TRECVID-2015 Semantic Indexing task: Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
Dakota Consulting - NIST

National Institute of Standards and Technology

Outline

- Task summary (Goals, Data, Run types, Concepts, Metrics)
- Evaluation details
    - Inferred average precision
    - Participants
- Evaluation results
    - Hits per concept
    - Results per run
    - Results per concept
    - Significance tests
- Progress task results
- Global Observations

Semantic Indexing task

- **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, searching and browsing.

- **Task:** Find shots that contain a certain concept, rank them according to confidence measure, submit the top 2000.

- Participants submitted one type of runs:
  - **Main run** Includes results for 60 concepts, from which NIST evaluated 30.

Semantic Indexing task (data)

- SIN testing dataset
  - Main test set (IACC.2.C): 200 hours, with durations between 10 seconds and 6 minutes.

- SIN development dataset
  - (IACC.1.A, IACC.1.B, IACC.1.C & IACC.1.tv10.training): 800 hours, used from 2010 – 2012 with durations between 10 seconds to just longer than 3.5 minutes.

- Total shots:
  - Development: 549,434
  - Test: IACC.2.C (113,046 shots)

- Common annotation for 346 concepts coordinated by LIG/LIF/Quaero from 2007-2013 made available.

Semantic Indexing task (Concepts)

- Selection of the 60 target concepts Were drawn from 500 concepts chosen from the TRECVID "high level features" from 2005 to 2010 to favor cross-collection experiments Plus a selection of LSCOM concepts.
- Generic-Specific relations among concepts 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).
- These concepts are expected to be useful for the content-based (instance) search task.

- Set of relations provided:
  - 427 "implies" relations, e.g. "Actor implies Person"
  - 559 "excludes" relations, e.g. "Daytime_Outdoor excludes Nighttime"

NIST
National Institute of Standards and Technology

Semantic Indexing task (training types)

- Six training types were allowed:
  - A – used only IACC training data (30 runs)
  - B – used only non-IACC training data (0 runs)
  - C – used both IACC and non-IACC TRECVID (S&V and/or Broadcast news) training data (2 runs)
  - D – used both IACC and non-IACC non-TRECVID training data(54 runs)
  - E – used only training data collected automatically using only the concepts' name and definition (0 runs)
  - F – used only training data collected automatically using a query built manually from the concepts' name and definition (0 runs)

30 Single concepts evaluated(1)

3 Airplane*
5 Anchorperson
9 Basketball*
13 Bicycling*
15 Boat_Ship*
17 Bridges*
19 Bus*
22 Car_Racing
27 Cheering*
31 Computers*
38 Dancing
41 Demonstration_Or_Protest
49 Explosion_fire
56 Government_leaders
71 Instrumental_Musician*

72 Kitchen
80 Motorcycle*
85 Office
86 Old_people
95 Press_conference
100 Running*
117 Telephones*
120 Throwing
261 Flags*
297 Hill
321 Lakes
392 Quadruped*
440 Soldiers
454 Studio_With_Anchorperson
478 Traffic

-The 14 marked with “*” are a subset of those tested in 2014

Evaluation

- The 30 evaluated single concepts were chosen after examining TRECVid 2013 60 evaluated concept scores across all runs and choosing the top 45 concepts with maximum score variation.
- Each feature assumed to be binary: absent or present for each master reference shot
- NIST sampled ranked pools and judged top results from all submissions
- Metrics: *inferred average precision per concept*
- Compared runs in terms of **mean** *inferred average precision* across the 30 concept results for main runs.

