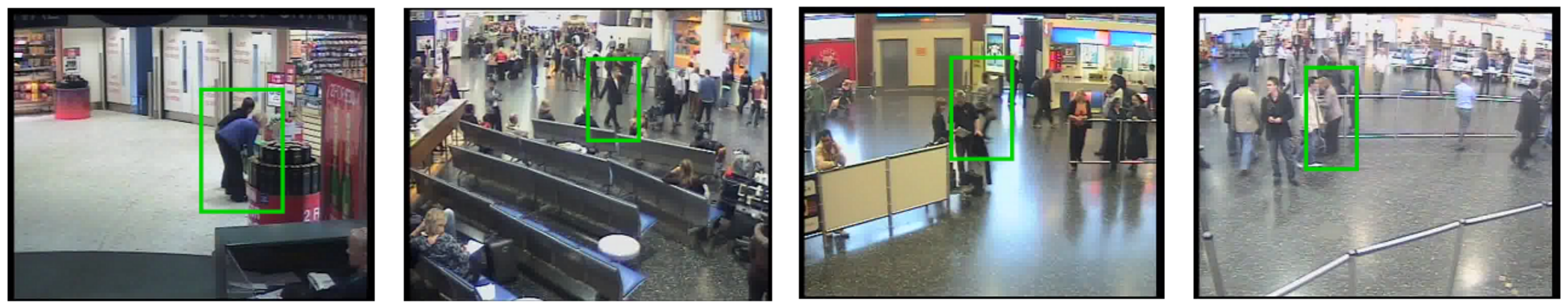
Team: UCSB_UCR_VCG
TRECVID 2012: Surveillance Event Detection

Mahmudul Hasan*, Yingying Zhu*, Santhoshkumar Sunderrajan**, Niloufar Pourian**, B.S.Manjunath**, and Amit Roy Chowdhury*
* University of California, Riverside, CA-92521
**University of California, Santa Barbara, CA-93106

Introduction

- Seven Activities: (1) CellToEar, (2) Embrace, (3) ObjectPut, (4) PeopleMeet, (5) PeopleSplitUp, (6) PersonRuns, and (7) Pointing.
- Challenges: background noise, clutter, difference of viewpoints, large crowd, illumination variation, occlusion, etc.



(a) Cam1: ObjectPut (b) Cam2: CellToEar (c) Cam3: PersonRuns (d) Cam5: Pointing

- Development Video Corpus: London Gatwick Airport, 5 Cameras, 100 hrs.
- Evaluation Video Corpus: London Gatwick Airport, 5 Cameras, 16 hrs.

