TRECVID-2010 Semantic Indexing task: Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
NIST

also with Franck Thollard, Andy Tseng, Bahjat Safadi (LIG) and Stéphane Ayache (LIF) and support from the Quaero Programme

Outline

- Task summary
- Evaluation details
  - Inferred average precision
  - Participants
- Evaluation results
  - Pool analysis
  - Results per category
  - Results per concept
  - Significance tests per category
- Global Observations
- Issues

Semantic Indexing task (1)

- Goal: Automatic assignment of semantic tags to video segments (shots)
- Secondary goals:
  - Encourage generic (scalable) methods for detector development.
  - Semantic annotation is important for filtering, categorization, browsing, searching, and browsing.
- Participants submitted two types of runs:
  - Full run Includes results for 130 concepts, from which NIST evaluated 30.
  - Lite run Includes results for 10 concepts.
- TRECVID 2010 SIN video data
  - Test set (IACC.1.A): 200 hrs, with durations between 10 seconds and 3.5 minutes.
  - Development set (IACC.1.tv10.training): 200 hrs, with durations just longer than 3.5 minutes.
  - Total shots: (Much more than in previous TRECVID years, no composite shots)
    - Development: 119,685
    - Test: 146,788
- Common annotation for 130 concepts coordinated by LIG/LIF/Quaero

Semantic Indexing task (2)

- Selection of the 130 target concepts
  - Include all the TRECVID "high level features" from 2005 to 2009 to favor cross-collection experiments
  - Plus a selection of LSCOM concepts so that:
    - we end up with a number of generic-specific relations among them for promoting research on methods for indexing many concepts and using ontology relations between them
    - we cover a number of potential subtasks, e.g. “persons” or “actions” (not really formalized)
  - It is also expected that these concepts will be useful for the content-based (known item) search task.
- Set of 116 relations provided:
  - 111 “implies” relations, e.g. “Actor implies Person”
  - 5 “excludes” relations, e.g. “Daytime_Outdoor excludes Nighttime”

Semantic Indexing task (3)

- NIST evaluated 20 concepts and Quaero evaluated 10 features
- 20 more concepts to be released by Quaero but not part of the official TRECVID 2010 results
- Four training types were allowed
  - A - used only IACC training data
  - B - used only non-IACC training data
  - C - used both IACC and non-IACC TRECVID (S&V and/or Broadcast news) training data
  - D - used both IACC and non-IACC non-TRECVID training data

Datasets comparison

	TV2007	TV2008= TV2007 + New	TV2009 = TV2008 + New	TV2010
Dataset length (hours)	~100	~200	~380	~400
Master shots	36,262	72,028	133,412	266,473
Unique program titles	47	77	184	N/A

Number of runs for each training type

REGULAR FULL RUNS	A	B	C	D
Only IACC data	87			
Only non-IACC data		1		
Both IACC and non-IACC TRECVID data			6	
Both IACC and non-IACC non-TRECVID data				7
LIGHT RUNS	**A**	**B**	**C**	**D**
Only IACC data	127			
Only non-IACC data		6		
Both IACC and non-IACC TRECVID data			7	
Both IACC and non-IACC non-TRECVID data				10
Total runs (150)	127 84.7%	6 4%	7 4.6%	10 6.6%

30 concepts evaluated

- 4 Airplane_flying*
- 6 Animal
- 7 Asian_People
- 13 Bicycling
- 15 Boat_ship*
- 19 Bus*
- 22 Car_Racing
- 27 Cheering
- 28 Cityscape*
- 29 Classroom*
- 38 Dancing
- 39 Dark-skinned_People
- 41 Demonstration_Or_Protest*
- 44 Doorway
- 49 Explosion_Fire
- 52 Female-Human-Face-Closeup
- 53 Flowers
- 58 Ground_Vehicles
- 59 Hand*
- 81 Mountain
- 84 Nighttime*
- 86 Old_People
- 100 Running
- 105 Singing*
- 107 Sitting_down
- 115 Swimming
- 117 Telephones*
- 120 Throwing
- 126 Vehicle
- 127 Walking

-The 10 marked with "*" are a subset of those tested in 2008 & 2009

Evaluation

- Each feature assumed to be binary: absent or present for each master reference shot
- Task: Find shots that contain a certain feature, rank them according to confidence measure, submit the top 2000
- NIST sampled ranked pools and judged top results from all submissions
- Evaluated performance effectiveness by calculating the *inferred average precision* of each feature result
- Compared runs in terms of **mean** *inferred average precision* across the:
  - 30 feature results for full runs
  - 10 feature results for lite runs

Inferred average precision (infAP)

- Developed* by Emine Yilmaz and Javed A. Aslam at Northeastern University
- Estimates average precision surprisingly well using a surprisingly small sample of judgments from the usual submission pools
- This means that more features can be judged with same annotation effort
- Experiments on previous TRECVID years feature submissions confirmed quality of the estimate in terms of actual scores and system ranking

* J.A. Aslam, V. Pavlu and E. Yilmaz, ***Statistical Method for System Evaluation Using Incomplete Judgments*** Proceedings of the 29th ACM SIGIR Conference, Seattle, 2006.

Motivation for xinfAP and pooling strategy

- to make the evaluation more sensitive to shots returned below the lowest rank (~100) previously pooled and judged
- to adjust the sampling to match the relative importance of highest ranked items to average precision
- to exploit more infAP's ability to estimate of AP well even at sampling rates much below the 50% rate used in previous years

2010: mean extended Inferred average precision (xinfAP)

- 3 pools were created for each concept and sampled as:
  - Top pool (ranks 1-10) sampled at 100%
  - Middle pool (ranks 11-100) sampled at 20%
  - Bottom pool (ranks 101-2000) sampled at 5%

30 concepts	10 lite concepts
117,058 total judgments	49,253 total judgments
6958 total hits	2237 total hits
2700 Hits at ranks (1-10)	970 Hits at ranks (1-10)
2235 Hits at ranks (11-100)	755 Hits at ranks (11-100)
2023 Hits at ranks (101-2000)	512 Hits at ranks (101-2000)

- Judgment process: one assessor per concept, watched complete shot while listening to the audio.
- infAP was calculated using the judged and unjudged pool by sample_eval
- Random run problem: evaluation of non-pooled submissions?

2010 : 39/69 Finishers

---	***	KIS	***	---	SIN	Aalto University School of Science and Technology
---	---	---	---	---	SIN	Aristotle University of Thessaloniki
CCD	INS	KIS	---	SED	SIN	Beijing University of Posts and Telecom.-MCPRL
CCD	***	---	***	---	SIN	Brno University of Technology
---	***	KIS	MED	SED	SIN	Carnegie Mellon University - INF
CCD	---	KIS	---	***	SIN	City University of Hong Kong
---	***	---	MED	---	SIN	Columbia University / UCF
---	***	---	---	---	SIN	DFKI-MADM
---	***	---	***	***	SIN	EURECOM
---	***	---	---	---	SIN	Florida International University
---	***	---	---	---	SIN	France Telecom Orange Labs (Beijing)
---	---	---	---	---	SIN	Fudan University
***	---	---	---	---	SIN	Fuzhou University
---	INS	KIS	MED	---	SIN	Informatics and Telematics Inst.
---	---	---	***	SED	SIN	INRIA-willow
---	***	---	---	---	SIN	Inst. de Recherche en Informatique de Toulouse - Equipe SAMoVA
---	INS	---	---	***	SIN	JOANNEUM RESEARCH
---	INS	KIS	MED	***	SIN	KB Video Retrieval
---	---	---	---	---	SIN	Laboratoire d'Informatique Fondamentale de Marseille
---	INS	***	***	---	SIN	Laboratoire d'Informatique de Grenoble for IRIM
---	---	---	---	---	SIN	LSIS / UMR CNRS & USTV
CCD	INS	***	***	***	SIN	National Inst. of Informatics
---	***	---	---	---	SIN	National Taiwan University
***	***	***	***	SED	SIN	NHK Science and Technical Research Laboratories
---	---	KIS	---	---	SIN	NTT Communication Science Laboratories-UT
---	***	***	---	---	SIN	Oxford/IIIT
---	---	---	***	---	SIN	Quaero consortium
---	---	***	---	---	SIN	Ritsumeikan University

