TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning, and Hyperlinking

George Awad {gawad@nist.gov} Asad A. Butt {asad.butt@nist.gov}
Jonathan Fiscus {jfiscus@nist.gov} David Joy {david.joy@nist.gov}
Andrew Delgado {andrew.delgado@nist.gov}
Willie McClinton {Multimodal information group student intern}
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

Martial Michel {martialmichel@datamachines.io}
Data Machines Corp., Sterling, VA 20166, USA

Alan F. Smeaton {alan.smeaton@dcu.ie}
Insight Research Centre, Dublin City University, Glasnevin, Dublin 9, Ireland

Yvette Graham {graham.yvette@gmail.com}
ADAPT Research Centre, Dublin City University, Glasnevin, Dublin 9, Ireland

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

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

Maria Eskevich {maria@clarin.eu}
CLARIN ERIC, Netherlands

Roeland Ordelman {roeland.ordelman@utwente.nl}
University of Twente, Netherlands

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

May 30, 2018

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2017 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 seventeen 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 2017 represented a continuation of five tasks from 2016, and the addition of a new pilot video to text description task. In total, 35 teams (see Table 1) from various research organizations worldwide completed one or more of the following six 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. Video to Text Description (pilot task) (VTT)

Table 2 represent organizations that registered but did not submit any runs.

This year TRECVID used again the same 600 hours of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC.3) that were used for ad-hoc Video Search in 2016. 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 again the 464 hours of the BBC (British Broadcasting Corporation) EastEnders video as used before since 2013 till 2016. 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 similarly to previous years, while 3,288 hours of blib.tv videos were used for the video Hyperlinking task. Finally, the new video to text description pilot task proposed last year was run again and used 1 880 Twitter vine videos collected through the online Twitter API public stream.

The Ad-hoc search, instance search, and multimedia event detection results were judged by NIST human 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 video-to-text task was annotated by NIST human assessors and scored automatically later on using Machine Translation (MT) metrics and Direct Assessment (DA) by Amazon Mechanical Turk workers on sampled runs.

This paper is an introduction to the evaluation framework, tasks, data, and measures used in the workshop. For detailed information about the approaches and results, the reader should see the various site reports and the results pages available at the workshop proceeding online page [TV17Pubs, 2017].

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
−−	−−	*VT*	−−	−−	−−	*NAm*	*Arete*	Arete Associates
IN	−−	−−	*MD*	*SD*	∗∗	*Asia*	*BUPT_MCPRL*	Beijing Univ. of Posts and TeleComm.s
−−	−−	−−	*MD*	−−	−−	*Asia*	*MCISLAB*	Beijing Institute of Technology
−−	−−	*VT*	−−	−−	−−	*NAm*	*CMUBOSCH*	Carnegie Mellon Univ. Robert Bosch LLC, Research Technology Center
−−	−−	*VT*	−−	−−	−−	*Aus*	*UTS_CAI*	Center of AI, Univ. of Technology Sydney
IN	−−	−−	−−	−−	−−	*Eur*	*TUC_HSMW*	Chemnitz Univ. of Technology Univ. of Applied Sciences Mittweida
−−	−−	*VT*	∗∗	−−	−−	*Asia*	*UPCer*	China university of Petroleum
−−	−−	*VT*	−−	−−	−−	*NAm*	*CCNY*	City College of New York, CUNY
−−	*HL*	*VT*	∗∗	−−	*AV*	*Asia*	*VIREO*	City Univ. of Hong Kong
−−	−−	*VT*	−−	−−	−−	*NAm*	*KBVR*	Etter Solutions
−−	∗∗	−−	−−	−−	*AV*	*Eur*	*EURECOM*	EURECOM
−−	−−	−−	−−	−−	*AV*	*NAm*	*FIU_UM*	Florida International Univ. Univ. of Miami
−−	−−	−−	−−	−−	*AV*	*Eur* + *Asia*	*kobe_nict_siegen*	Kobe Univ., National Institute of Information and Comm. Technology (NICT), Univ. of Siegen, Germany
IN	−−	−−	*MD*	*SD*	*AV*	*Eur*	*ITI_CERTH*	Information Technologies Institute, Centre for Research and Technology Hellas
−−	−−	*VT*	−−	−−	−−	*Eur*	*DCU.Insight.ADAPT*	Insight Centre for Data Analytics @ DCU Adapt Centre for Digital Content and Media
IN	−−	−−	−−	−−	−−	*Eur*	*IRIM*	EURECOM, LABRI, LIG, LIMSI, LISTIC
−−	−−	*VT*	−−	−−	−−	*Asia*	*KU_ISPL*	Intelligent Signal Processing Laboratory of Korea Univ.
−−	*HL*	−−	−−	−−	−−	*Eur*	*IRISA*	IRISA; CNRS; INRIA; INSA Rennes, Univ. Rennes 1
−−	∗∗	−−	−−	−−	*AV*	*Eur*	*ITEC_UNIKLU*	Klagenfurt Univ.
IN	∗∗	*VT*	∗∗	∗ ∗ ∗	*AV*	*Asia*	*NII_Hitachi_UIT*	National Institute of Informatics, Japan (NII); Hitachi, Ltd; Univ. of Information Technology, VNU-HCM, Vietnam (HCM-UIT)
IN	−−	−−	−−	−−	−−	*Asia*	*WHU_NERCMS*	National Engineering Research Center for Multimedia Software, Wuhan Univ.
IN	−−	−−	−−	−−	−−	*Asia*	*NTT_NII*	NTT Comm. Science Laboratories, National Institute of Informatics
IN	∗∗	∗∗	∗∗	∗∗	∗∗	*Asia*	*PKU_ICST*	Peking Univ.
−−	*HL*	−−	−−	−−	−−	*Eur*	*EURECOM_POLITO*	Politecnico di Torino and Eurecom
−−	−−	*VT*	*MD*	*SD*	*AV*	*NAm* + *Asia*	*INF*	Renmin Univ. Shandong Normal Univ. Chongqing university of posts and telecommunications
−−	−−	*VT*	−−	−−	−−	*NAm* + *Asia*	*RUC_CMU*	Renmin Univ. of China Carnegie Mellon Univ.
−−	−−	*VT*	−−	∗∗	−−	*Asia*	*SDNU_MMSys*	Shandong Normal Univ.
−−	−−	−−	∗∗	*SD*	−−	*Asia*	*BCMI*	Shanghai Jiao Tong Univ.
−−	−−	−−	−−	*SD*	−−	*Asia*	*SeuGraph*	Southeast Univ. Computer Graphics Lab
−−	−−	*VT*	−−	−−	−−	*Asia* + *Aus*	*DL* − 61 − 86	The Univ. of Sydney Zhejiang Univ.
−−	−−	*VT*	−−	−−	−−	*Asia*	*TJU_NUS*	Tianjin Univ.; National Univ. of Singapore
−−	−−	∗∗	*MD*	−−	∗∗	*Asia*	*TokyoTech_AIST*	Tokyo Institute of Technology, National Institute of Advanced Industrial Science and Technology
−−	−−	*VT*	*MD*	∗∗	*AV*	*Eur*	*MediaMill*	Univ. of Amsterdam
−−	−−	∗∗	∗∗	−−	*AV*	*Asia*	*Waseda_Meisei*	Waseda Univ.; Meisei Univ.
−−	−−	−−	−−	*SD*	−−	*Asia*	*WHU_IIP*	Wuhan Univ.

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

Table 2: Participants who did not submit any runs

Task						Location	TeamID	Participants
IN	HL	VT	MD	SD	AV			
−−	−−	**	**	−−	−−	NAm	burka	AFRL
−−	−−	−−	**	−−	−−	NAm	rponnega	Arizona State Univ.
**	**	**	**	**	**	Eur	ADVICE	Baskent Univ.
−−	−−	−−	**	**	−−	Asia	drBIT	Beijing Institute of Technology
**	−−	−−	−−	−−	−−	Asia	U_TK	Dept. of Information Science & Intelligent Systems, The Univ. of Tokushima
−−	−−	−−	−−	**	−−	Afr	EJUST_CPS	Egypt-Japan Univ. of Science and Technology.(EJUST)
−−	**	**	**	−−	−−	Afr	mounira	ENIG
−−	−−	**	−−	**	−−	NAm	UNCFSU	Fayetteville State Univ.
−−	−−	−−	**	−−	−−	Asia	Fudan	Fudan Univ.
−−	**	−−	−−	−−	−−	NAm	FXPAL	FX Palo Alto Laboratory, Inc.
−−	−−	−−	−−	−−	**	Asia	V.DO	Graduate School of Convergence Science and Technology (GSCST), Seoul National Univ.(SNU).
−−	−−	**	−−	−−	**	Asia	HFUT_Multimedia$_B$W	Hefei Univ. of technology
**	−−	−−	−−	−−	−−	Asia	Victors	IIT
**	−−	−−	−−	−−	**	Eur	JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH
−−	−−	−−	**	−−	−−	Asia	TCL_HRI_team	KAIST
−−	−−	−−	**	−−	**	Eur	LIG	LIG/MRIM
−−	−−	**	**	−−	−−	Asia	DreamVideo	Multimedia Research Center, HKUST
−−	−−	−−	−−	**	−−	Asia	mcmliangwengogo	Multimedia Communication Laboratory at MCM Inc.
**	−−	−−	−−	−−	−−	Asia	NTUROSE	Nanyang Technological Univ.
−−	−−	−−	−−	**	−−	Asia	DLMSLab20170109	National Central Univ. CSIE
−−	−−	−−	**	**	−−	Asia	NUSLV	National Univ. of Singapore
**	**	**	**	**	−−	Afr	REGIMVID	National Engineering School of Sfax (Tunisia)
−−	**	−−	**	−−	−−	Eur	NOVASearch	NOVA Laboratory for Computer Science and Informatics Universidade NOVA Lisboa
−−	−−	−−	**	−−	**	SAm	ORAND	ORAND S.A. Chile
−−	**	−−	−−	−−	−−	Eur	LaMas	Radboud Univ., Nijmegen
**	−−	**	**	**	−−	Asia	PKUMI	Peking Univ.
−−	−−	**	−−	−−	−−	NAm	prna	Philips Research North America
**	−−	−−	**	**	−−	Afr	SSCLL_Team	Sfax Smart City Living Lab (SSCLL)
−−	−−	−−	−−	**	−−	Asia	Texot	Shanghai Jiao Tong Univ.
−−	−−	−−	**	−−	−−	Asia	strong	srm university, india
−−	−−	**	−−	−−	−−	NAm	CVPIA	The Univ. of Memphis
−−	**	**	**	−−	**	Asia	UEC	The Univ. of Electro-Communiacations, Tokyo
**	**	**	**	**	−−	Asia	shiyue	TianJin Univ.
−−	−−	−−	**	−−	−−	Asia	Superimage2017	Tianjin Univ.
**	**	**	**	**	−−	NAm	IQ	Vapplica Group Llc
−−	−−	−−	**	−−	−−	Eur	MHUG	Univ. of Trento
**	−−	**	−−	**	−−	Eur + Asia	Sheff_UET	Univ. of Engineering and Technology Lahore, Pakistan The Univ. of Sheffield, UK
−−	−−	**	−−	**	−−	NAm	UNTCV	Univ. of North Texas
−−	−−	−−	−−	−−	**	Asia	Visionelites	Univ. of Moratuwa, Sri Lanka.
−−	−−	−−	**	**	−−	NAm	VislabUCR	Univ. of California, The Visualization and Intelligent Systems Laboratory (VISLab)
−−	−−	−−	−−	−−	**	Eur	vitrivr	Univ. of Basel
−−	−−	−−	**	−−	−−	Asia	YamaLab	Univ. of Tokyo Graduate School of Arts and Sciences
−−	−−	−−	**	−−	−−	Asia	SITE	VIT Univ., Vellore

Task legend. IN:instance search; MD:multimedia event detection; HL:Hyperlinking; VT:Video-to-Text; SD:surveillance event detection; AV:Ad-hoc search; −−: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 [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 MED 2017 systems used the same, HAVIC development materials as in 2016, which were distributed by NIST on behalf of the LDC. Teams were also able to use site-internal resources.

Exemplar videos provided for the Pre-Specified event condition for MED 2017 belong to the HAVIC corpus.

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 (200 000 Clips with a total duration of 2 050.46 h and total size of 703 GB) are used for evaluation. Exemplar videos provided for the Ad-Hoc event condition for MED 2017 were drawn from the YFCC100M dataset. Each MED participant was responsible for dereferencing and downloading the data, as they were only provided with the identifiers for each video used in the 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.

2.7 Twitter Vine Videos

The organizers collected about 50 000 video URL using the public Twitter stream API. Each video duration is about 6 sec. A list of 1 880 URLs were distributed to participants of the video-to-text pilot task. The 2016 pilot testing data were also available for training (a set of about 2000 Vine URLs and their ground truth descriptions).

3 Ad-hoc Video Search

This year we continued the Ad-hoc video search task that was resumed again last year. The task models the end user video search use-case, who is looking for segments of video containing people, 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 by 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 2017 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 (see Appendix A for the complete list).

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 video collection of about 600 h was used for testing. It contained 335 944 video clips in mp4 format and xml meta-data files. Throughout this report we does not differentiate between a clip and a shot and thus they may be used interchangeabily.

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 10 groups submitted a total of 52 runs, from which 19 runs were manually-assisted and 33 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 150 across all submissions after removing duplicates. The bottom pool sampled 2.5 % of ranked 150 to 1000 clips and not already included in a pool. 10 Human judges (assessors) were

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

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, 89 435 clips were judged while 370 616 clips fell into the unjudged part of the overall samples. Total hits across the 30 topics reached 9611 with 7209 hits at submission ranks from 1 to 100, 2013 hits at submission ranks 101 to 150 and 389 hits at submission ranks between 151 to 1000.

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 1 % from the total tested clips. Top 5 found queries were "a person wearing any kind of hat", "a person wearing a blue shirt", "a blond female indoors", "a person wearing a scarf", and "a man and woman inside a car". On the other hand, the bottom 5 found queries were "a person holding or opening a briefcase", "a person talking on a cell phone", "a crowd of people attending a football game in a stadium", "children playing in a playground", and "at least two planes both visible". 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 there is no correlation between top performance and finding unique clips as was the case in 2016. However top performing manually-assisted runs were among the least unique clips contributors which may conclude that humans helped those systems in retrieving more common clips but not necessarily unique clips. We notice as well that top unique clips' contributors where among the least performed teams which may indicate that their approaches may have been different than other teams to successful in retrieve unique clips but not the very common clips retreived by other teams as well.

