ICMR 2017 Mini-Tutorial

TRECVid Semantic Indexing

Georges Quénot
Multimedia Information Modeling and Retrieval Group

Laboratoire d'**I**nformatique de **G**renoble

with input from George Awad (Dakota Consulting, Inc and NIST)
and many others

June 6, 2017

Tutorial Outline

- Part I: general TRECVid introduction
- Part II: the Semantic Indexing (SIN) task
- PART III: the LIG / IRIM “baseline”, from “bag of things” to deep learning

Part I
General TRECVid introduction

G. Awad, J. Fiscus, D. Joy, M. Michel, A. F. Smeaton, W. Kraaij, G. Quénot, M. Eskevich, R. Aly, R. Ordelman, G. J. F. Jones, B. Huet, M. Larson. ***TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking.*** TRECVID 2016, NIST, USA.

http://www-nlpir.nist.gov/projects/tvpubs/tv16.papers/tv16overview.pdf

TREC Video Retrieval Evaluation

TRECVID 2016

George Awad#
Ian Soboroff *
Angela Ellis*
Darrin Dimmick*

Jonathan Fiscus**
David Joy**
Martial Michel**
Andrew Delgado**

Alan Smeaton (Dublin City University)
Wessel Kraaij (TNO, Radboud University Nijmegen)
Georges Quénot (Laboratoire d'Informatique de Grenoble)
Roeland Ordelman, Robin Aly (University of Twente)
Maria Eskevich, Martha Larson (Radboud University Nijmegen)
Gareth Jones (Dublin City University)
Benoit Huet (EURECOM)
Marc Ritter (Technische Universitat Chemnitz)
Stephanie Strassel+
Xuansong Li+ et al

* Retrieval Group / ** Multimodal Information Group
Information Access Division
Information Technology Laboratory
NIST

+ Linguistic Data Consortium

Dakota Consulting, Inc
Silver Spring, MD

NIST
National Institute of Standards and Technology

What is TRECVID?

Workshop series (2001 – present) → http://trecvid.nist.gov

to promote research/progress in content-based video analysis/exploitation

Foundation for large-scale laboratory testing

Forum for the

- ✓ exchange of research ideas
- ✓ discussion of approaches – what works, what doesn't, and why.

Focus: content-based approaches

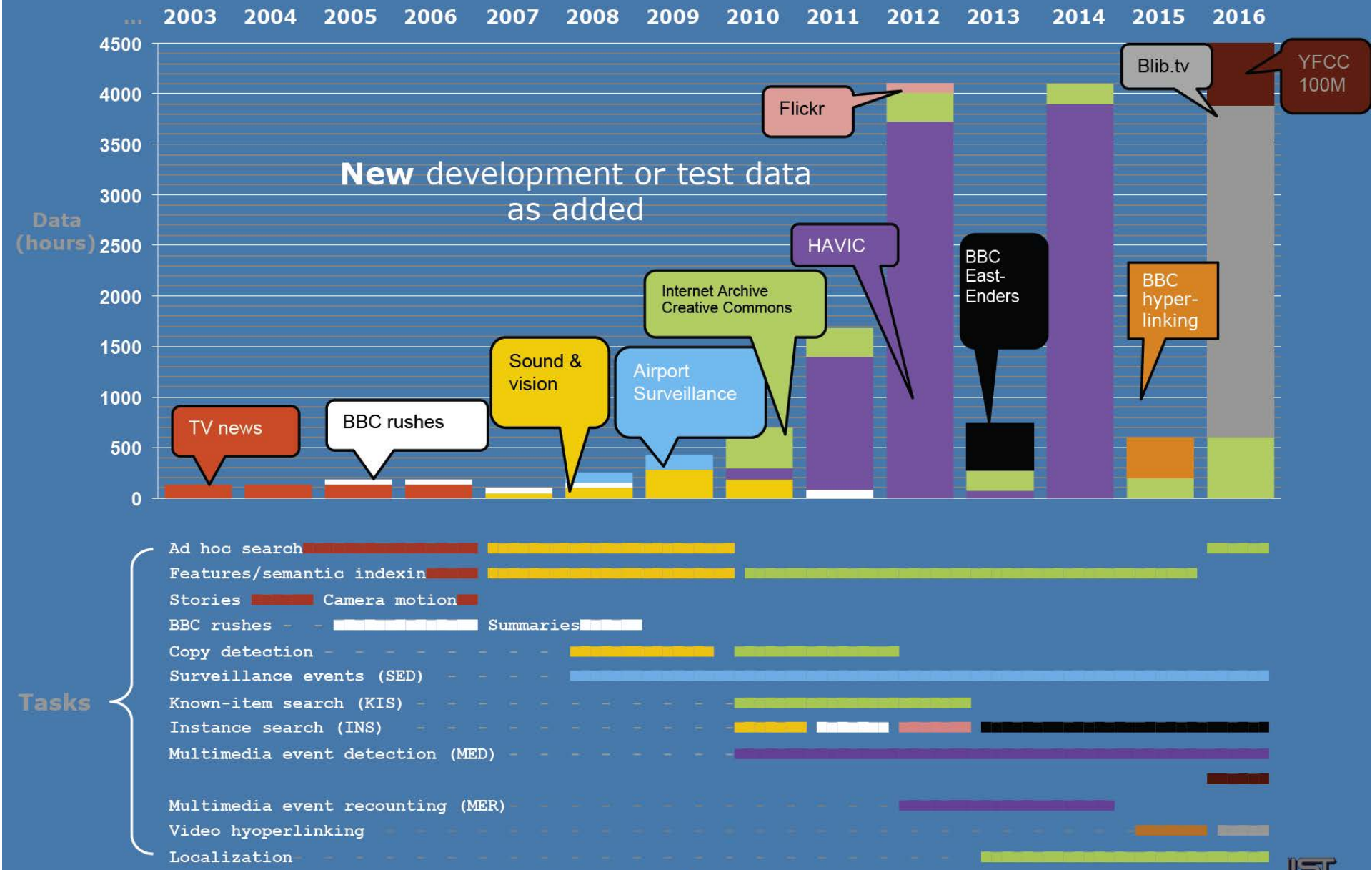
- ✓ search / detection / summarization / segmentation / …

Aims for realistic system tasks and test collections

- ✓ unfiltered data
- ✓ focus on relatively high-level functionality (e.g. interactive search)
- ✓ measurement against human abilities

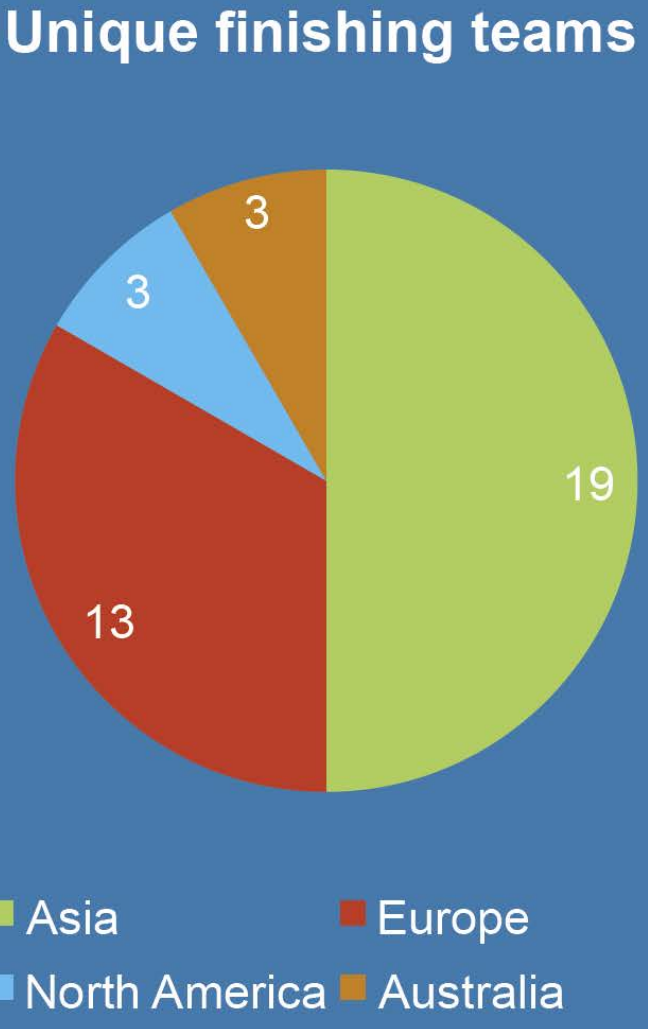
Provides data, tasks, and uniform, appropriate scoring procedures

TRECVID's Evolution



TV2016 Finishers

Groups Finished	Task code	Task name
8	SED	Surveillance event detection
13	AVS	Ad-hoc Video Search
13	INS	Instance search
12	MED	Multimedia event detection
5	LNK	Video hyperlinking
3	LOC	Localization
7	VTT	Pilot task (Video_to_Text)



Part II
the Semantic Indexing (SIN) task
TRECVid 2010-2015
(formerly “high level feature detection”)

G Awad, C. G. M. Snoek, A. F. Smeaton, G. Quénot. ***TRECVID Semantic Indexing of Video: A 6-Year Retrospective. Invited Paper***. ITE Transactions on Media Technology and Applications, 2016.

https://www.jstage.jst.go.jp/article/mta/4/3/4_187/_article

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

Semantic Indexing task

System task definition:

- Given
  - a test collection (~200 hours of video from the IACC collection)
  - a master shot reference (~120,000 shots)
  - a set of concept definitions (from 30 to 346)
- return for each target concept a list of at most 2000 shot IDs from the test collection ranked according to their likelihood of containing the target.

Data

- ~1400 hours from the “Internet Archive Creative Commons” (IACC collection).
- 7 slices of ~200 hours:
  - IACC.1.tv10.training: initial training data
  - IACC.1.[A-C]: 2010-2012 test data
  - IACC.2.[A-C]: 2013-2015 test data
- Approximately 120,000 video shots / slice
- Note: IACC.3 ~600 additional hours for the 2016-2018 AVS task

Additional data

- Common annotation for 346+ concepts coordinated by LIG / LIF / Quaero* from 2007-2013 made available (~28M annotations)
- Multilingual Automatic Speech Recognition (ASR) was provided by LIMSI / Vocapia**

* Stéphane Ayache and Georges Quénot. ***Video Corpus Annotation using Active Learning.*** In European Conference on Information Retrieval (ECIR), pages 187–198, Glasgow, Scotland, mar 2008.

** Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. ***The LIMSI Broadcast News transcription system.*** *Speech Communication*, 37(1-2):89–108, 2002.

Concepts

- 500 target concepts selected
  - TRECVid “high level features” from 2005 to 2010 to favor cross-collection experiments
  - Completed by a set of LSCOM concepts for covering a number of potential subtasks, e.g. “persons” or “actions” and for including a number of relations among concepts
  - These concepts were expected to be useful for the content-based (INS, AVS, MED …) search tasks
- One or two subsets were selected for evaluation each year

Relations

- Generic-Specific and exclusion relations among concepts for promoting research on methods for indexing many concepts and using ontology relations between them.
- Set of relations provided:
  - 427 “implies” relations, e.g. “Actor implies Person”
  - 559 “excludes” relations, e.g. “Daytime_Outdoor excludes Nighttime”
- Not exhaustive

