CMU @ TRECVID Event Detection

Ming-yu Chen & Alex Hauptmann
School of Computer Science
Carnegie Mellon University

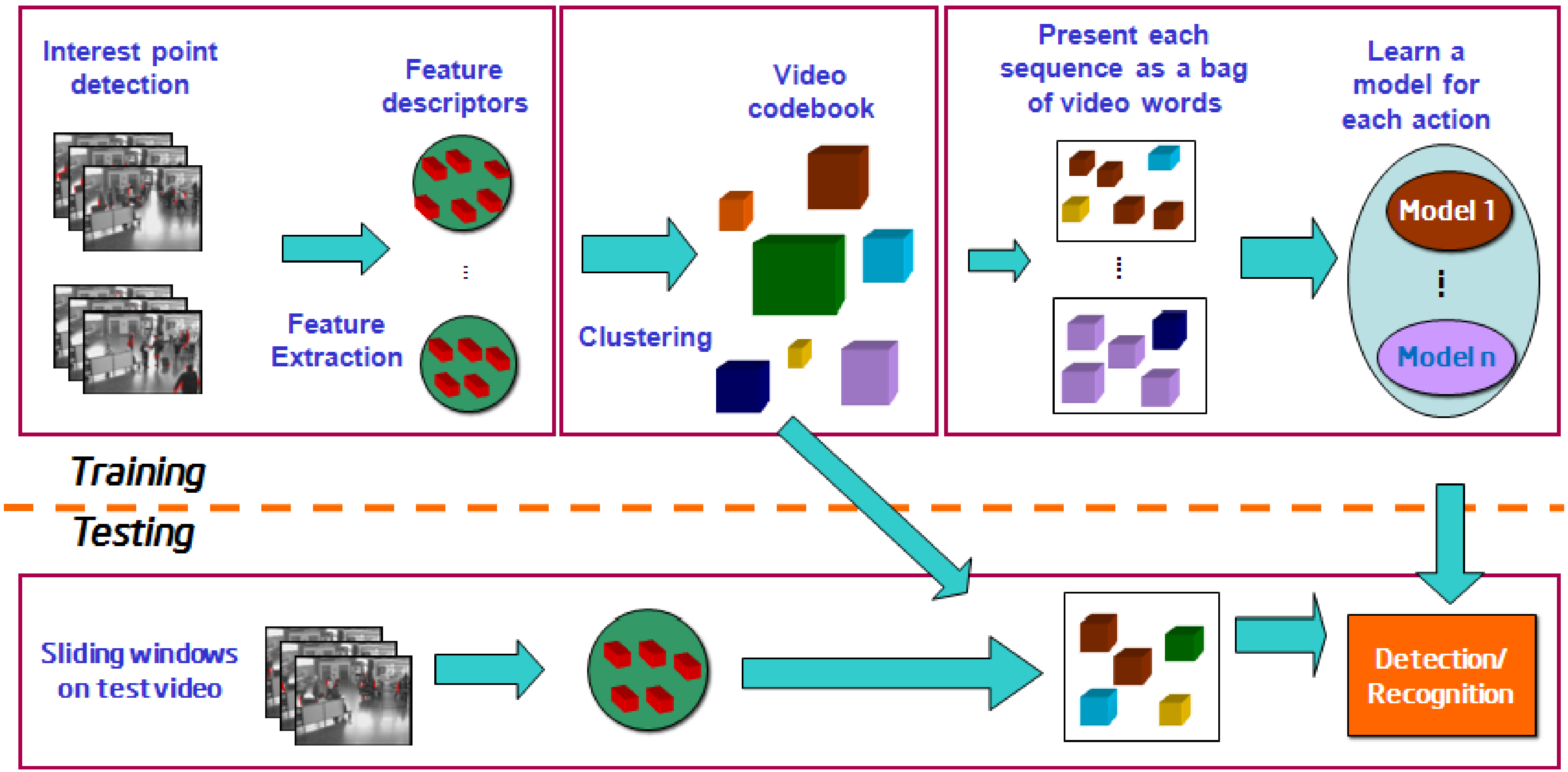
CMU @ TRECVID 2009 Event Detection

- CMU submitted all 10 event detection tasks

- Part-based generic approach
  - Local features extracted from videos
    - Local features describe both appearance and motion
    - Bag of word features represent video content
  - Robust to action deformation, occlusion and illumination

- Sliding window detection approach
  - Extend part-based method to detection tasks
  - False alarm reduction is a critical task

