TRECVID 2018

Ad-hoc Video Search Task : Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
Dakota Consulting, Inc;
National Institute of Standards and Technology

Outline

- Task Definition
- Video Data
- Topics (Queries)
- Participating teams
- Evaluation & results
- General observation

Task Definition

- Goal: promote progress in content-based retrieval based on end user **ad-hoc (generic) queries** that include persons, objects, locations, actions and their combinations.

- Task: Given a test collection, a query, and a master shot boundary reference, return a ranked list of at most 1000 shots (out of 335 944) which best satisfy the need.

- Testing data: 4593 Internet Archive videos (IACC.3), 600 total hours with video durations between 6.5 min to 9.5 min. Reflects a wide variety of content, style and source device.

- Development data: ≈1400 hours of previous IACC data used between 2010-2015 with concept annotations.

Query Development

- Test videos were viewed by 10 human assessors hired by the National Institute of Standards and Technology (NIST).
- 4 facet description of different scenes were used (if applicable):
  - Who : concrete objects and being (kind of persons, animals, things)
  - What : are the objects and/or beings doing ? (generic actions, conditions/state)
  - Where : locale, site, place, geographic, architectural
  - When : time of day, season
- In total assessors watched ≈35% of the IACC.3 videos
- 90 Candidate queries chosen from human written descriptions to be used between 2016-2018.

TV2018 Queries by complexity

- Person + Action + Object + Location

Find shots of exactly two men at a conference or meeting table talking in a room
Find shots of a person playing keyboard and singing indoors
Find shots of one or more people on a moving boat in the water
Find shots of a person in front of a blackboard talking or writing in a classroom
Find shots of people waving flags outdoors

- Person/being + Action + Location

Find shots of a dog playing outdoors
Find shots of people performing or dancing outdoors at nighttime
Find shots of one or more people hiking
Find shots of people standing in line outdoors

TV2018 Queries by complexity

- Person + Action/state + Object

Find shots of a person sitting on a wheelchair
Find shots of a person climbing an object (such as tree, stairs, barrier)
Find shots of a person holding, talking or blowing into a horn
Find shots of a person lying on a bed.
Find shots of a person with a cigarette
Find shots of a truck standing still while a person is walking beside or in front of it
Find shots of a person looking out or through a window
Find shots of a person holding or attached to a rope
Find shots of a person pouring liquid from one container to another

- Person + Action

Find shots of medical personnel performing medical tasks
Find shots of two people fighting
Find shots of a person holding his hand to his face

TV2018 Queries by complexity

- Action + Object + Location

Find shots of car driving scenes in a rainy day

- Person + Object

Find shots of two or more people wearing coats

Find shots of a person where a gate is visible in the background

- Person/being

Find shots of two or more cats both visible simultaneously

- Person + Location

Find shots of a person in front of or inside a garage

Find shots of one or more people in a balcony

- Object + Location

Find shots of an elevator from the outside or inside view

- Object

Find shots of a projection screen

Find shots of any type of Christmas decorations

Training and run types

Three run submission types:

- ✓ Fully automatic (F): System uses official query directly(33 runs)
- ✓ Manually-assisted (M): Query built manually (16 runs)
- ✓ Relevance Feedback (R): Allow judging top-5 once (2 runs)

Four training data types:

- ✓ A – used only IACC training data (0 runs)
- ✓ D – used any other training data (50 runs)
- ✓ E – used only training data collected automatically using only the query text (1 run)
- ✓ F – used only training data collected automatically using a query built manually from the given query text (0 runs)

Finishers : 13 out of 23

Team	Organization	Runs		
		M	F	R
INF	Carnegie Mellon University; Shandong Normal University; Renmin University; Beijing University of Technology	-	5	-
kobe_kindai	Graduate School of System Informatics, Kobe University; Department of Informatics, Kindai University	4	-	-
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas; Queen Mary University of London	-	4	-
NECTEC	National Electronics and Computer Technology Center	1	1	-
NII_Hitachi_UIT	National Institute of Informatics, Japan (NII); Hitachi, Ltd; University of Information Technology, VNU-HCM, Vietnam	-	3	-
MediaMill	University of Amsterdam	-	4	-
Waseda_Meisei	Waseda University; Meisei University	2	4	-
VIREO_NExT	National University of Singapore; City University of Hong Kong	4	3	2
NTU_ROSE_AVS	ROSE LAB, NANYANG TECHNOLOGICAL UNIVERSITY	-	1	-
FIU_UM	Florida International University, University of Miami	4	-	-
RUCMM	Renmin University of China	-	4	-
SIRET	SIRET Department of Software Engineering, Faculty of Mathematics and Physics, Charles University	1	-	-
UTS_ISA	University of Technology Sydney	-	4	-

Evaluation

Each query assumed to be binary: absent or present for each master reference shot.

NIST judged top tanked pooled results from all submissions 100% and sampled the rest of pooled results.

Metrics: *Extended inferred average precision per query.*

Compared runs in terms of **mean** extended *inferred average precision* across the 30 queries.

Mean Extended Inferred Average Precision (XInfAP)

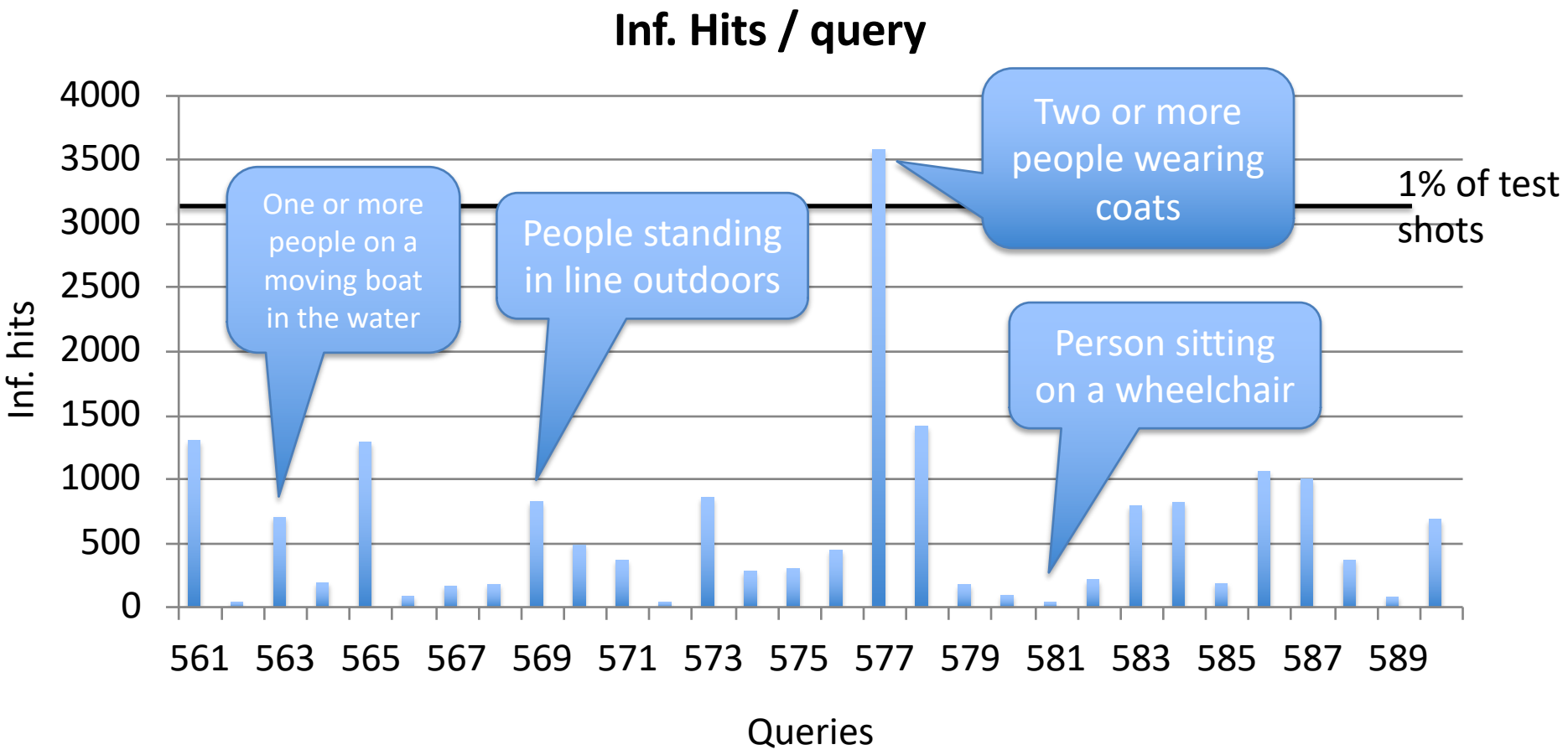
2 pools were created for each query and sampled as:

