TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking

George Awad {gawad@nist.gov} Jonathan Fiscus {jfiscus@nist.gov}
David Joy {david.joy@nist.gov} Martial Michel {martial.michel@nist.gov}

Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

Alan F. Smeaton {alan.smeaton@dcu.ie}
Insight Centre for Data Analytics, Dublin City University

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

Maria Eskevich {M.Eskevich@let.ru.nl}
Radboud University, Netherlands

Robin Aly {r.aly@utwente.nl} Roeland Ordelman {roeland.ordelman@utwente.nl}
University of Twente, Netherlands

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

Benoit Huet {benoit.huet@eurecom.fr}
EURECOM, Sophia Antipolis, France

Martha Larson {m.a.larson@tudelft.nl}
Radboud University;Delft University of Technology, Netherlands

November 8, 2016

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2016 was a TREC-style video analysis and retrieval evaluation, the goal of which remains to promote progress in content-based exploitation of digital video via open, metrics-based evaluation. Over the last fourteen years this effort has yielded a better understanding of how systems can effectively accomplish such processing and how one can reliably benchmark their performance. TRECVID is funded by the NIST and other US government agencies. In addition, many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2016 represented a continuation of five tasks from 2015, the replacement of the semantic in-

dexing task by a new Ad-hoc video search task and a new pilot video to text description task. 39 teams (see Table 1) from various research organizations worldwide completed one or more of seven tasks:

1. Ad-hoc Video Search
2. Instance search
3. Multimedia event detection
4. Surveillance event detection
5. Video hyperlinking
6. Concept localization
7. Video to Text Description (pilot task)

About 600 new hours of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC.3) were used for Ad-hoc Video Search. Unlike previously used professionally edited broadcast news and educational programming, the IACC videos reflect a wide variety of content, style, and source device determined only by the self-selected donors. About 464 hours of BBC EastEnders video was reused for the instance search task. Approximately 2.2 million I-frame images were used for testing in the localization task and 3,288 hours of blip.tv videos were used for the video Hyperlinking task. For the surveillance event detection task, 11 hours of airport surveillance video was used, and almost a total of 4 738 hours from the HAVIC collection of Internet videos in addition to a subset of Yahoo YFC100M videos were used in the multimedia event detection task. A new video to text pilot task was proposed this year. The task used about 2000 Twitter vine videos collected through the online API public stream.

Ad-hoc search, instance search, multimedia event detection, and localization results were judged by NIST assessors. The video hyperlinking results were assessed by Amazon Mturk workers after initial manual check for sanity while the anchors were chosen by media-researchers. Surveillance event detection was scored by NIST using ground truth created by NIST through manual adjudication of test system output. Finally, the new pilot task was annotated by collaboration with TUC Chemnitz group of Dr. Marc Ritter.

This paper is an introduction to the evaluation framework — the tasks, data, and measures for the workshop. For detailed information about the approaches and results, the reader should see the various site reports and the results pages available at the workshop proceeding online page.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities,materials, or equipment are necessarily the best available for the purpose.

2 Data

2.1 Video

BBC EastEnders video

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

Blip10000 Hyperlinking video

Blip10000 data set consists of 14,838 videos for a total of 3,288 hours from blip.tv. The videos cover a broad range of topics and styles. It has automatic speech recognition transcripts provided by LIMSI; user-contributed metadata and shot boundaries provided by TU Berlin. Also, video concepts based on the MediaMill MED Caffe models are provided by EURECOM.

Internet Archive Creative Commons (IACC.3) video

4593 Internet Archive videos (144 GB, 600 hours) with Creative Commons licenses in MPEG-4/H.264 format with duration ranging from 6.5 to 9.5 min and a mean duration of almost 7.8 min. Most videos will have some metadata provided by the donor available e.g., title, keywords, and description.

Approximately 1200 h of IACC.1 and IACC.2 videos used between 2010 to 2015 were available for system development.

As in the past, the Computer Science Laboratory for Mechanics and Engineering Sciences (LIMSI) and Vocapia Research provided automatic speech recognition for the English speech in the IACC.3 videos.

iLIDS Multiple Camera Tracking Data

The iLIDS Multiple Camera Tracking data consisted of ≈150 h of indoor airport surveillance video col-

Table 1: Participants and tasks

Task						Location	TeamID	Participants
−−	*HL*	−−	*MD*	*SD*	AV	NAm+Asia	INF	Beijing U. of Posts and Tele.;U. Autonoma de Madrid; Shandong U.; Xian JiaoTong U.
−−	−−	−−	*MD*	**	–	Asia	BIT_MCIS	Beijing Inst. of Tech., Media Computing and Intelligent System Lab.
−−	**	−−	*MD*	−−	AV	Asia	VIREO	City U. of Hong Kong
−−	*HL*	−−	−−	−−	–	Eur	IRISA	CNRS, IRISA, INSA, Universite de Rennes 1
−−	−−	**	−−	−−	AV	Asia	UEC	U. of Electro-Communications, Tokyo
IN	−−	−−	−−	−−	–	Asia	U_TK	U. of Tokushima
IN	−−	−−	−−	−−	–	Aus	UQMG	U. of Queensland - DKE Group of ITEE
IN	−−	−−	−−	**	**	Eur	insightdcu	Dublin City U.; Polytechnic U. of Catalonia
−−	−−	−−	*MD*	−−	–	NAm	Etter	Etter Solutions
−−	*HL*	−−	−−	−−	AV	Eur	EURECOM	EURECOM
−−	−−	−−	−−	−−	AV	NAm	FIU_UM	Florida International U.; U. of Miami
−−	*HL*	−−	−−	−−	–	NAm	FXPAL	FX PALO ALTO LABORATORY, INC
−−	−−	−−	−−	*SD*	–	Asia	HRI	Hikvision Research Institute
IN	−−	−−	*MD*	*SD*	AV	Eur	ITI_CERTH	Centre for Research and Tech. Hellas
IN	−−	−−	−−	−−	**	Eur	IRIM	EURECOM;LABRI;LIG;LIP6;LISTIC
IN	−−	−−	−−	−−	**	Eur	JRS	JOANNEUM RESEARCH
−−	−−	−−	−−	−−	AV	Eur	ITEC_UNIKLU	Klagenfurt University
−−	−−	−−	−−	−−	AV	Eur+Asia	kobe_nict_siegen	Kobe U.; Natl. Inst. of Inf. and Comm. Tech.;U. of Siegen
−−	−−	−−	*MD*	−−	–	Asia	KoreaUnivISPL	Korea U.
IN	−−	−−	*MD*	−−	–	NAm+Asia	PKU_MI	Peking U.; Rutgers U.
IN	−−	**	*MD*	*SD*	**	Asia	BUPT_MCPRL	Beijing U. of Posts and Telecommunications
IN	**	*LO*	*MD*	*SD*	AV	Asia	NII_Hitachi_UIT	Natl. Inst. of Inf.;Hitachi; U. of Inf. Tech.
IN	−−	−−	−−	−−	–	Asia	WHU_NERCMS	Natl. Eng. Research Center for Multimedia Software, Wuhan U.
−−	−−	−−	*MD*	**	–	Asia	nttfudan	NTT Media Intelligence Laboratories; Fudan U.
IN	**	**	**	**	**	NAm+Asia	PKU_ICST	Peking U.
−−	*HL*	−−	−−	−−	–	Eur	EURECOM_POLITO	Politecnico di Torino Eurecom
−−	−−	−−	−−	*SD*	–	Aus	WARD	U. of Queensland
IN	−−	−−	−−	−−	–	Asia	SIAT_MMLAB	Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
−−	−−	−−	−−	*SD*	–	Asia	SeuGraph	Southeast U. Computer Graphics Lab
IN	−−	−−	−−	−−	–	Asia	TRIMPS_SARI	Third Research Inst., Ministry of Public Security; Chinese Academy of Sciences
−−	−−	*LO*	*MD*	−−	**	Asia	TokyoTech	Tokyo Inst. of Tech.
IN	−−	−−	−−	−−	–	Eur	TUC	TU Chemnitz - Junior Professorship Media Computing - Chair Media Informatics
−−	−−	−−	−−	−−	AV	Eur	IMOTION	U. of Basel; U. of Mons; Koc U.
**	−−	**	*MD*	−−	AV	Eur	MediaMill	U. of Amsterdam
−−	−−	*LO*	−−	−−	–	Aus	UTS_CMU_D2DCRC	U. of Technology, Sydney D2DCRC
−−	−−	−−	−−	−−	AV	Eur	vitrivr	U. of Basel
−−	**	**	**	−−	AV	Asia	Waseda	Waseda U.
−−	**	−−	−−	*SD*	–	Asia	IIP_WHU	Wuhan U.

