The TRECVID 2010 Surveillance Event Detection (SED) Evaluation Overview and Results

Martial Michel
Jonathan Fiscus
Paul Over

Home Office Scientific Development Branch
CPNI Centre for the Protection of National Infrastructure
Homeland Security Science and Technology Directorate
NIST National Institute of Standards and Technology

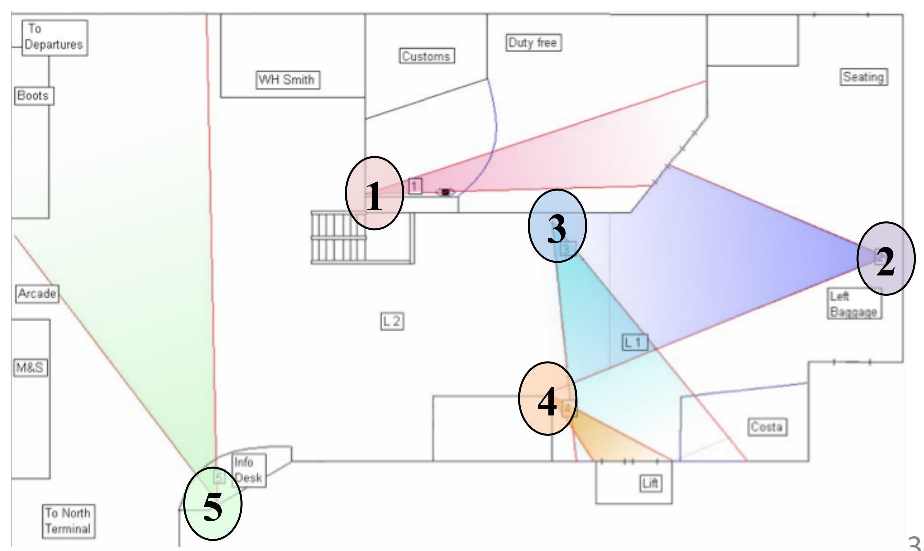
Motivation

- SED addresses the need for automatic detection of events in large amounts of surveillance video
- Challenges
  - requires application of several Computer Vision techniques
  - involves subtleties that are readily understood by humans, difficult to encode for machine learning approaches
  - can be complicated due to clutter in the environment, lighting, camera placement, traffic, etc.

Evaluation Source Data

Imagery Library for Intelligent Detection Systems

- Reused same data as the past year of SED evaluation
- UK Home Office collected CCTV video from 5 camera views at a busy airport
- Development Set
  - 100 hours of video
  - 10 events annotated on 100% of the data
- Evaluation Set
  - “iLIDS Multiple Camera Tracking Scenario Training set”
  - 45 hours of video
  - 10 events annotated on 1/3 of the data



Event Detection Task

- Given a textual description of an ***observable event of interest***, automatically detect all occurrences of the event in a non-segmented corpus of video

- Identify each event observation by:
  - The ***temporal extent*** *(beginning and end frames)*
  - A ***decision score***: a numeric score indicating how likely the event observation exists with more positive values indicating more likely observations
  - An ***actual decision***: A Boolean value indicating whether or not the event observation should be counted for the primary metric computation

Events and Instances per Hour (IpH)

Single Person events		
PersonRuns	7.02 IpH	Someone runs ← ***Lowest frequency***
Pointing	69.74 IpH	Someone points ← ***Highest frequency***
Single Person + Object events		
CellToEar	12.73 IpH	Someone puts a cell phone to his/her head or ear
ObjectPut	40.74 IpH	Someone drops or puts down an object
Multiple People events		
Embrace	11.48 IpH	Someone puts one or both arms at least part way around another person
PeopleMeet	29.46 IpH	One or more people walk up to one or more other people, stop, and some communication occurs
PeopleSplitUp	12.27 IpH	From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame

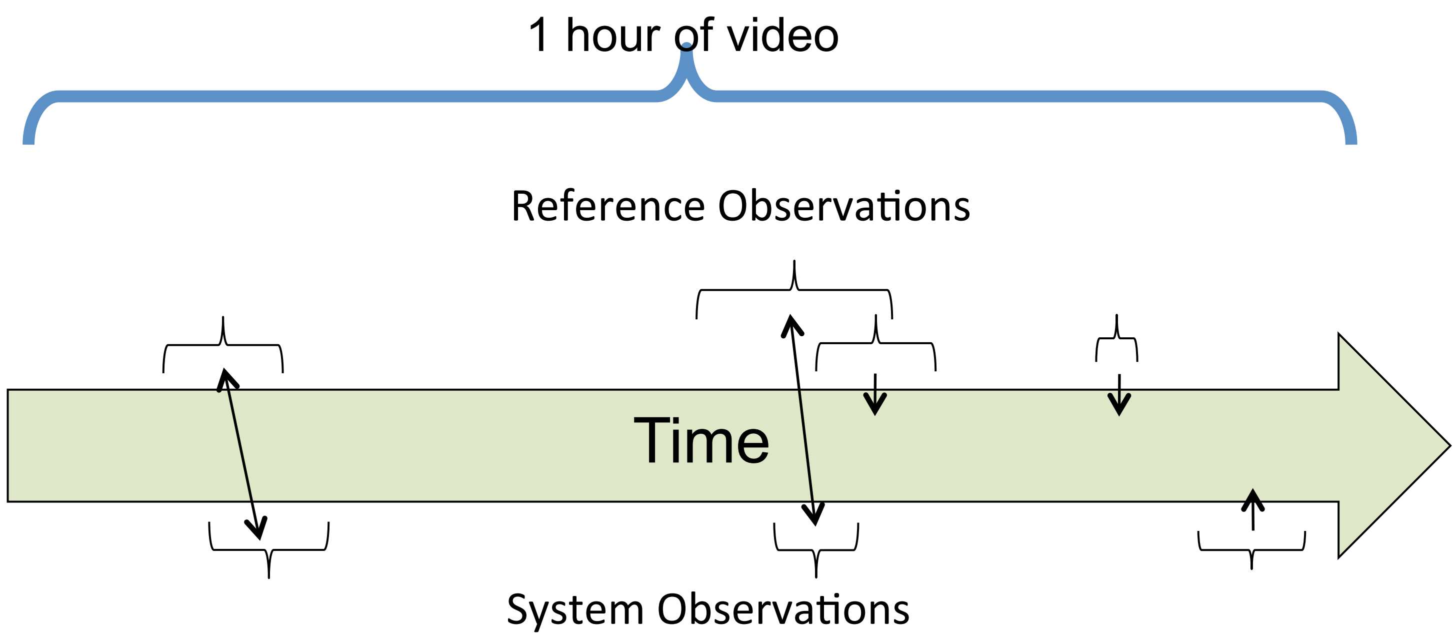
ElevatorNoEntry, OpposingFlow, and TakePicture events were not evaluated in 2010

Evaluation Protocol & Scoring Process

- Evaluation Plan
  http://www.nist.gov/itl/iad/mig/trecvid.cfm
- Framework for Detection Evaluation (F4DE) Toolkit
  http://www.nist.gov/itl/iad/mig/tools.cfm

