TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics

Paul Over {over@nist.gov}
Jon Fiscus {jfiscus@nist.gov}
Greg Sanders {gregory.sanders@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

George Awad
Dakota Consulting, Inc.
1110 Bonifant Street, Suite 310
Silver Spring, MD 20910
{gawad@nist.gov}

Alan F. Smeaton {Alan.Smeaton@dcu.ie}
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland

Wessel Kraaij {wessel.kraaij@tno.nl}
TNO
Delft, the Netherlands
Radboud University Nijmegen
Nijmegen, the Netherlands

Georges Quénot {Georges.Quenot@imag.fr}
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP /
CNRS, LIG UMR 5217, Grenoble, F-38041 France

April 11, 2016

1 Introduction

This introduction to the TREC Video Retrieval Evaluation (TRECVID) 2013 will be expanded and rereleased in early 2014 with a discussion of approaches and results.

TRECVID 2013 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 ten 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 National Institute of Standards and Technology (NIST) and other US government agencies. Many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2013 represented a continuation of five tasks from 2012. Fifty-one teams (see Tables 1 and 2) from various research organizations — 20 from Europe, 20 from Asia, 9 from North America, and 2 from South America — completed one or more of five tasks:

1. Semantic indexing
2. Instance search
3. Multimedia event detection
4. Multimedia event recounting
5. Surveillance event detection

Some 200 h of short videos from the Internet Archive (archive.org), available under Creative Commons licenses (IACC.2), were used for semantic indexing. 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 h of new BBC EastEnders video was used for the instance search task. 45 h of airport surveillance video was reused for the surveillance event detection task. Almost 5200 h from the Heterogeneous Audio Visual Internet Corpus (HAVIC) of Internet videos was used for development and testing in the multimedia event detection task.

Instance search results were judged by NIST assessors - similarly for the semantic indexing task with additional assessments done in France under the European Quaero program (QUAERO, 2010). Multimedia and surveillance event detection were scored by NIST using ground truth created manually by the Linguistic Data Consortium under contract to NIST. The multimedia event recounting task was judged by humans experts in an evaluation designed by NIST.

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, 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 with the European Union's AXES project (www.axes-project.eu) 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.

Internet Archive Creative Commons (IACC.2) video

The IACC.2 dataset comprises 7300 Internet Archive (archive.org) videos (144 GB, 600 h) with Creative Commons licenses in MPEG-4/H.264 format with duration ranging from 10 s to 6.4 min and a mean duration of almost 5 min. Most videos have some metadata provided by the donor available e.g., title, keywords, and description

For 2013, approximately 600 additional h of Internet Archive videos with Creative Commons licenses in MPEG-4/H.264 and with durations between 10 s and 6.4 min were used as new test data. This data was randomly divided into 3 datasets: IACC.2.A, IACC.2.B, and IACC.2.C. IACC.2.A is the test dataset for semantic indexing in 2013; IACC.2.B and IACC.2.C were available for gauging current systems against future test data under the "progress" option in the semantic indexing task. Most videos had some donor-supplied metadata available e.g., title, keywords, and description. Approximately 600 h of IACC.1 videos were available for system development.

As in the past, LIMSI and Vocapia Research provided automatic speech recognition for the English speech in the IACC.2 video.

iLIDS Multiple Camera Tracking Data

The iLIDS Multiple Camera Tracking data consisted of ≈150 h of indoor airport surveillance video collected 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 video consisted of the ≈100 h of data used for SED 2008 evaluation. The evaluation video consisted of the same additional ≈50 h of data from Imagery Library for Intelligent Detection System's (iLIDS) multiple camera tracking scenario data used for the 2009 - 2013 evaluations(UKHO-CPNI, 2007 (accessed June 30, 2009)).

One third of the evaluation video was annotated by the Linguistic Data Consortium using a triple-pass annotation procedure. Seven of the ten annotated events were used for the 2013 evaluation.

Heterogeneous Audio Visual Internet Corpus (HAVIC)

HAVIC ((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-4s Advanced Audio Coding (ACC) (ACC, 2010) encoded audio.

This year, the HAVIC system development materials were re-partitioned into the follow data components:

- Event kits [290 h] (event training material for 40 events),

Table 1: Participants and tasks

Task					Location	TeamID	Participants
−−	MD	MR	−−	SI	Eur	PicSOM	Aalto U.
−−	−−	−−	SD	−−	NAm	ATTLabs	AT&T Labs Research
−−	−−	−−	SD	−−	Asia	BIT	Beijing Institute of Techonolgy
−−	MD	MR	SD	SI	NAm	Inf	Carnegie Mellon U.
IN	−−	−−	−−	**	Eur	CEALIST	CEA LIST, Vision & Content Engineering Lab
IN	**	−−	−−	SI	Eur	IRIM	IRIM Consortium
IN	MD	MR	−−	SI	Asia	VIREO	City U. of Hong Kong
−−	−−	−−	SD	SI	Eur	dcu...	Dublin City U., Univ of Ulster,, Vicomtech-IK4
IN	MD	MR	−−	SI	Eur	AXES	Access to Audiovisual Archives
IN	−−	−−	−−	−−	Eur	iAD_DCU	Dublin City U., U. of Tromso
−−	−−	−−	−−	SI	Eur	EURECOM	EURECOM
−−	−−	−−	−−	SI	Eur	VIDEOSENSE	EURECOM, LIRIS, LIF, LIG, Ghanni
**	**	−−	−−	SI	Eur	TOSCA	EuropeOrganization(s)
−−	−−	−−	−−	SI	NAm	FIU_UM	Florida International Univ, Univ. of Miami
−−	−−	−−	−−	SI	Eur	FHHI	Fraunhofer Heinrich Hertz Institute, Berlin
−−	−−	−−	−−	SI	Asia	HFUT	Hefei U. of Tech.
**	MD	MR	SD	SI	NAm	IBM...	IBM T. J. Watson Research Center
IN	MD	MR	−−	SI	Eur	ITI_CERTH	Centre for Research & Tech. Hellas
−−	**	−−	−−	SI	Eur	Quaero	INRIA, LIG, KIT
IN	−−	−−	−−	−−	Eur	ARTEMIS	Institut Mines-Telecom; Telecom SudParis; ARTEMIS
−−	MD	−−	−−	−−	Eur,Asia	siegen...	U. of Siegen, Kobe U., Muroran Institute of Tech.
IN	**	−−	−−	SI	Eur	JRS	JOANNEUM RESEARCH FmbH
−−	MD	MR	−−	−−	NAm	GENIE	Kitware, Inc.
IN	**	−−	SD	**	Asia	BUPT...	Beijing Univ. of Posts & Telecommunications
IN	−−	−−	−−	−−	Asia	MIC_TJ	Tongji U.
IN	MD	**	**	SI	Asia	NII	National Institute of Informatics
−−	−−	−−	−−	SI	Asia	NHKSTRL	NHK (Japan Broadcasting Corp.)
−−	−−	−−	−−	SI	Asia	ntt	NTT Media Intelligence Labs, Dalian Univ of Tech.
IN	**	−−	−−	−−	Asia	NTT_NII	NTT, NII
IN	MD	−−	**	−−	SAm	ORAND	ORAND S.A. Chile
IN	**	**	−−	SI	Asia	FTRDBJ	Orange Labs International Centers China
IN	−−	−−	−−	−−	Asia	IMP	Osaka Prefecture U.
−−	−−	−−	SD	−−	Asia	PKU_OS	Peking U.-OS
IN	**	−−	**	**	Asia	PKU-ICST	Peking U.-ICST
−−	MD	−−	−−	−−	Eur	Vis_QMUL	Queen Mary, U. of London
−−	MD	MR	−−	−−	NAm	BBNVISER	Raytheon
−−	MD	MR	−−	−−	NAm,Eur	SRI_SESAME	SRI International U. of Amsterdam
−−	MD	MR	−−	SI	NAm	SRIAURORA	SRI, Sarnoff, Central Fl.U., U. Mass., Cycorp, ICSI, Berkeley
−−	−−	−−	SD	−−	NAm	ccnysri	The City College of New York SRI-International Sarnoff
IN	MD	−−	−−	−−	Eur	TNO_M3	TNO
IN	MD	**	**	SI	Asia	TokyoTech...	Tokyo Institute of Tech. Canon Inc.
IN	−−	−−	**	−−	Asia	thu.ridl	Tsinghua U.
IN	−−	−−	−−	SI	Eur,Asia	sheffield	U. of Sheffield, Harbin Engineering U., PRC U. of Eng. & Tech.
**	**	−−	**	SI	SAm	MindLAB	Universidad Nacional de Colombia
IN	MD	**	−−	SI	Eur	MediaMill	U. of Amsterdam
**	**	−−	−−	SI	Asia	UEC	U. of Electro-Communications
−−	MD	−−	−−	−−	NAm,Asia	UMass	U. of Massachusetts Amherst, U. of Science & Tech. Beijing
−−	−−	−−	SD	−−	NAm	VIVA...	U. of Ottawa, Ecole Polytechnique de Montreal
IN	−−	−−	−−	−−	Asia	NERCMS	Wuhan U.

Task legend. IN:instance search; MD:multimedia event detection; MR:multimedia event recounting; SD: surveillance event detection; SI:semantic indexing; −−:no run planned; **:planned but not submitted

Table 2: Participants who did not submit any runs

IN	MD	MR	SD	SI	Location	TeamID	Participants
−−	**	**	−−	−−	Eur	Bilkent_RETINA	Bilkent U.
**	−−	−−	**	−−	Eur	Brno	Brno U. of Tech.
−−	−−	−−	−−	**	Eur	ECL_LIRIS	Ecole Centrale de Lyon LIRIS UMR 5205 CNRS
−−	**	−−	−−	−−	NAm	KBVR	Etter Solutions LLC
**	−−	−−	−−	**	Asia	IRC_FZU	Fuzhou U.
−−	**	**	**	−−	Asia	IITH	Indian Institute of Tech. Hyderabad
−−	**	−−	**	−−	Eur	INRIA_STARS	INRIA - STARS
**	**	**	**	**	Asia	ECNU	Institute of Computer Applications
**	−−	−−	**	**	SAm	RECOD	U. of Campinas
−−	**	−−	**	−−	Asia	VCAX	Xi'an Jiaotong U.
**	**	**	**	**	Asia	IMG_THU	Tsinghua U.
**	−−	−−	−−	−−	Eur	Lincoln_scs	Lincoln U.
−−	−−	−−	**	**	NAm	LANL_DSGM	Los Alamos National Lab
**	−−	−−	−−	−−	Asia	MML	MML,(CITI) of Academia Sinica
−−	**	**	**	−−	Asia	MMM_TJU	MMM_TJU
−−	−−	−−	**	−−	NAm	Noblis	Noblis, Inc.
−−	**	−−	−−	−−	NAm,Asia	OMGA	OMRON Corporation, Georgia Institute of Tech.
−−	−−	−−	**	−−	NAm	CBSA	Canada Border Services Agency
−−	**	−−	**	−−	Asia	SYSU_IMC	Sun Yat-sen U.
−−	−−	−−	**	−−	Asia	TYUT_***	Taiyuan U. of Tech.
**	−−	−−	**	−−	NAm	UCSB_UCR	U. of California - Santa Barbara, U. of California - Riverside
−−	**	−−	−−	**	Eur	DCAPI	U. of Lincoln
**	−−	−−	−−	−−	Asia	U_tokushima	U. of Tokushima
−−	**	**	−−	−−	Eur	IMS	Vienna U. of Tech.
**	**	−−	−−	**	Eur	WIDEIO	WIDE IO LTD

Task legend. IN:instance search; MD:multimedia event detection; MR:multimedia event recounting; SD: surveillance event detection; SI:semantic indexing; −−: no run planned; **:planned but not submitted

- Research Resources [314 h] (development resources composed of MED11 (2011 MED data) Development data and a portion of the MED11 Test data that may be altered, amended or annotated in any way participants need to facilitate their research),

- MEDTest [837 h] (a site-internal testing data set composed of a subset of the MED11 Test data that is structured as fixed background [non-event] clip set and additional positive examples for test events),

- KindredTest [675 h] (an internal testing data structured as a fixed set of background [non-event] clips that contain a 'city building exterior' and the same event positives as used in the MEDTest collection.)

The evaluation corpus was the 3722 hour MED Progress Collection (PROGAll) and a new, 1243 hour subset or PROGAll (designated PROGSub) to give participants the option to process less test collection data.

3 Semantic indexing

A potentially important asset to help video search/navigation is the ability to automatically identify the occurrence of various semantic features/concepts such as "Indoor/Outdoor","People", "Speech" etc., which occur frequently in video information. The ability to detect features is an interesting challenge by itself but takes on added importance to the extent it can serve as a reusable, extensible basis for query formation and search. The semantic indexing task was a follow-on to the feature extraction task. It was coordinated by NIST and by Georges Quénot under the Quaero program.

3.1 System task

The semantic indexing (SIN) task was as follows. Given a standard set of shot boundaries for the semantic indexing test collection and a list of concept definitions, participants were asked to return for each concept in the full set of concepts, at most the top 2 000 video shots from the standard set, ranked according to the highest possibility of detecting the presence of the concept. The presence of each concept was assumed to be binary, i.e., it was either present or absent in the given standard video shot. If the concept 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. A pilot extension to the task in 2012 was the addition of a paired-concepts topics where the goal was to detect the presence of a pair of concepts that are visible in the video shot at the same time.

Three novelties were introduced as pilot extensions to the participants in 2013:

- Measurement of system progress for a fixed set of concepts and independent of the test data, across 3 years (2013-2015)

- To offer a new, optional system output in concept pairs to indicate the temporal sequence in which the two concepts occur in the video shot

- To offer a new, optional "localization" subtask with the goal of localizing 10 detected concepts inside the I-Frames of the video shots

Five hundred concepts were selected for the TRECVID 2011 semantic indexing task. In making this selection, the organizers drew from the 130 used in TRECVID 2010, the 374 selected by CU/Vireo for which there exist annotations on TRECVID 2005 data, and some from the LSCOM ontology. From these 500 concepts, 346 concepts were selected for the full task in 2011 as those for which there exist at least 4 positive samples in the final annotation. For 2013 the same list of 500 concepts has been used as a starting point for selecting the 60 single concepts for which participants must submit results in the main task and the 10 concept pairs in the paired concept task. The 10 concepts for localization will be a subset of the main task concepts.

This year the evaluated paired-concepts were as follows, listed by concept number:

911 Telephones + Girl
912 Kitchen + Boy
913 Flags + Boat_ship
914 Boat_ship + Bridges
915 Quadruped + Hand
916 Motorcycle + Bus
917 Chair + George_w_Bush
918 Flowers + Animal
919 Explosion_Fire + Dancing
920 Government_Leader + Flags

In 2013 the task will again support 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 retrieved by a general purpose search engine (e.g. Google) using only the concept name and/or definition with only automatic processing of the returned images. This will again be implemented by using additional categories for the training types besides the A to D ones (see below).

Four types of submissions are considered: "main" in which participants submitted results for 60 single concepts, "loc" in which main task participants submitted localization results for 10 concepts drawn from the 60 main concepts, "progress" in which participants submitted independent results for all and only the 60 main task concepts but against the IACC.2.A, IACC.2.B, and IACC.2.C data, and finally the "pair" submissions in which participants submitted results for 10 concept pairs and optionally, information on the time sequence in which the two concepts appears may be submitted.

TRECVID evaluated 38 of the 60 submitted single concept results and all of the 10 submitted paired concept results. Fifteen single concepts and 5 paired concepts were judged at NIST. Twenty-three single concepts and 5 paired concepts were judged under the Quaero program in France. NIST judged the localization submissions for 10 of the 38 single concepts. No time sequence results for the paired concepts were submitted. The 60 single concepts are listed below. Those that were evaluated in the main task are marked with an asterisk. The subset evaluated for localization are marked with ">".

3 * > Airplane
5 * Anchorperson
6 * Animal
9 Basketball
10 * Beach
13 Bicycling
15 * > Boat_Ship
16 * Boy
17 * > Bridges
19 * > Bus
22 Car_Racing
25 * > Chair
27 Cheering
29 Classroom
31 * Computers
38 * Dancing
41 Demonstration_Or_Protest
49 * Explosion_Fire
52 * Female-Human-Face-Closeup
53 * Flowers
54 * Girl
56 * Government-Leader
57 Greeting
59 * > Hand
63 Highway
71 * Instrumental_Musician
72 * Kitchen
77 Meeting
80 * > Motorcycle
83 * News_Studio
84 Nighttime
85 Office
86 * Old_People
89 * People_Marching
95 Press_Conference
97 Reporters
99 Roadway_Junction
100 * Running
105 * Singing
107 * Sitting_Down
112 Stadium
115 Swimming
117 * > Telephones
120 * Throwing
163 * Baby
227 * Door_Opening
254 * Fields
261 * > Flags
267 * Forest
274 * George_Bush
297 Hill
321 Lakes
342 * Military_Airplane
359 Oceans
392 * > Quadruped
431 * Skating
434 Skier
440 Soldiers
454 * Studio_With_Anchorperson
478 Traffic

Concepts were defined in terms which a human judge could understand. Some participating groups made their feature detection output available to participants in the search task which really helped in the search task and contributed to the collaborative nature of TRECVID.

The fuller concept definitions provided to system developers and NIST assessors are listed

with the detailed semantic indexing runs at the back of the workshop notebook and on the webpage: `http://www-nlpir.nist.gov/projects/tv2012/tv11.sin.500.concepts_ann_v2.xls`

Work at Northeastern University (Yilmaz & 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, Ianeva, Kraaij, & Smeaton, 2006). This year mean extended inferred average precision (mean xinfAP) was used which permits sampling density to vary (Yilmaz, Kanoulas, & Aslam, 2008). This allowed the evaluation to be more sensitive to shots returned below the lowest rank ($\tilde{1}00$) 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.

3.2 Data

The IACC.2.A collection was used for testing. It contained 112 677 shots. IACC.2.B-C collections used in the "Progress" task contained 107 806 and 113 467 shots. Automatic Speech Recognition (ASR) output on IACC.2 videos was provided by LIMSI (Gauvain, Lamel, & Adda, 2002) and a community annotation of concepts was organized by LIG and LIF groups (Ayache & Quénot, 2008).

3.3 Evaluation

Each group was allowed to submit up to 4 prioritized main runs and two additional if they are "no annotation" runs, one localization run was allowed with each main submission, up to 2 progress runs was allwed on each of the 2 progress datasets, and up to two paired-concept runs. Each participant in the paired concept task must submit a baseline run which just combines for each pair the output of group's two independent single-concept detectors. In fact 26 groups submitted a total of 98 main runs, 9 localization runs, 18 progress runs, and 21 paired-concept runs. No teams participated in the temporal sequence subtask for concept pairs.

Main and paired concepts

For each concept in the main and paired concept tasks, pools were created and randomly sampled as follows. The top pool sampled 100 % of shots ranked 1-200 across all submissions. The bottom pool sampled 6.7 % of ranked 201-2000 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, he or she was asked to rejudge all clips submitted by at least 10 runs at ranks 1 to 200 and was judged as not containing the concept by the assessor. In all, 336 683 were judged. 2 018 182 shots fell into the unjudged part of the overall samples.