Task legend. IN:instance search; MD:multimedia event detection; HL:Hyperlinking; LO:Localization; SD:surveillance event detection; AV:Ad-hoc; −−:no run planned; ∗∗:planned but not submitted

lected in a busy airport environment by the United Kingdom (UK) Center for Applied Science and Technology (CAST). The dataset utilized 5, frame-synchronized cameras.

The training videos consisted of the ≈100 h of data used for SED 2008 evaluation. The evaluation videos consisted of the same additional ≈50 hours of data from Imagery Library for Intelligent Detection System's (iLIDS) multiple camera tracking scenario data used for the 2009 - 2013 evaluations [UKHO-CPNI, 2009] .

Heterogeneous Audio Visual Internet (HAVIC) Corpus

The HAVIC Corpus [Strassel et al., 2012] is a large corpus of Internet multimedia files collected by the Linguistic Data Consortium and distributed as MPEG-4 (MPEG-4, 2010) formatted files containing H.264 (H.264, 2010) encoded video and MPEG-4 Advanced Audio Coding (AAC) (AAC, 2010) encoded audio.

The HAVIC systems used the same, LDC-provided development materials as in 2013 but teams were also able to use site-internal resources. Approximately 98003 clips with total duration of 3712.89 hours and total size of 1300G were reused from the MED15 task as an evaluation collection.

Table 2: Participants who did not submit any runs

Task						Location	TeamID	Participants
IN	*HL*	*LO*	*MD*	*SD*	*AV*			
−−	−−	**	−−	−−	**	Eur	PicSOM	Aalto U.
−−	−−	−−	−−	−−	**	Asia	ABZOOBA	Abzooba Inc. India
−−	−−	−−	**	−−	−−	NAm	fork	Arizona state U.
−−	−−	**	**	−−	**	Asia	SamHMS	Beijing Samsung Telecom R&D Center
−−	−−	**	−−	**	−−	NAm	BCTS	Brain Corporation Technical Services
−−	−−	−−	−−	−−	**	NAm	CCNY	City U. of New York; Graduate Center, City U. of New York; NVIDIA Research
**	−−	−−	−−	−−	−−	Asia	CVARL_WU	Computing Center of Computer School at Wuhan U.
**	−−	**	**	**	**	Eur	ADVICE	BASKENT U.
−−	−−	−−	**	−−	−−	Eur	HEU008	Harbin Engineering U.
−−	−−	**	**	−−	−−	Asia	hulustar	HULU LLC
−−	−−	−−	−−	**	−−	Asia	NP	IIT Hyderabad
−−	−−	**	**	−−	−−	Eur	INRIA_STARS	INRIA
−−	−−	−−	−−	−−	**	Asia	TAM	Intel
**	−−	−−	**	−−	**	Asia	Ravi	JNTUK
−−	−−	**	−−	−−	**	Eur	LIG	Laboratoire d'Informatique de Grenoble
**	−−	**	−−	−−	−−	Eur	MetuMedia	Middle East Technical U. Department of Electrical/Electronics Engineering
**	−−	−−	−−	**	−−	Asia	Mitsubishi_Electric	Mitsubishi Electric Corporation
−−	−−	−−	**	**	−−	Asia	MI_TJU	Multimedia Institute, Tianjin U.
−−	−−	−−	**	−−	−−	Asia	nus_action	National U. of Singapore
−−	−−	−−	**	−−	−−	NAm	NEU_MITLL	Northeastern U. and MIT Lincoln Laboratory
**	−−	−−	−−	−−	−−	Asia	NTT	NTT Communication Science Laboratories; NTT Media Intelligence Laboratories
**	**	−−	**	−−	−−	SAm	ORAND	ORAND S.A. Chile
−−	−−	−−	−−	**	−−	NAm	QUPROR	Private Research
**	−−	**	**	**	**	Asia	QUT	Qatar U.
−−	**	**	−−	−−	**	Asia	REGIMVID	REGIM; U. of Sfax
**	−−	−−	**	−−	−−	Asia	saricas	Shanghai Advanced Research Institute, Chinese Academy of Sciences
−−	−−	−−	−−	**	−−	Asia	sjtu_icl	Shanghai Jiao Tong U.
−−	−−	−−	−−	**	−−	Asia	zy_scu	Sichuan U.
**	**	**	**	**	**	Asia	Trimps	The Third Research Institute of the Ministry of Public Security
**	**	**	**	**	**	Asia	HAWKEYE	Tsinghua U.
**	−−	−−	−−	−−	−−	Asia	THSS_IMMG	Tsinghua U. School of Software
**	**	**	**	−−	−−	Eur	TUZ	TUBITAK UZAY
**	−−	−−	−−	−−	−−	Asia	BMC_UESTC	U. of Electronic Science and Technology of China
**	−−	−−	−−	−−	−−	Eur+Asia	Sheffield_UETLahore	U. of Sheffield; U. of Engineering & Technology
−−	−−	−−	**	−−	−−	Eur+Asia	trento_tokyo_univ	U. of Trento
−−	−−	−−	−−	**	−−	Eur+Asia	UniKent	U. of Kent
−−	−−	−−	**	−−	**	Asia	zjgsucvg	Zhejiang Gongshang U.