** : group didn't submit any runs

-- : group didn't participate

2010 : 39/69 Finishers

```
--- --- --- --- --- SIN   SHANGHAI JIAOTONG UNIVERSITY-IS
*** *** *** *** SED SIN   Tianjin University
--- *** --- *** SED SIN   Tokyo Inst. of Technology + Georgia Inst. of Technology
CCD *** --- --- *** SIN   TUBITAK - Space Technologies Research Inst.
--- --- --- --- --- SIN   Universidad Carlos III de Madrid
--- INS KIS *** *** SIN   University of Amsterdam
--- *** *** *** *** SIN   University of Electro-Communications
--- --- --- *** *** SIN   University of Illinois at Urbana-Champaign & NEC Labs.America
*** *** --- *** --- SIN   University of Marburg
*** *** *** --- *** SIN   University of Sfax
--- --- *** --- *** SIN   Waseda University
```

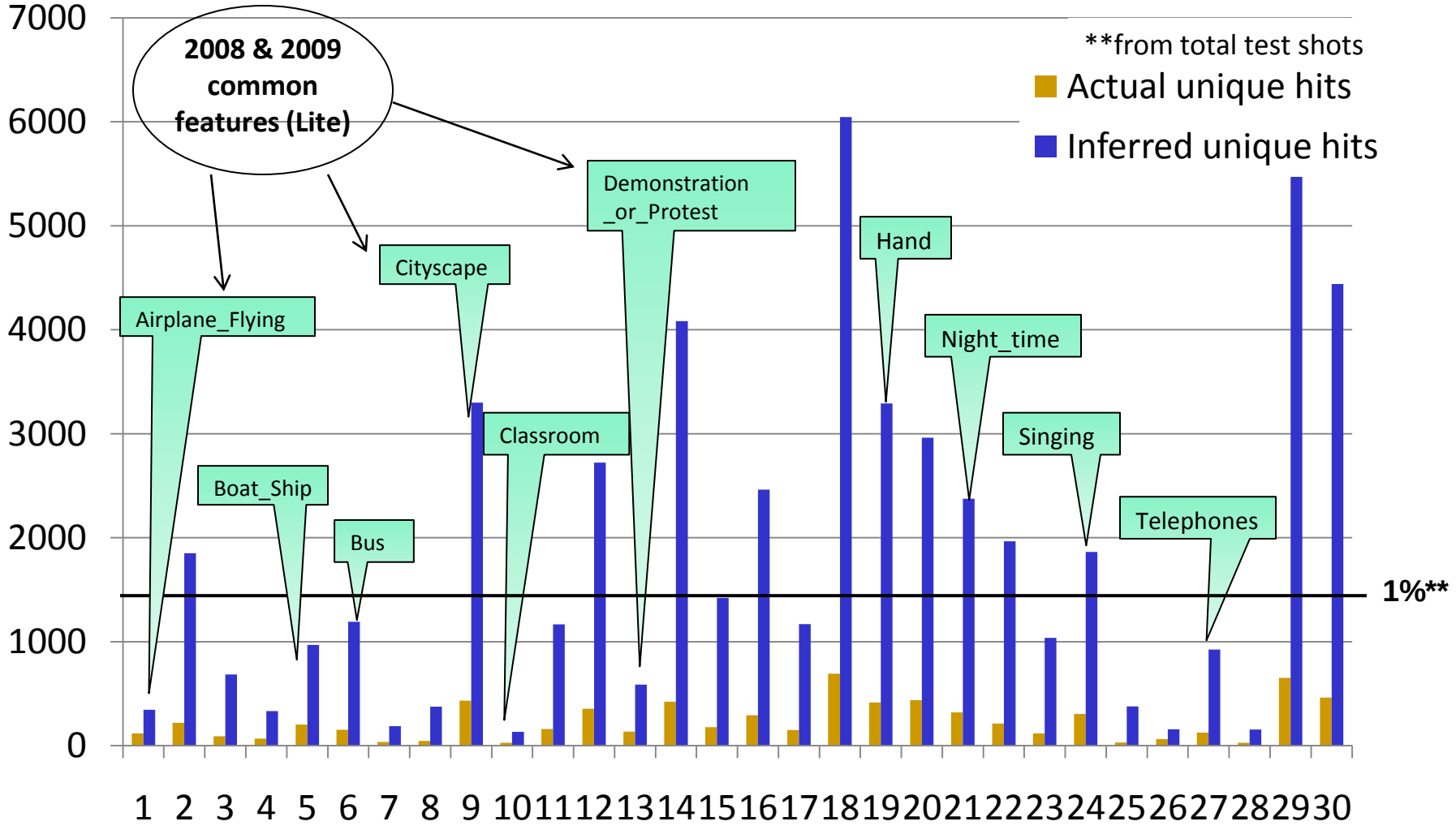
Almost same steady ratio of participation and finishing

	Task finishers	Participants
2010	39	69
2009	42	70
2008	43	64
2007	32	54
2006	30	54
2005	22	42
2004	12	33

** : group didn't submit any runs

-- : group didn't participate

Frequency of hits varies by feature



2 Animal	3 Asian_People	4 Bicycling	7 Car_Racing	8 Cheering	11 Dancing	12 Dark-skinned_People	14 Doorway	15 Explosion_Fire	16 Female-Human-Face-Closeup
17 Flowers	18 Ground_Vehicles	20 Mountain	22 Old_People	23 Running	25 Sitting_down	26 Swimming	28 Throwing	29 Vehicle	30 Walking

True shots contributed uniquely by team

Full runs

Team	No. of Shots	Team	No. of shots
CMU	67	Brn	15
MUG	59	Fuz	14
FIU	50	LIF	12
KBV	35	Mar	12
UEC	28	UC3	10
NII	28	VIR	10
DFK	24	CU	7
inr	24	Uza	6
TT+GT	21	NHK	5
MM	20	Pic	4
IIP	19	FTR	2
NEC	18	Fud	2

TV2009

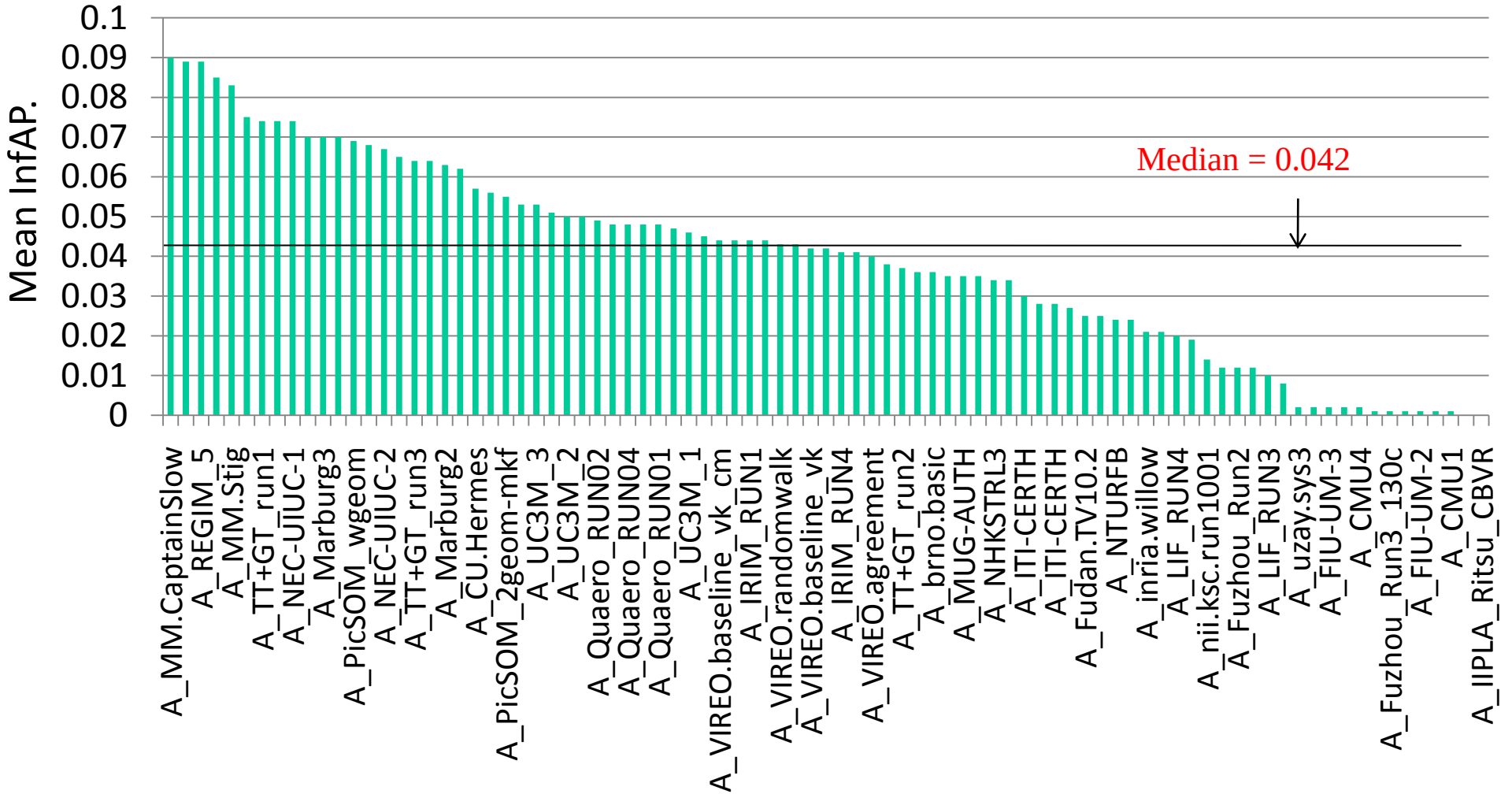
Team	Shots
BRN	2
FIU	4
FZU	4
IRI	1
ISM	3
ITI	3
LSI	10
NHK	5
NII	8
SJT	1
TIT	2
Tsi	2
UEC	2
UKA	1
VIT	2
VPU	1
XJT	3
ZJU	4
Uza	8

