数字视频编解码技术国家工程实验室
National Engineering Laboratory for Video Technology

PKU@TRECVID2009: Single-Actor and Pair-Activity Event Detection in Surveillance Video

General Coach: Wen Gao [a], Xihong Wu [b], Tiejun Huang [a]
Executive Coach: Yonghong Tian [a], Yaowei Wang [a] , Lei Qing [a]
Member: Zhipeng Hu [a*], Guangnan Ye [b*], Guochen Jia [a], Xibin Chen [b], Qiong Hu [c], Kaihua Jiang [b]

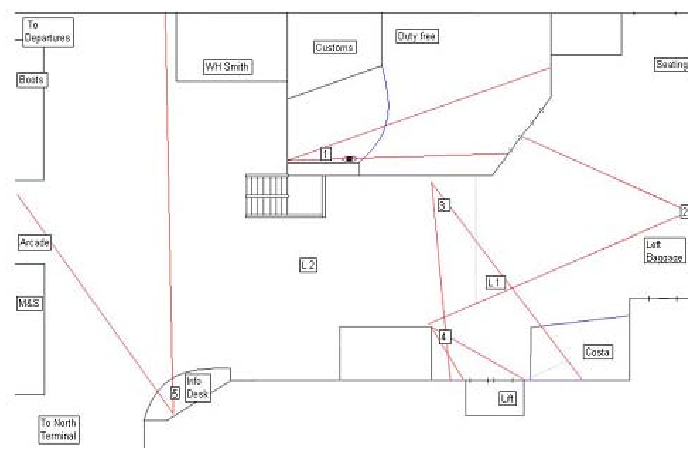
[a] National Engineering Laboratory for Video Technology, Peking University
[b] Speech and Hearing Research Center, Peking University
[c] Key Lab of Intel. Inf. Proc., Institute of Computing Technology, Chinese Academy of Sciences

Outline

- Overview
  - Introduction of TRECVID-ED Tasks
  - Summary of TRECVID-ED 2008
  - Our Results in TRECVID-ED 2009
- Our Solution in the eSur System
  - Background Modeling
  - Detection and Tracking
  - Event Classification
  - Post-processing
- Illustrative Results
- Summary

Overview of TRECVID-ED Tasks

- Task
  - To develop an automatic system to detect observable events in surveillance video
- Ten Events
  - PeopleMeet
  - PeopleSplitUp
  - Embrace
  - ElevatorNoEntry
  - PersonRun
  - CellToEar
  - ObjectPut
  - TakePicture
  - Pointing
  - OpposingFlow
- Challenges
  - Clutter scenes
  - Illumination variations
  - Occlusion
  - Different camera views
  - No clear event definition

