Evaluating Multiple Video Understanding and Retrieval Tasks at TRECVID 2021

George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, et al.

Evaluating Multiple Video Understanding and Retrieval Tasks at TRECVID 2021

George Awad {gawad@nist.gov}
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

Yvette Graham {graham.yvette@gmail.com}
ADAPT Centre, Trinity College Dublin, Ireland

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

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

June 15, 2022

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 2021 represented a continuation of six tasks. In total, 39 teams from various research organizations worldwide signed up to join the evaluation campaign this year, where 22 teams (Table 1) completed one or more of the following six tasks, and 17 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 (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 Commons 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 task started using a subset of 1977 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 Video Summarization task also made use of the BBC Eastenders dataset, while the DSDI task worked on public natural disaster 6 h videos collected from a Nepal earthquake event in 2015 and combined with additional drones footage donated by the University of Vermont.

The Ad-hoc search, Instance Search, and Video Summarization results were judged by NIST human assessors, while the Video to Text task ground-truth was created 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. Full ground-truth was also built for the Disaster Scene Description and Indexing tasks and later on used to score teams' 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 tasks, data, evaluation framework, and measures used in the 2021 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 [TV21Pubs, 2021]. Finally, we would like to acknowledge that all work presented here has been cleared by RPO (Research Protection Office)[1]

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[2]. In the following

[1] under RPO number: #ITL-17-0025

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

Table 1: Participants and tasks

Task						Location	TeamID	Participants
IN	*VT*	*AV*	*AH*	*DS*	*VS*			
−−	−−	−−	−−	−−	*VS*	*Eur*	*Adapt_team*	ADAPT Research Centre
IN	−−	*AV*	−−	*DS*	−−	*Asia*	*BUPT_MCPRL*	Beijing University of Posts and Telecommunications
−−	−−	−−	−−	*DS*	−−	*Eur*	*VCL_CERTH*	Center for Research and Technology Hellas Information Technologies Institute
−−	−−	*AV*	**	**	−−	*NAm*	*INF*	CMU
−−	−−	−−	*AH*	−−	−−	*Asia*	*DMT_CUC_*01	Communiation University of China
−−	−−	−−	−−	−−	*VS*	*Eur*	*EURECOM*	EURECOM
−−	−−	−−	**	*DS*	−−	*NAm*	*FIU_UM*	Florida International University University of Miami
−−	−−	**	*AH*	−−	−−	*Asia*	*kindai_ogu_osaka*	Kindai University, Osaka Gakuin University, Osaka University
−−	−−	*AV*	*AH*	−−	−−	*Eur*	*ITI_CERTH*	Information Technologies Institute, Centre for Research and Technology Hellas
**	**	−−	*AH*	−−	**	*Asia*	*GodSpeed*	Kuaishou Tech
−−	*VT*	−−	−−	−−	−−	*Asia*	*kslab*	Nagaoka University of Technology
**	**	**	**	**	*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*	*RUCMM*	Renmin University of China
−−	*VT*	−−	*AH*	−−	−−	*Asia*	*RUC_AIM*3	Renmin University of China
−−	−−	−−	*AH*	−−	−−	*Asia*	*VIREO*	Singapore Management University and City University of Hong Kong
**	*VT*	*AV*	−−	**	**	*Asia*	*UEC*	The University of Electro-Communications, Tokyo
−−	−−	*AV*	−−	−−	−−	*Asia*	*TokyoTech*	Tokyo Institute of Technology
−−	*VT*	−−	−−	−−	**	*Eur*	*MMCUniAugsburg*	University of Augsburg
−−	**	−−	*AH*	−−	−−	*Asia*	*WasedaMeiseiSoftbank*	Waseda University, Meisei University, SoftBank Corporation
IN	−−	−−	−−	−−	−−	*Asia*	*WHU_NERCMS*	Wuhan University
−−	−−	*AV*	−−	−−	−−	*NAm*	*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*			
−−	**	−−	−−	−−	−−	*Eur*	*PicSOM*	Aalto University
−−	**	−−	−−	−−	**	*NAm*	*ARETE*	ARETE ASSOCIATES
**	−−	−−	**	−−	−−	*Asia*	*zju_om_center*	Bingjing Institute of Zhejiang University Zhejiang University Carnegie Mellon University
−−	−−	−−	**	−−	−−	*Asia*	*DMT_CUC*02	Communiation University of China
−−	−−	−−	−−	−−	**	*Asia*	95	CUC
−−	**	−−	−−	−−	−−	*SAm*	*IMFD_IMPRESEE*	DCC, Univesity of Chile Millennium Institute Foundational Research on Data (IMFD) IMPRESEE
**	−−	**	**	−−	**	*NAm*	*drylwlsn_visual*	drylwlsn_visual
−−	**	−−	**	−−	−−	*Eur*	*fhg.iais.nm.map*	Fraunhofer IAIS (NetMedia)
−−	−−	−−	**	−−	−−	*Asia*	*DVA*	IIST DECU ISRO
**	**	**	−−	−−	**	*Asia*	*chandra*	Individual
−−	−−	−−	−−	**	−−	*Eur*	*LINKS*	Links Foundation
−−	−−	**	**	−−	−−	*Aus*	*RMIT_GORSE*	RMIT
−−	−−	−−	**	−−	−−	*Asia*	*SejongRCV*	Sejong University
−−	**	−−	−−	−−	−−	*Asia*	*mitju*	Tianjin University, China
−−	**	−−	−−	−−	−−	*Asia*	*TMGO*	Tongji University, Nanjing University of Information Science and Technology
−−	**	−−	−−	−−	−−	*Asia*	*ok*	University of Science and Technology of China and Institute of Automation, Chinese Academy of Sciences
**	−−	−−	**	−−	**	*Eur*	*YildizTeam*	Yildiz Technical University

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

sections we will give an overview of the main datasets used this year across the different tasks.

2.1 BBC EastEnders Instance Search Dataset

The BBC in collaboration with 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 Search task and also adopted for the Video Summarization task to summarize the major events in 5 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 subset of short videos from V3C2 dataset (also drawn from the V3C video dataset) was 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 spatio-temporal localization 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 the United States. A variety of camera viewpoints and resolutions were included, with different levels of cluttered backgrounds, and activities are performed by many ordinary people. The spatial resolution of the cameras is either 1920x1080 or 1920x1072. The VIRAT dataset is 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 montages of randomly selected videos. In addition, we have built 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 that are between 3 seconds and 10 seconds long. The video sources are from Twitter Vine, Flickr, and V3C2. The dataset is being updated annually and in total, there are 10862 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 4387 video files in webm format and Creative Commons License. Those 4387 videos have been extracted from Flickr and the V3C2 dataset. This year 1977 V3C2 videos were used as a testing set, out of which the ground truth for 1677 has been

made public, whereas 300 videos will be used to compare the progress of systems over a period of 3 years.

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 304.8 m (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 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 aiming 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.

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 1000 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 offered 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): The 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. The system takes the formulated query as input and produces results without further human intervention.
- Relevance-Feedback: The system takes the official query as input and produces 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 for 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

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 through 2021 so that teams can measure their progress against three years (2019-2021). In general, the 20 fixed progress topics are divided equally into two sets of 10 topics. The first set was evaluated in 2020 to measure system progress for two years (2019-2020).

A Novelty run type was also allowed to be submitted within the main task. The goal of this run type is to encourage systems to submit novel and unique relevant shots not easily discovered by other runs. In other words, to find rare true positive shots. Finally, teams were allowed to submit an optional explainability parameter with each shot. This was formulated as a keyframe and bounding box to localize the region that supports the query evidence.

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 7475 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, 8 groups submitted a total of 77 runs with 39 main runs and 38 progress runs. Two groups submitted novelty runs. The 39 main runs consisted of 29 fully automatic, and 10 manually-assisted runs, while the progress runs consisted of 29 fully automatic and 9 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 20 % 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. Ten 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 not. 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 that were 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. New to this year's process is adding two extra judgment buttons (Yes near miss, and No near hit). The idea was to also mark clips that were submitted as relevant to the topic but may be considered hard for a system to detect (hence, it's near miss), and vice versa with other clips that are not relevant but are very close to being correct.

In all, 140 309 clips were judged while 167 036 clips fell into the unjudged part of the overall samples. Total hits across the 30 topics reached 32 161 with 14 938 hits at submission ranks from 1 to 100, 10794 hits at submission ranks 101 to 250 and 6429 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.

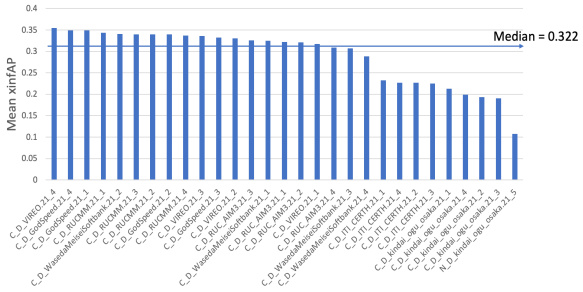

Figure 2: AVS: 29 Automatic Runs across 20 Main Queries

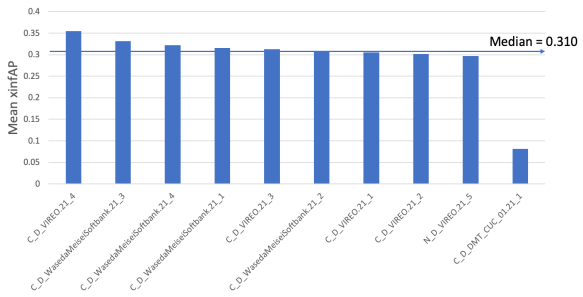

Figure 3: AVS: 10 Manually-Assisted Runs across 20 Main queries

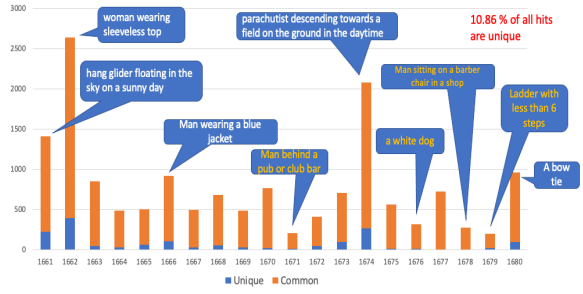

Figure 4: AVS: Unique vs overlapping results in main task

Ad-hoc Results

Figures 2 and 3 show the results of all the 29 fully automatic runs and 10 manually-assisted submissions respectively.

This is the third year for the ad-hoc task to work with the V3C1 dataset. As tested queries in the main task are different each year, we can not directly compare the performance the same way we do in the progress subtask. However, we can see that most automatic runs outperformed the top manually-assisted runs. In general, the 2021 median score is higher than the last two years for automatic runs. Similar results with respect to manually-assisted runs were observed as the maximum and median scores are higher than the last two years. The top-performing runs in both tasks come from the team VIREO and score almost similar (0.35 mean infAP)

We should also note here that all submissions were of type 'D', and no runs using category 'E' or 'F' were submitted. Also, while the evaluation supported relevance feedback run types, this year no submissions were received under this category.

To test if there were significant differences between the runs submitted, we applied a randomization test [Manly, 1997] on the top 10 runs for

[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 %
1592	111996	101233	90.39	6680	6.60	1019	15.25
1595	111624	98896	88.60	6277	6.35	808	12.87
1599	110962	102141	92.05	6953	6.81	1661	23.89
1600	111996	100915	90.11	15789	15.65	3498	22.15
1601	111996	104802	93.58	8214	7.84	1104	13.44
1603	111996	99481	88.83	7478	7.52	1638	21.90
1605	111996	99965	89.26	6781	6.78	5450	80.37
1607	111996	104662	93.45	8120	7.76	638	7.86
1608	111996	103015	91.98	7390	7.17	176	2.38
1609	111996	107051	95.58	8209	7.67	520	6.33
1661	39000	37604	96.42	3191	8.49	1408	44.12
1662	39000	36719	94.15	3749	10.21	2637	70.34
1663	39000	36267	92.99	2873	7.92	849	29.55
1664	39000	37507	96.17	2997	7.99	484	16.15
1665	39000	36570	93.77	3149	8.61	498	15.81
1666	39000	36611	93.87	2930	8.00	921	31.43
1667	39000	37532	96.24	2860	7.62	490	17.13
1668	39000	36618	93.89	3691	10.08	679	18.40
1669	39000	36515	93.63	2397	6.56	485	20.23
1670	39000	37607	96.43	3158	8.40	768	24.32
1671	39000	37400	95.90	2727	7.29	202	7.41
1672	39000	37640	96.51	3223	8.56	408	12.66
1673	39000	37665	96.58	3358	8.92	706	21.02
1674	39000	36684	94.06	3180	8.67	2081	65.44
1675	39000	36418	93.38	2113	5.80	564	26.69
1676	39000	36386	93.30	2179	5.99	318	14.59
1677	39000	37197	95.38	2554	6.87	723	28.31
1678	39000	37459	96.05	2735	7.30	272	9.95
1679	39000	37525	96.22	2913	7.76	194	6.66
1680	39000	36971	94.80	2441	6.60	962	39.41

manually-assisted and automatic run submissions using a significance threshold of $p<0.05$.

For automatic runs, the analysis showed that the only significant difference was between VIREO runs 4 and 3 (run 4 is better than run 3). All other runs ranked between rank 2 and 9 had no significant differences in their performance according to the test.

For manually-assisted runs, the analysis showed that there was no significant difference between VIREO run 4 and WasedaMeiseiSoftbank run 3, VIREO run 4 is better than all other VIREO runs, WasedaMeiseiSoftbank run 3 is better than all other WasedaMeiseiSoftbank runs, and all runs are better than run 1 of the team DMT_CUC_01.

Figure 4 shows for each topic the number of relevant and unique shots submitted by all teams combined (blue color). On the other hand, the orange bars show the total non-unique true shots submitted by at least 2 or more teams. The four topics: 1661, 1662, 1666, and 1674 achieved the most unique hits overall while also reporting a high number of hits overall, while the four topics: 1675, 1677, 1678, and 1679 reported the lowest unique hits. In general, topics that reported a high number of hits consisted of high number of unique as well as non-unique hits, while topics that reported low number of hits mainly only consisted of non-unique hits representing the difficulty of the query (with the exception of query

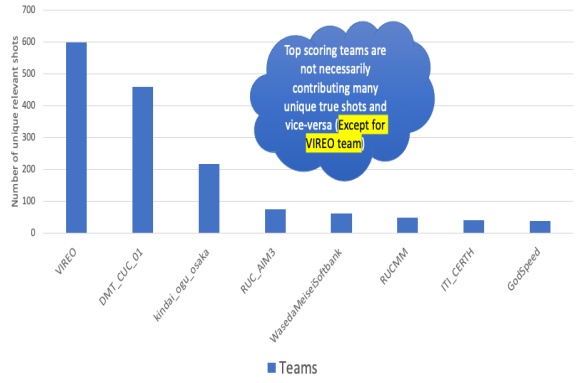

Figure 5: AVS: 1534 Unique shots contributed by teams in main task

1677).

From observing these results, it can be shown that performance drops when topics start asking for information that needs more complicated analysis or feature combinations such as Relational (Man behind a bar, wearing cap backwards), exact count (ladder with less than 6 steps), or conditions (white dog, person looking at themselves in mirror, adult wearing backpack, walking on side walk). On the other hand, performance is higher for objects (bow tie, hang glider), or common states (woman in sleeveless top).

We should also note here that high/low hits per topic don't necessarily mean high/low performance in InfAP as a good run must detect and rank results high as well.

Figure 5 shows the number of unique clips found by the different participating teams. From this figure and the overall scores in figure 2 and 3 it can be shown that there is no clear relation between teams who found the most unique shots and their total performance. One exception this year is team VIREO who managed to perform best overall and also report most unique relevant shots.

Figures 6 and 7 show the performance of the top 10 teams across the 20 main queries. Note that each series in this plot represents a rank (from 1 to 10) of the scores, but all scores at a given rank do not necessarily belong to a specific team. A team's scores can rank differently across the 20 queries. Some samples of top and bottom performing queries are highlighted with the query text.

From the figures, we can see a high similarity between automatic and manually-assisted systems in terms of query performance relative to each other. Harder queries are those that included non- traditional combinations of concepts (e.g. ladder with less than 6 steps), relational concepts (e.g. wearing a cap backwards), and fine-grained actions (pointing with finger).

In general, for automatic systems and for topics not performing well, usually all top 10 runs are condensed together with low spread between their scores. This is however not consistent with manually-assisted systems where there is a bigger spread across systems performance in most queries which may reflect how different systems reformulated the NIST queries based on their query interface.

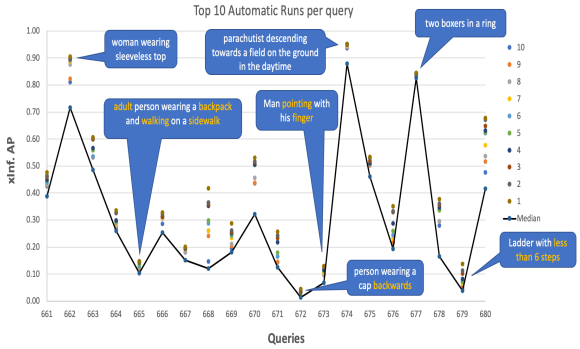

Figure 6: AVS: Top 10 runs (xinfAP) per query (fully automatic)

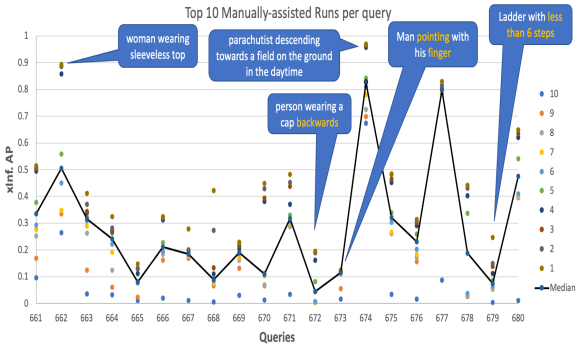

Figure 7: AVS: Top 10 runs (xinfAP) per query (manually assisted)

The novelty run type encourages submitting unique (hard to find) relevant shots. Systems were asked to label their runs as either of novelty type or common type runs. A new novelty metric was designed to score runs based on how good they are in

detecting unique relevant shots. A weight was given to each topic and shot pairs such as follows:

$$TopicX\_ShotY_{weight}(x) = 1 - \frac{N}{M}$$

where N is the number of times Shot Y was retrieved for topic X by any run submission, and M is the number of total runs submitted by all teams. For instance, a unique relevant shot weight will be close to 1.0 while a shot submitted by all runs will be assigned a weight of 0.

For Run R and for all topics, we calculate the summation S of all unique shot weights only and the final novelty metric score is the mean score across all evaluated 20 topics. Figure 8 shows the novelty metric scores. The red bars indicate the submitted novelty runs.

We should note here that in running this experiment, for a team that submitted a novelty run, we removed all its other common runs submitted. The reason for doing this was the fact that usually for a given team there would be many overlapping shots within all its submitted runs. For other teams who did not submit novelty runs, we chose the best (top scoring) run for each team for comparison purposes. As shown in the figure, one of the two novelty runs (by VIREO team) submitted scored best based on our metric while the other run achieved average performance. More runs are needed to conduct a better comparison within novelty systems.

