TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains

George Awad {gawad@nist.gov}
Georgetown University;
Information Access Division, National Institute of Standards and Technology, USA

Asad A. Butt {asad.butt@nist.gov}
Johns Hopkins University;
Information Access Division, National Institute of Standards and Technology, USA

Keith Curtis {keith.curtis@nist.gov}
Information Access Division, National Institute of Standards and Technology, USA

Jonathan Fiscus {jfiscus@nist.gov} Afzal Godil {godil@nist.gov}
Yooyoung Lee {yooyoung@nist.gov} Andrew Delgado {andrew.delgado@nist.gov}
Jesse Zhang {jesse.zhang@nist.gov} Eliot Godard {eliot.godard@nist.gov}
Baptiste Chocot {baptiste.chocot@nist.gov}
Information Access Division, National Institute of Standards and Technology, USA

Lukas Diduch {lukas.diduch@nist.gov}
Dakota-consulting, USA

Jeffrey Liu {jeffrey.liu@ll.mit.edu}
MIT Lincoln Laboratory, USA

Alan F. Smeaton {alan.smeaton@dcu.ie}
Insight Centre, School of Computing, Dublin City University, Ireland

Yvette Graham {graham.yvette@gmail.com}
ADAPT Centre, School of Computing, Dublin City University, Ireland

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

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

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

November 23, 2020

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) is a TREC-style video analysis and retrieval evaluation with the goal of promoting progress in research and development of content-based exploitation and retrieval of information from digital video via open, metrics-based evaluation.

Over the last twenty 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 has been funded by NIST (National Institute of Standards and Technology) and other US government agencies. In addition, many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2020 represented a continuation of four tasks and the addition of two new tasks. In total, 52 teams from various research organizations worldwide signed up to join the evaluation campaign this year where 29 teams (Table 1) completed one or more of the following six tasks and 23 teams registered but did not submit any runs (Table 2):

1. Ad-hoc Video Search (AVS)
2. Instance Search (INS)
3. Disaster Scene Description and Indexing (DSDI)
4. Video to Text Description (VTT)
5. Activities in Extended Video (ActEV)
6. Video Summarization (VSUM)

This year TRECVID continued the usage of the Vimeo Creative Commons collection dataset (V3C1) [Rossetto et al., 2019] of about 1000 hours in total and segmented into 1 million short video shots to support the Ad-hoc video search task. The dataset is drawn from the Vimeo video sharing website under the Creative Common licenses and reflects a wide variety of content, style, and source device determined only by the self-selected donors.

The Instance Search task continued working with the 464 hours of the BBC (British Broadcasting Corporation) EastEnders video as used before since 2013, while the Video to Text description task started using a subset of 1700 short videos from the Vimeo V3C2 dataset.

For the Activities in Extended Video task, about 10 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.

The new Video Summarization task also made use of the BBC Eastenders dataset, while the DSDI task worked on public natural disaster 5 h videos collected from a Nepal earthquake event in 2015.

The Ad-hoc search, Instance Search, and Video Summarization results were judged by NIST human assessors, while the Video to Text 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. The Disaster Scene Description and Indexing task was also annotated by human assessors and scored automatically using Mean Average Precision (MAP).

The systems 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 2020 evaluation campaign. 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 [TV20Pubs, 2020]. Finally we would like to acknowledge that all work presented here has been cleared by HSPO (Human Subject Protection Office) under HSPO number: #ITL-17-0025

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 (Intelligence Advanced Research Projects Activity), NIST, or the U.S. Government.

2 Datasets

Many datasets have been adopted and used across the years since TRECVID started in 2001 and all available resources and datasets from previous years can be accessed from our website[1]. In the following sections we will give an overview of the main datasets

[1] https://trecvid.nist.gov/past.data.table.html

Table 1: Participants and tasks

Task						Location	TeamID	Participants
IN	*VT*	*AV*	*AH*	*DS*	*VS*			
−−	−−	−−	−−	*DS*	−−	*Eur*	*VCL*	Information Technologies Institute (ITI) Centre of Research and Technology Hellas (CERTH)
−−	*VT*	−−	−−	−−	−−	*Eur*	*PicSOM*	Aalto University
IN	−−	*AV*	−−	∗∗	−−	*Asia*	*BUPT_MCPRL*	Beijing University of Posts and Telecommunications
−−	−−	∗∗	*AH*	−−	−−	*Asia*	*VIdeoREtrievalGrOup*	City University of Hong Kong
−−	*VT*	−−	∗∗	−−	−−	*SAm*	*IMFD_IMPRESEE*	University of Chile; Millennium Institute of Data Foundation (IMFD), Chile; Impresee Inc, Chile
−−	−−	−−	−−	−−	*VS*	*Eur*	*MeMAD*	Eurecom and Aalto for MeMAD
−−	−−	−−	*AH*	*DS*	−−	*NAm*	*FIU_UM*	Florida International University; University of Miami
−−	−−	*AV*	*AH*	−−	−−	*Asia*	$kindai_ogu$	Kindai University; Osaka Gakuin University
−−	−−	∗∗	−−	*DS*	−−	*Asia*	*VAS*	Hitachi, Ltd. R&D
−−	−−	−−	*AH*	−−	−−	*Asia*	*DVA_Researchers*	Indian Institute of Space Science & Technology (IIST), Thiruvananthapuram Development and Educational Communication Unit (DECU), Indian Space Research Organisation (ISRO)
−−	−−	*AV*	*AH*	*DS*	−−	*Eur*	*ITI_CERTH*	Information Technologies Institute , Centre for Research and Technology Hellas
−−	*VT*	−−	−−	−−	−−	*Asia*	*KU_ISPL*	korea university
−−	−−	−−	−−	*DS*	−−	*Eur*	*SHIELD*	LINKS Foundation
−−	*VT*	−−	−−	−−	−−	*Asia*	*KsLab_NUT*	Nagaoka University of Technology
−−	−−	−−	−−	*DS*	−−	*Asia*	*NIICT*	National Institute of Information and Communications Technology (Japan), and National Institute of Informatics (Japan)
IN	∗∗	∗∗	∗∗	*DS*	*VS*	*Asia*	*NII_UIT*	National Institute of Informatics , Japan; University of Information Technology, VNU-HCMC, Vietnam
IN	∗∗	−−	−−	∗∗	−−	*Asia*	*PKU_WICT*	Peking University
−−	*VT*	−−	*AH*	−−	−−	*Asia*	*RUC_AIM3*	Renmin University of China
−−	−−	−−	*AH*	−−	−−	*Asia*	*RUCMM*	Renmin University of China
IN	−−	*AV*	∗∗	−−	−−	*Asia*	*UEC*	The University of Electro-Communications, Tokyo
−−	−−	*AV*	−−	−−	−−	*Asia*	*TokyoTech_AIST*	Tokyo Institute of Technology , National Institute of Advanced Industrial Science and Technology (AIST)
−−	*VT*	−−	−−	−−	−−	*Eur*	*MMCUniAugsburg*	University of Augsburg
−−	∗∗	−−	−−	*DS*	∗∗	*Aus*	*UTSVideo*	University of Technology Sydney
∗∗	−−	−−	−−	*DS*	−−	*NAm*	*COVIS*	UNT College of Engineering; UNT Dept. of Computer Science and Engineering; UNT Dept. of Electrical Engineering
−−	∗∗	∗∗	*AH*	−−	∗∗	*Asia*	*WasedaMeiseiSoftbank*	Waseda University; Meisei University; SoftBank Corporation
IN	−−	−−	−−	−−	−−	*Asia*	*WHU_NERCMS*	Wuhan University
−−	∗∗	−−	*AH*	−−	−−	*Asia*	*ZY_BJLAB*	XinHuaZhiYun Technology
−−	−−	*AV*	−−	−−	−−	*NAm*	*INF*	Carnegie Mellon University
−−	−−	*AV*	−−	−−	−−	*NAm*	*CRCV_UCF*	University of Central Florida

Task legend. IN:Instance Search; VT:Video to Text; AV:Activities in Extended videos; AH:Ad-hoc search; DS: Disaster Scene Description and Indexing; VS: Video Summarization; −−:no run planned; ∗ ∗ ∗:planned but not submitted

