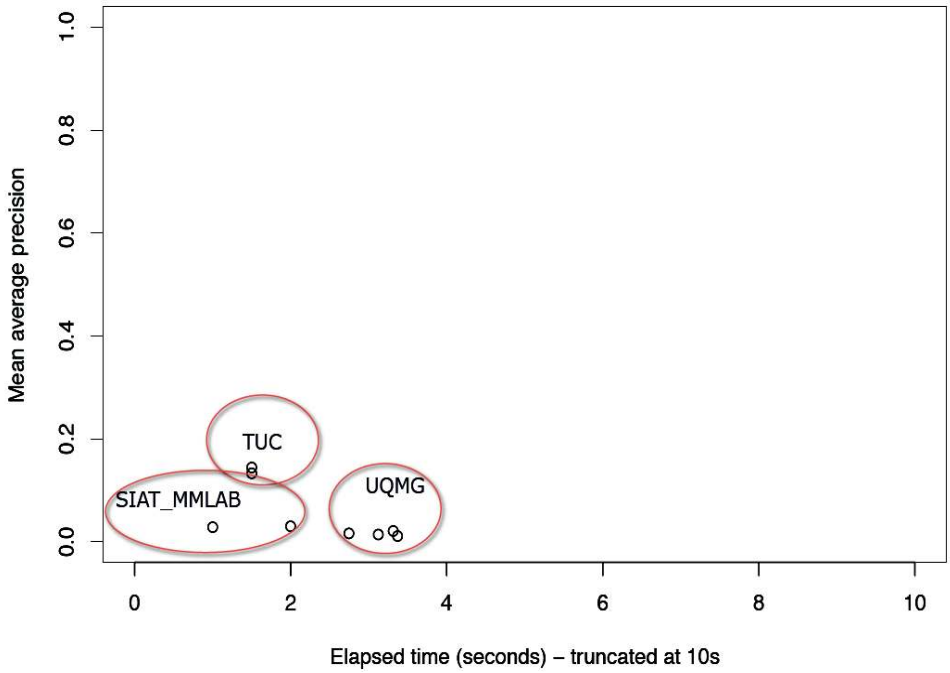

Figure 14: INS: Mean average precision versus time for fastest runs

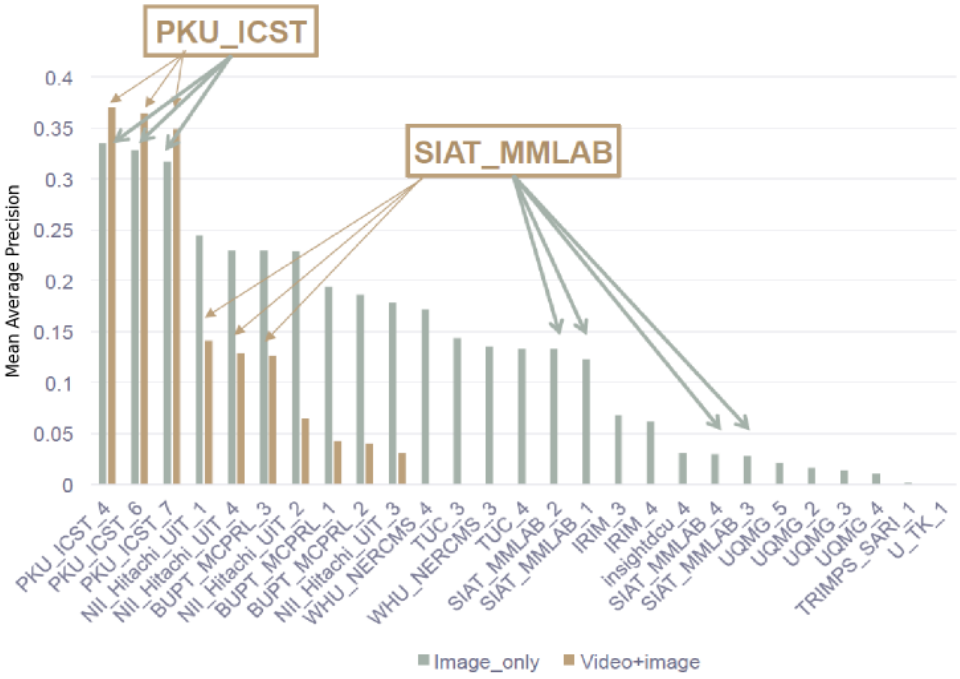

Figure 15: INS: Effect of number of topic example images used

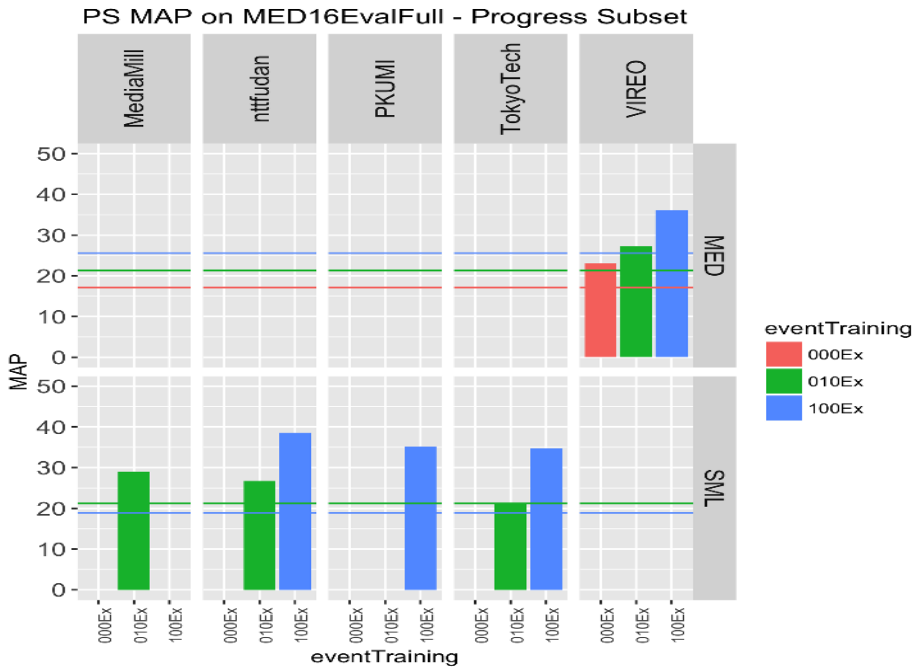

Figure 16: MED: MAP scores on the progress subset of MED16-EvalFull. Lines represent last year's high scores for the given evaluation condition

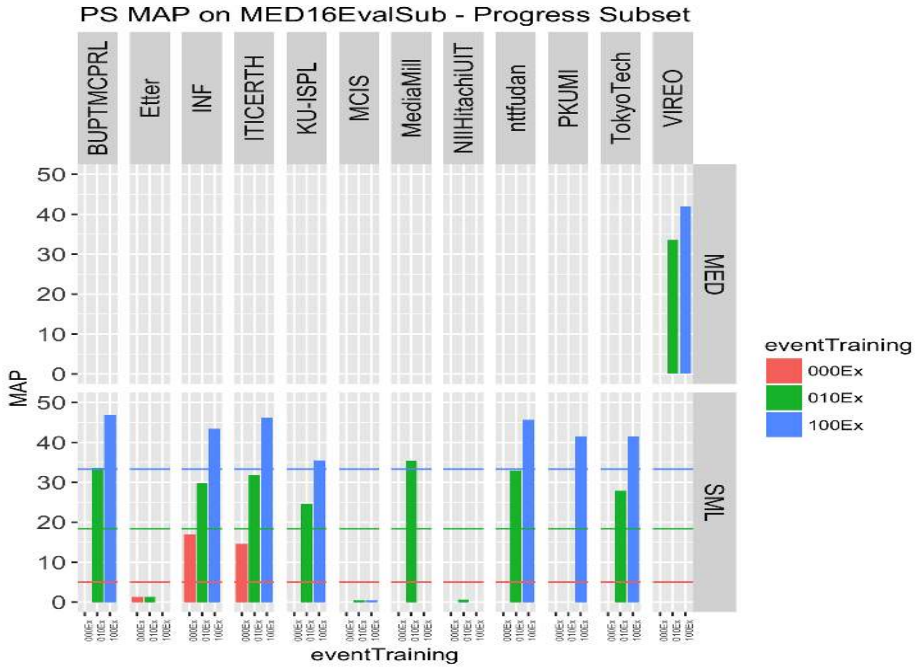

Figure 17: MED: MAP scores on the progress subset of MED16-EvalSub. Lines represent last year's high scores for the given evaluation condition

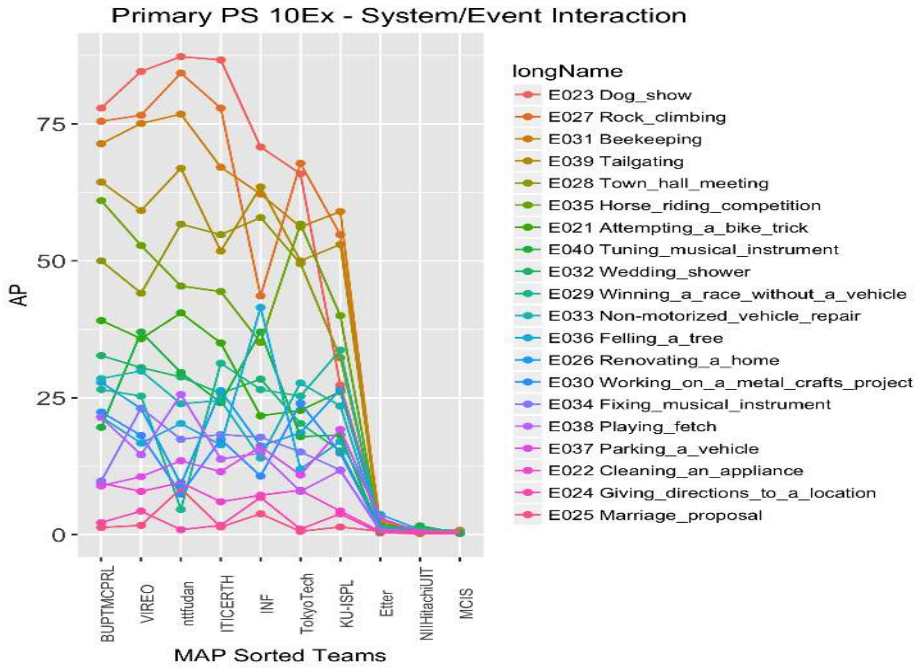

Figure 18: MED: Pre-Specified events vs. systems (progress subset of MED16-EvalSub; 10Ex)