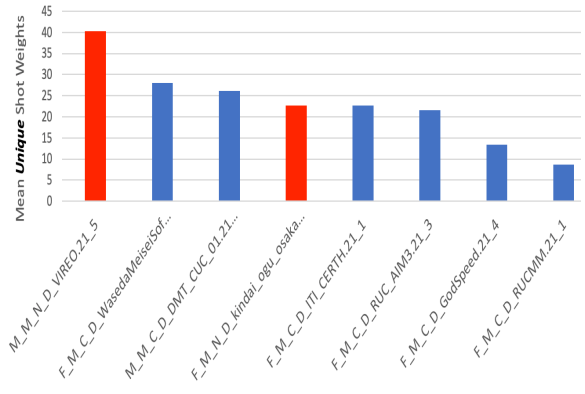

Figure 8: AVS: Novelty runs vs best common run from each team

Among the submission requirements, we asked teams to submit the processing time that was consumed to return the result sets for each query. Figures 9 and 10 plot the reported processing times vs the InfAP scores among all run queries for automatic and manually-assisted runs respectively.

It can be seen that spending more time did not necessarily help in many cases and few queries achieved high scores in less time. There is more work to be done to make systems efficient and effective at the same time. In general, most automatic systems reported processing time below 10 s. While most manually-assisted systems reported processing times above 10 s.

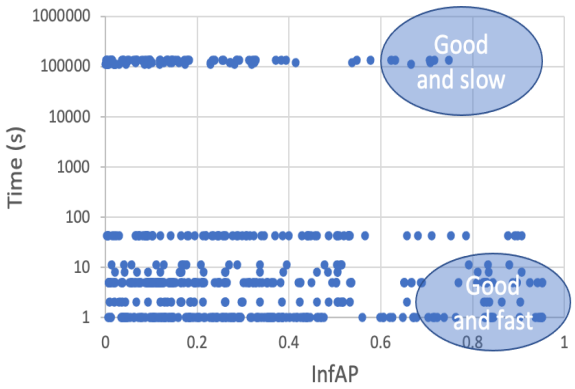

Figure 9: AVS: Processing time vs Scores (fully automatic)

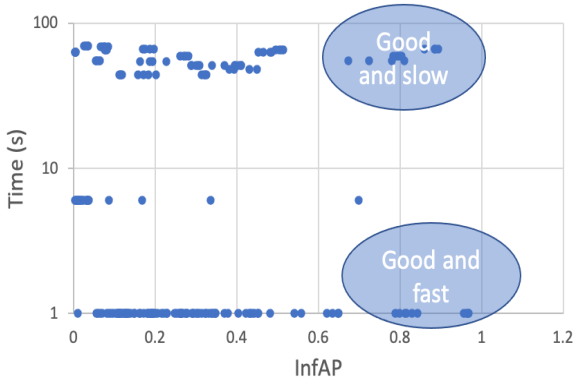

Figure 10: AVS: Processing time vs Scores (Manually assisted)

The progress task results are shown in figures 11 and 12 for automatic and manually-assisted systems respectively. In total, 14 teams participated in this progress task since 2019. Comparing the best run between 2019 to 2021 for each team, we can see that all teams achieved better performance in 2021 (gray

bar) for those who participated in the three years. There are some teams who participated in just one or two of the three years. The majority of teams who participated in two years had better performance in the second year. The results indicate that the task achieved its purpose to measure system progress and encouraged systems to develop better systems year over year.

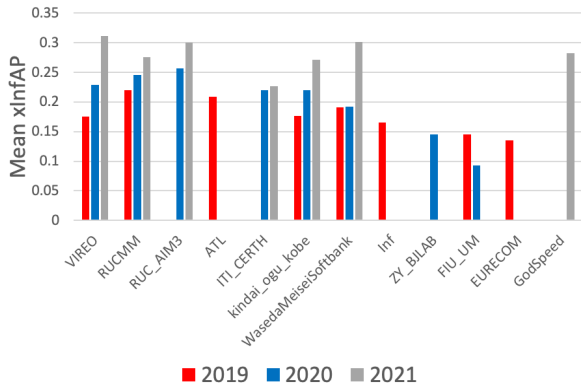

Figure 11: AVS: Max performance per team on 10 progress queries (automatic systems)

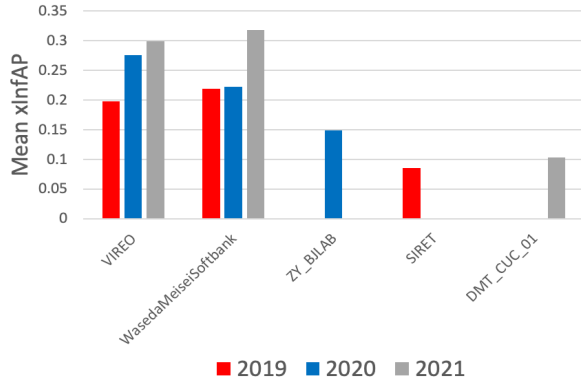

Figure 12: AVS: Max performance per team on 10 progress queries (manually-assisted systems)

To analyze in general which topics were the easiest and most difficult we sorted topics by the number of runs that scored xInfAP >= 0.5 for any given topic and assumed that those were the easiest topics, while topics with xInfAP < 0.5 were assumed hard topics. From this analysis, it can be concluded that the top 5 hard topics were: "Person wearing a cap backwards", "Ladder with less than 6 steps", "Man pointing with his finger", "Adult person wearing a backpack and walking on a sidewalk", and "Person looking at themselves in a mirror". On the other hand, the top 5 easiest topics were: "Two boxers in a ring", "Parachutist descending towards a field on the ground in the daytime", "Woman wearing sleeveless top", "A bow tie", and "Person with a tattoo on their arm". Similar to our observations about the number of unique hits per query and how it is related to query performance, the hard queries are those that combine different conditions, relationships, or states compared to simple visual concepts in easier queries (e.g. tattoo, bow tie, parachutist, boxers).

Sample results of frequently submitted false positive shots are demonstrated[4] in Figure 13.

Figure 13: AVS: Samples of frequent false positive results

Ad-hoc Observations and Conclusions

Compared to the semantic indexing task that was conducted to detect single concepts (e.g., airplane, animal, bridge) from 2010 to 2015 it can be seen from this year's results that the ad-hoc task is still very hard and systems still have a lot of room to research methods that can deal with unpredictable queries composed of one or more concepts including their interactions, relationships and conditions.

In 2018 we concluded 1-cycle of three years of Ad-hoc task using the Internet Archive (IACC.3) dataset [Awad et al., 2016a]. In 2019, a new dataset, Vimeo Creative Commons Collection (V3C1), was introduced and adopted for testing at least for a 3 year cycle (2019-2021). NIST Developed a set of 90 queries to be used for 3 years including a progress subtask.

[4]All figures are in the public domain and permissible under RPO #ITL-17-0025

To summarize major observations in 2021 we can see that overall, team participation and task completion rates are stable. Most submitted runs were of training type "D", and no runs of type "E" or "F" were submitted. Two novelty run types were submitted. Overall, 112 systems (81 automatic and 31 manually-assisted) were submitted between 2019 and 2021 in the progress task. The majority of 2021 systems performed higher than their 2019 and 2020 versions in the progress subtask.

Fully automatic and manually-assisted are almost similar in terms of query performance relative to each other. Top scoring teams did not necessarily report unique relevant shots (thus they are good in ranking relevant shots) with the exception of VIREO team. About 11% of all hits are unique, while the rest are common hits across the runs. Few systems are efficient and effective at retrieving fast (less than 10 sec) and accurate results. There is a much needed research into recognizing and detecting rare, unusual, and fine-grained queries especially those that are composed of underrepresented visual concepts in mainstream training datasets.

As a general high-level system overview, we can see that there are two main competing approaches among participating teams: "concept-based banks" and "visual-textual embedding spaces". There is a clear trend towards the embedding approaches due to their performance. Concept-based banks are often used as a complement to embedding approaches.

For detailed information about the approaches and results for individual teams, we refer the reader to the reports [TV21Pubs, 2021] 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 the work from previous years in the concept detection task [Awad et al., 2016b] the instance search task seeks to address some of these needs. For six years (2010-2015) the instance search task 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 was introduced. The task spanned 3 years till 2018 and since 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 could 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:

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 %
9280	33479	29124	86.99	260	2498	8.58	510	20.42
9281	34067	28842	84.66	520	5312	18.42	235	4.42
9283	34999	29458	84.17	300	3233	10.97	194	6.00
9286	34999	28185	80.53	520	5201	18.45	68	1.31
9288	34107	23063	67.62	800	5085	22.05	87	1.71
9289	34259	21179	61.82	440	3864	18.24	433	11.21
9291	34998	27931	79.81	520	5429	19.44	131	2.41
9293	34076	24465	71.80	300	3193	13.05	80	2.51
9296	34998	22481	64.24	540	4864	21.64	1024	21.05
9297	34039	20407	59.95	640	5412	26.52	129	2.38
9319	15000	13167	87.78	520	1262	9.58	435	34.47
9320	15000	12979	86.53	520	1212	9.34	601	49.59
9321	15000	13064	87.09	800	2178	16.67	644	29.57
9322	15000	13534	90.23	520	1457	10.77	412	28.28
9323	15000	13467	89.78	520	1420	10.54	463	32.61
9324	15000	13145	87.63	620	1612	12.26	474	29.40
9325	15000	13741	91.61	520	2050	14.92	144	7.02
9326	15000	13462	89.75	520	1603	11.91	232	14.47
9327	15000	13656	91.04	440	1595	11.68	158	9.91
9328	15000	13620	90.80	340	1438	10.56	83	5.77
9329	15000	13589	90.59	620	2537	18.67	329	12.97
9330	15000	13407	89.38	260	903	6.74	229	25.36
9331	15000	13395	89.30	520	1877	14.01	254	13.53
9332	15000	13474	89.83	520	1826	13.55	311	17.03
9333	15000	10432	69.55	520	1495	14.33	147	9.83
9334	15000	8454	56.36	560	1718	20.32	93	5.41
9335	15000	6964	46.43	580	1700	24.41	36	2.12
9336	15000	8942	59.61	520	1322	14.78	83	6.28
9337	15000	9147	60.98	520	1731	18.92	251	14.50
9338	15000	9121	60.81	380	1278	14.01	238	18.62

  - 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 (external sources or the NIST provided training data). The following training options were provided for evaluation:

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

could submit for evaluation:

1. Fully automatic (F) runs: The system takes official query as input and produces results without any human intervention.

2. Interactive humans in the loop (I) runs: The system takes official query as input and produces 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.

Query Topics

NIST reviewed 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 could 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 6 individuals and 8 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 type of examples used for training). In total, 3 groups submitted 22 runs including 16 automatic and 6 interactive runs. From the 22 runs, 7 runs belonged to the progress subtask, while 15 belonged to the main 2021 task. In addition to the 7 progress runs in 2021, a set of total 28 progress runs were submitted by teams in 2019 and 2020. All 35 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 stratum comprised of 20 rank levels (1-20, 21-40, 41-60, etc) up to rank 520. Finally, all duplicates in each stratum were removed.

For a given topic[5], 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 a few instances, more pools below rank 520 were generated for assessors who finished early their initial pools and had more time to continue working on the task.

In general, submissions were pooled and judged down to at least rank 260, resulting in 76 305 judged shots including 8 508 total relevant shots (11.1%). 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 run-time and location accuracy were also of interest here, of these two, only run-time was reported.

For detailed information about the approaches and results for individual teams' performance and runs, the reader should see the various site reports [TV21Pubs, 2021] in the online workshop notebook proceedings.

Results

Figures 14 and 15 show the sorted scores of runs for automatic and interactive systems. WHU_NERCMS achieve the best results for both automatic and interactive runs.

Figure 16 shows the progress topics scores for 2019, 2020, and 2021. From this chart we can see that all teams who submitted progress runs in multiple

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

years saw an improvement in performance year-on-year, with WHU_NERCMS achieving the largest improvement in performance.

Figure 17 shows the distribution of automatic run scores (average precision) by topic as a box plot. The topics are sorted by the maximum score with the best performing topic on the left. Median scores vary from 0.674 down to 0.057. The main factor affecting topic difficulty this year is the target action.

Figures 18 and 19 show the distribution of automatic run scores by character and action. These are sorted by maximum score with the best performing character and action on the left.

Figures 20 and 21 show the easiest and hardest topics, calculated by the number of runs with average precision scores above 0.3 and below 0.3 respectively. These figures show that holding a phone and sitting on couch were the easiest actions to find, while carrying bag and open door & enter were the hardest actions to find.

Figure 22 documents the raw scores of the top 10 automatic runs and the results of a partial randomization test [Manly, 1997] and sheds some light on which differences in ranking are likely to be statistically significant. One angled bracket indicates $p < 0.05$. There are little significant differences between the top runs this year.

The relationship between effectiveness (mean average precision) and elapsed processing time is depicted in Figure 23 for the automatic runs with elapsed times less than or equal to 300s. Of those reported times below 300s, we can see that in general the most accurate systems require longer processing times.

Figure 24 shows the relationship between the two categories of runs (images only for training OR video and images) and the effectiveness of the runs. These show that far more runs use image only examples rather than image and video examples, but the team that used both example types achieved slightly improved performance from using image + video examples.

Figure 25 shows the effect of the data source used for training, with participants being able to use an external data source instead of or in addition to the NIST provided training data. The use of external data only provides by far better performance in this years task than the use of external data in addition to NIST provided data.

Observations

This was the third year the task used the new query type of person+action, and the sixth year using the EastEnders dataset. There was a slight decrease in the number of participants who signed up for the task this year with 11 teams registering. There was also a decrease in the number of finishers this year, with 3 teams finishing the task compared to 5 teams who completed last year's task.

Once again for this year of the task, participating teams could use external data instead of or in addition to NIST provided data. Results this year showed that the use of external data only consistently provides better results. Teams could also again make use of video examples or image-only examples. The majority of teams used image-only examples rather than image and video examples, but the team that used both example types achieved slightly improved performance.

BUPT_MCPRL first adopted RetinaFace[Deng et al., 2019b] to detect faces in shot keyframes, then extract face landmarks using PFLD[Guo et al., 2019] for face alignment. DeepSORT[Wojke et al., 2017] was then used to track persons. FaceNet[Schroff et al., 2015] was then used to extract facial feature representation for cosine similarity matching. A threshold was set for cosine distance to filter shots. A query selection strategy was employed to remove low relevancy queries. For video-level action recognition a SlowFast[Feichtenhofer et al., 2019] model pre-trained on Kinetics-400[Kay et al., 2017] dataset to roughly judge whether the action occurs in INS video shots. For clip-level action recognition, the SlowFast model was trained on the AVA-Kinetics[Li et al., 2020] dataset.

They then used Cascade R-CNN[Cai and Vasconcelos, 2018] pre-trained on COCO dataset to locate persons. Using this pre-trained model the action scores of each detected person in keyframes were obtained. Frame-level HOI detection was used to recognize human-object interactions in a single frame. For this, the iCGPN model was trained on the HICO-DET[Chao et al., 2018] dataset. For Action feature retrieval, the SlowFast model pre-trained on AVA-Kinetics was adopted to extract action features of keyframes. Cosine similarities were calculated among them and maximal similarity set as the shot keyframes similarity. For emotion-related action retrieval, a lightweight CNN pre-trained on CK+[Lucey et al., 2010] and

FERPLUS[Barsoum et al., 2016] was applied to recognize emotion-related actions such as shouting, laughing, and crying. To obtain better performance, late-fusion was applied in the post-processing stage. Finally, they took the maximal score of the key-frames in each shot as the shot scores to generate the shot ranking lists for all person-action pairs.

PKU_WICT proposed a two-stage approach consisting of similarity computing and result re-ranking. They used four aspects for action specific recognition: frame-level action recognition, video-level action recognition (trained using Kinetics-700[Smaira et al., 2020] and Moments in Time[Monfort et al., 2019] datasets), object detection (pre-trained on MS-COCO[Lin et al., 2014], Visual Genome[Krishna et al., 2017], and Object365[Shao et al., 2019]) and facial expression recognition (Using VGGNet[Simonyan and Zisserman, 2014] trained on CK+[Lucey et al., 2010] and FER2013[kag, 2013]).

Finally, they computed the fusion value of all prediction scores of a shot as the final prediction score *ActScore*. For person specific recognition they first detected faces in query examples and filtered out bad faces of low detection confidence and complimented them with good faces of high detection confidence. Next, face features from queries and shots were extracted based on deep convolutional neural networks and calculated the similarity. Top N query expansion strategy was conducted for further improving the retrieval results. They then used two fusion strategies to fuse scores from action specific and person specific recognition.

For person retrieval, WHU_NERCMS used a face detection model RetinaFace[Deng et al., 2019b] and a face recognition model ArcFace[Deng et al., 2019a] to compute person retrieval scores. For action retrieval, they used two types of methods, a frame-level method to detect interactions between people and objects in the frame, and a video-level method to detect spatio-temporal information in the video. The final ranking list was obtained from result fusion of the respective ranking lists person and action, using weight fusion and filter fusion methods. Additionally, given that some actions have temporal continuity and can last more than one shot, they proposed Score Temporal Expansion (STE)[Yang et al., 2021] for re-ranking, which adjusts the fusion score for shots by fusing the scores of neighbouring shots.

Conclusions

This was the third year of the updated Instance Search task in which queries comprised of a specific person doing a specific action. The action recognition part of the task made this task a more challenging problem than before the updated task, with maximum and average results still below those of previous years for the specific person in a specific location queries. Results did however show a big improvement over the previous year of the task, which was the second year of the updated task. Those results had also shown an improvement over the first year of the task, given continuous improvements in performance over the three years.

There were a total of 3 finishers out of 11 participating teams in this year's task, a decrease from the previous year. All 3 finishers submitted notebook papers. All 3 teams also submitted runs for the progress queries, 2 of which can be directly compared against their progress runs from the previous two years and one which could be directly compared to their performance in the preceding year, all of which showed an improvement in performance year-on-year.

This was the final year of the INS person-action retrieval task. A new Deep Video Understanding (DVU) task will be beginning in its place which will make use of a new full movie dataset.