30 Single concepts evaluated in 2015

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

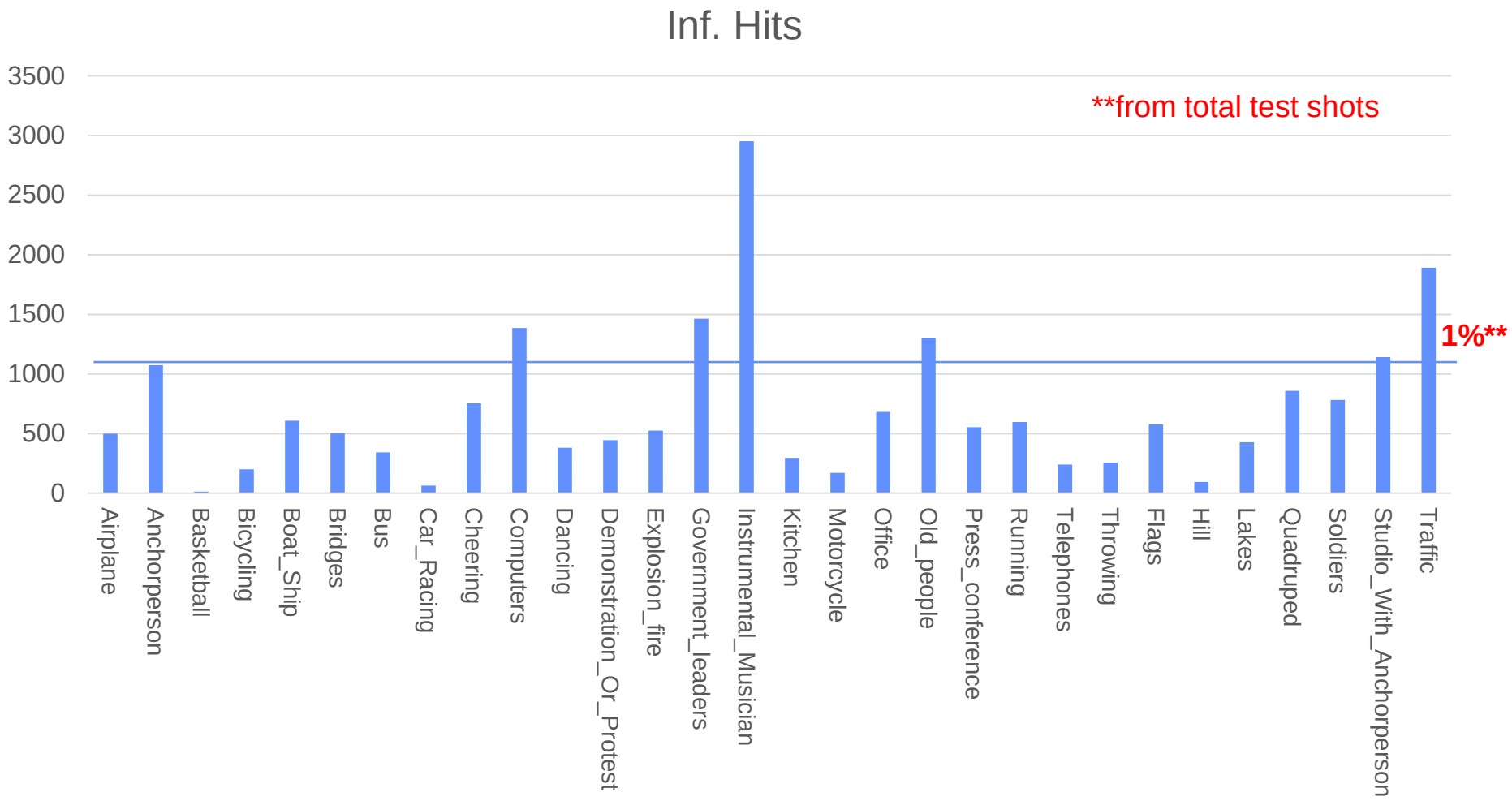
- Each feature assumed to be *binary*: absent or present for each master reference shot
- NIST sampled ranked pools and judged top results from all submissions
- Metrics: inferred Average Precision per concept
- Compared runs in terms of Mean inferred Average Precision across the 30 concept results for main runs.

Inferred average precision (infAP)

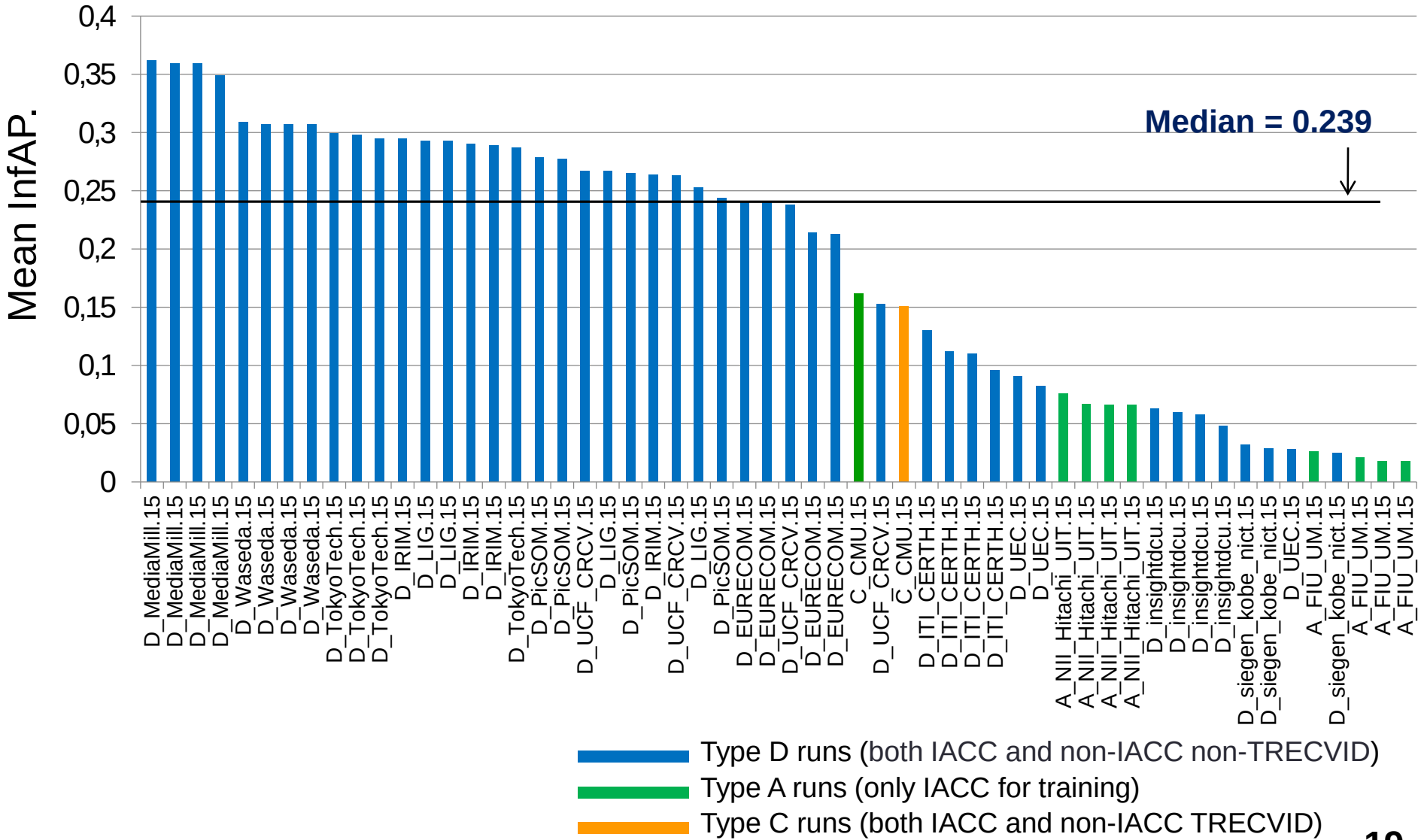
- Developed* by Emine Yilmaz and Javed A. Aslam at Northeastern University
- Estimates average precision surprisingly well using a surprisingly small sample of judgments from the usual submission pools
- More features can be judged with same effort
- Increased sensitivity to lower ranks
- Experiments 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.

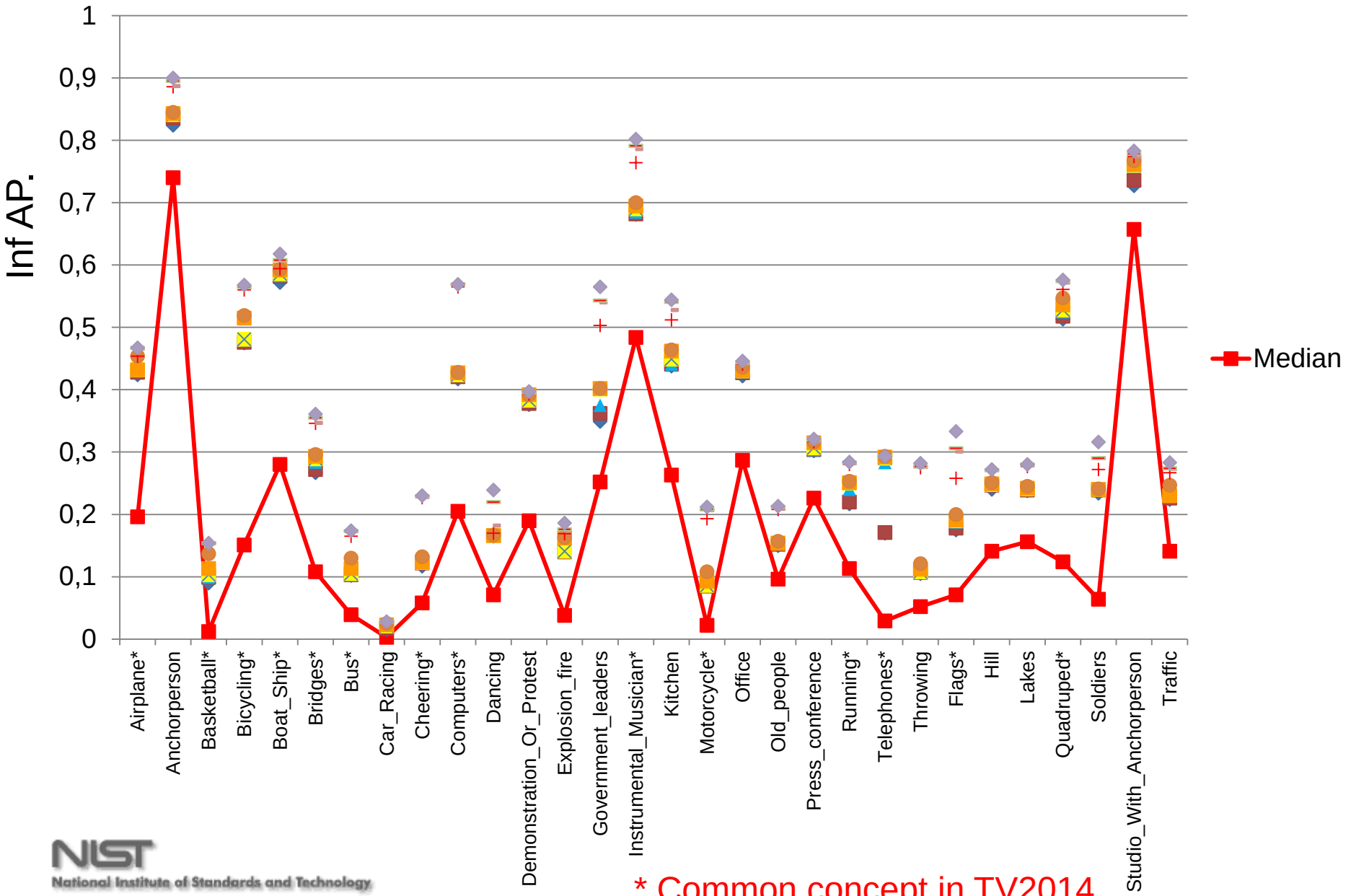
Inferred frequency of hits varies by concept



