TRECVID-2009 High-Level Feature task: Overview

Wessel Kraaij
TNO // Radboud University

George Awad
NIST

Outline

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

High-level feature task (1)

- Goal: Build benchmark collection for visual concept detection methods
- Secondary goals:
  - encourage generic (scalable) methods for detector development
  - semantic annotation is important for search/browsing
- Participants submitted runs for 10 features from those tested in 2008 and 10 new features for 2009.
- Common annotation for new features coordinated by LIG/LIF
- TRECVID 2009 video data
  - Netherlands Institute for Sound and Vision (~**380 hours** of news magazine, science news, news reports, documentaries, educational programming and archival video in MPEG-1).
  - ~100 hours for development (50 hrs TV2007 dev. + 50 hrs TV2007 test)
  - 280 hours for test (100 hrs TV2008 test + new 180 hrs TV2009 test)

High-level feature task (2)

- NIST evaluated 20 features using a 50% random sample of the submission pools (Inferred AP)
- Four training types were allowed
  - A :
    - Systems trained on only common TRECVID development collection data OR
    - (formerly B) systems trained on only common development collection data but not on (just) common annotation of it.
  - C : System is not of type A.
  - a : same as A but no training data specific to any sound and vision data has been used (TV6 and before).
  - c : same as C but no training data specific to any sound and vision data has been used.
- Training category B,b has been dropped allowing systems to focus on:
  - If training data was from the common development & annotation.
  - If training data belongs to S&V data.

Run type determined by sources of training data

	A	C	a	c
TV3-6 (Broadcast news)	🟩	🟩	🟩	🟩
TV7,8,9 (S&V)	🟩	🟩		
Other training data		🟩		🟩

TV2007 vs TV2008 vs TV2009 datasets

	TV2007	TV2008	TV2009 = TV2008 + New
Dataset length (hours)	~100	~200	[illegible]
Shots	18,142	35,766	93,902
Unique program titles	47	77	184

More diversity from the long tail

TV2009 10 new features selection

- Participants suggested features that include:
  - Parts of natural scenes.
  - Child.
  - Sports.
  - Non-speech audio component.
  - People and objects in action.
  - Frequency in consumer video.
- NIST basic selection criteria:
  - Features has to be moderately frequent
  - Has clear definition
  - Be of use in searching
  - No overlap with previously used topics/features

20 features evaluated

- 1 Classroom*
- 2 Chair
- 3 Infant
- 4 Traffic_intersection
- 5 Doorway
- 6 Airplane_flying*
- 7 Person_playing_musical_instrument
- 8 Bus*
- 9 Person_playing_soccer
- 10 Cityscape*
- 11 Person_riding_bicycle
- 12 Telephone*
- 13 Person_eating
- 14 Demonstration_Or_Protest*
- 15 Hand*
- 16 People_dancing
- 17 Nighttime*
- 18 Boat_ship*
- 19 Female_human_face_closeup
- 20 Singing*

-Features were selected to be better suited to sound and vision data

- The 10 marked with “*” are a subset of those tested in 2008

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 pooled 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 20 feature results.

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
- Cost is less detail and more variability for each feature result in a run
- Experiments on TRECVID 2005, 2006, 2007 & 2008 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.

2009: Inferred average precision (infAP)

- Submissions for each of 20 features were pooled down to about 100 items (so that each feature pool contained ~ 6500 - 7000 shots) (2008: 130 items, 6777 shots)
  - varying pool depth per feature
- A 50% random sample of each pool was then judged:
- 68,270 total judgments (TV8: 67,774)
- 7036 total hits
- Judgment process: one assessor per feature, watched complete shot while listening to the audio.
- infAP was calculated using the judged and unjudged pool by trec_eval

2009 : 42/70 Finishers

Asahikasei Co.	--	FE	--	CD
Brno University of Technology	ED	FE	SE	**
Beijing University of Posts and Telecom.-MCPRL	ED	FE	SE	CD
Carnegie Mellon University	ED	FE	SE	--
Institut EURECOM	ED	FE	--	--
Florida International University	--	FE	--	**
France Telecom research & development - Beijing	--	FE	--	--
Fudan University	--	FE	--	CD
Fuzhou University	ED	FE	--	**
IBM Watson Research Center	ED	FE	SE	CD
Tsinghua University-IMG	ED	FE	SE	CD
GDR ISIS - IRIM consortium	ED	FE	SE	**
UPS - IRIT - SAMoVA	--	FE	SE	--
The Institute of Statistical Mathematics	--	FE	--	--
IUPR - DFKI	--	FE	--	--
Chinese Academy of Sciences-IVA_NLPR_IA_CAS	ED	FE	--	**
Laboratoire d'Informatique Fondamentale de Marseille	--	FE	--	--
Laboratoire d'Informatique de Grenoble	--	FE	SE	--
LSIS, Université Sud Toulon Var	--	FE	--	--
University of Marburg	--	FE	SE	--
University of Amsterdam	ED	FE	SE	--
Centre for Research and Technology Hellas	--	FE	SE	--
Tsinghua University-MPAM	--	FE	--	CD
NHK Science and Technical Research Laboratories	**ED**	**FE**	**SE**	**
National Institute of Informatics	ED	FE	SE	CD
Oxford/IIIT	--	FE	--	--
Helsinki University of Technology TKK	--	FE	SE	--

** : group didn't submit any runs

bold: did not submit HLF runs in 2008

-- : group didn't participate

2009 : 42/70 Finishers

Peking University-PKU-ICST	ED	FE	SE	**
Laboratoire REGIM	ED	FE	SE	--
Shanghai Jiao Tong University-IICIP	ED	FE	SE	--
Shanghai Jiao Tong University-IS	--	FE	--	--
Universidad Carlos III de Madrid	--	FE	--	--
Tokyo Institute of Technology	ED	FE	--	--
TUBITAK UZAY	ED	FE	--	CD
University of Central Florida	--	FE	--	--
University of Electro-Communications	ED	FE	SE	--
University of Karlsruhe (TH)	--	FE	--	--
City University of Hong Kong	ED	FE	SE	CD
Aristotle University of Thessaloniki	--	FE	SE	--
Universidad Autónoma de Madrid	ED	FE	SE	**
Xi'an Jiaotong University	--	FE	SE	CD
Zhejiang University	--	FE	SE	--

