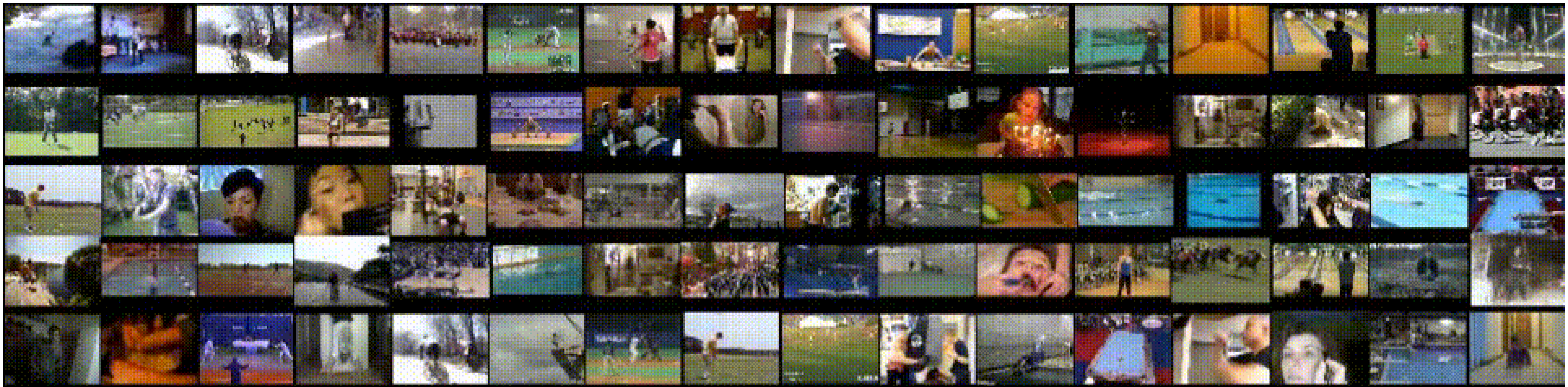


Overview of the detection challenge

THUMOS '14

Presented by Yu-Gang Jiang (Fudan University)
Zurich, Switzerland, Sept. 7th 2014

The THUMOS’13 Localization Challenge Dataset

- 24 action classes from UCF101; 3207 clips in total
- 3 training/test splits
  - 18 out of 25 groups for training, 7 for testing
- Selected 10 classes from UCF11 (a part of UCF101 that has local bbx annotations), annotated 14 more classes (1818 clips)
  - Hired 8 people; each spent ~40 hours

Basketball	BasketballDunk	Biking	CliffDiving
CricketBowling	Diving	Fencing	FloorGymnastics
GolfSwing	HorseRiding	IceDancing	LongJump
PoleVault	RopeClimbing	SalsaSpin	SkateBoarding
Skiing	Skijet	SoccerJuggling	Surfing
TennisSwing	TrampolineJumping	VolleyballSpiking	WalkingWithDog

Example Videos of THUMOS'13

- BasketballShooting

- BasketballDunk

The number of localization submissions we received in 2013:

0

The number of submissions we received in 2014:

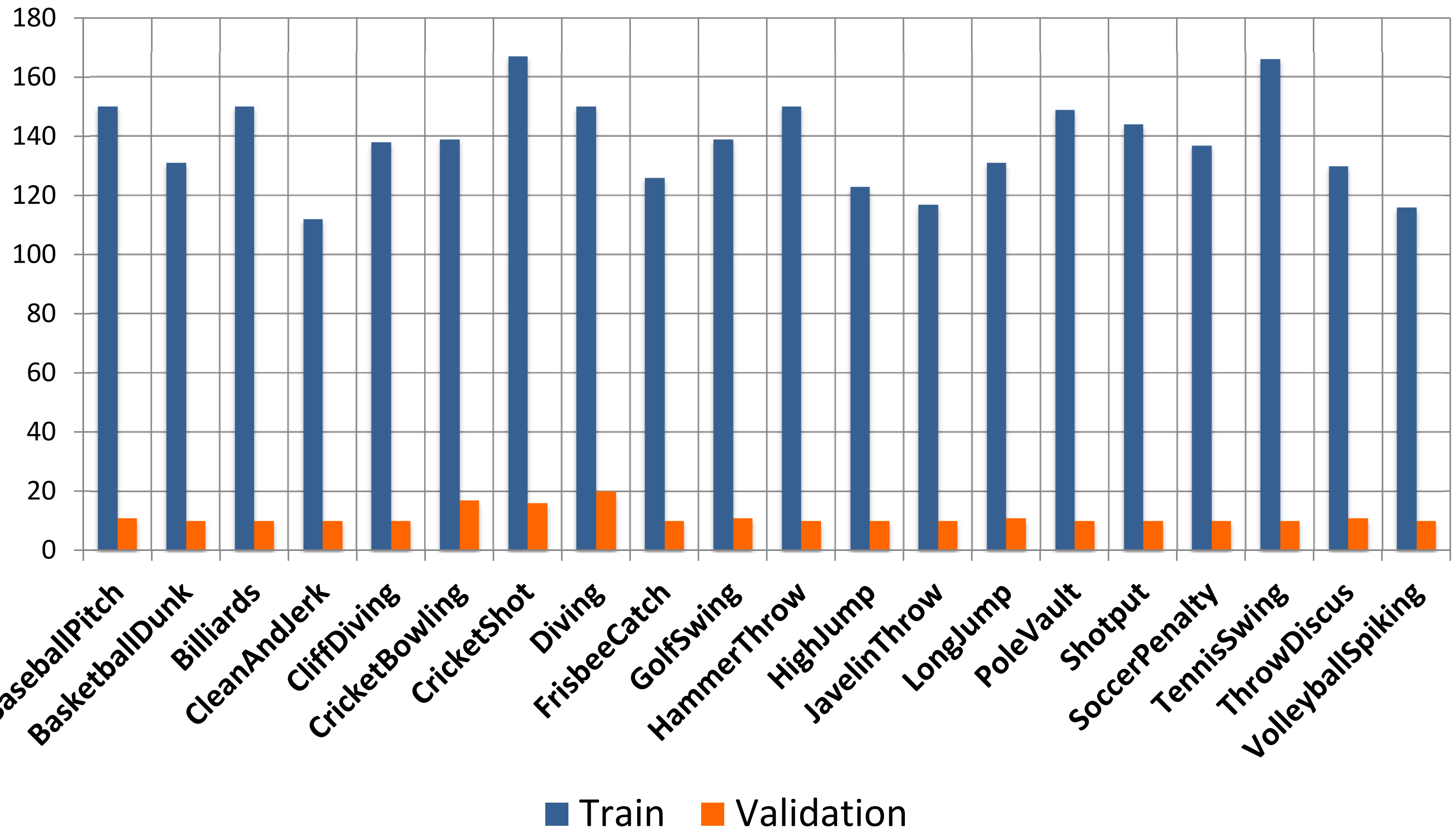
11 runs from 3 teams!!!

