TRECVID-2012 Semantic Indexing task: Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
Dakota Consulting, Inc

also with Franck Thollard, 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 three types of runs:
  - Full run Includes results for 346 concepts, from which NIST and Quaero evaluated 46.
  - Lite run Includes results for 50 concepts, subset of the above 346, 15 evaluated.
  - Pair run Includes results for 10 concept pairs, all evaluated. *NEW*
- TRECVID 2012 SIN video data
  - Test set (IACC.1.C): 200 hrs, with durations between 10 seconds and 3.5 minutes.
  - Development set (IACC.1.A, IACC.1.B & IACC.1.tv10.training): 600 hrs, with durations between 10 seconds to just longer than 3.5 minutes.
  - Total shots: (Much more than in previous TRECVID years, no composite shots)
    - Development: 403,800
    - Test: 145,634
- Common annotation for 346 concepts coordinated by LIG/LIF/Quaero

Semantic Indexing task (2)

- Selection of the 346 target concepts
  - Include all the TRECVID “high level features” from 2005 to 2010 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 relations provided:
  - 427 “implies” relations, e.g. “Actor implies Person”
  - 559 “excludes” relations, e.g. “Daytime_Outdoor excludes Nighttime”

Semantic Indexing task (3)

- NIST evaluated 20 concepts + 5 concept pairs and Quaero evaluated 26 concepts + 5 concept pairs.

- Six 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
  - E – used only training data collected automatically using only the concepts' name and definition *NEW*
  - F – used only training data collected automatically using a query built manually from the concepts' name and definition *NEW*

Datasets comparison

	TV2007	TV2008 = TV2007 + New	TV2009 = TV2008 + New	TV2010	TV2011 = TV2010 + New	TV2012 = TV2011 + New
Dataset length (hours)	~100	~200	~380	~400	~600	~800
Master shots	36,262	72,028	133,412	266,473	403,800	549,434
Unique program titles	47	77	184	N/A	N/A	N/A

Number of runs for each training type

REGULAR FULL RUNS (51 runs)	A	B	C	D	E	F
Only IACC data	47					
Only non-IACC data		0				
Both IACC and non-IACC TRECVID data			0			
Both IACC and non-IACC non-TRECVID data				3		
used only training data collected automatically using only the concepts’ name and definition					0	
used only training data collected automatically using a query built manually from concepts’ name and definition						1
LIGHT RUNS (91 runs)	**A**	**B**	**C**	**D**	**E**	**F**
Only IACC data	83					
Only non-IACC data		0				
Both IACC and non-IACC TRECVID data			0			
Both IACC and non-IACC non-TRECVID data				4		
used only training data collected automatically using only the concepts’ name and definition					1	
used only training data collected automatically using a query built manually from concepts’ name and definition						3

Number of runs for each training type

PAIR RUNS (16 runs)	A	B	C	D	E	F
Only IACC data	14					
Only non-IACC data		0				
Both IACC and non-IACC TRECVID data			0			
Both IACC and non-IACC non-TRECVID data				1		
used only training data collected automatically using only the concepts' name and definition					0	
used only training data collected automatically using a query built manually from concepts' name and definition						1
Total Runs (107)	97	0	0	5	1	4
	90%			5%	1%	4%

56 concepts evaluated

3 Airplane
4 Airplane_Flying
9 Basketball
13 Bicycling
15 Boat_Ship
16 Boy
17 Bridges
25 Chair
31 Computers
51 Female_Person*
54 Girl
56 Government_Leader
57 Greeting
63 Highway
71 Instrumental_Musician

72 Kitchen
74 Landscape
75 Male_Person*
77 Meeting
80 Motorcycle
84 Nighttime*
85 Office
95 Press_Conference
99 Roadway_Junction
101 Scene_Text*
105 Singing*
107 Sitting_down*
112 Stadium
116 Teenagers
120 Throwing

128 Walking_Running*
155 Apartments
163 Baby
198 Civilian_Person
199 Clearing
254 Fields
267 Forest
274 George_Bush
276 Glasses
297 Hill
321 Lakes
338 Man_Wearing_A_Suit
342 Military_Airplane
359 Oceans
434 Skier
440 Soldiers

901 Beach + Mountain
902 Old_people + Flags
903 Animal + Snow
904 Bird + Waterscape_waterfront
905 Dog + Indoor
906 Driver + Female_Human_face
907 Person + underwater
908 Table + Telephone
909 Two_People + Vegetation
910 Car + Bicycle

-The 7 marked with "*" are a subset of those tested in 2011

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:
  - 46 feature results for full runs
  - 15 feature results for lite runs
  - 10 feature results for concept-pairs 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.

2012: mean extended Inferred average precision (xinfAP)

- 3 pools were created for each concept and sampled as:
  - Top pool (ranks 1-200) sampled at 100%
  - Bottom pool (ranks 201-2000) sampled at 10%

56 concepts
282949 total judgments
35361 total hits
17739 Hits at ranks (1-100)
9783 Hits at ranks (101-200)
7839 Hits at ranks (201-2000)