HLF keeps attracting participants

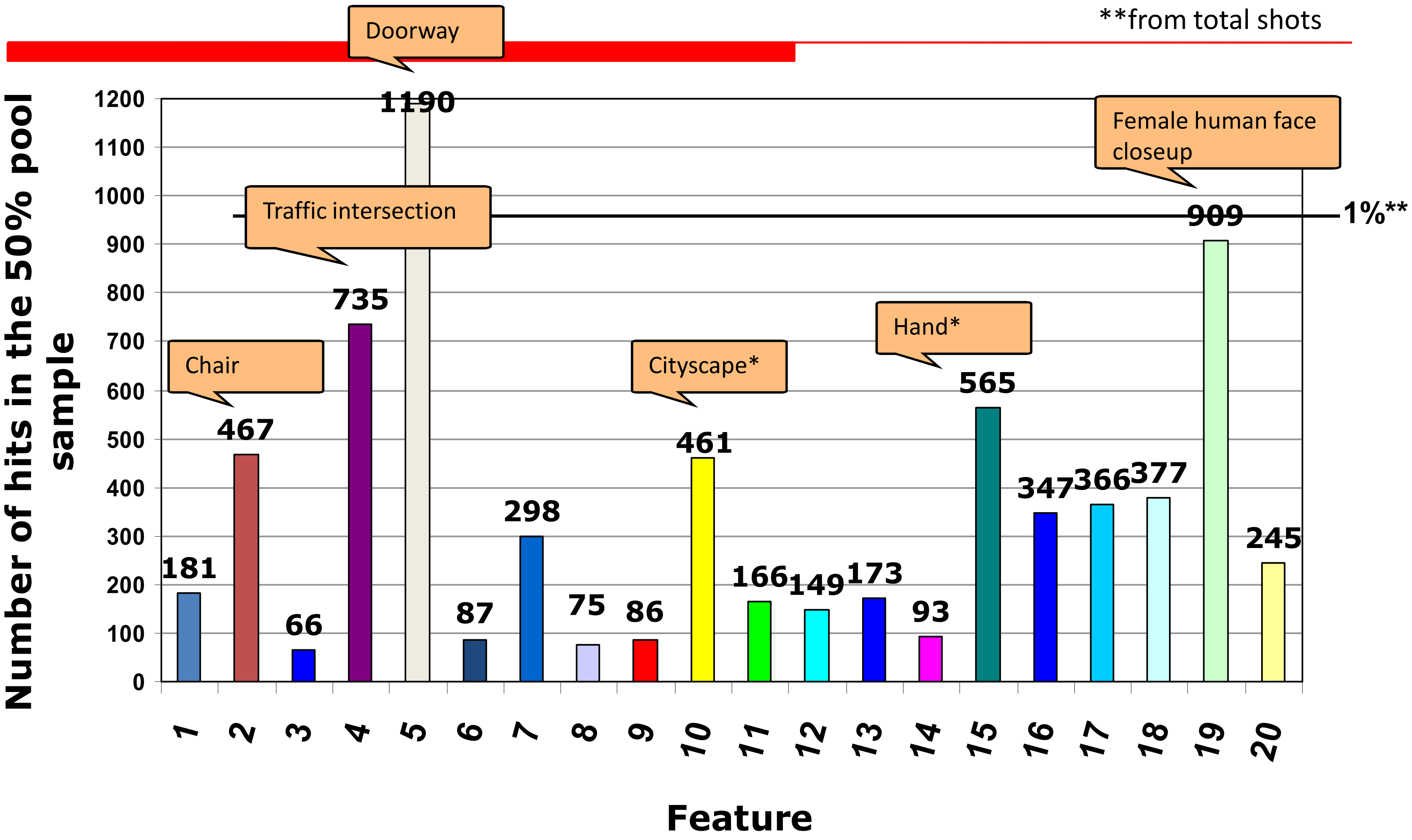
roughly 35% "new"

	HLF finisher	TV09 finisher
2009	42	70
2008	43	115
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



**from total shots

Doorway

Traffic intersection

Chair

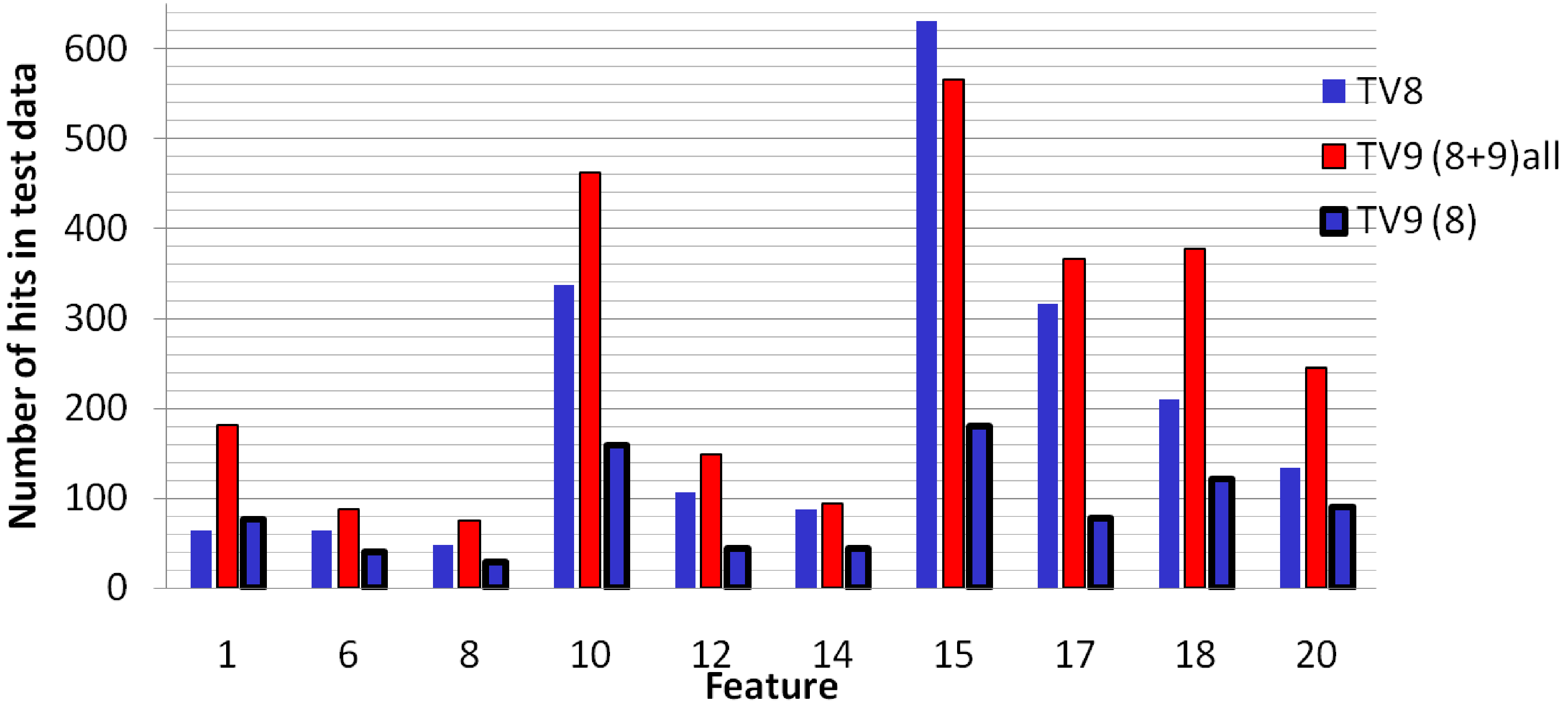
Cityscape*

Hand*

Female human face closeup

1%**

TV2008 vs TV2009 hits for common features



TV8 : Hits of tv8 runs on tv8 test data

TV9(8) : Hits of tv9 runs on shared tv8 test data

TV9(8+9) all : Hits of tv9 runs on tv9 test data + tv8 test data

Feature 1 : Classroom Feature 6 : Airplane_flying Feature 8 : Bus Fetaure 10 : Cityscape Feature 12: Telephone