Changes in 2014

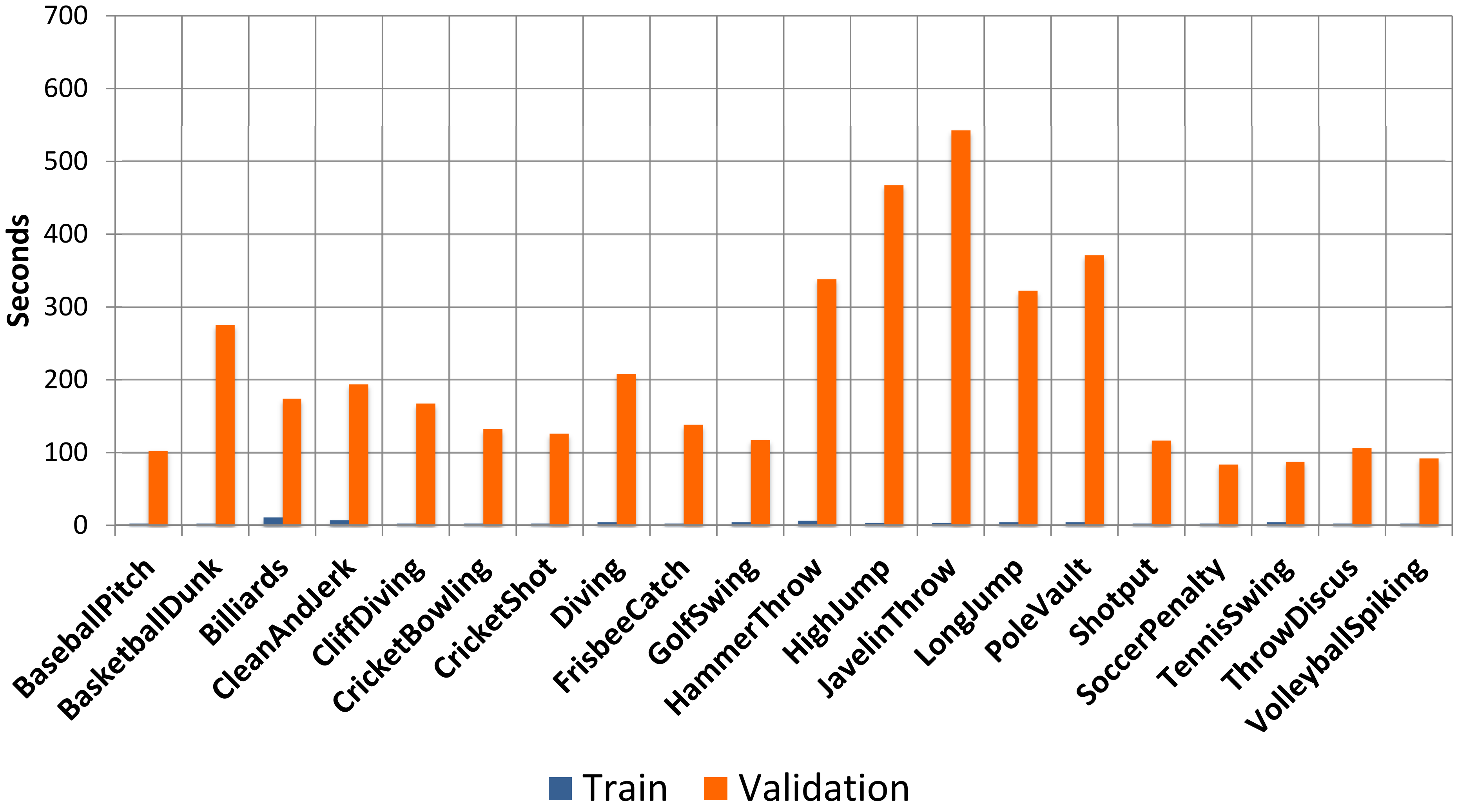
- Switched from spatial-temporal localization to temporal localization.
  - Temporal boundaries are more important
  - Lower computational complexity
  - Annotation is cheaper

- Adopted temporally untrimmed videos for validation and testing; UCF101 was still used, for training only.

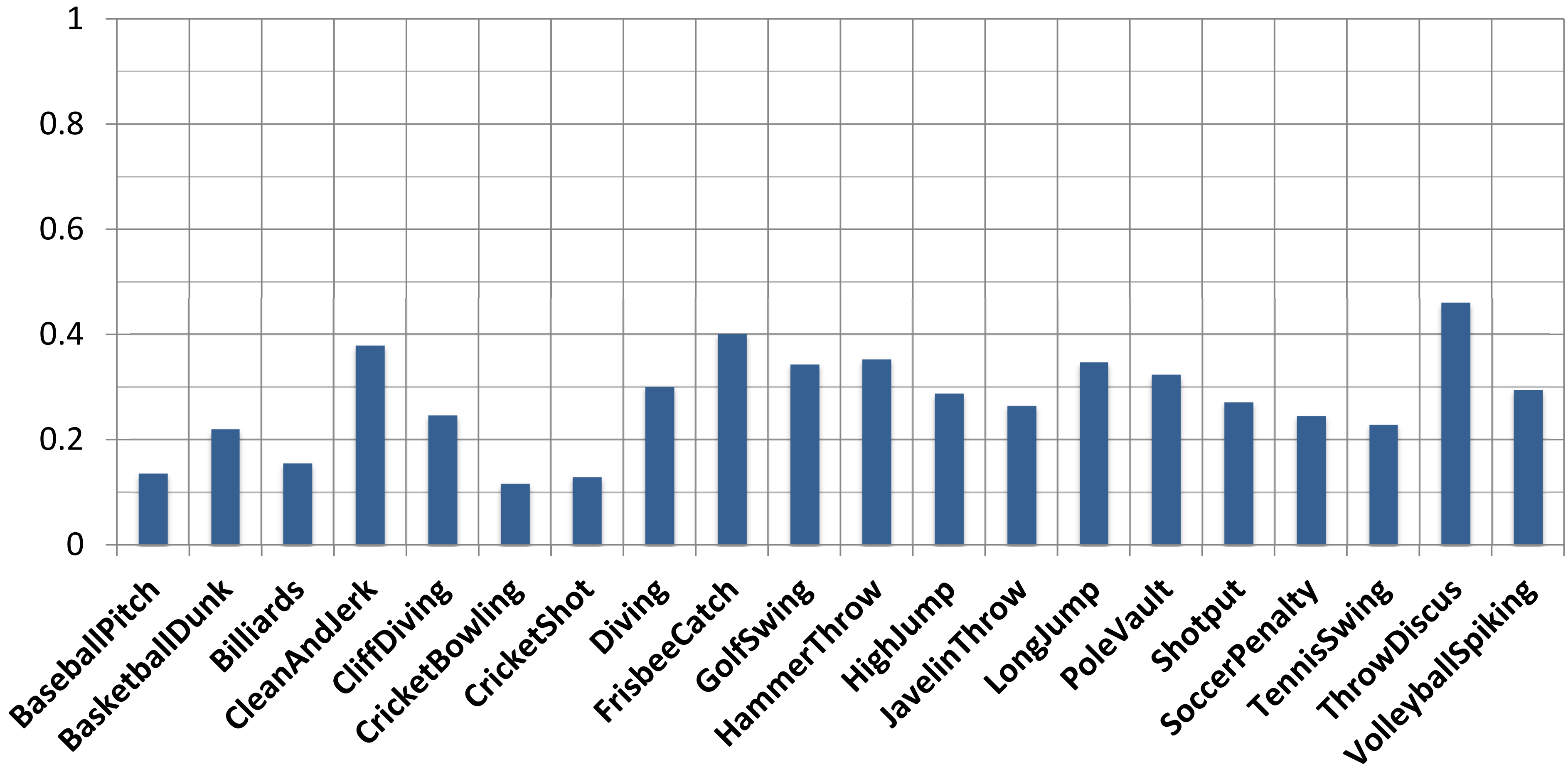
Number of Clips Per Class (Train & Validation)



