TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search

George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzad Godil, David Joy, Andrew Delgado, Alan Smeaton, Yvette Graham, et al.

TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search

George Awad {gawad@nist.gov} Asad A. Butt {asad.butt@nist.gov}
Keith Curtis {keith.curtis@nist.gov}
Yooyoung Lee {yooyoung@nist.gov} Jonathan Fiscus {jfiscus@nist.gov}
Afzal Godil {godil@nist.gov} David Joy {david.joy@nist.gov}
Andrew Delgado {andrew.delgado@nist.gov}
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, 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

Joao Magalhaes {jmag@fct.unl.pt}, David Semedo {df.semedo@campus.fct.unl.pt}
NOVA LINCS, Universidade NOVA de Lisboa, Portugal

Saverio Blasi {saverio.blasi@bbc.co.uk}
British Broadcasting Corporation (BBC R&D)

April 18, 2019

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2018 was a TREC-style video analysis and retrieval evaluation, the goal of which remains to promote progress in research and development of content-based exploitation and retrieval of information from digital video via open, metrics-based evaluation. Over the last eighteen 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 addi-

tion, many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2018 represented a continuation of three tasks from TRECVID 2017. In addition, three new pilot tasks: Social Media Video Storytelling Linking, Streaming Multimedia Knowledge-base Population, and Activities in Extended Video task were introduced. In total, 37 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. Streaming Multimedia Knowledge-base Population (SM-KBP)
4. Activities in Extended Video (ActEV)
5. Social Media Video Storytelling Linking (LNK)
6. Video to Text Description (VTT)

Table 2 represents 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 and 2017. 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 2017. While the video to text description task used 1 921 Twitter social media Vine videos collected through the online Twitter API public stream.

For the Activities in Extended Video task, about 7 hours of the VIRAT (Video and Image Retrieval and Analysis Tool) dataset was used which was designed to be realistic, natural and challenging for video surveillance domains in terms of its resolution, background clutter, diversity in scenes, and human activity/event categories.

About 200k images and videos were used from Twitter for development and testing by the Social Media Video Storytelling Linking task.

The new SM-KBP pilot task run by the TAC project asked participating systems to extract knowledge elements from a stream of heterogeneous documents containing multilingual multimedia sources including text, speech, images, videos, and pdf files; aggregate the knowledge elements from multiple documents without access to the raw documents themselves, and develop semantically coherent hypotheses, each of which represents an interpretation of the document stream. TRECVID participating teams only worked on the first part to extract knowledge elements from document streams.

The Ad-hoc search, instance search results were judged by NIST human assessors, while the Streaming Multimedia Knowledge-base Population task was assessed by human judges hired by the linguistic data consortium (LDC). The 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. Finally, the video storytelling linking results were assessed using Amazon Mechanical Turk workers.

The system submitted for the ActEV (Activities in Extended Video) evaluations were scored by NIST using reference annotations created by Kitware, Inc.

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 [TV18Pubs, 2018].

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. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, NIST, or the U.S. Government.

2 Datasets

2.1 BBC EastEnders Instance Search Dataset

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

Table 1: Participants and tasks

Task						Location	TeamID	Participants
IN	*VL*	*VT*	*MD*	*AE*	*AV*			
−−	−−	*VT*	−−	−−	−−	*Eur*	*PicSOM*	Aalto University
−−	*VL*	−−	−−	−−	−−	*Eur*	*ADAPT*	Adapt Centre School of Computer Science and Statistics of TCD
IN	−−	−−	−−	*AE*	−−	*Asia*	*BUPT_MCPRL*	Beijing University of Posts and Telecommunications
−−	−−	*VT*	**	*AE*	*AV*	*NAm + Asia*	*INF*	Carnegie Mellon University Shandong Normal University Renmin University Beijing University of Technology
−−	−−	*VT*	−−	−−	−−	*Aus*	*UTS_CETC_D2DCRC*	Centre for Artificial Intelligence, University of Technology Sydney
−−	**	*VT*	−−	−−	**	*Eur*	*EURECOM*	EURECOM
−−	−−	−−	**	−−	*AV*	*NAm*	*FIU_UM*	Florida International University University of Miami
−−	−−	−−	−−	−−	*AV*	*Asia*	*kobe_kindai*	Graduate School of System Informatics, Kobe University Department of Informatics, Kindai University
IN	−−	**	−−	*AE*	*AV*	*Eur*	*ITI_CERTH*	Information Technologies Institute / Centre for Research and Technology Hellas Queen Mary University of London
−−	−−	−−	−−	*AE*	−−	*NAm*	*JHUVAD*	Johns Hopkins University Amazon, Inc.
−−	−−	*VT*	−−	−−	−−	*Asia*	*kslab*	Knowledge Systems Laboratory, Nagaoka University of Technology
−−	−−	*VT*	−−	−−	−−	*Asia*	*KU_ISPL*	Korea University
−−	−−	−−	−−	*AE*	−−	*NAm*	*IBM − MIT − Purdue*	IBM;MIT;Purdue University
IN	−−	−−	−−	−−	−−	*Eur*	*IRIM*	Laboratoire d'Intgration des Systmes et des Technologies (CEA-LIST) Laboratoire Bordelais de Recherche en Informatique (LABRI) Laboratoire d'Informatique de Grenoble (LIG) Laboratoire d'Informatique pour la Mcanique et les Sciences de l'Ingnieur (LIMSI) Laboratoire d'Informatique, Systmes, Traitement de l'Information et de la Connaissance (LISTIC)
IN	−−	−−	−−	−−	−−	*Eur*	*PLUMCOT*	LIMSI KIT
−−	−−	−−	−−	−−	*AV*	*Asia*	*NECTEC*	National Electronics and Computer Technology Center NECTEC
IN	**	**	**	**	*AV*	*Asia*	*NII_Hitachi_UIT*	National Institute of Informatics, Japan Hitachi, Ltd., Japan University of Information Technology, VNU-HCMC, Vietnam
−−	−−	−−	**	−−	*AV*	*Asia*	*VIREO_NExT*	National University of Singapore City University of Hong Kong
IN	−−	−−	−−	−−	−−	*Asia*	*WHU_NERCMS*	National Engineering Research Center for Multimedia Software,Wuhan University
**	−−	*VT*	−−	−−	**	*SAm*	*ORAND*	ORAND S.A. Chile
IN	**	**	**	**	**	*Asia*	*PKU_ICST*	Peking University
−−	−−	*VT*	**	−−	−−	*Asia*	*NTU_ROSE*	Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University
**	−−	*VT*	−−	−−	*AV*	*Asia*	*RUCMM*	Renmin University of China
−−	−−	−−	−−	−−	*AV*	*Asia*	*NTU_ROSE_AVS*	ROSE LAB, NANYANG TECHNOLOGICAL UNIVERSITY
−−	**	*VT*	−−	**	−−	*Asia*	*MMsys_CCMIP*	Shandong Normal University Shandong University
−−	−−	−−	−−	−−	*AV*	*Eur*	*SIRET*	SIRET Department of Software Engineering, Faculty of Mathematics and Physics, Charles University
−−	−−	−−	−−	*AE*	−−	*Asia*	*SeuGraph*	Southeast University Computer Graphics Lab
−−	−−	−−	−−	*AE*	−−	*NAm*	*SRI*	SRI International
−−	−−	−−	−−	*AE*	−−	*NAm*	*STR*	Systems & Technology Research
−−	−−	−−	**	*AE*	−−	*Asia*	*VANT*	Tokyo Institute of Technology, National Institute of Advanced Industrial Science and Technology, Nanyang Technological University
−−	−−	−−	−−	*AE*	−−	*NAm*	*crcv*	UCF
−−	*VL*	−−	−−	−−	−−	*Eur*	*NOVASearch*	Universidade NOVA Lisboa
IN	−−	−−	−−	*AE*	−−	*Eur*	*HSMW_TUC*	University of Applied Sciences Mittweida; Chemnitz University of Technology
−−	−−	**	−−	*AE*	*AV*	*Eur*	*MediaMill*	University of Amsterdam
−−	**	*VT*	−−	−−	−−	*NAm*	*UCR_VCG*	University of California, Riverside
**	−−	−−	−−	*AE*	−−	*NAm*	*usfBULLS*	University of South Florida, Tampa
−−	−−	−−	−−	*AE*	−−	*NAm*	*UMD*	University of Maryland
−−	−−	−−	−−	*AE*	−−	*Aus*	*UTS − CETC*	University of Technology, Sydney
−−	−−	−−	−−	−−	*AV*	*Aus*	*UTS_ISA*	University of Technology Sydney
−−	−−	*VT*	−−	−−	−−	*Asia*	*UPCer*	UPC
−−	−−	**	**	−−	*AV*	*Asia*	*Waseda_Meisei*	Waseda University Meisei University

Task legend. IN:Instance search; MD:Streaming multimedia knowledge base population; VL:Video linking; VT:Video-to-Text; AE:Activities in Extended videos; 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	VL	VT	MD	AE	AV			
−−	−−	**	−−	−−	−−	NAm	$AreteEast$	Arete Associates
−−	−−	−−	−−	**	−−	$Asia$	$Mpl.bh$	Beihang university
−−	−−	**	−−	−−	−−	NAm	CMU_LSMA	Carnegie Mello University
**	−−	**	−−	−−	−−	Eur	$CEALIST$	Commissariat à l'énergie Atomique et aux énergies Alternatives Laboratoire d'Integration des Systemes et des Technologies
−−	−−	−−	−−	−−	**	$Asia$	$SogangDMV$	Dept. of Computer Science and Engineering, Sogang University
**	−−	−−	−−	−−	−−	$Asia$	U_TK	Dept. of Information Science & Intelligent Systems, The University of Tokushima
−−	−−	**	−−	−−	−−	Eur	$DCU.Insight$	Dublin City University
**	−−	−−	−−	−−	−−	NAm	$teamfluent$	Fluent.ai Inc.
**	**	**	**	**	**	$Asia$	GE	Graphic Era University
−−	−−	**	−−	−−	−−	$Asia$	$UDLT$	Tianjin University
−−	**	−−	−−	−−	−−	Eur	$SC4wTREC$	IBM Watson, IBM Ireland
−−	−−	−−	−−	−−	**	Eur	$ITEC\_UNIKLU$	Institute of Information Technology Klagenfurt University
−−	−−	−−	−−	−−	**	$Asia$	D_A777	Malla Reddy College of Engineering Technology, Department of Electronics and communication Engineering
−−	−−	−−	−−	**	−−	$Asia$	$TJUSMG$	Multimedia information processing center
**	**	**	**	**	**	$Asia$	ZJU_612	net media lab of ZJU
−−	−−	−−	−−	**	−−	NAm	$nVIDIA\_CamSol$	nVIDIA
−−	**	−−	−−	−−	−−	Eur	$EURECOM\_POLITO$	Politecnico di Torino and Eurecom
−−	−−	**	−−	−−	−−	$NAm + Asia$	RUC_CMU	Renmin Unversity of China Carnegie Mellon University
−−	−−	−−	−−	**	−−	NAm	sbu	Stony Brook University
**	−−	**	**	**	**	$Asia$	MDA	Department of electronic engineering, Tsinghua University
−−	**	**	−−	−−	−−	$Asia$	tju_nus	Tianjin University, China SeSaMe Research Centre, National University of Singapore, Singapore
−−	**	−−	−−	−−	−−	Eur	$IRISA$	Université de Rennes 1
−−	−−	−−	−−	**	−−	Eur	$IRIMAS_UHA_CrowdSurv$	University of Haute-Alsace
−−	−−	−−	**	−−	−−	Afr	$REGIMVID$	University of sfax
−−	−−	−−	−−	**	−−	Afr	$UJCV$	University of Johannesburg
−−	−−	−−	−−	−−	**	Eur	$vitrivr$	University of Basel, Switzerland

Task legend. IN:instance search; MD:Streaming multimedia knowledge base population; VL:Video linking; VT:Video-to-Text; AE:Activities in extended videos; AV:Ad-hoc search; −−:no run planned; ∗∗:planned but not submitted

additional metadata.

2.2 Internet Archive Creative Commons (IACC.3) Ad-hoc Search Videos

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 min to 9.5 min and a mean duration of ≈7.8 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 and 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 Activity Detection VIRAT Dataset

The VIRAT Video Dataset [Oh et al., 2011] is a large-scale surveillance video dataset designed to assess the performance of activity detection algorithms in realistic scenes. The dataset was collected to facilitate both detection of activities and to localize the corresponding spatio-temporal location of objects associated with activities from a large continuous video. The stage for the data collection data was a group of buildings, and grounds and roads surrounding the area. The VIRAT dataset are closely aligned with real-world video surveillance analytics. In addition, we are also building a series of even larger multi-camera datasets, to be used in the future to organize a series of Activities in Extended Video (ActEV) challenges. The main purpose of the data is to stimulate the computer vision community to develop advanced algorithms with improved performance and robustness of human activity detection of multi-camera systems that cover a large area.

2.4 SM-KBP task multimedia data

The Linguistic Data Consortium (LDC) distributed a set of about 10 000 training corpus documents including at least 1200 to 1500 topic-relevant and/or scenario relevant documents. For the 2018 pilot, the scenario was the Russian/Ukrainian conflict (2014-2015). In addition, a set of 6 training topics were also distributed. Documents in general included images, videos, web pages in text format, tweets, audio and pdf files. A set of 3 topics were used for evaluation along with 10 000 testing documents.

2.5 Social Media Video Storytelling Linking data

The data for the following events was crawled (Table 9):

The Edinburgh Festival (EdFest) consists of a celebration of the performing arts, gathering dance, opera, music and theatre performers from all over the world. The event takes place in Edinburgh, Scotland and has a duration of 3 weeks in August.

Le Tour de France (TDF) is one of the main road cycling race competitions. The event takes place in France (16 days), Spain (1 day), Andorra (3 days) and Switzerland (3 days).

The development data covers the 2016 editions of the above events and for each event there's 20 stories.The test data covers the 2017 editions of the above events and for each event there's 15 stories.