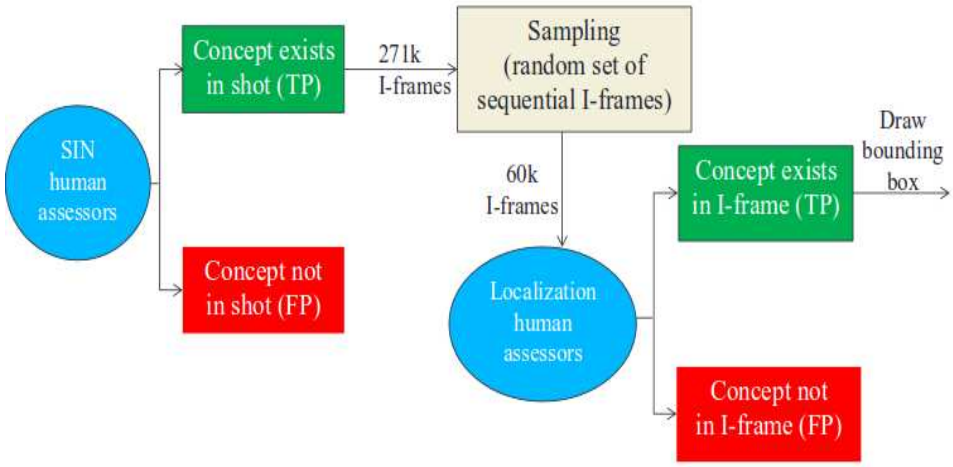
Localization

For the localization subtask judging proceeded as follows. For each shot found to contain a concept in the main task, a sequential 22 % subset of the included I-Frames beginning at a randomly selected point within the shot was selected and presented to an assessor. The seletion of a sequence of images rather than a random sample was intended to favor systems that used the video context of each image to do the localization rather than treating each image in isolation. 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. Figure **??** shows the evaluation framework. 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. Assessors were told 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.

Early in the assesssment process it became clear some additional guidelines were needed. Sometimes in a series of sequential images the assessor might know from context that a blurred area was in fact the concept. In this case we instructed the assessor to judge such an image as containing the concept and box the blurry area.

As this was the first running of this task and assessment, we planned a minimum of 5 assessor halfdays for each of the 10 topics to be judged. At NIST we tried the task ourselves with the software we de-

Figure 1: Concept Localization Evaluation Framework



veloped and based on our performance we estimated each assessor could judge roughly 6000 images in the time alloted.

The following table describes for each of the 10 localization concepts the number of shots judged to contain the concept and the number of I-Frames comprised by those shots.

Concept	*Name...*	*True shots*	*I-Frames*
3	Airplane	100	297
15	Boat_Ship	479	2917
17	Bridges	140	884
19	Bus	148	1095
25	Chair	1298	20064
59	Hand...	1598	18290
80	Motorcycle	289	1846
117	Telephones	152	1348
261	Flags	480	5980
392	Quadruped	448	6641

The larger numbers of I-Frames to be judged for concepts 25 and 59 within the time allotted caused us to assign some of those images to assessors who had not done the original shot judgments. Such additional assessors were told to make liberal judgments about the (non)presence of the concept (not worry about fringe cases) and focus on the localization. One concept presented particular problems when assigned for localization to multiple assessors.

```
Hand: A close-up view of one or more human
      hands, where the hand is the primary
      focus of the shot
```

While the definition of "a human hand" is relatively clear, the notions of "close-up" and "primary focus" are very fuzzy and invite differing judgments.

3.4 Measures

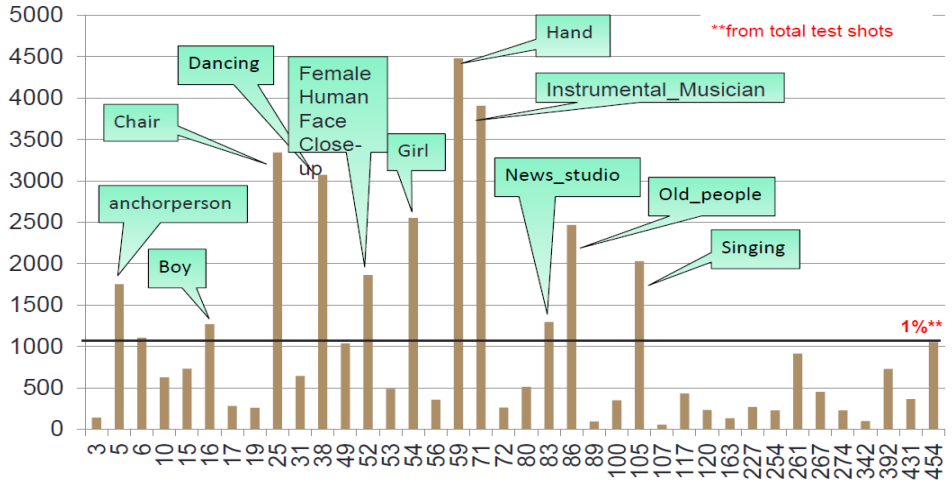
Main and paired concepts

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 concepts, runs can be compared in terms of the mean inferred average precision across all evaluated single concepts. The results also provide some information about "within concept" performance.

Localization

Temporal and spatial localization were evaluated using precision and recall based on the judged items at two levels - the frame and the pixel, respectively. NIST then calculated an average for each of these values for each concept and for each run. For each shot that is judged to contain a concept, a subset of the shot's I-Frames was viewed and annotated to locate the pixels representing the concept. The set of annotated I-Frames was then be used to evaluate the localization for the I-Frames submitted by the systems.

Figure 2: SIN: Histogram of shot frequencies by concept number



3.5 Results

Single Concepts

Performance varied greatly by concept. Figure 2 shows how many unique instances were found for each tested concept. The inferred true positives (TPs) of 12 concepts exceeded 1 % from the total tested shots. Top performing concepts were "Hand", "Chair", "Dancing", "Girl", "Old_People", "Singing","Instrumental_Musician","AnchorPerson", "News_Studio", "Female_Human_Face_Closeup", and "Boy".

On the other hand, features that had the fewest TPs were "Airplane*", "People_Marching", "Sitting_Down*","Military_Airplane*", "Bridges*", "Kitchen*", "Bus", "Government_Leader","Door_Opening", "Fields", "Throwing*", "Baby*","George_Bush", "Skating", "Forest", and "Telephones". It is worth mentioning here that there are 7 common concepts with TRECVID 2012 that share the low found TPs percentage. Those are the ones listed above and end with a star "*". It is not very clear if systems are really struggling to detect them or they are rare in the testing dataset used.

The top performing concepts were more generic by definition than the bottom performing ones which are more specific in category, location or action such as "sitting-down", "Kitchen", and "Baby". In addition, many of the low performing features are easily confusable by another visually similar features such as "Airplane","Military_Airplane".

Figure 3 shows the results of category A for the main run submissions. Category A runs used only IACC training data. The median score across all runs was 0.128 while maximum score reached 0.321. Also, the median baseline run score automatically generated by NIST is plotted on the graph with score 0.143.

Still category A runs were the most popular type and achieve top recorded performances. Only 8 runs from category E & F was submitted and achieved top scores of 0.048 and 0.046 respectively.

Figure 4 shows the performance of the top 10 teams across the 38 features. Few concepts reflected a medium spread between the scores of the top 10 such as feature "Anchorperson", "Animal", "Hand","Instrumental-Musician", "Quadruped", "Skating","Flags","Baby","Throwing","Dancing", "Computers", and "Beach". The spread in scores may indicate that there is still room for further improvement within used techniques. The majority of the rest of the concepts had a tight spread of scores among the top 10 which may indicate a small variation in used techniques performance. In general, the median scores ranged between 0.001 (feature "Sitting_down" and "Telephones") and 0.54 (feature "News-Studio"). As a general observation, feature "Sitting_down" had the minimum median score for the last 3 years which demonstrates how difficult this feature is for the systems to detect.

To test if there were significant differences between the systems' performance, we applied a randomiza-

Figure 3: xinfAP by run (cat. A)

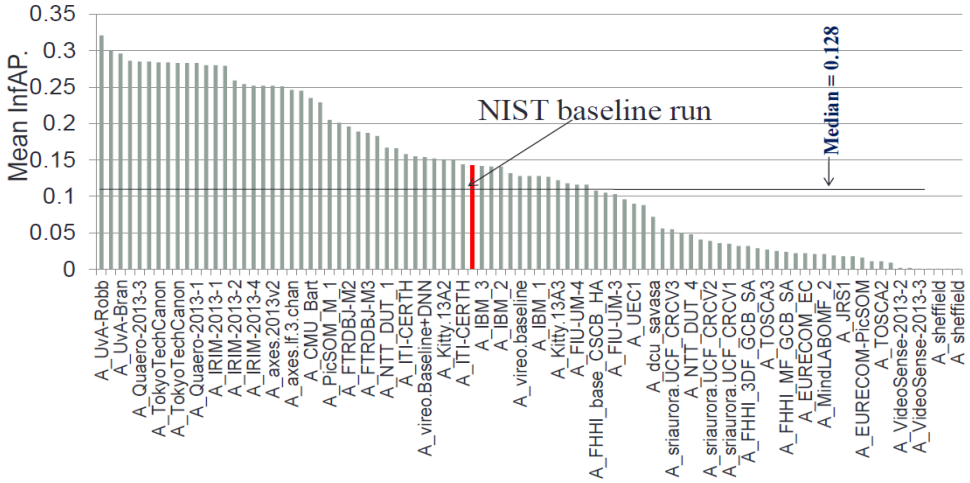


Figure 4: Top 10 runs (xinfAP) by concept number

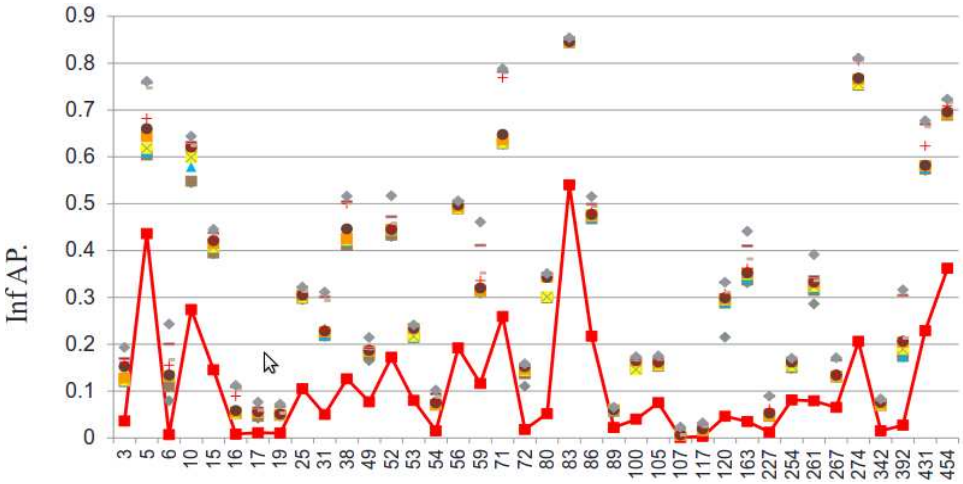
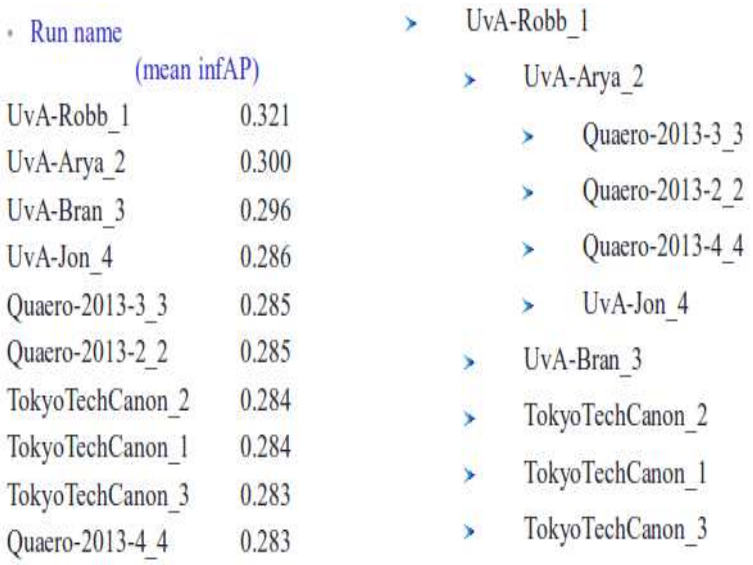


Figure 5: Significant differences among top A-category full runs



Run name	(mean infAP)
UvA-Robb_1	0.321
UvA-Arya_2	0.300
UvA-Bran_3	0.296
UvA-Jon_4	0.286
Quaero-2013-3_3	0.285
Quaero-2013-2_2	0.285
TokyoTechCanon_2	0.284
TokyoTechCanon_1	0.284
TokyoTechCanon_3	0.283
Quaero-2013-4_4	0.283

tion test (Manly, 1997) on the top 10 runs for training category A as shown in Figure 5. The figure indicates the order by which the runs are significant according to the randomization test. Different levels of indentation signifies a significant difference according to the test. Runs at the same level of indentation are indistinguishable in terms of the test. In this test the top ranked run was significantly better than other runs.

Concept-Pairs

Figure 6 shows the performance for the subtask of concept-pairs. In general the highest number of true positives (hits) came from the concepts "Flag + Boat_Ship" and fewest hits came from "Flowers + Animal". Compared to last year, hits are much less this year reaching maximum little above 100. It is not clear yet if this is due to that those concepts are rare in the testing set or systems are still learning how to combine the results of different independent detectors and fuse their results. One major issue is that some systems learn the presense of certain concepts using different features that can exist in a positive concept-pair shot. For example a system can learn the presence of a crowd of people when trained on a set of shots in a stadium setting and therefore can miss concept-pair shot if it asked to return a shot that contains a crowd of people AND a street. And

Figure 6: Concept-Pairs: Histogram of shot frequencies by concept number

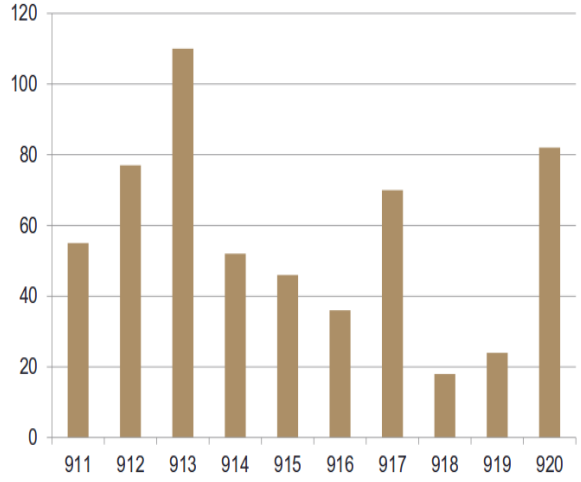


Figure 8: Significant differences among top A-category concept-pairs runs

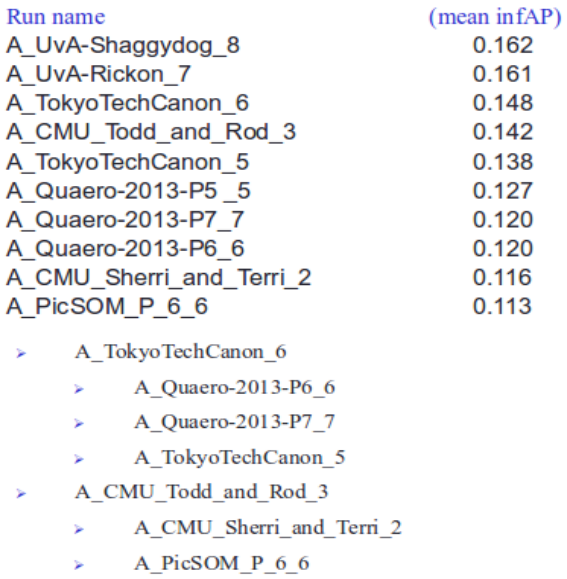
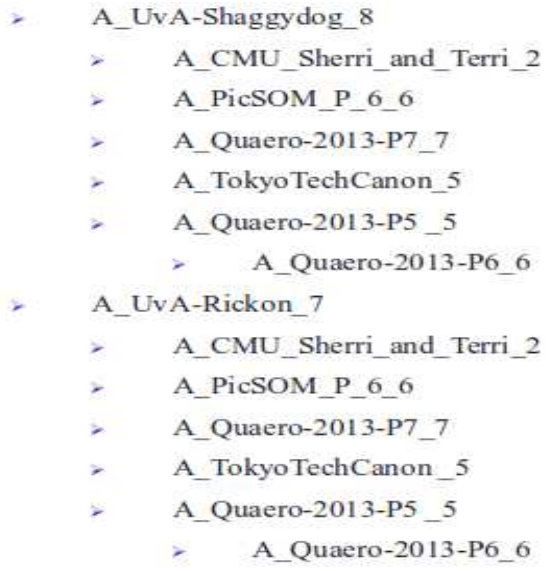


Figure 9: Significant differences among top A-category concept-pairs runs (continued)



thus we believe that for concept pairs learning, systems have to learn in a way that is less sensitive to the context of the surrounding visual appearance of non-relevant features.

Figure 7 show the range of scores for concept-pairs runs. The top run achieved score 0.164 (higher than last year) while the median score was 0.1125. This year systems were required to submit a baseline run which just combines for each pair the output of the group's two independent single-concept detectors. The goal was to achieve better scores in their other submitted regular runs. In fact most systems submitted a baseline run but not all. And among those who submitted we found that 3 teams had baseline runs that achieved better scores than regular runs while only 2 teams had all their regular runs better than the baseline. This again confirms that it is harder for systems to learn single concepts independently from other relevant context features and use their output for detecting paired concepts. Figures 8 and 9 show the randomization test on concept-pair runs.

Concept Localization

Figure 10 show the mean precision, recall and fscore of the returned I-frames by all runs across all 10 concepts. All runs reported much higher recall (reaching a maximum above 50 %) than precision or fscore except 1 team (FTRDBJ) which had close scores for the 3 measures. Lower precision scores (maximum 20 %) indicate that most runs returned a lot of non-relevant I-frames that didn't contain the concept. On the other hand figure 11 shows the same measure by run for spatial localization (correctly returning a bounding box around the concept). Here scores were much lower than the temporal measures and reaching hardly above 10 % precision. This indicates that finding the best bounding box was a much harder problem than just returning a correct I-frame.

The average true positive I-frames vs average false positive I-frames for each run can be shown in Figure 12. For many runs the average False positive I-frames are almost double the average true positive I-frames even for top runs. Systems that tried to be more conservative in reporting I-frames didn't gain much in terms of fscore measure. There is a big challenge for systems to try to balance the accuracy of the returned I-frames while still achieving high fscore measure.