Figures 3 and 4 show the results of all the 19 manually-assisted and 33 fully automatic run submissions respectively. This year the max and median scores are significantly higher than 2016 for both run submission types (e.g 3x times for automatic runs). We should also note here that 12 runs were submitted under the training category of E, while there was 0 runs using category F similarly to last year while the majority of runs were of type D. Compared to the semantic indexing task that was running to detect single concepts (e.g airplane, animal, bridge,...etc) from 2010 - 2015 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. A sample topics where highlighted by oval shapes to represent topics that manually-assisted runs achieved higher scores compared to their corresponding ones in the automatic runs. Surprisingly there are some topics as well where automatic runs achieved better than manually-assisted ones. A sample of top queries are highlighted in green while samples of bottom queries are highlighted in yellow.

A main theme among the top performing queries is their composition of more common visual concepts (e.g snow, kitchen, hat, etc) compared to the bottom ones which require more temporal analysis for some activities (e.g running, falling down, dancing, eating, opening/closing object, etc). 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.

In order to analyze which topics in general were the most easy or difficult we sorted topics by number of runs that scored xInfAP >= 0.7 for any given topic and assumed that those were the easiest topics, while xInfAP < 0.7 indicates a hard topic. Figure 7 shows a table with the easiest/hardest topics at the top rows. From that table it can be concluded that hard topics are associated with activities, actions and more dynamics or conditions that must be satisfied in the retrieved shots compared to just simple concepts within the easy topics.

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 and automatic run submissions as shown in Figures 8 and 9 respectively using significance threshold of $p<0.05$. The figure indicate the order by which the runs are significant according to the randomization test. Different levels of indentation means a significant difference according to the test. Runs at the same level of indentation are indistinguishable in terms of the test. For example, in this test the top 4 ranked runs were significantly better than all or most other runs while there is no significant difference between the four of them.

Among the submission requirements, we asked teams to submit the processing time that was consumed to return the result sets for each query. Figures 11 and 10 plots the reported processing time vs the InfAP scores among all run queries for automatic and manually-assisted runs respectively. 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 about 15 % (minimum 0 %) of submitted clips are common between any pair of runs. In comparison, the average was about 8 % in the previous year. These results show that although most submitted runs are diverse, systems compared to last year may be more similar in their approaches or at least trained on very similar datasets.

2017 Observations

A summary of general approaches by teams 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 understanding, expansion and/or transformation approaches to map concepts banks to queries. Concept scores fusion was investigated by most teams to combine useful results that satisfy the queries. Different approaches investigated video to text and unified text-image vector space approaches.

General task observations include that the Ad-hoc search is still more difficult than simple concept-based tagging. Maximum and median scores for manually-assisted and fully automatic runs are better than 2016 with manually-assisted runs performing slightly better suggesting more work needs to be done for query understanding and knowledge transfer between the human experience in formulating the query and the automatic systems.

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 dominant runs submitted where of type D and E with no runs submitted of type A or F.

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV17Pubs, 2017] in the online workshop notebook proceedings.

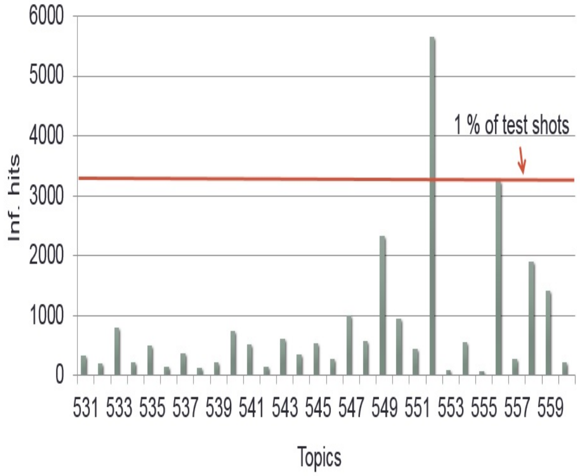


Figure 1: AVS: Histogram of shot frequencies by query number

4 Instance search

An important need in many situations involving video collections (archive video search/reuse, per-

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
9189	38009	12084	31.79	260	3367	27.86	60	1.78
9190	38032	7613	20.02	520	4000	52.54	1771	44.28
9191	38060	8188	21.51	480	3619	44.20	1488	41.12
9192	38056	9688	25.46	220	1979	20.43	442	22.33
9193	38038	11695	30.75	220	2501	21.39	142	5.68
9194	38038	11290	29.68	440	4874	43.17	387	7.94
9195	38029	12129	31.89	220	2603	21.46	258	9.91
9196	38046	7537	19.81	520	3627	48.12	1482	40.86
9197	38003	11243	29.58	120	1585	14.10	49	3.09
9198	38011	11027	29.01	140	1968	17.85	19	0.97
9199	38017	12483	32.84	160	2673	21.41	90	3.37
9200	38001	12310	32.39	120	1634	13.27	42	2.57
9201	38014	13242	34.83	200	2965	22.39	65	2.19
9202	38003	11894	31.30	300	2392	20.11	80	3.34
9203	38008	12909	33.96	160	2540	19.68	16	0.63
9204	38043	9744	25.61	420	4018	41.24	593	14.76
9205	38006	11573	30.45	100	1528	13.20	15	0.98
9206	38019	12078	31.77	200	3009	24.91	38	1.26
9207	38003	12116	31.88	140	2040	16.84	17	0.83
9208	38022	13496	35.50	140	2162	16.02	37	1.71
9209	31000	9945	32.08	240	2149	21.61	218	10.14
9210	31000	10223	32.98	320	2592	25.35	394	15.20
9211	31000	9435	30.44	220	2302	24.40	157	6.82
9212	31000	10226	32.99	200	1861	18.20	179	9.62
9213	31000	10027	32.35	240	2263	22.57	159	7.03
9214	31000	10399	33.55	120	1152	11.08	58	5.03
9215	31000	10604	34.21	200	1750	16.50	140	8.00
9216	31000	6929	22.35	400	2353	33.96	1174	49.89
9217	31000	7244	23.37	380	2227	30.74	984	44.19
9218	31000	9996	32.25	140	1432	14.33	50	3.49

sonal 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., 2016b] the instance search task seeks to address some of these needs. For six years (2010-2015) the instance search task has tested systems on retrieving specific instances of individual objects, persons and locations. Since 2016, a new query type, to retrieve specific persons in specific locations has been introduced.

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

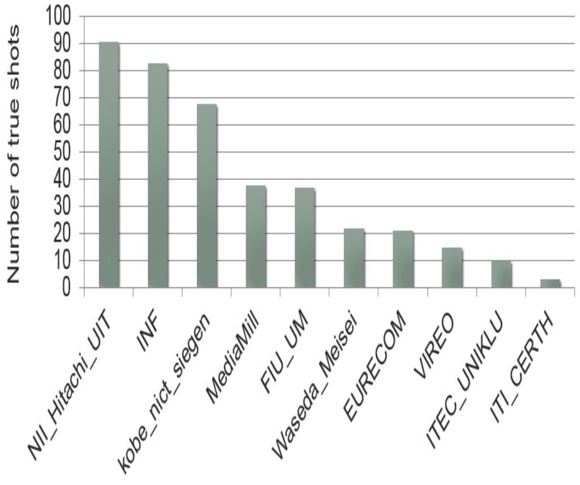


Figure 2: AVS: Unique shots contributed by team

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 different target persons across the same locations. In total this year we asked systems to find 8 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 8 groups submitted 31 automatic and 8 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. In general submissions were pooled and judged down to at least rank 100 resulting in 75 165 judged shots including 10 604 total relevant shots. Table 3[2] presents information about the pooling and judging.

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.

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

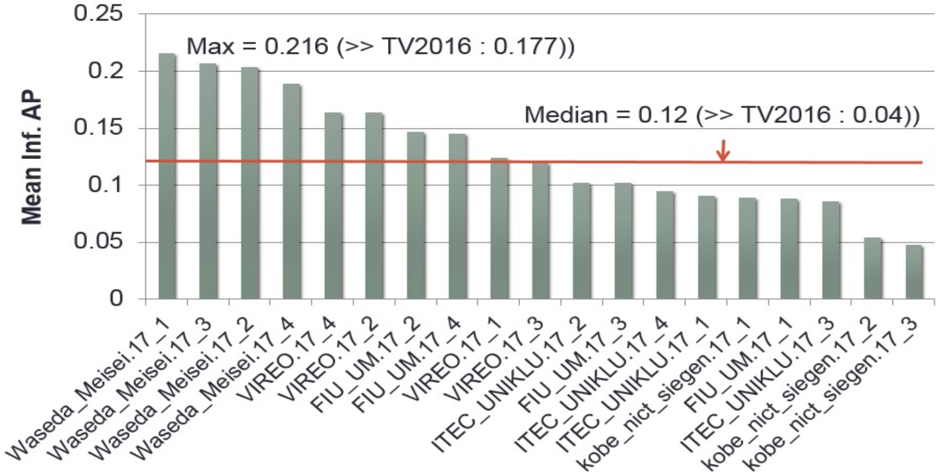


Figure 3: AVS: xinfAP by run (manually assisted)

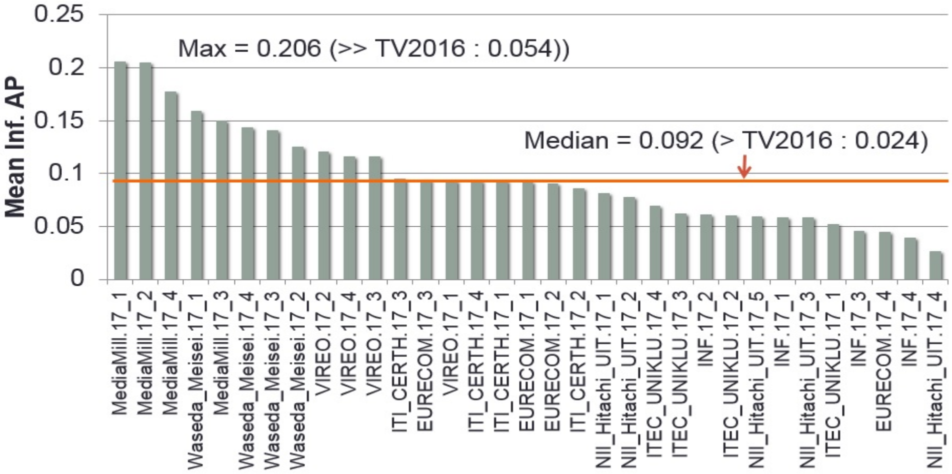


Figure 4: AVS: xinfAP by run (fully automatic)

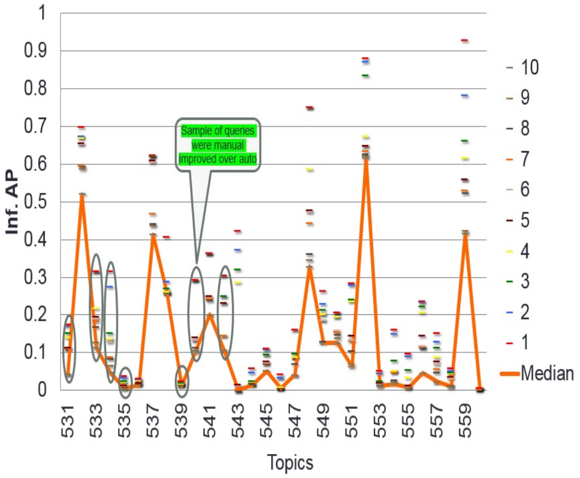


Figure 5: AVS: Top 10 runs (xinfAP) by query number (manually assisted)

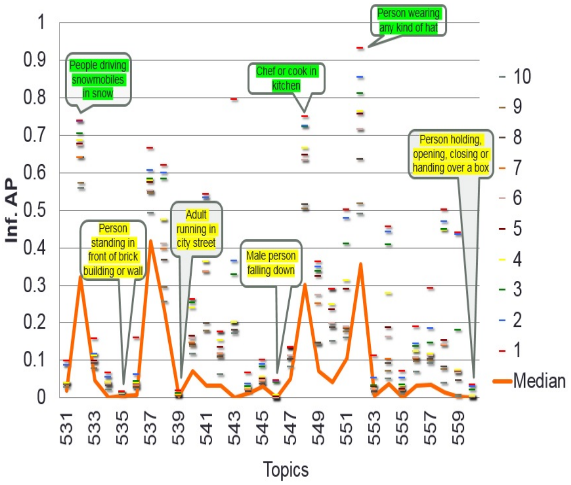


Figure 6: AVS: Top 10 runs (xinfAP) by query number (fully automatic)

4.6 Results

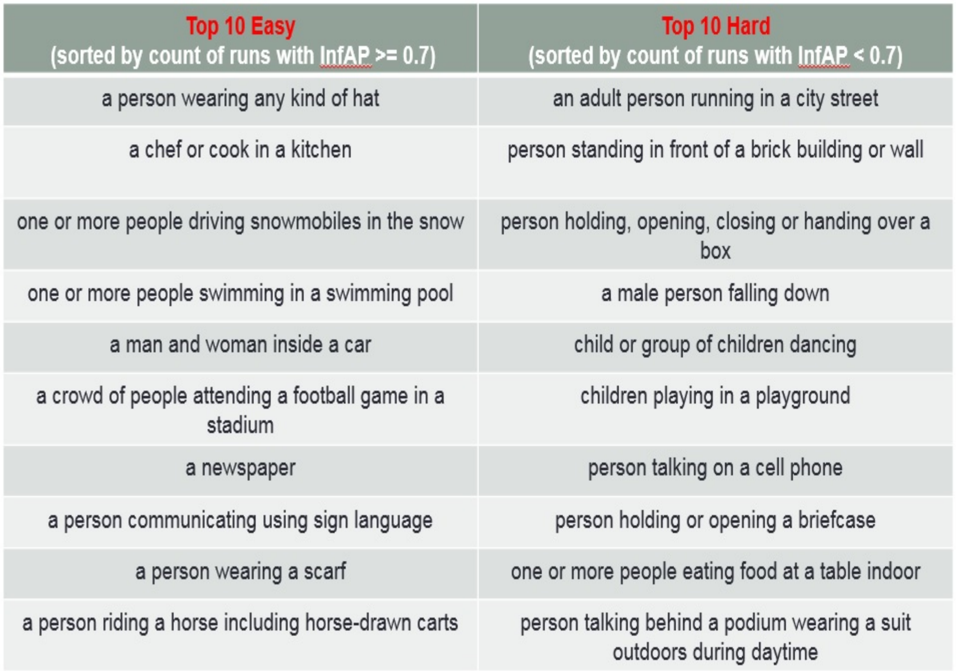
Figures 12 and 13 show the sorted scores of runs for automatic and interactive systems respectively. Both set of results show a significant increase in performance compared to 2016 results. Specifically maximum score in 2017 for automatic runs reached 0.549 compared to 0.370 in 2016 and maximum score in 2017 for interactive runs reached 0.677 compared to 0.484 in 2016.