Task legend. IN:instance search; MD:multimedia event detection; HL:Hyperlinking; LO:Localization; SD:surveillance event detection; AV:Ad-hoc; −−:no run planned; **:planned but not submitted

Yahoo Flickr Creative Commons 100M dataset (YFCC100M)

The YFCC100M dataset [Thomee et al., 2016] is a large collection of images and videos available on Yahoo! Flickr. All photos and videos listed in the collection are licensed under one of the Creative Commons copyright licenses. The YFCC100M dataset is comprised of 99.3 million images and 0.7 million videos. Only a subset of the YFCC100M videos (100,000 Clips with total duration of 1025.06 hours and total size of 352G) are used for evaluation.

3 Ad-hoc Video Search

The previous Semantic Indexing task run from 2010-2015 addressed the problem of automatic assignment of predefined semantic tags representing visual or multimodal concepts to video segments. More and more concepts were trained and developed over the course of those six years. However, testing individual visual concepts are not very realistic in a real-world setting as an average user would more likely be interested in searching for those concepts in a particular context or in a combined form. This year a new Ad-hoc search task was introduced to model the end user video search use-case, who is looking for segments of video containing persons,objects,activities,locations, etc. and combinations of the former.

It was coordinated by NIST and by Georges Quénot at the Laboratoire d'Informatique de Grenoble.

The Ad-hoc video search task was as follows. Given a standard set of shot boundaries for the IACC.3 test collection and a list of 30 Ad-hoc queries, participants were asked to return for each query, at most the top 1 000 video shots from the standard set, ranked according to the highest possibility of containing the target query. The presence of each query was assumed to be binary, i.e., it was either present or absent in the given standard video shot.

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

In 2016 the task again supported experiments using the "no annotation" version of the tasks: the idea is to promote the development of methods that permit the indexing of concepts in video shots using only data from the web or archives without the need of additional annotations. The training data could for instance consist of images or videos retrieved by a general purpose search engine (e.g. Google) using only the query definition with only automatic processing of the returned images or videos. This was implemented by adding the categories of "E" and "F" for the training types besides A and D:[1]

- A - used only IACC training data
- D - used any other training data
- E - used only training data collected automatically using only the official query textual description
- F - used only training data collected automatically using a query built manually from the given official query textual description

This means that even just the use of something like a face detector that was trained on non-IACC training data would disqualify the run as type A.

Two main submission types will be accepted:

- Fully automatic runs (no human input in the loop): System takes a query as input and produced result without any human intervention.
- Manually-assisted runs: where a human can formulate the initial query based on topic and query interface, not on knowledge of collection or search results. Then system takes the formulated query as input and produces result without further human intervention.

TRECVID evaluated 30 query topics listed in Appendix A.

Work at Northeastern University [Yilmaz and Aslam, 2006] has resulted in methods for estimating standard system performance measures using relatively small samples of the usual judgment sets so that larger numbers of features can be evaluated using the same amount of judging effort. Tests on past data showed the new measure (inferred average precision) to be a good estimator of average precision [Over et al., 2006]. This year mean extended inferred average precision (mean xinfAP) was used which permits sampling density to vary [Yilmaz et al., 2008]. This allowed the evaluation to be more sensitive to shots returned below the lowest rank (≈100) previously pooled and judged. It also allowed adjustment of the sampling density to be greater among the highest ranked items that contribute more average precision than those ranked lower.

3.1 Data

The IACC.3 collection was used for testing. It contained 335 944 shots.

3.2 Evaluation

Each group was allowed to submit up to 4 prioritized main runs and two additional if they are "no annotation" runs. In fact 13 groups submitted a total of 52 runs, from which 22 runs were Manually-assisted and 30 were fully automatic runs.

For each query topic, pools were created and randomly sampled as follows. The top pool sampled 100

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

% of shots ranked 1-200 across all submissions. The bottom pool sampled 11.1 % of ranked 201-1000 shots and not already included in a pool. Human judges (assessors) were presented with the pools - one assessor per concept - and they judged each shot by watching the associated video and listening to the audio. Once the assessor completed judging for a topic, her or she was asked to rejudge all clips submitted by at least 10 runs at ranks 1 to 200. In all, 187 918 shots were judged while 371 376 shots fell into the unjudged part of the overall samples.

3.3 Measures

The *sample_eval* software, a tool implementing xinfAP, was used to calculate inferred recall, inferred precision, inferred average precision, etc., for each result, given the sampling plan and a submitted run. Since all runs provided results for all evaluated topics, runs can be compared in terms of the mean inferred average precision across all evaluated query topics. The results also provide some information about "within topic" performance.

3.4 Results

Readers should see the online proceedings for individual team's performance and runs.

4 Instance search

An important need in many situations involving video collections (archive video search/reuse, personal video organization/search, surveillance, law enforcement, protection of brand/logo use) is to find more video segments of a certain specific person, object, or place, given one or more visual examples of the specific item. The instance search task seeks to address some of these needs. For the past six years (2010-2015) the instance search task has tested systems on retrieving specific instances of individual objects, persons and locations. This year systems were tested on a new query type, to retrieve specific persons in specific locations.

4.1 Data

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

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

4.2 System task

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

Each query consisted of a set of

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

Information about the use of the examples was reported by participants with each submission. The possible categories for use of examples were as follows:

A one or more provided images - no video used
E video examples (+ optionally image examples)

4.3 Topics

NIST viewed a sample of test videos and developed a list of recurring people, locations and the appearance of people at certain locations. In order to test the effect of persons or locations on the performance of

