# TRECVID-2011 Semantic Indexing task: Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
NIST

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 two types of runs:
  - Full run Includes results for 346 concepts, from which NIST evaluated 20.
  - Lite run Includes results for 50 concepts, subset of the above 346.
- TRECVID 2011 SIN video data
  - Test set (IACC.1.B): 200 hrs, with durations between 10 seconds and 3.5 minutes.
  - Development set (IACC.1.A & 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: 146,788 + 119,685
    - Test: 137,327
- Common annotation for 360 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 116 relations provided:
  - 559 “implies” relations, e.g. “Actor implies Person”
  - 10 “excludes” relations, e.g. “Daytime_Outdoor excludes Nighttime”

# Semantic Indexing task (3)

- NIST evaluated 20 concepts and Quaero evaluated 30 concepts

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

# Number of runs for each training type

| REGULAR FULL RUNS | A | B | C | D |
|---|---|---|---|---|
| Only IACC data | 62 | | | |
| Only non-IACC data | | 2 | | |
| Both IACC and non-IACC TRECVID data | | | 1 | |
| Both IACC and non-IACC non-TRECVID data | | | | 3 |
| **LIGHT RUNS** | **A** | **B** | **C** | **D** |
| Only IACC data | 96 | | | |
| Only non-IACC data | | 2 | | |
| Both IACC and non-IACC TRECVID data | | | 1 | |
| Both IACC and non-IACC non-TRECVID data | | | | 3 |
| Total runs (102) | 96<br>94% | 2<br>2% | 1<br>1% | 3<br>3% |

# 50 concepts evaluated

**2 Adult**
**5 Anchorperson**
**10 Beach**
**21 Car**
**26 Charts**
**27 Cheering***
**38 Dancing***
**41 Demonstration_Or_Protest***
**44 Doorway***
**49 Explosion_Fire***
**50 Face**
**51 Female_Person**
**52 Female-Human-Face-Closeup***
**53 Flowers***
**59 Hand***
**67 Indoor**
**75 Male_Person**
**81 Mountain***
**83 News_Studio**
**84 Nighttime***
**86 Old_People***
**88 Overlaid_Text**
**89 People_Marching**
**97 Reporters**
**100 Running***
**101 Scene_Text**
**105 Singing***
**107 Sitting_down***
**108 Sky**
**111 Sports**
**113 Streets**
**123 Two_People**
**127 Walking***
**128 Walking_Running**
**227 Door_Opening**
**241 Event**
**251 Female_Human_Face**
**261 Flags**
**292 Head_And_Shoulder**
**332 Male_Human_Face**
**354 News**
**392 Quadruped**
**431 Skating**
**442 Speaking**
**443 Speaking_To_Camera**
**454 Studio_With_Anchorperson**
**464 Table**
**470 Text**
**478 Traffic**
**484 Urban_Scenes**

-The 10 marked with "*" are a subset of those tested in 2010

# 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:
  - 50 feature results for full runs
  - 23 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.

# 2011: mean extended Inferred average precision (xinfAP)

- 2 pools were created for each concept and sampled as:
  - Top pool (ranks 1-100) sampled at 100%
  - Bottom pool (ranks 101-2000) sampled at 8%

| 50 concepts |
|---|
| 268156 total judgments |
| 52522 total hits |
| 6747 Hits at ranks (1-10) |
| 28899 Hits at ranks (11-100) |
| 16876 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

# 2011 : 28/56 Finishers

```
--- --- KIS --- --- SIN   Aalto University
--- --- --- --- --- SIN   Beijing Jiaotong University
CCD INS KIS --- SED SIN   Beijing University of Posts and Telecommunications-MCPRL
CCD --- --- *** *** SIN   Brno University of Technology
--- *** *** MED SED SIN   Carnegie Mellon University
--- --- KIS MED --- SIN   Centre for Research and Technology Hellas
--- INS KIS MED --- SIN   City University of Hong Kong
--- --- KIS MED --- SIN   Dublin City University
--- --- --- *** --- SIN   East China Normal University
--- --- --- --- --- SIN   Ecole Centrale de Lyon, Université de Lyon
--- --- *** *** --- SIN   EURECOM
--- INS --- --- --- SIN   Florida International University
CCD --- --- --- --- SIN   France Telecom Orange Labs (Beijing)
--- --- --- --- --- SIN   Institut EURECOM
*** *** *** *** *** SIN   Tsinghua University, Fujitsu R&D and Fujitsu Laboratories
--- INS --- *** --- SIN   JOANNEUM RESEARCH Forschungsgesellschaft mbH and Vienna
University of Technology
--- --- *** MED --- SIN   Kobe University
*** INS *** *** *** SIN   Laboratoire d'Informatique de Grenoble
*** INS *** MED *** SIN   National Inst. of Informatics
*** *** *** *** SED SIN   NHK Science and Technical Research Laboratories
--- --- --- --- --- SIN   NTT Cyber Solutions Lab
--- *** --- MED --- SIN   Quaero consortium
--- --- --- MED SED SIN   Tokyo Institute of Technology, Canon Corporation
CCD --- --- --- --- SIN   University of Kaiserslautern
*** *** --- *** --- SIN   University of Marburg
--- *** *** MED --- SIN   University of Amsterdam
--- *** *** MED --- SIN   University of Electro-Communications
CCD --- --- --- --- SIN   University of Queensland
```