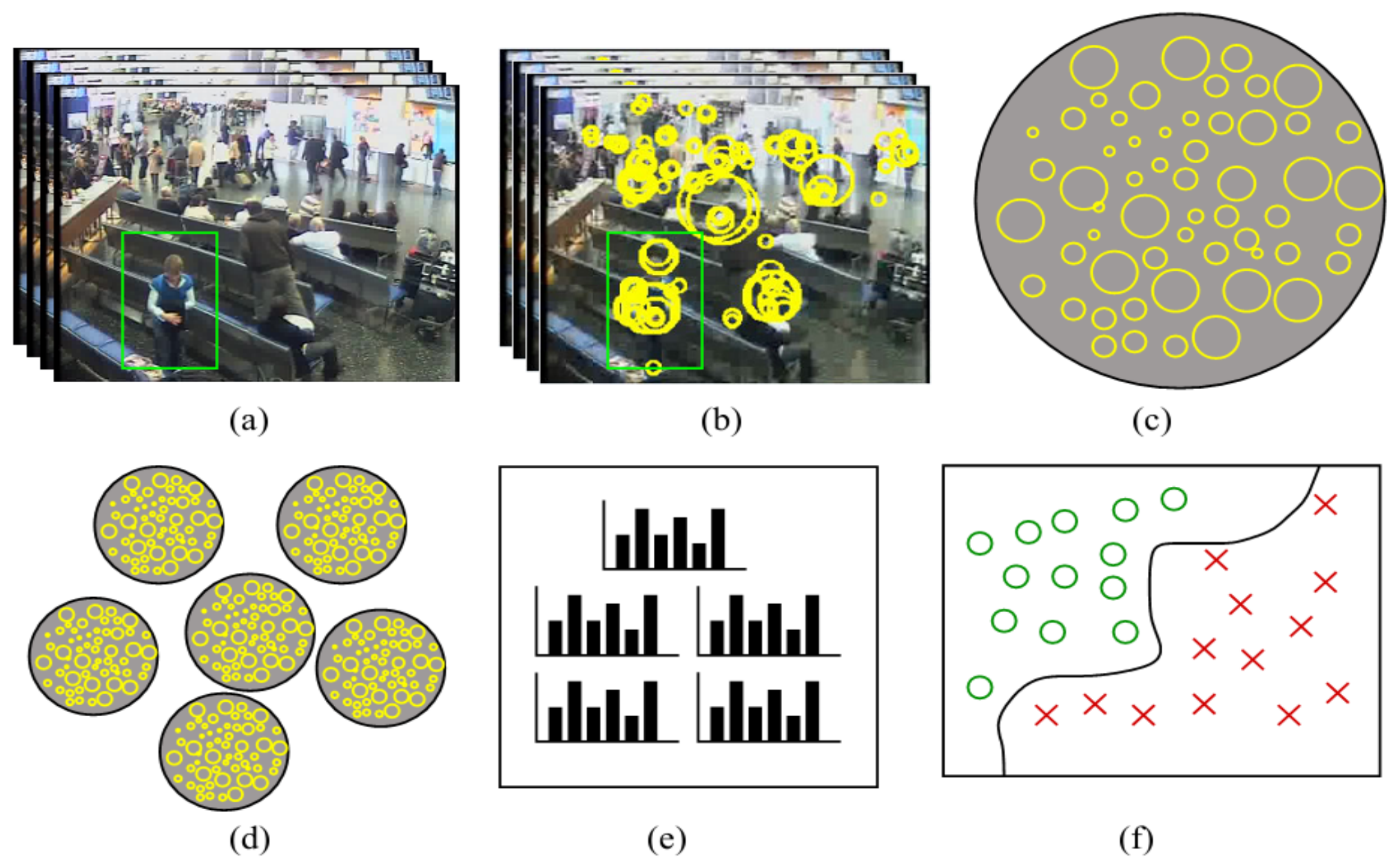
Approaches

- **Spatio-temporal cuboid based approach:** activities like CellToEar, Embrace, ObjectPut, and Pointing are the results of articulated motion of human parts. For these activities, we exploit spatio-temporal sliding cuboid based approach.
- **Track based approach:** In the activities like PeopleMeet, PeopleSplitUp and PersonRuns, the characteristics of trajectories of the persons of interest in the activities are discriminative. For these activities, we exploit track based approach.

Spatio-temporal Cuboid Based Approach: Feature extraction