- 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

2012 : 25 Finishers

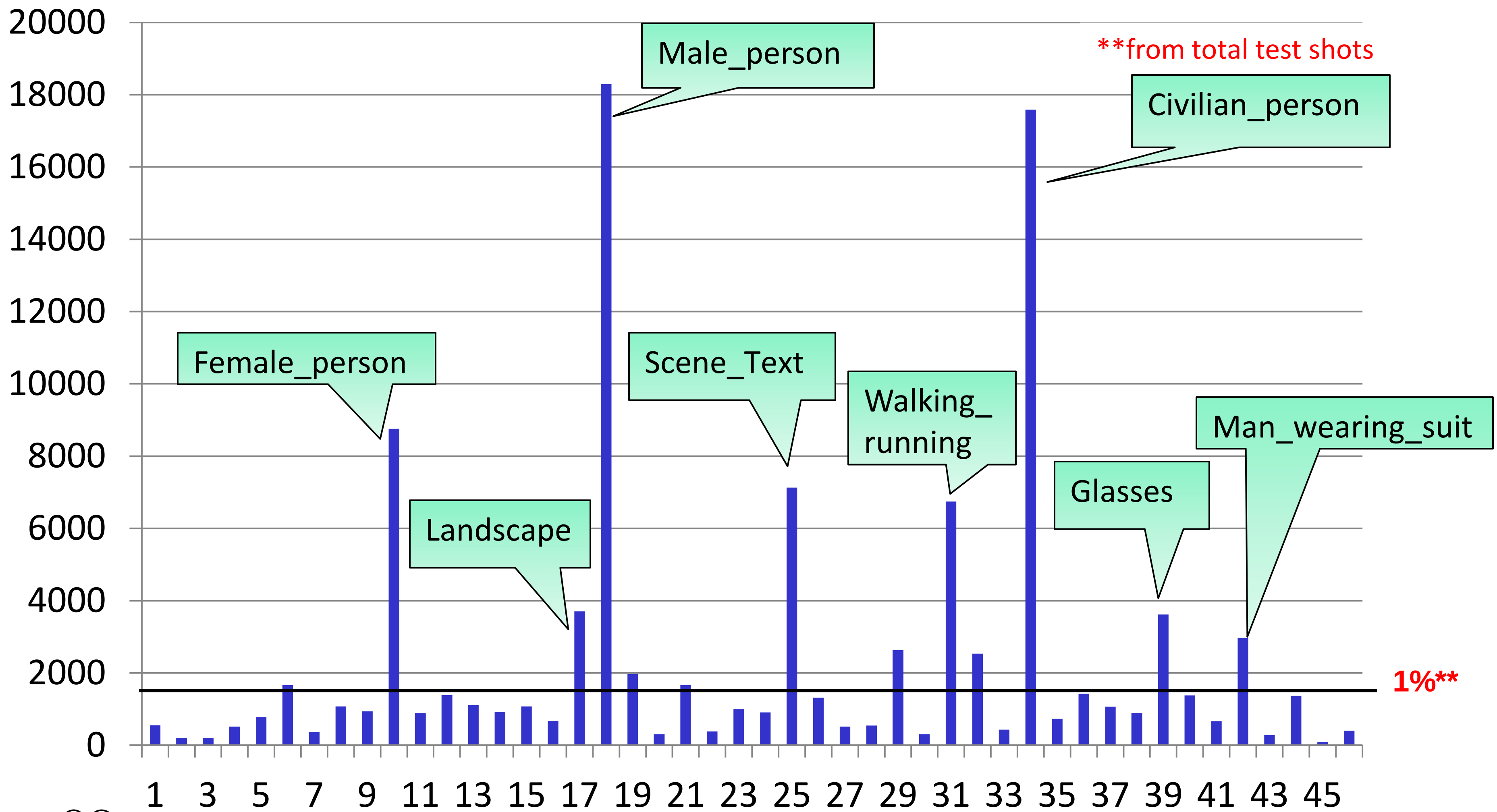
PicSOM	Aalto U.
INF	Carnegie Mellon U.
CEALIST	CEA
VIREO	City U. of Hong Kong
ECL_Liris	Ecole Centrale de Lyon, Universit de Lyon
EURECOM	EURECOM - Multimedia Communications
VideoSense	EURECOM VideoSense Consortium
FIU_UM	Florida International U. U. of Miami
FTRDBJ	France Telecom Orange Labs (Beijing)
kobe_muroran	Kobe U., Muroran Institute of Technology
IBM	IBM T. J. Watson Research Center
ITI_CERTH	Informatics and Telematics Institute (Centre for Research and Technology)
Quaero	INRIA, IRIT, LIG, U. Karlsruhe
ECNU	Institute of Computer Applications, East China Normal U.
JRS.VUT	JOANNEUM RESEARCH Forschungsgesellschaft mbH Vienna U. of Technology
IRIM	IRIM - Indexation et Recherche d'Information MultimÃ©dia GDR-ISIS
NII	National Institute of Informatics
NHKSTRL	NHK Science and Technical Research Laboratories
ntt	NTT Cyber Space Laboratories School of Software, Dalian U. of Technology
IRC_Fuzhou	School of Mathematics and Computer Science Fuzhou U.
stanford	Stanford U.
TokyoTechCanon	Tokyo Institute of Technology and Canon
MediaMill	U. of Amsterdam
UEC	U. of Electro-Communications
GIM	U. of Extremadura

2012 : 25/52 Finishers

Participation and finishing declined! Why?

	Task finishers	Participants
2012	25	52
2011	28	56
2010	39	69
2009	42	70
2008	43	64
2007	32	54
2006	30	54
2005	22	42
2004	12	33

Frequency of hits varies by feature



2011 common features

10	18	21	25	26	27	31	
Female_person	Male_person	Night_time	Scene_Text	Singing	Sitting_down	Walking_running	

True shots contributed uniquely by team

Full runs

Team	No. of Shots	Team	No. of shots
CEA	664	Qu	10
VIR	469		
FIU	464		
IBM	427		
UEC	271		
UvA	209		
nii	156		
ITI	127		
NHK	99		
FTR	63		
Tok	146		
Pic	46		
IRI	37		
CMU	16		

Less unique shots compared to TV2011

Lite runs

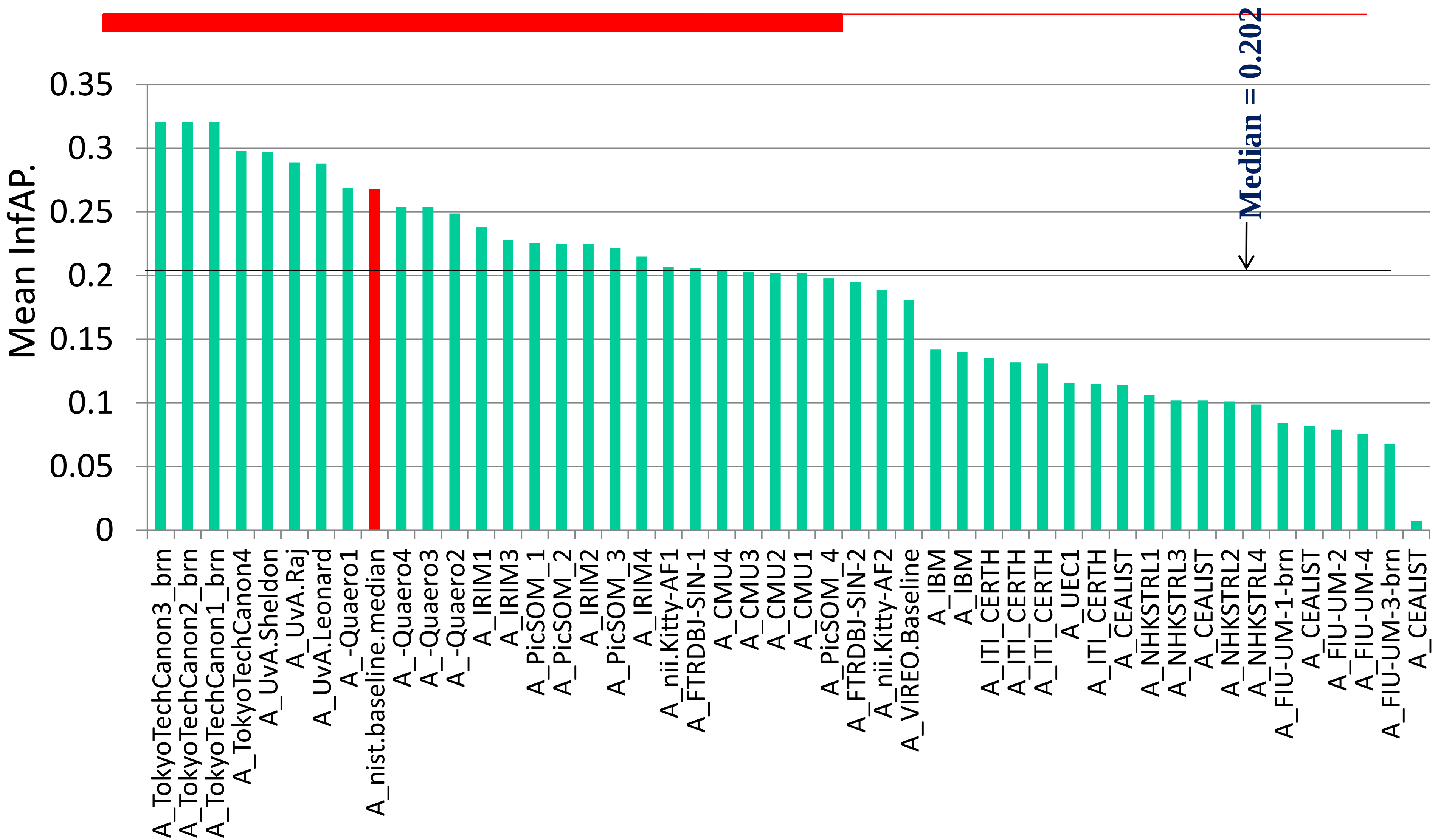
Team	No. of Shots	Team	No. of shots
Fud	363	Vid	32
CEA	218	Kob	31
nii	211	NTT	26
FIU	190	FTR	25
GIM	132	NHK	23
UvA	119	Ecl	20
JRS	103	ITI	18
IBM	95	IRI	10
Tok	79	CMU	5
UEC	71	Pic	4
VIR	71	ECN	2
sta	37		
Eur	33		

