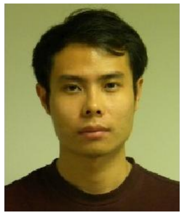
AT&T Research at TRECVID 2013: Surveillance Event Detection

Xiaodong Yang[†*], Zhu Liu[‡], Eric Zavesky[‡],
David Gibbon[‡], Behzad Shahraray[‡]

[†] City College of New York, CUNY
[‡] AT&T Labs - Research

*This work is carried out when the author worked as a research intern at AT&T Labs – Research.

Team Members



Xiaodong
Yang

Zhu
Liu

Eric
Zavesky

David
Gibbon

Behzad
Shahraray

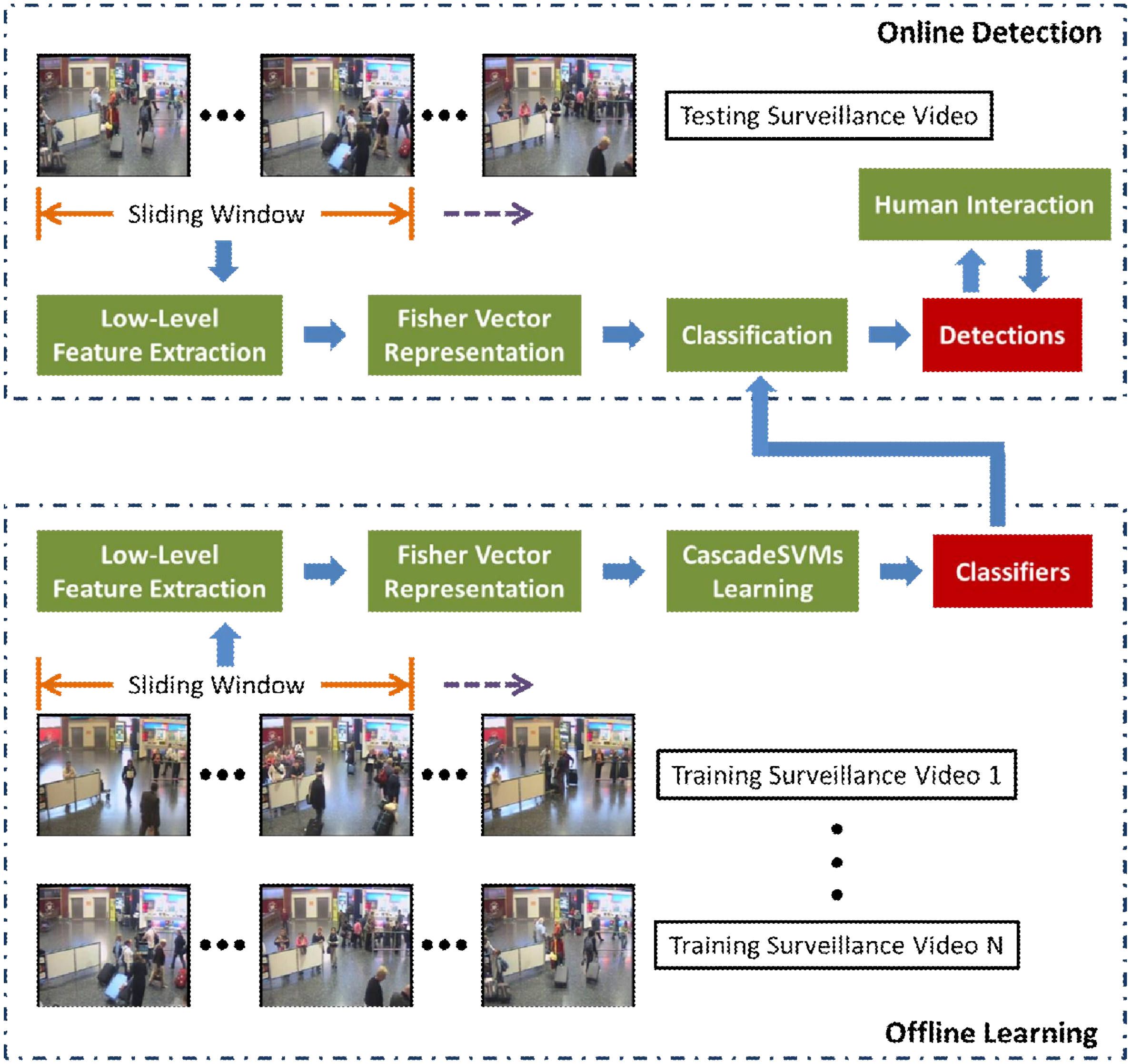
Outline

- System Overview
- Low-Level Features
- Video Representation
- CascadeSVMs
- Human Interactions
- Performance Evaluation
- Conclusion

Outline

- System Overview
- Low-Level Features
- Video Representation
- CascadeSVMs
- Human Interactions
- Performance Evaluation
- Conclusion

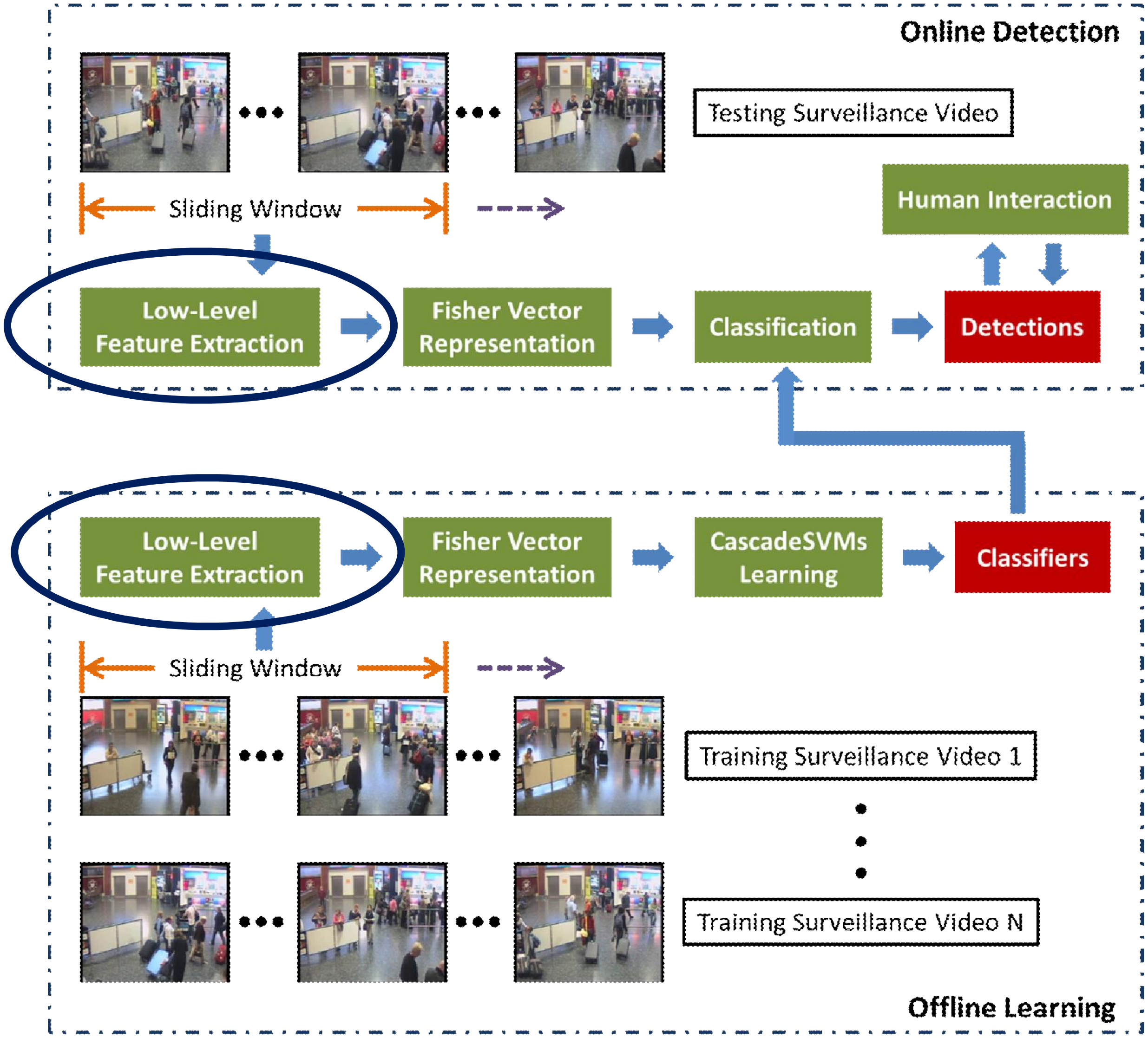
System Overview


Online Detection
Testing Surveillance Video
Sliding Window
Human Interaction
Low-Level Feature Extraction
Fisher Vector Representation
Classification
Detections
Low-Level Feature Extraction
Fisher Vector Representation
CascadeSVMs Learning
Classifiers
Sliding Window
Training Surveillance Video 1
Training Surveillance Video N
Offline Learning

Outline

- System Overview
- Low-Level Features
- Video Representation
- CascadeSVMs
- Human Interactions
- Performance Evaluation
- Conclusion

System Overview

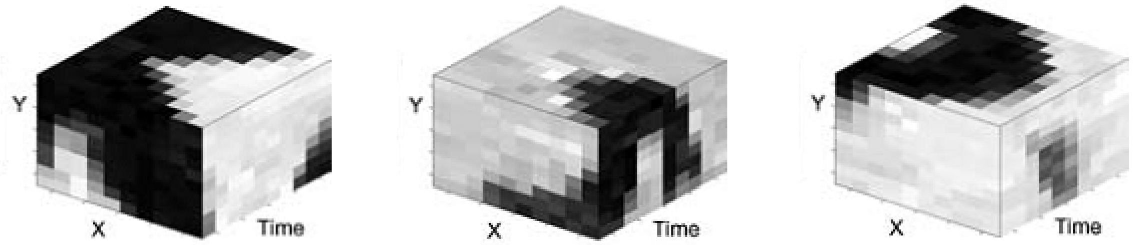

Online Detection
Testing Surveillance Video
Sliding Window
Human Interaction
Low-Level Feature Extraction
Fisher Vector Representation
Classification
Detections
Low-Level Feature Extraction
Fisher Vector Representation
CascadeSVMs Learning
Classifiers
Sliding Window
Training Surveillance Video 1
Training Surveillance Video N
Offline Learning