NIST
National Institute of Standards and Technology

2015: mean extended Inferred average precision (xinfAP)

- 2 pools were created for each concept and sampled as:
  - •Top pool (ranks 1-200) sampled at 100%
  - •Bottom pool (ranks 201-2000) sampled at 11.1%

30 concepts
195,500 total judgments
11,636 total hits
7489 Hits at ranks (1-100)
2970 Hits at ranks (101-200)
1177 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

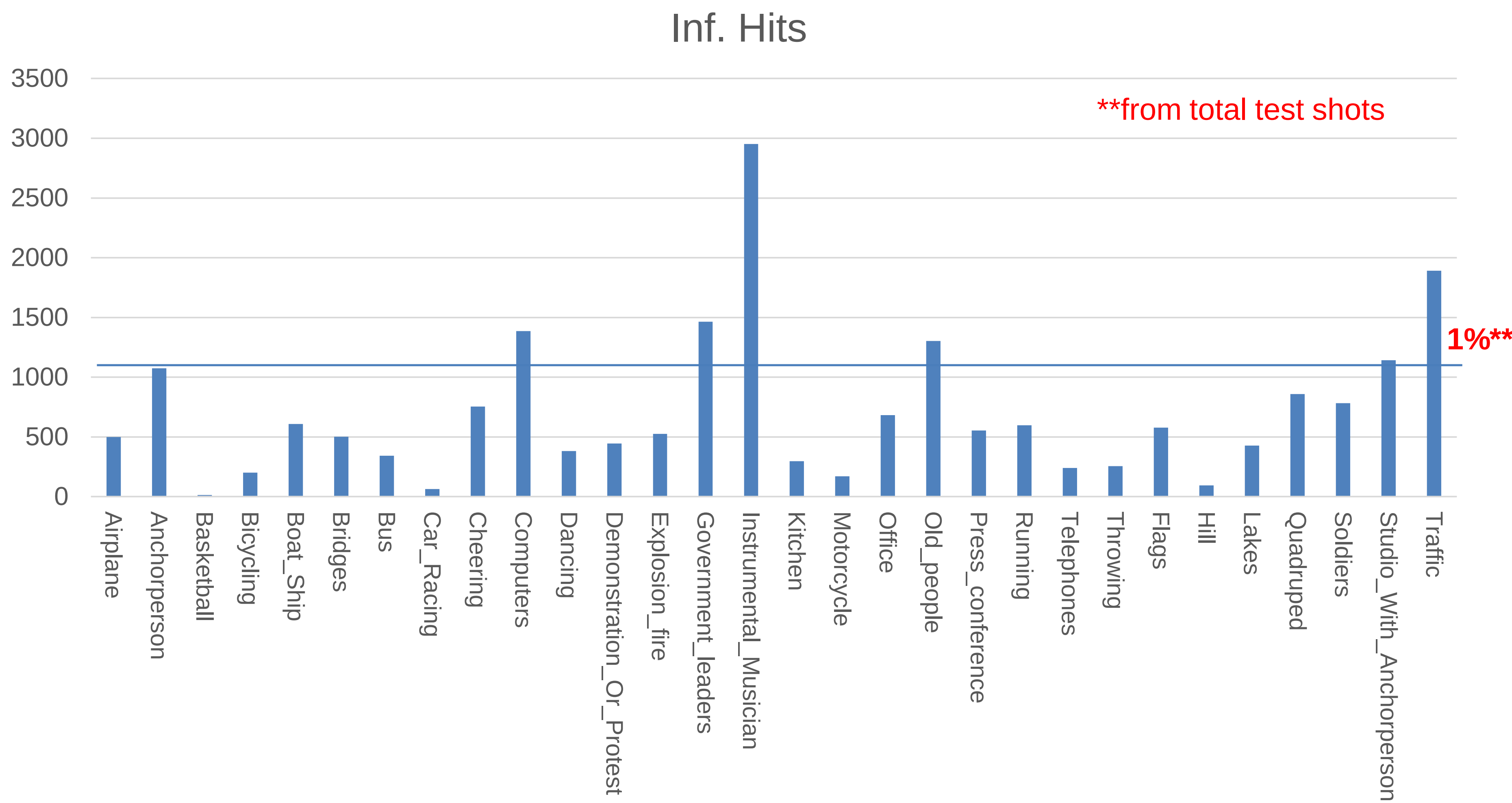
NIST
National Institute of Standards and Technology