** : group didn't submit any runs

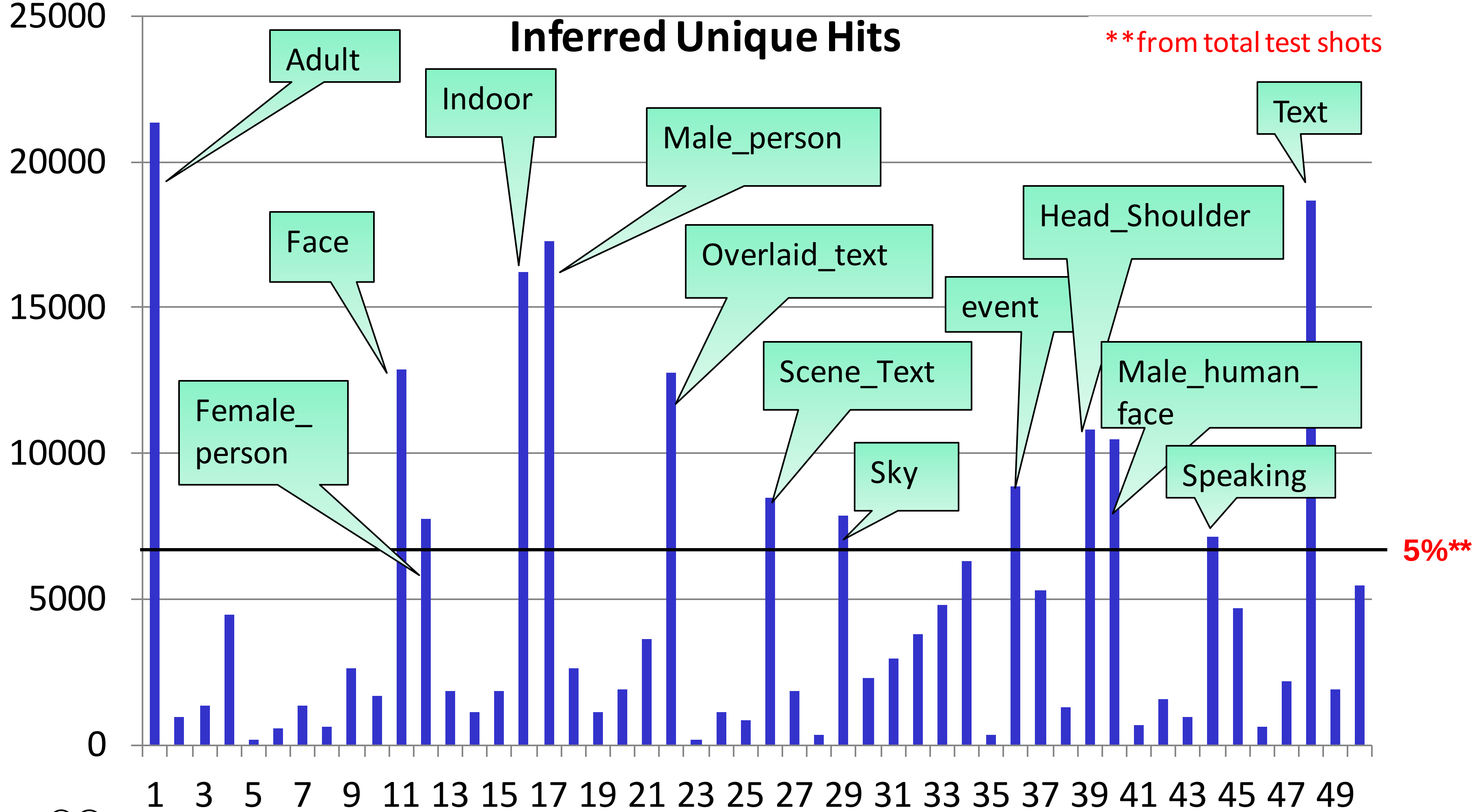
-- : group didn't participate

# 2011 : 28/56 Finishers

Participation and finishing declined! Why?

| | Task finishers | Participants |
|---|---|---|
| 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




2010 common features

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **6 Cheering** | **8 Demonstration_ Protest** | **10 Explostion_ Fire** | **14 Flowers** | **18 Mountain** | **21 Old_People** | **27 Singing** | **33 Walking** |
| 7 Dancing | 9 Doorway | 13 Female_face _closeup | 15 Hand | 20 Night_time | 25 Running | 28 Sitting_down | |

# True shots contributed uniquely by team

**Full runs**

| Team | No. of Shots | Team | No. of shots |
|---|---|---|---|
| Vid | 1130 | Mar | 69 |
| UEC | 965 | NHK | 49 |
| iup | 822 | dcu | 49 |
| vir | 749 | FTR | 42 |
| nii | 429 | Qua | 9 |
| CMU | 385 | FIU | 2 |
| ecl | 214 | | |
| brn | 185 | | |
| Pic | 177 | | |
| IRI | 154 | | |
| ITI | 151 | | |
| Tok | 140 | | |
| UvA | 72 | | |