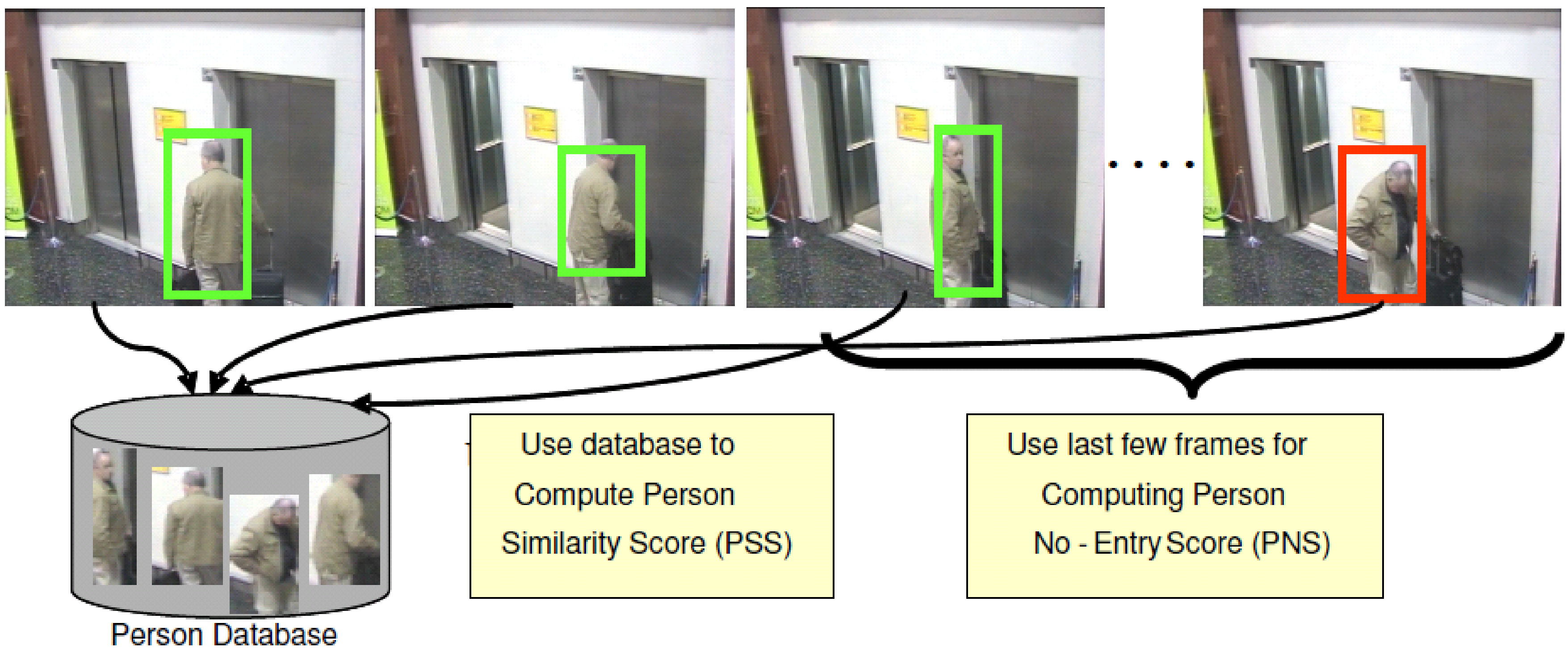


The Best Results of 2008

SITEID	Event	#Ref	#Sys	#CorDet	#FA	#Miss	Act.DCR
IFP-UIUC-NEC	CellToEar	349	15	1	14	348	0.999
Intuvision	ElevatorNoEntry	0	8	**0**	8	0	NA
DCU	Embrace	401	36193	91	5091	310	1.271
IFP-UIUC-NEC	ObjectPut	1944	83	6	77	1938	1.004
Intuvision	OpposingFlow	12	31	9	12	3	**0.251**
SJTU	PeopleMeet	1182	25033	270	5779	912	1.337
CMU	PeopleSplitUp	671	42415	185	42230	486	4.856
MCG-ICT-CAS	PersonRuns	314	662	23	639	291	0.989
SJTU	Pointing	2316	1005	35	970	2281	1.080
Intuvision	TakePicture	23	10	**0**	10	23	1.000

- Note:
  - There are much rooms for improvement.
  - OpposingFlow event has good detection performance.
  - ElevatorNoEntry and TakePicture events are zero CorDets.

Approaches in 2008



P. Yarlagadda, et. al, INTUVISION EVENT DETECTION SYSTEM FOR TRECVID 2008

- Elevator
- X. Yang, e... extraction
  - Haar c
  - Histog
- ...

University at TRECVID 2008

J.B. Guo et. al, TRECVID 2008 Event Detection By MCG-ICT-CAS

Classifier

+ histogram matching

2008

Our Results in TRECVID-ED2009 (1)

p-eSur_1

Event	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
PeopleMeet	449	125	7	118	442	1.023
PeopleSplitUp	187	198	7	191	180	1.025
Embrace	175	80	1	79	174	1.020
ElevatorNoEntry	3	4	2	2	1	0.334

p-eSur_2

Event	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
PeopleMeet	449	210	15	195	434	1.030
PeopleSplitUp	187	881	14	867	173	1.209
Embrace	175	164	3	161	172	1.036
PersonRuns	107	356	5	351	102	1.068

p-eSur_3

Event	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
PeopleMeet	449	210	15	195	434	1.030
PeopleSplitUp	187	881	14	867	173	1.209
Embrace	175	164	3	161	172	1.036
ElevatorNoEntry	3	0	0	0	3	1.000

Our Results in TRECVID-ED2009 (2)

- Compared with the best results in TRECVID-ED 2008
  - Directly on the reported results in terms of Act. DCR

Event	Our Best	Best 2008	Imp.
PeopleMeet	1.023	1.337	-0.314
PeopleSplitUp	1.025	4.856	-3.831
Embrace	1.020	1.271	-0.251
ElevatorNoEntry	0.334	N/A	-
PersonRuns	1.068	0.989	+0.079

Note: Our results are evaluated on the ED 2009 data by 2009 DCR metric, while the 2008 best results are evaluated on the ED 2008 data by 2008 DCR metric.