Figure 14 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 0.611 down to 0.024. Two main factors might be expected to affect topic difficulty. The target person or the location. From the analysis of the performance of topics, it can be shown that for example the persons ”Archie”, ”Peggy” and ”phil” were easier to find as 2 ”Archie” topics were among the top 15 topics compared to only 1 in the bottom 15 topics. Similarly, 3 ”Peggy” and ”Phil” topics were among the top 15 topics compared to only 1 in the bottom 15 topics. On the other hand the target persons ”Ryan” and ”Janine” are among the hardest persons to retrieve as most of their topics where in the bottom half. In addition, it seems that the public location ”Mini-Market” made it harder to find the target persons at as 4 out of the bottom 15 topics were at the location ”Mini-Market” compared to only 1 in the top 15 topics.

Figure 15 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 $p < 0.05$. For example the top 2 runs while significantly better than the rest of the other 8 runs, there is no significant difference among each of them.

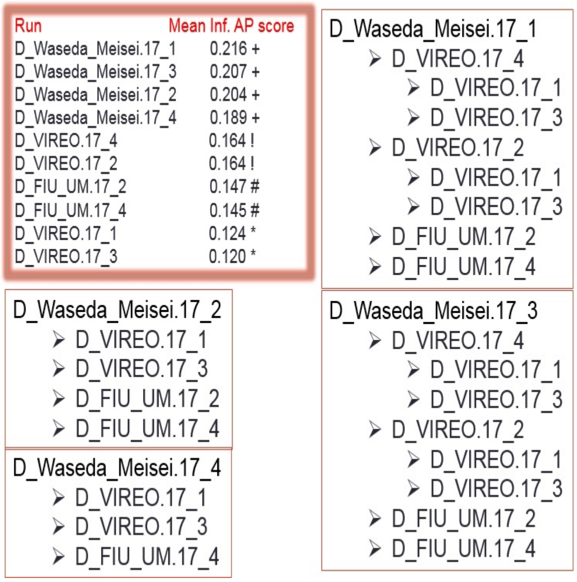
The relationship between the two main measures - effectiveness (mean average precision) and elapsed processing time is depicted in Figure 18 for the automatic runs with elapsed times less than or equal to 200 s. Only 1 team (TUC_HSMW) reported processing time below 10 s. In general there seem to be from the plot that there is a positive correlation between processing time and effectiveness.

Figure 16 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. We noticed that the location ”Mini-Market” seems to be easier when compared to auto-



Top 10 Easy (sorted by count of runs with InfAP >= 0.7)	Top 10 Hard (sorted by count of runs with InfAP < 0.7)
a person wearing any kind of hat	an adult person running in a city street
a chef or cook in a kitchen	person standing in front of a brick building or wall
one or more people driving snowmobiles in the snow	person holding, opening, closing or handing over a box
one or more people swimming in a swimming pool	a male person falling down
a man and woman inside a car	child or group of children dancing
a crowd of people attending a football game in a stadium	children playing in a playground
a newspaper	person talking on a cell phone
a person communicating using sign language	person holding or opening a briefcase
a person wearing a scarf	one or more people eating food at a table indoor
a person riding a horse including horse-drawn carts	person talking behind a podium wearing a suit outdoors during daytime

Figure 7: AVS: Easy vs Hard topics



Run	Mean Inf. AP score
D_Waseda_Meisei.17_1	0.216 +
D_Waseda_Meisei.17_3	0.207 +
D_Waseda_Meisei.17_2	0.204 +
D_Waseda_Meisei.17_4	0.189 +
D_VIREO.17_4	0.164 !
D_VIREO.17_2	0.164 !
D_FIU_UM.17_2	0.147 #
D_FIU_UM.17_4	0.145 #
D_VIREO.17_1	0.124 *
D_VIREO.17_3	0.120 *

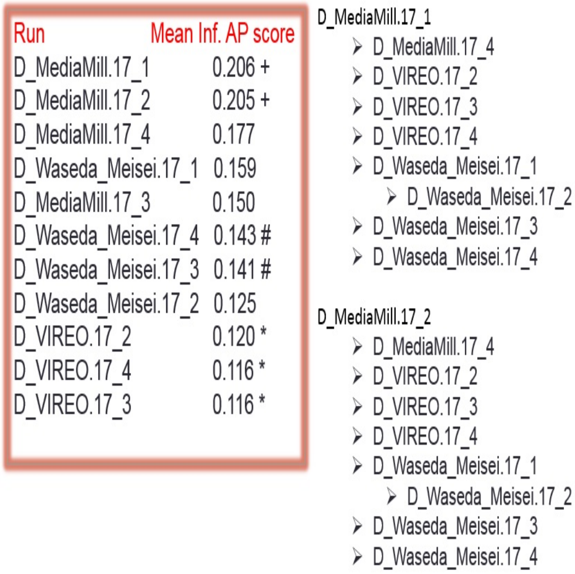
D_Waseda_Meisei.17_2
- D_VIREO.17_1
- D_VIREO.17_3
- D_FIU_UM.17_2
- D_FIU_UM.17_4

D_Waseda_Meisei.17_4
- D_VIREO.17_1
- D_VIREO.17_3
- D_FIU_UM.17_4

D_Waseda_Meisei.17_1
- D_VIREO.17_4
  - D_VIREO.17_1
  - D_VIREO.17_3
- D_VIREO.17_2
  - D_VIREO.17_1
  - D_VIREO.17_3
- D_FIU_UM.17_2
- D_FIU_UM.17_4

D_Waseda_Meisei.17_3
- D_VIREO.17_4
  - D_VIREO.17_1
  - D_VIREO.17_3
- D_VIREO.17_2
  - D_VIREO.17_1
  - D_VIREO.17_3
- D_FIU_UM.17_2
- D_FIU_UM.17_4

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



Run	Mean Inf. AP score
D_MediaMill.17_1	0.206 +
D_MediaMill.17_2	0.205 +
D_MediaMill.17_4	0.177
D_Waseda_Meisei.17_1	0.159
D_MediaMill.17_3	0.150
D_Waseda_Meisei.17_4	0.143 #
D_Waseda_Meisei.17_3	0.141 #
D_Waseda_Meisei.17_2	0.125
D_VIREO.17_2	0.120 *
D_VIREO.17_4	0.116 *
D_VIREO.17_3	0.116 *

D_MediaMill.17_1
- D_MediaMill.17_4
- D_VIREO.17_2
- D_VIREO.17_3
- D_VIREO.17_4
- D_Waseda_Meisei.17_1
  - D_Waseda_Meisei.17_2
- D_Waseda_Meisei.17_3
- D_Waseda_Meisei.17_4

D_MediaMill.17_2
- D_MediaMill.17_4
- D_VIREO.17_2
- D_VIREO.17_3
- D_VIREO.17_4
- D_Waseda_Meisei.17_1
  - D_Waseda_Meisei.17_2
- D_Waseda_Meisei.17_3
- D_Waseda_Meisei.17_4

Figure 9: AVS: Statistical significant differences (top 10 fully automatic runs)

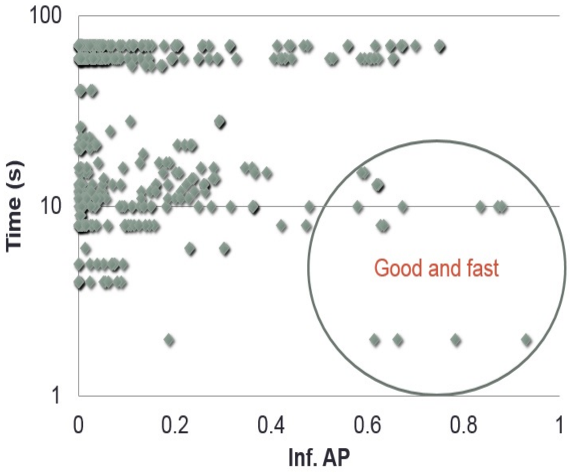


Figure 10: AVS: Processing time vs Scores (Manually assisted)

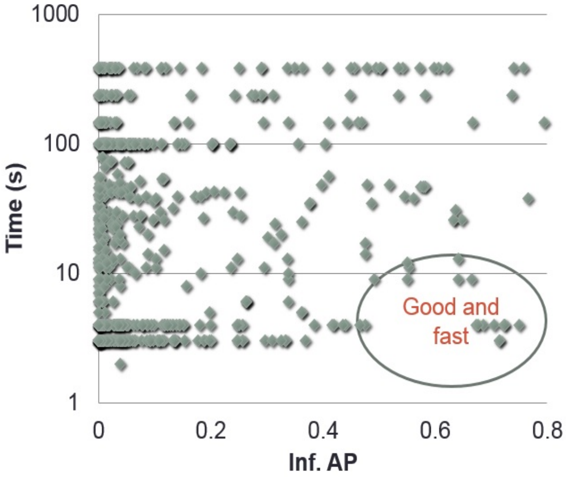


Figure 11: AVS: Processing time vs Scores (fully automatic)

matic run results. This may be due to the human in the loop effect. On the other hand, still a common pattern holds for target persons Archie and Peggy as they are still easy to spot, while "Ryan" and "Janine" are among the hardest. Figure 17 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).

Figure 19 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. We notice this year more teams are using video examples which is encouraging in order to take advantage of the full video frames for better training data instead of just few images.

4.7 Summary of observations

This is the second year the task is using the person+location query type and using the same Eastenders dataset. Although there was some decrease in number of participants who signed up for the task, the % of finishers are still the same. We should also note that this year a time consuming process was spent trying to get the data agreement set with the donor (BBC) which happened but may have affected number of teams who did not get enough time to work on and finish the task. The task guidelines were updated to give more clear rules about what is allowed or not allowed by teams (e.g using previous year's ground truth data, or manually editing the given query images). More teams used the E condition (training with video examples) which is encouraging to enable more temporal approaches (e.g. tracking characters). In general there was limited participation in the interactive systems while the overall performance for automatic systems has improved compared to last year.

To summarize the main approaches taken by different teams, NII_Hitachi_UIT team focused on improving face recognition using hard negative samples and Radial Basis Function (RBF) kernel instead of linear kernel for SVM. They also tried to improve recall using scene tracking backward and forward to reidentify persons. Finally, they did some experiments with person name mentions in the video transcripts but there was no gain noticed. The ITI_CERTH team focused on interactive runs where their system included several modes for navigation including visual similarity, scene similarity, face detection and visual concepts. Late fusion of scores where applied on the deep convolutional neural network (DCNN) face descriptors and scene descriptors but their conclusion was that performance is limited by the suboptimal face detection. The NTT team applied location search based on aggregated selective match kernel while the person search was based on OpenFace neural network models which is limited to frontal faces and fusion of results was based on ranks or scores. The OpenFace here as well influenced the results by its limitations. The WHU-NERCMS team had several components in their system including a filter to delete irrelevant shots, person search based on face recognition and speaker identification, scene retrieval based on landmarks and convolutional neural network (CNN) features and finally fusion based on multiplying scores. Their analysis concluded that the scene retrieval is limited by the pre-trained CNN models.

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV17Pubs, 2017] in the online workshop notebook proceedings.

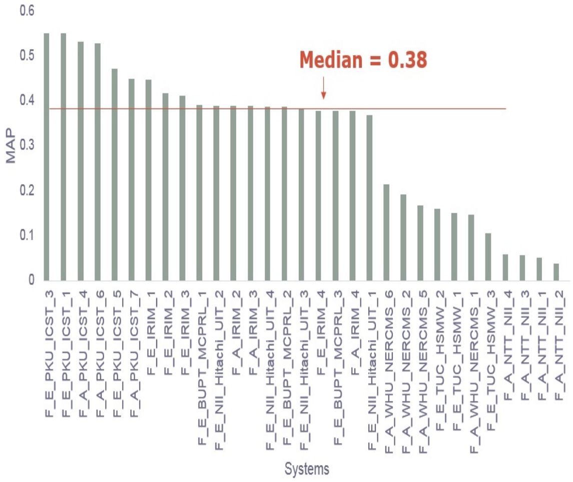


Figure 12: INS: Mean average precision scores for automatic systems

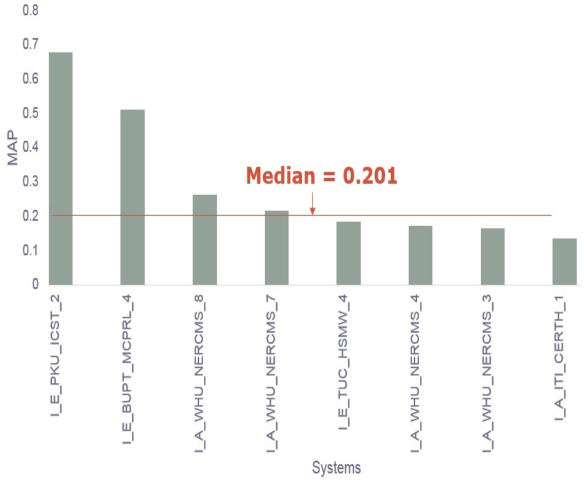


Figure 13: INS: Mean average precision scores for interactive systems

5 Multimedia event detection

The 2017 Multimedia Event Detection (MED) evaluation was the eighth evaluation of technologies that search multimedia video clips for complex events of interest to a user.

The MED 17 evaluation saw the introduction of several changes aimed at simplifying and reducing the cost of administering the evaluation. One major change, was that an additional set of clips from the Yahoo Flickr Creative Commons 100M dataset (YFCC100M) supplanted the HAVIC Progress portion of the test set from MED 16.