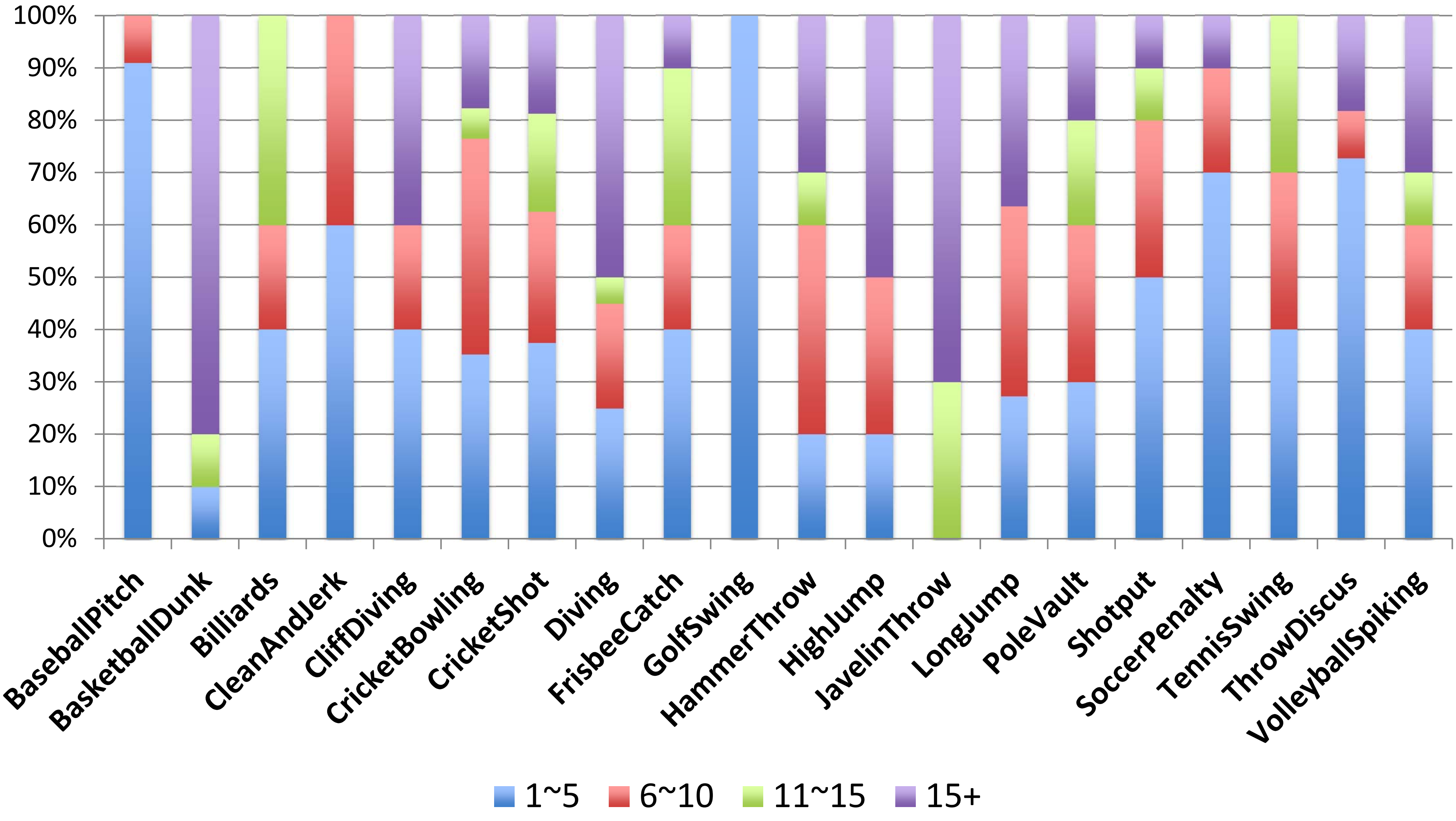
Clip Duration Per Class (Train & Validation)



Action Duration (%) Per Class (Validation)

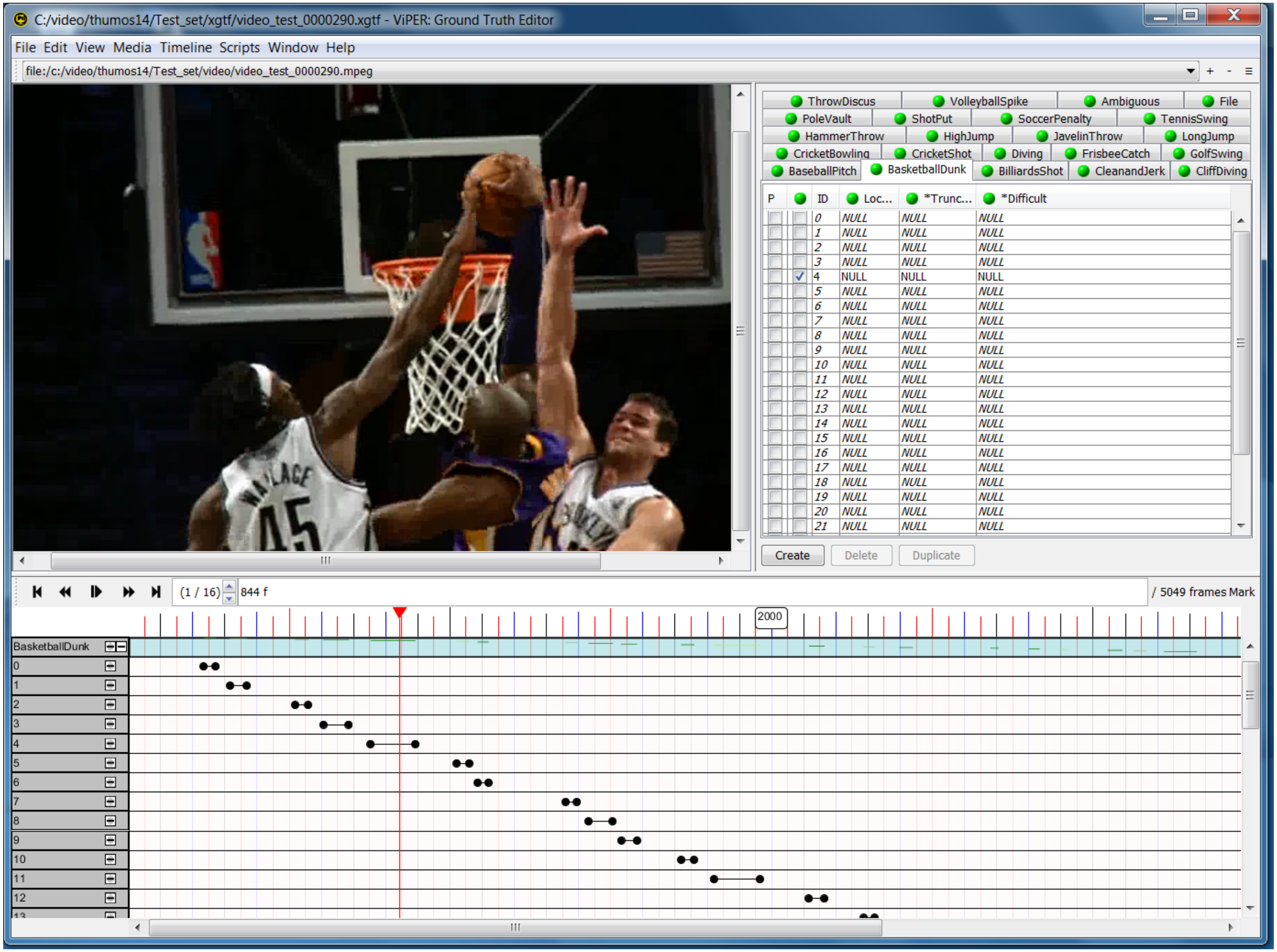


Number of Action Instances (Validation)