  - On the TRECVID-ED 2008 data in terms of 2008 Act. DCR

Event	Our Best	Best 2008	Imp.
PeopleMeet	1.245	1.337	-0.092
PeopleSplitUp	1.976	4.856	-2.880
Embrace	1.208	1.271	-0.063
ElevatorNoEntry	0.130	N/A	-
PersonRuns	1.249	0.989	+0.260

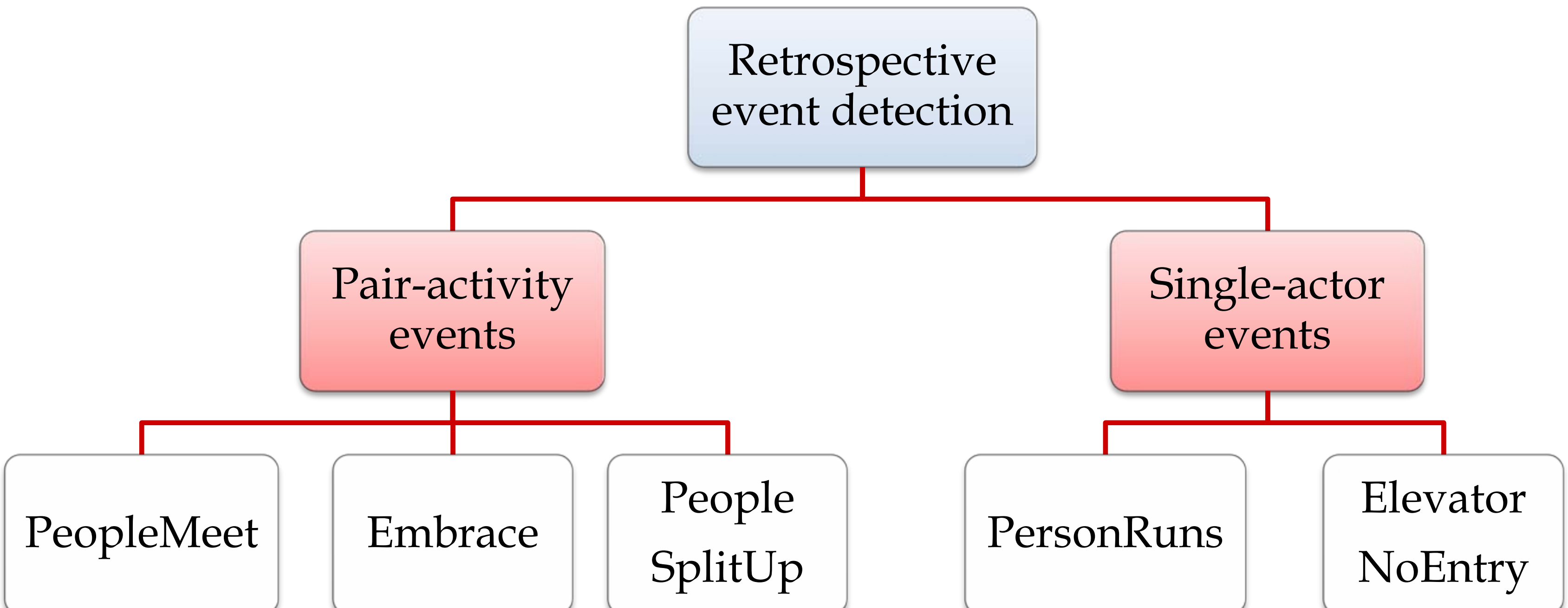
What are Improved?