Table 2: Participants who did not submit any runs

Task						Location	TeamID	Participants
IN	*VT*	*AV*	*AH*	*DS*	*VS*			
−−	∗∗	−−	∗∗	−−	∗∗	*Asia*	*ATL*	Alibaba group, ZheJiang University
−−	∗∗	−−	−−	−−	−−	*NAm*	*Arete*	Arete_Associates
−−	−−	∗∗	−−	−−	−−	*Eur + Asia*	*SYMBEN*	Athlone Institute of Technology, Ireland Aligarh Muslim University, India Lahore College for Women Univesity, Lahore, Pakistan Islamia University Bahawalpur, Pakistan
−−	∗∗	−−	−−	−−	−−	*Asia*	*BDVIDEO*	BAIDU
−−	∗∗	−−	−−	−−	∗∗	*Asia*	*NDKS*	Charotar University Of Science & Technology
−−	−−	∗∗	−−	−−	−−	*Asia*	*Byte_Karma*	CHARUSAT
−−	−−	−−	∗∗	−−	−−	*Asia*	*UPC_VIT*2020	China university of petroleum (east China)
∗∗	−−	−−	∗∗	∗∗	−−	*NAm*	*VCUB*	CSE Dept UB
−−	−−	−−	∗∗	−−	∗∗	*NAm*	*drylwlsn_visual*	drylwlsn_visual
−−	−−	∗∗	−−	∗∗	−−	*Eur*	*IOSBVID_TV*20	Fraunhofer IOSB Research Institute Karlsruhe Institute of Technology
−−	−−	−−	∗∗	−−	−−	*NAm*	*ark*_20	Huawei Noah's Ark lab
∗∗	∗∗	∗∗	∗∗	∗∗	∗∗	*Asia*	*aalekhn*	Independent Researcher
−−	−−	∗∗	−−	−−	−−	*NAm*	*usf_bulls*	Institute for Artificial Intelligence (AI+X), University of South Florida
−−	∗∗	−−	∗∗	−−	∗∗	*Asia*	*KNU.visual_lab*	Kangwon national university
−−	∗∗	−−	∗∗	−−	−−	*Eur*	*LIG*	Multimedia Information Modeling and Retrieval group of LIG Explainable and Responsible Artificial Intelligence Chair of the MIAI Institute.
∗∗	∗∗	∗∗	∗∗	∗∗	∗∗	*Asia*	*DAMILAB*	NIT Warangal
−−	−−	−−	−−	∗∗	∗∗	*Asia*	*PKUMI*	Peking University
−−	∗∗	−−	−−	−−	∗∗	*Afr*	*REGIM_Lab_VSUM*	Research Groups in Intelligent Machines
∗∗	−−	−−	∗∗	−−	−−	*Eur*	*AIT_SRI*_2020	Software Research Institute Athlone IT
∗∗	∗∗	−−	−−	−−	−−	*Eur + Asia*	*Sheffield_UETLahore*	University of Sheffield Department of Computer Science University of Engineering and Technology, Lahore Department of Computer Science
−−	−−	−−	∗∗	−−	−−	*Asia*	*ustcmcc*	University of Science and Technology of China officially Huawei Technologies Co. Ltd.
∗∗	−−	∗∗	∗∗	∗∗	−−	*Eur*	*Aptitude*	Universite de Mons
∗∗	−−	−−	∗∗	−−	−−	*NAm*	*VSR*	Visionary Systems and Research (VSR)

Task legend. IN:Instance Search; VT:Video to Text; AV:Activities in extended videos; AH:Ad-hoc search; DS: Disaster Scene Description and Indexing; VS: Video Summarization; −−:no run planned; ∗∗:planned but not submitted used this year across the different tasks.

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 since 2013. The data comprise 244 weekly "omnibus" broadcast files (divided into 471 527 shots), transcripts, and a small amount of additional metadata. This dataset was adopted to test systems on retrieving target persons (characters) doing specific everyday actions in the Instance Seaerch task and also adopted for the Video Summarization task to summarize the major events in 3 characters during a time period of about 6 to 8 weeks of episodes.

2.2 Vimeo Creative Commons Collection (V3C) Dataset

The V3C1 dataset (drawn from a larger V3C video dataset [Rossetto et al., 2019]) is composed of 7475 Vimeo videos (1.3 TB, 1000 h) with Creative Commons licenses and mean duration of 8 min. All videos have some metadata available such as title, keywords, and description in json files. The dataset has been segmented into 1 082 657 short video segments according to the provided master shot boundary files. In addition, keyframes and thumbnails per video segment have been extracted and made available. While the V3C1 dataset was adopted for testing the Ad-hoc video search systems, the previous Internet Archive datasets (IACC.1-3) of about 1800 h were available for development and training. In addition to the above, a small subset of 1700 short videos from V3C2 dataset (also drawn from the V3C video dataset) were used to test the Video to Text systems.

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 outdoor 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 data was collected at different buildings and parking lots at multiple sites distributed throughout America. A variety of camera viewpoints and resolutions were included, with different level of cluttered backgrounds, and activity are performed by many ordinary people. The spatial resolution of the cameras is 1920x1080 or 1920x1072. The VIRAT dataset are closely aligned with real-world video surveillance analytics. The 35 activities used for this evaluation could be broadly categorized as: person/multi-person activity, person object interaction, vehicle activity, and person vehicle/facility interaction. Figure 1 shows the different VIRAT image montage of randomly selected videos. In addition, we have build a larger Multiview Extended Video with Activities (MEVA) dataset [Kitware, 2020] which is used for different ActEV Sequestered Data Leaderboard (SDL) competitions [NIST, 2020]. The main purpose of the VIRAT 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.

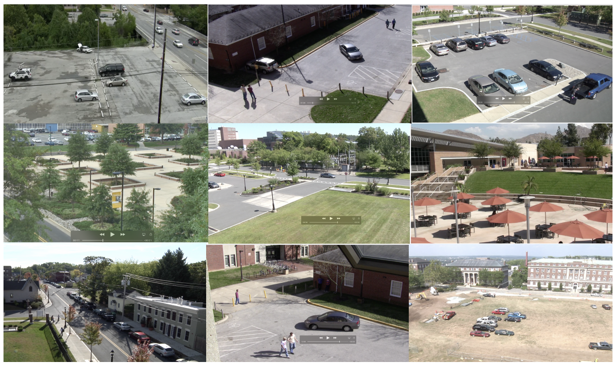


Figure 1: Shows the different VIRAT videos montage of few selected video clips.

2.4 TRECVID-VTT

This dataset contains short videos (ranging from 3 seconds to 10 seconds) previously used for the TRECVID VTT task since 2016. In total, there are 9185 videos with captions. Each video has between 2 and 5 captions, which have been written by dedicated annotators. The collection includes 6475 URLs from Twitter Vine and 2710 video files in webm format and have the Creative Commons License. Those 2710 videos belong to Flickr and the V3C2 dataset

(1700 V3C2 videos were used as a testing set this year).

2.5 Low Altitude Disaster Imagery (LADI)

The LADI dataset consists of over 20 000 annotated images, each at least 4 MB in size and was available as development dataset for the DSDI systems. The images are collected by the Civil Air Patrol from various natural disaster events. The raw images were previously released into the public domain. Two key distinctions are the low altitude (less than 1000 ft), oblique perspective of the imagery and disaster-related features, which are rarely featured in computer vision benchmarks and datasets. The dataset currently employs a hierarchical labeling scheme of a five coarse categories and then more specific annotations for each category. The initial dataset focuses on the Atlantic Hurricane and spring flooding seasons since 2015.

3 Evaluated Tasks

3.1 Ad-hoc Video Search

The Ad-hoc Video Search (AVS) task was resumed at TRECVID again in 2016 utilizing the Internet Archive Creative Commons (IACC.3) dataset and in 2019 a new Vimeo dataset (V3C1) was adopted instead. The task is trying to model 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 the Laboratoire d'Informatique de Grenoble[2].

The task for participants was defined as the following: given a standard set of master shot boundaries (about 1 Million shots) from the V3C1 test collection and a list of 30 ad-hoc textual queries (see Appendix A and B), participants were asked to return for each query, at most the top 1 000 video clips from the master shot boundary reference 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. For example, 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 addition, query definitions such as "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. Lastly, 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 (such as a television showing the target query) may be grounds for doing so. Three main submission types were accepted:

- Fully automatic runs (no human input in the loop): System takes a query as input and produces results 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 results 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-30 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 up to 3 iterations.

In general, runs submitted were allowed to choose any of the below four training types:

- 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

The training categories "E" and "F" are motivated by the idea of promoting the development of methods that permit the indexing of concepts in video

[2] Thanks to Georges Quénot

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.