System overviw

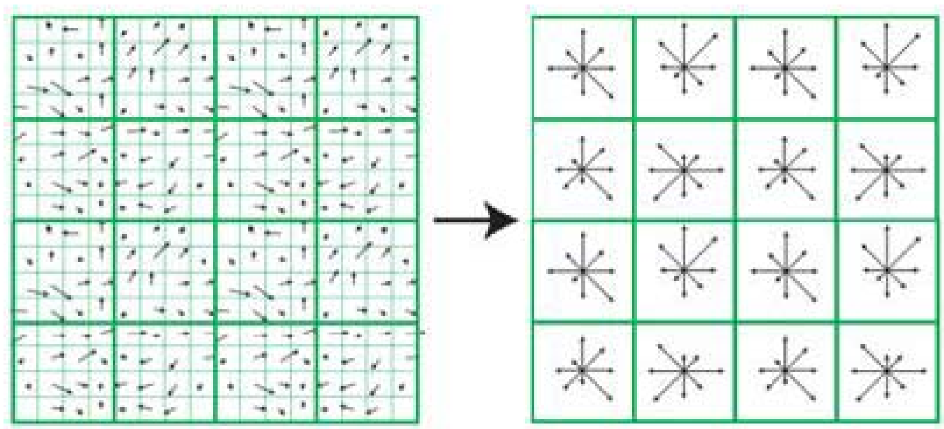


MoSIFT – feature detection

- MoSIFT detects spatial interest points in multiple scales
  - Local maximum of Difference of Gaussian (DoG)
- MoSIFT computes optical flow to detect moving areas
- MoSIFT detects video interest areas by local maximum of DoG and optical flows

MoSIFT – feature description

- Descriptor of shape
  - Histogram of Gradient (HoG)
  - Aggregate neighbor areas as 4x4 grids; each grid is described as 8 orientations
  - 4x4x8 = 128 dimensional vector to describe shape of interest areas
- Descriptor of motion
  - Histogram of Optical Flow (HoF); the same format as HoG
  - 128 dimensional vector to describe motion of interest areas
- 256 dimensional vectors as feature descriptors



Event detection

- K-mean cluster algorithm is applied to quantize feature points extracted from videos
  - K is chosen by cross-validation
- A video codebook is built by clustering result
  - A visual code is a category of similar video interest points
- Bag of word (BoW) feature is constructed for each video sequence
  - Soft weight is used to construct BoW feature
- Event models are trained by Support Vector Machine (SVM)
  - X^2 kernel is applied
- Sliding window approach creates video sequence in both training and testing sets

Evaluation metric - DCR

- Normalized Detection Cost Rate (NDCR) is used to evaluate performances.

$$DetectionCost(S,E) = Cost_{Miss} \bullet P_{Miss}(S,E) + Cost_{FA} \bullet R_{FA}(S,E)$$

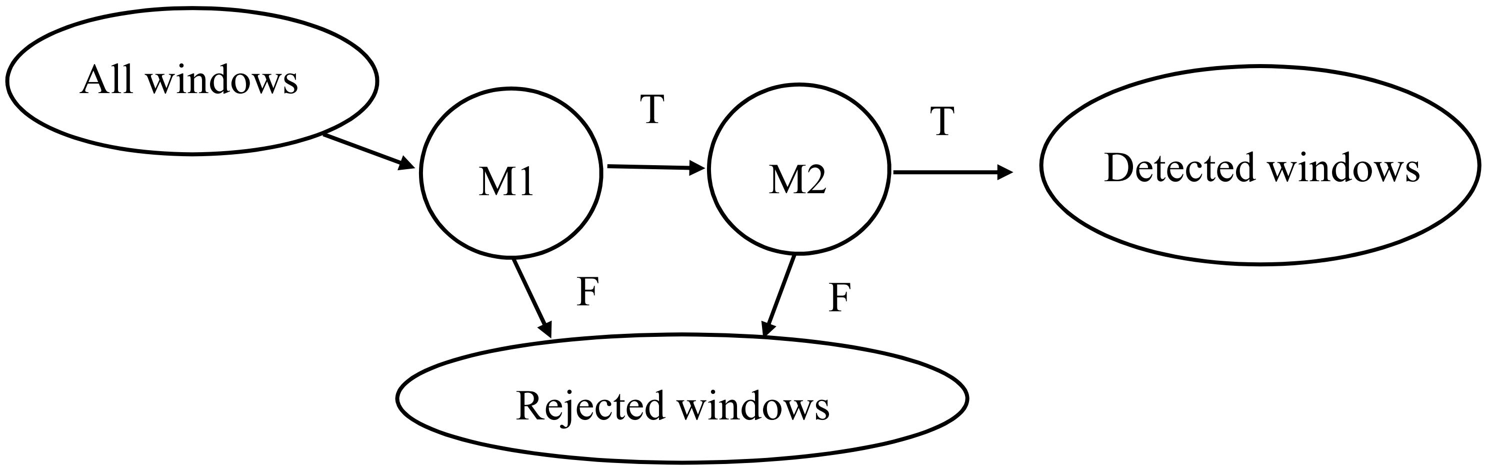
[0,1]