- ✓ Top pool (ranks 1 to 150) sampled at 100 %
- ✓ Bottom pool (ranks 151 to 1000) sampled at 2.5 %
- ✓ % of sampled and judged clips from rank 151 to 1000 across all runs and topics (min= 1.6 %, max = 62 %, mean = 28 %)

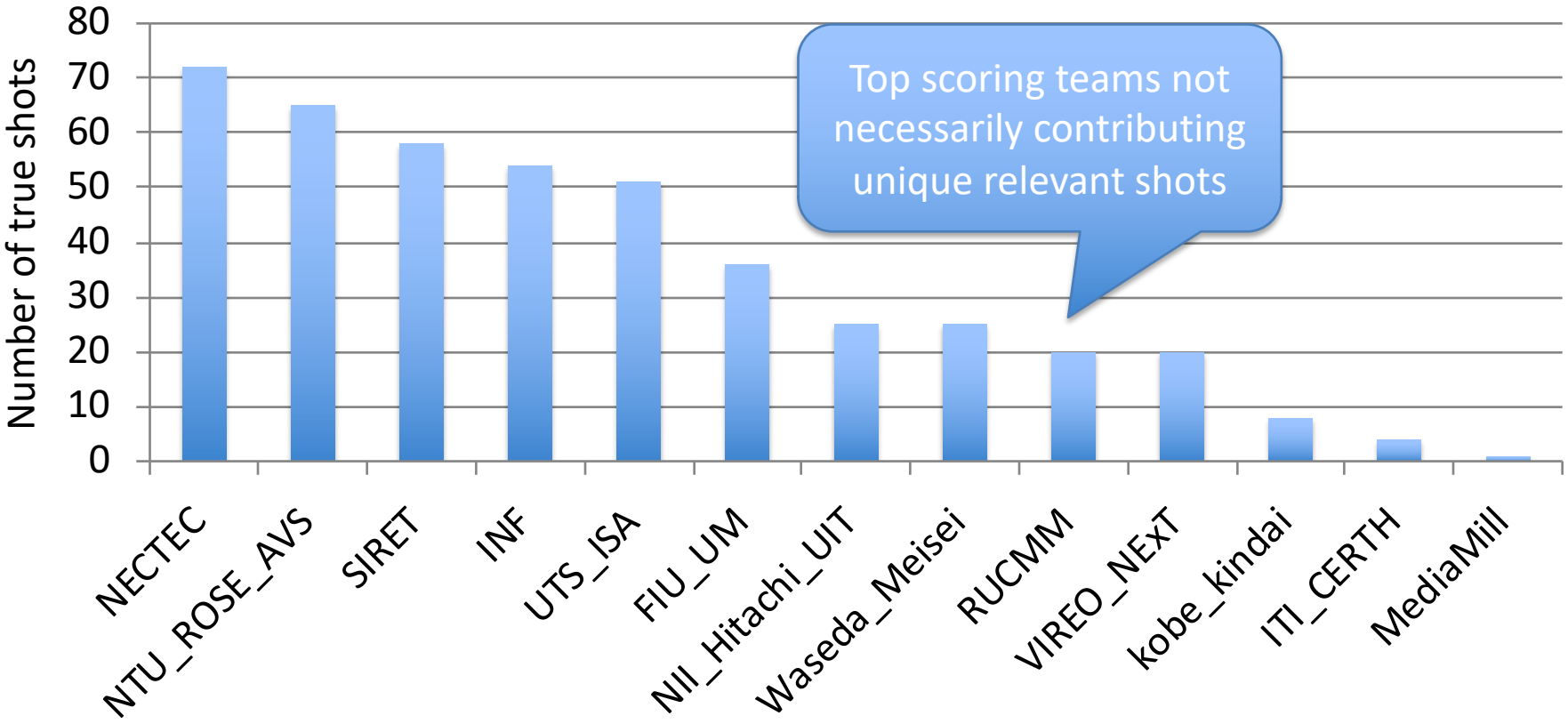
30 queries
92 622 total judgments
7381 total hits
5635 hits at ranks (1 to100)
1469 hits at ranks (101 to 150)
277 hits at ranks (151 to 1000)

Judgment process: one assessor per query, watched complete shot while listening to the audio. infAP was calculated using the judged and unjudged pool by sample_eval tool