Baseline run by NIST

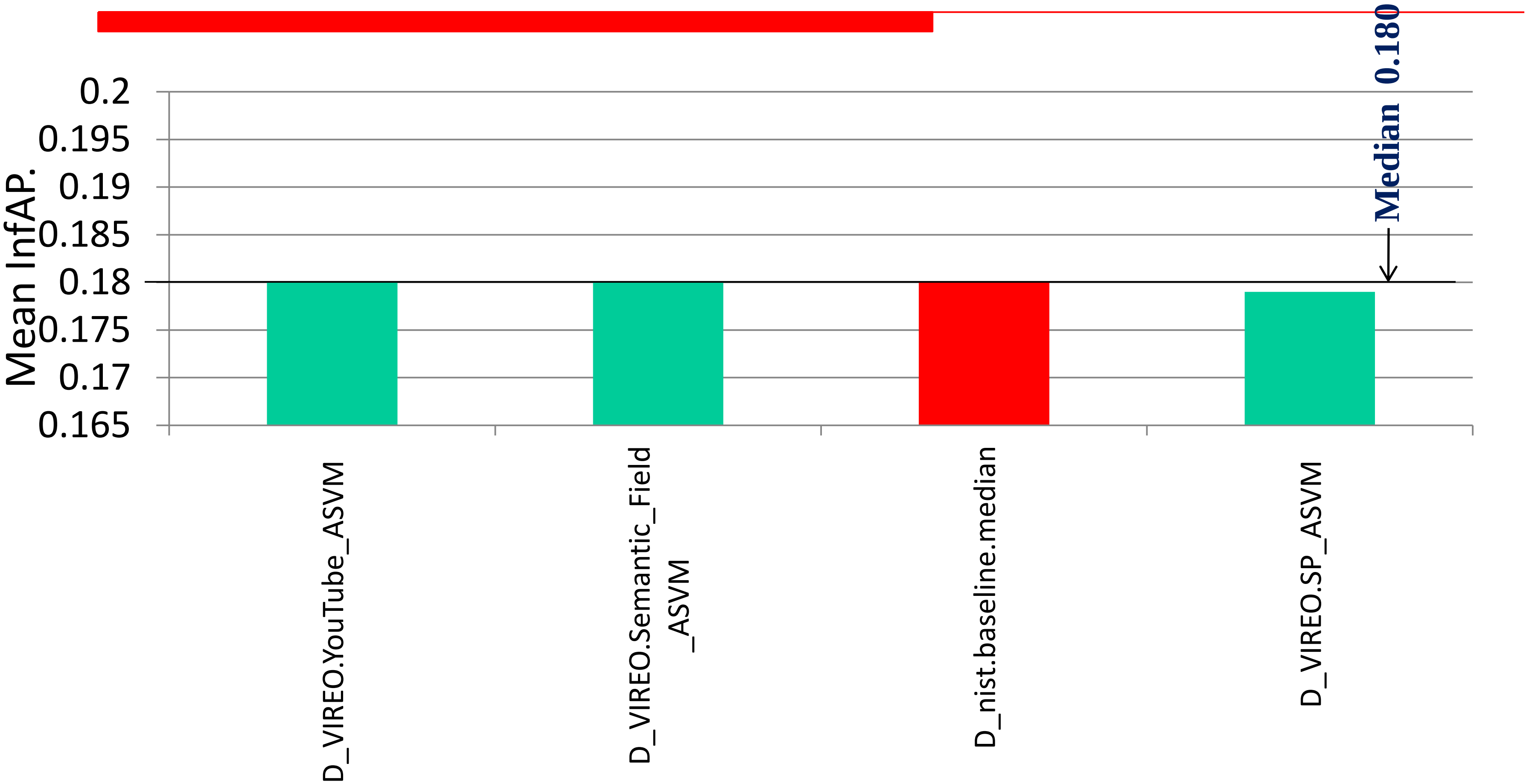
- A median baseline run is created for each run type and training category.
- Basic idea:
  - For each feature, find the median rank of each submitted shot calculated across all submitted runs in that run type and training category.
  - The final shot median rank value is weighted by the ratio of all submitted runs to number of runs that submitted that shot:

$$ShotX_{Median\_rank} = Median\_rank * \frac{TotalNumberOfRuns}{NumberOfRunsSubmittedX}$$

Category A results (Full runs)