Low-Level Feature Extraction

- STIP-HOG/HOF
- MoSIFT
- ActionHOG
- Dense Trajectories (DT)
  - Trajectory
  - HOG
  - HOF
  - Motion Boundary Histogram (MBH)

Low-Level Feature Extraction

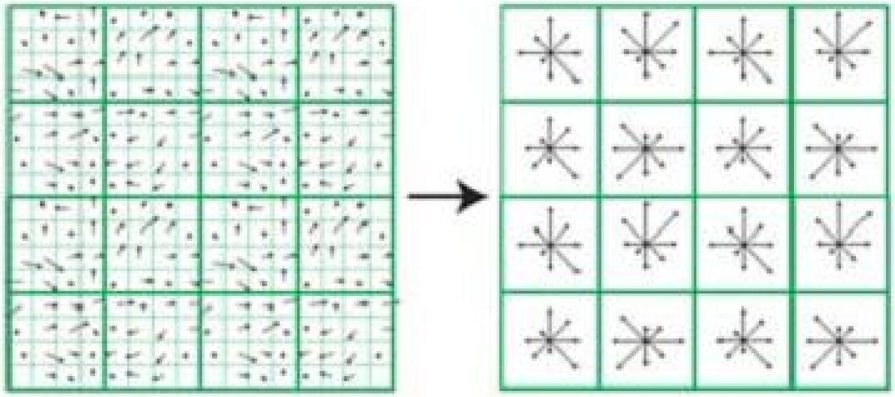
- STIP
  - 3D Harris corner detector
  - HOG-HOF descriptor



I. Laptev. On Space-Time Interest Points. *IJCV*, 2005.

Low-Level Feature Extraction

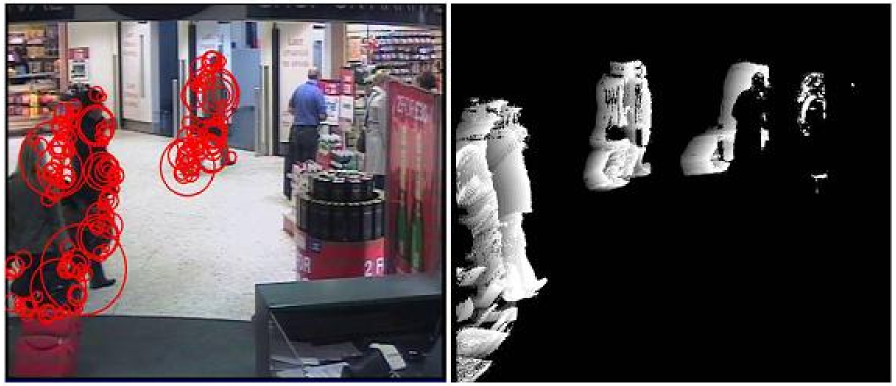
- MoSIFT
  - SIFT detector + motion
  - SIFT descriptor
    - image gradient
    - optical flow



M. Chen and A. Hauptmann. MoSIFT: Recognizing Human Actions in Surveillance Videos. *CMU-CS-09-161*, 2009.

Low-Level Feature Extraction

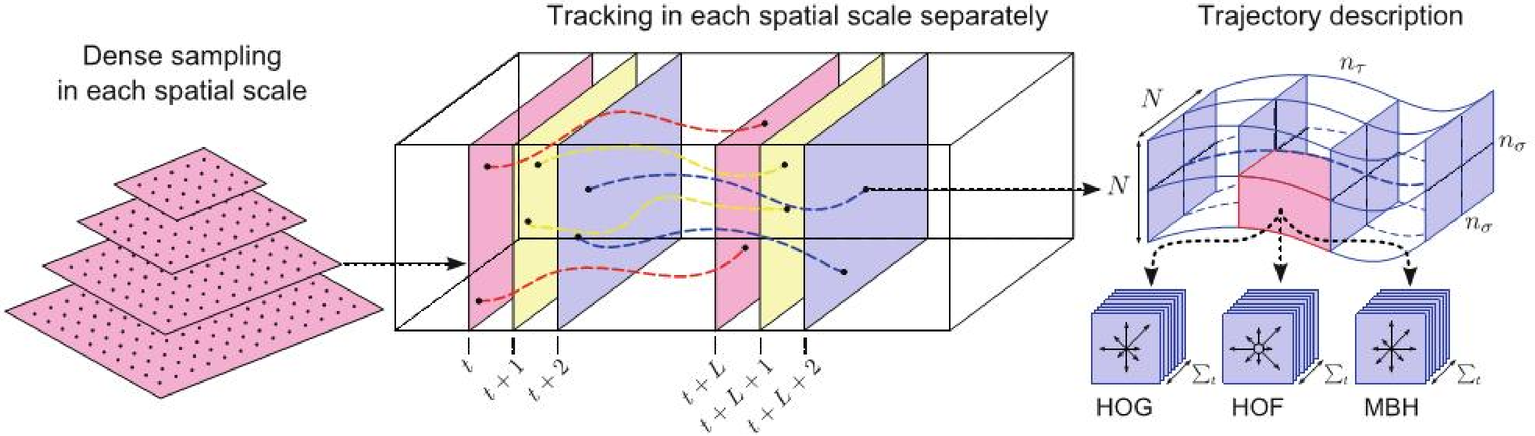
- ActionHOG
  - SURF detector + motion
  - HOG
    - image gradient
    - motion history image
    - optical flow



X. Yang, C. Yi, L. Cao, and Y. Tian. MediaCCNY at TRECVID 2012: Surveillance Event Detection. *NIST TRECVID Workshop*, 2012.

Low-Level Feature Extraction

- Dense Trajectories
  - dense sampling + tracking
  - Trajectory
  - HOG
  - HOF
  - MBH

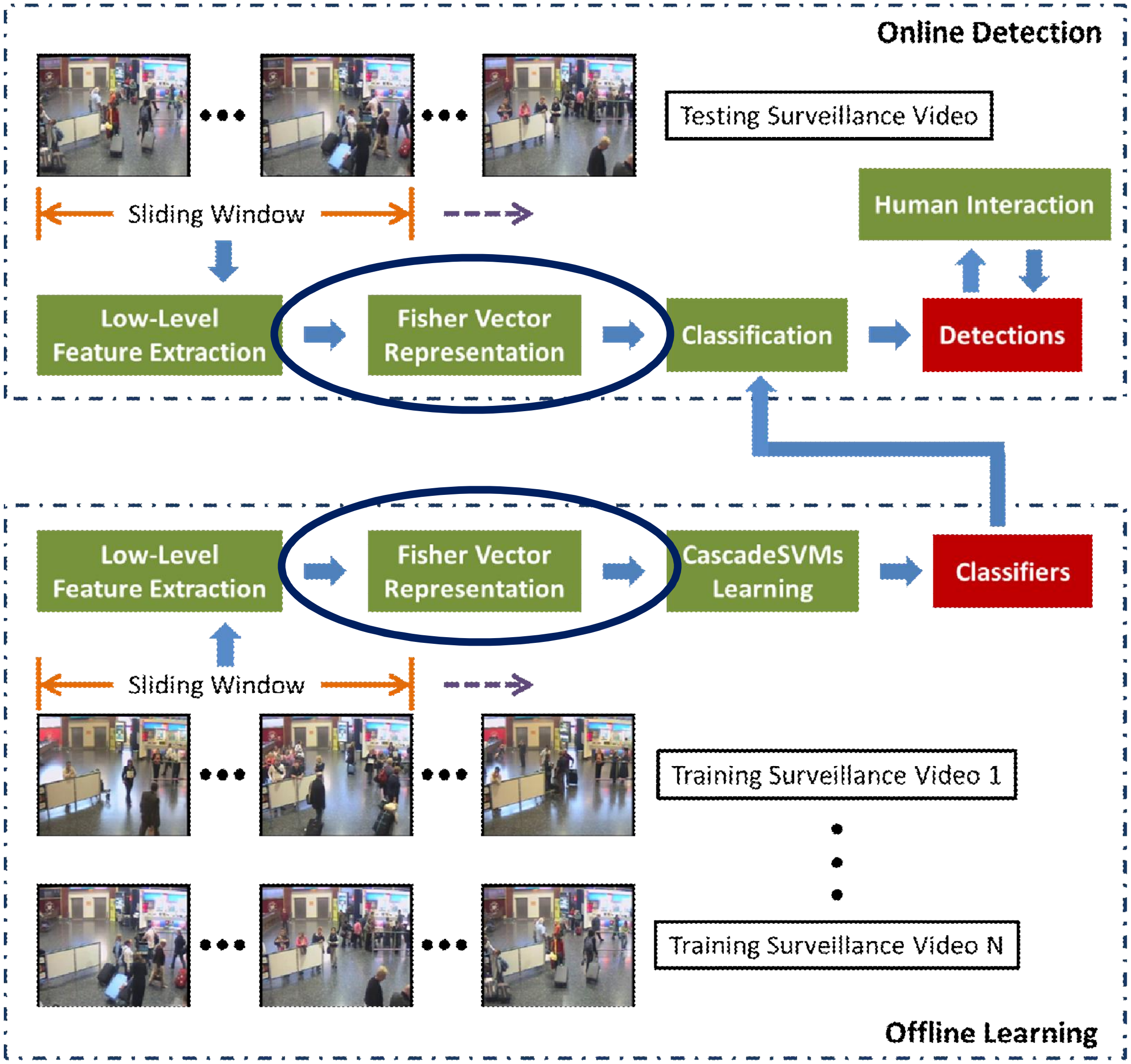


H. Wang, A. Klaser, C. Schmid, and C. Liu. Action Recognition by Dense Trajectories. *CVPR*, 2011.

Outline

- System Overview
- Low-Level Features
- Video Representation
- CascadeSVMs
- Human Interactions
- Performance Evaluation
- Conclusion

System Overview


Online Detection
Testing Surveillance Video
Sliding Window
Human Interaction
Low-Level Feature Extraction
Fisher Vector Representation
Classification
Detections
Low-Level Feature Extraction
Fisher Vector Representation
CascadeSVMs Learning
Classifiers
Sliding Window
Training Surveillance Video 1
Training Surveillance Video N
Offline Learning

Video Representation

- Fisher Vector
  - low-level features $X = \{x_t, t = 1 \dots T\}$
  - GMM $u_\lambda(x) = \sum_{i=1}^{K} w_i u_i(x)$

    $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1 \dots K\}$
  - gradient wrt. mean

$$\mathcal{G}_{\mu,i}^{X} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right)$$

  - gradient wrt. variance

$$\mathcal{G}_{\sigma,i}^{X} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right]$$

F. Perronnin, J. Sanchez, and T. Mensink. Improving The Fisher Kernel for Large-Scale Image Classification. *ECCV*, 2010.