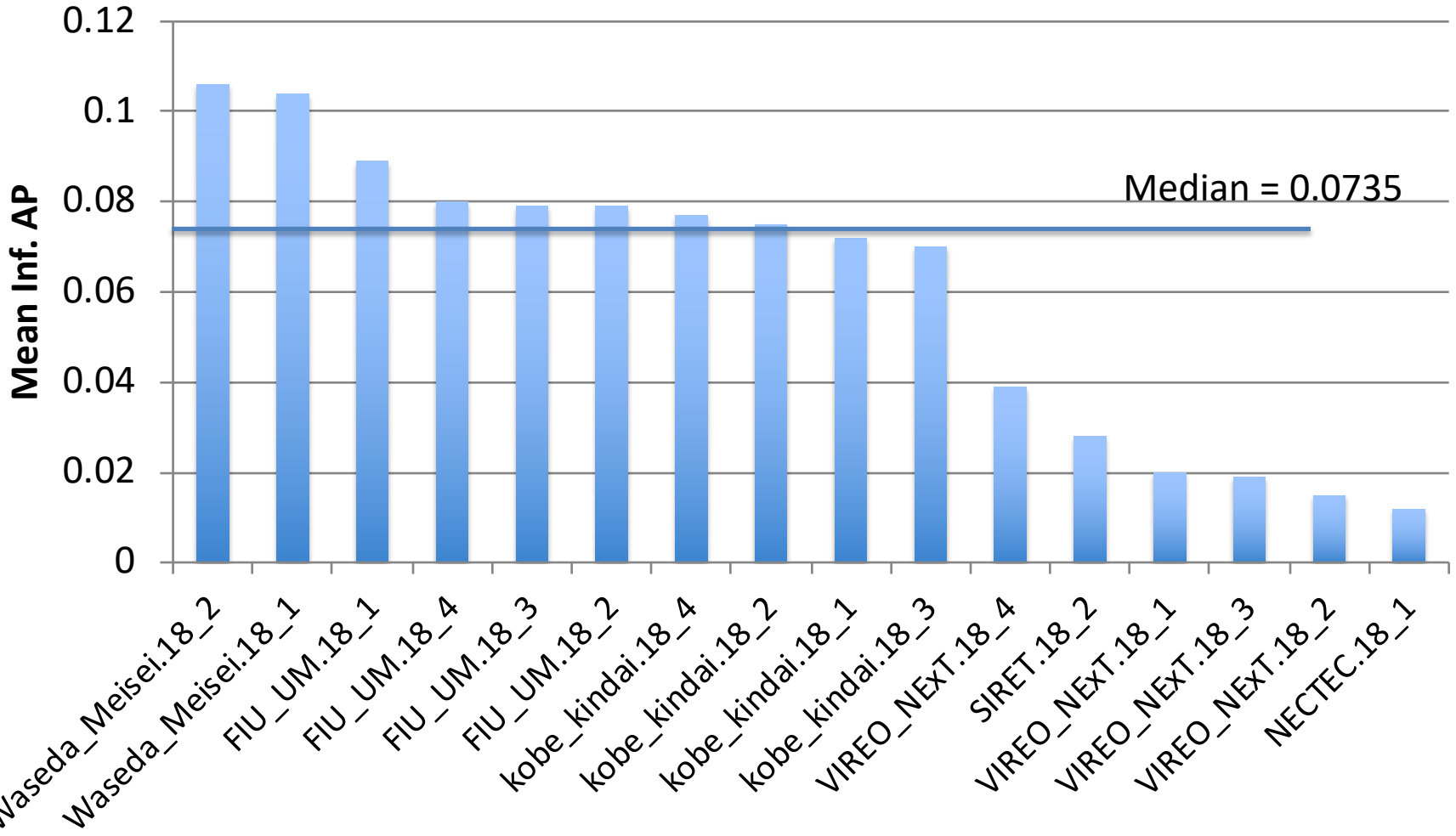
Inferred frequency of hits varies by query



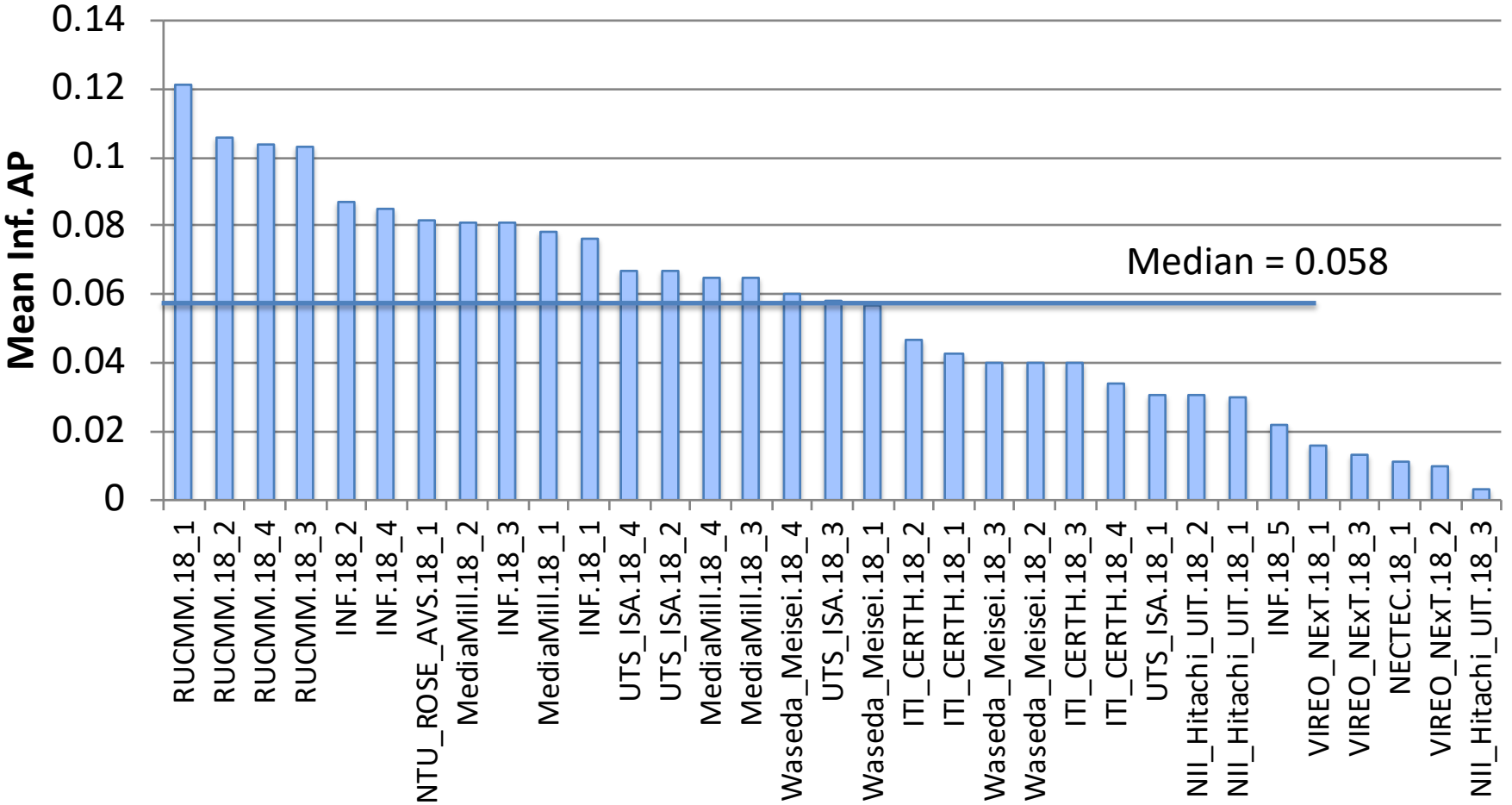
Total true shots contributed uniquely by team



Sorted scores
(16 Manually-assisted runs, 6 teams)



Sorted scores
(33 Fully automatic runs, 10 teams)