Main runs scores – 2015 submissions



Top 10 InfAP scores by concept



NIST
National Institute of Standards and Technology

* Common concept in TV2014

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

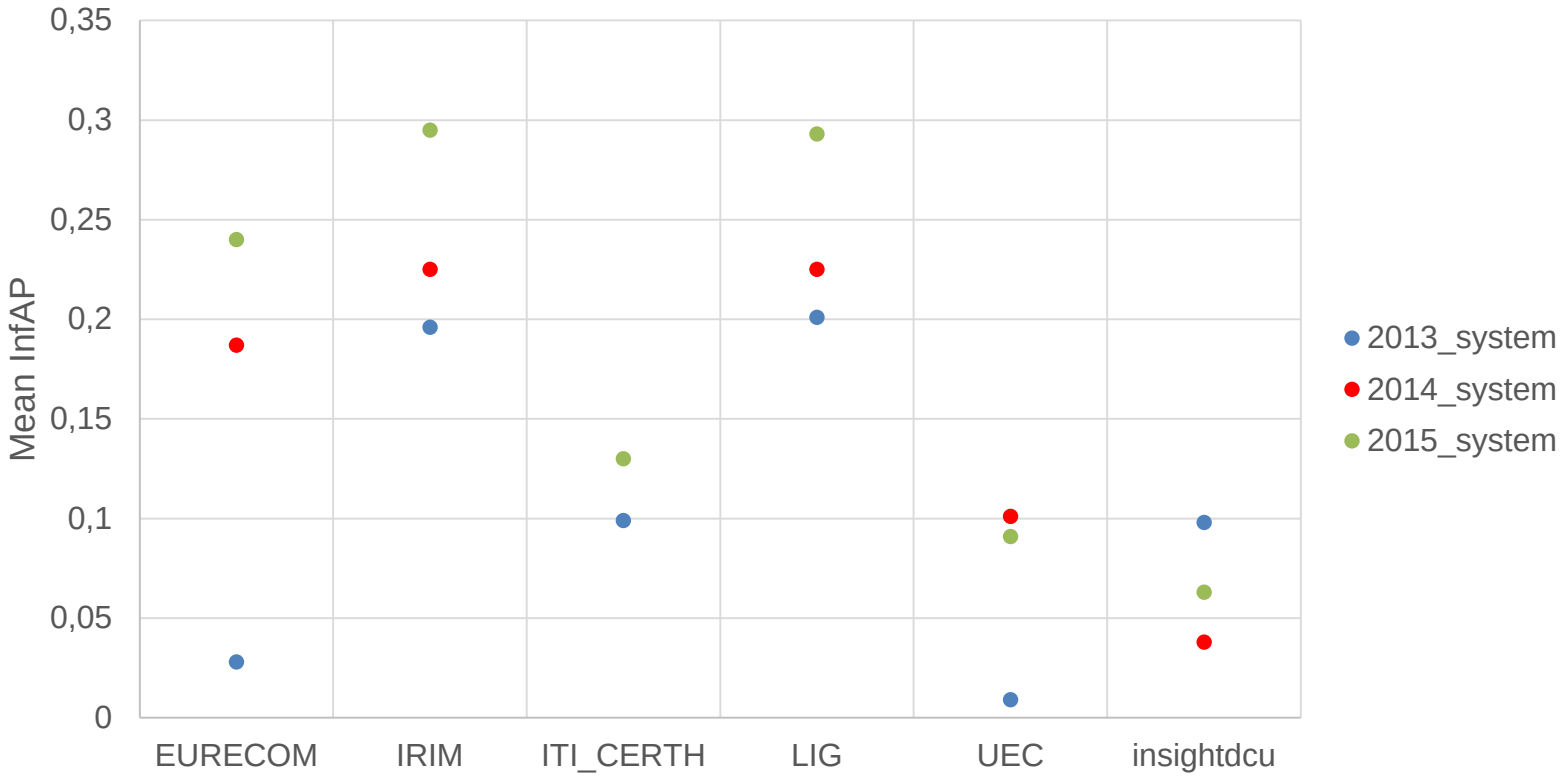
- D_MediaMill.15_1
  - D_MediaMill.15_3
    - D_Waseda.15_1
      - D_Waseda.15_3
      - D_Waseda.15_4
    - D_Waseda.15_2
    - D_TokyoTech.15_1
    - D_TokyoTech.15_2

- 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)

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.

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

TRECVid SIN versus ImageNet LSVRC

- Video shots
- Moderate resolution
- Non-exclusive labels
- Hierarchy of labels
- Highly imbalanced labels
- Examples “from the wild”
- Find shots for a label
- Average precision at 2000
- Platt’s normalization

- Still images
- Medium to high resolution
- Exclusive labels
- Only leaves
- Well balanced labels
- Typical examples
- Find labels for an image
- Top-N error (N = 1 or 5)
- Soft-max normalization

PART III
The LIG / IRIM “baseline”
From “bag of things” to deep learning

M. Budnik, E.-L. Gutierrez-Gomez, B. Safadi, D. Pellerin, G. Quénot. ***Learned features versus engineered features for multimedia indexing***. Multimedia Tools and Applications, (2017) 76: 11941.

doi:10.1007/s11042-016-4240-2

https://link.springer.com/article/10.1007/s11042-016-4240-21

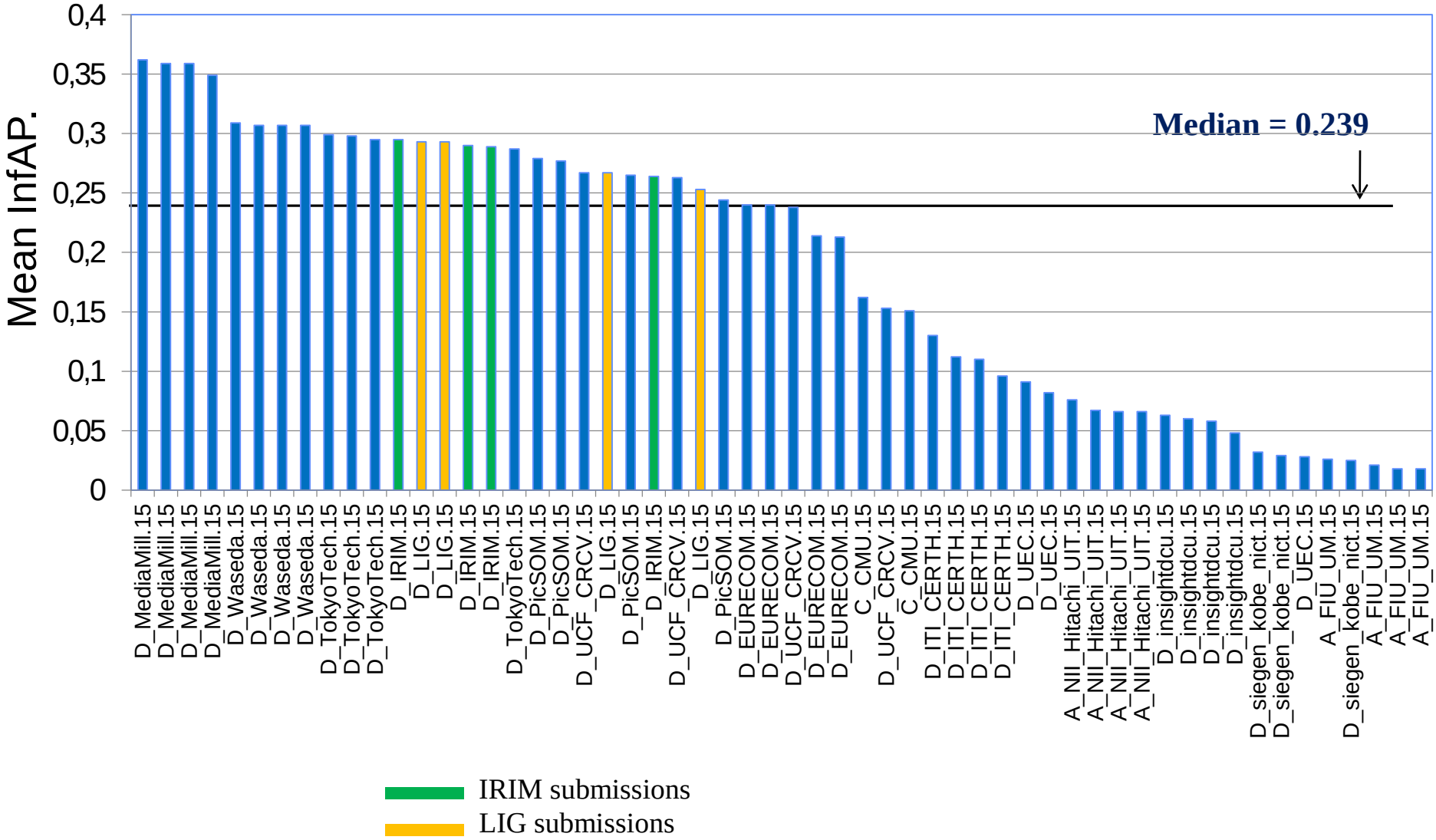
IRIM

- Multimedia Information Indexing and Retrieval group of GDR 720 ISIS (CNRS)
- 11 participants to the SIN task 2012-2015:
  - CEA-LIST, ETIS, Eurecom, INRIA-TEXMEX, LABRI, LIF, LIG, LIMSI-TLP, LIP6, LIRIS, LISTIC
- Contributed descriptors (features), classification scores, fusion results or methods and more

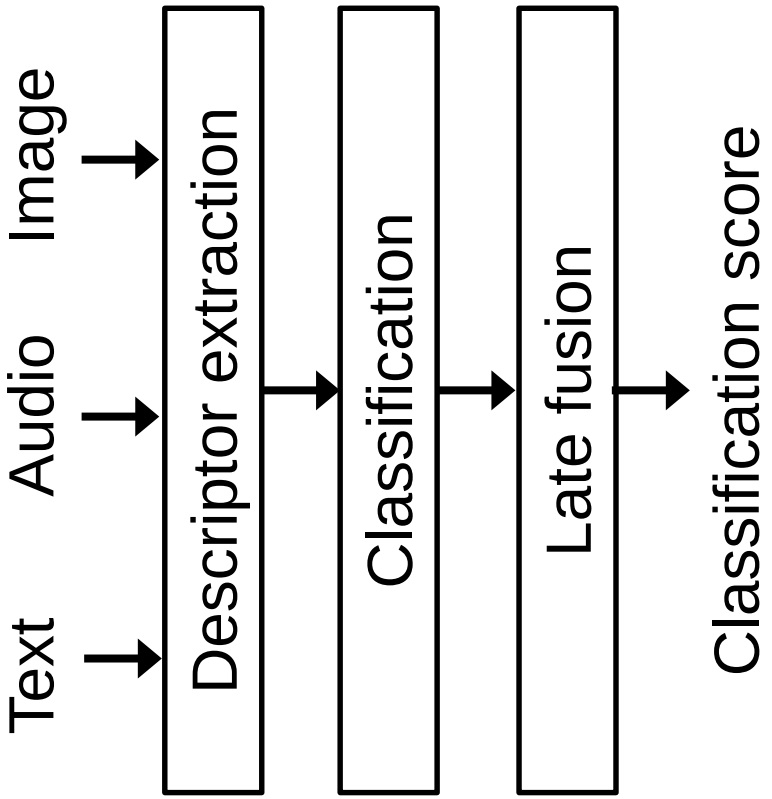
Outline

- System overview and features
- Contrast experiments
  - Engineered versus learned features
  - Temporal re-scoring
  - Conceptual feedback
  - I-frames
  - Retrained versus use of features
- Conclusion

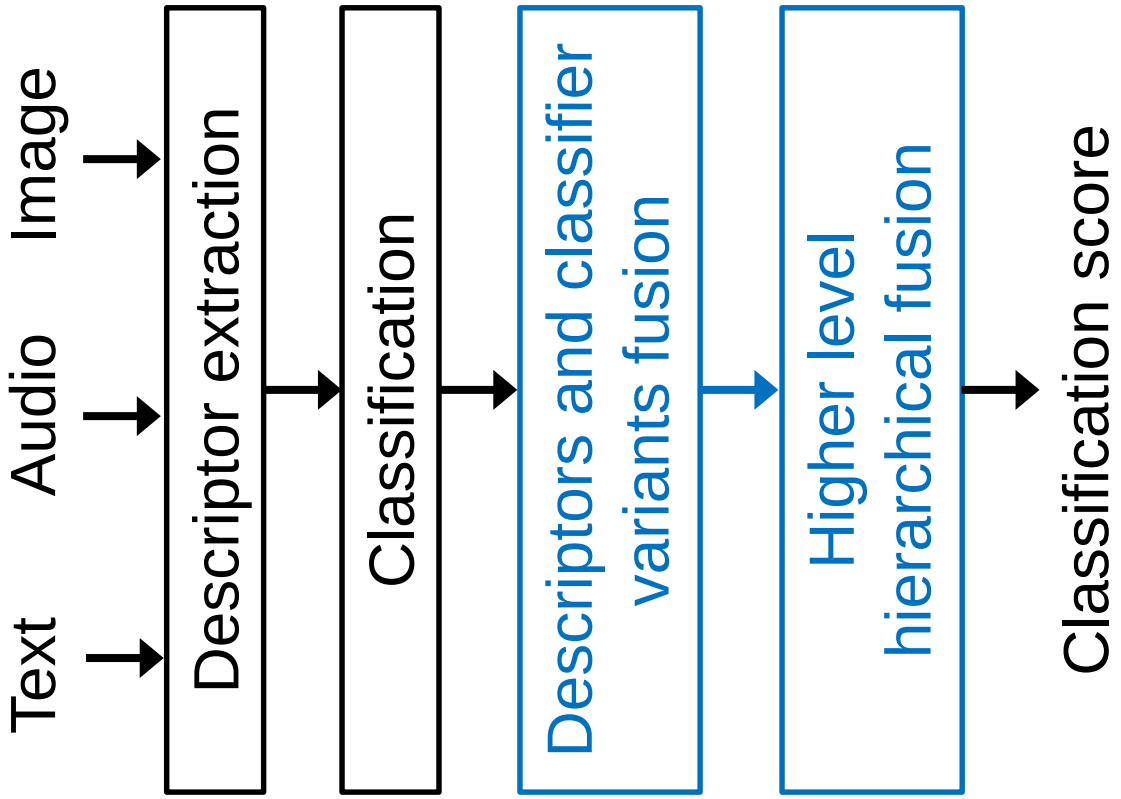
Main runs scores 2015 (from NIST)