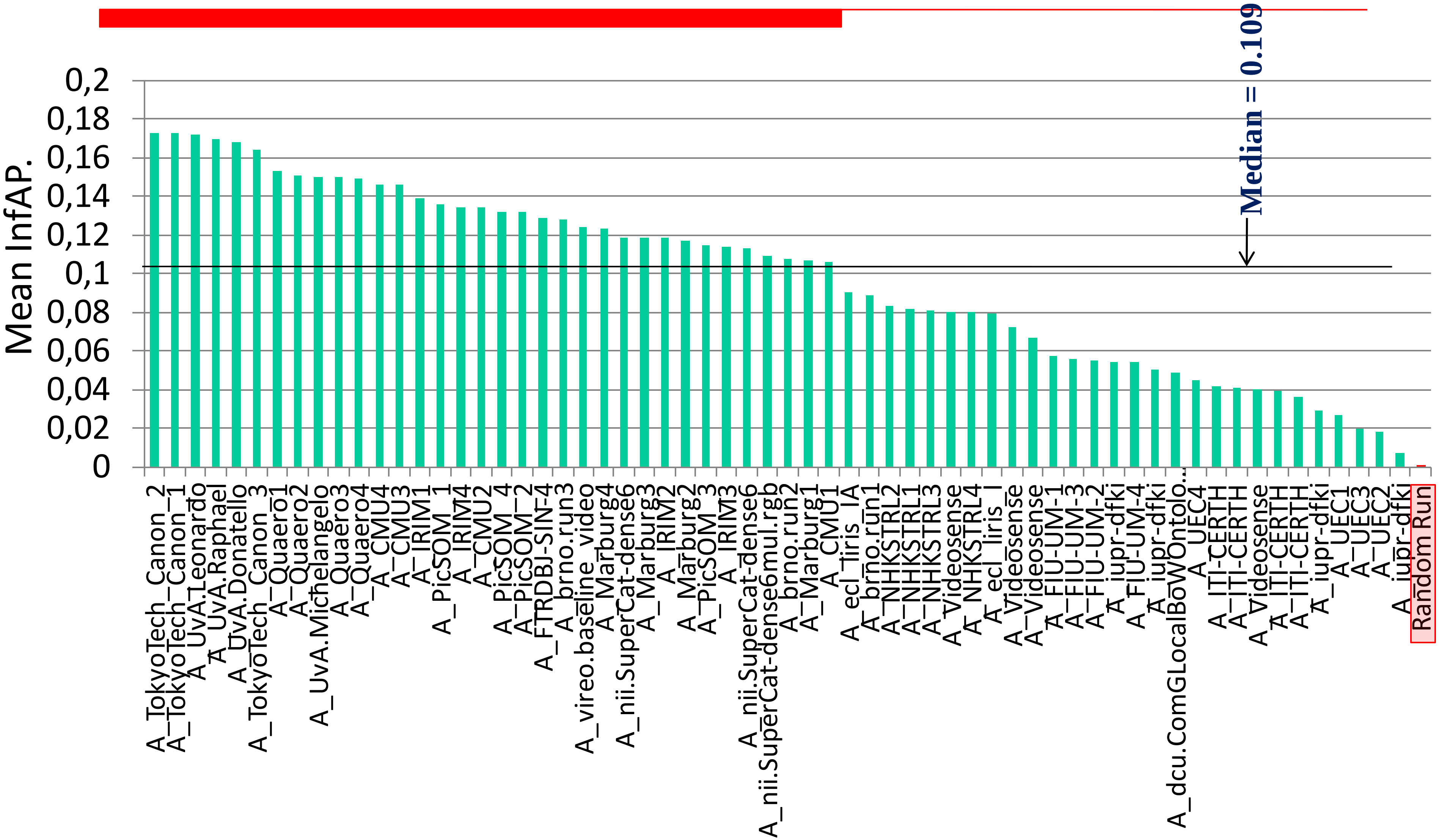
More unique shots compared to TV2010

**Lite runs**

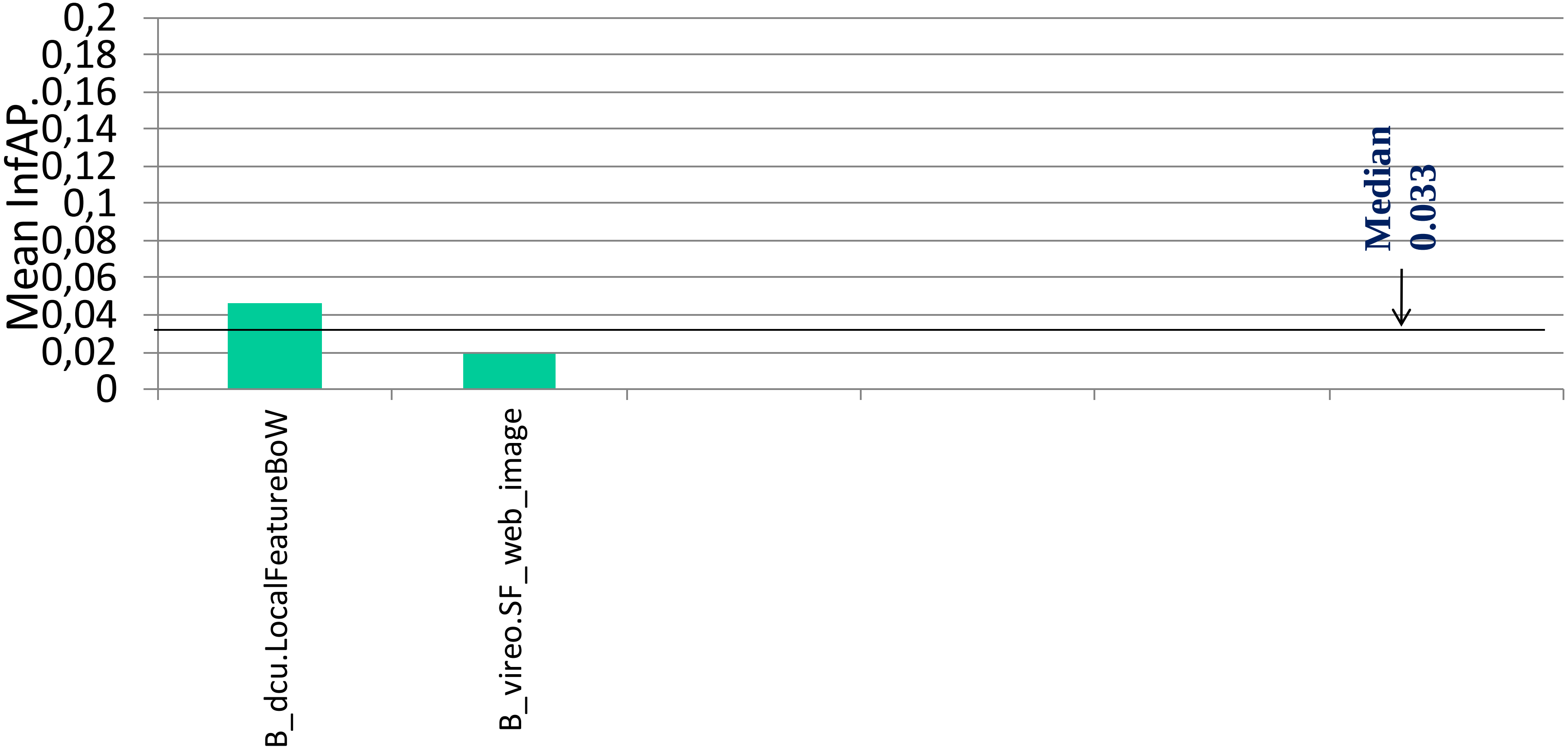
| Team | No. of Shots | Team | No. of shots |
|---|---|---|---|
| UEC | 506 | ITI | 41 |
| JRS | 404 | brn | 41 |
| Vid | 337 | FTR | 30 |
| iup | 318 | Tok | 25 |
| vir | 257 | UvA | 19 |
| BJT | 245 | UQM | 16 |
| MCP | 149 | Eur | 11 |
| nii | 145 | Mar | 9 |
| cs2 | 120 | ECN | 3 |
| CMU | 102 | Qua | 2 |
| IRI | 50 | | |
| thu | 48 | | |
| Pic | 45 | | |

No. of unique shots found are **MORE** than what was found in TV2010 (more shots this year)