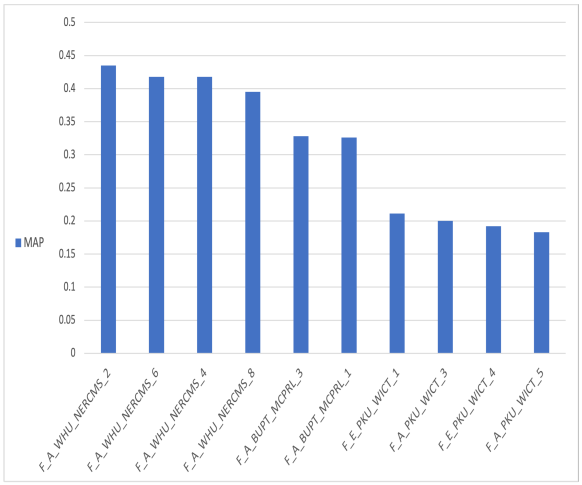

Figure 14: INS: Mean average precision scores for automatic systems

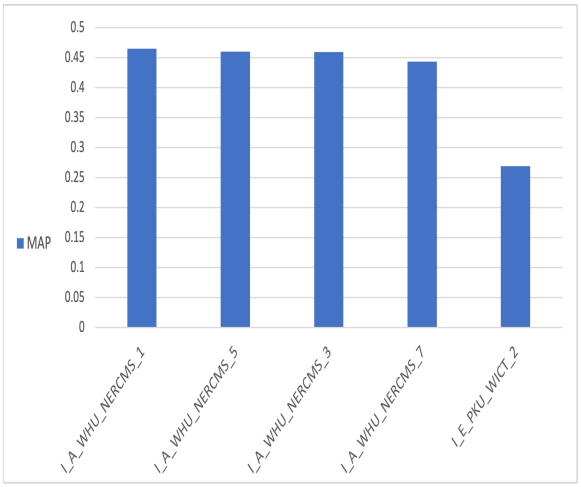

Figure 15: INS: Mean average precision scores for interactive systems

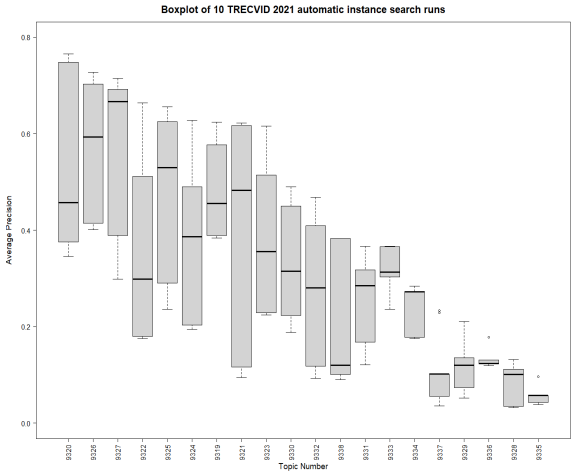

Figure 17: INS: Boxplot of average precision by topic for automatic runs.

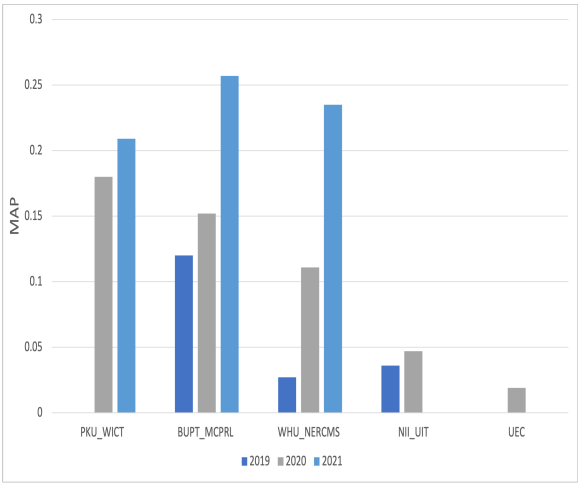

Figure 16: INS: Mean average precision scores comparing results on 2019, 2020, and 2021 progress topics

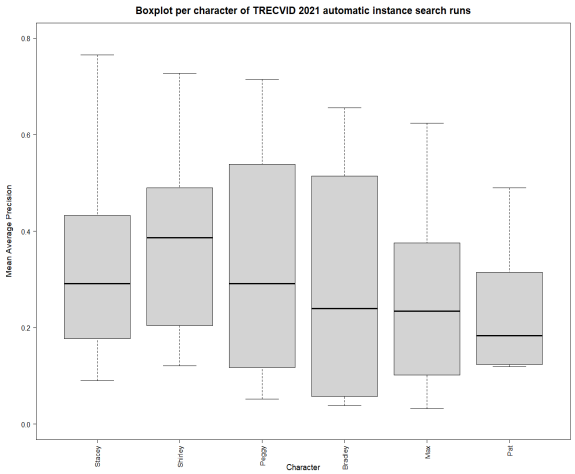

Figure 18: INS: Boxplot of average precision by character for automatic runs.

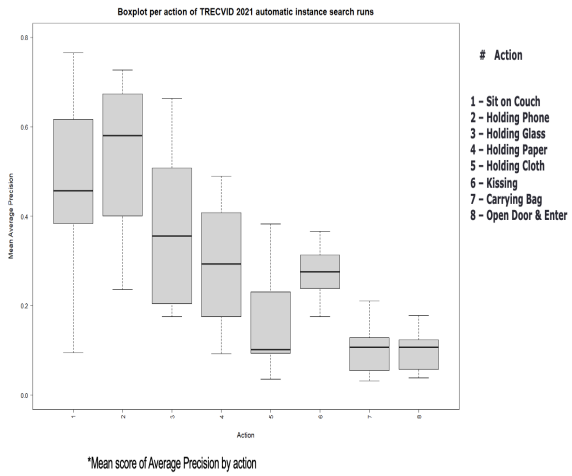

Figure 19: INS: Boxplot of average precision by action for automatic runs.

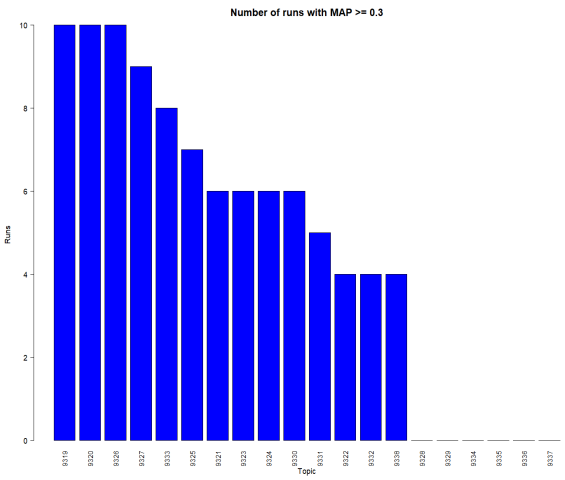

Figure 20: INS: Easiest topics for automatic systems

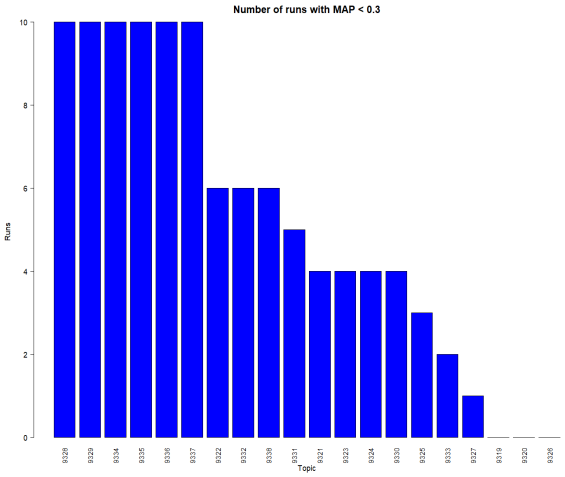

Figure 21: INS: Hardest topics for automatic systems

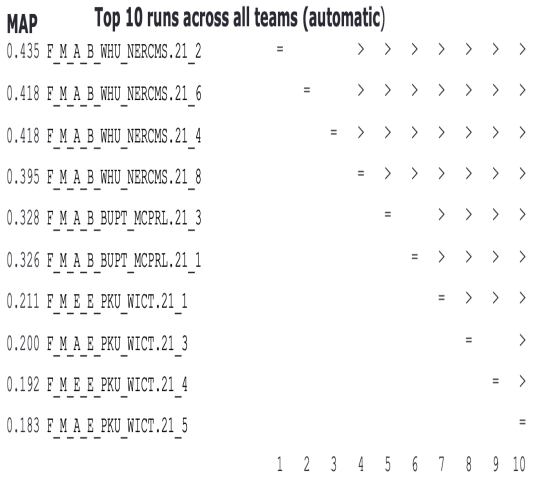

Figure 22: INS: Randomization test results for top automatic runs. "E":runs used video examples. "A":runs used image examples only.

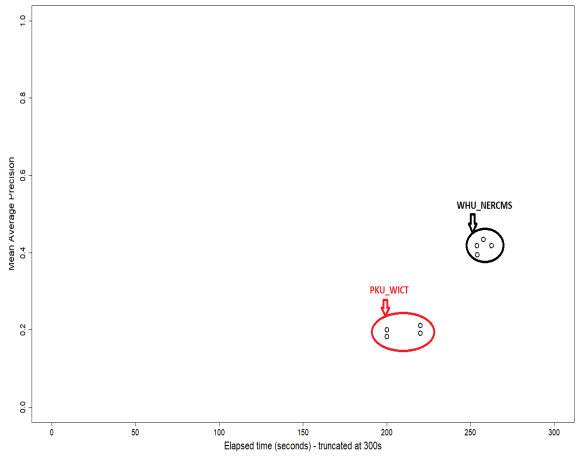

Figure 23: INS: Mean average precision versus time for fastest runs

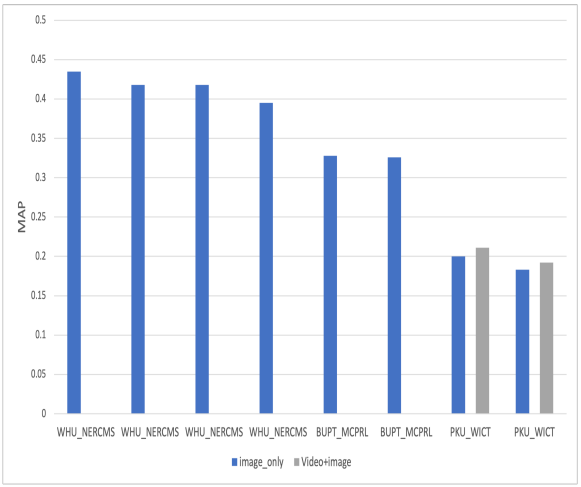

Figure 24: INS: Effect of image vs video data type

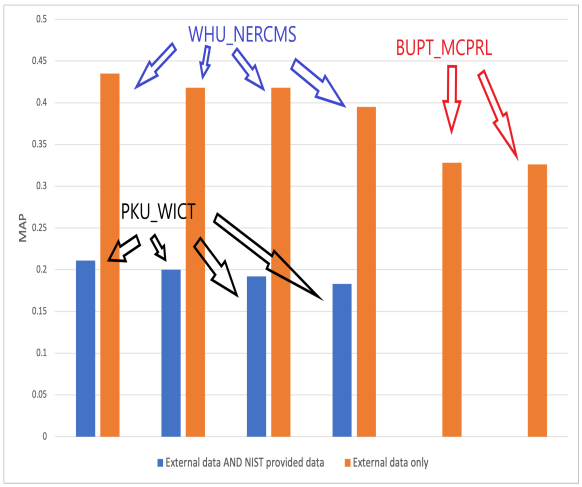

Figure 25: INS: Effect of data source used

3.3 Disaster Scene Description and Indexing

Computer vision capabilities have rapidly been advancing and are expected to become an important component for 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 New Jersey Office of Homeland Security and MIT Lincoln Laboratory 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

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 test dataset has 5 coarse categories, each divided into 4-9 more specific labels.

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.

DSDI was introduced in TRECVID in 2020, and this is the second iteration of the task.

Datasets

Training Dataset The training dataset is based on the LADI dataset hosted as part of the AWS Public Dataset program along with the DSDI video test dataset used in 2020.

The LADI dataset consists of 20 000+ human annotated images and about 500 000 machine annotated images. The images are from locations with FEMA major disaster declaration for a hurricane, earthquake, or flooding[6]. The lower altitude criterion distinguishes the LADI dataset from satellite datasets to support the development of computer vision capabilities with small drones operating at low altitudes. A minimum image size (4MB) was selected to maximize the efficiency of the crowd source workers, since lower resolution images are harder to annotate.

The ground truth for the DSDI test set for 2020 was made public after completion of the task and is available to be used as training dataset. It consisted of about 5 hours of video that were segmented into small video clips (or shots) of a maximum duration 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.

Test Dataset The test dataset for the task this year consists of about 6.7 hours of video. The test dataset was segmented into small video clips (or shots) of a maximum duration of 20.85 seconds. The videos are from earthquake, flooding, fire, and erosion affected areas. They have been collected from both domestic and international sources. There are a total of 2801 shots with a median length of 8.34 seconds. We also included some location metadata with the videos, which included the start and end coordinates, and the path of the aircraft.

Categories The categories used for the test dataset are the same as those used for the LADI training dataset [Liu et al., 2019]. 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.

As can be expected from a real-world dataset, features appear with varied frequency within the videos. Some features such as grass, trees, buildings, roads, etc. appear much more frequently than others. The lava feature does not appear in any of the shots in the test dataset. Figure 27 shows the number of shots that contain each feature.

Annotation The video annotation was done using full-time annotators instead of crowd sourcing. It is

[6]https://www.fema.gov/disaster/declarations

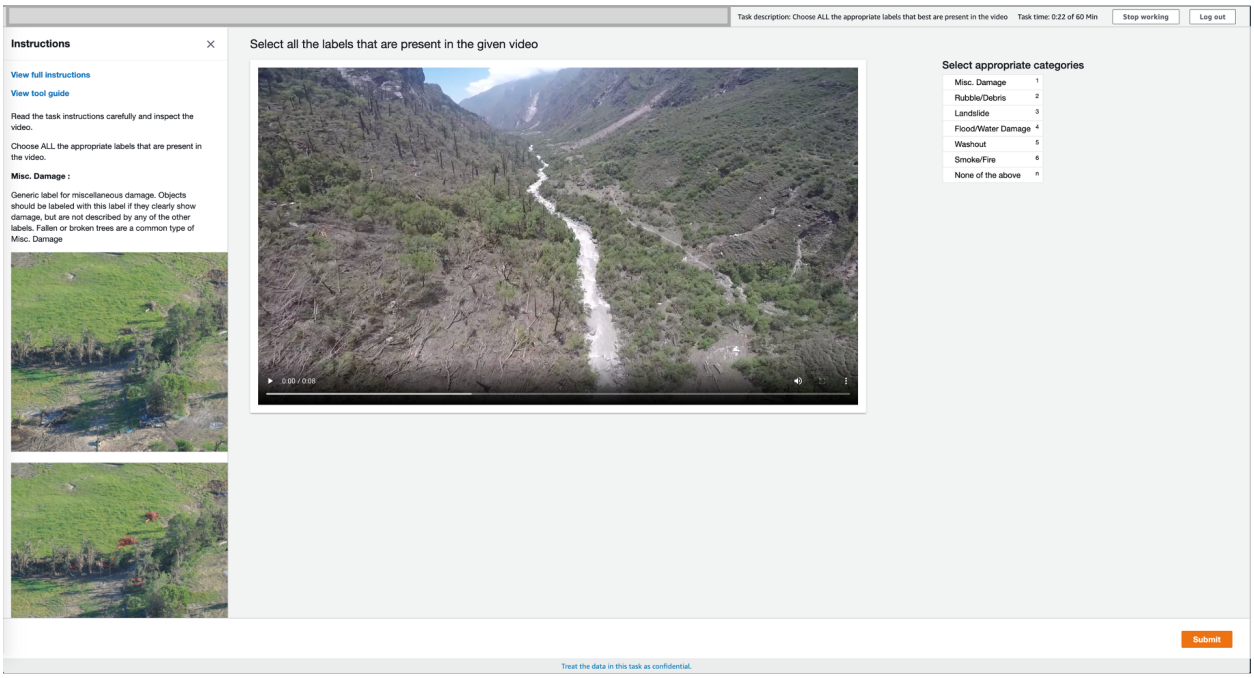

Figure 26: DSDI: Screenshot of a video being annotated for the Damage category. The annotator watches the video and marks all the labels that are visible in the video.

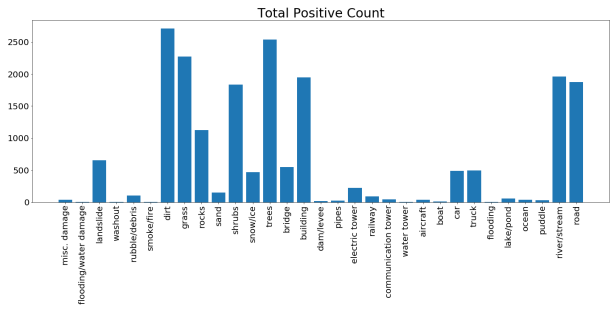

Figure 27: DSDI: Number of shots containing each feature (excluding Lava, which does not appear in any shots).

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 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 confusion during the labeling of the dataset.

Two full-time annotators labeled the testing dataset. Both annotators had worked on the task previously in 2020 and were familiar with it. The Amazon Augmented AI (Amazon A2I) tool was used during the process. The annotators worked independently on each category. Figure 26 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 was 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 is mean average precision (MAP). The average precision is calculated for each feature, and the mean average precision is 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.

Results

This year 6 teams signed up to join the task and finally 3 teams submitted runs. In total, we received 12 runs including 4 LADI+Others (O) runs and 8 LADI-based (L) runs. For detailed information about the approaches and results for individual teams' performances and runs, we refer the reader to the site reports [TV21Pubs, 2021] in the online workshop notebook proceedings. We present the overall results in this section.

None of the videos in the testing dataset had any occurrences of the lava feature, and so that feature was removed from all result calculations.

Figures 28 and 29 show the box and whisker plot of average precision scores for each feature for systems with run types L and O respectively. Systems tend to perform well on features that are commonly seen in training data, such as grass, trees, buildings, etc.

Figures 30 and 31 show the average precision values organized by categories for run types L and O respectively. These charts show how the systems perform on features within each category.

Finally, Figures 32 and 33 show the mean average precision score for each run with training type L and O respectively.

We also reported the true positives, true negatives, false positives, and false negatives for each run. The F-measure using these values is shown in Figure 34.

Conclusion and Future Work

This was the second iteration of the DSDI task. While the participation in the task decreased from last year, teams performed reasonably well. Some known issues with the training data are:

1. The LADI dataset labels can be noisy due to crowd-sourced annotation.

2. There is a class imbalance as certain labels are far more prevalent than others.

3. The datasets are mostly limited to a certain types of disasters. It is not simple to have representation for all disaster labels since data acquisition requires multiple sources.

The DSDI test dataset was labeled by dedicated annotators, which resulted in cleaner annotation. As these become part of the training dataset in the coming years, we will get high quality videos to help systems train.

Some possible improvements in the LADI dataset, subject to available funding, are as follows:

1. Labels for additional disaster events will be added.

2. Bounding boxes or segmentation will be provided for a subset of classes to enable object detection and localization.

3. Improved documentation and tutorials will be made available.

Teams participating in the DSDI task have worked hard to clean and refine annotations for LADI. We will work with the teams to make the annotations and code available for public use.

The task will continue next year with a similar amount of testing video data. However, we will attempt to focus on other disaster categories.