- What?
  1. Effectively reduce the false alarms of detection
  2. Obtain comparable detection accuracy, and much better results for ElevatorNoEntry
- Why?
  1. Adaptive background modeling
  2. Effective human detection and tracking
  3. Ensemble of one-vs.-all SVM and automata-based classifiers
  4. Effective event merging and post-processing

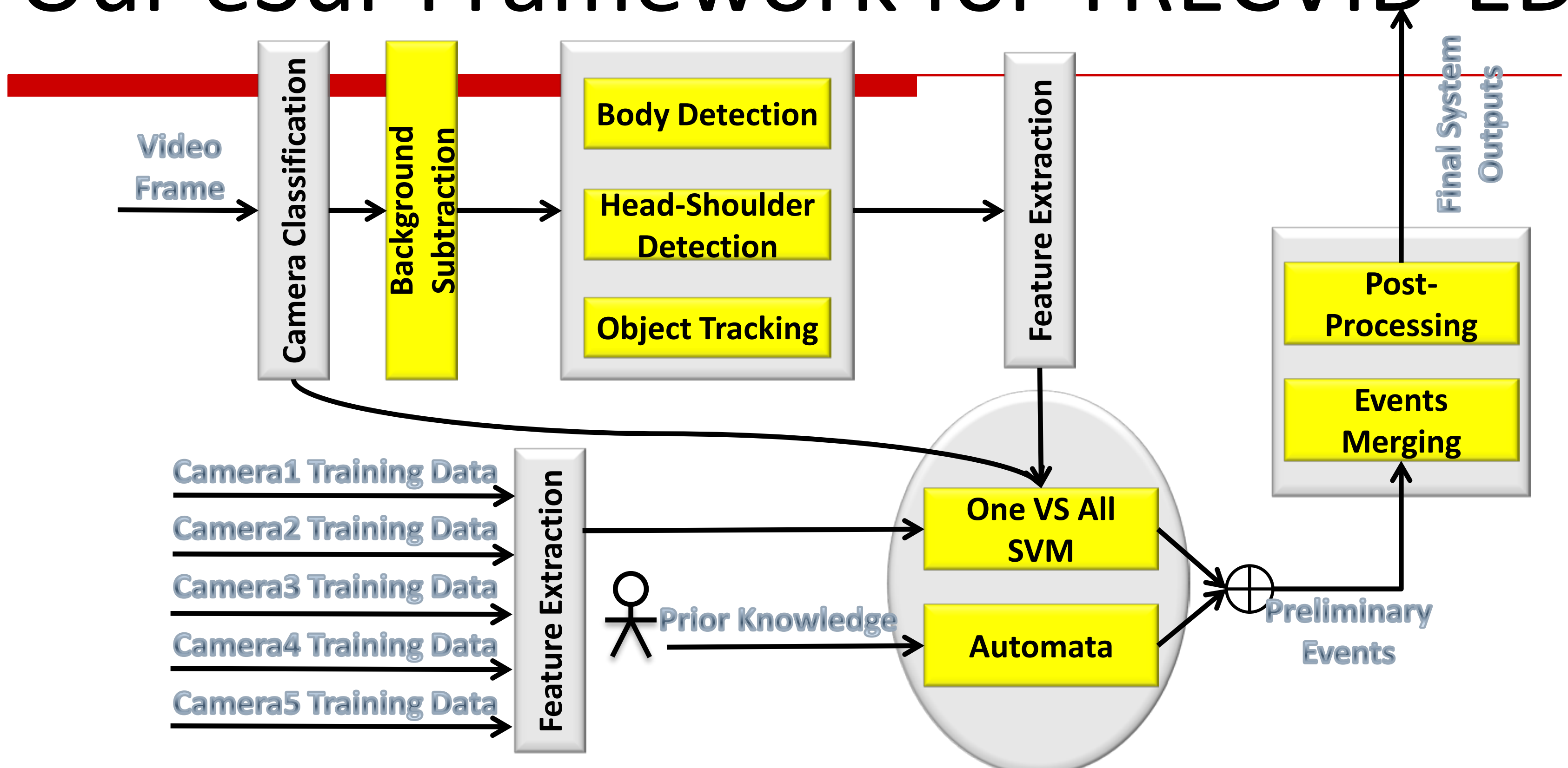
Our Solution:

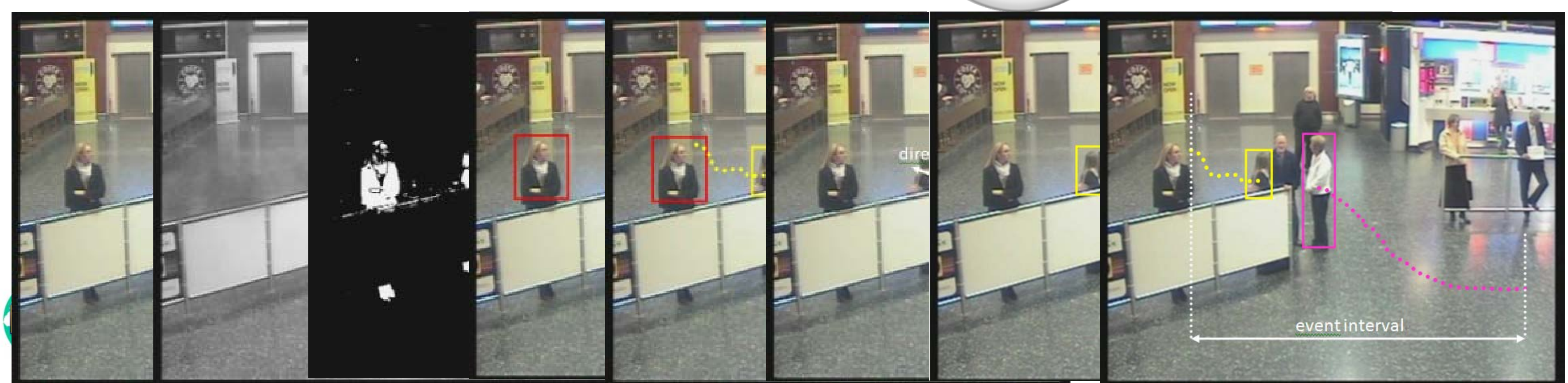
Treatments for Different Event Categories

- **Pair-activity Event:**
  - One people interact with another people
- **Single-actor Event:**
  - No interaction with other people



Our eSur Framework for TRECVID-ED





Our Solution (1):
Background Modeling

- **Mixture of Gaussian (MoG):**
  - To accurately extract the foreground while effectively decreasing detection false alarms.

- **Block-wise PCA Model:**
  - To identify which camera the video belongs to
    - Also used in the ElevatorNoEntry event detection.
  - "block" : segment each frame into blocks
  - "wise" : adaptively select the principle component for background reconstruction

MoG

- Key Idea
  - Randomly select 1000 frames from each camera
  - Manually label the foreground objects
  - Use EM algorithm to estimate the model
- Results of Background Reconstruction


Cam1 Background


Cam2 Background


Cam3 Background


Cam5 Background

- Disadvantage: Computation time-consuming

Block-wise PCA

- General PCA
  - Model a whole frame
  - Problems
    - high spatio-temporal computation complexity
    - high miss ratio (especially for static objects).
- Block-wise PCA
  - Segment a frame into blocks, and model each block respectively.
    - Lower spatio-temporal computation complexity
  - Adaptively select principle component by the MMSE to the mean background
    - Lower miss ratio and less block effect.

$$B = \arg\min_{B_i} \left\| \overline{I} - B_i \right\|^2 \qquad B_i = \phi_i \phi_i^T I$$

where $\overline{I}$ is the trained mean background, ϕ_i is the *ith* principle component and B_i is the ith reconstructed background

