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

George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzad Godil, David Joy, Andrew Delgado, Alan F. 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)

November 12, 2018

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 content-based exploitation of 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 addition, many organizations and individuals worldwide contribute sig-

nificant time and effort.

TRECVID 2018 represented a continuation of three tasks from 2017, the addition of a new pilot task Social Media Video Storytelling Linking, jointly participating in a new pilot task "Streaming Multimedia Knowledge-base Population" with the Text Analysis Conference (TAC) project, and introducing the new Activities in Extended Video task as a continuation to the previous surveillance event detection task that ended in 2017. 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 represent organizations that registered but did not submit any runs.

This year TRECVID used again the same 600 hours of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC.3) that were used for ad-hoc Video Search in 2016 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 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 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.

2 Video Data

2.1 BBC EastEnders video

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

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

2.2 Internet Archive Creative Commons (IACC.3) video

The IACC.3 dataset consists of 4 593 Internet Archive videos (144 GB, 600 h) with Creative Commons licenses in MPEG-4/H.264 format with duration ranging from 6.5 to 9.5 min and a mean duration of ≈7.8 min. Most videos will have some metadata provided by the donor available e.g. title, keywords, and description. Approximately 1 200 h of IACC.1 and IACC.2 videos used between 2010 to 2015 were available for system development. As in the past, the Computer Science Laboratory for Mechanics and Engineering Sciences (LIMSI) and Vocapia Research provided automatic speech recognition for the English speech in the IACC.3 videos.

2.3 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 the 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 7):

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 coverts 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 were 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 started 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 possibility of containing the target query. The presence of each query was assumed to be binary, i.e., it was either present or absent in the given standard video shot.

Judges at NIST followed several rules in evaluating system output. If the query was true for some frame (sequence) within the shot, then it was true for the shot. This is a simplification adopted for the benefits it afforded in pooling of results and approximating the basis for calculating recall. In query definitions, "contains x" or words to that effect are short for "contains x to a degree sufficient for x to be recognizable as x by a human". This means among other things that unless explicitly stated, partial visibility or audibility may suffice. The fact that a segment contains video of a physical object representing the query target, such as photos, paintings, models, or toy versions of the target (e.g picture of Barack Obama vs Barack Obama himself), was NOT grounds for judging the query to be true for the segment. Containing video of the target within video may be grounds for doing so.

Like it's predecessor, in 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.

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 Data

The IACC.3 video collection of about 600 h was used for testing. It contained 335 944 video clips in mp4 format and xml meta-data files. Throughout this report we does not differentiate between a clip and a shot and thus they may be used 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

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

% 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 Results

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.

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

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

4.1 Data

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

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

4.2 System task

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

Each query consisted of a set of

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

Information about the use of the examples was reported by participants with each submission. The

possible categories for use of examples were as follows:

A one or more provided images - no video used
E video examples (+ 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)
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) and in fact 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.

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

4.5 Measures

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

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 (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 multiple document elements in multiple modalities; therefore, 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 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)

A set of 3 evaluation topics and about 10,000 documents were released as testing data. The 3 testing topics were:

- Suspicious Deaths and Murders in Ukraine (January-April 2015)
- Odessa Tragedy (May 2, 2014)
- 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 received as well 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 (National Institute of Standards and Technology) supported by IARPA (Intelligence Advanced Research Projects Activity) launched ActEV (Activities in Extended Video) evaluations to promote a video analytics technology. The purpose of the evaluations is to develop a system that can automatically detect a target activity and identify and track objects associated with the activity. With both retrospective analysis and real-time analysis applications in mind, the challenges include activity detection in a multi-camera streaming environment and its temporal (and spatio-temporal) localization of the activity for reasoning.

To understand current state-of-the-art algorithms, we initiated the ActEV18 1.A evaluation with the following reference temporal segmentation and leaderboard evaluations. ActEV18 is an extension of TRECVID Surveillance Event Detection (SED) [Michel et al., 2017] and the evaluations are conducted under TRECVID [Awad et al., 2016a]. In this paper, we present a brief overview of the challenge tasks and performance measures for the ActEV18 evaluations. Further, we discuss the target application for the evaluations along with the evaluation type and conditions.

6.1 Data

For the ActEV18 1.A evaluation, we used a subset of the 12 activities from the VIRAT V1 dataset [Oh et al., 2011] that were annotated by Kitware, Inc in 2017. The dataset is a large-scale surveillance video dataset designed to assess the performance of activity detection algorithms in realistic scenes. The VIRAT dataset was collected to facilitate both detection of activities and to localize the corresponding temporal segment of the activity and spatio-temporal location of objects associated with activities from a large continuous video. The VIRAT dataset are closely aligned with real-world video analytics.