2015 : 15 Finishers

PicSOM	Aalto U., U. of Helsinki
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas
CMU	Carnegie Mellon U.; CMU-Affiliates
Insightdcu	Dublin City Un.; U. Polytechnica Barcelona
EURECOM	EURECOM
FIU_UM	Florida International U., U. of Miami
IRIM	CEA-LIST, ETIS, EURECOM, INRIA-TEXMEX, LABRI, LIF, LIG, LIMSI-TLP, LIP6, LIRIS, LISTIC
LIG	Laboratoire d'Informatique de Grenoble
NII_Hitachi_UIT	Natl.Inst. Of Info.; Hitachi Ltd; U. of Inf. Tech.(HCM-UIT)
TokyoTech	Tokyo Institute of Technology
MediaMill	U. of Amsterdam Qualcomm
siegen_kobe_nict	U. of Siegen; Kobe U.; Natl. Inst. of Info. and Comm. Tech.
UCF_CRCV	U. of Central Florida
UEC	U. of Electro-Communications
Waseda	Waseda U.

National Institute of Standards and Technology

Inferred frequency of hits varies by concept

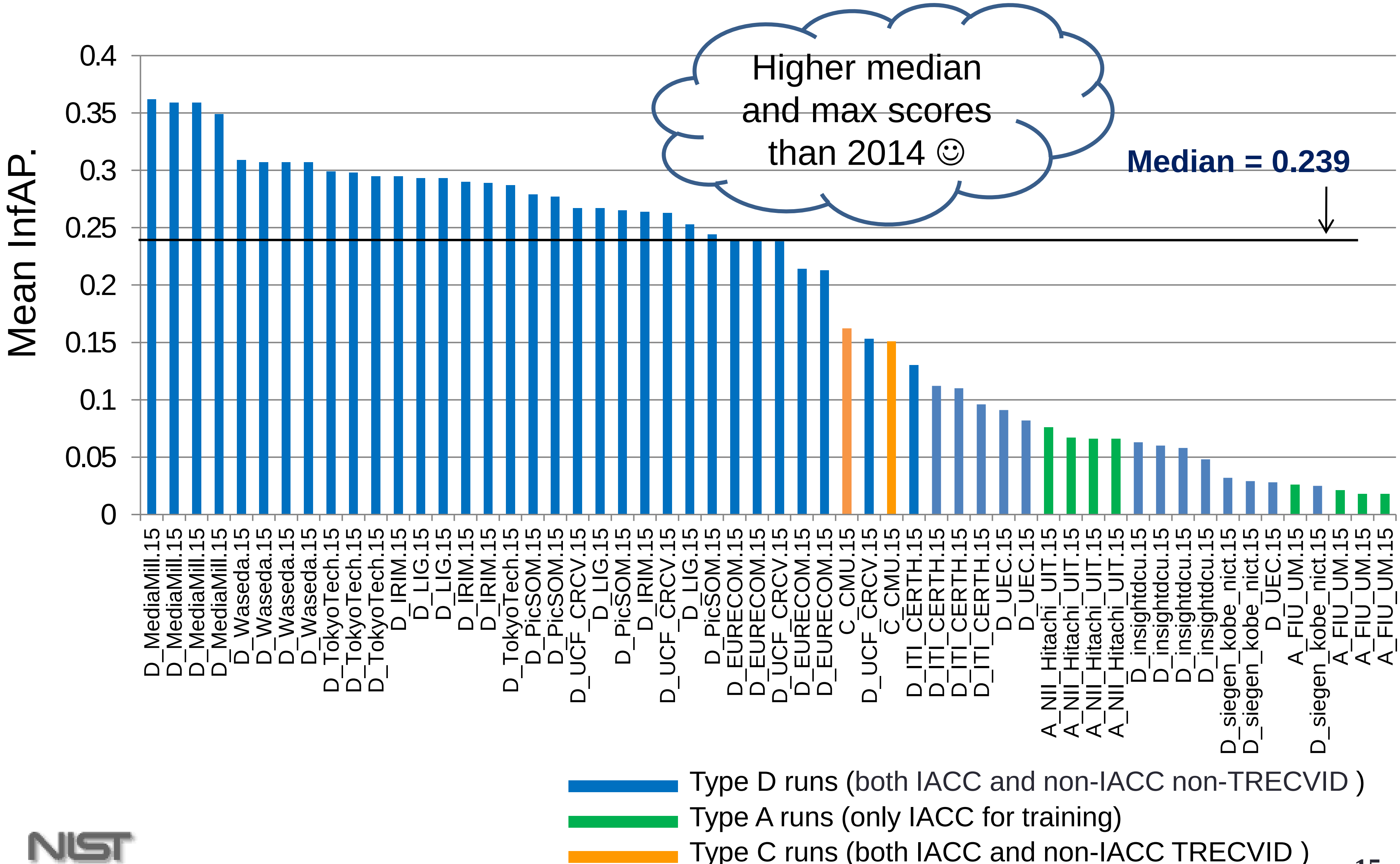


Total true shots contributed uniquely by team

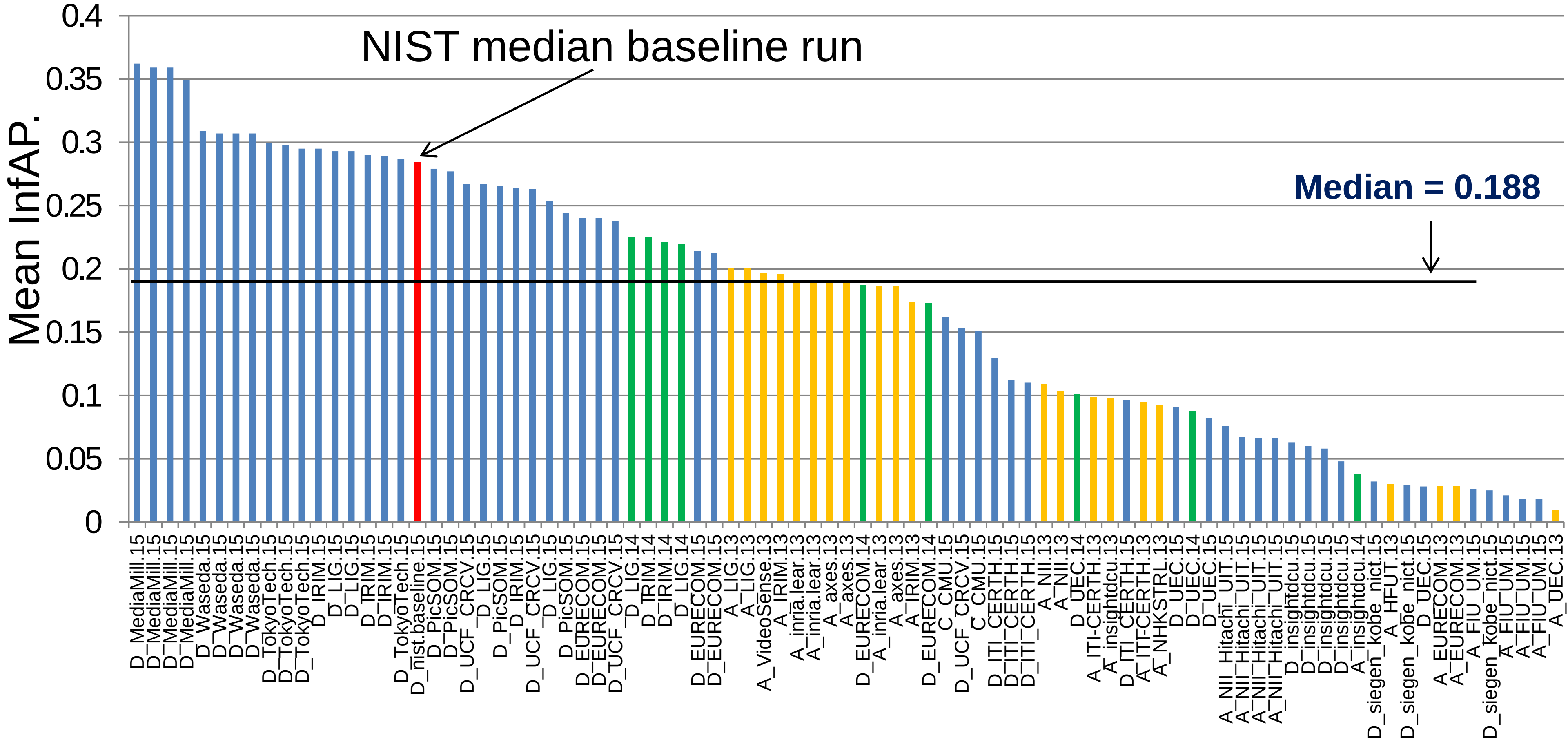
Team	No. of Shots	Team	No. of shots
Insightdcu	27	Mediamill	8
NII	19	NHKSTRL	7
UEC	17	ITI_CERTH	6
siegen_kobe_nict	13	HFUT	4
EURECOM	10	CMU	3
FIU	10	LIG	2
UCF	10	IRIM	1