3.4 Video to Text

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, to mention a few. In recent years there have been major advances in computer vision techniques that enabled researchers to start

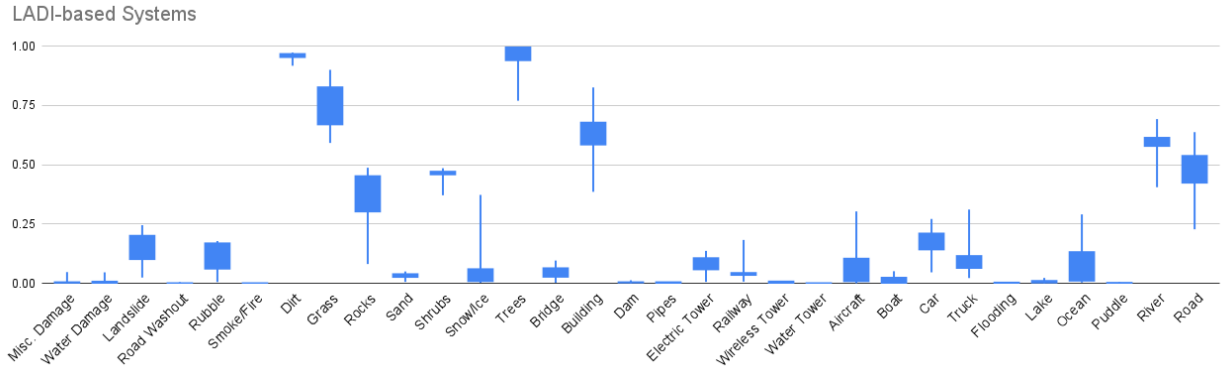

Figure 28: DSDI: Box and whisker plot of average precision values for each feature for systems with training type L.

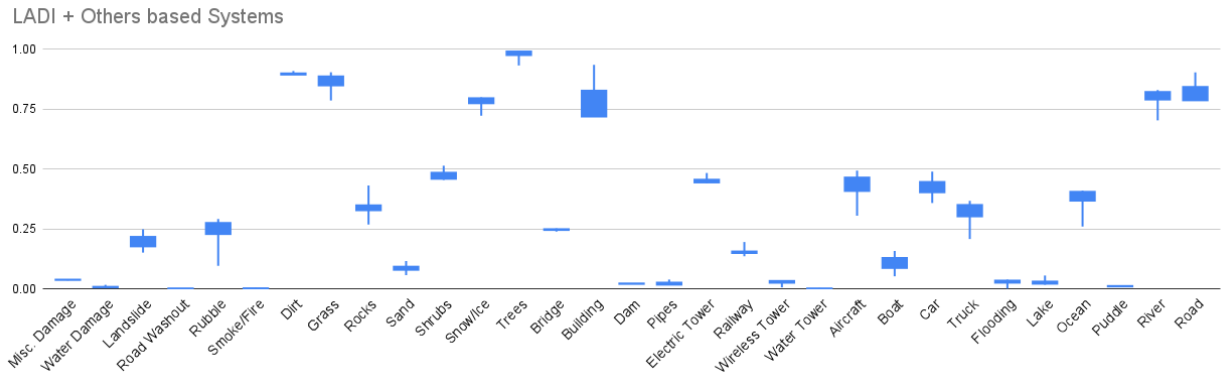

Figure 29: DSDI: Box and whisker plot of average precision values for each feature for systems with training type O.

	Fill-in-the-Blanks (3 Runs)	Description Generation (15 Runs)
KSLAB		X
MMCUniAugsburg		X
RUC_AIM3	X	X
RUCMM	X	X
UEC		X

Table 6: VTT: List of teams participating in each of the subtasks.

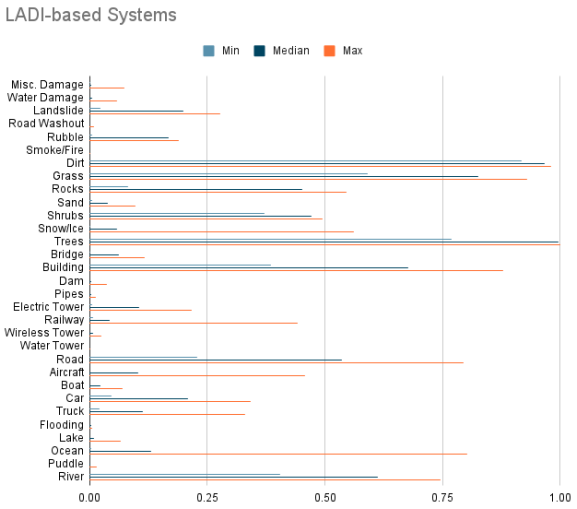

Figure 30: DSDI: Average precision values organized by categories for systems with training type L.

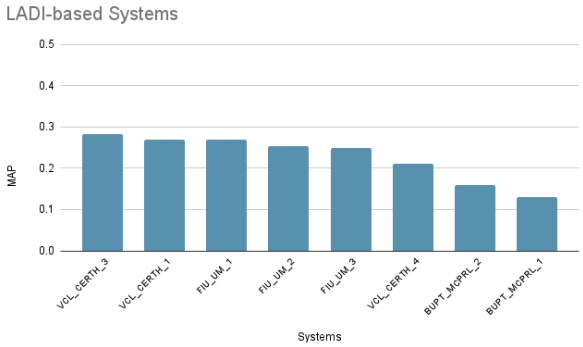

Figure 32: DSDI: Mean average precision score for each run with training type L.

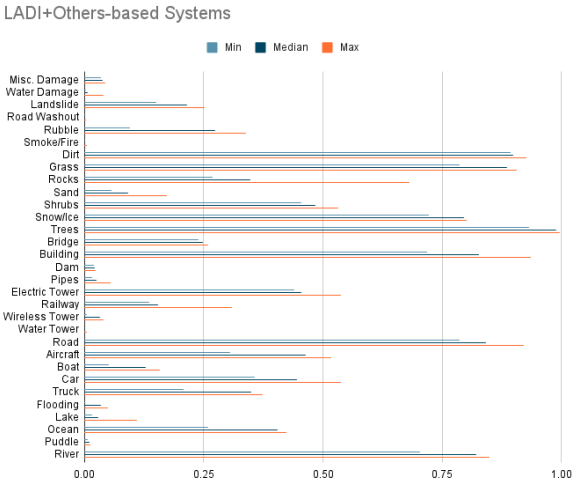

Figure 31: DSDI: Average precision values organized by categories for systems with training type O.

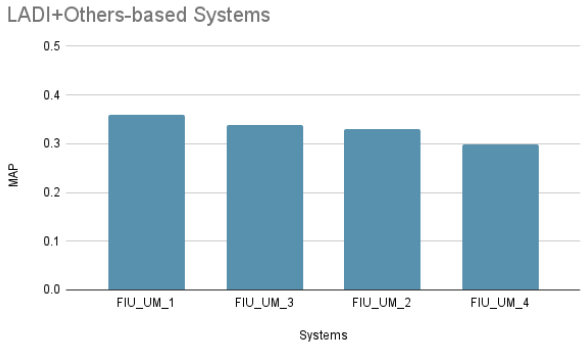

Figure 33: DSDI: Mean average precision score for each run with training type O.

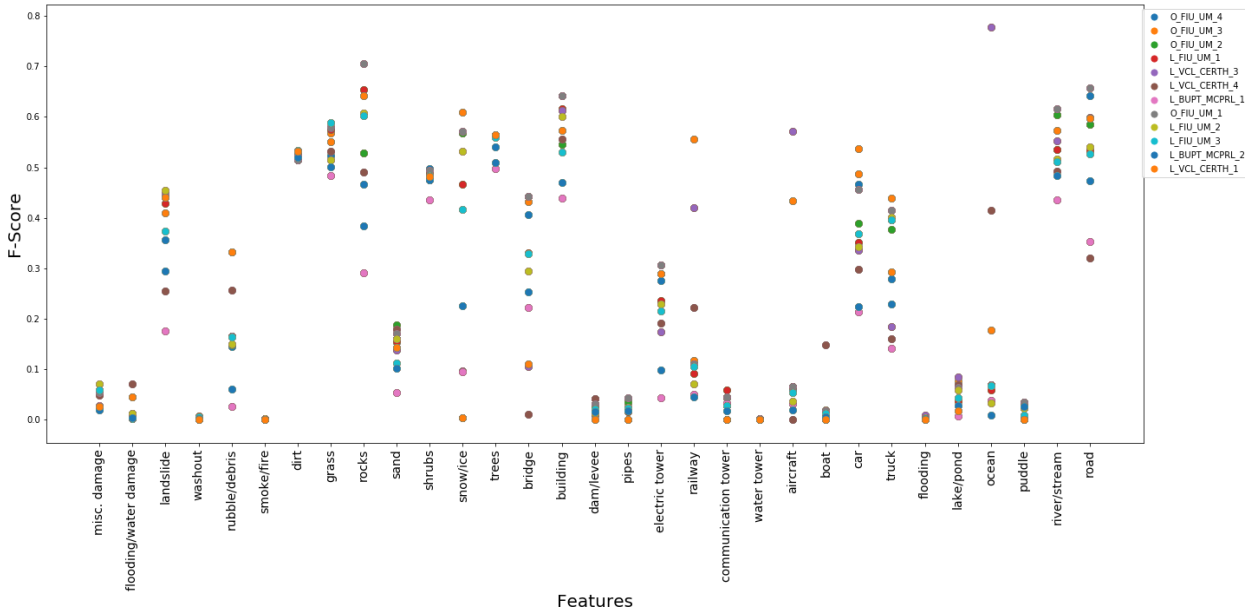

Figure 34: DSDI: F-measure for all the runs.

practical work on solving the challenges posed in automatic video captioning.

There are many use-case application scenarios that 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 the prediction of future events from the video.

The Video to Text (VTT) task was introduced in TRECVID 2016. Since then, there have been substantial improvements in the dataset and evaluation. The major changes for this year include:

1. A new subtask called "Fill-in-the-Blanks" has been introduced and the previous "Matching and Ranking" subtask has been discontinued.

2. We will start system progress monitoring for 3 years. This is similar to AVS and INS tasks. We have selected a subset of "progress" videos for which we will currently withhold the ground truth. Participants will then be able to compare their systems for 2022 and 2023 to measure improvement over the years on the same set of videos.

System Task

The VTT task is divided into two subtasks:

- Description Generation Subtask
- Fill-in-the-Blanks Subtask

The subtasks are independent of each other, which means that teams may participate in either one or both.

Details of the two subtasks are as follows:

- **Description Generation**: 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. Up to 4 runs are allowed per team.

- **Fill-in-the-Blanks**: For each video a corresponding description sentence is provided with a blank denoting a missing word or words. Return the most appropriate word or words to fill in the blank and complete each sentence. Up to 2 runs are allowed per team.

The new Fill-in-the-Blanks subtask is a variation of visual question answering (VQA) and requires systems to understand both the visual and textual information to find the missing word(s). This can play an

important part in video understanding using multimodal information.

For this year, 5 teams participated in the VTT task. The 5 teams submitted a total of 15 runs for the description generation subtask. 2 teams participated in the fill-in-the-blanks subtask and submitted a total of 3 runs. A summary of participating teams is shown in Table 6.

Data

Starting in 2020, the VTT data is 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 1977 video segments were annotated manually by multiple annotators for this year's task. Of these, we have selected 300 videos for our progress set. Hence, our results will be reported for 1677 videos.

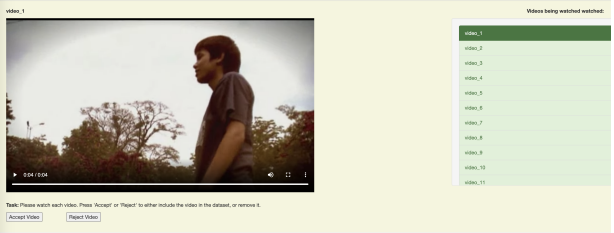

Figure 35: VTT: Screenshot of video selection tool.

It is important for a good dataset to have a diverse set of videos. We reviewed around 9000 videos and selected 1977 videos. Figure 35 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	17.27	867
2	18.77	867
3	18.97	810
4	19.14	810
5	19.50	834
6	20.00	810
7	20.33	843
8	25.42	810
9	27.78	867
10	32.24	867

Table 7: VTT: Average number of words per sentence for all the annotators. The table also shows the number of videos watched by each annotator.

Annotation Process The videos were divided among 10 annotators, with each video being annotated by exactly 5 people.

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** was the video taken (e.g., locale, site, place, geographic location, architectural)?
- **When** was the video taken (e.g., time of day, season)?

Different annotators provide varying amounts 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 17.27 words to 32.24 words, emphasizing the difference in descriptions provided by the annotators. The overall average sentence length for the dataset is 21.99 words.

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 2.71 (on a scale of 1 to 5), showing that the annotators thought the videos were close to medium level of difficulty on average. The average score for the second question was 2.15 (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 -0.59.

Submissions

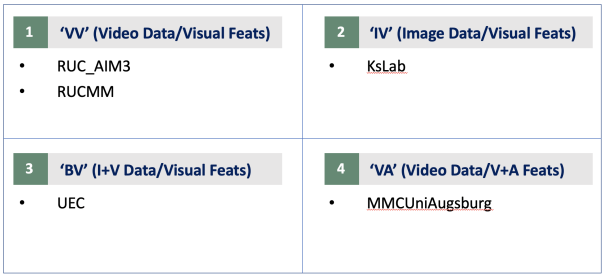

Figure 36: VTT: Run types for description generation submissions.

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.

Figure 36 shows the run types submitted by the teams for the description generation subtask. The run types for the 3 fill-in-the-blanks runs were all 'VV'.

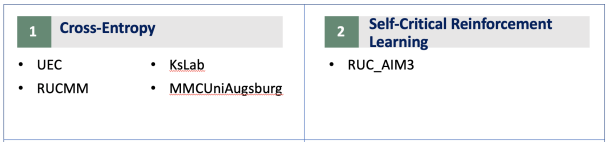

Figure 37: VTT: Loss functions for description generation submissions.

Teams were also asked to specify the loss function used for their runs, and Figure 37 shows the loss functions used by the teams for the description generation task. For the fill-in-the-blanks subtask, all runs used the cross-entropy loss function.

Evaluation and Metrics

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 and Natural Language Generation [Graham et al., 2016, Mille et al., 2020]. DA now constitutes the official method of ranking in main MT benchmark evaluations [Bojar et al., 2017, Barrault et al., 2020].

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), 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. For instance, 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.

The fill-in-the-blank subtask was evaluated using manual scoring in a manner similar to DA. Human assessors watched a video and the corresponding sentence with a blank followed by a word or words chosen by a system. The assessors ranked the missing words on a scale of 1 - 100. The final score is reported as a z-score, where the raw score is standardized per individual assessor's mean and standard deviation.

Overview of Approaches

For detailed information about the approaches and results for individual teams' performance and runs, we refer the reader to the site reports [TV21Pubs, 2021] in the online workshop notebook proceedings. Here we present a high-level overview of the different systems.

RUC_AIM3 proposes a concept enhanced pretraining-based Transformer model (CE-PTM) that has four parts, namely video encoder, text encoder, concept encoder, and multimodal transformer. For Fill-in-the-Blanks subtask, they propose four approaches based on pretraining-based Transformer model. Hybrid reranking allows them to create an ensemble of the methods.

RUCMM improves the encoder and decoder of the Bottom-Up-Top-Down (BUTD) model [Anderson et al., 2018]. They use scene-level as well as object-level video features, and two attention LSTMs for them. Finally, they also use reinforced adversarial learning so that the generated captions have similar fluency and language style as ground truth descriptions.

KsLab uses image-only datasets to train the network. They use DenseNet to extract visual features and report a small improvement over ResNet. BURTSUM is used for text aggregation. The resulting sentences are relatively short leading to lower evaluation scores.

MMCUniAugsburg uses a model based on the Transformer architecture [Vaswani et al., 2017]. They use image and audio embedding layers along with positional encoding. Unlike the previous year, they use both 2D and 3D features that were extracted

using I3D. The system is trained on VATEX and 90% of TRECVID-VTT data. The models are fine-tuned with self critical sequence training that optimizes CIDEr and BLEU-4 metrics.

UEC fine-tunes the image captioning model of [Luo et al., 2018]. They use pretrained ResNet-101 model to extract visual features. The captioning model uses attention mechanism. The model is trained on COCO and TRECVID datasets.

Results

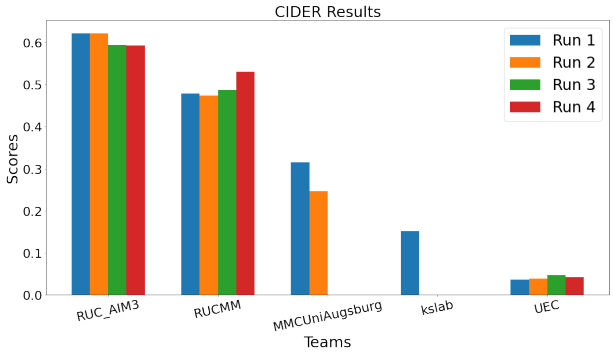

Figure 38: VTT: Comparison of all runs using the CIDEr metric.

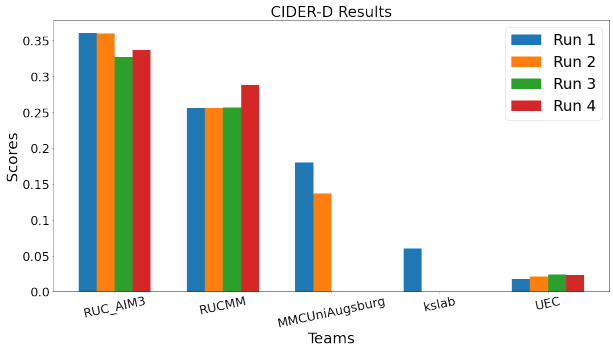

Figure 39: VTT: Comparison of all runs using the CIDEr-D metric.

Description Generation The description generation subtask scoring was done using popular automatic metrics that compare the system generation captions with ground truth captions as provided by assessors. We also continued the use of Direct Assessment, which was introduced in TRECVID 2017, to compare the submitted runs.

The metric score for each run is calculated as the average of the metric scores for all the descriptions

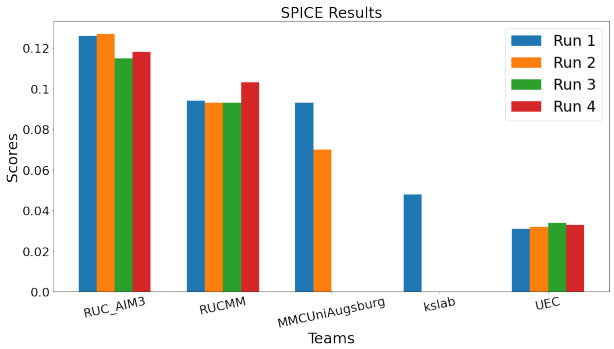

Figure 40: VTT: Comparison of all runs using the SPICE metric.