2.6 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 1 903 URLs was distributed to participants of the video-to-text pilot task. The 2016 and 2017 pilot testing data were also available for training (a set of about 3800 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 in 2016. 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 probability 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 2018 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.

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

Three 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.
- Relevance-Feedback: System takes the official query as input and produce initial results, then a human judge can assess the top-5 results and input this information as a feedback to the system to produce a final set of results. This feedback loop is strictly permitted only once.

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 (≈150) 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 Ad-hoc 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 do not differentiate between a clip and a shot and thus they may be used interchangeably.

3.2 Evaluation

Each group was allowed to submit up to 4 prioritized main runs per submission type and two additional

if they were “no annotation” runs. In fact 13 groups submitted a total of 52 runs, from which 16 runs were manually-assisted, 33 were fully automatic runs and 2 relevance-feedback.

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 presented with the pools - one assessor per topic - 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, 92 622 clips were judged while 380 835 clips fell into the unjudged part of the overall samples. Total hits across the 30 topics reached 7381 with 5635 hits at submission ranks from 1 to 100, 1469 hits at submission ranks 101 to 150 and 277 hits at submission ranks between 151 to 1000.

3.3 Measures

The *sample_ eval* software [2], 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 Ad-hoc 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 "Two or more people wearing coats", "a person in front of or inside a garage", "exactly two men at a conference or meeting table talking in a room", "people waving flags outdoors", and "truck standing still while a person is walking beside or in front of it".

On the other hand, the bottom 5 found queries were "person playing keyboard and singing indoors", "two or more cats both visible simultaneously", "person sitting on a wheelchair", "car driving scenes in a rainy day", and "a dog playing outdoors". 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 and 2017.

Top performing manually-assisted runs were as well 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.

Figures 3 and 4 show the results of all the 16 manually-assisted and 33 fully automatic run submissions respectively.

This year the max and median scores (10 to 12 % mean Inferred average precision) are significantly lower than 2017 for both run submission types but better than 2016 with the exception of manual runs in 2016 where their max score are better than 2018. We should also note here that only 1 run were submitted under the training category of E, while there was 0 runs using category F 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 to 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 including their interactions.

The two relevance feedback run types scored between 1.6 % to 1.8 % mean inferred average precision with no statistical significant difference between the two of them. As it is the first year to introduce such run types, there is not much conclusions that can be drawn about those runs.

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. Some samples of top queries are highlighted in green while samples

[2] http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/sample_eval/

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 boat, flags, white coat, horn, bed, etc) compared to the bottom ones which require more temporal analysis for some activities and combination of one or more facets of who,what and where/when (e.g dancing outdoors at night, climbing, gate in the background, holding a robe, standing in line, etc). In general there is a noticeable spread in score ranges among the top 10 runs specially with high performing topics which may indicate the variation in the performance of the used techniques and that there is still room for further improvement. However for low performing topics, usually all top 10 runs are condensed together with low spread between their scores.

In order to analyze which topics in general were the most easy or difficult we sorted topics by number of runs that scored xInfAP $>= 0.3$ for any given topic and assumed that those were the easiest topics, while xInfAP < 0.1 indicates a hard topic. In retrospective, when we tested the same threshold used in 2017 (0.7) it turned out that all 2018 topics were hard. Therefore, we had to reduce the used threshold to 0.3 and 0.1 to cluster the hard/easy topics.

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 easily identifiable visual 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 $p<0.05$. These figures 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, there is no significant difference between the 3 top manually assisted runs.

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 8 % (minimum 6 %) of submitted clips are common between any pair of runs. In comparison, the average was about 15 % in the previous year. These results suggest that although most submitted runs are diverse, systems compared to last year may be more diverse in their approaches or trained on different training data corpus.

3.5 Ad-hoc Observations and Conclusions

This year concluded 1-cycle of three years of the Ad-hoc task applying generic queries on internet videos with stable team participation and completion percentage. The task is still more difficult than simple concept-based tagging with the observation that 2018 queries seem to be the most hard compared to 2016 or 2017. Maximum and median scores for manually-assisted and fully automatic runs are lower than 2017 with automatic runs performing slightly better than manually-assisted runs which is little surprising, although encouraging, as the last two years manually-assisted runs performed better than automatic. Among high scoring topics, there is more room for improvement among systems, while among low scoring topics, most systems' scores are collapsed in small narrow range. 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.

A summary of general approaches by team can be drawn to show that many methods relied on concept banks integration, using different strategies for concept selections, linguistic analysis of queries, and joint image and text representation (embedding approaches) spaces using available image/video captioning datasets. Deep learning approaches dominated teams' methods and used pretrained models.

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV18Pubs, 2018] in the online workshop notebook proceedings.

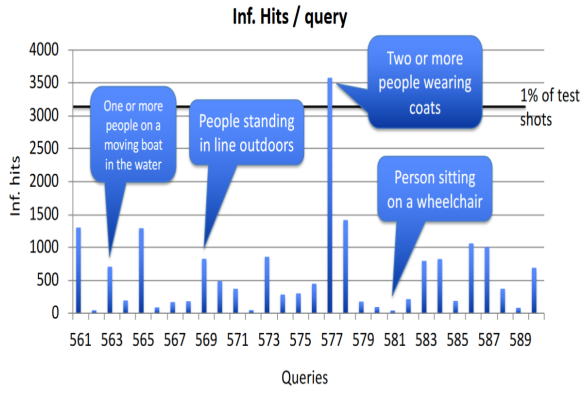


Figure 1: AVS: Histogram of shot frequencies by query number

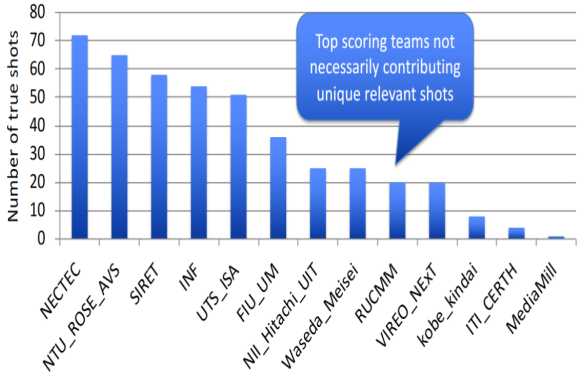


Figure 2: AVS: Unique shots contributed by team

4 Instance search

An important need in many situations involving video collections (archive video search/reuse, personal video organization/search, surveillance, law enforcement, protection of brand/logo use) is to find more video segments of a certain specific person, object, or place, given one or more visual examples of the specific item. Building on work from previous years in the concept detection task [Awad et al., 2016] the instance search task seeks to address some of these needs. For 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 Instance Search Data

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

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

4.2 System task

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

Each query consisted of a set of:

- The name of the target person

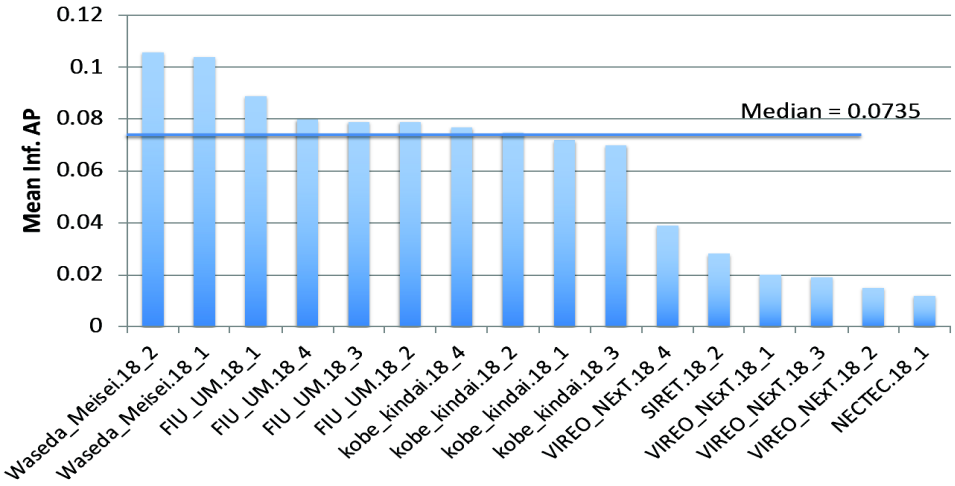


Figure 3: AVS: xinfAP by run (manually assisted)

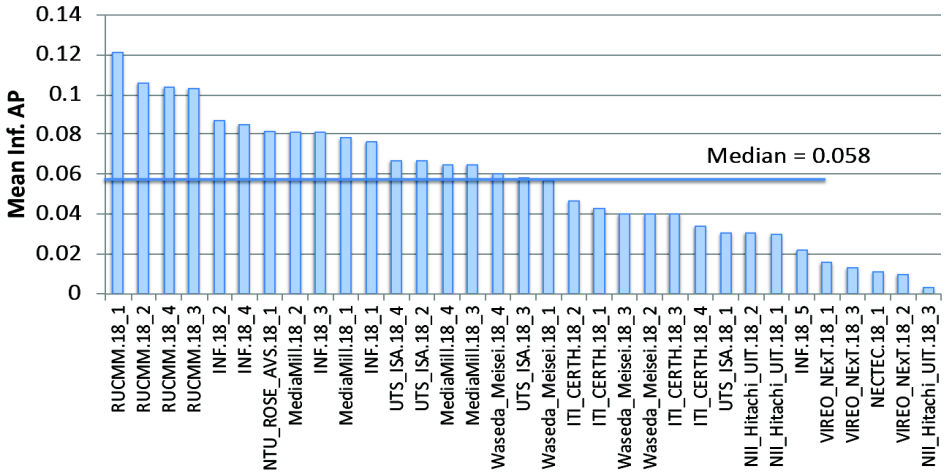


Figure 4: AVS: xinfAP by run (fully automatic)

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 %
9219	38720	7649	19.75	520	4457	58.27	850	19.07
9220	39054	9717	24.88	520	5426	55.84	456	8.40
9221	37301	7977	21.39	520	4729	59.28	73	1.54
9222	38635	9597	24.84	520	5645	58.82	132	2.34
9223	39018	7719	19.78	520	4040	52.34	658	16.29
9224	38750	6330	16.34	520	3085	48.74	173	5.61
9225	38488	8967	23.30	520	5561	62.02	105	1.89
9226	39164	6556	16.74	520	3361	51.27	1165	34.66
9227	36794	7861	21.36	520	4552	57.91	55	1.21
9228	38730	8872	22.91	520	5512	62.13	51	0.92
9229	39226	7148	18.22	520	3638	50.89	819	22.51
9230	37920	7895	20.82	520	4593	58.18	30	0.65
9231	38532	9098	23.61	520	5532	60.80	144	2.60
9232	39029	7787	19.95	520	4556	58.51	928	20.37
9233	37915	8083	21.32	520	4447	55.02	84	1.89
9234	38835	8300	21.37	520	4573	55.10	135	2.95
9235	39036	8225	21.07	520	4437	53.94	675	15.21
9236	38300	7922	20.68	520	4358	55.01	163	3.74
9237	38815	8081	20.82	520	4510	55.81	376	8.34
9238	38217	7523	19.68	520	4387	58.31	57	1.30
9239	37347	5555	14.87	520	2927	52.69	431	14.72
9240	30727	7481	24.35	520	4019	53.72	442	11.00
9241	30382	4017	13.22	520	2122	52.82	1195	56.31
9242	29996	4506	15.02	520	1924	42.70	654	33.99
9243	31000	6204	20.01	520	3233	52.11	1340	41.45
9244	29413	7517	25.56	520	4181	55.62	68	1.63
9245	28855	7409	25.68	520	4375	59.05	51	1.17
9246	30484	8228	26.99	520	4600	55.91	240	5.22
9247	29493	8636	29.28	520	4959	57.42	49	0.99
9248	28662	8056	28.11	520	4378	54.34	118	2.69

- 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 (+ optional image examples)

Each run was also required to state the source of the training data used. This year participants were allowed to use training data from an external source, instead of, or in addition to the NIST provided training data. The following are the options of training data to be used:

A Only sample video 0
B Other external data
C Only provided images/videos in the query
D Sample video 0 AND provided images/videos in the query (A+C)

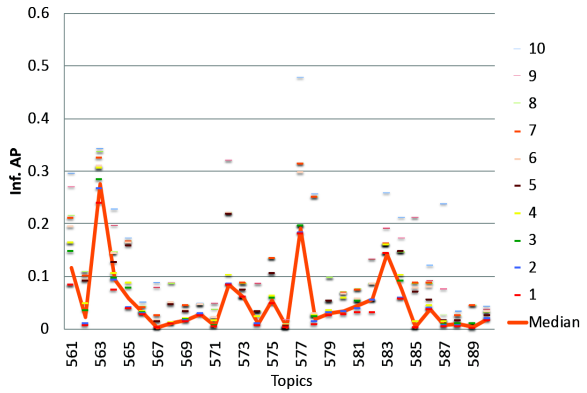


Figure 5: AVS: Top 10 runs (xinfAP) by query number (manually assisted)

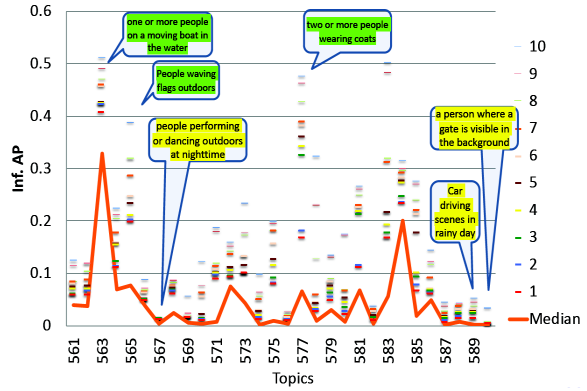


Figure 6: AVS: Top 10 runs (xinfAP) by query number (fully automatic)

E External data AND NIST provided data (sample video 0 OR query images/videos)

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 10 target persons across 4 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). In total, 8 groups submitted 31 automatic and 9 interactive runs (using only the first 21 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[3], 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 128 117 judged shots including 11 717 total relevant shots. Table 3 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 of interest here, of these two, only speed was reported.