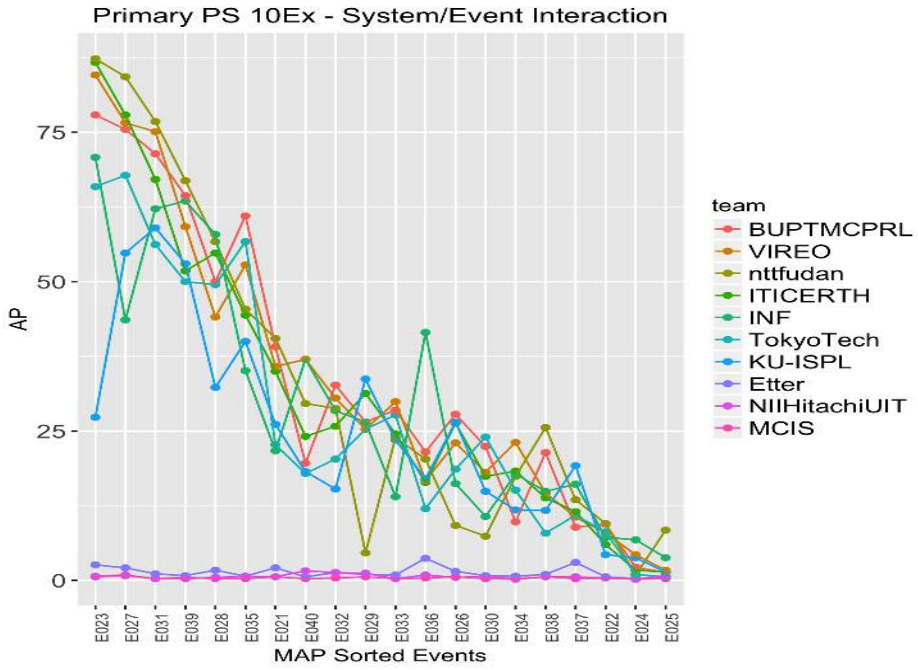

Figure 19: MED: Systems vs. Pre-Specified events (progress subset of MED16-EvalSub; 10Ex)

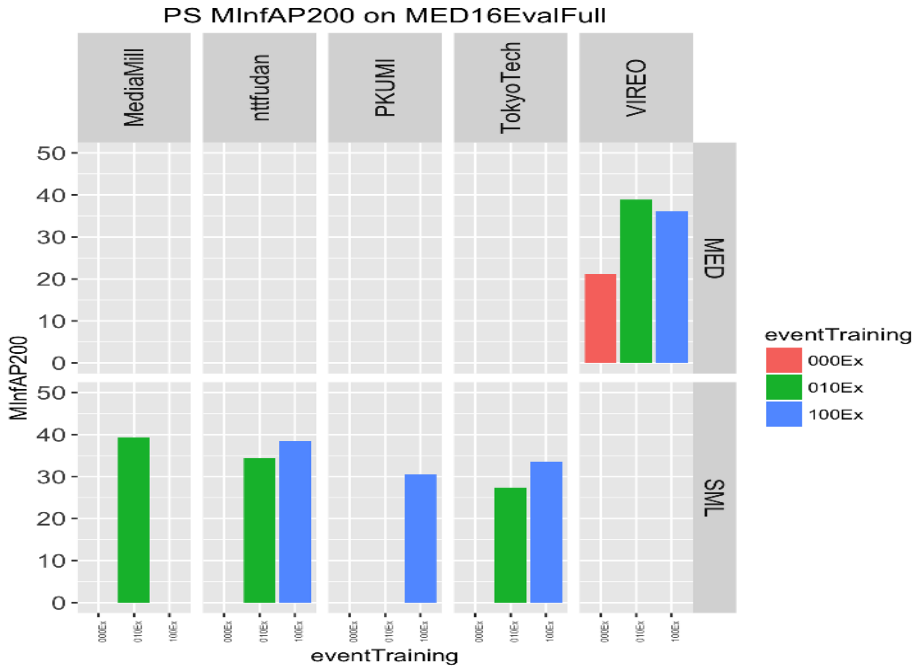

Figure 20: MED: MInfAP200 scores on MED16-EvalFull for Pre-Specified events

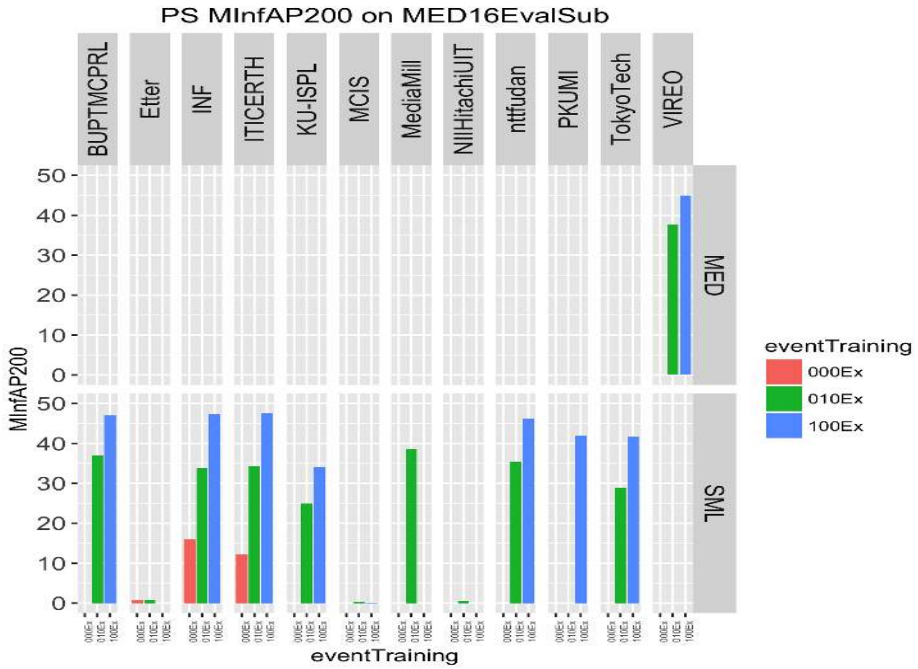

Figure 21: MED: MInfAP200 scores on MED16-EvalSub for Pre-Specified events

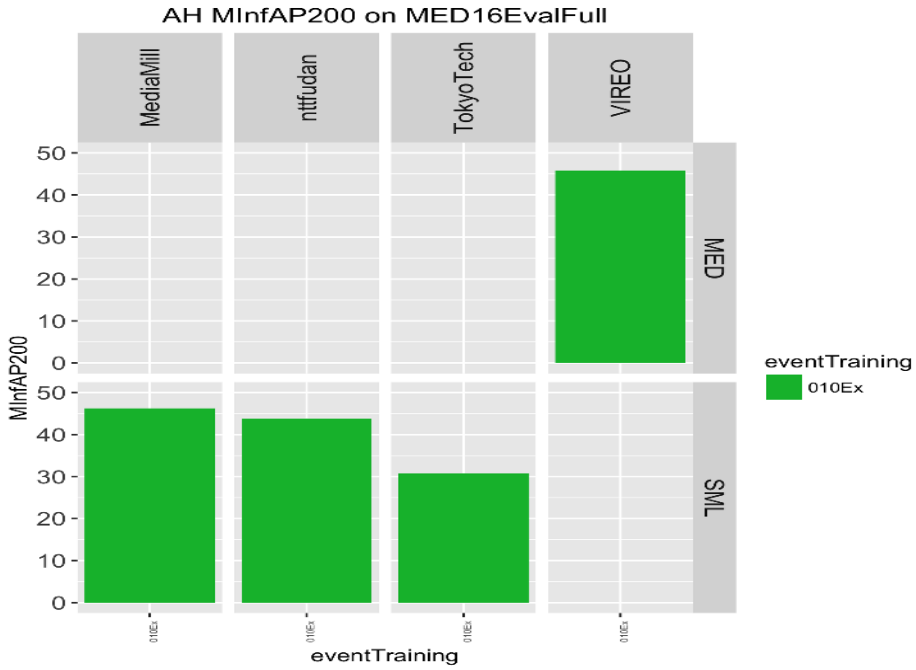

Figure 22: MED: MInfAP200 scores on MED16-EvalFull for ad-Hoc events

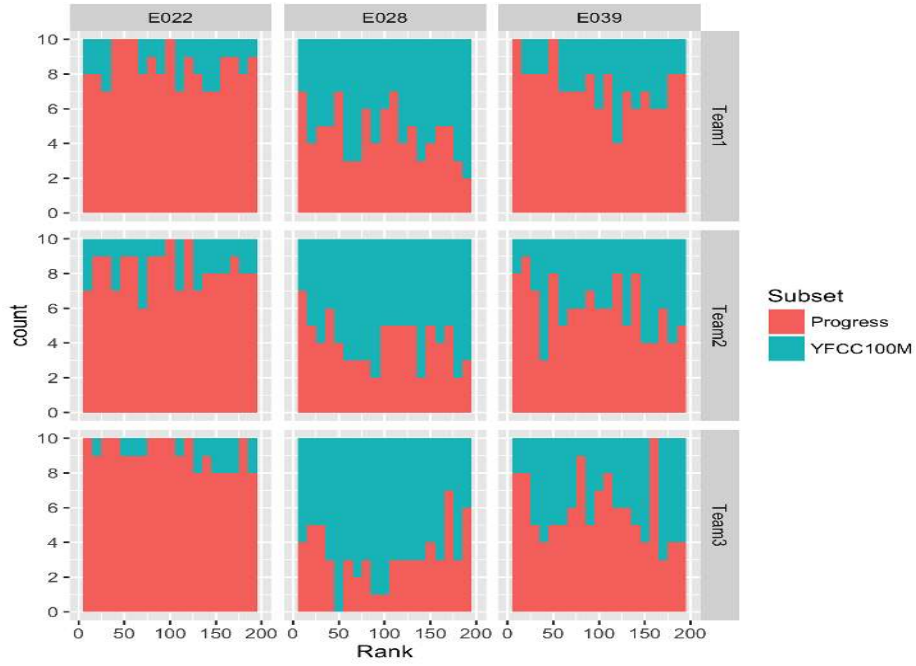

Figure 23: MED: Top 200 ranked clips by dataset for a sample of systems and events