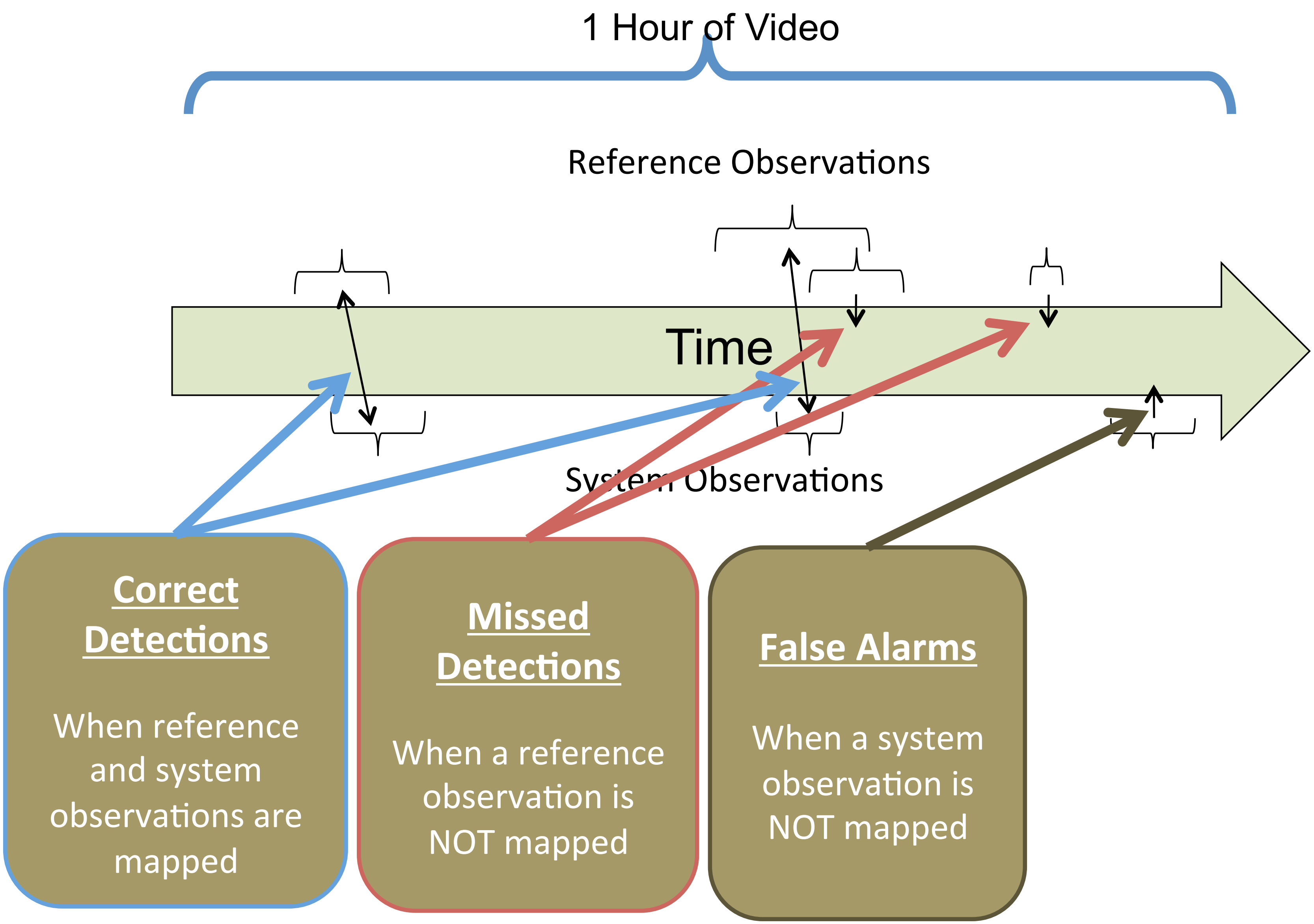
- Four step evaluation process (for each Event)
  1. Segment mapping
  2. Segment scoring
  3. Error metric calculation
  4. Error visualization

Step 1: Segment Mapping



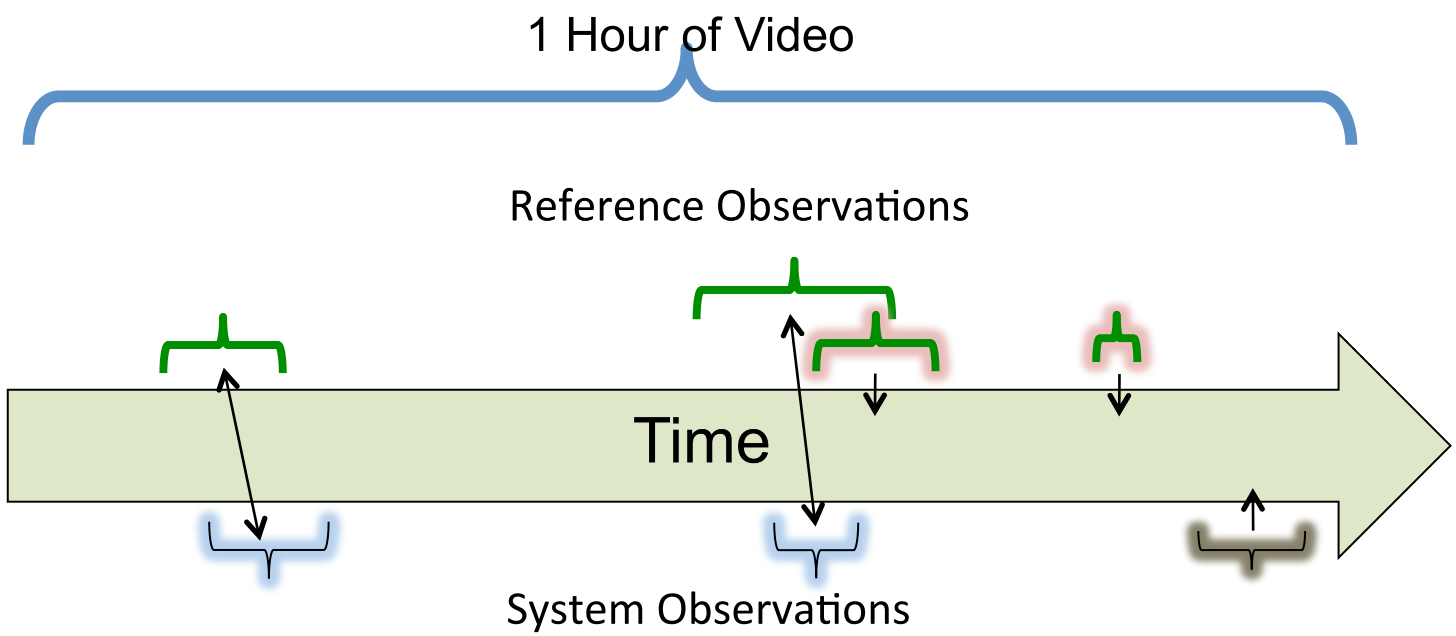
Utilizes the Hungarian Solution to Bipartite Graph Matching