Basic classification pipeline

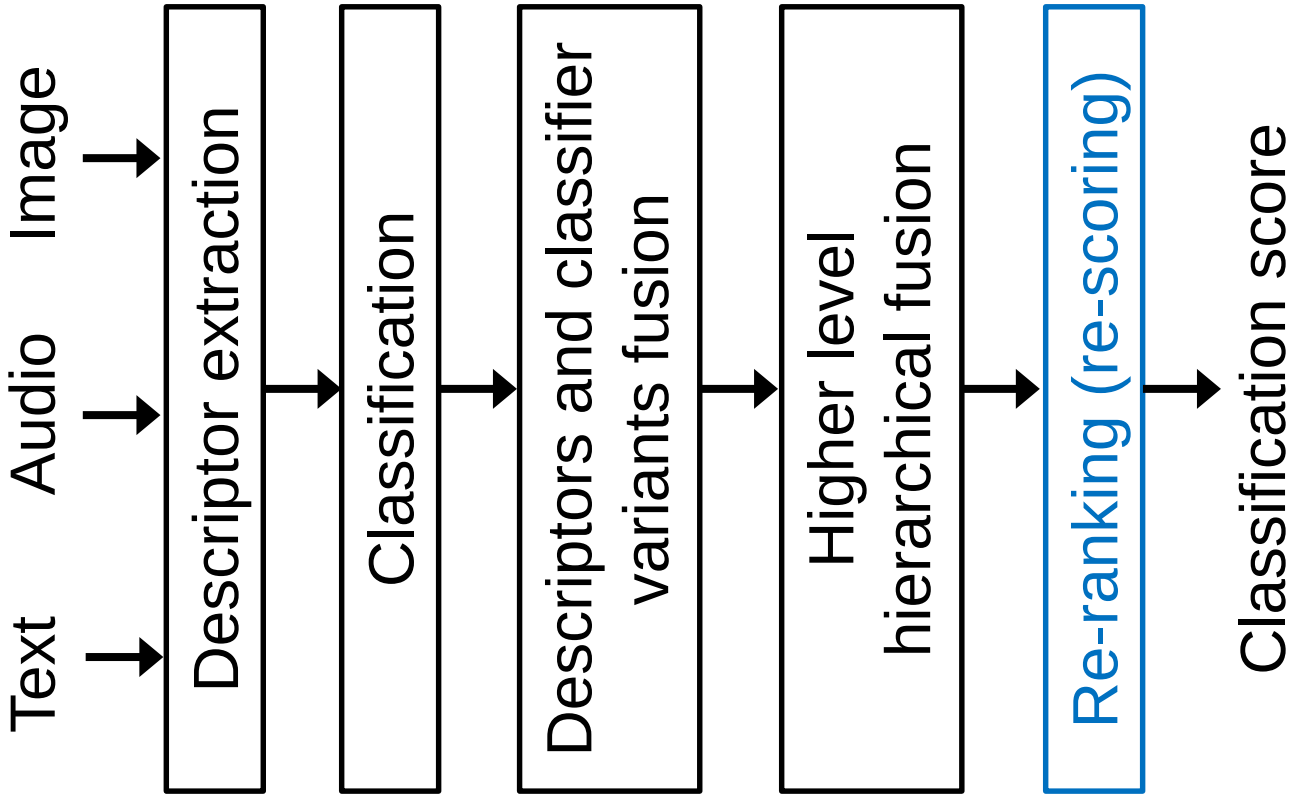


LIG/Quaero/IRIM classification pipeline



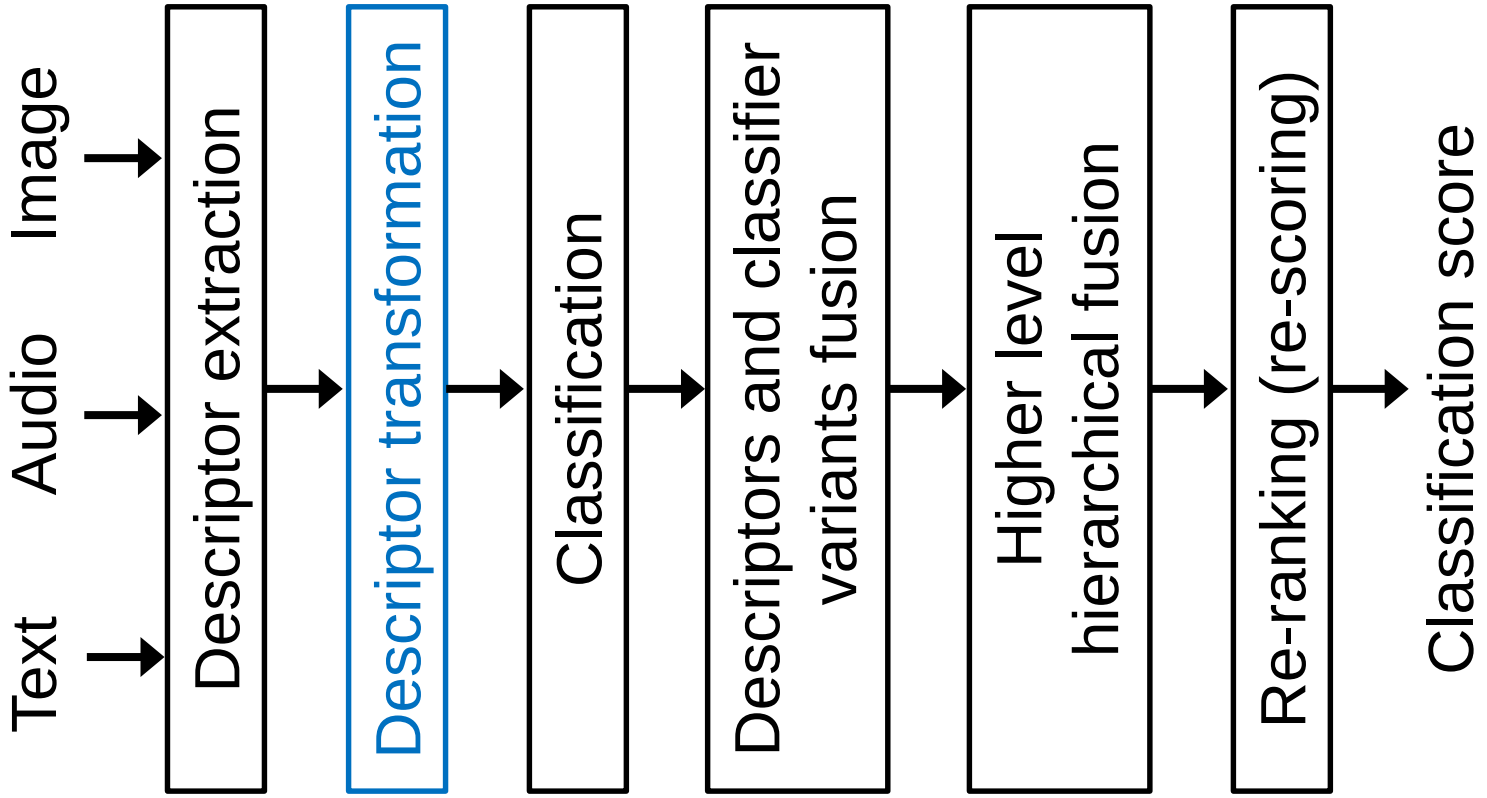
+ hierarchical fusion [Strat et al., ECCV/IFCVCR workshop 2012, Springer 2014]

LIG/Quaero/IRIM classification pipeline



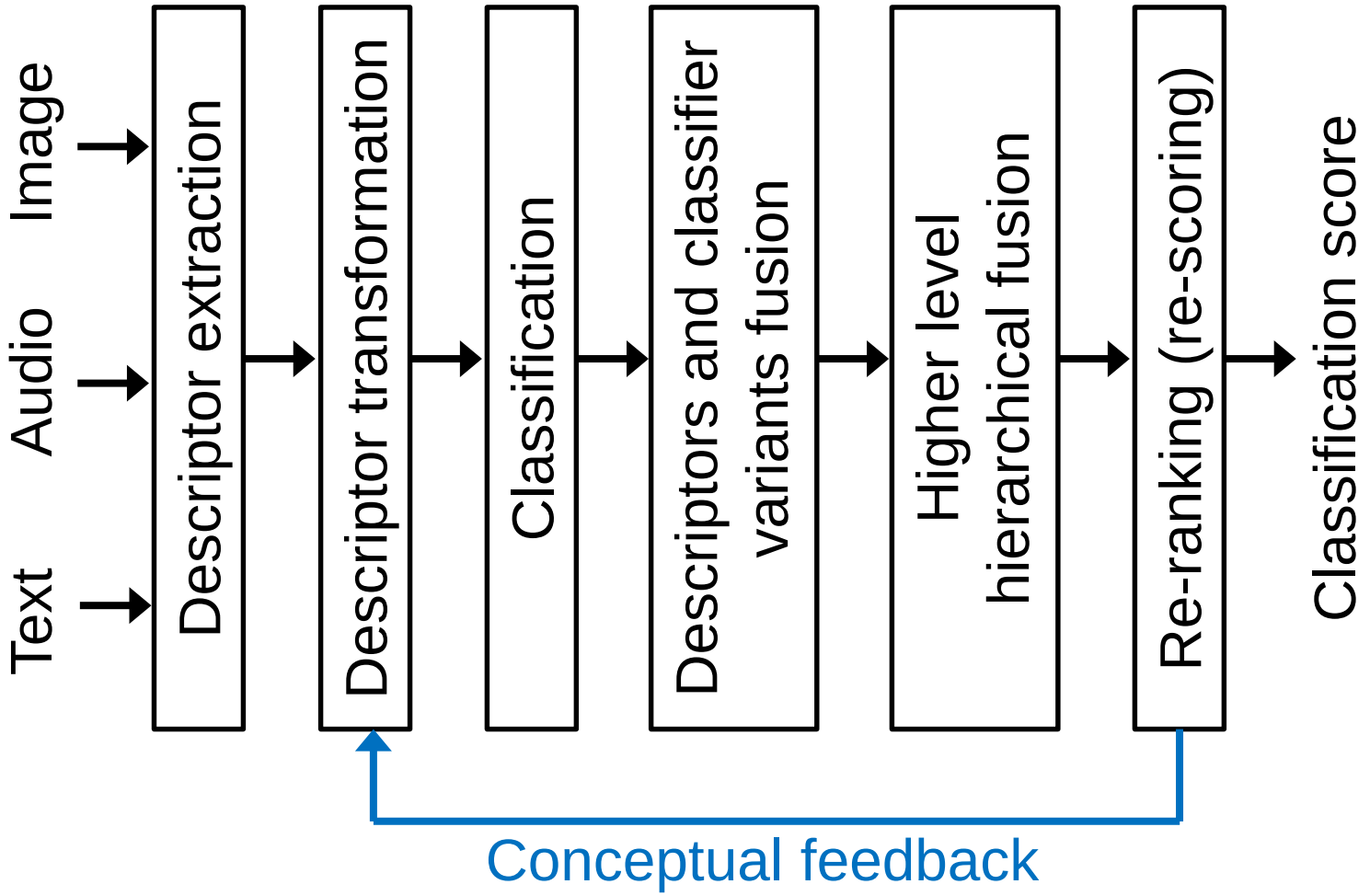
+ Temporal re-ranking [Safadi et al., CIKM 2011; Wang et al, TV 2009]: update shot scores considering other shots' scores for a same concept