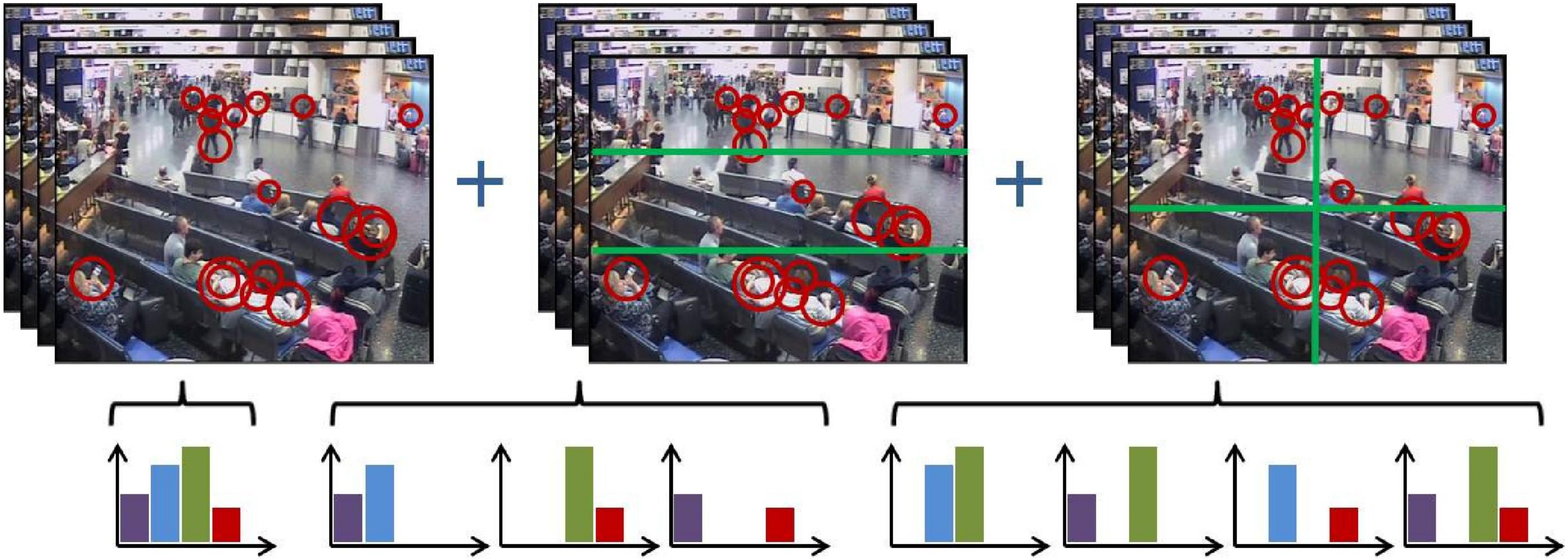
Video Representation

- Fisher Vector
  - $\mathcal{G}_{\lambda}^{X}$ concatenation of $\mathcal{G}_{\mu,i}^{X}$ and $\mathcal{G}_{\sigma,i}^{X}$ $i = 1 \ldots K$
  - dimension of $2KD$
  - GMM-128

Feature	STIP	MoSIFT	ActionHOG	DT-HOG	DT-HOF	DT-MBH	DT-Traj
Feat-Dim	162	256	216	96	108	192	30
FV-Dim	330K	520K	440K	200K	220K	400K	60K

Video Representation

- Spatial Pyramids

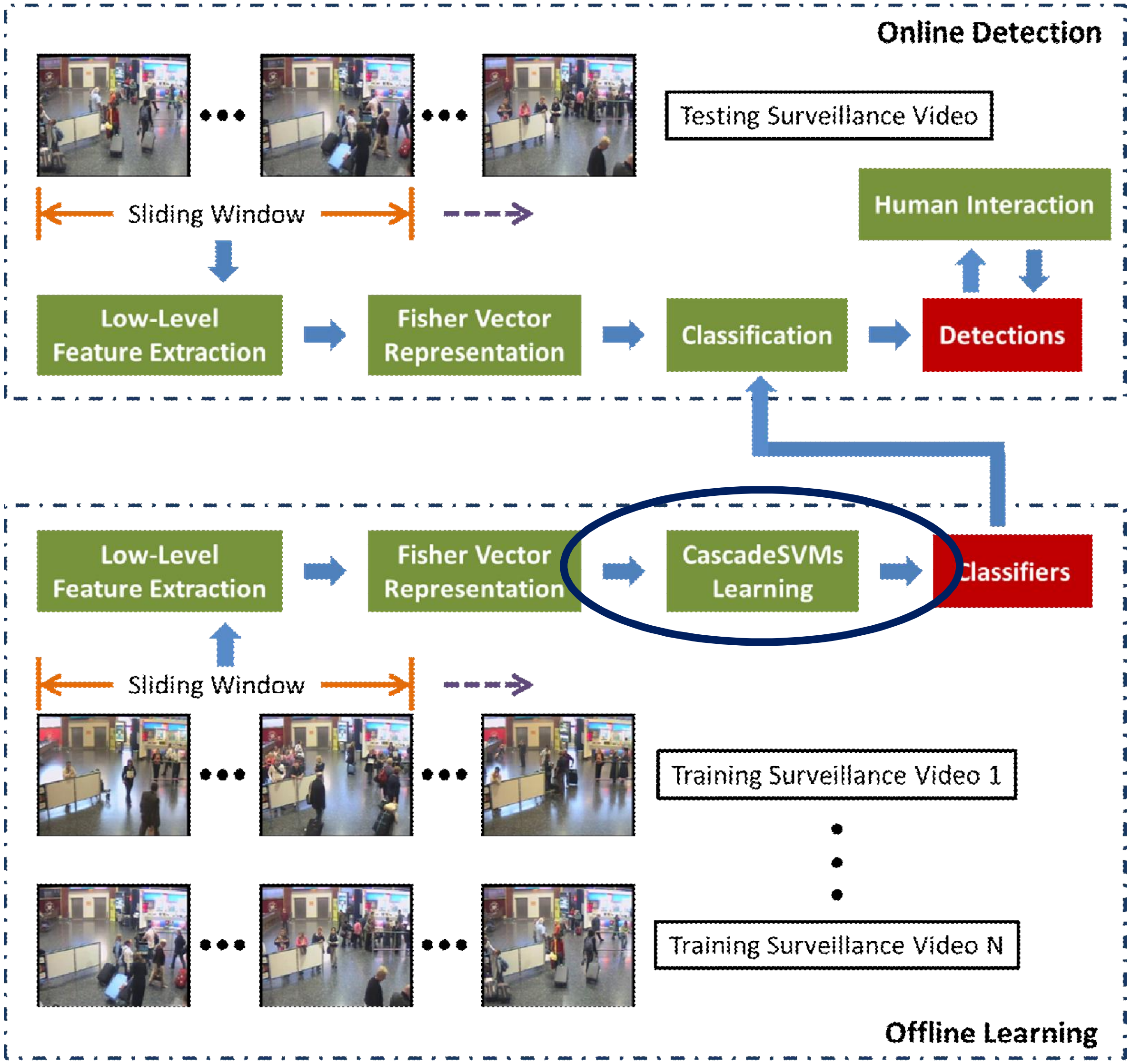


S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bag of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *CVPR*, 2006.

Outline

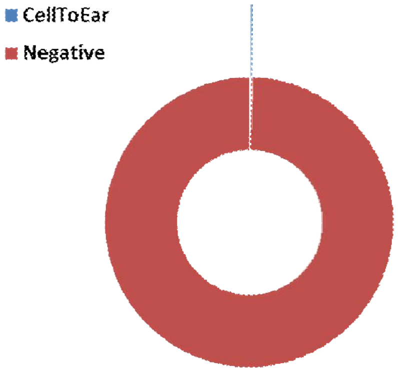
- System Overview
- Low-Level Features
- Video Representation
- CascadeSVMs
- Human Interactions
- Performance Evaluation
- Conclusion

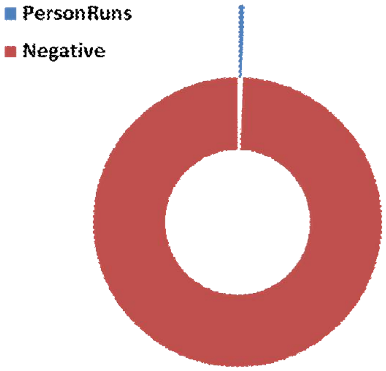
System Overview

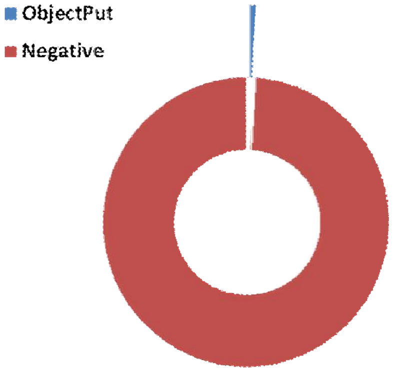

Online Detection
Testing Surveillance Video
Sliding Window
Human Interaction
Low-Level Feature Extraction
Fisher Vector Representation
Classification
Detections
Low-Level Feature Extraction
Fisher Vector Representation
CascadeSVMs Learning
Classifiers
Sliding Window
Training Surveillance Video 1
Training Surveillance Video N
Offline Learning