- Event video clips are segmented from the video corpus and spatial extent of the activity regions are drawn.
- STIP features are generated and collected those, belong to the activity regions.
- STIP features are clustered into visual words using k-mean (400) algorithm.
- Video clips are represented using histograms of visual words.
- Discriminative classifiers are trained for each camera-activity pair using SVM.

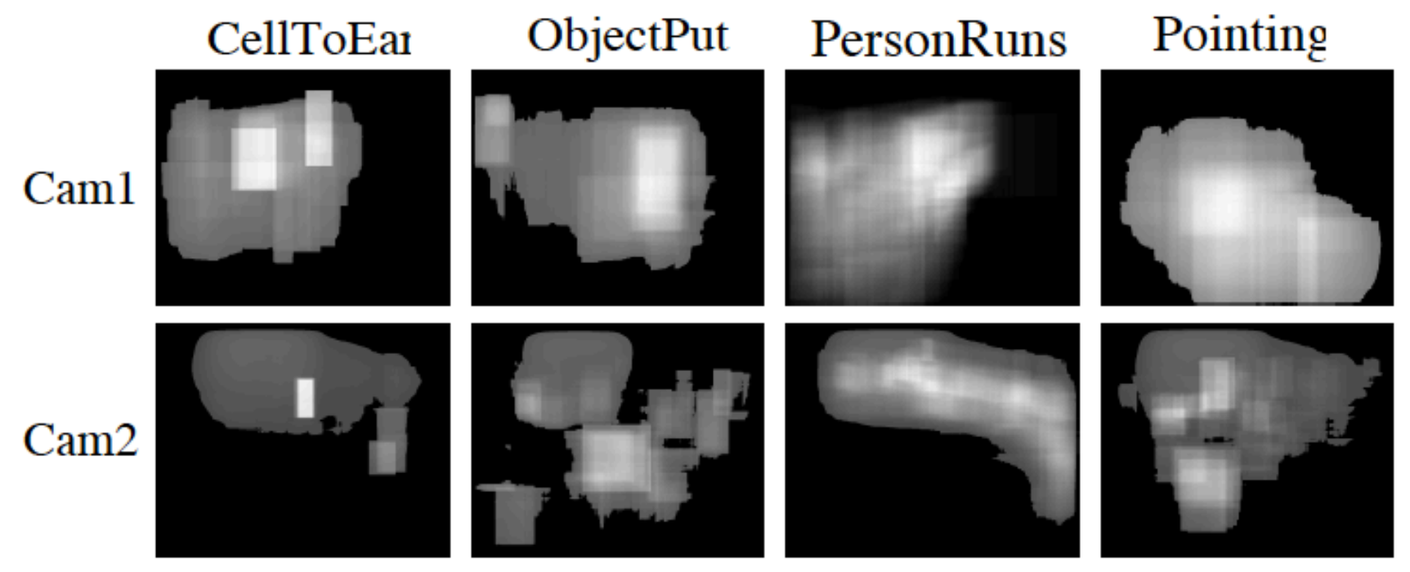
	CellToEar	Embrace	ObjectPut	Pointing	Total
CAM1	25	19	199	277	520
CAM2	94	82	280	304	760
CAM3	107	240	185	291	823
CAM4	2	2	9	18	31
CAM5	51	50	59	230	390
Total	279	393	732	1120	2524



