TRECVID-2013 Semantic Indexing task: Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
Dakota Consulting, Inc

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

- 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.
- Participants submitted four types of runs:
  - Main run Includes results for 60 concepts, from which NIST and Quaero evaluated 38
    - Localization run includes results for 10 pixel-wise localized concepts from the 60 evaluated concepts in main runs. *NEW*
  - Progress run Includes results for 60 concept for 3 non-overlapping datasets, from which 2 datasets will be evaluated the next 2 years. *NEW*
  - Pair run Includes results for 10 concept pairs, all evaluated.

Semantic Indexing task (data)

- SIN testing dataset
  - Main test set (IACC.2.A): 200 hrs, with durations between 10 seconds and 6 minutes.
  - Progress test set (IACC.2.B, IACC.2.C): each 200 hrs and non overlapping from IACC.2
- SIN development dataset
  - (IACC.1.A, IACC.1.B, IACC.1.C & IACC.1.tv10.training): 800 hrs, used from 2010 – 2012 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: 549,434
  - Test: IACC.2.A (112,677), IACC.2.B (107,806), IACC.2.C (113,467)
- 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 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 (instance) 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 (training types)

- Six training types were allowed:
  - A - used only IACC training data (110 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 (0 runs)
  - D - used both IACC and non-IACC non-TRECVID training data (0 runs)
  - E – used only training data collected automatically using only the concepts' name and definition (6 runs)
  - F – used only training data collected automatically using a query built manually from the concepts' name and definition (3 runs)
- E & F results inconclusive
  - E & F hardly represented - 9 runs
  - only 1 team system provided an E vs F pair
  - no clear difference.

38 concepts evaluated(1)

Single Concepts

3 Airplane*
5 Anchorperson
6 Animal
10 Beach
15 Boat_Ship*
16 Boy*
17 Bridges*
19 Bus
25 Chair*
31 Computers*
38 Dancing
49 Explosion_Fire
52 Female-Human-Face-Closeup
53 Flowers
54 Girl*
56 Government_Leader*
59 Hand
71 Instrumental_Musician*
72 Kitchen*
80 Motorcycle*
83 News_Studio
86 Old_People
89 People_Marching
100 Running
105 Singing*
107 Sitting_down*
117 Telephones
120 Throwing*
163 Baby*
227 Door_Opening
254 Fields*
261 Flags
267 Forest*
274 George_Bush*
342 Military_Airplane*
392 Quadruped
431 Skating
454 Studio_With_Anchorperson

-The 19 marked with "*" are a subset of those tested in 2012

Concepts evaluated (2)

- Concept pairs
  - [911] Telephones + Girl
  - [912] Kitchen + Boy
  - [913] Flags + Boat_Ship
  - [914] Boat_Ship + Bridges
  - [915] Quadruped + Hand
  - [916] Motorcycle + Bus
  - [917] Chair + George_[W_]Bush
  - [918] Flowers + Animal
  - [919] Explosion_Fire + Dancing
  - [920] Government-Leader + Flags
- Localization concepts
  - [3] Airplane
  - [15] Boat_ship
  - [17] Bridges
  - [19] Bus
  - [25] Chair
  - [59] Hand
  - [80] Motorcycle
  - [117] Telephones
  - [261] Flags
  - [392] Quadruped

Evaluation

- NIST evaluated 15 concepts + 5 concept pairs and Quaero evaluated 23 concepts + 5 concept pairs.
- 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
- Metrics : *inferred average precision per concept*
- Compared runs in terms of **mean** *inferred average precision* across the:
  - 38 feature results for main 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
- More features can be judged with same effort
- Increased sensitivity to lower ranks
- 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.

2013: 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 6.7%

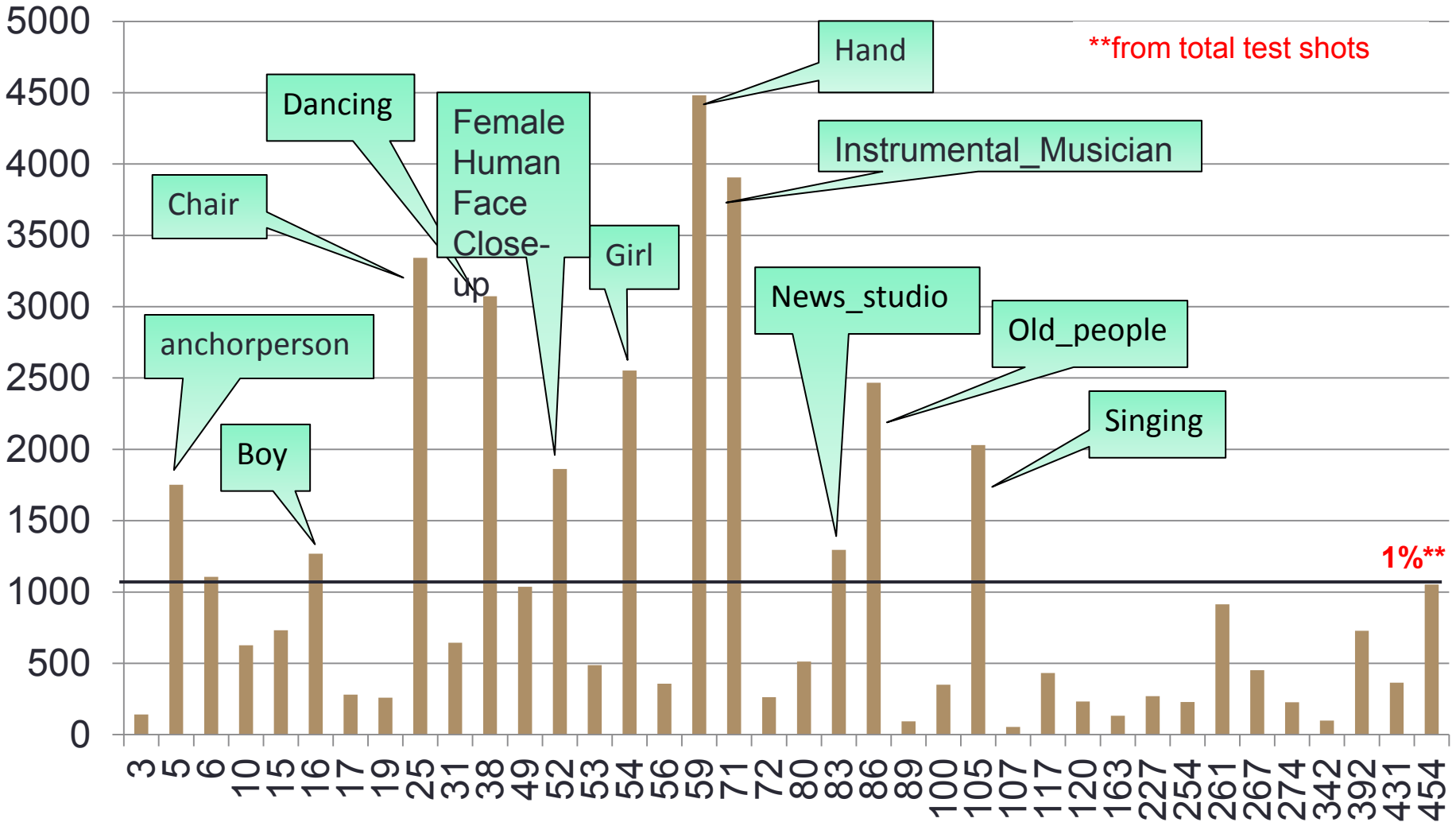
48 concepts
336,683 total judgments
12006 total hits
8012 Hits at ranks (1-100)
3239 Hits at ranks (101-200)
755 Hits at ranks (201-2000)