Step 2: Segment Scoring



Step 3: Error Metric Computation

Compute Normalized Detection Cost Rate (1/2)



$$P_{Miss} = \frac{\# MissedObs}{\# TrueObs}$$

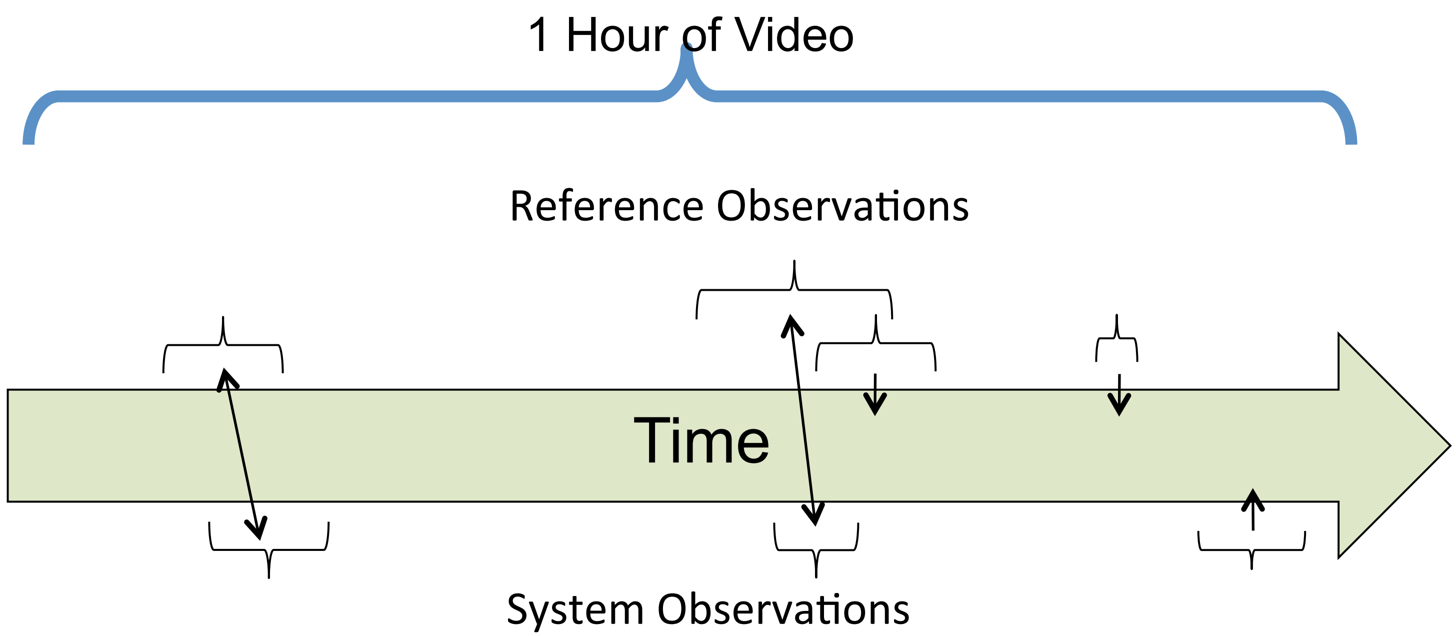
$$Rate_{FA} = \frac{\# FalseAlarms}{SignalDuration}$$

$$P_{Miss} = \frac{2}{4} = .50$$

$$Rate_{FA} = \frac{1}{1Hr} = 1FA / Hr$$

Step 3: Error Metric Computation

Compute Normalized Detection Cost Rate (2/2)



Primary Metric

$$NDCR = P_{Miss} + \underbrace{\frac{Cost_{FA}}{Cost_{Miss} * R_{TARGET}}}_{\text{Beta}} * R_{FA}$$

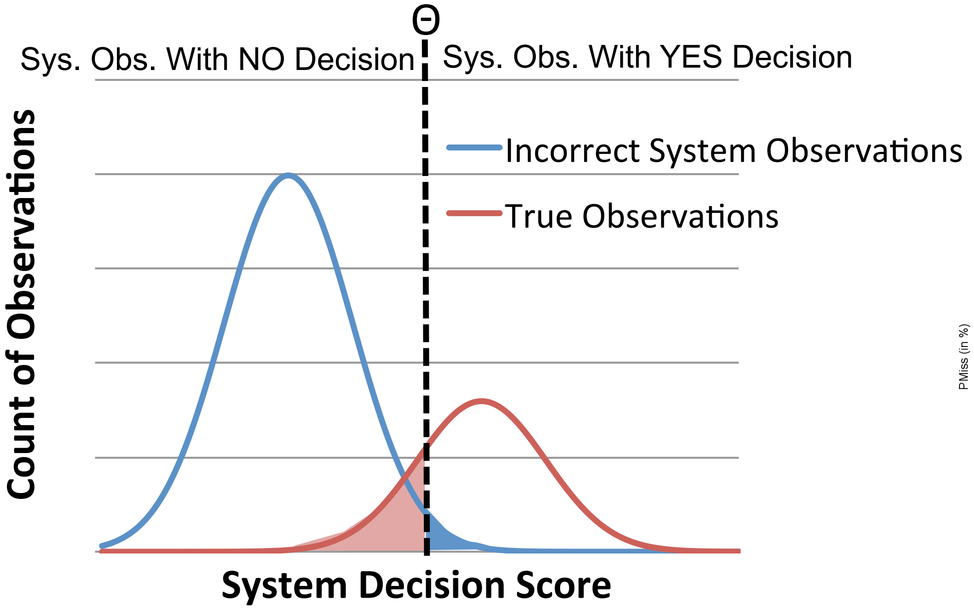
$$NDCR = 0.5 + \frac{1}{10 * 20} * 1 = .505$$

$$Cost_{Miss} = 10$$
$$Cost_{FA} = 1$$
$$R_{TARGET} = 20$$

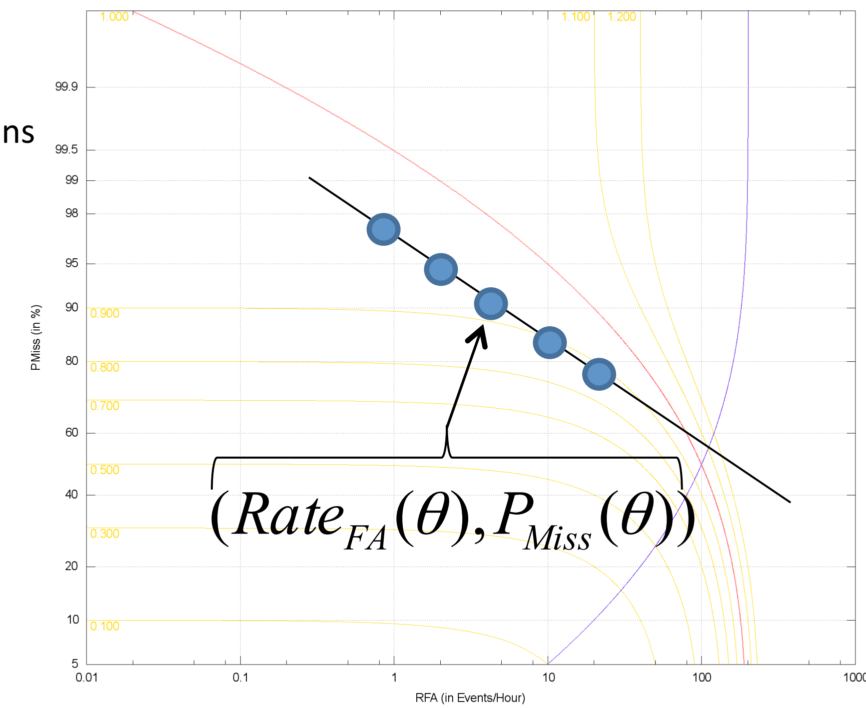
Range of NDCR() is [0:∞)
- NDCR = 0.0 is a perfect system
- NDCR = 1.0 is equivalent to a system that outputs nothing

Step 4: Error Visualization

Decision Error Tradeoff Curves *(Prob$_{Miss}$ vs. Rate$_{FA}$)*



Compute Rate$_{FA}$ and P$_{Miss}$ for all Θ



$$MinimumNDCR(\theta) = \arg\min_{\theta} \left[P_{Miss}(\theta) + \frac{Cost_{FA}}{Cost_{Miss} * R_{TARGET}} * R_{FA}(\theta) \right]$$

$$ActualNDCR(Act.Dec.) = P_{Miss}(Act.Dec.) + \frac{Cost_{FA}}{Cost_{Miss} * R_{TARGET}} * R_{FA}(Act.Dec.)$$