LIG/Quaero/IRIM classification pipeline



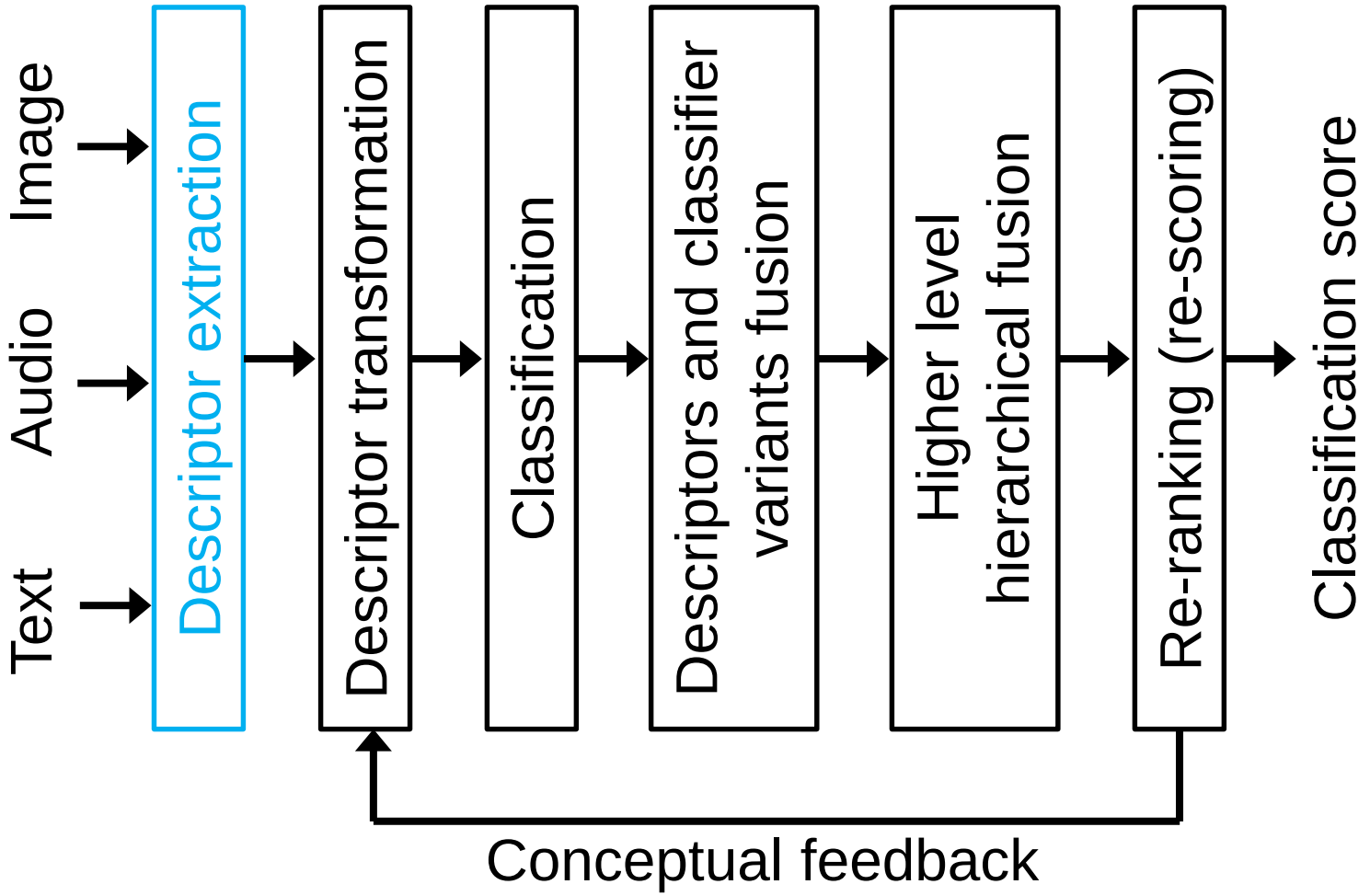
+ Descriptor optimization [Safadi et al., MTAP 2015]: combination of PCA-based dimensionality reduction and pre- and post- power transformations

LIG/Quaero/IRIM classification pipeline



+ conceptual feedback [Hamadi et al., MTAP, 2015]

LIG/Quaero/IRIM classification pipeline



+ semantic descriptors [TRECVid 2013 and 2014]

IRIM contributed “engineered” features

- “Top 6” types by cross-validation:
  - LIRIS OC-LBP
  - LIG BoW opponent SIFT
  - CEA-LIST pyramidal BoW dense SIFT
  - ETIS pyramidal BoW lab colors and quaternionic wavelets
  - LISTIC BoW retina SIFT
  - ETIS Vectors of Locally Aggregated Tensors
- Seven more from Labri, LIF, LIG, LIRIS and LISTIC (including audio and motion)

Xerox “engineered” semantic features

- Fisher Vector based descriptor [Perronnin, IJCV 2013]:
  - XEROX/ilsvrc2010: vectors of 1000 scores trained on ILSVRC10 and applied to key frames, kindly produced by Florent Perronnin from Xerox (XRCE)
  - XEROX/imagenet10174: same with10174 concepts scores trained ImageNet

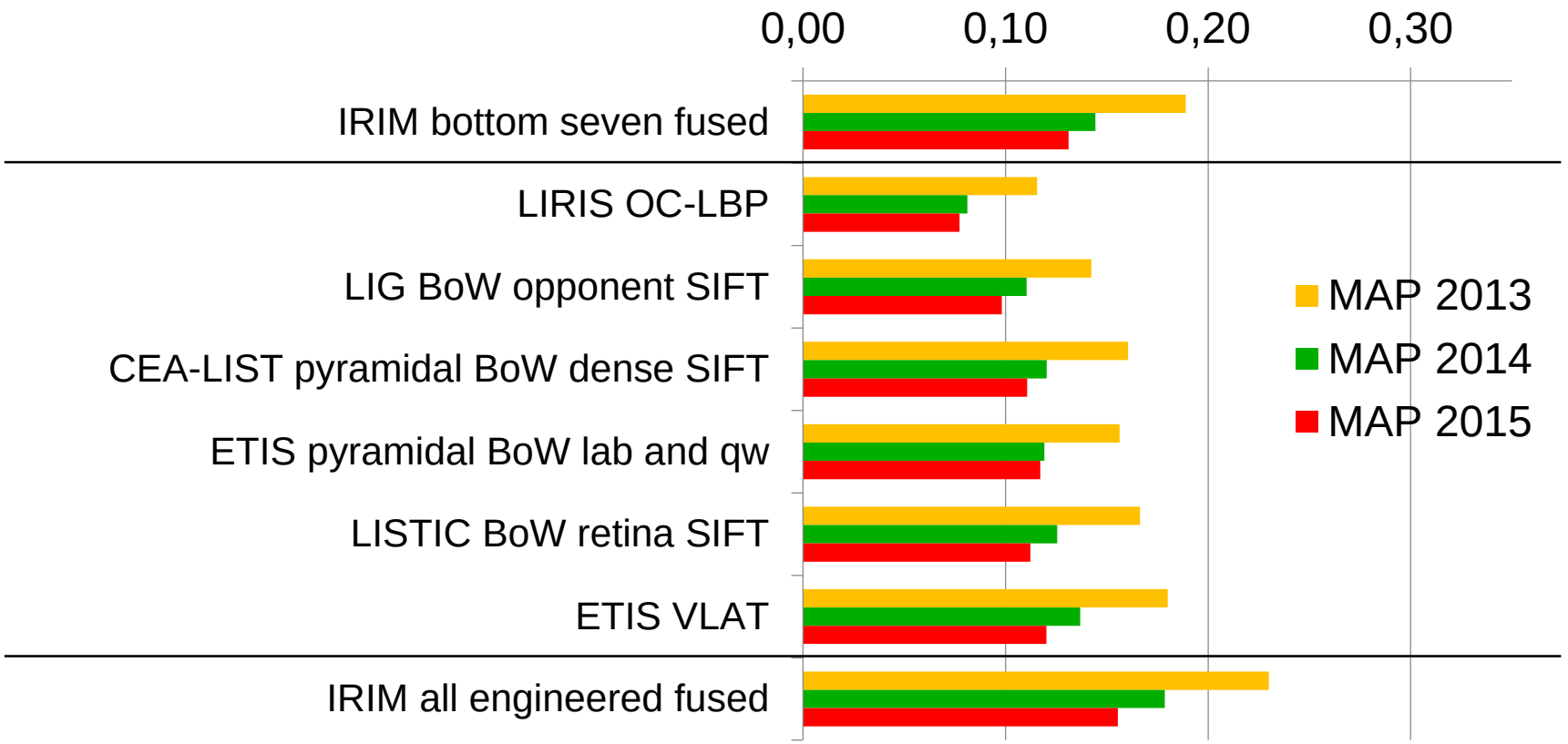
“Deep” (learned) semantic features

- All extracted by LIG using Berkeley caffe tool [Jia et al, 2013] and provided pre-trained models:
  - AlexNet [Krizhevsky et al., 2012] final output layer and last three hidden layer outputs (before normalization or ReLU)
  - GoogLeNet [Szegedy et al., 2014] last hidden layer output (before RELU)
  - VGG-19 [Simonyan and Zisserman, 2014] final output layer (before normalization)

Experiments

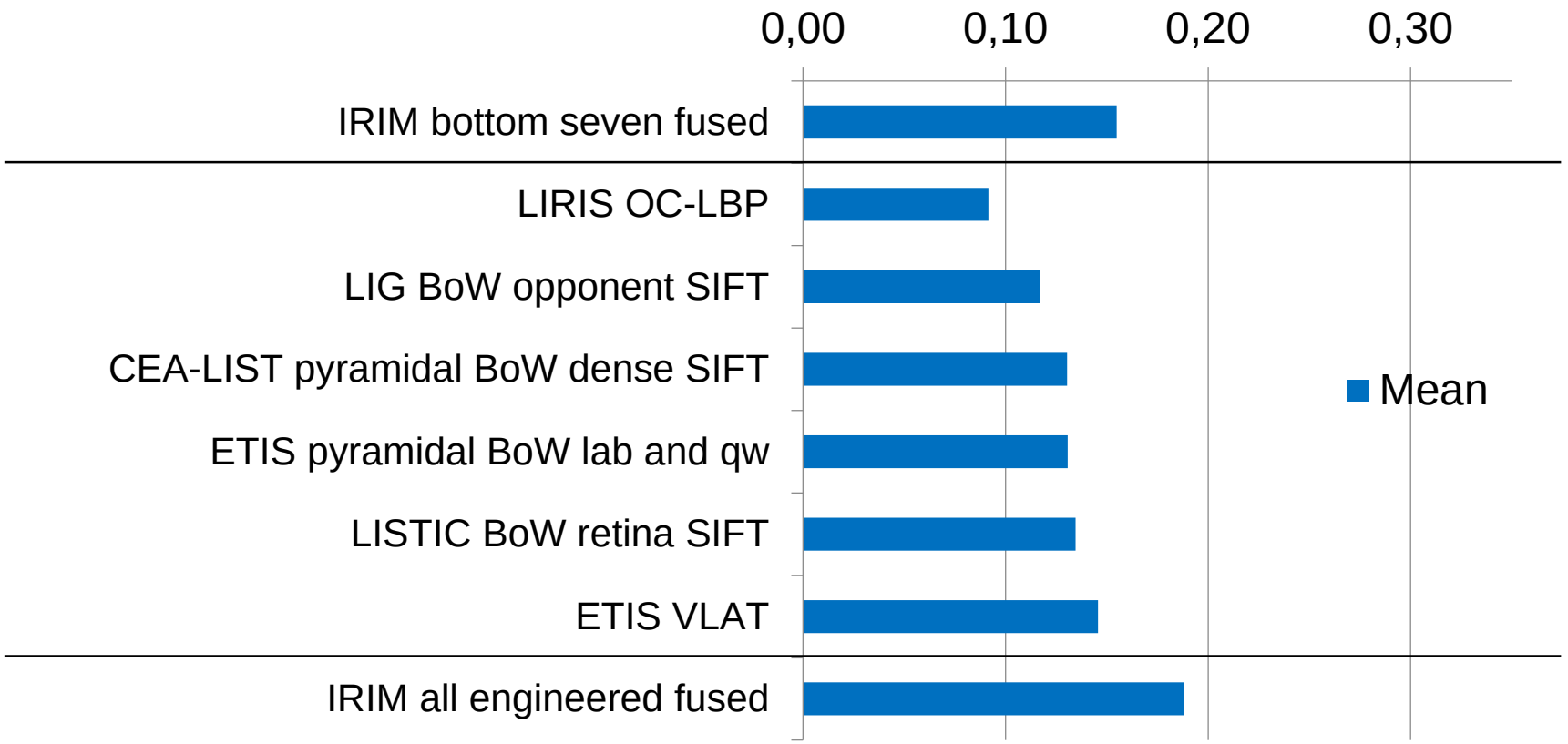
- Use of SIN 2013 development data only (no tuning on SIN 2013 or 2014 test data) and
- Various components using indirectly ImageNet annotated data → D type submissions
- Evaluation on SIN 2013, 2014 and 2015 test data
- Use of a combination of kNN and MSVM for classification [Safadi, RIAO 2010]
- Hierarchical late fusion in all combinations (linear combinations of classification scores)
- Unless otherwise noted, results are computed using only one key frame per shot
- Multiple frames: all I-frames with max pooling