Table 3: Instance search pooling and judging statistics

Topic number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number relevant	% judged that were relevant
9159	45016	14915	33.1	280	5226	35	92	1.8
9160	44880	15726	35	140	2580	16.4	68	2.6
9161	44919	14866	33.1	120	2318	15.6	13	0.6
9162	44557	16898	37.9	200	4029	23.8	31	0.8
9163	44899	13997	31.2	180	2867	20.5	305	10.6
9164	45354	15231	33.6	460	8647	56.8	890	10.3
9165	45352	11726	25.9	360	4337	37	1169	27
9166	45377	13275	29.3	280	3841	28.9	849	22.1
9167	45420	10905	24	520	5669	52	1614	28.5
9168	45825	14533	31.7	200	3691	25.4	763	20.7
9169	45796	13118	28.6	440	5708	43.5	715	12.5
9170	45833	12968	28.3	200	2902	22.4	247	8.5
9171	45809	14152	30.9	520	7272	51.4	546	7.5
9172	45843	12827	28	340	4123	32.1	785	19
9173	45818	15837	34.6	360	6792	42.9	1135	16.7
9174	45817	15147	33.1	380	6213	41	428	6.9
9175	45787	14446	31.6	220	4097	28.4	99	2.4
9176	45835	16249	35.5	200	3581	22	231	6.5
9177	45732	15322	33.5	280	4786	31.2	321	6.7
9178	45887	14243	31	460	7217	50.7	896	12.4
9179	39734	13280	33.4	180	3047	22.9	49	1.6
9180	39733	12201	30.7	220	3462	28.4	144	4.2
9181	39256	14320	36.5	520	8504	59.4	574	6.7
9182	39221	11973	30.5	200	3152	26.3	134	4.3
9183	39207	13000	33.2	220	3507	27	116	3.3
9184	39786	13438	33.8	420	6379	47.5	1243	19.5
9185	39741	14009	35.3	220	3655	26.1	88	2.4
9186	39751	12827	32.3	180	3139	24.5	81	2.6
9187	39784	14885	37.4	140	2677	18	38	1.4
9188	39743	13271	33.4	220	3326	25.1	136	4.1

a given query, the topics tested target persons across the same locations. In total this year we asked systems to find 7 target persons across 5 target locations. 30 test queries (topics) were then created (Appendix B).

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

4.4 Evaluation, Measures

Each group was allowed to submit up to 4 runs (8 if submitting pairs that differ only in the sorts of examples used) and in fact 13 groups submitted 41 automatic and 7 interactive runs (using only the first 20 topics). Each interactive search was limited to 5 minutes.

The submissions were pooled and then divided into strata based on the rank of the result items. For a given topic, the submissions for that topic were judged by a NIST assessor who played each submitted shot and determined if the topic target was present. The assessor started with the highest ranked stratum and worked his/her way down until too few relevant shots were being found or time ran out. Table 3 presents information about the pooling and judging.

This task was treated as a form of search and

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

4.5 Results

Readers should see the online proceedings for individual team's performance and runs.

5 Multimedia event detection

The 2016 Multimedia Event Detection (MED) evaluation was the sixth evaluation of technologies that search multimedia video clips for complex events of interest to a user.

The focus of MED 15 was to make MED less costly to both participate in and administer. MED 16 continues that trend by replacing a portion of the test set with an equal number of videos from the Yahoo Flickr Creative Commons 100M data set (YFCC100M), which is new to MED this year. The YFCC100M dataset is more readily accessible and contains shorter duration videos than the HAVIC dataset.

The MED 16 evaluation protocol is identical to MED 15, with the following modifications:

- Replaced roughly half of the test set with a subset of the YFCC100M dataset videos.
- Introduced 10 new Ad-Hoc (AH) events.
- Scored both Pre-Specified (PS) and AH event sets using Inferred Mean Average Precision [Yilmaz et al., 2008], reference generated through pooled assessment.

A user searching for events, complex activities occurring at a specific place and time involving people interacting with other people and/or objects, in multimedia material may be interested in a wide variety of potential events. Since it is an intractable task to build special purpose detectors for each event a priori, a technology is needed that can take as input a human-centric definition of an event that developers (and eventually systems) can use to build a search query.

The events for MED were defined via an event kit which consisted of:

- An event name which was an mnemonic title for the event.
- An event definition which was a textual definition of the event.
- An event explication which was an expression of some event domain-specific knowledge needed by humans to understand the event definition.
- An evidential description which was a textual listing of the attributes that are indicative of an event instance. The evidential description provides a notion of some potential types of visual and acoustic evidence indicating the event's existence but it was not an exhaustive list nor was it to be interpreted as required evidence.
- A set of illustrative video examples containing either an instance of the event or content "related" to the event. The examples were illustrative in the sense they helped form the definition of the event but they did not demonstrate all the inherent variability or potential realizations.

Within the general area of finding instances of events, the evaluation included three styles of system operation. The first is for Pre-Specified event systems where knowledge of the event(s) was taken into account during generation of the metadata store for the test collection. This style of system has been tested in MED since 2010. The second style is the Ad-Hoc event task where the metadata store generation was completed before the events where revealed. This style of system was introduced in MED 2012. The third style is a variation of Ad-Hoc event detection with 15 minutes of human interaction to search the evaluation collection in order to build a better query. As with MED 15, no one participated in this task.

5.1 Data

A development and evaluation collection of Internet multimedia (i.e., video clips containing both audio and video streams) clips was provided to MED participants.

The HAVIC data, which was collected and distributed by the Linguistic Data Consortium, consists of publicly available, user-generated content posted to the various Internet video hosting sites. Instances of the events were collected by specifically searching for target events using text-based Internet search engines. All video data was reviewed to protect privacy,

Table 4: MED '16 Pre-Specified Events

— MED'12 event re-test
Attempting a bike trick
Cleaning an appliance
Dog show
Giving directions
Marriage proposal
Renovating a home
Rock climbing
Town hall meeting
Winning a race without a vehicle
Working on a metal crafts project
— MED'13 event re-test
Beekeeping
Wedding shower
Non-motorized vehicle repair
Fixing a musical instrument
Horse riding competition
Felling a tree
Parking a vehicle
Playing fetch
Tailgating
Tuning a musical instrument

Table 5: MED '16 Ad-Hoc Events

E051 - Camping
E052 - Crossing a Barrier
E053 - Opening a Package
E054 - Making a Sand Sculpture
E055 - Missing a Shot on a Net
E056 - Operating a Remote Controlled Vehicle
E057 - Playing a Board Game
E058 - Making a Snow Sculpture
E059 - Making a Beverage
E060 - Cheerleading

remove offensive material, etc., prior to inclusion in the corpus. Video clips were provided in MPEG-4 formatted files. The video was encoded to the H.264 standard. The audio was encoded using MPEG-4's Advanced Audio Coding (AAC) standard.

The YFCC100M data, collected and distributed by Yahoo!, consists of photos and videos licensed under one of the Creative Commons copyright licenses. While the entire YFCC100M dataset consists of 99.3 million images and 0.7 million videos, only a subset of 100,000 randomly selected videos were chosen for this years evaluation.

MED participants were provided the data as specified in the HAVIC and YFCC100M data sections of this paper. The MED '16 Pre-Specified event names are listed in Table 4, and Table 5 lists the MED '16 Ad-Hoc Events.