The full list of changes to the MED evaluation protocol for 2017 are as follows:

- HAVIC Progress portion of the test set supplanted by additional YFCC100M clips
- Introduced 10 new Ad-Hoc (AH) events
- Discontinued the 0 Exemplar (0Ex), and 100 Exemplar (100Ex) training conditions
- Discontinued the interactive Ad-Hoc subtask
- All participants were required to process the full test set

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

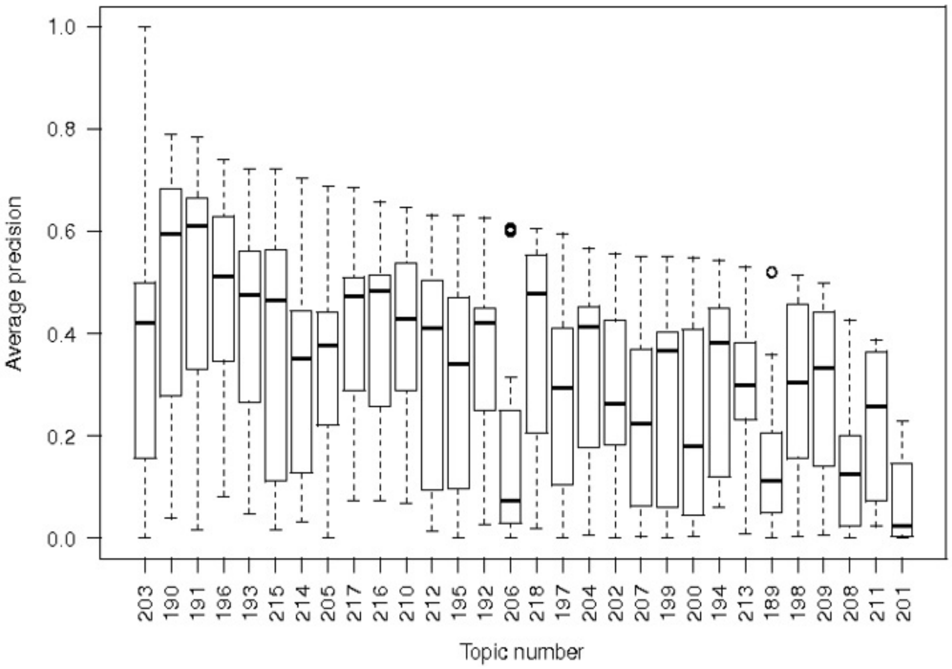
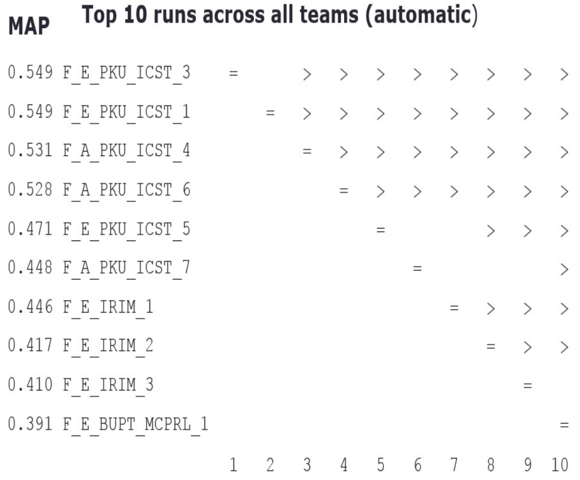


Figure 14: INS: Boxplot of average precision by topic for automatic runs.



MAP	Top 10 runs across all teams (automatic)	1	2	3	4	5	6	7	8	9	10
0.549	F_E_PKU_ICST_3	=		>	>	>	>	>	>	>	>
0.549	F_E_PKU_ICST_1		=	>	>	>	>	>	>	>	>
0.531	F_A_PKU_ICST_4			=	>	>	>	>	>	>	>
0.528	F_A_PKU_ICST_6				=	>	>	>	>	>	>
0.471	F_E_PKU_ICST_5					=			>	>	>
0.448	F_A_PKU_ICST_7						=				>
0.446	F_E_IRIM_1							=	>	>	>
0.417	F_E_IRIM_2								=	>	>
0.410	F_E_IRIM_3									=	
0.391	F_E_BUPT_MCPRL_1										=

Figure 15: INS: Randomization test results for top automatic runs. "E":runs used video examples. "A":runs used image examples only.

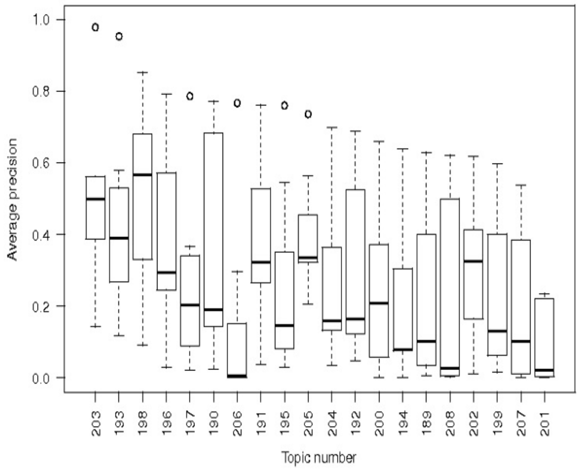
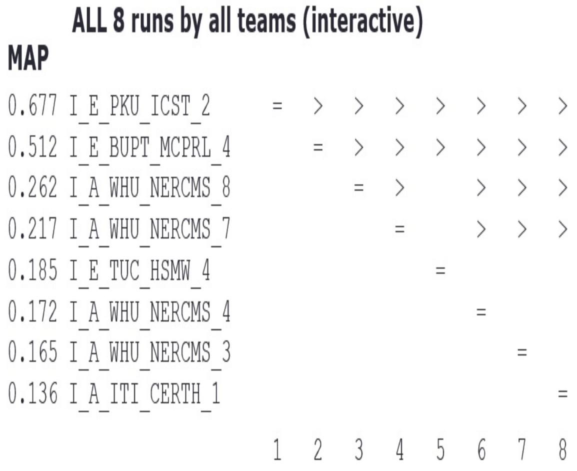


Figure 16: INS: Boxplot of average precision by topic for interactive runs



ALL 8 runs by all teams (interactive)

MAP									
0.677	I_E_PKU_ICST_2	=	>	>	>	>	>	>	>
0.512	I_E_BUPT_MCPRL_4		=	>	>	>	>	>	>
0.262	I_A_WHU_NERCMS_8			=	>		>	>	>
0.217	I_A_WHU_NERCMS_7				=		>	>	>
0.185	I_E_TUC_HSMW_4					=			
0.172	I_A_WHU_NERCMS_4						=		
0.165	I_A_WHU_NERCMS_3							=	
0.136	I_A_ITI_CERTH_1								=
		1	2	3	4	5	6	7	8

Figure 17: INS: Randomization test results for top interactive runs. "E":runs used video examples. "A":runs used image examples only.

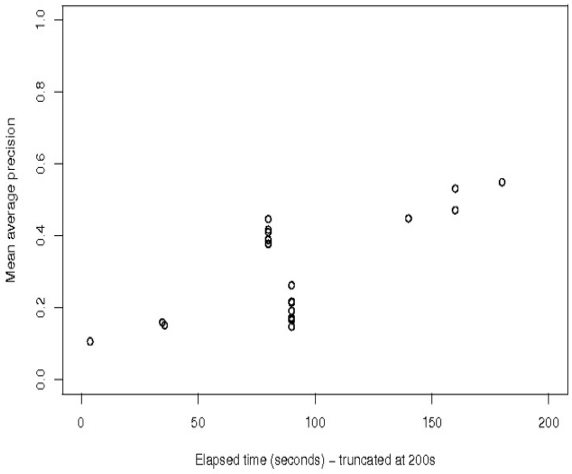


Figure 18: INS: Mean average precision versus time for fastest runs

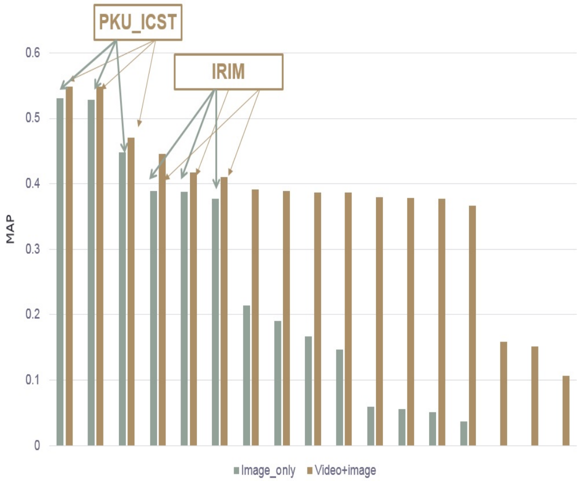


Figure 19: INS: Effect of number of topic example images used

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 two 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. In past years evaluations, a third style, interactive Ad-Hoc event detection was offered, which was 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 no teams had chosen to participate in the interactive Ad-Hoc task for both MED 2015 and MED 2016, it's no longer supported.

5.1 Data

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

The HAVIC data, which was collected by the Linguistic Data Consortium, consists of publicly available, user-generated content posted to the various Internet video hosting sites. Instances of the events were collected by specifically searching for target events using text-based Internet search engines. All video data was reviewed to protect privacy, remove offensive material, etc., prior to inclusion in the corpus. Video clips were provided in MPEG-4 formatted files. The video was encoded to the H.264 standard. The audio was encoded using MPEG-4s 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. In MED 2016, 100 000 randomly selected[3] videos from the YFCC100M dataset were included in the test set. This year, those same 100 000 videos, along with 100 000 new videos, selected in the same way from the YFCC100M dataset comprise the test set.

MED participants were provided the data as specified in the HAVIC and YFCC100M data sections of this paper. The MED '17 Pre-Specified event names are listed in Table 4, and Table 5 lists the MED '17 Ad-Hoc Events.

Table 4: MED '17 Pre-Specified Events

—— MED'16 event re-test
Camping
Crossing a Barrier
Opening a Package
Making a Sand Sculpture
Missing a Shot on a Net
Operating a Remote Controlled Vehicle
Playing a Board Game
Making a Snow Sculpture
Making a Beverage
Cheerleading

Table 5: MED '17 Ad-Hoc Events

Fencing
Reading a book
Graduation ceremony
Dancing to music
Bowling
Scuba diving
People use a trapeze
People performing plane tricks
Using a computer
Attempting the clean and jerk

5.2 Evaluation

The participating MED teams tested their system outputs on the following dimensions:

- Events: all 10 Pre-Specified events (PS17) and/or all 10 Ad-Hoc events (AH17).
- 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 PS and AH systems.

[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/

For each event search, a system generated a rank for each video in the test set, where a rank is a value from 1 (best) 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 evaluation measure for performance was Inferred Mean Average Precision[Yilmaz et al., 2008]. While Mean Average Precision (MAP) was used as a measure in the past, specificially over the HAVIC test set data, this is not possible for MED 17, as the test set is comprised entirely YFCC100M video data, which has not been fully annotated with respect to the MED 17 events.

5.4 Results

6 teams participated in the MED '17 evaluation. All teams participated in the Pre-Specified (PS) Event condition, processing the 10 PS events. 4 teams chose to participate in the Ad-Hoc (AH) portion of the evaluation, which was optional, processing the 10 AH events. This year, all teams submitted runs for only "Small" (SML) sized systems.

For the Mean Inferred Average Precision metric, 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. We define two strata 1-60 with a sampling rate of 100 %, and 61-200 at 20 %. We refer to Inferred Average Precision, and Mean Inferred Average Precision measures using these parameters as infAP200, and MinfAP200 respectively. These parameters were selected for the MED 2015 evaluation as they produced MinfAP scores highly correlated with MAP (R^2 of 0.989 [Over et al., 2015]), a trend which was also observed in MED 2016.

This year, we introduced 10 new AH events, with exemplars sourced from the YFCC100M dataset. A different scouting method was used this year for the AH events. We used a Multimedia Event Detection developed for the Intelligence Advanced Research Projects Activity (IARPA) Aladdin program, which was trained on prospective event kits with exemplars sourced from the fully annotated HAVIC dataset found with a simple text search. We then processed a subset of the YFCC100M dataset, disjoint from the evaluation set, and hand selected exemplars from the returned ranked lists, prioritizing diversity. This approach allowed us to create event kits with exemplars taken from an unannotated collection of video.

Figures 20 and 21 show the MinfAP200 scores for the PS and AH event conditions respectively. Figure 22 shows the infAP200 scores on the PS event condition broken down by event and system. Figure 23 shows this same breakdown for the AH event condition, an interesting system effect can be observed for the INF team on several events. According to the system descriptions provided by teams, the system submitted by INF ignored the exemplar videos, effectively submitting as a 0Ex system (official support for the 0Ex evaluation condition was dropped this year). Figures 24, and 25 show the PS and AH event conditions, respectively, broken down by system and event.

Figures 28 and 29 show the size of the assessment pools by event, and the target richness within each pool. Note that for event E076, "Scuba diving", the assessment pool is almost completely saturated with targets, at 97.6 %. To contrast, figures 26 and 27 show the assessment pool size, and target richness by event for the PS event condition.

5.5 Summary

In summary, all 6 teams participated in the Pre-Specified (PS) test, processing all 10 PS events, with MinfAP200 scores ranging from 0.003 to 0.406 (median of 0.112). For the Ad-Hoc (AH) event condition, 4 of 6 teams participated, processing all 10 AH events, where MinfAP200 scores ranged from 0.316 to 0.636 (median of 0.455).

This year saw the introduction of 10 new AH events, scouted with a MED system in the loop instead of a simple text search of human annotations, and with exemplar videos sourced from YFCC100M instead of HAVIC. While the infAP200 scores appear to be higher in absolute terms for the AH event condition, over PS, the authors would like to caution against making direct comparisons between the two because of these differences. For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV17Pubs, 2017] in the online workshop notebook proceedings.

The MED task will not continue in 2018, citing declining interest in the task. However, we intend to release the test set annotations for this year, and prior evaluation years for continued research. We would like to thank task participants for their interest, and IARPA for their support of the task through 2015.