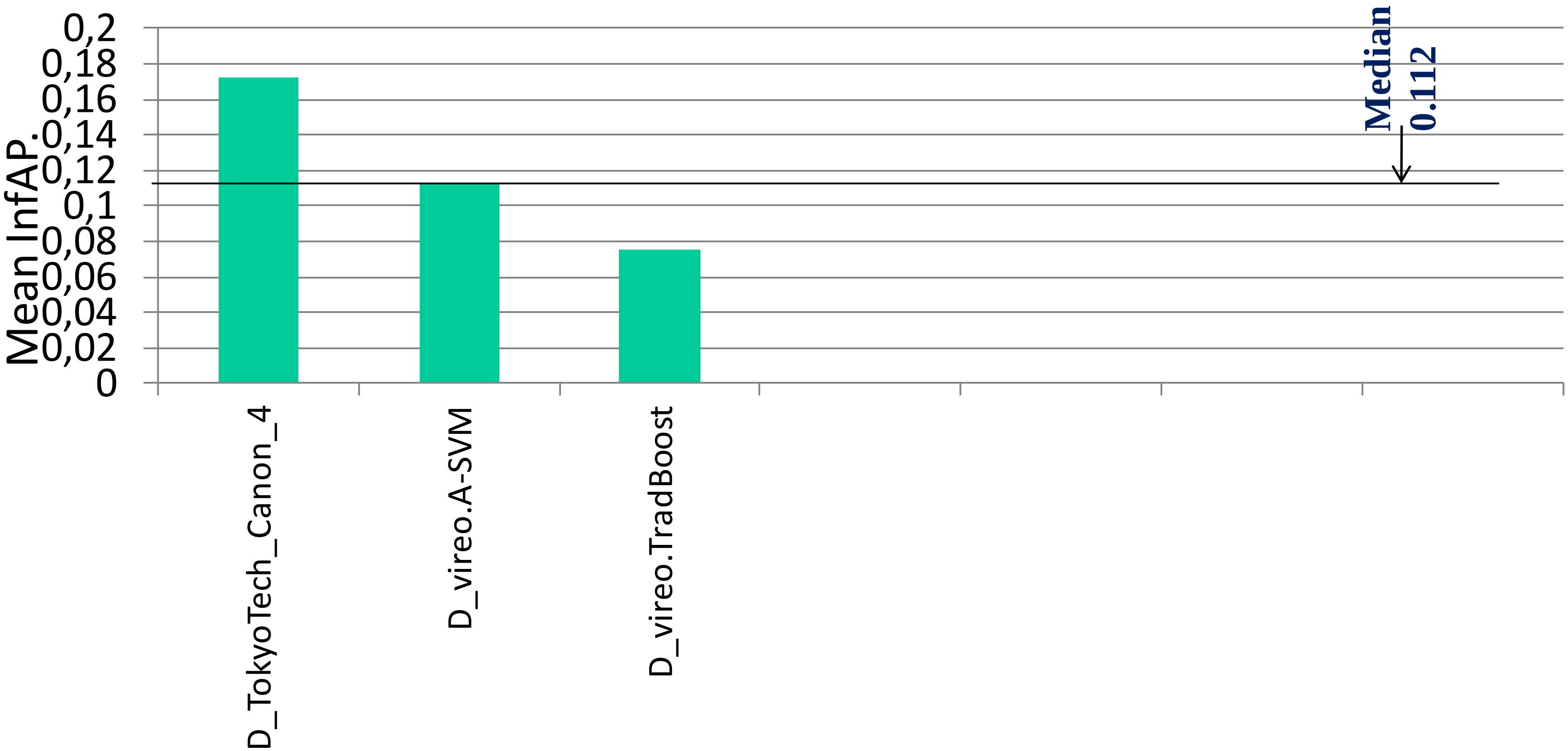
# Category A results (Full runs)