For more information about DETCurves: http://www.nist.gov/speech/publications/storage_paper/det.pdf

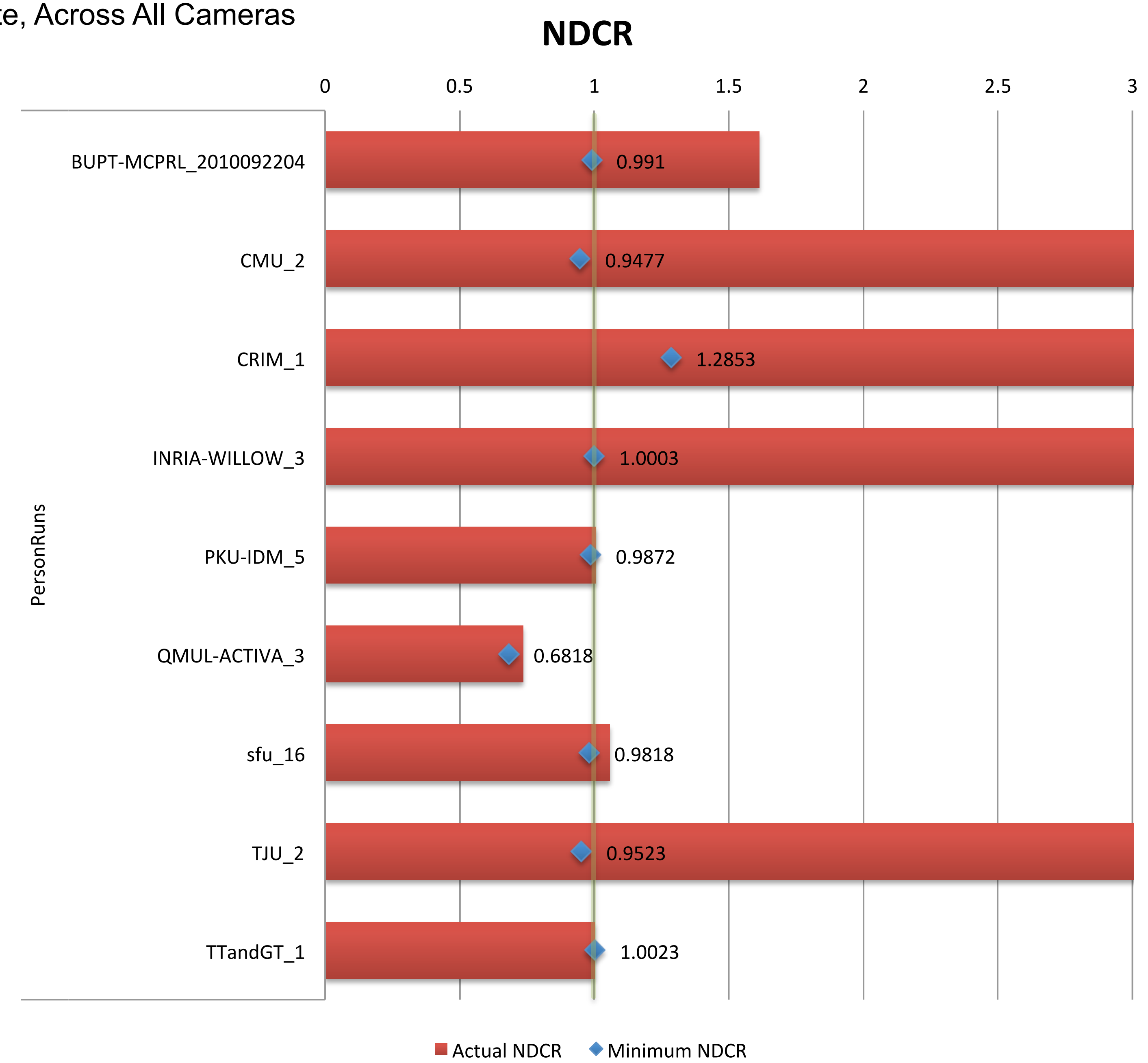
2010 SED Participants

11 Sites

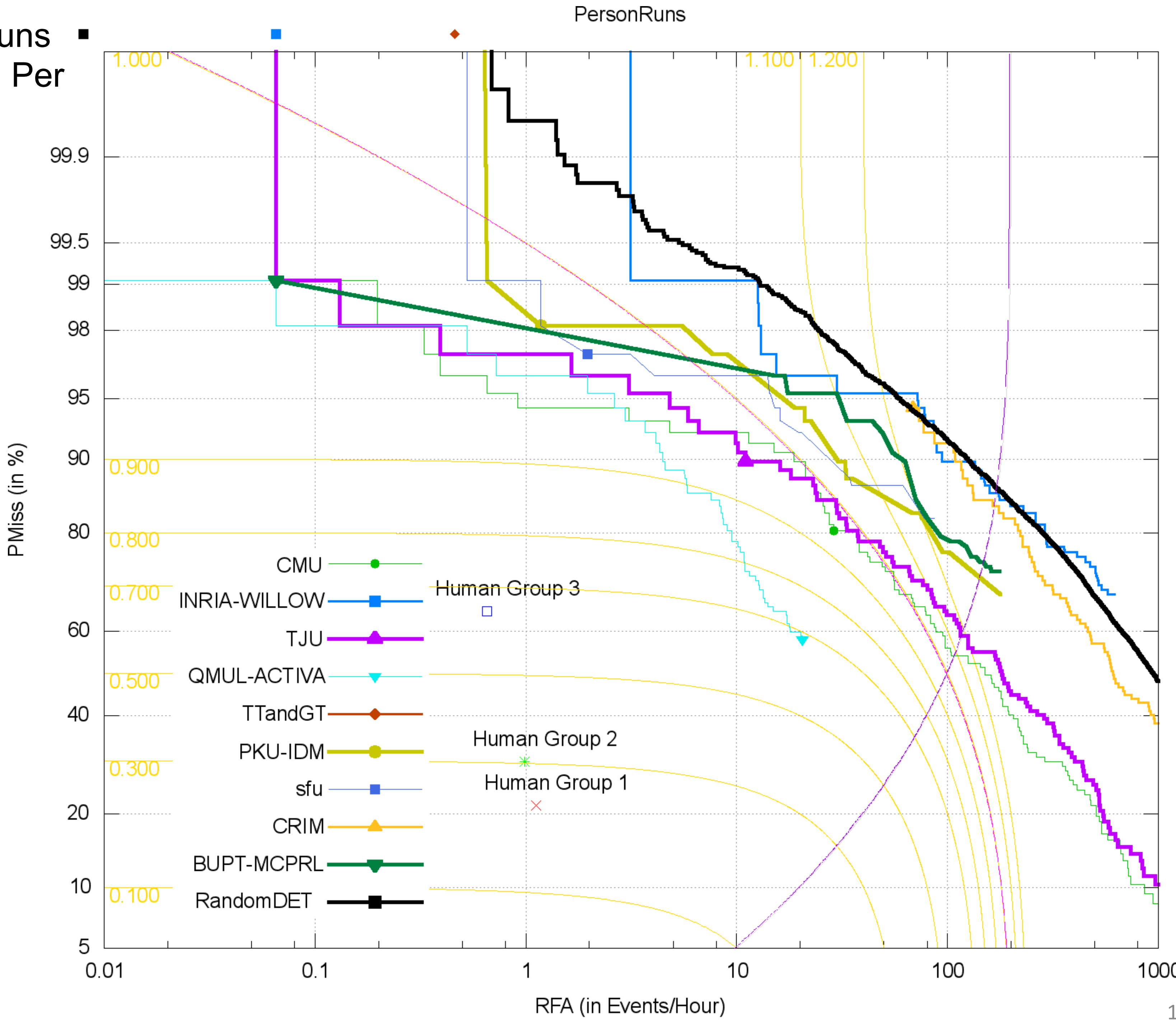
		Single Person		Single Person + Object		Multiple People		
		PersonRuns	Pointing	CellToEar	ObjectPut	Embrace	PeopleMeet	PeopleSplitUp
3 years in a row	Carnegie Mellon University [CMU]	11	11	11	11	11	11	11
	NHK Science and Technical Research Laboratories [NHKSTRL]		1	1	1			
2 years in a row	Beijing University of Posts and Telecommunications, MCPRL [BUPT-MCPRL]	2	2		2	2		
	Peking University, IDM [PKU-IDM]	4				4	4	4
	Simon Fraser University [SFU]	1						
	Tokyo Institute of Technology and Georgia Institute of Technology [TTandGT]	1					1	1
new	Centre de Recherche Informatique de Montréal [CRIM]	1	1		1			
	Institut National de Recherche en Informatique et en Automatique, WILLOW [INRIA-WILLOW]	6	6	6	6	6	6	6
	Intelligent Perception Group of Beijing JiaoTong University [IPG-BJTU]		1	1	1	1		
	Queen Mary University of London [QMUL-ACTIVA]	1						
	Tianjin University [TJU]	8	8	8	8	8	8	8
	Total Participants per Event	9	7	5	7	6	5	5