Table 4 lists a number of instances for each activity for the train and validation—due to ongoing evaluations, the test sets are not included in the table. A total of 2.7 video hours were annotated for the 1.A eval test set for given 12 activities. Note that the numbers of instances are not balanced across activities, which may affect the system performance results.

For the reference temporal segmentation (RefSeg) evaluation, we released the annotations of the reference temporal segments for a randomly chosen half of the test set 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. Therefore, a total of 19 activities selected from the VIRAT V1 data set were used.

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

Table 4: A list of 12 activities and its number of instances for the ActEV18 1.A evaluation

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

Table 5: A list of additional 7 activities and its number of instances for the ActEV18 Leaderboard

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

6.2 Tasks and Measures

The general purpose of the ActEV18 evaluation is to promote the development of a system that automatically 1) identifies a target activity along with the time span of the activity (activity detection), 2) detects objects associated with the target activity (activity and object detection), and 3) tracks multiple objects across multiple cameras (activity and object detection/tracking). In the following subsections, we define each task and describe its performance measure.

Activity Detection (AD)

In this task, given a target activity, a system automatically detects its presence and then temporally localizes all instances of the activity in video sequences. The system should provide the start and end frames indicating the temporal segment of the target activity and a presence confidence score that indicates how likely the activity occurred. To evaluate system performance, we modified the metrics from TRECVID: Surveillance Event Detection [Michel et al., 2017] and CLEAR [Bernardin and Stiefelhagen, 2008]. The primary metric addresses how correctly the system detected the occurrence of the activity. The scoring procedure between reference annotation and system output can be divided into four distinctive steps: 1) instance alignment, 2) confusion matrix calculation, 3) summary performance metrics, and 4) result visualization.

The goal of the alignment step is to find a one-to-one correspondence of the instances between the reference and the system output. 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. Miss Detection (MD) indicates that an instance in the reference has no correspondence to the system output instance while False Alarm (FA) indicates that an instance in the system output has no correspondence to the reference. The following step is to 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 (P_{Miss}) and a rate of false alarms (R_{FA})to be computed at a given threshold:

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

$$\text{Rate}_{\text{FA}}(\tau) = \frac{N_{\text{FA}}(\tau)}{\text{VideoDurInMinutes}}$$

where $N_{\text{MD}}(\tau)$ is the number of missed detections at the threshold τ while $N_{\text{FA}}(\tau)$ is the number of false alarms. $N_{\text{TrueInstance}}$ is the number of reference instances annotated in the sequence. Lastly, the Detection Error Trade-off (DET) curve is used to visualize system performance.

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

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 1.

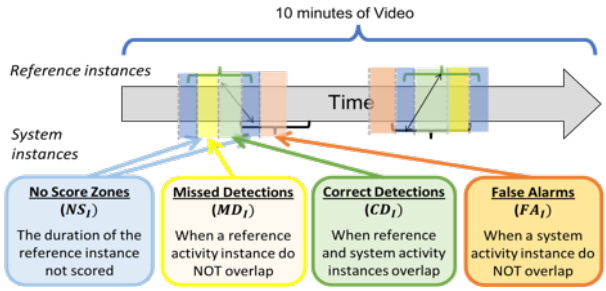


Figure 1: Confusion matrix computation of instance-pairs for temporal localization

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:

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

where

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

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 $R_{\text{FA}} = 0.15$ and N_{MIDE} at $R_{\text{FA}} = 1$.

Activity and Object Detection (AOD)

In this task, a system not only detects 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 should 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 term for the object detection congruence function to yield potential corresponding instances between the reference and system output—see further detail in the evaluation plan [Lee et al., 2018].

For the object detection (secondary metric), we employed the N_MODE (Normalized Multiple Object Detection Error) metric described in [Kasturi et al., 2009][Bernardin and Stiefelhagen, 2008]. N_MODE evaluates the relative number of false positives and missed detections for all objects per activity instance. Note that the metrics 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. $\text{CD}_t(\tau)$ thus counts the reference and system output object bounding boxes that are correctly mapped for frame t at threshold (τ). $\text{MD}_t(\tau)$ counts the reference bounding boxes not mapped to a system object bounding box at threshold τ. $\text{FA}_t(\tau)$ counts the system bounding boxes that are not aligned to reference bounding boxes.