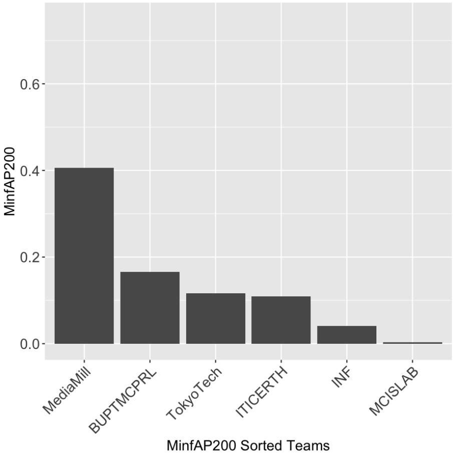


Figure 20: MED: Mean infAP200 scores of primary systems submitted for the Pre-Specified event condition

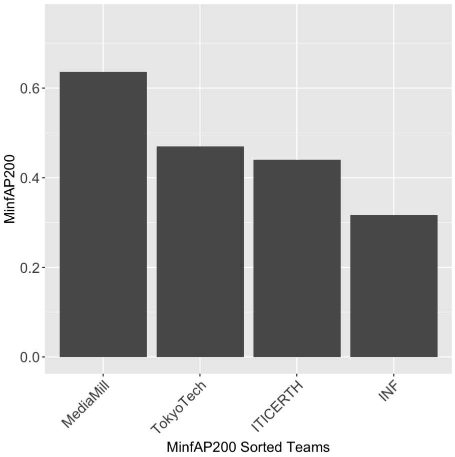


Figure 21: MED: Mean infAP200 scores of primary systems submitted for the Ad-Hoc event condition

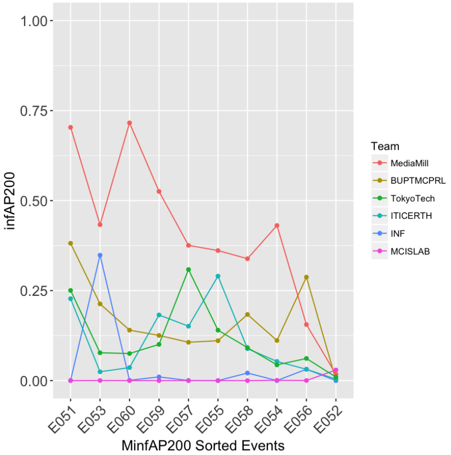


Figure 22: MED: Pre-Specified systems vs. events

6 Surveillance event detection

The 2017 Surveillance Event Detection (SED) evaluation was the tenth 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 has occurred every year. 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 camera views.

For 2017, the evaluation test data used a 10-hour subset (EVAL17) 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)). EVAL17 is identical to the evaluation set for 2016.

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

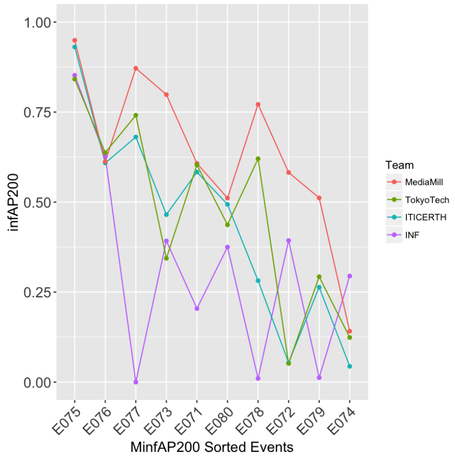


Figure 23: MED: Ad-Hoc systems vs. events

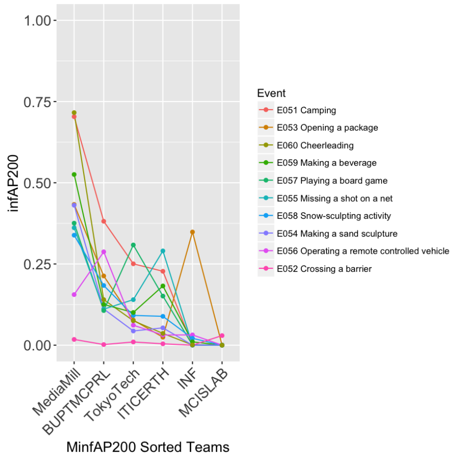


Figure 24: MED: Pre-Specified events vs. systems

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 high-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

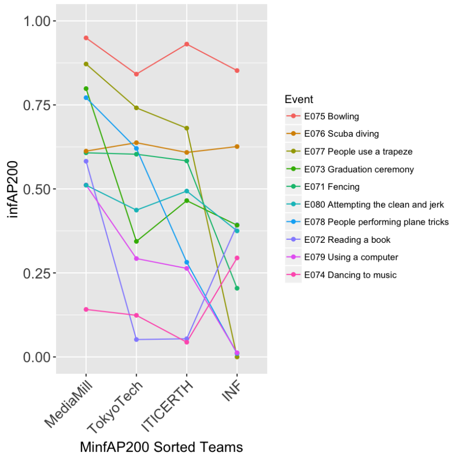


Figure 25: MED: Ad-Hoc events vs. systems

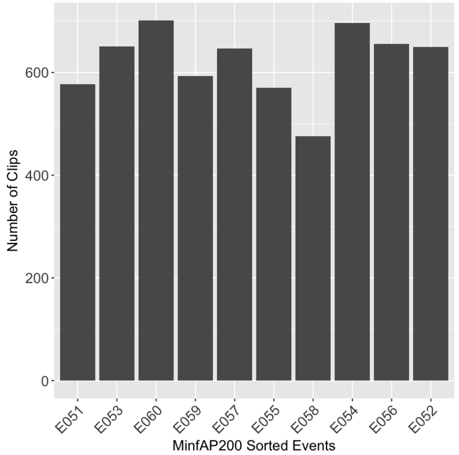


Figure 26: MED: Pre-Specified assessment pool size

- 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

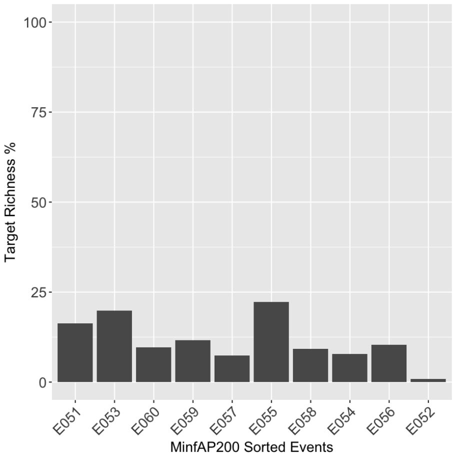


Figure 27: MED: Pre-Specified assessment pool target richness

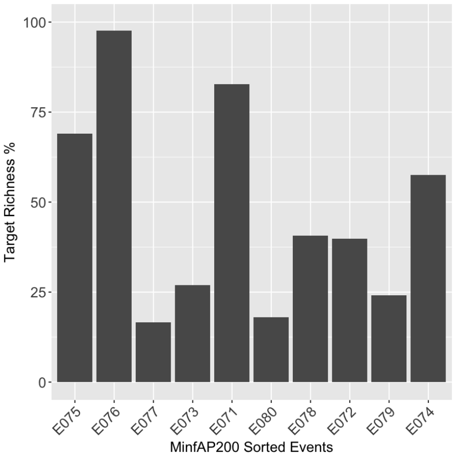


Figure 29: MED: Ad-Hoc assessment pool target richness

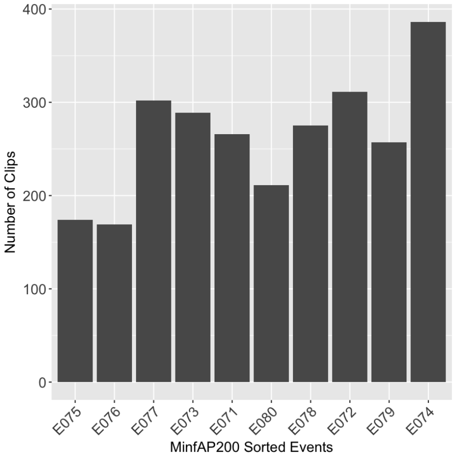


Figure 28: MED: Ad-Hoc assessment pool size

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 2017 as SUB17.

In 2017, 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 EVAL17 and/or SUB17. 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

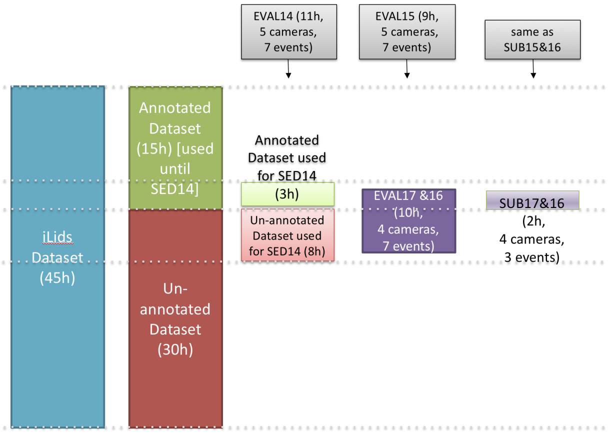

EVAL14 (11h, 5 cameras, 7 events)
EVAL15 (9h, 5 cameras, 7 events)
same as SUB15&16
iLids Dataset (45h)
Annotated Dataset (15h) [used until SED14]
Un-annotated Dataset (30h)
Annotated Dataset used for SED14
(3h)
Un-annotated Dataset used for SED14 (8h)
EVAL17 &16 (10h, 4 cameras, 7 events)
SUB17&16 (2h, 4 cameras, 3 events)

Figure 30: SED17 Data Source

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, 2016 and 2017 data uses subsets of this data.

Figure 30 provides a visual representation of the annotated versus annotated information in the dataset, and how this dataset was used over the years of the SED program.

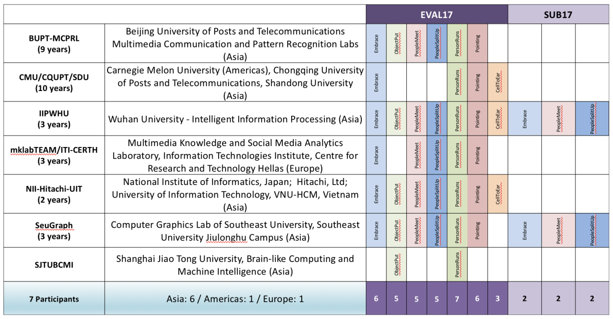
Events were represented in the Video Performance Evaluation Resource (ViPER) format using an annotation schema that specified each event observation's time interval.

6.2 Evaluation

Figure 31 shows the 7 participants to SED17.

For EVAL17, 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.

SUB17 followed the same concept, but only using 3 possible events (Embrace, PeopleMeet and PeopleSplitUp).



		EVAL17							SUB17		
BUPT-MCPRL **(9 years)**	Beijing University of Posts and Telecommunications Multimedia Communication and Pattern Recognition Labs (Asia)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing				
CMU/CQUPT/SDU **(10 years)**	Carnegie Melon University (Americas), Chongqing University of Posts and Telecommunications, Shandong University (Asia)	Embrace				PersonRuns	Pointing	CellToEar			
IIPWHU **(3 years)**	Wuhan University - Intelligent Information Processing (Asia)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing	CellToEar	Embrace	PeopleMeet	PeopleSplitUp
mklabTEAM/ITI-CERTH **(3 years)**	Multimedia Knowledge and Social Media Analytics Laboratory, Information Technologies Institute, Centre for Research and Technology Hellas (Europe)	Embrace		PeopleMeet	PeopleSplitUp	PersonRuns	Pointing				
NII-Hitachi-UIT **(2 years)**	National Institute of Informatics, Japan; Hitachi, Ltd; University of Information Technology, VNU-HCM, Vietnam (Asia)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing	CellToEar			
SeuGraph **(3 years)**	Computer Graphics Lab of Southeast University, Southeast University Jiulonghu Campus (Asia)	Embrace	ObjectPut	PeopleMeet	PeopleSplitUp	PersonRuns	Pointing		Embrace	PeopleMeet	PeopleSplitUp
SJTUBCMI	Shanghai Jiao Tong University, Brain-like Computing and Machine Intelligence (Asia)		ObjectPut			PersonRuns					
7 Participants	Asia: 6 / Americas: 1 / Europe: 1	6	5	5	5	7	6	3	2	2	2

Figure 31: SED17 Participants. Columns: Short name (years participating), Site name (Location), EVAL17 Events (from left to right: Embrace, ObjectPut, PeopleMeet, PeopleSplitUp, PersonRuns, Pointing, CellToEar), and SUB17 Events (Embrace, PeopleMeet, PeopleSplitUp)

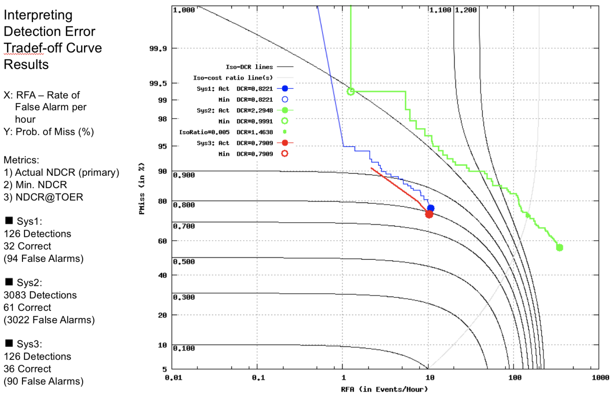


Figure 32: Interpreting DETCurve Results

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.

Figure 32 present a DETCurve with three systems, with the abscissa of the graph being the rate of false alarms (in Error/hour), and in ordinate the probability of miss (in percents). A few systems are present on that curve, Sys1, Sys2 and Sys3. Sys1 has 126 decisions, 32 of which are correct decisions, leaving 94 False Alarms. Sys2 has 3083 decisions, 61 of which are correct decisions, leaving 3022 False Alarms. Only Sys2 crosses the balancing line. Sys3 has 126 decisions, 36 of which are correct decisions, and 90 False Alarms. On the graph is shown that Sys3 has the lowest Act NDCR and lowest Min NDCR.

SED17 results are 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.

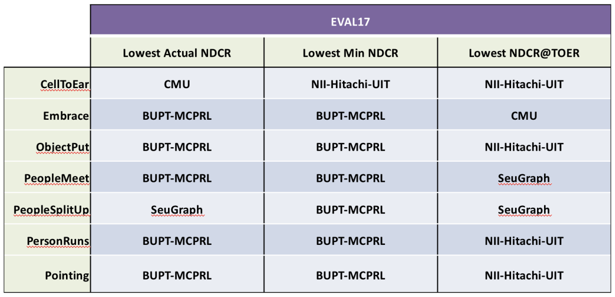
3. **N**DCR at Target Operating Error Ratio (NDCR@TOER, Secondary Metric), is another diagnostic metric. It is found by searching the DET curve for the point where it crosses the theoretical balancing point where two error types (Miss Detection and False Alarm) contribute equally to the measured NDCR. 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 SED17 Evaluation Plan [4].

6.4 Results

Figure 33 shows, per Event and per Metric the systems with the lowest NDCR for the 2017 SED Evaluation (only on primary submissions).

Figure 34, 35, 36 and 37 present the EVAL17 primary submission results for the CellToEar, PersonRuns, PeopleSplitUp and Embrace events. For additional individual results, please see the TRECVID SED proceedings.



