TRECVID-2007 High-Level Feature task: Overview

Wessel Kraaij
TNO
&
Paul Over & George Awad
NIST

Outline

- Task summary
- Evaluation details
  - Inferred Average precision
  - Participants
- Evaluation results
  - Results per category
  - Results per feature
  - Significance tests category A
  - Comparison with TV2006
- Global Observations
  - Site summaries
  - Preliminary metadata analysis
    - Looking at efficiency data
- Issues

High-level feature task description

- Goal: Build benchmark collection for visual concept detection methods
- Secondary goals:
  - encourage generic (scalable) methods for detector development
  - feature-indexing could help search/browsing
- Participants submitted runs for all 39 LSCOM-lite features
- TRECVID 2007 video data
  - Netherlands Institute for Sound and Vision (~**100 hours** of news magazine, science news, news reports, documentaries, educational programming and archival video in MPEG-1).
  - 50 hours for development.
  - 50 hours for test.
  - TRECVID 2005 & TRECVID 2006 annotated data.
- NIST evaluated 20 features from the 39 using a 50% random sample of the submission pools (Inferred AP)

High-level feature 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.

TV2006 vs TV2007 dataset

	TV2006	TV2007
Dataset length (hours)	~158	~100
Number of shots	79,484	18,142
Average shot length	7 sec	20 sec
Number of unique program titles	11	47

HLF became even more challenging for machine learning

- Small imbalanced training collection
- Large variation in examples
- Noisy Annotations
- Decisions to be made:
  - find suitable representations
  - find optimal fusion strategies

- TV2007:
  - Lower scores:
    - new genres
    - less redundancy in the collection (no commercials, few “easy” weather and sports shots),
    - collection is much more heterogeneous,
    - b/w clips
    - smaller development set (# shots)

20 LSCOM-lite features evaluated (-22: corporate leader, +33: boat/ship)

1 sports

3 weather

5 office

6 meeting

10 desert

12 mountain

17 waterscape/ waterfront

23 police security

24 military personnel

26 animal

27 computer tv screen

28 us flag

29 airplane

30 car

32 truck

33 boat/ship

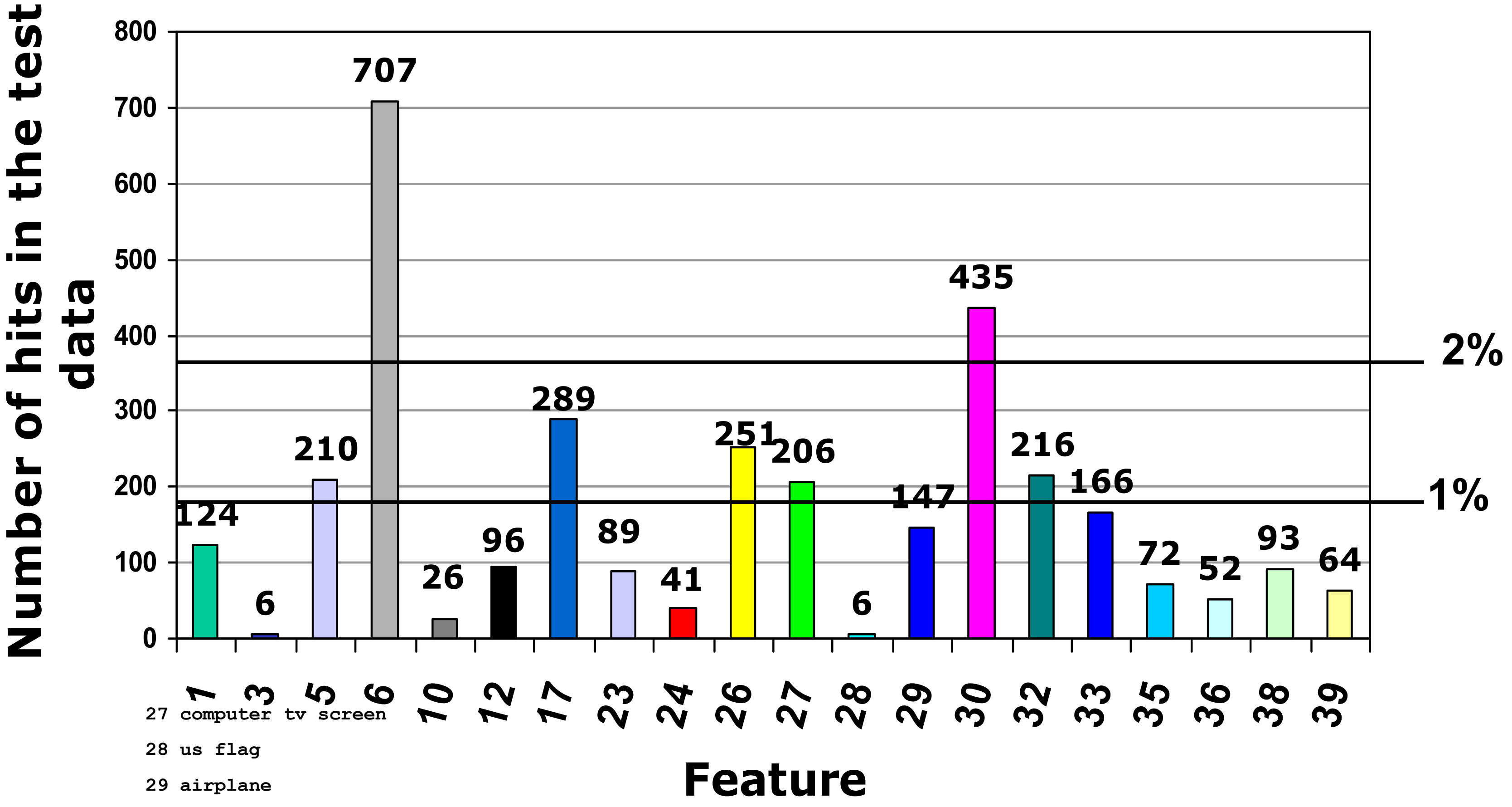
35 people marching

36 explosion fire

38 maps

39 charts

Frequency of hits varies by feature (tv7)