- 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

2013 : 26 Finishers

PicSOM	Aalto U.
INF	Carnegie Mellon U.
IRIM	CEA-LIST, ETIS, EURECOM, INRIA-TEXMEX, LABRI, LIF, LIG, LIMSI-TLP, LIP6, LIRIS, LISTIC, CNAM
VIREO	City U. of Hong Kong
Dcu_savasa	Dublin City U. (Ireland), U. of Ulster (UK), Vicomtech-IK4 (Spain)
EURECOM	EURECOM - Multimedia Communications
VIDEOSENSE	EURECOM,LIRIS, LIF, LIG, Ghanni
TOSCA	EuropeOrganization(s)
FIU_UM	Florida International U., U. of Miami
FHHI	Fraunhofer Heinrich Hertz Institute, Berlin
HFUT	Hefei U. of Technology
IBM	IBM T. J. Watson Research Center
ITI_CERTH	Information Technologies Institute(Centre for Research and Technology Hellas)
Quaero	INRIA, LIG, KIT
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH
AXES	DCU,UTwente,Oxford,INRIA,Fraunhofer,KULeuven,Technicolor,ErasmusU, Cassidian,BBC,DW,NISV,ERCIM
NII	National Institute of Informatics
NHKSTRL	NHK (Japan Broadcasting Corp.)
ntt	NTT Media Intelligence Labs, Dalian U. of Technology
FTRDBJ	Orange Labs International Centers China
SRIAURORA	SRI, Sarnoff, Central Fl.U., U. Mass., Cycorp, ICSI, Berkeley
TokyoTechCanon	Tokyo Institute of Technology and Canon
Sheffield	U. of Sheffield, UK Harbin Engineering U., PRC U. of Engineering & Technology, Lahore, Pakistan
MindLAB	U. Nacional de Colombia
MediaMill	U. of Amsterdam
UEC	U. of Electro-Communications

Inferred frequency of hits varies by concept



Total true shots contributed uniquely by team

Main runs

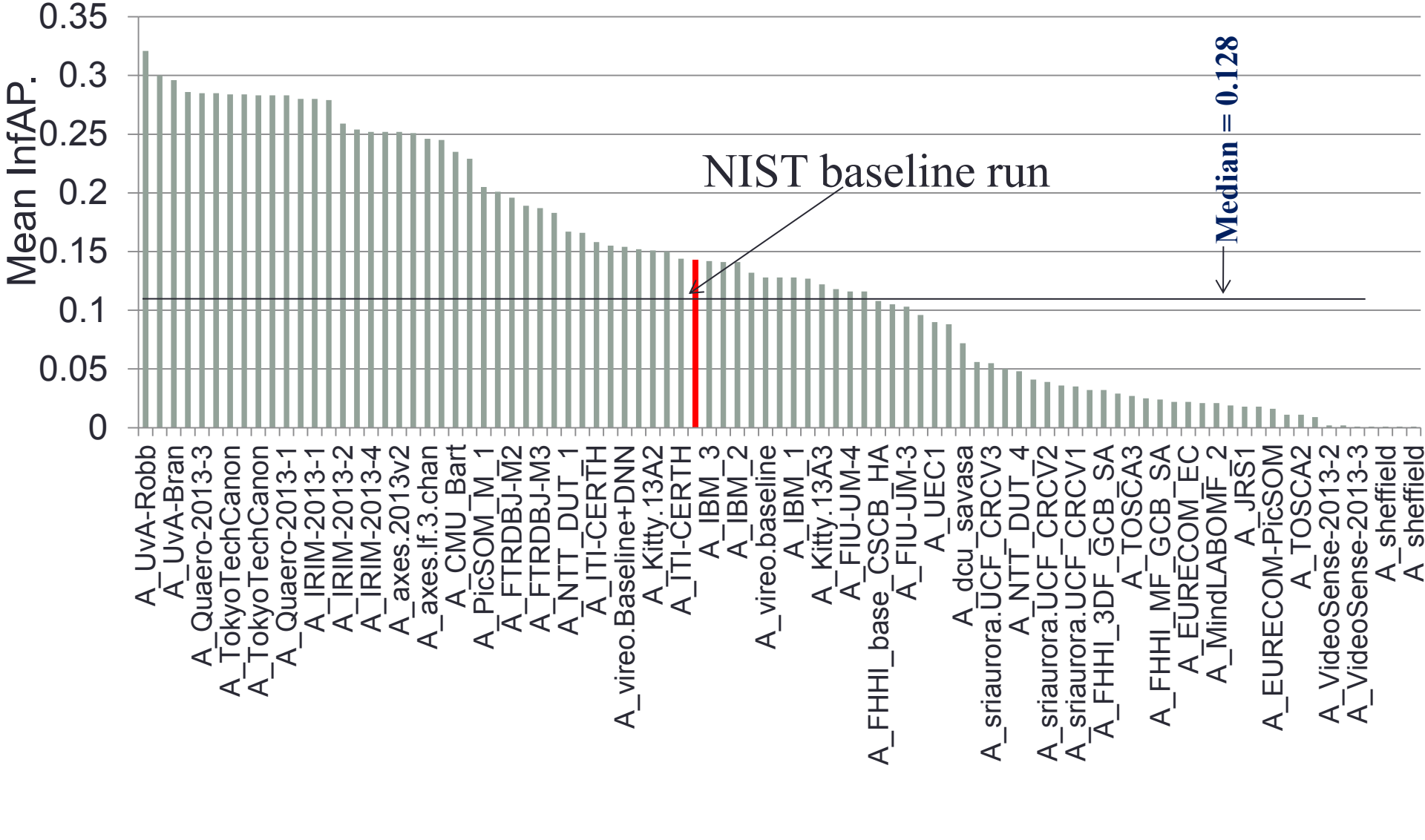
Team	No. of Shots	Team	No. of shots
NTT	65	FIU	10
Min	51	Kit	10
sri	49	FTR	8
EUR	38	ITI	8
FHH	32	Dcu	7
UEC	30	TOS	6
UvA	25	IBM	2
JRS	22	She	1
CMU	18	Tok	1
HFU	14		
vir	14		
NHK	13		
Pic	11		

Fewer unique shots compared to TV2012

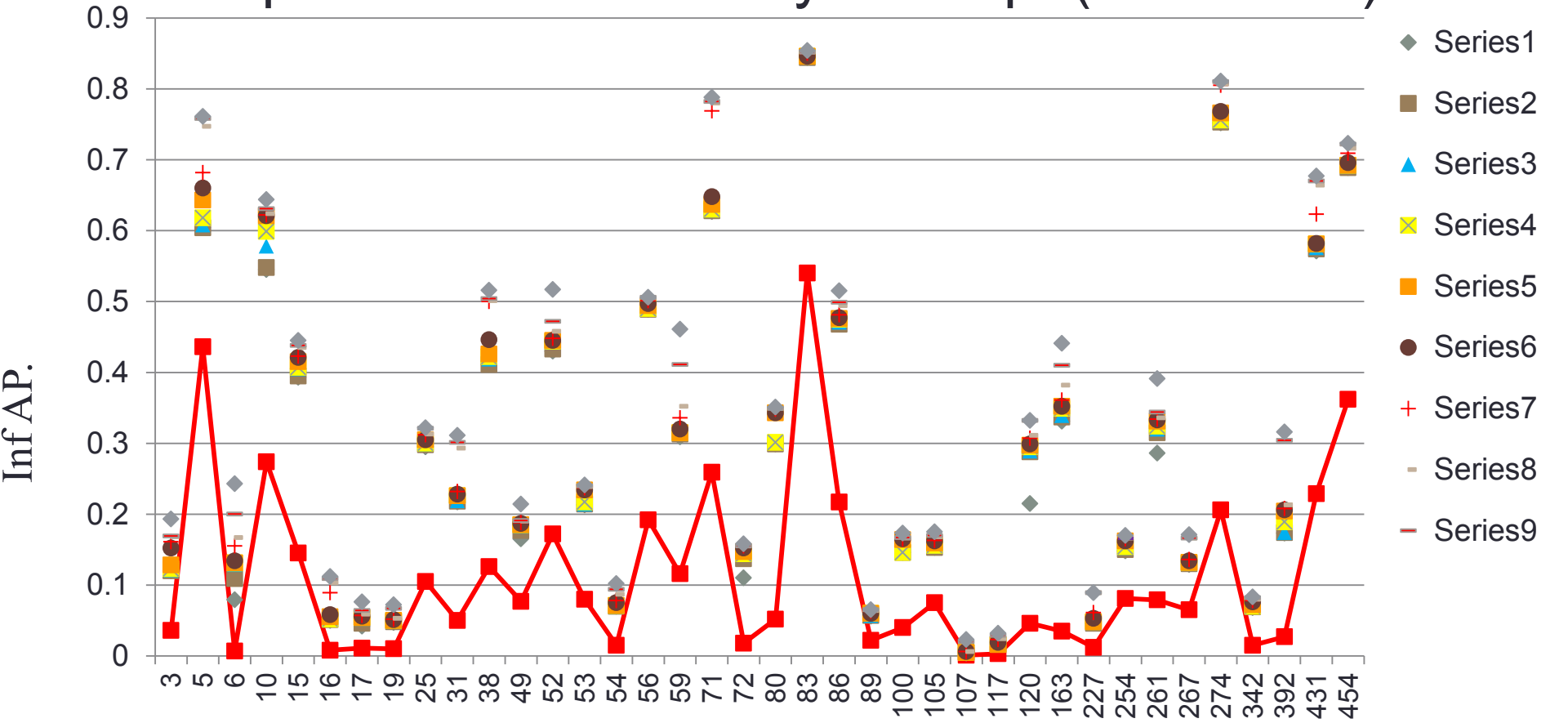
Pair runs

Team	No. of Shots
Sri	3
CMU	2
HFU	1

Category A results (Main runs)