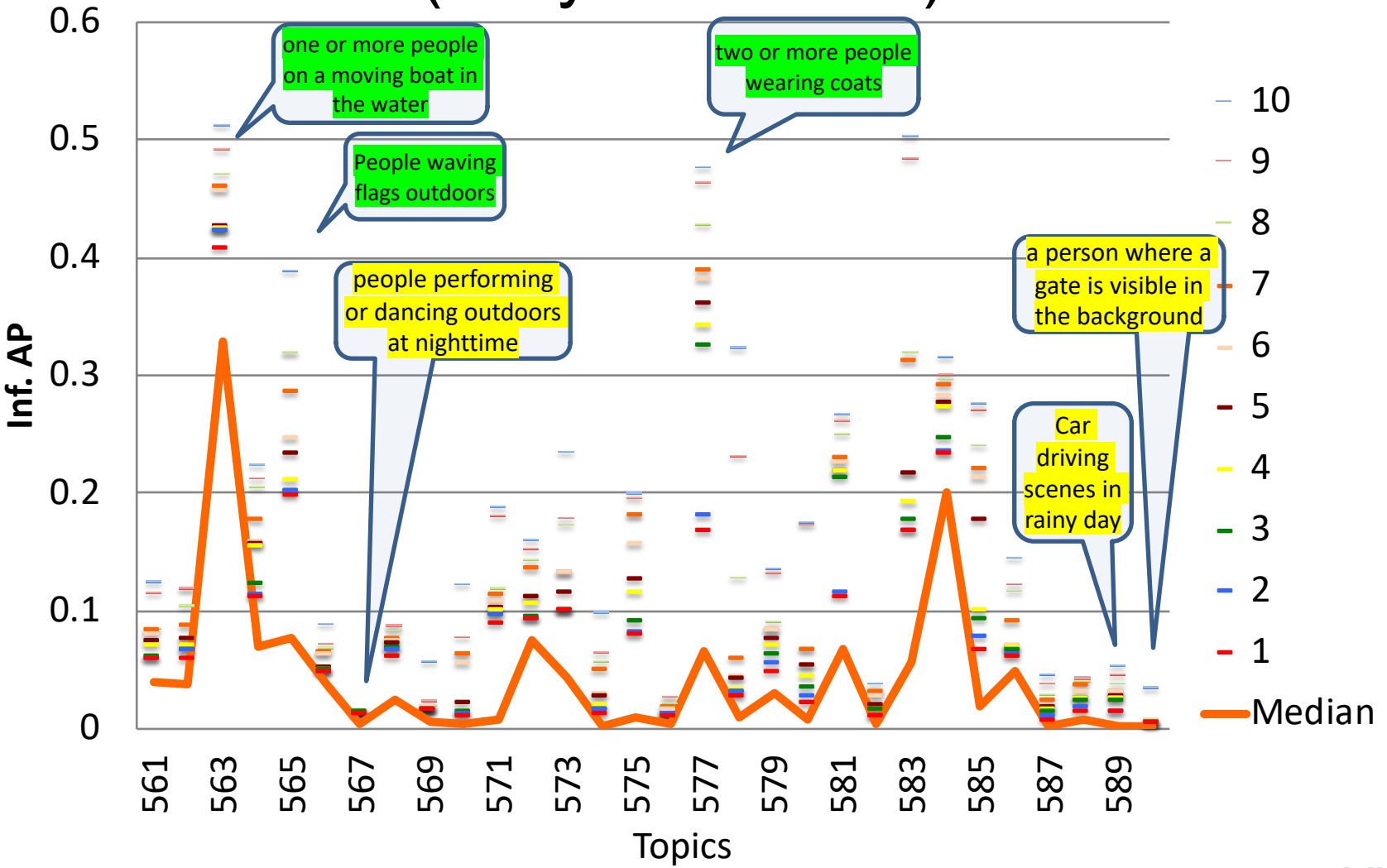
2 Relevance feedback runs, 1 team

- VIREO_NExT.18_1 0.018
- VIREO_NExT.18_2 0.016

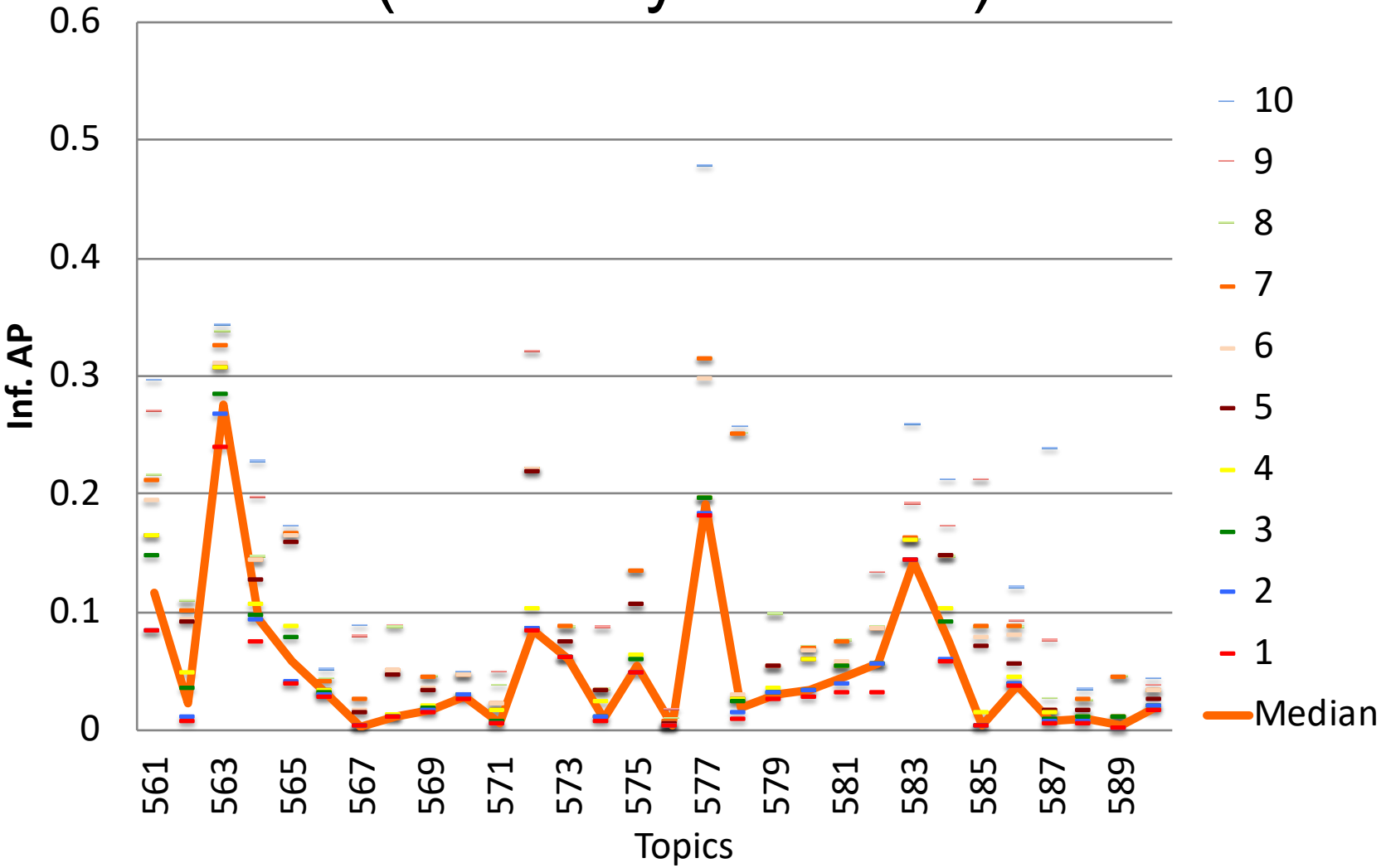
** New run type in 2018

** No significant difference between the two runs based on the randomization testing

Top 10 infAP scores by query (Fully Automatic)



Top 10 infAP scores by queries (Manually-Assisted)



Performance in the last 3 years ?