# Category B results (Full runs)

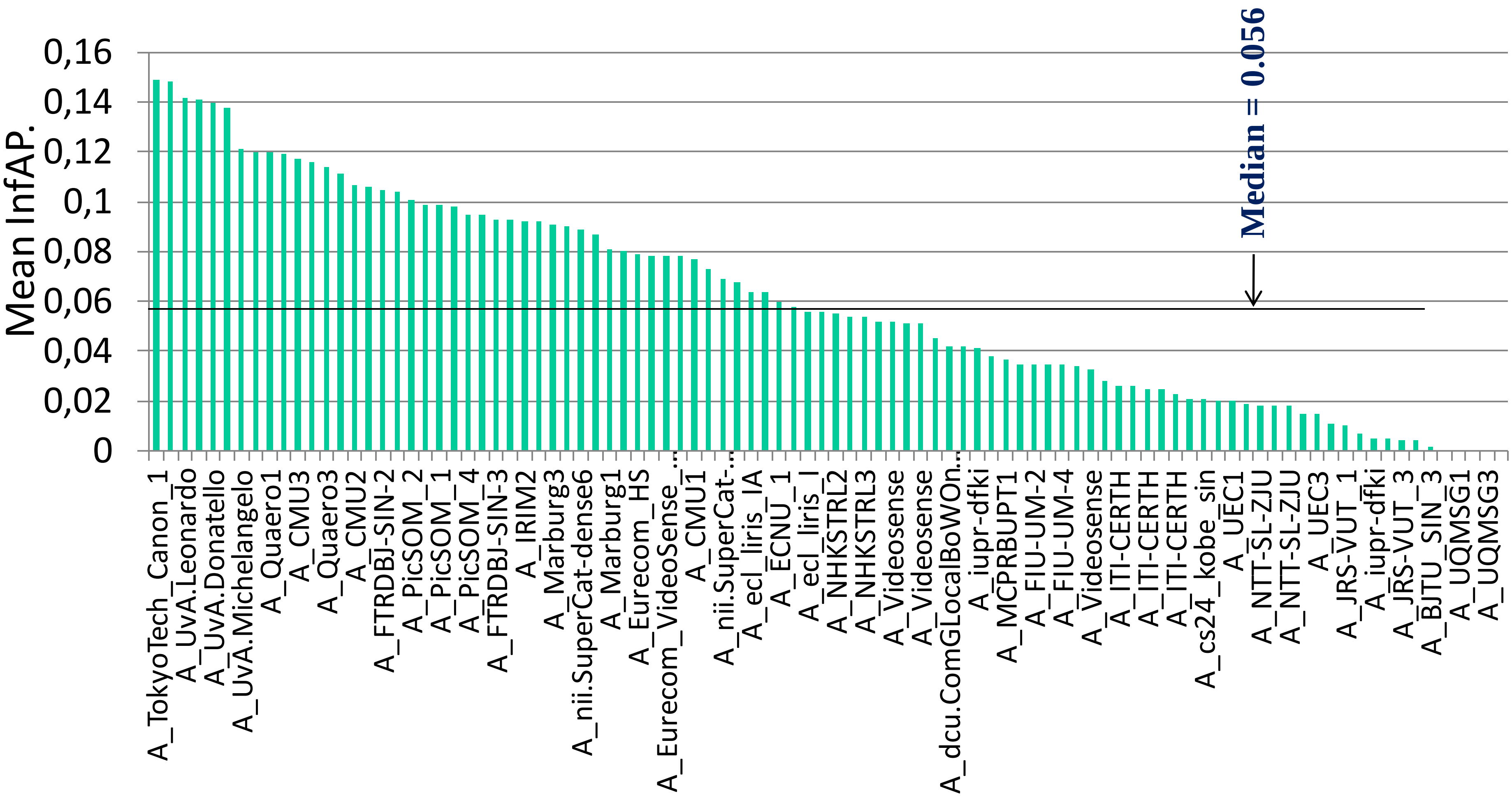


# Category D results (Full runs)

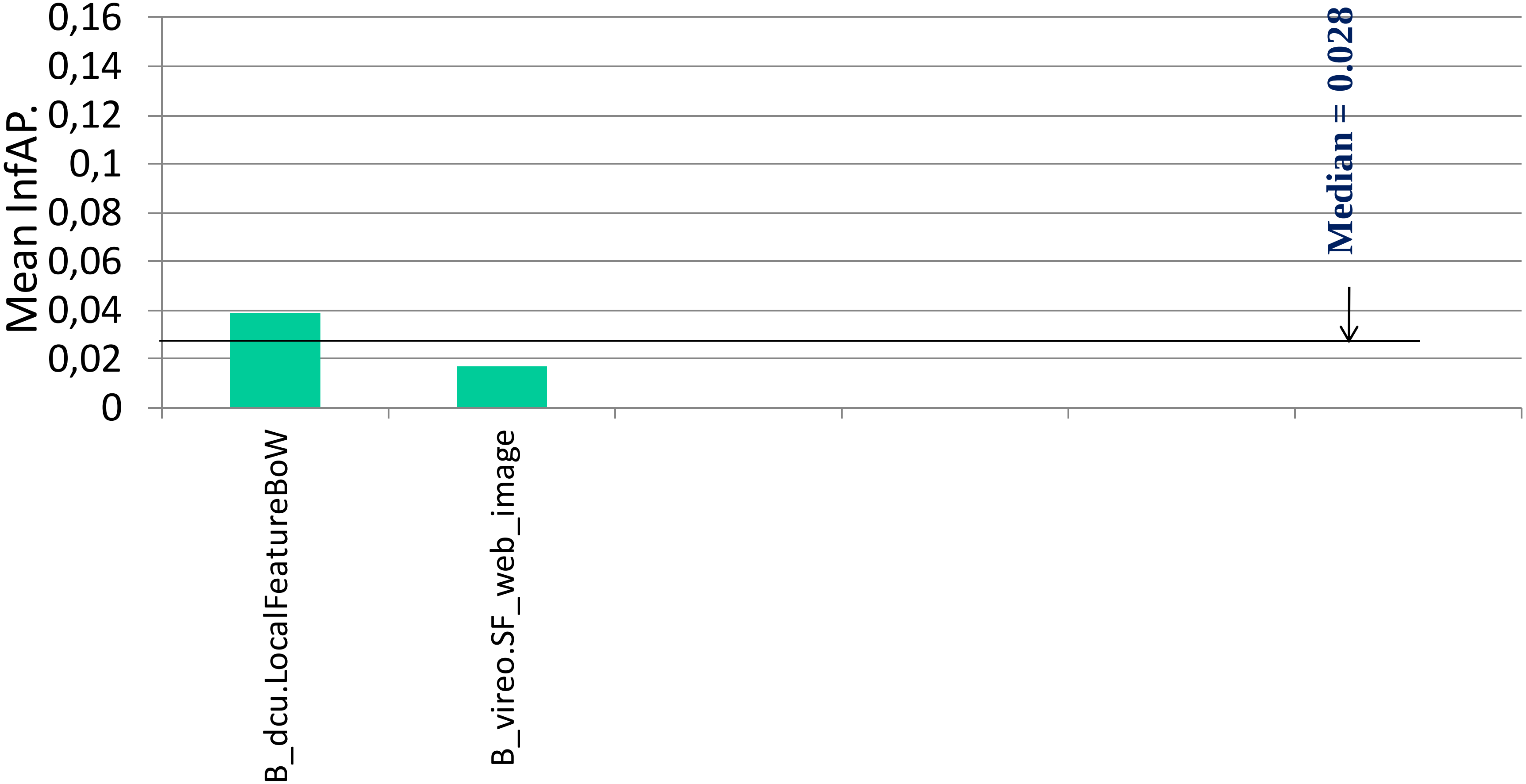