Feature 14 : Demonstration_or_protest Feature 15 : Hand Feature 17: Nighttime Feature 18 : Boat_ship Feature 20 : Hand

Number of runs of each training type

The common data (A) still is the most popular by far

Tr-Type	2009	2008	2007
A	203 **(91.4%)**	152 **(76%)**	146 **(89.5%)**
B*	--	15 **(7.5%)**	7 **(4.3%)**
C	13 **(5.8%)**	22 **(11%)**	6 **(3.7%)**
a	3 **(1.3%)**	9 **(4.5%)**	4 **(2.5%)**
b*	--	0	0
c	3 **(1.3%)**	2 **(1%)**	0
Total runs	222	200	163

S&V-specific training predominates

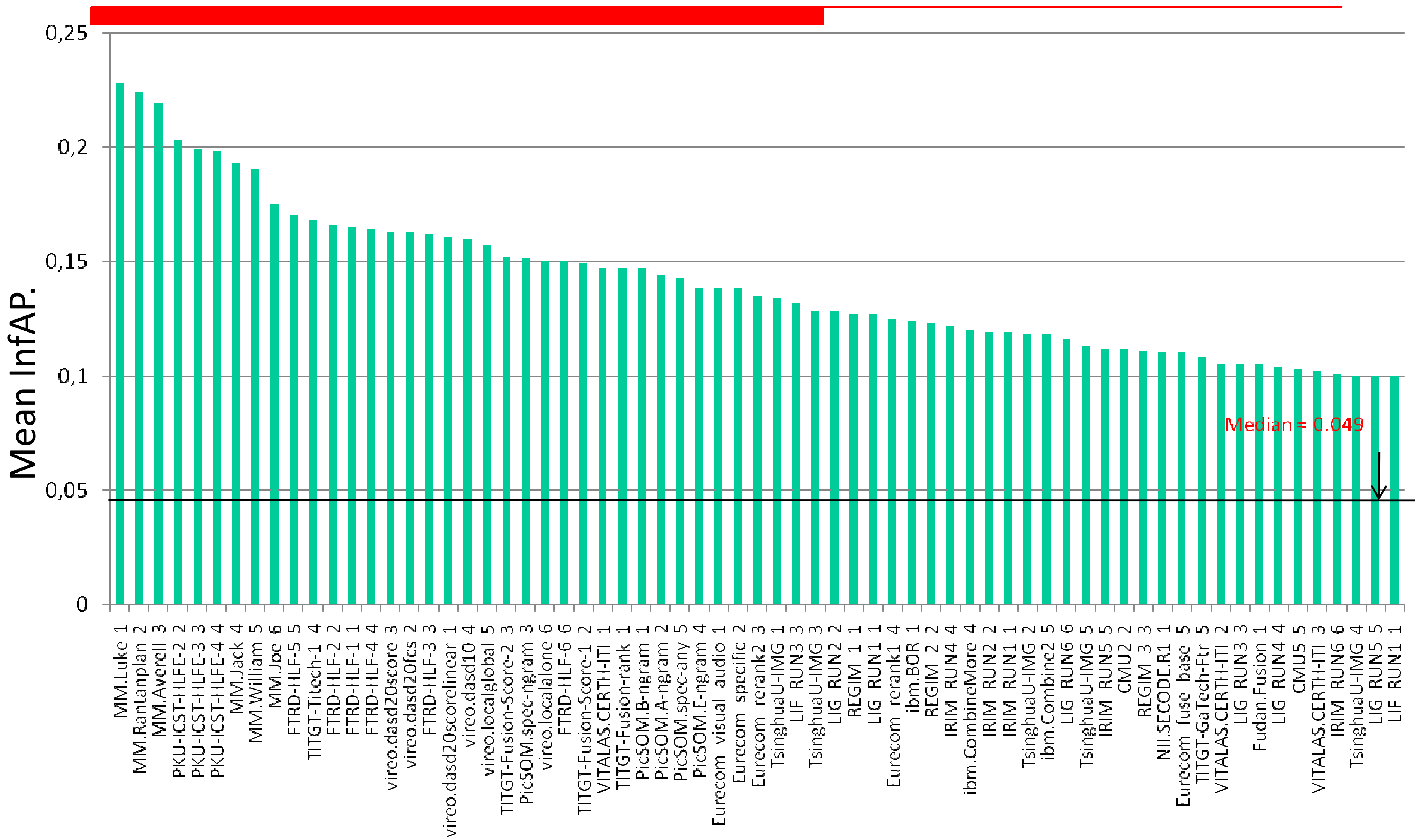
Non-S&V-specific training rare

Any reasons for the rare submissions in non-S&V training categories?

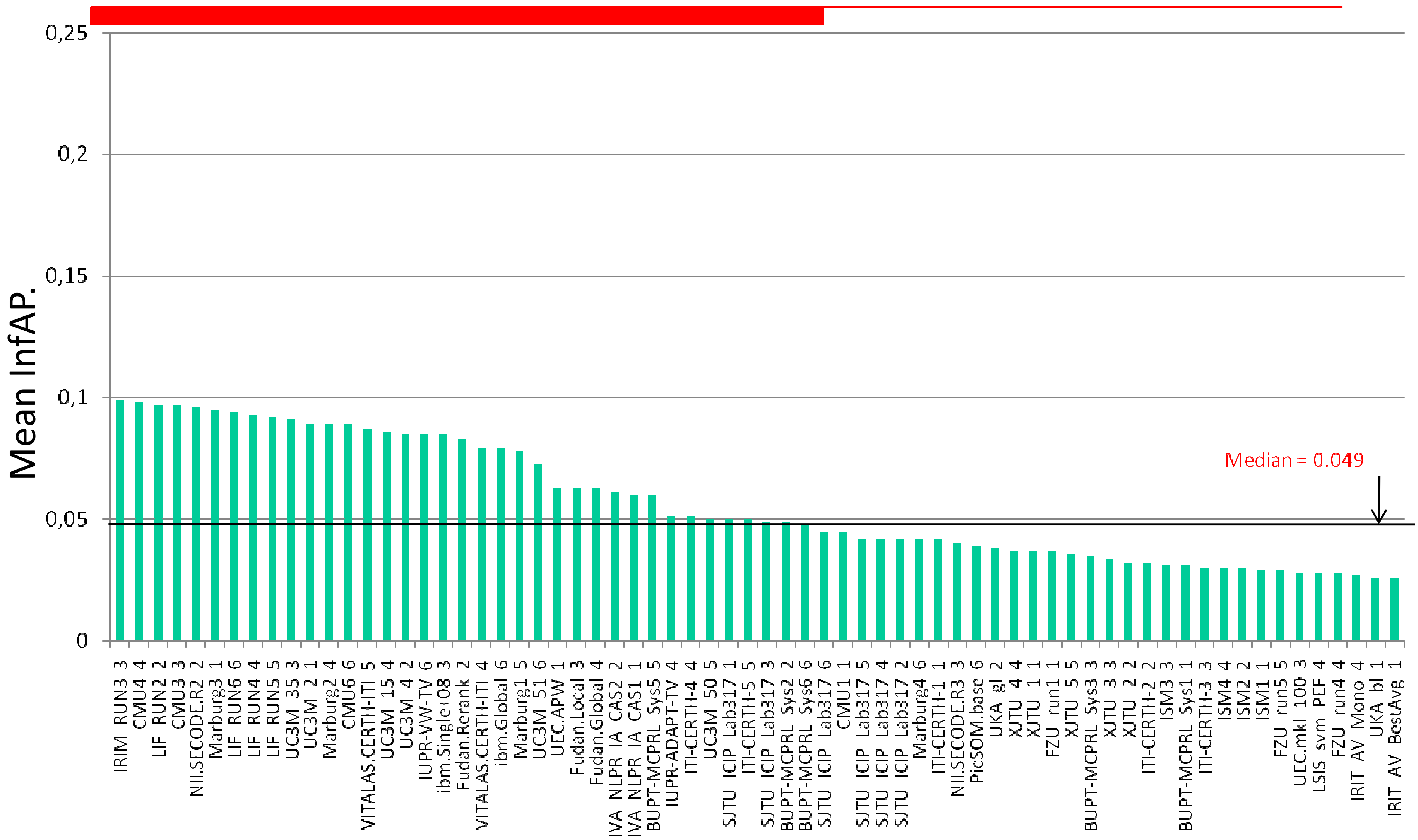
True shots contributed uniquely by team for each feature