[0, ∞)

- Strongly penalize false alarms
  - NDCR doesn't encourage to detect more positive examples as much as reducing false alarms
  - Reducing false alarms is then extremely important to improve NDCR scores

False alarm reduction

- Cascade architecture is highly used to reduce false alarm in detect tasks
- We applied the idea of cascade algorithm in test phase to reduce false alarm
  - Two positive biased classifiers are built (due to computation, it can extend to more layers)
  - Windows pass both classifiers will be predicted as positive



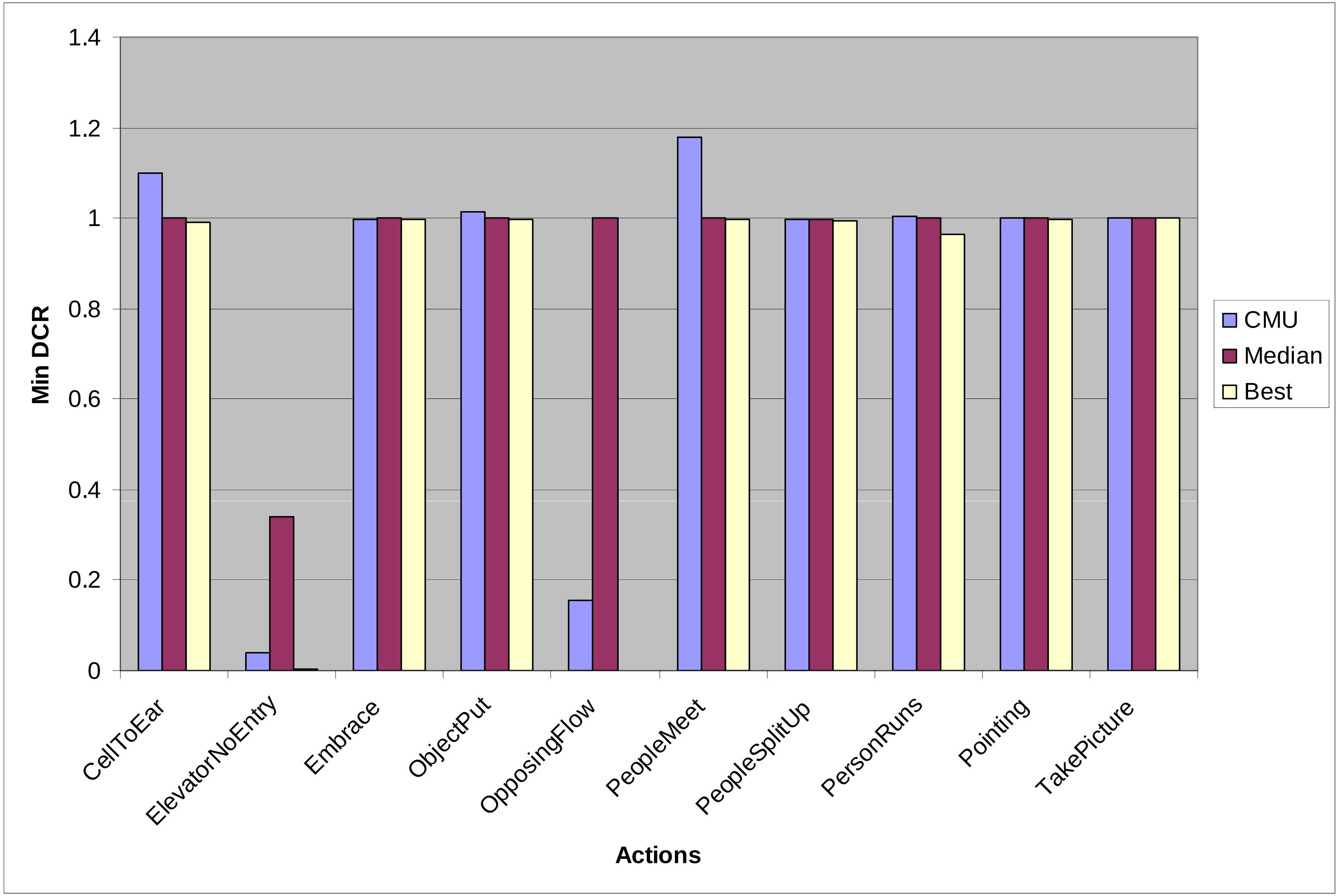
False alarm reduction (Cont.)