A new progress subtask was introduced in 2019 with the objective of measuring system progress on a set of 20 fixed topics (Appendix B). As a result, 2019 systems were allowed to submit results for 20 common topics (not evaluated in 2019) that will be fixed for three years (2019-2021). This year NIST evaluated progress runs submitted in 2019 and 2020 so that teams can measure their progress against two years (2019-2020) while in 2021 they can measure their progress against three years. In general, the 20 fixed progress topics are divided equally into two sets of 10 topics to be evaluated in 2020 and 2021.

A "Novelty" run type was also allowed to be submitted within the main task. The goal of this run is to encourage systems to submit novel and unique relevant shots not easily discovered by other runs.

Dataset

The V3C1 dataset (drawn from a larger V3C video dataset [Rossetto et al., 2019]) was adopted as a testing dataset. It is composed of 7 475 Vimeo videos (1.3 TB, 1000 h) with Creative Commons licenses and mean duration of 8 min. All videos have some metadata available e.g., title, keywords, and description in json files. The dataset has been segmented into 1 082 657 short video segments according to the provided master shot boundary files. In addition, keyframes and thumbnails per video segment have been extracted and made available. For training and development, all previous Internet Archive datasets (IACC.1-3) with about 1 800 h were made available with their ground truth 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.

Evaluation

Each group was allowed to submit up to 4 prioritized runs per submission type, and per task type (main or progress) and two additional if they were of training type "E" or "F" runs. In addition, one novelty run type was allowed to be submitted within the main task.

In fact, 9 groups submitted a total of 75 runs with 39 main runs and 36 progress runs. Two groups submitted a novelty runs. The 39 main runs consisted of 26 fully automatic, and 13 manually-assisted runs. While the progress runs consisted of 24 fully automatic and 12 manually-assisted runs.

To prepare the results from teams for human judgments, a workflow was adopted to pool results from runs submitted. For each query topic, a top pool was created using 100 % of clips at ranks 1 to 250 across all submissions after removing duplicates. A second pool was created using a sampling rate at 11.1 % of clips at ranks 251 to 1000, not already in the top pool, across all submissions and after removing duplicates. Using these two master pools, we divided the clips in them into small pool files with about 1000 clips in each file. 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 then voting if the clip contained the query topic or no. 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 and was voted as false positive by the assessor. This final step was done as a secondary check on the assessors judging work and to give them an opportunity to fix any judgment mistakes. In all, 147 950 clips were judged while 226 097 clips fell into the unjudged part of the overall samples. Total hits across the 30 topics reached 22 859 with 12 210 hits at submission ranks from 1 to 100, 7969 hits at submission ranks 101 to 250 and 2725 hits at submission ranks between 251 to 1000. Table 3 presents information about the pooling and judging per topic.

Measures

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 measure inferred average precision (infAP) 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 (≈250) 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. The *sample_eval* software [3], 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.

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV20Pubs, 2020] in the online workshop notebook proceedings.

3.2 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 tested systems on retrieving specific instances of individual objects, persons and locations. A more challenging task and important goal in some applications is to combine two or more entities. Therefore, starting in 2016 a new query type, to retrieve specific persons in specific locations had been introduced. The task spanned 3 years till 2018 and starting in 2019 a similar query type has been adopted to retrieve instances of named persons doing named actions.

Dataset

Finding realistic test data, which contains sufficient recurrences of various specific objects/persons/locations under varying conditions has been difficult. Initially, 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), British Broadcasting Corporation (BBC) rushes (2011), and Flickr (2012).

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). One dedicated video (Id 0) was provided for development where participants can use it in any way they wish, while the rest of the dataset episodes were used for testing. The usage of the BBC Eastenders proved to be very useful and adequate for the task and TRECVID has been using this same dataset since 2013.

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 predefined actions with example videos, and a collection of topics (queries) that delimit a specific person in some example images and videos, locate for each topic up to the 1000 clips most likely to contain a recognizable instance of the person performing one of the predefined named actions. Each query consisted of a set of:

- The name of the target person
- The name of the target action
- 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
- 4 - 6 short sample video clips of the target action
- A text description of the target action

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)

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

Table 3: Ad-hoc search pooling and judging statistics

Topic number	Total submitted	Unique submitted	total that were unique %	Number judged	unique that were judged %	Number relevant	judged that were relevant %
1591	72692	64555	88.81	6115	9.47	705	11.53
1593	73856	70481	95.43	7705	10.93	345	4.48
1594	72936	65249	89.46	6043	9.26	547	9.05
1596	73996	67095	90.67	5321	7.93	57	1.07
1597	73996	66281	89.57	5355	8.08	213	3.98
1598	73936	62872	85.04	5675	9.03	230	4.05
1602	73996	68596	92.70	6238	9.09	1585	25.41
1604	73996	64148	86.69	6495	10.13	905	13.93
1606	73996	61256	82.78	9626	15.71	277	2.88
1610	72942	64411	88.30	7072	10.98	953	13.48
1641	39000	32867	84.27	3416	10.39	723	21.17
1642	39000	31640	81.13	2602	8.22	1042	40.05
1643	39000	34885	89.45	5287	15.16	302	5.71
1644	39000	33874	86.86	4041	11.93	1152	28.51
1645	37502	30863	82.30	4344	14.08	1339	30.82
1646	38734	33868	87.44	4319	12.75	461	10.67
1647	39000	36846	94.48	5094	13.83	1678	32.94
1648	39000	32881	84.31	4331	13.17	826	19.07
1649	39000	30802	78.98	3026	9.82	1804	59.62
1650	39000	33807	86.68	3879	11.47	322	8.30
1651	39000	34875	89.42	3772	10.82	518	13.73
1652	38592	31836	82.49	3363	10.56	597	17.75
1653	39000	33888	86.89	4178	12.33	972	23.26
1654	37502	36810	98.15	4778	12.98	529	11.07
1655	38756	36879	95.16	5139	13.93	569	11.07
1656	39000	31773	81.47	5158	16.23	1234	23.92
1657	39000	31930	81.87	5535	17.33	837	15.12
1658	39000	35877	91.99	5011	13.97	832	16.60
1659	39000	31813	81.57	2155	6.77	441	20.46
1660	39000	32758	83.99	2877	8.78	900	31.28

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)

The task supported 2 types of runs that teams can submit for evaluation:

1. Fully automatic (F) runs: System takes official query as input and produced results without any human intervention.
2. Interactive humans in the loop (I) runs: System takes official query as input and produced results where humans can filter or re-rank search results for up to a period of 5 elapsed minutes per search and 1 user per system run.

In the above both run types, all provided official query image/video examples should be frozen with no human modifications to them.

Table 4: Instance search pooling and judging statistics

Topic number	Total submitted	Unique submitted	total that were unique %	Max. result depth pooled	Number judged	unique that were judged %	Number relevant	judged that were relevant %
9279	26880	25315	94.18	400	4260	16.83	533	12.51
9282	27102	23137	85.37	320	2897	12.52	105	3.62
9284	27999	24769	88.46	200	2068	8.35	75	3.63
9285	27999	21807	77.88	280	2783	12.76	110	3.95
9287	27378	21046	76.87	260	2593	12.32	264	10.18
9290	26950	18193	67.51	240	1922	10.56	73	3.80
9292	27999	21836	77.99	340	3282	15.03	107	3.26
9294	27110	21762	80.27	220	2336	10.73	37	1.58
9295	28000	20386	72.81	340	3327	16.32	389	11.69
9298	27090	18549	68.47	200	1882	10.15	22	1.17
9299	16563	14323	86.48	440	3484	24.32	389	11.17
9300	15170	13708	90.36	300	2038	14.87	261	12.81
9301	16999	12451	73.25	280	2036	16.35	237	11.64
9302	16999	12379	72.82	200	1385	11.19	127	9.17
9303	16969	12680	74.72	500	3977	31.36	270	6.79
9304	17000	13818	81.28	500	2590	18.74	187	7.22
9305	16998	13781	81.07	280	1775	12.88	83	4.68
9306	16937	10061	59.40	200	1474	14.65	28	1.90
9307	16997	10823	63.68	220	1490	13.77	122	8.19
9308	16979	15452	91.01	320	2675	17.31	92	3.44
9309	16999	15411	90.66	520	3373	21.89	191	5.66
9310	16960	11998	70.74	240	1870	15.59	67	3.58
9311	16978	12093	71.23	480	3407	28.17	397	11.65
9312	16978	12025	70.83	200	1653	13.75	28	1.69
9313	16982	14349	84.50	300	1887	13.15	123	6.52
9314	16965	14127	83.27	200	1319	9.34	103	7.81
9315	17000	12307	72.39	340	2135	17.35	48	2.25
9316	16999	11794	69.38	200	1378	11.68	12	0.87
9317	17000	14124	83.08	420	2295	16.25	285	12.42
9318	16998	13231	77.84	280	1660	12.55	155	9.34

Query Topics

NIST viewed a sample of test videos and developed a list of recurring actions and the persons performing these actions. In order to test the effect of persons or actions on the performance of a given query, the topics tested different target persons performing the same actions. Besides the main task with unique queries each year, starting in 2019, a progress subtask was introduced to measure system progress on a set of fixed queries. In total, 20 common queries were released in 2019 and participating systems were allowed to submit results against those queries such that in 2020 NIST will evaluate 10 of those 20 queries to measure progress across two years (2019 - 2020) and evaluate the other 10 queries in 2021 measuring progress across 3 years (2019 - 2021). The 20 common queries comprised of 9 individual persons and 10 specific actions (Appendix D).