Top 10 InfAP scores by concept (Main runs)



3* Airplane	5 Anchorperson	6 Animal	10 Beach	15* Boat_ship	16* Boy	17* Bridges	19 Bus	25* Chair	31* Computers	38 Dancing	49 Explosion_ Fire	52 Female_human _face_closeup	53 Flowers	54* Girl
56* Government _ Leader	59 Hand	71* Instrumental _ Musician	72* Kitchen	80* Motorcycle	83 Niews_studio	86 Old_people	100 Running	105* Singing	107* Sitting_ down	117 Telephones	120* Throwing	163* Baby	227 Door_openin g	254* Fields
261 Flags	267* Forest	274* George_Bu sh	342* Military_Airplan e	392 Quadruped	431 Skating	454 Studio_with_ anchorperson								

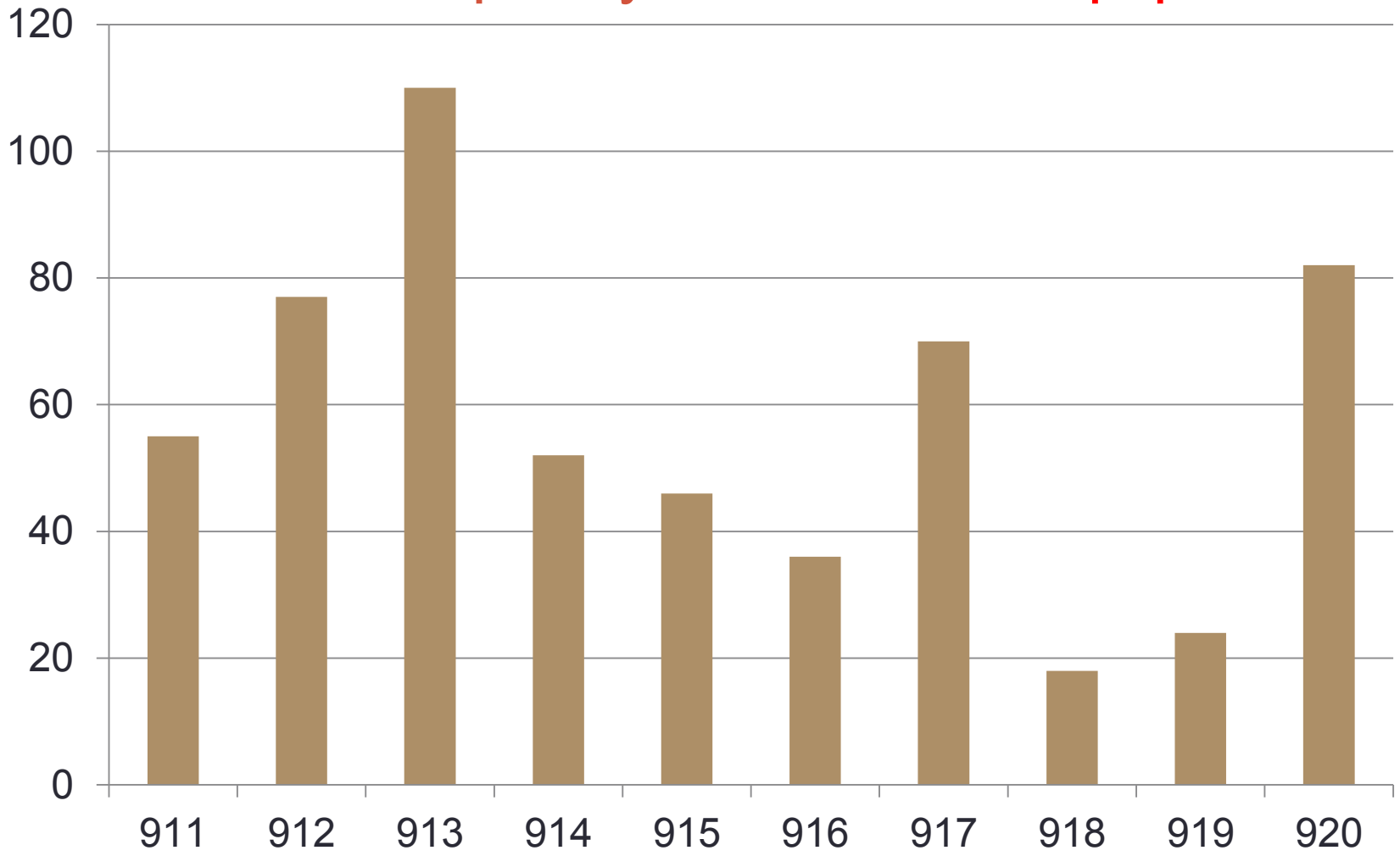
* Common concept in TV2012

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

Run name	(mean infAP)
UvA-Robb_1	0.321
UvA-Arya_2	0.300
UvA-Bran_3	0.296
UvA-Jon_4	0.286
Quaero-2013-3_3	0.285
Quaero-2013-2_2	0.285
TokyoTechCanon_2	0.284
TokyoTechCanon_1	0.284
TokyoTechCanon_3	0.283
Quaero-2013-4_4	0.283

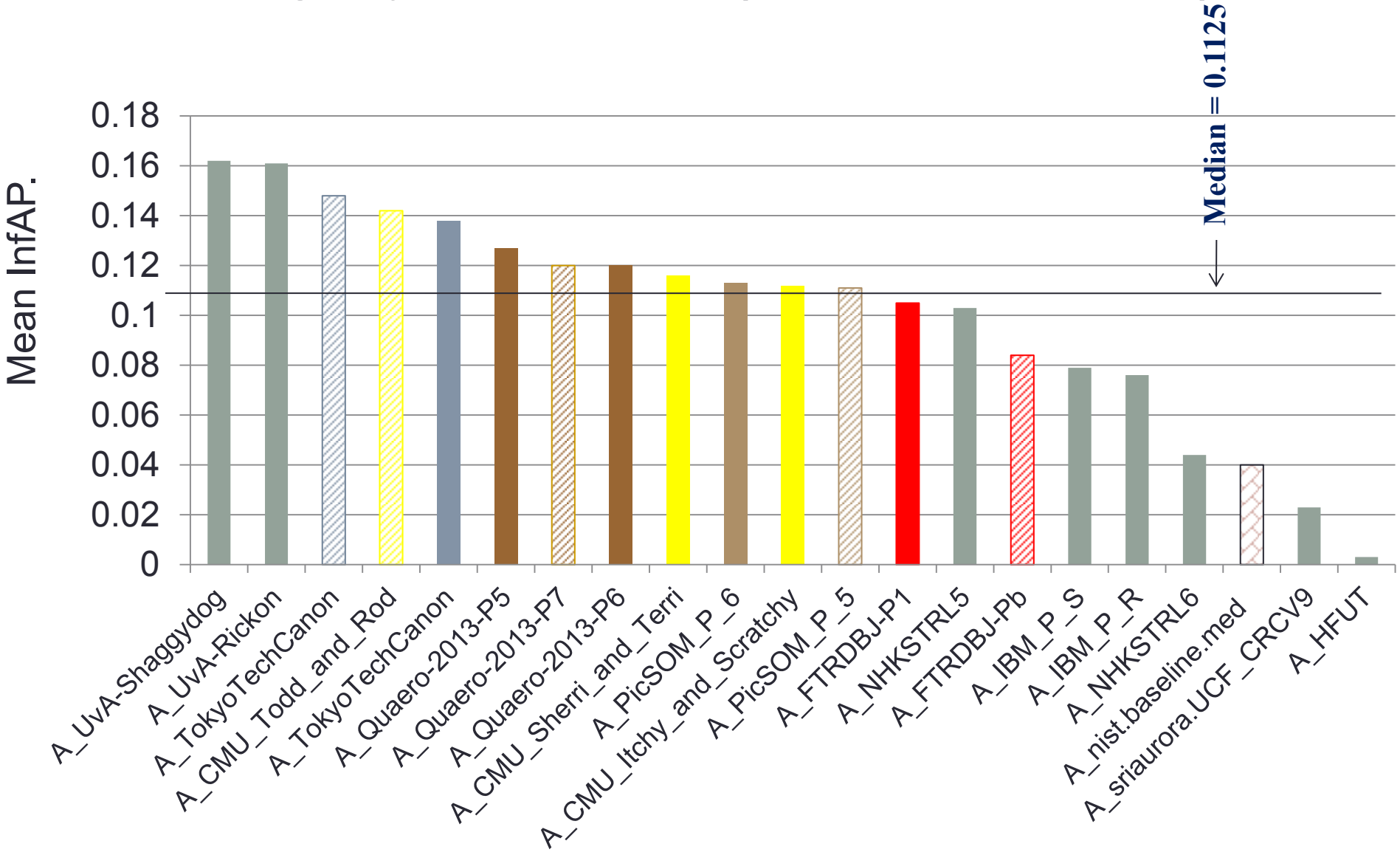
- UvA-Robb_1
  - UvA-Arya_2
    - Quaero-2013-3_3
    - Quaero-2013-2_2
    - Quaero-2013-4_4
    - UvA-Jon_4
  - UvA-Bran_3
  - TokyoTechCanon_2
  - TokyoTechCanon_1
  - TokyoTechCanon_3

