TRECVID 2019: An evaluation campaign to benchmark Video Activity Detection, Video Captioning and Matching, and Video Search & retrieval

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} Andrew Delgado {andrew.delgado@nist.gov}
Jesse Zhang {jesse.zhang@nist.gov}
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

Eliot Godard {eliot.godard@nist.gov}
Guest Researcher, NIST, USA

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

November 7, 2019

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2019 was a TREC-style video analysis and retrieval evaluation, the goal of which remains to promote progress in research and development of content-based exploitation and retrieval of information from digital video via open, metrics-based evaluation.

Over the last nineteen years this effort has yielded a better understanding of how systems can effectively accomplish such processing and how one can reliably benchmark their performance. TRECVID is funded by NIST (National Institute of Standards and Technology) and other US government agencies. In addition, many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2019 represented a continuation of four tasks from TRECVID 2018. In total, 27 teams (see Table 1) from various research organizations worldwide completed one or more of the following four tasks:

1. Ad-hoc Video Search (AVS)
2. Instance Search (INS)
3. Activities in Extended Video (ActEV)
4. Video to Text Description (VTT)

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

This year TRECVID used a new vimeo creative commons collection dataset (V3C1) of about 1000 hours in total and segmented into 1 million short video shots. The dataset is drawn from the vimeo video sharing website under the creative common licenses and 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, while the video to text description task used a combination of 1044 Twitter social media Vine videos collected through the online Twitter API public stream and another 1010 short Flicker videos.

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 Ad-hoc search, instance search results were judged by NIST human assessors, while 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.

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 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 [TV19Pubs, 2019].

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, NIST, or the U.S. Government.

2 Datasets

2.1 BBC EastEnders Instance Search Dataset

The BBC in collaboration the European Union's AXES project made 464 h of the popular and long-running soap opera EastEnders available to TRECVID for research 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 actions.

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 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 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 available. While the V3C1 dataset was adopted for testing, the previous Internet Archive datasets (IACC.1-3) of about 1 800 h were available for development and training.

2.3 Activity Detection VIRAT Dataset

The VIRAT Video Dataset [Oh et al., 2011] is a large-scale surveillance video dataset designed to as-

Table 1: Participants and tasks

Task				Location	TeamID	Participants
INS	VTT	ActEv	AV			
− − −	VTT	− − − − −	AVS	Eur	EURECOM	EURECOM
− − −	VTT	− − − − −	− − −	Asia	FDU	Fudan University
− − −	VTT	− − − − −	− − −	Asia	KU_ISPL	Korea University
− − −	∗ ∗ ∗	ActEv	∗ ∗ ∗	Aus	MUDSML	Monash University
− − −	VTT	− − − − −	− − −	Eur	PicSOM	Aalto University
INS	∗ ∗ ∗	− − − − −	− − −	Asia	PKU_ICST	Peking University
− − −	− − −	− − − − −	AVS	Eur	SIRET	Charles University
INS	− − −	ActEv	− − −	Eur	HSMW_TUC	University of Applied Sciences Mittweida Chemnitz University of Technology
− − −	VTT	− − − − −	− − −	Aus	UTS_ISA	Centre for Artificial Intelligence, University of Technology Sydney
− − −	VTT	− − − − −	− − −	Eur	Insight_DCU	Insight Dublin City University
− − −	VTT	− − − − −	AVS	NAm + SAm	IMFD_IMPRESEE	Millennium Institute Foundational Research on Data (IMFD) Chile; Impresee Inc ORAND S.A. Chile
∗ ∗ ∗	− − −	ActEv	∗ ∗ ∗	Eur	ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas
− − −	− − −	∗ ∗ ∗ ∗ ∗	AVS	Asia	kindai_kobe	Dept. of Informatics, Kindai University Graduate School of System Informatics, Kobe University
− − −	− − −	ActEv	− − −	Asia	NTT_CQUPT	NTT Media Intelligence Laboratories Chongqing University of Posts and Telecommunications
− − −	∗ ∗ ∗	− − − − −	AVS	Asia	WasedaMeiseiSoftbank	Waseda University; Meisei University; SoftBank Corporation
INS	− − −	ActEv	− − −	Asia	BUPT_MCPRL	Beijing University of Posts and Telecommunications
− − −	VTT	− − − − −	− − −	Asia	KsLab	Nagaoka University of Technology
INS	∗ ∗ ∗	ActEv	∗ ∗ ∗	Asia	NII_Hitachi_UIT	National Institute of Informatics; Hitachi, Ltd; University of Information Technology, VNU-HCM
− − −	VTT	− − − − −	− − −	Asia	RUC_AIM3	Renmin University of China
− − −	VTT	− − − − −	AVS	Asia	RUCMM	Renmin University of China; Zhejiang Gongshang University
− − −	− − −	ActEv	AVS	Asia	VIREO	City University of Hong Kong
INS	− − −	− − − − −	− − −	Asia	WHU_NERCMS	National Engineering Research Center for Multimedia Software
− − −	− − −	− − − − −	AVS	NAm	FIU_UM	Florida Intl. University; University of Miami
− − −	− − −	ActEv	− − −	NAm	UCF	University of Central Florida
− − −	− − −	ActEv	− − −	Eur	FraunhoferIOSB	Fraunhofer IOSB and Karlsruhe Institute of Technology (KIT)
INS	∗ ∗ ∗	ActEv	AVS	NAm + Asia + Aus	Inf	Monash University; Renmin University; Shandong University
− − −	∗ ∗ ∗	− − − − −	AVS	Asia	ATL	Alibaba group, ZheJiang University