1 sports

3 weather

5 office

6 meeting

10 desert

12 mountain

17 waterscape/ waterfront

23 police security

24 military personnel

26 animal

27 computer tv screen

28 us flag

29 airplane

30 car

32 truck

33 boat/ship

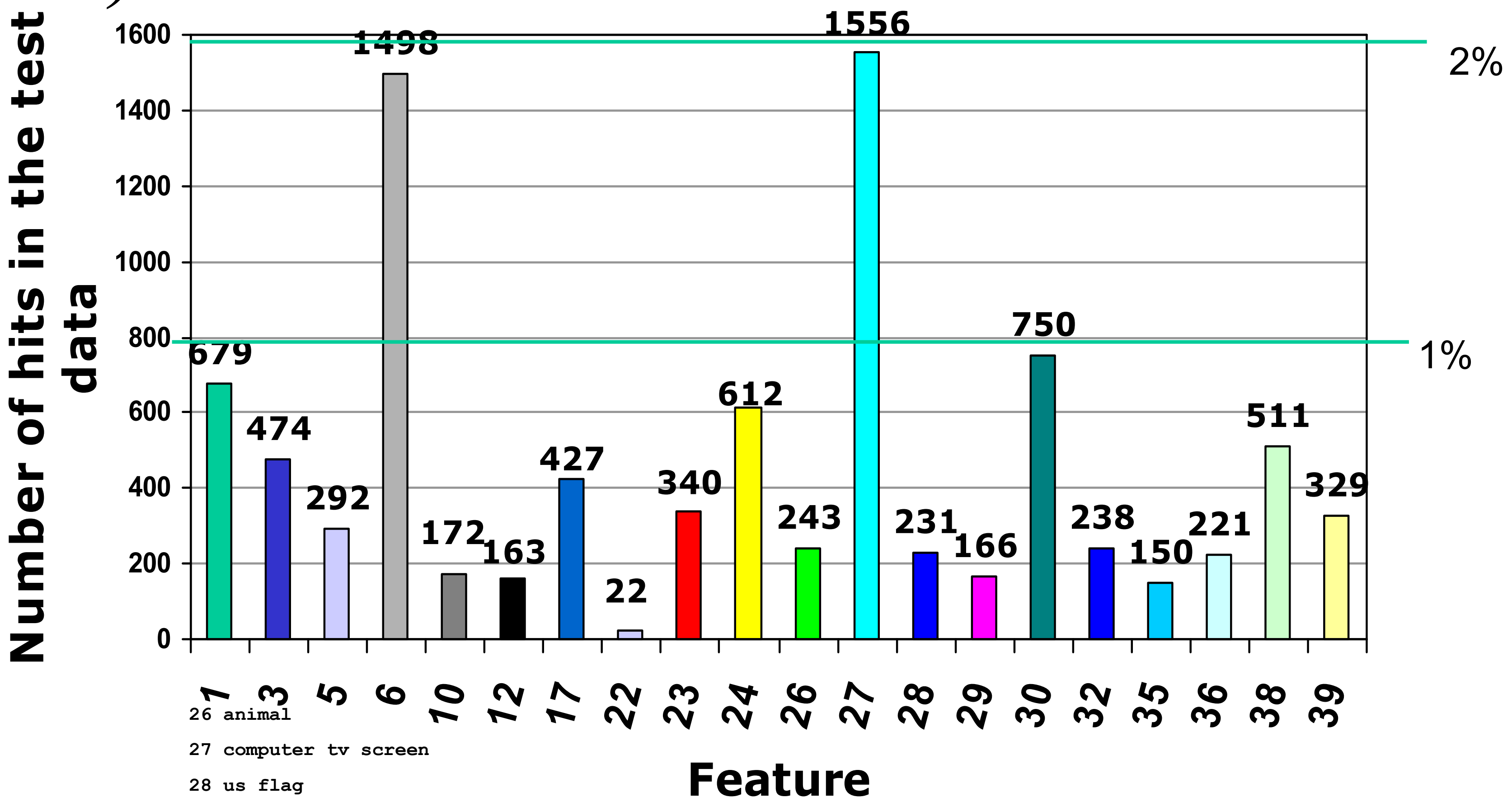
35 people marching

36 explosion fire

38 maps

39 charts

Frequency of hits varies by feature (tv6)



1 sports

3 weather

5 office

6 meeting

10 desert

12 mountain

17 waterscape/ waterfront

22 corporate leader

23 police security

24 military personnel

26 animal

27 computer tv screen

28 us flag

29 airplane

30 car

32 truck

35 people marching

36 explosion fire

38 maps

39 charts

This year, only very few true shots were uniquely found.

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 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.

2007: Inferred average precision (infAP)

- Submissions for each of 20 features were pooled down to about average 154 items (so that each feature pool contained ~ 6500 shots)
  - varying pool depth per feature
- A 50% random sample of each pool was then judged:
- 66,293 total judgements (~ 50 hr of video)
- Judgement 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

2007: 32/54 Participants (2006: 30/54, 2005: 22/42, 2004: 12/33)

Bilkent University	**	FE	SE	**
Brno University of Technology	SB	FE	**	SU
City University of Hong Kong (CityU)	--	FE	SE	SU
Columbia University	--	FE	**	SU
COST292 Team	SB	FE	SE	SU
École Nationale Supérieure des Télécommunications / TSI	--	FE	--	--
Fudan University	--	FE	SE	–
Helsinki University of Technology	**	FE	SE	SU
IBM T. J. Watson Research Center	**	FE	SE	**
Institut EURECOM	--	FE	--	SU
JOANNEUM RESEARCH Forschungsgesellschaft mbH	**	FE	--	SU
KDDI R&D Labs, Inc./ Tokushima U. / Tokyo U.	**	FE	--	SU
K-Space	--	FE	SE	–
LIG (Laboratoire d'Informatique de Grenoble)	SB	FE	**	**
LIP6 - Laboratoire dInformatique de Paris 6	--	FE	--	SU
Microsoft Research Asia	--	FE	SE	**

** : group didn't submit any runs

-- : group didn't participate

2007: 32 Participants (continued)

Multimedia Content Analysis Group (CAS)	--	FE	--	--
Institute of Computing Technology (MCG,CAS)	--	FE	SE	**
National Institute of Informatics	--	FE	--	SU
National Taiwan University	--	FE	**	SU
Oxford University	--	FE	SE	–
Philipps University Marburg	SB	FE	**	**
Tokyo Institute of Technology	**	FE	**	**
Tsinghua University / Intel Chinese Research Center	SB	FE	SE	SU
University of Karlsruhe (TH)	SB	FE	--	--
University of Amsterdam (MediaMill Team)	--	FE	SE	--
University of California, Berkeley	--	FE	**	--
University of California, Santa Barbara	--	FE	SE	SU
University of Central Florida	--	FE	SE	**
University of Electro-Communications	--	FE	**	--
University of Iowa	**	FE	SE	--
University of Louisville	--	FE	--	--

Number of runs of each training type

Tr-Type	2007	2006	2005	2004	2003
A	146 (89.5%)	86 (68.8%)	79 (71.8%)	45 (54.2%)	22 (36.7%)
B	7 (4.3%)	32 (25.6%)	24 (21.8%)	27 (32.5%)	20 (33.3%)
C	6 (3.7%)	7 (5.6%)	7 (6.3%)	11 (13.3%)	18 (30.0%)
a	4 (2.5%)	N/A	N/A	N/A	N/A
b	0	N/A	N/A	N/A	N/A
c	0	N/A	N/A	N/A	N/A
Total runs	163	125	110	83	60

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.

# runs using (common) annotation resource (out of 110 runs)

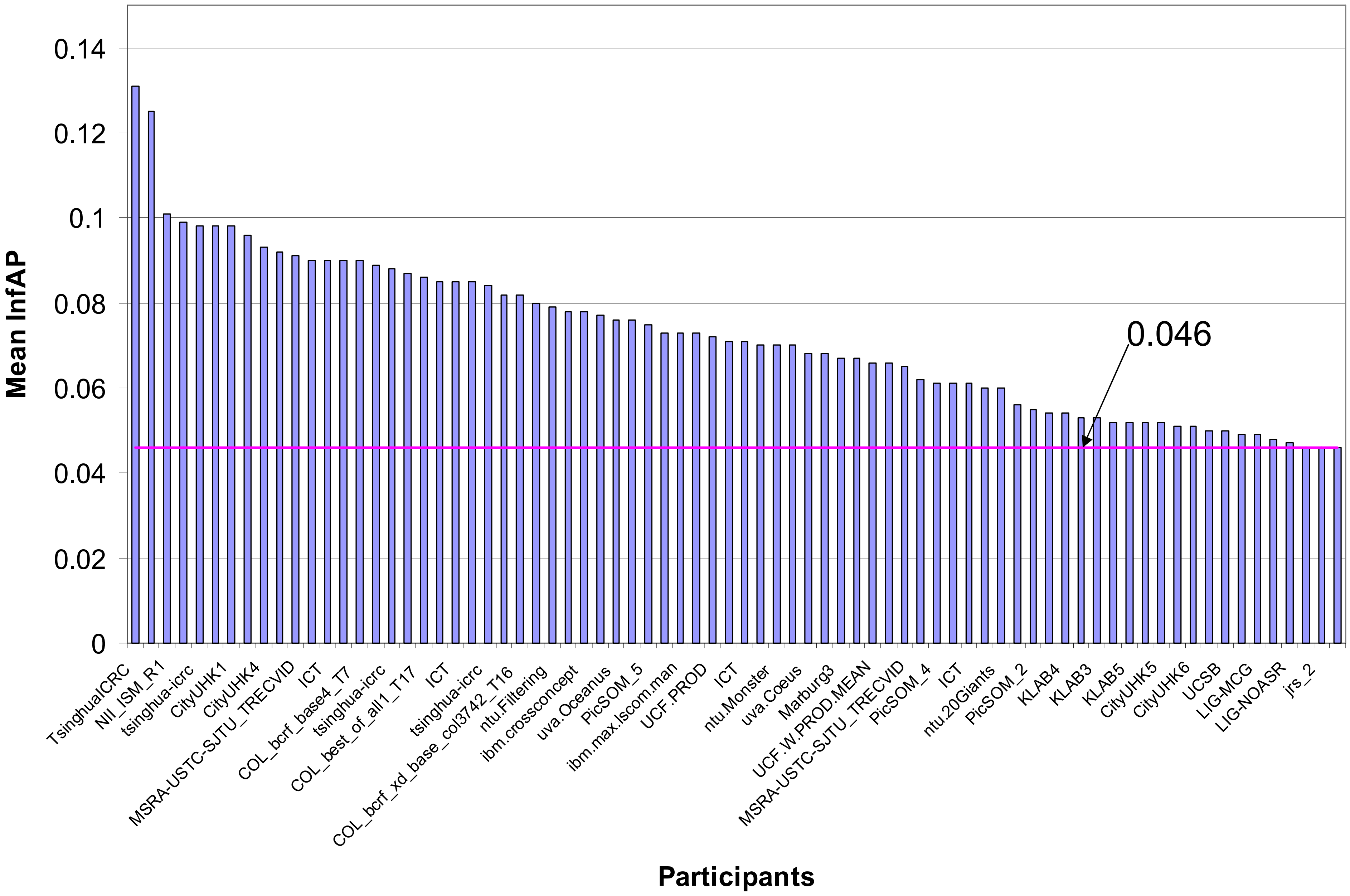
- CAS: 65
- LIG: 69
- TV2005: 17
- TV2003: 4
- MediaMill: 5
- LSCOM: 11
- Labelme: 3
- Top 10 runs only use a combination of LIG/CAS/Labelme