The F-score performance by concept is shown in figures 13 and 14 for temporal and spatial respectively across all runs. In general temporal scores are higher than spatial scores with the concept "Flags" reporting maximum score of 0.6 for temporal localization and about 0.3 for spatial localization. We notice

Figure 7: xinfAP by run (cat. A) - concept-pairs

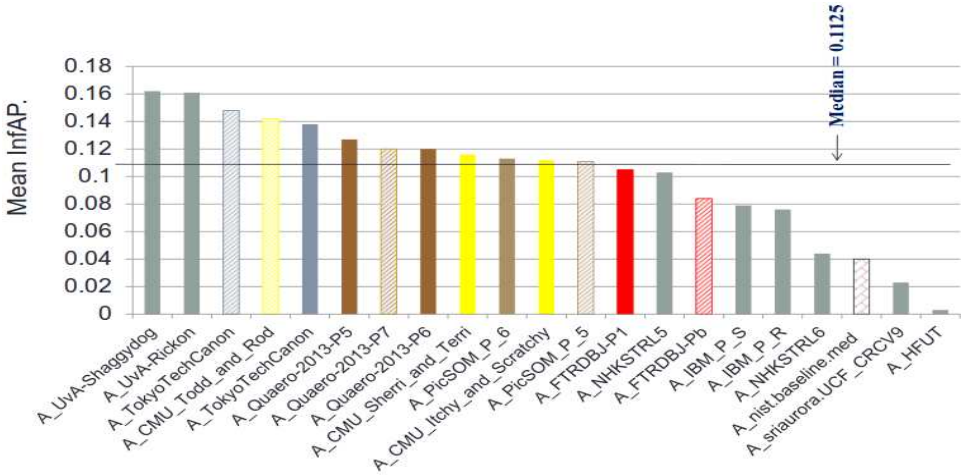


Figure 10: Concept Localization: Temporal localization results by run

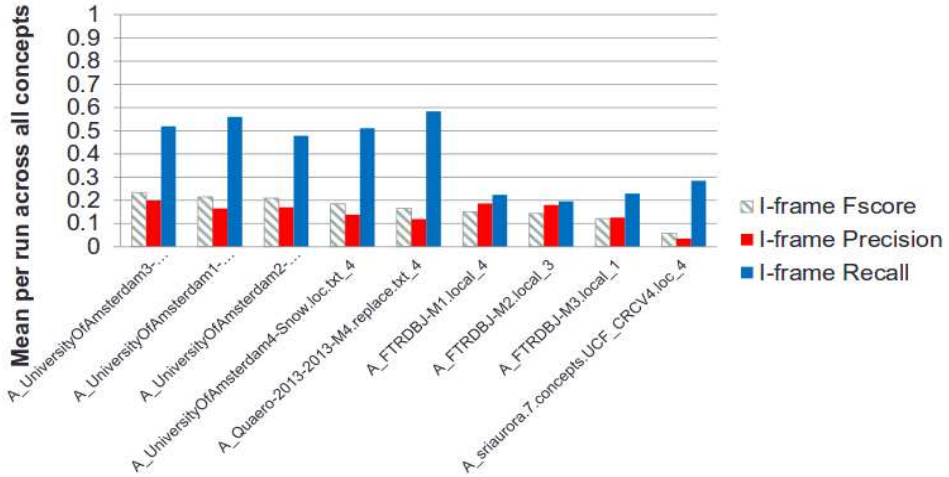


Figure 11: Concept Localization: Spatial localization results by run

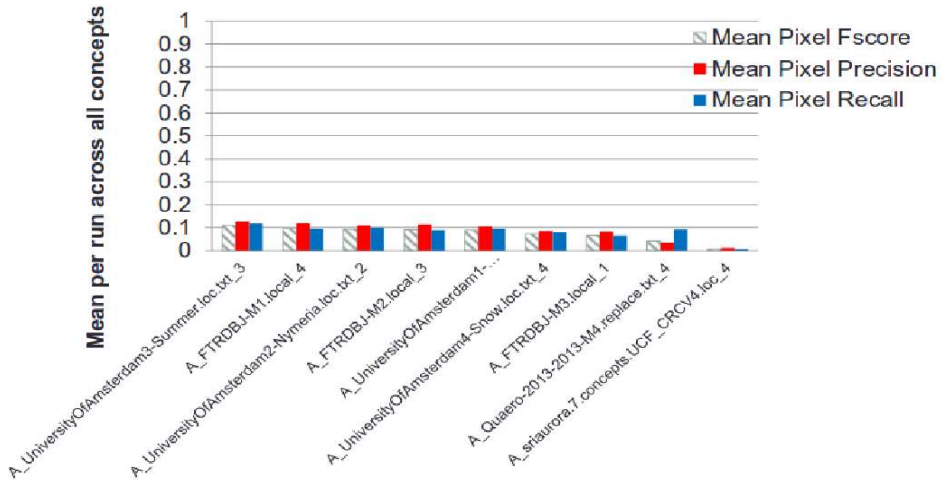


Figure 12: Concept Localization: TP vs FP I-frames by run

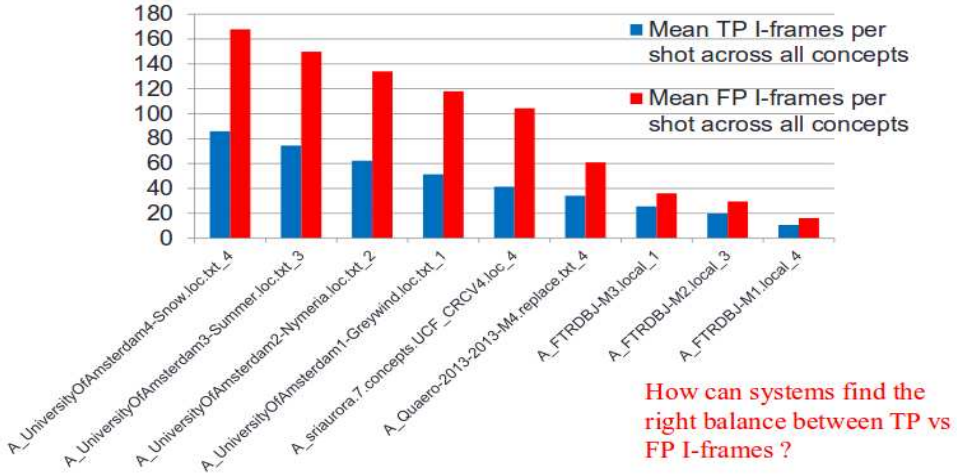


Figure 13: Concept Localization: Temporal localization by concept

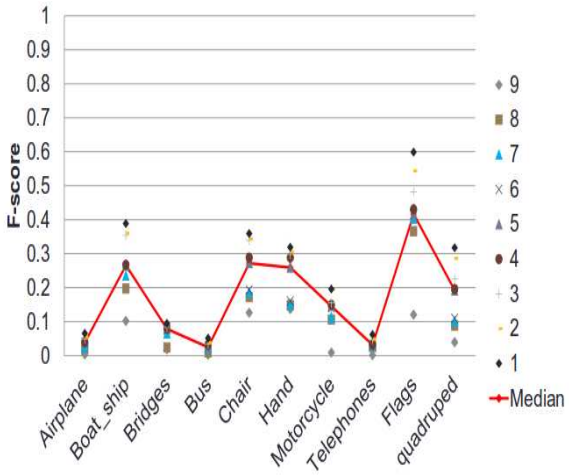


Figure 14: Concept Localization: Spatial localization by concept

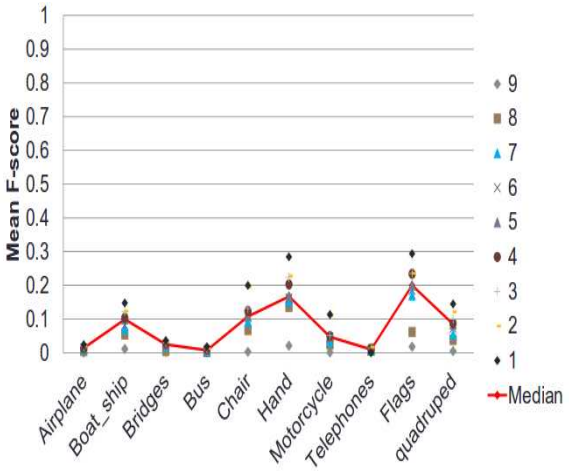
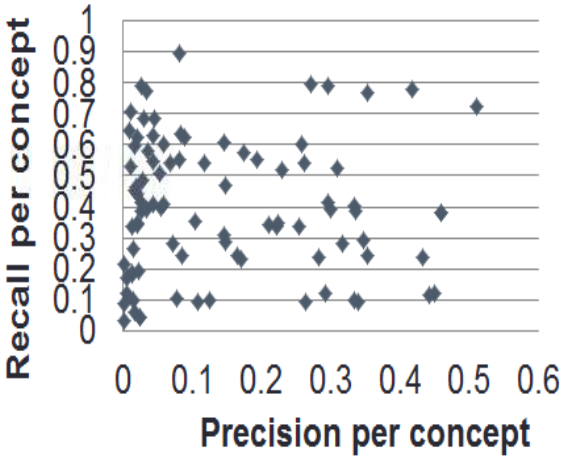


Figure 15: Concept Localization: temporal precision and recall per concept for all teams

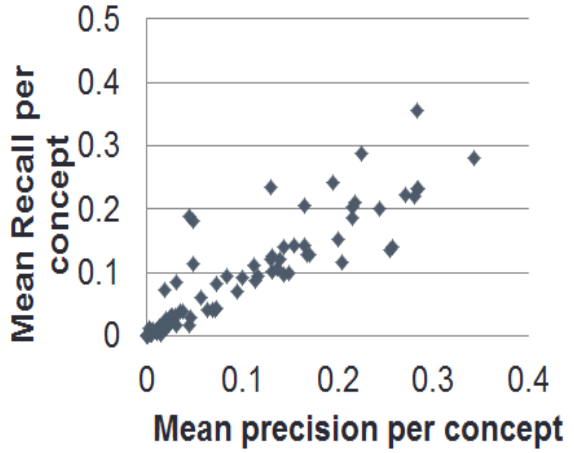


very low maximum scores for 4 concepts in both localization types: "Airplane", "Bridges", "Bus", "telephones" compared to the other 6 concepts which scores are spread among the 9 submitted runs by at least 0.1.

To visualize the distribution of recall vs precision for both localization types we plotted the results of recall and precision for each submitted concept and run in Figures 15 and 16 for temporal and spatial localization respectively. We can see in Figure 15 that the majority of systems submitted a lot of non-target I-frames achieving high recall and low precision while very few found a balance. An interesting observation in Figure 16 shows that systems are good in submitting an accurate approximate bounding box size which overlaps with the ground truth bounding box coordinates. This is indicated by the cloud of points in the direction of positive correlation between the precision and recall for spatial localization.

Finally, to summarize some observations after running the SIN task in 2013 we can conclude that training type A is dominating the submissions while training type E & F still very few with zero submissions for types B, C, & D. Number of unique shots found is less than 2012. Concept-pairs subtask is very challenging for systems with baseline runs in many cases are still better than regular runs. In addition, no teams submitted any results for feature sequence in the concept-pairs runs. For localization subtask, finding the I-frames only was easier than finding the correct

Figure 16: Concept Localization: spatial precision and recall per concept for all teams



bounding box around the concepts in the I-frames and systems can find a good approximate bounding box size that overlaps with the ground truth box but still not with high precision.

In regard to site experiments, the following is a non-exhaustive summary of the main highlights in the submitted papers: There is a focus on robustness and merging many different feature representations. Some sites applied spatial pyramids, improved bag of word approaches, Fisher/super-vectors, VLADs (vector of locally aggregated descriptors), and VLATs (Vectors of Locally Aggregated Tensors). There was some experiments in audio analysis, consideration of scalability issues, improved rescoring methods and use of semantic features. Also, there is work on the kernel size parameter of the SVM-RBF kernel, no annotation conditions included use of socially tagged videos or images and develop strategies for positive example selection. Few sites started exploring the deep convolutional neural networks (deep learning).

For detailed information about each participating research team experiments, results and their conclusions, please see the workshop notebook papers: `www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html`.

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.

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 what will likely be a multi-year effort using 464 h of the BBC soap opera EastEnders. Two hundred forty-four 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, and a collection of queries that delimit a person, object, or place entity in some example video, locate for each query the 1000 shots most likely to contain a recognizable instance of the entity. Each query consisted of a set of

- a brief phrase identifying the target of the search
- 4 example frame images drawn at intervals from videos containing the item of interest. For each frame image:
  - a binary mask of an inner region of interest within the rectangle
- an indication of the target type taken from this set of strings (OBJECT, PERSON)

Topics

NIST viewed every 10th test video and developed a list of recurring objects, people, and locations. Thirty test queries (topics) were then created. Candidate topic targets were chosen to exhibit various kinds of variation, including:

- inherent - boundedness, size, rigidity, planarity, mobility, ...
- locale - mutiplicity, variability, complexity, ...
- camera view - distance, angle, lighting, ...

Half of the topics looked for stationary objects and half for non-stationary objects or people. Of the four topics looking for people, two concerned named characters whose names were provided and two concerned unnamed extras. The guidelines for the task allowed the use of metadata assembled by the EastEnders fan community as long as this use was documents by participants and shared with other teams.

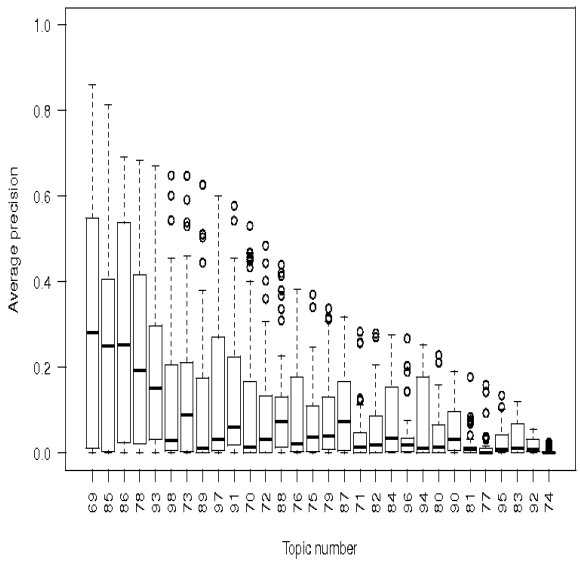
4.3 Evaluation, Measures

Each group was allowed to submit up to 4 runs and in fact 22 groups submitted 65 automatic and 9 interactive runs (using only the first 24 topics). Each interactive search was limited to 15 min.

Shots from which topic example images were taken, were filtered from all submissions. Then the submissions were pooled and 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. All topic pools were judged down to at least rank 120 (on average 253, maximum 460) resulting in 209 302 judged shots (in 600 person-hours). 13 907 clips (on average 463.6 per topic) contained the topic target (6.6 %).

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 measured in the pilot.

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



4.4 Results

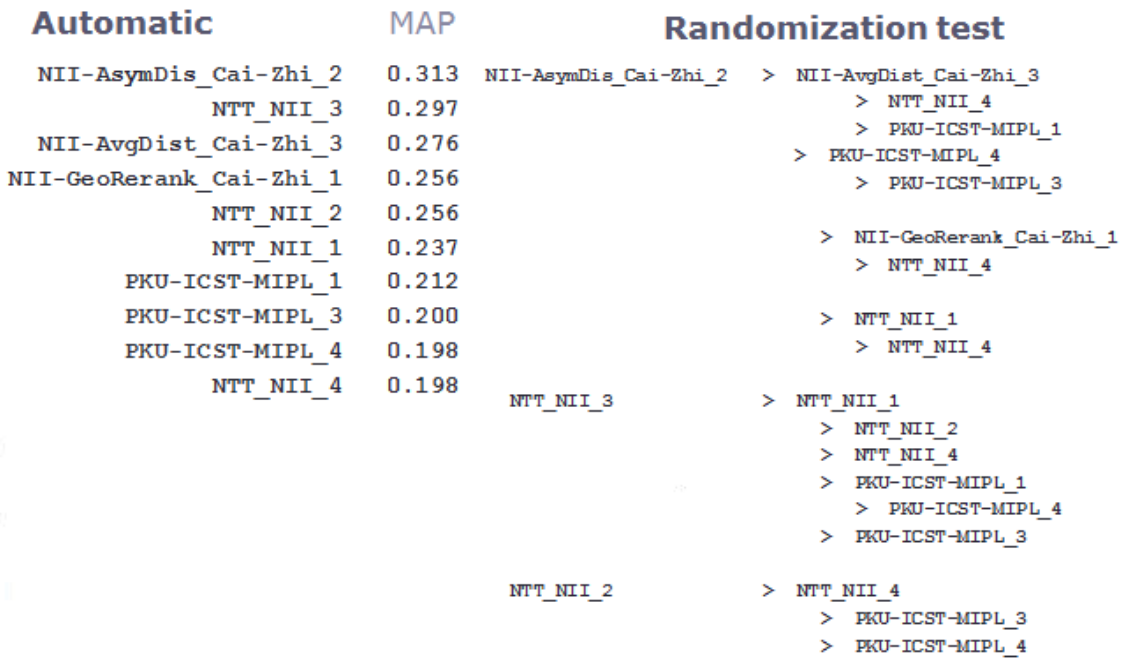
Figure 17 shows the distribution of automatic run scores (average precision) by topic as a boxplot. Topics are sorted by maximum score with the best performing topic at the left. Median scores vary from about 0.3 down to almost 0.0. Per topic variance varies as well with the largest values being associated with the topics that have the best performance. Many factors might be expected to affect topic difficulty. All things being equal one might expect targets with less variability to be easier to find. Rigid, stationary objects would fall into that category. In fact for the automatic runs topics with targets that are stationary, rigid objects make up 8 of the 15 with the best scores, while such targets make up only 3 of the bottom 15. Figure 18 documents the raw performance of the top 10 automatic runs and the results of a randomization test (Manly, 1997) some light on which differences in the ranking are likely to be statistically significant ($p < 0.05$). The right angled bracket indicates a significant difference.