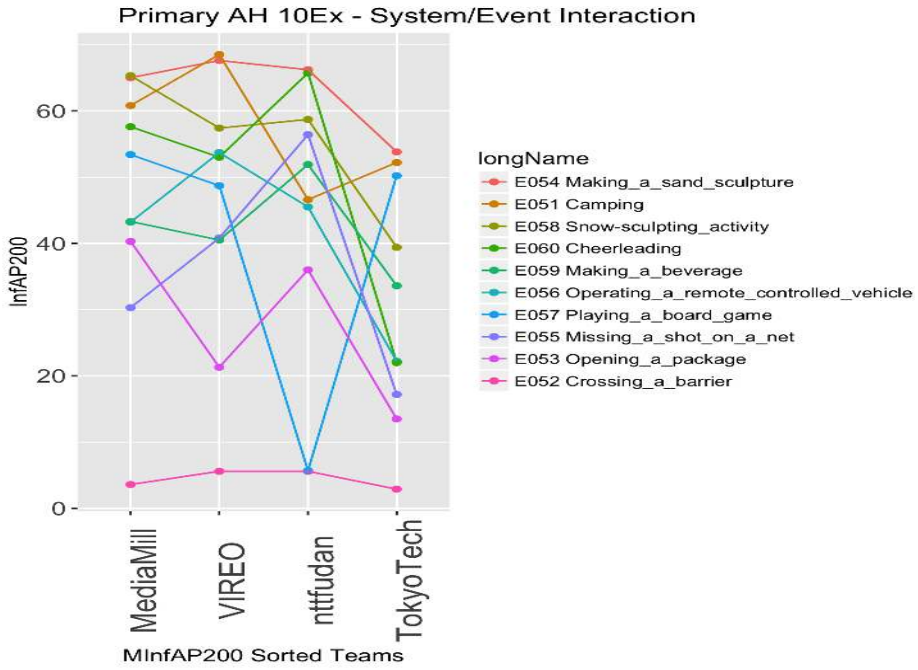

Figure 24: MED: Ad-Hoc events vs. systems (MED16-EvalFull; 10Ex)

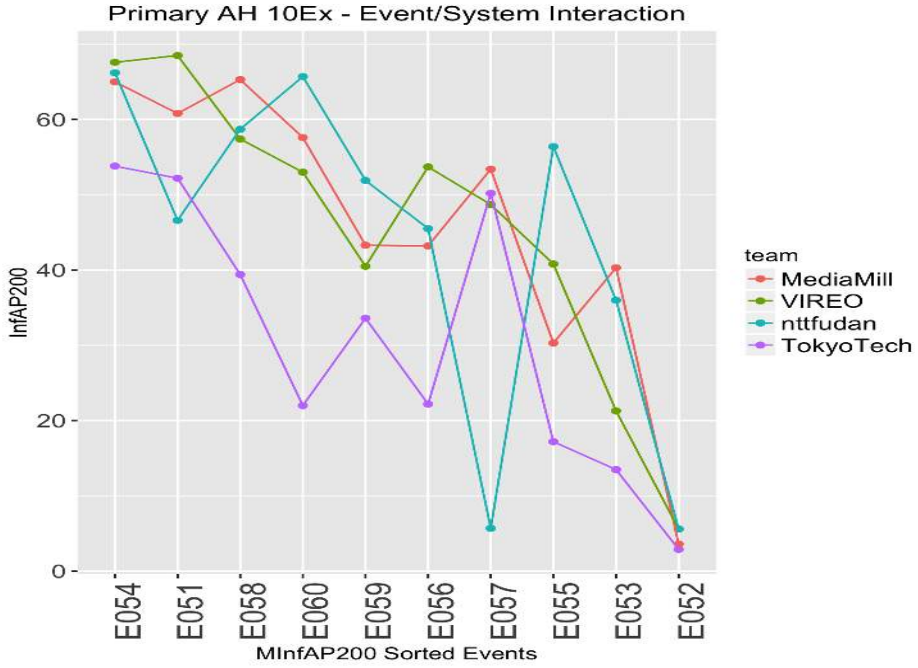

Figure 25: MED: Systems vs. ad-Hoc events (MED16-EvalFull; 10Ex)

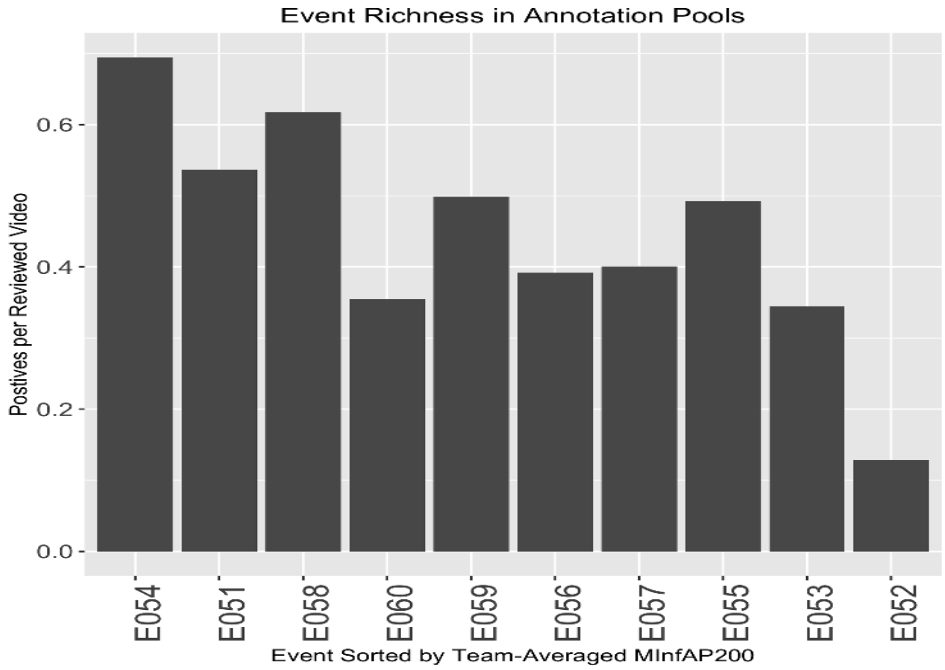

Figure 26: MED: Event richness

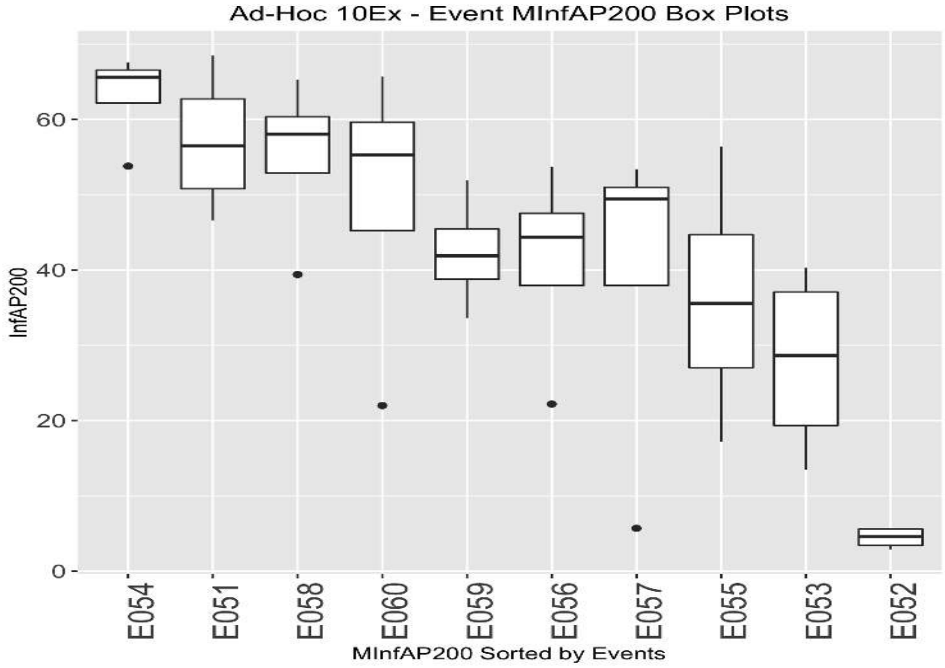

Figure 27: MED: Event richness

		EVAL16							SUB16		
		Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing	CellToEar	Embrace	PeopleMeet	PeopleSplitUp
BUPT-MCPRL (8 years)	Beijing University of Posts and Telecommunications (China)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing				
INF (9 years)	Informedia Team (Carnegie Mellon University, University of Technology Sydney, Renmin University of China, Shandong University) (USA)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing	CellToEar			
HRI (1 year)	Hikvision Research Institute (HRI) (China)	Embrace				PersonRuns	Pointing		Embrace		
IIPWHU (2 years)	Wuhan University - Intelligent Information Processing (China)			PeopleMeet	PeopleSplitUp					PeopleMeet	PeopleSplitUp
ITI-CERTH (2 years)	Information Technologies Institute, Centre for Research and Technology Hellas (Greece)	Embrace		PeopleMeet	PeopleSplitUp	PersonRuns	Pointing				
NII-Hitachi-UIT (1 year)	National Institute of Informatics, Japan (NII); Hitachi, Ltd; University of Information Technology, VNU-HCM, Vietnam (HCM-UIT) (Japan, Vietnam)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing	CellToEar			
SeuGraph (2 years)	Computer Graphics Lab of Southeast University, Southeast University Jiulonghu Campus (China)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing	CellToEar			
WARD (2 years)	ITEE, The University of Queensland (Australia)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing	CellToEar			
8 Participants:	China: 4, Australia: 1, Greece: 1, Japan: 1, USA: 1, Vietnam: 1	7	5	7	7	7	7	4	1	1	1

Figure 28: SED16 Participants

	EVAL16		
	Lowest Actual NDCR	**Lowest Min NDCR**	**Lowest NDCR@TOER**
CellToEar	**CMUTEAM**	**4 way tie: CMUTEAM, NII-Hitachi-UIT, SeuGraph, WARD**	**NII-Hitachi-UIT**
Embrace	**BUPT-MCPRL**	**BUPT-MCPRL**	**HRI**
ObjectPut	**BUPT-MCPRL**	**BUPT-MCPRL**	**WARD**
PeopleMeet	**BUPT-MCPRL**	**BUPT-MCPRL**	**WARD**
PeopleSplitUp	**BUPT-MCPRL**	**CMUTEAM**	**WARD**
PersonRuns	**BUPT-MCPRL**	**BUPT-MCPRL**	**WARD**
Pointing	**BUPT-MCPRL**	**BUPT-MCPRL**	**HRI**