2010 NDCRs for “PersonRuns” Event

Best Run Per Site, Across All Cameras

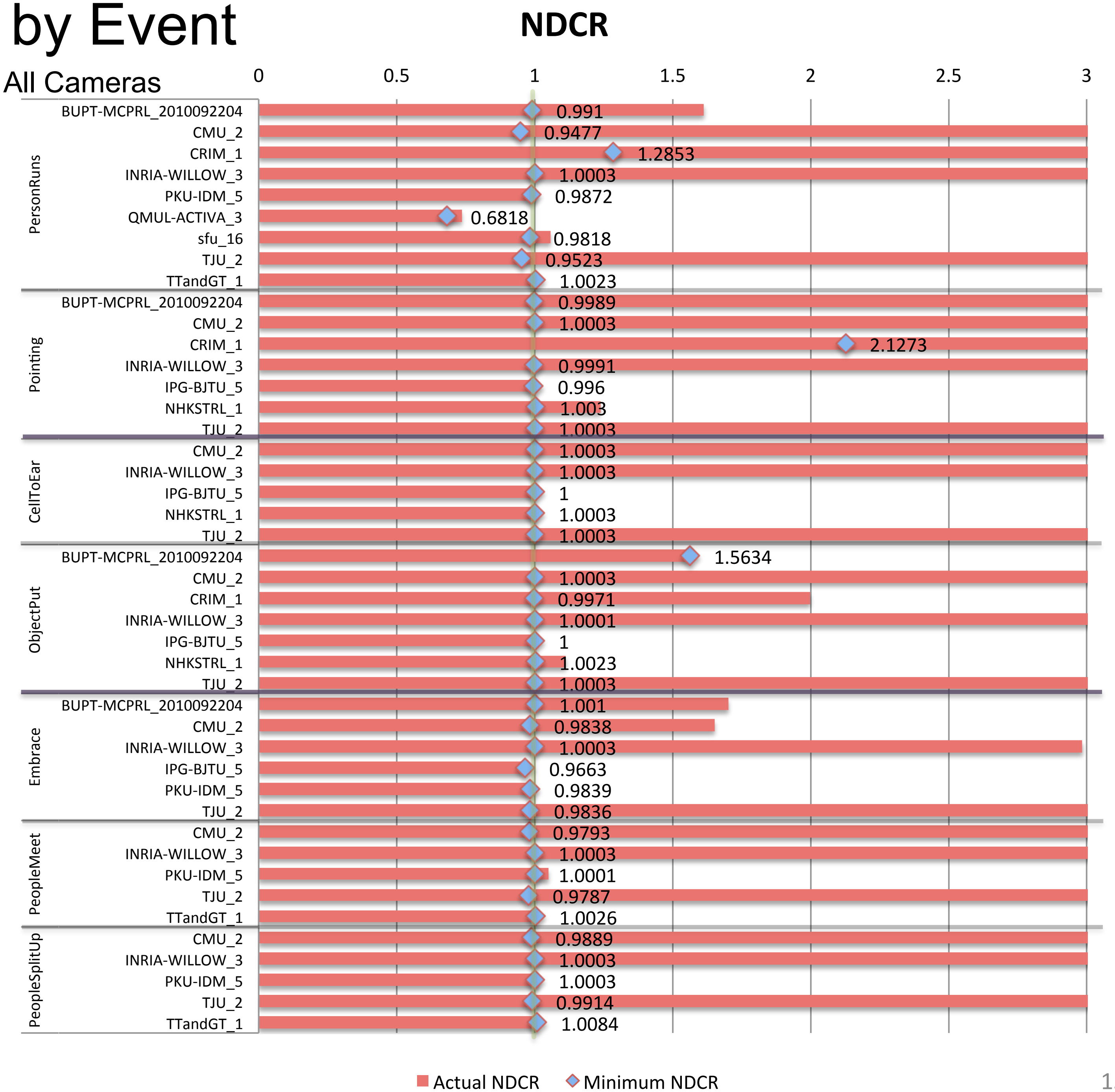


PersonRuns (Best Run Per Site)