Lite runs

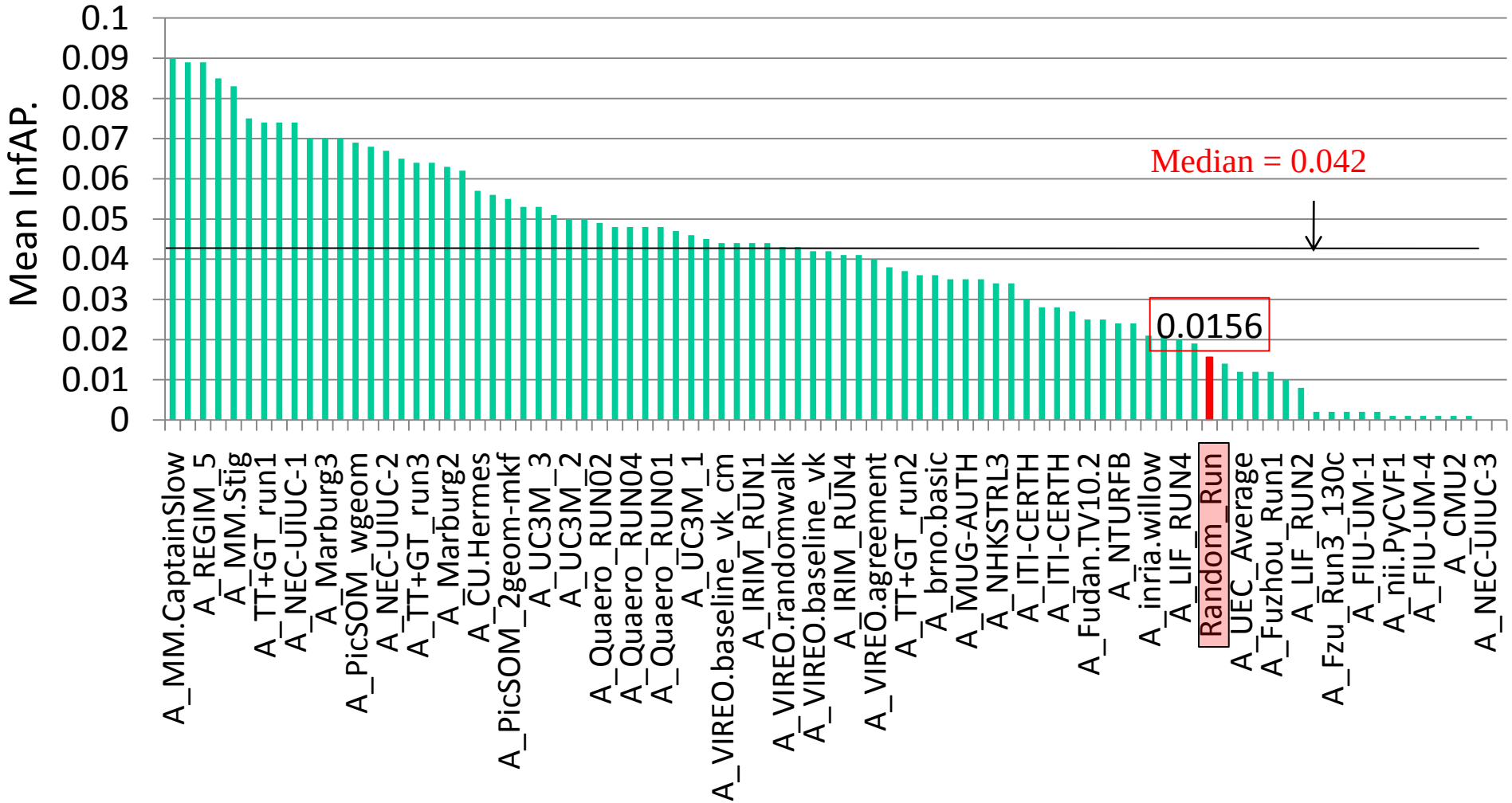
Team	No. of Shots	Team	No. of shots
CMU	16	DFK	2
IRI	10	FTR	2
kml	10	Fuz	2
MUG	8	IIP	2
KBV	6	NEC	2
MMM	5	ntt	2
TT+GT	5	CU	1
XJT	5	JRS	1
FIU	4	LIF	1
nii	4	MCP	1
Eur	3	NHK	1
inr	3	UEC	1
Oxi	3	Uza	1
SJT	3	VIR	1
UC3	3		

No. of unique shots found are **MORE** than what was found in TV2009 (more shots this year)

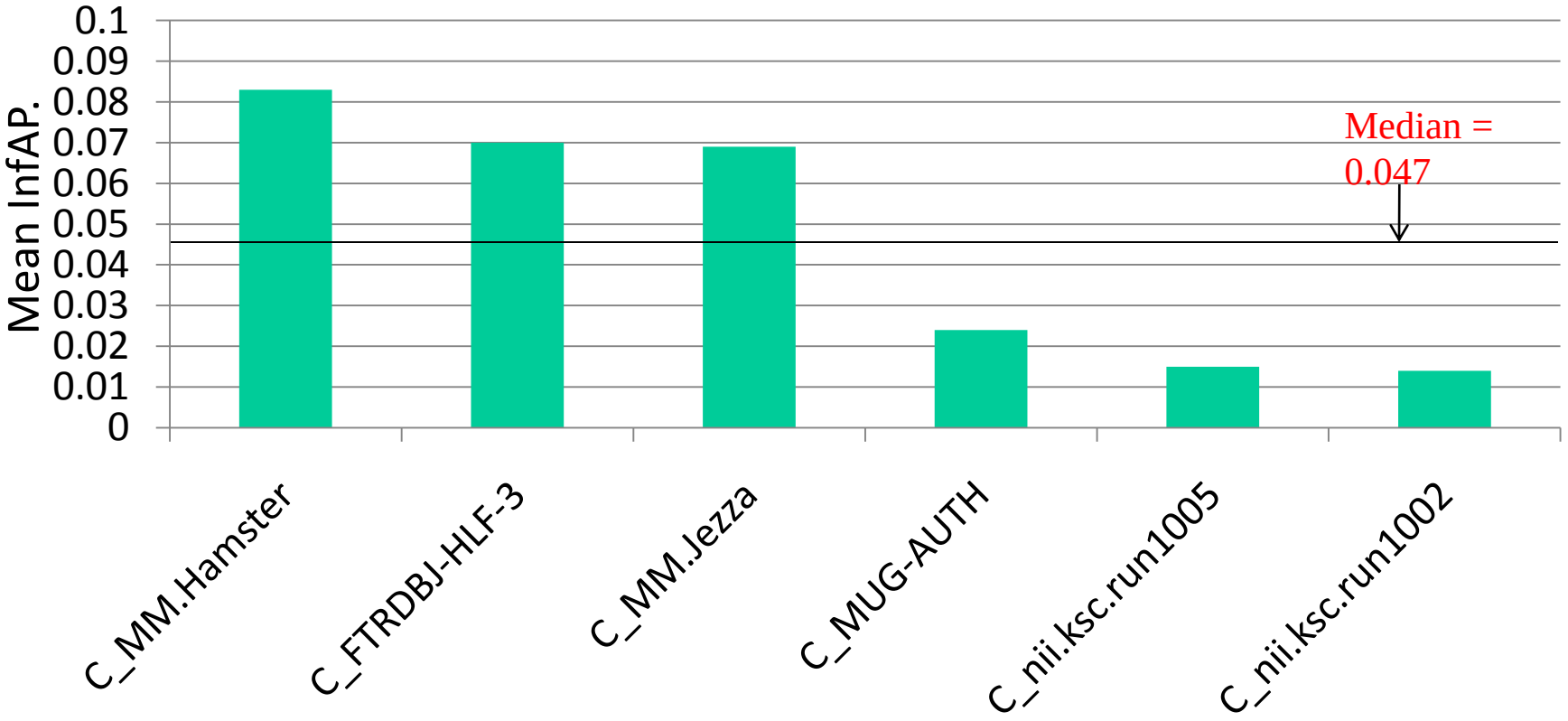
Category A results (Full runs)