Annotation Tool

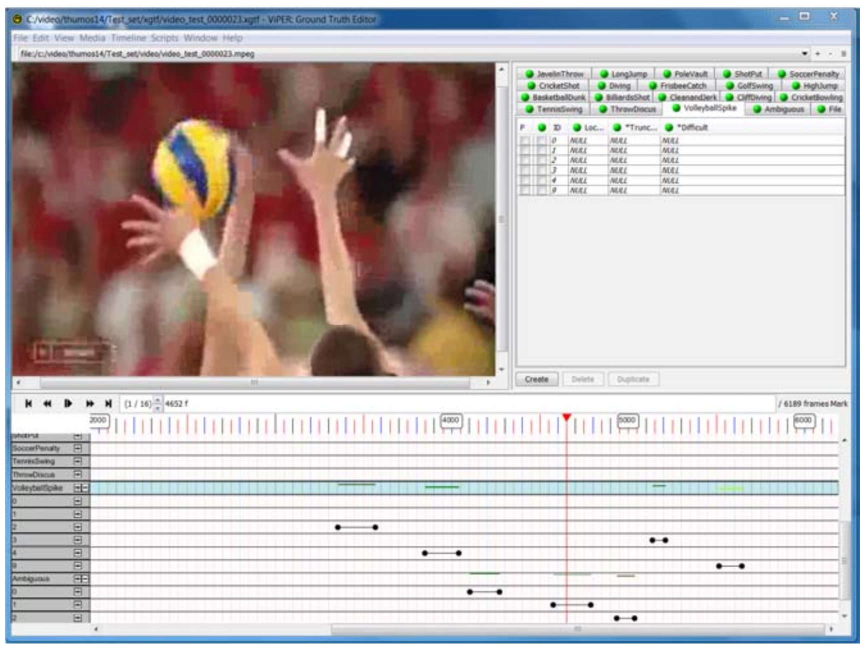
http://viper-toolkit.sourceforge.net/



Annotation Tool

http://viper-toolkit.sourceforge.net/

- Mostly annotated by one person, who did two passes.
    - Additional checks made by at least two other people.
- An action interval was annotated as Ambiguous in several cases including heavy temporal or spatial crops (below), graphics animations, etc.

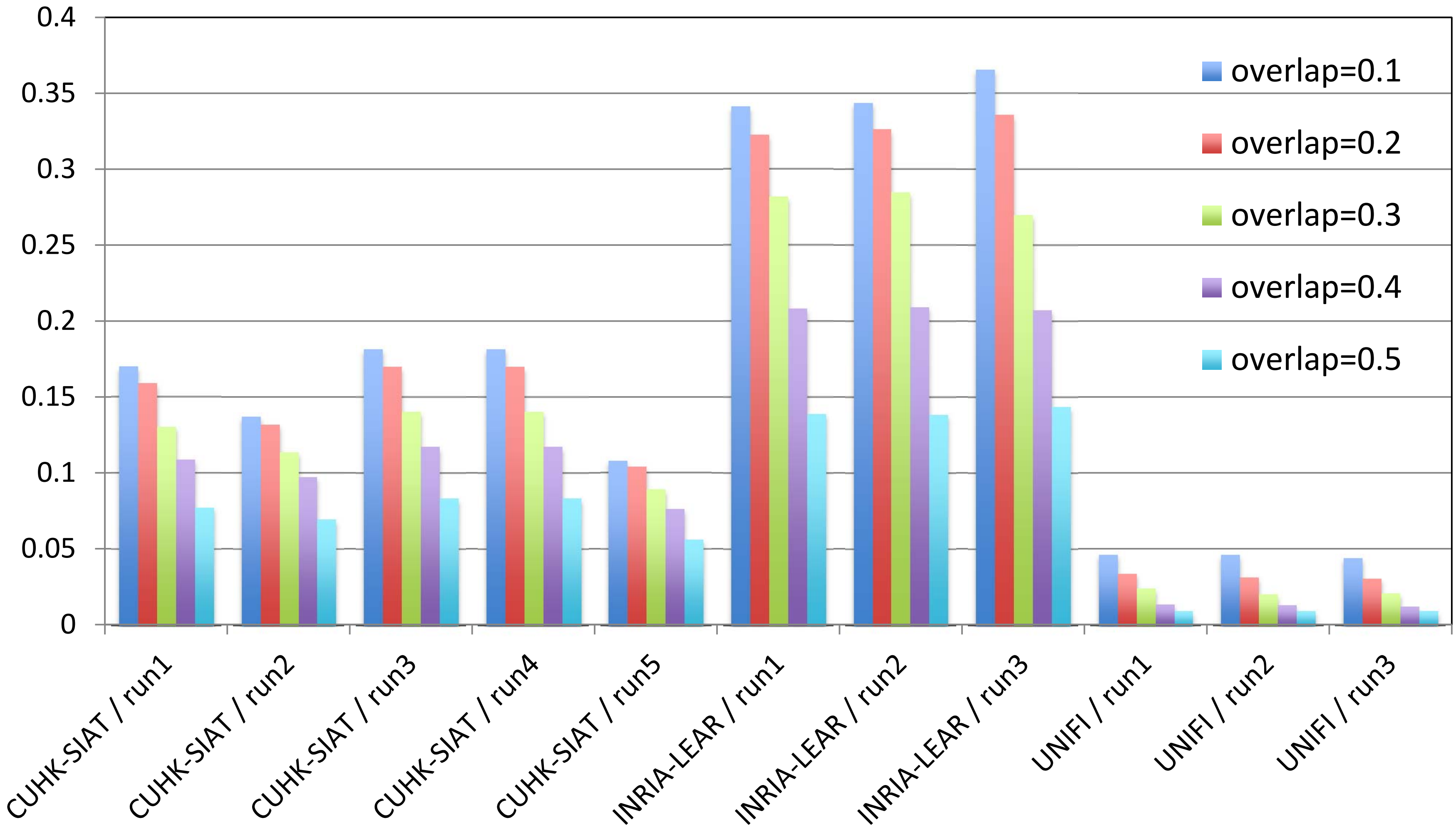


Evaluation Measure

- The traditional “intersection-over-union” criterion
  - A detection is correct if the predicted class label is correct and the overlapping criterion is larger a threshold (0.5)
- AP / mAP