Figure 29: SED16: Systems with the lowest NDCR, per event, per metric

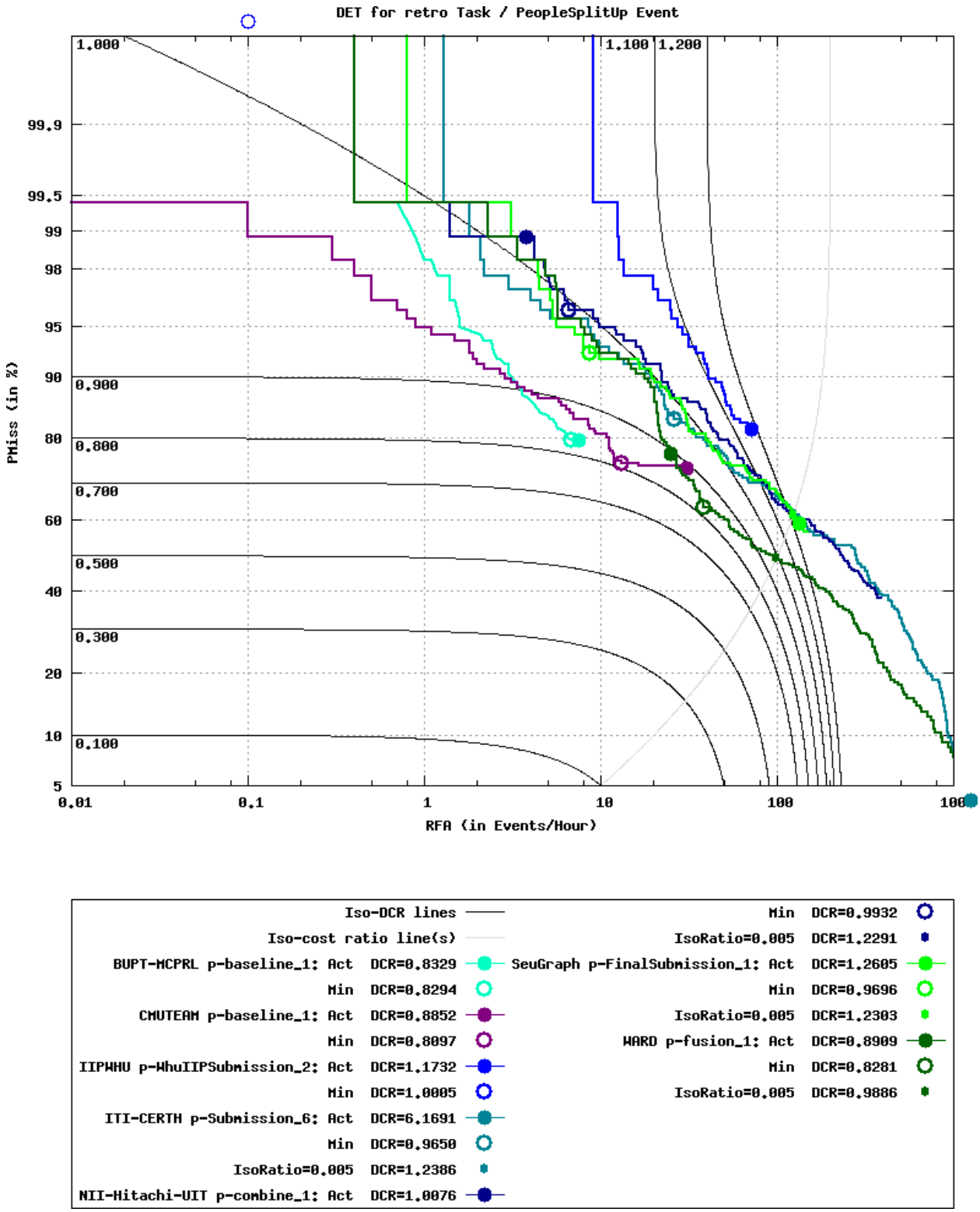

Figure 30: SED16: PeopleSplitUp

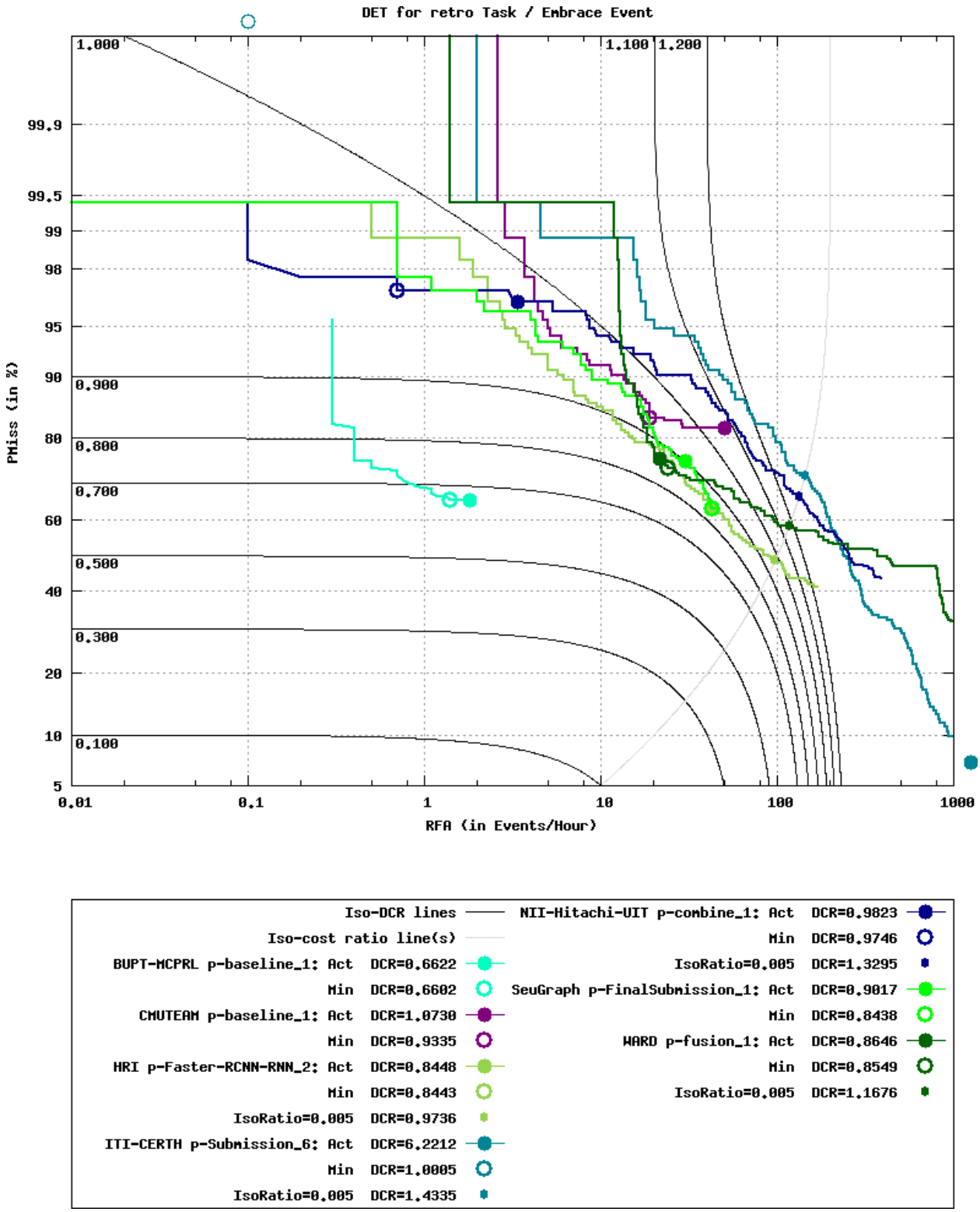

Figure 31: SED16: Embrace

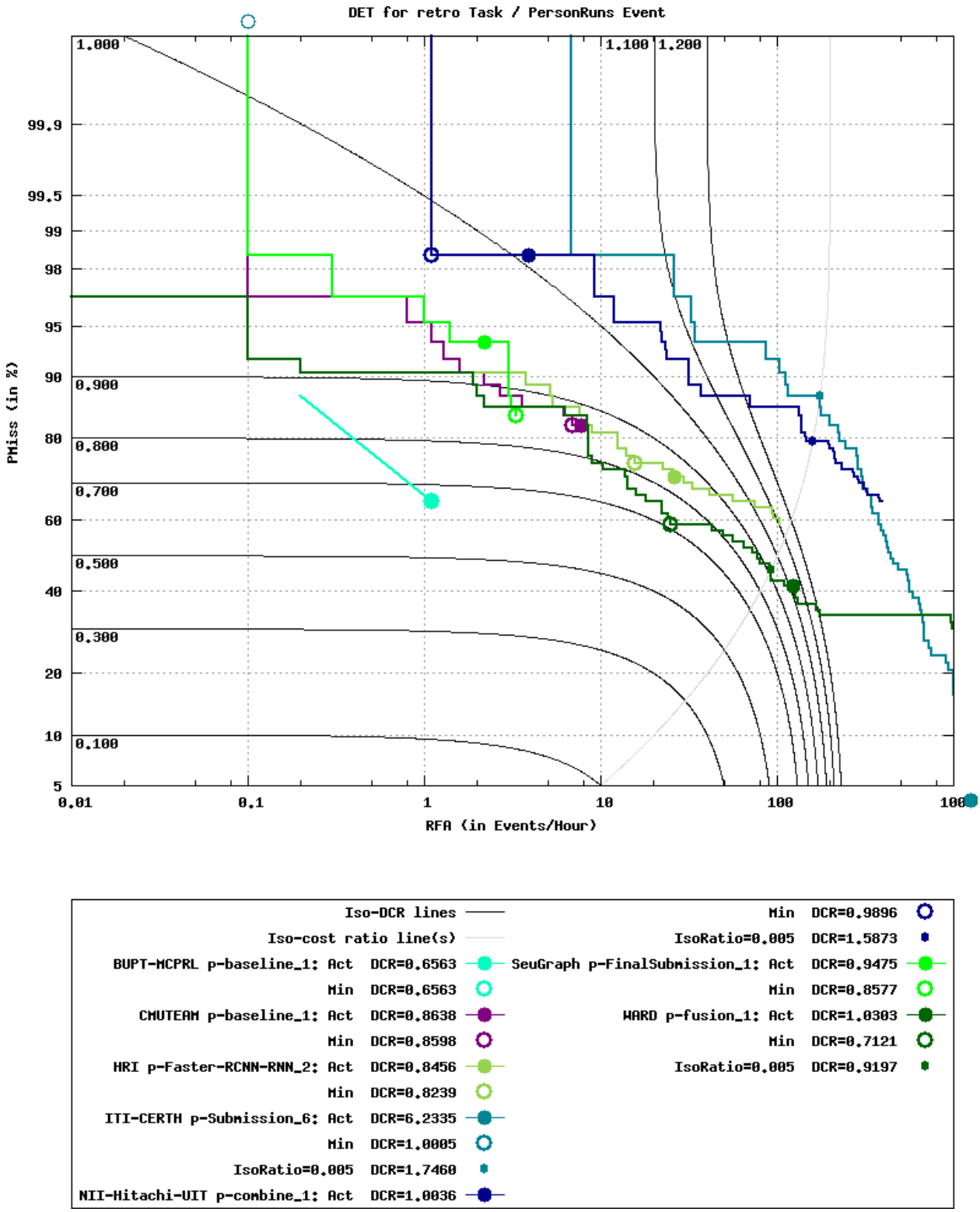