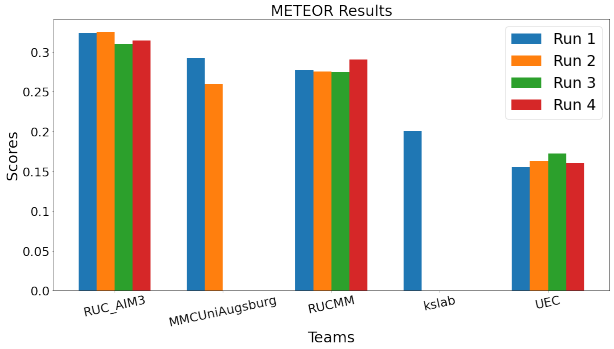

Figure 41: VTT: Comparison of all runs using the METEOR metric.

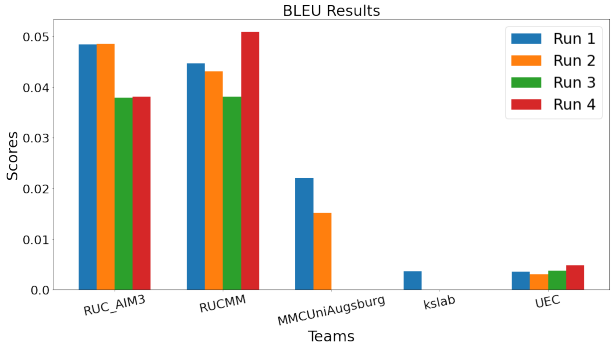

Figure 42: VTT: Comparison of all runs using the BLEU metric.

	CIDER	CIDER-D	SPICE	METEOR	BLEU	STS
CIDER	1.000	0.997	0.984	0.960	0.960	0.986
CIDER-D	0.997	1.000	0.990	0.964	0.956	0.980
SPICE	0.984	0.990	1.000	0.988	0.929	0.982
METEOR	0.960	0.964	0.988	1.000	0.897	0.984
BLEU	0.960	0.956	0.929	0.897	1.000	0.947
STS	0.986	0.980	0.982	0.984	0.947	1.000

Table 8: VTT: Correlation between overall run scores for automatic metrics.

	CIDEr	CIDErD	SPICE	METEOR	BLEU	STS
CIDEr	1.0000	0.9060	0.6730	0.7010	0.6330	0.6910
CIDErD	0.9060	1.0000	0.6500	0.6950	0.6290	0.6070
SPICE	0.6730	0.6500	1.0000	0.7340	0.6230	0.7110
METEOR	0.7010	0.6950	0.7340	1.0000	0.6460	0.7240
BLEU	0.6330	0.6290	0.6230	0.6460	1.0000	0.5330
STS	0.6910	0.6070	0.7110	0.7240	0.5330	1.0000

Table 9: VTT: Correlation between individual description scores for automatic metrics.

	CIDER	CIDER-D	SPICE	METEOR	BLEU	STS	DA_Z
CIDER	1.000	0.998	0.969	0.922	0.975	0.979	0.990
CIDER-D	0.998	1.000	0.975	0.928	0.961	0.976	0.980
SPICE	0.969	0.975	1.000	0.986	0.901	0.989	0.948
METEOR	0.922	0.928	0.986	1.000	0.846	0.976	0.910
BLEU	0.975	0.961	0.901	0.846	1.000	0.942	0.991
STS	0.979	0.976	0.989	0.976	0.942	1.000	0.979
DA_Z	0.990	0.980	0.948	0.910	0.991	0.979	1.000

Table 10: VTT: Correlation between overall run scores for the primary runs.

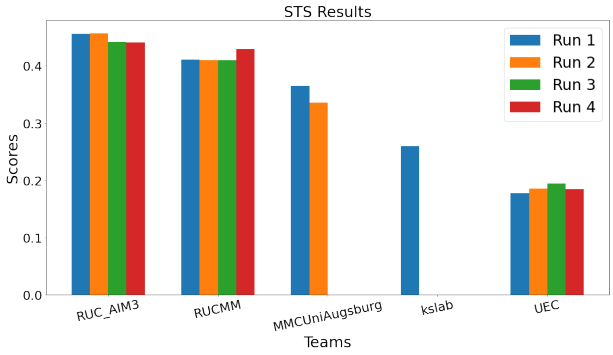

Figure 43: VTT: Comparison of all runs using the STS metric.

within that run. Figure 38 shows the performance comparison of all teams using the CIDEr metric. All runs submitted by each team are shown in the graph. Figure 39 shows the scores for the CIDEr-D metric, which is a modification of CIDEr. Figure 40 shows the SPICE metric scores. Figures 41 and 42 show the scores for METEOR and BLEU metrics respectively. The STS metric allows the comparison between two sentences. For this reason, the captions are compared to a single ground truth description at a time, resulting in 5 STS scores. We will report the average of these scores as the STS score, and Figure 43 shows how the runs compare on this metric.

Table 8 shows the correlation between the different metric scores for all the runs. The metrics correlate very well, which shows that they agree on the overall scoring of the runs. However, if we look at the description level metric scores, as shown in Table 9, we find that the metrics do not correlate as well. CIDEr and CIDEr-D have a very high correlation score since they are based on the same method. However, the correlation scores between all other metrics range between 0.5 and 0.75.

Figure 44 shows how the systems compare according to the CIDEr-D metric. The green squares indicate that the system in the row is significantly better ($p < 0.05$) than the system in the column. SPICE, BLEU, and STS significance tests show the same results. It can be seen that for these metrics, RUC_AIM3 is significantly better than the other teams.

Figure 45 shows the average DA score $[0-100]$ for each system. The score is micro-averaged per caption, and then averaged over all videos. Figure 46 shows the average DA score per system after it is standardized per individual AMT worker's mean and

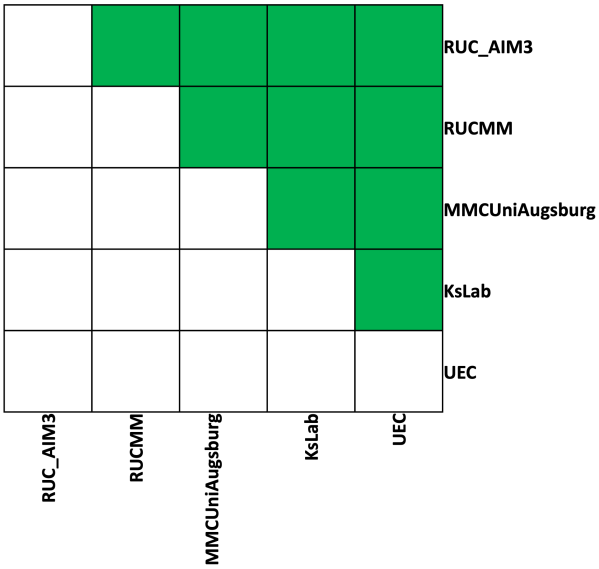

Figure 44: VTT: Comparison of the primary runs of each team with respect to the CIDEr-D score. Green squares indicate a significantly better result for the row over the column.

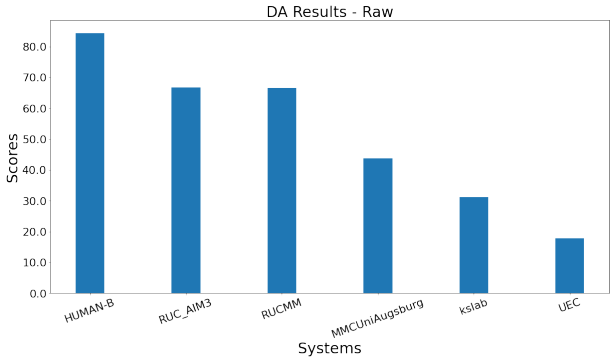

Figure 45: VTT: Average DA score for each system. The systems compared are the primary runs submitted, along with a manually generated system labeled as HUMAN_B.

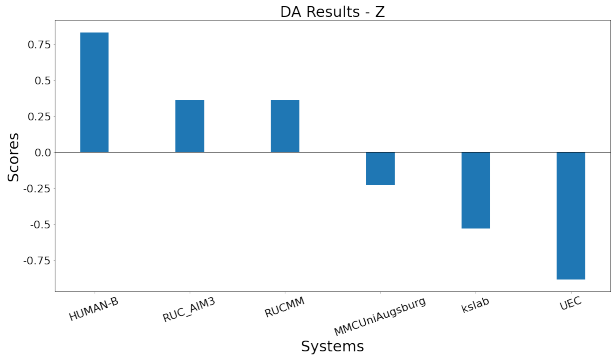

Figure 46: VTT: Average DA score per system after standardization per individual worker's mean and standard deviation score.

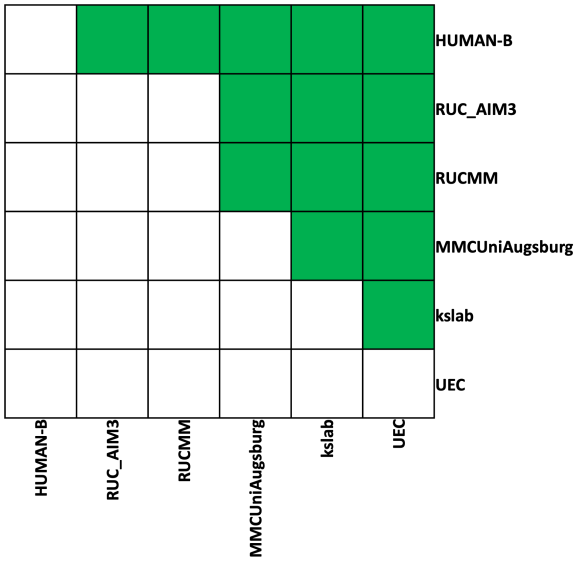

Figure 47: VTT: Comparison of the primary runs of each team with respect to the DA score. The 'HUMAN' system is ground truth captions. Green squares indicate a significantly better result for the row over the column.

standard deviation score. The HUMAN system represents manual captions provided by assessors. As expected, captions written by assessors outperform the automatic systems. Figure 47 shows how the systems compare according to DA. The green squares indicate that the system in the row is significantly better than the system shown in the column ($p < 0.05$). The figure shows that no system reaches the level of human performance. Among the systems, RUC_AIM3 and RUCMM are significantly better than the other systems.

Table 10 shows the correlation between different overall metric scores for the primary runs of all teams. The 'DA_Z' metric is the score generated by humans. The score correlates very well with the other metrics.

Teams were asked to provide a confidence score for each generated sentence. We expected these confidence scores to have a positive correlation with the metric scores. Figure 48 shows the correlation of the sentence confidence scores reported by the systems and the various metric scores. RUC_AIM3 and RUCMM show better correlation than the others with most metrics, especially CIDEr. BLEU seems to be the least correlated with the confidence scores. While most runs have a positive correlation, KsLab has negative correlation for some metrics.

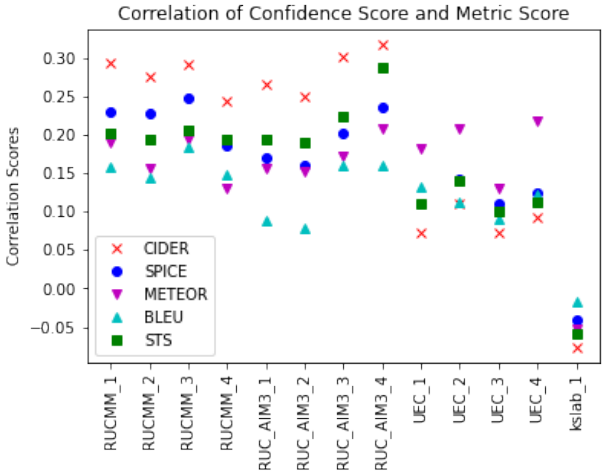

Figure 48: VTT: Correlation of system reported sentence confidence scores and the various metric scores. The two top performing teams, RUC_AIM3 and RUCMM, show a better correlation with most metrics.

Table 11 shows the average length of sentences for each run. The table also shows the average length when only unique words are counted. These lengths

	Run	Avg Length	Avg Length (Unique)
1	KSLAB_1	11.50566	9.673226
2	RUCMM_2	13.45617	11.90757
3	RUCMM_3	13.57305	11.80262
4	RUCMM_1	13.61419	12.04114
5	RUCMM_4	13.91354	12.11032
6	RUC_AIM3_3	14.6297	11.26297
7	UEC_1	14.96959	13.0316
8	UEC_3	15.07633	12.30769
9	UEC_2	15.09243	12.77519
10	UEC_4	15.13596	12.97018
11	RUC_AIM3_4	15.57245	11.83125
12	RUC_AIM3_1	15.87001	12.27549
13	RUC_AIM3_2	16.00537	12.35957
14	MMCUniAugsburg_2	16.44902	12.36315
15	MMCUniAugsburg_1	17.46094	13.18903

Table 11: VTT: The table shows the average length of sentences for each run. The Avg Length (Unique) column shows the average length when only unique words are counted.

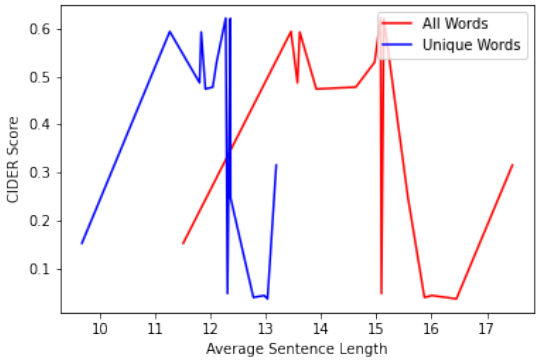

Figure 49: VTT: Correlation of average sentence length and the CIDEr score.

can vary significantly for certain runs where words and phrases repeat often. Figure 49 shows how the average sentence length correlates with the CIDEr score. There does not seem to be any pattern to suggest that longer sentences score better.

Fill-in-the-Blanks The Fill-in-the-Blanks subtask was introduced this year. It runs independently from the description generation subtask. Participants are provided with a video and a corresponding sentence with word or words missing. The goal is to predict the best word or words to complete the sentence. Given that each video has a single ground truth, most automatic metrics are unsuitable for scoring runs. We, therefore, used manual evaluation. Table 12 shows the scores for the 3 team runs and a Human system, which has ground truth solutions. The teams are divided into clusters, where the higher clusters significantly outperform lower clusters. The RUC_AIM3 runs outperform the one by RUCMM.

We also attempted a character n-gram F-score metric [Popović, 2015] for automatic evaluation of the runs. The scores for 4-gram and 6-gram are shown in table 13. Due to the small number of runs, it is hard to judge the quality of the results. However, since the metric only looks at characters present in the ground truth, the metric cannot be a substitute for manual evaluation.

Figure 50 shows an example video and caption of a fill-in-the-blanks subtask. It illustrates the need for

Teams	Average-Z
HUMAN	0.420
RUC_AIM3_RUN2	0.173
RUC_AIM3_RUN1	0.130
RUCMM	-0.102

Table 12: VTT: The table shows the Average-Z scores for all the systems, including a Human system. The double lines separate clusters, where the lower ranked clusters are are significantly outperformed by higher ranked clusters.

Teams	C-4 Scores	C-6 Scores
RUC_AIM3_RUN1	44.06	36.91
RUC_AIM3_RUN2	41.89	34.98
RUCMM	9.94	7.00

Table 13: VTT: The table shows the chrf scores for 4-gram and 6-gram.

- Sentence:

Male and female university athletes chant together on _____ to show that they are united.

- Ground Truth Answer:
  - a sports field
- System Answers:
  - a sports field
  - huddle

Figure 50: VTT: An example video and caption in the fill-in-the-blanks subtask. One of the system outputs, 'a sports field', matches exactly with the ground truth. However, the answer 'huddle' is not entirely wrong. Despite a minor grammatical error in the sentence, this answer does make sense in the context.

manual evaluation as the sentence can be completed in many different ways, and the generated word need not match the ground truth at all.

Conclusion and Future Work

The VTT task continues to have healthy participation. Given the challenging nature of the task, and the increasing interest in video captioning in the computer vision community, we hope to see improvements in performance.

This year we continued with the V3C2 dataset and plan to continue with this dataset for the next year. With increasing interest in video captioning, participants have a number of open datasets available to train their systems.

Some of the major changes this year included:

1. A progress task was introduced, where 300 videos were selected. The ground truth for these videos will be withheld to compare the performance of systems over three years on the same set of videos.

2. The matching and ranking subtask was removed.

3. The fill-in-the-blanks subtask was introduced.

For the next year, we will decide on whether to continue with the fill-in-the-blanks subtask. We are exploring other possible subtasks including VQA and dense video captioning. Possible improvements in dataset may include object localization in videos.

3.5 Activities in Extended Video

The Activities in Extended Video (ActEV) evaluation series is designed to accelerate the development of robust, multi-camera, automatic human activity detection systems for forensic and real-time alerting applications. 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) associated with the activity. This year we continued with the ActEV task same as in 2020 with 35 target activities. NIST TRECVID ActEV series

was initiated in 2018 to support the Intelligence Advanced Research Projects Activity (IARPA) Deep Intermodal Video Analytics (DIVA) Program.

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 Video (MEVA) dataset [Kitware, 2020] with 37 activities. 2) The TRECVID 2021 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 datasets [Lee et al., 2018]. For the self-reported evaluation, the participants ran their software on their hardware and configurations and submitted the system outputs with the defined format to the NIST scoring server. For the leaderboard evaluation, the participants submitted their runnable systems to the NIST scoring server, which was independently evaluated on the sequestered data using the NIST hardware.

The ActEV18 evaluation addressed 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 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. The TRECVID 2021 ActEV (ActEV21) is also based on the same 35 activities as ActEV20 and on the VIRAT V1 and V2 datasets and systems are evaluated on the activity detection (AD) task only.

Figure 51 illustrates an example of representative activities that were used in the TRECVID 2021 ActEV.

All these evaluations are primarily targeted for the forensic analysis that processes an entire corpus prior to returning a list of detected activity instances.

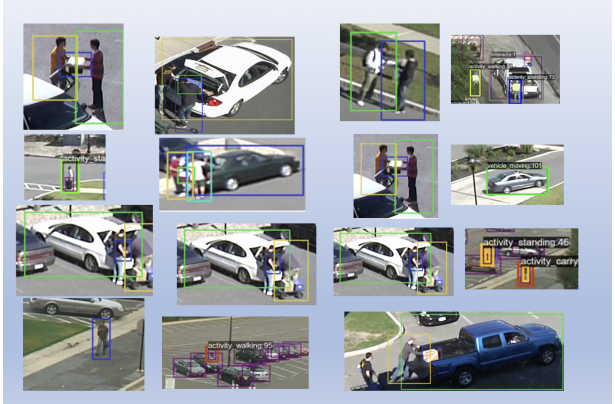

Figure 51: Example of activities for ActEV series. IRB (Institutional Review Board): ITL-00000755

In this section, we first discuss the task and datasets used and introduce a new metric to evaluate algorithm performance. In addition, we present the results for the TRECVID ActEV21 submissions and discuss observations and conclusions.

Task and Dataset

In the ActEV21 leaderboard evaluation, we addressed activity detection (AD) task for detecting and localizing activities; a system was required to automatically detect and localize 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 ActEV21 was an open leaderboard evaluation. The challenge participants were required to run their systems locally and submit the outputs in a pre-specified format to the NIST scoring server. The systems were supposed to detect target activities that visibly occurred in a single-camera video, as well as the frame span (the start and end frames) of the detected activity instance along with a confidence score indicating the likelihood of the presence of the activity within the frame boundaries.