Team	Shots	Feature(s)
BRN	2	Doorway
FIU	4	Doorway, Chair
FZU	4	Doorway, Female_face_closeup
IRI	1	Doorway
ISM	3	Traffic_intersection, Cityscape
ITI	3	Person_eating, Chair
LSI	10	Doorway, Chair, Traffic_intersection, Cityscape, Telephone, Nighttime
NHK	5	Doorway, Chair, Traffic_intersection, Hand
NII	8	Doorway, Traffic_intersection, Hand, Boat_ship, Female_face_closeup
SJT	1	Doorway
TIT	2	Traffic_intersection, Cityscape
Tsi	2	Traffic_intersection, Female_face_closeup
UEC	2	Doorway
UKA	1	Hand
VIT	2	Classroom, Traffic_intersection
VPU	1	Doorway
XJT	3	Doorway
ZJU	4	Doorway, Boat_ship
Uza	8	Chair, Traffic_intersection, Doorway, Boat_ship, Telephone, Cityscape

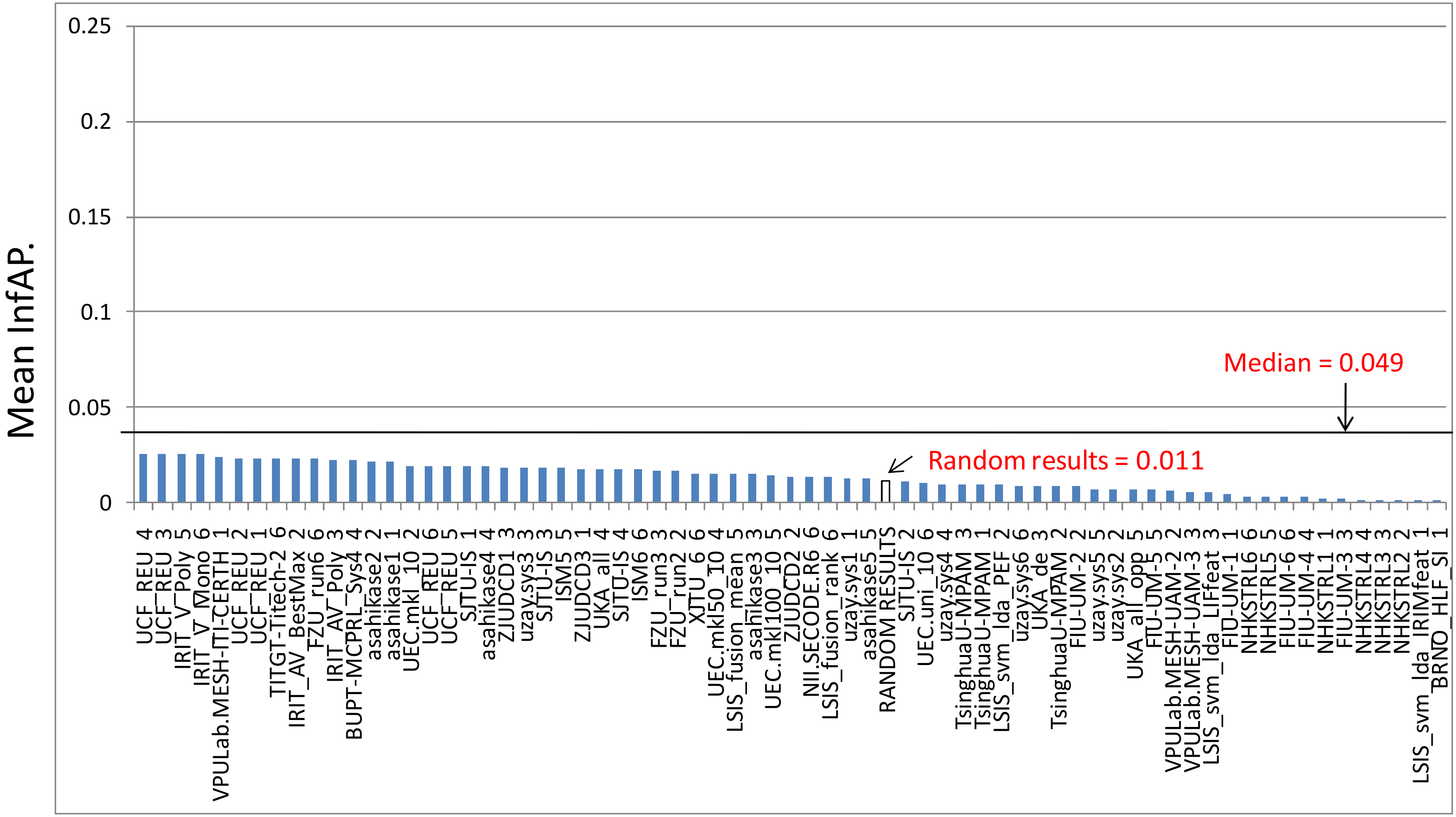
Category A results - Top (1- 67)