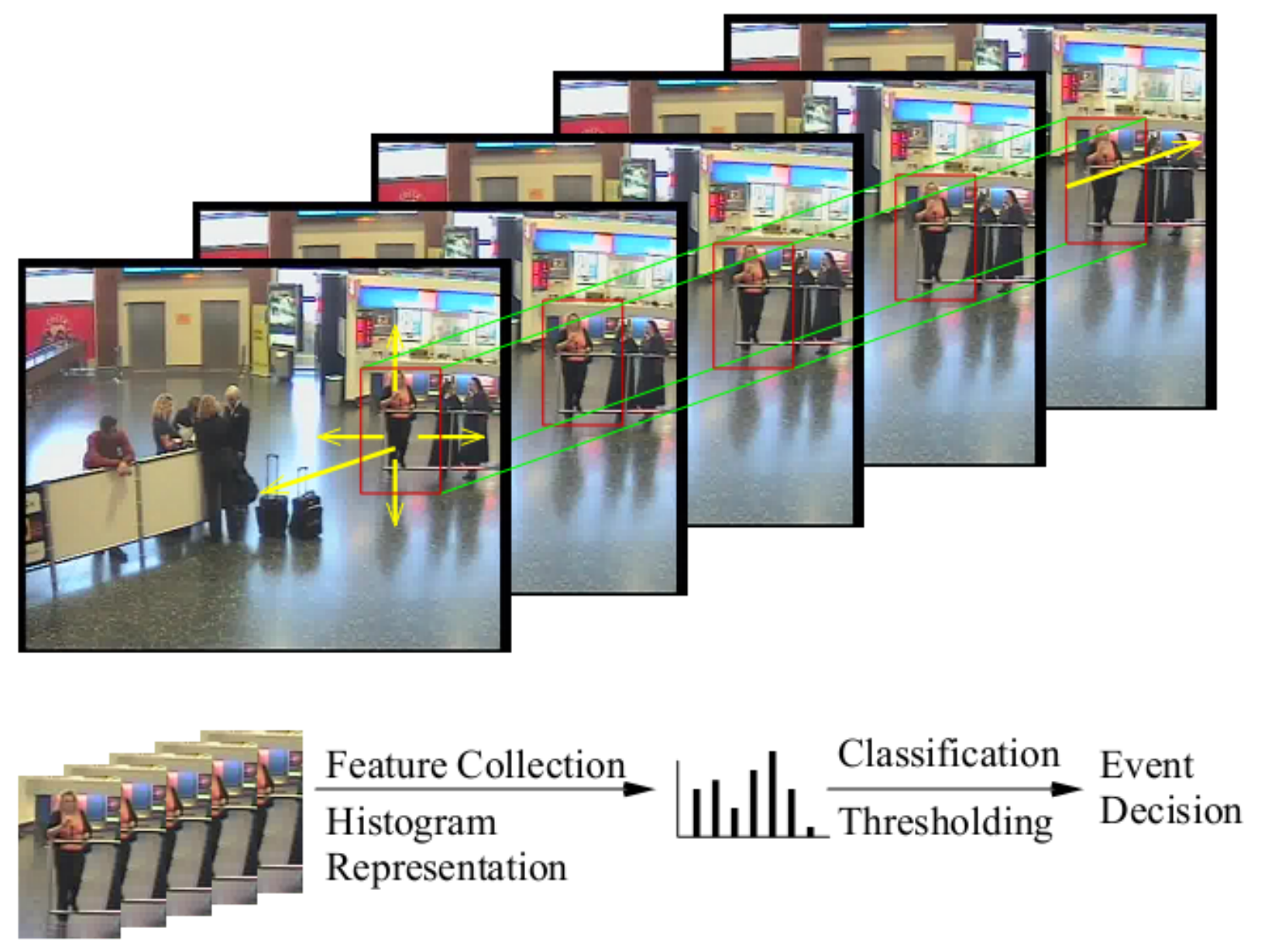
(a) (b) (c) (d) (e) (f)