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

Top 10 Easy (sorted by count of runs with InfAP >= 0.3)	Top 10 Hard (sorted by count of runs with InfAP < 0.1)
Find shots of one or more people on a moving boat in the water	Find shots of two people fighting
Find shots of two or more people wearing coats	Find shots of a person holding or attached to a rope
Find shots of a person holding, talking or blowing into a horn	Find shots of one or more people hiking
Find shots of people waving flags outdoors	Find shots of car driving scenes in a rainy day
Find shots of two or more cats both visible simultaneously	Find shots of people performing or dancing outdoors at nighttime
Find shots of a person lying on a bed	Find shots of a person where a gate is visible in the background
Find shots of a person in front of or inside a garage	Find shots of people standing in line outdoors
	Find shots of a dog playing outdoors
	Find shots of a person holding his hand to his face

Figure 7: AVS: Easy vs Hard topics

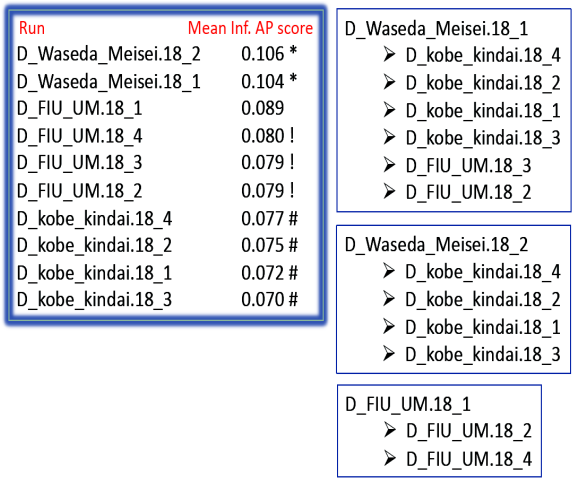


Figure 8: AVS: Statistical significant differences (top 10 manually-assisted runs). The symbols #,! and * denotes that there is no statistical significance between those runs for a given team

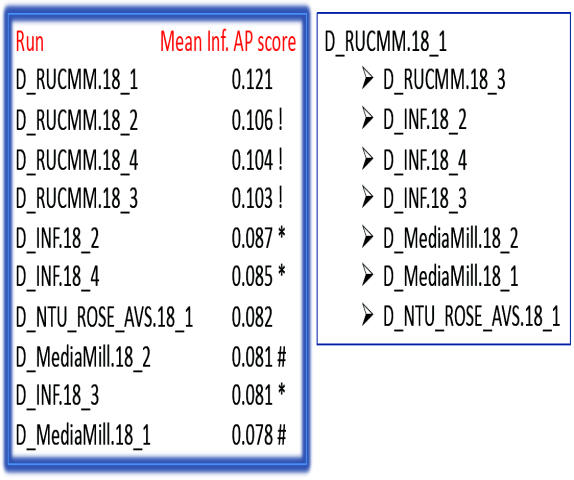


Figure 9: AVS: Statistical significant differences (top 10 fully automatic runs). The symbols #,! and * denotes that there is no statistical significance between those runs for a given team

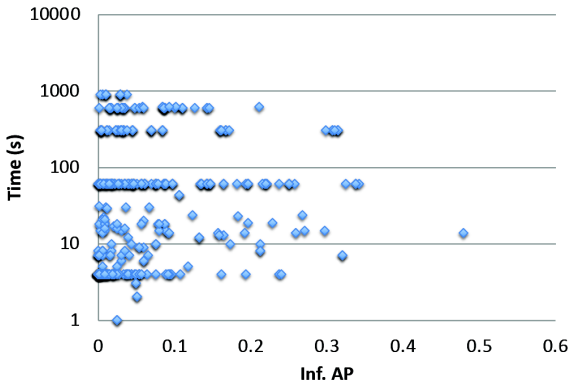


Figure 10: AVS: Processing time vs Scores (Manually assisted)

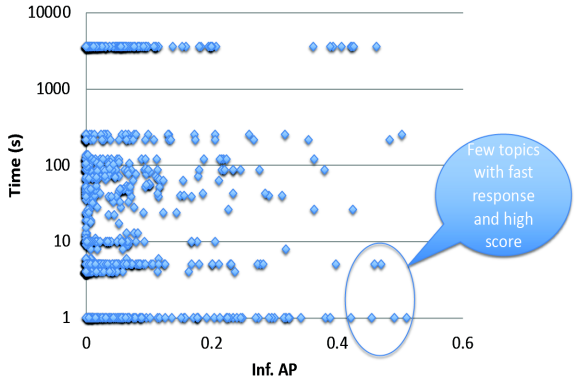


Figure 11: AVS: Processing time vs Scores (fully automatic)

4.6 Instance Search Results

Figures 12 and 13 show the sorted scores of runs for automatic and interactive systems respectively. Both set of results show a slight decrease in best performances compared to 2017 and an increase in worst performances, with the mean results remaining around the same. Specifically the maximum score in 2018 for automatic runs reached 0.463 compared to 0.549 in 2017 and maximum score in 2017 for interactive runs reached 0.524 compared to 0.677 in 2017. However, the lowest score in 2018 for automatic runs reached 0.096 compared to 0.037 in 2017.

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 0.6172 down to 0.1347. Two main factors might be expected to affect topic difficulty: the target person and the location. From the analysis of the performance of topics, it can be shown that for example the persons "Zainab", "Heather" and "Garry" were easier to find as 3 "Zainab" topics were among the top 15 topics with none in the bottom 15 topics. Also, 2 "Heather" and "Garry" topics were among the top 15 topics compared to only 1 in the bottom 15 topics. On the other hand the target persons "Chelsea", "Jane" and "Mo" 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 "Pub" made it harder to find the target persons as 5 out of the bottom 15 topics were at the location "Pub" compared to only 2 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 $p < 0.05$. There are little significant differences between the top runs this year.

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. Two teams (HSMW_TUC and IRIM) reported processing times of 0 s. The next lowest reported processing time is 81 s. This raises the question of whether or not the lowest processing times were reported correctly.

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. The location "Pub" seems to be slightly easier when compared to automatic 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 "Garry" and "Zainab" as they are still easy to spot, while "Chelsea" and "Jane" are still among the most difficult.

Figure 17 shows the results of a partial randomization test for the 9 submitted interactive runs. 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). This shows much more significance between the interactive runs than for the top 10 automatic runs.

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 that a lot less teams are using video examples compared to previous years. We feel that the use of video examples is to be encouraged in order to take advantage of the full video frames for better training data instead of just a few images.

Figure 20 shows the effect of the data source used for training, this year being the first in which participants are able to use an external data source instead of or in addition to the NIST provided training data. The use of external data in addition to the NIST provided data gives by far the best results. These are followed by the NIST provided data which gives the next best results. By comparison, the use of external data instead of the NIST provided data gives quite poor results.

4.7 Instance Search Observations

This is the third 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 number of finishers remained the same, given a slight increase in the percentage of finishers.

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. This was the first year in which participants were allowed to use external data instead of, or in addition to the NIST provided data. Results have shown that the use of external data in addition to the NIST provided data consistently gives best results. However, results also show that the use of external data instead of the NIST provided data gives quite poor results.

To summarize the main approaches taken by different teams, NII_Hitachi_UIT team focused on improving the person INS module using FaceNet and on improving the location INS module using Bag of Words (BoW) and Deep Image Retrieval (DIR). They also propose Progressive Extension and Intersection Pushing (PEIP) to obtain suggestive returned shots. This method consists of multiple iteration processes. 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 it is essential that they change their fusion strategy. The IRIM team used a combination of two person recognition methods and one location recognition method. Late fusion was performed on the person recognition methods, followed by additional late fusion to mix in the location scores. Their main focus this year was on the person recognition aspect, with data augmentation and faces re-ranking having a positive impact. The ICST team used Bag of Words (BoW) and convolutional neural networks (CNN) for location search, query pre-processing based on super resolution, deep models for face recognition, and text based refinement for person search. Their fusion method was based on a combination of score and rank based fusion.

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV18Pubs, 2018] in the online workshop notebook proceedings.

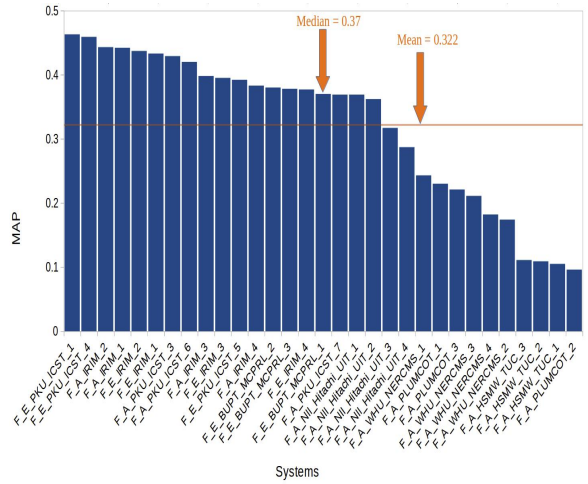


Figure 12: INS: Mean average precision scores for automatic systems

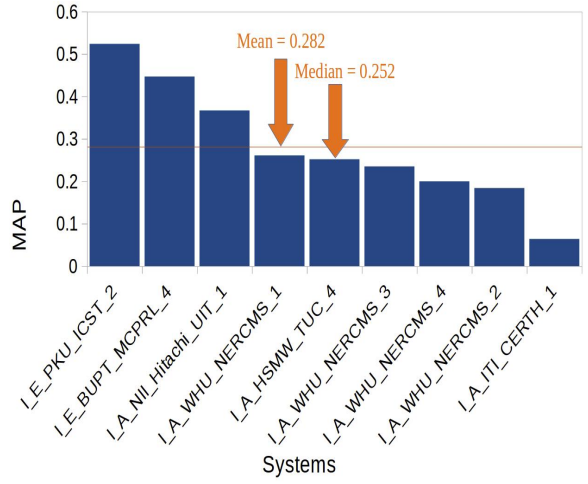


Figure 13: INS: Mean average precision scores for interactive systems

5 Streaming Multimedia Knowledge Base Population

The 2018 Streaming Multimedia Knowledge Base Population (SM-KBP) evaluation is a new pilot task jointly run by the Text Analysis Conference. The task tries to address the need for technologies to analyze and extract knowledge from multimedia to support answering questions and queries to respond to the situations such as natural disasters or international conflicts. In such situations, analysts and the public are often confronted with a variety of information coming through multiple media sources. The streaming multimedia extraction task asks systems to extract knowledge elements (KEs) from heterogeneous multimedia sources such as text documents, images, videos, audio, social media sites, etc. Although the big picture of the task is to use those knowledge elements to populate a knowledge base and later on to answer questions, TRECVID participants only had the opportunity to work on the first stage (denoted here as "TA1") of the task and mainly to analyze the video data stream to extract detectable knowledge elements based on a provided ontology.

5.1 Task Definition

TA1 systems are expected to process one document at a time (single document processing) and produce a set of KEs for each input document from the document stream. This is referred to as a document-level knowledge graph. A knowledge graph (KG) represents all knowledge, whether it comes from the document stream or some shared background knowledge, or via insertion of knowledge by a human user. A KE is a node or edge in a knowledge graph. Knowledge element types are defined in the ontology. A node in the knowledge graph represents an Entity/Filler, Event, or Relation, and an edge links an event or relation to one of its arguments. A KE represents a single entity, a single event or a single relation. The KE maintains a cluster (or a node) of all mentions from within the document of the same entity, and a cluster (or node) of one or more mentions of each event or relation. An entity cluster should group together entity mentions that are referring to the same real-world entity. The same is true with events and relation clusters, though the definition of equality (coreference) may be fuzzier than for entities.

A document may contain document elements in multiple modalities and multiple languages; there-

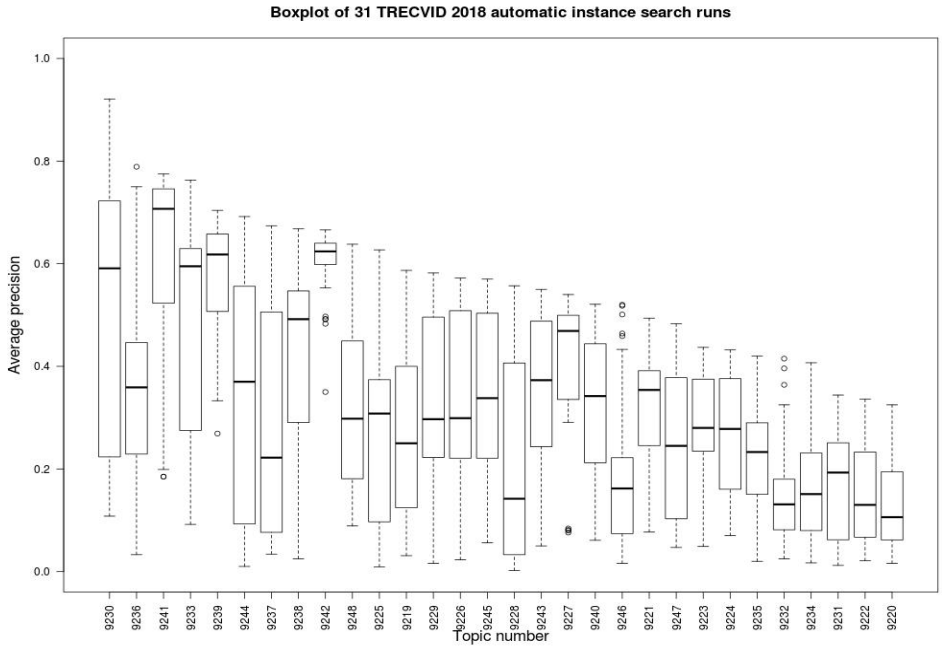


Figure 14: INS: Boxplot of average precision by topic for automatic runs.