For this evaluation, we used 35 activities from the VIRAT dataset and the activities were annotated by Kitware, Inc. The VIRAT dataset consists of 29 hours of video and more than 43 activity types. A total of 10 hours of video 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 14 lists the number of instances for each activity for the training and validation sets. Due to

Table 14: A list of activity names for TRECVID ActEV, for ActEV19 there were 18 activities and for ActEV20 and ActEV21 there were 35 activities based on the VIRAT dataset and their associated number of instances for the training and validation sets are also listed.

VIRAT19 (18 Activities)	VIRAT20, ActEV21 (35 Activities)	Train	Validate
Closing	person_closes_facility_or_vehicle_door	141	130
Closing_Trunk	person_closes_trunk	21	31
x	vehicle_drops_off_person	0	4
Entering	person_enters_facility_or_vehicle	77	70
Exiting	person_exits_facility_or_vehicle	66	72
x	person_interacts_object	101	88
Loading	person_loads_vehicle	38	38
Open_Trunk	person_opens_trunk	22	35
Opening	person_opens_facility_or_vehicle_door	137	128
x	person_person_interaction	11	17
x	person_pickups_object	19	12
x	vehicle_picks_up_person	9	5
Pull	person_pulls_object	23	43
	person_pushs_object	4	6
Riding	person_rides_bicycle	22	21
x	person_sets_down_object	12	11
Talking	person_talks_to_person	41	67
Transport_HeavyCarry	person_carries_heavy_object	31	44
Unloading	person_unloads_vehicle	32	44
activity_carrying	person_carries_object	237	364
x	person_crouches	1	9
x	person_gestures	82	148
x	person_runs	14	18
x	person_sits	21	11
x	person_stands	398	819
x	person_walks	761	901
specialized_talking_phone	person_talks_on_phone	17	16
specialized_texting_phone	person_texts_on_phone	5	20
x	person_uses_tool	7	11
x	vehicle_moves	718	797
x	vehicle_starts	259	239
x	vehicle_stops	292	295
vehicle_turning_left	vehicle_turns_left	152	176
vehicle_turning_right	vehicle_turns_right	149	172
vehicle_u_turn	vehicle_makes_u_turn	9	13

possible future evaluations based on the dataset, the information about the test sets is not included in the table. The frequency of instances are not balanced across activities, which may affect the system performance results.

Measures

Activity detection in extended video is not a discrete detection task unlike speaker recognition [Greenberg et al., 2020] and fingerprint identification [Karu and Jain, 1996], it is a streaming detection task where multiple activity instances can overlap temporally or spatially and is similar to keyword spotting in audio [Le et al., 2014]. From a metrology perspective the difference between discrete and streaming detection tasks is that non-target trials (i.e., test probes not belonging to the class) are not countable for streaming detection because the number of unique temporal/spatial instances are near infinite. To account for this difference, the ActEV evaluations used two methods to normalize the measured false alarm performance. The first, "Rate of False Alarms", is an instance-based false alarm measure that uses the number of video minutes as an estimate of the number of non-target trials as the false alarm denominator. The second, "Time-based False Alarms", is a time-based false alarm measure that used the sum of non-target time as the denominator. The two variations correspond to two views concerning the impact false alarms have on a user reviewing detection. The former is instance-based which implies the user effort would scale linearly with the detected instances and the latter time-based which implies the user effort would scale linearly with the duration of video reviewed.

The primary measure of performance for TRECVID ActEV21 is the normalized, partial Area Under the DET Curve ($nAUDC$) from 0 to a fixed, Time-based False Alarm (T_{fa}) nAUDC T_{FA} value a , denoted $nAUDC_a$, which is the same as the metric used for the TRECVID ActEV20 and ActEV19 evaluations. All ActEV performance measurements were on a per-activity basis and then performance was aggregated by averaging over activities. While presence confidences scores were used to compute performance, cross-activity presence confidences score normalization was not required nor evaluated.

For TRECVID ActEV18, the primary metric was instance-based measures for both missed detections and false alarms (as illustrated in Figure 52). The metric evaluates how accurately a system detects instance occurrences of the activity.

As shown in Figure 52, the detection confusion matrix is 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 scores are not used as a decision threshold. Rather, a decision threshold is applied on the scores to determine the error counts (N_{FA} and N_{miss}).

In the ActEV21 evaluation (same as for ActEV19 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 the video duration in minutes. $N_{\text{TrueInstance}}$ is the number of reference instances annotated in the sequence per activity. Lastly, the Detection Error Tradeoff (DET) curve [Martin et al., 1997] is used to visualize system performance. For the TRECVID ActEV18 challenge, we evaluated algorithm performance for two 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 ActEV21, 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:

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

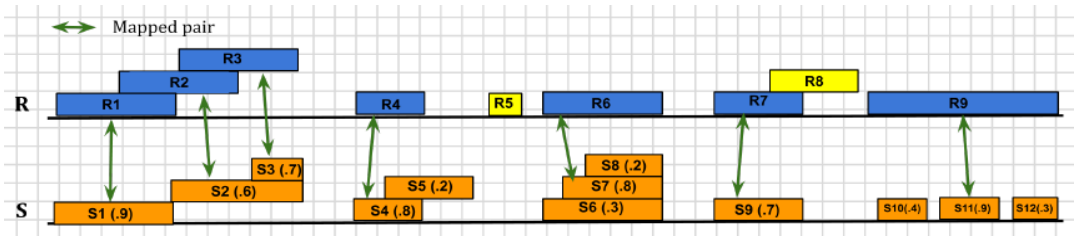

Figure 52: Illustration of activity instance alignment and P_{miss} calculation (R is the reference instances and S is the system instances. In S, the first number indicates instance id and the second indicates presence confidence score. For example, $S1(.9)$ represents the instance $S1$ with corresponding confidence score (.9). Green arrows indicate aligned instances between R and S)

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 53.

The non-reference duration (NR) of the video where no target activities occur 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 53); 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.

Figure 54 shows a summary of performance metric calculation. For given reference annotation and system output, the steps are 1) Align the reference activity instance with each relevant system's instance; 2)Compute detection confusion matrix; 3)Compute summary performance metrics; and 4) Visualize the results such as DET curve shown here, which the x-axis is Time-based False Alarm (TFA) Rate and y-axis is probability of missed detection. For ActEV21 our primary metric is mean Normalized partial Area Under the DET Curve $\mu nAUDC$. 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 ActEV19-21 metric, we also used Time-based false alarms and calculated $\mu nAUDC$ from T_{FA} 0 to 0.2.

ActEV Results

A total of 6 teams from academia and industry from 4 countries participated in the ActEV21 evaluation. Each participant was allowed to submit multiple system outputs and a total of 104 submissions were received. Table 15 lists the participating teams along with results ordered by $\mu nAUDC$ scores for the best performing system per team along with $mean\_P_{miss}@.15T_{FA}$ values. The best performance on activity detection is by BUPT-MCPRL at 40.9% followed by UCF at 43.1% and INF-CMU is third at 44.4

Figure 55 shows the ranking of activities over the top 3 systems. The x-axis is the activities; The y-axis is $\mu nAUDC$ and a smaller value is considered better performance. The color-coded points represent $\mu nAUDC$ values for each system and the hor-

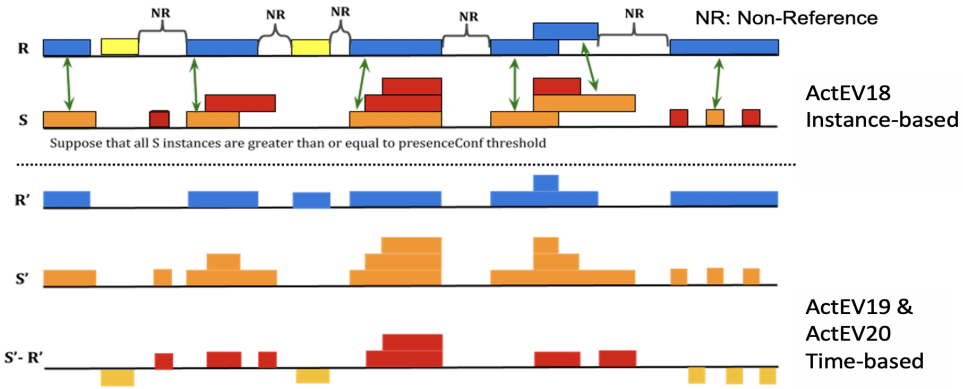

Figure 53: Comparison of instance-based and time-based false alarms. R is the reference instances and S is the system instances. R' is the histogram of the count of reference instances and S' is the histogram of the count of system instances for the target activity. S shows a depiction of instance-based false alarms while $S' - R'$ illustrates time-based false alarms as marked in red.

Table 15: Summary of participants information and and results ordered by $\mu nAUDC$ values, along with $mean\_P_{miss}@T_{FA}.15$ values. Each team was allowed to have multiple submissions.

Team Name	**Organization**	$\mu nAUDC$	$\mu P_{miss}@.15T_{fa}$
BUPT-MCPRL	Beijing University of Posts and Telecommunications, China	0.409	0.325
UCF	University of Central Florida, USA	0.431	0.341
INF-CMU	Carnegie Mellon University, USA	0.444	0.351
M4D-2021	Information Technologies Institute, Greece	0.847	0.794
TokyoTech-AIST	Tokyo Institute of Technology, Japan	0.852	0.82
Team UEC	The University of Electro-Communications, Japan	0.964	0.95

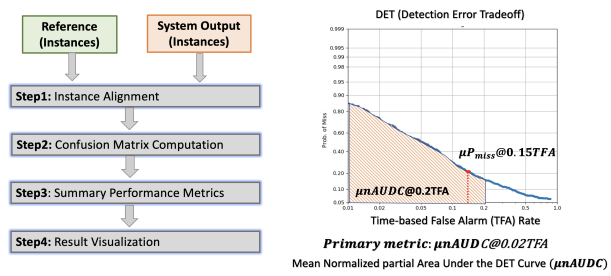

Figure 54: Performance Measure Calculation and Detection Error Tradeoff (DET) Curves

izontal black bar indicates the median. The systems are ordered by mean $nAUDC$ over 35 activities. We observed that activity-level performance is vertically sparse for some activities (e.g., vehicle_turns_right) while the others have a smaller gap (e.g., person_pulls_object). In addition, bottom point colors changes which indicate each system has different strength depending on activities.

We observed that, given datasets and systems, person_uses_tool is the easiest to detect while person_pickup_object is the most difficult activity to detect across the systems.

Figure 56 illustrates the Activity-Level comparison over the top three systems. The 35 activities are shown in the different boxes, for each box the x-axis is the three team name and the y-axis is $\mu nAUDC$ value. A lower value is considered a better performance. Although BUPT-MCPRL performs better in general, each system has its own detection strength depending on activity type. For example, given dataset and systems, we observed that each system can better detect the following activities marked in blue. Some activities have a lower error rate across the top-three systems marked in red. This indicates that a fusion of the detection strength for multiple systems can potentially increase overall system performance further.

Figure 57 shows a comparison of Activity Detection Difficulty; the blue heatmap shows the activity ranking while the orange heatmap shows the instance count for the training and validation set. Where x-axis, shows the top three team names and average activity ranking (AVG) and y-axis, shows the 35 activities Numbers in the matrix, shows the ranking of 35 activities per system. If you look at the top 3 easiest and hardest activities, they are different, however, we can observe common activities such as person_rides_bicycle being easier and person_pickups_object being harder. Interestingly, both person_uses_tool and person_pickups_object have a low instance count for both training and validation set which provide a high uncertainty on system performance.

To examine the performance improvement from ActEV20 through ActEV21, Figure 58 shows the leaderboard evaluation results from ActEV20 and ActEV21. The same 6 teams participated in either ActEV20 or ActEV21 evaluations. Our results showed that some teams improved their system performance over the years. For example, both BUPT-MCPRL and UCF teams reduced the detection error rate by nearly 15% from the last year.

Summary

In this section, we presented the TRECVID ActEV21 evaluation task, the performance metric and results for human activity detection. We primarily focused on the activity detection task only and the time-based false alarms were used to have a better understanding of the system's behavior and to be more relevant to the use cases. The metric based on instance-based false alarms was used in ActEV18, ActEV19 and ActEV20 evaluations. The ActEV21 and ActEV20 activity names are the same and are consistent with the MEVA [Kitware, 2020] dataset and have 35 target activities in total. Six teams from 4 countries participated in the ActEV21 evaluation and made a total of 112 submissions. We observed that, given the datasets and systems, person_uses_tool is the easiest to detect while person_pickup_object is the most difficult activity to detect across the systems.

The TRECVID ActEV21 evaluation provided researchers an opportunity to evaluate their activity detection algorithms on a self-reported leaderboard. The competition also resulted in progress, BUPT-MCPRL and UCF teams reduced the activity detection error rate by nearly 15% from the last year. The INF-CMU submission missing the deadline slightly had the best performance. We hope the TRECVID ActEV21 evaluation, and the associated datasets will facilitate the development of activity detection algorithms. This will in turn provide an impetus for more research worldwide in the field of activity detection in videos.

3.6 Video Summarization

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

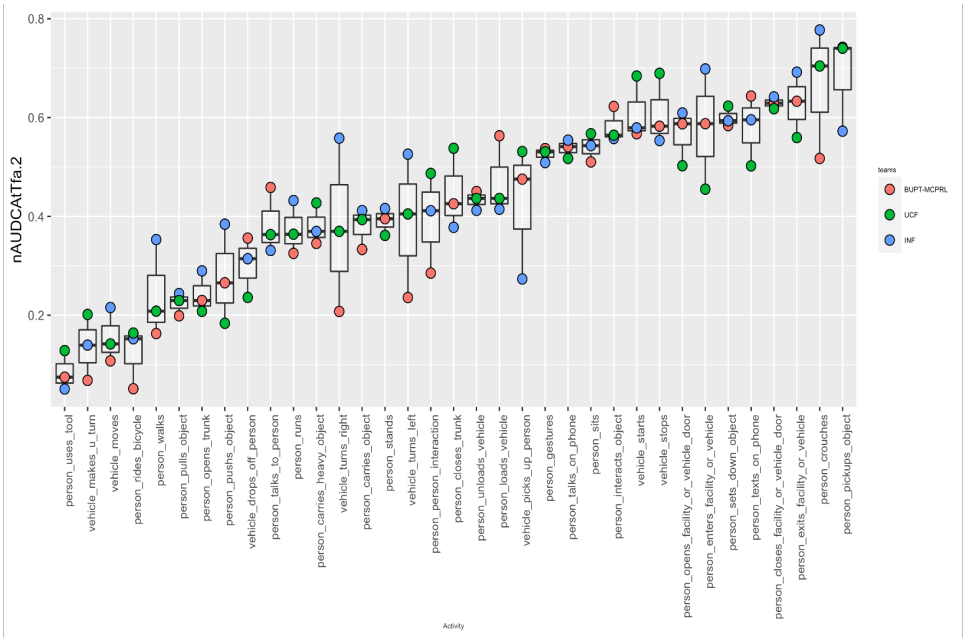

Figure 55: Ranking of Activities over the Top Three Systems. The 35 activities are shown in the different boxes, for each x-axis is the team name and the y-axis is $\mu nAUDC$ value. A lower value is considered a better performance.

Figure 56: Activity-Level comparison over the top three systems. The 35 activities are shown in the different boxes, for each box the x-axis is the three team names and the y-axis is $\mu nAUDC$ value. A lower value is considered a better performance

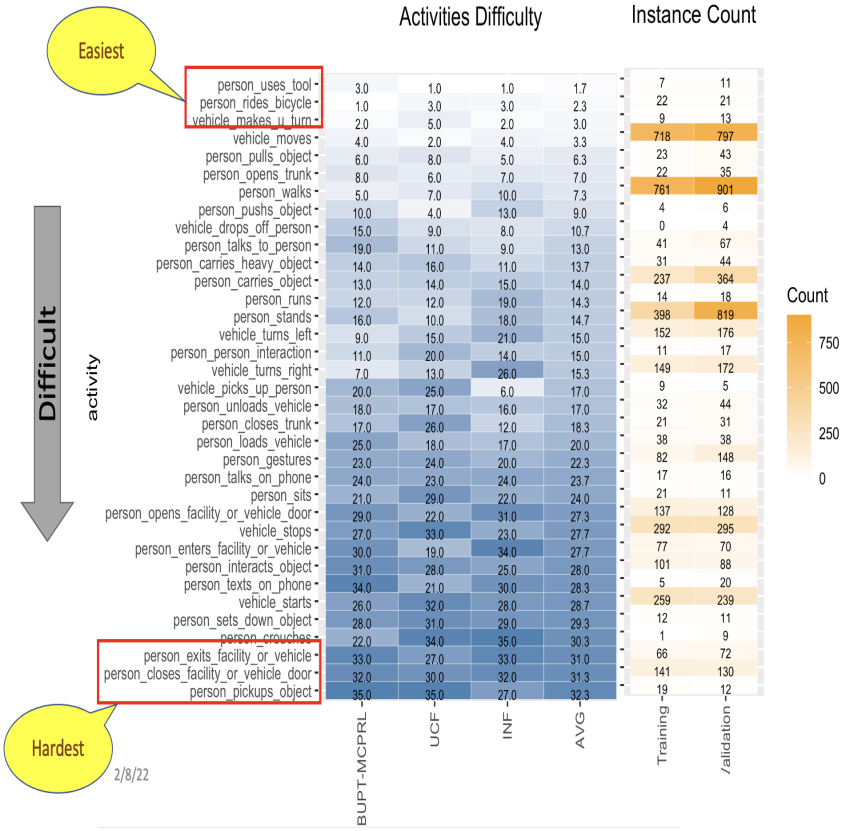

Figure 57: Which activities are easier or more difficult to detect?

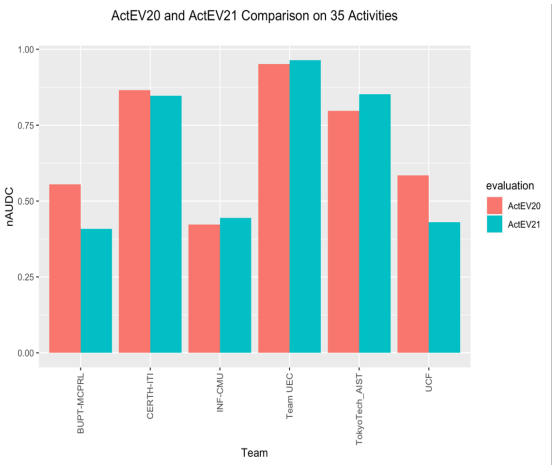

Figure 58: Comparison of ActEV20 vs ActEV21 results for the 35 activities

sonal 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. In 2020 we introduced a new video summarization track in TRECVID in which the task was to summarize the major life events of specific characters over a number of weeks of programming on the BBC Eastenders TV series. The plan is to, every year, choose a few characters from a specific period of the show, and to ask participating teams to produce summaries for the character's major life events in that period.