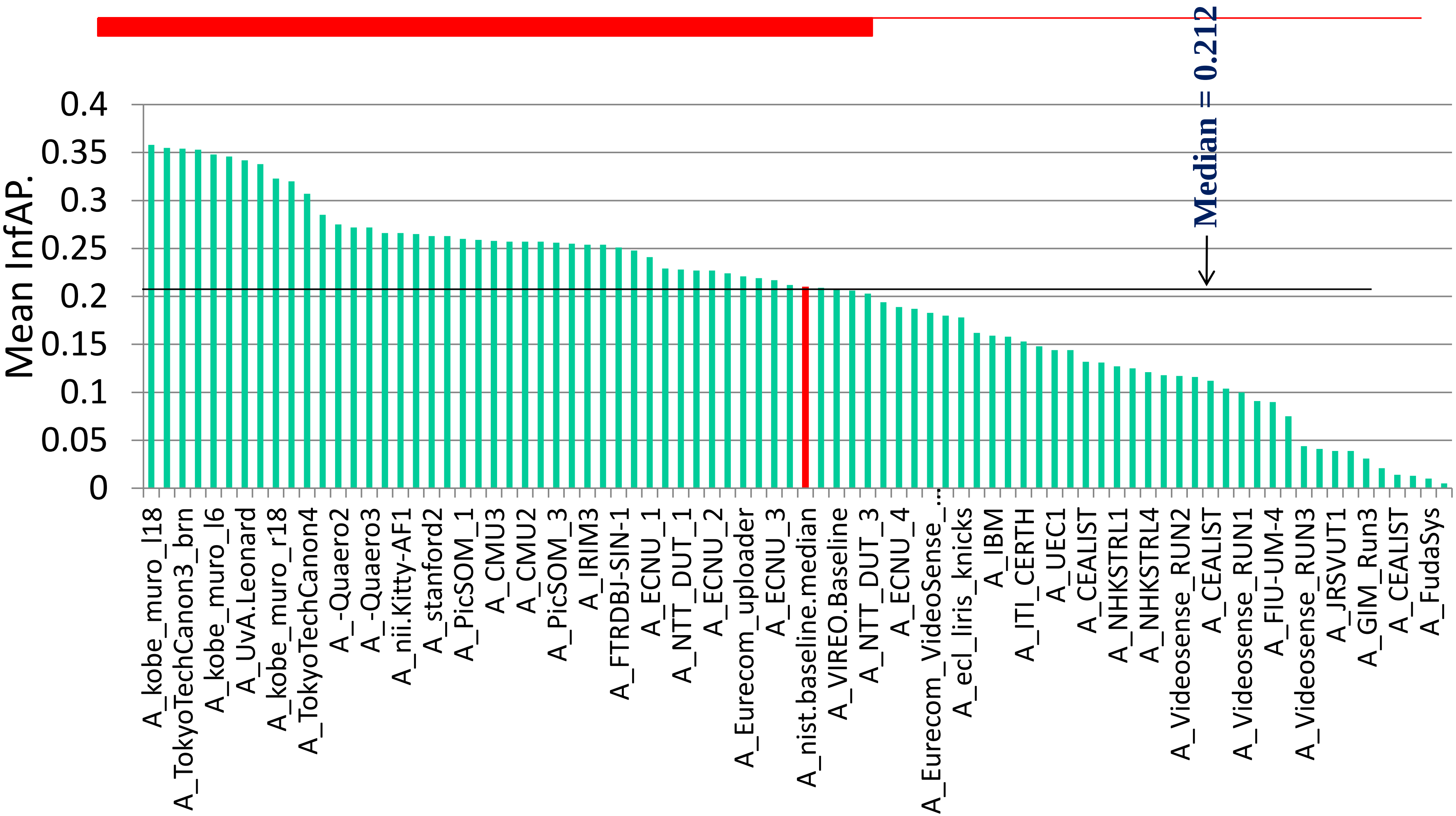


Category D results (Full runs)