5.2 Evaluation

Sites submitted MED system outputs testing their systems on the following dimensions:

- Events: all 20 Pre-Specified events (PS15) and/or all 10 Ad-Hoc events (AH15).
- Interactivity: Human interaction with query refinement using the search collection.
- Test collection: either the MED16 Full Evaluation collection (MED16-EvalFull) or a 783 hour subset (MED16-EvalSub) collection.
- Query Conditions: 0 Ex (the event text and the 5,000-clip Event Background collection 'EventBG'), 10 Ex (the event text, EventBG, and 10 positive and 10 miss clips per event), 100 Ex (the event text, EventBG, and 100 positive and 50 miss clips per event. Only for the PS condition).
- Hardware Definition: Teams self-reported the size of their computation cluster as the closest match to the following three standards:
  - SML - Small cluster consisting of 100 CPU cores and 1,000 GPU cores
  - MED - Medium cluster consisting of 1,000 CPU cores and 10,000 GPU cores
  - LRG - Large cluster consisting of 3,000 CPU cores and 30,000 GPU cores

Full participation requires teams to submit both 10Ex, PS and AH systems.

For each event search, a system generated:

- A rank for each search clip in the evaluation collection: A value from 1 (best rank) to N representing the best ordering of clips for the event.

Rather than submitting detailed runtime measurements to document the computational resources, participants labeled their systems as the closest match to one of three cluster sizes: small, medium and large. (See above.)

Submission performance was computed using the Framework for Detection Evaluation (F4DE) toolkit.

5.3 Measures

System output was evaluated by how well the system retrieved and detected MED events in evaluation search video metadata. The determination of correct detection was at the clip level, i.e. systems provided a response for each clip in the evaluation search video set. Participants had to process each event independently in order to ensure each event could be tested independently.

The primary evaluation measure for performance was Inferred Mean Average Precision.

5.4 Results

Readers should see the online proceedings for individual team's performance and runs.

6 Surveillance event detection

The 2016 Surveillance Event Detection (SED) evaluation was the ninth evaluation focused on event detection in the surveillance video domain. The first such evaluation was conducted as part of the 2008 TRECVID conference series [Rose et al., 2009] and again in 2009, 2010, 2011, 2012, 2013, 2014 and 2015. It was designed to move computer vision technology towards robustness and scalability while increasing core competency in detecting human activities within video. The approach used was to employ real surveillance data, orders of magnitude larger than previous computer vision tests, and consisting of multiple, synchronized camera views.

For 2016, the evaluation test data used a 10-hour subset (EVAL16) from the total 45 hours available of the test data from the Imagery Library for Intelligent Detection System's (iLIDS)[UKHO-CPNI, 2009] Multiple Camera Tracking Scenario Training (MCTTR) data set collected by the UK Home Office Centre for Applied Science and Technology (CAST) (formerly Home Office Scientific Development Branch's (HOSDB)). EVAL16 added 1 hour to the EVAL15 set.

This 10 hours contains a subset of the 11-hour SED14 Evaluation set that was generated following a crowdsourcing effort in order to generate the reference data. Since 2015, "camera4" is not used, as it had little events of interest.

In 2008, NIST collaborated with the Linguistics Data Consortium (LDC) and the research community to select a set of naturally occurring events with varying occurrence frequencies and expected difficulty. For this evaluation, we define an event to be an observable state change, either in the movement or interaction of people with other people or objects. As such, the evidence for an event depends directly on what can be seen in the video and does not require higher level inference. The same set of seven 2010 events were used for the 2011, 2012, 2013, 2014, 2015 and 2016 evaluations.

Those events are:

- CellToEar: Someone puts a cell phone to his/her head or ear
- Embrace: Someone puts one or both arms at least part way around another person
- ObjectPut: Someone drops or puts down an object
- PeopleMeet: One or more people walk up to one or more other people, stop, and some communication occurs
- PeopleSplitUp: From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame
- PersonRuns: Someone runs
- Pointing: Someone points

Introduced in 2015 was a 2-hour "Group Dynamic Subset" subset (SUB15) limited to three specific events: Embrace, PeopleMeet and PeopleSplitUp. This dataset was reused in 2016 as SUB16.

In 2016, only the retrospective event detection was supported. The retrospective task is defined as follows: given a set of video sequences, detect as many event observations as possible in each sequence. For this evaluation, a single-camera condition was used as the required condition (multiple-camera input was allowed as a contrastive condition). Furthermore, systems could perform multiple passes over the video prior to outputting a list of putative events observations (i.e., the task was retrospective).

The annotation guidelines were developed to express the requirements for each event. To determine if the observed action is a taggable event, a *reasonable interpretation rule* was used. The rule was, "if according to a reasonable interpretation of the video, the event must have occurred, then it is a taggable event". Importantly, the annotation guidelines were designed to capture events that can be detected by

human observers, such that the ground truth would contain observations that would be relevant to an operator/analyst. In what follows we distinguish between event types (e.g., parcel passed from one person to another), event instance (an example of an event type that takes place at a specific time and place), and an event observation (event instance captured by a specific camera).

6.1 Data

The development data consisted of the full 100 hours data set used for the 2008 Event Detection [Rose et al., 2009] evaluation. The video for the evaluation corpus came from the approximate 50 hour iLIDS MCTTR data set. Both data sets were collected in the same busy airport environment. The entire video corpus was distributed as MPEG-2 in Phase Alternating Line (PAL) format (resolution 720 x 576), 25 frames/sec, either via hard drive or Internet download.

System performance was assessed on EVAL16 and/or SUB16. Like SED 2012 and after, systems were provided the identity of the evaluated subset.

In 2014, event annotation was performed by requesting past participants to run their algorithms against the entire subset of data. A confidence score obtained from the participant's systems was created. A tool developed at NIST was then used to review event candidates. A first level bootstrap data was created out of this process and refined as actual test data evaluation systems from participants were received to generate a second level bootstrap reference which was then used to score the final SED results. The 2015 and 2016 data uses subsets of this data.

Events were represented in the Video Performance Evaluation Resource (ViPER) format using an annotation schema that specified each event observation's time interval.

6.2 Evaluation

For EVAL16, sites submitted system outputs for the detection of any of 7 possible events (PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, and Pointing). Outputs included the temporal extent as well as a confidence score and detection decision (yes/no) for each event observation. Developers were advised to target a low miss, high false alarm scenario, in order to maximize the number of event observations.

SUB16 followed the same concept, but only using 3 possible events (Embrace, PeopleMeet and PeopleSplitUp).

Teams were allowed to submit multiple runs with contrastive conditions. System submissions were aligned to the reference annotations scored for missed detections / false alarms.