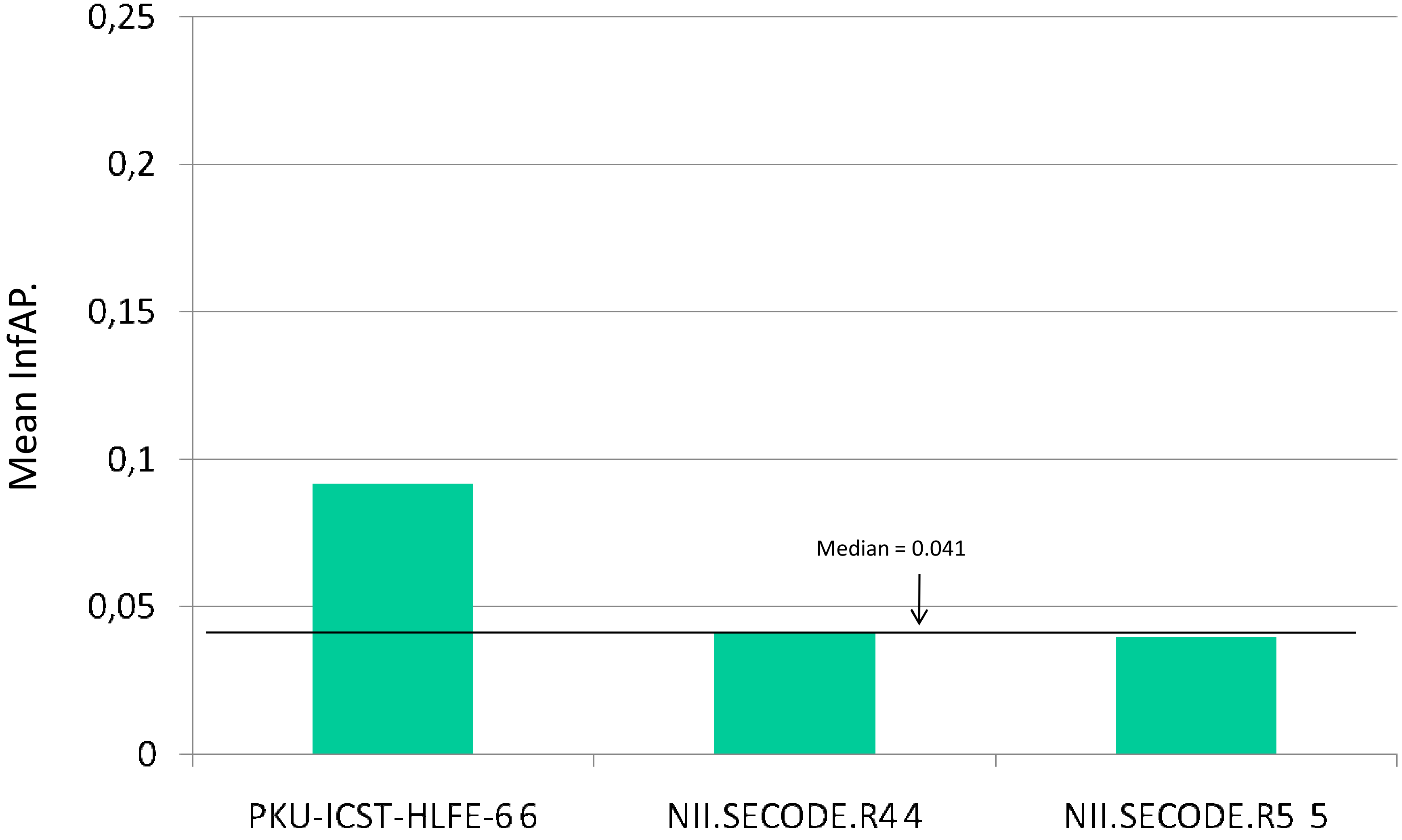
Category A results - Middle (68 - 135)



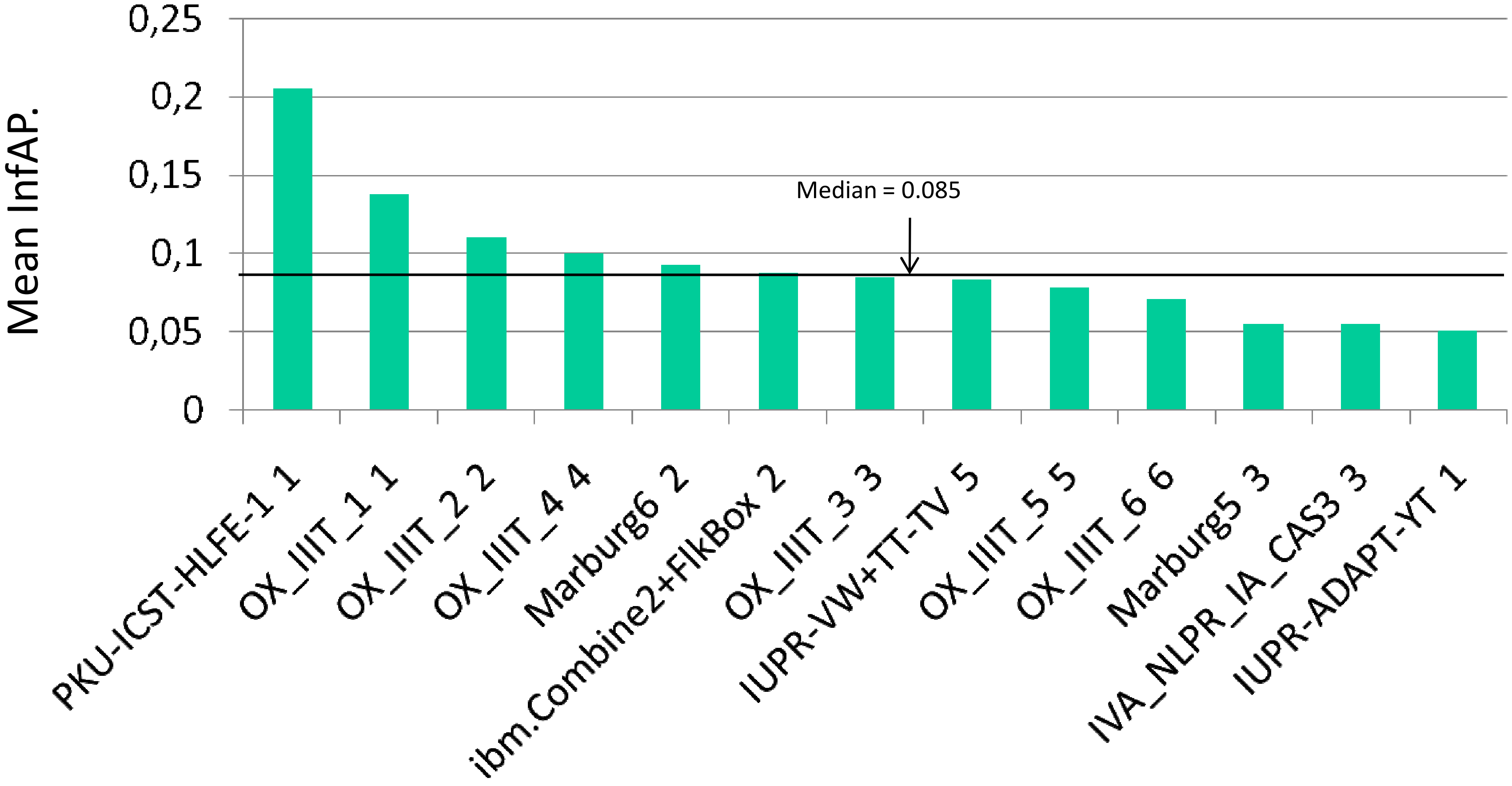
Category A results - Bottom (136-203)