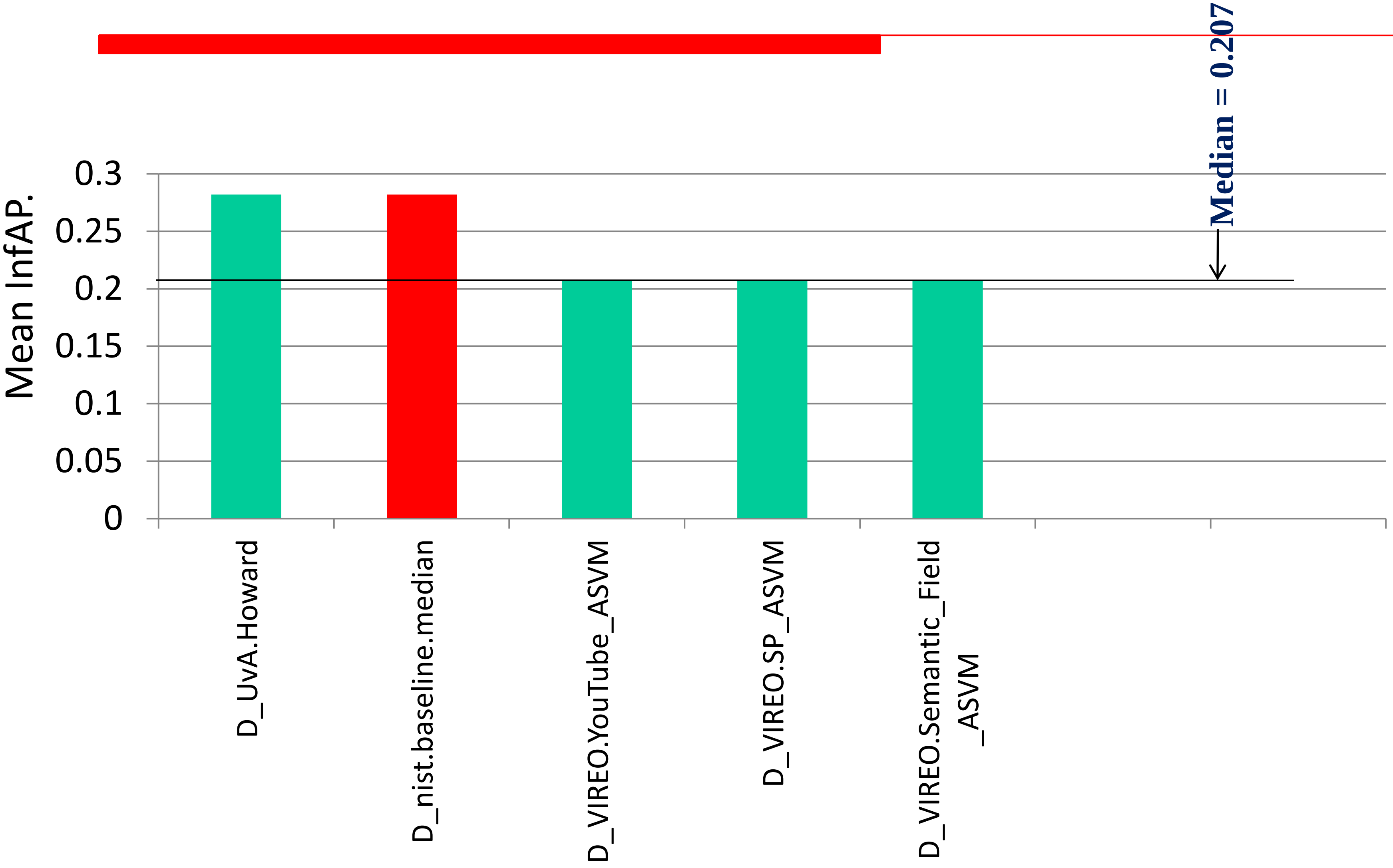


Note: Category F has only 1 run (F_VIREO.Semantic_Pooling) with score = 0.048

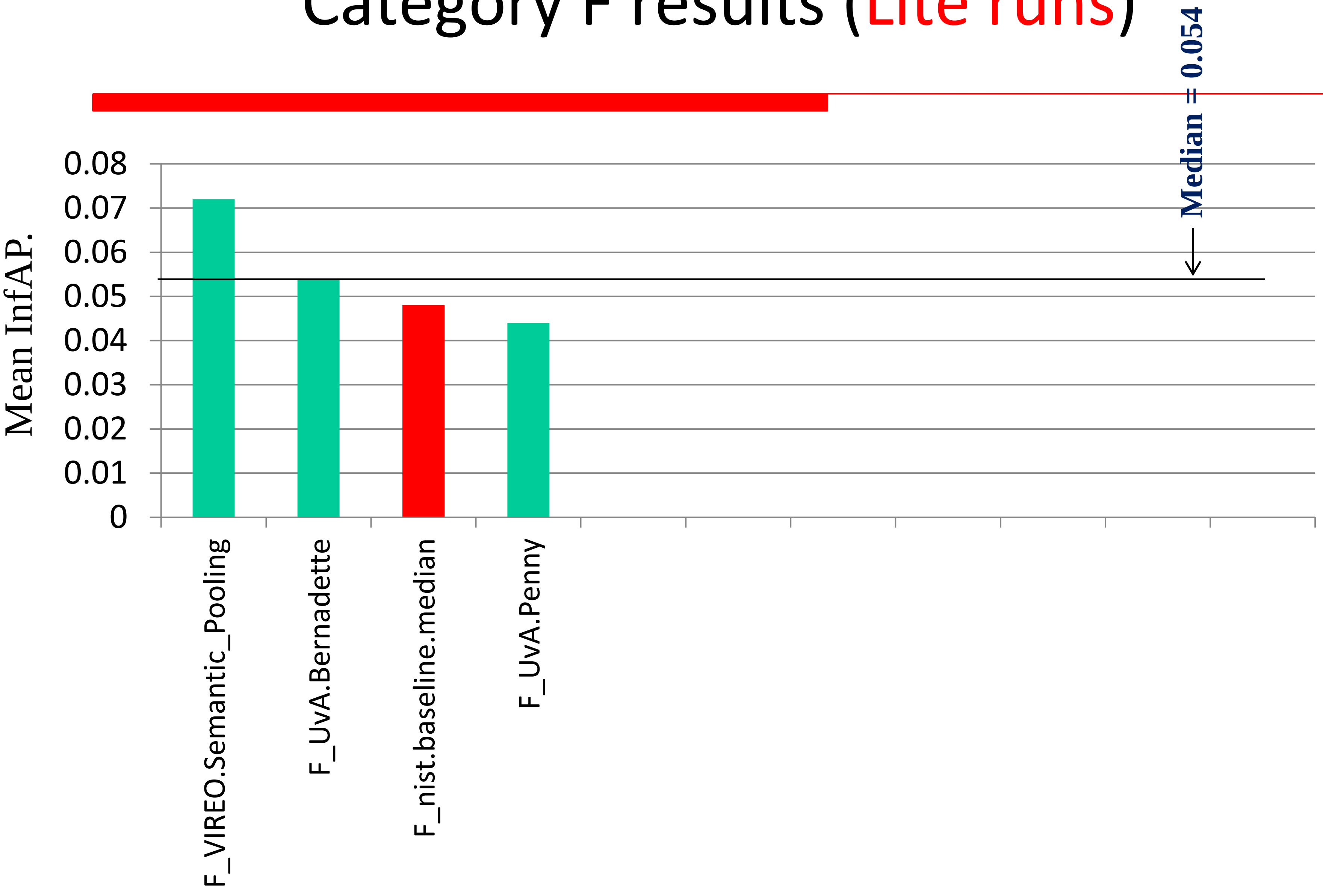
Category A results (Lite runs)



Category D results (Lite runs)

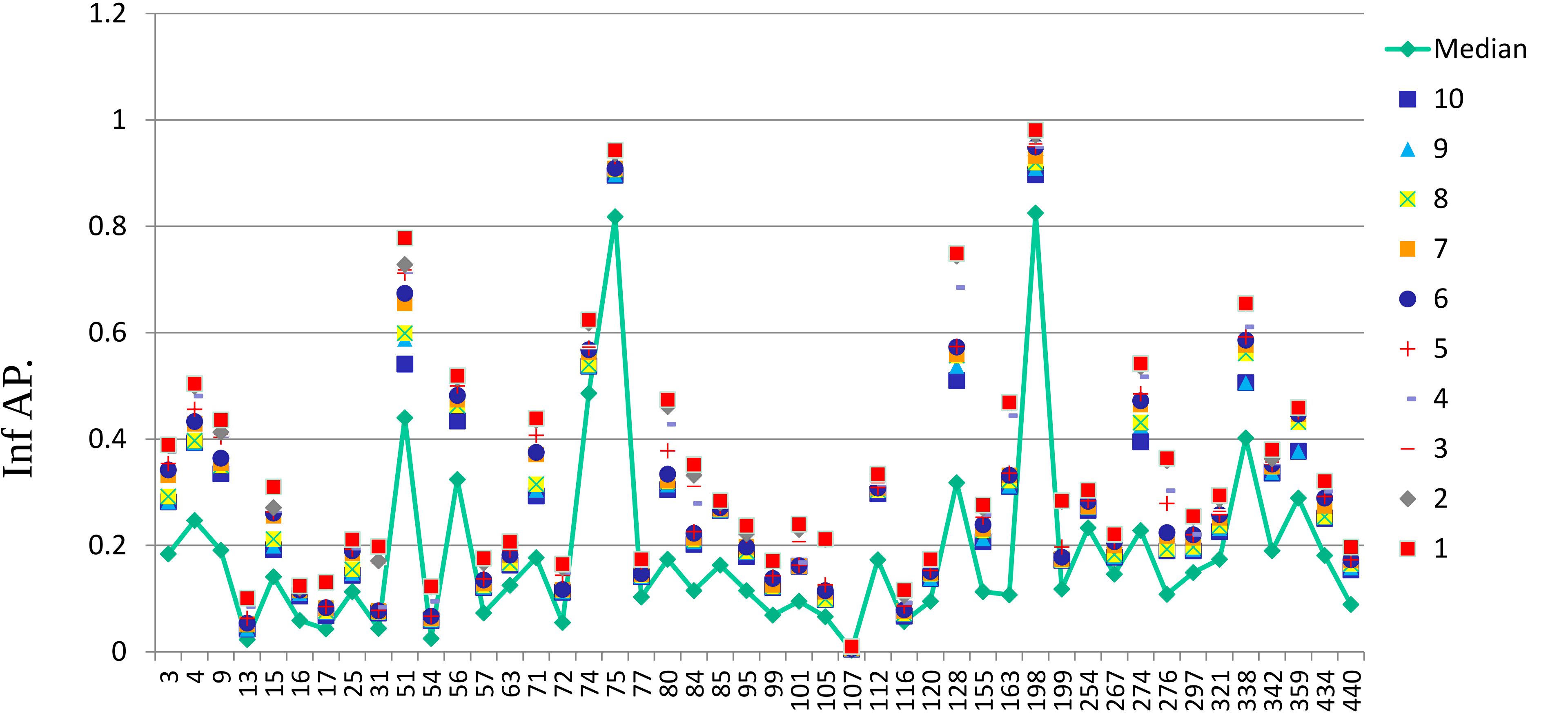


Category F results (Lite runs)



