# TRECVID-2008 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 category A
  - comparison with TV2007
- 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 20 LSCOM features
- TRECVID 2008 video data
  - Netherlands Institute for Sound and Vision (~**200 hours** of news magazine, science news, news reports, documentaries, educational programming and archival video in MPEG-1).
  - 100 hours for development.
  - 100 hours for test.
  - TRECVID 2003, 2005 & TRECVID 2007 annotated data.

# High-level feature task (2)

- NIST evaluated 20 features using a 50% random sample of the submission pools (Inferred AP)
- Six training types were allowed
  - A : Systems trained on only common TRECVID development collection data
  - B : Systems trained on only common development collection data but not on (just) common annotation of it.
  - C : System is not of type A or B
  - a : same as A but no training data specific to any sound and vision data has been used
  - b : same as B but no training data specific to any sound and vision data has been used
  - c : same as C but no training data specific to any sound and vision data has been used

# TV2007 vs TV2008 dataset

| | TV2007 | TV2008 |
|---|---|---|
| Dataset length (hours) | ~100 | ~200 |
| Number of shots | 18,142 | 35,766 |
| Number of unique program titles | 47 | 77 |

# 20 LSCOM features evaluated

- 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

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

# 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)

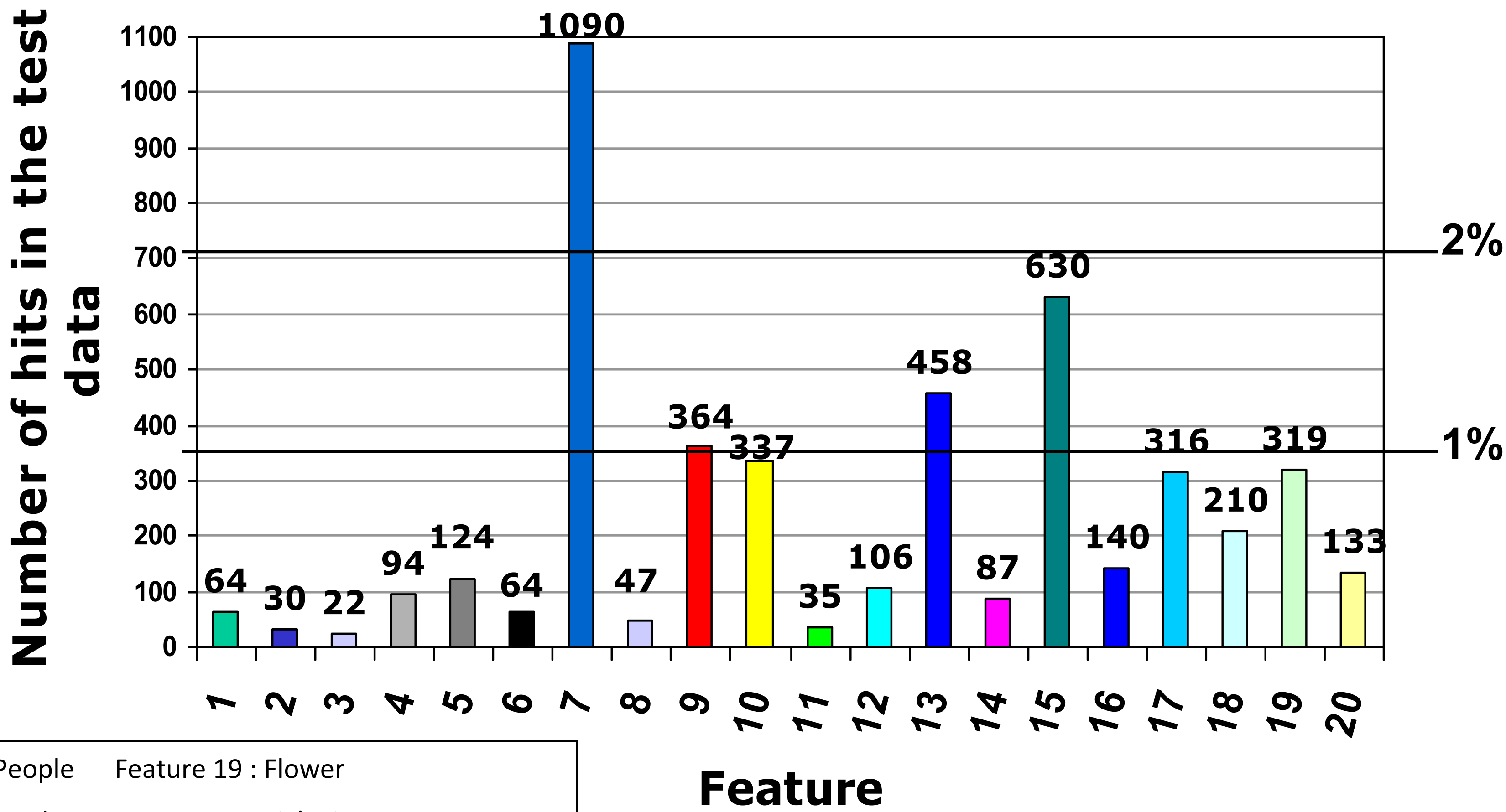
- Developed* by Emine Yilmaz and Javed A. Aslam at Northeastern University
- Estimates average precision surprisingly well using a surprisingly small sample of judgments from the usual submission pools
- Experiments on TRECVID 2005 & 2006 & 2007 feature submissions confirmed quality of the estimate in 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.

# 2008: Inferred average precision (infAP)

- Submissions for each of 20 features were pooled down to about average 130 items (so that each feature pool contained ~ 6777 shots)
  - varying pool depth per feature
- A 50% random sample of each pool was then judged:
- 67,774 total judgments  (TV7: 66,293)
- 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

# Frequency of hits varies by feature




Feature 7 : 2 People    Feature 19 : Flower

Feature 15 : Hand    Feature 17 : Nighttime

Fetaure 13 : Street

Feature 9 : Driver

Feature 10 : Cityscape

# 2008 : 43/115 Participants

| | | | | | |
|---|---|---|---|---|---|
| Asahikasei Co. | -- | ** | FE | RU | -- |
| Bilkent University | CD | -- | FE | -- | ** |
| Brno University of Technology | CD | ED | FE | ** | SE |
| Beijing University of Posts and Telecommunications | CD | ** | FE | -- | -- |
| Carnegie Mellon University | -- | ED | FE | RU | ** |
| Columbia University | CD | -- | FE | -- | SE |
| COST292 Team (Delft Univ.) | CD | ** | FE | RU | SE |
| Florida International Univ. | -- | ** | FE | -- | -- |
| Fudan University | CD | ED | FE | -- | SE |
| IBM T. J. Watson Research Center | CD | ** | FE | ** | SE |
| INRIA-LEAR | CD | -- | FE | -- | -- |
| MMIS (Open Univ.) | ** | -- | FE | -- | SE |
| Microsoft Research Asia | ** | ** | FE | ** | SE |
| NHKSTRL | ** | ED | FE | RU | ** |
| National Institute of Informatics | CD | ** | FE | RU | SE |
| IRIM | ** | ** | FE | RU | ** |
| ISM (The Institute of Statistical Mathematics) | -- | -- | FE | -- | -- |
| IUPR-DFKI | ** | -- | FE | -- | -- |