True shots contributed uniquely by team for each feature

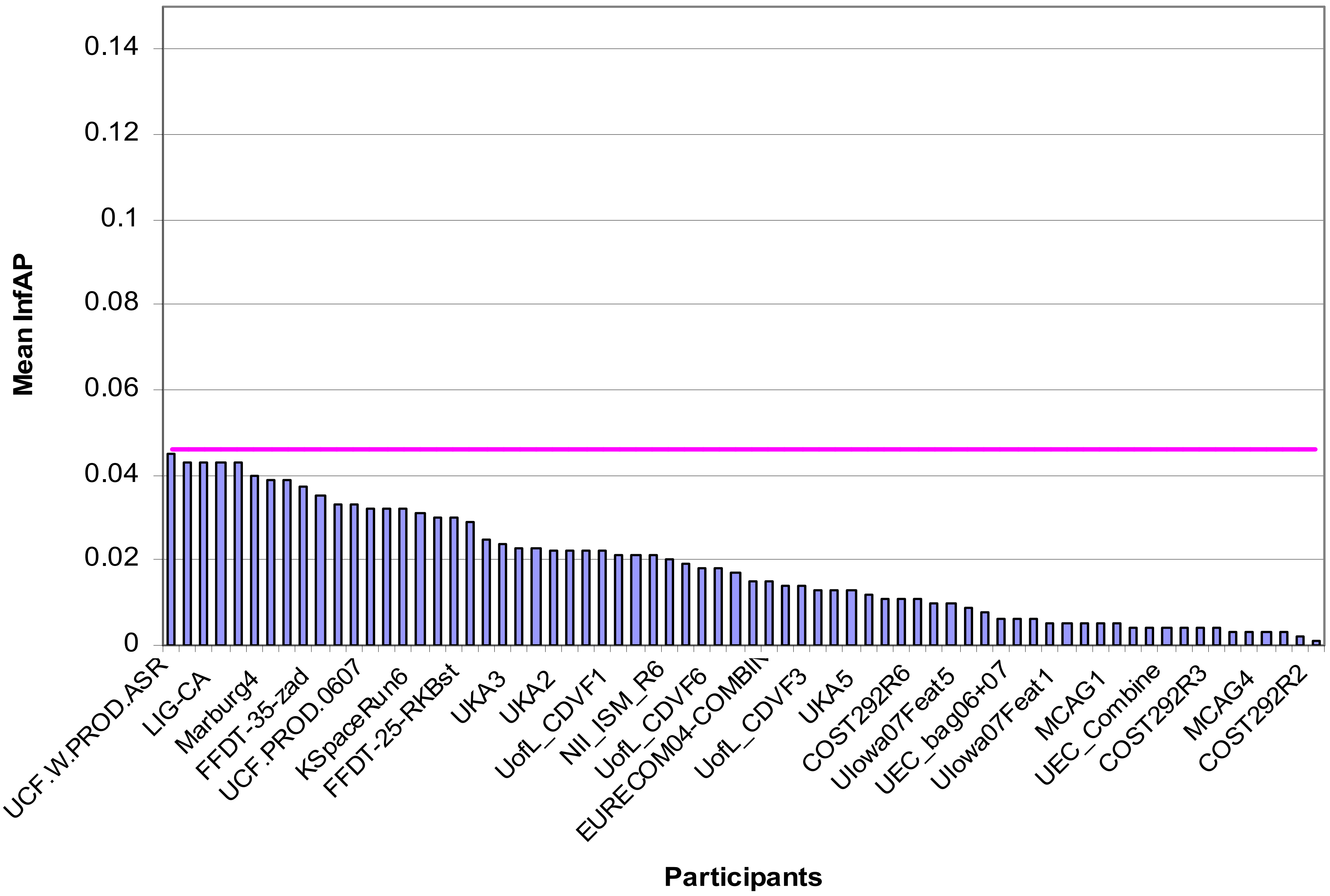
- UEC
  - Feature 6 (Meeting)
- UvA
  - Feature 33 (boat or ship)

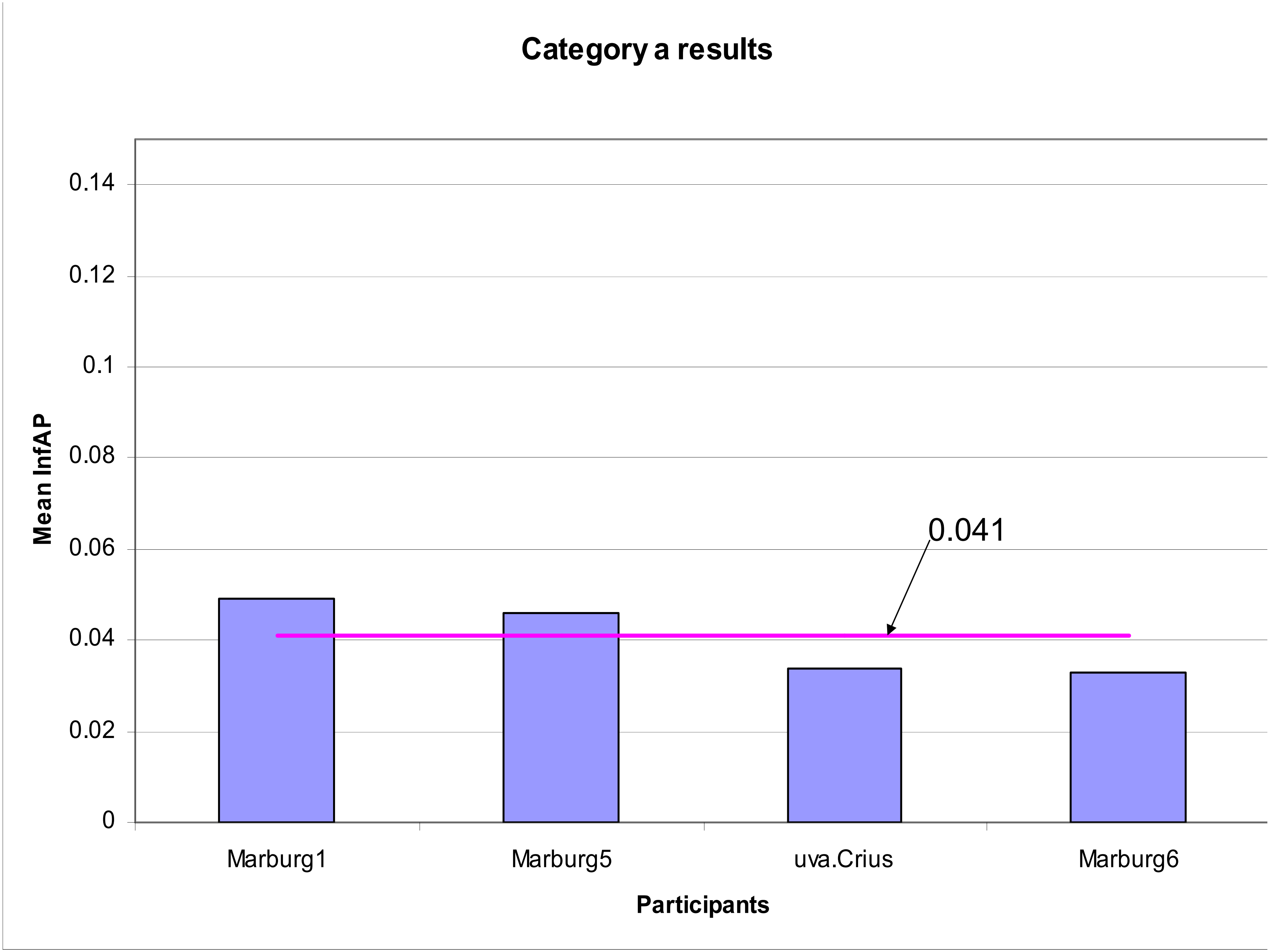
- Unlike TRECVID 2006 where many groups found different unique true shots.