Fewer unique shots compared to TV2014, TV2013 & TV2012

Main runs scores – 2015 submissions

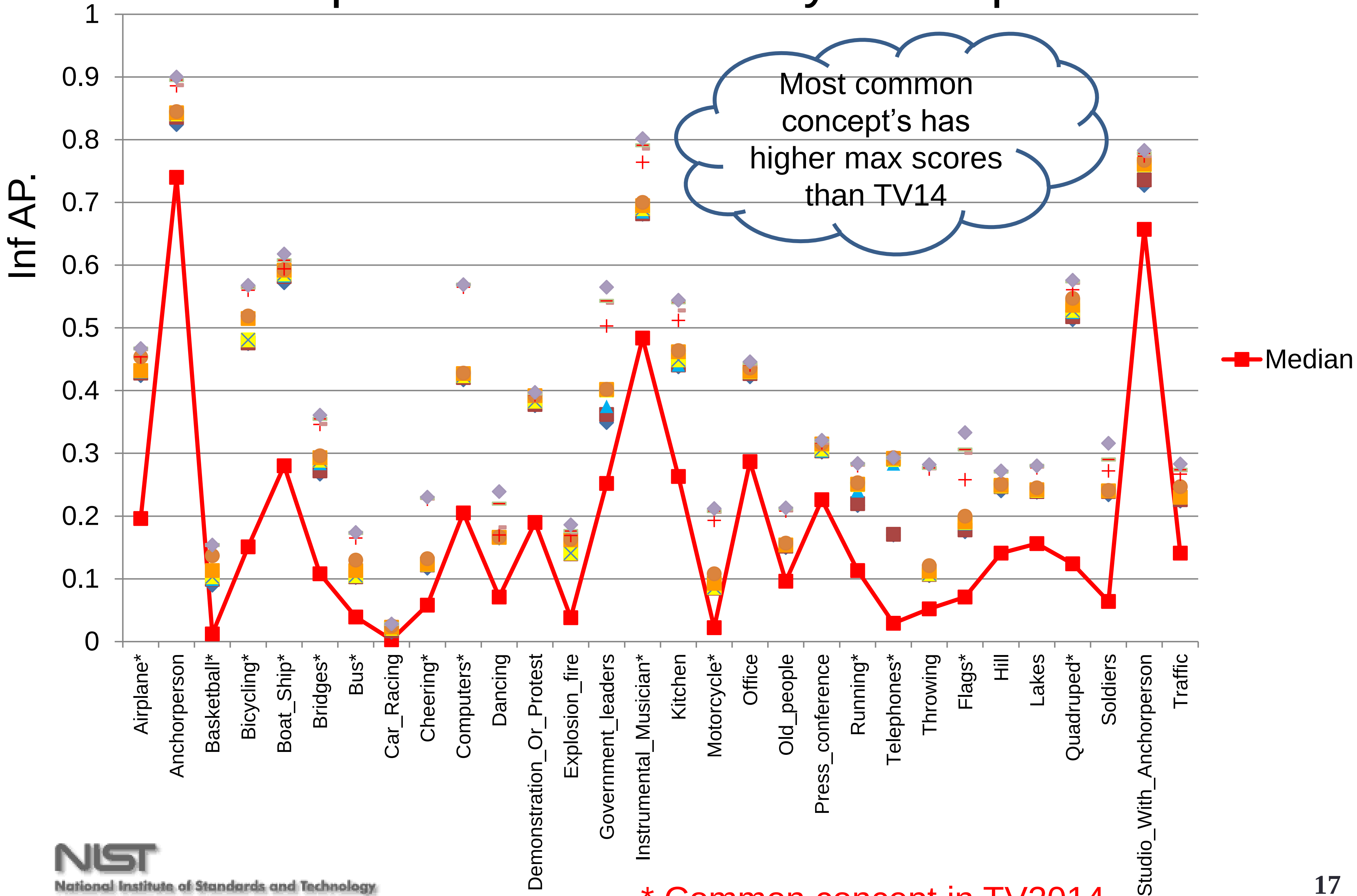


Main runs scores – Including progress



* Submitted runs in 2013 against 2015 testing data (Progress runs)
* Submitted runs in 2014 against 2015 testing data (Progress runs)

Top 10 InfAP scores by concept



* Common concept in TV2014

NIST
National Institute of Standards and Technology

Statistical significant differences among top 10 Main runs (using randomization test, p < 0.05)

•Run name	(mean infAP)
D_MediaMill.15_4	0.362
D_MediaMill.15_2	0.359
D_MediaMill.15_1	0.359
D_MediaMill.15_3	0.349
D_Waseda.15_1	0.309
D_Waseda.15_4	0.307
D_Waseda.15_3	0.307
D_Waseda.15_2	0.307
D_TokyoTech.15_1	0.299
D_TokyoTech.15_2	0.298

- D_MediaMill.15_4
  - D_MediaMill.15_3
    - D_TokyoTech.15_1
    - D_TokyoTech.15_2
    - D_Waseda.15_1
      - D_Waseda.15_3
      - D_Waseda.15_4
    - D_Waseda.15_2