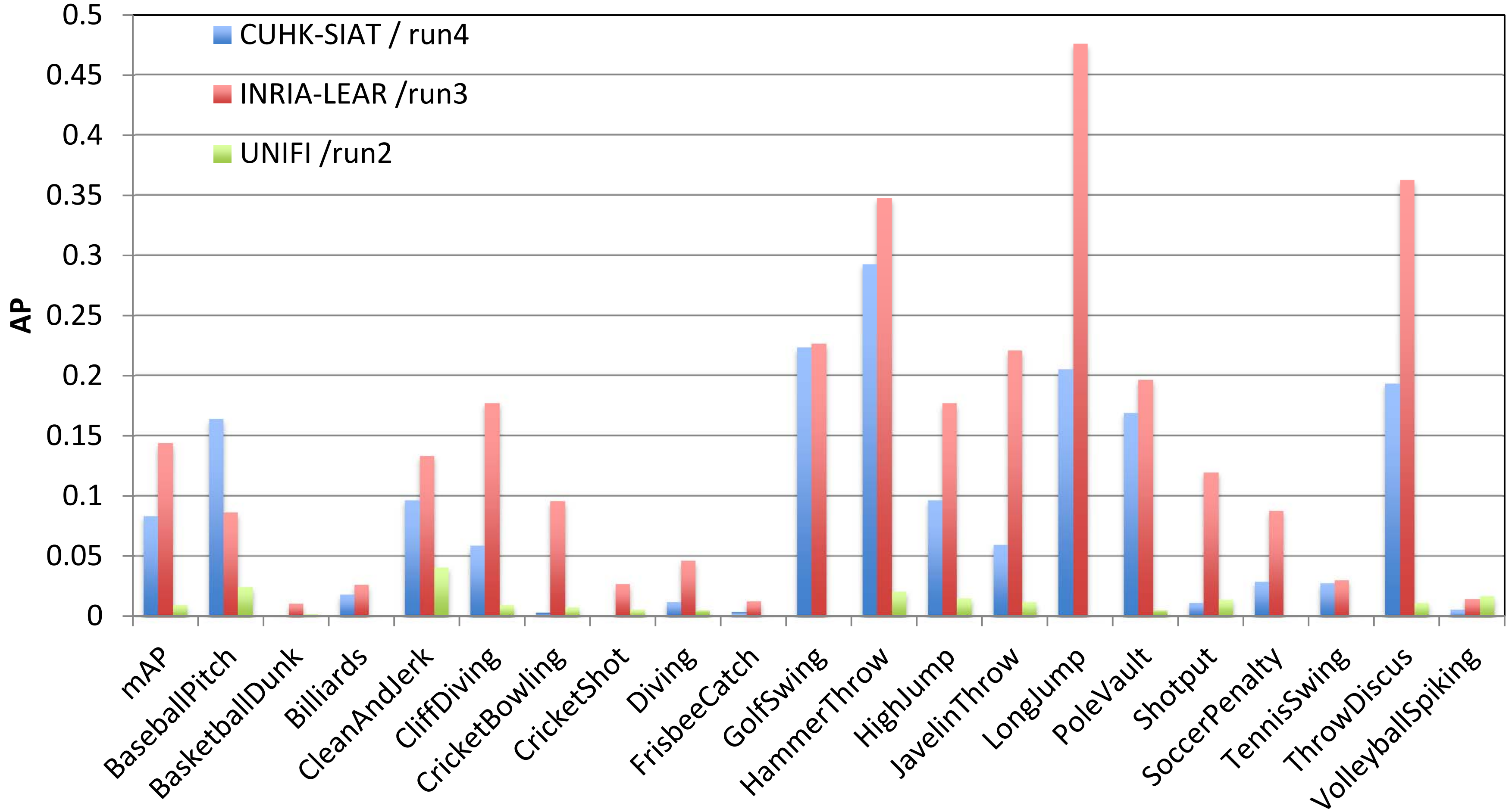
Results

Overall (mAP)



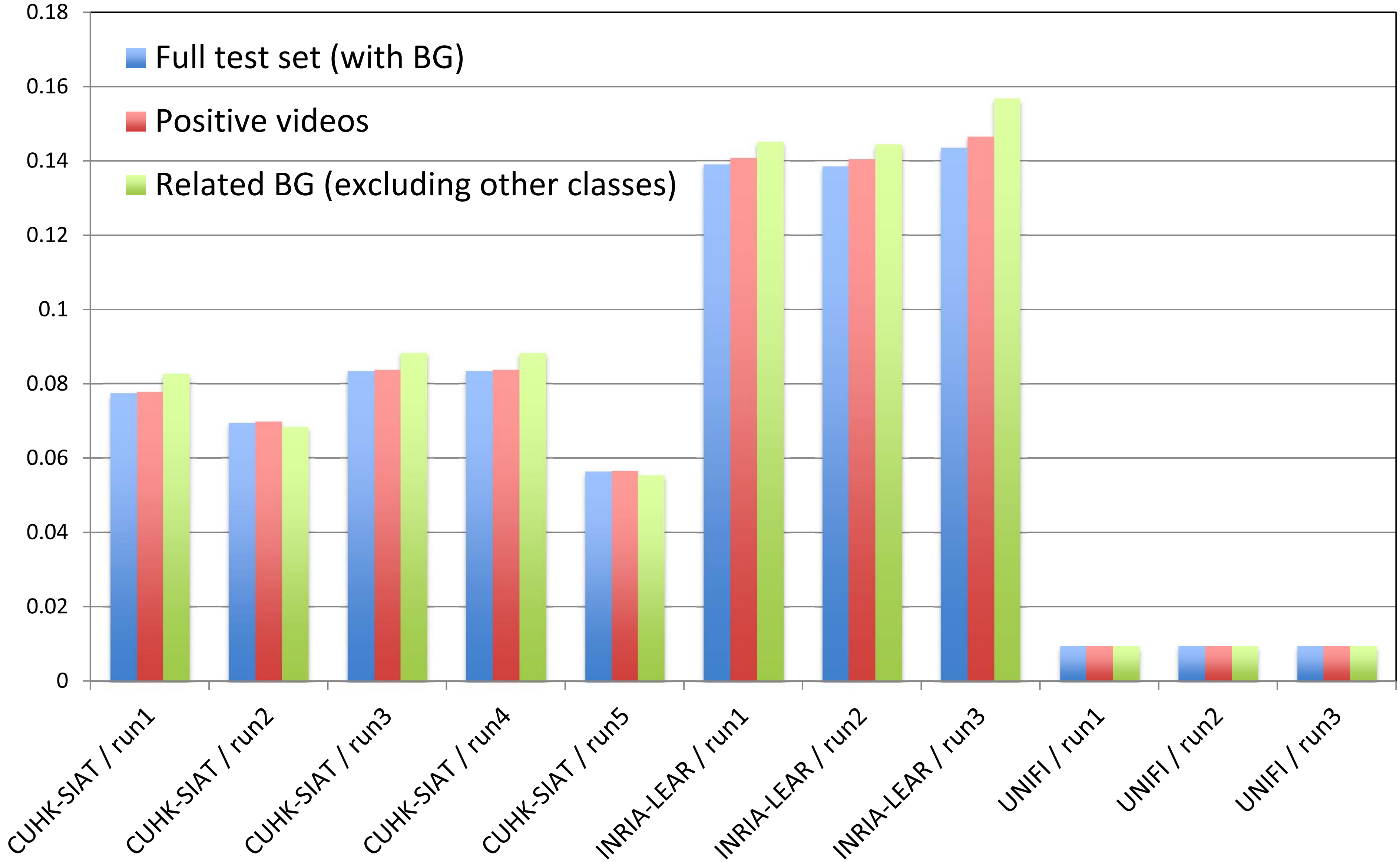
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
overlap=0.1
overlap=0.2
overlap=0.3
overlap=0.4
overlap=0.5
CUHK-SIAT / run1
CUHK-SIAT / run2
CUHK-SIAT / run3
CUHK-SIAT / run4
CUHK-SIAT / run5
INRIA-LEAR / run1
INRIA-LEAR / run2
INRIA-LEAR / run3
UNIFI / run1
UNIFI / run2
UNIFI / run3

Per-class (AP)