数字视频编解 National Enginee

Comparative Results

Blocking vs. No Blocking

Result with no blocking

Result with blocking

Method	No-blocking	Blocking
Training time (for 300 frames)	361.332s	150.406s

** Experiment platform : Intel Xeon E5410 2.33GHz , 8G*

Block PCA vs. Block-wise PCA

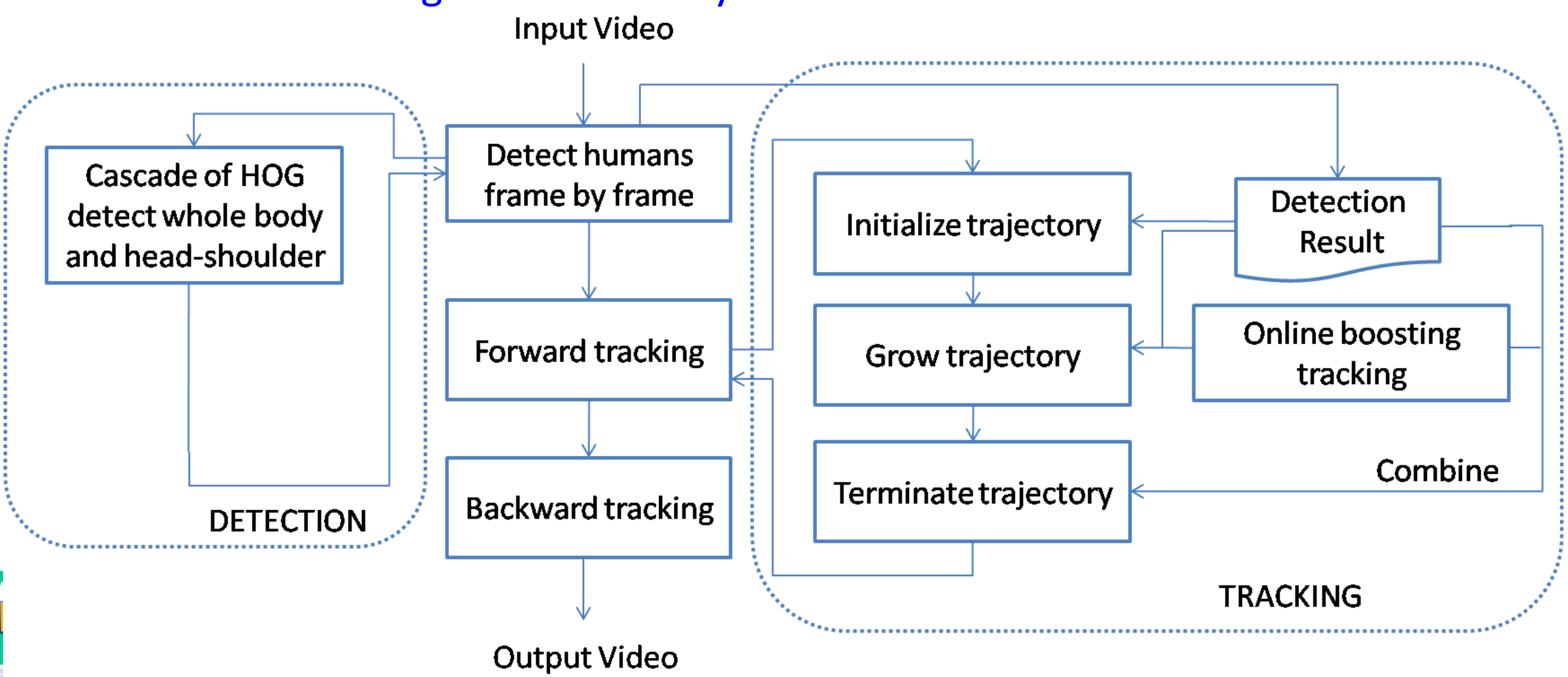
original image

Block PCA

Block-wise PCA

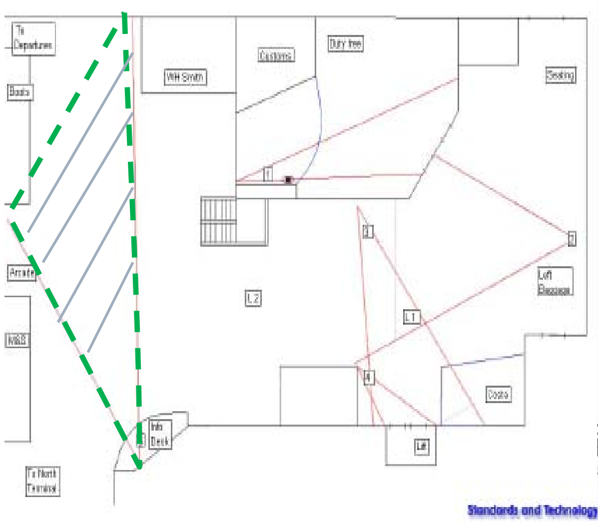
Our Solution (2): Detection and Tracking