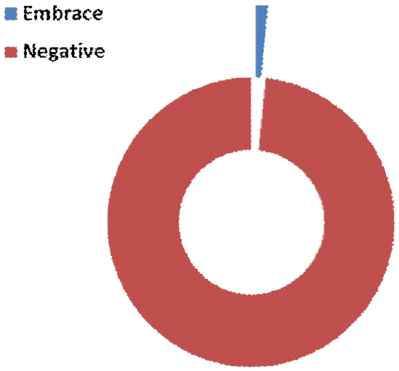
CascadeSVMs

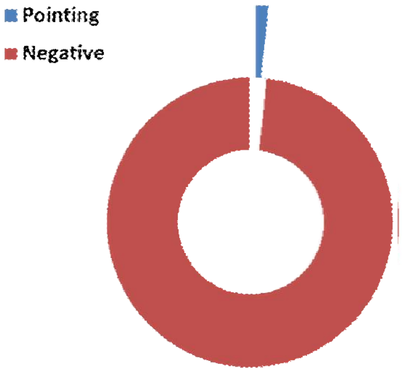
- Imbalanced Data

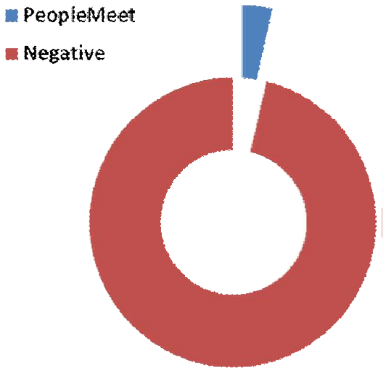


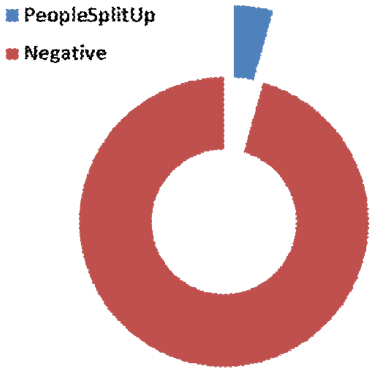






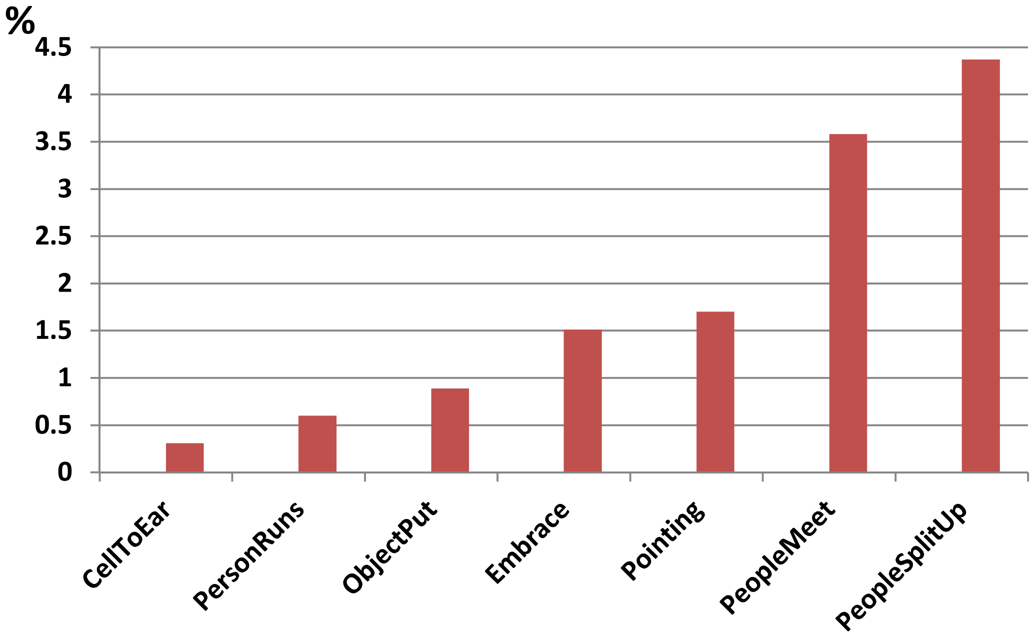




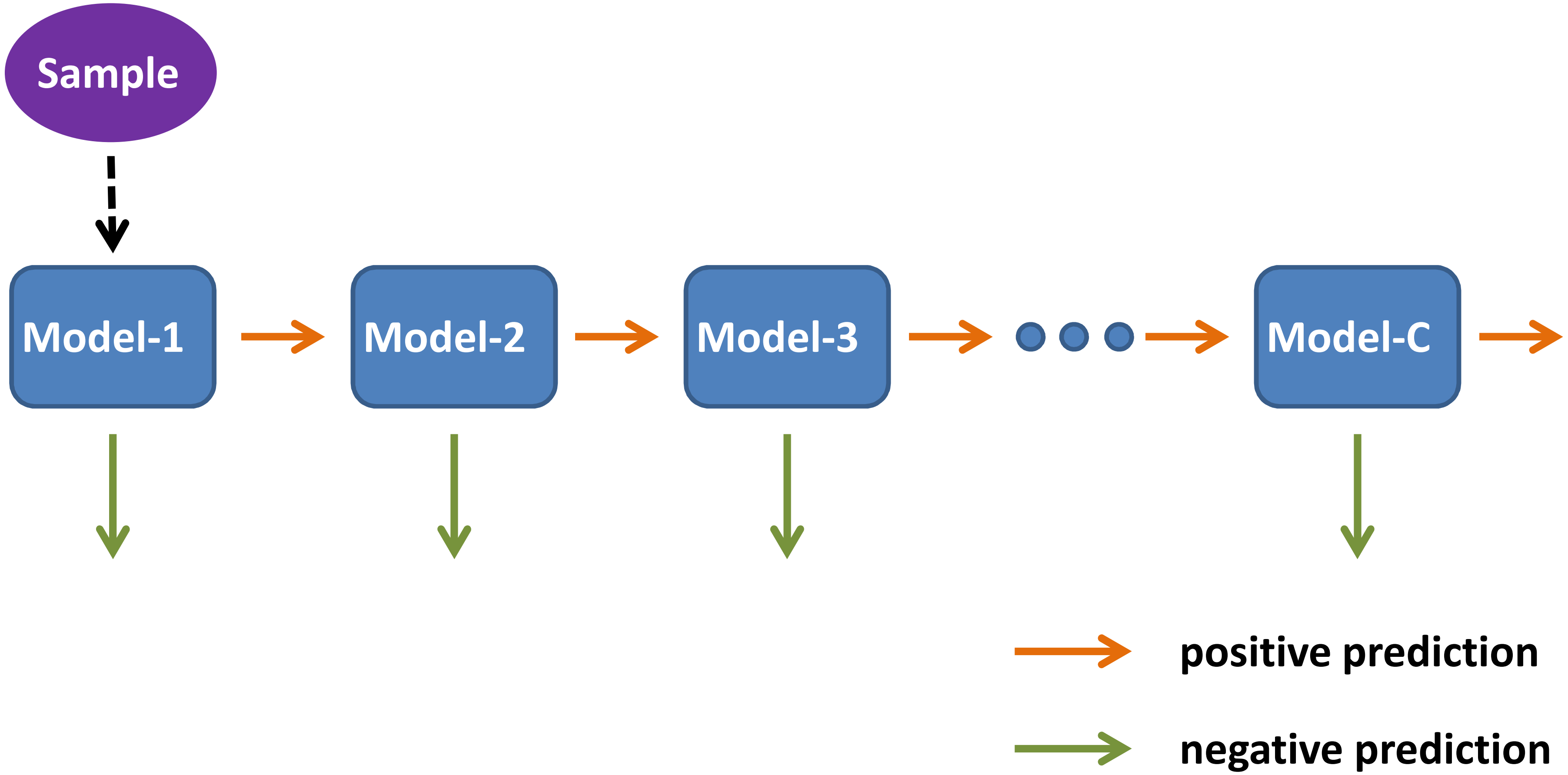


CascadeSVMs

- Imbalanced Data



CascadeSVMs



Sample

Model-1

Model-2

Model-3

Model-C

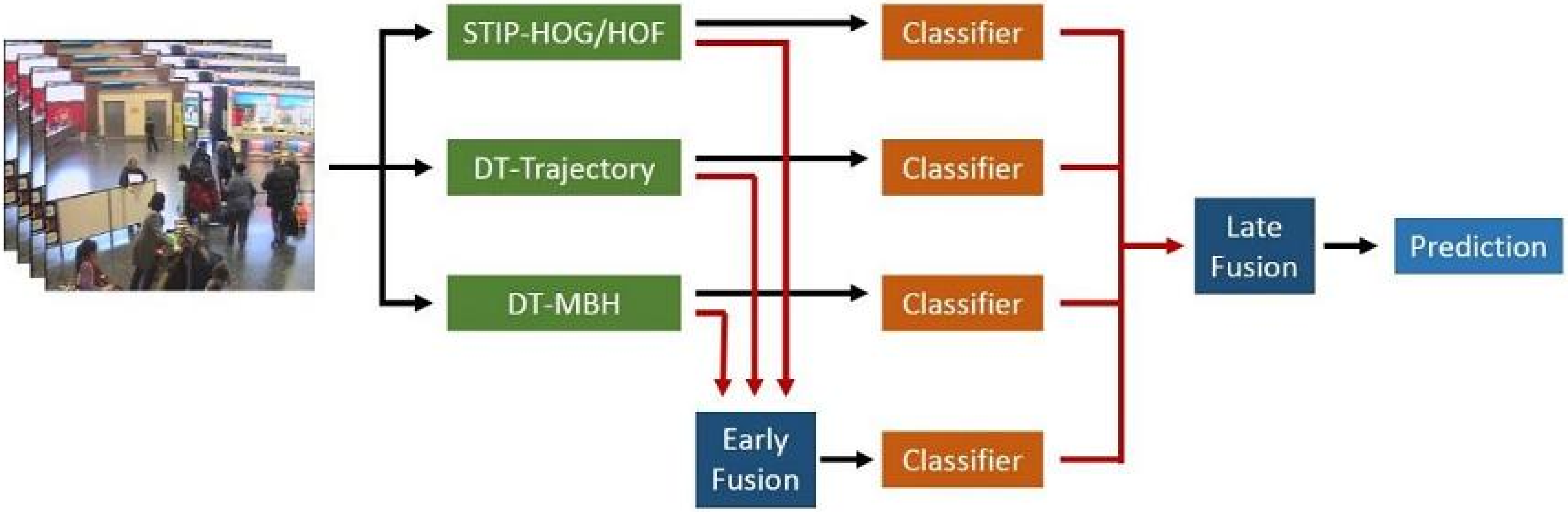
positive prediction

negative prediction

X. Yang, C. Yi, L. Cao, and Y. Tian. MediaCCNY at TRECVID 2012: Surveillance Event Detection. *NIST TRECVID Workshop*, 2012.