Performance of low-level “engineered” features



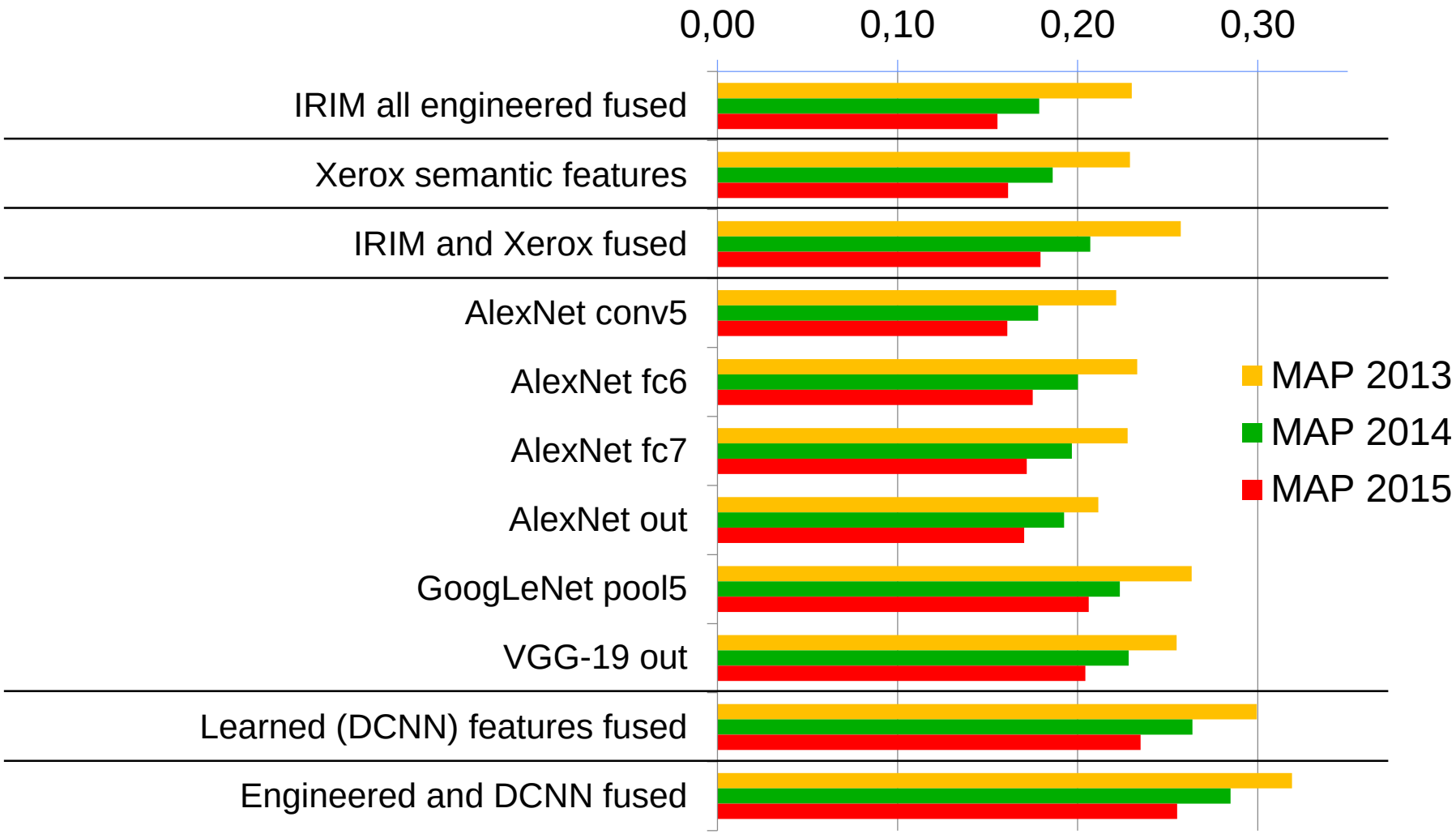
The more you add, the better performance you get

Performance of low-level “engineered” features

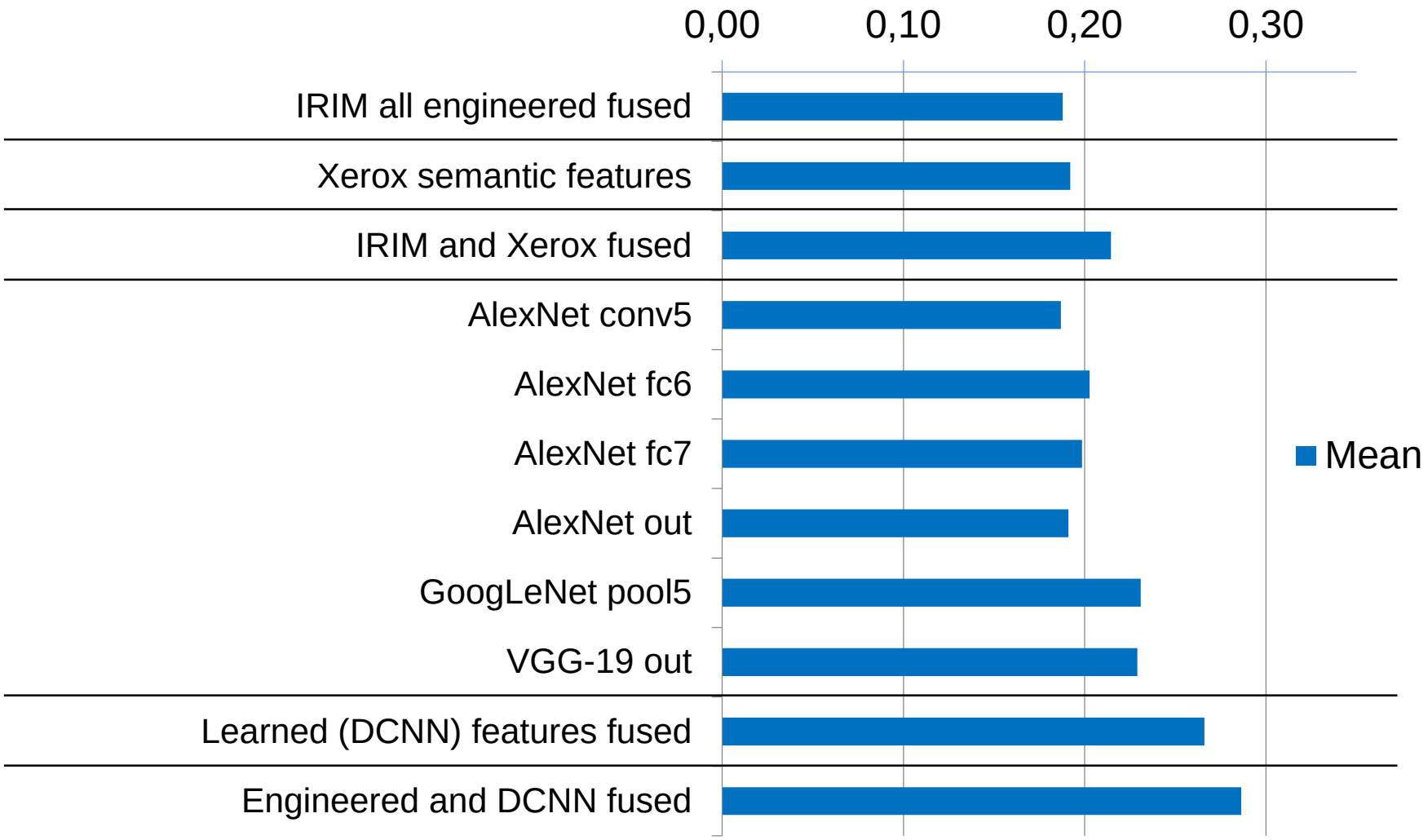


The more you add, the better performance you get

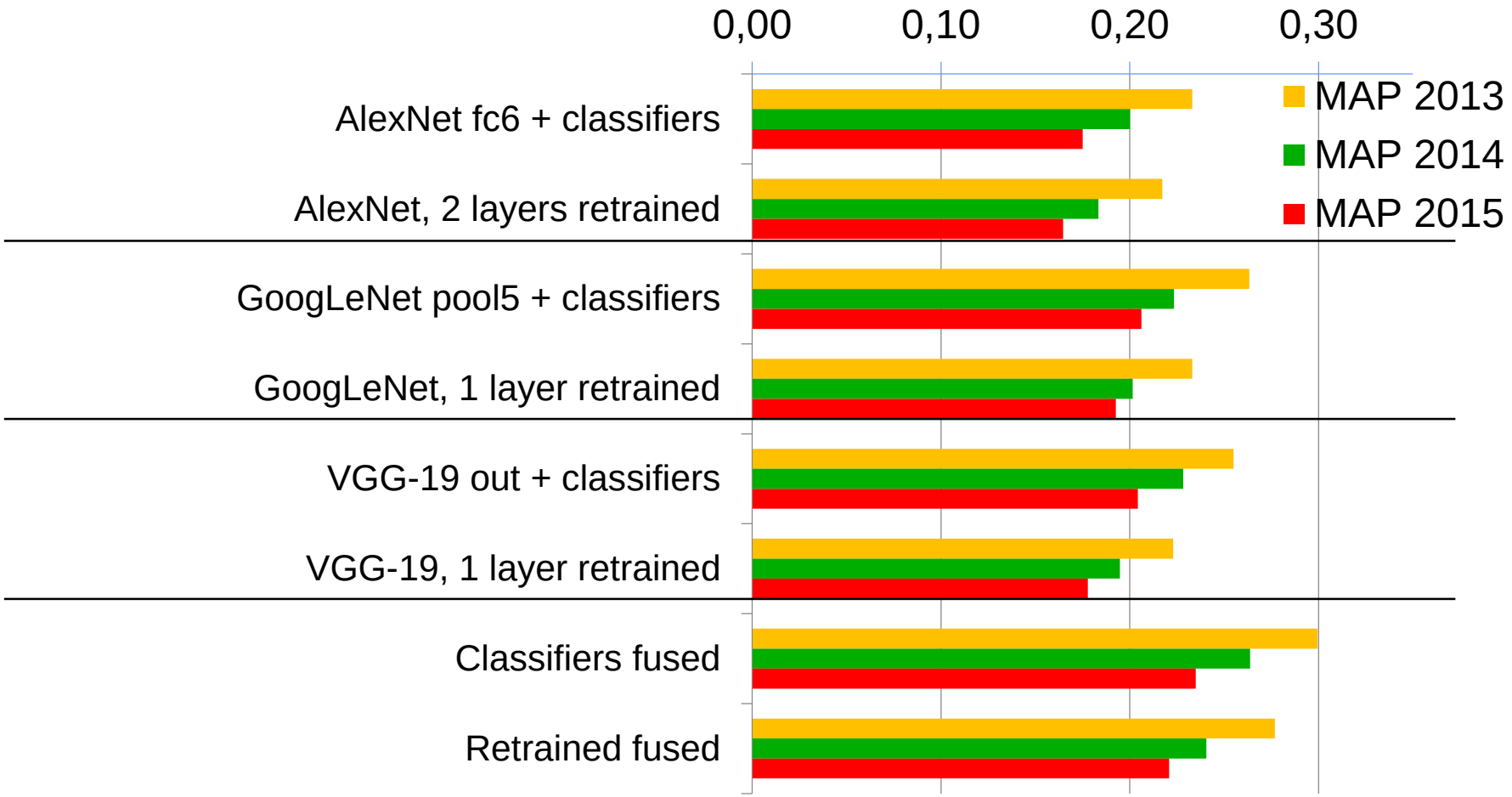
Performance of engineered and learned features