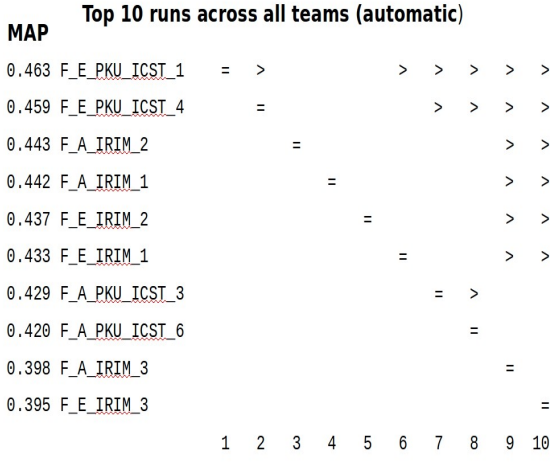


Figure 15: INS: Randomization test results for top automatic runs. "E":runs used video examples. "A":runs used image examples only.

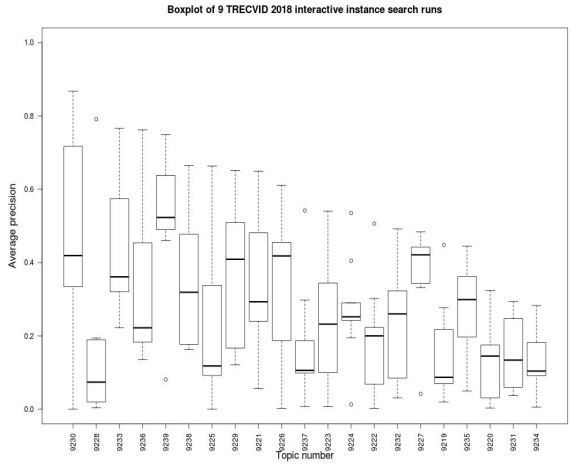


Figure 16: INS: Boxplot of average precision by topic for interactive runs

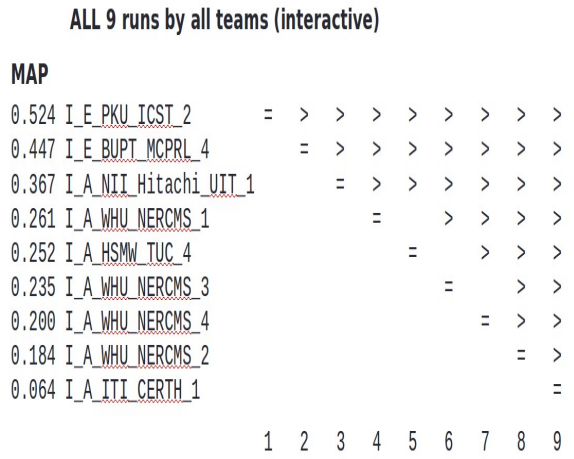


Figure 17: INS: Randomization test results for top interactive runs. "E":runs used video examples. "A":runs used image examples only.

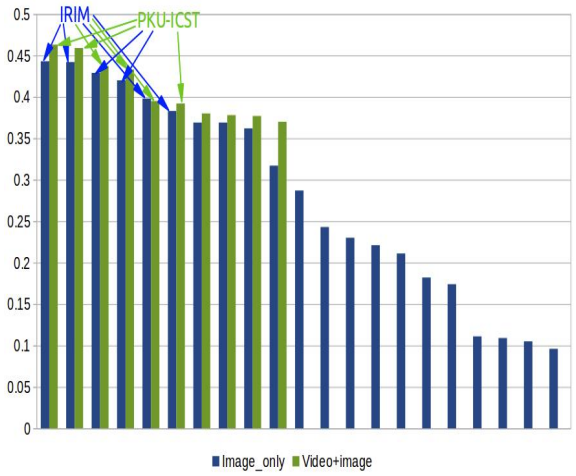


Figure 19: INS: Effect of number of topic example images used

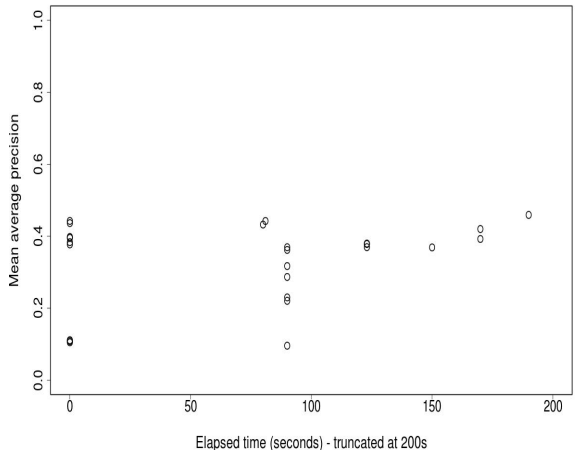


Figure 18: INS: Mean average precision versus time for fastest runs

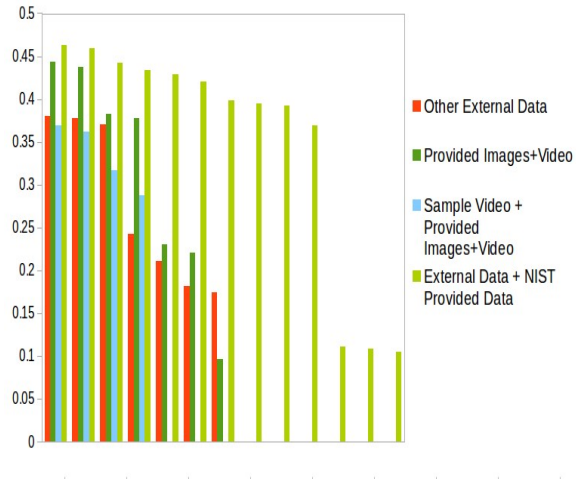


Figure 20: INS: Effect of data source used

fore, cross-lingual and cross-modal entity and event coreference are required. Conceptually, TA1 system must process each document in the order given in the document stream and must freeze all output for a document before starting to process the next document in the stream; however, because TA1 is stateless across documents (i.e., TA1 must process each document independently), in practice for the pilot evaluation, TA1 may choose to parallelize processing of documents for efficiency. NIST will evaluate output for only selected documents in the data stream, via pooling and assessment.

5.2 SM-KBP Data

For the 2018 pilot, the conflict scenario chosen was the Russian/Ukrainian conflict (2014-2015). A training corpus of 10 000 documents were released by LDC and included at least between 1200 and 1500 topic-relevant and/or scenario relevant documents. The training corpus included data addressing a set of 6 training topics as follows:

- Crash of Malaysian Air Flight MH17 (July 17, 2014)
- Flight of Deposed Ukrainian President Viktor Yanukovych (February 2014)
- Who Started the Shooting at Maidan? (February 2014)
- Ukrainian War Ceasefire Violations in Battle of Debaltseve (January-February 2015)
- Humanitarian Crisis in Eastern Ukraine (July-August 2014)
- Donetsk and Luhansk Referendum, aka Donbas Status Referendum (May 2014)

One evaluation topic and about 10 000 documents were released as testing data:

- Siege of Sloviansk and Battle of Kramatorsk (April-July 2014)

The distributed corpus included different modalities such as videos, images, html web pages, tweets, audio and pdf files. For all the video data, NIST released a shot boundary reference table that maps each whole video to several shot segments to be used by systems in their run submissions.

Task participants also received an ontology of entities, events, event arguments, relations, and SEC (sentiment, emotion, and cognitive state), defining the KEs that are in scope for the evaluation tasks.

5.3 Evaluation Queries

NIST distributed a set of evaluation queries to TA1 participants to apply to their output knowledge graphs. The queries in general tested a system for its effectiveness in determining the presence of a knowledge element or knowledge graph in the document collection, where a document may contain multiple document elements, and each document element can be text, video, image, or audio. Broadly, queries may be one of three types:

- Class level queries: The query will provide a type from the ontology, and the teams will be asked to return all mentions of the class corresponding to the given type (e.g Person, Organization, Geopolitical Entity, Facility, Location, Weapon, Vehicle).
- Instance level queries (a.k.a. "zero-hop queries"): The query will provide a mention of an entity or filler from the ontology, and the teams will be asked to return all mentions of that particular entity/filler. For e.g., the query may ask for all mentions of "Jack Bauer" referred to in document 32 at offset 100-110.
- Graph queries: The query will be composed of a combination of ontology types and their instances and ask for a connected graph with at least one edge.

Teams were provided queries in two formats, which are intended to be semantically equivalent:

- Simplified: Simplified query in an XML format that teams may apply to their KBs using any automatic technique that they choose. These simplified queries will be expressed in the domain ontology and are intended to be human-readable but will not be executable using standard tools.
- Executable: Executable SPARQL query that teams should apply using a dockerized tool provided by NIST; subsequently, NIST would use the same tool to apply executable queries in a uniform way to all KBs from all teams.

5.4 Measures

Teams were asked to submit the whole knowledge base (KB) in addition to xml response files to the evaluation queries. All responses required a justification grounded in the documents (e.g text span,

video shot, image ID, etc). All mentions returned in response to a class-based query requesting a particular type (e.g., "Location") or to a zero-hop query requesting mentions of a particular entity/filler (e.g., "Vladimir Putin") in a "core" subset of the evaluation documents, were assessed by LDC for correctness. Graphs returned in response to graph queries are broken into assessment triples (subject justification, object justification, predicate justification) for assessment by LDC. Evaluation scores are based on F1 of Precision and Recall. For more details on the guidelines and evaluation measures and procedures please refer to the detailed evaluation plan of the task(s) provided by TAC [4]

6 Activities in Extended Video

NIST developed the Activities in Extended Video (ActEV) evaluations series to support the Intelligence Advanced Research Projects Activity (IARPA) Deep Intermodal Video Analytics (DIVA) Program. ActEV is an extension of the TRECVID Surveillance Event Detection (SED) [Michel et al., 2017] evaluations where systems only detected and temporally localized activities. The goal of ActEV is to evaluate performance of video analytic technologies that automatically detect a target activity and identify and track objects associated with the activity via a task-driven evaluation. With both retrospective analysis and real-time analysis applications in mind, the challenges include activity detection in a multi-camera streaming environment and temporal (and spatio-temporal) localization of the activity for reasoning. To understand current state-of-the-art, we initiated the ActEV18 challenge that includes three evaluations: activity-level, reference temporal segmentation, and leaderboard.

6.1 Tasks and Measures

The purpose of the ActEV challenge is to promote the development of systems that automatically: 1) identify a target activity along with the time span of the activity (activity detection), 2) detect objects associated with the activity occurrence (activity and object detection), and 3) track multiple objects associated with the activity instance (activity and object detection and tracking). In the following subsections, we define each task and describe its performance measure.

[4] https://tac.nist.gov/2018/SM-KBP/guidelines/SM-KBP_2018_Evaluation_Plan_V0.8.pdf

Activity Detection (AD)

Given a target activity, a system automatically detects and temporally localizes all instances of the activity in extended video sequences. The system should provide the start and end frames indicating the temporal location of the target activity and a presence confidence score with higher values indicating the instance is more likely to have occurred. To evaluate system performance, we utilized the streaming detection evaluation protocols and metrics from TRECVID: SED [Michel et al., 2017] and Classification of Events Activities and Relationships (CLEAR) [Bernardin and Stiefelhagen, 2008]. The primary metric evaluates how accurately the system detected the occurrences of the activity. The scoring procedure between reference and system output can be divided into four distinctive steps: 1) instance alignment, 2) confusion matrix calculation, 3) summary performance metrics, and 4) graphical analysis of the Type I/II error trade-off space. The goal of the alignment step is to find a one-to-one correspondence of the instances between the reference and the system output. This step is required because in the unsegmented, streaming detection scenario, activities can occur at any time for any duration. We utilize the Hungarian algorithm [Munkres, 1957] to find an optimal mapping while reducing the computational complexity—this is covered in further detail in the equations in the evaluation plan [Lee et al., 2018]. The next step is to calculate the detection confusion matrix for activity instance occurrence. Correct Detection (CD) indicates that the reference and system output instances are correctly mapped. Missed Detection (MD) indicates that an instance in the reference has no correspondence in the system output while False Alarm (FA) indicates that an instance in the system output has no correspondence in the reference. After calculating the confusion matrix, we summarize system performance: for each instance, a system output provides a confidence score that indicates how likely the instance is associated with the target activity. The confidence score can be used as a decision threshold, enabling a probability of missed detections (P_{Miss}) and a rate of false alarms (R_{FA}) to be computed at a given threshold:

$$P_{\text{miss}}(\tau) = \frac{N_{\text{MD}}(\tau)}{N_{\text{TrueInstance}}}$$

$$\mathrm{R_{FA}}(\tau) = \frac{N_{\mathrm{FA}}(\tau)}{\mathrm{VideoDurInMinutes}}$$

where $N_{\mathrm{MD}}(\tau)$ is the number of missed detections at the threshold τ , $N_{\mathrm{FA}}(\tau)$ is the number of false alarms, and *VideoDurInMinutes* is number of minutes of video. $N_{\mathrm{TrueInstance}}$ is the number of reference instances annotated in the sequence. Lastly, the Detection Error Tradeoff (DET) curve [Martin and Przybocki, 1997] is used to visualize system performance.

In this paper, we evaluate performance on the operating points; P_{miss} at $R_{\mathrm{FA}} = 0.15$ and P_{miss} at $R_{\mathrm{FA}} = 1$.