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 character's 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

The primary task for this track was, 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, summarize the major life events of each character within the specified time frame of the series. Some examples of major life events were as follows: 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. Summaries were limited to a maximum number of unique shots, thus the main challenge was 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 in bitmap (bmp) file format 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 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 assumed the searcher wanted to keep the process simple. So, the ROI could contain non-target pixels, e.g., non-target regions visible through the target or occluding regions.

Sub-task

A sub-task was introduced in 2021 in which teams had prior knowledge of the questions which were to be used for evaluation. All other requirements remained the same as for the main task.

Topics

By analyzing metadata of the full set of BBC Eastenders omnibus episodes, NIST selected queries of five characters who were shown to play a big part in the series over a ten week period. Three characters were selected from one ten-week period, and another two characters were selected from a different ten-week period of the series. The following five characters were chosen for the 2021 task:

- Max
- Jack
- Tanya
- Archie
- Peggy

In addition to specifying this year's query characters, the time frame of the series (Start Shot # and End Shot #) from which to generate summaries for each character, 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 16.

Evaluation

Each team was asked to submit 4 runs, with the maximum number of shots and maximum summary length

Table 16: Video Summarization Queries and Specifics

Character	Max	Jack	Tanya	Archie	Peggy
Start Shot #	shot60_1	shot60_1	shot60_1	shot79_1	shot79_1
End Shot #	shot70_2040	shot70_2040	shot70_2040	shot89_2036	shot89_2036
Max # Shots Run 1	5	5	5	5	5
Max Time Run 1	50 seconds	50 seconds	50 seconds	50 seconds	50 seconds
Max # Shots Run 2	10	10	10	10	10
Max Time Run 2	100 seconds	100 seconds	100 seconds	100 seconds	100 seconds
Max # Shots Run 3	15	15	15	15	15
Max Time Run 3	150 seconds	150 seconds	150 seconds	150 seconds	150 seconds
Max # Shots Run 4	20	20	20	20	20
Max Time Run 4	200 seconds	200 seconds	200 seconds	200 seconds	200 seconds

as specified in Table 16. In total, 3 participating teams submitted 12 runs, both for the main task and for the sub-task, giving 24 runs to be evaluated. Each run contained video summaries for each of the 5 specified queries, giving a total of 120 video summaries to be evaluated.

Submissions were evaluated by the TRECVID team at NIST, with two people responsible for evaluating summaries for two of the queries, and another person responsible for evaluating summaries for one query. Assessors answered 5 content-based questions for each of the 24 video summaries they had been asked to evaluate for each query. Content questions were created by the TRECVID team after watching each episode from the two specified time-frames of the series for the chosen queries, 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:

- Max
    1. What was the cause of Max's serious injuries which left him in hospital?
    2. What is/was the relationship between Max and Tanya?
    3. What kind of weapon does Max obtain from Phil?
    4. Where are Max and Jack during the violent confrontation between them when a gun is drawn?
    5. Who is responsible, or who does Max believe is responsible, for the serious injuries which left him in hospital?
- Jack
    1. What happens when police break in the door of Jack and Tanya's home?
    2. Where are Max and Jack during the violent confrontation between them when a gun is drawn?
    3. Who does Jack offer to pay in order to withdraw their statement to the police?
    4. Why is Jack a suspect in the hit and run on Max?
    5. What does Jack reveal to Tanya about his dodgy past?
- Tanya
    1. What does Tanya reveal to the police while being interviewed at the station?
    2. What is/was the relationship between Max and Tanya?
    3. What does Jack reveal to Tanya about his dodgy past?
    4. What does Tanya discover in the sink and on Jack's clothes?
    5. What big move were Tanya and Jack planning for the future?
- Archie
    1. What happens when Phil throws Archie into a pit?

2. What happens after Danielle reveals to Archie that Ronnie is her mother?
3. Where do Peggy and Archie get married?
4. What happens when Archie arrives at the pub after Peggy invited him?
5. What happens when Archie is kidnapped?

- Peggy
  1. Who does Peggy ask to kill Archie?
  2. Where do Peggy and Archie get married?
  3. Show one of the challenges which Peggy faces in her election run.
  4. What does Peggy overhear Archie saying, which causes his marriage to be over?
  5. What is Janine doing to irritate or anger Peggy?

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 better, were flipped. This gave a total of 21 possible marks available for subjective quality scores. The remainder was calculated by taking 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.

Results - Main Task

Table 17 and continued 18 show the main task individual results for each submission query and run on all metrics and content questions. Team ADAPT achieves the best results on the main task.

Figure 59 shows the average scores for each target query by team. Scores are averaged across all runs. **Archie** is shown to be the easiest character to summarize the major life events of, with **Peggy** the most difficult. **Tanya** obtains the most consistent results across teams.

Figure 60 shows the average scores for each target query by team. Scores are averaged across all target queries. Run 3 conditions (max 15 shots, 150 seconds) gives the best results for the main task. This is consistent with the previous year of the task.

Figure 61 shows the average scores for each team. Scores are averaged across all runs and target queries. Teams ADAPT and EURECOM achieve similar results, with ADAPT achieving slightly higher marks. NII_UIT achieves lower scores for the main task.

Figure 62 shows the individual scores for all teams, runs and target queries. This chart visualizes the final results shown in table 17, from which it can be seen that team ADAPT scores higher for **Archie** run 1 and 2 than for all other submitted summaries. From this chart we can also see remarkably consistent results across all teams for **Tanya** run 4.

Results - Sub Task

Table 19 and continued 20 show the sub task individual results for each submission query and run on all metrics and content questions. Team EURECOM achieves the best results on the main task, just about exceeding results for NII_UIT, who achieved the biggest improvement in performance for the subtask.

Figure 63 shows the sub-task average scores for each target query by team. Scores are averaged across all runs. **Tanya** is shown to be the easiest character to summarize the major life events of, with all teams achieving their best marks for this character. **Max** was the most difficult character to summarize for the sub-task.

Figure 64 shows the average scores for each target query by team. Scores are averaged across all target

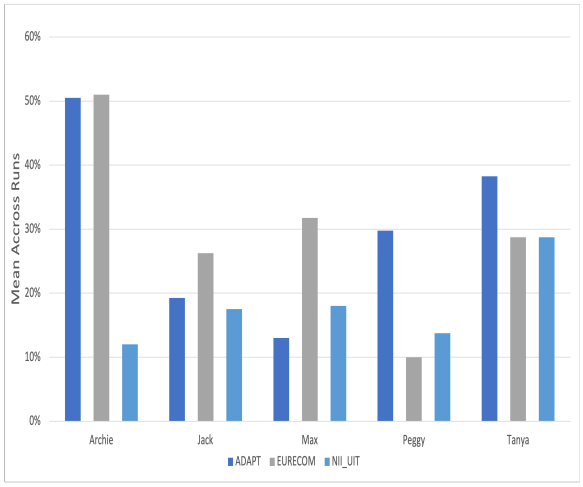

Figure 59: VSUM Main Task: Average scores by Character

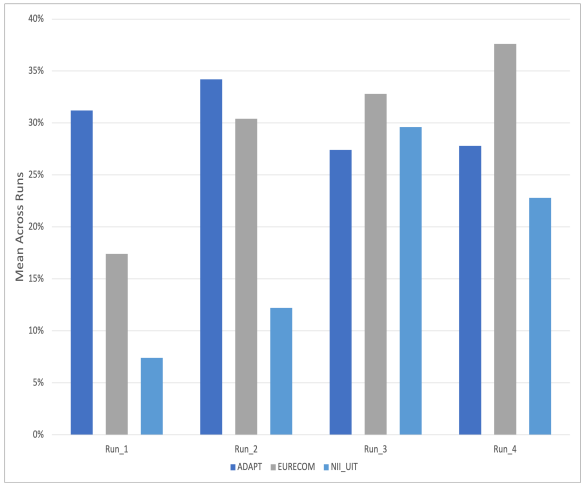

Figure 60: VSUM Main Task: Average scores for each run

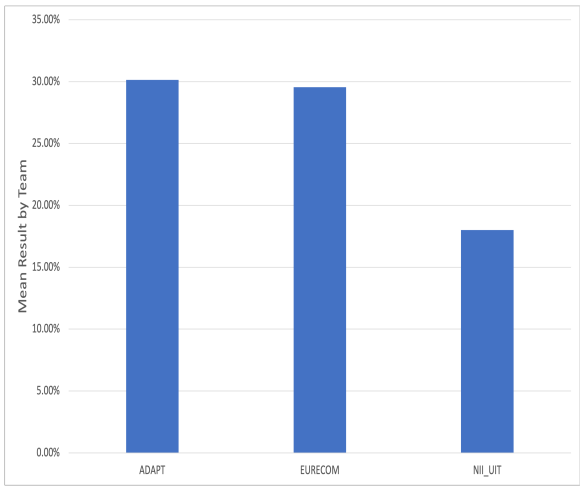

Figure 61: VSUM Main Task: Average scores by team

queries. Run 4 conditions (max 20 shots, 200 seconds) gives the best results for the sub-task. This is seen for all teams who participated in the sub-task.

Figure 65 shows the average scores for each team. Scores are averaged across all runs and target queries. Teams EURECOM and NII_UIT achieve similar results, with EURECOM achieving slightly higher marks. ADAPT achieves lower scores for the main task. NII_UIT shows by far the largest performance improvement for the sub task of known questions.

Figure 66 shows the individual scores for all teams, runs and target queries. This chart visualizes the final results shown in Table 19, from which it can be seen that team NII_UIT scores higher for **Tanya** run 4 and **Archie** run 2 than for all other submitted summaries, with nearly perfect scores for **Tanya** run 4. From this chart we can also see consistent results across all teams for **Archie** run 1 and **Max** run 1.

Observations

This is the second year of the video summarization task. Due to this, the decision was taken to require that teams submit results for 4 different runs, specified by a maximum number of shots and maximum summary length in seconds. It was found that the conditions for run 3 scores higher for the main task, which is consistent with the previous year of the task, however run 4 conditions score higher for the sub-task of known questions. The maximum allowed length for summaries was reduced this year by two-thirds, or

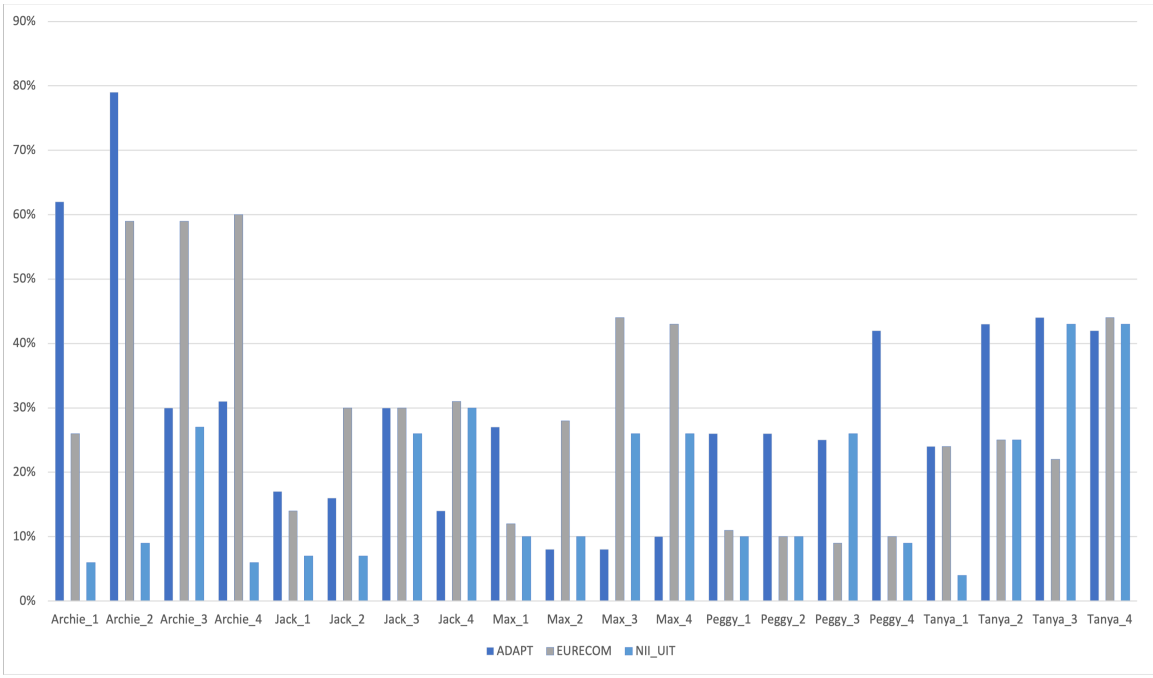

Figure 62: VSUM Main Task: Individual scores

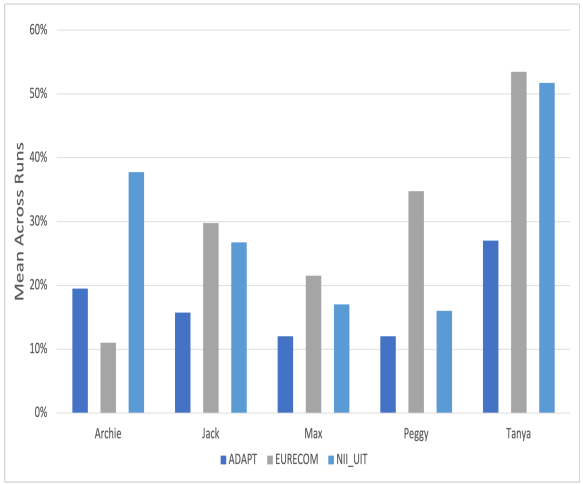

Figure 63: VSUM Sub Task: Average scores by Character

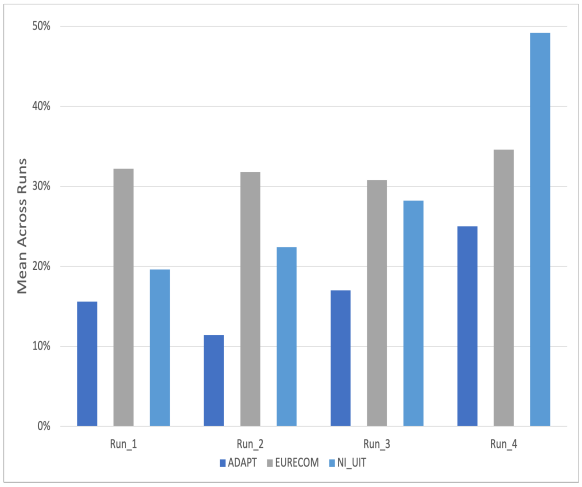

Figure 64: VSUM Sub Task: Average scores for each run

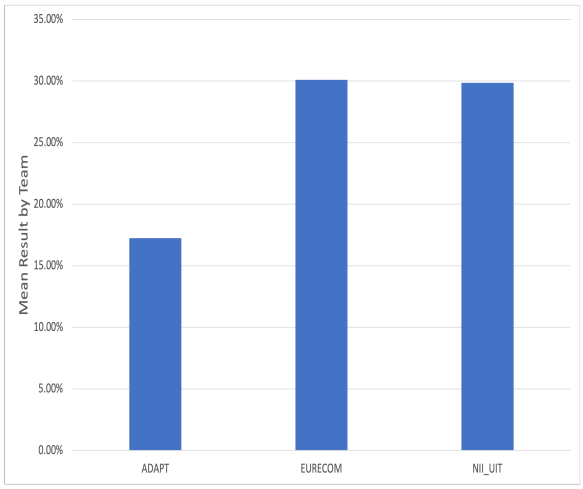

Figure 65: VSUM Sub Task: Average scores by team

66.6%. All submitted summaries used all of the maximum number of shots allowed, and all were shorter than the maximum allowed summary length.

We now summarize the approaches taken by teams. Team ADAPT used short fan videos from YouTube and archived videos for image dataset preparation. Frames with characters from the task were selected and OpenCV [Bradski, 2000] methods used to sample only frames with faces. By using Keras [Chollet et al., 2018] augmentation methods, several datasets were created with different augmentation options. Tensorflow API [Abadi et al., 2016] was used for model development and Keras API methods were used to solve the problem of overfitting. OpenCV methods were used to filter out frames not containing characters. Characters were then detected using a pre-trained TensorFlow model. Following this, they performed scraping of the synopsis from fan sites and video metadata, with the hypothesis that if a character is not mentioned in an episode's synopsis, there will be no important scenes for that character in that episode. An audio track was then extracted from the clip, and the speech was transcribed using Deep Speech [Hannun et al., 2014]. Specific keywords were then searched to help determine the importance of a clip.

Team EURECOM proposed an approach based on zero-shot classification of named events. Faces were extracted and recognized using the Face Celebrity Recognition library [Lisena et al., 2021]. Faces were first detected using an MTCNN [7], and the FaceNet [8] model was applied to get face embeddings. A multi-class SVM classifier was then applied to output predictions. XML transcripts were then aligned with the given shot segmentation. A model was then constructed with the hypothesis that the least likely events are also likely to be the most interesting and should likely be included in the summary and weighting assigned to events as an inverse of their perceived likelihood to appear in soap operas. The weighting for events was further multiplied by the confidence score obtained from the zero-shot classifier. Finally, in order to extract informative scenes which should therefore be sufficiently long, the score per scene was further multiplied by the log of the length of the shot dialogue. The max score was then selected on all event labels, and the N shots with the highest score were kept.

Team NII_UIT proposed a framework to generate final summaries by casting the problem as a text-matching problem. For face score, MTCNN [Zhang et al., 2016] was used for face detection, VGGFace2 [Cao et al., 2018] for face representation, and cosine similarity for face matching. A co-appearance face score was then generated to model the co-appearance of the target character with another character. Text-matching scores were then generated using EastEnders Wiki content and in the sub-task using the provided questions. An Event Detection model was trained using EfficientNet [Tan and Le, 2019] for the detection of major life events in image data. An importance score was then calculated by combining all of the above scores. They then applied the Knapsack Multiple Constraints problem for shot selection due to the maximum number of shots and maximum summary length constraints.

Conclusions

This was the second year of the Video Summarization task. Teams were asked to produce summaries of the major life events of five target characters from within a specified time frame of the BBC Eastenders series. For the main task content questions were not known to teams in advance. For the sub-task teams had prior knowledge of the content questions on which summaries would be evaluated. The major challenges of this task were to locate only shots for the target queries and to identify those shots most likely to have been considered major life events.