Automatic	**2016**	**2017**	**2018**
Teams	9	8	10
Runs	30	33	33
Min *xInfAP*	0	**0.026**	0.003
Max *xInfAP*	0.054	**0.206**	0.121
Median *xInfAP*	0.024	**0.092**	0.058

Manually-Assisted	**2016**	**2017**	**2018**
Teams	8	5	6
Runs	22	19	16
Min *xInfAP*	0.005	**0.048**	0.012
Max *xInfAP*	0.169	**0.207**	0.106
Median *xInfAP*	0.043	**0.111**	0.072

Easy vs difficult topics overall (2017)

Top 10 Easy (sorted by count of runs with InfAP >= 0.7)	Top 10 Hard (sorted by count of runs with InfAP < 0.7)
a person wearing any kind of hat	an adult person running in a city street
a chef or cook in a kitchen	person standing in front of a brick building or wall
one or more people driving snowmobiles in the snow	person holding, opening, closing or handing over a box
one or more people swimming in a swimming pool	a male person falling down
a man and woman inside a car	child or group of children dancing
a crowd of people attending a football game in a stadium	children playing in a playground
a newspaper	person talking on a cell phone
a person communicating using sign language	person holding or opening a briefcase
a person wearing a scarf	one or more people eating food at a table indoor
a person riding a horse including horse-drawn carts	person talking behind a podium wearing a suit outdoors during daytime

Easy vs difficult topics overall (2018)

Top 10 Easy (sorted by count of runs with InfAP >= 0.7)	Top 10 Hard (sorted by count of runs with InfAP < 0.7)
Nothing	ALL topics

Threshold of infAP = 0.7 (same used in 2017) is too high for 2018 topics → 2018 topics are more harder ?

Easy vs difficult topics overall (2018)

Top 10 Easy (sorted by count of runs with InfAP >= 0.3)	Top 10 Hard (sorted by count of runs with InfAP < 0.1)
Find shots of one or more people on a moving boat in the water	Find shots of two people fighting
Find shots of two or more people wearing coats	Find shots of a person holding or attached to a rope
Find shots of a person holding, talking or blowing into a horn	Find shots of one or more people hiking
Find shots of people waving flags outdoors	Find shots of car driving scenes in a rainy day
Find shots of two or more cats both visible simultaneously	Find shots of people performing or dancing outdoors at nighttime
Find shots of a person lying on a bed	Find shots of a person where a gate is visible in the background
Find shots of a person in front of or inside a garage	Find shots of people standing in line outdoors
	Find shots of a dog playing outdoors
	Find shots of a person holding his hand to his face

Statistical significant differences among top 10 "M" runs (using randomization test, p < 0.05)

Run	Mean Inf. AP score
D_Waseda_Meisei.18_2	0.106 *
D_Waseda_Meisei.18_1	0.104 *
D_FIU_UM.18_1	0.089
D_FIU_UM.18_4	0.080 !
D_FIU_UM.18_3	0.079 !
D_FIU_UM.18_2	0.079 !
D_kobe_kindai.18_4	0.077 #
D_kobe_kindai.18_2	0.075 #
D_kobe_kindai.18_1	0.072 #
D_kobe_kindai.18_3	0.070 #

D_Waseda_Meisei.18_1
- D_kobe_kindai.18_4
- D_kobe_kindai.18_2
- D_kobe_kindai.18_1
- D_kobe_kindai.18_3
- D_FIU_UM.18_3
- D_FIU_UM.18_2

D_Waseda_Meisei.18_2
- D_kobe_kindai.18_4
- D_kobe_kindai.18_2
- D_kobe_kindai.18_1
- D_kobe_kindai.18_3

D_FIU_UM.18_1
- D_FIU_UM.18_2
- D_FIU_UM.18_4

!#* : no significant difference among each set of runs

- Runs higher in the hierarchy are significantly better than runs more indented.

Statistical significant differences among top 10 "F" runs (using randomization test, p < 0.05)