The equation for N_MODE follows:

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

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 goal of this task is to address whether the system 1) correctly detects/localizes the target activity, 2) correctly detects/localizes the required objects in that activity, and 3) correctly identifies/tracks these objects over time.

Although the AODT task and performance measures are defined in the evaluation plan [Lee et al., 2018], this paper primarily focuses on the AD and AOD tasks with the single camera view and at the activity observation level.

6.3 Evaluation Framework

For any benchmark or evaluation, it is essential to ask what target applications the system will be utilized for [Seltzer et al., 1999]. The collection of data sets and the development of performance metrics should reflect the expected system behavior of a particular use case across a range of hardware or software platforms. In this section, we describe the target applications of the evaluations and a brief evaluation framework for the ActEV18 challenges.

Target Application

The technology developed for the evaluations is expected to be applied to both forensic analytics and real-time alerting applications: the forensic analytics tool is a system that processes vast collections for repeated investigation while the real-time alerting tool is a system that processes many video streams with detection occurring within defined latency during collection. For this evaluation, we define detection latency to be the time from the onset of the activity to the last frame processed by the system before being able to declare the activity is occurring.

With forensic and alerting applications in mind, the evaluation differentiates two types of systems: forensic systems and low latency systems. In this paper, the ActEV18 evaluations target the forensic applications only.

Evaluation Type

For ActEV18, there are the two evaluation types: 1) self-reported and 2) independent. For the self-reported evaluation, the performers 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 performers 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.

Evaluation Conditions

To examine the ability of systems in different aspects, the ActEV18 evaluations conducted a series of the three evaluations; 1) 1.A activity-level, 2) reference temporal segmentation (RefSeg), and 3) leaderboard evaluations

The 1.A 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 for each activity. The purpose of this evaluation is to examine the systems' ability to classify activity instances without being hampered by the system's instance localization results. The leaderboard evaluation provides overall performance after aggregating system performance across all target activities. We summarize the results and analyses for all three evaluations.

6.4 Results

A total of 15 teams from the academic and industrial sectors participated in the ActEV18 evaluations. In the ActEV18 1.A phase evaluation, a total of 20 systems from 13 teams (including the baseline algorithm) were submitted for AD, while a total of 16 systems from 11 teams were submitted for AOD.

In the RefSeg evaluation, a total of 11 systems were submitted and some teams who participated in the 1.A evaluation did not submit their systems to the RefSeg evaluation. The results are computed on the teams who participated in both the 1.A and RefSeg evaluations.

In the leaderboard evaluation, we evaluated 7 more activities in addition to the 12 listed in the 1.A and RefSeg evaluations. Each team was limited to uploading 50 submissions maximum. We picked the best-performed result (based on P_{miss} at $R_{\text{FA}} = 0.15$) out of all submissions.

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

7 Social-media video storytelling linking

The new *social-media video storytelling linking* task seeks to advance the area of visual story-telling with collaborative videos, images and texts available in social-media.

7.1 System task

The goal 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 also preserving a good flow of the whole visual story.

A news story topic is an actual news narrative and the news segments correspond to particular sentences of the news, that a journalist may wish to illustrate. For each story segment (a sentence query with a strong visual

component), systems should propose the *single video or image* that satisfy the two requirements:

- Best illustrates the news segment;
- Makes the best transition from the previous video/image illustration.

In this task, a visual storyline is composed of a set of images/videos organised in a sequence to provide a cohesive narrative. Tackling the task of illustrating a storyline means taking into account not only the relevance of the individual pieces of content, but also the way they transition from one to the other, Figure 2. As such, assuring the quality and meaningfulness of these transitions is an important component of the editing process.

7.2 Data

To enable social media visual storyline illustration, a data collection strategy was designed to create a suitable corpora, limiting the number of retrieved documents to those posted during the span of the event. Events adequate for storytelling were selected, namely those 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 like music festivals, sports competitions, etc., are examples of good candidates of storylines. Taking the aforementioned aspects into account, the data for the following events was crawled (Table 7):

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 *keyword-based* approach, consists of querying the social media APIs with a set of 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 grouping content of the same topic was also manually defined. The data collected is detailed in Table 7.

Development data

The development data coverts the 2016 editions of the above events and for each event there's 20 stories. Several stories were generated with simple baselines and evaluated with crowd-sourcing. 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 story quality on 1 to 5 scale.