[7] https://github.com/ipazc/mtcnn

[8] https://github.com/davidsandberg/facenet

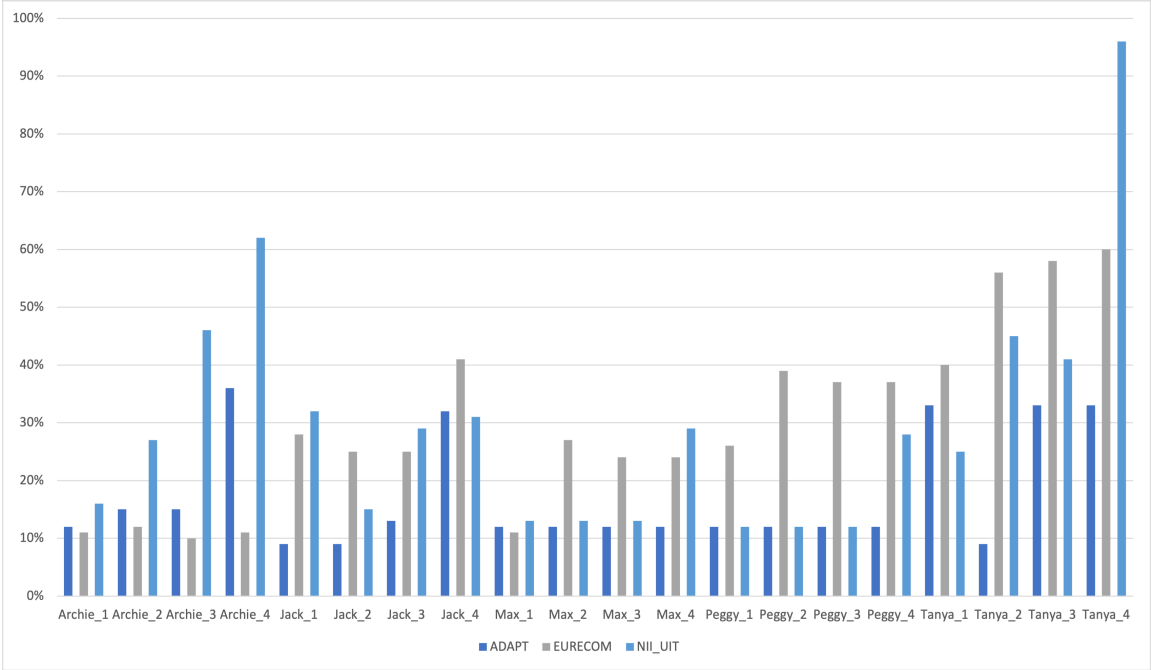

Figure 66: VSUM Sub Task: Individual scores

There were a total of 3 finishing teams out of 11 participating teams in this year's task. All 3 finishing teams submitted notebook papers and presented their approaches at the TRECVID workshop. This is to be the final year of the existing VSUM task using the current EastEnders dataset and the task will be incorporated into a new Movie Summarization (MSUM) task which will make use of a new full movie dataset.

4 Summing up and moving on

In this overview paper to TRECVID 2021, 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 [TV21Pubs, 2021]. Finally, we are looking forward to continuing a new evaluation cycle in 2022 after refining the current tasks and introducing any potential new tasks.

5 Authors' note

TRECVID would not have happened in 2021 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 license 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 Dublin City University (12/RC/2289) and grant number 13/RC/2106 (ADAPT Centre for Digital Content Technology, `www.adaptcentre.ie`) at Trinity College Dublin. We would like to thank Tim Finin and Lushan Han of University of Maryland, Baltimore County for providing access to the semantic similarity metric. Finally, the TRECVID team at NIST would like to thank all external coordinators for their efforts across the different tasks they helped to coordinate.

References

[kag, 2013] (2013). Challenges in representation learning: Facial expression recognition challenge.

[Abadi et al., 2016] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

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

[Anderson et al., 2018] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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

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

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

[Barrault et al., 2020] Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

[Barsoum et al., 2016] Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283.

[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.

[Bradski, 2000] Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123.

[Cai and Vasconcelos, 2018] Cai, Z. and Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162.

[Cao et al., 2018] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.

[Chao et al., 2018] Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., and Deng, J. (2018). Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE.

[Chollet et al., 2018] Chollet, F. et al. (2018). Keras: The python deep learning library. *Astrophysics Source Code Library*, pages ascl–1806.

[Deng et al., 2019a] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019a). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

[Deng et al., 2019b] Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., and Zafeiriou, S. (2019b). Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*.

[Feichtenhofer et al., 2019] Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

[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.

[Greenberg et al., 2020] Greenberg, C. S., Mason, L. P., Sadjadi, S. O., and Reynolds, D. A. (2020). Two decades of speaker recognition evaluation at the national institute of standards and technology. *Computer Speech & Language*, 60:101032.

[Guo et al., 2019] Guo, X., Li, S., Yu, J., Zhang, J., Ma, J., Ma, L., Liu, W., and Ling, H. (2019). Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*.

[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.

[Hannun et al., 2014] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[Karu and Jain, 1996] Karu, K. and Jain, A. K. (1996). Fingerprint classification. *Pattern recognition*, 29(3):389–404.

[Kay et al., 2017] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

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

[Krishna et al., 2017] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

[Le et al., 2014] Le, V.-B., Lamel, L., Messaoudi, A., Hartmann, W., Gauvain, J.-L., Woehrling, C., Despres, J., and Roy, A. (2014). Developing stt and

kws systems using limited language resources. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[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`.

[Li et al., 2020] Li, A., Thotakuri, M., Ross, D. A., Carreira, J., Vostrikov, A., and Zisserman, A. (2020). The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

[Lisena et al., 2021] Lisena, P., Laaksonen, J., Troncy, R., et al. (2021). Facerec: an interactive framework for face recognition in video archives. In *2nd International Workshop on Data-driven Personalisation of Television (DataTV) Collocated with the ACM International Conference on Interactive Media Experiences (IMX 2021)*, pages 21–23.

[Liu et al., 2019] Liu, J., Strohschein, D., Samsi, S., and Weinert, A. (2019). Large scale organization and inference of an imagery dataset for public safety. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–6.

[Lucey et al., 2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE.

[Luo et al., 2018] Luo, R., Price, B., Cohen, S., and Shakhnarovich, G. (2018). Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

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

[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.

[Mille et al., 2020] Mille, S., Belz, A., Bohnet, B., Castro Ferreira, T., Graham, Y., and Wanner, L. (2020). The third multilingual surface realisation shared task (SR'20): Overview and evaluation results. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 1–20, Barcelona, Spain (Online). Association for Computational Linguistics.

[Monfort et al., 2019] Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al. (2019). Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508.

[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.

[Popović, 2015] Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

[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.

[Schroff et al., 2015] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[Shao et al., 2019] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., and Sun, J. (2019). Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Smaira et al., 2020] Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., and Zisserman, A. (2020). A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*.

[Tan and Le, 2019] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

[TV21Pubs, 2021] TV21Pubs (2021). `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.21.org.html`.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[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.

[Wojke et al., 2017] Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE.

[Yang et al., 2021] Yang, J., Liang, C., Niu, Y., Huang, B., and Wang, Z. (2021). A spatio-temporal identity verification method for person-action instance search in movies. *arXiv preprint arXiv:2111.00228*.

[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.

[Zhang et al., 2016] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Table 17: Video Summarization Queries and Specifics - Main Task

Team_Run_Query	Tempo	Contextuality	Redundancy	Q1	Q2	Q3	Q4	Q5	Score
ADAPT_1_Archie	5	3	2	Yes	No	Yes	No	Yes	62%
ADAPT_2_Archie	6	5	4	Yes	Yes	Yes	No	Yes	79%
ADAPT_3_Archie	4	6	4	No	Yes	No	No	No	30%
ADAPT_4_Archie	5	5	3	No	Yes	No	No	No	31%
EURECOM_1_Archie	3	4	5	No	Yes	No	No	No	26%
EURECOM_2_Archie	3	4	4	Yes	Yes	No	No	Yes	59%
EURECOM_3_Archie	3	5	5	Yes	Yes	No	No	Yes	59%
EURECOM_4_Archie	3	5	4	Yes	Yes	No	No	Yes	60%
NII_UIT_1_Archie	3	2	7	No	No	No	No	No	6%
NII_UIT_2_Archie	3	3	5	No	No	No	No	No	9%
NII_UIT_3_Archie	4	3	4	No	No	No	Yes	No	27%
NII_UIT_4_Archie	2	2	6	No	No	No	No	No	6%
ADAPT_1_Jack	6	5	2	No	No	No	No	No	17%
ADAPT_2_Jack	6	4	2	No	No	No	No	No	16%
ADAPT_3_Jack	5	5	4	No	No	No	Yes	No	30%
ADAPT_4_Jack	4	5	3	No	No	No	No	No	14%
EURECOM_1_Jack	6	3	3	No	No	No	No	No	14%
EURECOM_2_Jack	5	5	4	No	No	No	No	Yes	30%
EURECOM_3_Jack	4	4	2	No	No	No	No	Yes	30%
EURECOM_4_Jack	5	4	2	No	No	No	No	Yes	31%
NII_UIT_1_Jack	2	2	5	No	No	No	No	No	7%
NII_UIT_2_Jack	3	2	6	No	No	No	No	No	7%
NII_UIT_3_Jack	4	3	5	No	No	No	Yes	No	26%
NII_UIT_4_Jack	6	4	4	No	No	No	Yes	No	30%
ADAPT_1_Max	3	3	3	No	Yes	No	No	No	27%
ADAPT_2_Max	2	3	5	No	No	No	No	No	8%
ADAPT_3_Max	2	4	4	No	No	No	No	No	10%
ADAPT_4_Max	3	3	4	No	No	No	No	No	10%
EURECOM_1_Max	4	3	3	No	No	No	No	No	12%
EURECOM_2_Max	4	3	3	No	No	Yes	No	No	28%
EURECOM_3_Max	4	3	3	No	Yes	Yes	No	No	44%
EURECOM_4_Max	4	3	4	No	Yes	Yes	No	No	43%
NII_UIT_1_Max	3	3	4	No	No	No	No	No	10%
NII_UIT_2_Max	3	3	4	No	No	No	No	No	10%
NII_UIT_3_Max	3	3	4	No	Yes	No	No	No	26%
NII_UIT_4_Max	3	3	4	No	Yes	No	No	No	26%
ADAPT_1_Peggy	2	3	3	No	Yes	No	No	No	26%
ADAPT_2_Peggy	2	3	3	No	Yes	No	No	No	26%
ADAPT_3_Peggy	2	3	4	No	No	Yes	No	No	25%
ADAPT_4_Peggy	2	3	3	No	No	Yes	No	Yes	42%
EURECOM_1_Peggy	3	3	3	No	No	No	No	No	11%
EURECOM_2_Peggy	3	3	4	No	No	No	No	No	10%
EURECOM_3_Peggy	3	3	5	No	No	No	No	No	9%
EURECOM_4_Peggy	3	3	4	No	No	No	No	No	10%
NII_UIT_1_Peggy	2	3	3	No	No	No	No	No	10%
NII_UIT_2_Peggy	3	3	4	No	No	No	No	No	10%
NII_UIT_3_Peggy	3	3	4	No	No	Yes	No	No	26%
NII_UIT_4_Peggy	2	3	4	No	No	No	No	No	9%

Table 18: Video Summarization Queries and Specifics - Main Task Continued

Team_Run_Query	Tempo	Contextuality	Redundancy	Q1	Q2	Q3	Q4	Q5	Score
ADAPT_1_Tanya	3	2	5	No	Yes	No	No	No	24%
ADAPT_2_Tanya	4	4	5	No	No	No	Yes	Yes	43%
ADAPT_3_Tanya	4	4	4	No	Yes	Yes	No	No	44%
ADAPT_4_Tanya	3	4	5	No	Yes	No	No	Yes	42%
EURECOM_1_Tanya	4	2	6	Yes	No	No	No	No	24%
EURECOM_2_Tanya	2	4	5	Yes	No	No	No	No	25%
EURECOM_3_Tanya	2	2	6	Yes	No	No	No	No	22%
EURECOM_4_Tanya	5	4	5	Yes	Yes	No	No	No	44%
NII_UIT_1_Tanya	2	1	7	No	No	No	No	No	4%
NII_UIT_2_Tanya	3	3	5	No	Yes	No	No	No	25%
NII_UIT_3_Tanya	4	4	5	No	Yes	Yes	No	No	43%
NII_UIT_4_Tanya	4	4	5	No	Yes	Yes	No	No	43%

Table 19: Video Summarization Queries and Specifics - Sub Task

Team_Run_Query	Tempo	Contextuality	Redundancy	Q1	Q2	Q3	Q4	Q5	Score
ADAPT_1_Archie	6	3	5	No	No	No	No	No	12%
ADAPT_2_Archie	6	4	3	No	No	No	No	No	15%
ADAPT_3_Archie	6	4	3	No	No	No	No	No	15%
ADAPT_4_Archie	7	5	2	No	Yes	No	No	No	36%
EURECOM_1_Archie	3	4	4	No	No	No	No	No	11%
EURECOM_2_Archie	4	5	5	No	No	No	No	No	12%
EURECOM_3_Archie	3	4	5	No	No	No	No	No	10%
EURECOM_4_Archie	4	5	6	No	No	No	No	No	11%
NII_UIT_1_Archie	6	5	3	No	No	No	No	No	16%
NII_UIT_2_Archie	4	4	5	No	Yes	No	No	No	27%
NII_UIT_3_Archie	4	5	3	Yes	No	No	No	Yes	46%
NII_UIT_4_Archie	5	5	4	Yes	Yes	No	No	Yes	62%
ADAPT_1_Jack	2	3	4	No	No	No	No	No	9%
ADAPT_2_Jack	2	3	4	No	No	No	No	No	9%
ADAPT_3_Jack	4	4	3	No	No	No	No	No	13%
ADAPT_4_Jack	5	5	2	No	No	No	Yes	No	32%
EURECOM_1_Jack	4	4	4	No	No	No	No	Yes	28%
EURECOM_2_Jack	3	3	5	No	No	No	No	Yes	25%
EURECOM_3_Jack	3	3	5	No	No	No	No	Yes	25%
EURECOM_4_Jack	3	4	6	No	No	No	Yes	Yes	41%
NII_UIT_1_Jack	5	5	2	No	No	No	Yes	No	32%
NII_UIT_2_Jack	5	5	3	No	No	No	No	No	15%
NII_UIT_3_Jack	5	5	5	No	No	No	Yes	No	29%
NII_UIT_4_Jack	5	5	3	No	No	No	Yes	No	31%

Table 20: Video Summarization Queries and Specifics - Sub Task Continued

Team_Run_Query	Tempo	Contextuality	Redundancy	Q1	Q2	Q3	Q4	Q5	Score
ADAPT_1_Max	5	3	4	No	No	No	No	No	12%
ADAPT_2_Max	5	3	4	No	No	No	No	No	12%
ADAPT_3_Max	5	3	4	No	No	No	No	No	12%
ADAPT_4_Max	5	3	4	No	No	No	No	No	12%
EURECOM_1_Max	5	3	5	No	No	No	No	No	11%
EURECOM_2_Max	5	3	5	No	No	Yes	No	No	27%
EURECOM_3_Max	3	3	6	No	No	Yes	No	No	24%
EURECOM_4_Max	3	3	6	No	No	Yes	No	No	24%
NII_UIT_1_Max	4	4	3	No	No	No	No	No	13%
NII_UIT_2_Max	4	4	3	No	No	No	No	No	13%
NII_UIT_3_Max	4	4	3	No	No	No	No	No	13%
NII_UIT_4_Max	4	4	3	No	Yes	No	No	No	29%
ADAPT_1_Peggy	4	4	4	No	No	No	No	No	12%
ADAPT_2_Peggy	4	4	4	No	No	No	No	No	12%
ADAPT_3_Peggy	4	4	4	No	No	No	No	No	12%
ADAPT_4_Peggy	4	4	4	No	No	No	No	No	12%
EURECOM_1_Peggy	4	4	6	No	No	Yes	No	No	26%
EURECOM_2_Peggy	2	4	7	No	No	Yes	No	Yes	39%
EURECOM_3_Peggy	1	3	7	No	No	Yes	No	Yes	37%
EURECOM_4_Peggy	1	3	7	No	No	Yes	No	Yes	37%
NII_UIT_1_Peggy	4	4	4	No	No	No	No	No	12%
NII_UIT_2_Peggy	4	4	4	No	No	No	No	No	12%
NII_UIT_3_Peggy	4	4	4	No	No	No	No	No	12%
NII_UIT_4_Peggy	4	4	4	No	No	No	No	Yes	28%
ADAPT_1_Tanya	7	4	2	No	No	Yes	No	No	33%
ADAPT_2_Tanya	7	1	7	No	No	No	No	No	9%
ADAPT_3_Tanya	7	4	2	No	No	No	No	Yes	33%
ADAPT_4_Tanya	7	4	2	No	No	No	No	Yes	33%
EURECOM_1_Tanya	2	4	6	No	Yes	No	No	Yes	40%
EURECOM_2_Tanya	2	4	6	No	Yes	Yes	No	Yes	56%
EURECOM_3_Tanya	2	6	6	No	Yes	Yes	No	Yes	58%
EURECOM_4_Tanya	5	5	6	No	Yes	Yes	No	Yes	60%
NII_UIT_1_Tanya	5	3	7	No	No	Yes	No	No	25%
NII_UIT_2_Tanya	5	5	5	No	Yes	Yes	No	No	45%
NII_UIT_3_Tanya	4	3	6	Yes	No	Yes	No	No	41%
NII_UIT_4_Tanya	6	6	3	Yes	Yes	Yes	Yes	Yes	96%

A Ad-hoc query topics - 20 unique

661 Find shots of a hang glider floating in the sky on a sunny day
662 Find shots of a woman wearing sleeveless top
663 Find shots of a person with a tattoo on their arm
664 Find shots of city street where ground is covered by snow
665 Find shots of an adult person wearing a backpack and walking on a sidewalk
666 Find shots of a man wearing a blue jacket
667 Find shots of a person looking at themselves in a mirror
668 Find shots of a person wearing an apron indoors
669 Find shots of a woman holding a book
670 Find shots of a person painting on a canvas
671 Find shots of a man behind a pub bar or club bar
672 Find shots of a person wearing a cap backwards
673 Find shots of a man pointing with his finger
674 Find shots of a parachutist descending towards a field on the ground in the daytime
675 Find shots of two or more ducks swimming in a pond
676 Find shots of a white dog
677 Find shots of two boxers in a ring
678 Find shots of a man sitting on a barber chair in a shop
679 Find shots of a ladder with less than 6 steps
680 Find shots of a bow tie

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

9319 Find Max sitting on couch

9320 Find Stacey sitting on couch

9321 Find Peggy sitting on couch
9322 Find Stacey Holding drinking glass/cup/bottle/can
9323 Find Bradley Holding drinking glass/cup/bottle/can
9324 Find Shirley Holding drinking glass/cup/bottle/can
9325 Find Bradley holding a phone / handset including talking on phone
9326 Find Shirley holding a phone / handset including talking on phone
9327 Find Peggy holding a phone / handset including talking on phone
9328 Find Max Holding/carrying a bag/purse/backpack
9329 Find Peggy Holding/carrying a bag/purse/backpack
9330 Find Pat Holding paper including photos/envelope,notebooks, magazines, etc
9331 Find Shirley Holding paper including photos/envelope,notebooks, magazines, etc
9332 Find Peggy Holding paper including photos/envelope,notebooks, magazines, etc
9333 Find Max Kissing
9334 Find Stacey Kissing
9335 Find Bradley opening a door and entering a room/building
9336 Find Pat opening a door and entering a room/building
9337 Find Max Holding cloth including jackets, coats, kitchen towels, cleaning towels, etc
9338 Find Stacey Holding cloth including jackets, coats, kitchen towels, cleaning towels, etc

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