Performance of engineered and learned features, mean on 2013-2015

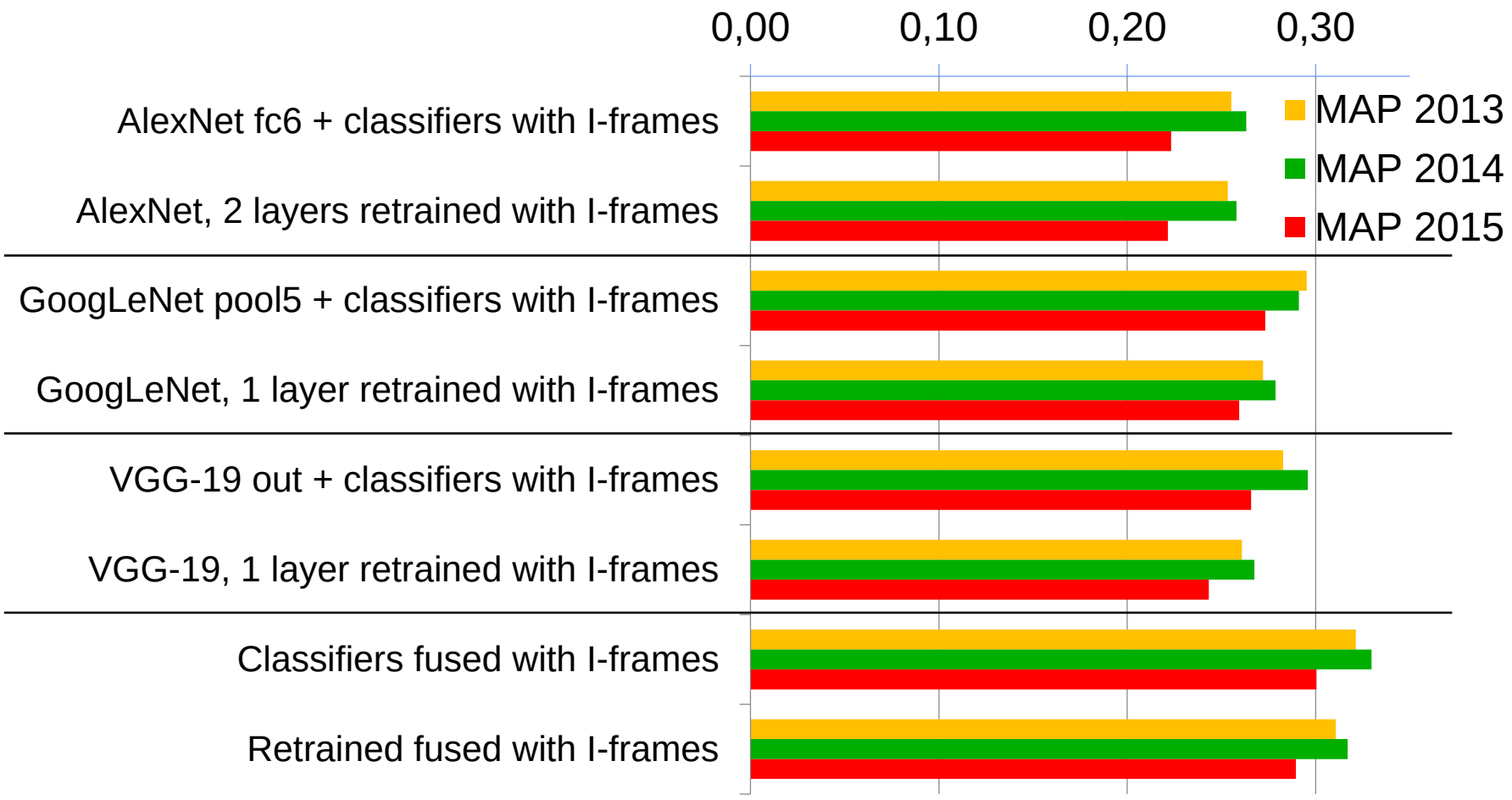


Use of extracted features versus network retraining



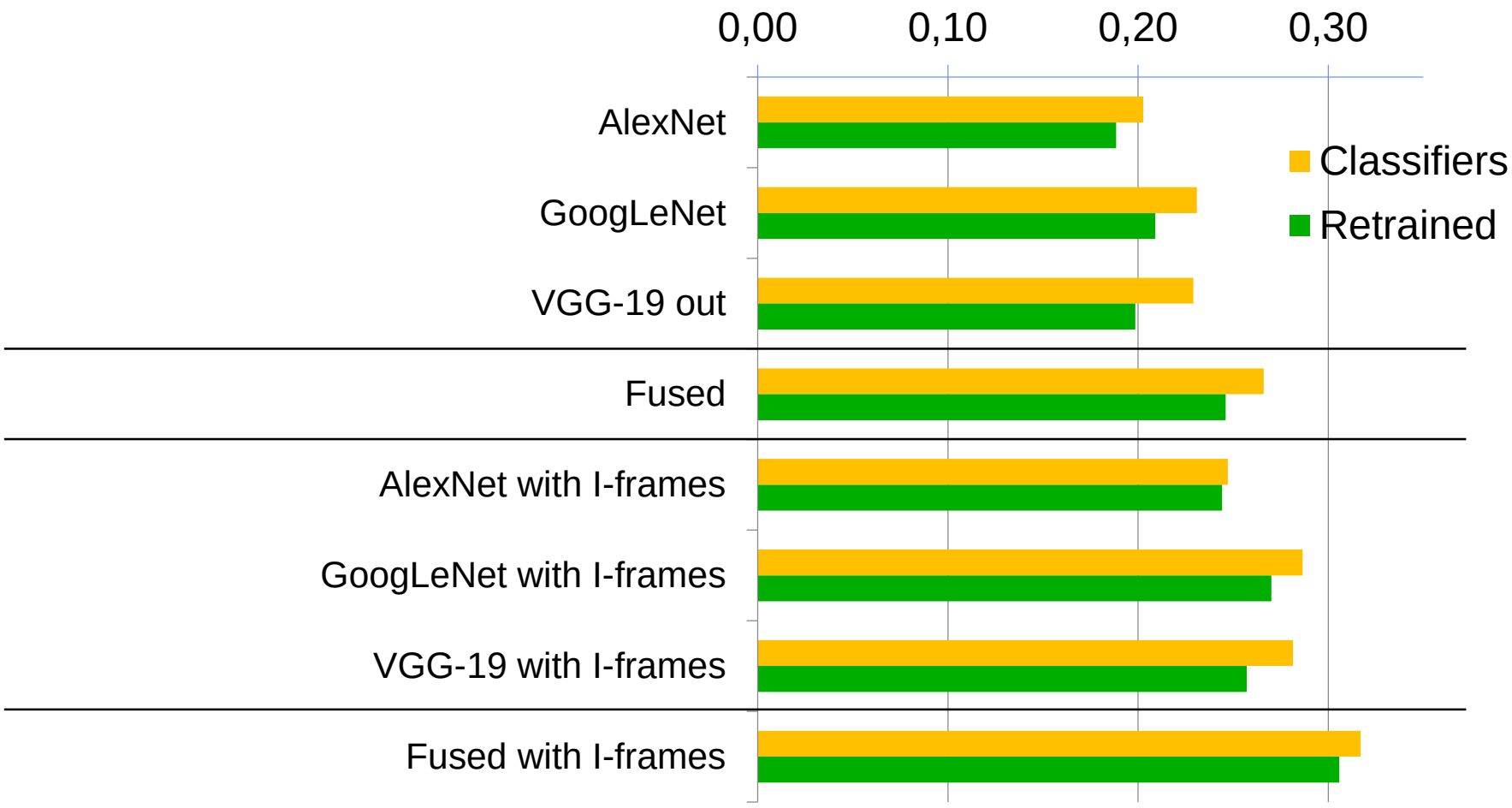
Re-training is always less good (not enough data?)

Use of extracted features versus network retraining with I-frames



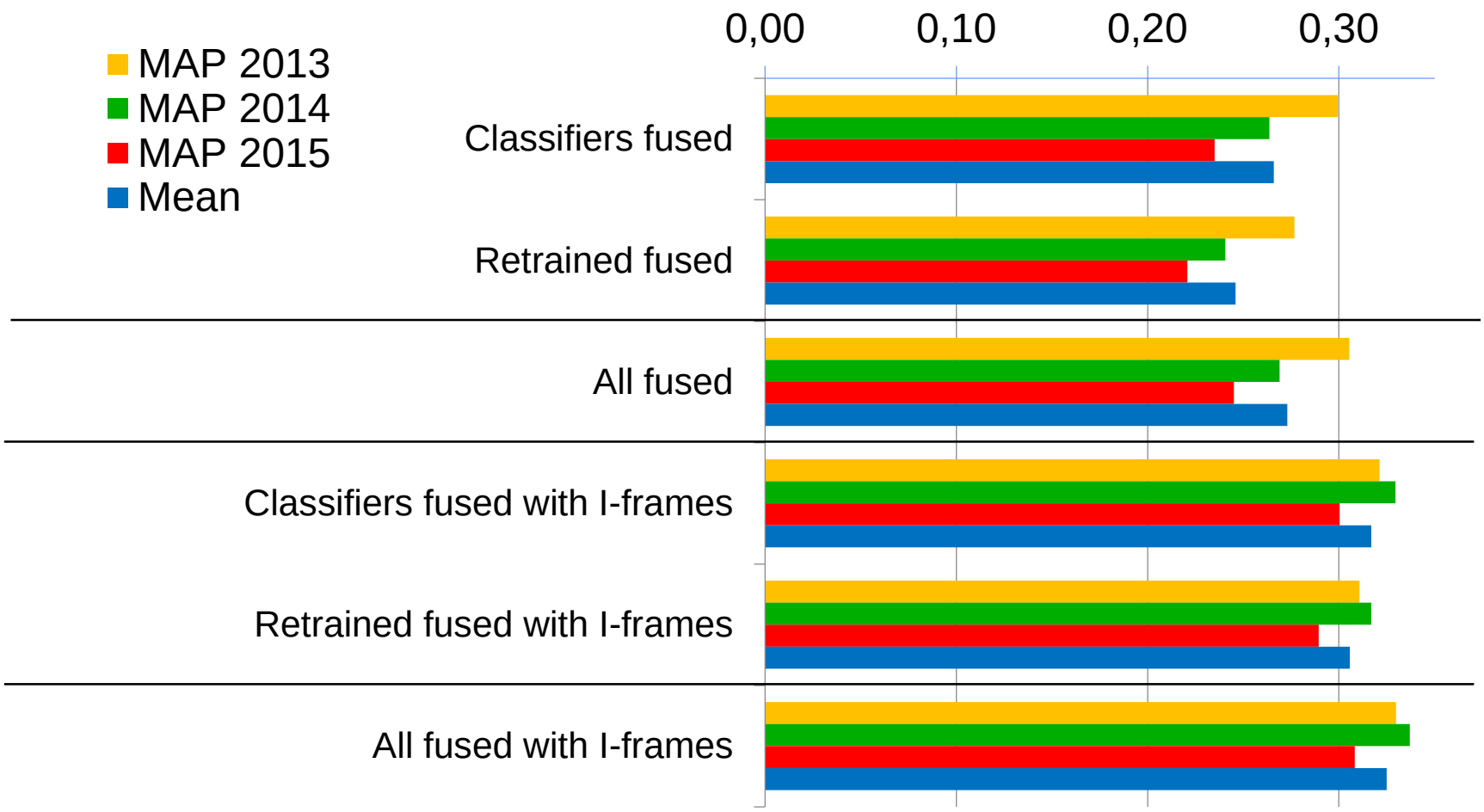
Less differences between collections with I-frames