6.3 Measures

Since detection system performance is a tradeoff between probability of miss vs. rate of false alarms, this task used the Normalized Detection Cost Rate (NDCR) measure for evaluating system performance. NDCR is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per time unit). At the end of the evaluation cycle, participants were provided a graph of the Decision Error Tradeoff (DET) curve for each event their system detected; the DET curves were plotted over all events (i.e., all days and cameras) in the evaluation set.

6.4 Results

Readers should see the online proceedings for individual team's performance and runs.

7 Video hyperlinking

7.1 System task

The high-level definition of the Video Hyperlinking (LNK) task in 2016 is the same as that of the 2015 edition of the task [Over et al., 2015]. The task requires the automatic generation of hyperlinks between given manually defined *anchors* within source videos and *target* videos from within a substantial collection of videos. Both targets and anchors are video segments with a start time and an end time. The result of the task for each anchor is a ranked list of target videos in decreasing likelihood of being *about* the content of the given anchor. Targets have to fulfill the following requirements: i) they must be from different videos than the anchor, ii) they may not overlap with other targets in the same anchor, finally iii), in order to facilitate ground truth annotation, the targets must be between 10 and 120 seconds in length.

The 2016 edition of the LNK task has the following main differences from the 2015 edition:

- The task switched from the professionally generated content of the BBC broadcast collection to a subset of the Blip10000 collection [Schmiedeke et al., 2013] crawled from the blip.tv website.

- The anchors created were ensured to be multimodal, i.e. the information about suitable targets is verbal-visual, by the use of an additional crowdsourcing anchor verification stage [Eskevich et al., 2017].

- The relevance assessment framework was split into 2 steps: general vetting of the submitted target video segments, and collection of detailed relevance descriptions.

7.2 Data

The Blip10000 dataset used for the 2016 task consists of 14,838 semi-professionally created videos [Schmiedeke et al., 2013]. As part of the task release, automatically detected shot boundaries were provided [Kelm et al., 2009], together with automatic speech recognition (ASR) transcripts [Lamel, 2012] originally provided with this dataset.

Additionally new versions of ASR transcripts and visual features were made available for the task. The new set of ASR transcripts were created by LIMSI using the 2016 version of their neural network acoustic models in their ASR system. The visual concepts were obtained using the BLVC CaffeNet implementation of the so-called AlexNet [Krizhevsky et al., 2012], which was trained by Jeff Donahue (@jeffdonahue) with minor variation from the version described in [Krizhevsky et al., 2012]. The model is available with the Caffe distribution [2]. In total, detection scores for 1000 visual concepts were extracted, with the five most likely concepts for each keyframe being released.

Data inconsistencies

Two issues were identified in the distributed version of the collection.

- The first issue is that for one video the wrong ASR file was provided. Here, we blacklisted the video, totally excluding it from the results and evaluation.

- The second issue is related to the metadata creation history, i.e. not all types of metadata were created using the original files, some made use of intermediate extracted content, i.e. extracted audio for the ASR transcripts. This led to the issue that for some video files, the length of the provided *'.ogv'* encoding was shorter than the encoding for which the shot cut detection and keyframe extraction was performed. If this was the case, it was possible for a run that used visual data only to return segments that did not exist in the ASR transcripts, which were derived from the *'.ogv'* video files. For 416 video files, circa 3% of all the data, the keyframes extended more than five minutes over the supplied *'.ogv'* video, which corresponds to 138 hours of extension. To make the evaluation comparable, we ignored all results after the of the *'.ogv'* video files.

7.3 Anchors

Anchors in the video hyperlinking task are essentially comparable to the search topics used in a standard video retrieval tasks. As in the 2015 edition of the task, we define an anchor to be the triple of: video (v), start time (s) and end time (e).

In 2016, we focused on the multimodal anchors, i.e. we selected anchors where the maker, who created the video, was using both audio and video modalities. These video segments cannot be properly understood by a potential viewer, if they are exposed only to one of the channels. In order to find segments that satisfied this criteria, we compiled a list of the following speech cues that can be associated with situations where people are showing something: 'can see', 'seeing here', 'this looks','looks like', 'showing', and 'want to show'. For practical reasons, we also limited anchors to be between 10 and 60 seconds long. Anchor creators used the mentioned speech cues to find potential anchors and decided whether to include the anchor by watching it. In total, 2 creators generated 94 anchors and corresponding descriptions of potentially relevant targets, i.e. information request descriptions that were further used in the evaluation process. 4 of these 94 anchors were later discarded from the evaluation because the crowdsourcing anchor verification test proved that they were not truly multimodal [Eskevich et al., 2017].

[2] see http://caffe.berkeleyvision.org/ for details

7.4 Evaluation

Ground truth

The ground truth was generated by pooling the top 5 results of all formally submitted participant runs (20), and running the assessment tasks on the Amazon Mechanical Turk (AMT)[3] platform. Overall, the ground truth creation proceeded in two stages:

- 'Target Vetting': top 5 targets for each anchor from the participants runs were assessed using the so-called forced choice approach being imposed on the crowdworkers. The forced choice means that the crowdworkers were given a target video segment and 5 textual targets descriptions (one of them being taken from the actual anchor that the target in question has been retrieved for). The task for the workers was to chose a definition that they felt was best suited to a given video segment. In case they chose the target description of the original anchor, this was considered to be a judgment of relevance. In case the target was unsuitable for any of the anchors, the crowdworkers were expected not to be comfortable making the choice among the 5 given options. For each top 5 anchor–target pair we collected 3 crowdworkers judgments. The final relevance decision was made based on the majority of the relevance judgments.

- 'Video-to-Video Relevance Analysis': the crowdworkers were shown both the anchor and target video segments, and were asked to give a textual description (2-3 natural language sentences) of the relevance relationship, i.e. what made the target relevant to the anchor.

Target vetting stage for all the participants submissions involves large scale crowdsourcing submissions processing which is not plausible to be carried out manually. Therefore, we ran a manual check of a small subset of crowdworkers submissions to the Target Vetting stage in order to confirm that the task was understood correctly. Further, the submissions were accepted or rejected automatically, following the algorithm that checked whether all the required decision metadata fields have been filled in, and whether the answer to the test questions were correct.

Initially we aimed at providing ground truth from the top 10 results of the 20 submitted runs. However, the top 10 ranks contained a total of 12,758 non-overlapping segments. Due to limited assessment resources we focused on the top 5 ranks from each run, comprising 7,216 targets. Of these targets, 2526 were identified as relevant and 4690 non-relevant.

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

Evaluation metrics

The evaluation metrics used were standard Mean Average Precision (MAP) and an adaptation of MAP called Mean Average interpolated Segment Precision (MAiSP) which is based on previously proposed adaptations of MAP for this task [Racca and Jones, 2015]. For MAP computation, we assume that a result segment is relevant if it overlaps with a segment that was judged relevant (see also [Aly et al., 2013]).