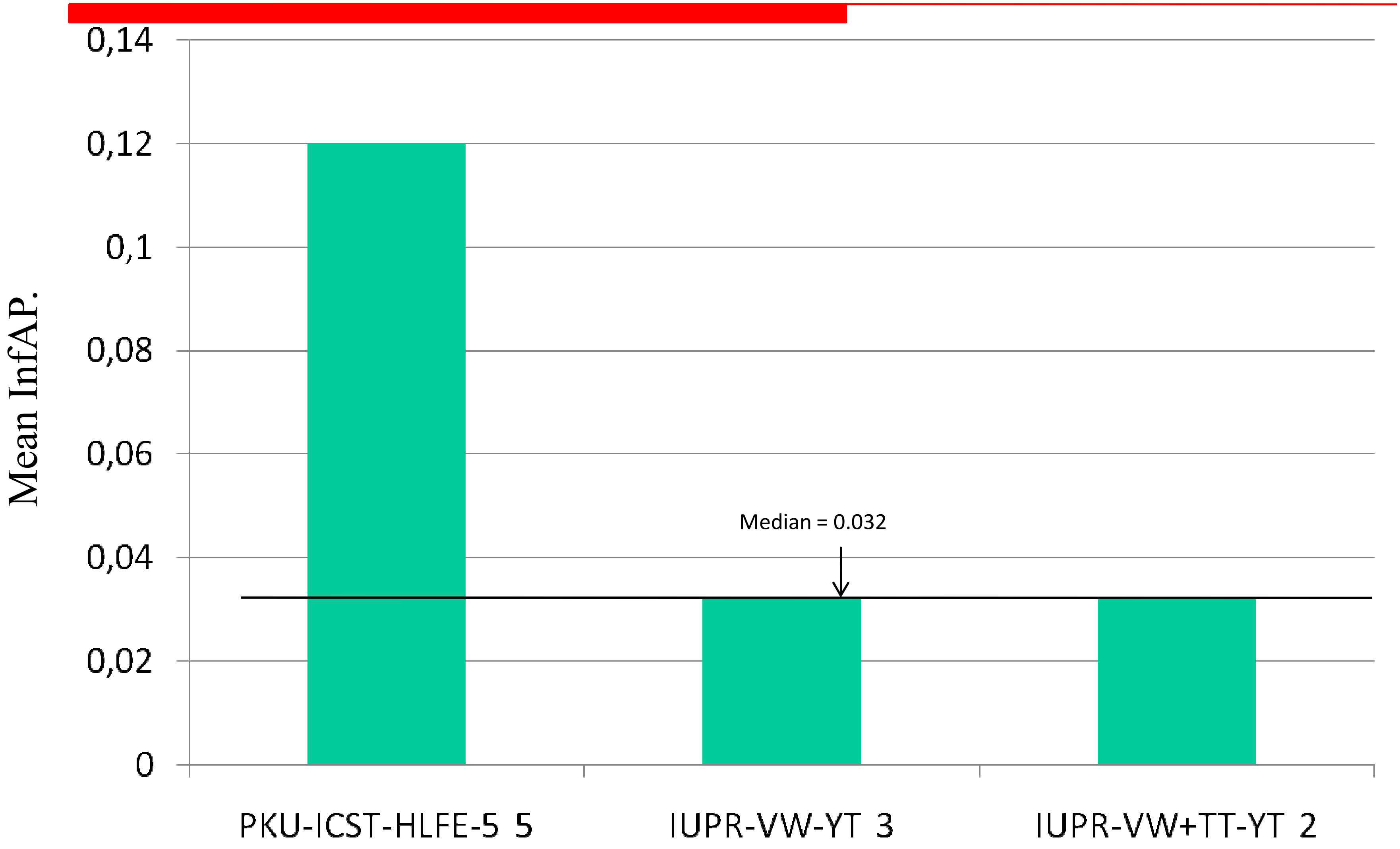
Category a results



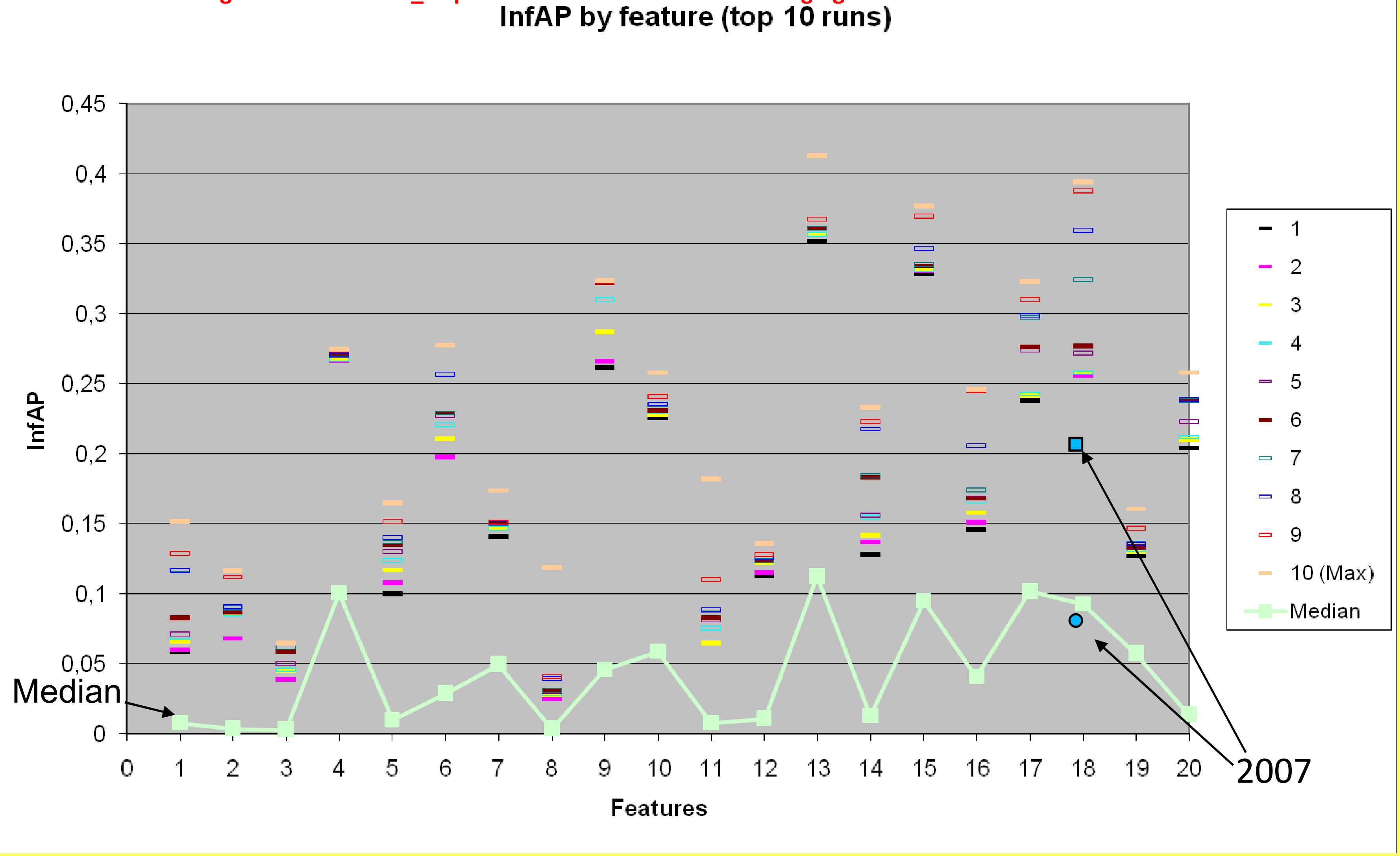
Category C results



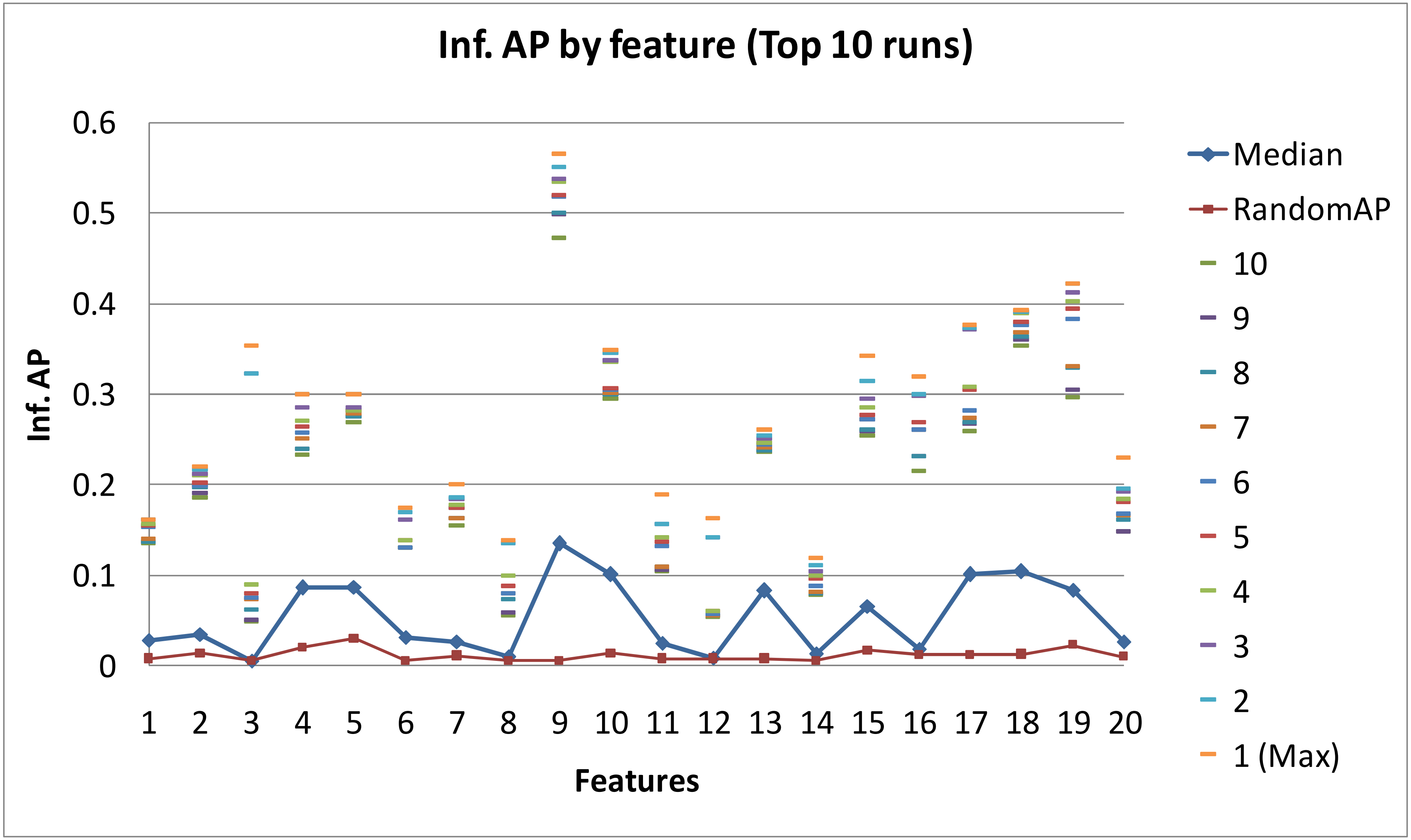
Category c results



1 Classroom 2 Bridge 3 Emergency_Vehicle 4 Dog 5 Kitchen 6 Airplane_flying 7 Two people 8 Bus
9 Driver 10 Cityscape 11 Harbor 12 Telephone 13 Street 14 Demonstration_Or_Protest 15 Hand
16 Mountain 17 Nighttime 18 Boat_ship 19 Flower 20 Singing



TV 2008 results



1 Classroom*
2 Chair
3 Infant
4 Traffic_intersection
5 Doorway
6 Airplane_flying*
7 Person_playing_musical_instrument
8 Bus*
9 Person_playing_soccer
10 Cityscape*
11 Person_riding_bicycle
12 Telephone*
13 Person_eating
14 Demonstration_Or_Protest *
15 Hand*
16 People_dancing
17 Nighttime*
18 Boat_ship*
19 Female_human_face_closeup
20 Singing*

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

Run name	(mean infAP)
MM.Luke_1	(0.228)
MM.Rantanplan_2	(0.224)
MM.Averell_3	(0.219)
PKU-ICST-HLFE-2_2	(0.203)
PKU-ICST-HLFE-3_3	(0.199)
PKU-ICST-HLFE-4_4	(0.198)
MM.Jack_4	(0.193)
MM.William_5	(0.190)
MM.Joe_6	(0.175)