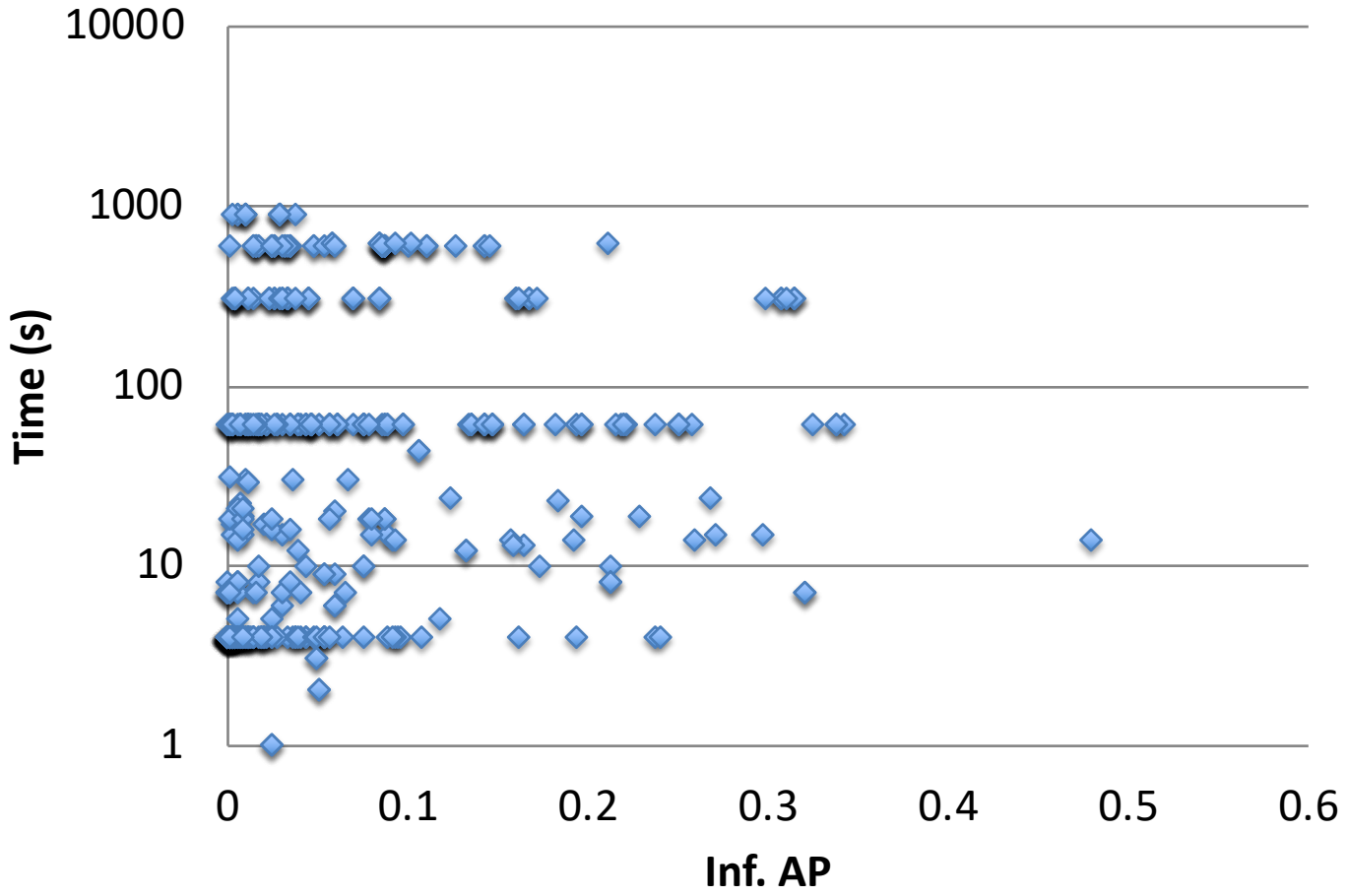
Run	Mean Inf. AP score
D_RUCMM.18_1	0.121
D_RUCMM.18_2	0.106 !
D_RUCMM.18_4	0.104 !
D_RUCMM.18_3	0.103 !
D_INF.18_2	0.087 *
D_INF.18_4	0.085 *
D_NTU_ROSE_AVS.18_1	0.082
D_MediaMill.18_2	0.081 #
D_INF.18_3	0.081 *
D_MediaMill.18_1	0.078 #

D_RUCMM.18_1
- D_RUCMM.18_3
- D_INF.18_2
- D_INF.18_4
- D_INF.18_3
- D_MediaMill.18_2
- D_MediaMill.18_1
- D_NTU_ROSE_AVS.18_1

!#* : no significant difference among each set of runs

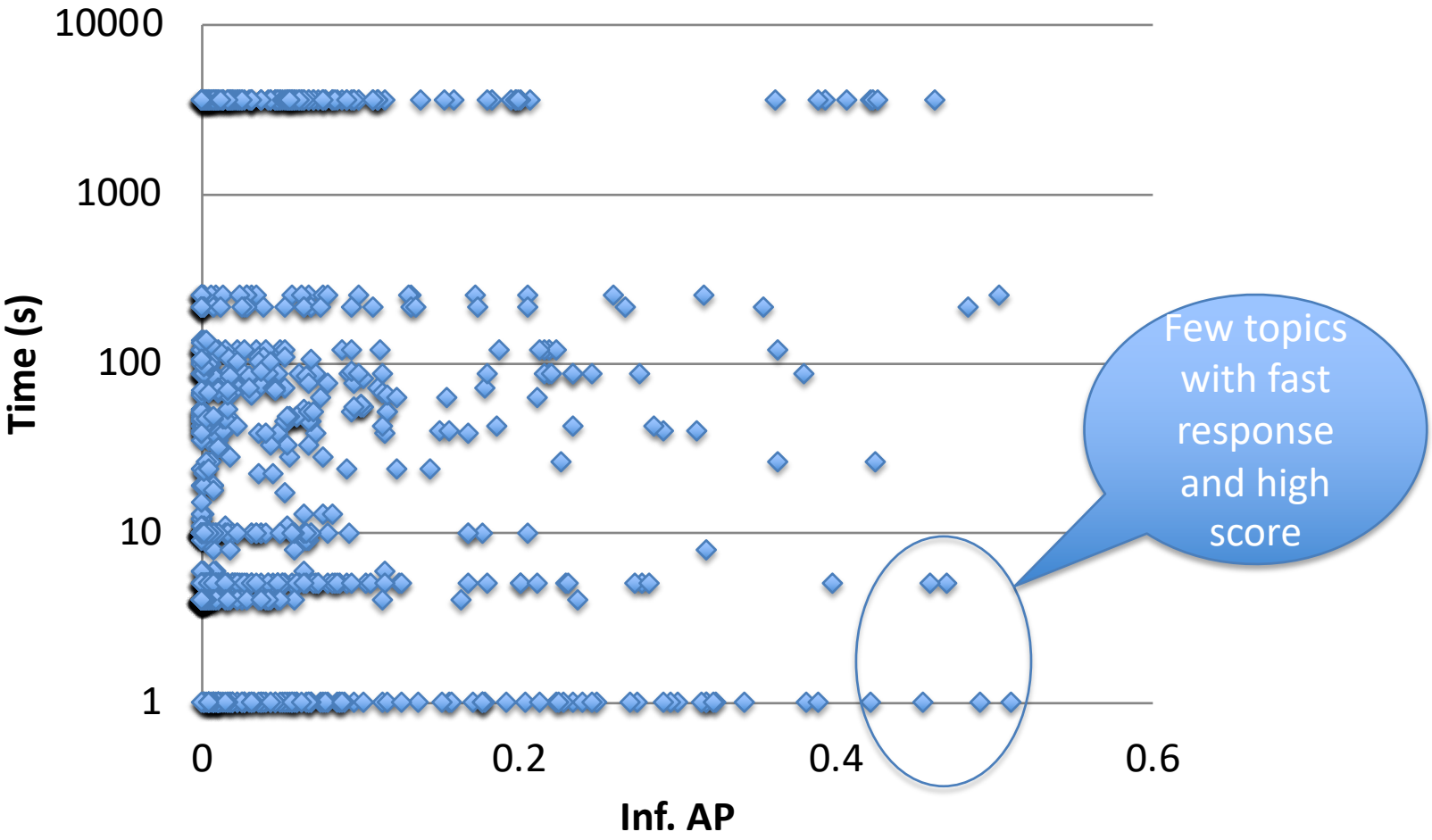
- Runs higher in the hierarchy are significantly better than runs more indented.