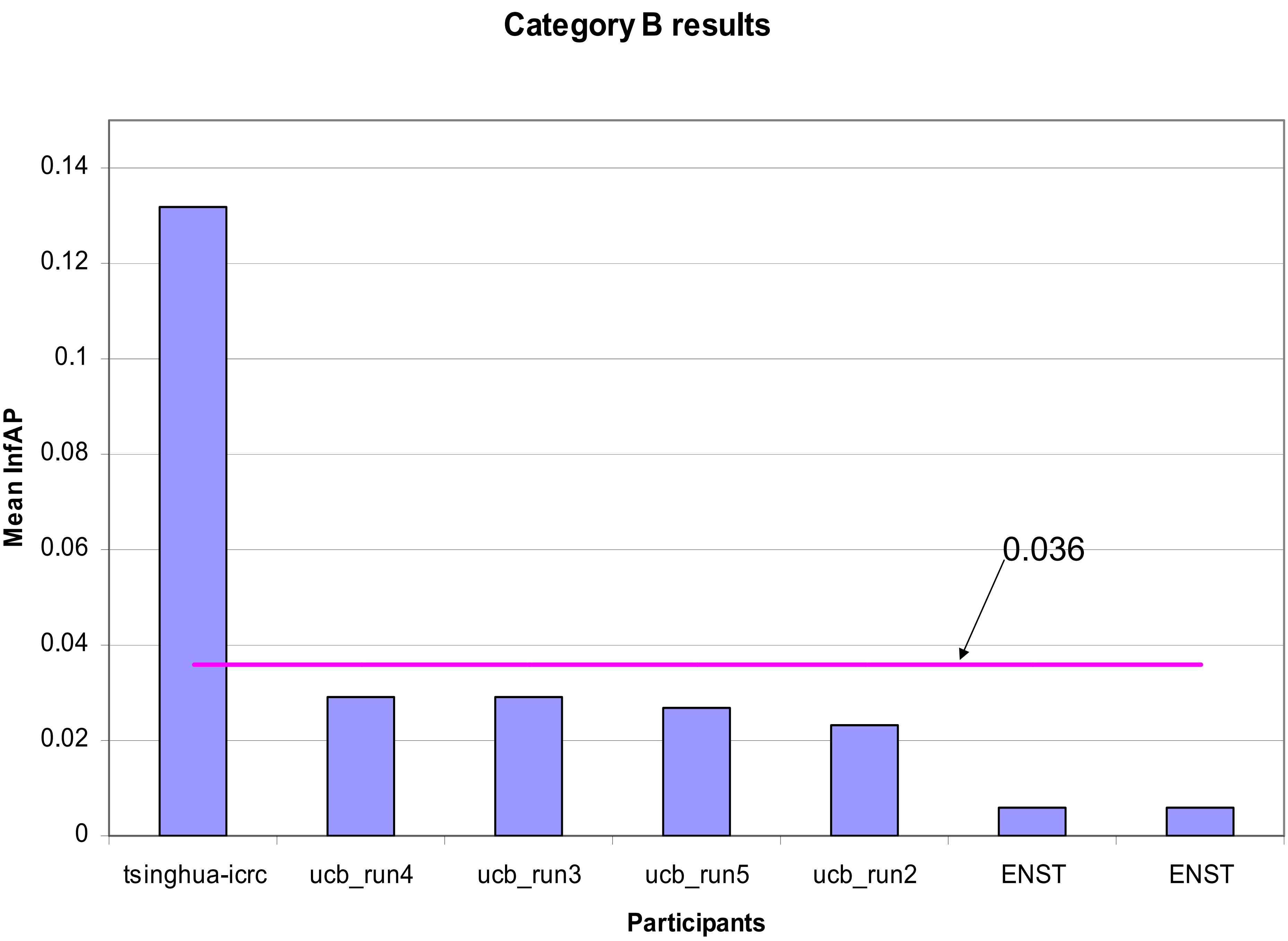
Category A results (top half)

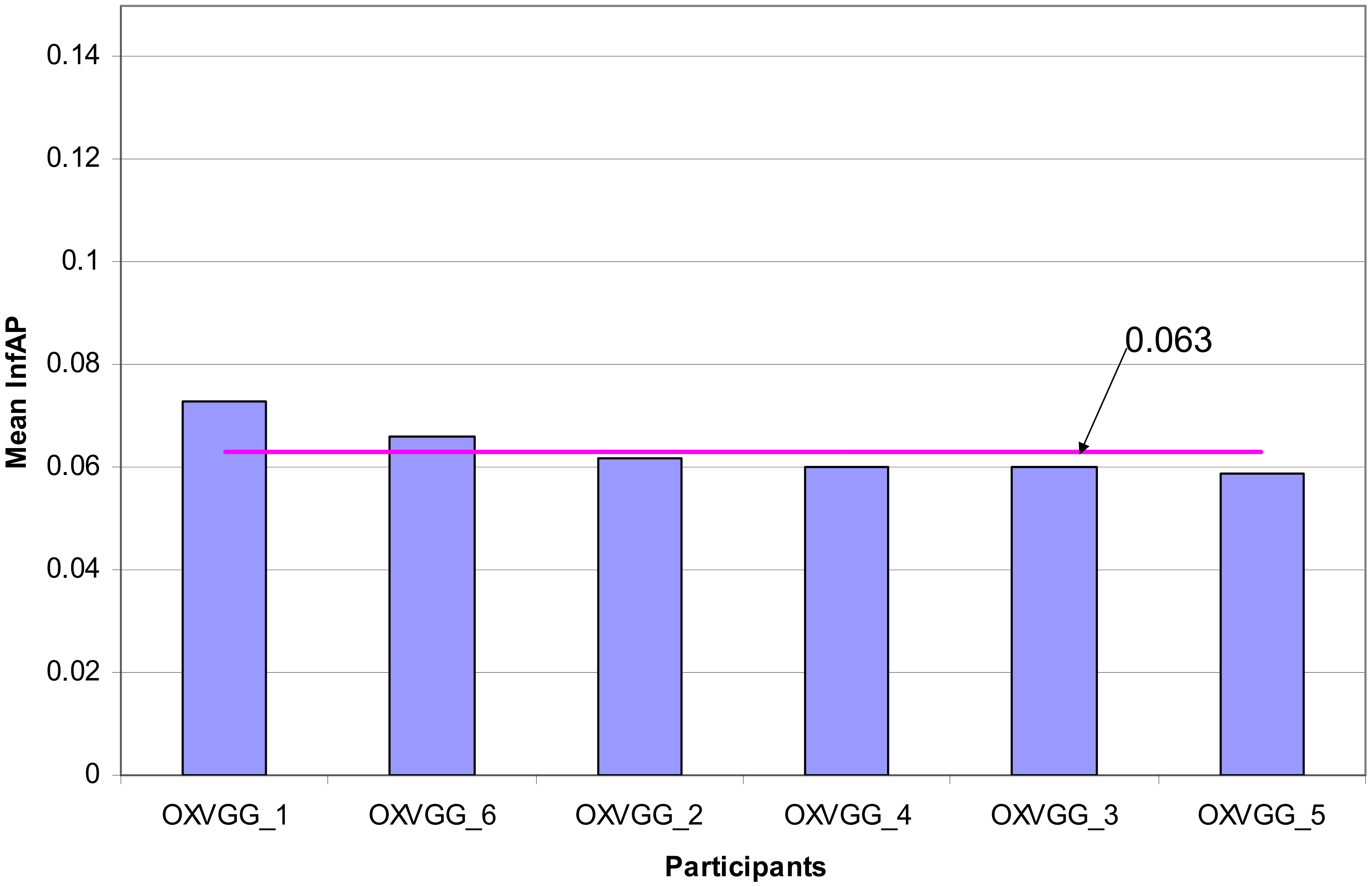


Category A results (bottom half)