Task legend. INS:Instance search; VTT:Video-to-Text; ActEv:Activities in Extended videos; AVS:Ad-hoc search; − −:no run planned; ∗ ∗ ∗:planned but not submitted

Table 2: Participants who did not submit any runs

Task				Location	TeamID	Participants
INS	*VTT*	*ActEv*	*AVS*			
* * *	− − −	− − − − −	* * *	*Eur*	*JRS*	JOANNEUM RESEARCH
− − −	* * *	* * * * *	* * *	*Eur*	*MediaMill*	University of Amsterdam
* * *	− − −	− − − − −	− − −	*Asia*	*IOACAS*	University of Chinese Academy of Sciences
* * *	− − −	− − − − −	* * *	*Asia*	*D_A777*	Malla Reddy College of Engineering Technology, Department of Electronics and communication Engineering
− − −	* * *	* * * * *	− − −	*NAm*	*Arete*	Scientific Computing Data Analytics Image Processing and Computer Vision
− − −	* * *	− − − − −	− − −	*Asia*	*GDGCV*	G D Goenka University
− − −	* * *	− − − − −	− − −	*Asia*	*MAGUS_ITAI.Wing*	Nanjing University ITAI
− − −	− − −	* * * * *	* * *	*Asia*	*TokyoTech_AIST*	Tokyo Institute of Technology, National Institute of Advanced Industrial Science and Technology
* * *	− − −	* * * * *	* * *	*NAm + Asia*	*TeamCRN*	Microsoft Research; Singapore Management University; University of Washington
− − −	− − −	* * * * *	− − −	*NAm*	*USF*	University of South Florida, USF
* * *	− − −	− − − − −	* * *	*Aus*	*MIAOTEAM*	University of Technology Sydney
− − −	− − −	− − − − −	* * *	*Asia*	*MET*	Sun Yet-sen University

Task legend. INS:instance search; VTT:Video-to-Text; ActEv:Activities in extended videos; AVS:Ad-hoc search; − −:no run planned; **:planned but not submitted

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

2.4 Twitter Vine Videos

A dataset of about 50 000 video URL using the public Twitter stream API have been collected by NIST. Each video duration is about 6 sec. A list of 1 044 URLs was distributed to participants of the video-to-text task. The previous years' testing data from 2016-2018 were also available for training (a set of about 5700 Vine URLs and their ground truth descriptions).

2.5 Flicker Videos

Robin Aly at the University of Twente worked in consultation with NIST to collect Flickr video dataset available under a Creative Commons license for research. The videos were then divided into segments of about 10s in duration. A set of 91 videos divided into 74 958 files was chosen independently by NIST. This year a set of about 1000 segmented video clips were selected randomly to complement the Twitter vine videos for the video-to-text task testing dataset.

3 Ad-hoc Video Search

This year we continued the Ad-hoc video search task that was resumed again in 2016 but adopted a new dataset (V3C1). 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 V3C1 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 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. 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 2019 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] 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

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.