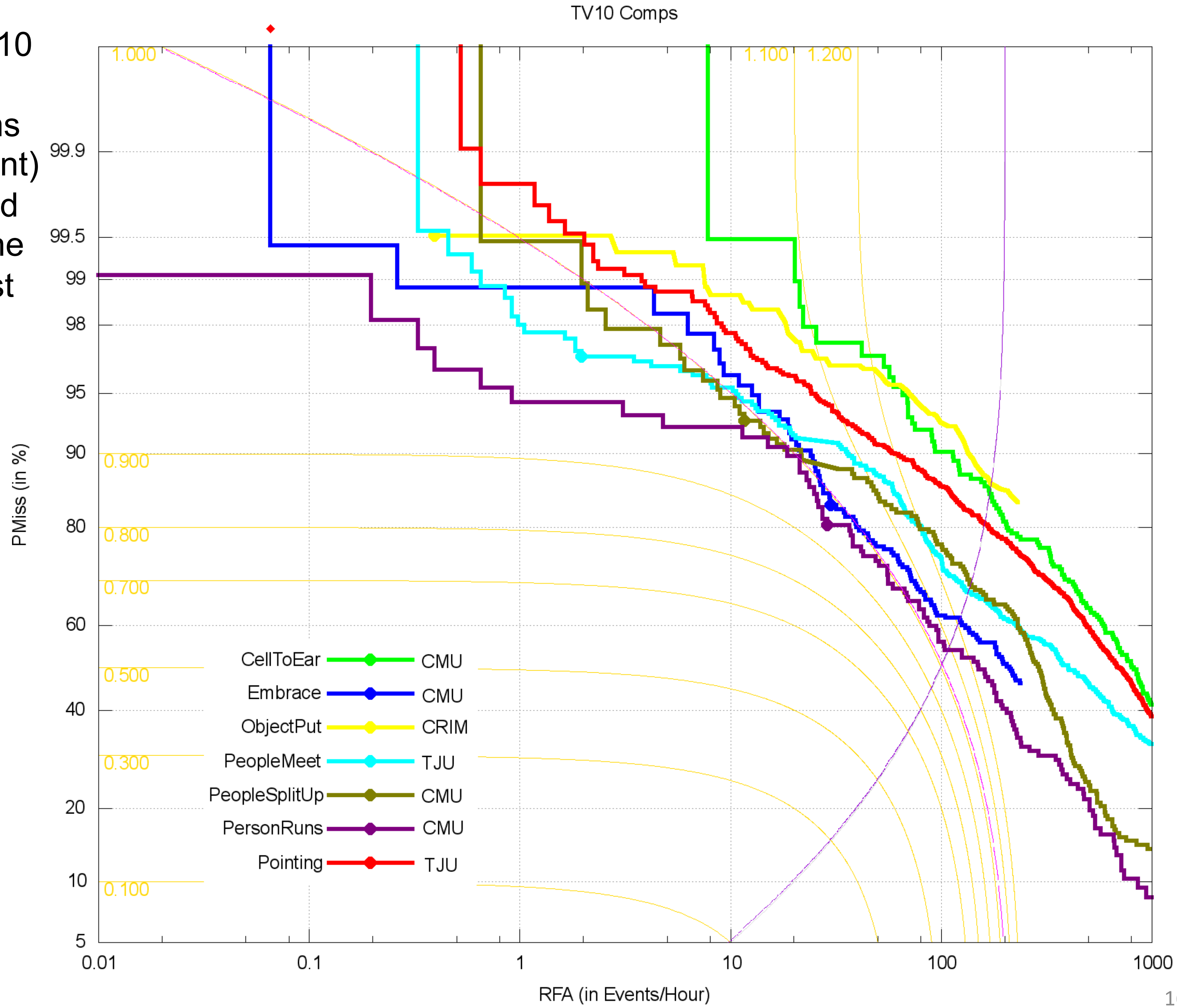


2010 NDCRs by Event

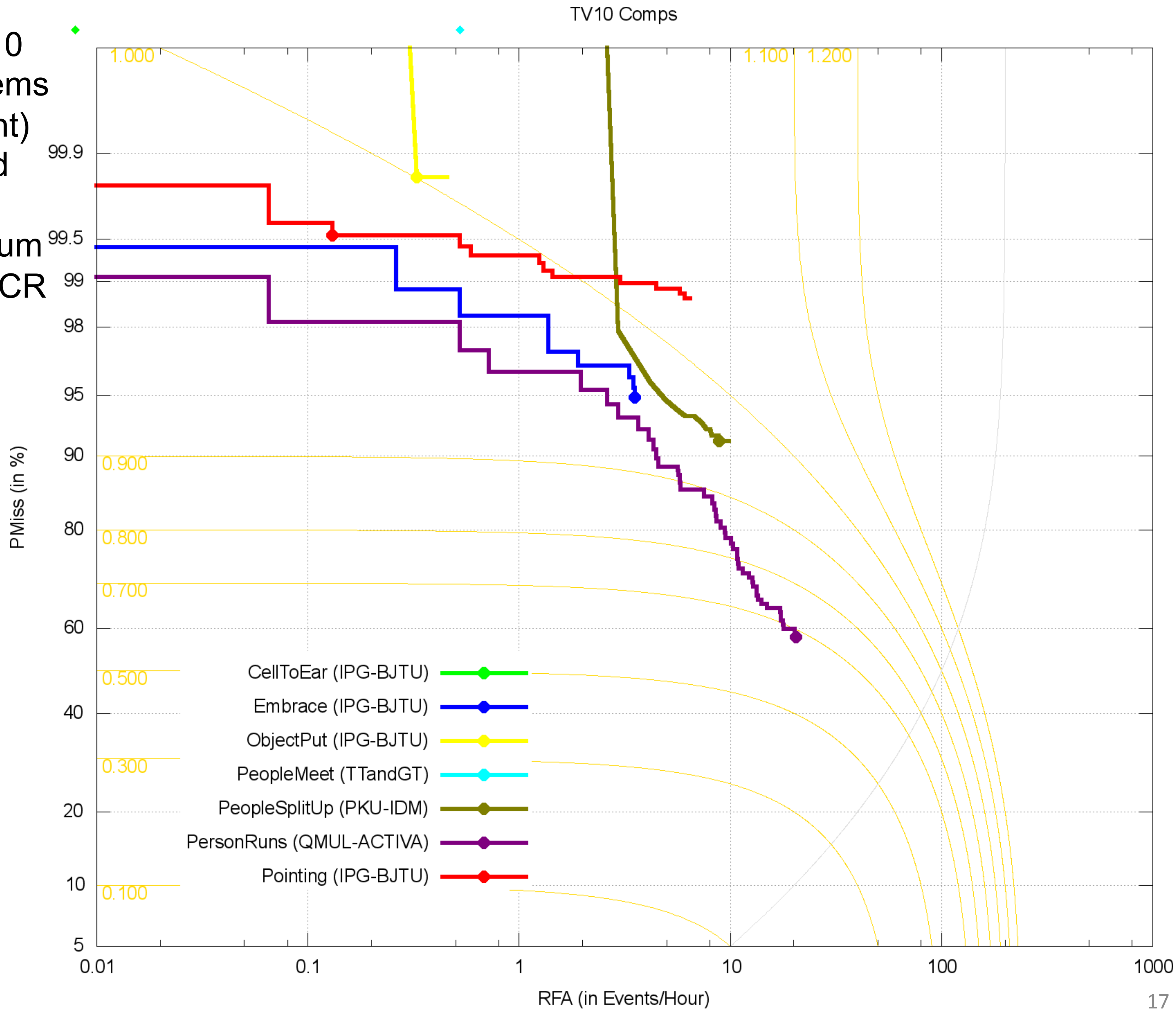
Best Run Per Site, Across All Cameras



SED 2010 Best Systems (per Event) selected using the iso-cost line



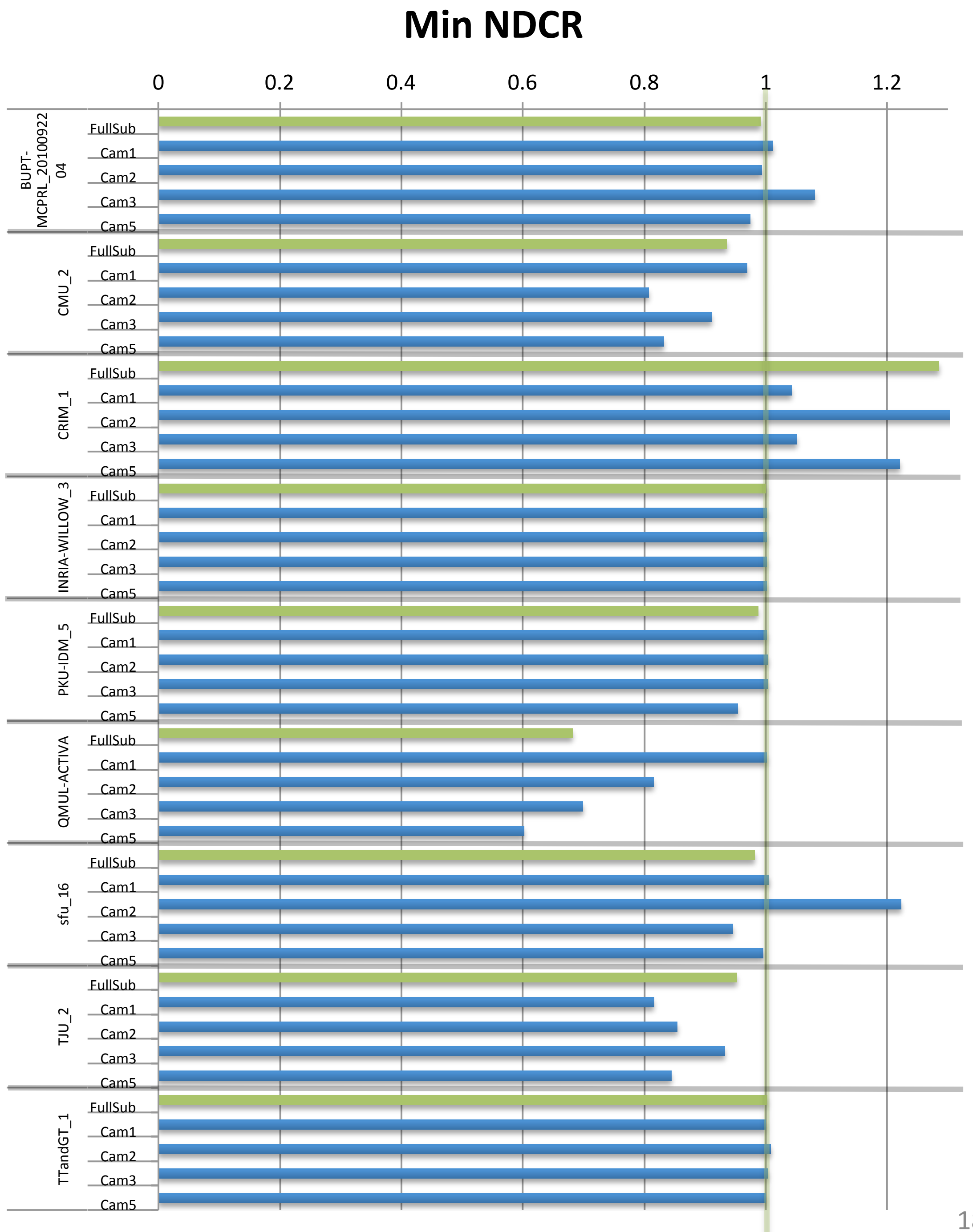
SED 2010 Best Systems (per Event) selected using the minimum Actual NDCR