A set of 20 unique queries (Appendix C) were released in the main task comprising of 8 individual persons and 9 specific actions. In total, we evaluated those 20 queries in addition to 10 queries from the progress subtask set.

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

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, 5 groups submitted 33 runs including 31 automatic and 2 interactive runs. From the 33 runs, 16 runs belonged to the progress subtask, while 17 belonged to the main 2020 task. In addition to the 16 progress runs in 2020, a set of 12 progress runs were submitted by 3 separate teams in 2019. All 28 runs were evaluated and scored on 10 queries this year.

All run submissions were pooled and then divided into strata based on the rank of the result items. Each strata comprised of 20 rank levels (1-20, 21-40, 41-60, etc) up to rank 520. Finally, all duplicates in each stratum was removed.

For a given topic[4], the submissions for that topic were judged by a NIST human assessor who played each submitted shot and determined if the topic target was present (the target person was seen doing the specific action). 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 200, resulting in 71 251 judged shots including 4 920 total relevant shots (6.9%). Table 4 presents information about the pooling and judging.

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 (MAP) over all queries. While speed and location accuracy were also of interest here, of these two, only speed was reported.

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV20Pubs, 2020] in the online workshop notebook proceedings.

3.3 Disaster Scene Description and Indexing

Computer vision capabilities have rapidly been advancing and are expected to become an important component to incident and disaster response. Having prior knowledge about affected areas can be very helpful for the first responders. Communication systems often go down in major disasters, which makes it very difficult to get any information regarding the damage. Automated systems, such as robots or low flying drones, can therefore, be used to gather information before rescue workers enter the area.

With the popularity of deep learning, computer vision research groups have access to very large image and video datasets for various tasks and the performances of systems have dramatically improved. However, the majority of computer vision capabilities are not meeting public safety's needs, such as support for search and rescue, due to the lack of appropriate training data and requirements. Most current datasets do not have public safety hazard labels due to which state-of-the-art systems trained on these datasets fail to provide helpful labels in disaster scenes.

In response, the MIT Lincoln Lab developed a dataset of images collected by the Civil Air Patrol of various natural disasters. The Low Altitude Disaster Imagery (LADI) dataset was developed as part of a larger NIST Public Safety Innovator Accelerator Program (PSIAP) grant. Two key properties of the dataset are as follows:

1. Low altitude
2. Oblique perspective of the imagery and disaster-related features.

These are rarely featured in computer vision benchmarks and datasets. The LADI dataset acted as a starting point to help label a new video dataset with disaster-related features to be used as testing data in the DSDI task. The image dataset could be used for the training and development of systems for the DSDI task.

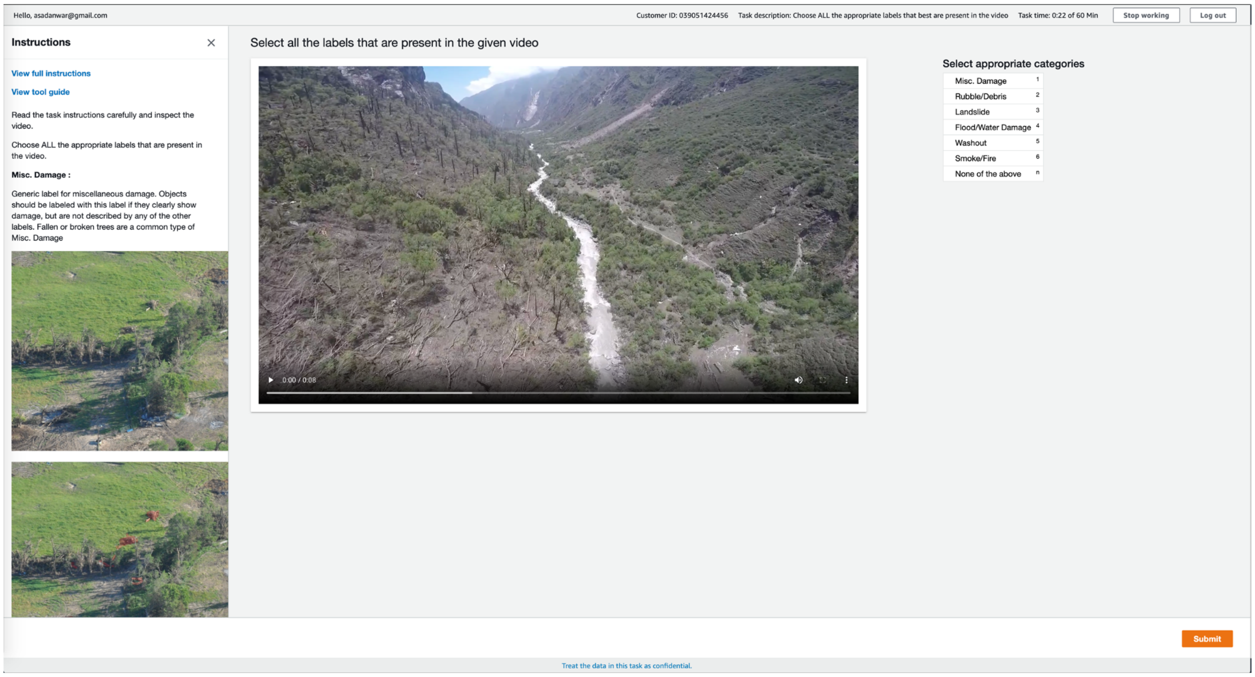
Datasets

Training Dataset The training dataset is based on the LADI dataset hosted as part of the AWS Public Dataset program. It consists of 20 000+ annotated images. The images are from locations with FEMA major disaster declaration for a hurricane or flooding. The lower altitude criteria distinguishes the LADI dataset from satellite datasets to support development of computer vision capabilities with small drones operating at low altitudes. A minimum image size (4MB) was selected to maximize the efficiency

[4] Please refer to Appendix C and D for query descriptions.

Damage	Environment	Infrastructure	Vehicles	Water
Misc. Damage	Dirt	Bridge	Aircraft	Flooding
Flooding/Water Damage	Grass	Building	Boat	Lake/Pond
Landslide	Lava	Dam/Levee	Car	Ocean
Road Washout	Rocks	Pipes	Truck	Puddle
Rubble/Debris	Sand	Utility or Power Lines/Electric Towers		River/Stream
Smoke/Fire	Shrubs	Railway		
	Snow/Ice	Wireless/Radio Communication Towers		
	Trees	Water Tower		
		Road		

Table 5: DSDI: The testing dataset has 5 coarse categories, each divided into 4-9 more specific labels.


Hello, asadanwar@gmail.com

Customer ID: 039051424456 Task description: Choose ALL the appropriate labels that best are present in the video Task time: 0:22 of 60 Min

Instructions

Read the task instructions carefully and inspect the video.

Choose ALL the appropriate labels that are present in the video.

Misc. Damage :

Generic label for miscellaneous damage. Objects should be labeled with this label if they clearly show damage, but are not described by any of the other labels. Fallen or broken trees are a common type of Misc. Damage

Select all the labels that are present in the given video

Select appropriate categories

Treat the data in this task as confidential.

Figure 2: Screenshot of a video being annotated for the Damage category. The annotator watches the video and marks all the labels that are visible in the video.

of the crowd source workers, since lower resolution images are harder to annotate.