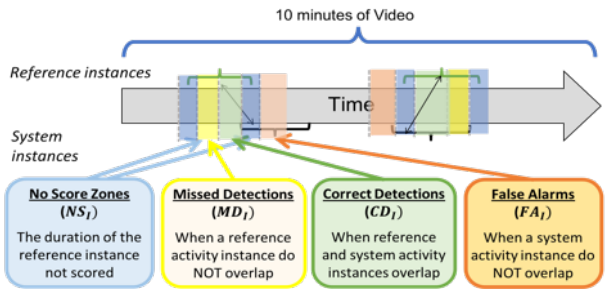


Figure 21: Confusion matrix computation of instance-pairs for temporal localization

The secondary metric for the AD task evaluates how precisely the system temporally localizes activity instances. In this measure, the confusion matrix is first calculated in the instance pair-level as illustrated in Figure 22. Due to annotation error or ambiguity of the start and end frames for the activity, we utilize the No-Score (NS_I) zone (blue): the duration of NS are not scored. To summarize system performance on temporal localization in activity instances, the Normalized Multiple Instance Detection Error (N_MIDE) is computed:

$$N_{\mathrm{MIDE}} = \frac{1}{N_{\mathrm{mapped}}} \sum_{I=1}^{N_{\mathrm{mapped}}} \left(C_{\mathrm{MD}} \times P_{MD} + C_{\mathrm{FA}} \times P_{FA} \right) \quad (1)$$

where

$$P_{MD} = \frac{MD_I}{MD_I + CD_I}$$
$$P_{FA} = \frac{FA_I}{\mathrm{Dur}_V - (MD_I + CD_I + NS_I)} \quad (2)$$

where C_{MD} and C_{FA} are the cost functions for the missed detections and false alarms respectively. N_{mapped} is the number of mapped instance pairs between reference and system output and Dur_V is the duration of the reference video V. For the ActEV18 evaluation, C_{MD} and C_{FA} are both equal to 1 and multiple N_{MIDE} values (since instance-pairs are changed at different decision thresholds) are calculated at different operating points; for instance, N_{MIDE} at $R_{\mathrm{FA}} = 0.15$ and N_{MIDE} at $R_{\mathrm{FA}} = 1$. The NS_I default value is zero.

Activity and Object Detection (AOD)

In this task, a system not only detects/localizes the target activity, but also detects the presence of target objects and spatially localizes the objects that are associated with a given activity. In addition to the activity information, the system must provide the coordinates of object bounding boxes and object presence confidence scores.

The primary metric is similar to AD, however, the instance alignment step uses an additional alignment term for object detection congruence to optimally map reference and system output instances—this is covered in further detail in the evaluation plan [Lee et al., 2018].

For the object detection (secondary) metric, we employed the Normalized Multiple Object Detection Error (N_MODE) described in [Kasturi et al., 2009] and [Bernardin and Stiefelhagen, 2008]. N_MODE evaluates the relative number of false alarms and missed detections for all objects per activity instance. Note that the metric is applied only to the frames where the system overlaps with the reference. The metric also uses the Hungarian algorithm to align objects between the reference and system output at the frame level. The confusion matrix for each frame t is calculated from the confidence scores of the objects' bounding boxes, referred to as the object presence confidence threshold τ. $\mathrm{CD}_t(\tau)$ is the count of reference and system output object bounding boxes that are correctly mapped for frame t at threshold τ. $\mathrm{MD}_t(\tau)$ is the count of reference bounding boxes not mapped to a system object bounding box at threshold τ. $\mathrm{FA}_t(\tau)$ is the count of system bounding boxes that are not aligned to reference bounding boxes. The equation for N_MODE follows:

$$N_{\mathrm{MODE}(\tau)} = \sum_{t=1}^{N_{\mathrm{frames}}} \frac{(C_{\mathrm{MD}} \times MD_t(\tau) + C_{\mathrm{FA}} \times FA_t(\tau))}{\sum_{t=1}^{N_{\mathrm{frames}}} N_R^t}$$

N_{frames} is the number of frames in the sequence for the reference instance and N_R^t is the number of reference objects in frame t. For each instance-pair, the minimum N_MODE value (minMODE) is calculated for object detection performance and P_{Miss} at R_{FA} points are reported for both activity-level and object-level detections. For the activity-level detection, we used the same operating points P_{miss} at $R_{\text{FA}} = 0.15$ and P_{miss} at $R_{\text{FA}} = 1$ while P_{miss} at $R_{\text{FA}} = 0.5$ was used for the object-level detection. We used 1- minMODE for the object detection congruence term to align the instances for the target activity detection. In this evaluation, the spatial object localization (that is, how precisely system can localize the objects) is not addressed.

Activity Object Detection/Tracking (AODT)

The goals of this task are to address whether the system correctly detects/localizes the target activity, correctly detects/localizes the required objects in that activity (object type and bounding box), and correctly tracks these objects over time.

Although the AODT task and performance measures are defined in the ActEV evaluation plan [Lee et al., 2018], the ActEV18 evaluations did not include this task.

6.2 Evaluation Framework

For ActEV challenges, there are the two evaluation types: 1) self-reported and 2)independent. For the self-reported evaluation, the participants run their software on their hardware and configurations and submit the system output with the defined format to the NIST scoring server. For the independent evaluation, the participants submit their runnable system, which is independently evaluated on the sequestered data using the evaluator's hardware. The following ActEV18 evaluation results are based on the *self-reported* evaluation only.

To examine the ability of systems in different aspects, the ActEV18 evaluations conducted a series of the three evaluations, 1) activity-level, 2) reference temporal segmentation (RefSeg), and 3) leaderboard evaluations. The activity-level evaluation measures accuracy and robustness of activity detection and temporal localization. For the RefSeg evaluation, systems are given the reference temporal segment information of the instances and then only activity type labeling is performed. The purpose of this evaluation is to examine the systems' ability to classify activity instances when the presence of an activity is known. The leaderboard evaluation provides overall performance after aggregating system performance across all target activities where developers can process the test collection multiple times and receive performance scores immediately. In the following section, we summarize the results for all three evaluations.

6.3 ActEV Dataset

Table 4: A list of 12 activities and their associated number of instances

Activity Type	Train	Validation
Closing	126	132
Closing_trunk	31	21
Entering	70	71
Exiting	72	65
Loading	38	37
Open_Trunk	35	22
Opening	125	127
Transport_HeavyCarry	45	31
Unloading	44	32
Vehicle_turning_left	152	133
Vehicle_turning_right	165	137
Vehicle_u_turn	13	8

For the ActEV18 activity-level evaluation, we used 12 activities from the VIRAT V1 dataset [Oh et al., 2011] that were annotated by Kitware, Inc. The detailed definition of each activity is described in the evaluation plan [Lee et al., 2018]. Table 4 lists the number of instances for each activity for the train and validation sets. Due to ongoing evaluations, the test sets are not included in the table. A total of 2.7 video hours were annotated for the test set across 12 activities. The numbers of instances are not balanced across activities, which may affect the system performance results.

For the RefSeg evaluation, we released the annotations of the reference temporal segments for half of the test set, randomly chosen, for the 12 activities shown in Table 4.

For the leaderboard evaluation, we added 7 more activities, listed in Table 5, on top of the 12 activities.

Table 5: A list of additional 7 activities and their associated number of instances (provided for Leaderboard only)

Activity Type	Train	Validation
Interacts	88	101
Pull	21	22
Riding	21	22
Talking	67	41
Activity_carrying	364	237
Specialized_talking_phone	16	17
Specialized_texting_phone	20	5

6.4 ActEV Results

In the ActEV18 activity-level evaluation, 14 teams (including baseline) from academia and industry participated. For the given 12 activities, Table 6 summarizes the performances of each team for both the AD and AOD tasks. The teams were limited to two submissions (primary and secondary). For AD, 20 systems from 13 teams (including the baseline algorithm) were submitted while 16 systems from 11 teams were submitted for AOD. For performance measures, the P_{miss} at $R_{\text{FA}} = 0.15$ and P_{miss} at $R_{\text{FA}} = 1$ were used for activity detection while N_{MIDE} at $R_{\text{FA}} = 0.15$ and N_{MIDE} at $R_{\text{FA}} = 1$ were reported for temporal localization. For simplicity, in Table 6, we listed the values of the metrics with the average values across all 12 activities for each system. The systems are ordered by the mean P_{miss} at $R_{\text{FA}} = 0.15$ (labeled 'PR.15')—a smaller value denotes a better performance. For the AOD task, there are two scoring protocols: AOD_AD and AOD_AOD. For the AOD_AD scoring protocol, the system is scored without additional terms of the object detection congruence, while for AOD_AOD the system is scored taking object detection into account. For AOD, we only list the P_{miss} at $R_{\text{FA}} = 0.15$ (labeled 'PR.15') and $\text{ObjectP}_{\text{miss}}$ at $R_{\text{FA}} = 0.5$ (labeled 'OPR.5').

Figure 22 shows the ranking of the 20 systems (ordered by P_{miss} at $R_{\text{FA}} = 0.15$) for the AD task. The x-axis lists the systems and the y-axis is P_{miss} at $R_{\text{FA}} = 0.15$. The rectangle dotted line (blue) represents the ranking of system performance on the metric P_{miss} at $R_{\text{FA}} = 0.15$ (activity occurrence detection) while the circle dotted line (circle) is the corresponding N_{MIDE} at $R_{\text{FA}} = 0.15$ values (temporal localization).

The general trend on system performance between activity detection and temporal localization is quite different. Note that the temporal localization performance is based on activity instances that were correctly detected—the activity instances detected may not be the same across systems. The results show that for activity detection, UMD achieved the lowest error P_{miss} at $R_{\text{FA}} = 0.15$ (PR.15: 61.8 %) followed by SeuGraph (PR.15: 62.4 %), while UCF achieved the lowest localization error (NR.15: 65.4 %).

Figure 23 illustrates the ranked list of the AOD systems. In addition to the activity detection and temporal localization, the graph includes system performance on object detection (marked in purple triangle). Again, the general trend between activity detection (rectangle) and object detection (triangle) has some differences. The results indicate that for activity detection, SeuGraph and UMD have the lowest error P_{miss} at $R_{\text{FA}} = 0.15$ (PR.15: 66.4 % and 68 % respectively) while the IBM_MIT _PURDUE team has the lowest object detection error (OPR.5:11%).

To examine the system's ability to classify activities when presence and temporal extent are known, the test set was divided into the two partitions, TestPart1 and TestPart2. The temporal localization reference data of each activity instance in the video was provided for the TestPart1 test set. To enable direct comparison, the systems submitted for the activity-level evaluation were scored on the TestPart1 test set only (termed EvalPart1 evaluation) and the systems that were submitted after releasing reference temporal segments were scored on the TestPart1 (termed RefSeg evaluation). Eleven systems were submitted for the RefSeg evaluation. The following results are computed on the teams who participated in both the EvalPart1 and RefSeg evaluations. Figure 24 compares the RefSeg (rectangle) and EvalPart1 (circle) results—ordered by the RefSeg results. With few exceptions, system performance with reference segment information is better than system performance without.

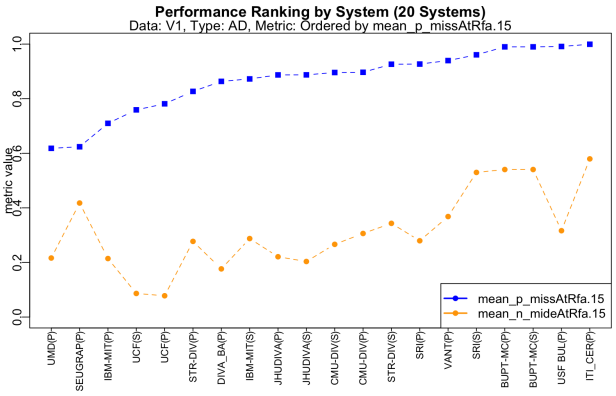


Figure 22: The ranked list of performance (AD)

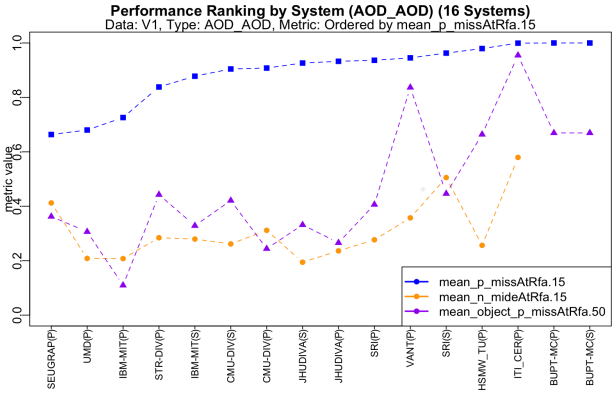


Figure 23: The ranked list of performance (AOD)

Table 6: ActEV18 activity-level evaluation results (P: Primary, S: Secondary)

System and Version		AD				AOD		
						AOD_AD	AOD_AOD	
		PR.15	NR.15	PR1	NR1	PR.15	PR.15	OPR.5
UMD	P	0.618	0.441	0.216	0.223	0.618	0.68	0.306
SeuGraph	P	0.624	0.621	0.418	0.416	0.624	0.664	0.362
IBM-MIT-Purdue	P	0.71	0.603	0.214	0.23	0.71	0.726	0.11
UCF	S	0.759	0.624	0.086	0.129	n/a	n/a	n/a
UCF	P	0.781	0.654	0.078	0.112	n/a	n/a	n/a
STR-DIVA Team	P	0.827	0.722	0.277	0.321	0.827	0.838	0.443
DIVA_Baseline	P	0.863	0.72	0.176	0.196	n/a	n/a	n/a
IBM-MIT-Purdue	S	0.872	0.704	0.288	0.282	0.872	0.878	0.329
JHUDIVATeam	P	0.887	0.829	0.221	0.219	0.887	0.933	0.266
JHUDIVATeam	S	0.887	0.813	0.203	0.24	0.887	0.926	0.332
CMU-DIVA	S	0.896	0.831	0.266	0.317	0.896	0.904	0.421
CMU-DIVA	P	0.897	0.766	0.306	0.349	0.897	0.908	0.244
STR-DIVA Team	S	0.926	0.905	0.343	0.355	n/a	n/a	n/a
SRI	P	0.927	0.856	0.279	0.282	0.927	0.936	0.406
VANT	P	0.94	0.918	0.368	0.385	0.94	0.945	0.837
SRI	S	0.961	0.885	0.53	0.49	0.961	0.963	0.446
BUPT-MCPRL	P	0.99	0.839	0.54	0.248	0.99	1	0.669
BUPT-MCPRL	S	0.99	0.839	0.54	0.248	0.99	1	0.669
USF Bulls	P	0.991	0.949	0.316	0.375	n/a	n/a	n/a
ITI_CERTH	P	0.999	0.998	0.579	0.667	0.999	0.999	0.955
HSMW_TUC	P	n/a	n/a	n/a	n/a	0.961	0.968	0.502

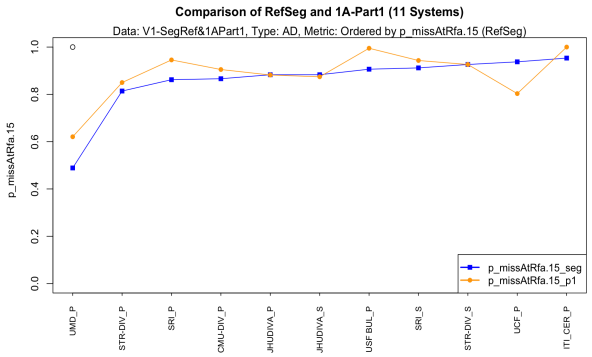


Figure 24: Performance comparison of RefSeg and EvalPart1 on the AD task

Table 7: Leaderboard results (as of 11/08/18) for AD and AOD ordered by the AD task

Teams	AD		AOD		
			AD	AOD	
	PR.15	NR.15	PR.15	PR.15	OPR.5
Team_Vision	0.709	0.252	0.709	0.752	0.175
UCF	0.733	0.179	0.774	0.934	0.753
BUPTMCPRL	0.749	0.215	0.751	0.786	0.324
INF	0.844	0.283	0.857	0.951	0.421
VANT	0.882	0.392	n/a	n/a	n/a
DIVABaseline	0.895	0.369	0.906	0.941	0.747
UTS-CETC	0.925	0.177	n/a	n/a	n/a
NIIHitachiUIT	0.925	0.177	0.931	0.941	0.728
USF Bulls	0.934	0.306	n/a	n/a	n/a

For leaderboard, we evaluated the 12 activities (used in the activity-level and RefSeg evaluations) plus 7 additional activities for a total of 19 activities. Each team was limited to uploading 50 submissions maximum. We picked each team's submission with the lowest detection error (based on P_{miss} at $R_{FA} = 0.15$) out of all of their submissions.