In Figure 19, a boxplot of the interactive runs' performance, the best median is actually slightly below that for the automatic runs. Topics with targets that are stationary, rigid objects make up 5 of the 12 with the best scores, but such targets also make up 4 of the bottom 12 topics. Figure 20 documents the raw per-

Table 3: Instance search pooling and judging statistics

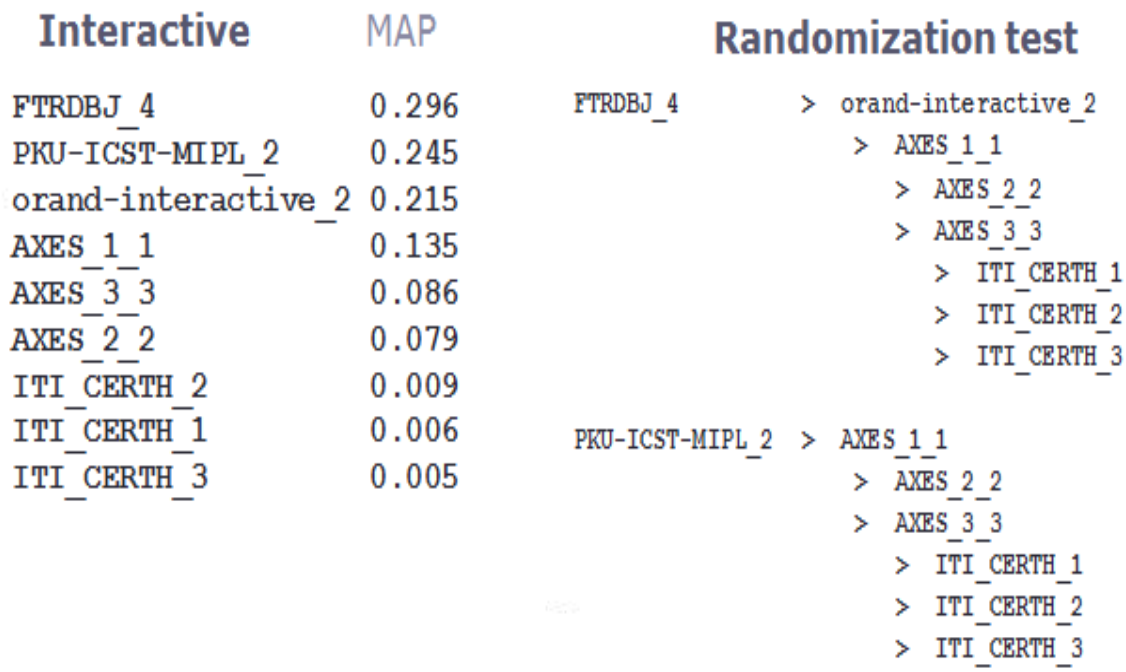
Topic number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number relevant	% judged that were relevant
9069	66 478	20 269	30.5	460	9593	47.3	2290	23.9
9070	65 452	28 565	43.6	240	7184	25.1	735	10.2
9071	65 549	27 920	42.6	180	5594	20.0	31	0.6
9072	65 915	29 988	45.5	300	9004	30.0	261	2.9
9073	65 573	27 991	42.7	280	7366	26.3	673	9.1
9074	65 139	32 323	49.6	180	6563	20.3	97	1.5
9075	65 473	33 608	51.3	160	5742	17.1	78	1.4
9076	65 745	26 867	40.9	360	10066	37.5	825	8.2
9077	65 326	34 641	53.0	120	4777	13.8	31	0.6
9078	66 231	24 029	36.3	400	8346	34.7	876	10.5
9079	64 580	25 513	39.5	240	6440	25.2	385	6.0
9080	65 236	30 133	46.2	220	7021	23.3	250	3.6
9081	64 489	25 568	39.6	240	7202	28.2	211	2.9
9082	64 721	31 225	48.2	160	5298	17.0	61	1.2
9083	65 336	30 016	45.9	300	9563	31.9	115	1.2
9084	65 711	29 419	44.8	180	5730	19.5	28	0.5
9085	65 966	26 109	39.6	300	6708	25.7	440	6.6
9086	65 837	21 813	33.1	280	5370	24.6	759	14.1
9087	64 785	30 486	47.1	180	5703	18.7	25	0.4
9088	66 664	26 852	40.3	340	8643	32.2	1605	18.6
9089	65 423	27 553	42.1	280	7790	28.3	1265	16.2
9090	64 680	24 525	37.9	200	4948	20.2	363	7.3
9091	65 465	24 470	37.4	320	7598	31.1	761	10.0
9092	66 069	22 767	34.5	260	5972	26.2	164	2.7
9093	62 247	31 191	50.1	200	5896	18.9	70	1.2
9094	61 922	27 120	43.8	240	6789	25.0	163	2.4
9095	62 617	28 770	45.9	240	7605	26.4	439	5.8
9096	62 017	30 118	48.6	220	6637	22.0	161	2.4
9097	62 571	25 486	40.7	280	7412	29.1	251	3.4
9098	62 749	29 301	46.7	240	6742	23.0	383	5.7

Figure 18: INS: top automatic run rankings with randomization tests results



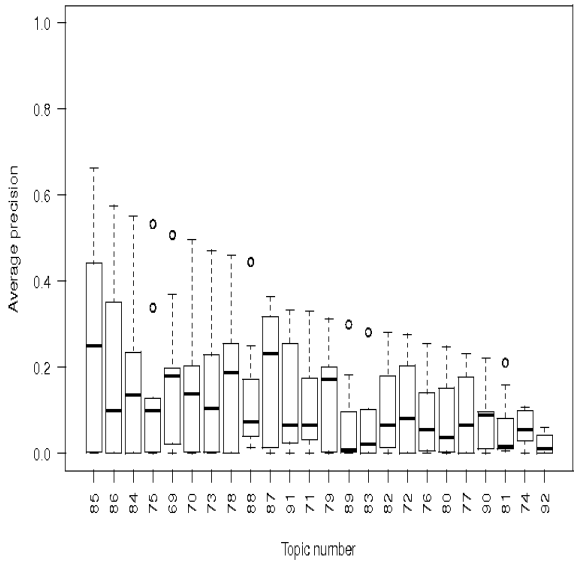
Automatic	MAP
NII-AsymDis_Cai-Zhi_2	0.313
NTT_NII_3	0.297
NII-AvgDist_Cai-Zhi_3	0.276
NII-GeoRerank_Cai-Zhi_1	0.256
NTT_NII_2	0.256
NTT_NII_1	0.237
PKU-ICST-MIPL_1	0.212
PKU-ICST-MIPL_3	0.200
PKU-ICST-MIPL_4	0.198
NTT_NII_4	0.198

Figure 20: INS: top interactive run rankings with randomization tests results



Interactive	MAP
FTRDBJ_4	0.296
PKU-ICST-MIPL_2	0.245
orand-interactive_2	0.215
AXES_1_1	0.135
AXES_3_3	0.086
AXES_2_2	0.079
ITI_CERTH_2	0.009
ITI_CERTH_1	0.006
ITI_CERTH_3	0.005

Figure 19: INS: Boxplot of interactive runs - average precision by topic



formance of the top interactive runs and the results of a randomization test (Manly, 1997) to shed some light on which differences in the ranking are likely to be statistically significant ($p < 0.05$). The right angled bracket indicates a significant difference.

The relationship between the two main measures - effectiveness (mean inferred average precision) and mean elapsed topic processing time is depicted in Figure 21. Higher effectiveness does not require more processing time. The relationship between maximum average precision for a topic and the number of true positives for that topic found in the test collection can be seen in Figure 22. Although there appears to be some tendency for some topics with larger numbers of true positives to get higher scores, the relationship is not consistent.

For detailed results please see the online workshop notebook (TV13Notebook, 2013).

Regarding approaches taken, systems typically processed the test video to choose keyframes and then analyzed each keyframe using a form of local scale-invariant feature transforms (SIFT) together with global features. Matching of the topic to the test clips was carried out using object recognition based on keypoint matches, bag-of-visual-words (BovW) clustering of keypoints to a codebook with similarity function, and spatial verification. Fusion of scores fol-

Figure 21: INS: average elapsed time versus mean average precision

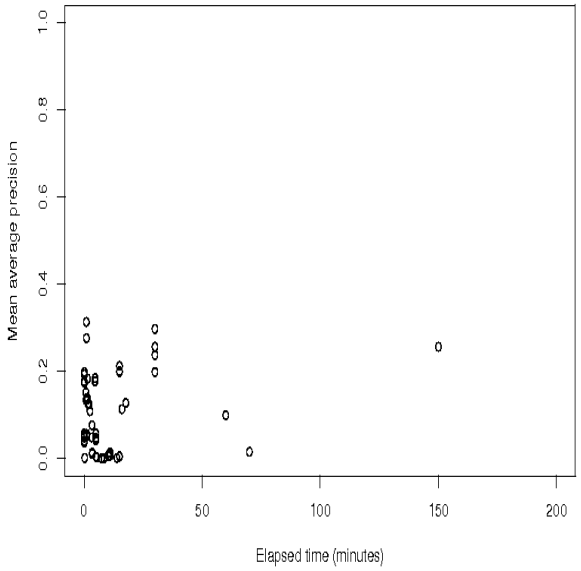
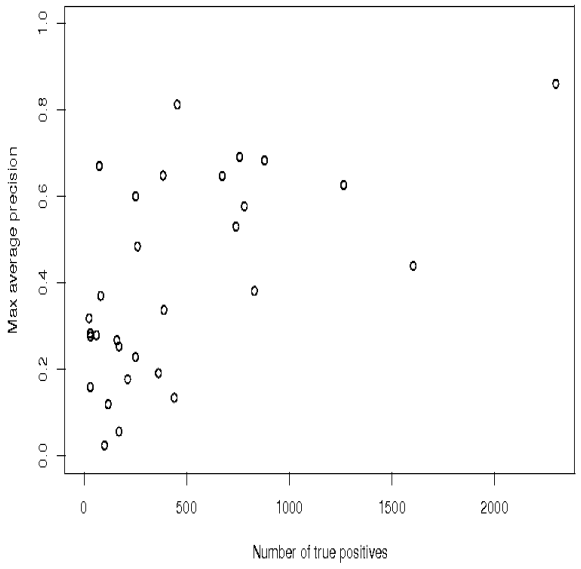


Figure 22: INS: true positives per topic versus maximum average precision



lowed.

Issues explored included how to exploit the focus versus background of the topic example images (University of Amsterdam, VIREO:City University of Hong Kong), the effect of adding extra sample images from Internet sources (AXES:Access to Multimedia), and different levels of fusion, combining different feature types (local, global)(CEA, University of Sheffield, Beijing University of Posts and Telecommunications), Vlad quantization (AXES, ITI-CERTH:Informatics and Telematcis Institute Greece), combining multiple keypoint detectors and multiple descriptors (NII:National Institute of Informatics Japan, NTT:Nippon Telegraph and Telephone). The AXES team experimented with finding additional faces using Google image search to enhance the training data. Orange Labs Beijing incorporated a face classifier which helped with some topics at a cost for processing time.

Various groups experimented with system architectures and efficiency. TNO used Hadoop to speed up their searches. Johanneum Research (JRS) employed a graphic processing unit (GPU) for object search. The Multimedia and Intelligent Computing Lab at Tongji University team implemented hybrid parallelization using GPUs and map/reduce. A number of systems incorporated techniques from text information retrieval including inverted files for fast lookup, use of collection statistics (BM25 weighting enhancements NTT-NII), and pseudo-relevance feedback (Peking University, NTT-NII, IAD-DCU:Dublin City University.

Interactive (human-in-the-loop) experiments were carried out by several teams. For the Orange Labs Beijing team their interactive runs outperformed their automatic runs (due to multiple feedback cycles). Similarly for the Peking University team. The AXES group looked at fusion of query-time subsystems (closed captions, Google image visual model, face recognition, object/location retrieval and their experiments focused on different user types. Three interactive runs from ITI-CERTH found Vlad quantization outperformed BovW and that their user interface benefited from a scene segmentation module that linked related shots.

For detailed information about the experiments each participating research team performed and their conclusions, please see the workshop notebook papers: `www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html`.

5 Multimedia event detection

The 2013 Multimedia Event Detection (MED) evaluation was the third evaluation of technologies that search multimedia video clips for complex events of interest to a user. The 2013 included many important changes:

- Events tested: 10 new events were added for the Ad-Hoc evaluation.
- Evaluation conditions: the first Ad-Hoc event evaluation task was supported which tested systems on an additional 10 new events.
- A new 0-video exemplar event training condition was introduced and the maximum number of event training exemplars was reduced to 100 from 130.
- Developers were asked to build a single best system and then asked to run a prescribed set contrastive conditions.
- Indexing collections: the MED Progress Collection, which is 3722 h in duration, was used again this year as planned but a 1/3 subset (referred to as PROGSub) was introduced as a smaller test collection for new participants.
- The primary performance metric was changed to Mean Average Precision.

An event for MED:

- is a complex activity occurring at a specific place and time;
- involves people interacting with other people and/or objects;
- consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and semantic relationships to the overarching activity;
- is directly observable.

A user searching for events 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 is an mnemonic title for the event.
- An event definition which is a textual definition of the event.
- An event explication which is a textual listing of some attributes that are often 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 is not an exhaustive list nor is it to be interpreted as required evidence.
- An evidential description which is 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 is not an exhaustive list nor is 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 are illustrative in the sense they help form the definition of the event but they do not demonstrate all the inherent variability or potential realizations.

Developers built Pre-Specified event systems where knowledge of the event(s) was taken into account during generation of the metadata store for the test collection. In 2013, the initial Ad-Hoc event task was conducted where the metadata store generation was completed before the events where revealed.

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 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, 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.

Table 4: MED '13 Pre-Specified Events

Testing Events
MED'11 event re-test
Birthday Party
Changing a vehicle tire
Flash mob gathering
Getting a vehicle unstuck
Grooming an animal
Making a sandwich
Parade
Parkour
Repairing an appliance
Working on a sewing project
MED'12 event re-test
Attempting a bike trick
Cleaning an appliance
Dog show
Giving directions to a location
Marriage proposal
Renovating a home
Rock climbing
Town hall meeting
Winning a race without a vehicle
Working on a metal crafts project

MED participants were provided the data as specified in the HAVIC data section of this paper. The MED '13 Pre-Specified event names are listed in Table 4 and Table 5 lists the MED '13 Ad-Hoc Events.

5.2 System task

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

- Events: either all 20 Pre-Specified events (PS13) and/or all 10 Ad-Hoc events (AH13).
- Subsystems: a single full system (FullSys) and up to 4 reduced input systems that included: Optical Character Recognition (OCR) only, Automatic Speech Recognition (ASRSys) only, Non-OCR visual (VisualSys) only, Non-ASR audio (AudioSys) only.
- Test collection: either the full Progress collection (PROGFull) or the 1/3 subset of the Progress collection (PROGSub).
- Event exemplar training: 100 Ex (100 positive and 50 miss clips), 10 Ex (10 positive and 10 miss clips), and 0 Ex (0 positive and 0 miss clips).

Table 5: MED '13 Ad-Hoc Events

Testing Events
Beekeeping
Wedding shower
Non-motorized vehicle repair
Fixing musical instrument
Horse riding competition
Felling a tree
Parking a vehicle
Playing fetch
Tailgating
Tuning musical instrument

Full participation would mean teams would submit 30 runs, (5 subsystems * 2 event sets * 3 event exemplar conditions).

For each event search a system generates:

- A Score for each search collection clip: A probability value between 0 (low) and 1 (high) representing the system's confidence that the event is present in the clip.
- A Detection Threshold for the event: A probability value between 0 and 1 - an estimation of the detection score at or above which the system will assert that the event is detected in the clip.
- The event agent execution time: The number of seconds used to search for the event in the metadata store.

System developers also reported the hardware components used and computation times of the metadata generation, event query generation, and event search modules as well as the metadata store size.

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

5.3 Evaluation, measures

System output will be evaluated by how well the system retrieves and detects MED events in evaluation search video metadata and by the computing resources used to do so. The determination of correct detection was at the clip level, i.e. systems will provide a response for each clip in the evaluation search video set. Participants must process each event independently in order to ensure each event was be tested independently.

The primary evaluation measures for performance will be Mean Average Precision.

There are three primary measures for computational speed expressed as real-time factors. Real-time factor (RT) is the total processing time divided by the number of hours of video in the test collection. The three aspects computed were: (1) Metadata Generation Processing Speed, (2) Event Query Generation Processing Speed, and (3) Event Search Processing Speed.

5.4 Results

18 teams participated in the MED '13 evaluation, 6 teams were new. All teams participated in the Pre-Specified (PS), 100 Exemplar (100Ex) test processing all 20 events. Sixteen teams participated in the Ad-Hoc (AH) task. Three teams chose to process the PROGSub Subset.

The MED13 evaluation re-used the MED Progress Evaluation collection. Since the Progress set will be used through 2015 MED evaluations, protecting the statistic of the Progress set is of the utmost importance, NIST reported only Mean Average Precision to prevent revealing statistics of the Progress set for each run.

Table 6 presents the MAP (averaged over events) for the Pre-Specified event and Ad-Hoc task submissions for all training exemplar conditions and system/sub-systems. The two box plots in Figure 23 shows the same data and illustrates many findings.

First, the range of MAPs for Pre-Specified vs. Ad-Hoc events were surprisingly similar despite the difference in event population and evaluated systems. For the required PS,100Ex condition, Full system (FullSys) subsystems, the MAPs ranged from 0.2 % to 34.6 % with a median of 23.9 % across teams. For the AH, 100Ex, FullSys runs, the MAPs ranged from 0.2 % to 40.7 % with a median of 24.4 % across teams.

Second, systems were able to improve performance using additional event exemplars. There was a 330 % and 130 % relative median MAP improvement going from 0Ex to 10Ex and 10Ex to 100Ex runs respectively for the PS-FullSys runs. For the AH runs, there was a larger relative improvement of median MAPs between 0Ex and 10Ex runs of 946 % whereas a lower relative improvement (79 %) between 10Ex and 100Ex runs. The relative difference between the PS vs. AH for the 0Ex to 10Ex change in performance, tracks the lower median MAP scores for AH, 0Ex runs.

Third, the MAPs for the Visual subsystems were closer to the FullSys runs indicating systems made most use of their system's visual processing components to perform MED. The visual system alone did not account for all the performance. The Audio subsystem, and to a lesser degree the ASR and OCR subsystems, contributed as well.

Participants were asked to report three runtimes so that three speeds could be computed. The speeds are: (1) Metadata Generation Processing Speed (MGPS), (2) Event Query Generation Processing Speed, and (3) Event Search Processing Speed.

Figure 24 presents two measurements of metadata generation speeds measured in multiples of realtime of video files). Figure 24 graph A shows the MGPS for the Search Collection for 18 teams (though not all teams provided measurements for all subsystems). Since the system populations are unbalanced, the median speeds are a more reasonable statistics to analyze than the mean. As expected, RTs for the FullSys and VisualSys subsystems that used exemplar training (100Ex and 10EX) required the most runtime ranging from 0.04 to 0.15 median RT. The median RTs for other subsystem/Ex combinations ranged from 0.004 for the Audio, 0Ex condition to 0.052 for the ASRSys, 100Ex condition.

Figure 24 graph B shows the Event Query Generation Processing speed based on the metadata extraction speeds on the Event Background Collection for reporting teams. In general, the speeds for metadata generation on the event background data were slower than for the search collection. This is expected because teams were instructed to use a Commercial Off-the-Shelf PC to perform Event Query computation. The general trends with regard to subsystems and exemplar training conditions are analogous to the metadata generation speed for the search collection.

Figure 25 shows the Event Search Processing speeds for the Pre-Specified and Ad-Hoc events. The speeds presented here are the event-averaged speeds for the reporting teams. Median search execution RT speeds are similar for both Pre-Specified and Ad-Hoc events. For the pre-specified events, the slowest median search speed was $1.30\text{x}10^{-3}$ and for the FullSys/100Ex condition while the fastest was $9.32\text{x}10^{-6}$ for the OCRSys/0Ex condition.

There is evidence of improvement between the MED '12 and MED '13 evaluation. Six teams, (BBNVISER, CMU, GENIE, IBMCU, Sesame, SRIAURORA), participated in the Pre-Specified, 10Ex task which was the identical test with respect to the event training condition and test videos. The relative MAP improvements were 141 %, 37 %, 84 %, -46 %, 41 %, and 22 %.

In summary, 18 teams participated in the MED '13 evaluation. All teams participated in the Pre-Specified (PS), 100 Exemplar (100Ex) test processing all 20 events. 16 teams participated in the Ad-Hoc (AH) task. 3 teams chose to process the PROG-Sub Subset. The division of the types of information sources brought to bear for the systems show that by far, the VisualSys content alone provided the majority of evidence to detect events but that the other components, non-ASR Audio, ASR, and OCR, provided additional evidence for detection. There is evidence of improvement for five the six teams that participated in matching evaluation conditions between MED '12 and MED '13.

TRECVID '14 evaluation will include the MED Track. Proposed changes include the introduction of 10 new Ad-Hoc events, support for a pseudo-relevance feedback evaluation condition, and introduce a larger test collection.

6 Multimedia event recounting

The 2013 Multimedia Event Recounting (MER) evaluation was the second evaluation of technologies that recount the multimedia video events detected by MED systems.

In more detail, the purpose, of the 2013 Multimedia Event Recounting (MER) track, was to stimulate the development of technologies that state the *important evidence* that led a Multimedia Event Detection (MED) system to decide that a multimedia clip contains an instance of a specific event and to allow human users to rapidly and accurately find clips of interest via the recountings. The 2013 TRECVID MER evaluation assessed just the recounting of the evidence.