** : group didn't submit any runs

-- : group didn't participate

# 2008: 43 Participants (continued)

| | | | | | |
|---|---|---|---|---|---|
| JOANNEUM RESEARCH Forschungsgesellschaft mbH | ** | ** | FE | RU | -- |
| LIG (Laboratoire d'Informatique de Grenoble) | ** | -- | FE | -- | ** |
| Laboratoire LIRIS (LYON) | -- | ** | FE | ** | ** |
| University of Twente and CWI | -- | -- | FE | -- | SE |
| LSIS_GLOT (CNRS LSIS) | -- | -- | FE | -- | -- |
| Marburg | ** | ** | FE | ** | ** |
| MCG-ICT-CAS | CD | ED | FE | -- | SE |
| Mediamill (Univ. of Amsterdam) | -- | ** | FE | -- | SE |
| MESH | -- | -- | FE | - | SE |
| National Taiwan University | ** | ** | FE | -- | SE |
| Oxford Univ. | ** | -- | FE | -- | SE |
| PKU-ICST (Peking Univ.) | ** | ** | FE | ** | SE |
| PicSom(Helsinki University of Technology) | CD | -- | FE | RU | SE |
| Queensland University of Technology | -- | -- | FE | RU | -- |
| REGIM | -- | ** | FE | RU | SE |
| SJTU | -- | ED | FE | -- | SE |
| SP-UC3M (Universidad Carlos III de Madrid) | -- | -- | FE | -- | SE |
| Thu-intel | CD | ** | FE | RU | SE |
| Tokyo Institute of Technology | -- | ED | FE | RU | -- |
| University of Electro-Communications | ** | ** | FE | RU | ** |
| University of Karlsruhe (TH) | -- | -- | FE | -- | -- |

# 2008: 43 Participants (continued)

| | |
|---|---|
| Universite Pierre et Marie Curie - LIP6 | ** ** FE RU -- |
| VIREO (City University of Hong Kong) | CD ** FE RU SE |
| VITALAS (CERTH-ITI (GR), CWI(NL), U.Sunderland (UK)) | -- -- FE -- SE |
| XJTU (Xi'an Jiaotong University) | ** -- FE -- -- |

*HLF keeps attracting more participants, most of them come back the next year.*

| | | |
|---|---|---|
| 2008 | 43 | 115 |
| 2007 | 32 | 54 |
| 2006 | 30 | 54 |
| 2005 | 22 | 42 |
| 2004 | 12 | 33 |

# Number of runs of each training type

| Tr-Type | 2008 | 2007 |
|---|---|---|
| A | 152 (76%) | 146 (89.5%) |
| B | 15 (7.5%) | 7 (4.3%) |
| C | 22(11%) | 6 (3.7%) |
| a | 9(4.5%) | 4 (2.5%) |
| b | 0 | 0 |
| c | 2(1%) | 0 |
| Total runs | 200 | 163 |

1- More interest in special training data (B, C).
2- More interest in un-related Data (a, c).
3- The common data (A) still is the most popular.

System training type:

**A** - Only on common dev. collection and the common annotation.

**B** - Only on common dev. collection but not on (just) the common annotation.

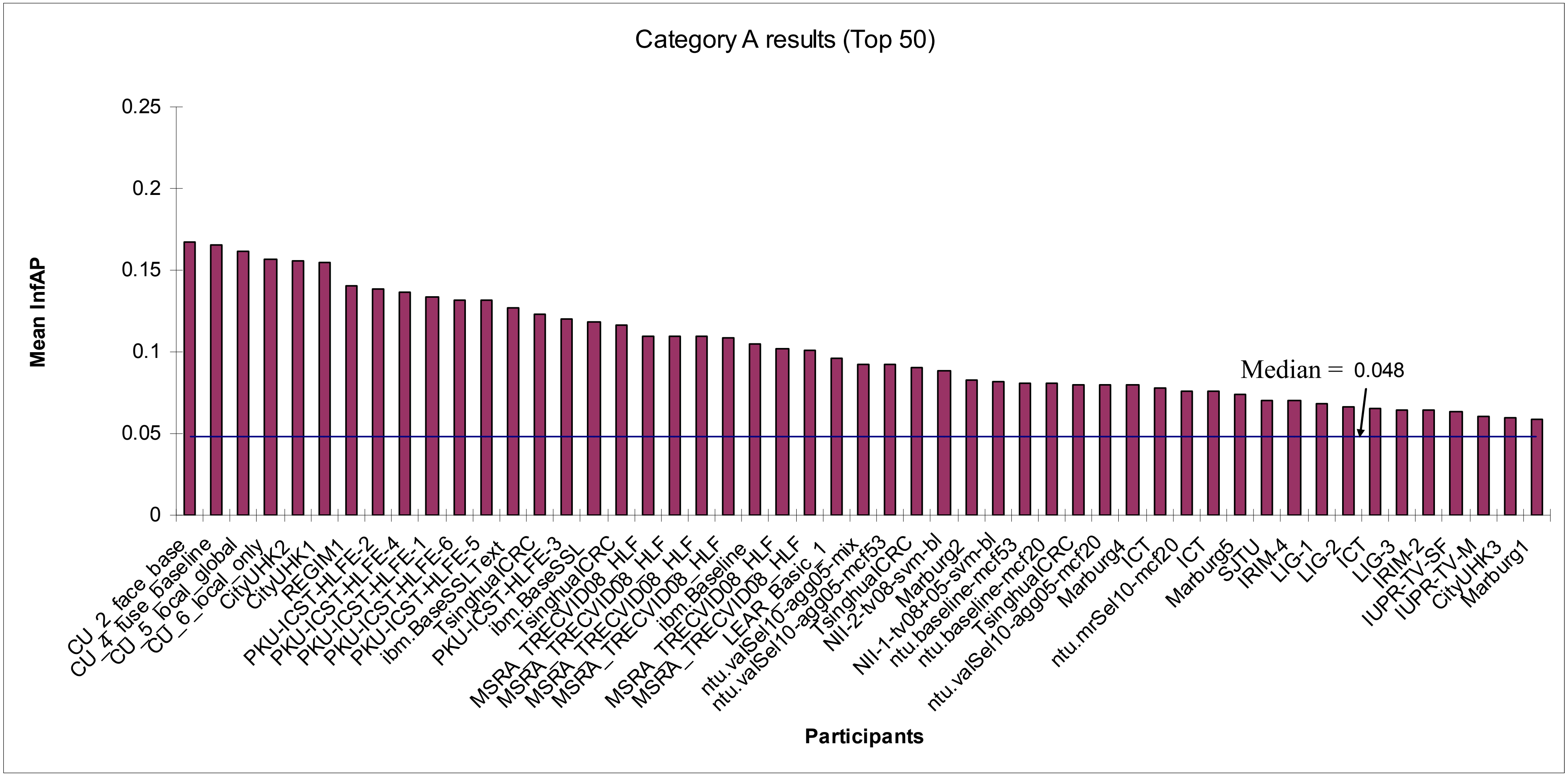
**C** - not of type A or B.

**a , b, c** – Same as A, B, & C respectively but without using any specific training data from Sound and Vision dataset.