A new progress subtask was introduced this year with the objective of measuring system progress on a set of 20 fixed topics. As a result, this year systems were allowed to submit results for 30 query topics (see Appendix A for the complete list) to be evaluated in 2019 and additional results for 20 common topics (not evaluated in 2019) that will be fixed for three years (2019-2021). Next year in 2020 the evaluated progress runs can measure their system progress against two years (2019-2020) while in 2021 they can measure their progress against three years.

A new extra one "Novelty" run type was 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.

3.1 Ad-hoc Data

The V3C1 dataset (drawn from a larger V3C video dataset [Rossetto et al., 2019]) was adopted as a testing data. 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 will 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.

3.2 Evaluation

Each group was allowed to submit up to 4 prioritized runs per submission type, and per task type (main or

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

progress) and two additional if they were "no annotation" runs. In addition, one novelty run type was allowed to be submitted within the main task.

In fact 10 groups submitted a total of 85 runs with 47 main runs and 38 progress runs. The 47 main runs consisted of 37 fully automatic, and 10 manually-assisted runs.

For each query topic, pools were created and randomly sampled as follows. The top pool sampled 100 % of clips ranked 1 to 250 across all submissions after removing duplicates. The bottom pool sampled 11.1 % of ranked 251 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, 181 649 clips were judged while 256 753 clips fell into the unjudged part of the overall samples. Total hits across the 30 topics reached 23 549 with 10 910 hits at submission ranks from 1 to 100, 8428 hits at submission ranks 101 to 250 and 4211 hits at submission ranks between 251 to 1000.

3.3 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 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 (≈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 [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.

3.4 Ad-hoc Results

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV19Pubs, 2019] in the online workshop notebook proceedings.

4 Instance search

An important need in many situations involving video collections (archive video search/reuse, personal video organization/search, surveillance, law enforcement, protection of brand/logo use) is to find more video segments of a certain specific person, object, or place, given one or more visual examples of the specific item. Building on work from previous years in the concept detection task [Awad et al., 2016] the instance search task seeks to address some of these needs. For six years (2010-2015) the instance search task tested systems on retrieving specific instances of individual objects, persons and locations. From 2016 to 2018, a new query type, to retrieve specific persons in specific locations had been introduced. From 2019, a new query type has been introduced to retrieve instances of named persons doing named actions.

4.1 Instance Search Data

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

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

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

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 %
9249	27122	7343	27.07	520	4360	59.38	439	10.07
9250	27225	8100	29.75	520	4827	59.59	367	7.60
9251	27029	7324	27.10	520	4178	57.05	241	5.77
9252	27228	7225	26.54	520	4332	59.96	352	8.13
9253	27031	7144	26.43	520	4086	57.19	575	14.07
9254	27092	7615	28.11	520	4461	58.58	524	11.75
9255	27278	8835	32.39	520	5153	58.32	275	5.34
9256	27220	9359	34.38	520	5309	56.73	250	4.71
9257	27073	8456	31.23	520	4979	58.88	178	3.58
9258	27418	8169	29.79	520	4894	59.91	41	0.84
9259	27344	8483	31.02	520	5322	62.74	91	1.71
9260	27212	7102	26.10	520	4350	61.25	56	1.29
9261	27162	6627	24.40	520	4185	63.15	234	5.59
9262	27543	8174	29.68	520	4766	58.31	229	4.80
9263	28000	9524	34.01	520	5801	60.91	46	0.79
9264	28000	7964	28.44	520	4895	61.46	91	1.86
9265	27759	7471	26.91	520	4677	62.60	196	4.19
9266	27964	7627	27.27	520	4565	59.85	499	10.93
9267	27122	7701	28.39	520	4697	60.99	35	0.75
9268	27140	8661	31.91	520	4924	56.85	39	0.79
9269	25085	8122	32.38	520	4505	55.47	139	3.09
9270	25070	7454	29.73	520	4543	60.95	273	6.01
9271	25040	9929	39.65	520	5478	55.17	101	1.84
9272	26000	9073	34.90	520	5268	58.06	115	2.18
9273	25905	8515	32.87	520	4816	56.56	139	2.89
9274	25167	6410	25.47	520	3847	60.02	487	12.66
9275	25641	7192	28.05	520	4550	63.28	471	10.35
9276	25940	8995	34.68	520	4905	54.53	29	0.59
9277	25068	7749	30.91	520	4589	59.22	40	0.87
9278	25059	7242	28.90	520	4337	59.89	40	0.92

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 action 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 performing one of the 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)

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 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. In total, this year we provided 30 unique queries to be evaluated this year, in addition to 20 common queries which will be stored and evaluated in later years and used to measure teams progress year-on-year (10 will be evaluated in 2020 to measure 2019-2020 progress, 10 remaining queries will be evaluated in 2021 to measure 2019-2021 progress). 12 progress runs were submitted by 3 separate teams in 2019. The 30 unique queries provided for this years task comprised of 10 individual persons and 12 specific actions. The 20 common queries which will be evaluated in later years comprised of 9 individual persons and 10 specific actions (Appendix B).