The 2013 evaluation of MER consisted of three metrics, described briefly here and in more detail later. The first, *Percent Recounting Review Time*, is the total time for a judge to assess the recounting, divided by the total duration of the clips to be assessed. The second, *Accuracy*, is the fraction of the clips where the recounting alone allowed the judge to determine whether the clip contained an instance of the event of interest. The third, *Precision of the Observation Text* is the mean grade (across judges) on a five-point scale of "A: Excellent" through "F: Fails." The choices on that five-point scale were as-

Table 6: MED '13 Mean Average Precisions for Pre-Specified Event and Ad-Hoc Event Systems

			PROGAll					PROGSub Only	
			MAP					MAP	
			FullSys	ASRSys	AudioSys	OCRSys	VisualSys	FullSys	VisualSys
AH	100Ex	BBNVISER	35.9 %	8.4 %	15.4 %	5.1 %	27.2 %		
		CERTH-ITI						9.6 %	9.8 %
		CMU	40.1 %	6.2 %	16.7 %	3.9 %	32.0 %		
		Genie	22.9 %	4.5 %	10.6 %		19.2 %		
		IBM-Columbia	3.2 %		0.2 %		3.1 %		
		INRIA-LEAR	40.7 %	0.9 %	12.6 %	1.1 %	33.5 %		
		MediaMill	28.7 %		6.0 %		27.3 %		
		NII	28.3 %		9.1 %		22.9 %		
		ORAND	4.4 %				4.4 %		
		PicSOM	0.7 %		0.2 %		0.7 %		
		SRIAURORA	27.4 %	4.3 %	10.2 %	4.6 %	23.3 %		
		Sesame	28.9 %	4.2 %	6.0 %	0.2 %	26.7 %		
		TNO						9.0 %	5.8 %
		TokyoTechCanon	25.9 %						
		UMass	5.6 %						
		VisQMUL	0.2 %		0.2 %		0.2 %		
	10Ex	BBNVISER	16.8 %	4.3 %	6.0 %	2.2 %	12.9 %		
		CERTH-ITI						3.0 %	3.0 %
		CMU	24.1 %	2.7 %	9.3 %	1.1 %	18.6 %		
		Genie	13.1 %	2.4 %	3.7 %		8.7 %		
		IBM-Columbia	1.7 %		0.2 %		1.8 %		
		MediaMill	16.1 %		2.7 %		15.5 %		
		PicSOM	2.3 %		0.2 %		2.4 %		
		SRIAURORA	16.6 %	4.3 %	5.6 %	4.6 %	11.8 %		
		Sesame	14.1 %	1.4 %	2.7 %	0.2 %	15.1 %		
		VisQMUL	0.2 %		0.2 %		0.4 %		
	0Ex	BBNVISER	8.9 %	2.6 %	0.6 %	3.1 %	5.7 %		
		CMU	10.8 %	3.1 %	0.2 %	3.0 %	5.7 %		
		Genie	0.5 %	0.5 %	0.5 %		1.3 %		
		IBM-Columbia	1.3 %		0.2 %		1.6 %		
		SRIAURORA	1.5 %	4.3 %	0.2 %	4.6 %	0.6 %		
		Sesame	3.0 %	2.3 %		2.2 %	1.5 %		
		TNO						0.3 %	
		UMass	1.0 %	2.5 %		4.3 %	0.5 %		
		VisQMUL	+0.2 %		0.2 %		+0.2 %		
PS	100Ex	BBNVISER	33.0 %	7.6 %	12.0 %	4.8 %	28.2 %		
		CERTH-ITI						10.5 %	10.2 %
		CMU	30.6 %	7.8 %	12.6 %	3.1 %	26.4 %		
		Genie	23.3 %	0.6 %	7.8 %		19.9 %		
		IBM-Columbia	3.0 %		0.3 %		3.0 %		
		INRIA-LEAR	34.6 %						
		MediaMill	28.1 %		5.9 %		26.0 %		
		NII	28.2 %		7.1 %		24.9 %		
		ORAND						0.6 %	
		PicSOM	6.4 %				6.4 %		
		SRIAURORA	24.7 %	3.0 %	0.8 %	3.7 %	22.5 %		
		Sesame	27.6 %	4.0 %	5.9 %	0.2 %	26.1 %		
		SiegenKobeMuro	4.1 %				4.1 %		
		TNO						10.3 %	5.2 %
		TokyoTechCanon	24.5 %						
		UMass	13.0 %						
		VIREO	26.5 %				25.5 %		
		VisQMUL	0.2 %						
	10Ex	BBNVISER	16.6 %	3.5 %	4.4 %	3.2 %	13.3 %		
		CERTH-ITI						3.0 %	3.0 %
		CMU	12.6 %	2.0 %	4.7 %	0.8 %	11.2 %		
		Genie	10.4 %	0.3 %	2.6 %		10.3 %		
		IBM-Columbia	2.2 %		0.3 %		2.2 %		
		MediaMill	15.0 %		2.6 %		14.0 %		
		PicSOM	3.2 %				3.2 %		
		SRIAURORA	13.7 %	3.0 %	0.9 %	3.7 %	12.4 %		
		Sesame	10.3 %	1.4 %	2.6 %	0.2 %	11.6 %		
		VisQMUL	0.2 %						
	0Ex	BBNVISER	5.2 %	1.4 %	0.5 %	2.8 %	3.5 %		
		CMU	3.7 %	1.8 %	0.3 %	2.1 %	2.4 %		
		Genie	1.3 %	1.7 %	1.1 %		1.0 %		
		IBM-Columbia	1.6 %		0.2 %		1.8 %		
		SRIAURORA	7.0 %	3.0 %	0.2 %	3.7 %	6.5 %		
		Sesame	2.4 %	1.7 %		2.3 %	1.3 %		
		TNO						0.4 %	
		UMass	5.6 %	2.3 %		3.3 %	5.1 %		
		VisQMUL	0.2 %						

Figure 23: MED: Pre-specified and Ad-Hoc SubSystems MAP scores by Event Training Exemplar

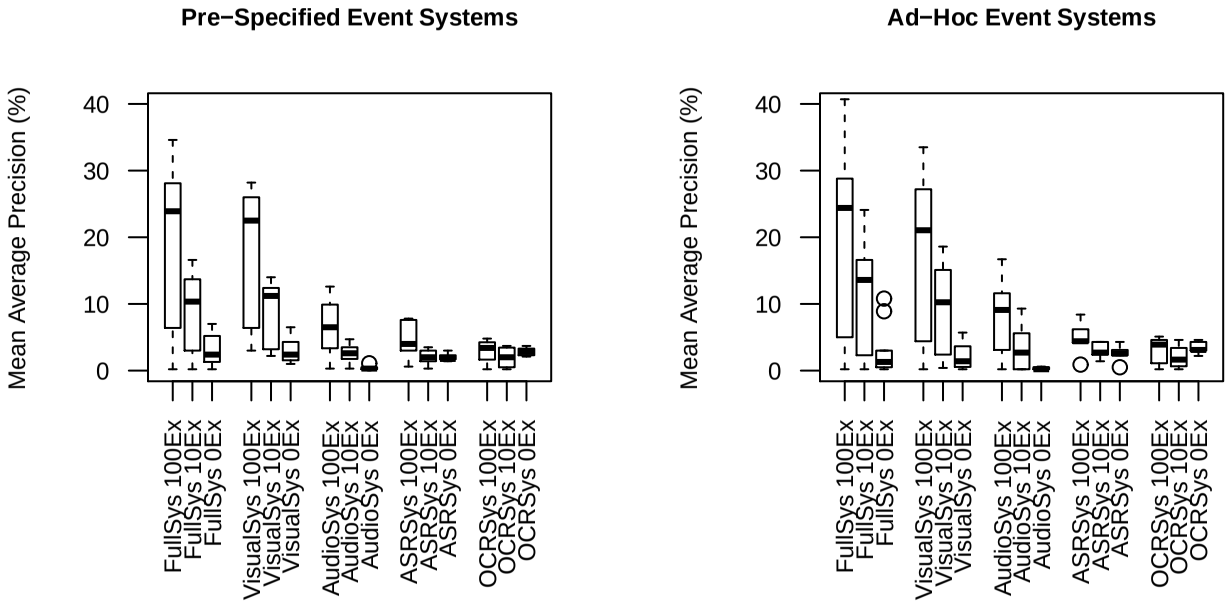


Figure 24: MED: Metadata generation speeds for the search collection and event background data sets

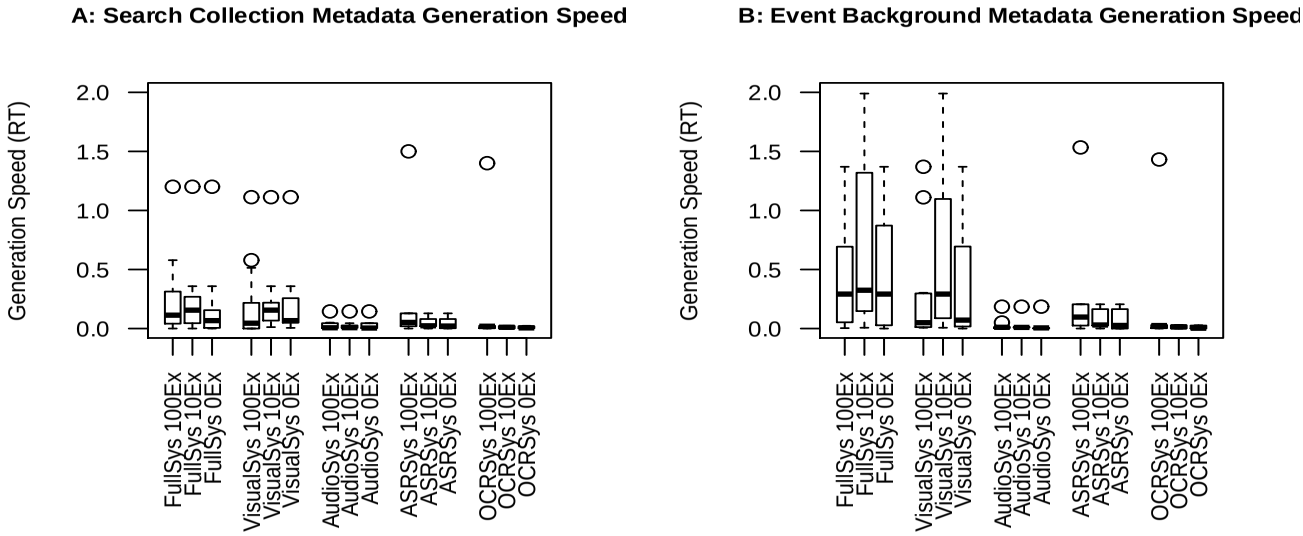
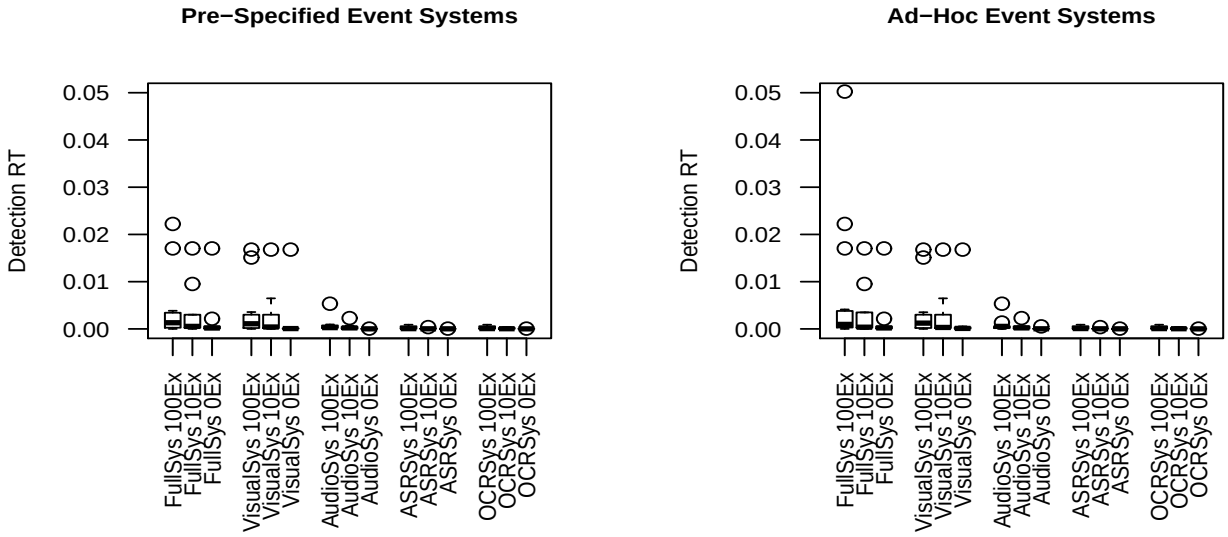


Figure 25: MED: Search speeds for the Pre-Specified and Ad-Hoc evens broken down by subsystem type and training exemplar



signed numeric values, so that the mean and other statistics could be calculated.

Each *event* was explicitly defined by an *Event Kit*. A clip that is *positive* for an event contains an *instance* of that event.

Each event in this evaluation

- is a complex activity occurring at a specific place and time;
- involves people interacting with other people and/or objects;
- consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and semantic relationships to the over-arching activity; and
- is directly observable.

Participation in MER 2013 was open to all 2013 TRECVID MED participants whose system always produced a recounting for each clip that their MED system deemed to be positive (that is, identified as being above their MED system's decision threshold for being positive).

Input data formats were as in existing HAVIC data. MER output data formats used ASCII XML text. NIST provided a rendering tool and a MER DTD schema to be used to specify and validate system output.

The systems recountings were evaluated by a panel of judges. NIST created a MER Workstation to view and judge the recountings, and NIST provided access to that workstation to the MER participants and the judges.

We are interested in recountings that state the evidence in a way that human readers find easily understandable.

6.1 Data

The MER task drew from the same data as the MED task. See the MED Data section for more information.

6.2 System task

Given an event kit and a test video clip that the team's MED system deemed to contain an instance of the event, the MER system was to produce a recounting summarizing the key evidence for the event in the clip. Evidence means observations such as scene/context, persons, animals, objects, activities,

text, non-linguistic audio, and other evidence supporting the detection of the event. Each observation was associated with an indication of the system's confidence that the observation is correct and an indication of how important the system believes the observation to be. The system indicated, in the recounting, the order in which the observations should be presented to the user, so that a system could exploit order of presentation to achieve its goals.

For each observation, the recounting was to include a list of pointers to the evidence in the clip, indicating

- temporally, where in the clip the piece of evidence occurs, and
- spatially, where in the frame the evidence occurs (if visible evidence).

We refer to these pieces of evidence (excerpts from the clip) as *snippets*. Snippets are described more fully below.

In addition, the recounting was to include a list of the source(s) of the observation, drawn from the following list:

- *video*: (not involving OCR)
- *visible_text*: (text via OCR)
- *speech*: (transcribed via ASR)
- *non_speech_audio*: (without ASR textual transcription)

Systems produced an XML element for each *observation*, and that element included attributes that gave the following information.

id a unique identifier that can be used in other XML elements to associate elements, e.g., to associate an object or person with an activity

description a textual statement of the observation (For example, if the *type* is *object*, the *description* might be *red Toyota Camry*.) The description may be used to, for example, state only what is observable (e.g., red Camry) or could, for example, also include semantic inferences (e.g., the getaway vehicle).

source(s) as described above

confidence in the range 0.0 through 1.0, with 1.0 indicating highest confidence

importance in the range 0.0 through 1.0, with 1.0 indicating highest importance

presentation_order a number (1, 2, 3, etc.)

one or more snippets

A *snippet* is a spatio-temporal pointer to the piece of evidence. It contains a *start_time* and an *end_time*, given either as times or frame numbers. If the piece of evidence is not purely auditory, the snippet also gives initial and final bounding boxes within the frame, consisting of pixel coordinates of the upper-left and lower-right corners of the bounding box, relative to the upper-left corner of the frame. The bounding box was free to encompass the entire frame, and in many systems did so. A snippet could optionally also state the source (as described above). For implementation reasons, each snippet was required to explicitly include a snippet type (audio_video, audio_only, or keyframe).

The MER Evaluation was performed on the MED 100-Ex pre-specified event condition. NIST chose, for evaluation, ten events and six clips for each of the ten events. If a system did not generate a recounting for some of the selected clips (because the MED system did not deem them to be positive), the system was evaluated on only the clips for which it had produced a recounting. Two of the systems, however, were so conservative in their MED decisions that this approach did not result in "enough" recountings being selected, and NIST chose recountings for additional clips for those two systems. These two systems are identified in the results.

The ten 2013 MER evaluation events, chosen from the MED pre-specified events, included the five evaluated in 2012:

- E022 Cleaning an appliance,
- E026 Renovating a home,
- E027 Rock climbing,
- E028 Town hall meeting, and
- E030 Working on a metal crafts project.

Five additional MED pre-specified events were evaluated as MER events for the first time in 2013:

- E007 Changing a vehicle tire,
- E009 Getting a vehicle unstuck,
- E010 Grooming an animal,
- E013 Parkour, and
- E015 Working on a sewing project.

6.3 Evaluation procedures

Using the MER workstation, the judge studied the event kit text (not the example videos) and then assessed the recounting by:

1. Reading the entire list of textual descriptions
2. Viewing/hearing all the snippets defined by the spatiotemporal pointers
3. On the basis of the recounting, classifying the clip as one of the following:
    - The clip *contains* an instance of the event
    - The clip *does not contain* an instance of the event
    - I *do not know* because the *recounting* does not allow me to tell whether the clip contains an instance of the event
    - I *do not know* because the *event kit* does not allow me to tell whether the clip contains an instance of the event

The MER workstation makes the event kit text continuously available to the judge for reference. The MER Workstation does not display the source(s) of the information—neither for each piece of evidence, nor for the snippet(s) associated with each piece of information. The stated sources of information can be used for post-hoc understanding of the system.

6.4 Measures

In this section we discuss the metrics and link to graphs showing the results. We wanted to score all teams on the same clip:event set. We we able to do so for eight of the ten systems (because they made a positive MED decision on sufficiently many clip:event pairs in common). However, for IBM there was not enough overlap with those eight teams and we chose a different set of clips for which to judge the recountings. Likewise for Vireo, we were forced to choose a different set of recountings. Therefore, the results, by team, for IBM and for Vireo probably should not be directly compared to each other or to the results for the other eight teams.

For each submission and each event, NIST measured the following characteristics of the recountings, for each system.

Percent Recounting Review Time:

The percentage of clip time the judges took to perform steps 1 through 3 above.

(*Total time needed to perform steps 1–3*) / (*Total duration of clips to be assessed*)

The results by team for Percent Recounting Review Time are shown in Figure 26 and the corresponding results by event are shown in Figure 27.

Accuracy:

The degree to which the judges assessments (step 3 above) agree with the MED ground truth.

(*Number of correctly classified clips*) / (*Number of clips to be assessed*)

A clip was scored as correctly classified if either

- the clip really is positive and the judge indicated the clip contains an instance of the event, or
- the clip really is negative and the judge indicated the clip does not contain an instance of the event.

The *number of clips to be assessed* does not count any clips where the judge indicated the *event kit* did not allow him/her to tell whether the clip contains an instance of the event.

The results by team for Accuracy are shown in Figure 28 and the corresponding results by event are shown in Figure 29.

Precision of the observation text:

The mean (across judges) of the judges' scores on the following question, which was asked for each observation: "How well does the text of this observation describe the snippet(s)?"