Note: Category C has only 1 run (C_dcu.GlobalFeature ) with score = 0.01

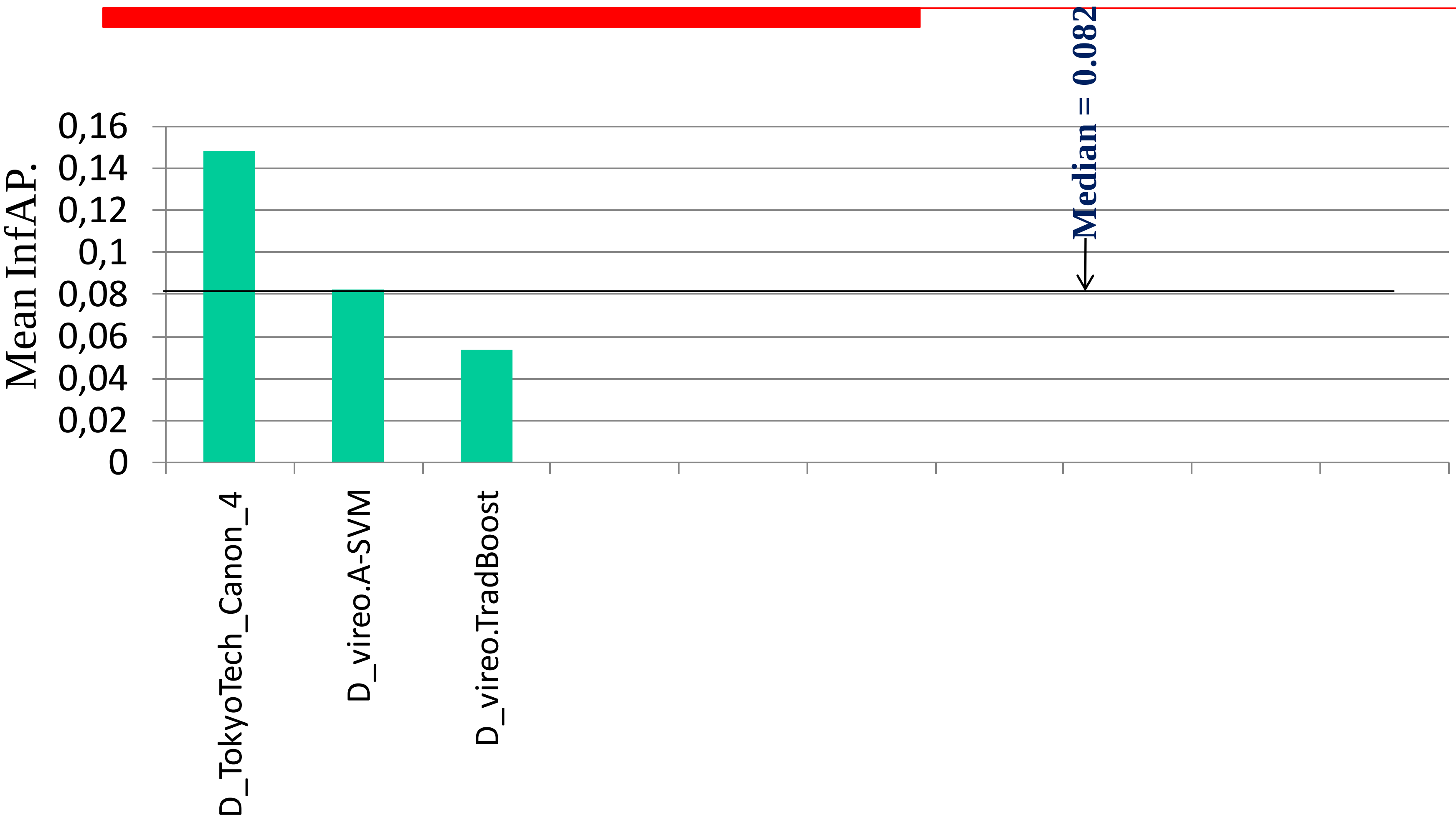
# Category A results (Lite runs)