Test data

The test data covers the 2017 editions of the above events and for each event there's 15 stories. The topics are available for download, but the ground truth will only be available after submissions.

7.3 Story topics

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

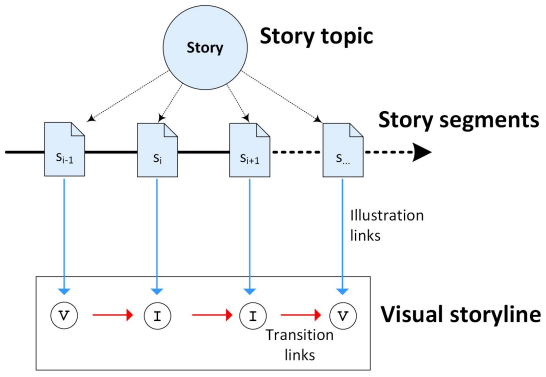
Figure 2 illustrates the visual storyline quality assessment framework. In particular, storyline illustrations are assessed in terms of *relevance of illustrations* (blue links in Figure 2) and *coherence of transitions* (red links in Figure 2). Once a visual storyline is generated, annotators will judge 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.

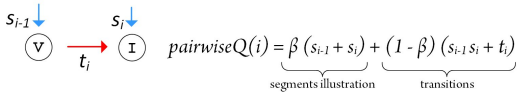
Similarly with respect to the *coherence* of a visual storyline, each story transition is judged by annotators as the degree of affinity between pairs of story segment illustrations:

- t_i=0: there is no relation between the segment illustrations;
- t_i=1: there is a relation between the two segments;

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



(a) Visual story editing assessment framework.



(b) Visual story quality assessment metric.

Figure 2: Methodology for evaluating visual storyline illustration.

- t_i=2: there is an appealing semantic and visual coherence between the two segment illustrations.

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 formalised 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 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 represents 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 penalised. 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 a boost to the 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.

7.5 Relevance judgments

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

7.6 Results

Two groups submitted five runs each, resulting in 10 run submissions, which were used for ground truth creation and assessment using the metrics described above.

8 Video to Text Description

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

There are many use case application scenarios which can greatly benefit from technology such as video summarization in the form of natural language, facilitating the search and browsing of video archives using such descriptions, describing videos as an assistive technology, etc. In addition, learning video interpretation and temporal relations among events in a video will likely contribute to other computer vision tasks, such as prediction of future events from the video.

The "Video to Text Description" (VTT) task was introduced in TRECVid 2016 as a pilot. Since then, there have been substantial improvements in the dataset and evaluation.

8.1 Data

Over 50k Twitter Vine videos have been collected automatically, and each video has a total duration of about 6 seconds. In the task this year, a dataset of 1 903 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

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

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

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)

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 thus there is no corpus to work from, so 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 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".

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 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) [8], with strict quality control measures applied to filter out or downgrade the weightings from workers unable to demonstrate the ability to rate good captions higher than lower quality captions. This is achieved by deliberately "polluting" some of the manual (and correct) captions with linguistic substitutions to generate captions whose semantics are questionable. Thus we might substitute a noun for another noun and turn the manual caption "A man and a woman are dancing on a table" into "A *horse* and a woman are dancing on a table", where "horse" has been substituted for "man". We expect such automatically-polluted captions to be rated poorly and when an AMT worker correctly does this, the ratings for that worker are improved.

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 8.

8.4 Results

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.

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

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 EastEnders video available for use in TRECVID. Finally, Rob Cooper at BBC facilitated the copyright licences issues.

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

11 Acknowledgments

The Video-to-Text work has been partially supported by Science Foundation Ireland (SFI) as a part of the Insight Centre at DCU (12/RC/2289). We would like to thank Tim Finin and Lushan Han of University of Maryland, Baltimore County for providing access to the semantic similarity metric.

References

[Awad et al., 2016a] Awad, G., Fiscus, J., Joy, D., Michel, M., Kraaij, W., Smeaton, A. F., Quénot, G., Eskevich, M., Aly, R., Ordelman, R., Ritter, M., Jones, G. J., , Huet, B., and Larson, M. (2016a). TRECVID

2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA.

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

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

[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`.

[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.

[Seltzer et al., 1999] Seltzer, M., Krinsky, D., Smith, K., and Zhang, X. (1999). The case for application-specific benchmarking. In *Hot Topics in Operating Systems, 1999. Proceedings of the Seventh Workshop on*, pages 102–107. IEEE.

[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.

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

Table 6: 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 7: 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

Table 8: 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	

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