CascadeSVMs

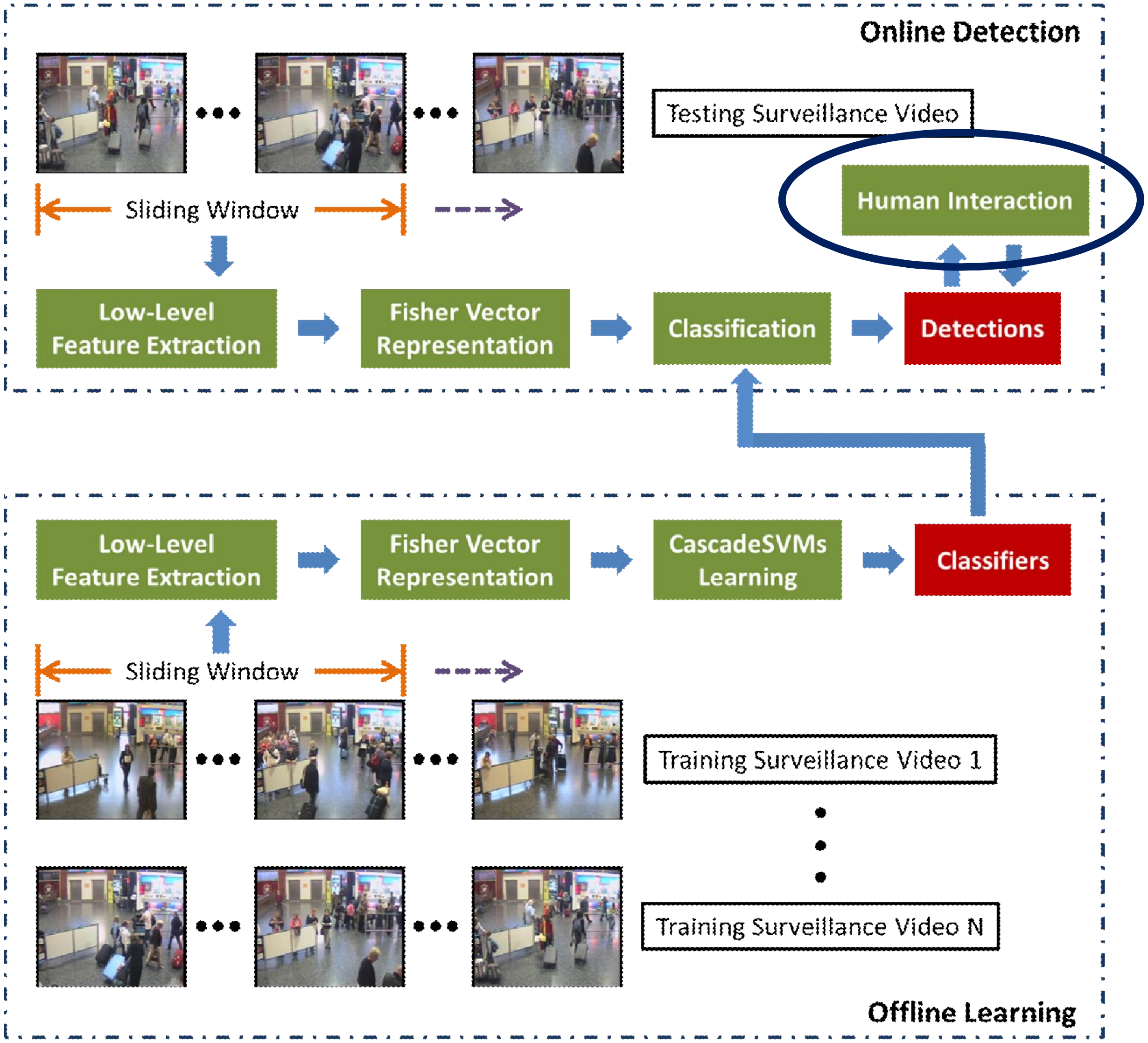
- Feature Fusion



Outline

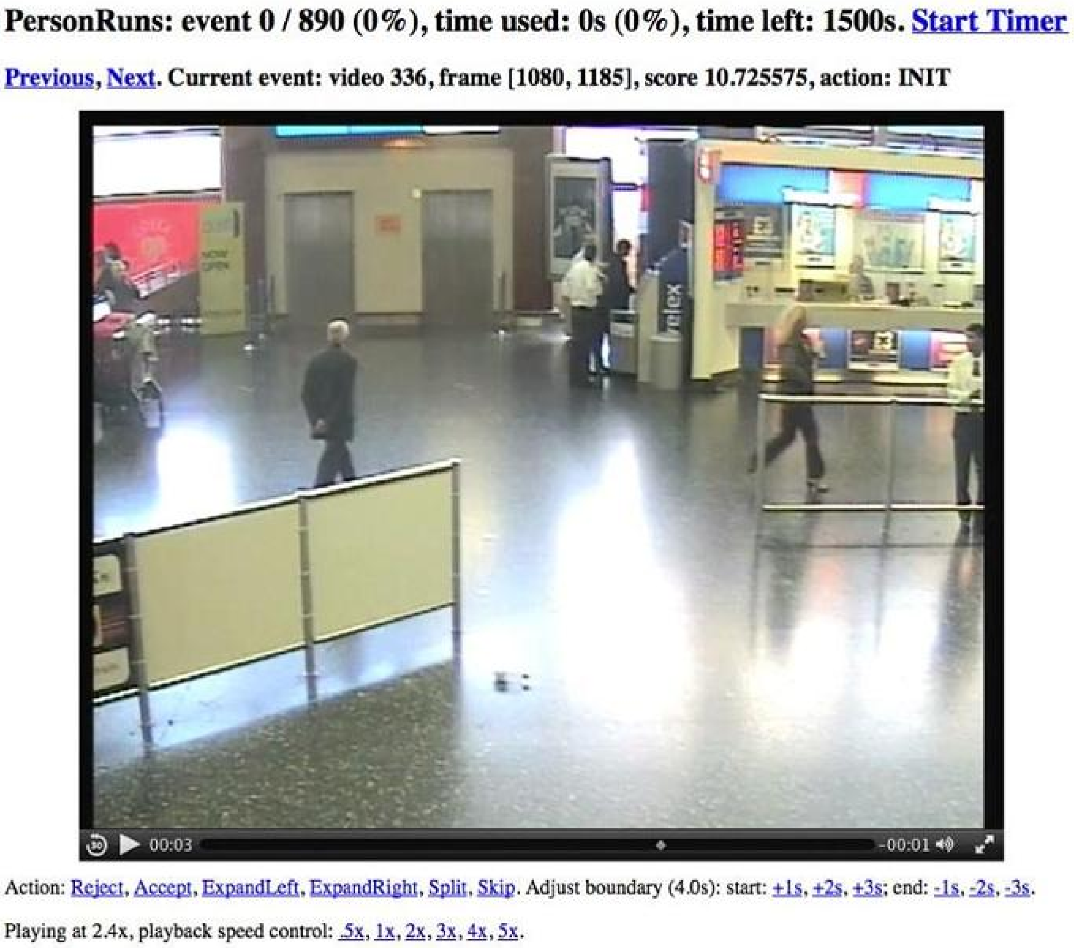
- System Overview
- Low-Level Features
- Video Representation
- CascadeSVMs
- Human Interactions
- Performance Evaluation
- Conclusion

System Overview


Online Detection
Testing Surveillance Video
Sliding Window
Human Interaction
Low-Level Feature Extraction
Fisher Vector Representation
Classification
Detections
Low-Level Feature Extraction
Fisher Vector Representation
CascadeSVMs Learning
Classifiers
Sliding Window
Training Surveillance Video 1
Training Surveillance Video N
Offline Learning

Human Interactions

- High Throughput UI



PersonRuns: event 0 / 890 (0%), time used: 0s (0%), time left: 1500s. Start Timer