	EVAL17		
	Lowest Actual NDCR	Lowest Min NDCR	Lowest NDCR@TOER
CellToEar	CMU	NII-Hitachi-UIT	NII-Hitachi-UIT
Embrace	BUPT-MCPRL	BUPT-MCPRL	CMU
ObjectPut	BUPT-MCPRL	BUPT-MCPRL	NII-Hitachi-UIT
PeopleMeet	BUPT-MCPRL	BUPT-MCPRL	SeuGraph
PeopleSplitUp	SeuGraph	BUPT-MCPRL	SeuGraph
PersonRuns	BUPT-MCPRL	BUPT-MCPRL	NII-Hitachi-UIT
Pointing	BUPT-MCPRL	BUPT-MCPRL	NII-Hitachi-UIT

Figure 33: SED17 Systems with the lowest NDCR

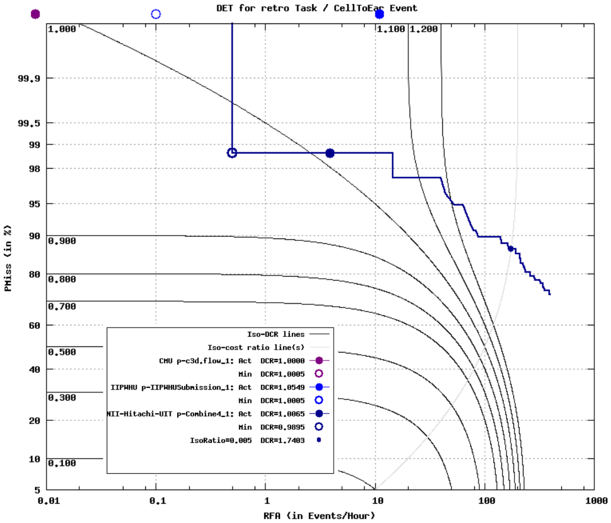


Figure 34: SED17 CellToEar Results

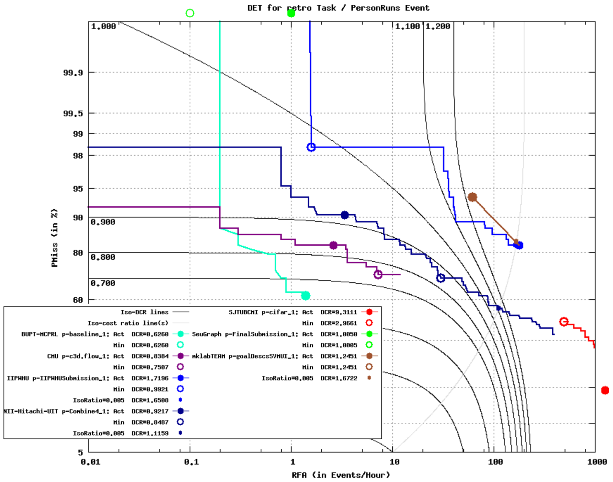


Figure 35: SED17 PersonRuns Results

[4] ftp://jaguar.ncsl.nist.gov/pub/SED17/SED17_EvalPlan_v2.pdf

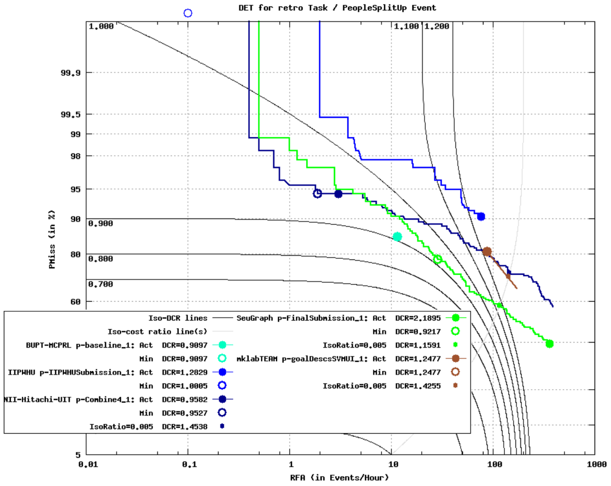


Figure 36: SED17 PeopleSplitUp Results

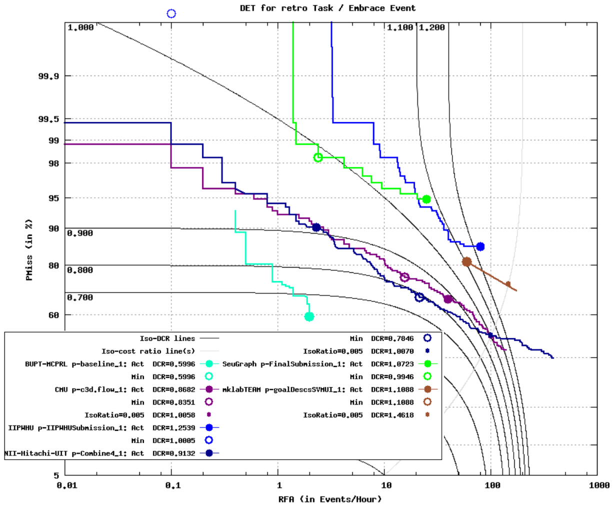


Figure 37: SED17 Embrace Results

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV17Pubs, 2017] in the online workshop notebook proceedings.

7 Video hyperlinking

7.1 System task

In 2017, we follow the high-level definition of the Video Hyperlinking (LNK) task edition 2015 [Over et al., 2015], while reusing the dataset that was introduced in 2016 [Awad et al., 2016a], and thus carrying out the comparison both between the 2017 systems, and their 2016 counterparts. 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 2017 edition of the LNK task has used the 2016 subset of the Blip10000 collection [Schmiedeke et al., 2013] crawled from blip.tv, a website that hosted semi-professional user-generated content. The 2017 anchors were multimodal, i.e., the information about suitable targets, or the information request, is a combination of both audio and visual streams.

7.2 Data

The Blip10000 dataset used for the 2017 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]. There are two sets of automatic speech recognition (ASR) transcripts: 2012 version that was originally provided with this dataset [Lamel, 2012], and 2016 version that was 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[5]. 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 order to being able to compare systems performances with 2016 results, we created the anchors of the same multimodal nature. 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.

In 2017, the anchor creators had to browse through the collection videos in the collection, and manually select the anchors. In order to optimize their search for anchors, and to ensure their representativity, we checked the genre labels that are available for the dataset, discarding the videos with genres that did not convey multimodal combinations, e.g. 'music_and_entertainment', 'literature'. For practical reasons of further assessment, we also limited anchors to be between 10 and 60 seconds long. In total, two creators generated 25 anchors and corresponding descriptions of potentially relevant targets, i.e., information request descriptions that were further used in the evaluation process.

7.4 Evaluation

Ground truth

The ground truth was generated by pooling the top 10 results of all formally submitted participant runs (12), and running the assessment task on the Amazon Mechanical Turk (AMT)[6] platform[7]. 'Target Vetting' task was organised as follows: The top 10 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.

The Target Vetting stage for all the participants' submissions involves large-scale crowdsourcing submissions processing, which is not feasible to carry out manually. Therefore, after a small scale manual check, we proceeded with automatic acceptance/rejection framework tested in previous years: the script checks whether all the required decision

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

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

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

metadata fields had been filled in, and whether the answers to the test questions were correct.

The answers thus collected are further transformed into positive/negative relevance judgments following this logic depicted in Table 6:

- In case the target description provided by task participants is clearly relevant, or clearly non-relevant, the workers should feel comfortable with their decisions (Cases 1 and 3);

- In cases where the relevance/non-relevance is less obvious, the workers indicate that they are uncomfortable with their decision (Cases 2 and 4).

For each top-10 anchor–target pair we collected three crowdworkers' judgments. The final relevance decision was made based on the majority of the relevance judgments.

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

Three groups submitted four runs each, resulting in 12 run submissions, which were used for ground truth creation and assessment using the metrics described above. They also submitted the results of their systems on the development set. An overall comparison of the systems' performance according to Precision at rank 5 and MAiSP are given in Figures 38-39.

In terms of Precision@5, all teams achieved scores well above 0.5. The order of the teams changes when results are evaluated using the MAiSP measure.

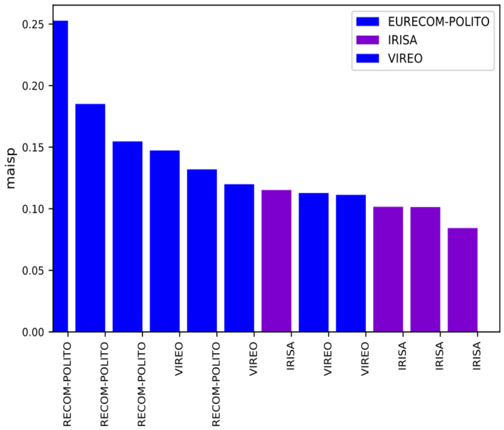


Figure 38: LNK MAiSP Results

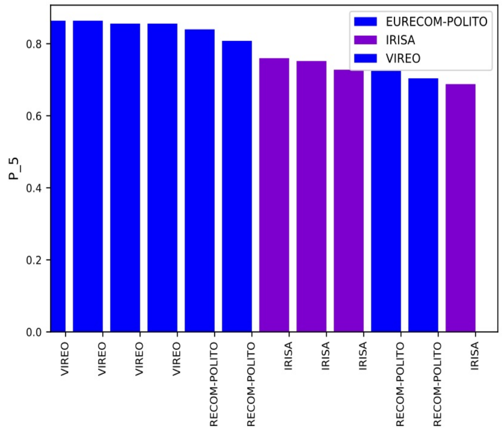


Figure 39: LNK P5 Results

Table 6: LNK'17: Automatic relevance assessment procedure of the MTurk submissions.

Case ID	MTurk worker's choice of target description	MTurk worker's feedback on decisionmaking process	Relevance decision	Number of cases
1	Correct	Positive	Relevant	547
2	Correct	Negative	Relevant	3849
3	Other	Positive	Non-relevant	1021
4	Other	Negative	Non-relevant	864

8 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 automatic 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, facilitating the search and browsing of video archives using such descriptions, describing videos as an assistive technology, 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.

The "Video to Text Description" (VTT) task was introduced in TRECVid 2016 as a pilot. This year, we continue the task with some modifications to the dataset.

8.1 Data

Over 50k Twitter Vine videos have been collected automatically, and each video has a total duration of about 6 seconds. In the task this year, a dataset of 1 880 Vine videos was selected and annotated manually by multiple assessors. An attempt was made to create a diverse dataset by removing any duplicates or similar videos as a preprocessing step. The videos were divided amongst 10 assessors, with each video being annotated by at least 2 assessors, and at most 5 assessors. The assessors 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)

Furthermore, the assessors were also asked the following question to rate the difficulty of each video on a scale of 1 to 5:

> "Please rate how difficult it was to describe the video on a scale of 1 (very easy) to 5 (very difficult)".

The videos are divided into 4 groups, based on the number of descriptions available for them. Hence, we have groups of videos with 2, 3, 4, and 5 descriptions. These groups are referred to as G2, G3, G4, and G5, respectively. Each group has multiple sets of descriptions, with each set containing a description for all the videos in that group. Therefore, videos in G2 have 2 sets (A, B) while videos in G3 have 3 sets (A, B, C), and so forth. Since all 1 880 videos have at least 2 descriptions, they are all in G2. Each group with higher number of descriptions is a subset of lower groups.

8.2 System task

The participants were asked to work on and submit results for at least one of two subtasks:

- Matching and Ranking: For each video URL in a group, return a ranked list of the most likely text

Group	No. of Videos in Set
G2	1613
G3	795
G4	388
G5	159

Table 7: Number of videos in each set for the matching and ranking task.

Subtask	Group	Runs Submitted
Matching and Ranking	G2	68
	G3	90
	G4	124
	G5	155
Description Generation	-	43

Table 8: Number of runs for each subtask.

description that corresponds (was annotated) to the video from each of the sets. Here the number of sets is equal to the number of descriptions for videos in the group.

- Description Generation: Automatically generate for each video URL a text description (1 sentence) independently and without taking into consideration the existence of any annotations.

The number of videos in each group for the matching and ranking subtask are shown in Table 7. A number of videos in the complete dataset have very similar descriptions, which can lead to confusion for systems regarding the matching and ranking task. For this reason, we removed such videos to reduce the number of videos in each group for this particular subtask. The entire dataset of 1 880 videos was used for the second subtask of description generation.

8.3 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 a number of metrics.

METEOR [Banerjee and Lavie, 2005] and BLEU [Papineni et al., 2002] are standard metrics in machine translation (MT). BLEU (bilingual evaluation understudy) is a metric used in MT and was one of the first metrics to achieve a high correlation with human judgments 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.

This year the CIDEr (Consensus-based Image Description Evaluation) metric [Vedantam et al., 2015] was used for the first time. It computes TD-IDF (term frequency inverse document frequency) for each n-gram to give a sentence similarity score. The CIDEr metric has been reported to show high agreement with consensus as assessed by humans.

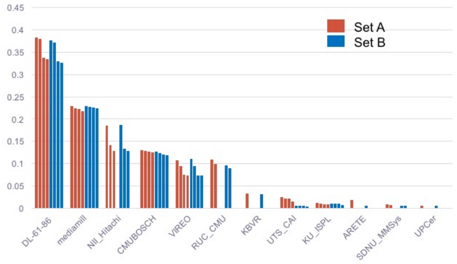


Figure 40: VTT: Matching and Ranking results across all runs for Group 2

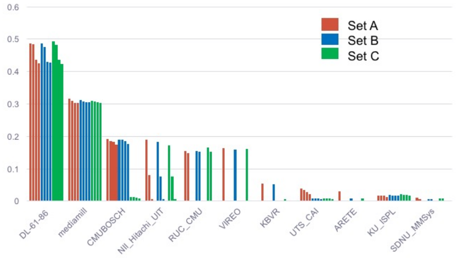


Figure 41: VTT: Matching and Ranking results across all runs for Group 3

The semantic similarity metric (STS) [Han et al., 2013] was also applied to the results, as in the previous year of this task. This metric measures how semantically similar the submitted description is to one of the ground truth descriptions.