Inferred frequency of hits for concept pairs



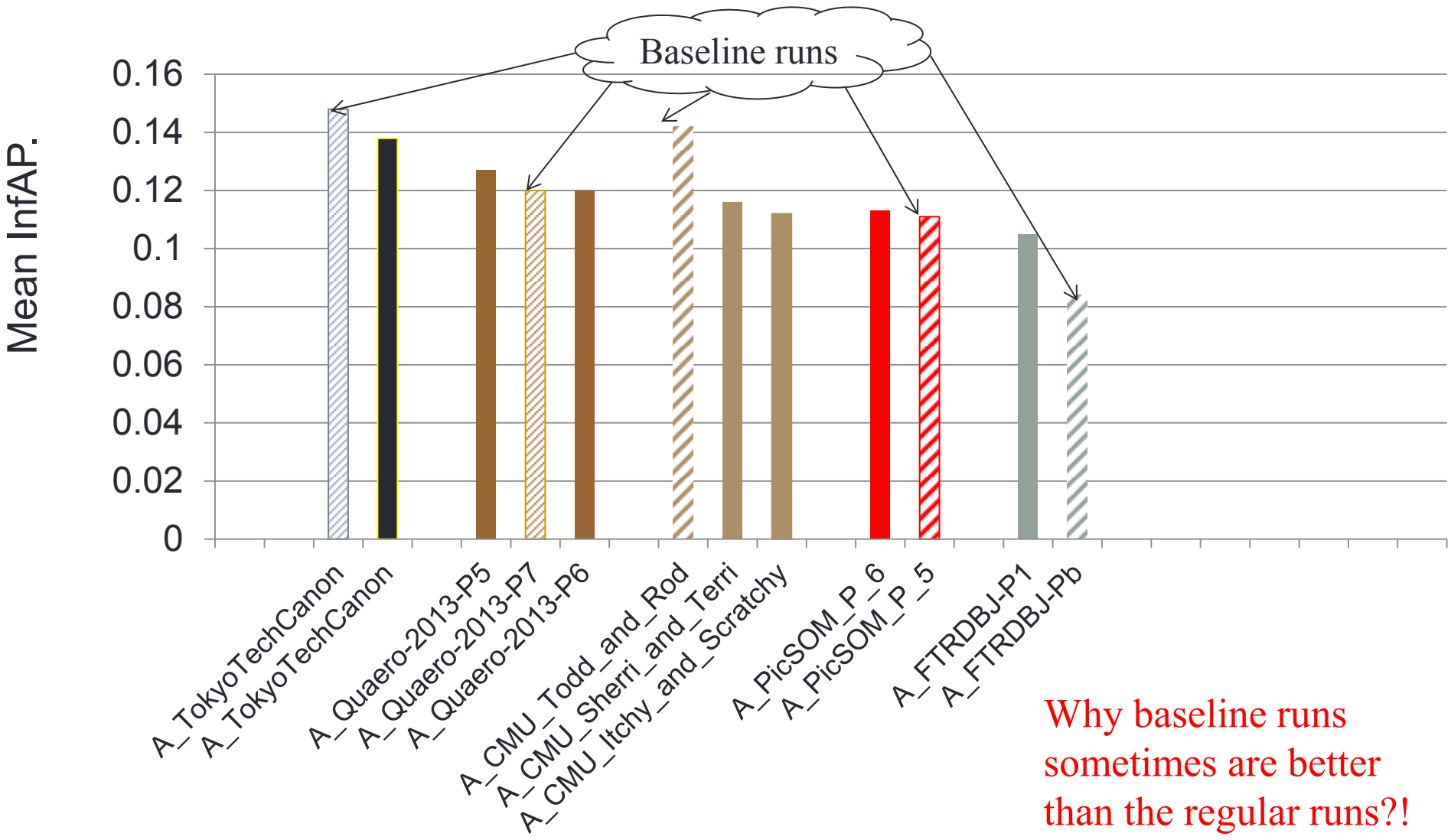
911	912	913	914	915	916	917	918	919	920
Telephones + Girl	Kitchen + Boy	Flags + Boat_ship	Boat_ship + Bridges	Quadruped + Hand	Motorcycle + Bus	Chair + George_W_Bush	Flowers + Animal	Explosion_Fire + Dancing	Government_leader + Flags

Category A results (Concept Pairs)



Only 1 ‘E’ run submitted with score 0!

Category A results (regular vs baseline runs by group)



Why baseline runs sometimes are better than the regular runs?!

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

Run name	(mean infAP)
A_UvA-Shaggydog_8	0.162
A_UvA-Rickon_7	0.161
A_TokyoTechCanon_6	0.148
A_CMU_Todd_and_Rod_3	0.142
A_TokyoTechCanon_5	0.138
A_Quaero-2013-P5 _5	0.127
A_Quaero-2013-P7_7	0.120
A_Quaero-2013-P6_6	0.120
A_CMU_Sherri_and_Terri_2	0.116
A_PicSOM_P_6_6	0.113

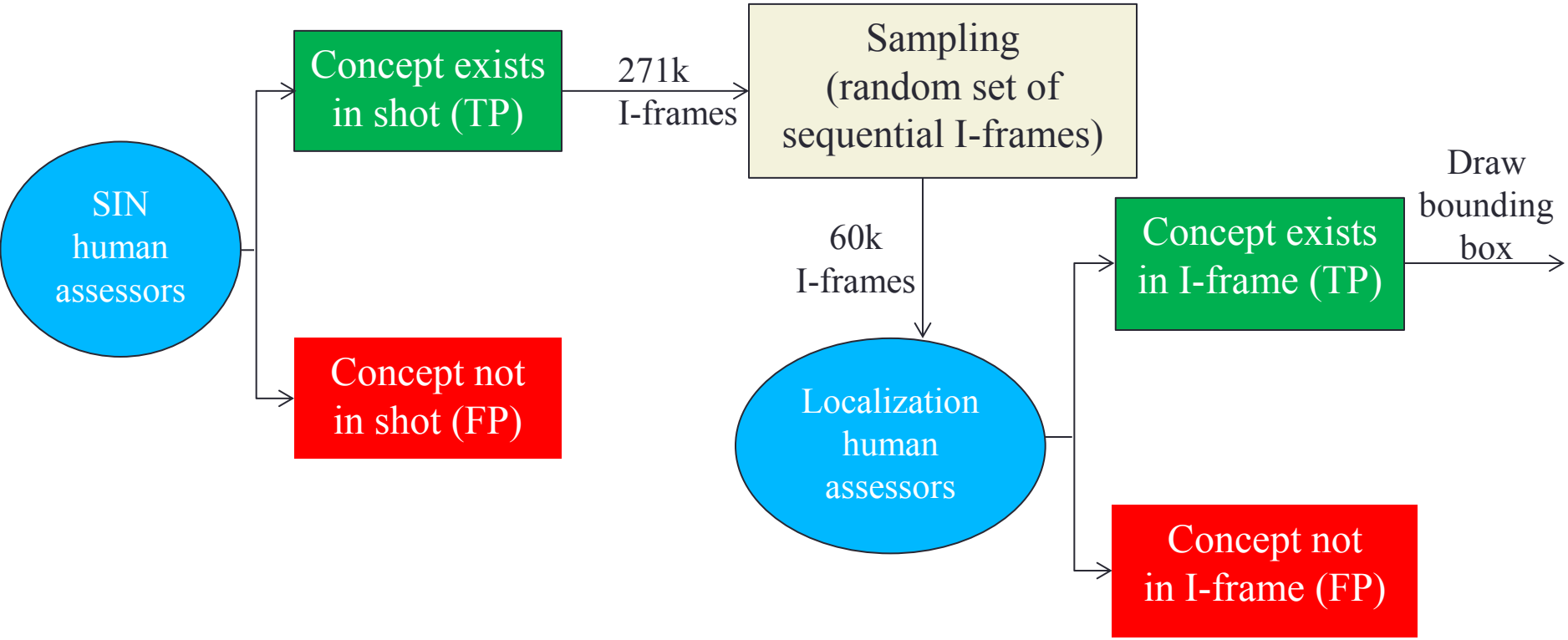
- A_TokyoTechCanon_6
  - A_Quaero-2013-P6_6
  - A_Quaero-2013-P7_7
  - A_TokyoTechCanon_5
- A_CMU_Todd_and_Rod_3
  - A_CMU_Sherri_and_Terri_2
  - A_PicSOM_P_6_6