- Lesson from last year, multi-scale sliding window approach has a lot of false alarm
- We do not apply multi-scale this year
- Instead of several short positive predictions, we aggregated consecutive positive predictions as a long positive segment
  - Reduce number of positive predictions
- Performance improves 80% by cascade algorithm
- Performance improves 40% by concatenating short predictions to long predictions

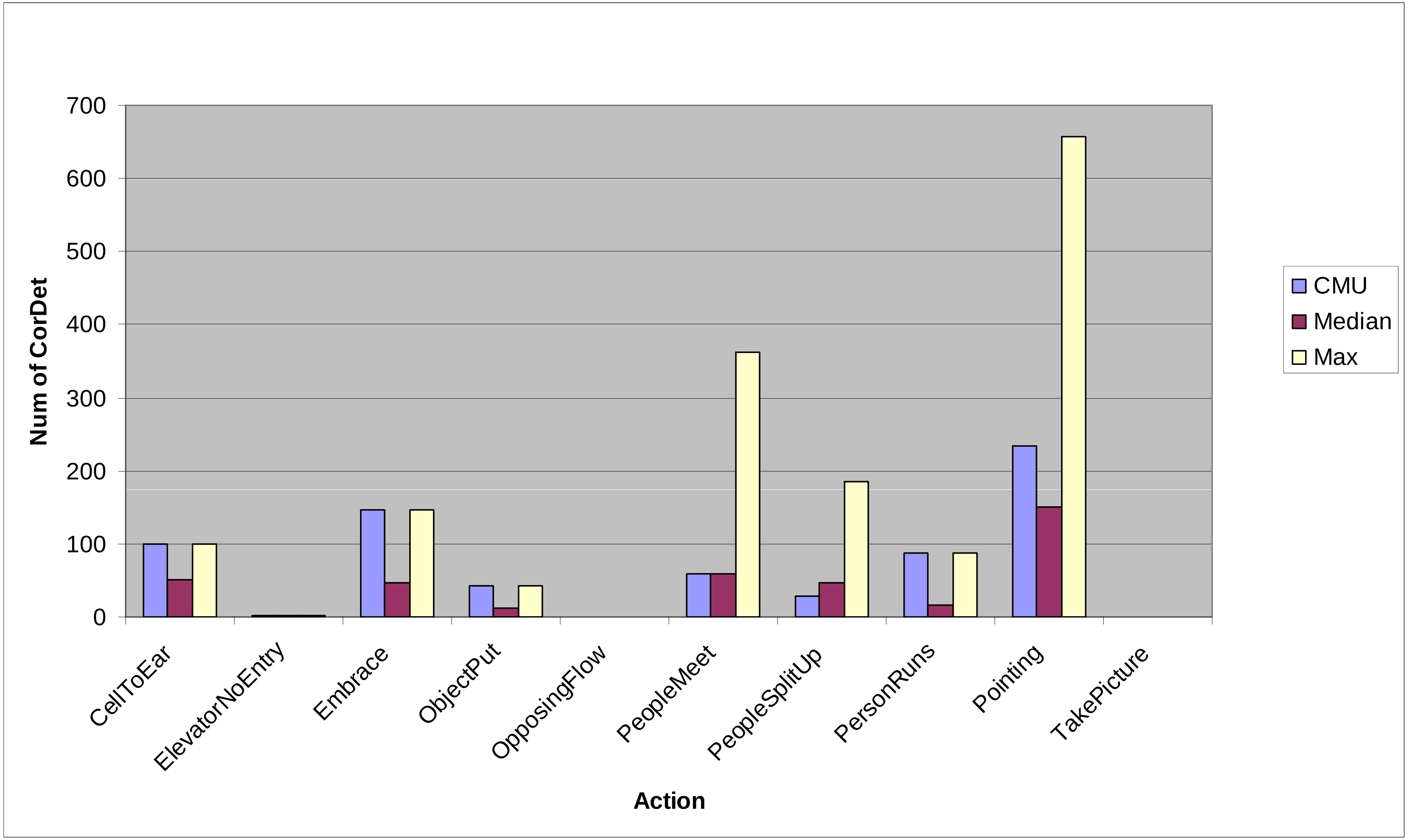
System set up

- MoSIFT features are extracted via 3 different scales every 5 frames
    - approximate 2160 hours for a single core to extract MoSIFT features
- A sliding window (25 frames) slides every 5 frames
- 1000 video codes
- Soft weighted BoW feature representation (4 nearest clusters)
- One against all SVM model for each action of each camera view
    - 50 models are built (10 actions * 5 camera views)

Performance comparison



Correct detection comparison



Performance (2008 v.s. 2009)

	#Ref	#Sys	#CorDet	#FA	#Miss	Act. RFA	Act. PMiss	Act. DCR	Min RFA	Min PMiss	Min DCR
CellToEar Event	364	24993	30	24963	334	488.4908	0.9176	3.3600	9.6669	0.9973	1.0456
ElevatorNoEntry Event	5	50	0	50	5	0.9784	1.0000	1.0049	0.9784	1.0000	1.0049
Embrace Event	405	19997	84	19913	321	389.6694	0.7926	2.7409	23.1496	0.9975	1.1133
ObjectPut Event	1958	54923	319	54604	1639	1068.5236	0.8371	6.1797	66.6702	0.9995	1.3328
OpposingFlow Event	17	150	0	150	17	2.9353	1.0000	1.0147	2.9353	1.0000	1.0147
PeopleMeet Event	1249	69898	382	69516	867	1360.3305	0.6942	7.4958	3.4245	0.9992	1.0163
PeopleSplitUp Event	681	42415	195	42220	486	826.1861	0.7137	4.8446	0.0000	0.9985	0.9985
PersonRuns Event	321	19981	39	19942	282	390.2369	0.8785	2.8297	1.7807	0.9969	1.0058