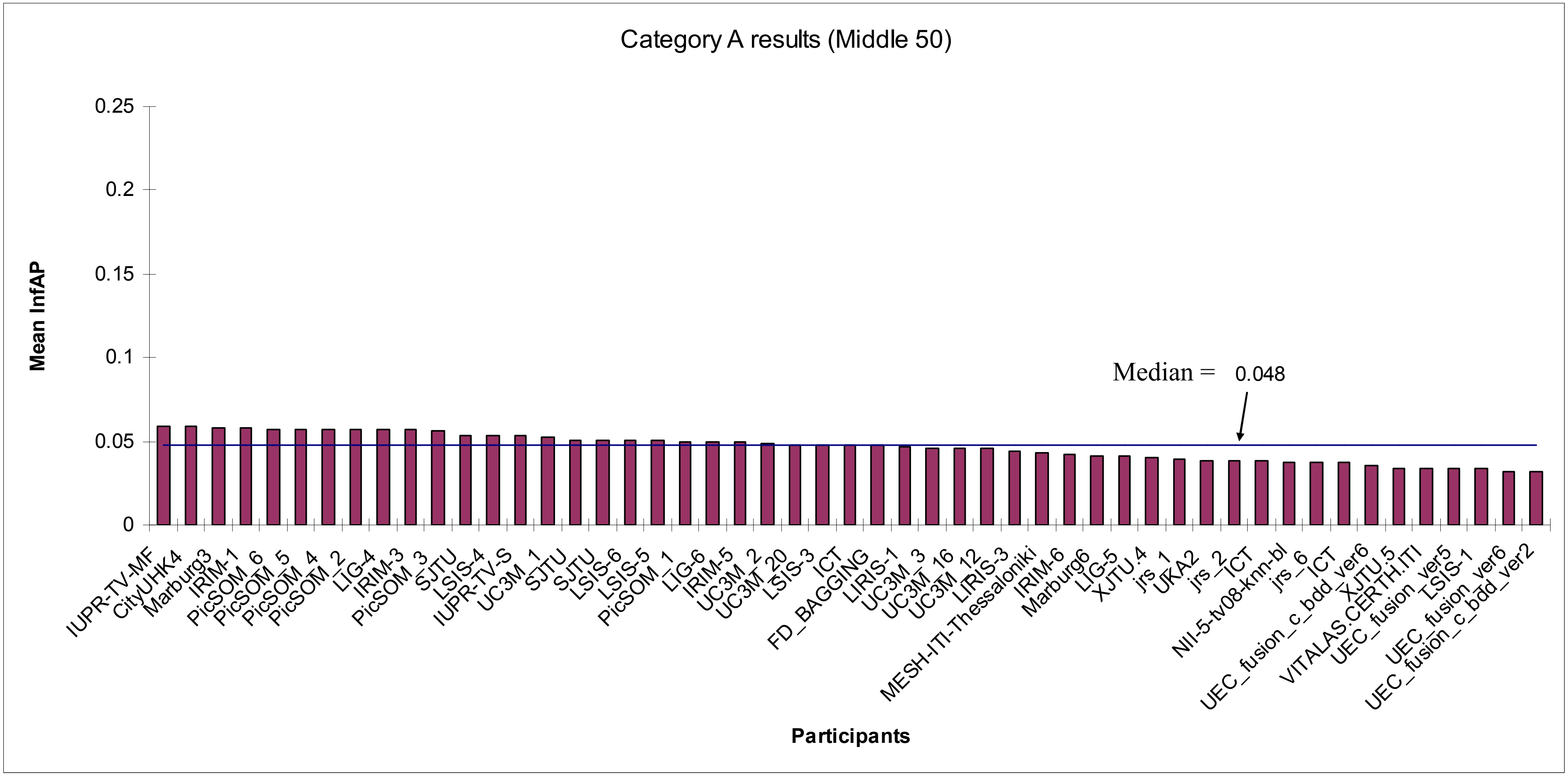
# True shots contributed uniquely by team for each feature

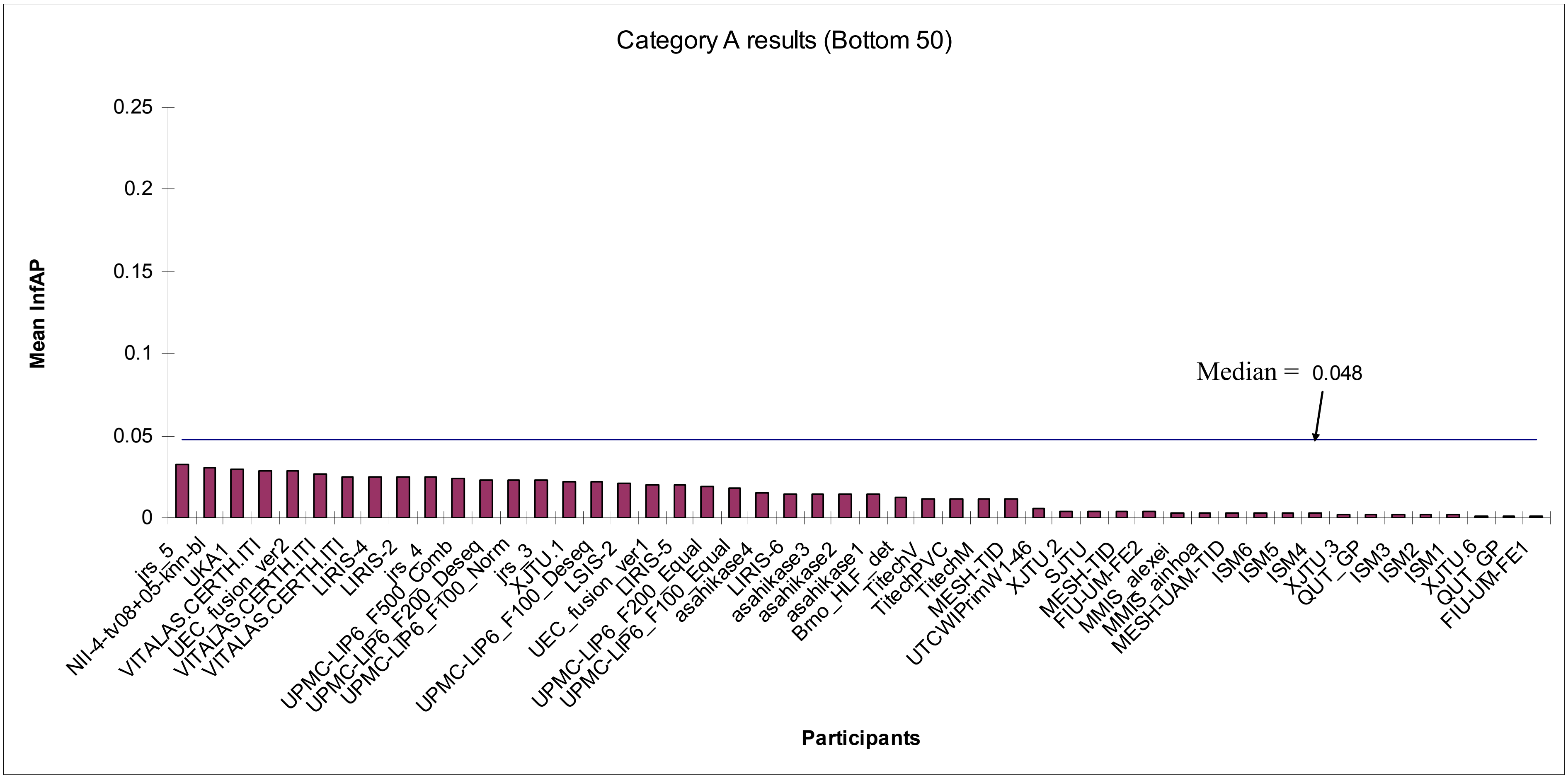
- BUPT - 1 shot (feature 7 : 2 people)
- IUPR-DFKI – 1 shot (feature 15 : hand)
- NHKSTRL – 1 shot (feature 17 : Nighttime)
- Queensland University – 1 shot (feature 18 : Boat_Ship)
- Institute of Image Communication and Information Processing – 1 shot (feature 8 : Bus)
- Asahikasei – 1 shot (feature 7 : 2 people)
- Delft University of Technology – 2 shots (feature 12 : Telephone , feature 13 : Street)
- CNRS LSIS – 2 shots (feature 7 : 2 people)
- MESH – 2 shots (feature 15 : hand , feature 16 : Mountain)
- UTC – 2 shots (feature 7 : 2 people)
- École Nationale d'Ingénieurs de Sfax ENIS – 6 shots (feature 17 : nighttime , feature 4 : dog , feature 12 :  Telephone)