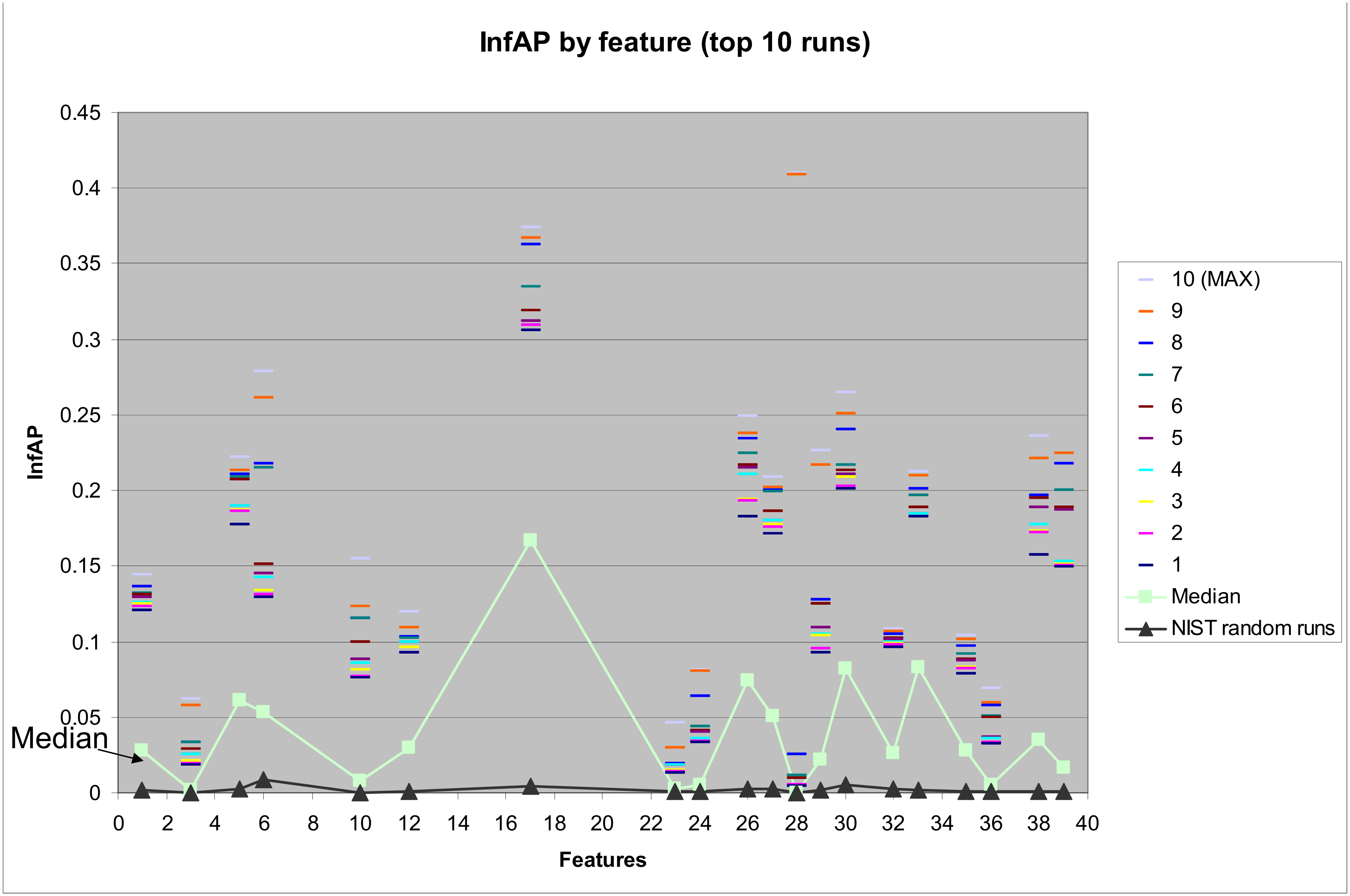



Category a results
Mean InfAP
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0.041
Marburg1
Marburg5
uva.Crius
Marburg6
Participants


Category B results
Mean InfAP
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0.036
tsinghua-icrc
ucb_run4
ucb_run3
ucb_run5
ucb_run2
ENST
ENST
Participants



Category C results



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)

- TsinghualCRC_1 (0.131)
- tsinghua-icrc_2 (0.125)
- NII_ISM_R1_1 (0.101)
- CityUHK2_2 (0.099)
- tsinghua-icrc_6 (0.098)
- CityUHK3_3 (0.098)
- CityUHK1_1 (0.098)
- MSRA-USTC-SJTU_TRECVID_1(0.096)
- CityUHK4_4 (0.093)
- MSRA-USTC-SJTU_TRECVID_2 (0.092)

- TsinghualCRC_1
  - tsinghua-icrc_2
    - CityUHK2
    - tsinghua-icrc_6
    - CityUHK3_3
    - CityUHK1_1
      - CityUHK4_4
    - MSRA-USTC-SJTU_TRECVID_1
      - MSRA-USTC-SJTU_TRECVID_2

The influence of tv7 specific training data

- a_uva.Crius_6 (0.034) baseline tv2005
- A_uva.Iapetus_3 (0.050) baseline tv2007 (+47%)

- a_Marburg1_4 (0.049) baseline tv2005
- A_Marburg2_3 (0.070) baseline tv2007 (+43%)

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

Run name (mean infAP)

- Marburg1_4 (0.049)
- Marburg5_6 (0.046)
- uva.Crius_6 (0.034)
- Marburg6_5 (0.033)

- Marburg1_4
  - Marburg6_5
  - Uva.Crius_6

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

Run name (mean infAP)

- A_Marburg2_3 (0.070)
- A_Marburg3_2 (0.067)
- a_Marburg1_4 (0.049)
- a_Marburg5_6 (0.046)
- A_Marburg4_1 (0.039)
- a_Marburg6_5 (0.033)

- A_Marburg2_3
  - a_Marburg5_6
  - a_Marburg1_4
    - A_Marburg4_1
    - a_Marburg6_5
- A_Marburg3_2
  - A_Marburg4_1
  - a_Marburg5_6
  - a_Marburg6_5

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

Run name (mean infAP)