Use of extracted features versus network retraining, mean 2013-2015



Less differences between methods with I-frames

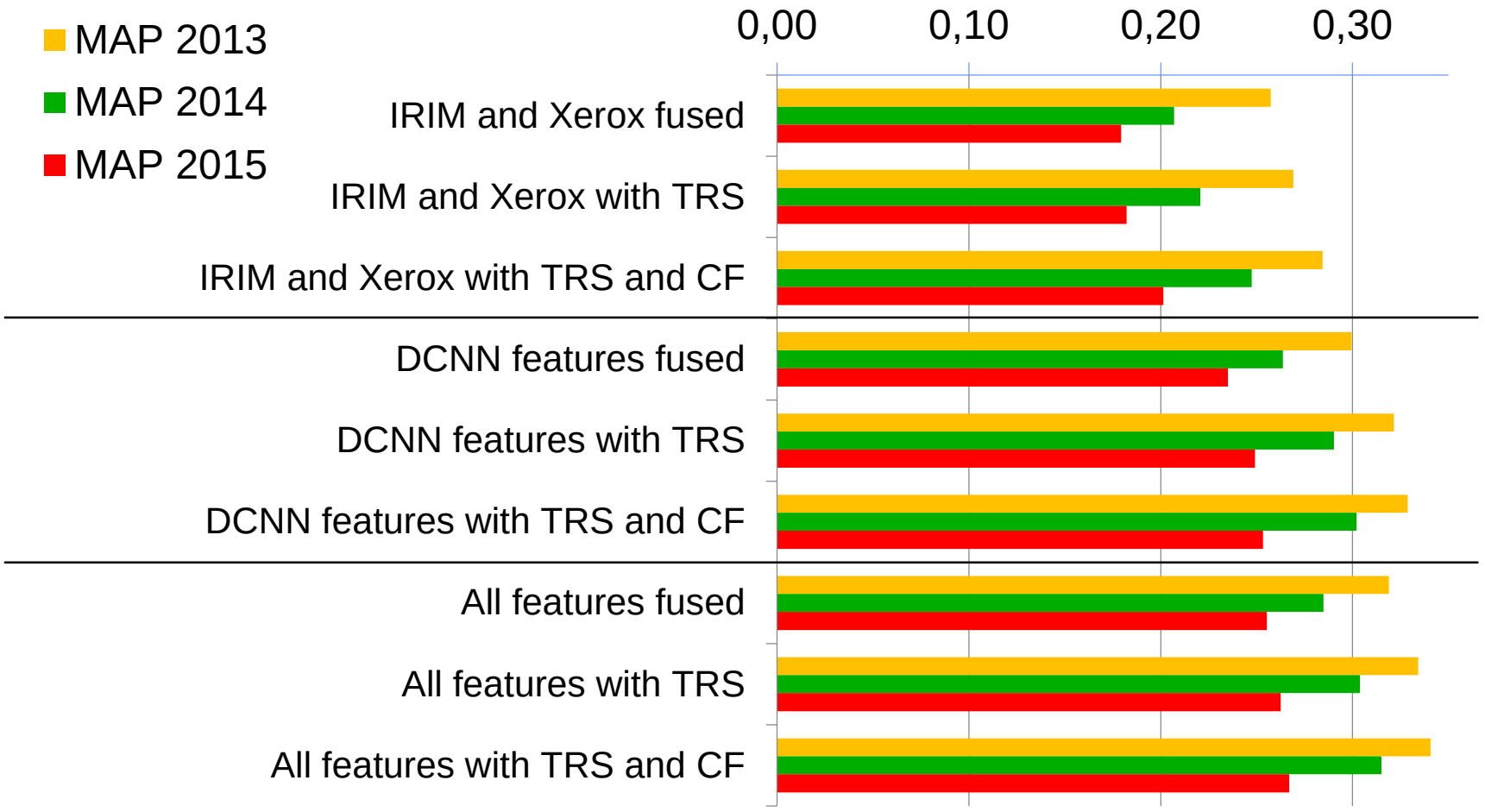
Use of I-frames alone (no TRS or CF)



Less differences between collections with I-frames

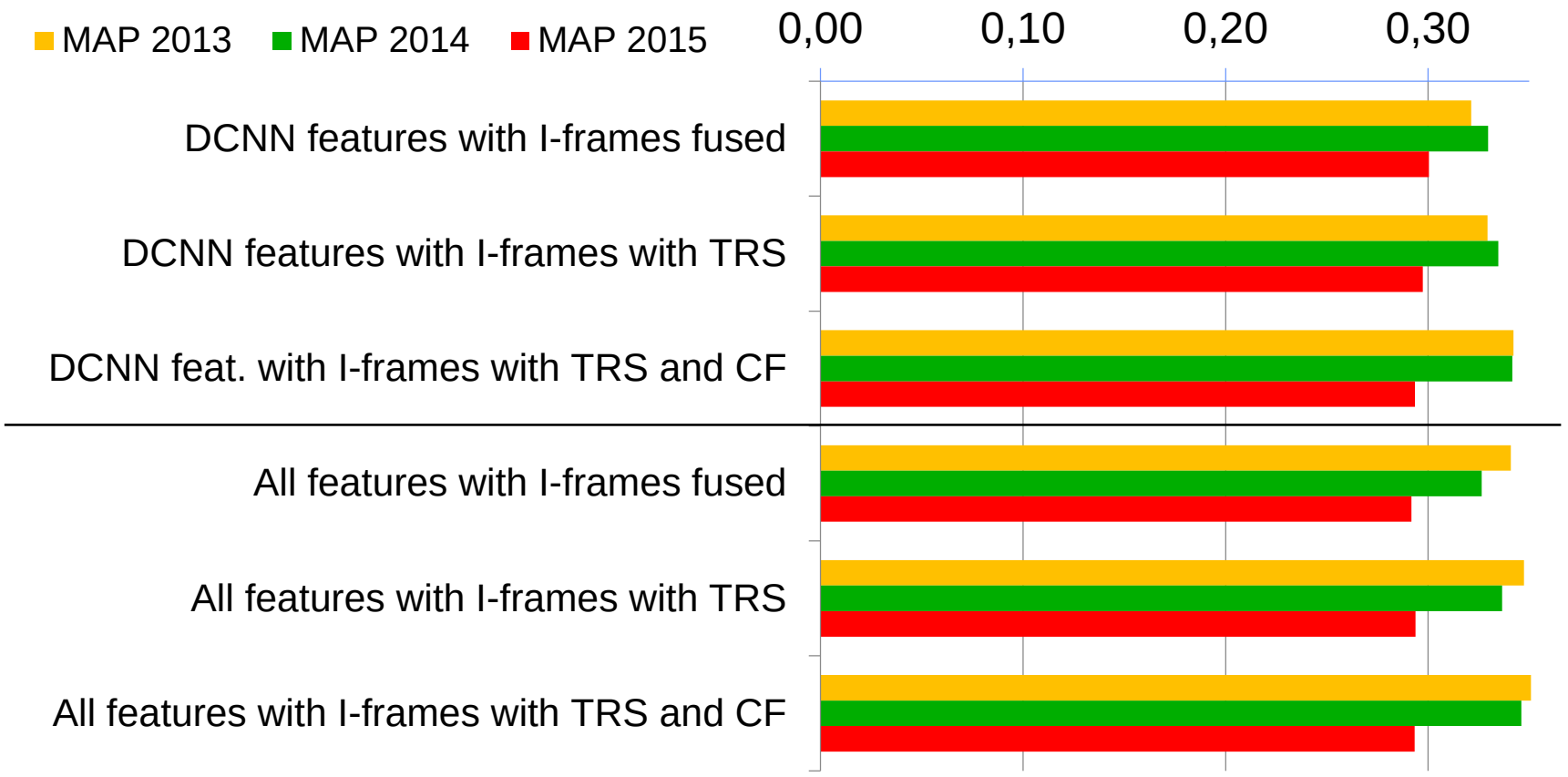
Gains of 8%, 25%, 26% and 19% on 2013, 2014, 2015 and mean MAPs respectively (all fused)

Use of temporal re-scoring (TRS) and conceptual feedback (CF)



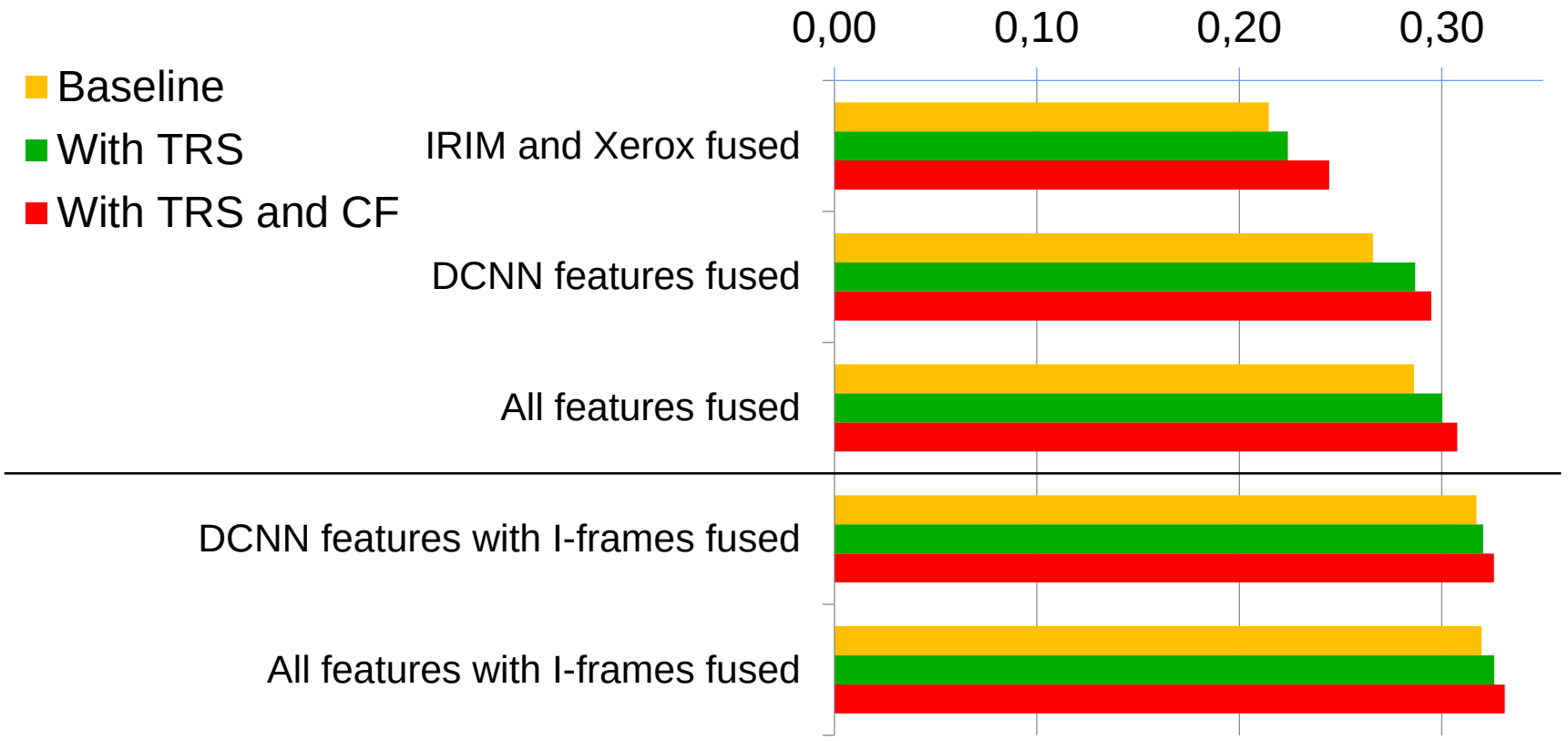
Both TRS and CF always improve performance

Use of temporal re-scoring (TRS) and conceptual feedback (CF) with I-frames



Small (if any) performance gain from TRS and CF with I-frames (redundancy between multiple shots and frames)

Use of (TRS) and (CF) mean on 2013-2015



Less improvement from TRS and CF with I-frames

Less improvement from engineered features with I-frames

Use of semantic features for the semantic indexing task

- Learned features are better than engineered ones but the combination of both is even better
- Applying a KNN/MSVM combination to extracted learned features does significantly better than retraining the deep networks they come from but fusion does even better
- Temporal re-ranking and conceptual feedback improves in all cases but not much when multiple frames are used
- Using multiple frames per shot significantly improves the performance but the gain does not cumulate well with temporal re-scoring (neighbor frames vs. neighbor shots)
- Most effective (and efficient) solution: multiple DCNN features with classical learning and use of I-frames
- All these results are relative to TRECVid SIN data and annotations (but this worked well at VOC 2012 too).

Is semantic indexing a solved problem?

- Huge progresses over 15 years
- Official TRECVid SIN Metric: MAP@2000
  - MAP on all is significantly higher though still far from 1
  - P@N is often very good even if the MAP is not
- Really impressive performance on some concepts
- Quite good performance on many concepts
- Still poor or very poor performance on many concepts
- Many useful concepts not available or easily derivable, e.g. “red taxi” (NTCIR-lifelog)

Is semantic indexing a solved problem?

- TRECVid SIN stopped after 2015
  - Decrease in the number of participants
  - Move toward other challenges like ImageNet
  - Most approaches rely on frame (still image) classification
  - The specificity of video shots compared to still images not really exploited beyond pooling on multiple frames despite some exceptions
  - Low effort in audio or motion analysis: not much benefit as quite few concepts benefit from them, also moved toward other challenges (e.g. Hollywood)

Thanks

Slides available from:

http://mrim.imag.fr/georges.quenot/icmr2017/SIN.pdf