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.

4.4 Evaluation

Each group was allowed to submit up to 4 runs (8 if submitting pairs that differ only in the sorts of examples used). In total, 6 groups submitted 26 automatic and 2 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 141 599 judged shots including 6 592 total relevant shots (4.66%). Table 3 presents information about the pooling and judging.

4.5 Measures

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

4.6 Instance Search Results

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV19Pubs, 2019] in the online workshop notebook proceedings.

5 Activities in Extended Video

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. ActEV is an extension of the TRECVID Surveillance Event Detection (SED) [Michel et al., 2017] evaluations where systems only detected and temporally localized activities. The ActEV series are designed to accelerate development of robust automatic activity detection in multi-camera views for forensic and real-time alerting applications in mind. The previous TRECVID 2018 ActEV ran on 12 activities from the VIRAT V1 dataset [Lee et al., 2018] and addressed the two different tasks: 1) identify a target activity along with the time span of the activity (AD: activity detection),

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

2) detect objects associated with the activity occurrence (AOD: activity and object detection). For the TRECVID 2019 ActEV evaluation, we increased the number of activities to 18 from both VIRAT V1 and V2 datasets and focused on the activity detection (AD) task only. 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 9 different organizations were participated in this year evaluation and over 1000 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 TRECVID19 ActEV submissions and discuss observations and conclusions.

5.1 Task and Dataset

In this evaluation, there is one activity detection (AD) task for detecting and localizing activities. That is, given a target activity, a system 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 must be correct, and the temporal overlap must fall within a minimal requirement(see details in the evaluation plan [Godil et al., 2019]). The type of the TRECVID 2019 ActEV challenge is called open leaderboard evaluation; the challenge participants should run their software on their systems and configurations and submit the system output to the TRECVID 2019 ActEV Scoring Server. The leaderboard evaluation are expected 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, and the Presence Confidence value indicating the system's "confidence score" that the activity is present.

As shown in Table 4, for the TRECVID19 ActEV leaderboard evaluation, we used 18 activities from the VIRAT V1 and V2 dataset [Oh et al., 2011] that were annotated by Kitware, Inc. The VIRAT dataset consists of 29 video hours and 23 activity types. A total of 10 video hours were annotated for the test set across 18 activities. The detailed definition of each activity is described in the evaluation plan [Godil et al., 2019].

Table 4: A list of 18 activities and their associated number of instances on VIRAT V1 and V2

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
Pull	21	22
Riding	21	22
Talking	67	41
Activity_carrying	364	237
Specialized_talking_phone	16	17
Specialized_texting_phone	20	5

The numbers of instances are not balanced across activities, which may affect the system performance results.Table 4 lists the number of instances for each activity for the train and validation sets. Due to ongoing evaluations, the test sets are not included in the table.

5.2 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, while an instance indicates an occurrence (time span of the start and end frames) in associated with the activity.

For the past year TRECVID18 ActEV, the primary metric was instance-based measures (as illustrated in Figure 1 and evaluated how accurately the system detected the instance occurrences of the activity.

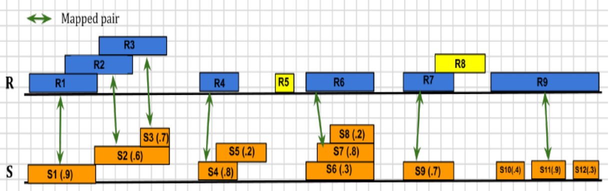

Figure 1: depiction of activity instance alignment and Pmss calculation (In S, the first number indicates instance id and the second indicates presenceConf score. For example, S1 (.9) represents the instance S1 with corresponding confidence score 0.9. Green arrows indicate aligned instances between R and S)