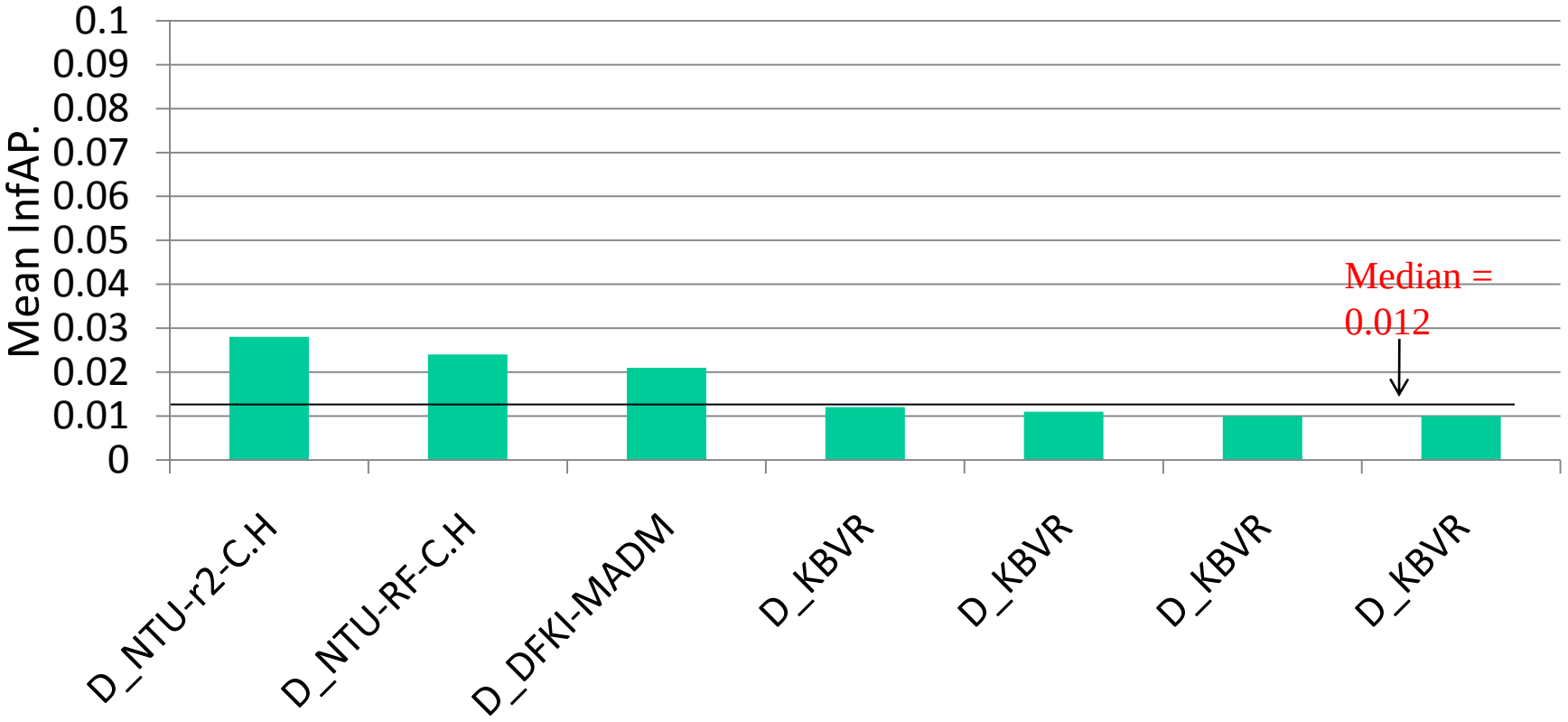
Category A results (Full runs)



Category C results (Full runs)

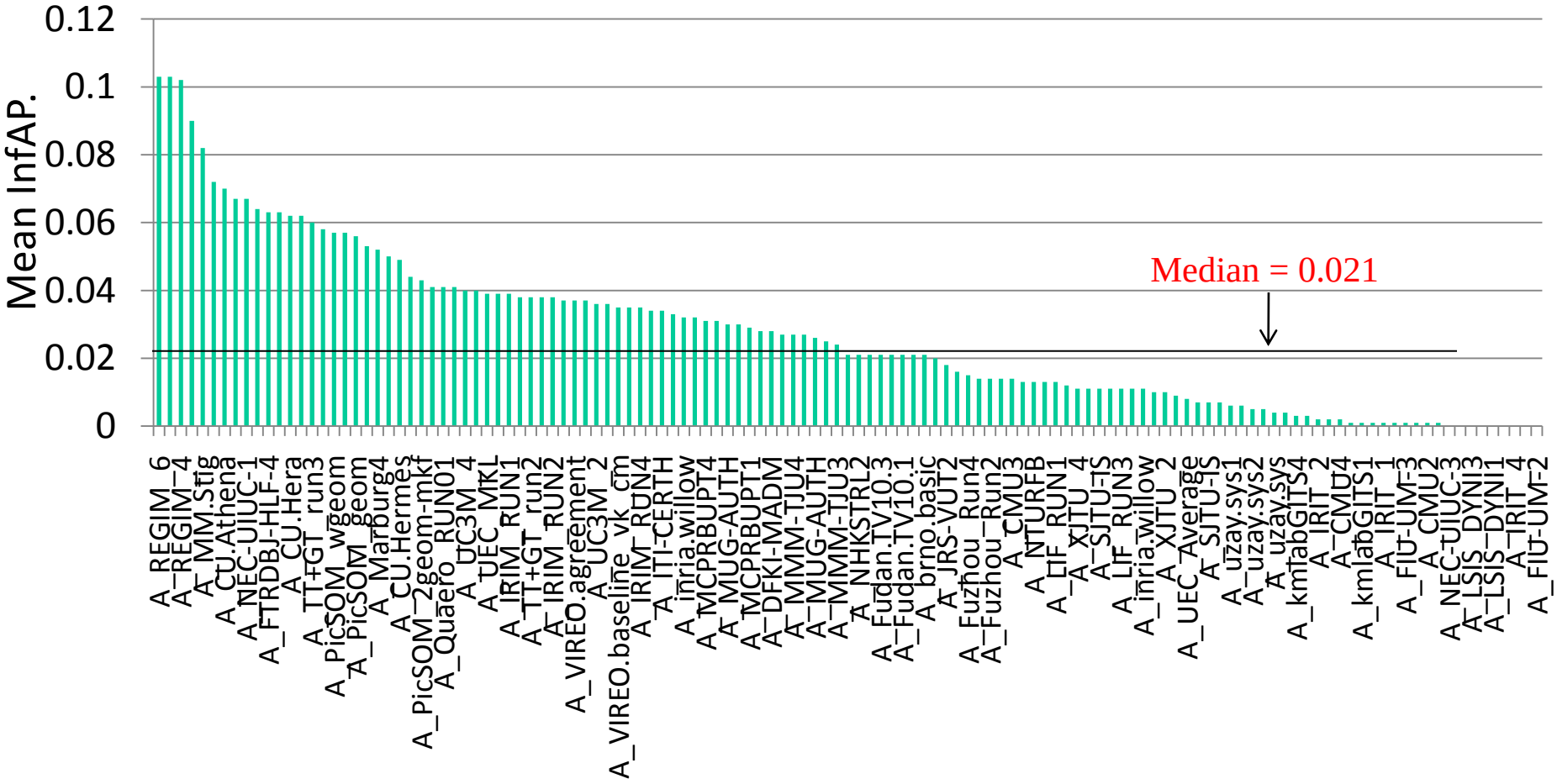


Category D results (Full runs)