# Category B results (Lite runs)

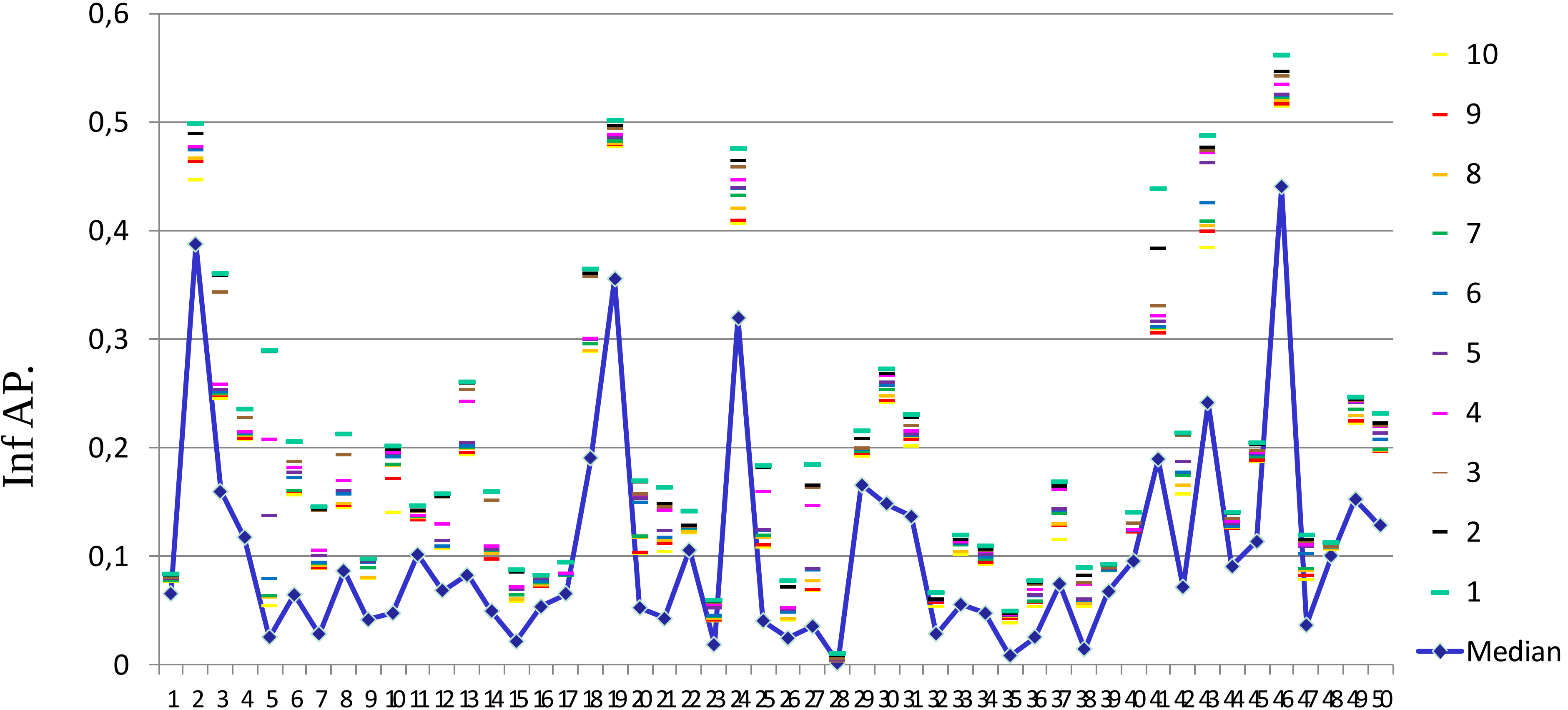


# Category D results (Lite runs)



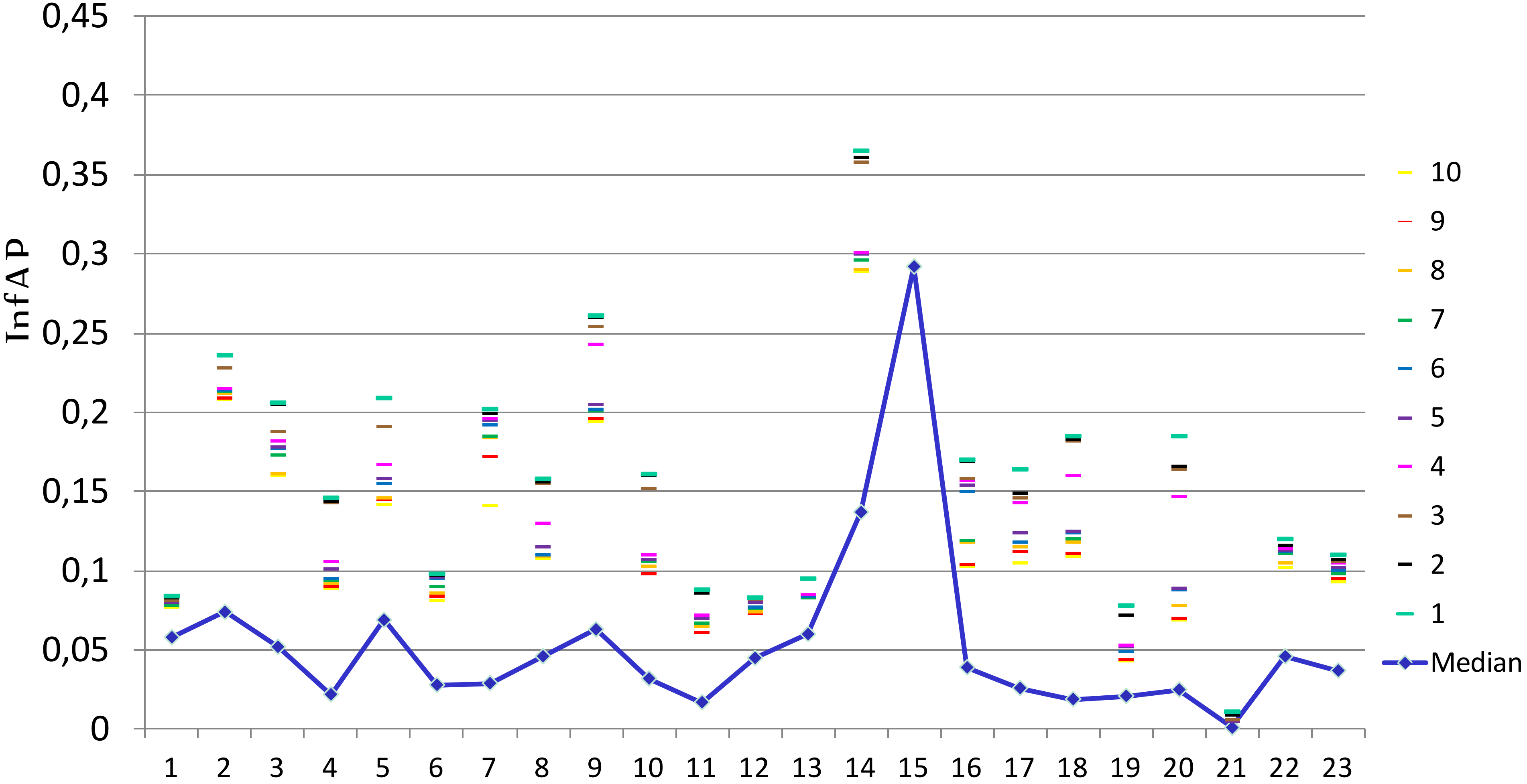

Note: Category C has only 1 run (C_dcu.GlobalFeature ) with score = 0.017

# Top 10 InfAP scores by feature (Full runs)




| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Adult | 2 Anchorperson | 3 Beach | 4 car | 5 Charts | 6 Cheering | 7 Dancing | 8 Demonstratio n/protest | 9 Doorway | 10 Explosion | 11 Face | 12 Female_person | 13 Female_face_ closeup | 14 Flowers | 15 Hand |
| 16 Indoor | 17 Male_person | 18 Mountain | 19 News_studio | 20 Nighttime | 21 old_people | 22 overlaid_text | 23 people_ marching | 24 Reporters | 25 Running | 26 Scene_Text | 27 Singing | 28 Sitting_down | 29 Sky | 30 Sports |
| 31 Streets | 32 Two_people | 33 Walking | 34 Walking_ Running | 35 Door_ opening | 36 Event | 37 Female_face | 38 Flags | 39 Head& Shoulders | 40 Male_human _face | 41 News | 42 Quadruped | 43 Skating | 44 Speaking | 45 Speaking_to_ camera |
| 46 Studio_with _anchorpers on | 47 Table | 48 Text | 49 Traffic | 50 Urban_scenes | | | | | | | | | | |

# Top 10 InfAP scores for 23 common features (Lite AND Full runs)




| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Adult | 2 Car | 3 Cheering | 4 Dancing | 5 Demonstratio n/protest | 6 Doorway | 7 Explosion | 8 Female_person | 9 Female_face_ Closeup | 10 Flowers | 11 Hand | 12 Indoor |
| 13 Male_person | 14 Mountain | 15 News_studio | 16 Nighttime | 17 old_people | 18 Running | 19 Scene_Text | 20 Singing | 21 Sitting_down | 22 Walking | 23 Walking_ Running | |

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

| Run name | (mean infAP) |
|---|---|
| A_TokyoTech_Canon_2 | 0.173 |
| A_TokyoTech_Canon_1 | 0.173 |
| A_UvA.Leonardo_1 | 0.172 |
| A_UvA.Raphael_3 | 0.170 |
| A_UvA.Donatello_2 | 0.168 |
| A_TokyoTech_Canon_3 | 0.164 |
| A_Quaero1 | 0.153 |
| A_Quaero2 | 0.151 |
| A_UvA.Michelangelo_4 | 0.150 |
| A_Quaero3 | 0.150 |

- A_UvA.Leonardo_1
  - A_UvA.Raphael_3
    - A_Quaero1
      - A_Quaero2
    - A_Quaero3
    - A_UvA.Michelangelo_4
  - A_UvA.Donatello_2
    - A_Quaero1
      - A_Quaero2
    - A_Quaero3
    - A_UvA.Michelangelo_4
- A_TokyoTech_Canon_2
  - A_TokyoTech_Canon_3
    - A_UvA.Michelangelo_4
    - A_Quaero1
      - A_Quaero2
    - A_Quaero3
- A_TokyoTech_Canon_1
  - A_TokyoTech_Canon_3
    - A_UvA.Michelangelo_4
    - A_Quaero1
      - A_Quaero2
    - A_Quaero3