It calculates the detection confusion matrix for activity instance; Correct Detection (CD) indicates that the reference and system output instances are correctly mapped. Missed Detection (MD) indicates that an instance in the reference has no correspondence in the system output while False Alarm (FA) indicates that an instance in the system output has no correspondence in the reference. After calculating the confusion matrix, we summarize system performance: for each instance, a system output provides a confidence score that indicates how likely the instance is associated with the target activity. The confidence score can be used as a decision threshold, enabling a probability of missed detections (P_{Miss}) and a rate of false alarms (R_{FA}) to be computed at a given threshold:

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

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

where N_{MD} (τ) is the number of missed detections at the threshold τ , N_{FA} (τ) 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 and Przybocki, 1997] is used to visualize system performance. For the TRECVID18 ActEV challenges, we evaluated algorithm performance on the operating points; P_{miss} at $R_{\text{FA}} = 0.15$ and P_{miss} at $R_{\text{FA}} = 1$.

For the TRECVID19 ActEV evaluation, however, 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} and the time-based false alarm T_{fa} are 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.

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

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 2.

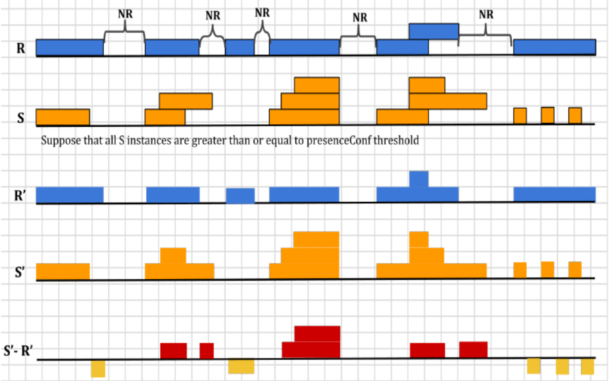
NR
NR
NR
NR
NR
R
S
Suppose that all S instances are greater than or equal to presenceConf threshold
R'
S'
S'- R'

Figure 2: Time-based false alarms T_{fa} calculation

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 the figure above); that is the duration of falsely detected system instances. This value is normalized by the non-reference duration of the video to provide the T_fa value in Equation above.

5.3 ActEV Results

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [Godil et al., 2019] in the online workshop notebook proceedings.

6 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.

6.1 VTT Data

The VTT data for 2019 consisted of two video sources.

- **Twitter Vine**: Since the inception of the VTT task, the testing data has comprised of Vine videos. Over 50k Twitter Vine videos have been collected automatically, and each video has a total duration of about 6 seconds. We selected 1044 Vine videos for this year's task.
- **Flickr**: Flickr video was collected under the Creative Commons License. Videos from this dataset have previously been used for the Instance Search Task at TRECVID. A set of 91 videos was collected, which was divided into 74 958 segments of about 10 seconds duration. A subset of 1010 segments was used for this year's VTT task.

A total of 2054 videos were selected and annotated manually by multiple assessors. An attempt was made to create a diverse dataset by removing any duplicates or similar videos as a preprocessing step. The videos were divided amongst 10 assessors, with each video being annotated by exactly 5 assessors. 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 previous experience 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.

6.2 System task

The VTT task is divided into two subtasks:

- Description Generation Subtask
- Matching and Ranking Subtask

Starting in 2019, the description generation subtask has been designated as a core/mandatory subtask, whereas the matching and ranking subtask is optional for all VTT task participants. Details of the two subtasks are as follows:

- **Description Generation** (Core): For each video, automatically generate a text description (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 the video from each of the 5 sets.

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

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

- Run type 'I' : Training using image captioning datasets only.
- Run type 'V' : Training using video captioning datasets only.
- Run type 'B' : Training using both image and video captioning datasets.

During the 2018 VTT task, it was observed that a number of teams only used image captioning datasets for training, since these datasets can be very large. Specifying the run types can help us compare the different systems based on their training data.

6.3 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 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.

Table 5: List of teams participating in each of the VTT subtasks. Description Generation was a core subtask in 2019.

	Matching & Ranking (11 Runs)	Description Generation (30 Runs)
IMFD_IMPRESEE	X	X
KSLAB	X	X
RUCMM	X	X
RUC_AIM3	X	X
EURECOM_MeMAD		X
FDU		X
INSIGHT_DCU		X
KU_ISPL		X
PICSOM		X
UTS_ISA		X

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) [4], 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, 10 teams participated in the VTT task this year. There were a total of 11 runs submitted by 4 teams for the matching and ranking subtask, and 30 runs submitted by 10 teams for the description generation subtask. A summary of participating teams is shown in Table 5.