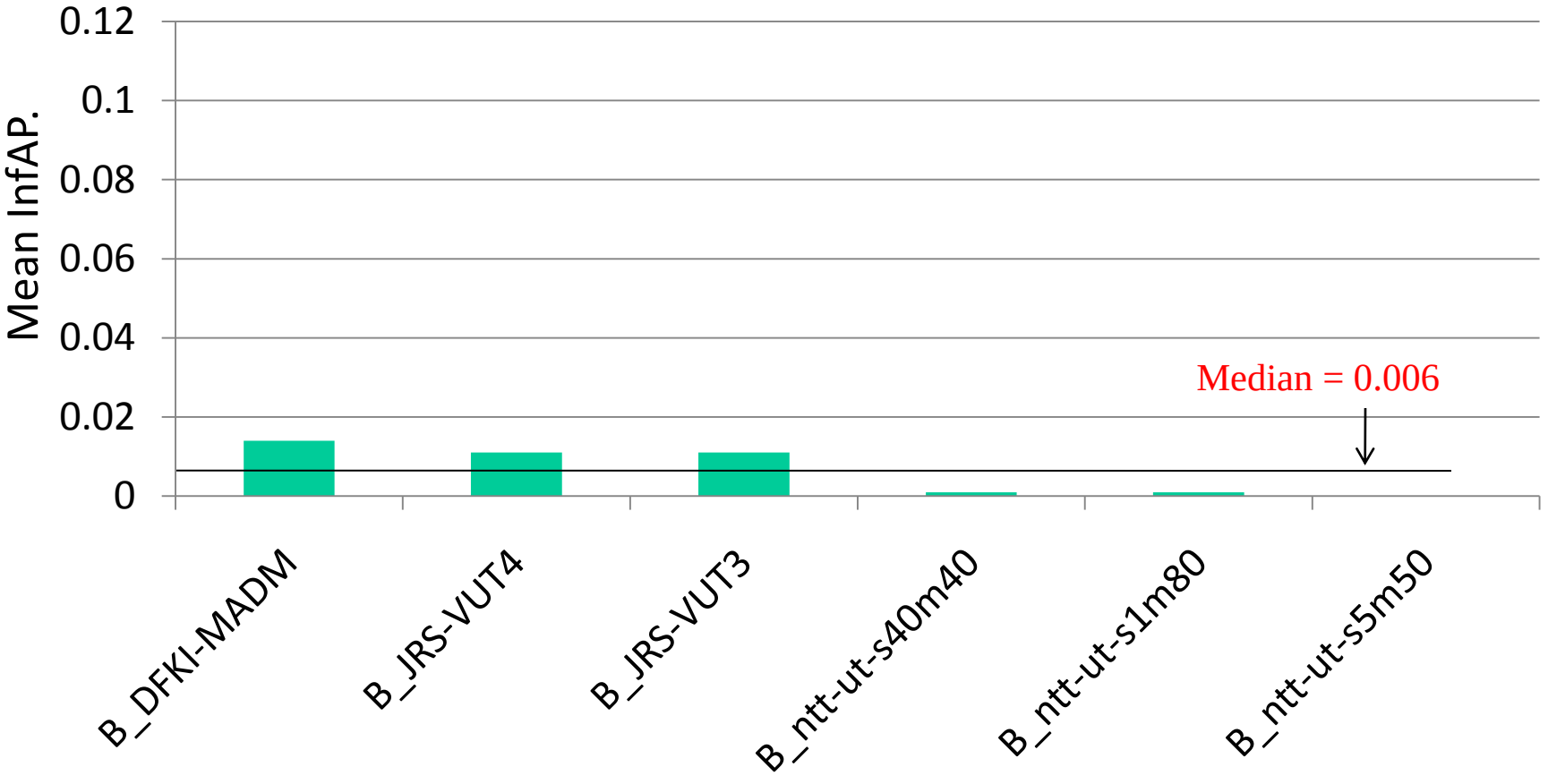


Note: Category B has only 1 run (B_DFKI-MADM) with score = 0.013

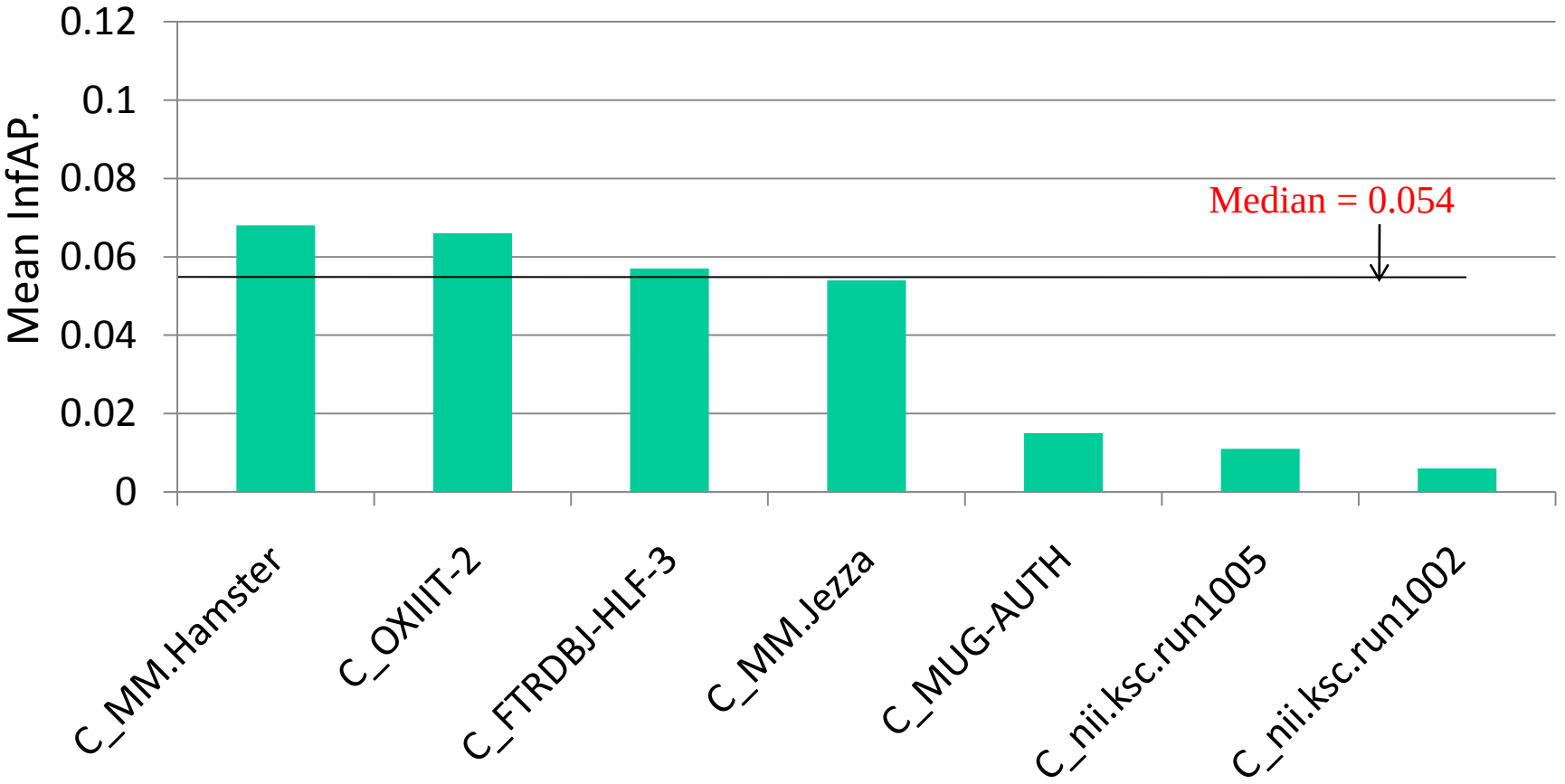
Category A results (Lite runs)