- A_UvA-Shaggydog_8
  - A_CMU_Sherri_and_Terri_2
  - A_PicSOM_P_6_6
  - A_Quaero-2013-P7_7
  - A_TokyoTechCanon_5
  - A_Quaero-2013-P5 _5
    - A_Quaero-2013-P6_6
- A_UvA-Rickon_7
  - A_CMU_Sherri_and_Terri_2
  - A_PicSOM_P_6_6
  - A_Quaero-2013-P7_7
  - A_TokyoTechCanon _5
  - A_Quaero-2013-P5 _5
    - A_Quaero-2013-P6_6

Concept localization subtask

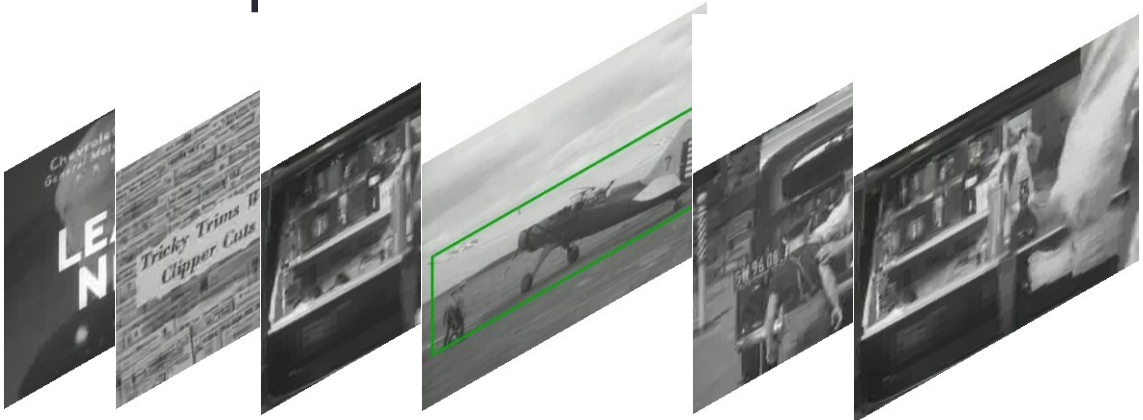
- Goal
  - Make concept detection more precise in time and space than current shot-level evaluation.
- Task
  - For each of the 10 concepts
    - For each of the top 1000 main run shots
      - For each I-Frame within the shot that contains the target, return
        - the x,y coordinates of the (UL,LR) vertices of a bounding rectangle containing all of the target concept and as little more as possible.
- Systems were allowed to submit more than 1 bounding box per I-frame but only one with maximum fscore were judged.

NIST Evaluation framework



Evaluation metrics

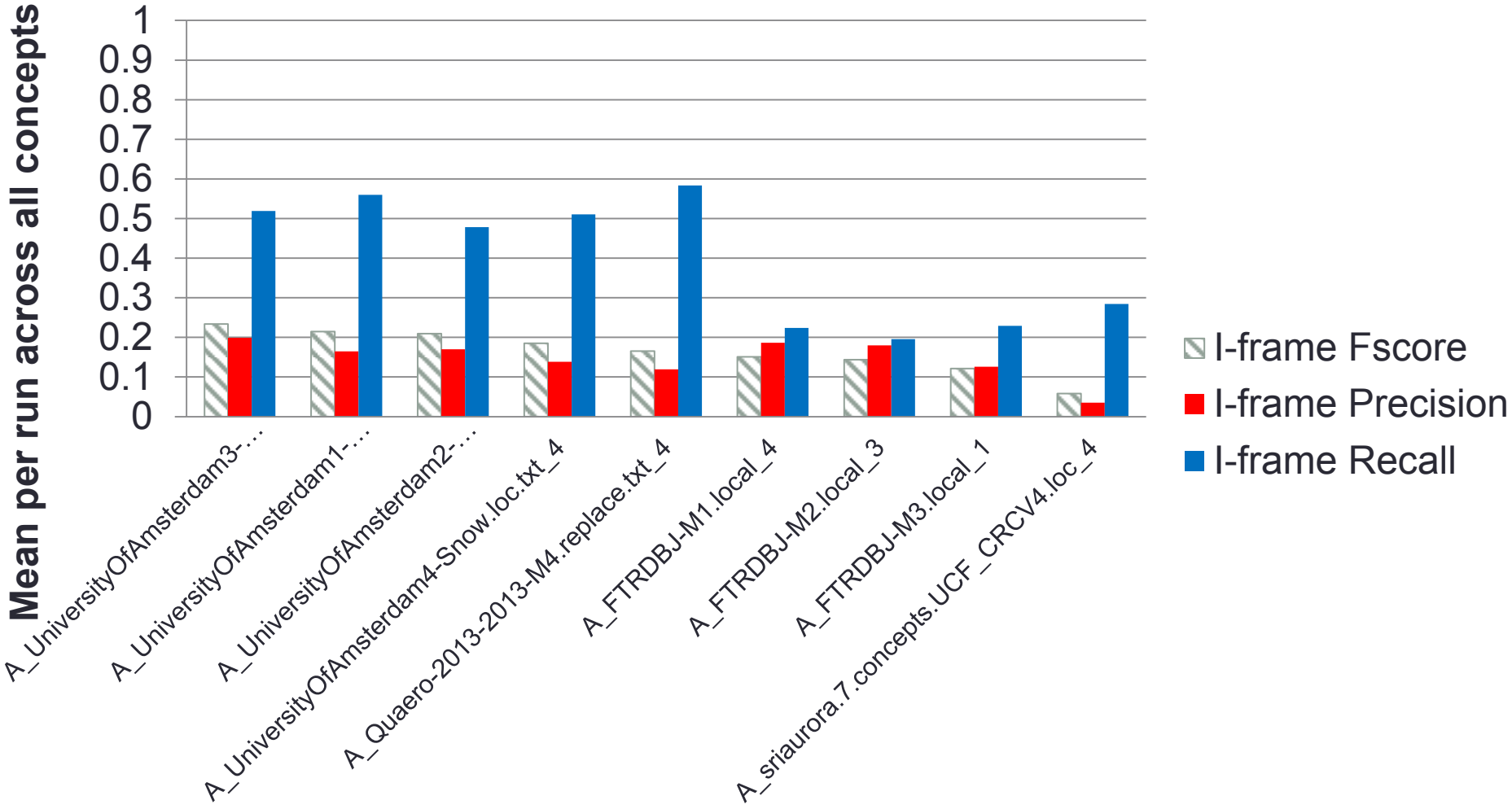
- Temporal localization: precision, recall and fscore based on the judged I-frames.
- Spatial localization: precision, recall and fscore based on the located pixels representing the concept.
- An average of precision, recall and fscore for temporal and spatial localization across all I-frames for each concept and for each run.