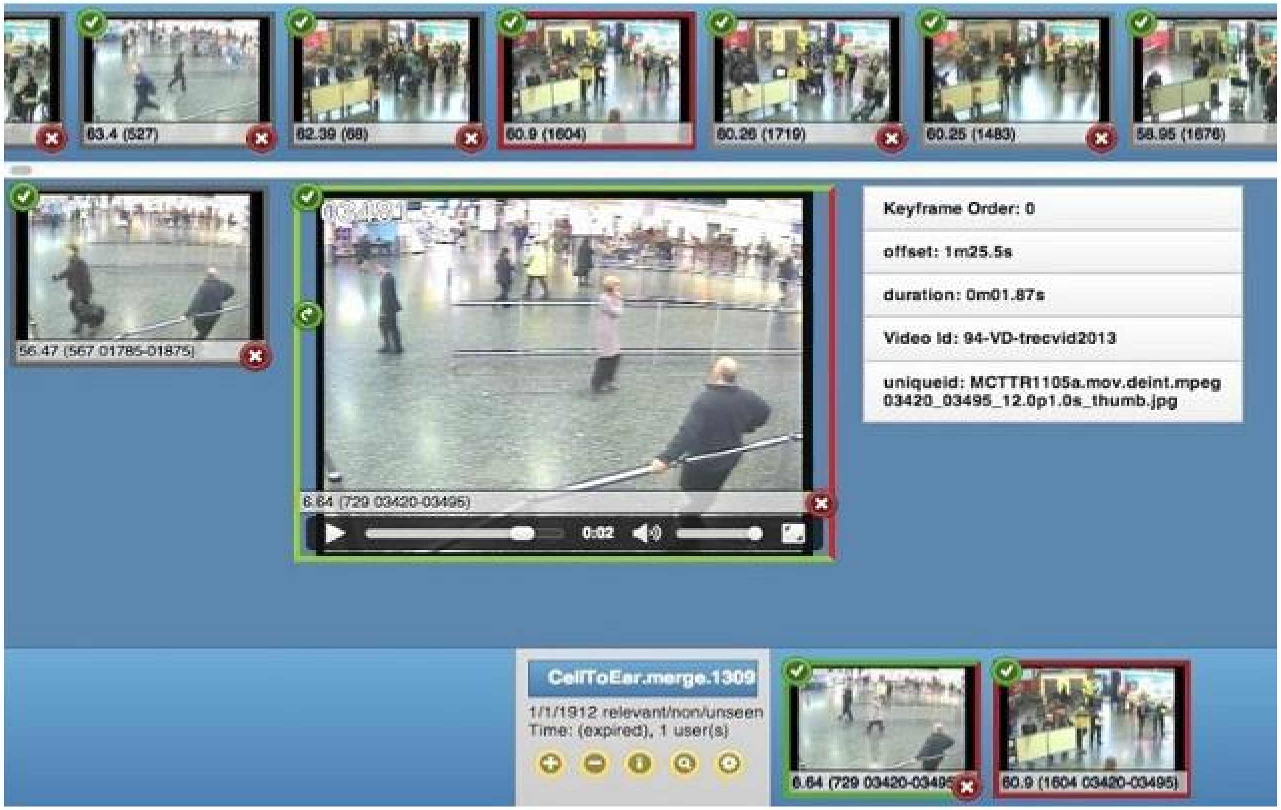
Previous, Next. Current event: video 336, frame [1080, 1185], score 10.725575, action: INIT

Action: Reject, Accept, ExpandLeft, ExpandRight, Split, Skip. Adjust boundary (4.0s): start: +1s, +2s, +3s; end: -1s, -2s, -3s.

Playing at 2.4x, playback speed control: .5x, 1x, 2x, 3x, 4x, 5x.

Human Interactions

- Triage UI



Keyframe Order: 0

offset: 1m25.5s

duration: 0m01.87s

Video Id: 94-VD-trecvid2013

uniqueid: MCTTR1105a.mov.deint.mpeg 03420_03495_12.0p1.0s_thumb.jpg

CellToEar.merge.1309

1/1/1912 relevant/non/unseen
Time: (expired), 1 user(s)

Outline

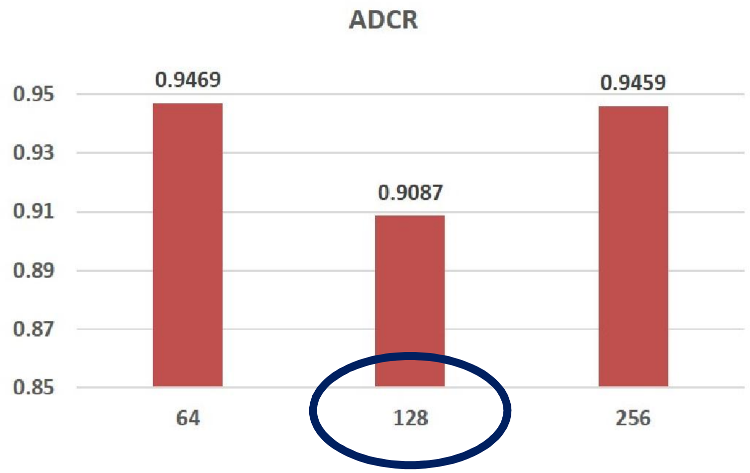
- System Overview
- Low-Level Features
- Video Representation
- CascadeSVMs
- Human Interactions
- Performance Evaluation
- Conclusion

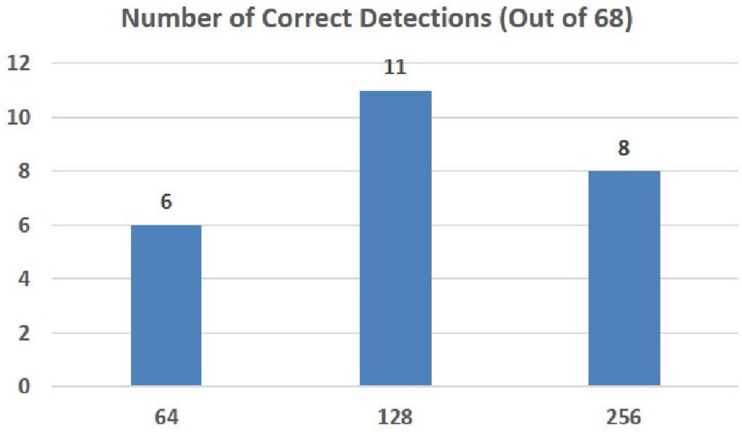
Performance Evaluation

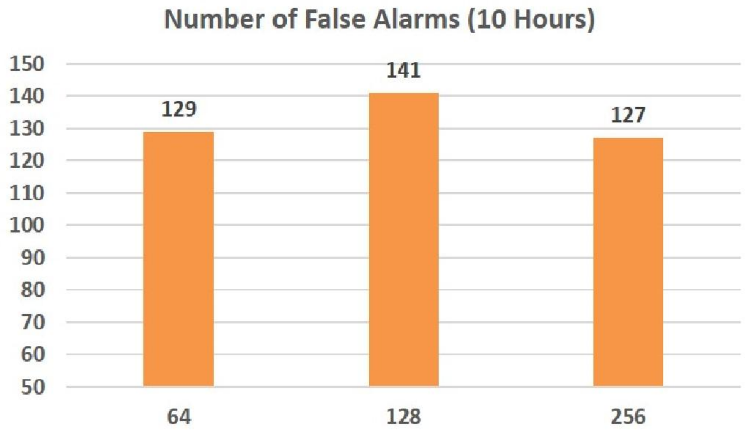
- Experimental Setup
  - PersonRuns
  - Fisher Vector
  - CascadeSVMs
  - 40-hour videos for training
  - 10-hour videos for testing