Processing time vs Inf. AP ("M" runs) Across all topics and runs



Processing time vs Inf. AP (“F” runs)

Across all topics and runs



2018 Main approaches

Renmin University of China: Automatic (0.121)

- Presentation to follow

Florida International University; University of Miami: Manual (0.089)

- Presentation to follow

Carnegie Mellon University; Shandong Normal University; Renmin University; Beijing University of Technology: Automatic (0.087)

- Presentation to follow

University of Amsterdam: Automatic (0.078)

- No notebook paper yet

2018 Main approaches

Waseda University, Meisei University: Manual (0.106), Automatic (0.060)

- Lot of work on concept bank integration
- Method 1 : Word-based keyword selection
- Method 2 : Similarity calculation between the word definition sentence and the whole query sentence
- Method 3 : Phrase-based concept selection
- Method 1 for manual, weighted combination for automatic (best with high weight on Method 3)

ROSE LAB, NANYANG TECHNOLOGICAL UNIVERSITY: Automatic (0.082)

- Image-based visual semantic embedding approach training from image/caption pairs (joint text-image representation space)
- No concept bank

2018 Main approaches

Kobe University, Kindai University: Manual (0.077)

- 5 concept banks
- Manual selection of concepts, different strategies
- Cascade filtering (did not work well)

Information Technologies Institute, Centre for Research and Technology Hellas; Queen Mary University of London: Automatic (0.043)

- Multiple concept banks
- Linguistic analysis of the query
- Use of sematic embedding's (text-based common representation)

2018 Task observations

- Finished 1-cycle of 3 years of Ad-hoc generic queries.
- Run training types are dominated by “D” runs.
- Stable team participation.
- Max and Median scores are < 2017 for both automatic and manually-assisted runs.
- In general manually-assisted runs perform better than automatic runs.
- Among high scoring topics, there is more room for improvement among systems.
- Among low scoring topics, most systems scores are collapsed in small narrow range.
- Most systems are slow. Few topics scored high in fast time.
- In general 2018 topics seem to be harder than 2017.
- Task is still challenging!

Interactive Video Retrieval subtask will be held as part of the Video Browser Showdown (VBS)

At MMM 2019
25th International Conference on Multimedia Modeling, January 8-11, 2019 Thessaloniki, Greece

- 10 Ad-Hoc Video Search (AVS) topics : Each AVS topic has several/many target shots that should be found.
- 10 Known-Item Search (KIS) tasks, which are selected completely random on site. Each KIS task has only one single 20 s long target segment.
- Registration for the task is now closed

9:30 – 12:00 : Ad-hoc Video Search

9:30 - 10:00, Word2VisualVec++ for Ad-hoc Video Search
(**RUCMM - Renmin University of China**)

10:00 - 10:30, Two approaches for cross-modal retrieval
(**INF - Carnegie Mellon University; Shandong Normal University; Renmin University; Beijing University of Technology**)

10:30 - 11:00, **Break** with refreshments

11:00 - 11:30, Learning Unknown Concepts and Exploring Concept Hierarchy for Ad-hoc Video Search Task
(**FIU_UM - Florida International University; University of Miami**)

11:30 - 12:00, AVS discussion

2018 Questions and 2019 plans

- Was the task/queries realistic enough?!
- Do we need to change/add/remove anything to the task in 2019 ?
  - Query language – (add alternative sentences per query)
- Is there any specific reason for the low submissions in “E” & “F” training type runs? (training data collected automatically from the given query text)
- Did any team run their 2018 system on TV2016 & TV2017 topics ?
- “Long tail blindness” (from unique hits)?
  - May be add metric to award unique (diverse) shot finders, penalize near duplicates.
- Engineering versus research efforts?
- Shared “consolidated” concept banks?
  - New effort to be built to encourage teams to share resources/concept models,…etc
- Current plan is to continue the task but using *New* dataset Vimeo Creative Common Collections (V3C1) for potentially 3 more years.
- Proposal for also a “progress subtask”.

AVS Progress subtask

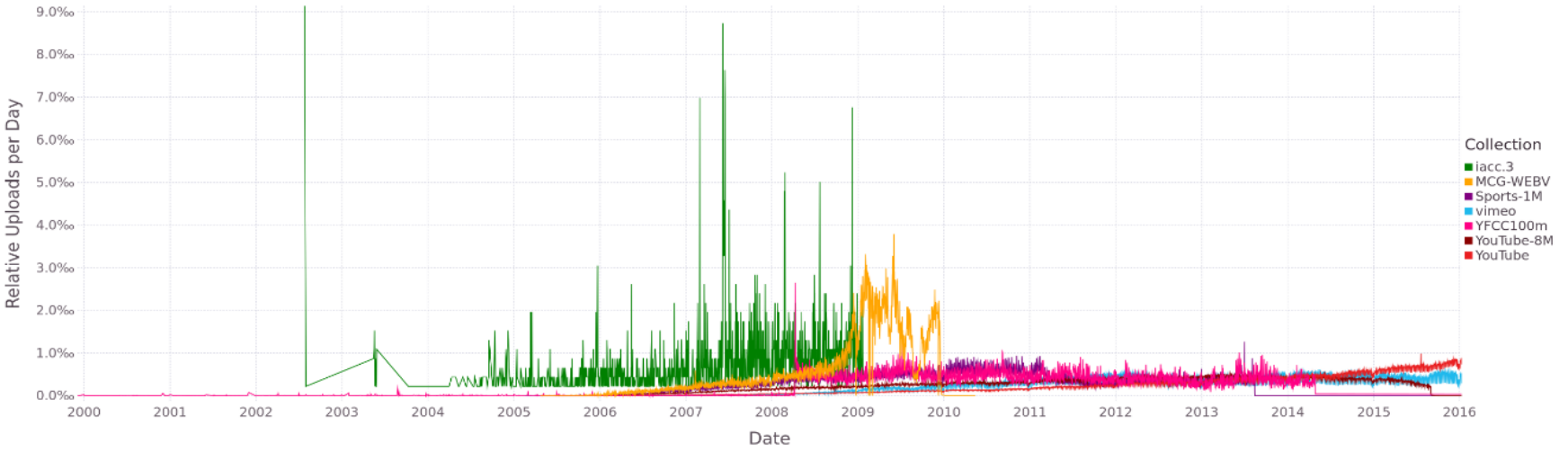
		Evaluation year		
Submission year		2019	2020	2021
	2019	Submit 50 queries (30 new + 20 common) Eval 30 new Queries		
	2020		Submit 40 queries (20 new + 20 common) Eval 30 (20 new + 10 common)	
	2021			Submit 40 queries (20 new + 20 common) Eval 30 (20 New + 10 common)

Goals : Evaluate 10 (set A) common queries submitted in 2 years (2019, 2020)
Evaluate 10 (set B) common queries submitted in 3 years (2019, 2020, 2021)
Evaluate 20 common queries submitted in 3 years (2019 , 2020, 2021)
Ground truth for 20 common queries can be released only in 2021