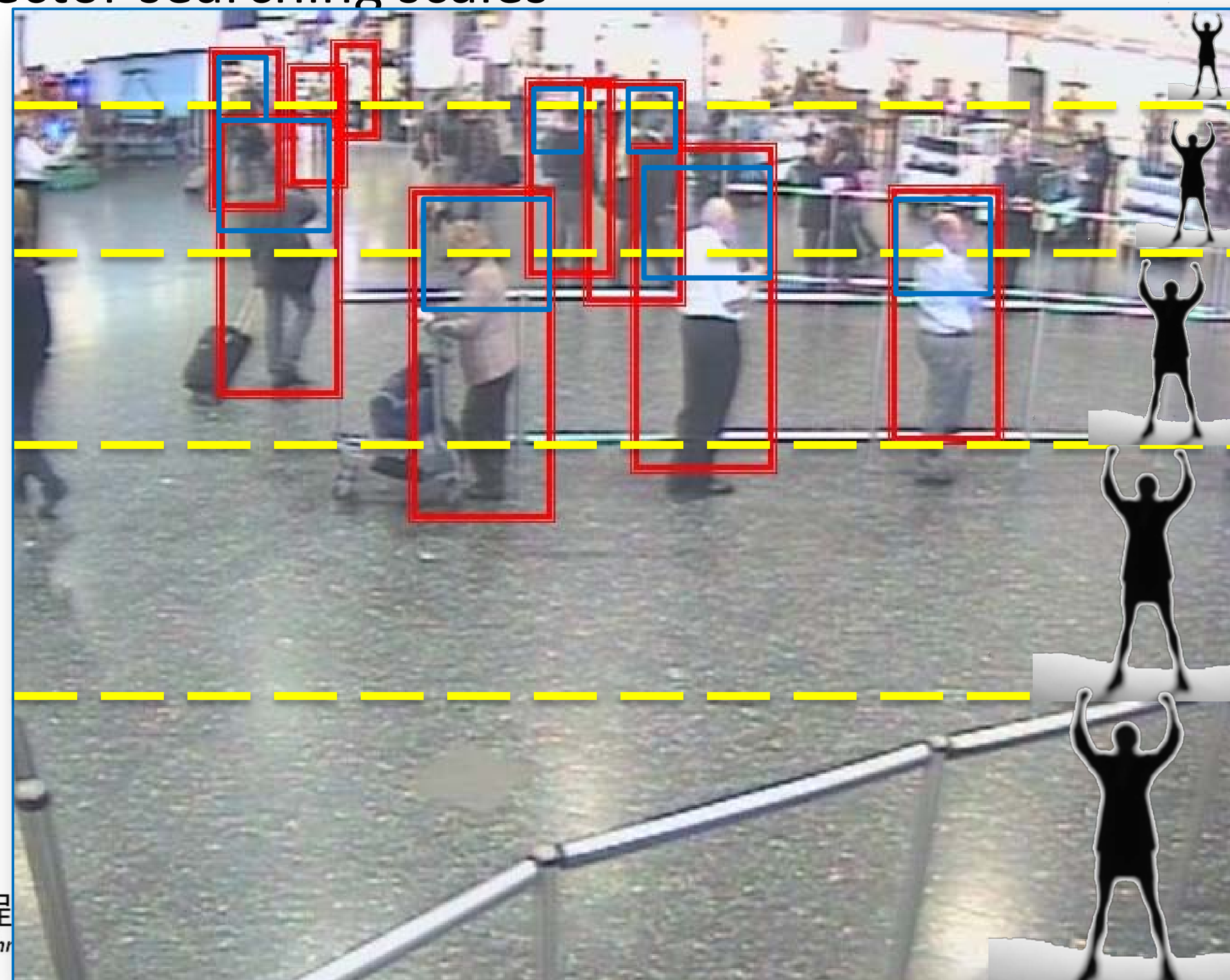
- Detection: Histogram of oriented gradients (HOG) for both whole body and head-shoulder
- Tracking: Online boosting
  - Forward and backward tracking
  - Combining color similarity to reduce drift