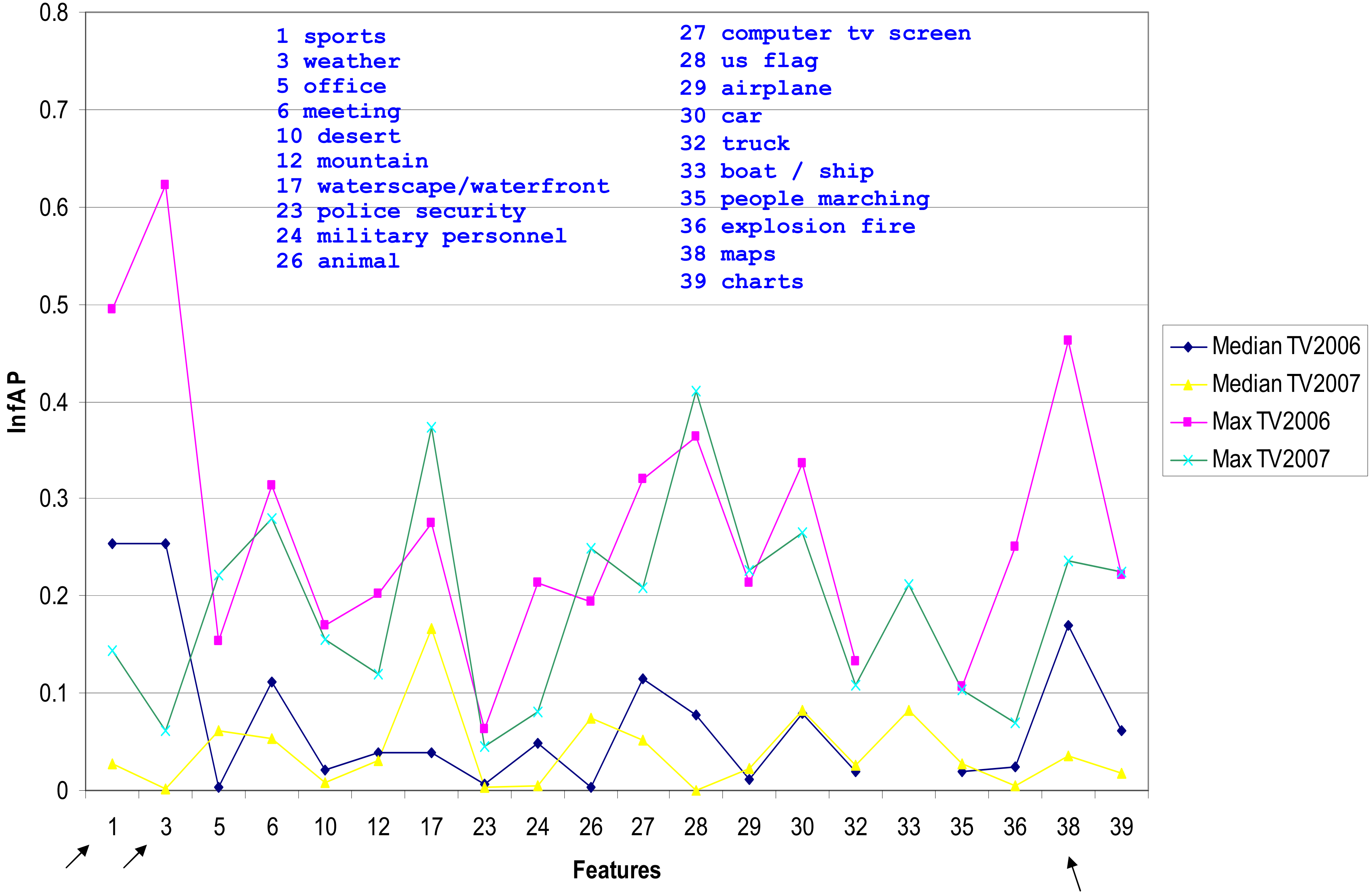
- A_uva.Hyperion_2 (0.085)
- A_uva.Oceanus_1 (0.076)
- A_uva.Coeus_4 (0.068)
- A_uva.Iapetus_3 (0.050)
- a_uva.Crius_6 (0.034)
- A_uva.Kronos_5 (0.011)

- A_uva.Hyperion_2
  - A_uva.Oceanus_1
    - A_uva.Coeus_4
      - A_uva.Iapetus_3
        - a_uva.Crius_6
          - A_uva.Kronos_5

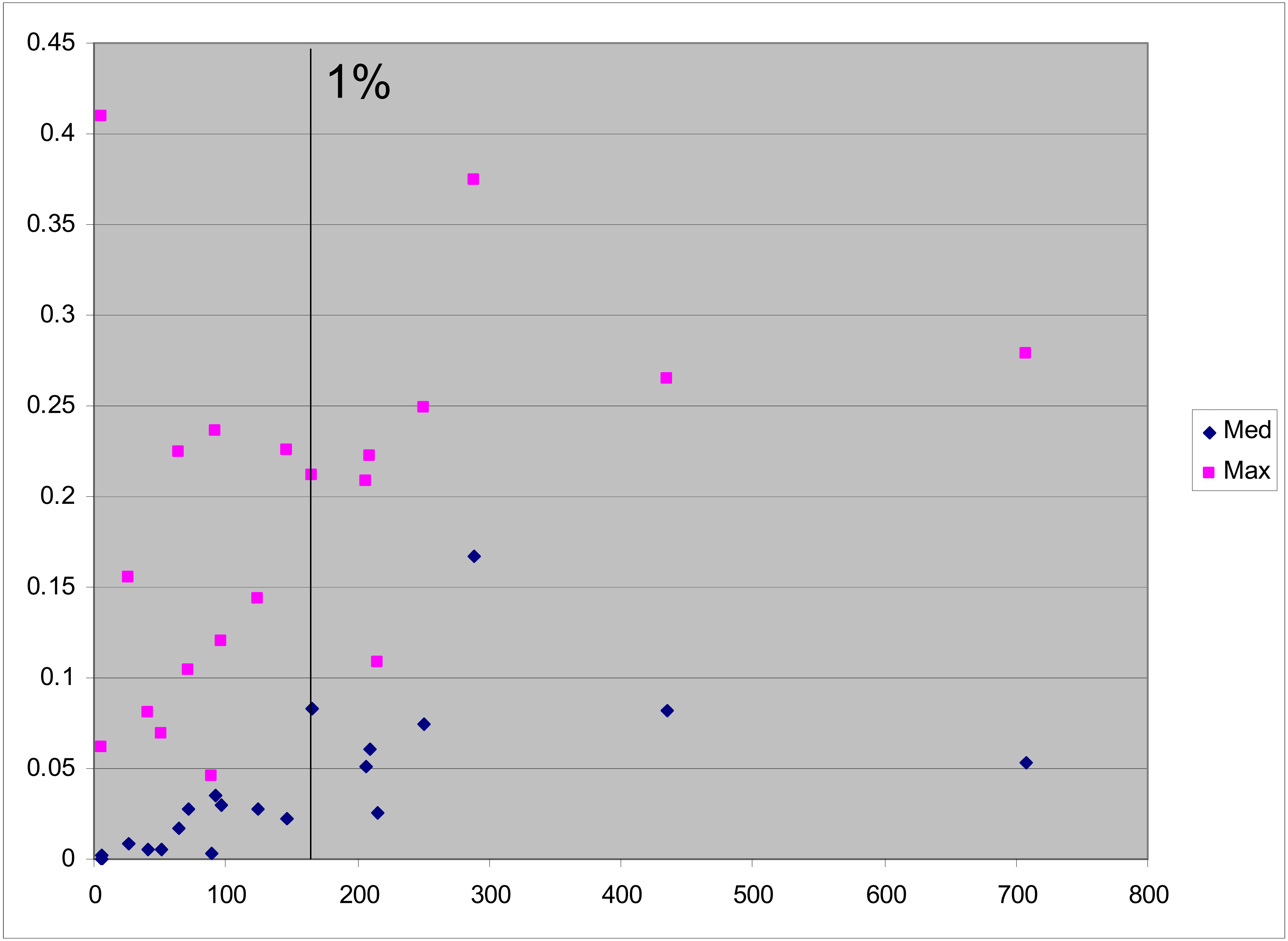
What is the best system for each feature?