Performance Evaluation

- Number of Gaussian Components
  - STIP

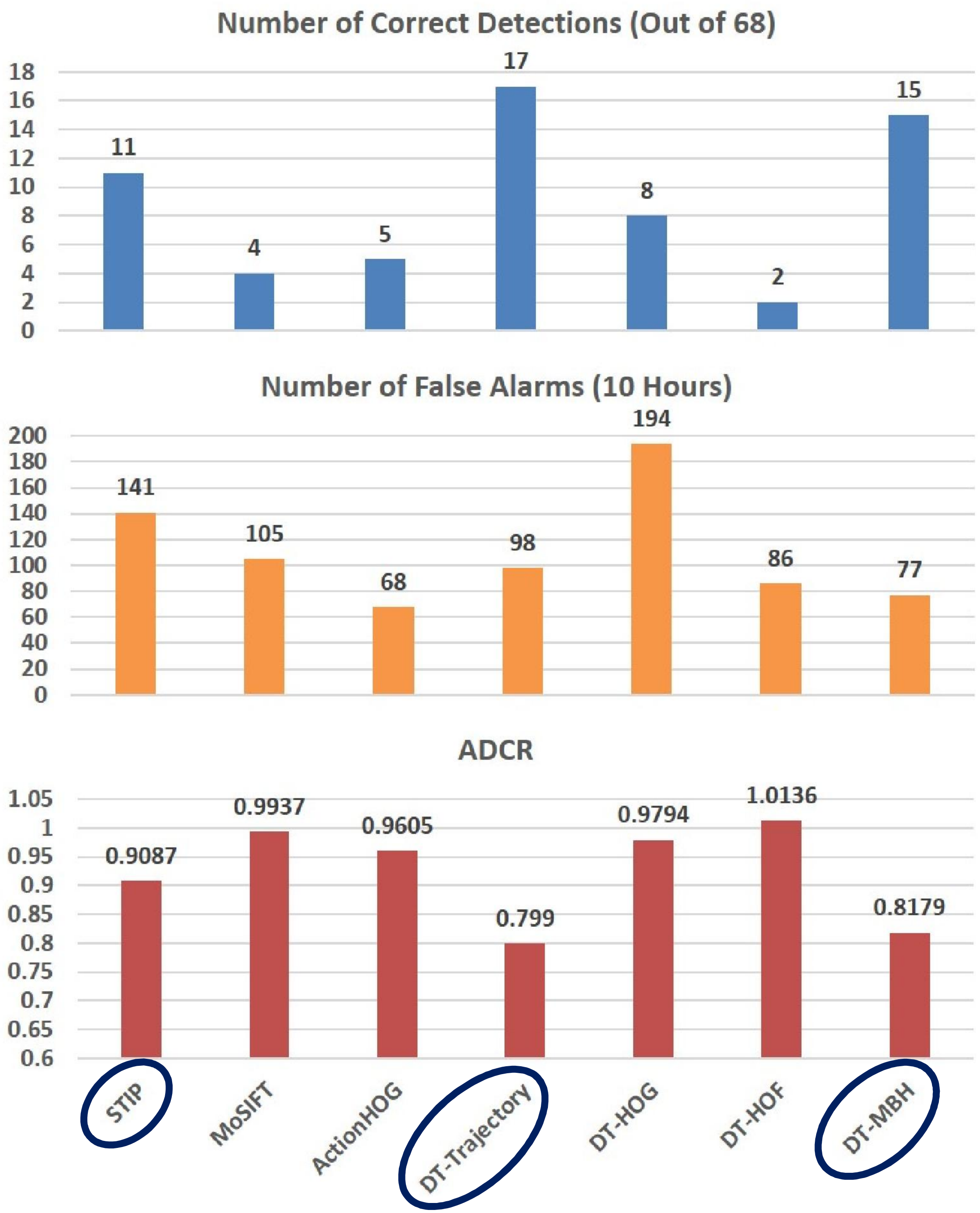






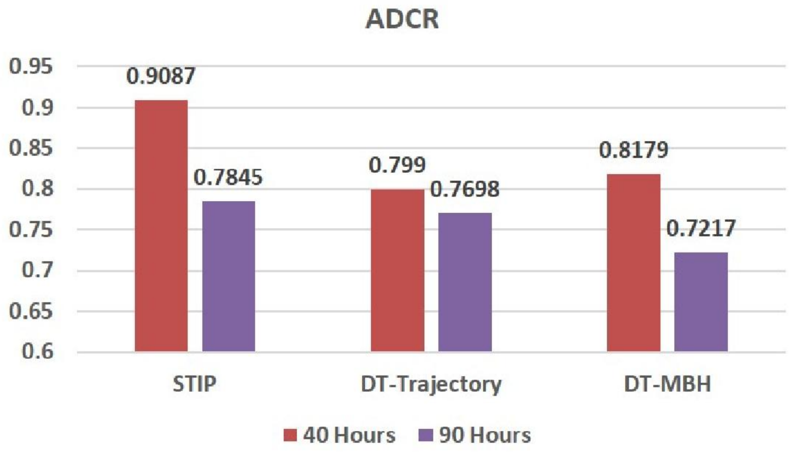
Performance Evaluation

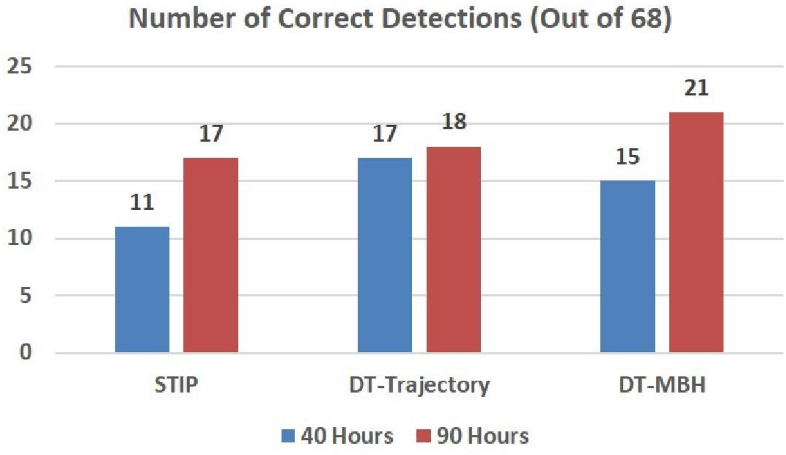
- Comparisons of Low-Level Features
  - STIP
  - MoSIFT
  - ActionHOG
  - DT-Trajectory
  - DT-HOG
  - DT-HOF
  - DT-MBH

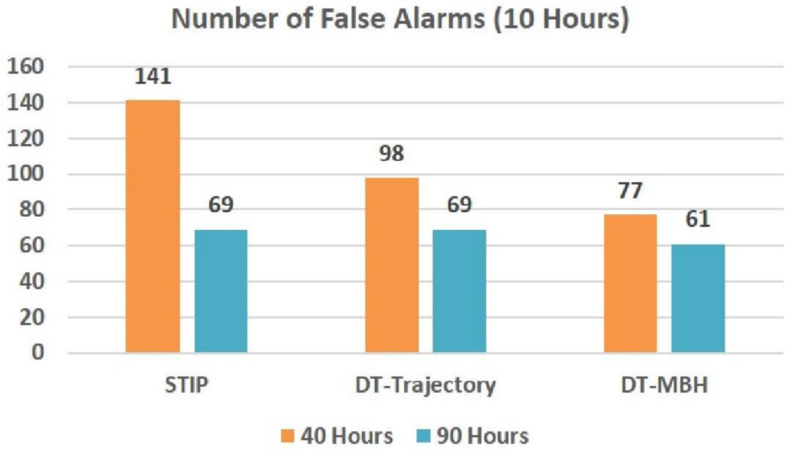

Number of Correct Detections (Out of 68)
18
16
14
12
10
8
6
4
2
0
11
4
5
17
8
2
15
Number of False Alarms (10 Hours)
200
180
160
140
120
100
80
60
40
20
0
141
105
68
98
194
86
77
ADCR
1.05
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.9087
0.9937
0.9605
0.799
0.9794
1.0136
0.8179
STIP
MoSIFT
ActionHOG
DT-Trajectory
DT-HOG
DT-HOF
DT-MBH