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

| Run name | (mean infAP) |
|---|---|
| B_dcu.LocalFeatureBoW_2 | 0.046 |
| B_vireo.SF_web_image_4 | 0.019 |

- B_dcu.LocalFeatureBoW_2
  - B_vireo.SF_web_image_4

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

| Run name | (mean infAP) |
|---|---|
| D_TokyoTech_Canon_4 | 0.172 |
| D_vireo.A-SVM _3 | 0.112 |
| D_vireo.TradBoost_2 | 0.076 |

- D_TokyoTech_Canon_4
  - D_vireo.A-SVM _3
    - D_vireo.TradBoost_2

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

| Run name | (mean infAP) |
|---|---|
| A_TokyoTech_Canon_1 | 0.149 |
| A_TokyoTech_Canon_2 | 0.148 |
| A_UvA.Leonardo_1 | 0.142 |
| A_UvA.Raphael_3 | 0.141 |
| A_UvA.Donatello_2 | 0.140 |
| A_TokyoTech_Canon_3 | 0.138 |
| A_UvA.Michelangelo_4 | 0.121 |
| A_Quaero1 | 0.120 |
| A_CMU4 | 0.120 |
| A_Quaero2 | 0.119 |

- A_UvA.Leonardo_1
  - A_UvA.Donatello_2
    - A_CMU4
    - A_Quaero1
    - A_Quaero2
    - A_UvA.Michelangelo_4
  - A_UvA.Raphael_3
    - A_CMU4
    - A_Quaero1
    - A_Quaero2
    - A_UvA.Michelangelo_4
- A_TokyoTech_Canon_1
  - A_TokyoTech_Canon_3
    - A_CMU4
    - A_Quaero1
    - A_Quaero2
    - A_UvA.Michelangelo_4
- A_TokyoTech_Canon_2
  - A_TokyoTech_Canon_3
    - A_CMU4
    - A_Quaero1
    - A_Quaero2
    - A_UvA.Michelangelo_4

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

| Run name | (mean infAP) |
|---|---|
| B_dcu.LocalFeatureBoW_2 | 0.039 |
| B_vireo.SF_web_image_4 | 0.017 |

- B_dcu.LocalFeatureBoW_2
  - B_vireo.SF_web_image_4

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

| Run name | (mean infAP) |
|---|---|
| D_vireo.TradBoost_2 | 0.054 |
| D_vireo.A-SVM_3 | 0.082 |
| D_TokyoTech_Canon_4 | 0.148 |

- D_TokyoTech_Canon_4
  - D_vireo.A-SVM_3
    - D_vireo.TradBoost_2

# Observations

- Site experiments include:
  - focus on robustness, merging many different representations
  - use of spatial pyramids
  - improved bag of word approaches
  - improved kernel methods
  - sophisticated fusion strategies
  - combination of low and intermediare/gigh features
  - 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 external (ImageNet 1000-concept) data
- Still not many experiments using external training data (main focus on category A)
- No improvement using external training data

# Presentations to follow

- 2:40 - 3:00, Tokyo Institute of Technology, Canon Corporation
- 3:00 - 3:20, PicSOM - Aalto University
- 3:20 - 3:40, CMU-Informedia - Carnegie Mellon University
- 3:40 - 4:00, Break in the NIST West Square Cafeteria
- 4:00 - 4:20, Quaero - Quaero Consortium
- 4:20 - 4:40, Discussion

# Less participation – poll results – this year

- Has the task become too big considering video data?
  - No (3).
  - Close to the limit.
  - Yes.
- Has the task become too big considering the number of concepts?
  - No (3).
  - Yes (2), we did not participate for this reason; at least the full task
- Did the task not brought enough novelty compared to previous years?
  - Yes, this is a concern, the task lacks excitement.
  - Not so much.
  - We found it sufficiently interesting to participate
  - Yes. A challenging topic for this year's task was the increasing of the number of concepts.
- Any other reason or issue with the task?
  - US Aladdin program / MED task competition?
  - Only 50 (of 346) concepts are evaluated in the testing phase. We would like to know how the Mean InfAP will change if the number of testing concepts is increased (lite versus full results already show some consistency)

# Poll results – next year

- Should we continue to increase the number of concepts for the full task?
  - Why increase? What is the underlying scientific question?
  - Possibly but slowly.
  - Slightly or keep the current size.
  - Yes, but the selected concepts should not be dropped out like this year. It's okay to keep the number of concepts.
  - No.
- Should we keep, reduce or increase the number of concepts for the light task?
  - No opinion.
  - Reduce the number. It is important to be able to annotate the data with ground truth. This is not possible if there are too many concepts.
  - Preferably less.
  - Keep the current size (3).
- Should we continue increasing the diversity of target concepts or not?
  - Again, what is the scientific rationale?
  - Maybe another task.
  - Yes, definitely.
  - Yes. How about increasing concepts of human emotion?
  - Yes.

# Poll results – next year

- Any other suggestion for introducing novelty in this task?
  - Perhaps collecting training data in an automatic fashion, rather than using the collaborative annotations.
  - Increase the diversity of video sources, in terms of countries and languages.
  - Increase the diversity of evaluation measures, not confine to MAP.
  - How about having multiple levels of appearance for positive samples?
  - Consider an online variant.
- Additional comments
  - 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.
  - The data size might be too big and it seems that computation power and storage play a key role to get promising results.
  - Improve the quality of the videos.
  - Low number of positive samples is a problem.
  - Provide clearer specification on all concepts.
  - Some concepts have very few positive instances.
  - Suggest change data type every year.
- **Many thanks for the feedback!**

# SIN 2012

- A maximum number of participants is good but not the goal; we want people to be happy with the proposed task.
- What is the scientific rationale for many and diverse concepts?
  - Potential applications require a large number of concepts and very diverse ones.
  - Scalability at the computing power level is not the only issue.
  - Relations between concepts (both explicit and implicit) may have a key role to play; this can be exploited and evaluated only at a sufficient scale.
- Another possible novelty:
  - Multiple levels of relevance for positive samples or ranking of positive samples
- Same or similar task; same type of data; similar volume of data.
- Comparable or slightly reduced number of concepts.
- Better definition of concepts, better annotation.
- Encourage and provide infrastructure for sharing contributed elements: low-level features, detection scores, ...