- A: Excellent (4 points)
- B: Good (3 points)
- C: Fair (2 points)
- D: Poor (1 point)
- F: Fails (zero points)

The metric about the precision of the observation text is intended to provide the system developers with useful feedback about their recountings.

The results by team for Precision of the Observation Text are shown in Figure 30 and the corresponding results by event are shown in Figure 31.

Comparisons among the above three metrics

A scatter plot for Percent Recounting Review Time compared to the Precision of the Observation Text is shown in Figure 32. A scatter plot for Accuracy compared to Precision of the Observation Text is shown in Figure 33. And a scatter plot for Accuracy compared to Percent Recounting Review Time is shown in Figure 34.

Figure 26: Percent Recounting Review Time, by system

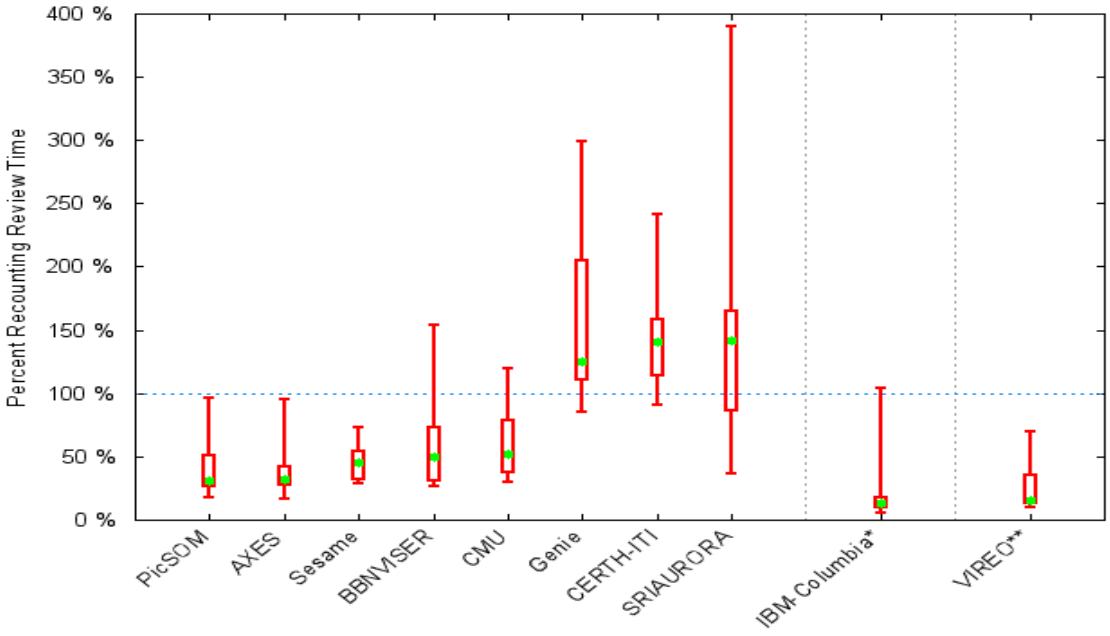


Figure 27: Percent Recounting Review Time, by event

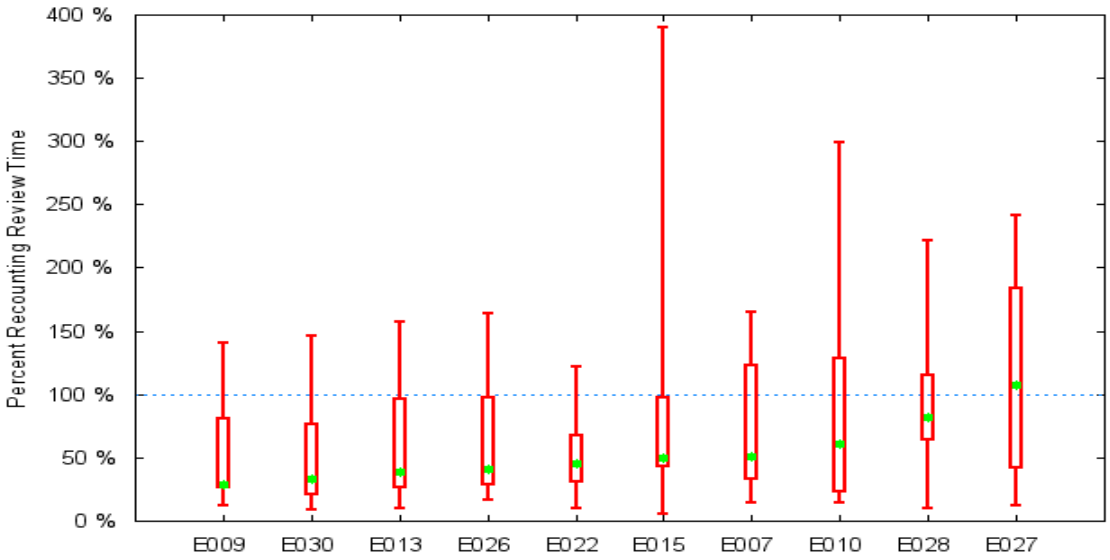


Figure 28: Accuracy, by system

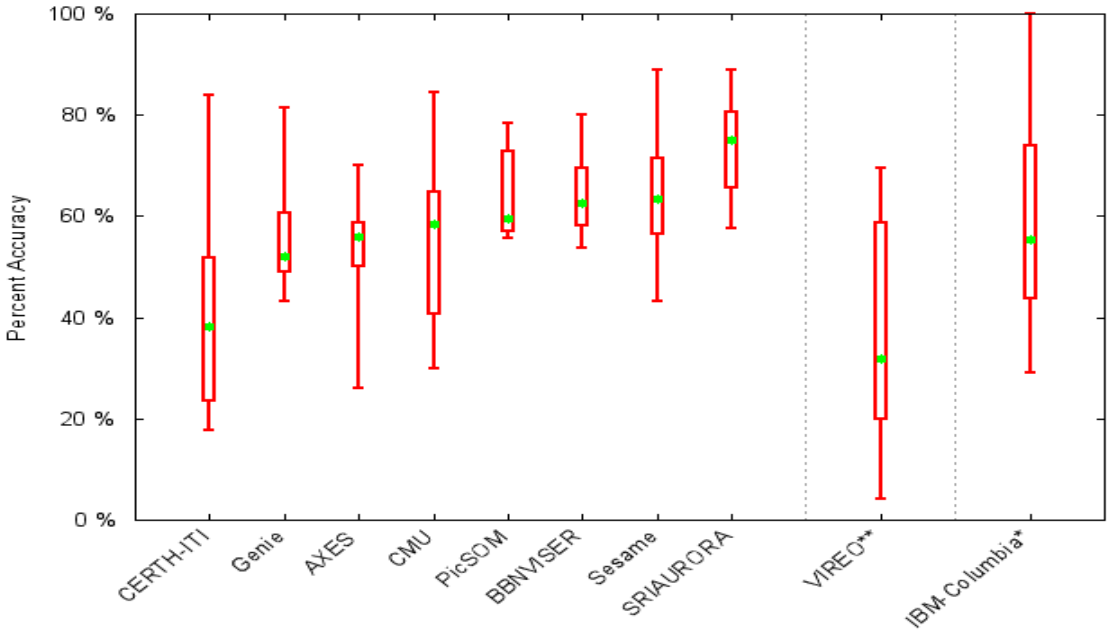


Figure 29: Accuracy, by event

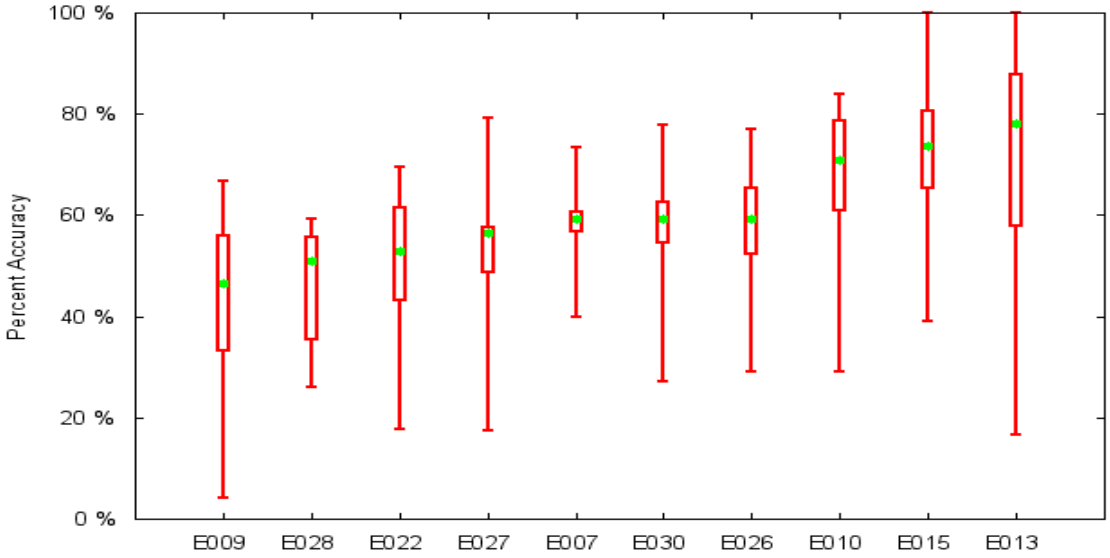


Figure 30: Precision of the Observation Text, by system

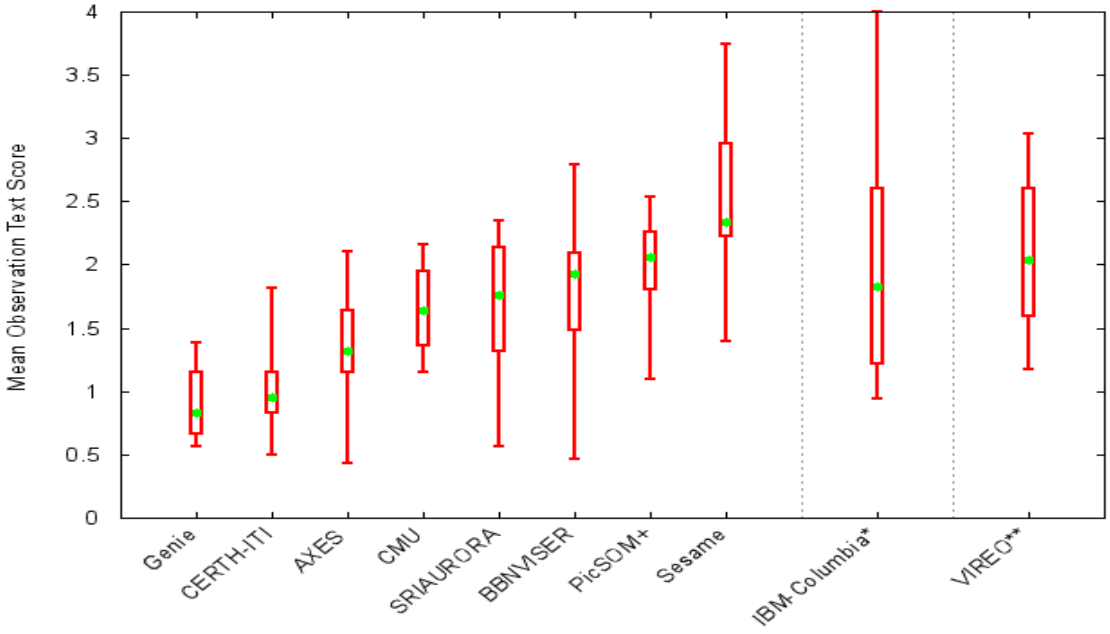


Figure 31: Precision of the Observation Text, by event ID

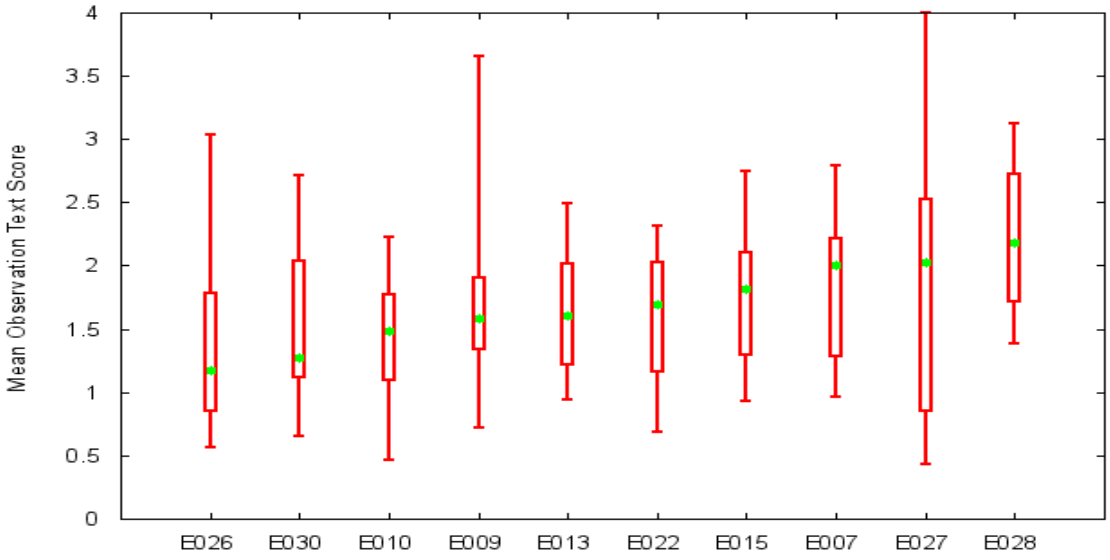


Figure 32: Precision of the Observation Text vs. Percent Recounting Review Time

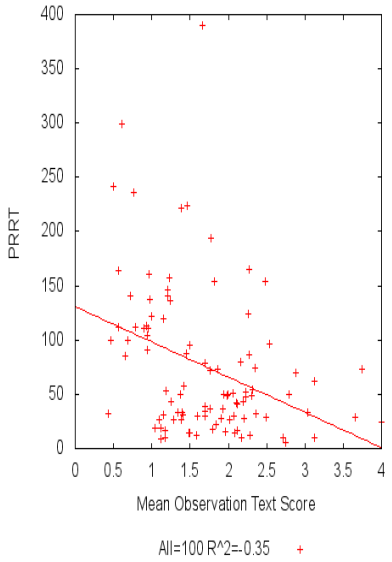


Figure 33: Accuracy vs. Precision of the Observation Text

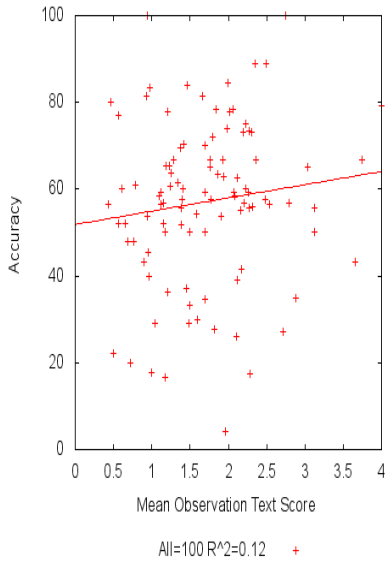
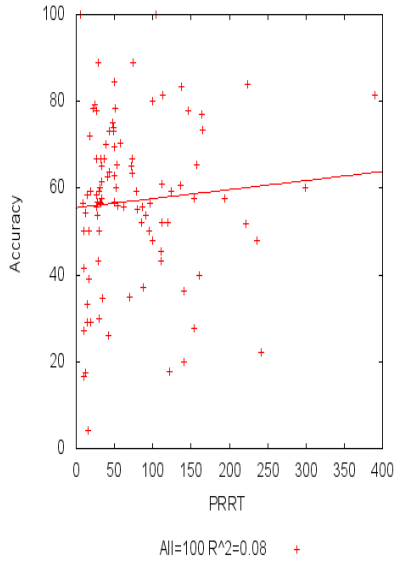


Figure 34: Accuracy vs. Percent Recounting Review Time



We have plotted a least-squares linear regression line for each of those three scatterplots. We have done so as a way of asking ourselves whether there is actually a linear relationship. Our belief in Figure 32 is that with increasing x-values (increasing Mean Observation Text Score) we see decreasing dispersion of the y-values (PRRT values). The two points in the upper-left quadrant of the plot contribute substantially to the R or R-squared values (R = 0.6 and R-squared = 0.35). However, it is not clear to us whether the relationship between x and y is linear or curvilinear in that plot. In contrast, in Figure 34 the x values appear to have no association with the y values, and the PRRT values can be clearly seen to be dense near zero, falling off rapidly to the right. That lack of linear association is also reflected in fact that the slope of the regression line is close to zero. Finally, in Figure 33, although there are many points in the middle of the graph that are near the regression line, the reader is invited to consider how a line on the diagonal from the upper-left corner of the graph to the bottom-right corner would also appear to reflect patterns in the scatter. Our parting observation about these three scatterplots is that, in each of the three, trimming the 10 % to 20 % most extreme values of the x-values and y-values would not bring out other patterns, nor would various robust regres-

sion approaches such as relplots. In short, apparent patterns in these three scatterplots may be just the eye playing tricks, and repetitions of the experiments would be needed in order to judge whether that is the case.

6.5 Results

For detailed results on each run's performance, see the on-line workshop notebook (TV13Notebook, 2013) and the workshop papers accessible from the publications webpage (TV13Pubs, 2013). That level of voluminous detail is omitted from this paper.

7 Interactive surveillance event detection

The 2013 Surveillance Event Detection (SED) evaluation was the sixth 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, Fiscus, Over, Garofolo, & Michel, 2009) and annually since. 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 2013, the evaluation re-used the 2009 test corpus and 2010 events. We also continued the Interactive SED Task introduced in 2012.

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. We have reused the same set of seven events that were selected in 2010.

For 2013, the evaluation re-used the 2009 test corpus. The test data was the Imagery Library for Intelligent Detection System's (iLIDS) (UKHO-CPNI, 2007 (accessed June 30, 2009)) Multiple Camera Tracking Scenario Training (MCTTR) data set collected by the United Kingdom's Home Office Science and Development Branch.

7.1 System task

In 2013, the Retrospective Surveillance Event Detection (rSED) and Interactive Surveillance Event Detection (iSED) tasks were 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 retrospective task addresses the need for automatic detection of events in large amounts of surveillance video. It requires application of several Computer Vision techniques, involves subtleties that are readily understood by humans, yet difficult to encode for machine learning approaches, and can be complicated due to clutter in the environment, lighting, camera placement, traffic, etc.

- The interactive task is defined as follows: given a collection of surveillance video data files (e.g. that from an airport, or commercial establishment) for preprocessing, at test time detect observations of events based on the event definition and for each return the elapsed search time and a list of video segments within the surveillance data files, ranked by likelihood of meeting the need described in the topic. Each search for an event by a searcher can take no more than 25 elapsed minutes, measured from the time the searcher is given the event to look for until the time the result set is considered final. Note that iSED is not a short latency task. Systems can make multiple passes over the data prior to presentation to the user.

  The Motivation for an interactive task is that SED remains a difficult task for humans and systems. Also, Interactivity and relevance feedback have been effectively employed in other tasks.