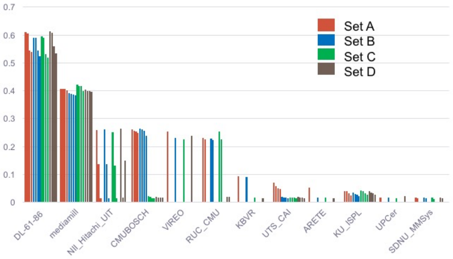


Figure 42: VTT: Matching and Ranking results across all runs for Group 4

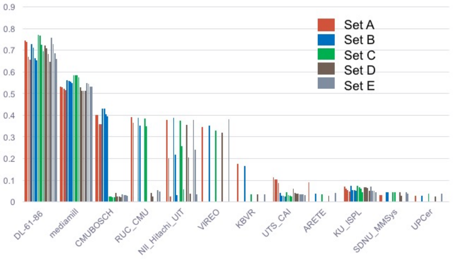


Figure 43: VTT: Matching and Ranking results across all runs for Group 5

In addition to automatic metrics, this year's description generation task includes human evaluation of the quality of automatically generated captions. Recent developments in Machine Translation evaluation have seen the emergence of Direct Assessment (DA), a method shown to produce highly reliable human evaluation results for MT [Graham et al., 2016]. DA now constitutes the official method of ranking in main MT benchmark evaluations [Bojar et al., 2017]. With respect to DA for evaluation of video captions (as opposed to MT output), human assessors are presented with a video and a single caption. After watching the video, assessors rate how well the caption describes what took place in the video on a 0–100 rating scale [Graham et al., 2017]. Large numbers of ratings are collected for captions, before ratings are combined into an overall average system rating (ranging from 0 to 100%). Human assessors are recruited via Amazon's Mechanical Turk (AMT) [8], with strict quality control measures applied to filter out or downgrade the weightings from workers unable to demonstrate the ability to rate good captions higher than lower quality captions. This is achieved by deliberately "polluting" some of the manual (and correct) captions with linguistic substitutions to generate captions whose semantics are questionable. Thus we might substitute a noun for another noun and turn the manual caption "A man and a woman are dancing on a table" into "A *horse* and a woman are dancing on a table", where "horse" has been substituted for "man". We expect such automatically-polluted captions to be rated poorly and when an AMT worker correctly does this, the ratings for that worker are improved.

Experiments have shown DA scores collected in this way for TRECVID 2016 video-captioning systems to be highly reliable, with scores from two separate data collection runs showing a close to perfect Pearson correlation of 0.997 [Graham et al., 2017]. In addition, included in the human evaluation is a hidden system made up of captions produced by human annotators. The purpose of this is to reveal at what point state-of-the-art performance in video captioning may be approaching human performance.

In total, 34 teams signed up for the task and 16 of those finished. The individual runs submitted for the subtasks and groups are shown in Table 8.

8.4 Results

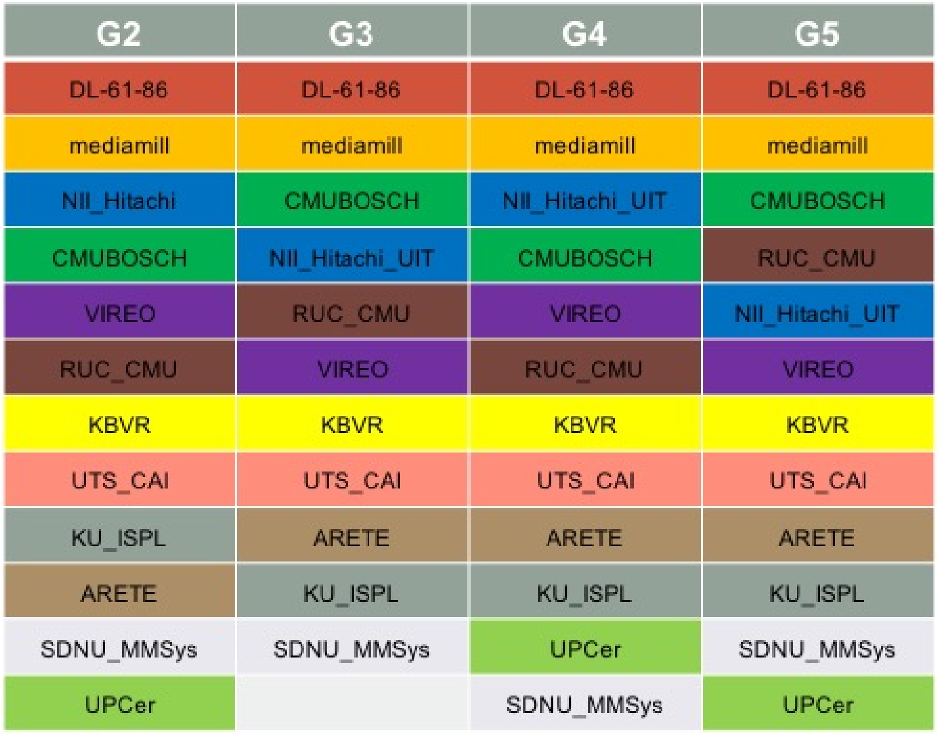
Readers should see the online proceedings for individual teams' performance and runs but here we present a high-level overview.

Matching and Ranking Sub-task

The results for the caption-ranking sub-task are shown in Figures 40 - 43. Figure 40 shows the mean inverted rank scores for all the submitted runs in G2. The runs are grouped together by teams, and results are color-coded for Set A and Set B. As expected, in most cases, the scores for a particular run are similar on Set A and Set B. However, in some cases *e.g.UTS_CAI*, the runs tend to perform much better over Set A as compared to Set B.

Figure 41 shows the mean inverted rank scores for all runs submitted for G3. Again, the scores for Sets A, B, and C are shown in different colors. Figures 42 and 43 show the scores for runs submitted for G4 and G5 respectively. The observations regarding the similarity of scores for the same run over different sets holds in each of the shown graphs for most cases. However, the runs from some teams have an anomalous behavior where they perform better on

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



G2	G3	G4	G5
DL-61-86	DL-61-86	DL-61-86	DL-61-86
mediamill	mediamill	mediamill	mediamill
NII_Hitachi	CMUBOSCH	NII_Hitachi_UIT	CMUBOSCH
CMUBOSCH	NII_Hitachi_UIT	CMUBOSCH	RUC_CMU
VIREO	RUC_CMU	VIREO	NII_Hitachi_UIT
RUC_CMU	VIREO	RUC_CMU	VIREO
KBVR	KBVR	KBVR	KBVR
UTS_CAI	UTS_CAI	UTS_CAI	UTS_CAI
KU_ISPL	ARETE	ARETE	ARETE
ARETE	KU_ISPL	KU_ISPL	KU_ISPL
SDNU_MMSys	SDNU_MMSys	UPCer	SDNU_MMSys
UPCer		SDNU_MMSys	UPCer

Figure 44: VTT: Ranking of teams with respect to the different groups

some sets as compared to others.

Figure 44 shows the ranking of the various teams with respect to the different groups. For each team, the scores for the best runs are used. The figure allows us to see which teams are performing well consistently.

Figure 45 shows the top 3 videos for each group. These videos are matched correctly in a consistent manner among runs. Most of the videos are of a short continuous scene, making it easier to describe. Figure 46 shows the bottom 3 videos for each group. In general, these videos either have lots of scenes combined, which makes them complex to describe, or contain very unusual actions.

Description Generation Sub-task

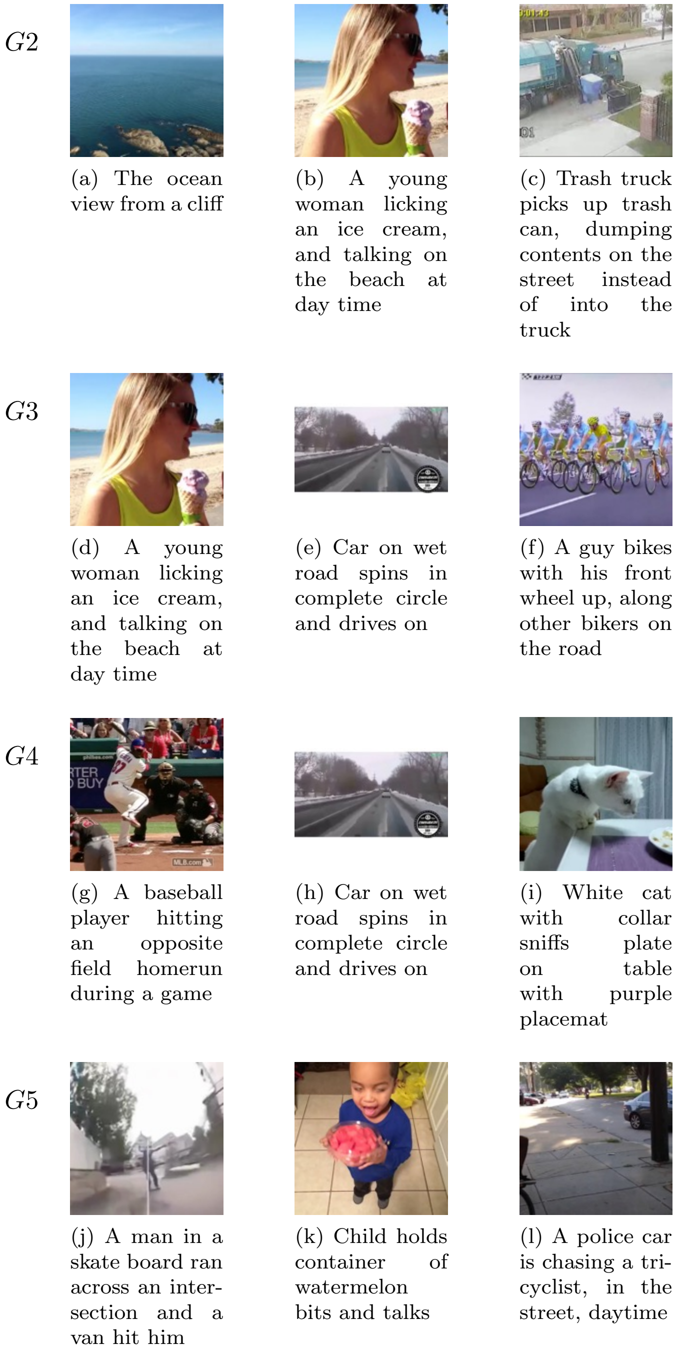
The description generation sub-task scoring was done using popular automatic metrics that compare the system generation captions with groundtruth captions as provided by assessors. We also used Direct Assessment this year to compare the submitted runs.

Figure 47 shows the comparison of all teams using the CIDEr metric. All runs submitted by each team are shown in the graph. Each team identified one run as their 'primary' run. Interestingly, the primary run was not necessarily the best run for each team.

For the remaining metrics, each run was scored separately for each group due to input limitation that the number of reference sentences need to be equal for all videos. Figure 48 shows the METEOR scores for the best runs for each team in each group. Figures 49 and 50 show the BLEU and STS results respectively.

Figure 51 shows the average DA score $[0-100]$ for each system. The score is micro-averaged per caption, and then averaged over all videos. Figure 52 shows the average DA score per system after it is standardized per individual AMT worker's mean and standard deviation score. The HUMAN-b system represents manual captions provided by assessors. As expected, captions written by assessors outperform the automatic systems. Figure 53 shows how the systems compare according to DA. The green squares indicate that the system in the row is significantly better than the system shown in the column. The figure shows that no system reaches the level of the human performance. Among the sytems, RUC_CMU clearly outperforms all the other systems.

Figure 54 shows the comparison of the various



(a) The ocean view from a cliff

(b) A young woman licking an ice cream, and talking on the beach at day time

(c) Trash truck picks up trash can, dumping contents on the street instead of into the truck

(d) A young woman licking an ice cream, and talking on the beach at day time

(e) Car on wet road spins in complete circle and drives on

(f) A guy bikes with his front wheel up, along other bikers on the road

(g) A baseball player hitting an opposite field homerun during a game

(h) Car on wet road spins in complete circle and drives on

(i) White cat with collar sniffs plate on table with purple placemat

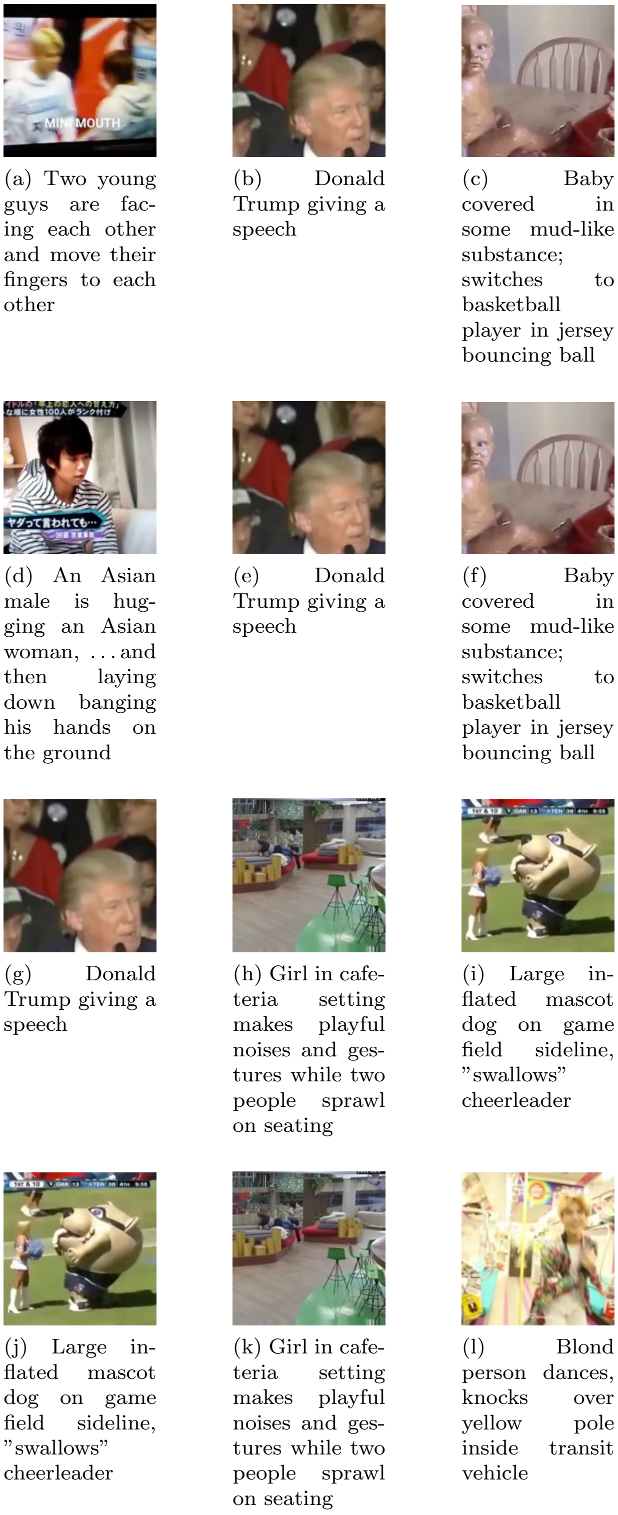
(j) A man in a skate board ran across an intersection and a van hit him

(k) Child holds container of watermelon bits and talks

(l) A police car is chasing a tricyclist, in the street, daytime

Figure 45: VTT: Rows 1 - 4 show the top 3 videos for G2 - G5 respectively. The video captions are from the manual groundtruth.

teams with respect to the different metrics used in the description generation subtask.