Figure 32: SED16: PersonRuns

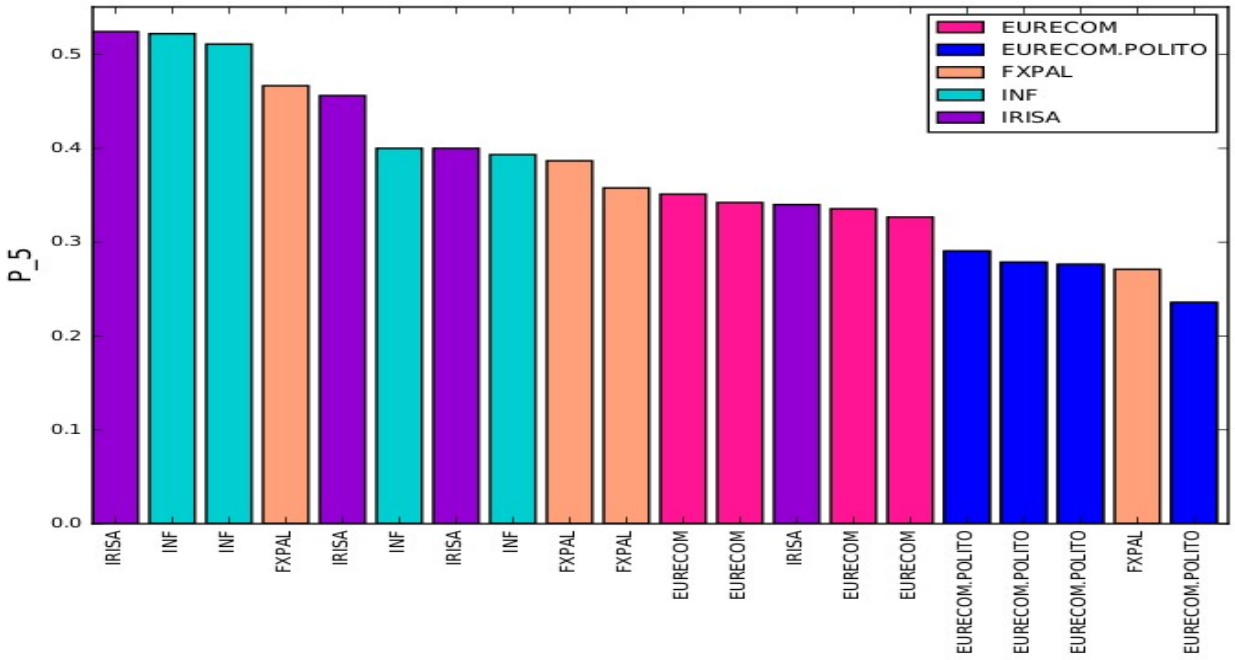

Figure 33: LNK16: Precision at rank 5

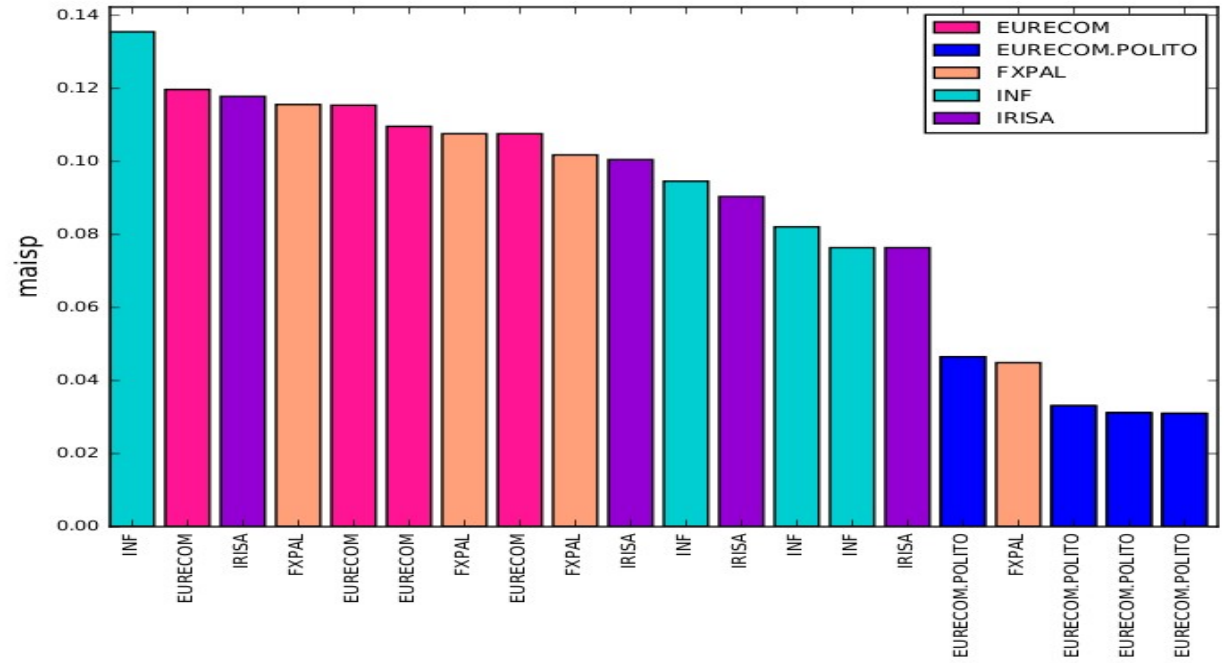

Figure 34: LNK16: Mean Average interpolated Segment Precision (MAiSP)