“PersonRuns” NDCRs

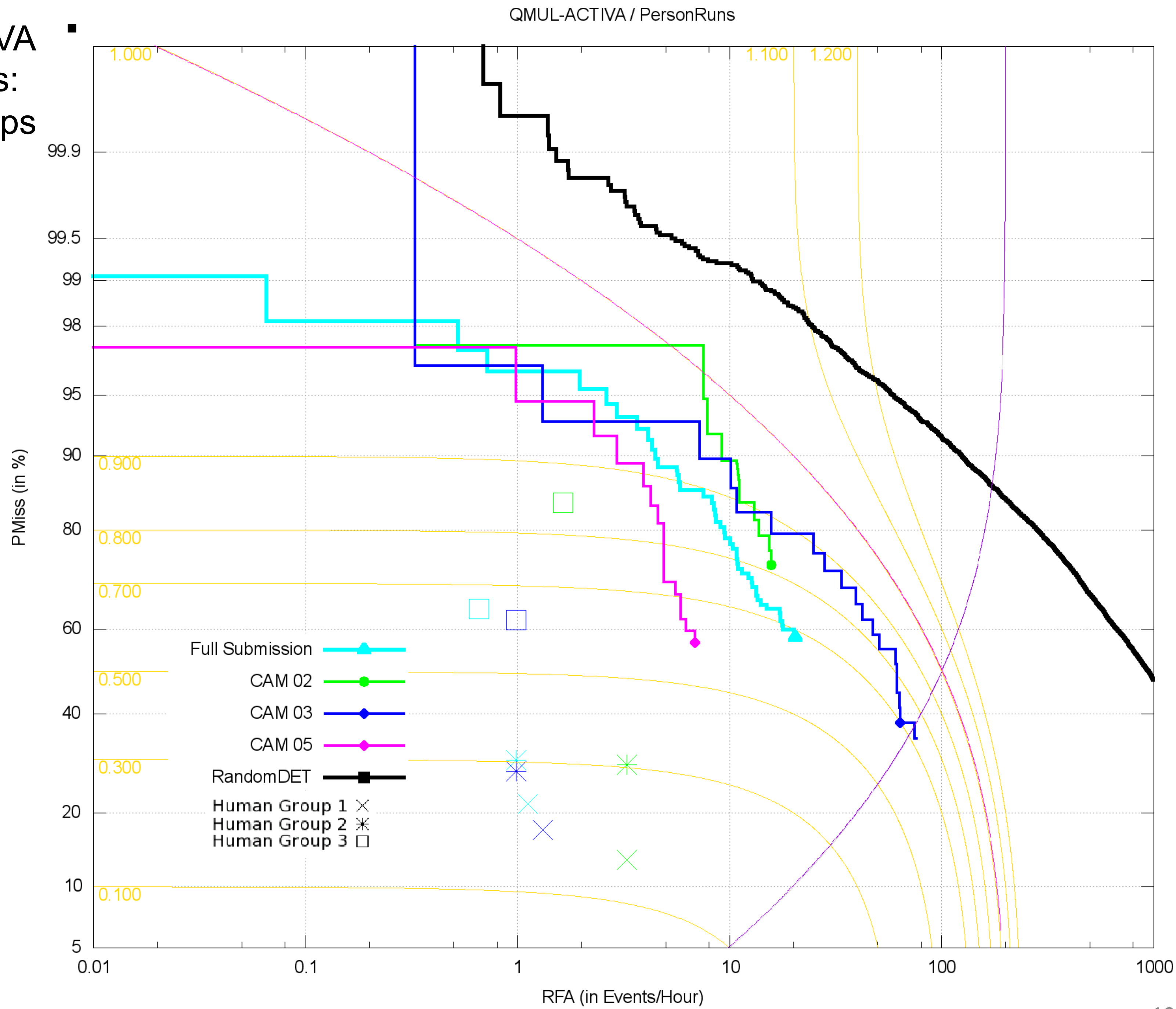
All sites, Best Run Per Event, Per cameras

- In most cases, the “Full Submission” score in close the the average of each individual camera scores
  - Use of “Per Event, Per Camera normalization”