FTRD-HLF-5_5	(0.170)

- MM.Luke_1
  - PKU-ICST-HLFE-2_2
    - MM.Joe_6
    - FTRD-HLF-5_5
  - PKU-ICST-HLFE-3_3
    - FTRD-HLF-5_5
  - PKU-ICST-HLFE-4_4
    - FTRD-HLF-5_5
  - MM.William_5
    - FTRD-HLF-5_5
  - MM.Jack_4
    - FTRD-HLF-5_5

- MM.Rantanplan_2
  - PKU-ICST-HLFE-2_2
    - MM.Joe_6
    - FTRD-HLF-5_5
  - PKU-ICST-HLFE-3_3
    - FTRD-HLF-5_5
  - PKU-ICST-HLFE-4_4
    - FTRD-HLF-5_5
  - MM.William_5
    - FTRD-HLF-5_5
  - MM.Jack_4
    - FTRD-HLF-5_5

- MM.Averell_3
  - MM.Joe_6
  - MM.William_5
    - FTRD-HLF-5_5
  - MM.Jack_4
    - FTRD-HLF-5_5

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

Run name (mean infAP)

PKU-ICST-HLFE-6_6 (0.092)
NII.SECODE.R4_4 (0.041)
NII.SECODE.R5_5 (0.040)

- PKU-ICST-HLFE-6_6
  - NII.SECODE.R4_4
  - NII.SECODE.R5_5

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

Run name (mean infAP)

PKU-ICST-HLFE-1_1 (0.205)
OX_IIIT_1_1 (0.138)
OX_IIIT_2_2 (0.110)
OX_IIIT_4_4 (0.100)
Marburg6_2 (0.093)
ibm.Combine2+FlkBox_2 (0.088)
OX_IIIT_3_3 (0.085)
IUPR-VW+TT-TV_5 (0.083)
OX_IIIT_5_5 0.078)
OX_IIIT_6_6 (0.071)