(a) Two young guys are facing each other and move their fingers to each other

(b) Donald Trump giving a speech

(c) Baby covered in some mud-like substance; switches to basketball player in jersey bouncing ball

(d) An Asian male is hugging an Asian woman, . . . and then laying down banging his hands on the ground

(e) Donald Trump giving a speech

(f) Baby covered in some mud-like substance; switches to basketball player in jersey bouncing ball

(g) Donald Trump giving a speech

(h) Girl in cafeteria setting makes playful noises and gestures while two people sprawl on seating

(i) Large inflated mascot dog on game field sideline, ”swallows” cheerleader

(j) Large inflated mascot dog on game field sideline, ”swallows” cheerleader

(k) Girl in cafeteria setting makes playful noises and gestures while two people sprawl on seating

(l) Blond person dances, knocks over yellow pole inside transit vehicle

Figure 46: VTT: Rows 1 - 4 show the bottom 3 videos for G2 - G5 respectively. The video captions are from the manual groundtruth.

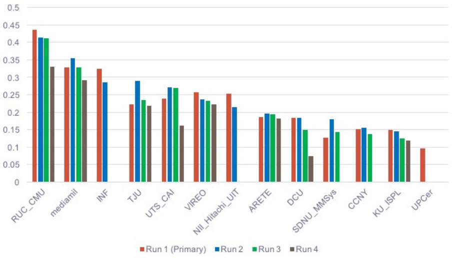


Figure 47: VTT: Comparison of all runs submitted by teams using the CIDEr metric

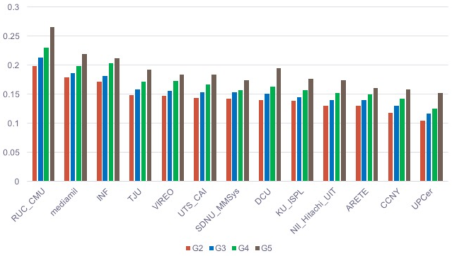


Figure 48: VTT: Comparison of the best run submitted by each team evaluated on each group using the METEOR metric

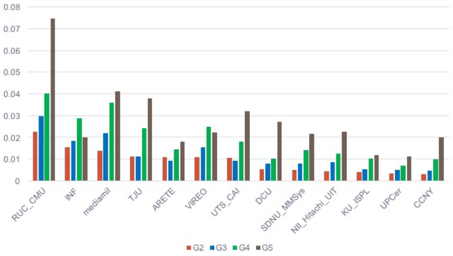


Figure 49: VTT: Comparison of the best run submitted by each team evaluated on each group using the BLEU metric

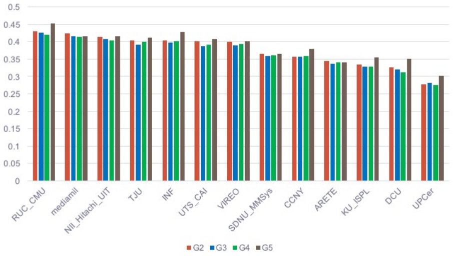


Figure 50: VTT: Comparison of the best run submitted by each team evaluated on each group using the STS metric

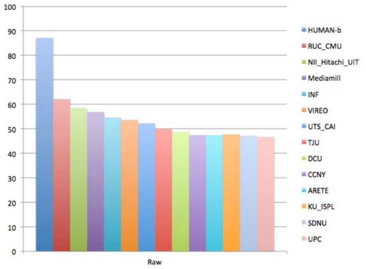


Figure 51: VTT: Average DA score for each system. The systems compared are the primary runs submitted, along with a manually generated set labeled as HUMAN-b

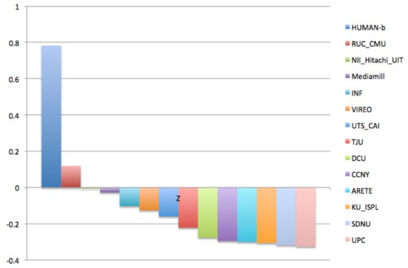


Figure 52: VTT: Average DA score per system after standardization per individual worker's mean and standard deviation score

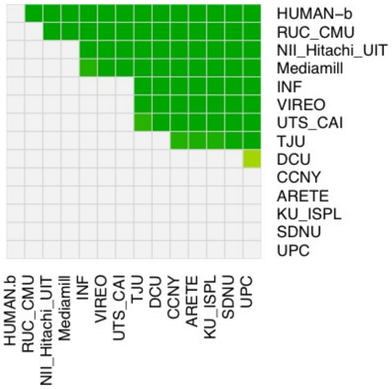
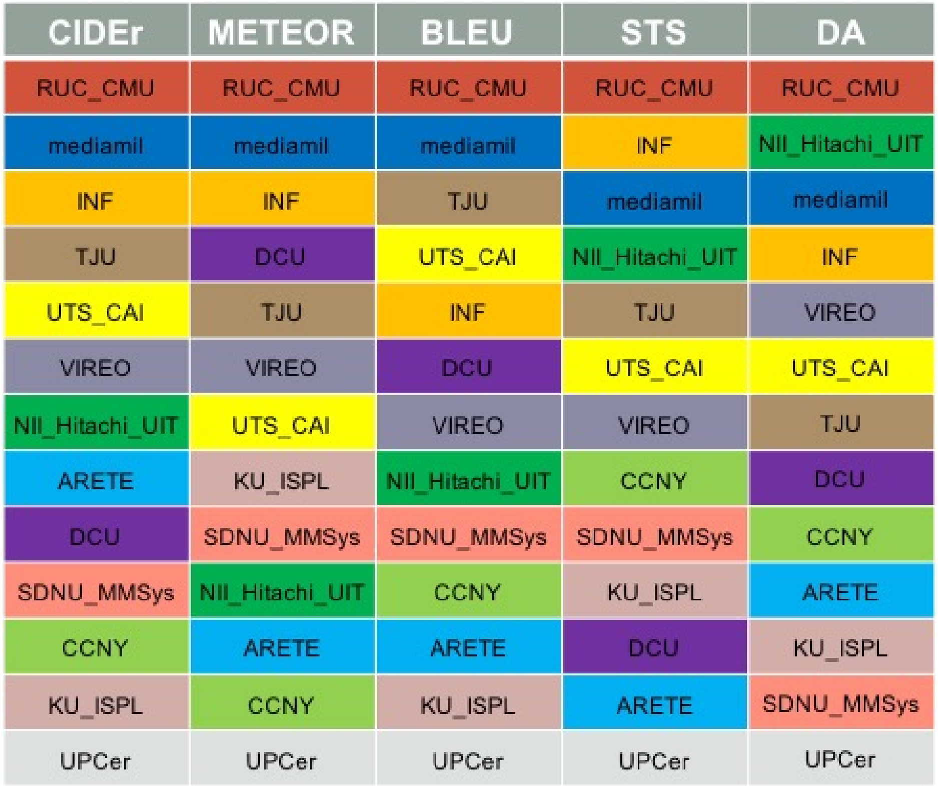


Figure 53: VTT: Comparison of systems with respect to DA. Green squares indicate a significantly better result for the row over the column



CIDEr	METEOR	BLEU	STS	DA
RUC_CMU	RUC_CMU	RUC_CMU	RUC_CMU	RUC_CMU
mediamil	mediamil	mediamil	INF	NII_Hitachi_UIT
INF	INF	TJU	mediamil	mediamil
TJU	DCU	UTS_CAI	NII_Hitachi_UIT	INF
UTS_CAI	TJU	INF	TJU	VIREO
VIREO	VIREO	DCU	UTS_CAI	UTS_CAI
NII_Hitachi_UIT	UTS_CAI	VIREO	VIREO	TJU
ARETE	KU_ISPL	NII_Hitachi_UIT	CCNY	DCU
DCU	SDNU_MMSys	SDNU_MMSys	SDNU_MMSys	CCNY
SDNU_MMSys	NII_Hitachi_UIT	CCNY	KU_ISPL	ARETE
CCNY	ARETE	ARETE	DCU	KU_ISPL
KU_ISPL	CCNY	KU_ISPL	ARETE	SDNU_MMSys
UPCer	UPCer	UPCer	UPCer	UPCer

Figure 54: VTT: Ranking of teams with respect to the different metrics for the description generation task

8.5 Conclusions and Observations

The number of teams participating in the VTT task increased this year, showing the interest in this area as computer vision algorithms continue to improve. The task this year evolved from last year's pilot owing to different number of manual descriptions. Each video was annotated by at least 2 assessors, and up to 5 assessors. This provided a richer dataset with varying number of captions per video. However, the variance in the number of descriptions resulted in extra submissions for the matching and ranking subtask, as well as different evaluations for some metrics in the description generation subtask. In the future, we will try to standardize the number of annotations per video in order to make the evaluation more uniform.

We also asked the assessors to rate the difficulty of describing each video. Only 33 of the 1880 videos were marked as hard, which did not provide much insight into determining the relationship between what was thought to be difficult by humans and systems.

This year, for the description generation subtask the CIDEr metric was used in addition to the other automatic metrics (BLEU, METEOR, and STS). Additionally, we also evaluated one run from each team using the direct assessment methodology, where humans rated how well a generated description matched the video.

During the creation of this task, we tried to remove redundancy and create a diverse set. This was done as a preprocessing step where videos were clustered based on similarity, and then a diverse set collected for annotation. Furthermore, videos which were given very similar captions by assessors were removed to create a dataset with little or no ambiguity for the matching subtask.

For the description generation subtask, in general systems scored higher on videos with higher number of annotations. This is the case since a larger number of groundtruth descriptions can result in the possibility of a higher number of word matches. Given that people can describe the same video in very different ways, a large number of annotations per video will help us evaluate systems better.

9 Summing up and moving on

This overview to TRECVID 2017 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 [TV17Pubs, 2017].

10 Authors' note

TRECVID would not have happened in 2017 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. Finally, Rob Cooper at BBC facilitated the copyright licences issues.
- Maria Eskevich, Roeland Ordelman, Gareth Jones, and Benoit Huet at Radboud University, University of Twente, Dublin City University, and EURECOM for coordinating the Video hyperlinking task.

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

11 Acknowledgments

The video hyperlinking work has been partially supported by: 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 Tim Finin and Lushan Han of University of Maryland, Baltimore County for providing access to the semantic similarity metric.

References

[Awad et al., 2016a] Awad, G., Fiscus, J., Joy, D., Michel, M., Kraaij, W., Smeaton, A. F., Quénot, G., Eskevich, M., Aly, R., Ordelman, R., Ritter, M., Jones, G. J., , Huet, B., and Larson, M. (2016a). TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA.

[Awad et al., 2016b] Awad, G., Snoek, C. G., Smeaton, A. F., and Quénot, G. (2016b). 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.

[Bojar et al., 2017] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages

169–214, Copenhagen, Denmark. Association for Computational Linguistics.

[Graham et al., 2017] Graham, Y., Awad, G., and Smeaton, A. (2017). Evaluation of Automatic Video Captioning Using Direct Assessment. *ArXiv e-prints.*

[Graham et al., 2016] Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2016). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

[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.

[TV17Pubs, 2017] TV17Pubs (2017). `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.17.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.

[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.

A Ad-hoc query topics

531 Find shots of one or more people eating food at a table indoors
532 Find shots of one or more people driving snowmobiles in the snow
533 Find shots of a man sitting down on a couch in a room
534 Find shots of a person talking behind a podium wearing a suit outdoors during daytime
535 Find shots of a person standing in front of a brick building or wall
536 Find shots of children playing in a playground
537 Find shots of one or more people swimming in a swimming pool
538 Find shots of a crowd of people attending a football game in a stadium
539 Find shots of an adult person running in a city street
540 Find shots of vegetables and/or fruits
541 Find shots of a newspaper
542 Find shots of at least two planes both visible
543 Find shots of a person communicating using sign language
544 Find shots of a child or group of children dancing
545 Find shots of people marching in a parade
546 Find shots of a male person falling down
547 Find shots of a person with a gun visible
548 Find shots of a chef or cook in a kitchen
549 Find shots of a blond female indoors
550 Find shots of a map indoors
551 Find shots of a person riding a horse including horse-drawn carts
552 Find shots of a person wearing any kind of hat
553 Find shots of a person talking on a cell phone
554 Find shots of a person holding or operating a tv or movie camera
555 Find shots of a person holding or opening a briefcase
556 Find shots of a person wearing a blue shirt
557 Find shots of person holding, throwing or playing with a balloon
558 Find shots of a person wearing a scarf
559 Find shots of a man and woman inside a car
560 Find shots of a person holding, opening, closing or handing over a box

B Instance search topics

9189 Find Peggy in this Cafe1

9190 Find Peggy in this LivingRoom 2

9191 Find Peggy in this Kitchen 2

9192 Find Billy in this Cafe1

9193 Find Billy in this Laundrette

9194 Find Billy in this Living Room 2

9195 Find Billy in this Kitchen 2

9196 Find Ian at this Cafe 1

9197 Find Ian in this Laundrette

9198 Find Ian in this Mini-Market

9199 Find Janine in this Cafe 1

9200 Find Janine in this Laundrette

9201 Find Janine in this Kitchen 2

9202 Find Janine in this Mini-Market

9203 Find Archie in this Laundrette

9204 Find Archie in this Living Room 2

9205 Find Archie in this Mini-Market

9206 Find Ryan in this Cafe 1

9207 Find Ryan in this Laundrette

9208 Find Ryan in this Kitchen 2

9209 Find Shirley in this Cafe 1

9210 Find Shirley in this Laundrette

9211 Find Shirley in this Living Room 2

9212 Find Shirley in this Kitchen 2

9213 Find Shirley in this Mini-Market

9214 Find Peggy in this Laundrette

9215 Find Phil in this Cafe 1

9216 Find Phil in this Living Room 2

9217 Find Phil at this Kitchen 2

9218 Find Phil in this Mini-Market