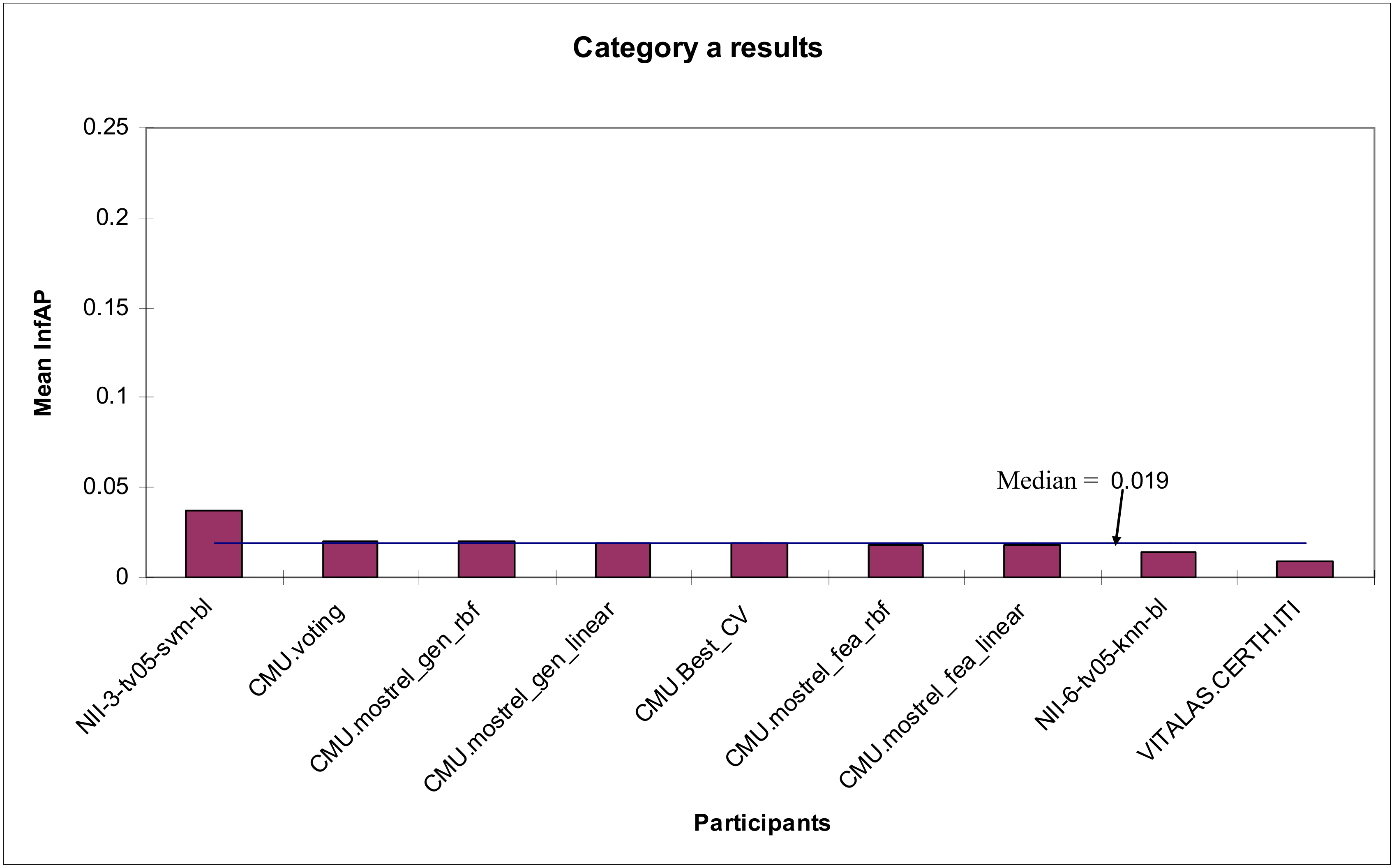
-Unlike TRECVID 2007 where <u>only two</u> groups found different unique true shots.