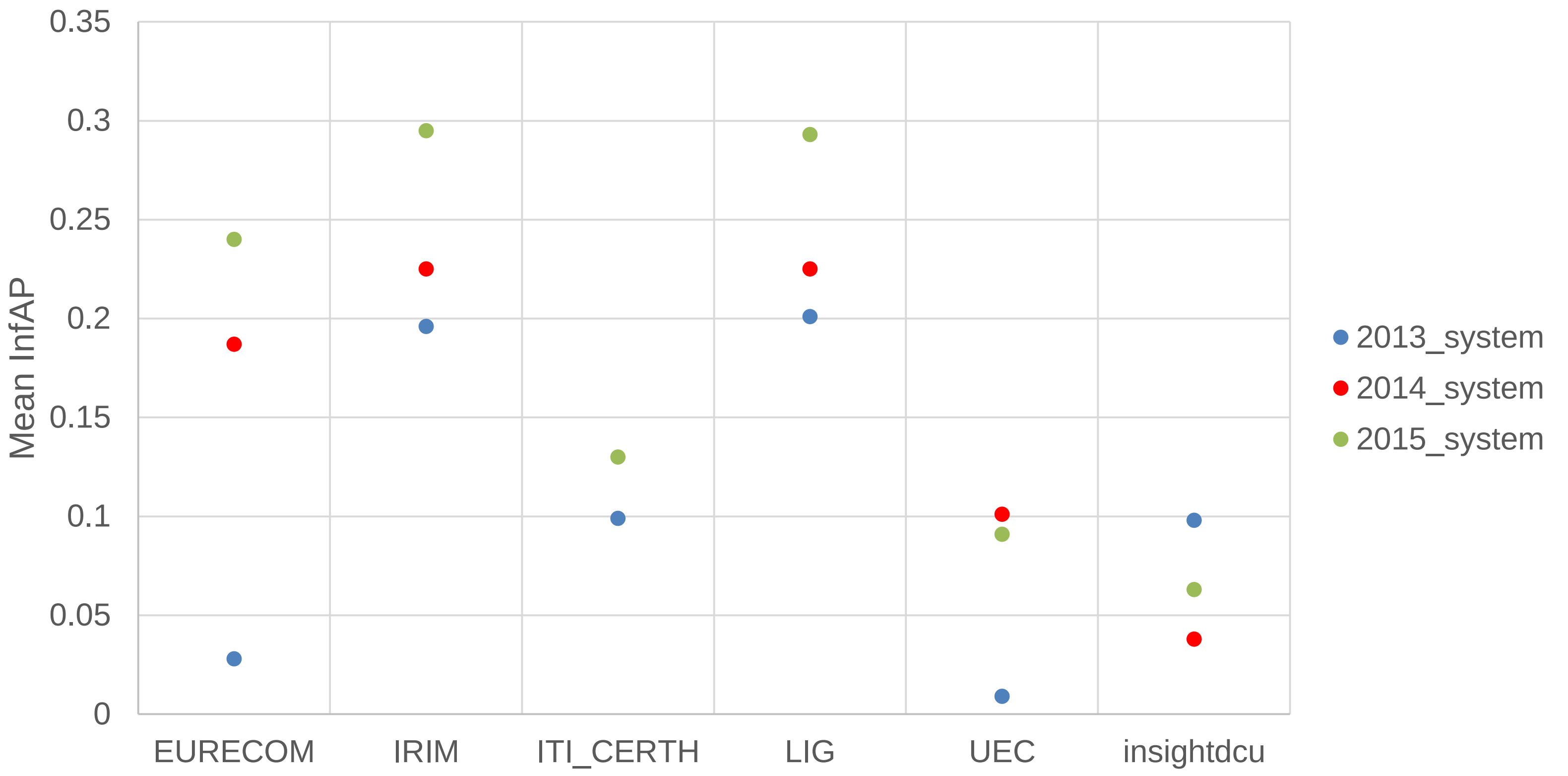
- D_MediaMill.15_1
  - D_MediaMill.15_3
    - D_Waseda.15_1
      - D_Waseda.15_3
      - D_Waseda.15_4
    - D_Waseda.15_2
    - D_TokyoTech.15_1
    - D_TokyoTech.15_2

- D_MediaMill.15_2
  - D_MediaMill.15_3
    - D_Waseda.15_1
      - D_Waseda.15_3
      - D_Waseda.15_4
    - D_Waseda.15_2
    - D_TokyoTech.15_1
    - D_TokyoTech.15_2

Progress subtask

- Measuring progress of 2013, 2014, & 2015 systems on IACC.2.C dataset.
- 2015 systems used same training data and annotations as in 2013 & 2014.
- Total 6 teams submitted progress runs against IACC.2.C dataset.