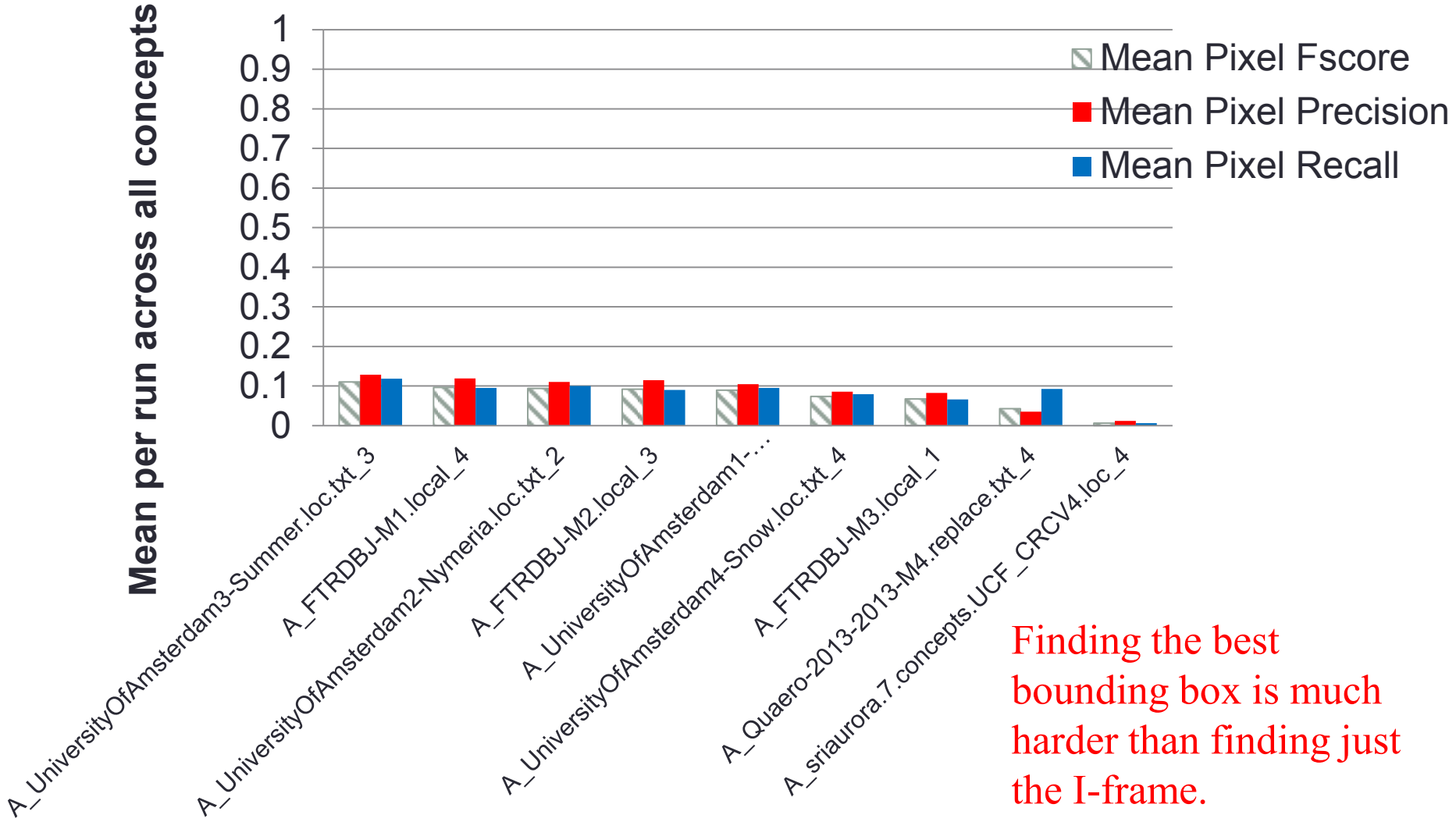
Participants (Finishers)

- 4 teams submitted 9 runs
  - UvA (University Of Amsterdam)
  - SRIAURORA (SRI, Sarnoff, Central Fl.U., U. Mass., Cycorp, ICSI, Berkeley)
  - FTRDBJ (Orange Labs International Centers China)
  - QUAERO (INRIA, LIG, KIT)