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,

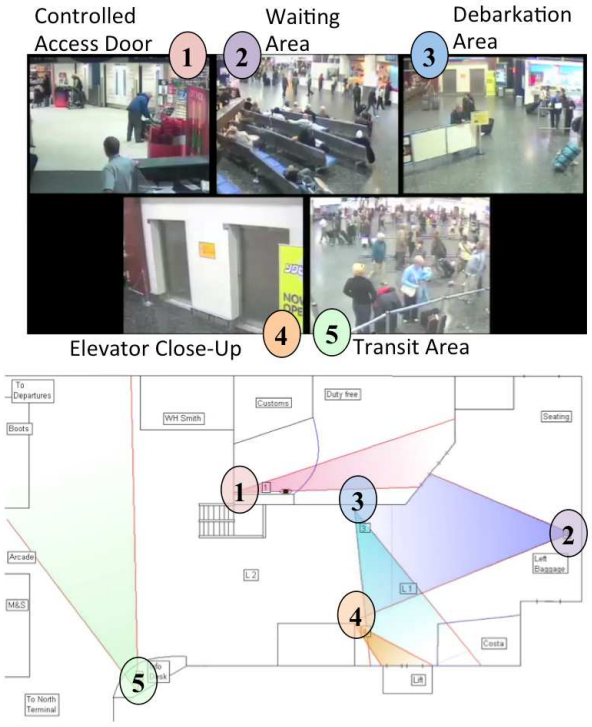


Figure 35: Camera views and coverage

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).

7.2 Data

The development data consisted of the full 100 h data set used for the 2008 Event Detection (Rose et al., 2009) evaluation. The video for the evaluation corpus came from the approximate 50 h 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. Figure 35 shows the coverage and views from the different cameras used for data collection.

Single Person events		
PersonRuns	7.02 / 2.80	Someone runs ← ***Lowest frequency***
Pointing	69.74 / 1.53	Someone points ← ***Highest frequency***
Single Person + Object events		
CellToEar	12.73 / 0.78	Someone puts a cell phone to his/her head or ear
ObjectPut	40.74 / 1.07	Someone drops or puts down an object
Multiple People events		
Embrace	11.48 / 6.13	Someone puts one or both arms at least part way around another person
PeopleMeet	29.46 / 4.89	One or more people walk up to one or more other people, stop, and some communication occurs
PeopleSplitUp	12.27 / 10.36	From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame

Figure 36: Event name, their rate of occurrences in Instances per Hour (IpH) / their average duration (in seconds) and Definition

System performance was assessed on the same 15-hour subset of the evaluation corpus as the 2009 Evaluation. Similar to the 2012 SED evaluation, systems were provided the identity of the evaluated subset so that searcher time for the interactive task was not expended on non-evaluated material. Event annotation was performed by the LDC using a three-pass annotation scheme. The multi-pass process improves the human annotation recall rates.

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

7.3 Evaluation, measures

Sites submitted system outputs for the detection of any 3 of 7 possible events (PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, and Pointing). Additional details for the list of event used can be found in Figure 36. For each instance observation, sites are asked to identify each detected event observation by:

- the temporal extent (beginning and end frames)
- a decision score: a numeric score indicating how likely the event observation exists with more positive values indicating more likely observations (normalized)
- an actual decision: a boolean value indicating whether or not the event observation should be counted for the primary metric computation

10 SED 2013 Participants
(with number of systems per event)

		Single Person				Person + object				Multiple People					
		PersonRuns		Pointing		CellToEar		ObjectPut		Embrace		PeopleMeet		PeopleSplitUp	
		iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED	iSED	rSED
6 years in a row	Carnegie Mellon University [CMU]	1	5	1	5	1		1	5	1	5	1	5	1	5
5 years in a row	Multimedia Communication and Pattern Recognition Labs, Beijing University of Posts and Telecommunications [BUPT-MCPRL]	1	2	1	2			1	2	1	5	1	5	1	5
2 years in a row	Brno University of Technology [BrnoUT]	1	2	1	2	1	2	1	2	1	2	1	2	1	2
	Dublin City University [dcu-savasa]	3	1	3	7	3	6	3	7						
	IBM Thomas J. Watson Research Center [IBM]	4	3	4	3	4	3	4	3	4	3	4	3	4	3
	The City College of New York Media Lab and SRI [CCNY-SRI]	2		2		2		2		2		2		2	
	Institute of Computer Science and Technology, Peking University [PKU-OS]	1	1	1	1			1	1						
	University of Ottawa(VIVA) and Polytechnique Montréal (LITIV) [VIVAuOttawaLITIVpoly]	1	2	1	2	1	2	1	2	1	2	1	2	1	2
NEW	AT&T Labs Research [ATT]	5		5		5		5		5		5		5	
	Beijing Institute of Technology [BIT]							2							
		19	16	19	22	17	13	21	22	15	17	15	17	15	17

Figure 37: TRECVID 2013 SED Participants Chart

Developers were advised to target a low miss, high false alarm scenario, in order to maximize the number of event observations.

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

Since detection system performance is a trade-off 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 unit time). Participants were provided a graph of the Decision Error Trade-off (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.

7.4 Results

There were 10 participants in 2013 (see figure 37), for a total of 122 Interactive Event Runs and 126 Retrospective Event Runs.

Figure 38 presents the event-averaged lowest Actual NDCR by site's rSED vs iSED for the 7 sites that submitted both types of runs. Out of those 7 sites, 6 show some reduction in their NDCR, with three large reductions (BrnoUT by 30 %, BUPT-MCPRL by 45 % and VIVAuOttawaLITIVpoly by 84 %).

From the 2012 evaluation (which used the same data set), we can see that some improvement for repeat performers for the rSED (figure 39) and iSED (figure 40) tasks. We can also see that iSED filters rSED results.

Comparable results since 2009 for rSED, and since for 2012 iSED are present in Figures 41-47. In those