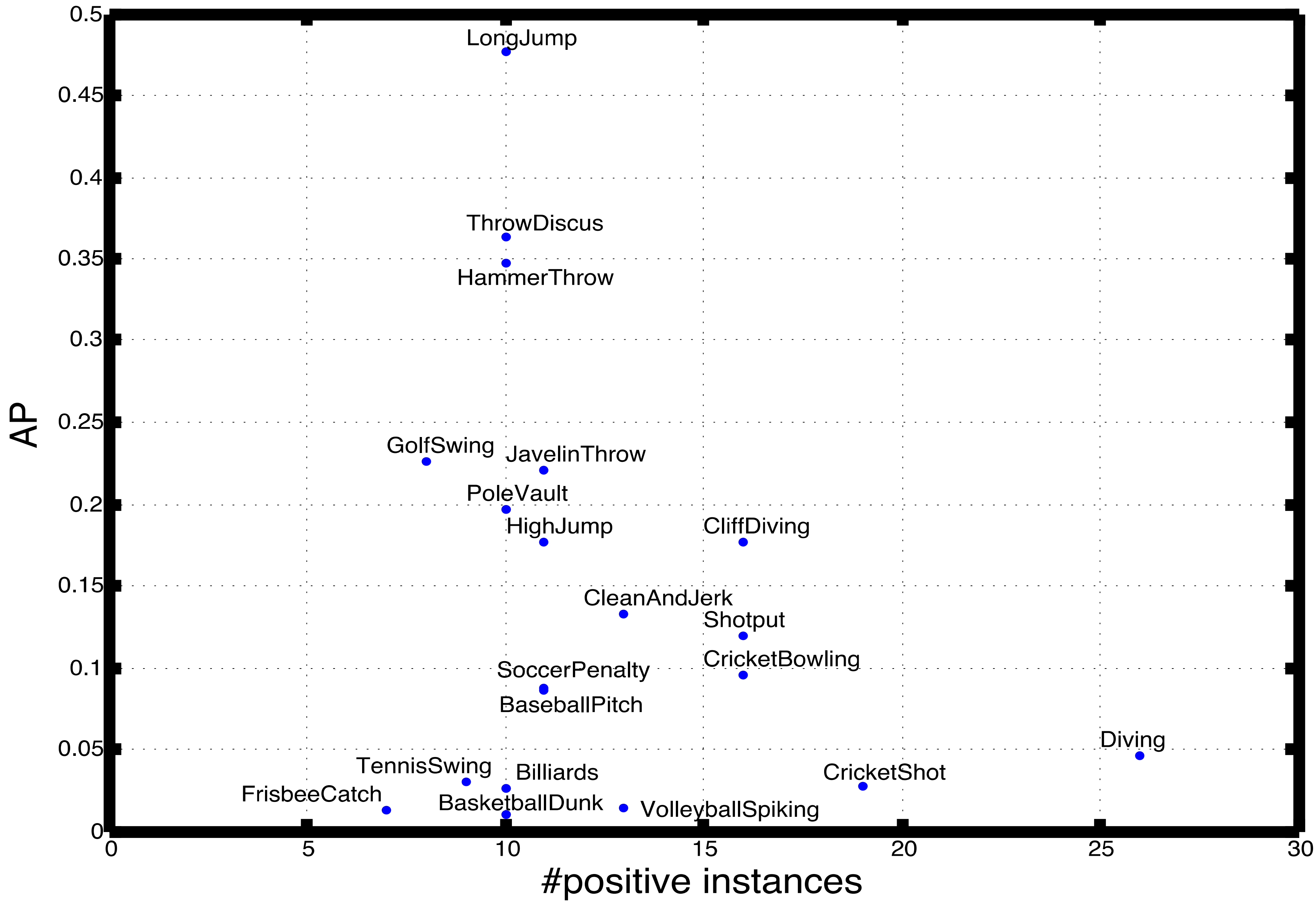


INRIA-LEAR achieved top results for 18 classes; CUHK-SIAT and UNIFI one class each

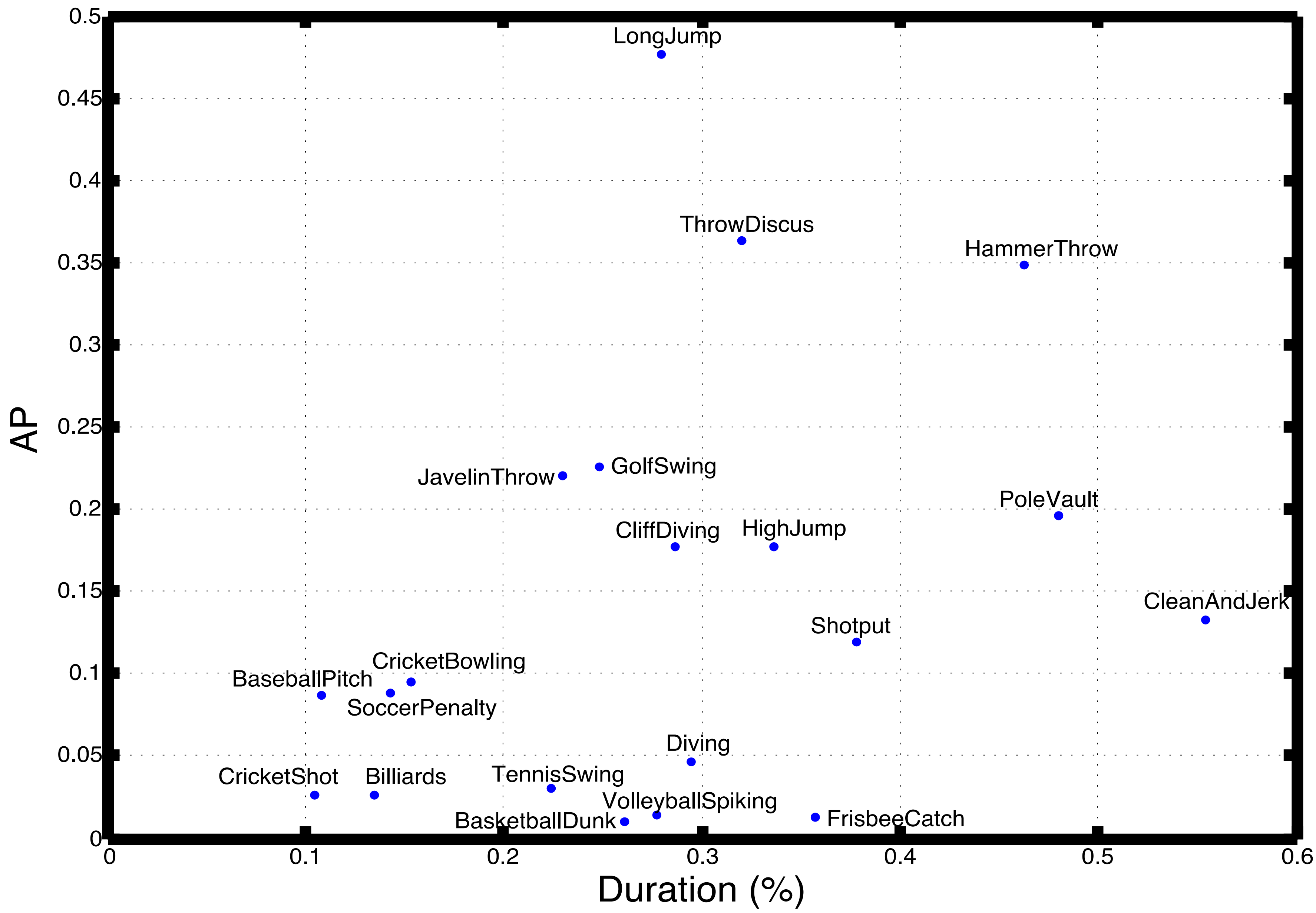
Overall (mAP) after excluding background videos or videos of the other classes



AP vs. # positive instances (INRIA R3)



AP vs. # action duration (%) (INRIA R3)



Approaches

CUHK-SIAT

- Feature:
  - FV encoding of IDT features
  - CNN features
  - Early fusion (?)
- Classifier:
  - 1-vs-rest (Linear?) SVM over temporal windows
  - fixed sliding window size (150 frames)
  - step size (100 frames)
- Post-processing:
  - Thresholds on both video and temporal clip window levels.