Table 7 summarizes the AD and AOD leaderboard results (as of 11/08/18) across all 19 activities; the metrics were first calculated on each activity and averaged across all activities. Out of 9 participants, the Team-Vision (IBM-MIT-Purdue) team has the lowest error μP_{miss} at $R_{FA} = 0.15$ for both AD and AOD.

6.5 ActEV Conclusion

This year fifteen teams participated in a series of ActEV18 evaluations, where the experiments were conducted using a structured evaluation framework.

We provided a ranked list of system performance for each task. For given target activities in test sets and the set of participants, our results showed that, for the AD task, the performance for P_{miss} at $R_{FA} = 0.15$ was 62 % for activity detection, and for temporal localization, N_{MIDE} at $R_{FA} = 0.15$ was 25 %. For the AOD task, the performance for activity detection for P_{miss} at $R_{FA} = 0.15$ is 68 % and for object detection ObjectP_{miss} at $R_{FA} = 0.5$ was 18 %.

We found that the activity detection and temporal localization performance trend differently—which implies that a better activity detection may not imply better activity localization. We observed that, with a few ex-

ceptions, activity detection performance with reference segment information is better than system performance without it.

The results of the ActEV18 evaluations will provide researchers an opportunity to obtain insight and direction for their system development and guide the next phase of ActEV evaluations to promote video analytics technology.

7 Social-media video storytelling linking

The new *social-media video storytelling linking* (LNK) task focuses on advancing the area of visual storytelling using collaborative videos, images and texts available in social media.

7.1 System task

The main objective is to illustrate a news story with social-media visual content. Starting from a news story topic and a stream of social-media video and images, the goal is to link a story-segment to image and video material, while preserving a good flow of the visual story.

A news story topic is an actual news narrative where the news segments correspond to particular sentences of the news that a journalist may wish to illustrate. In particular, a story segment can be defined as a sentence query with a strong visual component. For each story segment, systems should detect the *single video or image* that satisfies these two requirements:

- It best illustrates the news segment;
- It provides the optimal transition from the previous video/image illustration.

In this task, a visual storyline is composed of a set of images and/or videos organized in a sequence, to provide a cohesive narrative. This means that analyzing the relevance of the individual pieces of content is not enough when illustrating a storyline. Conversely, the way the pieces of content transition from one another, should also be taken into account, as shown in Figure 25. As such, assuring the quality and meaningfulness of these transitions is an important component of the editing process.

7.2 LNK Data

To enable social media visual storyline illustration, a data collection strategy was designed to create a suitable corpus, around major events, with considerable social-media activity. The number of retrieved documents was limited to those that were made available online during the span of the event. Events adequate for storytelling were selected, namely events with strong social-dynamics in terms of temporal variations with respect to their semantics (textual vocabulary and visual content). In other words, the unfolding of the event stories is encoded in each collection. Events that span over multiple days, such as music festivals, sports competitions, etc., are examples of good storylines. Taking the aforementioned aspects into account, the data for the following events was crawled (Table 9):

The Edinburgh Festival (EdFest) , an annual event consisting of a celebration of the performing arts, gathering dance, opera, music and theatre performers from all over the world. The event takes place in Edinburgh, Scotland and has a duration of 3 weeks in August.

Le Tour de France (TDF) , one of the most popular road cycling race competitions. The event usually takes place in France (16 d), Spain (1 d), Andorra (3 d) and Switzerland (3 d).

A *keyword-based* approach as adopted, consisting of querying the social media APIs with a set of event-related keyword terms. Thus, a curated list of keywords was manually selected for each event. Furthermore, hashtags in social media play the essential role of grouping similar content (e.g. content belonging to the same event) [Laniado and Mika, 2010]. Therefore, a set of relevant hashtags that group content belonging to the same topic was also manually defined. The data collected is detailed in Table 9.

Development data

The development data made use of content from the 2016 editions of the aforementioned events. Twenty stories were defined for each event, using simple baselines. These were then manually annotated with crowdsourcing. Three annotators were presented with each story title, and asked to rate each segment illustration as relevant or non-relevant, as well as rate the transitions between each of the segments. Finally, using the subjective assessment of the annotators, the score proposed in Section 7.4 was calculated for each story.

For each visual storyline, annotators were asked to rate the transitions between each sequential pair of images with a score of 0 (*"bad"*), 1 (*"acceptable"*) or 2 (*"good"*); they were also asked to rate the overall story quality on 1 to 5 scale.

Test data

The test data made use of content from the 2017 editions of the above events. For each event, 15 stories were defined. The topics and the ground truth are available for download.

7.3 Story topics

For the identification of event storylines, along with a focused crawling of social-media data about particular

Table 8: Development data covers the 2016 editions (relevance judgments available).

Event	Stories	Docs	Docs w/images	Docs w/videos	Crawling span	Crawling seeds	
EdFest2016	20	34,297	Twitter: 29,558	Twitter: 4,739	From: 2016-07-01	Terms	Edinburgh Festival, Edfest, Edinburgh Festival 2016, Edfest 2016
					Until: 2017-01-01	Hashtags	#edfest, #edfringe, #EdinburghFestival, #edinburghfest
TDF2016	20	75,385	Twitter: 67,032	Twitter: 8,353	From: 2016-06-01	Terms	le tour de france, le tour de france 2016, tour de france
					Until: 2017-01-01	Hashtags	#TDF2016, #TDF

Table 9: Test data covers the 2017 event editions (no relevance judgments available).

Event	Stories	Docs	Docs w/images	Docs w/videos	Crawling span	Crawling seeds	
EdFest2017	15	39,022	Twitter: 34,302	Twitter: 4,720	From: 2017-07-01	Terms	Edinburgh Festival, Edfest, Edinburgh Festival 2017, Edfest 2017
					Until: 2017-10-19	Hashtags	#edfest, #edfringe, #EdinburghFestival, #edinburghfest, #BBCedfest, #EdinburghFringe, #edinburghfringefestival
TDF2017	15	69,089	Twitter: 59,534	Twitter: 9,555	From: 2017-07-01	Terms	le tour de france, le tour de france 2017, tour de france
					Until: 2017-10-19	Hashtags	#TDF2017, #TDF, #TourdeFrance

events, a set of professional news[5] stories covering these same events was also collected. Two requirements were established regarding the identified storylines: general interestingness (i.e. news worthy and/or informative storylines), and availability of enough relevant supporting documents and media elements on the collected data.

7.4 Evaluation metric

Figure 25 illustrates the visual storyline quality assessment framework. In particular, storyline illustrations are assessed in terms of *relevance of illustrations* (blue links in Figure 25) and *coherence of transitions* (red links in Figure 25). Once a visual storyline is generated, annotators will judge the relevance of the illustration to the story segment as:

- s_i=0: the image/video is not relevant to the story segment;
- s_i=1: the image/video is relevant to the story segment;
- s_i=2: the image/video is highly relevant to the story segment.

Similarly with respect to the *coherence* of a visual storyline, each story transition is judged by annotators as the degree of affinity between pairs of story segment illustrations:

- t_i=0: there is no relation between the segment illustrations;
- t_i=1: there is a relation between the two segments;
- t_i=2: there is an appealing semantic and visual coherence between the two segment illustrations.

[5]We collected news from BBC, The Guardian and Reuters.

These two dimensions can be used to obtain an overall expression of the *"quality"* of a given illustration for a story of N segments. This is formalized by the expression:

$$Quality = \alpha \cdot s_1 + \frac{(1-\alpha)}{2(N-1)} \sum_{i=2}^{N} pairwiseQ(i) \quad (3)$$

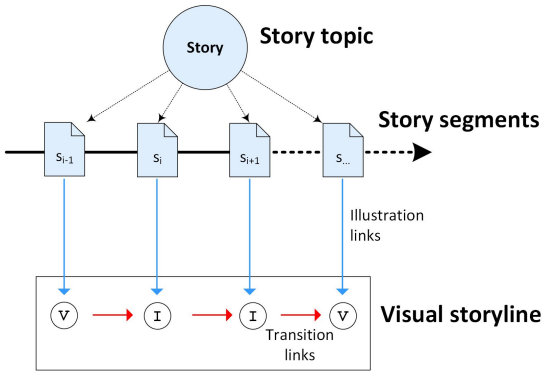
The function $pairwiseQ(i)$ defines quantitatively the perceived quality of two neighbouring segment illustrations based on their relevance and transition:

$$pairwiseQ(i) = \underbrace{\beta \cdot (s_i + s_{i-1})}_{\text{segments illustration}} \quad (4)$$

$$+ \underbrace{(1-\beta) \cdot (s_{i-1} \cdot s_i + t_{i-1})}_{\text{transition}} \quad (5)$$

where α weights the importance of the relevance of the first segment, and β weights the trade-off between *relevance of segment illustrations* and *coherence of transitions* towards the overall quality of the story.

Given the underlying subjectivity of the task, the values of α or β that optimally represent the human perception of visual stories are, in fact, average values. Nevertheless, we posit the following two reasonable criteria: (i) illustrating with non-relevant elements ($s_i = 0$) completely breaks the story perception and should be penalized. Thus, we consider values of $\beta > 0.5$; and (ii) the first image/video perceived is assumed to be more important, as it should grab the attention towards consuming the rest of the story. Thus, α is used to boost the overall storyline quality according to the relevance of first story segment s_1. It was empirically found that $\alpha = 0.1$ and $\beta = 0.6$ adequately represent human perception of visual stories editing.



Visual story editing assessment framework.

$$pairwiseQ(i) = \beta\,(s_{i-1} + s_i) + (1 - \beta)\,(s_{i-1}\,s_i + t_i)$$

segments illustration | transitions

Visual story quality assessment metric.

Figure 25: Methodology for evaluating visual storyline illustration.

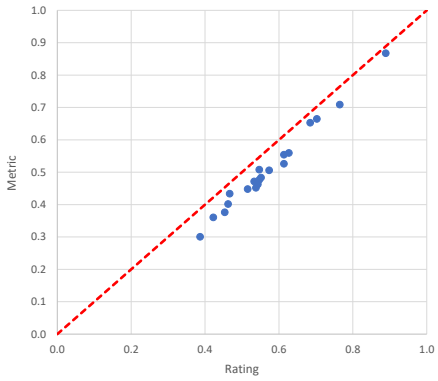


Figure 26: Results for the Edinburgh Festival 2017 and the Tour de France 2017 events.

7.5 Relevance judgments

Two participants submitted five runs each, resulting in 10 run submissions, which were used for ground truth creation and assessment using the metrics described above. The ground truth was generated by assessing the 150 stories that each participating team submitted. The stories were consumed using a dedicated prototype player that presented the segments text and illustration in a sequence.

All story segments were assessed by 3 annotators. Annotators rated the story illustration quality as a whole (in a scale from 1 to 5), the relevance of each segment (not relevant, relevant, very relevant) and the transition between segments (not relevant, relevant, very relevant). It should be noticed that some segments were illustrated by very long videos, some of more than 30 minutes in length. A relevance bias was identified towards longer stories.

7.6 Metric stability

We also examined the performance of the metric in terms of its stability. We computed the metric based on the relevance of segments and transitions between segments, and related it to the overall story rating assigned by annotators. Figure 26 compares the annotator rating to the quality metric. As can be seen, the relation is strong and relatively stable, which is a good indicator of the metric stability.

7.7 LNK Results

Results are illustrated in the Figure 27, Table 10 and Table 11. The run marked in red is the only manual run.

It is interesting to note that in EdFest, the manual run was outperformed by ADAPT's run *ed17_crun0*. This may be due to the fact that ADAPT's system trained concept detectors for the test queries with data collected from Google. It is also worth noticing that, in the EdFest event, all ADAPT's runs performed better than the NOVASearch runs. In the TDF stories, the manual run performed better than the methods proposed by both participants. When comparing the results across both events, it can be seen that all methods perform consistently. Within each participant's runs, the relation between methods is the same across both events.

7.8 LNK Conclusions and Observations

The new format of the TRECVID 2018 linking task aimed at creating visual summaries of live events using social media content. Two teams participated in the task achieving very good results. One of the events (Tour de France) is clearly easier that the other (The Edinburgh Festival), which allowed the participants to improve their own methods.

In terms of evaluation, the methodology was sound and the metric results were stable, and strongly correlated with the user perception of the visual stories generated by the participants. There were some idiosyncrasies in the participants submissions (e.g., the duration of the submitted video segments), which are being investigated to improve the next year task. Entertainment related events and emergency related events are planned to be used in the 2019 task.