Testing Dataset A pilot testing dataset of about 5 hours of video was distributed for this task. The testing dataset was segmented into small video clips (or shots) of a maximum of 20 seconds. The videos were from earthquake, hurricane, and flood affected areas. There were a total of 1825 shots with a median length of 16 seconds.

Categories The categories used for the testing dataset are the same as those used for the LADI training dataset. Five coarse categories were selected based on their importance for the task, and each of these categories is divided into 4-9 more specific labels. The hierarchical labeling scheme is shown in Table 5.

Annotation The video annotation was done using full time annotators instead of crowdsourcing. It is essential that the annotators become familiar with the task and the labels before they start a category. For this reason, we created a practice page for each category with multiple examples for each label within that category. The annotators were also given 2 videos as a test to mark the labels visible to them, and the answers were compared to ours. We also had regular discussions with the annotators to understand their process and clarify any confusions during the labeling of the dataset.

Two full time annotators labeled the testing dataset. The Amazon Augmented AI (Amazon A2I) tool was used during the process. The annotators worked independently on each category. Figure 2 shows a screenshot of the annotation page as visible to annotators. To create the final ground truth, for each shot, the union of the labels were used.

System Task

Systems were required to return a ranked list of up to 1000 shots for each of the 32 features. Each submitted run specified its training type:

- LADI-based (L): The run only used the supplied LADI dataset for development of its system.
- Non-LADI (N): The run did not use the LADI dataset, but only trained using other dataset(s).
- LADI + Others (O): The run used the LADI dataset in addition to any other dataset(s) for training purposes.

Evaluation and Metrics

The evaluation metric used for the task was mean average precision (MAP). The average precision is calculated for each feature, and the mean average precision reported for each submission. Furthermore, the true positive, true negative, false positive, and false negative rates are also reported. Teams self reported the clock time per inference to compare the speeds of the various systems.

In this first year for the task, 17 teams signed up to join the task and finally 9 teams submitted runs. In total, we received 30 runs including 9 LADI+Others (O) runs and 21 LADI-based (L) runs. For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV20Pubs, 2020] in the online workshop notebook proceedings.

3.4 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 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 the technology, such as video summarization in the form of natural language, facilitating the searching 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. Since then, there have been substantial improvements in the dataset and evaluation.

	Matching & Ranking (4 Runs)	Description Generation (19 Runs)
IMFD_IMPRESEE		X
KSLAB		X
KU_ISPL		X
MMCUniAugsburg		X
PICSOM		X
RUC_AIM3	X	X

Table 6: VTT: List of teams participating in each of the subtasks. Description Generation is a core task, whereas Matching and Ranking is optional.

System Task

The VTT task is divided into two subtasks:

- Description Generation Subtask
- Matching and Ranking Subtask

The description generation subtask has been designated as core/mandatory, which means that teams participating in the VTT task must submit at least one run to this subtask. The matching and ranking subtask is optional for the participants. This subtask was initially introduced to ease teams into the difficult video description task. However, with improvements over subsequent years, the subtask was made optional.

Details of the two subtasks are as follows:

- **Description Generation** (Core): For each video, automatically generate a text description of 1 sentence independently and without taking into consideration the existence of any annotated descriptions for the videos.
- **Matching and Ranking** (Optional): In this subtask, 5 sets of text descriptions are provided along with the videos. Each set contains a description for each video in the dataset, but the order of descriptions is randomized. The goal of the subtask is to return for each video a ranked list of the most likely text description that corresponds (was annotated) to that video from each of the 5 sets.

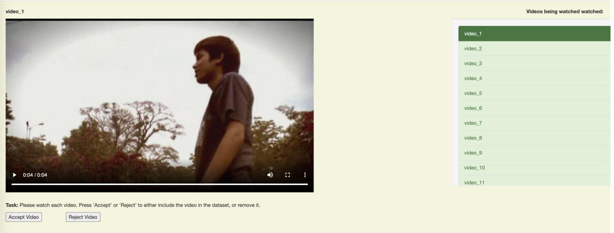
Up to 4 runs were allowed per team for each of the subtasks.

For this year, 6 teams participated in the VTT task. Only 1 team participated in the optional matching and ranking subtask with a total of 4 runs. There were 19 runs submitted for the description generation subtask. A summary of participating teams is shown in Table 6.

Data

The VTT data for 2020 was taken from the V3C2 data collection. In previous years, the VTT testing dataset consisted of Twitter Vine videos, which generally had a duration of 6 seconds. In 2019, we supplemented the dataset with videos from Flickr. The V3C dataset [Rossetto et al., 2019] is a large collection of videos from Vimeo. It also provides us with the advantage that we can distribute the videos rather than links, which may not be available in the future.

For the purpose of this task, we only selected video segments with lengths between 3 and 10 seconds. A total of 1700 video segments were annotated manually by multiple annotators for this year's task.



Task: Please watch each video. Press 'Accept' or 'Reject' to either include the video in the dataset, or remove it.

Figure 3: VTT: Screenshot of video selection tool.

It is important for a good dataset to have a diverse set of videos. We watched over 8000 videos and selected 1700 videos. Figure 3 shows a screenshot of the video selection tool that was used to decide whether a video was to be selected or not. We tried to ensure that the videos covered a large set of topics. If we came across a large number of videos that looked similar to previously selected clips, they were rejected. We also removed the following types of videos:

- Videos with multiple, unrelated segments that are hard to describe, even for humans.

- Any animated videos.
- Other videos that may be considered inappropriate or offensive.

Annotator	Avg. Length	Total Videos Watched
1	16.60	825
2	16.65	875
3	17.67	1700
4	19.62	825
5	21.22	875
6	22.61	875
7	22.71	875
8	24.14	825
9	25.81	825

Table 7: VTT: Average number of words per sentence for all the annotators. A large variation is observed between average sentence lengths for the different annotators. The table also shows the number of videos watched by each annotator. Annotator #3 watched all 1700 videos.

Annotation Process The videos were divided amongst 10 annotators, with each video being annotated by exactly 5 people. One of the annotators had to drop out and their workload was taken by an existing annotator, who wrote descriptions for all 1700 videos.

The annotators were asked to include and combine into 1 sentence, if appropriate and available, four facets of the video they are describing:

- **Who** is the video showing (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)?

Different annotators provide varying amount of detail when describing videos. Some people try to incorporate as much information as possible about the video, whereas others may write more compact sentences. Table 7 shows the average number of words per sentence for each of the annotators. The average sentence length varies from 16.60 words to 25.81 words, emphasizing the difference in descriptions provided by the annotators.

Furthermore, the annotators were also asked the following questions for each video:

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

The average score for the first question was 2.53 (on a scale of 1 to 5), showing that in general the annotators thought the videos were on the easier side to describe. The average score for the second question was 2.24 (on a scale of 1 to 3), meaning that they thought that other people would write a similar description as them for most videos. The two scores are negatively correlated as annotators are more likely to think that other people will come up with similar descriptions for easier videos. The Pearson correlation coefficient between the two questions is -0.61.

Submissions

Systems were required to specify the run types based on the types of training data and features used.

The list of training data types is as follows:

- ‘I’: Training using image captioning datasets only.
- ‘V’: Training using video captioning datasets only.
- ‘B’: Training using both image and video captioning datasets.

The feature types can be one of the following:

- ‘V’: Only visual features are used.
- ‘A’: Both audio and visual features are used.

Evaluation and Metrics

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 was one of the first metrics to achieve a high correlation with human judgments of quality. It is known to perform 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 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 TF-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 SPICE (Semantic Propositional Image Caption Evaluation) metric [Anderson et al., 2016] is another metric that has gained popularity in image captioning evaluation. The metric uses scene graph similarity between generated captions and the ground truth instead of n-grams.

The STS (Semantic Textual Similarity) metric [Han et al., 2013] was also applied to the results, as in the previous years 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) [5] with 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. This metric has been used every year since then to rate each team's primary run, as well as 4 human systems.

3.5 Activities in Extended Video

This year we continue with the ActEV task with 35 target activities that we had started from 2018. NIST TRECVID Activities in Extended Video (ActEV) series was initiated in 2018 to support the Intelligence Advanced Research Projects Activity (IARPA) Deep Intermodal Video Analytics (DIVA) Program. The Activities in Extended Video (ActEV) series of evaluations is designed to accelerate development of robust, multi-camera, automatic activity detection systems for forensic and real-time alerting applications. ActEV began with the Summer 2018 Blind and Leaderboard evaluations and has currently progressed to the running of two concurrent evaluations: 1) the ActEV Sequestered Data Leaderboard (ActEV SDL) based on the Multiview Extended

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

Video (MEVA) dataset [Kitware, 2020] with 37 activities. 2) The TRECVID 2020 ActEV TRECVID self-reported leaderboard based on the VIRAT V1 and V2 datasets [Oh et al., 2011] with 35 activities.