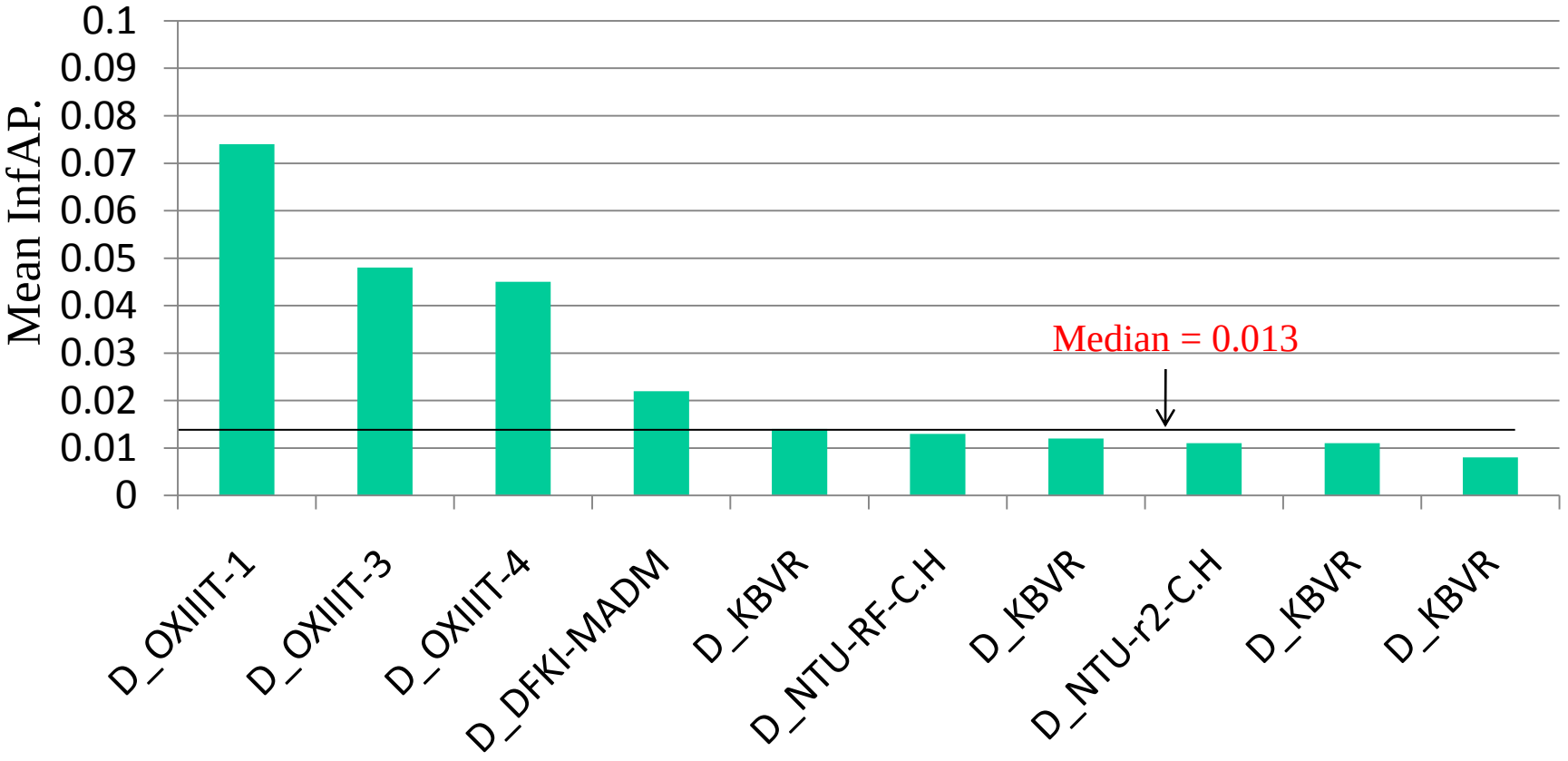
Category B results (Lite runs)



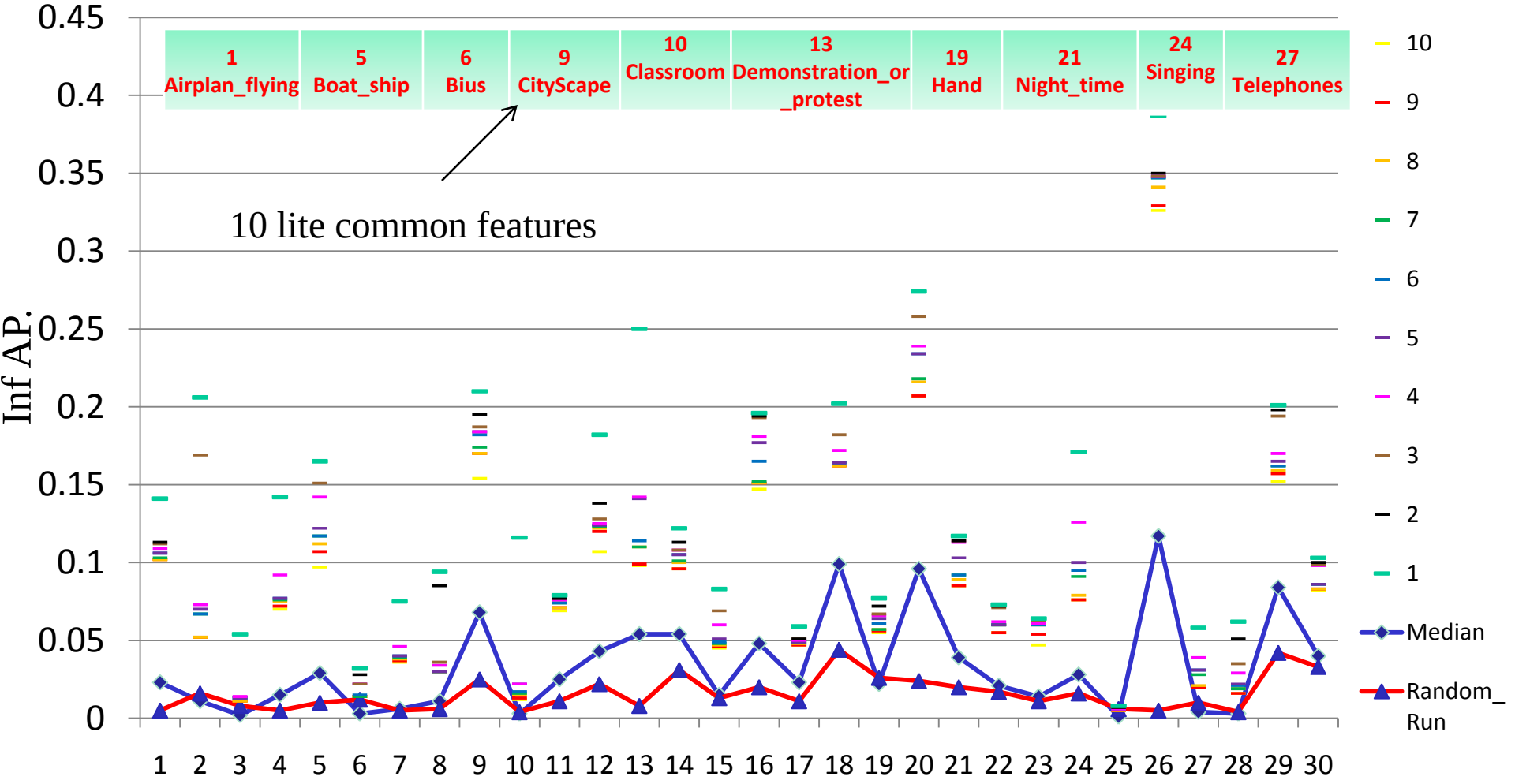
Category C results (Lite runs)