6.4 VTT Results

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV19Pubs, 2019] in the online workshop notebook proceedings.

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

7 Summing up and moving on

This overview to TRECVID 2019 has provided basic information on the goals, data, evaluation mechanisms, 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 [TV19Pubs, 2019].

8 Authors' note

TRECVID would not have happened in 2019 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.

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

9 Acknowledgments

The ActEV NIST work was supported by the Intelligence Advanced Research Projects Activity (IARPA), agreement IARPA-16002, order R18-774-0017. The authors would like to thank Kitware, Inc. for annotating the dataset. The Video-to-Text work has been partially supported by Science Foundation Ireland (SFI) as a part of the Insight Centre at DCU (12/RC/2289). We would like to thank Tim Finin and Lushan Han of University of Maryland, Baltimore County for providing access to the semantic similarity metric.

References

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

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

[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., 2019] Godil, A., Lee, Y., and Fiscus, J. (2019). Trecvid 2019 actev evaluation plan. `https://actev.nist.gov/pub/ActEV_TRECVID_EvaluationPlan_081219.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.

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

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

[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.

[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.

[TV19Pubs, 2019] TV19Pubs (2019). http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.19.org.html.

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

[Yilmaz and Aslam, 2006] Yilmaz, E. and Aslam, J. A. (2006). Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management (CIKM)*, Arlington, VA, USA.

[Yilmaz et al., 2008] Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, New York, NY, USA. ACM.

A Ad-hoc query topics

611 Find shots of a drone flying
612 Find shots of a truck being driven in the daytime
613 Find shots of a door being opened by someone
614 Find shots of a woman riding or holding a bike outdoors
615 Find shots of a person smoking a cigarette outdoors
616 Find shots of a woman wearing a red dress outside in the daytime
617 Find shots of one or more picnic tables outdoors
618 Find shots of coral reef underwater
619 Find shots of one or more art pieces on a wall
620 Find shots of a person with a painted face or mask
621 Find shots of person in front of a graffiti painted on a wall
622 Find shots of a person in a tent
623 Find shots of a person wearing shorts outdoors
624 Find shots of a person in front of a curtain indoors
625 Find shots of a person wearing a backpack
626 Find shots of a race car driver racing a car
627 Find shots of a person holding a tool and cutting something
628 Find shots of a man and a woman holding hands
629 Find shots of a black man singing
630 Find shots of a man and a woman hugging each other
631 Find shots of a man and a woman dancing together indoors
632 Find shots of a person running in the woods
633 Find shots of a group of people walking on the beach
634 Find shots of a woman and a little boy both visible during daytime
635 Find shots of a bald man
636 Find shots of a man and a baby both visible
637 Find shots of a shirtless man standing up or walking outdoors
638 Find shots of one or more birds in a tree
639 Find shots for inside views of a small airplane flying
640 Find shots of a red hat or cap

B Instance search topics - 30 unique

9249 Find Max Holding a glass

9250 Find Ian Holding a glass

9251 Find Pat Holding a glass

9252 Find Denise Holding a glass

9253 Find Pat Sitting on a couch

9254 Find Denise Sitting on a couch

9255 Find Ian Holding phone

9256 Find Phil Holding phone

9257 Find Jane Holding phone

9258 Find Pat Drinking

9259 Find Ian Opening door and entering room / building

9260 Find Dot Opening door and entering room / building

9261 Find Max Shouting
9262 Find Phil Shouting
9263 Find Ian Eating
9264 Find Dot Eating
9265 Find Max Crying
9266 Find Jane Laughing
9267 Find Dot Opening door and leaving room / building
9268 Find Phil Going up or down stairs
9269 Find Jack Sitting on a couch
9270 Find Stacey Carrying a bag
9271 Find Bradley Carrying a bag
9272 Find Stacey Drinking
9273 Find Jack Drinking
9274 Find Jack Shouting
9275 Find Stacey Crying
9276 Find Bradley Laughing
9277 Find Jack Opening door and leaving room / building
9278 Find Stacey Going up or down stairs

Instance search topics - 20 common

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