Temporal localization results by run

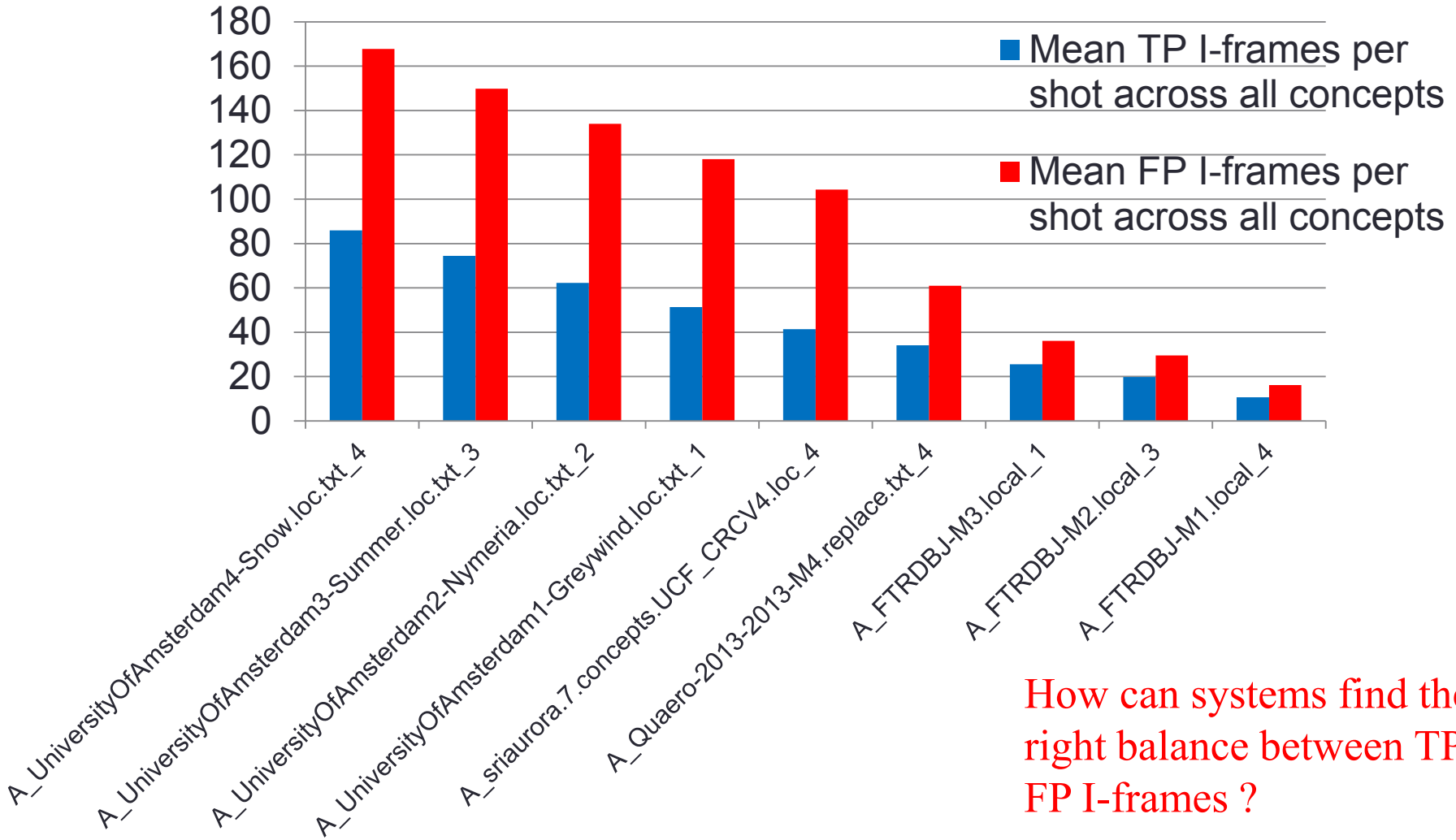


Spatial Localization results by run



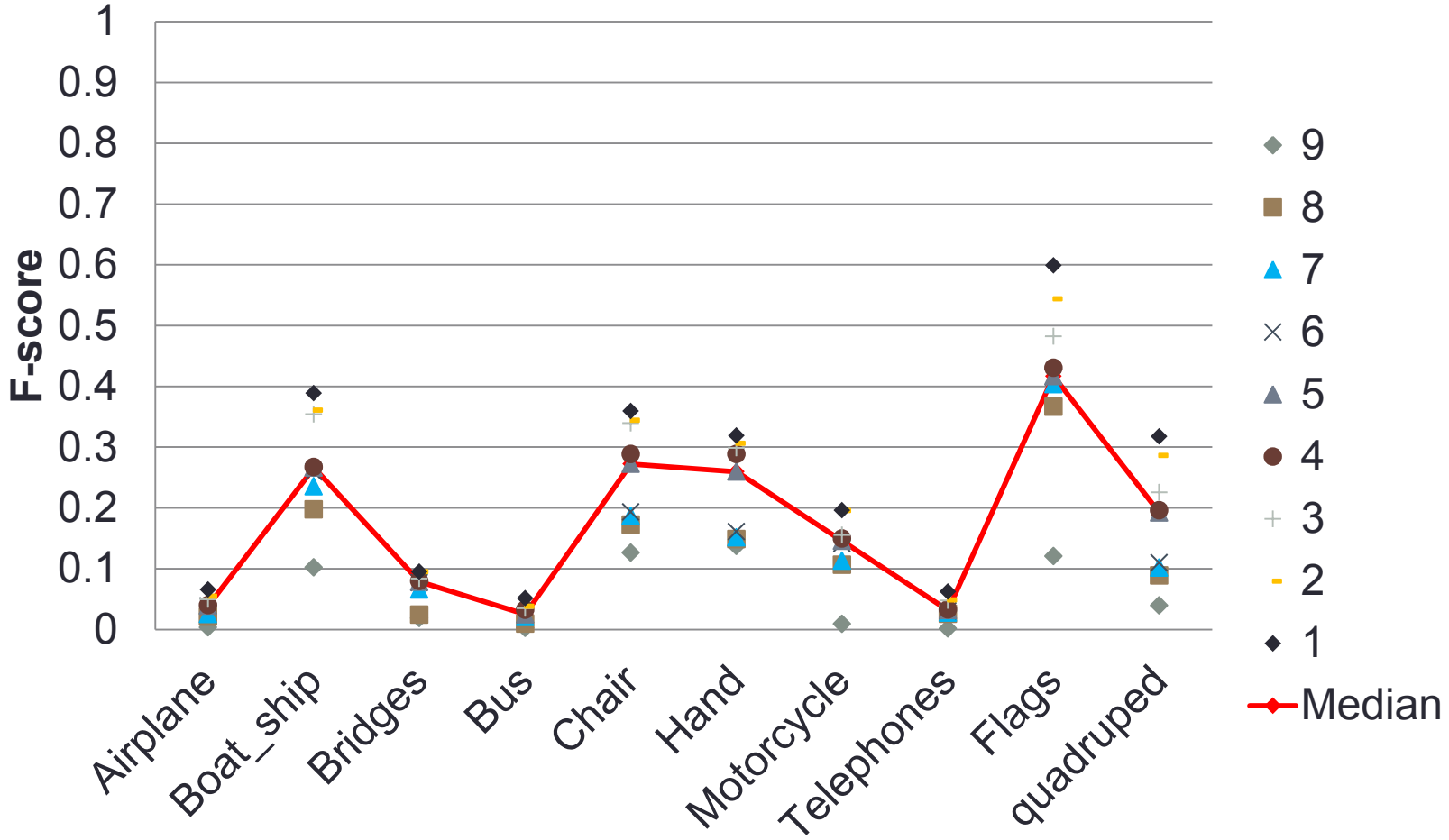
Finding the best bounding box is much harder than finding just the I-frame.

TP vs FP submitted I-frames by run

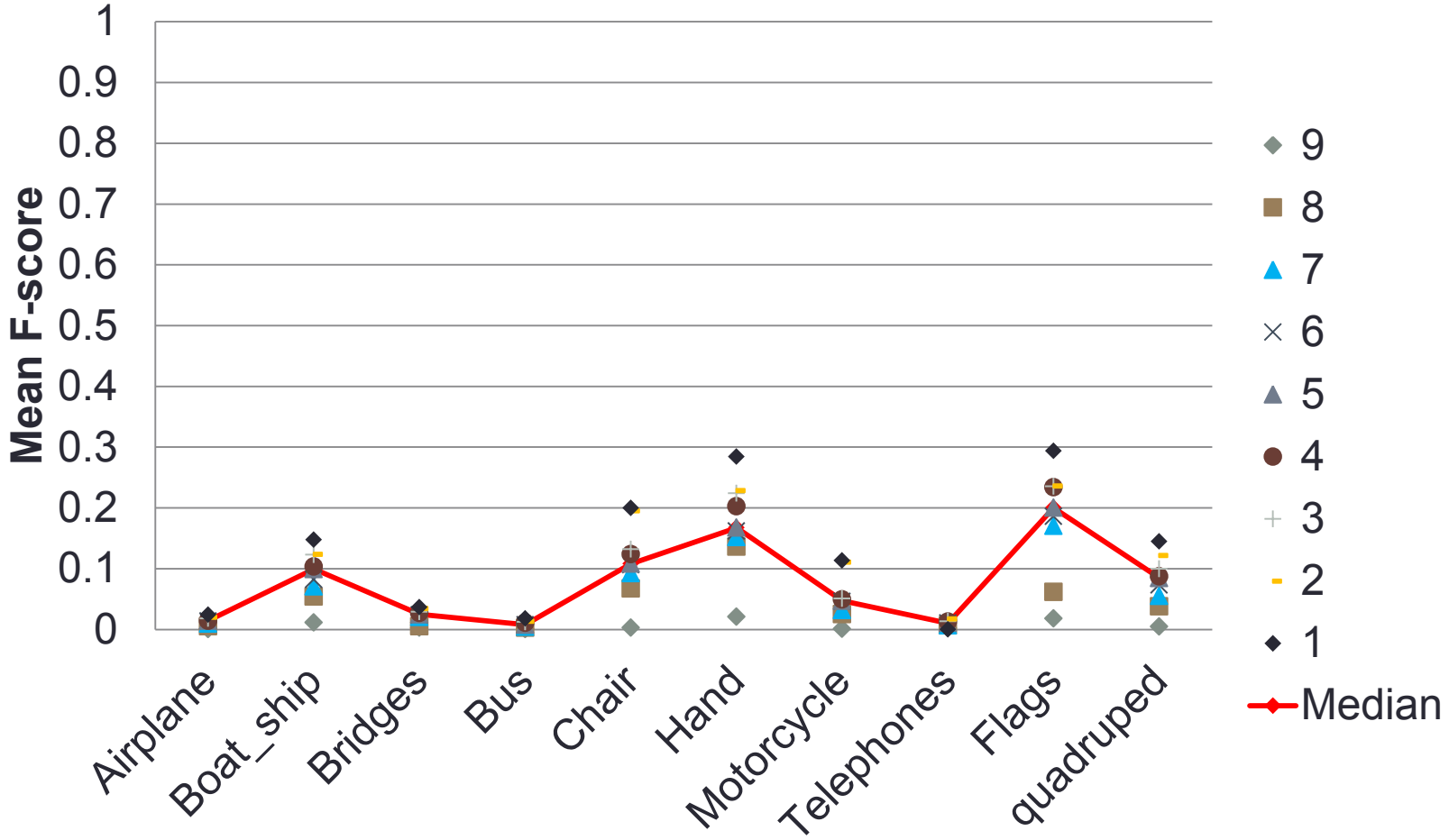


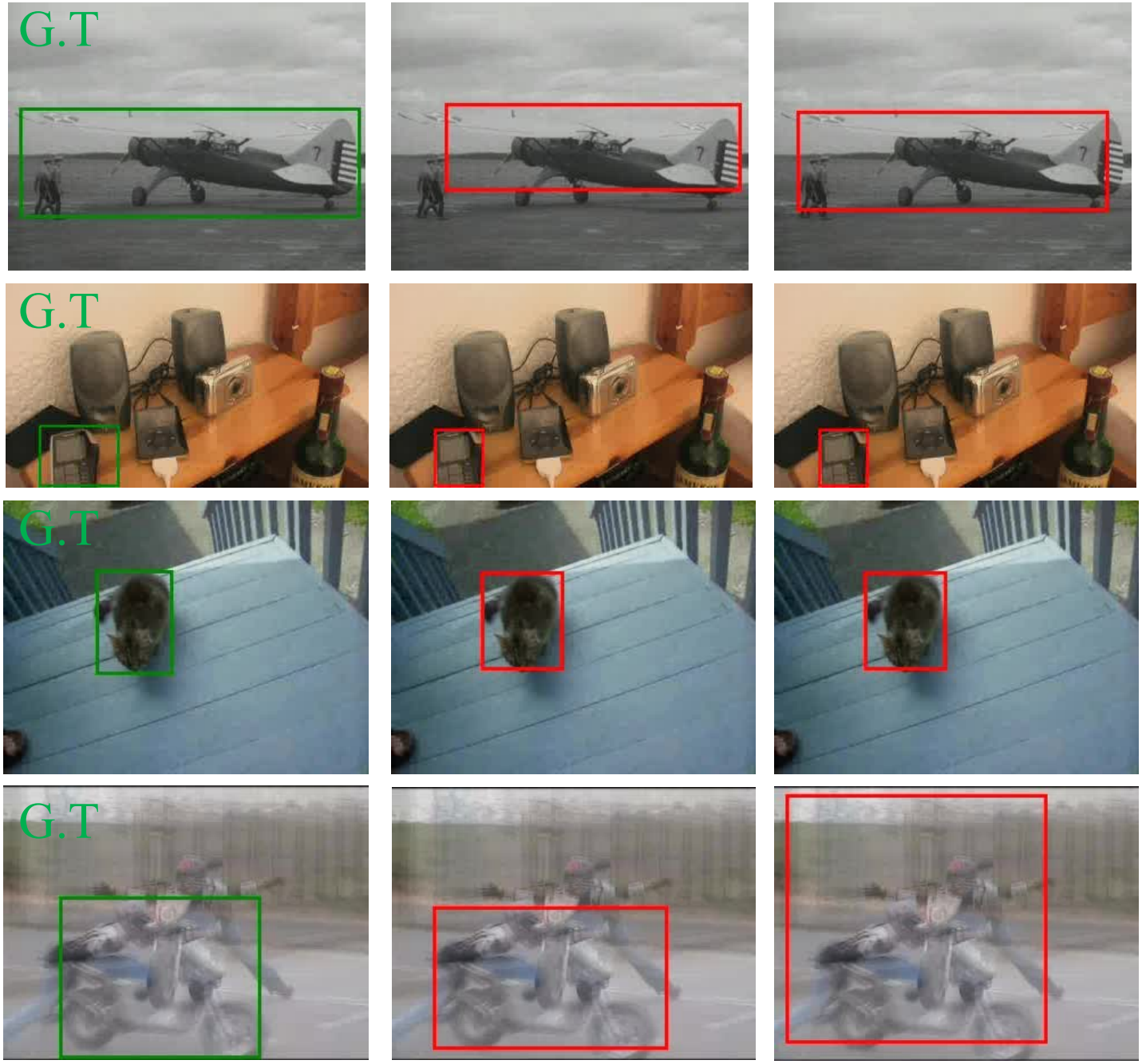
How can systems find the right balance between TP vs FP I-frames ?