Category D results (Lite runs)

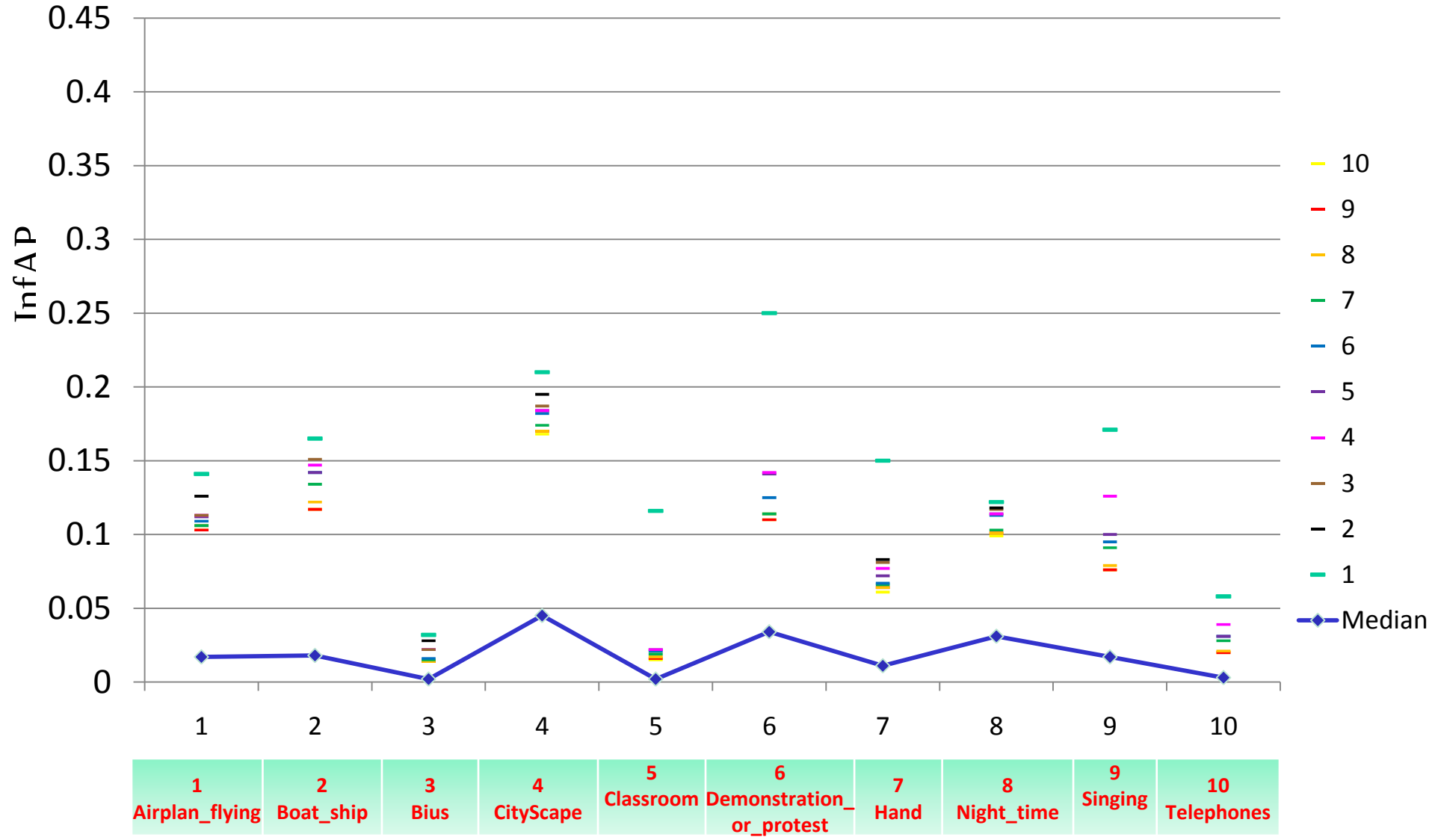


Top 10 InfAP scores by feature (Full runs)



2 Animal	3 Asian_People	4 Bicycling	7 Car_Racing	8 Cheering	11 Dancing	12 Dark-skinned_People	14 Doorway	15 Explosion_Fire	16 Female-Human-Face-Closeup
17 Flowers	18 Ground_Vehicles	20 Mountain	22 Old_People	23 Running	25 Sitting_down	26 Swimming	28 Throwing	29 Vehicle	30 Walking

Top 10 InfAP scores for 10 common features (Lite AND Full runs)



Significant differences among top 10 A-category full runs (using randomization test, $p < 0.05$)

Run name	(mean infAP)
A_MM.CaptainSlow _4	(0.090)
A_REGIM_6 _3	(0.089)
A_REGIM_5_1	(0.089)
A_REGIM_4 _2	(0.085)
A_MM.Stig _1	(0.083)
A_FTRDBJ-HLF-2_2	(0.075)
A_TT+GT_run1_1	(0.074)
A_NEC-UIUC-4_4	(0.074)
A_NEC-UIUC-1_1	(0.074)
A_PicSOM_2geom-max_2	(0.070)

- A_MM.CaptainSlow _4
  - A_FTRDBJ-HLF-2_2
  - A_MM.Stig _1
    - A_NEC-UIUC-4_4
    - A_NEC-UIUC-1_1
    - A_PicSOM_2geom-max_2

Some runs have higher scores but not significantly better than others with lower scores!

Significant differences among top 10 C-category full runs (using randomization test, p < 0.05)

Run name	(mean infAP)
C_MM.Hamster_3	(0.083)
C_FTRDBJ-HLF-3_3	(0.070)
C_MM.Jezza_2	(0.069)
C_MUG-AUTH_4	(0.024)
C_nii.ksc.run1005_1	(0.015)
C_nii.ksc.run1002_2	(0.014)

- C_MM.Hamster_3
  - C_FTRDBJ-HLF-3_3
    - C_MUG-AUTH_4
    - C_nii.ksc.run1005_1
    - C_nii.ksc.run1002_2
  - C_MM.Jezza_2
    - C_MUG-AUTH_4
    - C_nii.ksc.run1005_1
    - C_nii.ksc.run1002_2

Significant differences among top D-category full runs (using randomization test, p < 0.05)

Run name	(mean infAP)
D_NTU-r2-C.H_2	(0.028)
D_NTU-RF-C.H_1	(0.024)
D_DFKI-MADM_1	(0.021)
D_KBVR_2	(0.012)
D_KBVR_3	(0.011)
D_KBVR_4	(0.010)
D_KBVR_1	(0.010)

- D_NTU-r2-C.H_2
  - D_KBVR_2
    - D_KBVR_1
  - D_KBVR_3
  - D_KBVR_4
- D_NTU-RF-C.H_1
  - D_KBVR_1
  - D_KBVR_3
  - D_KBVR_4
- D_DFKI-MADM_1
  - D_KBVR_1
  - D_KBVR_4

Significant differences among top 10 A-category lite runs (using randomization test, $p < 0.05$)

Run name	(mean infAP)
A_REGIM_6_3	(0.103)
A_REGIM_5_1	(0.103)
A_REGIM_4_2	(0.102)
A_MM.CaptainSlow_4	(0.090)
A_MM.Stig_1	(0.082)
A_Eurecom_Weight_HE_3	(0.072)
A_CU.Athena_3	(0.070)
A_NEC-UIUC-4_4	(0.067)
A_NEC-UIUC-1_1	(0.067)
A_TT+GT_run1_1	(0.064)

- A_MM.CaptainSlow_4
  - A_CU.Athena_3
  - A_MM.Stig_1
  - A_NEC-UIUC-1_1
  - A_NEC-UIUC-4_4
- A_REGIM_5_1
  - A_NEC-UIUC-1_1
  - A_NEC-UIUC-4_4

Significant differences among top B-category lite runs (using randomization test, $p < 0.05$)

Run name	(mean infAP)
B_DFKI-MADM _2	(0.014)
B_JRS-VUT4_3	(0.011)
B_JRS-VUT3_4	(0.011)
B_ntt-ut-s40m40_1	(0.001)
B_ntt-ut-s1m80_3	(0.001)
B_ntt-ut-s5m50_2	(0.000)