Pointing Event	2369	79865	371	79494	1998	1555.5859	0.8434	8.6213	142.4985	0.9911	1.7036
TakePicture Event	27	50	0	50	27	0.9784	1.0000	1.0049	0.9784	1.0000	1.0049

Analysis Report	#Ref	#Sys	#CorDet	#FA	#Miss	Act. RFA	Act. PMiss	Act. DCR	Min RFA	Min PMiss	Min DCR
CellToEar	194	22658	100	22558	94	1479.483	0.484	7.882	20.660	0.995	1.098
ElevatorNoEntry	3	1041	3	1038	0	68.078	0.000	0.340	7.739	0.000	0.039
Embrace	175	20080	146	19934	29	1307.386	0.166	6.703	1.377	0.989	0.996
ObjectPut	621	2353	42	2311	579	151.569	0.932	1.690	3.017	0.998	1.014
OpposingFlow	1	2195	1	2194	0	143.895	0.000	0.720	30.956	0.000	0.155
PeopleMeet	449	2130	58	2072	391	135.894	0.871	1.550	36.466	0.998	1.180
PeopleSplitUp	187	10184	28	10156	159	666.088	0.850	4.181	0.721	0.995	0.998
PersonRuns	107	23721	87	23634	20	1550.053	0.187	7.937	2.427	0.991	1.003
Pointing	1063	7941	234	7707	829	505.469	0.780	3.307	0.066	0.999	0.999
TakePicture	12	0	0	0	12	0.000	1.000	1.000	0.000	1.000	1.000

High level feature extraction

- Motion related high level features
  - 7 motion related concepts
  - Airplane flying, Person playing soccer, Hand, Person playing a musical instrument, Person riding a bicycle, Person eating, People dacing

	MAP
MM	0.24
PKU	0.21
TITG	0.20
CMU	0.18
FTRD	0.18
VIREO	0.18
Eurecom	0.18

Conclusion & future work

- Conclusion:
  - A generic approach to detect events
  - MoSIFT features captures both shape and motion information
  - Perform robust over all tasks
  - False alarm reduction is critical to improve DCR

- Future work:
  - The approach can’t localize where the action is
  - The approach can further fuse with people tracking and global features
  - Bag of word representation is lack of spatial constraints