Spatio-temporal Cuboid Based Approach: Evaluation

- Activities tend to occur more in some parts of the video frame, which are distinct for different cameras and activities.
- We utilize this prior information from the training videos in the evaluation phase in order to reduce the number of false alarms.
- In order to construct the activity probability map, we employ background subtraction algorithm and manually drawn bounding boxes around the activity regions.

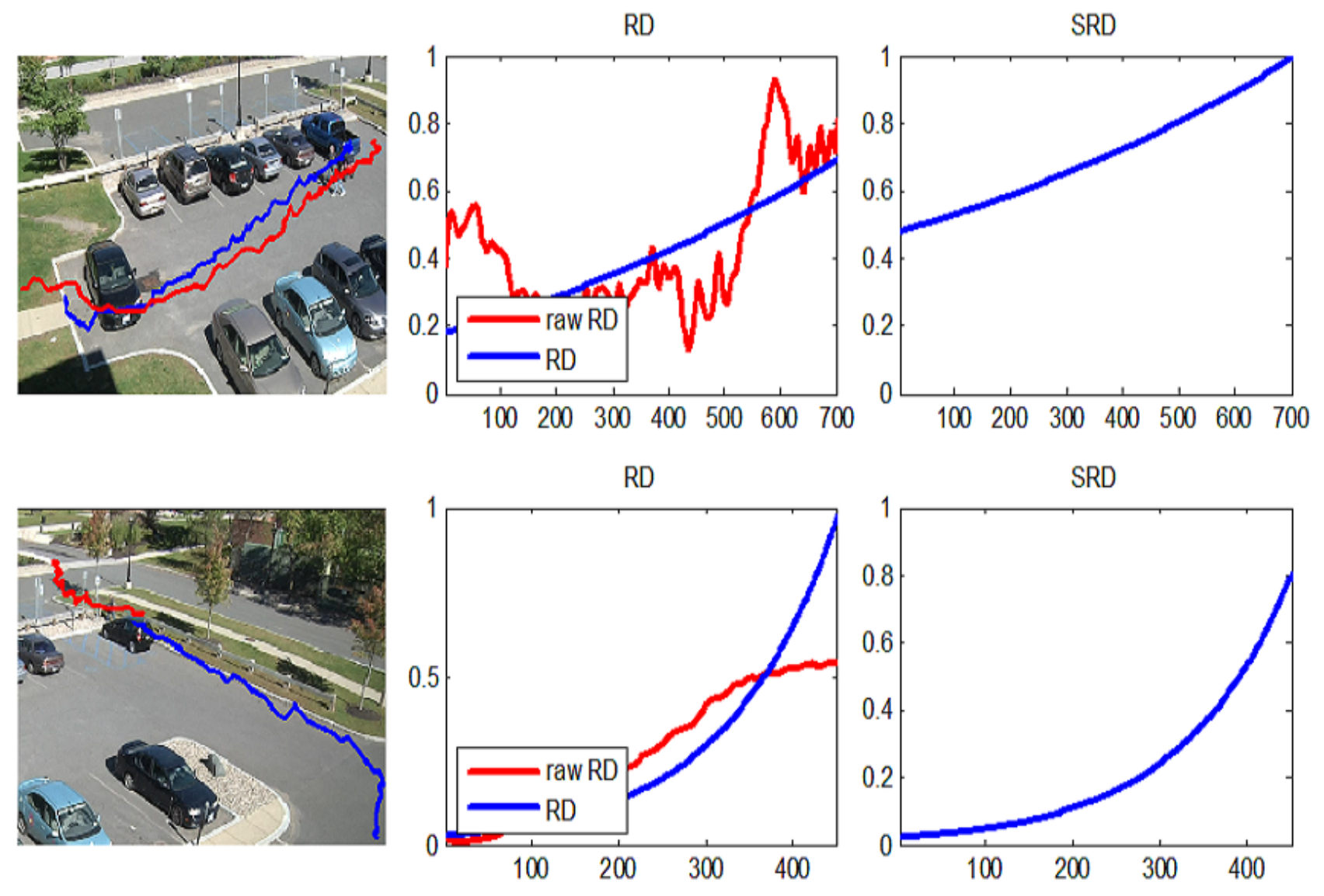


- We search the whole video using over-lapping spatio-temporal cuboids. Size of these cuboids are determined from training video corpus for each camera-activity pair.
- Features are extracted and histogram are constructed for each cuboid.
- Pre-trained classifier is used to obtain a probability. We use activity probability map to re-weight the original probability.



Track Based Approach: Feature Descriptor

- We use background subtraction and mean-shift tracker to generate tracks of moving objects.
- For PersonRuns, velocity of a trajectory and the range of the trajectory are used as the feature.
- For PeopleMeet and PeopleSplitUp, given two tracks, we introduce Slope of smoothed relative distance (SRD) to describe the convergence and divergence trends of two tracks.



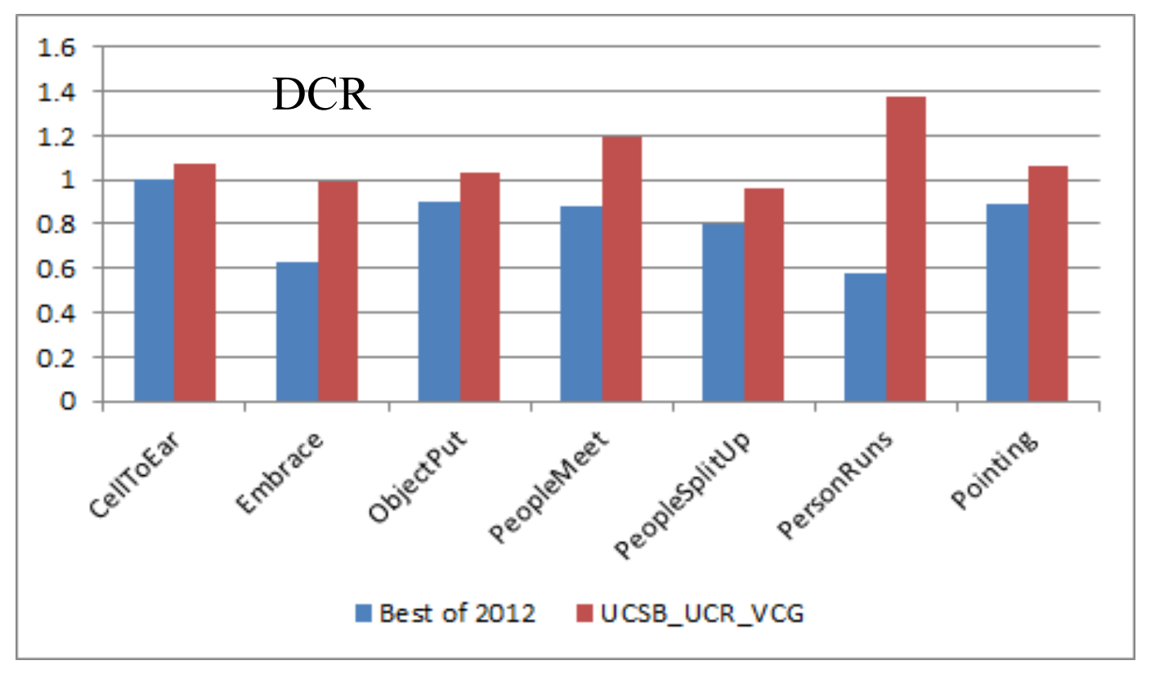
Track Based Approach: Feature Graph Matching