- PKU-ICST-HLFE-1_1
  - OX_IIIT_1_1
    - OX_IIIT_2_2
      - IUPR-VW+TT-TV_5
      - OX_IIIT_3_3
      - OX_IIIT_6_6
    - OX_IIIT_4_4
      - OX_IIIT_6_6
    - Marburg6_2

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

Run name (mean infAP)
PKU-ICST-HLFE-5_5 (0.120)
IUPR-VW-YT_3 (0.032)
IUPR-VW+TT-YT_2 (0.032)

- PKU-ICST-HLFE-5_5
  - IUPR-VW-YT_3
  - IUPR-VW+TT-YT_2

Significant differences among A/a category runs by group (using randomization test, p < 0.05)

Run name (mean infAP)

A_PKU-ICST-HLFE-2_2 (0.203)

A_PKU-ICST-HLFE-3_3 (0.199)

A_PKU-ICST-HLFE-4_4 (0.198)

a_PKU-ICST-HLFE-6_6 (0.092)

A_NII.SECODE.R1_1 (0.110)

A_NII.SECODE.R2_2 (0.096)

A_NII.SECODE.R3_3 (0.040)

A_NII.SECODE.R6_6 (0.013)

a_NII.SECODE.R4_4   (0.041)

a_NII.SECODE.R5_5 (0.040)

- A_PKU-ICST-HLFE-2_2
  - a_PKU-ICST-HLFE-6_6
- A_PKU-ICST-HLFE-3_3
  - a_PKU-ICST-HLFE-6_6
- A_PKU-ICST-HLFE-4_4
  - a_PKU-ICST-HLFE-6_6

- A_NII.SECODE.R1_1
  - A_NII.SECODE.R2_2
    - A_NII.SECODE.R3_3
      - A_NII.SECODE.R6_6
    - a_NII.SECODE.R4_4
      - A_NII.SECODE.R6_6
    - a_NII.SECODE.R5_5
      - A_NII.SECODE.R6_6

A/a: Influence of S&V specific training data

Significant differences among C/c category runs by group (using randomization test, $p < 0.05$)

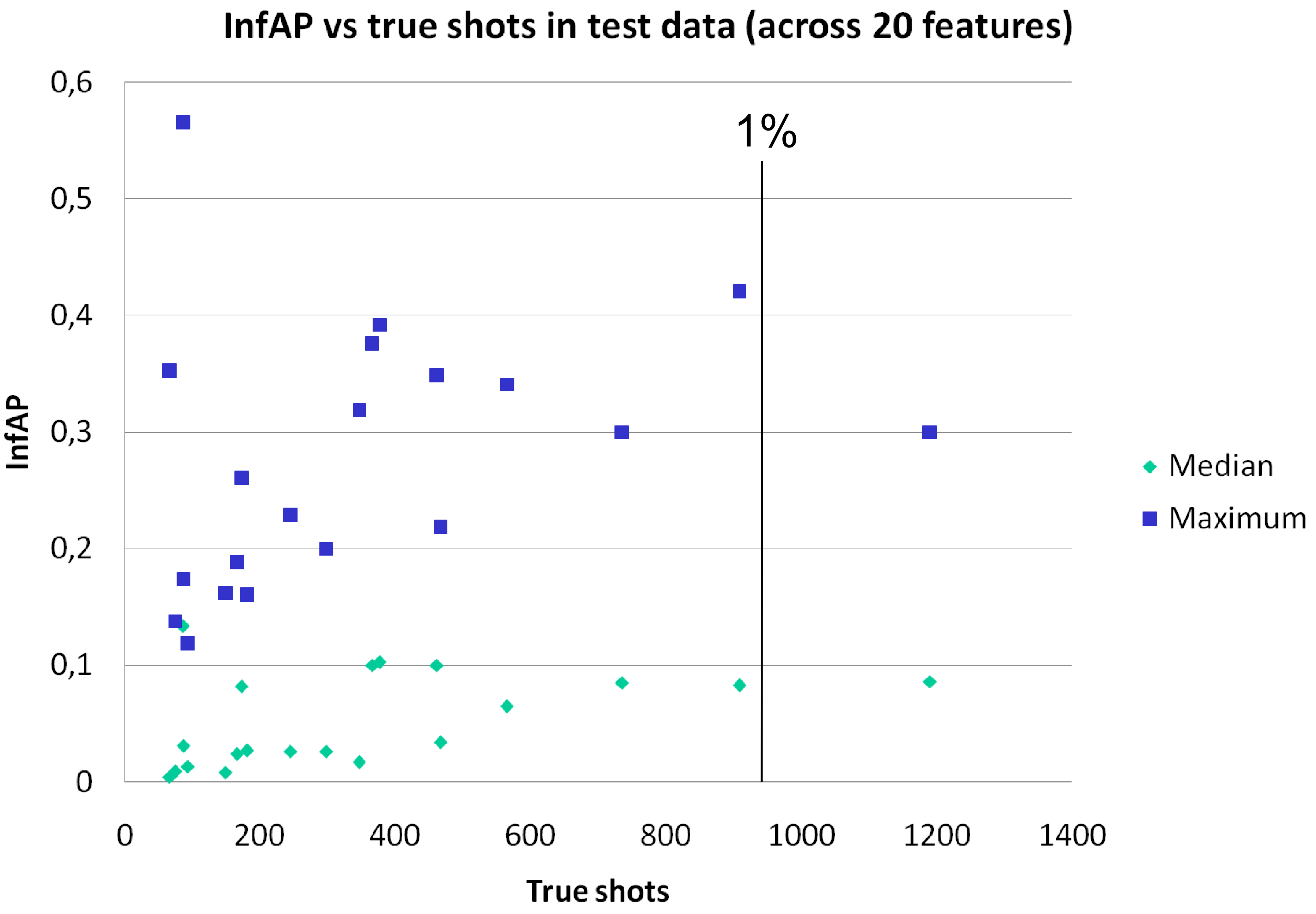
Run name	(mean infAP)
C_IUPR-ADAPT-YT_1	(0.051)
C_IUPR-VW+TT-TV_5	(0.083)
c_IUPR-VW+TT-YT_2	(0.032)
c_IUPR-VW-YT_3	(0.032)

- C_IUPR-VW+TT-TV_5
  - C_IUPR-ADAPT-YT_1
    - c_IUPR-VW+TT-YT_2
    - c_IUPR-VW-YT_3

C_PKU-ICST-HLFE-1_1	(0.205)
c_PKU-ICST-HLFE-5_5	(0.120)

- C_PKU-ICST-HLFE-1_1
  - c_PKU-ICST-HLFE-5_5

C/c: Influence of S&V specific training data (but including other)


InfAP vs true shots in test data (across 20 features)
0,6
0,5
0,4
0,3
0,2
0,1
0
InfAP
0
200
400
600
800
1000
1200
1400
True shots
1%
◆ Median
■ Maximum

Observations

- 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)

Questions to participants:

- How do we know whether the community as a whole achieves better results over the years?
  - Did any run their TV2008 system on TV2009 test data?
  - Did any run their system on tv2008 common 10 features?
- Did anyone use non-speech audio training data? (person_playing_musical_instrument, singing).
- Maybe the a and c categories should be retired?
- Should we also look at detector training and testing speed?