Feature	System (InfAP)
1 sports	A_tsinghua-icrc_6 (0.144)
3 weather	A_MSRA-USTC-SJTU_TRECVID_6 (0.062)
5 office	A_CityUHK3_3 (0.222)
6 meeting	A_PicSOM_6_1 (0.279)
10 desert	B_tsinghua-icrc_5 (0.155)
12 mountain	C_OXVGG_4_4 (0.12)
17 waterscape/waterfront	B_tsinghua-icrc_5 (0.374)
23 police security	A_ICT_3 (0.046)
24 military personnel	B_tsinghua-icrc_5 (0.081)
26 animal	A_CityUHK2_2 (0.249)
27 computer tv screen	A_TsinghuaICRC_1 (0.209)
28 us flag	A_NII_ISM_R2_2 (0.41)
29 airplane	A_ibm.max.hog.text.max_3 (0.226)
30 car	A_TsinghuaICRC_1 (0.265)
32 truck	A_ibm.max.hog.text.max_3 (0.108)
33 boat/ship	A_CityUHK4_4 (0.212)
35 people marching	A_TsinghuaICRC_1 (0.104)
36 explosion fire	A_tsinghua-icrc_6 (0.069)
38 maps	A_TsinghuaICRC_1 (0.236)
39 charts	A_MSRA-USTC-SJTU_TRECVID_2 (0.225) A_MSRA-USTC-SJTU_TRECVID_3 (0.225)

TV2006 vs TV2007



infAP vs. # true shots in test data



Site summaries (1)

Order: reception of metadata description	comparisor
MSRA-USTC-SJTU (Microsoft Research Asia, Univ. of Science and Technology of China, Shanghai Jiaotong Univ.)	
For high-level feature extraction, we investigated the benefit of unlabeled data by semi-supervised learning, and the multi-layer (ML) multi-instance (MI) relation embedded in video by MLMI kernel, as well as the correlations between concepts by correlative multi-label learning.	>>
LIG	
Nothing new, same system as last year. Just comparing various learning set : 2005 ; 2005+2007 ; 2007 ; no asr ; with mt	>
CityUHK	
Our main focus is to explore the upper limit of bag-of-visual-words (BoW) approach based upon local appearance features. We study and evaluate several factors which could impact the performance of BoW. By considering these important factors, we show that a local feature only system already yields top performance (MAP= 0.0935).	>>
JOANNEUM RESEARCH	
Various visual features: color, texture, edges, visual activity, camera motion, faces. Early vs. late fusion. Applying score correction by concept correlation (co-occurences).	>
tsinghua-icrc	
We try a novel approach, Multi-Label Multi-Feature learning (MLMF learning) to learn a joint-concept distribution on the regional level as an intermediate representation. Besides, we improve our Video diver indexing system by designing new features, comparing learning algorithms and exploring novel fusion algorithms. The two baselines of Yingying and Huanhuan are designed for comparing different learning algorithms. The run Beibei is a floating search for fusion. In the run Jingjing, we used SFFS to select best low-level features for each topic.In the run NiNi, we tried simulated annealing and PMSRA fusion approaches. In the run Olympic2008, we combine all these efforts.	>>

Site summaries (2)

National Taiwan University (NTU)

To optimize the efficiency, we extended LIBSVM to cut down the required training time.
We reused existent classifiers to boost detection accuracy by using late aggregation.
To exploit contextual relationship and temporal dependency, we proposed a novel post-processing framework.

>

Institute of Computing Technology (MCG-ICT-CAS)

The Average Precision Performances of the A_ICT_2 and the visual baseline A_ICT_5 show that the inferred average precision of 20 concepts benefit a lot from SIFT features. Except for 4 concepts – office, meeting, police_security, military – each concept has some boost in a different degree, especially for concepts such as desert, waterscape_waterfront, boat_ship, people-marching, explosion_fire, maps, charts.

>

Helsinki Univ. of Tech.

This year, we introduced a temporal and inter-concept co-occurrence analysis stage to our existing SOM-based density estimation method for
concept modeling. In addition, we studied the effect of optimizing the kernel width parameter for each concept separately.

>

k-space

Our major contibution this year was our run number 3 - which was a lightweight multi-modal run. We used a colour feature, texture feature, motion and audio, early fused through logistic regression and achieved decent results given the very fast training times. This emphasized for us the advantage of incorporating audio into the HLFE process.

<

University of Louisville

multi-modal context-dependent fusion of classifiers
relational fuzzy clustering and membership transformation

<

NII-ISM

There are two approaches: the first one combines several simple features such as color moments, edge orientation histogram and local binary patterns trained by SVM with RBF kernel; and the second one studies combination of global alignment (GA) kernel and penalized logistic regression machine (PLRM).

>>

Site summaries (3)

KDDI labs, Univ. of Tokushima, Tokyo Univ. of Tech.

key-frame extraction using a frame clustering method and two types of feature extractions, a color-based image retrieval method and SVM-based method, were tested. >

University of Marburg

Several experiments investigating the generalization capabilities of our system trained on broadcast news videos were conducted. We applied transductive learning to adapt the appearance models based on news videos to the sound and vision data. Furthermore the impact of seperate training for color and gray-scale shots was investigated. >

IBM

Efficiency, cross domain detectors, concept fusion >

UEC

In this year, we adopted late fusion of several types of features and the spatial pyramid method. >

MediaMill - University of Amsterdam

We extract region-based image features, on grid, keypoint, and segmentation level, which we combine with various supervised learners. In addition, we explore the utility of temporal image features. A late fusion approach of all region-based analysis methods using geometric mean was our most successful run. What is more, using MediaMill Challenge and LSCOM annotations, our visual-only approach generalizes to a set of 572 concept detectors. >

ENST

This is the first test of a 2-level GMM based representation <

Site summaries (4)

Uni Karlsruhe

Our system is a combination of best performing systems from the previous evaluation. Mainly, it is the fusion of IBM and Berkeley system. In our first participation in TRECVID we just wanted to build a baseline system. <

LIP6

A new way to extract features from keyframes, and shot reference files has been introduced. The sampling and the construction of the forests of Fuzzy Decision Trees (FDT) are new too and they have been introduced this year. Moreover, new t-norms have been used to classify test shots by means of a FDT. Results from all the FDT are aggregated to obtain a single value for a shot to have the HLF and to rank shots by means of their values. Various new techniques have been tested to optimize the obtained ranking: the RankBoost algorithm, and a weighted aggregation of the results of the FDT. <

Bilkent

i) KNN ii) bag of regions, Bayesian classifier >

COST292

The framework developed for the HLFE task comprises four systems. The first system transforms a set of low-level descriptors into the semantic space using Latent Semantic Analysis and utilizes neural networks for classification. The second system uses a Bayesian classifier trained with a “bag of regions”. The third system uses a multi-feature classifier based on SVMs and several descriptors. The fourth system uses two image classifiers based on ant colony optimization and particle swarm optimization respectively. <

Columbia University

Efficient and effective model adaptation for a new domain >

Institut Eurécom

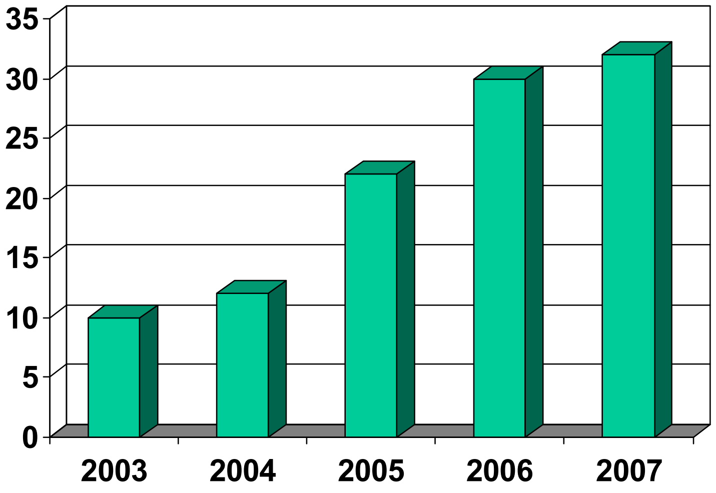
Comparison of global, region and audio based representations: global representation is strong, audio is weak <

Site summaries (5)