The TRECVID 2018 ActEV (ActEV18) evaluated system detection performance on 12 activities for the self-reported evaluation and 19 activities for the leaderboard evaluation using the VIRAT V1 and V2 dataset [Lee et al., 2018]. For the self-reported evaluation, the participants ran their software on their hardware and configurations and submitted the system output with the defined format to the NIST scoring server. For the leaderboard evaluation, the participants submitted their runnable system to the NIST scoring server, which was independently evaluated on the sequestered data using the NIST hardware.

The ActEV18 evaluation addressed the two different tasks: 1) identify a target activity along with the time span of the activity (AD: activity detection), 2) detect objects associated with the activity occurrence (AOD: activity and object detection).

For the TRECVID 2019 ActEV (ActEV19) evaluation, we primarily focused on the 18 activities and increased the number of instances for each activity. ActEV19 included the test set from both VIRAT V1 and V2 datasets and the systems were evaluated on the activity detection (AD) task only.

The TRECVID 2020 ActEV (ActEV20) self-reported leaderboard is based on the VIRAT V1 and V2 datasets with 35 activities with updated names to make it easier to use the MEVA dataset to train systems for TRECVID ActEV leaderboard.

Figure 4 illustrates an example of representative activities that were used in the TRECVID 2020 ActEV. The evaluation primarily targeted on the forensic analysis that processes the full corpus prior to returning a list of detected activity instances. A total of 4 different organizations participated in this year evaluation (ActEV20) and over ?XY different algorithms were submitted.

In this paper, we first discuss task and dataset used and introduce a new metric to evaluate algorithm performance. In addition, we present the results for the TRECVID20 ActEV submissions and discuss observations and conclusions.

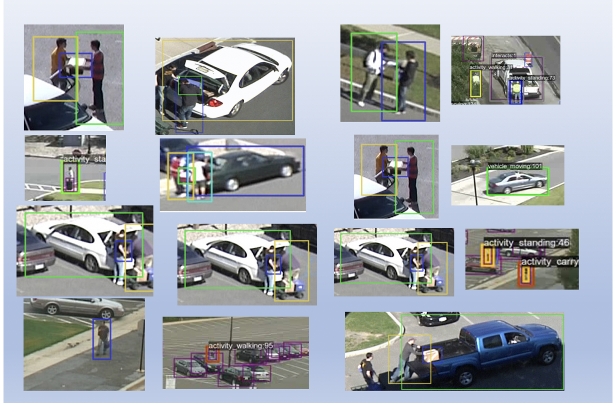


Figure 4: Example of activities for ActEV series. IRB (Institutional Review Board): 00000755

3.6 Task and Dataset

In the ActEV20 leaderboard evaluation, we addressed activity detection (AD) task for detecting and localizing activities; a system required to automatically detects and temporally localizes all instances of the activity. For a system-identified activity instance to be evaluated as correct, the type of activity should be correct, and the temporal overlap should fall within a minimal requirement. The type of the ActEV20 challenge was called an open leaderboard evaluation; the challenge participants should run their software on their systems and configurations and submit the defined system output to the NIST Scoring Server. The leaderboard evaluation should submit a system to report activities that visibly occur in a single-camera video by identifying the video file, the frame span (the start and end frames) of the activity instance, and the presence confidence value indicating the system's "confidence score" how likely the activity is present.

For this evaluation, we used 35 activities from the VIRAT dataset and the activities were annotated by Kitware, Inc. The VIRAT dataset consisted of 29 video hours and more than 43 activity types. A total of 10 video hours were annotated for the test set across 35 activities. The detailed definition of each activity and evaluation requirements are described in the evaluation plan [Godil et al., 2020].

Table 8 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. The numbers of instances are not balanced across activities, which may affect the system performance results.

3.7 Measures

In this evaluation, an activity is defined as "one or more people performing a specified movement or interacting with an object or group of objects (including driving and flying)", while an instance indicates an occurrence (time span of the start and end frames) in associated with the activity.

The primary measure of performance for TRECVID ActEV20 is the normalized, partial Area Under the DET Curve ($nAUDC$) from 0 to a fixed, Time-based False Alarm (T_{fa}) nAUDC Tfa value a , denoted $nAUDC_a$, which is the same as for the TRECVID ActEV19 evaluation.

For TRECVID ActEV18, the primary metric was instance-based measures for both missed detections and false alarms (as illustrated in Figure 5. The metric evaluated how accurately the system detected the instance occurrences of the activity.

As shown in Figure 5, the detection confusion matrix are calculated with alignment between reference and system output on the target activity instances; Correct Detection (CD) indicates that the reference and system output instances are correctly mapped (instances marked in blue). Missed Detection (MD) indicates that an instance in the reference has no correspondence in the system output (instances marked in yellow) while False Alarm (FA) indicates that an instance in the system output has no correspondence in the reference (instances marked in red). 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.

In the ActEV20 evaluation (same as for AvtEV19 evaluation), a probability of missed detections (P_{miss}) and a rate of false alarms (R_{FA}) were used and computed at a given decision threshold:

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

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

where $N_{\text{MD}}\ (\tau)$ is the number of missed detections at the threshold τ , $N_{\text{FA}}\ (\tau)$ is the number of false alarms, and *VideoDurInMinutes* is number of minutes of video. $N_{\text{TrueInstance}}$ is the number of reference instances annotated in the sequence. Lastly, the Detection Error Tradeoff (DET) curve [Martin et al., 1997] is used to visualize system performance. For the TRECVID ActEV18 challenges two years ago, we evaluated algorithm performance on the operating points; P_{miss} at $R_{\text{FA}}\ =\ 0.15$ and P_{miss} at $R_{\text{FA}} = 1$.

To understand system performance better and to be more relevant to the user cases, for ActEV20 and, we used the normalized, partial area under the DET curve ($nAUDC$) from 0 to a fixed time-based false alarm (T_{fa}) to evaluate algorithm performance. The partial area under DET curve is computed separately for each activity over all videos in the test collection and then is normalized to the range [0, 1] by dividing by the maximum partial area $nAUDC_a = 0$ is a perfect score. The $nAUDC_a$ is defined as:

$$nAUDC_a = \frac{1}{a}\int_{x=0}^{a} P_{miss}(x)dx, x = T_{fa}$$

where x is integrated over the set of T_{fa} values. The instance-based probability of missed detections P_{miss} is defined as:

$$P_{miss}(x) = \frac{N_{md}(x)}{N_{TrueInstance}}$$

where $N_{md}(x)$ is the number of missed detections at the presence confidence threshold that result in $T_{fa}\ =\ x$ (see the below equation for the details). $N_{TrueInstance}$ is the number of true instances in the sequence of reference.

The time-based false alarm T_{fa} is defined as:

$$T_{fa} = \frac{1}{NR}\sum_{i=1}^{N_{frames}} \max(0, S'_i - R'_i)$$

where N_{frames} is the duration of the video and NR is the non-reference duration; the duration of the video without the target activity occurring. S'_i is the total count of system instances for frame i while R'_i is the total count of reference instances for frame i. The detailed calculation of T_{fa} is illustrated in Figure 6.

The non-reference duration (NR) of the video where no target activities occurs is computed by constructing a time signal composed of the complement of the union of the reference instances duration. R is the reference instances and S is the system instances. R' is the histogram of the count of reference instances and S' is the histogram of the count of system instances for the target activity. R' and S' both have N_{frames} bins, thus R'_i is the value of the i^{th} bin R' while S'_i is the value of the i^{th} bin S'. S' is the total count of system instances in frame i and R' is the total count of reference instances in frame i. False alarm time is computed by summing over positive difference of $S' - R'$(shown in red in Figure 6); that is the duration of falsely detected system instances.