As the ground truth judgments were collected for the top 5 ranks of all submitted runs, Precision at rank 5 was another official evaluation metric.

7.5 Results

Five groups submitted four runs each, which resulted in 20 run submissions which were used for ground truth creation and assessment using the metrics described above. Readers should see the online proceedings for individual team's performance and runs.

8 Concept localization

The localization task challenges systems to make their concept detection more precise in time and space. Currently other video search task such as Adhoc and instance search systems are accurate to the level of the shot. In the localization task, systems are asked to determine the presence of the concept temporally within the shot, i.e., with respect to a subset of the frames comprised by the shot, and, spatially, for each such frame that contains the concept, to a bounding rectangle.

The localization is restricted to a subset of 10 concepts from those chosen and used in the semantic Indexing task between 2012 to 2015. This year a different set of concepts were tested than those tested in the past 3 years. In addition, most of the concepts were dynamic in nature compared to the object concepts used before in previous years.

For each concept from the list of 10 designated for localization, NIST distributed, about 5 weeks before the localization submissions are due at NIST for evaluation, a subset list of up to 1000 shots where each

Concept	*Name*	*shots*	*IFrames*
6	Animal	997	31330
13	Bicycling	998	21912
16	Boy	998	34230
38	Dancing	983	31584
49	Explosion_fire	983	20816
71	Instrument_Musician	1000	30374
100	Running	1000	24842
107	Sitting_Down	1000	52779
434	Skier	1000	32900
163	Baby	1000	17298

Table 6: Evaluated localization concepts

video shot may or may not contain the concept.

For each I-Frame within each shot in the list that contains the target, systems were asked to return the x,y coordinates of the upper left and lower right vertices of a bounding rectangle which contains all of the target concept and as little more as possible. Systems may find more than one instance of a concept per I-Frame and then may include more than one bounding box for that I-Frame, but only one will be used in the judging since the ground truth will contain only 1 per judged I-Frame, one chosen by the NIST assessor that is most prominent.

Table 6 describes for each of the 10 localization concepts the number of shots NIST distributed to systems and the number of I-Frames comprised by those shots:

8.1 Data

In total, 2 205 140 jpeg I-frames were extracted from the IACC.2 collection. 9 959 total shots were distributed and included total of 298 065 I-frames.

8.2 Evaluation

For each shot that contains a concept and selected and distributed by NIST, all I-frames were selected and displayed to the assessors and for each image the assessor was asked to decide first if the frame contained the concept or not, and, if so, to draw a rectangle on the image such that all of the visible concept was included and as little else as possible. In total, 55 789 I-frames were judged.

In accordance with the guidelines, if more than one instance of the concept appeared in the image, the assessor was told to pick just the most prominent one and box it in and stick with selecting it unless its prominence changed and another target concept has to be selected.

Assessors were instructed that in the case of occluded concepts, they should include invisible but implied parts only as a side effect of boxing all the visible parts.

In total, 11 runs were submitted this year by 3 teams.

8.3 Measures

Temporal and spatial localization were evaluated using precision and recall based on the judged items at two levels - the frame as the basis for temporal localization and the pixel bounding box for spatial localization. NIST then calculated an average for each of these values for each concept and for each run.

The set of annotated I-Frames was then used to evaluate the localization for the I-Frames submitted by the systems.

8.4 Results

Readers should see the online proceedings for individual team's performance and runs.

9 Video to Text Description

Automatic annotation of videos using natural language text descriptions has been a long-standing goal of computer vision. The task involves understanding of many concepts such as objects, actions, scenes, person-object relations, temporal order of events and many others. In recent years there has been major advances in computer vision techniques which enabled researchers to start practically to work on solving such problem. A lot of use case application scenarios can greatly benefit from such technology such as video summarization in the form of natural language, facilitating the search and browsing of video archives using such descriptions, describing videos to the blind, etc. In addition, learning video interpretation and temporal relations of events in the video will likely contribute to other computer vision tasks, such as prediction of future events from the video. This year a new showcase/pilot task has been proposed and a launched.

9.1 Video Dataset

A dataset of more than 30k Twitter Vine videos have been collected. Each has a total duration of about 6 sec long. In this showcase/pilot task a subset of 2000 Vine videos were randomly selected and annotated. Each video was annotated twice by two different annotators. In total, 4 sets of non-overlapping 500 videos were given to 8 annotators to generate a total of 4000 text descriptions. Those 4000 text descriptions were split into 2 sets corresponding to the original 2000 videos. Annotators were asked to include and combine in 1 sentence, if appropriate and available, four facets of the video they are describing:

- Who is the video describing such as concrete objects and beings (kinds of persons, animals, things)
- What are the objects and beings doing? (generic actions, conditions/state or events)
- Where such as locale,site,place,geographic, architectural (kind of place, geographic or architectural)
- When such as time of day, season

9.2 System Task

Given a set of about 2000 URLs of Vine videos and two sets (A and B) of text descriptions (each composed of 2000 sentences), systems are asked to work and submit results for at least one of two subtasks:

- Matching and Ranking: Return for each video URL a ranked list of the most likely text description that correspond (was annotated) to the video from each of the sets A and B.
- Description Generation: Automatically generate for each video URL a text description (1 sentence) independently and without taking into consideration the existence of sets A and B.

9.3 Evaluation

The matching and ranking subtask scoring was done automatically against the ground truth using mean inverted rank at which the annotated item is found or equivalent. While the Description generation subtask scoring was done automatically using the standard metrics from machine translation (MT) such as METEOR [Banerjee and Lavie, 2005] and BLEU [Papineni et al., 2002]. Systems were encouraged to take into consideration and use the four facets that annotators used as a guideline to generate their automated descriptions. In addition to using MT metrics, an experimental semantic similarity metric (STS) [Han et al., 2013] was applied. This metric measures how semantically similar is the submitted description to the ground truth description. In total, 11 teams signed up to the pilot task and seven finished by submitting 46 runs to the matching and ranking subtask and 16 runs to the description generation subtask.

9.4 Results

Readers should see the online proceedings for individual team's performance and runs.

10 Summing up and moving on

This introduction to TRECVID 2016 has provided basic information 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.