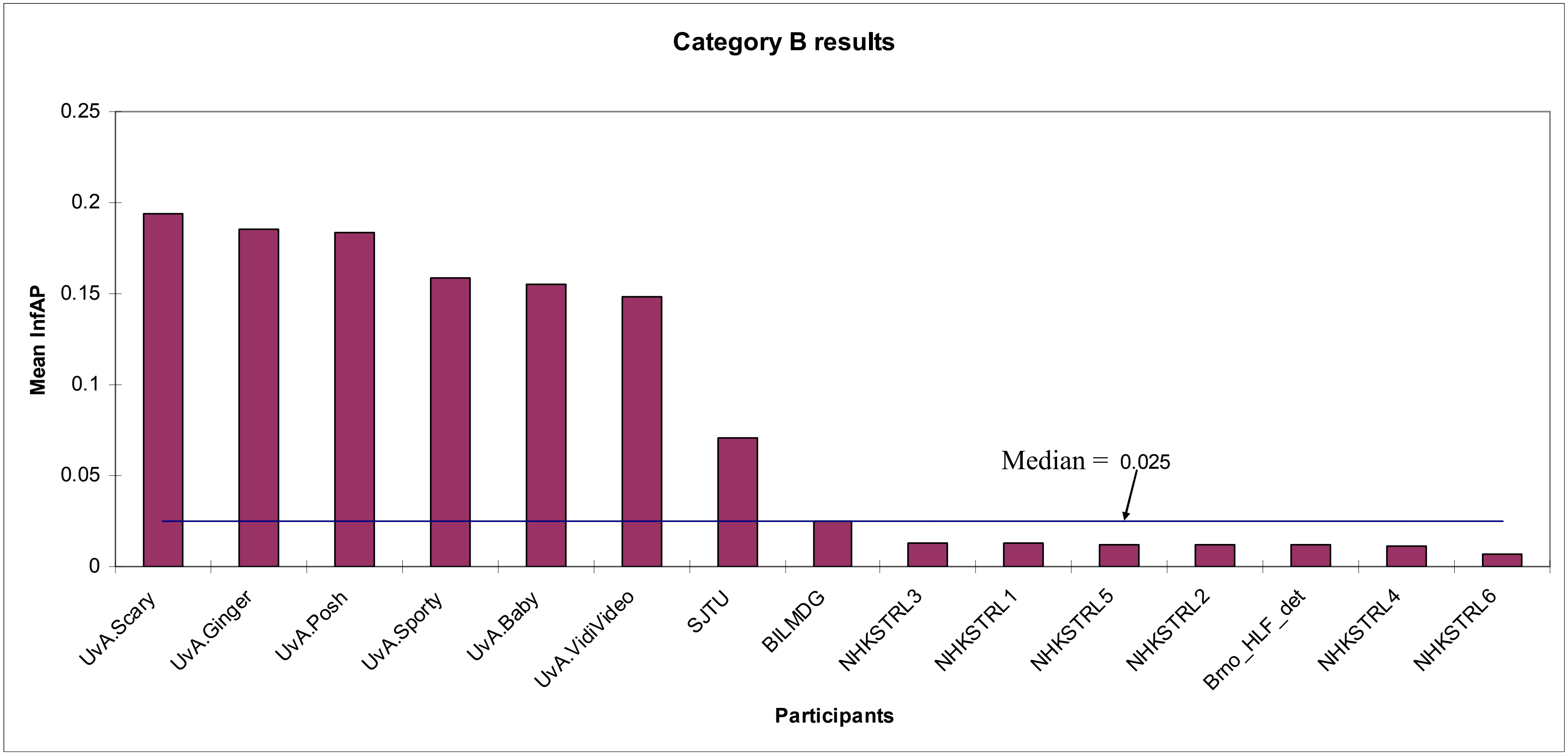


Category A results (Top 50)
Mean InfAP
0.25
0.2
0.15
0.1
0.05
0
Median = 0.048
Participants
CU_2_face_base
CU_4_fuse_baseline
CU_5_local_global
CU_6_local_only
CityUHK2
CityUHK1
REGIM1
PKU-ICST-HLFE-2
PKU-ICST-HLFE-4
PKU-ICST-HLFE-1
PKU-ICST-HLFE-6
PKU-ICST-HLFE-5
ibm.BaseSSLText
TsinghuaICRC
PKU-ICST-HLFE-3
ibm.BaseSSL
TsinghuaICRC
MSRA_TRECVID08_HLF
MSRA_TRECVID08_HLF
MSRA_TRECVID08_HLF
MSRA_TRECVID08_HLF
ibm.Baseline
MSRA_TRECVID08_HLF
MSRA_TRECVID08_HLF
LEAR_Basic_1
ntu.valSel10-agg05-mix
ntu.valSel10-agg05-mcf53
TsinghuaICRC
NII-2-tv08-svm-bl
Marburg2
NII-1-tv08+05-svm-bl
ntu.baseline-mcf53
ntu.baseline-mcf20
TsinghuaICRC
ntu.valSel10-agg05-mcf20
Marburg4
ICT
ntu.mrSel10-mcf20
ICT
Marburg5
SJTU
IRIM-4
LIG-1
LIG-2
ICT
LIG-3
IRIM-2
IUPR-TV-SF
IUPR-TV-M
CityUHK3
Marburg1

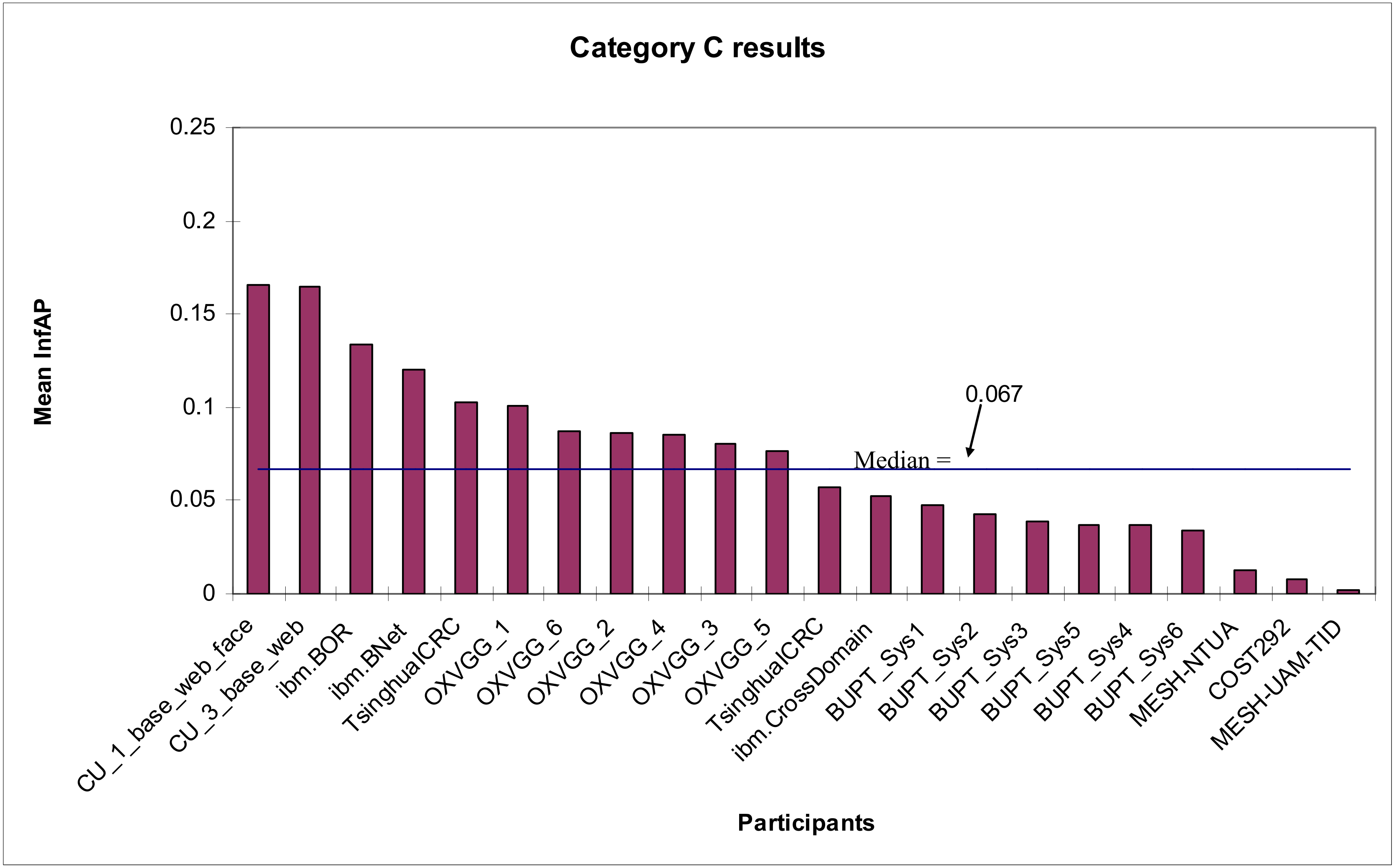

Category A results (Middle 50)
Mean InfAP
0.25
0.2
0.15
0.1
0.05
0
Median = 0.048
IUPR-TV-MF
CityUHK4
Marburg3
IRIM-1
PicSOM_6
PicSOM_5
PicSOM_4
PicSOM_2
LIG-4
IRIM-3
PicSOM_3
SJTU
LSIS-4
IUPR-TV-S
UC3M_1
SJTU
SJTU
LSIS-6
LSIS-5
PicSOM_1
LIG-6
IRIM-5
UC3M_2
UC3M_20
LSIS-3
ICT
FD_BAGGING
LIRIS-1
UC3M_3
UC3M_16
UC3M_12
LIRIS-3
MESH-ITI-Thessaloniki
IRIM-6
Marburg6
LIG-5
XJTU.4
jrs_1
UKA2
jrs_2
ICT
NII-5-tv08-knn-bl
jrs_6
ICT
UEC_fusion_c_bdd_ver6
XJTU.5
VITALAS.CERTH.ITI
UEC_fusion_ver5
LSIS-1
UEC_fusion_ver6
UEC_fusion_c_bdd_ver2
Participants