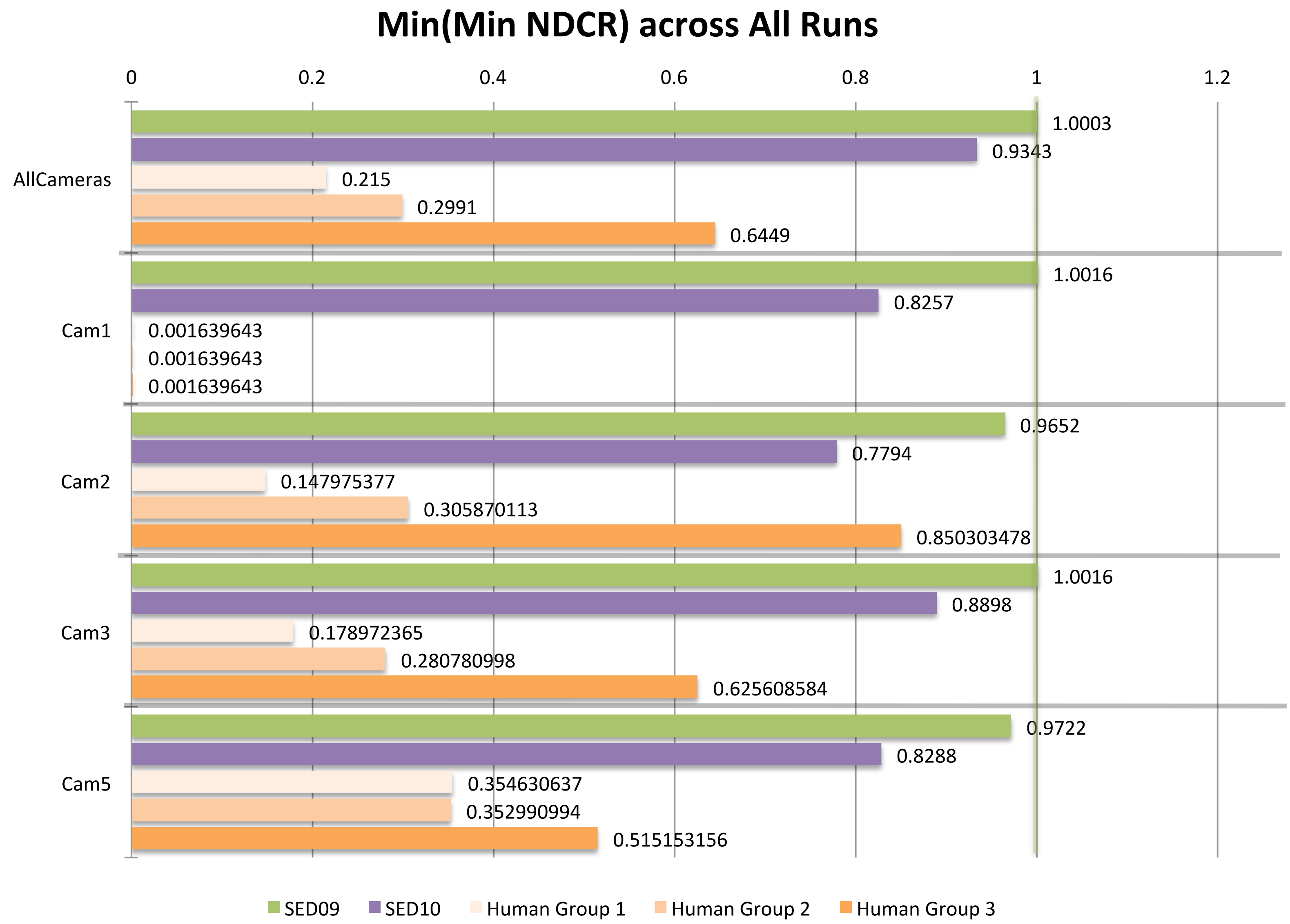


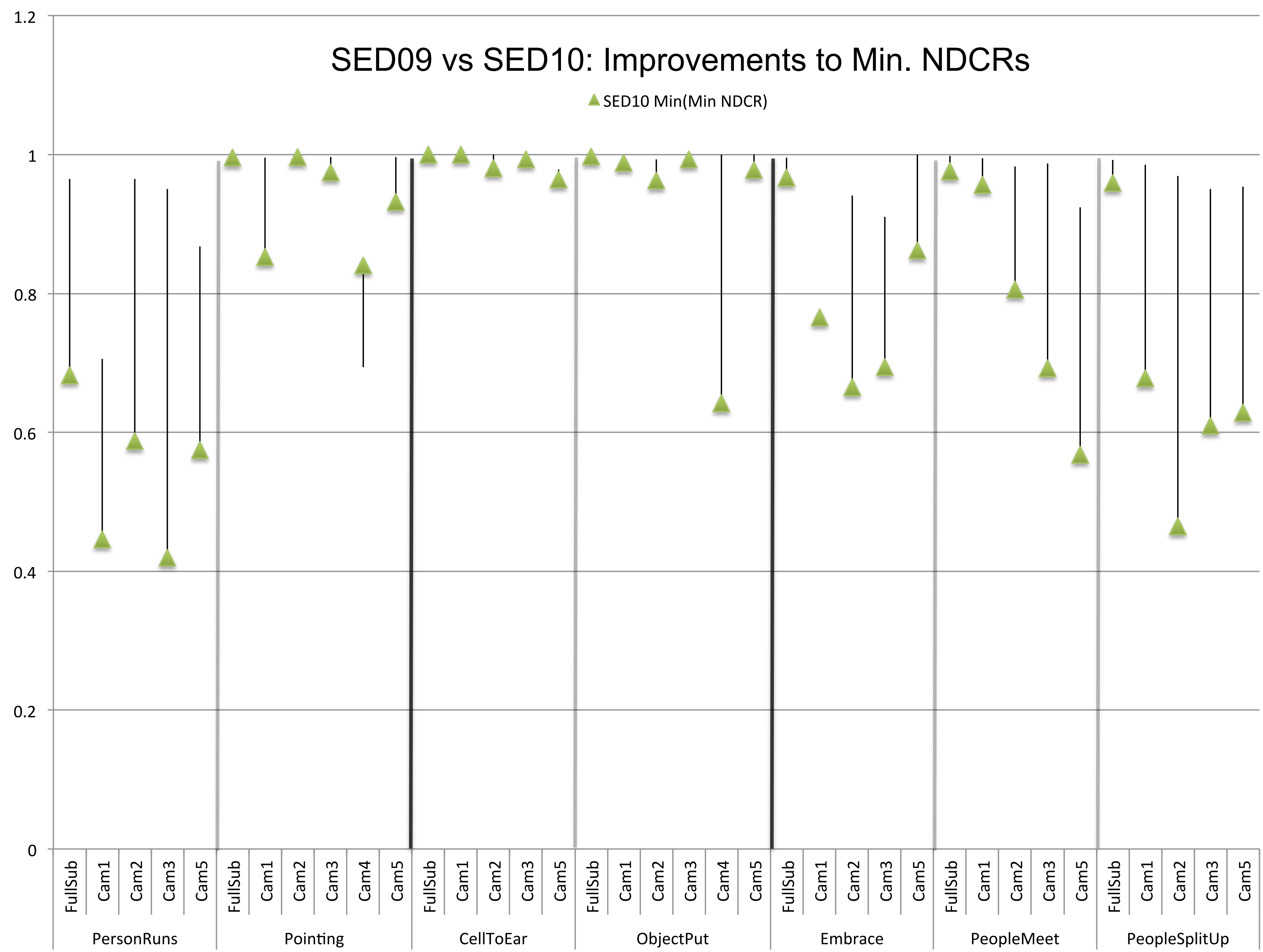
QMUL-ACTIVA PersonRuns: Camera comps

▪



CMU PersonRuns:
SED09 vs SED10 Cameras Comps




SED09 vs SED10: Improvements to Min. NDCRs
SED10 Min(Min NDCR)
1.2
1
0.8
0.6
0.4
0.2
0
FullSub
Cam1
Cam2
Cam3
Cam5
Cam4
PersonRuns
Pointing
CellToEar
ObjectPut
Embrace
PeopleMeet
PeopleSplitUp

Conclusions and Lessons Learned

- Improvement can be seen in most of the events
  - Not only by repeat participants, but results from new participants show improvement over 2009's scores
- Scoring improvement
  - Decision score normalization is important
  - Non-optimal normalization obscures performance gains
- Still areas for improvement
  - Insure the normalization of detection scores among cameras for a given event
  - Systems are still not at "Human Group" level