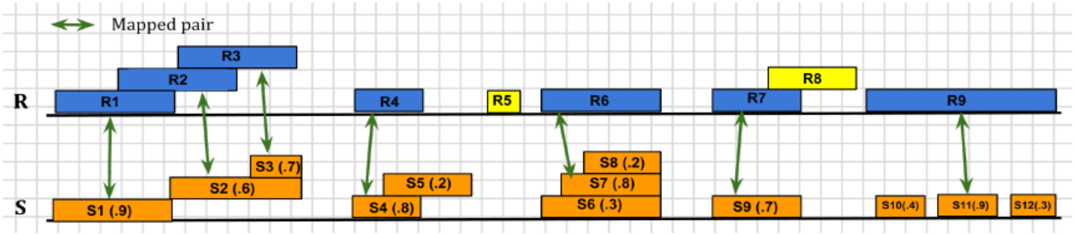


Figure 5: Illustration of activity instance alignment and P_{miss} calculation (R is the reference instances and S is the system instances. In S, the first number indicates instance id and the second indicates presence confidence score. For example, $S1(.9)$ represents the instance $S1$ with corresponding confidence score .9. Green arrows indicate aligned instances between R and S)

This value is normalized by the non-reference duration of the video to provide the T_fa value in Equation above.

Figure 7 shows visual representations of the major differences between the ActEV18 and ActEV19/ActEV20 metrics. For the ActEV18 metric, we used Instance-based Rate of false alarms and system performance was evaluated at the specific operating point as illustrated in the left DET. For the ActEV19/ActEV20 metric, we used Time-based false alarms and calculated $nAUDC$ from T_{fa} 0 to 0.2.

3.8 Video Summarization

An important need in many situations involving video collections (archive video search/reuse, personal video organization/search, movies, tv shows, etc.) is to summarize the video in order to reduce the size and concentrate the amount of high value information in the video track. From 2020 we begin a new video summarization track in TRECVID in which the task is to summarize the major life events of specific characters over a number of weeks of programming on the BBC Eastenders TV series. Typically, three characters will be chosen for this task every year, and summaries of their major life events must be between the selected period of the show, which will be specified to participants in advance of the task.

The use case for this task is to generate an automatic summary, using a predefined maximum number of unique shots, of the significant life events of a given character from the Eastenders series over a given number of episodes. The generated summaries should be enough to gain a clear and concise overview of that characters major life events over the course of 8 - 12 weeks of programming in the series, and to see how they intertwine with the major life events of other specified characters in that time frame of the series.

Video Summarization Data

In 2020 this 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).

System task

Given a collection of BBC Eastenders videos, a master shot boundary reference, a list of characters from the series, and a time frame of the series for which to use for summarization, summarize the major life events of each character within the specified time frame of the series. Some examples of major life events are more likely to be: The birth of a child rather than a short illness, A divorce rather than an argument with a loved one, the passing of a loved one rather than the passing of someone loosely known to you, etc., etc. Summaries are limited to a maximum number of unique shots, thus the main challenge is to select those shots most likely to be considered a major life event by human assessors.

Each topic consisted of a set of 4 example frame images (bmp) drawn from test videos containing the person of interest in a variety of different appearances to the extent possible.

For each frame image (of a target person) there

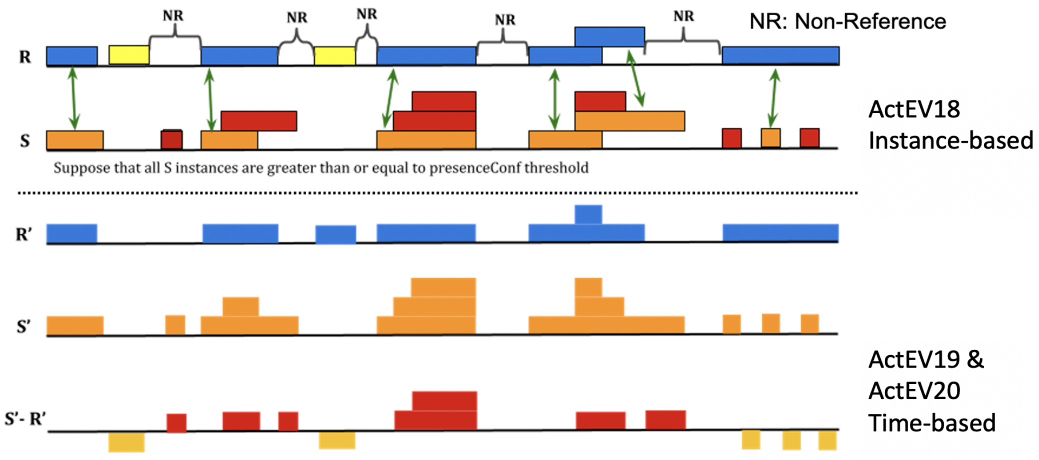


Figure 6: Comparison of instance-based and time-based false alarms. R is the reference instances and S is the system instances. R' is the histogram of the count of reference instances and S' is the histogram of the count of system instances for the target activity. S shows a depiction of instance-based false alarms while $S' - R'$ illustrates time-based false alarms as marked in red.

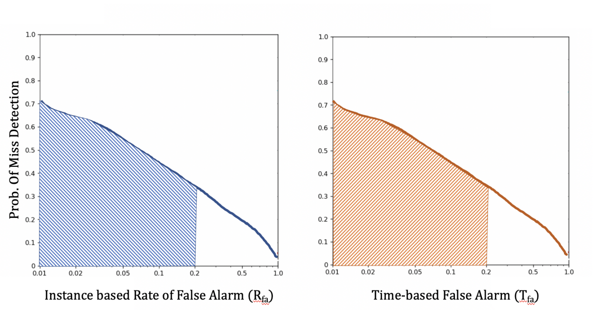


Figure 7: Comparison of ActEV18 (R_{fa}) and ActEV20 (T_{fa}) measures using the Detection Error Tradeoff (DET) curves

was a binary mask of the region of interest (ROI), as bounded by a single polygon and the ID from the master shot reference of the shot from which the image example was taken. In creating the masks (in place of a real searcher), we assume the searcher wants to keep the process simple. So, the ROI may contain non-target pixels, e.g., non-target regions visible through the target or occluding regions.

Topics

By analysing meta-data of the full set of BBC Eastenders omnibus episodes, NIST selected queries of three characters who were shown to play a big part in the series over a ten week period. The following three characters were selected:

- Janine
- Ryan
- Stacey

In addition to specifying this years query characters, the time frame of the series (Start Shot # and End Shot #), links to images of the query characters, and the maximum length and number of shots for each run were also disseminated to participating teams. These are indicated in Table 9.

Evaluation

Each group was asked to submit 4 runs, with the maximum number of shots and maximum summary length as specified in Table 9. In total, 2 groups submitted 8 runs, with each run containing video summaries for each of the 3 specified queries, giving a total of 24 video summaries to be evaluated.

Submissions were evaluated by the TRECVID team at NIST, with one person responsible for evaluating summaries for a single query. Assessors answered 5 content based questions for each of the 8 video summaries they had been asked to evaluate. Content questions were created by the TRECVID team after watching each episode of the specified time frame of the series, marking those scenes they considered to be important, reducing these to 5 specific scenes based on what they considered to be the 5 most important scenes for each query, and finally voting on these as a group to establish the final 5 most important scenes for each character. From each of these, a question was worded to ask if the submitted video summary *could be said* to have answered that question. The content questions for each character are specified below:

Janine

1. What is causing Ryan to be sick in bed?
2. How does Janine attempt to kill Ryan while in the hospital?
3. What happens when Janine attempts to play recording of Stacey?
4. Who stabbed Janine?
5. Who gives Janine the recording of Stacey?

Ryan

1. How does Janine attempt to kill Ryan in the hospital?
2. What does Ryan do when Janine is lying in the hospital?
3. Where is Ryan trapped?
4. What does Ryan tell Phil he can do for him?
5. Who is Ryan with when going to put his name on the baby's birth cert?

Stacey

1. Who climbs up to the roof to talk Stacey out of jumping off?
2. What does Stacey reveal when in a cell with Janine, Kat, and Pat?
3. What does Stacey admit to her mum in bedroom when mum is upset?
4. Who confronts Stacey in restroom where Stacey finally admits to killing Archie?
5. Who calls to Stacey's door to tell her to get her stuff and go, after Stacey's mum had called the police?