Note: Category E has only 1 run (E_nii.Kitty-EL4) with score = 0.044

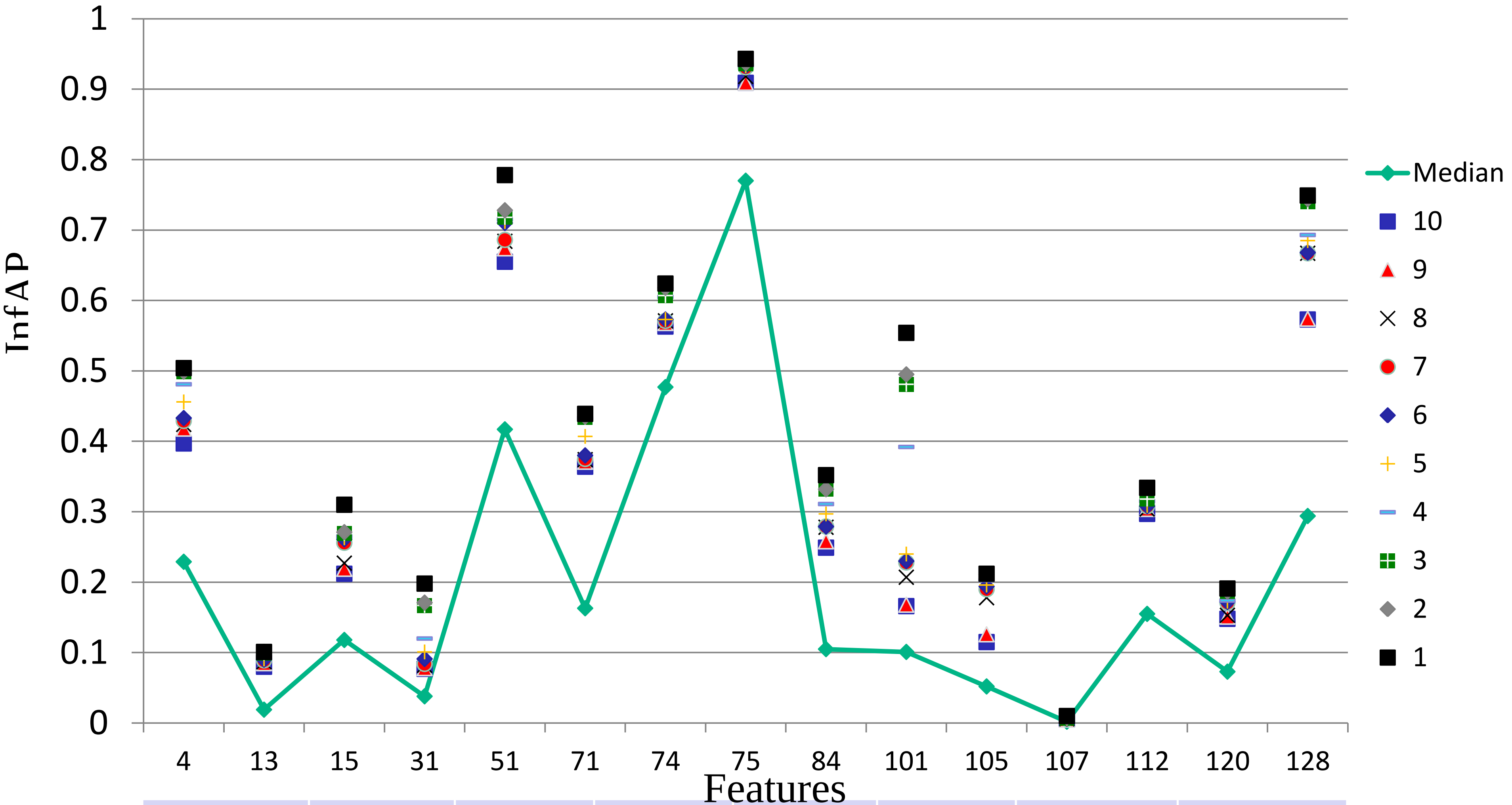
Top 10 InfAP scores by feature (Full runs)



3 Airplane	4 Airplane_flying	9 Basketball	13 Bicycling	15 Boat_ship	16 Boy	17 Bridges	25 Chair	31 Computers	51 Female_person	54 Girl	56 Government_Leader	57 Greeting	63 Highway	71 Instrumental_Musician
72 Kitchen	74 Landscape	75 Male_person	77 Meeting	80 Motorcycle	84 Nighttime	85 Office	95 Press_Conference	99 Roadway_Junction	101 Scene_Text	105 Singing	107 Sitting_down	112 Stadium	116 Teenagers	120 Throwing
128 Walking_Running	155 Apartments	163 Baby	198 Civilian_Person	199 Clearing	254 Fields	267 Forest	274 George_Bush	276 Glasses	297 Hill	321 Lakes	338 Man_Wearing_A_Suit	342 Military_Airplane	359 Oceans	434 Skier
440 Soldiers														

Top 10 InfAP scores for 15 common features

(Lite AND Full runs)



4 Airplane	13 Bicycling	15 Boat_ship	31 Computers	51 Feamale_ Person	71 Instrumental_ Musician	74 Landscape	75 Male_person
112 Stadium	120 Throwing	128 Walking_ Running	84 Nighttime	101 Scene_text	105 Singing	107 Sitting_down	

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

Run name	(mean infAP)
F_A_TokyoTechCanon3_brn_3	0.321
F_A_TokyoTechCanon2_brn_2	0.321
F_A_TokyoTechCanon1_brn_1	0.321
F_A_TokyoTechCanon4_4	0.298
F_A_UvA.Sheldon_1	0.297
F_A_UvA.Raj_2	0.289
F_A_UvA.Leonard_4	0.288
F_A_-Quaero1_1	0.269
F_A_-Quaero4_4	0.254
F_A_-Quaero3_3	0.254

Statistical significant differences among top 10 A-category full runs (using randomization test, p < 0.05) (2)

- A_TokyoTechCanon3_brn_3
  - A_UvA.Raj_2
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Sheldon_1
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Leonard_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_TokyoTechCanon4_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4

- A_TokyoTechCanon2_brn_2
  - A_UvA.Raj_2
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Sheldon_1
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Leonard_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_TokyoTechCanon4_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4

- A_TokyoTechCanon1_brn_1
  - A_UvA.Raj_2
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Sheldon_1
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_UvA.Leonard_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4
  - A_TokyoTechCanon4_4
    - F_A_-Quaero1_1
      - F_A_-Quaero3_3
      - F_A_-Quaero4_4

Statistical significant differences among top 10 D-category full runs (using randomization test, p < 0.05)

Run name	(mean infAP)
F_D_VIREO.Semantic_Field_ASVM_5	0.180
F_D_VIREO.YouTube_ASVM_3	0.180
F_D_VIREO.SP_ASVM_4	0.179