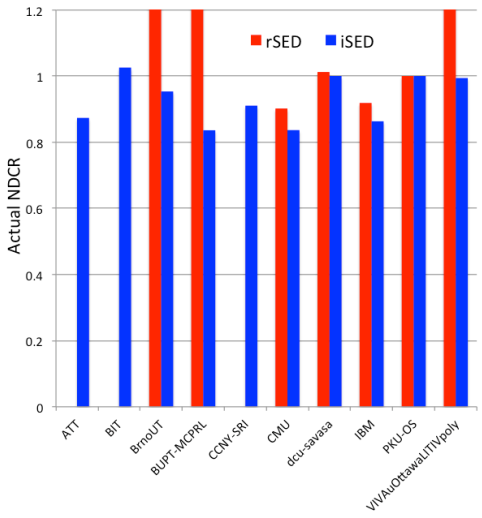


Figure 38: Event-Averaged, Lowest Act NDCR by Site: rSED vs. iSED

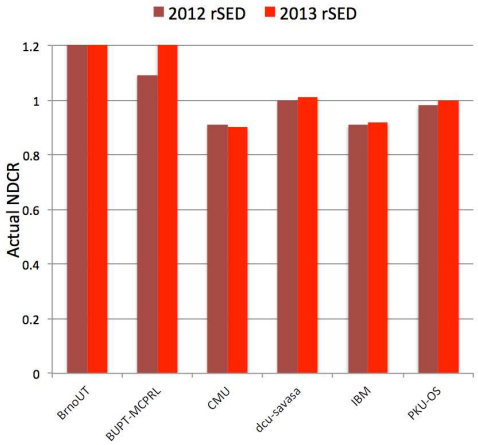


Figure 39: Event-Averaged, Lowest Act NDCR by Site: rSED 2012 vs 2013

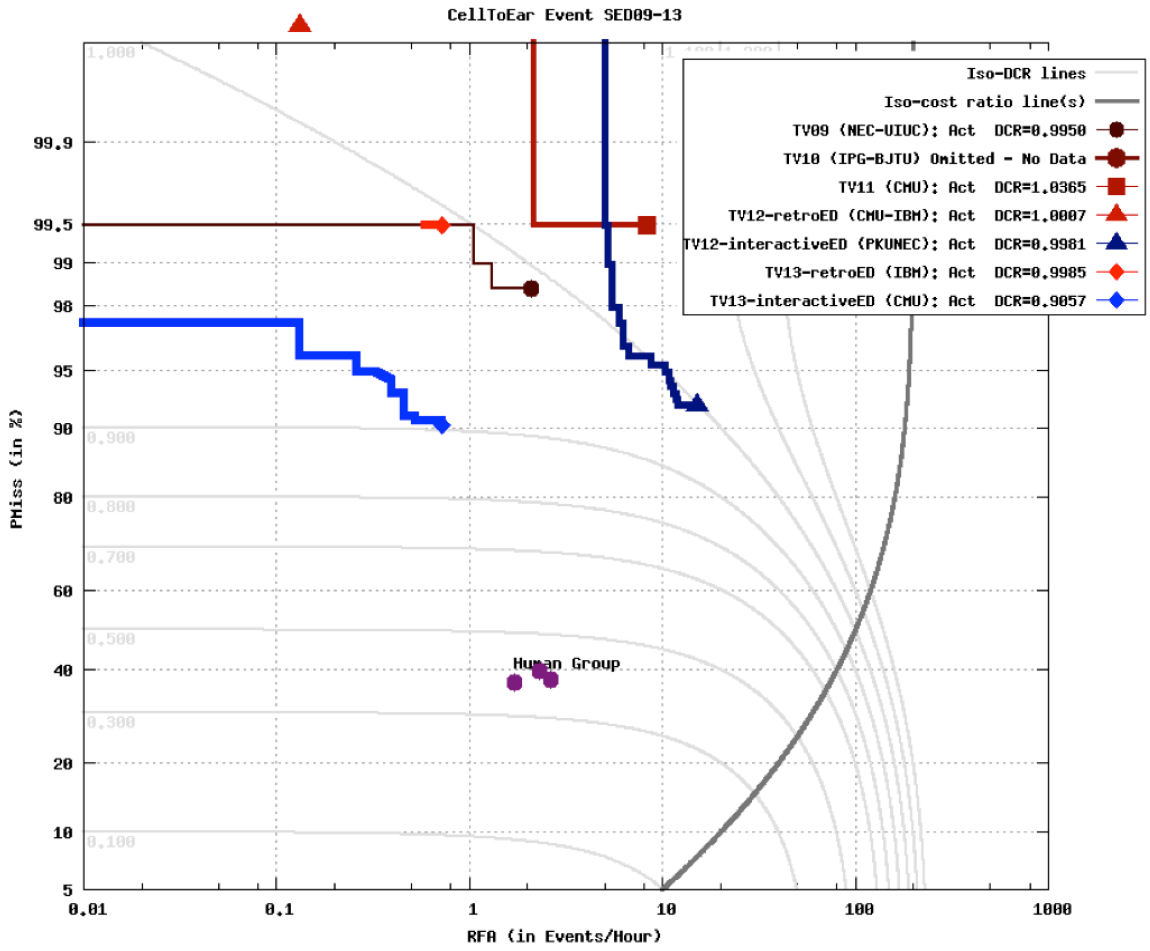


Figure 41: SED '09-13 : rSED and iSED - CellToEar

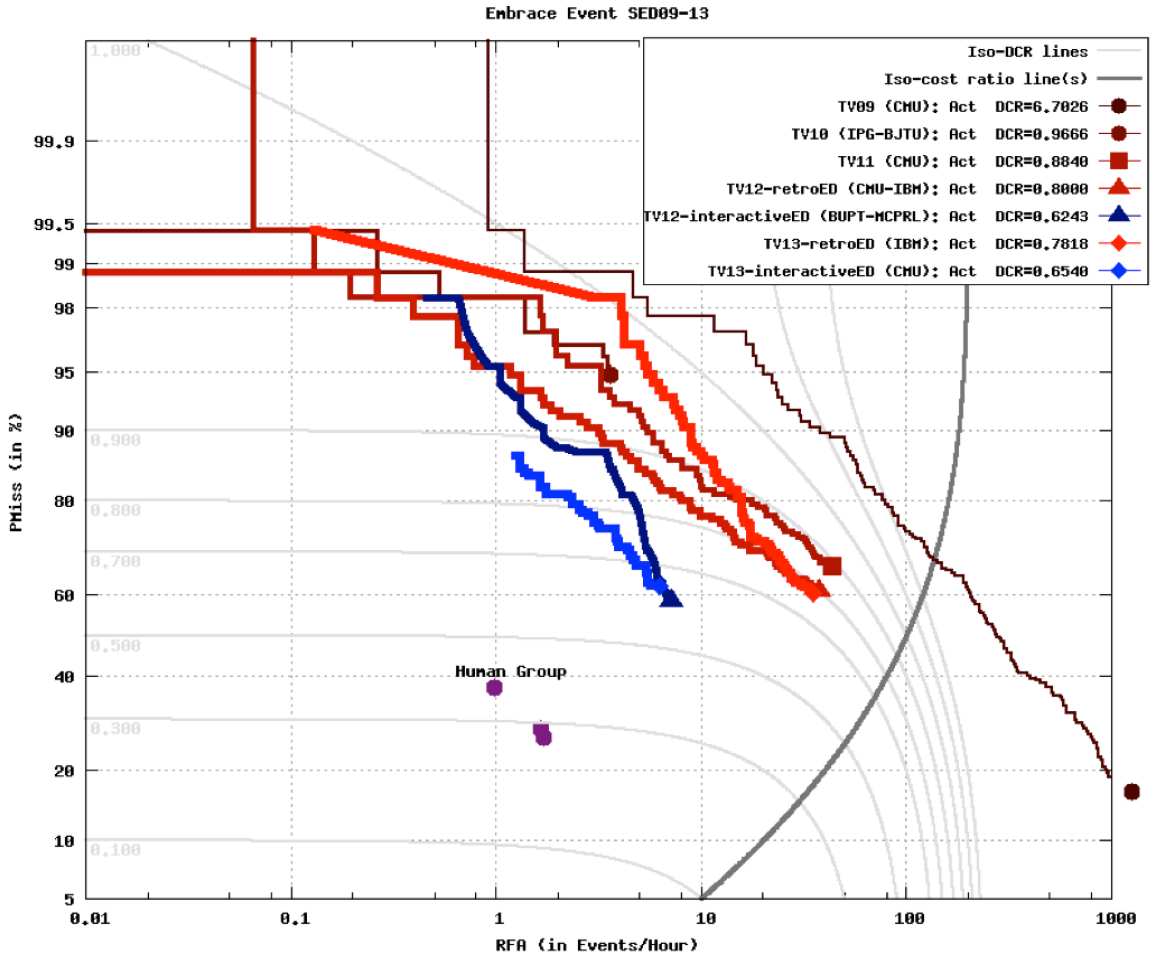


Figure 42: SED '09-13 : rSED and iSED - Embrace

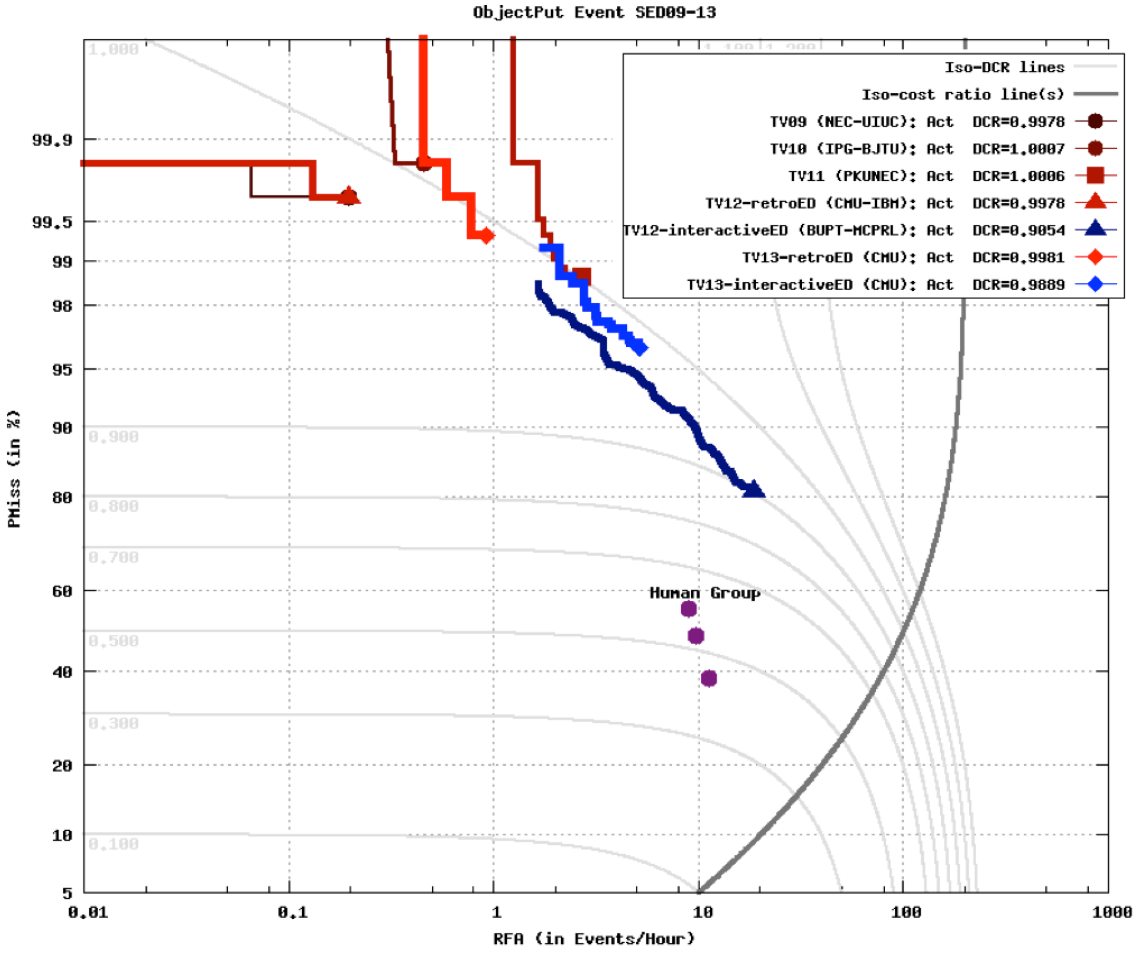


Figure 43: SED '09-13 : rSED and iSED - ObjectPut

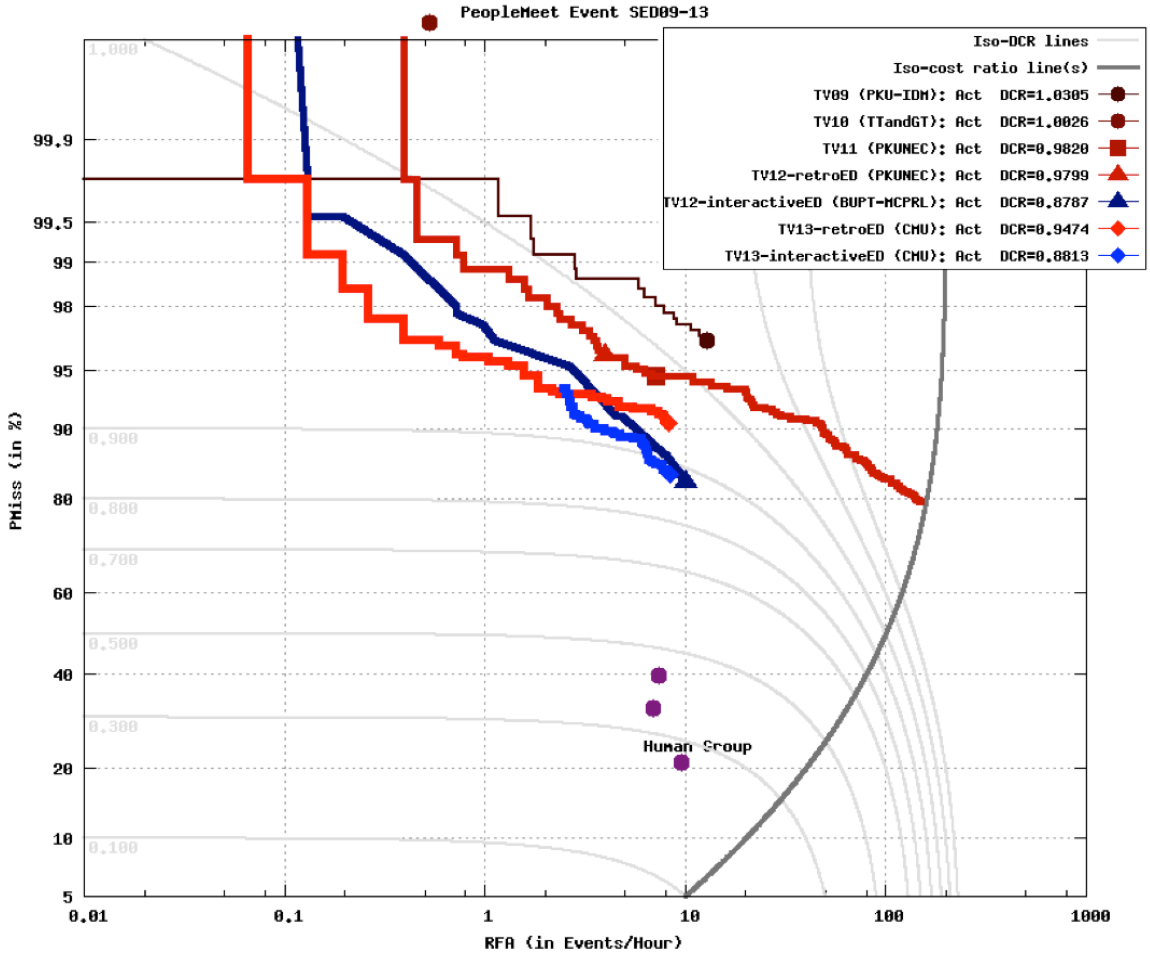


Figure 44: SED '09-13 : rSED and iSED - PeopleMeet

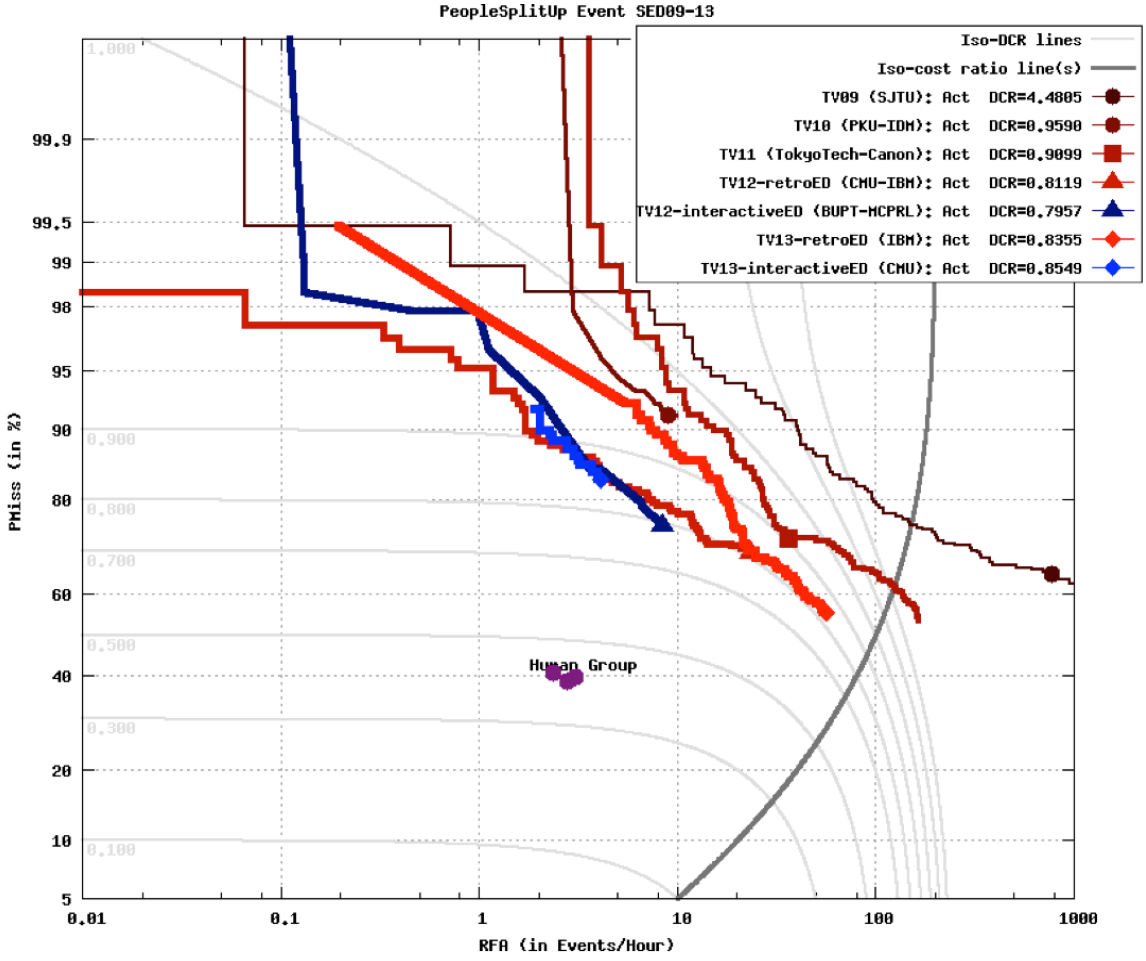


Figure 45: SED '09-13 : rSED and iSED - PeopleSplitUp

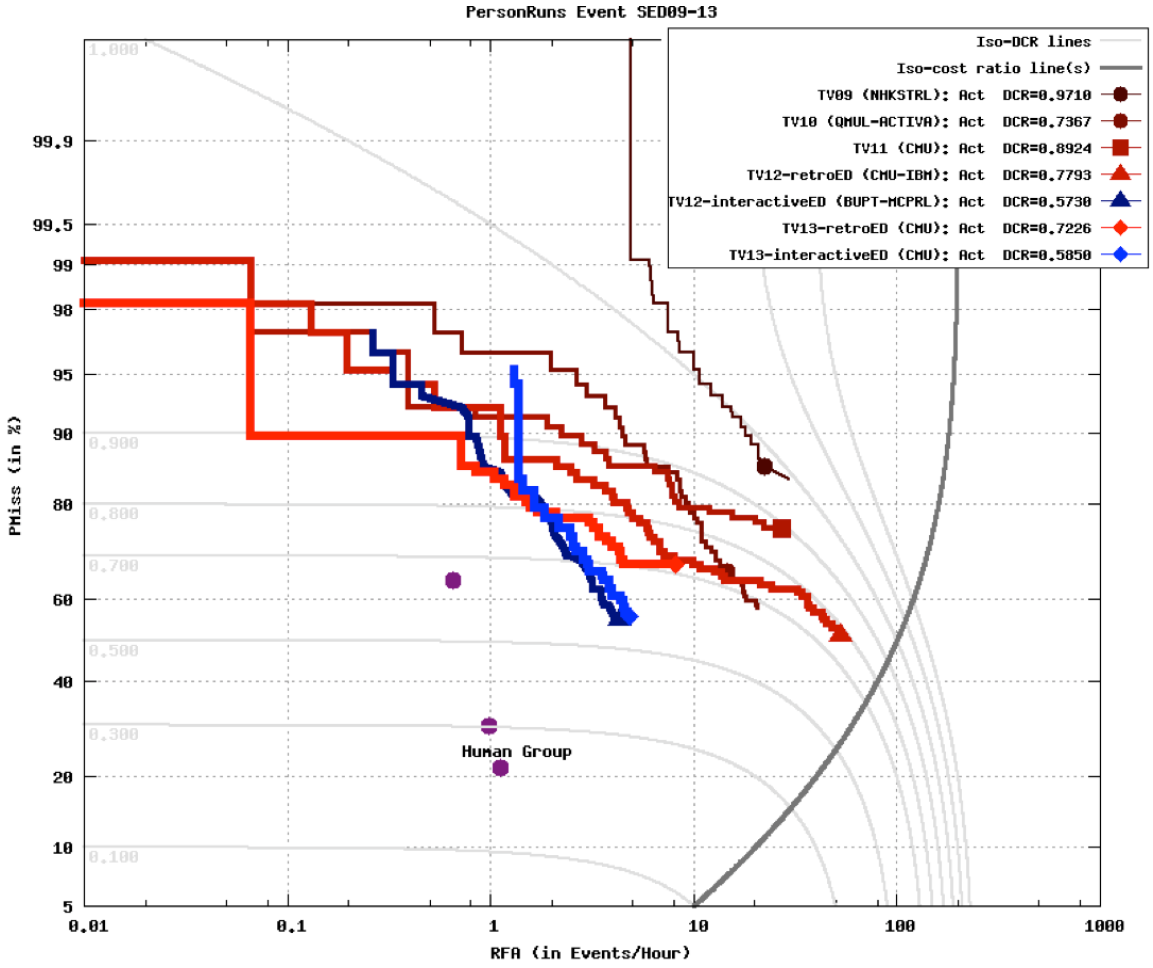


Figure 46: SED '09-13 : rSED and iSED - PersonRuns

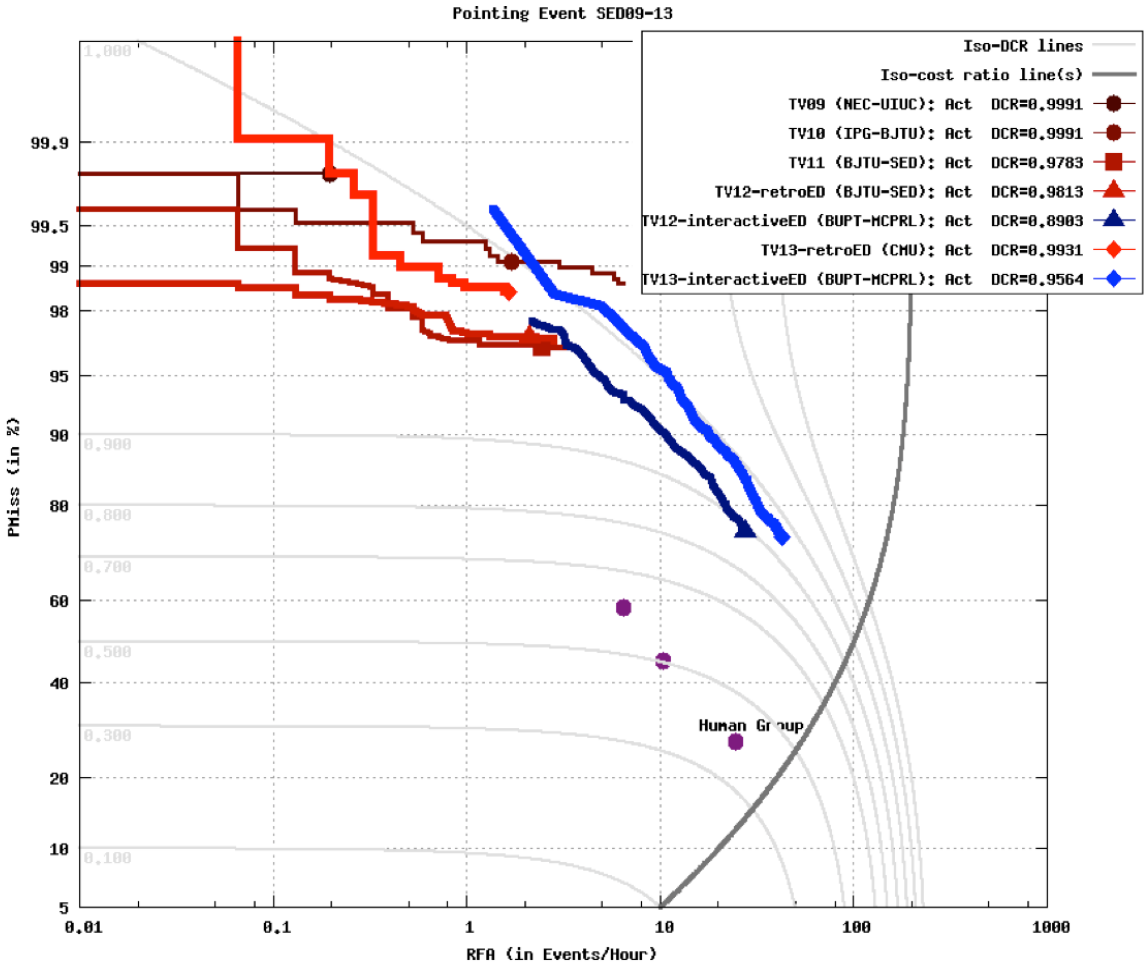


Figure 47: SED '09-13 : rSED and iSED - Pointing

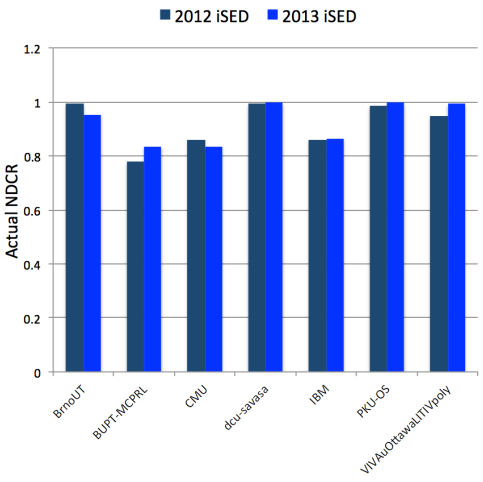


Figure 40: Event-Averaged, Lowest Act NDCR by Site: iSED 2012 vs 2013

plots, one can see that Single-person (PersonRuns, PeopleSplitUp, Pointing) and Multi-Person (PeopleMeet, Embrace) events show evidence of yearly improvements, yet not yet approaching human performance. Person+Object (ObjectPut, CellToEar) events remain difficult.

Readers are asked to see the results pages at the back of the workshop notebook and on the TRECVID website for information about each run's performance.

8 Summing up

This introduction to TRECVID 2013 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 in the results section at the back of the notebook.

As we reflect on a decade of TRECVid, which launched as a standalone annual benchmarking activity in 2003, we can see how some of the tasks we address have matured significantly. Shot Boundary Detection, story bound detection and automatic summarisation have all seen great progress over the decade while there are other tasks we are still working on because these are hard problems. Search, in all its varieties from known item (re-)location to broad search, event detection and automatic detection of concepts remain as a focus for our benchmarking and we will continue to throw hard problems like these as well as increasingly larger video collections, at the video research community because these are the drivers that push out the boundaries of achievement.

While changing tasks and increasing video collection sizes are evolutions of TRECVid, less obvious than these are the fact that in the last decade, scientific reproducibility has come into focus, in particular the open access to research data on which scientific experimentation is based. The ethos behind TRECVid, and TREC, is to make the reproduction of experiments easy and that includes making data, queries or topics, relevance judgments and scoring methods, open, transparent and available. There is a growing disquiet throughout the sciences about scientific reproducibility, highlighted in an article in The Economist on Oct 19, 2013, entitled "Unreliable Research: Trouble at the Lab"(*). This pointed at the many reasons why some scientific experimentation, right across the scientific disciplines, cannot be faithfully reproduced, including the rush to publish resulting from pressures from industry and funding agencies to see results, poor descriptions of scientific methods, carelessness, or sometimes fraud.

In TRECvid, because of our in-built ethos of easy reproducibility, we are somewhat insulated from this but just because we have our data, to the extent we are legally able, and our scoring methods available for everyone does not mean we are immune from this. We must ensure that we continue to provide enough detail in our descriptions of TRECVid work, through our workshop papers and other publications, so that our methods can be reproduced for others to compare their own work against.

9 Authors' note

TRECVID would not have happened in 2013 without support from the National Institute of Standards and Technology (NIST) and the Intelligence Advanced Research Projects Activity (IARPA). The research community is very grateful for this. Beyond that, various individuals and groups deserve special thanks:

- Koichi Shinoda of the TokyoTechCanon team agreed to host a copy of IACC.2 data
- Georges Quénot with Franck Thollard, Andy Tseng, Bahjat Safadi from LIG and Stéphane Ayache from LIF shared coordination of the semantic indexing task, organized the community

annotation of concepts, and provided judgments for 50 concepts under the Quaero program.

- Georges Quénot provided the master shot reference for the IACC.2 videos.
- The LIMSI Spoken Language Processing Group and VexSys Research provided ASR for the IACC.2 videos.
- Cees Snoek helped choose the SIN concept pairs
- 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 we want to thank all the participants and other contributors on the mailing list for their energy and perseverence.

10 Appendix A: Instance search topics

9069 OBJECT - a circular 'no smoking' logo

9070 OBJECT - a small red obelisk

9071 OBJECT - an Audi logo

9072 OBJECT - a Metropolitan Police logo

9073 OBJECT - this ceramic cat face

9074 OBJECT - a cigarette

9075 OBJECT - a SKOE can

9076 OBJECT - this monochrome bust of Queen Victoria

9077 OBJECT - this dog

9078 OBJECT - a JENKINS logo

9079 OBJECT - this CD stand in the market

9080 OBJECT - this public phone booth

9081 OBJECT - a black taxi

9082 OBJECT - a BMW logo

9083 OBJECT - a chrome and glass cafetiere

9084 PERSON - this man

9085 OBJECT - this David refrigerator magnet

9086 OBJECT - these scales

9087 OBJECT - a VW logo

9088 PERSON - Tamwar

9089 OBJECT - this pendant

9090 OBJECT - this wooden bench with rounded arms

9091 OBJECT - a Kathy's menu with stripes

9092 PERSON - this man

9093 OBJECT - these turnstiles

9094 OBJECT - a tomato-shaped ketchup dispenser

9095 OBJECT - a green public trash can

9096 PERSON - Aunt Sal

9097 OBJECT - these checkerboard spheres

9098 OBJECT - a P (parking automat) sign

References

Ayache, S., & Quénot, G. (2008, March). Video Corpus Annotation Using Active Learning,. In *Proceedings of the 30th european conference on information retrieval (ecir'08)* (pp. 187–198). Glasgow, UK.

Gauvain, J., Lamel, L., & Adda, G. (2002). The LIMSI Broadcast News Transcription System. *Speech Communication*, *37*(1-2), 89—108.

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

Over, P., Ianeva, T., Kraaij, W., & Smeaton, A. F. (2006). *TRECVID 2006 Overview.* `www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6overview.pdf`.

QUAERO. (2010). *QUAERO homepage.* `www.quaero.org/modules/movie/scenes/home/`.

Rose, T., Fiscus, J., Over, P., Garofolo, J., & Michel, M. (2009, December). The TRECVid 2008 Event Detection Evaluation. In *IEEE Workshop on Applications of Computer Vision (WACV).* IEEE.

Strassel, S., Morris, A., Fiscus, J., Caruso, C., Lee, H., Over, P., et al. (2012, may). 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).

TV13Notebook. (2013). `http://www-nlpir.nist.gov/projects/tv2013/active/workshop.notebook`.

TV13Pubs. (2013). `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.13.org.html`.

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, E., & Aslam, J. A. (2006, November). 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, E., Kanoulas, E., & 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* (pp. 603–610). New York, NY, USA: ACM.