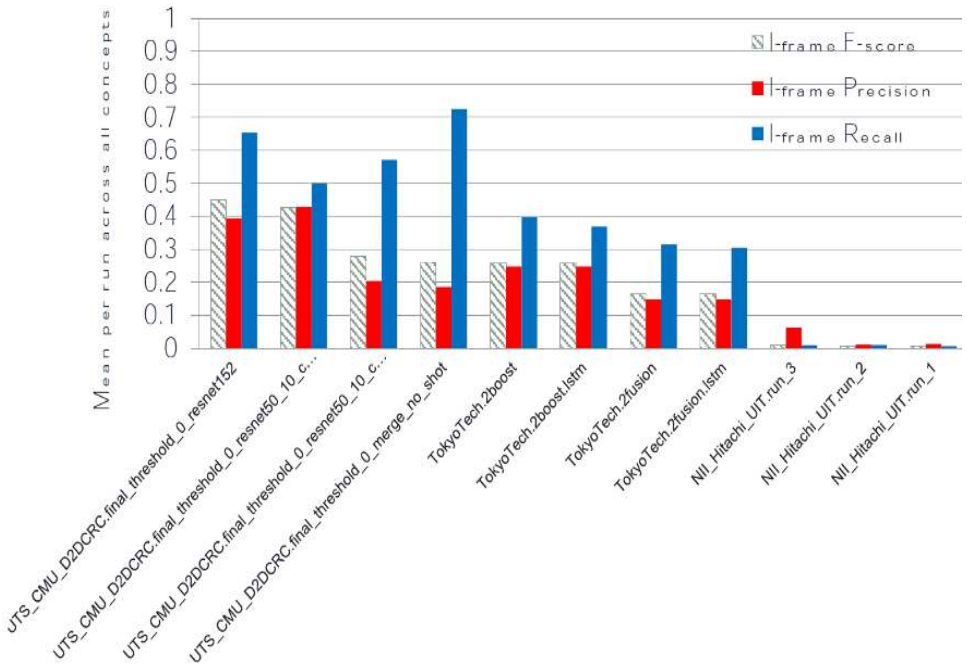

Figure 35: LOC: Temporal localization results by run

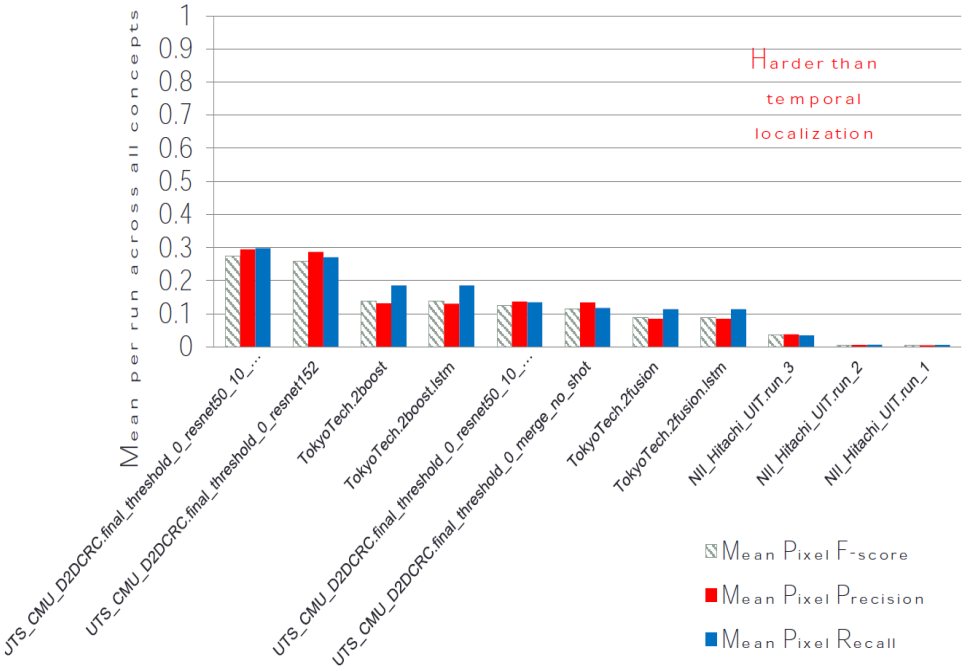

Figure 36: LOC: Spatial localization results by run

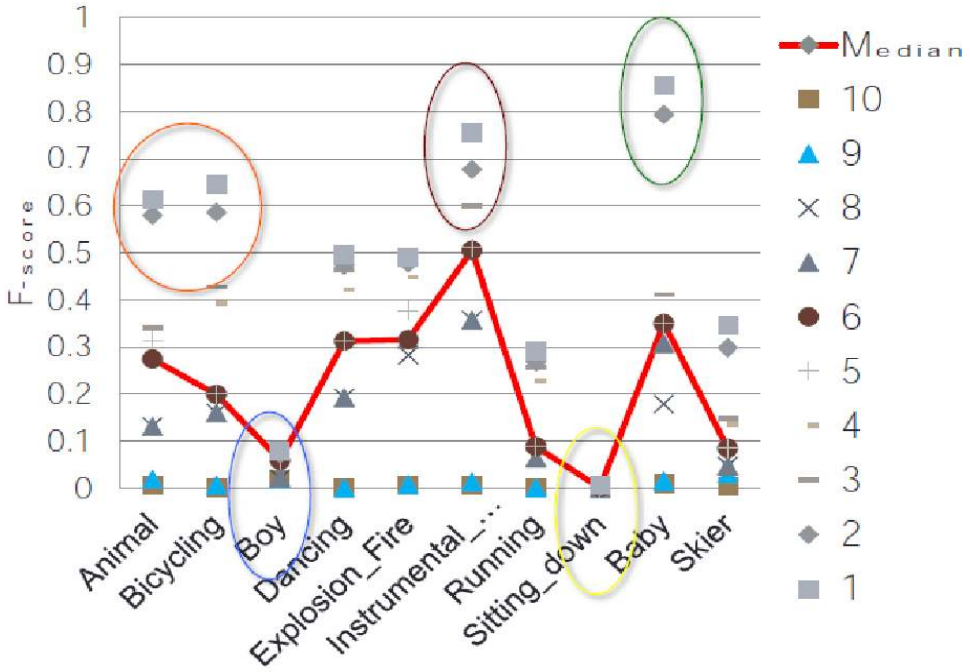

Figure 37: LOC: Temporal localization by concept

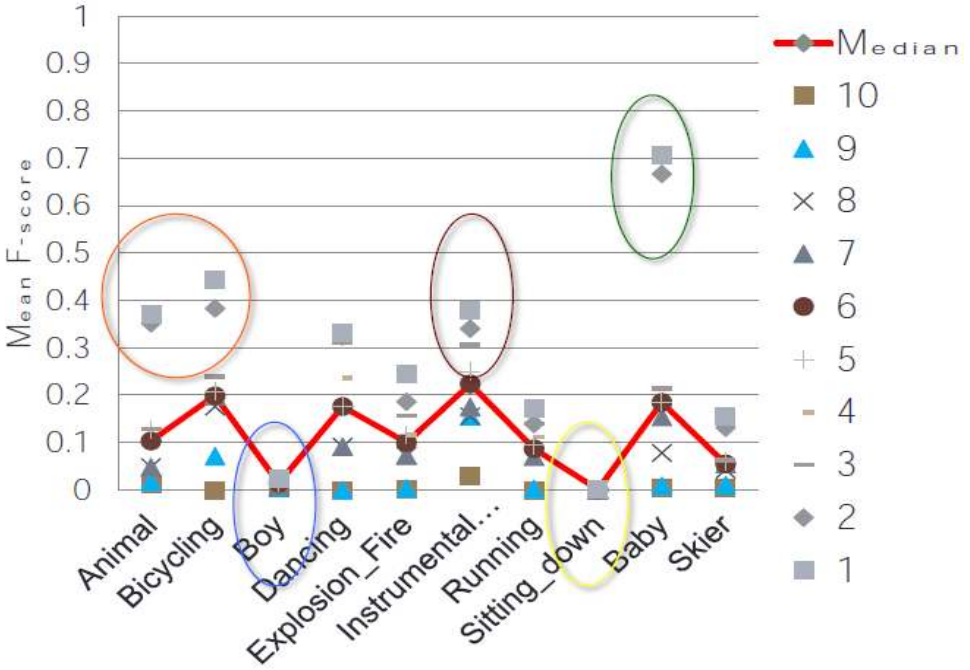

Figure 38: LOC: Spatial localization by concept

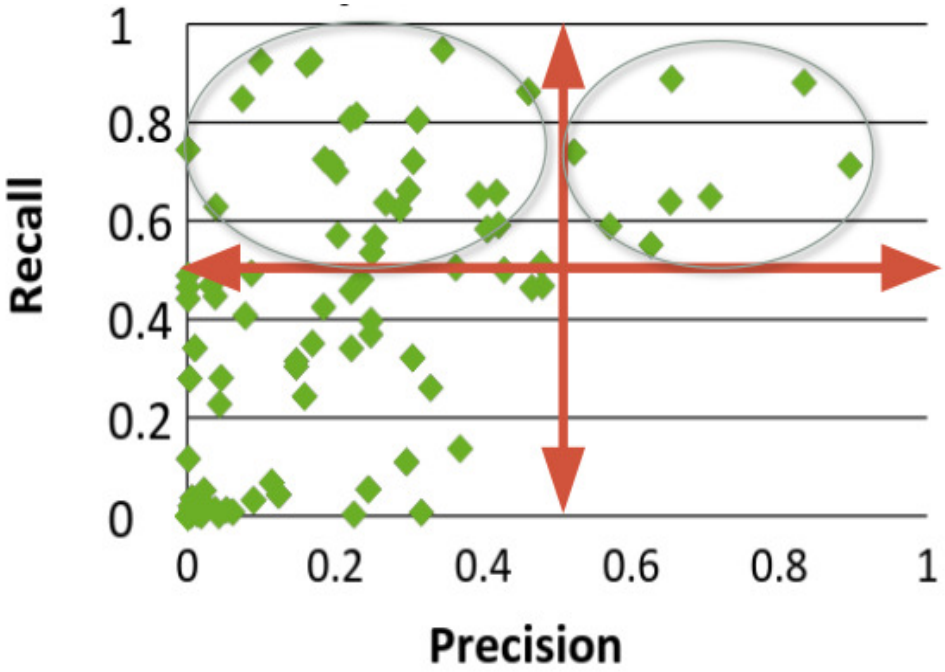

Figure 39: LOC: Temporal precision and recall per concept for all teams

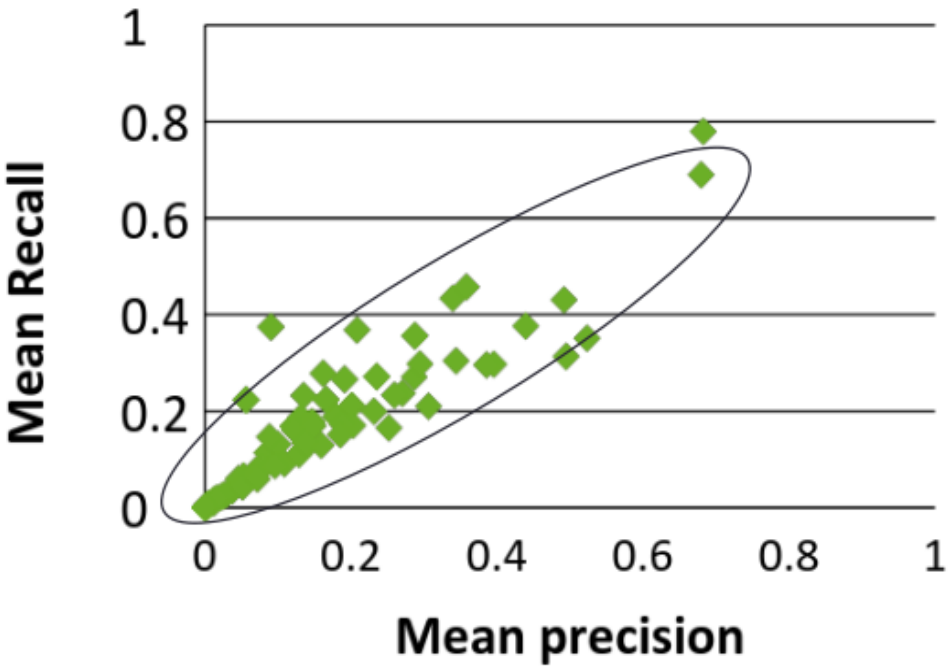

Figure 40: LOC: Spatial precision and recall per concept for all teams

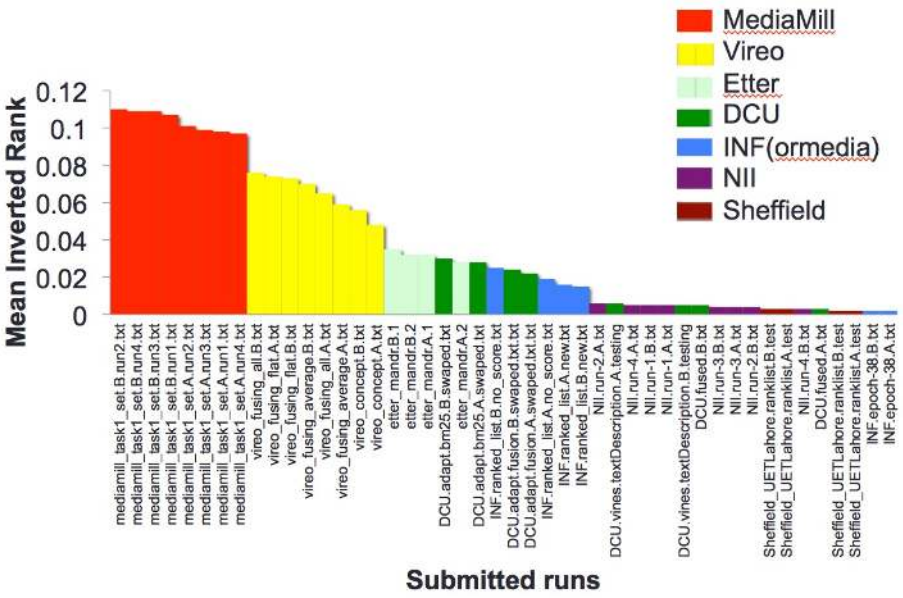

Figure 41: VTT: Matching and Ranking results across all runs

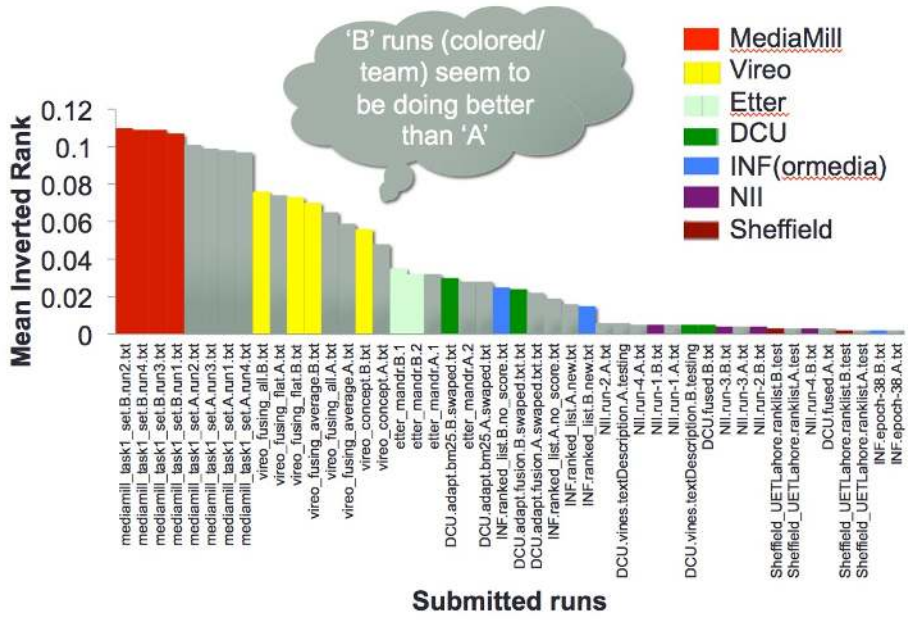

Figure 42: VTT: Matching and Ranking results across all runs

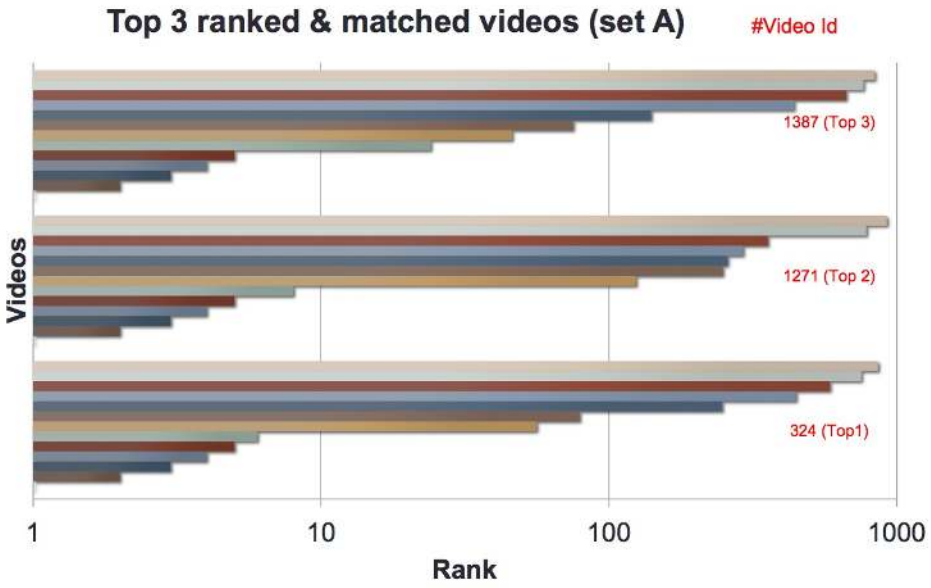

Figure 43: VTT: Top 3 matched videos from set A

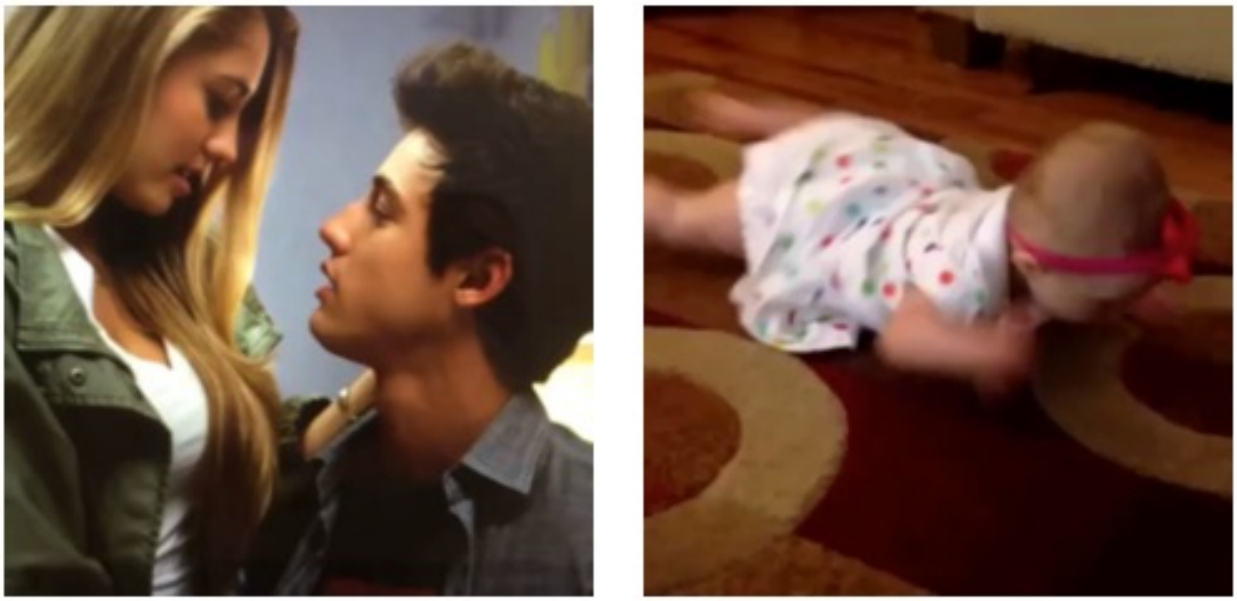

Figure 44: VTT: Samples of easy videos

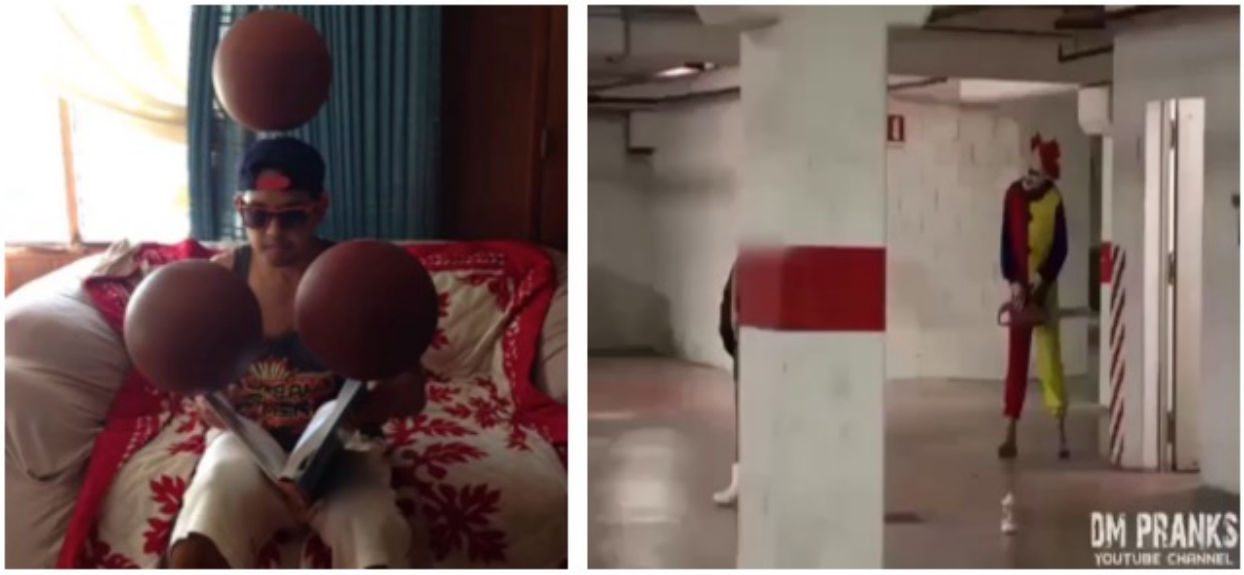

Figure 45: VTT: Samples of hard videos

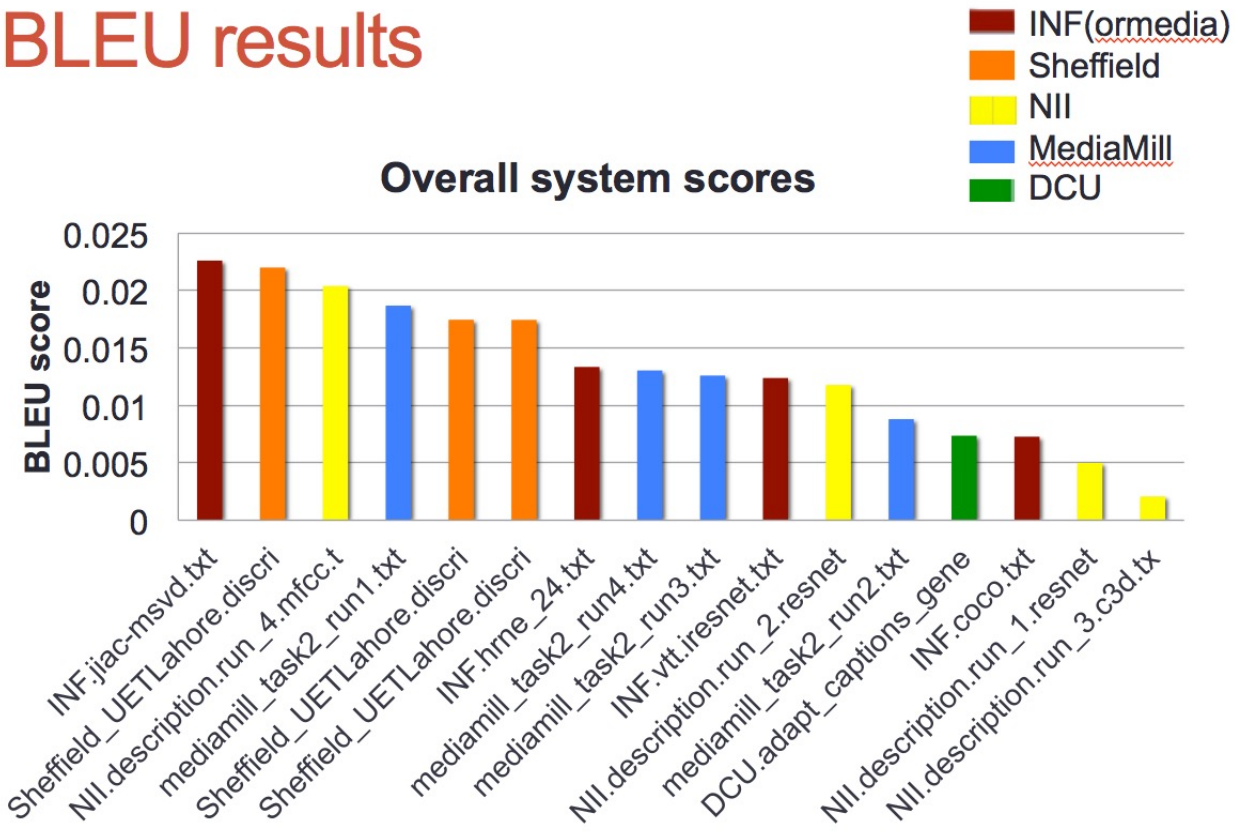

Figure 46: VTT: Results using the BLEU metric

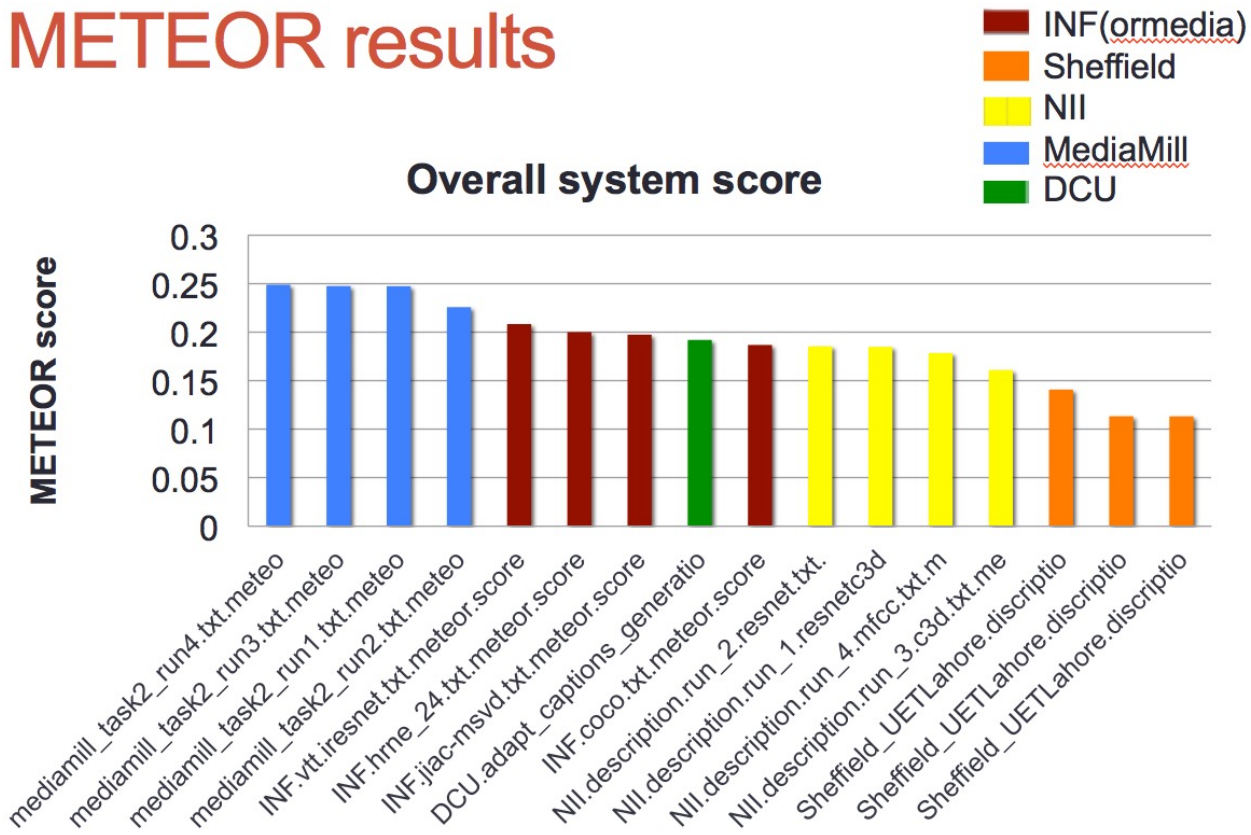

Figure 47: VTT: Results using the METEOR metric

Figure 48: VTT: Sample frame from a video

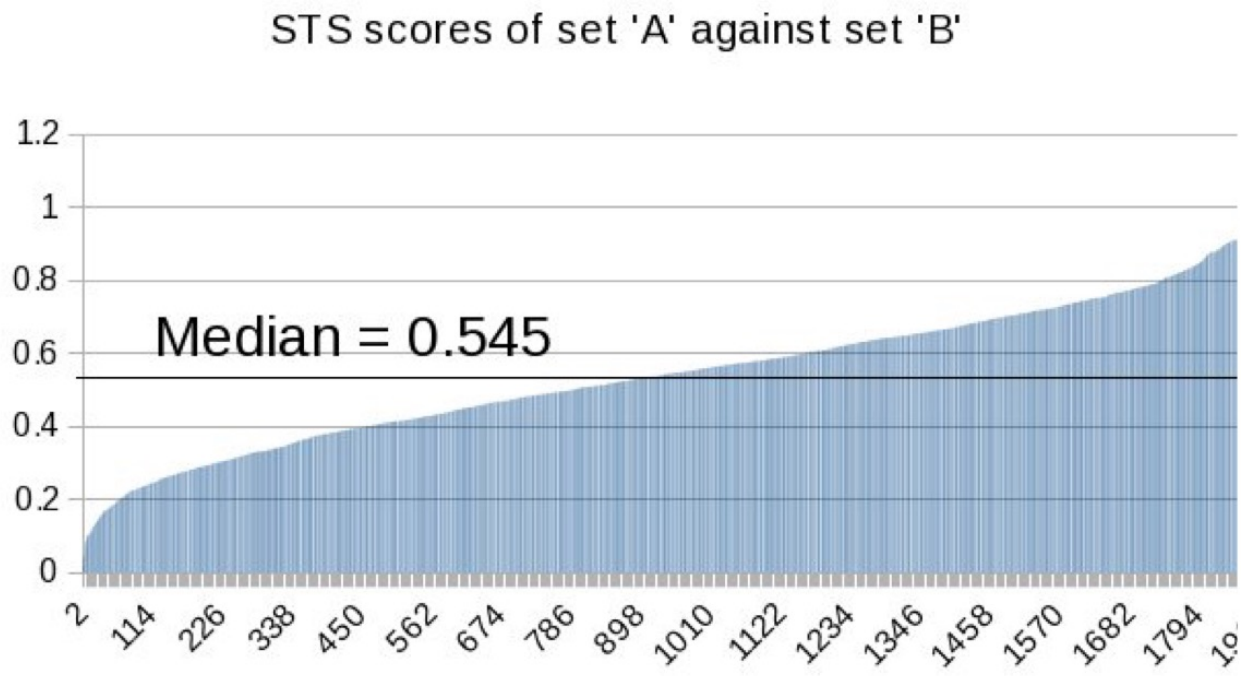