Category A results (Bottom 50)
Mean InfAP
0.25
0.2
0.15
0.1
0.05
0
Median = 0.048
jrs_5
NII-4-tv08+05-knn-bl
UKA1
VITALAS.CERTH.ITI
UEC_fusion_ver2
VITALAS.CERTH.ITI
VITALAS.CERTH.ITI
LIRIS-4
LIRIS-2
jrs_4
UPMC-LIP6_F500_Comb
UPMC-LIP6_F200_Deseq
UPMC-LIP6_F100_Norm
jrs_3
XJTU.1
UPMC-LIP6_F100_Deseq
LSIS-2
UEC_fusion_ver1
LIRIS-5
UPMC-LIP6_F200_Equal
UPMC-LIP6_F100_Equal
asahikase4
LIRIS-6
asahikase3
asahikase2
asahikase1
Brno_HLF_det
TitechV
TitechPVC
TitechM
MESH-TID
UTCWIPrimW1-46
XJTU.2
SJTU
MESH-TID
FIU-UM-FE2
MMIS_alexei
MMIS_ainhoa
MESH-UAM-TID
ISM6
ISM5
ISM4
XJTU.3
QUT_GP
ISM3
ISM2
ISM1
XJTU.6
QUT_GP
FIU-UM-FE1
Participants


Category a results
0.25
0.2
0.15
0.1
0.05
0
Mean InfAP
Median = 0.019
NII-3-tv05-svm-bl
CMU.voting
CMU.mostrel_gen_rbf
CMU.mostrel_gen_linear
CMU.Best_CV
CMU.mostrel_fea_rbf
CMU.mostrel_fea_linear
NII-6-tv05-knn-bl
VITALAS.CERTH.ITI
Participants