Fudan University	*i) Concept ontology for Bayesian inference, ii) learn a cluster of related features simultaneously*	>
University of Iowa	*comparison of color, edge and wavelet representations: no distinction found*	<
Oxford University	*Vision only, based on SIFT points. LDA based dimension reduction, combinin+B13g generic approach with feature specific techniques did help!*	>
Tokyo Institute of Technology	*SIFT points and motion features: combination yielded small gain*	<
University of Central Florida	*multiple keyframes per shot, multiple fusion strategies of visual and textual representations*	>
University of California at Santa Barbara	*Comparison of visual and audio runs. SIFT points were strong, audio weak (silent movies in testset)*	>
University of Brno	*?????? No paper, No metadata*	<
University of California at Berkeley	*???????? No paper, no metadata*	<
MCA-CAS	*???????? No paper, no metadata*	>

General observations (1)

- Participation is still increasing



- Maintained focus on cat A
- Most groups built a generic feature detector
- Top scores come from the usual suspects plus a few new groups

General observations

- Many groups did visual only runs
- Exploiting audio yielded mixed results across sites
- A few groups did experiment with alternative keyframe extraction methods
- Increasing activity on temporal analysis (9/21)
- Efficiency is an issue of active research
- Learning from unlabelled data
- some gray-scale specific approaches

Metadata collection

- Goal: provide rough summary data for
  - providing a standardized way to describe experiments
  - Enabling some meta-analysis

- Auto-annotation should be more reliable

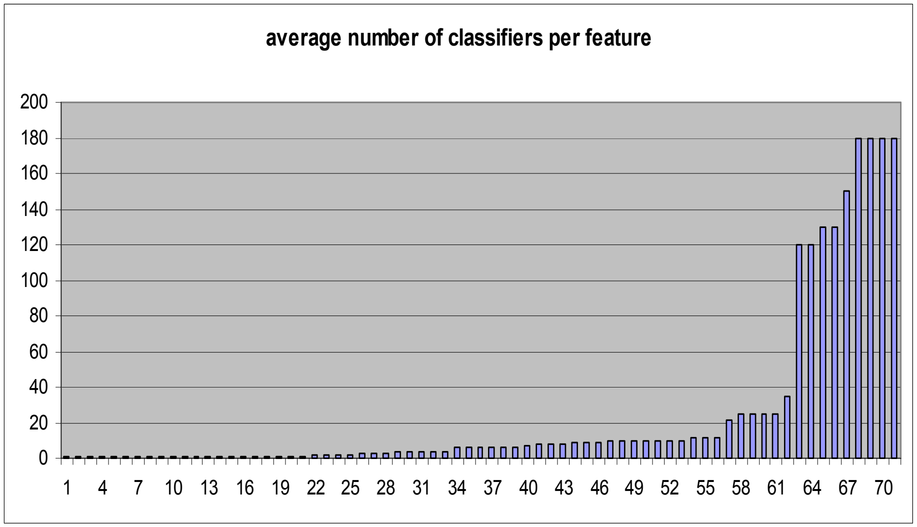
- 21 of 32 sites provided metadata on a last minute request (thanks ☺)

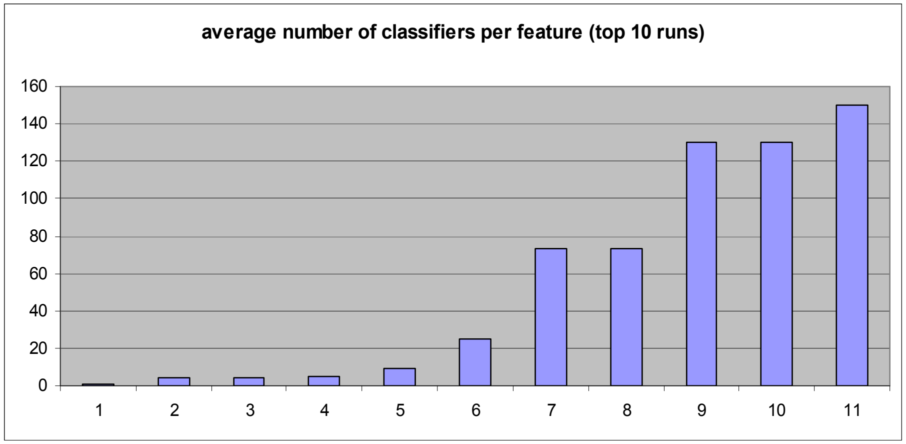
- Some sites reported that some of the data had not been captured (especially efficiency data)
- Only a preliminary analysis could be reported in this overview

Metadata collection (2)

- Standard metadata: run tag, training data category
- Keyframe selection method *(not provided this year)*
- Annotation resources (*better: labelled training data sets*)
- feature types
  - c: color, t: texture, s:shape, e:edges, a:acoustic, f:face, T: text
  - *Maybe we should make a distinction between OCR and ASR text?*
  - *Maybe we should add HLF as well (concept fusion)*
- granularity (local, region, global)
- Temporal analysis
- classifier techniques
- fusion
- Efficiency: training time, testing time, memory footprint, nr of classifiers, hardware platform
- generic vs. feature specific
- focus of site experiments *(textual and/or by highlighting)*

Classifier architecture



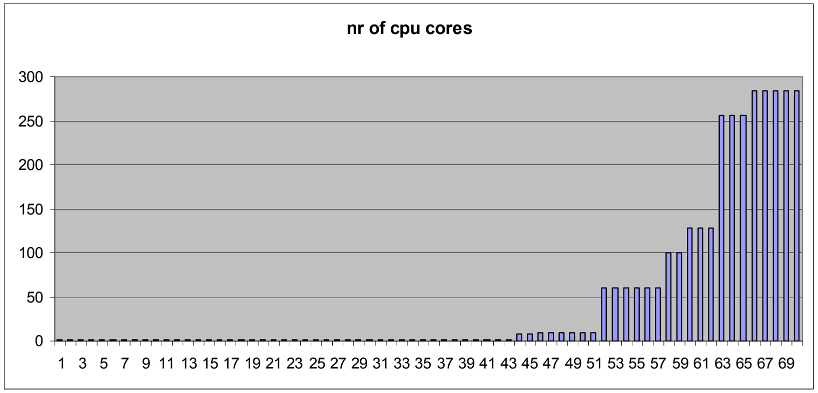


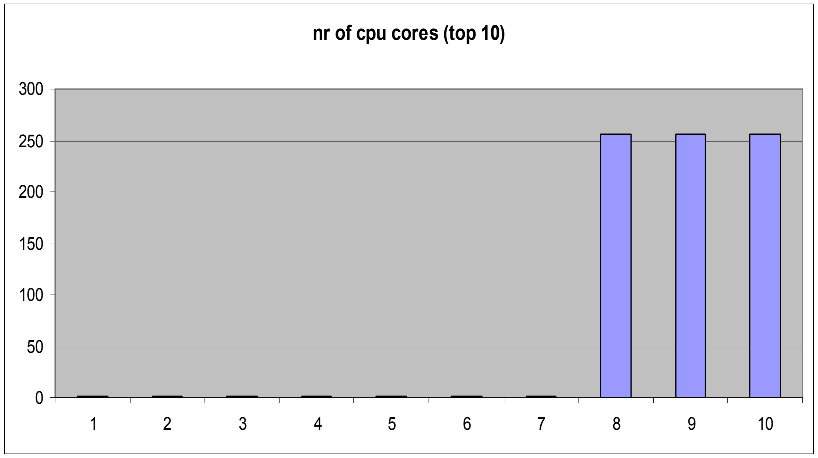
City U HK

Tsinghua
ICRC

- Only 9 out of 100 runs include feature specific techniques
- All top 10 runs have a generic architecture

Hardware Platform





City U HK, MSRA, NII

Tsinghua
ICRC

- Most groups use a single cpu
- Several groups use medium and large clusters
- Univ of Amsterdam runs are not included in the graph, but they use a large cluster (>200 nodes)

Efficiency

- Metadata definition was not entrirely clear (is feature exctraction time included?, per feature?!, wallclock or CPU time?)

- Training times reported:
  - all runs: between 00:25:00 (Tshinghua) and 25:00:00 (Tshinghua)
  - top 10: betwen 00:25:00 (Tshinghua) and 21:00:00 (MSRA)

- Testing times reported:
  - all runs: between 00:01:00 (Tshinghua) and 03:00:00 (ICT-CAS)
  - top 10: between 00:01:00 (Tshinghua) and 02:00:00 (MSRA)