- B_DFKI-MADM_2
  - B_ntt-ut-s40m40_1
  - B_ntt-ut-s1m80_3
  - B_ntt-ut-s5m50_2
- B_JRS-VUT4_3
  - B_ntt-ut-s40m40_1
  - B_ntt-ut-s1m80_3
  - B_ntt-ut-s5m50_2
- B_JRS-VUT3_4
  - B_ntt-ut-s40m40_1
  - B_ntt-ut-s1m80_3
  - B_ntt-ut-s5m50_2

Significant differences among top 10 C-category lite runs (using randomization test, $p < 0.05$)

Run name	(mean infAP)
C_MM.Hamster_3	(0.068)
C_OXIIIT-2_2	(0.066)
C_FTRDBJ-HLF-3_3	(0.057)
C_MM.Jezza_2	(0.054)
C_MUG-AUTH_4	(0.015)
C_nii.ksc.run1005_1	(0.011)
C_nii.ksc.run1002_2	(0.006)

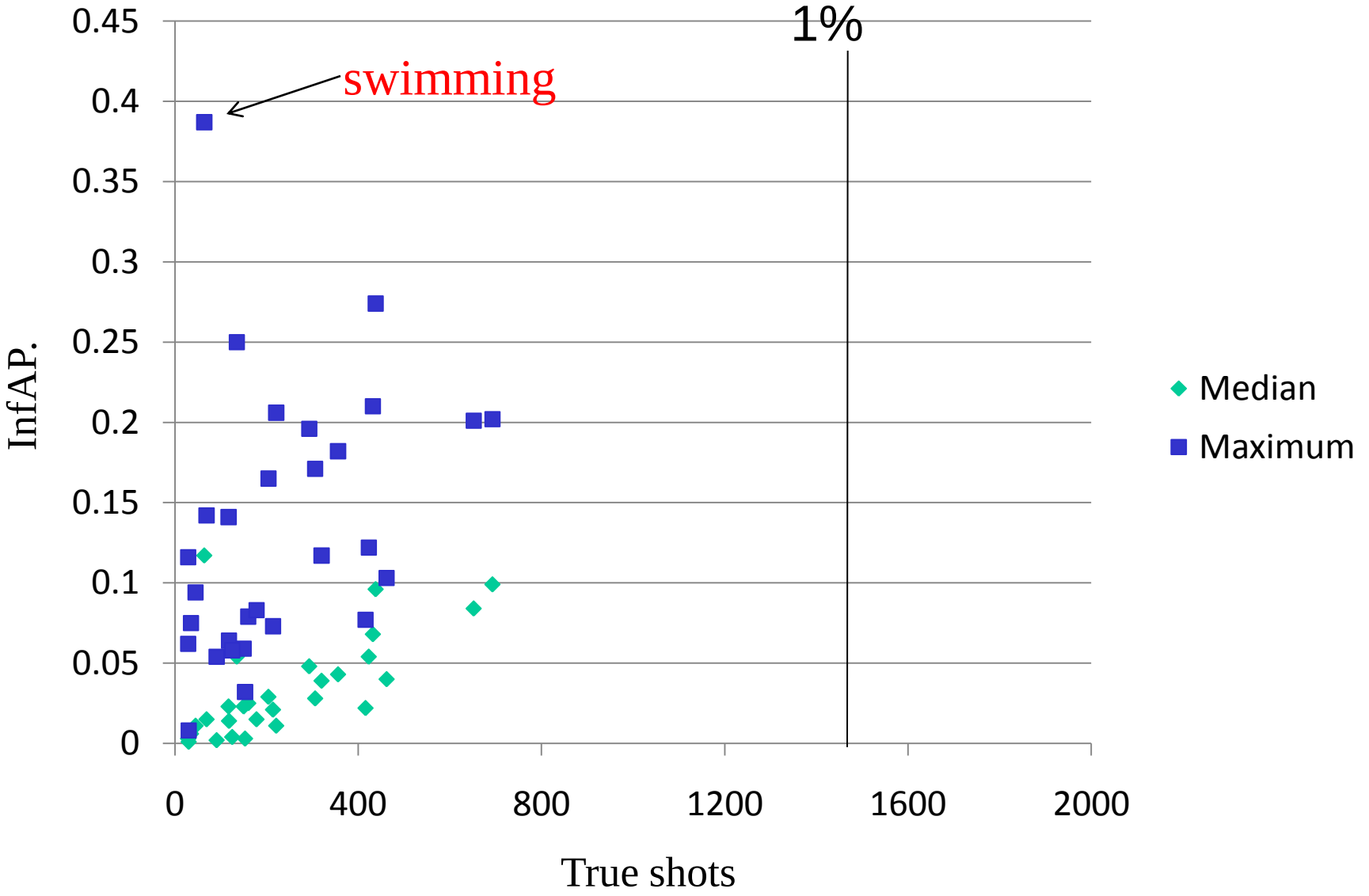
- C_MM.Hamster_3
  - C_MM.Jezza_2
    - C_MUG-AUTH_4
    - C_nii.ksc.run1005_1
    - C_nii.ksc.run1002_2
- C_OXIIIT-2_2
  - C_MUG-AUTH_4
  - C_nii.ksc.run1005_1
  - C_nii.ksc.run1002_2
- C_FTRDBJ-HLF-3_3
  - C_MUG-AUTH_4
  - C_nii.ksc.run1005_1
  - C_nii.ksc.run1002_2

Significant differences among top D-category lite runs (using randomization test, $p < 0.05$)

Run name	(mean infAP)
D_OXIIIT-1_1	(0.074)
D_OXIIIT-3_3	(0.048)
D_OXIIIT-4_4	(0.045)
D_DFKI-MADM_1	(0.022)
D_KBVR_3	(0.014)
D_NTU-RF-C.H_1	(0.013)
D_KBVR_4	(0.012)
D_NTU-r2-C.H_2	(0.011)
D_KBVR_2	(0.011)
D_KBVR_1	(0.008)

- D_OXIIIT-1_1
  - D_OXIIIT-3_3
    - D_KBVR_3
      - D_KBVR_1
      - D_KBVR_2
    - D_KBVR_4
    - D_NTU-RF-C.H_1
    - D_NTU-r2-C.H_2
    - D_DFKI-MADM_1
  - D_OXIIIT-4_4
    - D_KBVR_3
      - D_KBVR_1
      - D_KBVR_2
    - D_KBVR_4
    - D_NTU-RF-C.H_1
    - D_NTU-r2-C.H_2
    - D_DFKI-MADM_1

InfAp. Vs true shots in test data across 30 features



Observations (2009)

- Site experiments include:
  - focus on robustness, merging many different representations
  - comparing fusion strategies
  - efficiency improvements (e.g. GPU implementations)
  - analysis of more than one keyframe per shot
  - audio analysis
  - using temporal context information
  - analyzing motion information
  - automatic extraction of Flickr training data
- Fewer experiments using external training data (increased focus on category A)

Observations (2010)

- Site experiments include:
  - focus on robustness, merging many different representations
  - use of spatial pyramids
  - sophisticated fusion strategies
  - efficiency improvements (e.g. GPU implementations)
  - analysis of more than one keyframe per shot
  - audio analysis
  - using temporal context information
  - not so much use of motion information, metadata or ASR
  - use of training data from YouTube (not Flickr)

- Still not many experiments using external training data (main focus on category A)
- No improvement using external training data

Questions to participants

- How was the effect of the new data domain compared to S&V dataset in the past 3 years?
- How scalable was the systems dealing with the huge increase in no. of shots and concepts?
- How do we know whether the community as a whole achieves better results over the years?
  - Did any run their TV2009 system on TV2010 test data?
  - Did any run their system on tv2009 common 10 features?
- How to encourage submitting in category B, C, & D?
- Should we also look at detector training and testing speed?
- Any comments on the choice of the 130 concepts?

SIN 2011

- Same or similar task.
- Same type of data.
- Similar volume of data? Or still more?
- A third (Large scale, ~1000) set of concepts?
- Subtasks, e.g. persons, events, actions, locations, genres ...?
- Other classes of concepts? Emotions?
- Multiple levels of relevance for positive samples?
- Or ranking of positive samples?
- Encourage and provide infrastructure for sharing contributed elements: low-level features, detection scores, ...
- Possibility to submit unpooled runs to encourage the evaluation of the effect of many parameters.
- Derived measure: GMAP to better recognize work on difficult concepts?