No Significant difference

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

Run name	(mean infAP)
L_A_kobe_muro_l18_3	0.358
L_A_TokyoTechCanon1_brn_1	0.355
L_A_TokyoTechCanon3_brn_3	0.354
L_A_TokyoTechCanon2_brn_2	0.353
L_A_kobe_muro_l6_1	0.348
L_A_UvA.Sheldon_1	0.346
L_A_UvA.Leonard_4	0.342
L_A_UvA.Raj_2	0.338
L_A_kobe_muro_r18_2	0.323
L_A_kobe_muro_l5_4	0.320

- A_kobe_muro_l18_3
  - L_A_kobe_muro_l6_1
    - L_A_kobe_muro_l5_4
  - L_A_kobe_muro_r18_2

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

Run name	(mean infAP)
L_D_UvA.Howard_3	0.282
L_D_VIREO.Semantic_Field_ASVM_5	0.207
L_D_VIREO.SP_ASVM_4	0.207
L_D_VIREO.YouTube_ASVM_3	0.207

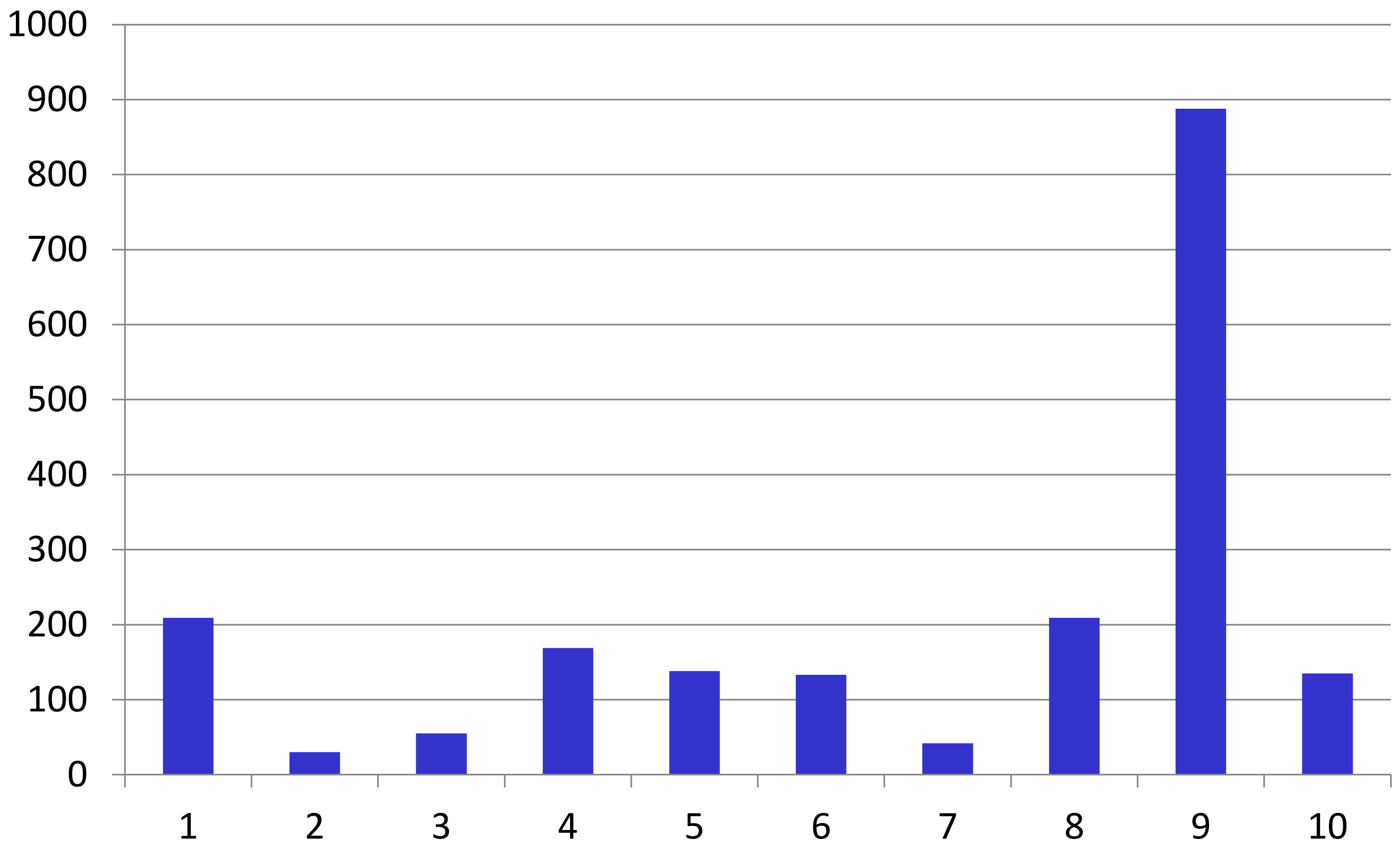
- L_D_UvA.Howard_3
  - L_D_VIREO.Semantic_Field_ASVM_5
  - L_D_VIREO.SP_ASVM_4
  - L_D_VIREO.YouTube_ASVM_3

Statistical significant differences among top F-category lite runs (using randomization test, p < 0.05)

Run name	(mean infAP)
L_F_VIREO.Semantic_Pooling_1	0.072
L_F_UvA.Bernadette_5	0.054
L_F_UvA.Penny_7	0.044