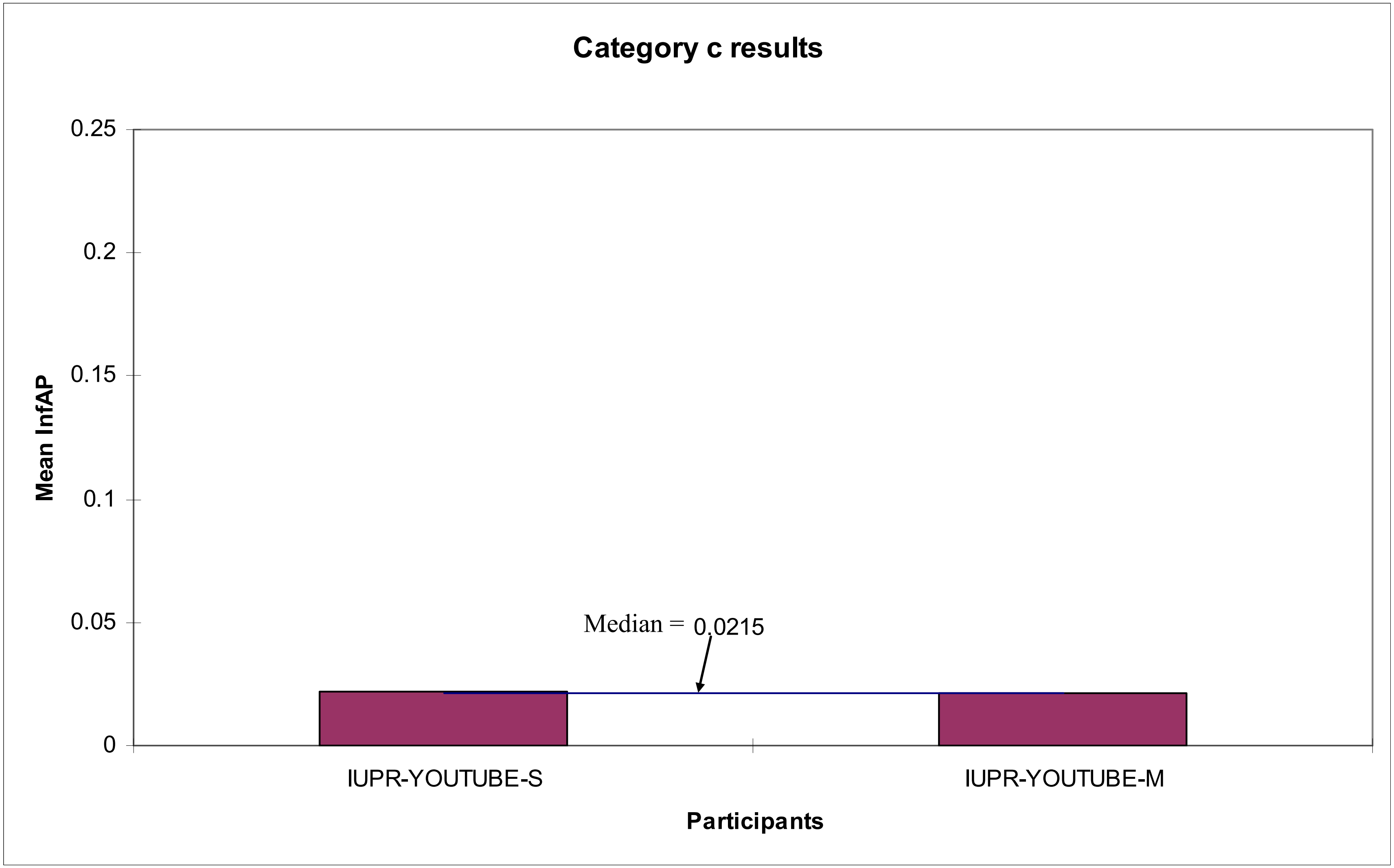



First year where B runs are better than A runs

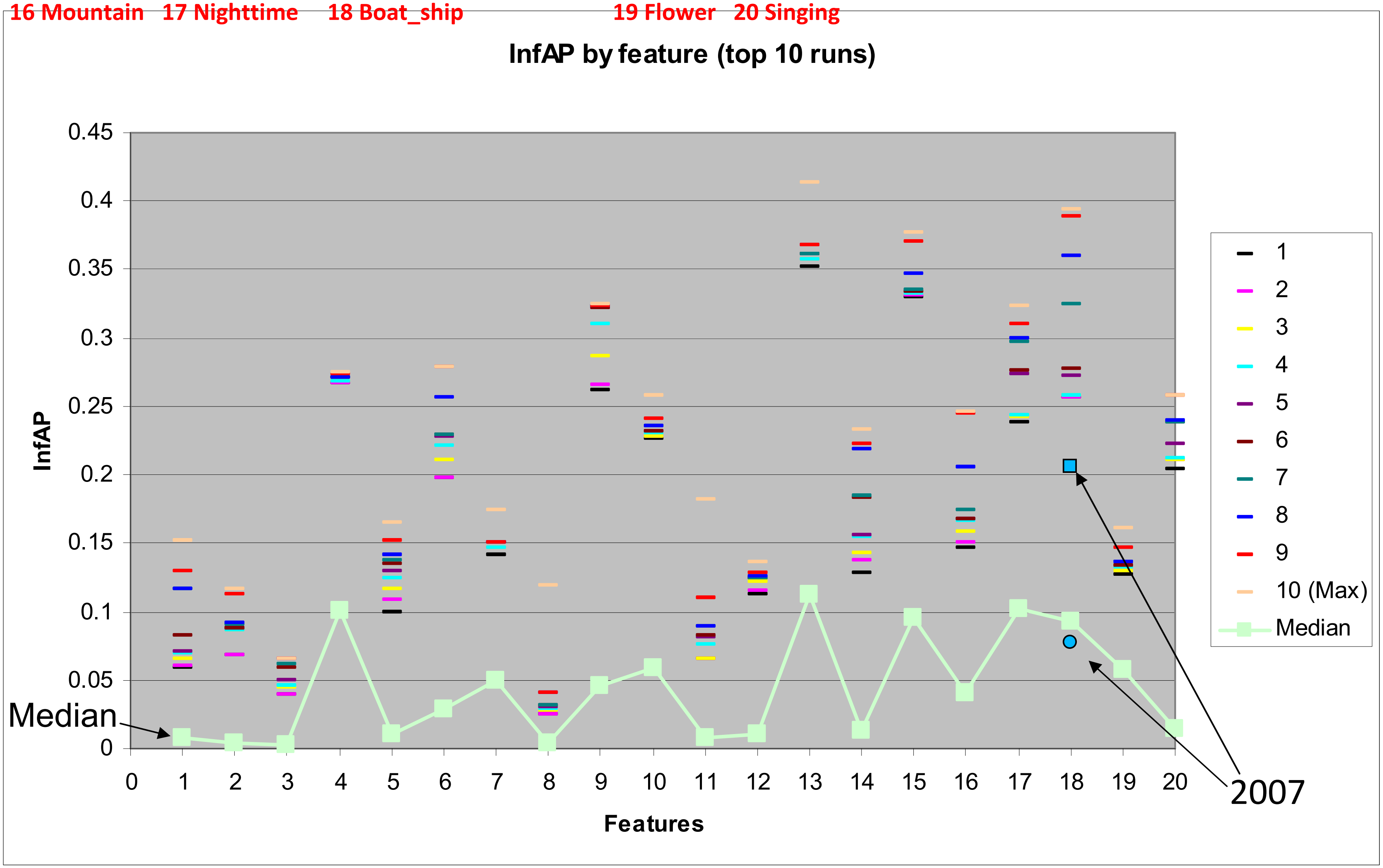



First year where C runs are on par with A runs

Data from e.g. Flickr, Youtube, Peekaboom


Category c results
0.25
0.2
0.15
0.1
0.05
0
Mean InfAP
Median = 0.0215
IUPR-YOUTUBE-S
IUPR-YOUTUBE-M
Participants

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




Which, if any, differences are significant, i.e. not due to chance?

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

## Run name (mean infAP)

- CU_2_face_base_2 (0.167)
- CU_4_fuse_baseline_4 (0.165)
- CU_5_local_global_5 (0.162)
- CU_6_local_only_6 (0.157)
- CityUHK2_2 (0.156)
- CityUHK1_1 (0.155)
- REGIM1_1 (0.140)
- PKU-ICST-HLFE-2_2 (0.138)
- PKU-ICST-HLFE-4_4 (0.137)
- PKU-ICST-HLFE-1_1 (0.134)

- CU_2_face_base_2
  - CU_6_local_only_6
    - PKU-ICST-HLFE-1_1
    - PKU-ICST-HLFE-2_2
    - PKU-ICST-HLFE-4_4
- CU_4_fuse_baseline_4
  - CU_6_local_only_6
    - PKU-ICST-HLFE-1_1
    - PKU-ICST-HLFE-2_2
    - PKU-ICST-HLFE-4_4
- CU_5_local_global_5
  - CU_6_local_only_6
    - PKU-ICST-HLFE-1_1
    - PKU-ICST-HLFE-2_2
    - PKU-ICST-HLFE-4_4
- CityUHK2_2
  - PKU-ICST-HLFE-1_1
  - PKU-ICST-HLFE-4_4
- CityUHK1_1
  - PKU-ICST-HLFE-1_1

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

Run name (mean infAP)

- NII-3-tv05-svm-bl_3 (0.037)
- CMU.voting_6 (0.020)
- CMU.mostrel_gen_rbf_2 (0.020)
- CMU.mostrel_gen_linear_1 (0.019)
- CMU.Best_CV_5 (0.019)
- CMU.mostrel_fea_linear_3 (0.018)
- CMU.mostrel_fea_rbf_4 (0.018)
- NII-6-tv05-knn-bl_6 (0.014)
- VITALAS.CERTH.ITI_5 (0.009)

- NII-3-tv05-svm-bl_3
  - CMU.voting_6
    - cMU.mostrel_fea_linear_3
    - cMU.mostrel_fea_rbf_4
    - VITALAS.CERTH.ITI_5
  - CMU.Best_CV_5
    - VITALAS.CERTH.ITI_5
  - cMU.mostrel_gen_rbf_2
    - cMU.mostrel_fea_linear_3
    - VITALAS.CERTH.ITI_5
  - cMU.mostrel_gen_linear_1
  - NII-6-tv05-knn-bl_6

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

Run name (mean infAP)

- UvA.Scary_1 (0.194)
- UvA.Ginger_4 (0.185)
- UvA.Posh_3 (0.184)
- UvA.Sporty_2 (0.159)
- UvA.Baby_5 (0.155)
- UvA.VidiVideo_6 (0.148)
- SJTU_5 (0.071)
- BILMDG_1 (0.025)
- NHKSTRL3_3 (0.013)
- NHKSTRL1_1 (0.013)

UvA.Scary_1

- UvA.Ginger_4 , UvA.Posh_3
  - UvA.Sporty_2
    - SJTU_5
      - BILMDG_1
      - NHKSTRL3_3
      - NHKSTRL1_1
  - UvA.Baby_5
    - SJTU_5
      - BILMDG_1
      - NHKSTRL3_3
      - NHKSTRL1_1
  - UvA.VidiVideo_6
    - SJTU_5
      - BILMDG_1
      - NHKSTRL3_3
      - NHKSTRL1_1

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

Run name (mean infAP)

- CU_1_base_web_face_1 (0.166)
- CU_3_base_web_3 (0.165)
- ibm.BOR_1 (0.134)
- Ibm.BNet_2 (0.120)
- TsinghuaICRC_4 (0.103)
- OXVGG_1_1 (0.101)
- OXVGG_6_6 (0.087)
- OXVGG_2_2 (0.086)
- OXVGG_4_4 (0.085)
- OXVGG_3_3 (0.080)

- CU_1_base_web_face_1 , CU_3_base_web_3
  - Ibm.BOR_1
    - OXVGG_1_1
  - Ibm.BNet_2
    - OXVGG_6_6
      - OXVGG_3_3
    - OXVGG_4_4
    - OXVGG_2_2
      - OXVGG_3_3
    - TsinghuaICRC_4
      - OXVGG_3_3

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

Run name (mean infAP)