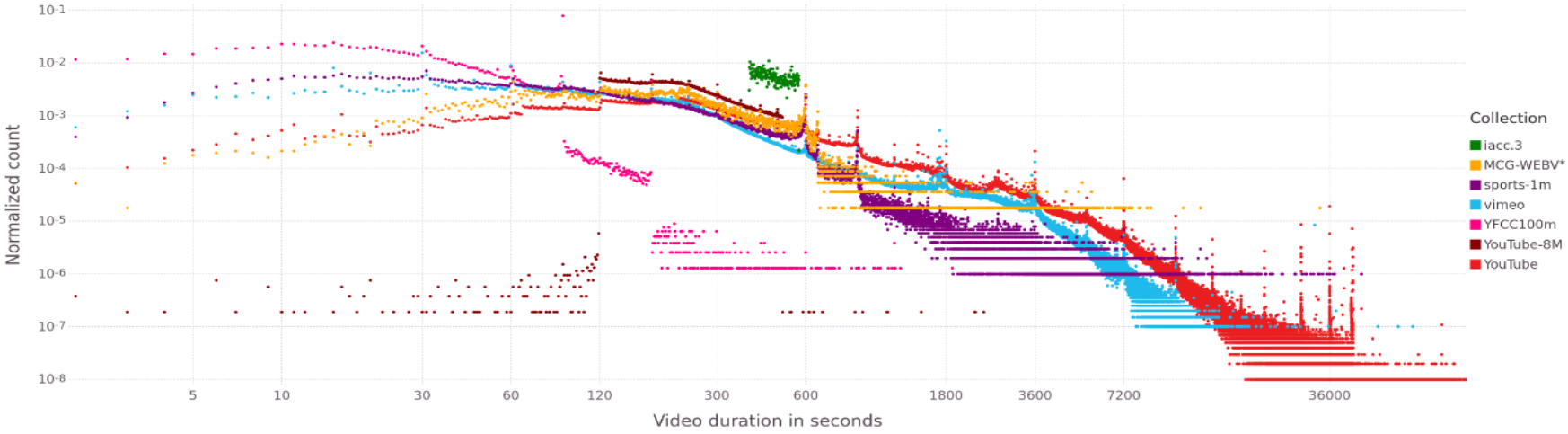
The state of Web Video

- In order for research to be reproducible, standardized datasets are necessary which can be shared freely
- The current state of Web Video in the wild is not or no longer represented accurately by research video collections [1]
- Other datasets exist, but they largely focus on a particular research question and are hence not widely applicable
- A new dataset of free contemporary and representative general purpose video material is necessary

[1] Rossetto, L., & Schuldt, H. (2017). Web video in numbers-an analysis of web-video metadata. *arXiv preprint arXiv:1707.01340*.



Age-distribution of common video collections vs what is found in the wild [1]



Duration-distribution of common video collections vs what is found in the wild [1]

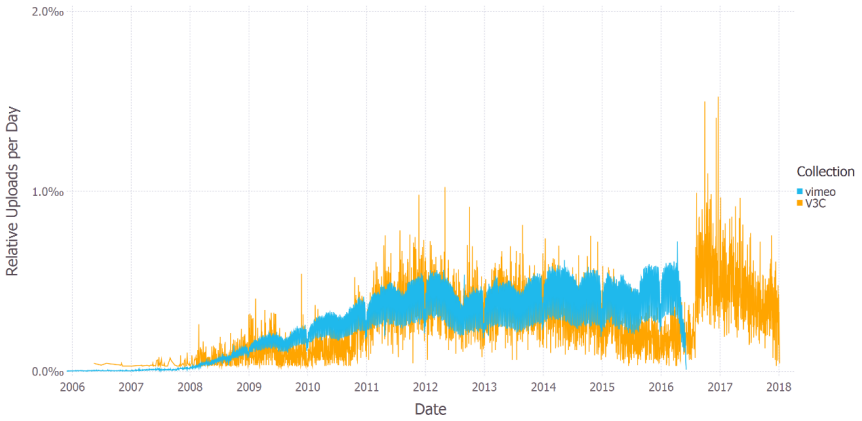
Vimeo Creative Commons Collection

The Vimeo Creative Commons Collection (V3C) [2] consists of '**free**' video material sourced from the web video platform **vimeo.com**. *It is designed to contain a wide range of content which is representative of what is found on the platform in general*. All videos in the collection have been released by their creators under a **Creative Commons License** which allows for unrestricted redistribution.

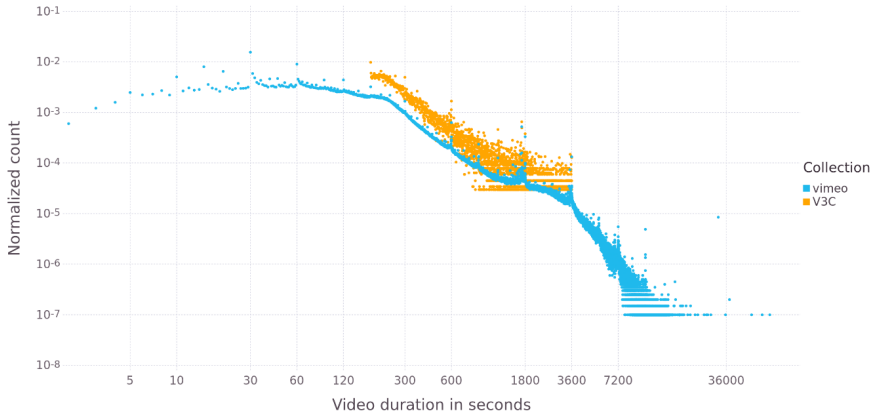
Partition	V3C1	V3C2	V3C3	**Total**
File Size	2.4TB	3.0TB	3.3TB	**8.7TB**
Number of Videos	7'475	9'760	11'215	**28'450**
Combined Video Duration	1000 hours, 23 minutes, 50 seconds	1300 hours, 52 minutes, 48 seconds	1500 hours, 8 minutes, 57 seconds	**3801 hours, 25 minutes, 35 seconds**
Mean Video Duration	8 minutes, 2 seconds	7 minutes, 59 seconds	8 minutes, 1 seconds	**8 minutes, 1 seconds**
Number of Segments	1'082'659	1'425'454	1'635'580	**4'143'693**

[2] Rossetto, L., Schuldt, H., Awad, G., & Butt, A. (2019). V3C – a Research Video Collection. *Proceedings of the 25th International Conference on MultiMedia Modeling*.

V3C Uploads and Duration



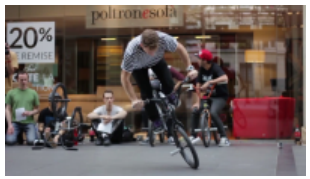
Age-distribution of the V3C in comparison with the vimeo data from [1]



Duration-distribution of the V3C in comparison with the vimeo data from [1]

V3C Content

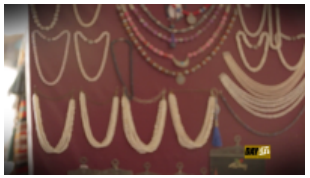
- Original Videos
- Video metadata from vimeo
- Automatically generated [3] video shot boundaries
- Lossless video keyframes for every segment
- Thumbnail image for every keyframe



#00001


#00072

#00314


#00885


#01411


#01976


#02539


#03827

[3] Rossetto, L., Giangreco, I., & Schuldt, H. (2014, December). Cineast: a multi-feature sketch-based video retrieval engine. In *Multimedia (ISM), 2014 IEEE International Symposium on*.

V3C1 demo-reel video

https://youtu.be/_k7Ksl8gPyU