Figure 49: VTT: STS scores against the two reference ground truth sets

13 Appendix A: Ad-hoc query topics

501 Find shots of a person playing guitar outdoors
502 Find shots of a man indoors looking at camera where a bookcase is behind him
503 Find shots of a person playing drums indoors
504 Find shots of a diver wearing diving suit and swimming under water
505 Find shots of a person holding a poster on the street at daytime
506 Find shots of the 43rd president George W. Bush sitting down talking with people indoors
507 Find shots of a choir or orchestra and conductor performing on stage
508 Find shots of one or more people walking or bicycling on a bridge during daytime
510 Find shots of a sewing machine
511 Find shots of destroyed buildings
512 Find shots of palm trees
514 Find shots of soldiers performing training or other military maneuvers
515 Find shots of a person jumping
516 Find shots of a man shake hands with a woman
517 Find shots of a policeman where a police car is visible
518 Find shots of one or more people at train station platform
519 Find shots of two or more men at a beach scene
520 Find shots of any type of fountains outdoors
521 Find shots of a man with beard talking or singing into a microphone
522 Find shots of a person sitting down with a laptop visible
523 Find shots of one or more people opening a door and exiting through it
525 Find shots of a person holding a knife
526 Find shots of a woman wearing glasses
527 Find shots of a person drinking from a cup, mug, bottle, or other container
528 Find shots of a person wearing a helmet
529 Find shots of a person lighting a candle
530 Find shots of people shopping

14 Appendix B: Instance search topics

9159 "Find Jim in the Pub"
9160 "Find Jim in this Kitchen"
9161 "Find Jim in this Laundrette"
9162 "Find Jim at this Foyer"
9163 "Find Jim in this Living Room"
9164 "Find Dot in the Pub"
9165 "Find Dot in this Kitchen"
9166 "Find Dot at this Foyer"
9167 "Find Dot in this Living Room"
9168 "Find Brad in the Pub"
9169 "Find Brad in this Kitchen"
9170 "Find Brad in this Laundrette"
9171 "Find Brad at this Foyer"
9172 "Find Brad in this Living Room"
9173 "Find Stacey in the Pub"

9174 "Find Stacey in this Kitchen"
9175 "Find Stacey in this Laundrette"
9176 "Find Stacey at this Foyer"
9177 "Find Stacey in this Living Room"
9178 "Find Patrick in the Pub"
9179 "Find Patrick in this Kitchen"
9180 "Find Patrick in this Laundrette"
9181 "Find Fatboy in the Pub"
9182 "Find Fatboy in this Laundrette"
9183 "Find Fatboy in this Living Room"
9184 "Find Pat in the Pub"
9185 "Find Pat in this Kitchen"
9186 "Find Pat in this Laundrette"
9187 "Find Pat at this Foyer"
9188 "Find Pat in this Living Room"