- A_NII-2-tv08-svm-bl_2 (0.088)
- A_NII-1-tv08+05-svm-bl_1 (0.082)
- A_NII-5-tv08-knn-bl_5 (0.037)
- a_NII-3-tv05-svm-bl_3 (0.037)
- A_NII-4-tv08+05-knn-bl_4 (0.031)
- a_NII-6-tv05-knn-bl_6 (0.014)

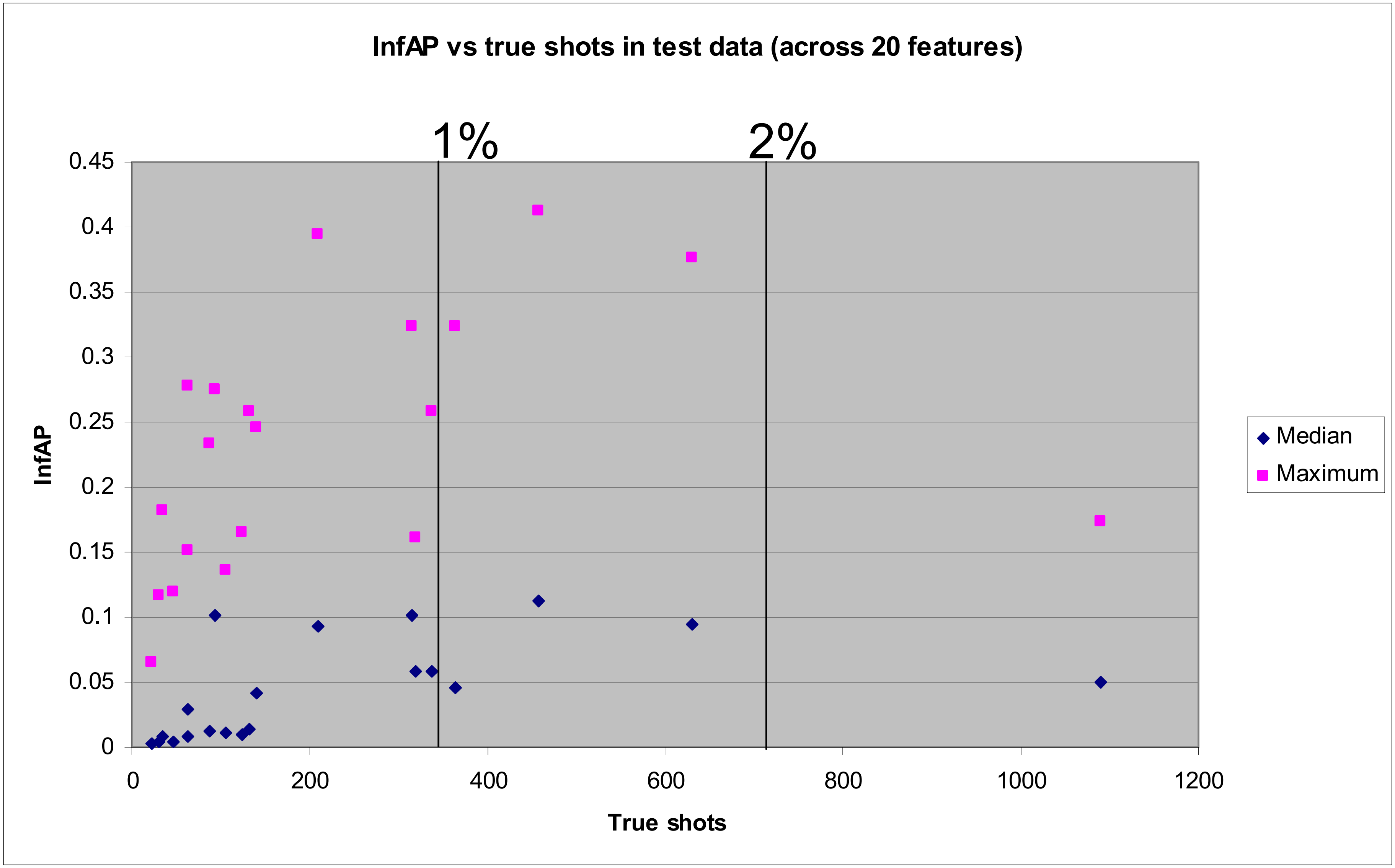
- A_NII-1-tv08+05-svm-bl_1, A_NII-2-tv08-svm-bl_2
  - A_NII-4-tv08+05-knn-bl_4
    - a_NII-6-tv05-knn-bl_6
  - A_NII-5-tv08-knn-bl_5
    - a_NII-6-tv05-knn-bl_6
  - a_NII-3-tv05-svm-bl_3
    - a_NII-6-tv05-knn-bl_6

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

Run name (mean infAP)

- A_VITALAS.CERTH.ITI_2 (0.034)
- A_VITALAS.CERTH.ITI_1 (0.029)
- A_VITALAS.CERTH.ITI_4 (0.027)
- A_VITALAS.CERTH.ITI_3 (0.025)
- a_VITALAS.CERTH.ITI_5 (0.009)

- A_VITALAS.CERTH.ITI_2
  - A_VITALAS.CERTH.ITI_3
    - a_VITALAS.CERTH.ITI_5
  - A_VITALAS.CERTH.ITI_1
    - a_VITALAS.CERTH.ITI_5
  - A_VITALAS.CERTH.ITI_4
    - a_VITALAS.CERTH.ITI_5

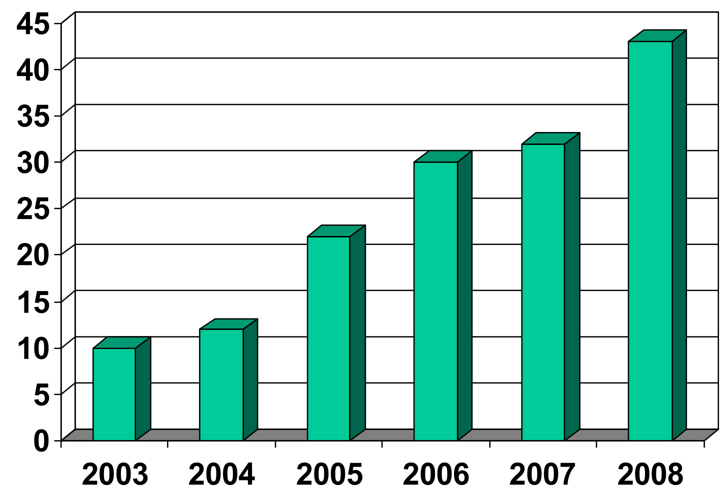


InfAP vs true shots in test data (across 20 features)
1%
2%
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
InfAP
0
200
400
600
800
1000
1200
True shots
Median
Maximum

# General observations (1)

- Participation is still increasing
- Accepted as an important building block for search

- More interest in cat B and C submissions (using e.g. web data)
- Submissions in cat B achieve best performance
- Submissions in cat C are on par with cat A



# General observations (2)

- Hardly any feature specific approaches
- Large variety in classifier architectures and choices of feature representations
- Hardware: usually a single, cpu, but several medium and larger clusters
- Nr of classifiers used for fusion ranges between 1 and >1160
- Testing times vary between 10m and 150h per feature.
- Approx. 30% of the runs do some form of temporal analysis
- Approx 50% of the runs use salient/sift points
- Compiled metadata will be made available to participants