11 Authors' note

TRECVID would not have happened in 2016 without support from the National Institute of Standards and Technology (NIST). The research community is very grateful for this. Beyond that, various individuals and groups deserve special thanks:

- Koichi Shinoda of the TokyoTech team agreed to host a copy of IACC.2 data
- Georges Quénot provided the master shot reference for the IACC.3 videos.
- The LIMSI Spoken Language Processing Group and Vocapia Research provided ASR for the IACC.3 videos.
- Noel O'Connor and Kevin McGuinness at Dublin City University along with Robin Aly at the University of Twente worked with NIST and Andy O'Dwyer plus William Hayes at the BBC to make the BBC EastEnders video available for use in TRECVID

- Roeland Ordelman, Maria Eskevich, Robin Aly, Gareth Jones, Benoit Huet, and Martha Larson at University of Twente, Radboud University, Dublin City University, EURECOM and Delft University of Technology for coordinating the Video hyperlinking task
- Marc Ritter at TUC Chemnitz for supporting the Video to Text task annotations

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

12 Acknowledgments

The video hyperlinking work has been partially supported by: ESF Research Networking Programme ELIAS (travel grants for Serwah Sabetghadam and Maria Eskevich); BpiFrance within the NexGenTV project, grant no. F1504054U; Science Foundation Ireland (SFI) as a part of the ADAPT Centre at DCU (13/RC/2106). We would like to thank Martha Larson, Ayana Yaegashi, and Serwah Sabetghadam for their work on anchor creation and crowdsourcing anchor verification and target vetting assessments.

13 Appendix A: Ad-hoc query topics

501 Find shots of a person playing guitar outdoors
502 Find shots of a man indoors looking at camera where a bookcase is behind him
503 Find shots of a person playing drums indoors
504 Find shots of a diver wearing diving suit and swimming under water
505 Find shots of a person holding a poster on the street at daytime
506 Find shots of the 43rd president George W. Bush sitting down talking with people indoors
507 Find shots of a choir or orchestra and conductor performing on stage
508 Find shots of one or more people walking or bicycling on a bridge during daytime
509 Find shots of a crowd demonstrating in a city street at night
510 Find shots of a sewing machine
511 Find shots of destroyed buildings
512 Find shots of palm trees
513 Find shots of military personnel interacting with protesters
514 Find shots of soldiers performing training or other military maneuvers
515 Find shots of a person jumping
516 Find shots of a man shake hands with a woman
517 Find shots of a policeman where a police car is visible
518 Find shots of one or more people at train station platform
519 Find shots of two or more men at a beach scene
520 Find shots of any type of fountains outdoors
521 Find shots of a man with beard talking or singing into a microphone
522 Find shots of a person sitting down with a laptop visible
523 Find shots of one or more people opening a door and exiting through it
524 Find shots of a man with beard and wearing white robe speaking and gesturing to camera
525 Find shots of a person holding a knife
526 Find shots of a woman wearing glasses
527 Find shots of a person drinking from a cup, mug, bottle, or other container
528 Find shots of a person wearing a helmet
529 Find shots of a person lighting a candle
530 Find shots of people shopping

14 Appendix B: Instance search topics

9159 "Find Jim in the Pub"
9160 "Find Jim in this Kitchen"
9161 "Find Jim in this Laundrette"
9162 "Find Jim at this Foyer"
9163 "Find Jim in this Living Room"
9164 "Find Dot in the Pub"
9165 "Find Dot in this Kitchen"
9166 "Find Dot at this Foyer"
9167 "Find Dot in this Living Room"
9168 "Find Brad in the Pub"
9169 "Find Brad in this Kitchen"
9170 "Find Brad in this Laundrette"
9171 "Find Brad at this Foyer"
9172 "Find Brad in this Living Room"
9173 "Find Stacey in the Pub"
9174 "Find Stacey in this Kitchen"

9175 "Find Stacey in this Laundrette"

9176 "Find Stacey at this Foyer"

9177 "Find Stacey in this Living Room"

9178 "Find Patrick in the Pub"

9179 "Find Patrick in this Kitchen"

9180 "Find Patrick in this Laundrette"

9181 "Find Fatboy in the Pub"

9182 "Find Fatboy in this Laundrette"

9183 "Find Fatboy in this Living Room"

9184 "Find Pat in the Pub"

9185 "Find Pat in this Kitchen"

9186 "Find Pat in this Laundrette"

9187 "Find Pat at this Foyer"

9188 "Find Pat in this Living Room"

References

[Aly et al., 2013] Aly, R., Eskevich, M., Ordelman, R., and Jones, G. J. F. (2013). Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval Tasks. *CoRR*, abs/1312.1913.

[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.

[Eskevich et al., 2017] Eskevich, M., Larson, M., Aly, R., Sabetghadam, S., Jones, G. J. F., Ordelman, R., and Huet, B. (2017). Multimodal video-to-video linking: Turning to the crowd for insight and evaluation. In *23rd International Conference on Multimedia Modeling(MMM)*, Reykjavik, Iceland.

[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.

[Kelm et al., 2009] Kelm, P., Schmiedeke, S., and Sikora, T. (2009). Feature-based video key frame extraction for low quality video sequences. In *10th Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2009, London, United Kingdom, May 6-8, 2009*, pages 25–28.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114.

[Lamel, 2012] Lamel, L. (2012). Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data. In Tavast, A., Muischnek, K., and Koit, M., editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012*, volume 247 of *Frontiers in Artificial Intelligence and Applications*, pages 1–8. IOS Press.

[Over et al., 2015] Over, P., Fiscus, J., Joy, D., Michel, M., Awad, G., Kraaij, W., Smeaton, A. F., Quénot, G., and Ordelman, R. (2015). TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2015*. NIST, USA.

[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.

[Racca and Jones, 2015] Racca, D. N. and Jones, G. J. F. (2015). Evaluating Search and Hyperlinking: An Example of the Design, Test, Refine Cycle for Metric Development. In *Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany.

[Rose et al., 2009] Rose, T., Fiscus, J., Over, P., Garofolo, J., and Michel, M. (2009). The TRECVid 2008 Event Detection Evaluation. In

IEEE Workshop on Applications of Computer Vision (WACV). IEEE.

[Schmiedeke et al., 2013] Schmiedeke, S., Xu, P., Ferrané, I., Eskevich, M., Kofler, C., Larson, M. A., Estève, Y., Lamel, L., Jones, G. J. F., and Sikora, T. (2013). Blip10000: a social video dataset containing SPUG content for tagging and retrieval. In *Multimedia Systems Conference 2013 (MMSys '13)*, pages 96–101, Oslo, Norway.

[Strassel et al., 2012] Strassel, S., Morris, A., Fiscus, J., Caruso, C., Lee, H., Over, P., Fiumara, J., Shaw, B., Antonishek, B., and Michel, M. (2012). Creating havic: Heterogeneous audio visual internet collection. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Thomee et al., 2016] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73.

[UKHO-CPNI, 2009] UKHO-CPNI (2007 (accessed June 30, 2009)). Imagery library for intelligent detection systems. `http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids/`.

[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.