INRIA-LEAR

- Feature:
  - FV encoding of IDT features
- Classifier:
  - 1-vs-rest (linear?) SVM over temporal windows, with hard negative mining
  - sliding window size: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, and 150 frames.
  - Step size: 10 frames
- Rescore by:
  - Detection window size
  - Class-specific duration prior estimated from training data
- Context:
  - Combining window's detection score with video's classification score for the same action class
  - Classification used additional features: SIFT (FV), Color Moments, CNN, MFCC (FV), ASR

UNIFI

- Features:
  - FV encoding of IDT features (weighted based on saliency predicted by BING Objectness)
  - CNN features using Decaf
  - Late Fusion
- Classifier:
  - 1-vs-rest linear SVM over temporal windows
  - sliding window size: 200 frames.
  - Step size: 100 frames
  - Combined with classification score (similar to INRIA-LEAR)

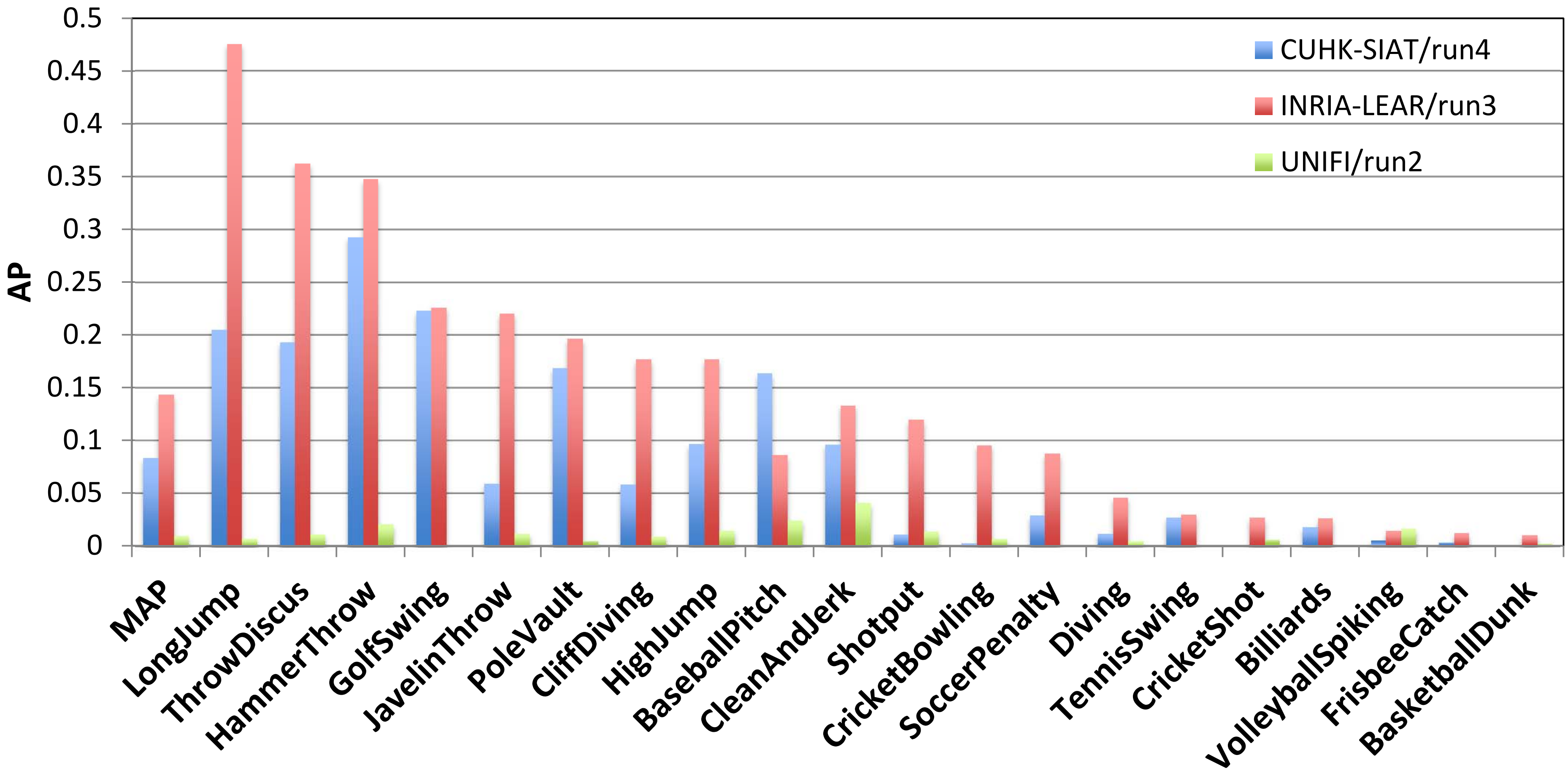
Summary

- Common techniques:
  - Features:
    - FV encoding of IDT feature
    - CNN feature
  - Classifier:
    - 1-vs-rest SVM over temporal windows
- Differences:
    - Early or late fusion
    - Window size, step size, hard negatives
    - Post-processing (rescoring, thresholding)
    - Combination with classification scores (context)

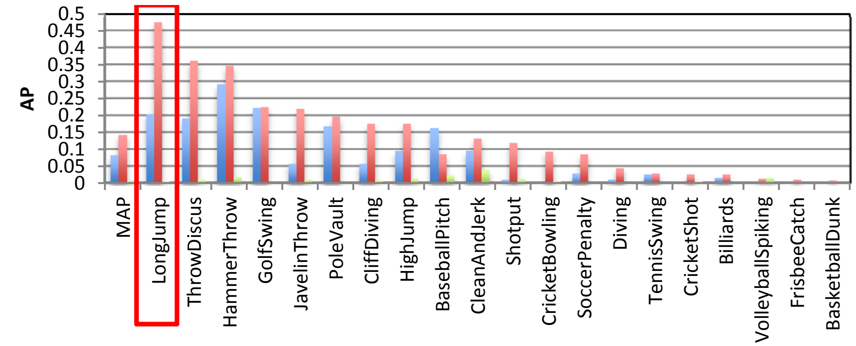
Examples

Per-class (AP)