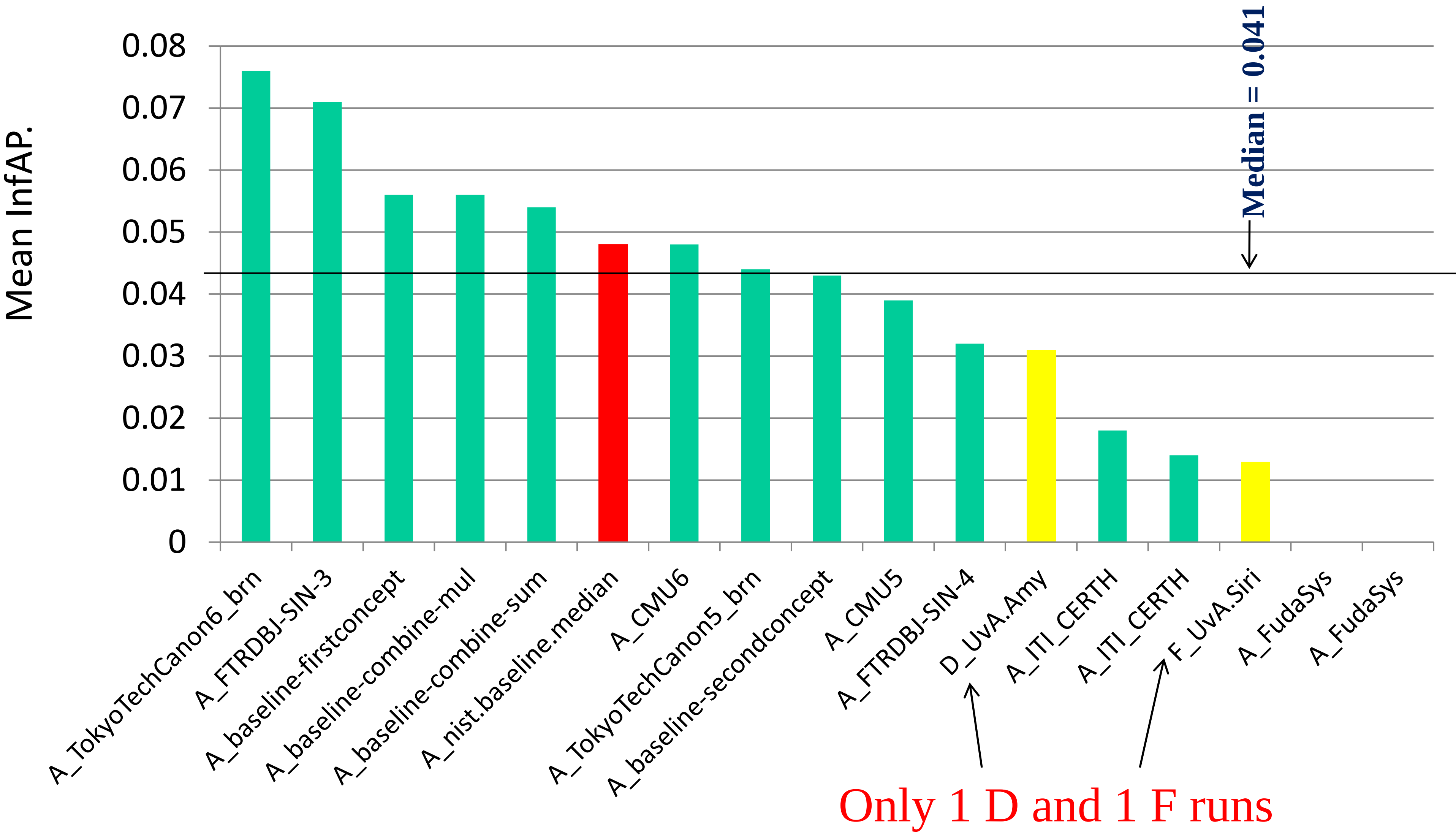
No Significant difference

Frequency of hits for concept pairs



1 Beach + Mountain	2 Old_people + Flags	3 Animal + Snow	4 Bird + Waterscape_ waterfront	5 Dog + Indoor	6 Driver + Female_human_ Face	7 Person + Underwater	8 Table + Telephone	9 Two_people + Vegetation	10 Car + Bicycle

Category A results (Concept Pairs)



Statistical significant differences among top 10 A-category Concept Pairs runs (using randomization test, $p < 0.05$)

Run name	(mean infAP)
P_A_TokyoTechCanon6_brn_6	0.076
P_A_FTRDBJ-SIN-3_3	0.071
P_A_baseline-firstconcept_3	0.056
P_A_baseline-combine-mul_1	0.056
P_A_baseline-combine-sum_2	0.054
P_A_CMU6_1	0.048
P_A_TokyoTechCanon5_brn_5	0.044
P_A_baseline-secondconcept_4	0.043
P_A_CMU5_2	0.039
P_A_FTRDBJ-SIN-4_4	0.032

- A_TokyoTechCanon6_brn_6
  - A_CMU5_2
  - A_FTRDBJ-SIN-4_4
- A_TokyoTechCanon5_brn_5
  - A_FTRDBJ-SIN-4_4
- A_FTRDBJ-SIN-3_3
  - A_baseline-secondconcept_4

Observations

- Site experiments include:
  - focus on robustness, merging many different representations
  - use of spatial pyramids
  - improved bag of word approaches
  - Fisher/super-vectors, VLADs, VLATs
  - sophisticated fusion strategies (IRIM presentation to follow)
  - combination of low and intermediate/high features
  - analysis of more than one keyf rame per shot
  - audio analysis
  - using temporal context information
  - use of metadata (Eurecom presentation to follow)
  - machine learning: automatic evaluation of modeling strategies
  - consideration of scalability issues
- Some participation on the concept pair task (see Mediamill presentation to follow)
- Still no improvement using external training data

Presentations to follow

- 2:10 - 2:30, Eurecom - Multimedia Communications (EURECOM)
- 2:30 - 2:50, IBM Research (IBM)
- 2:50 - 3:10, Kobe University; Muroran Insitute of Technology (kobe_muroran)

- 3:10 - 3:30, Break with refreshments served in the NIST West Square Cafeteria

- 3:30 - 3:50, Indexation et Recherche d'Information Multimédia GDR-ISIS (IRIM)
- 3:50 - 4:20, University of Amsterdam (MediaMill)
- 4:20 - 4:40, Discussion

- 4:50 p.m. NIST bus to Holiday Inn, Gaithersburg

Less participation again

- Poll last year:
  - Task becoming too big?
    - No new increase except for the development set.
  - Not enough novelty?
    - Concept pair and “no annotation”.
  - US Aladdin program / MED task competition?
  - “Too much time was spent on extracting features but more effort should be on developing new frameworks and learning methods”, “Provide more auxiliary information, such as speech recognition results, or others”:
    - IRIM initiative: sharing descriptors, classifier outputs and more (see IRIM’s presentation to follow)
    - too late and too few for 2012 but ready for 2013 and more.
- Maybe the number is hidden in joint participations?

SIN 2013

- Globally keep the task similar and of similar scale
- Further explore the “concept pair” and “no annotation” variants
- Common training data for the “no annotation” variant is likely to be delivered LIG (F type)
- Sharing of data proposed by IRIM
- Possible method for measuring progress over years
- New subtask about concept localization under consideration $\longrightarrow$ annotation issue
- Collaborative annotation available much earlier (end of February)
- Feedback welcome

Sharing of data for TRECVID SIN

- Organized by the IRIM groups of CNRS GRD ISIS.
- IRIM proposes its data sharing organization for the TRECVID SIN task. This comprises:
  - a wiki with read-write access for all
  - a data repository with read access for all and currently a write access only via one of the organizers
  - a small set of simple file formats
  - a (quite) simple directory structure
- Shared data mostly consist in descriptors and classification scores.
- Rewarding principle (same as for other contributions)
  - share and be cited and evaluated
  - use freely and cite

Sharing of data for TRECVID SIN

- Wiki (access with tv12 active participant login/password):
  - http://mrim.imag.fr/trecvid/wiki
  - http://mrim.imag.fr/trecvid/wiki/doku.php?id=sin_2012_task
- Associated data for SIN 2012 (access with IACC collection login/password):
  - http://mrim.imag.fr/trecvid/sin12
- Related actions:
  - Sharing of low-level descriptors by CMU for TRECVID 2003-2004
  - Mediamill challenge (101 concepts) using TRECVID 2005 data
  - Sharing of detection scores by CU-Vireo on TRECVID 2008-2010 data
- Possible extension to other TRECVID tasks, e.g. MED.