8 Video to Text Description

Automatic annotation of videos using natural language text descriptions has been a long-standing goal of com-

Table 10: Results for the Edinburgh Festival 2017 event data.

RUN	Quality	Team
ed17_crun0	0.665	ADAPT
ed17_crun1	0.472	ADAPT
ed17_crun2	0.508	ADAPT
ed17_crun3	0.434	ADAPT
ed17_crun4	0.448	ADAPT
ns_manual	0.653	NOVASearch
ns_sequential_without_relevance	0.376	NOVASearch
ns_sequential_with_relevance	0.360	NOVASearch
ns_fully_connected_without_relevance	0.402	NOVASearch
ns_fully_connected_with_relevance	0.301	NOVASearch

Table 11: Results for the Tour de France 2017 event data.

RUN	Quality	Team
ed17_crun0	0.709	ADAPT
ed17_crun1	0.526	ADAPT
ed17_crun2	0.560	ADAPT
ed17_crun3	0.452	ADAPT
ed17_crun4	0.477	ADAPT
ns_manual	0.868	NOVASearch
ns_sequential_with_relevance	0.463	NOVASearch
ns_sequential_without_relevance	0.484	NOVASearch
ns_fully_connected_with_relevance	0.506	NOVASearch
ns_fully_connected_without_relevance	0.554	NOVASearch

puter 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. Since then, there have been substantial improvements in the dataset and evaluation.

8.1 VTT 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 1 903 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 exactly 5 assessors. This is in contrast to the previous year's task where the number of annotations ranged between 2 and 5. 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)

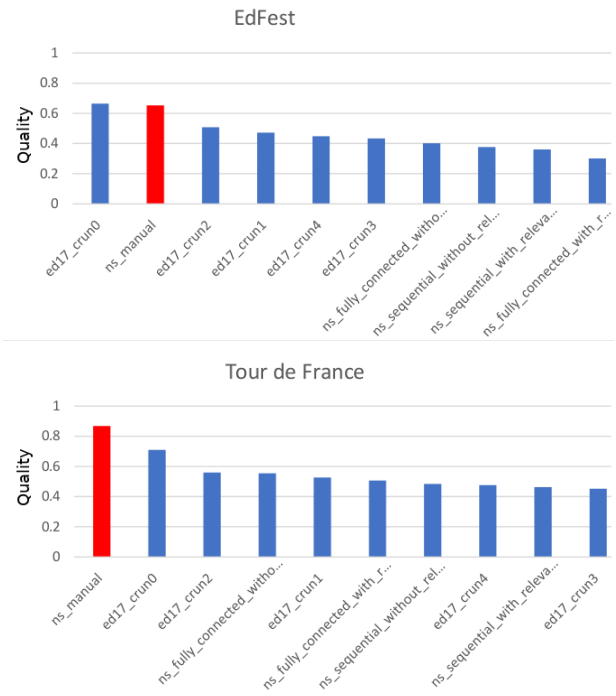


Figure 27: Results for the Edinburgh Festival 2017 and the Tour de France 2017 events.

Furthermore, the assessors were also asked the following questions:

- Please rate how difficult it was to describe the video.
  - Very Easy
  - Easy
  - Medium
  - Hard
  - Very Hard
- How likely is it that other assessors will write similar descriptions for the video?
  - Not Likely
  - Somewhat Likely
  - Very Likely

We carried out data preprocessing to ensure a usable dataset. Firstly, we clustered videos based on visual similarity. We used a tool called SOTU [Ngo, 2012], which uses visual bag of words, to cluster videos with 60 % similarity for at least 3 frames. This allowed us to remove any duplicate videos, as well as videos which were very similar visually (e.g., soccer games). However, we learned from last year's task that this automated procedure is not sufficient to create a clean and diverse dataset. For this reason, we manually went through a large set of videos, and removed the following types of videos:

- Videos with multiple, unrelated segments that are hard to describe, even for humans.
- Any animated videos.
- Other videos which may be considered inappropriate or offensive.

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 description that corresponds (was annotated) to the video from each of the 5 sets. Here the number of sets is equal to the number of groundtruth descriptions for videos.
- Description Generation: Automatically generate for each video URL a text description (1 sentence) independently and without taking into consideration the existence of any annotations.

Up to 4 runs were allowed per team for each of the subtasks.

This year, systems were also required to choose between two run types based on the type of training data they used:

- Run type 'V' : Training using Vine videos (can be TRECVID provided or non-TRECVID Vine data).
- Run type 'N' : Training using only non Vine videos.

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. We also used a human evaluation metric on selected runs to compare with the automatic metrics.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Banerjee and Lavie, 2005] and BLEU (BiLingual Evaluation Understudy) [Papineni et al., 2002] are standard metrics in machine translation (MT). BLEU 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 and there is no corpus to work from. Thus, our expectations are lowered when it comes to evaluation by BLEU. METEOR is based on the harmonic mean of unigram or n-gram precision and recall in terms of overlap between two input sentences. It

Table 12: List of teams participating in each of the VTT subtasks.

	Matching & Ranking (26 Runs)	Description Generation (24 Runs)
INF	X	X
KSLAB	X	X
KU_ISPL	X	X
MMSys_CCMIP	X	X
NTU_ROSE	X	X
PicSOM		X
UPCer		X
UTS_CETC_D2DCRC_CAI	X	X
EURECOM	X	
ORAND	X	
RUCMM	X	
UCR_VCG	X	

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.

The CIDEr (Consensus-based Image Description Evaluation) metric [Vedantam et al., 2015] is borrowed from image captioning. 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. We also report scores using CIDEr-D, which is a modification of CIDEr to prevent "gaming the system".

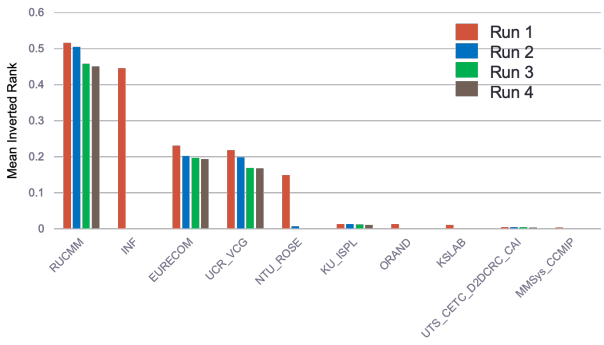


Figure 28: VTT: Matching and Ranking results across all runs for Set A

The STS (Semantic Similarity) metric [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.

In addition to automatic metrics, the description generation task includes human evaluation of the quality of automatically generated captions. Recent developments

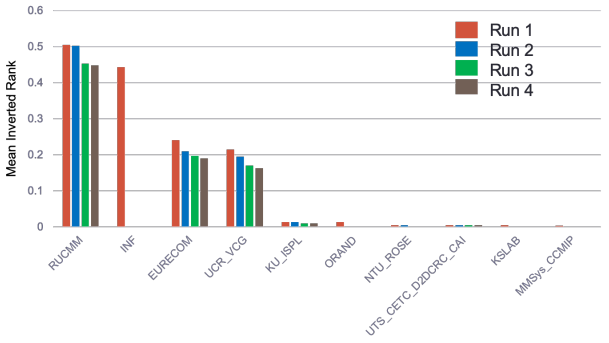


Figure 29: VTT: Matching and Ranking results across all runs for Set B

in Machine Translation evaluation have seen the emergence of DA (Direct Assessment), 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., 2018]. 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) [6], 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

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

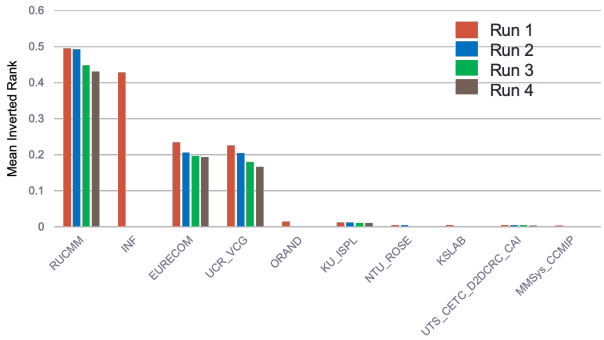


Figure 30: VTT: Matching and Ranking results across all runs for Set C

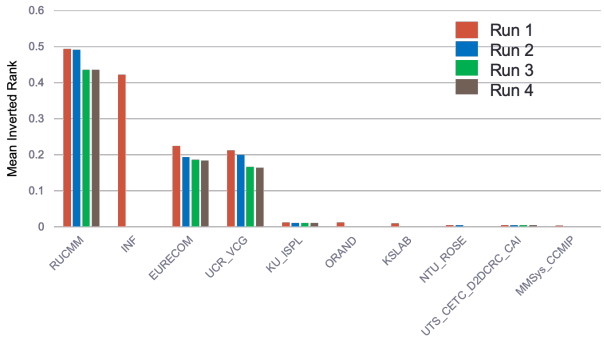


Figure 31: VTT: Matching and Ranking results across all runs for Set D

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.

DA was first used as an evaluation metric in TRECVID 2017. We have used this metric again this year to rate each team's primary run, as well as 4 human systems.

In total, 12 teams participated in the VTT task this year. There were a total of 26 runs submitted by 10 teams for the matching and ranking subtask, and 24 runs submitted by 8 teams for the description generation subtask. A summary of participating teams is shown in Table 12.

8.4 VTT Results

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

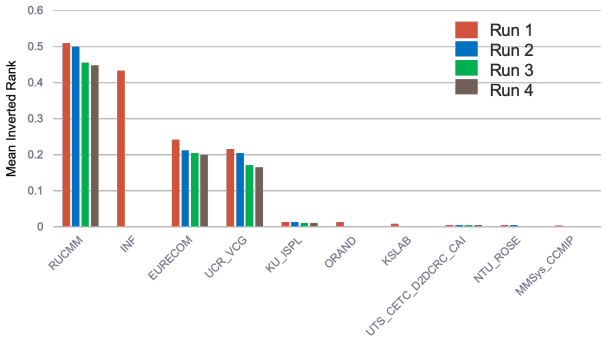


Figure 32: VTT: Matching and Ranking results across all runs for Set E

Matching and Ranking Sub-task

The results for the matching and ranking sub-task are shown for each of the 5 sets (A-E) in Figures 28 - 32. The graphs show the mean inverted rank scores for all runs submitted by the teams for each of the description sets.

Figure 33 shows the ranking of the various teams with respect to the different sets. For each team, the scores for the best runs are used. The figure allows us to compare the teams across all sets. It is worth noting that the top 4 teams are consistent across all the sets. For the remaining teams, there is not much difference between the scores of runs, and so even though we see fluctuation between teams across sets, there is not much to differentiate between their scores.

Videos consisting of continuous scenes with little camera movement were more likely to be matched in a consistent manner among runs. For example, one of the top videos matched showed a single shot of a woman playing a guitar and singing while sitting. Systems had more difficulty matching videos that consisted of complex actions or scene cuts. In some cases, the annotations contained an interpretation which was hard for systems to describe, such as "crying in pain".

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 continued the use of Direct Assessment, which was introduced in TRECVID 2017, to compare the submitted runs.

Figure 34 shows the comparison of all teams using the CIDEr metric. All runs submitted by each team are shown in the graph. Figure 35 shows the scores for the CIDEr-D metric, which is a modification of CIDEr. Figures 36, 37, 38 show the scores for METEOR, BLEU, and STS metrics respectively. Each team identified one run as their 'primary' run. Interestingly, the primary run was

A	B	C	D	E
RUCMM	RUCMM	RUCMM	RUCMM	RUCMM
INF	INF	INF	INF	INF
EURECOM	EURECOM	EURECOM	EURECOM	EURECOM
UCR_VCG	UCR_VCG	UCR_VCG	UCR_VCG	UCR_VCG
NTU_ROSE	KU_ISPL	ORAND	KU_ISPL	KU_ISPL
KU_ISPL	ORAND	KU_ISPL	ORAND	ORAND
ORAND	NTU_ROSE	NTU_ROSE	KSLAB	KSLAB
KSLAB	UTS_CETC_D2DCRC_CAI	KSLAB	NTU_ROSE	UTS_CETC_D2DCRC_CAI
UTS_CETC_D2DCRC_CAI	KSLAB	UTS_CETC_D2DCRC_CAI	UTS_CETC_D2DCRC_CAI	NTU_ROSE
MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP

Figure 33: VTT: Ranking of teams with respect to the different sets

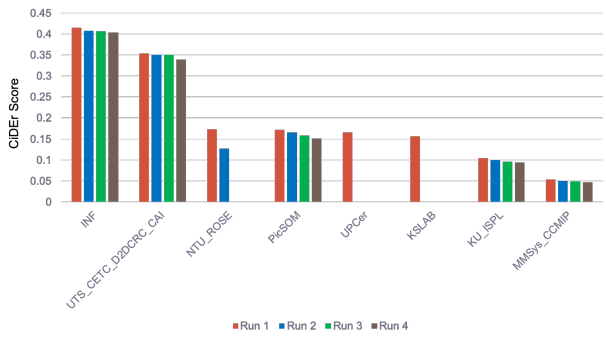


Figure 34: VTT: Comparison of all runs using the CIDEr metric

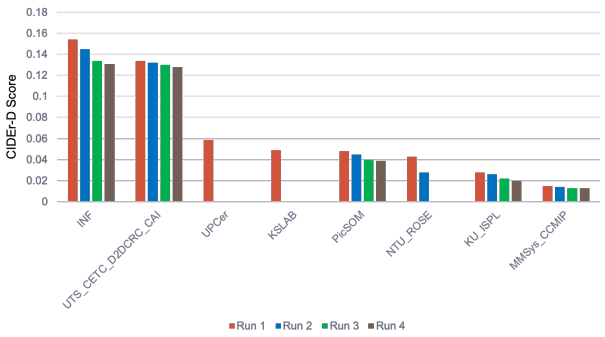


Figure 35: VTT: Comparison of all runs using the CIDEr-D metric

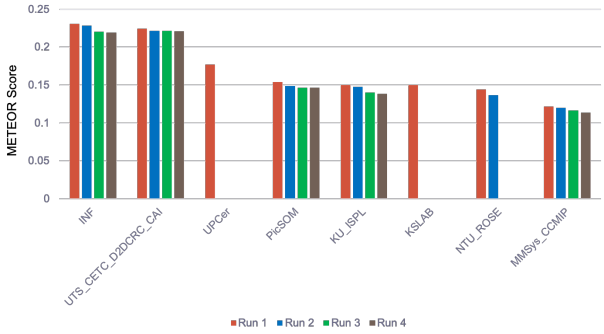


Figure 36: VTT: Comparison of all runs using the METEOR metric

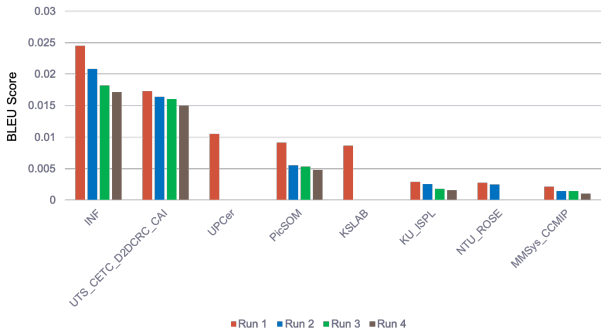


Figure 37: VTT: Comparison of all runs using the BLEU metric

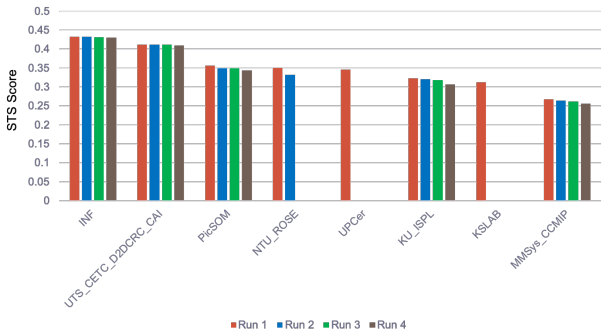


Figure 38: VTT: Comparison of all runs using the STS metric

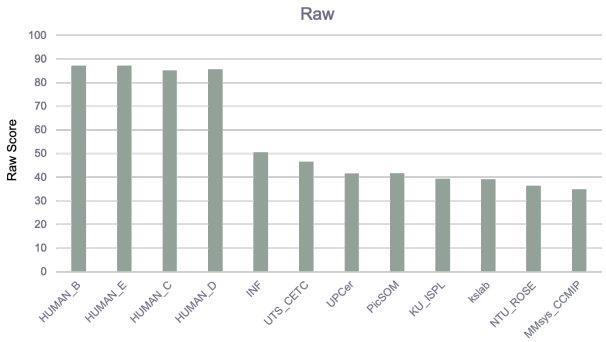


Figure 39: VTT: Average DA score for each system. The systems compared are the primary runs submitted, along with 4 manually generated system labeled as HUMAN_B to HUMAN_E

Figure 40: VTT: Average DA score per system after standardization per individual worker's mean and standard deviation score

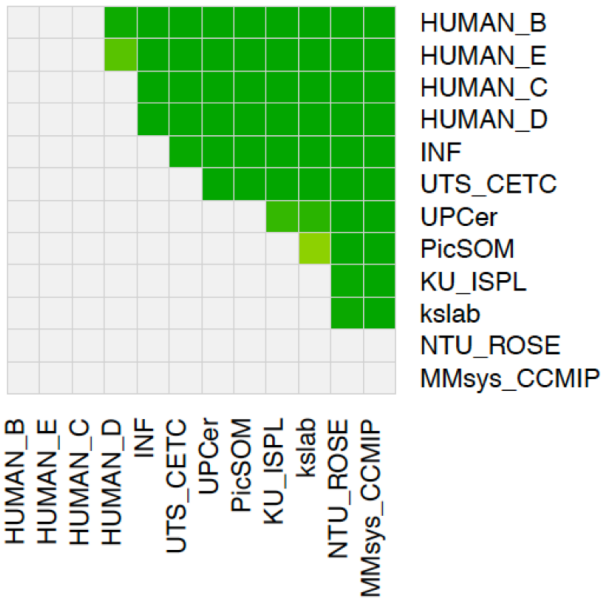


Figure 41: VTT: Comparison of systems with respect to DA. Green squares indicate a significantly better result for the row over the column

not necessarily the best run for each team. It was also observed that there was no significant advantages of using either of the run types (N or V).

Figure 39 shows the average DA score $[0 - 100]$ for each system. The score is micro-averaged per caption, and then averaged over all videos. Figure 40 shows the average DA score per system after it is standardized per individual AMT worker's mean and standard deviation score. The HUMAN systems represent manual captions provided by assessors. As expected, captions written by assessors outperform the automatic systems. Figure 41 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 systems, INF clearly outperforms all the other systems. An interesting observation is that HUMAN_B and HUMAN_E statistically perform better than HUMAN_D. This could be due to the difference in average sentence lengths in the different sets of annotations. However, a more detailed analysis is required to determine the cause of the significant difference.

Figure 42 shows the comparison of the various teams with respect to the different metrics used in the description generation subtask.

8.5 VTT Conclusions and Observations

The VTT task continues to have healthy participation. Given the challenging nature of the task, and the increasing interest in video captioning in the computer vision community, we hope to see improvements in performance. The task continues to evolve as the number of annotations per video was standardized to 5. This seems to be a reasonable number of annotations as it sufficiently captures the variation in the ways humans describe short videos. Efforts were made to create a cleaner dataset than previous years. To this end, a pre-processing step was done where videos were clustered based on similarity, and then a diverse set collected for annotation. Furthermore, an additional manual pass was made to remove any unwanted videos remaining after the previous step.

For the description generation subtask, we used multiple automatic evaluation metrics: CIDEr, CIDEr-D, METEOR, BLEU, 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.

9 Summing up and moving on

This overview to TRECVID 2018 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 [TV18Pubs, 2018].

10 Authors' note

TRECVID would not have happened in 2018 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

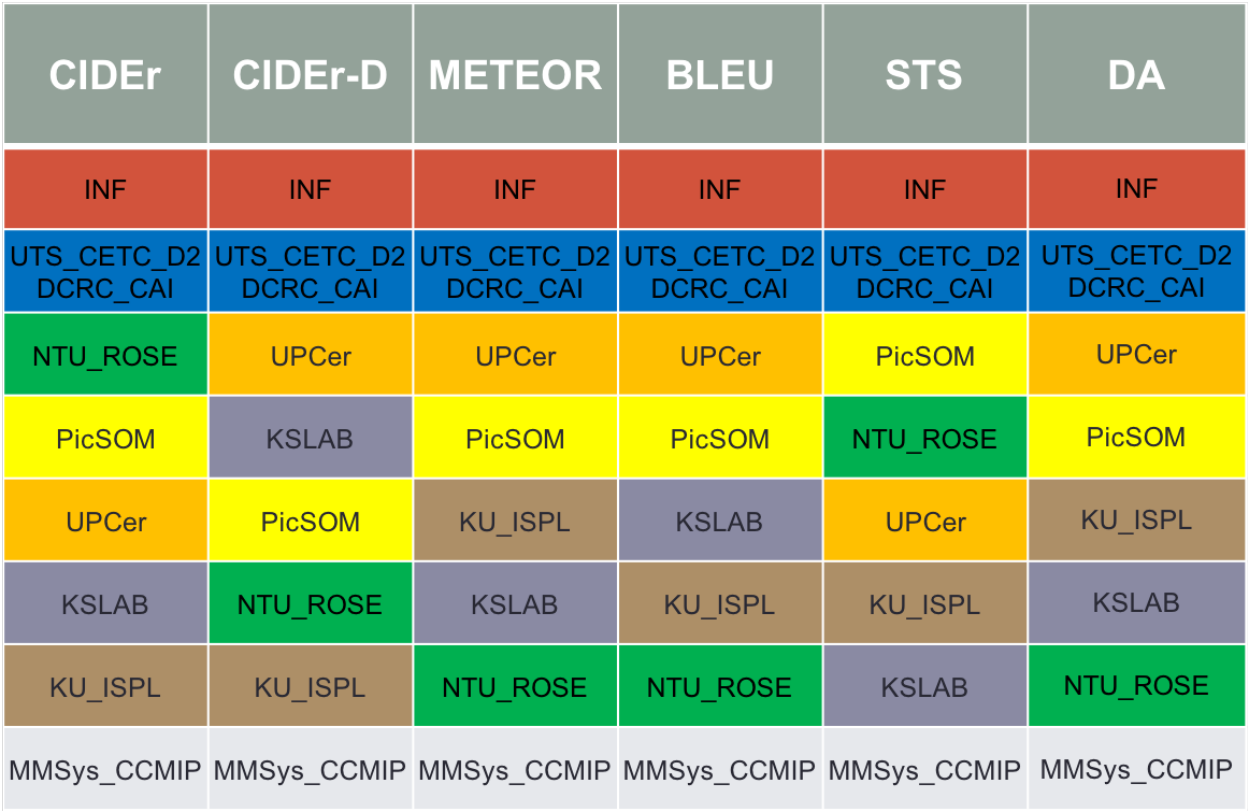


Figure 42: VTT: Ranking of teams with respect to the different metrics for the description generation task

EastEnders video available for use in TRECVID. Finally, Rob Cooper at BBC facilitated the copyright licence agreement for the Eastenders data.

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

11 Acknowledgments

The ActEV NIST work was supported by the Intelligence Advanced Research Projects Activity (IARPA), agreement IARPA-16002, order R18-774-0017. The authors would like to thank Kitware, Inc. for annotating the dataset. 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., 2016] Awad, G., Snoek, C. G., Smeaton, A. F., and Quénot, G. (2016). Trecvid Semantic Indexing of Video: A 6-year retrospective. *ITE Transactions on Media Technology and Applications*, 4(3):187–208.

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

[Bernardin and Stiefelhagen, 2008] Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008:1.

[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., 2018] Graham, Y., Awad, G., and Smeaton, A. (2018). Evaluation of automatic video captioning using direct assessment. *PloS one*, 13(9):e0202789.

[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.

[Kasturi et al., 2009] Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., and Zhang, J. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336.

[Laniado and Mika, 2010] Laniado, D. and Mika, P. (2010). Making sense of twitter. In *Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I*, ISWC'10, pages 470–485, Berlin, Heidelberg. Springer-Verlag.

[Lee et al., 2018] Lee, Y., Godil, A., Joy, D., and Fiscus, J. (2018). Actev 2018 evaluation plan. `https://actev.nist.gov/pub/Draft_ActEV_2018_EvaluationPlan.pdf`.

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

[Martin and Przybocki, 1997] Martin, A., D. G. K.-T. O. M. and Przybocki, M. (1997). The det curve in assessment of detection task performance. In *Proceedings*, pages 1895–1898.

[Michel et al., 2017] Michel, M., Fiscus, J., and Joy, D. (2017). Trecvid 2017 surveillance event detection evaluation. `https://www.nist.gov/itl/iad/mig/trecvid-surveillance-event-detection-evaluation-track`.

[Munkres, 1957] Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38.

[Ngo, 2012] Ngo, W.-L. Z. C.-W. (2012). Sotu in action.

[Oh et al., 2011] Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 3153–3160. IEEE.

[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.

[TV18Pubs, 2018] TV18Pubs (2018). `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.18.org.html`.

[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

561 Find shots of exactly two men at a conference or meeting table talking in a room
562 Find shots of a person playing keyboard and singing indoors
563 Find shots of one or more people on a moving boat in the water
564 Find shots of a person in front of a blackboard talking or writing in a classroom
565 Find shots of people waving flags outdoors
566 Find shots of a dog playing outdoors
567 Find shots of people performing or dancing outdoors at nighttime
568 Find shots of one or more people hiking
569 Find shots of people standing in line outdoors
570 Find shots of a projection screen
571 Find shots of any type of Christmas decorations
572 Find shots of two or more cats both visible simultaneously
573 Find shots of medical personnel performing medical tasks
574 Find shots of two people fighting
575 Find shots of a person pouring liquid from one container to another
576 Find shots of a person holding his hand to his face
577 Find shots of two or more people wearing coats
578 Find shots of a person in front of or inside a garage
579 Find shots of one or more people in a balcony
580 Find shots of an elevator from the outside or inside view
581 Find shots of a person sitting on a wheelchair
582 Find shots of a person climbing an object (such as tree, stairs, barrier)
583 Find shots of a person holding, talking or blowing into a horn
584 Find shots of a person lying on a bed
585 Find shots of a person with a cigarette
586 Find shots of a truck standing still while a person is walking beside or in front of it
587 Find shots of a person looking out or through a window
588 Find shots of a person holding or attached to a rope
589 Find shots of car driving scenes in a rainy day
590 Find shots of a person where a gate is visible in the background

B Instance search topics

9219 Find Jane in this Cafe 2
9220 Find Jane in this Pub
9221 Find Jane in this Mini-Market
9222 Find Chelsea in this Cafe 2
9223 Find Chelsea in this Pub
9224 Find Chelsea in this Mini-Market
9225 Find Minty in this Cafe 2
9226 Find Minty at this Pub
9227 Find Minty in this Mini-Market
9228 Find Garry in this Cafe 2
9229 Find Garry in this Pub
9230 Find Garry in this Laundrette
9231 Find Mo in this Cafe 2
9232 Find Mo in this Pub
9233 Find Mo in this Laundrette
9234 Find Darren in this Cafe 2
9235 Find Darren in this Pub
9236 Find Darren in this Laundrette

9237 Find Zainab in this Cafe 2
9238 Find Zainab in this Laundrette
9239 Find Zainab in this Mini-Market
9240 Find Heather in this Cafe 2
9241 Find Heather in this Laundrette
9242 Find Heather in this Mini-Market
9243 Find Jack in this Pub
9244 Find Jack in this Laundrette
9245 Find Jack in this Mini-Market
9246 Find Max in this Cafe 2
9247 Find Max at this Laundrette
9248 Find Max in this Mini-Market