sorted by performance



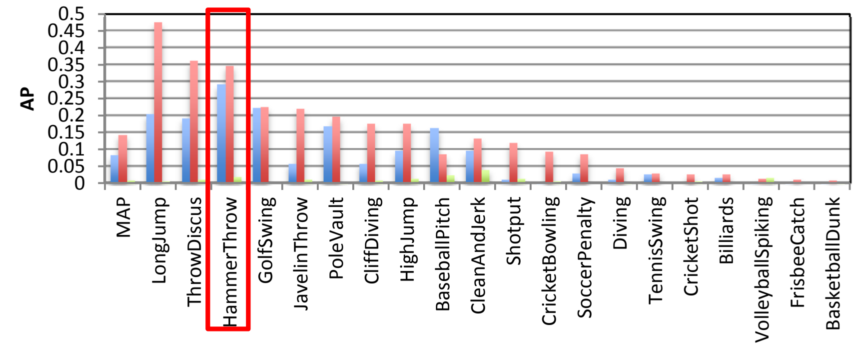
LongJump



Easy

Hard

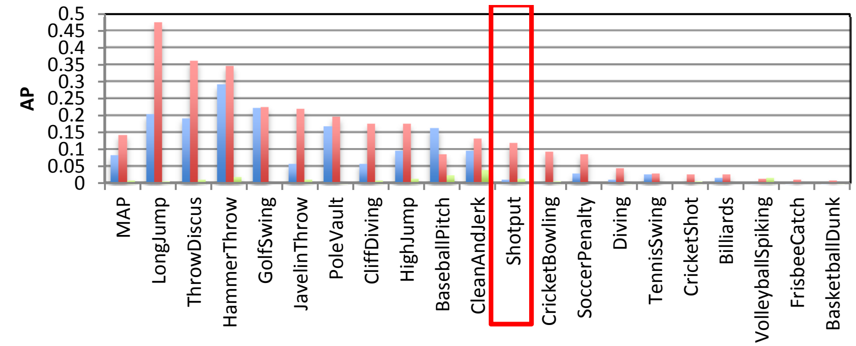
HammerThrow



Easy

Hard

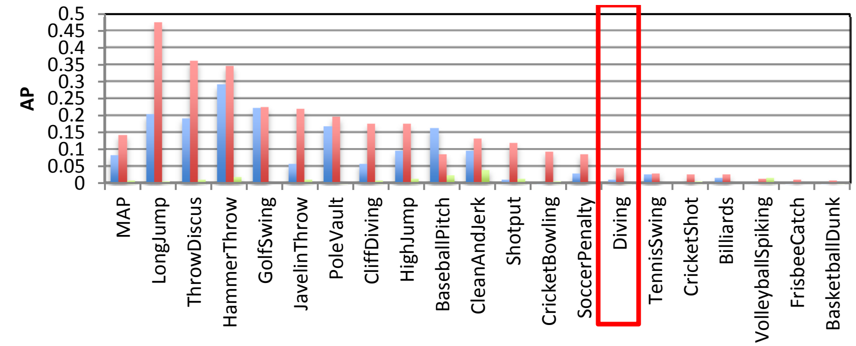
Shotput



Easy


Hard

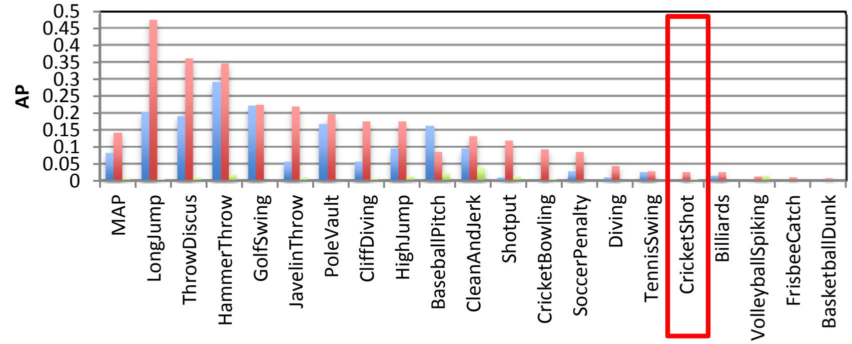

Diving



Easy

Hard

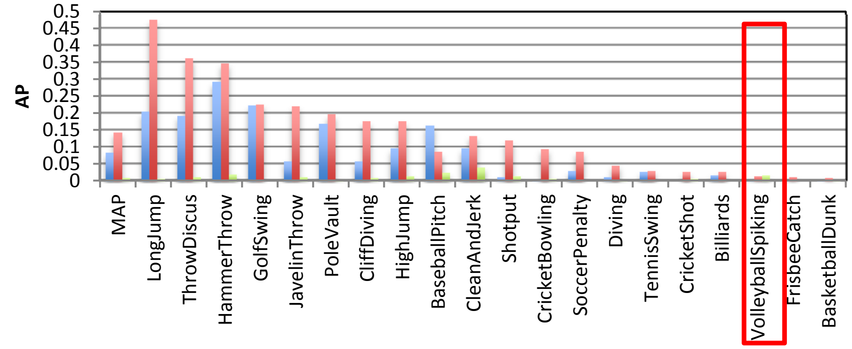
CricketShot



Easy

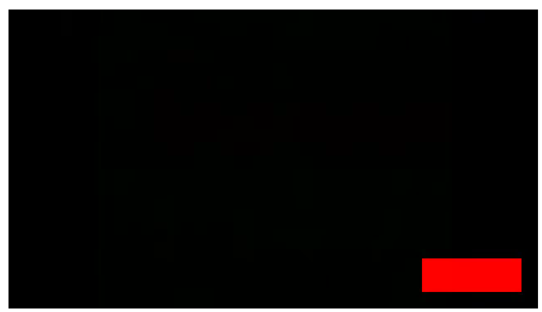
Hard

VolleyballSpiking

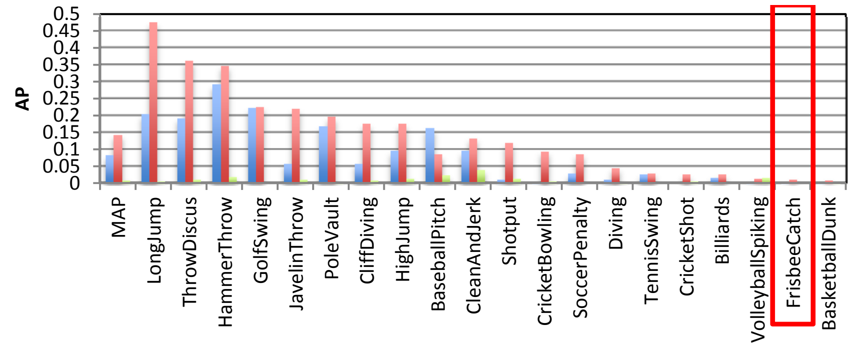


Easy

Hard



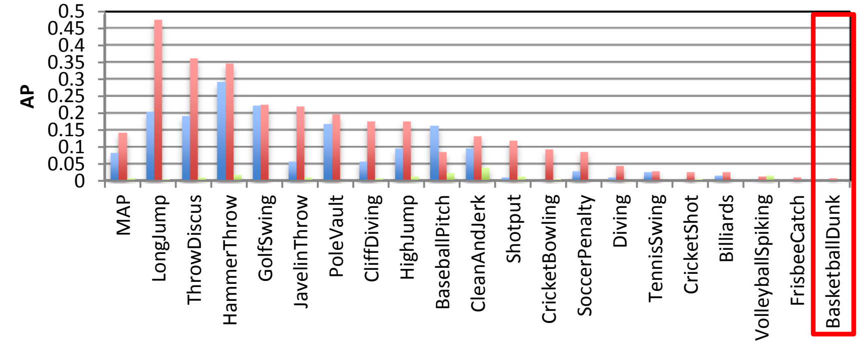
FrisbeeCatch



Easy

Hard

BasketballDunk



Easy

Hard

Thank you!

THUMOS’14

Zurich, Switzerland, Sept. 7th 2014