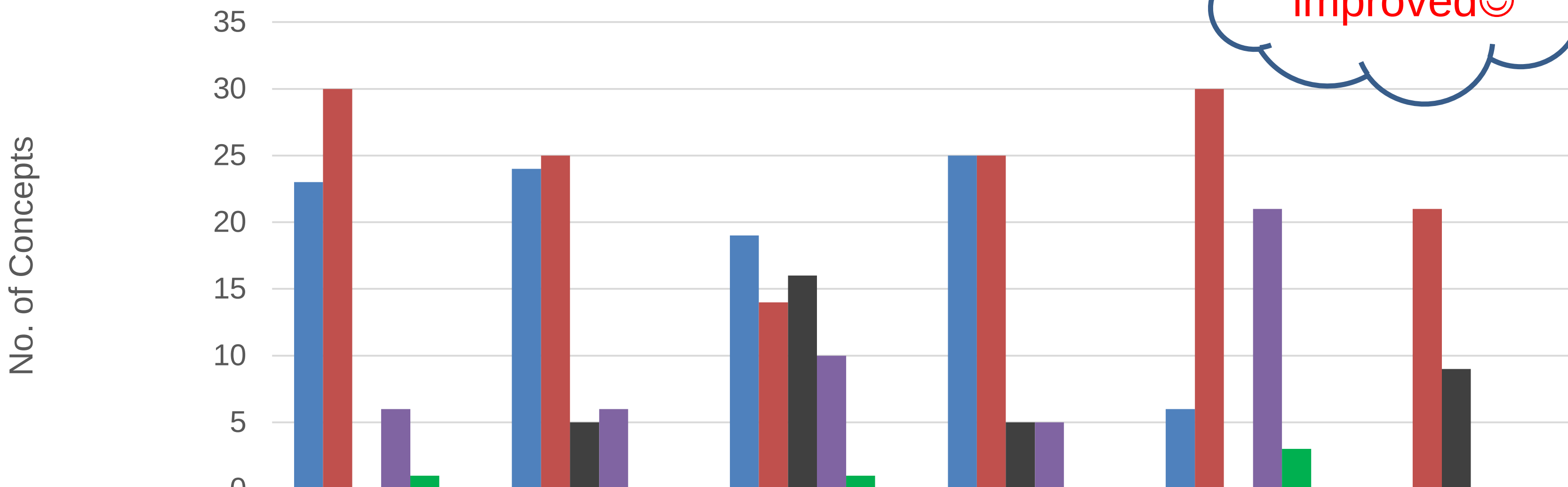
Progress subtask: Comparing best runs in 2013, 2014 & 2015 by team



Randomization tests show that 2015 systems are better than 2013 & 2014 systems (except for UEC, 2014 is better)

Progress subtask: Concepts improved vs weaken by team

Most 2015 concepts improved☺



	EURECOM	IRIM	insightdcu	LIG	UEC	ITI_CERTH
better than 2014	23	24	19	25	6	
better than 2013	30	25	14	25	30	21
worse than 2013	0	5	16	5	0	9
worse than 2014	6	6	10	5	21	
same as 2014	1	0	1	0	3	
same as 2013	0	0	0	0	0	0

2015 Observations

- 2015 main task was harder than 2014 main task that was itself harder than 2013 main task (different data and different set of target concepts)
- Raw system scores have higher Max and Median compared to TV2014 and TV2103, still relatively low but regularly improving
- Most common concepts with TV2015 have higher median scores.
- Most Progress systems improved significantly from 2014 to 2015 as this was also the case from 2013 to 2014.
- Stable participation (15 teams) between 2014 and 2015 (but was 26 teams for TV2013).

2015 Observations - methods

- Further moves toward deep learning
- More “deep-only” submissions
- Retraining of networks trained on ImageNet
- Use of many deep networks in parallel
- Data augmentation for training
- Use of multiple frames per shot for predicting
- Feeding of DCNNs with gradient and motion features
- Use of “deep features” (either final or hidden) with “classical” learning
- Hybrid DCNN-based/classical systems
- Engineered features still used as a complement (mostly Fisher Vectors, SuperVectors, improved BoW, and similar) but no new development
- Use of re-ranking or equivalent methods

SIN 2016 ?

- No SIN task is planned for 2016
- Resuming the ad hoc video retrieval task is considered instead