Assessors also marked video summaries on the subjective metrics of tempo/rhythm, contextuality, and redundancy, on a 7-point Likert-scale, with the following definitions. **Tempo/rhythm** was defined as: *How well do the video shots flow together? Do shots cut mid-sentence (indicating poor tempo/rhythm)? Do they flow together nicely so it wouldn't be obvious that this is an automatically generated summary (high tempo/rhythm)? (High is best).* **Contextuality** was defined as: *Does the content provide the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed? (High is best).* **Redundancy** was defined as: *Does the video contain content considered to be unnecessary or superfluous? (Low is best).*

Metrics

Scores were calculated as a percentage using marks for the 5 content based questions and the 3 subjective quality based questions. Base Likert-scale scores for Tempo/rhythm and contextuality were taken as assessed by human annotators. Scores for redundancy, where a lower score is best, were flipped. This gave a total of 21 possible marks available for subjective quality scores. The remainder was calculated by taken the remaining 79 possible marks and dividing by the 5 content based questions, giving a total of 15.8 possible marks for each correct content based question which was to be rounded to the nearest integer. This would give a perfect summary 100 points. A summary with no relevant content but all perfect scores for the other factors would get 21 points. Overall this gave summaries a maximum score of 100 down to a minimum score of 3.

4 Summing up and moving on

In this overview paper to TRECVID 2020, we provided basic information for all tasks we run this year and particularly on the goals, data, evaluation mechanisms, and metrics used. 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 [TV20Pubs, 2020]. Finally, we are looking forward to continue a new evaluation cycle in 2021 after refining the current tasks and introducing any potential new tasks.

5 Authors' note

TRECVID would not have happened in 2020 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.
- Luca Rossetto of University of Basel for providing the V3C dataset collection.
- 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 licence agreement for the Eastenders data.
- Jeffrey Liu and Andrew Weinert of MIT Lincoln Laboratory for supporting the DSDI task by making the LADI dataset available and helping with the testing dataset preparations.

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

6 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

[Anderson et al., 2016] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *ECCV*.

[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.

[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.

[Godil et al., 2020] Godil, A., Lee, Y., and Fiscus, J. (2020). TRECVID 2020 actev evaluation plan. https://actev.nist.gov/pub/TRECVID_2020_ActEV_EvaluationPlan.pdf.

[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.

[Kitware, 2020] Kitware (2020). MEVA Data Website. https://www.mevadata.org. Accessed: 2020-03-12.

[Lee et al., 2018] Lee, Y., Godil, A., Joy, D., and Fiscus, J. (2018). TRECVID 2019 actev evaluation plan. https://actev.nist.gov/pub/Draft_ActEV_2018_EvaluationPlan.pdf.

[Martin et al., 1997] Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proceedings*, pages 1895–1898.

[NIST, 2020] NIST (2020). ActEV Sequestered Data Leaderboard Website. https://actev.nist.gov/sdl. Accessed: 2020-03-12.

[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.

[Rossetto et al., 2019] Rossetto, L., Schuldt, H., Awad, G., and Butt, A. A. (2019). V3C–a research video collection. In *International Conference on Multimedia Modeling*, pages 349–360. Springer.

[TV20Pubs, 2020] TV20Pubs (2020). http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.20.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 8: A list of 35 activities on the VIRAT dataset and their associated number of instances for the train and validation sets

Activity Type	Train	Validate
person_closes_facility_or_vehicle_door	141	130
person_closes_trunk	21	31
vehicle_drops_off_person	0	4
person_enters_facility_or_vehicle	77	70
person_exits_facility_or_vehicle	66	72
person_interacts_object	101	88
person_loads_vehicle	38	38
person_opens_trunk	22	35
person_opens_facility_or_vehicle_door	137	128
person_person_interaction	11	17
person_pickups_object	19	12
vehicle_picks_up_person	9	5
person_pulls_object	23	43
person_pushs_object	4	6
person_rides_bicycle	22	21
person_sets_down_object	12	11
person_talks_to_person	41	67
person_carries_heavy_object	31	44
person_unloads_vehicle	32	44
person_carries_object	237	364
person_crouches	1	9
person_gestures	82	148
person_runs	14	18
person_sits	21	11
person_stands	398	819
person_walks	761	901
person_talks_on_phone	17	16
person_texts_on_phone	5	20
person_uses_tool	7	11
vehicle_moves	718	797
vehicle_starts	259	239
vehicle_stops	292	295
vehicle_turns_left	152	176
vehicle_turns_right	149	172
vehicle_makes_u_turn	9	13

Table 9: Video Summarization Queries and Specifics

Character	Janine	Ryan	Stacey
Start Shot #	shot175_1	shot175_1	shot175_1
End Shot #	shot185_1736	shot185_1736	shot185_1736
Max # Shots Run 1	5	5	5
Max Summary Length Run 1	150 seconds	150 seconds	150 seconds
Max # Shots Run 2	10	10	10
Max Summary Length Run 2	300 seconds	300 seconds	300 seconds
Max # Shots Run 3	15	15	15
Max Summary Length Run 3	450 seconds	450 seconds	450 seconds
Max # Shots Run 4	20	20	20
Max Summary Length Run 4	600 seconds	600 seconds	600 seconds

A Ad-hoc query topics - 20 unique

641 Find shots showing an aerial view of buildings near water in the daytime
642 Find shots of a person paddling kayak in the water
643 Find shots of people dancing or singing while wearing costumes outdoors
644 Find shots of sailboats in the water
645 Find shots of a person wearing a necklace
646 Find shots of a woman sitting on the floor
647 Find shots of people or cars moving on a dirt road
648 Find shots of a man in blue jeans outdoors
649 Find shots of someone jumping while snowboarding
650 Find shots of one or more people drinking wine
651 Find shots of one or more people skydiving
652 Find shots of a little boy smiling
653 Find shots of group of people clapping
654 Find shots of one or more persons exercising in a gym
655 Find shots of one or more persons standing in a body of water
656 Find shots of a long haired man
657 Find shots of a woman with short hair indoors
658 Find shots of two or more people under a tree
659 Find shots of a church from the inside
660 Find shots of train tracks during the daytime

B Ad-hoc query topics - 20 progress topics

591 Find shots of a person holding an opened umbrella outdoors
592 Find shots of a person reading a paper including newspaper
593 Find shots of one or more women models on a catwalk demonstrating clothes
594 Find shots of people doing yoga
595 Find shots of a person sleeping
596 Find shots of fishermen fishing on a boat
597 Find shots of a shark swimming under the water
598 Find shots of a man in a clothing store
599 Find shots of a person in a bedroom
600 Find shots of a person's shadow
601 Find shots of a person jumping with a motorcycle
602 Find shots of a person jumping with a bicycle
603 Find shots of people hiking
604 Find shots of bride and groom kissing
605 Find shots of a person skateboarding
606 Find shots of people queuing
607 Find shots of two people kissing who are not bride and groom
608 Find shots of two people talking to each other inside a moving car
609 Find shots of people walking across (not down) a street in a city
610 Find shots showing electrical power lines

C Instance search topics - 20 unique

9299 Find Ian sitting on couch

9300 Find Billy sitting on couch

9301 Find Ian Holding paper - including photos/envelope,notebooks, magazines, etc

9302 Find Bradley Holding paper - including photos/envelope,notebooks, magazines, etc

9303 Find Billy Holding paper - including photos/envelope,notebooks, magazines, etc

9304 Find Max Drinking

9305 Find Dot Drinking

9306 Find Pat Holding cloth - including jackets, coats, kitchen towels, cleaning towels, etc

9307 Find Heather Holding cloth - including jackets, coats, kitchen towels, cleaning towels, etc

9308 Find Ian Crying

9309 Find Heather Crying

9310 Find Max smoking a cigarette - including holding a cigarette between fingers

9311 Find Dot smoking a cigarette - including holding a cigarette between fingers

9312 Find Pat smoking a cigarette - including holding a cigarette between fingers

9313 Find Stacey Laughing

9314 Find Pat Laughing

9315 Find Max Going up or down the stairs

9316 Find Bradley Going up or down the stairs

9317 Find Max holding a phone / handset - including talking on phone

9318 Find Stacey holding a phone / handset - including talking on phone

D Instance search topics - 20 progress topics

9279 Find Phil Sitting on a couch

9280 Find Heather Sitting on a couch

9281 Find Jack Holding phone

9282 Find Heather Holding phone

9283 Find Phil Drinking

9284 Find Shirley Drinking

9285 Find Jack Kissing

9286 Find Denise Kissing

9287 Find Phil Opening door and entering room / building

9288 Find Sean Opening door and entering room / building

9289 Find Shirley Shouting

9290 Find Sean Shouting

9291 Find Stacey Hugging

9292 Find Denise Hugging

9293 Find Max Opening door and leaving room / building

9294 Find Stacey Opening door and leaving room / building

9295 Find Max Standing and talking at door

9296 Find Dot Standing and talking at door

9297 Find Jack Closing door without leaving

9298 Find Dot Closing door without leaving