Performance Evaluation

- How A Larger Training Set Helps
  - 40 vs. 90 hours training videos

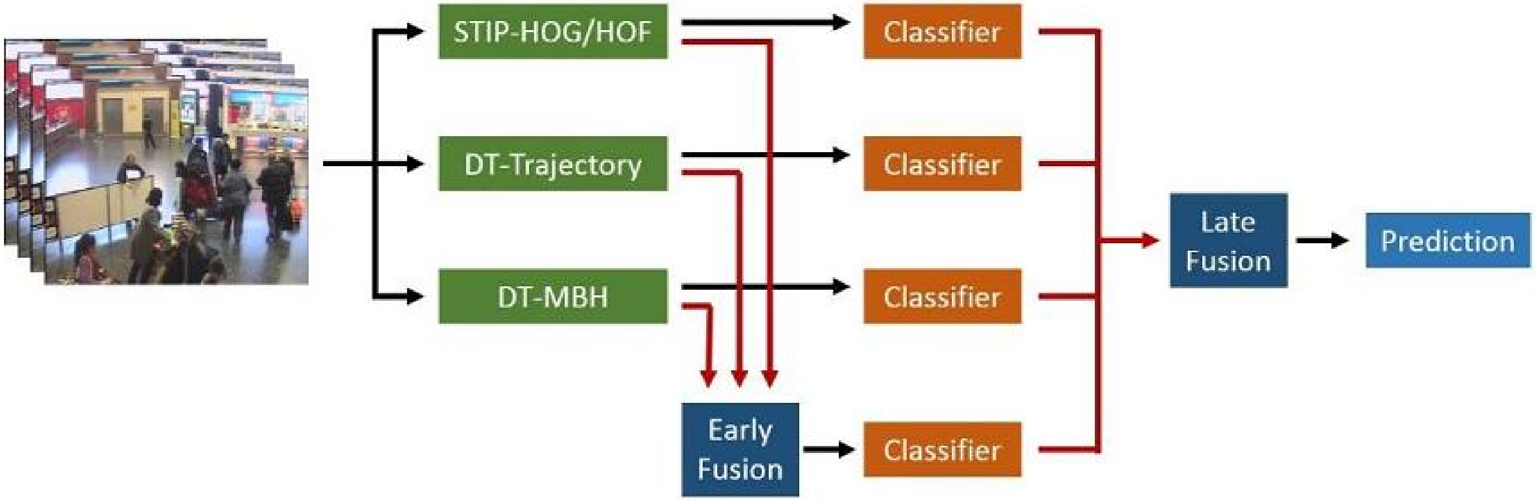


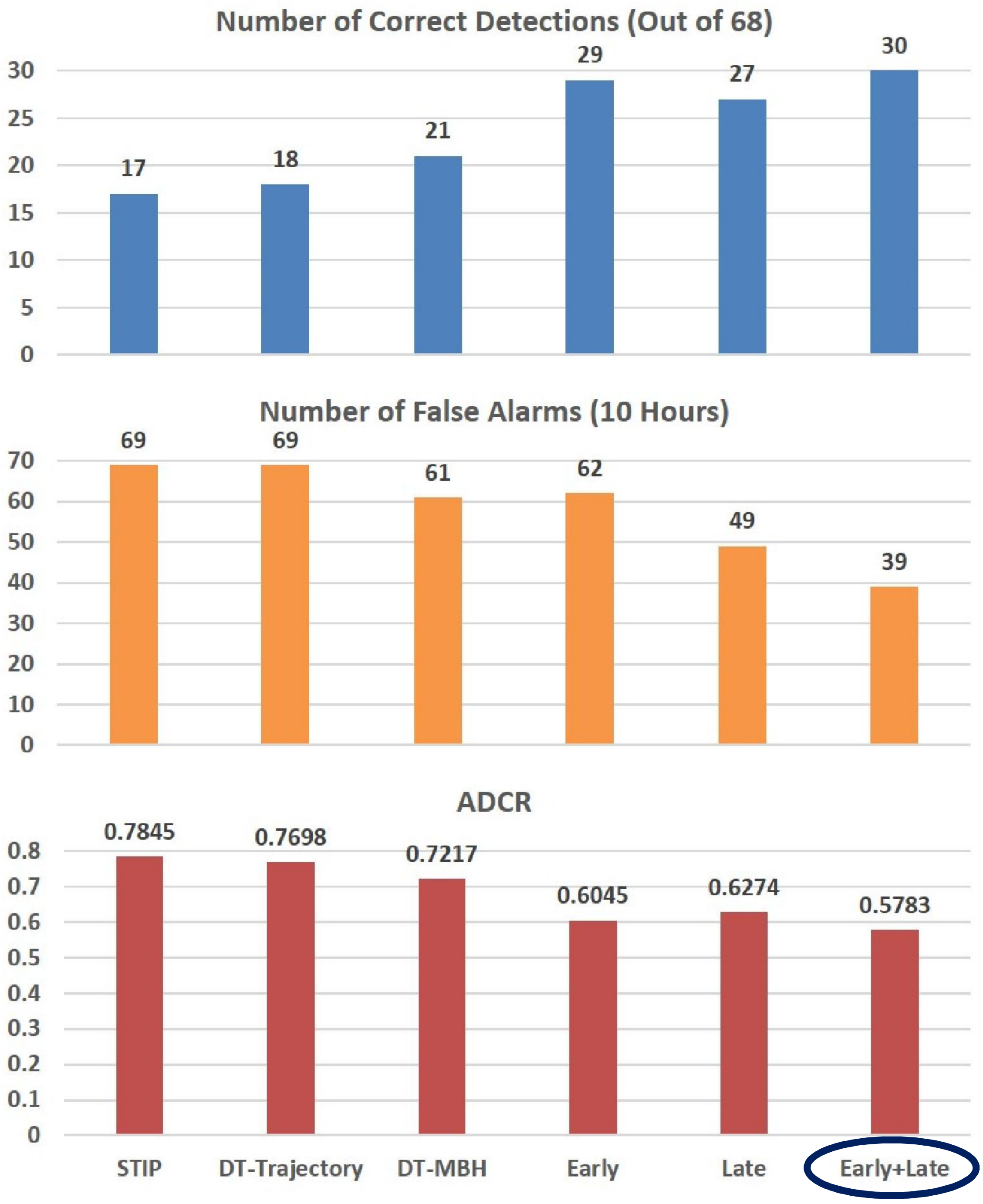




Performance Evaluation

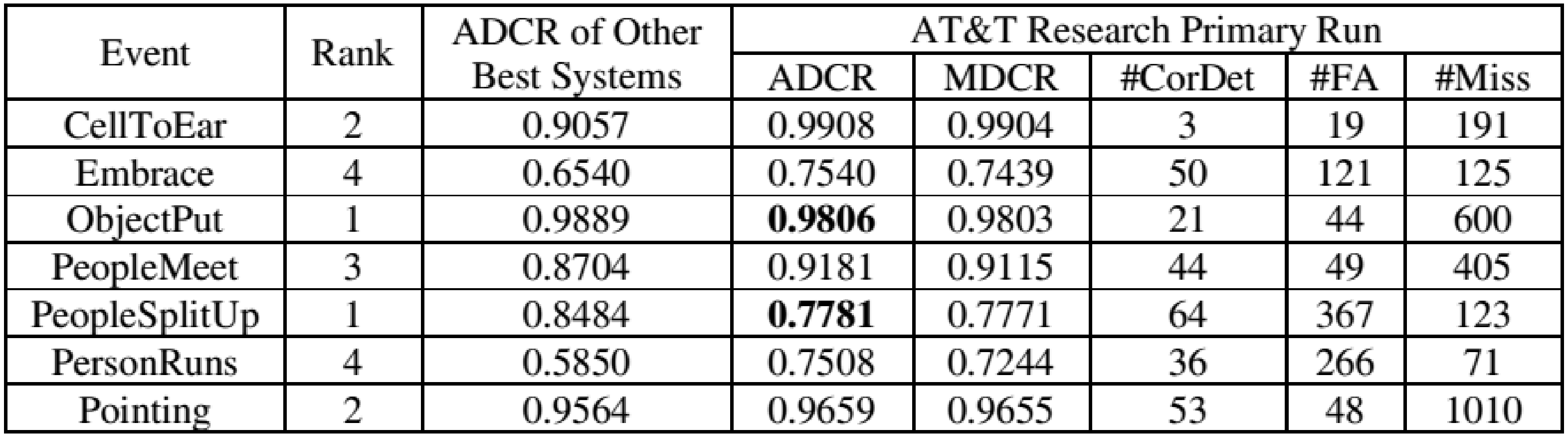
- Feature Fusion
  - 90 hours training videos
  - STIP, DT-Trajectory, DT-MBH
  - Early Fusion
  - Late Fusion
  - Early + Late Fusion




Number of Correct Detections (Out of 68)
30
25
20
15
10
5
0
17
18
21
29
27
30
Number of False Alarms (10 Hours)
70
60
50
40
30
20
10
0
69
69
61
62
49
39
ADCR
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.7845
0.7698
0.7217
0.6045
0.6274
0.5783
STIP
DT-Trajectory
DT-MBH
Early
Late
Early+Late

Performance Evaluation

- Formal Evaluation
  - Comparative Results



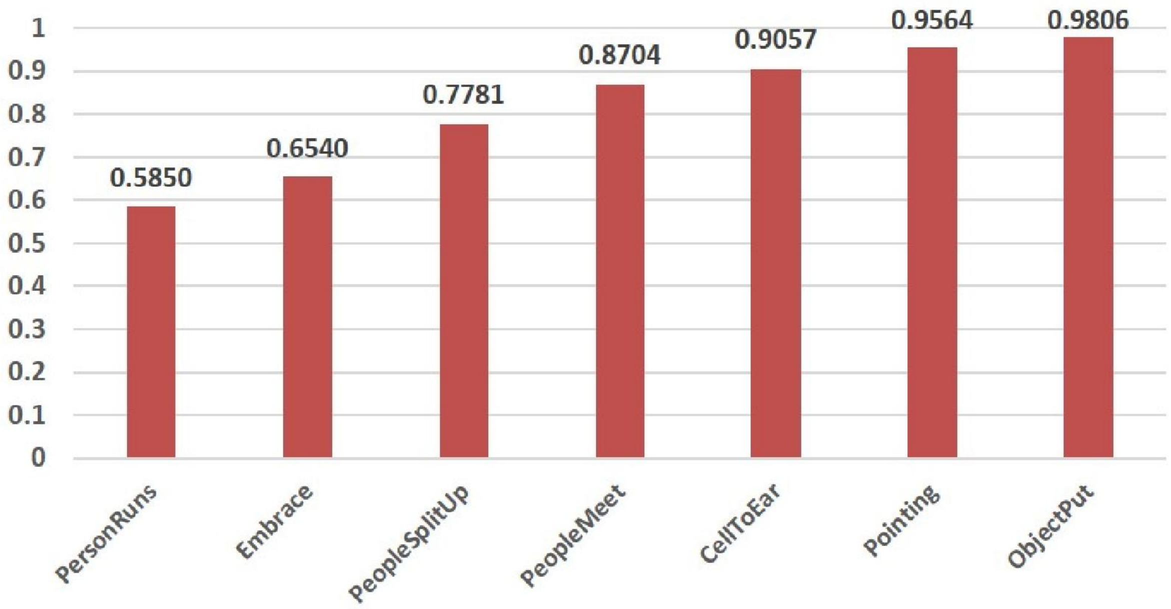
Event	Rank	ADCR of Other Best Systems	AT&T Research Primary Run				
			ADCR	MDCR	#CorDet	#FA	#Miss
CellToEar	2	0.9057	0.9908	0.9904	3	19	191
Embrace	4	0.6540	0.7540	0.7439	50	121	125
ObjectPut	1	0.9889	**0.9806**	0.9803	21	44	600
PeopleMeet	3	0.8704	0.9181	0.9115	44	49	405
PeopleSplitUp	1	0.8484	**0.7781**	0.7771	64	367	123
PersonRuns	4	0.5850	0.7508	0.7244	36	266	71
Pointing	2	0.9564	0.9659	0.9655	53	48	1010

Outline

- System Overview
- Low-Level Features
- Video Representation
- CascadeSVMs
- Human Interactions
- Performance Evaluation
- Conclusion

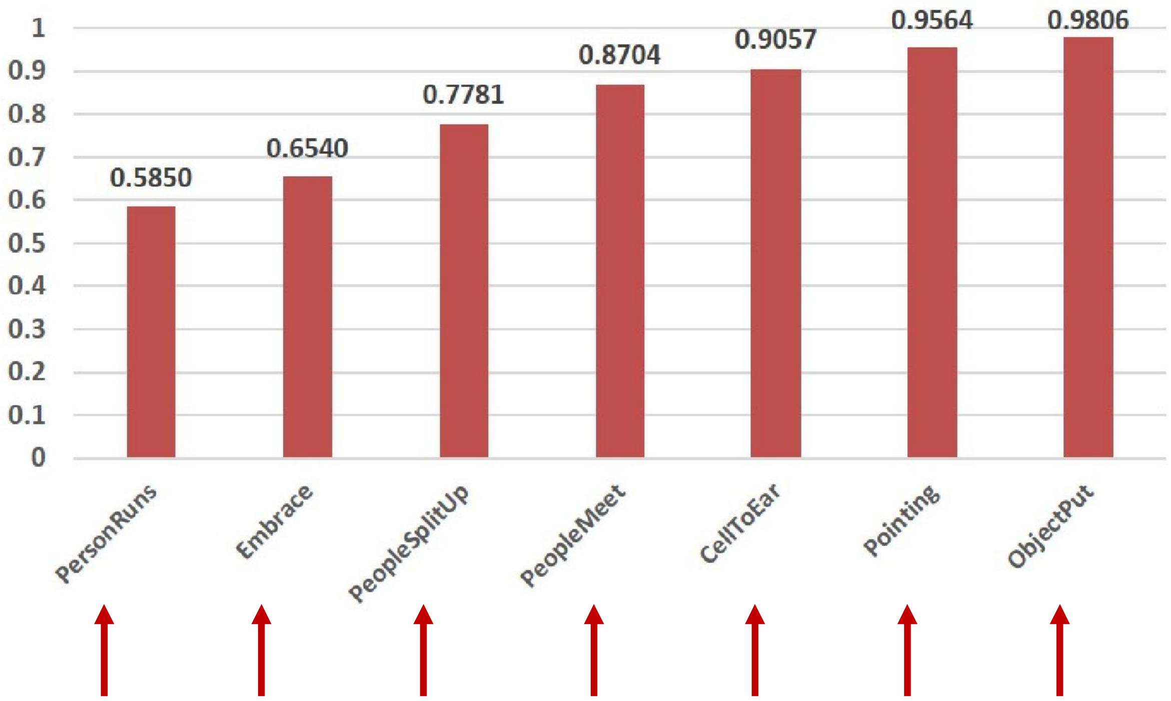
Conclusion

- Best ADCR



Conclusion

- Best ADCR



Single Person

Multiple People

Multiple People

Multiple People

Person Object

Single Person

Person Object

Conclusion

- Multiple Features
  - fusion scheme
  - ranking and selection
  - event-specific investigation
- Fisher Vector
  - accuracy and computation
- Human Interaction
  - collaborative mode
  - cross-event mode
  - static gesture detection

Thank You!