Temporal localization results per concept



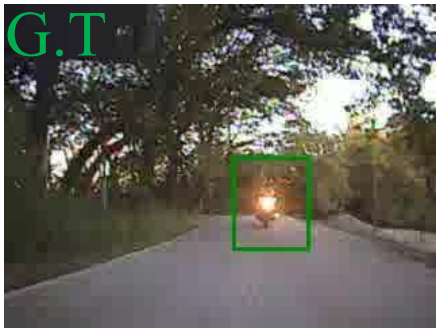
Spatial localization results per concept





Samples of good localization



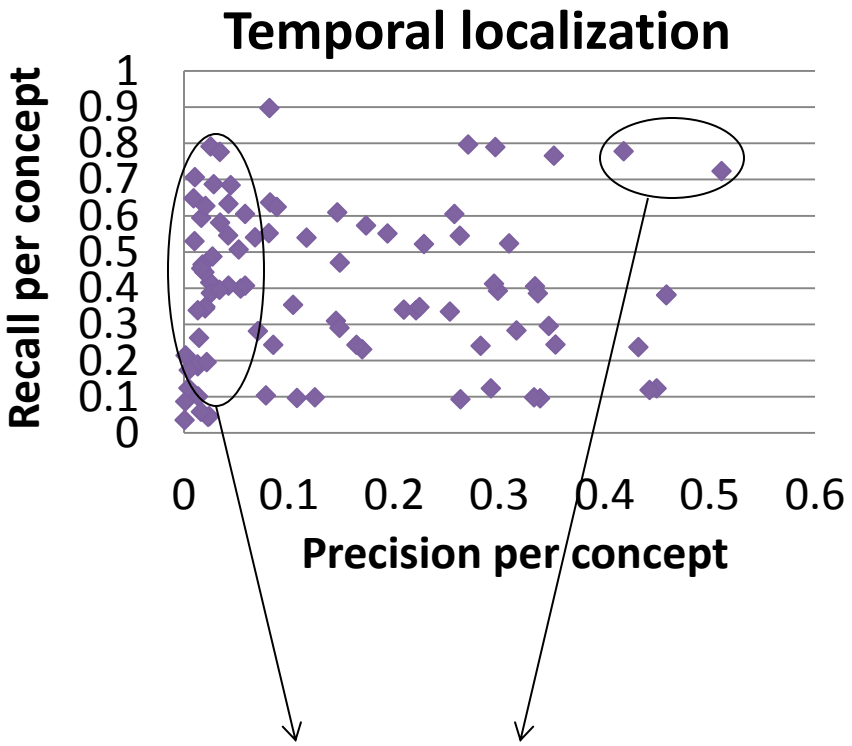




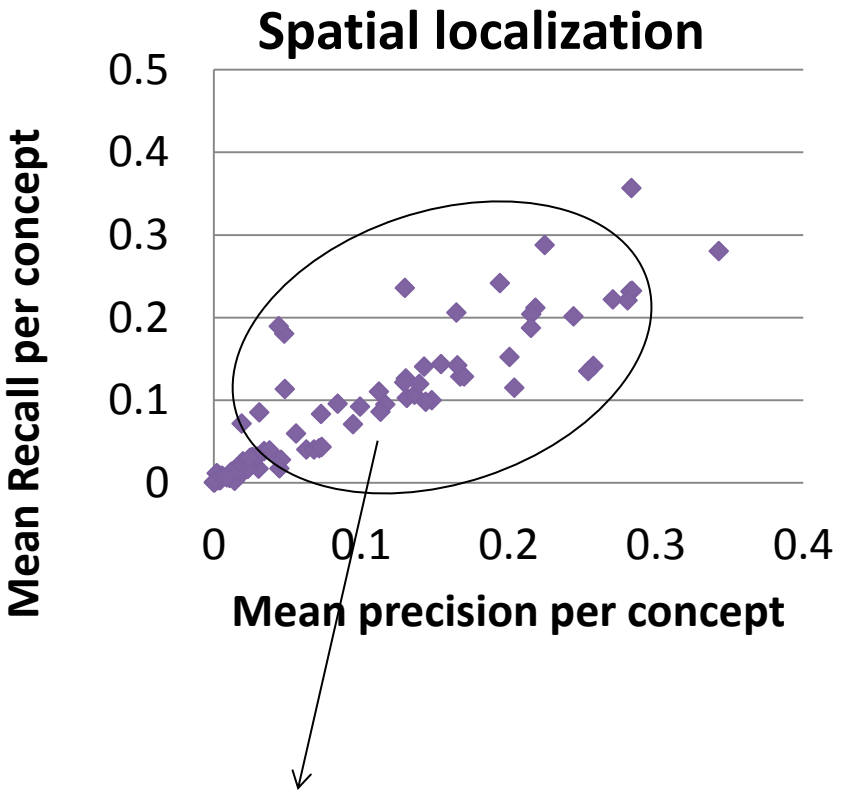


Samples of less good localization

Results per concept across all teams



Majority of systems submitted a lot of non-target I-frames. While few found a balance



Most systems submitted bounding boxes ~= G.T boxes AND overlaps. Systems are good in finding the real box sizes☺

2013 Observations

- No submissions for training types B, C, & D
- Training types E & F still very few
- Fewer unique shots found vs TV2012
- No teams submitted any results for feature sequence in concept pairs!! Why?
- Concept-pairs baseline submissions are better than regular runs! (why? How to improve learning concept pairs?)
- For most localization systems, finding the correct I-frame is much easier than finding the bounding box

2013 Observations

- Site experiments include (not exhaustive):
  - focus on robustness, merging many different representations
  - use of spatial pyramids
  - improved bag of word approaches
  - Fisher/super-vectors, VLADs, VLATs
  - audio analysis
  - consideration of scalability issues

  - improved rescoring methods
  - use of semantic features
  - work on the kernel size parameter of the SVM-RBF kernel
  - work on the “no annotation” conditions: use of socially tagged videos or images and develop strategies for positive example selection
  - deep convolutional neural networks (deep learning)

Announcements

- The full set of the 60 single concepts judgments are now available
- New qrels will be made available on the website
- No significant change in systems ranking are observed

SIN 2014

- Globally keep the task similar and of similar scale
- Further explore the “concept pair” and “no annotation” and “localization” variants
- Common training data for the “no annotation” variant is likely will be delivered LIG (F type)
- Sharing of data still proposed by IRIM
- Method for measuring progress over years
- Collaborative annotation unchanged
- 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 tv13 active participant login/password):
  - http://mrim.imag.fr/trecvid/wiki
  - http://mrim.imag.fr/trecvid/wiki/doku.php?id=sin_2013_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.