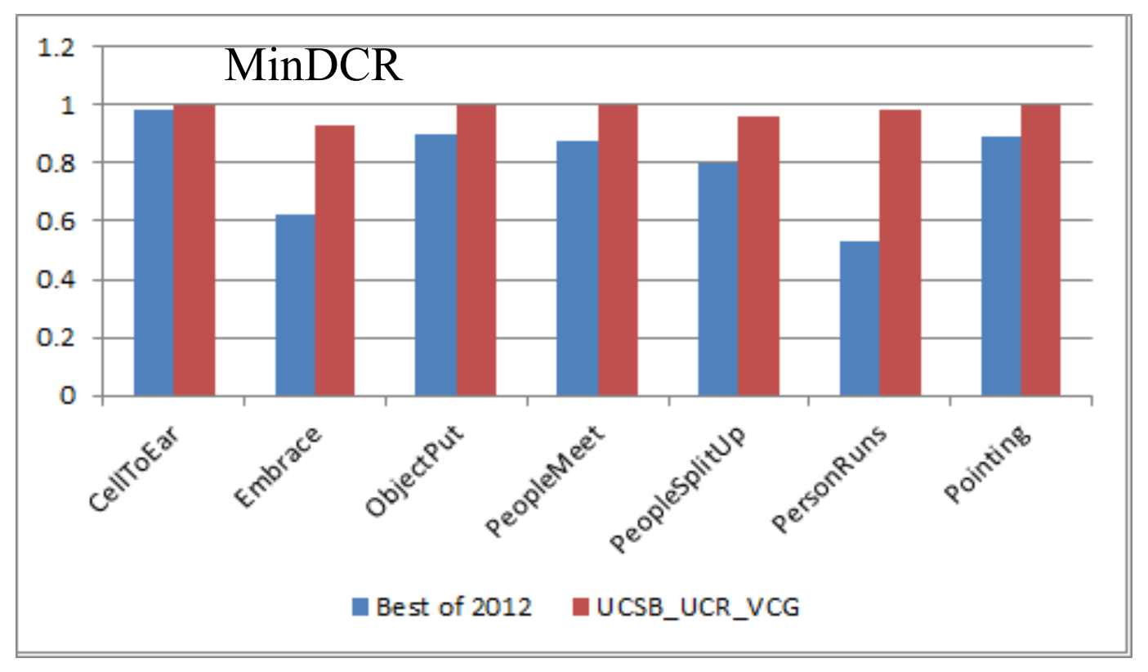
- Tracks are segmented into tracklets by concatenating equal-length time windows (size of 5 frame is used).
- Each tracklet forms a node in the feature graph. The edge features quantize the interaction between the two underlying objects.

$$d_n(i,i') = 0$$

$$d_e(\vec{ij}, \vec{i'j'}) = \frac{\|f^{SRD}_{(i)\vec{(i')}} - f^{SRD}_{(j)\vec{(j')}}\|}{s},$$

Experiments and Results





	Inputs			Actual Decision DCR Analysis								Minimum DCR Analysis			
Title	#Targ	#NTarg	#Sys	#CorDet	#Cor!Det	#FA	#Miss	RFA	PMiss	DCR	Dec. Tresh	RFA	PMiss	DCR	Dec. Thresh
CellToEar	194	260	263	3	0	260	191	17.05229	0.985	1.0698	0.3002	0.06559	1.000	1.0003	0.682
Embrace	175	338	358	20	0	338	155	22.16797	0.886	0.9966	0.6002	7.67353	0.891	0.9298	0.801
ObjectPut	621	112	116	4	0	112	617	7.34560	0.994	1.0303	0.6502	0.06559	1.000	1.0003	0.805
PeopleMeet	449	1007	1068	56	48	959	393	62.89670	0.875	1.1898	0.5031	0.06559	0.998	0.9981	0.998
PeopleSplitUp	187	335	360	24	56	279	163	18.29842	0.872	0.9631	0.5011	13.96976	0.888	0.9575	0.785
PersonRuns	107	1827	1851	24	0	1827	83	119.82510	0.776	1.3748	0.5061	3.73839	0.963	0.9813	0.982
Pointing	1063	221	230	9	0	221	1054	14.49444	0.992	1.0640	0.5700	0.19676	0.999	1.0000	0.816

- We keep five frame temporal and twenty pixel spatial distance between two overlapping cuboids.
- For PeopleMeet and PeopleSplitUp, the current system uses training instances from VIRAT Dataset release 1.
- Tracks with 5% highest velocity are classified as PersonRuns.