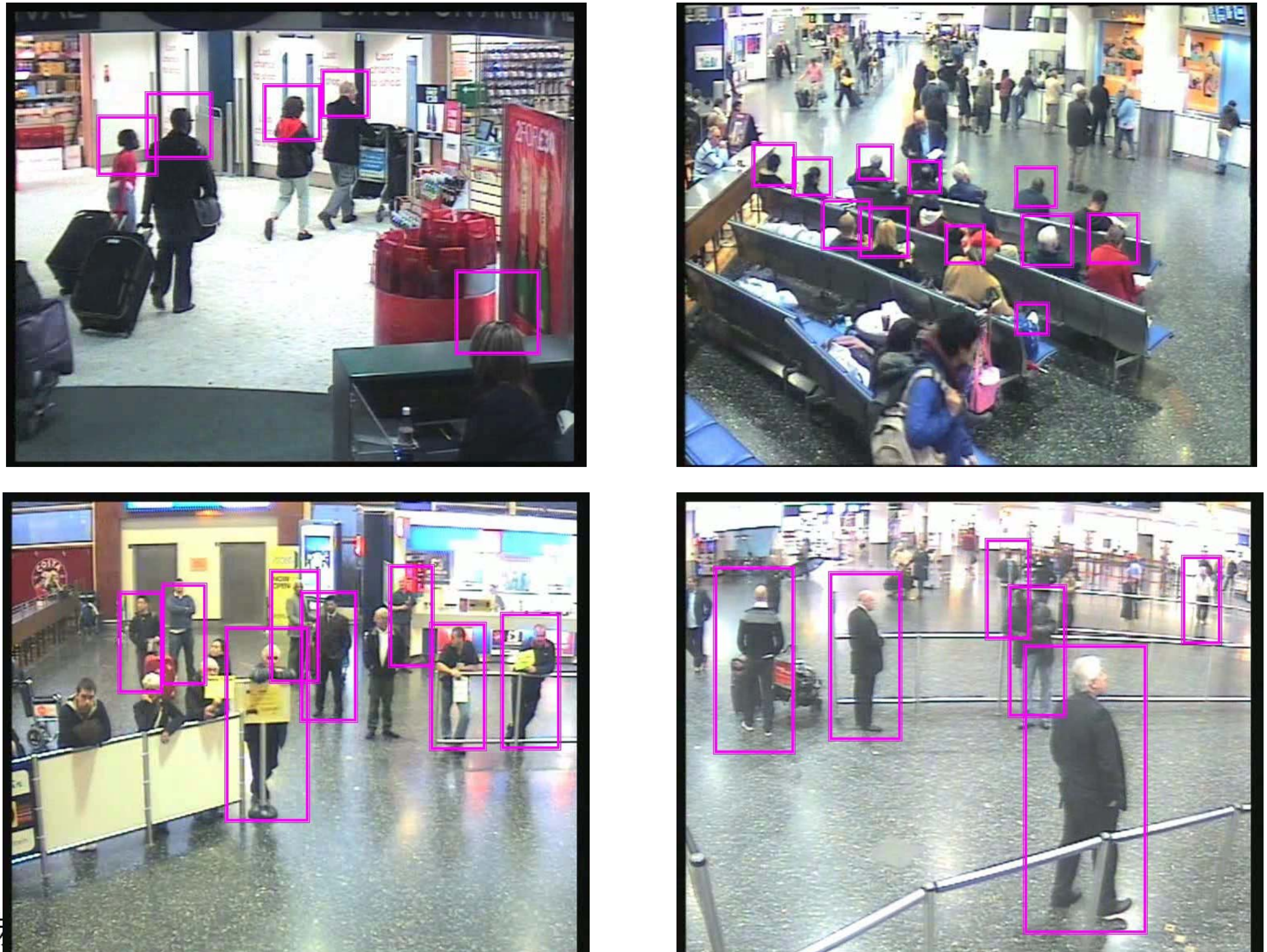
HOG Detector

- Fusion of Head-shoulder and Body detection
- Adjust the detector searching scales

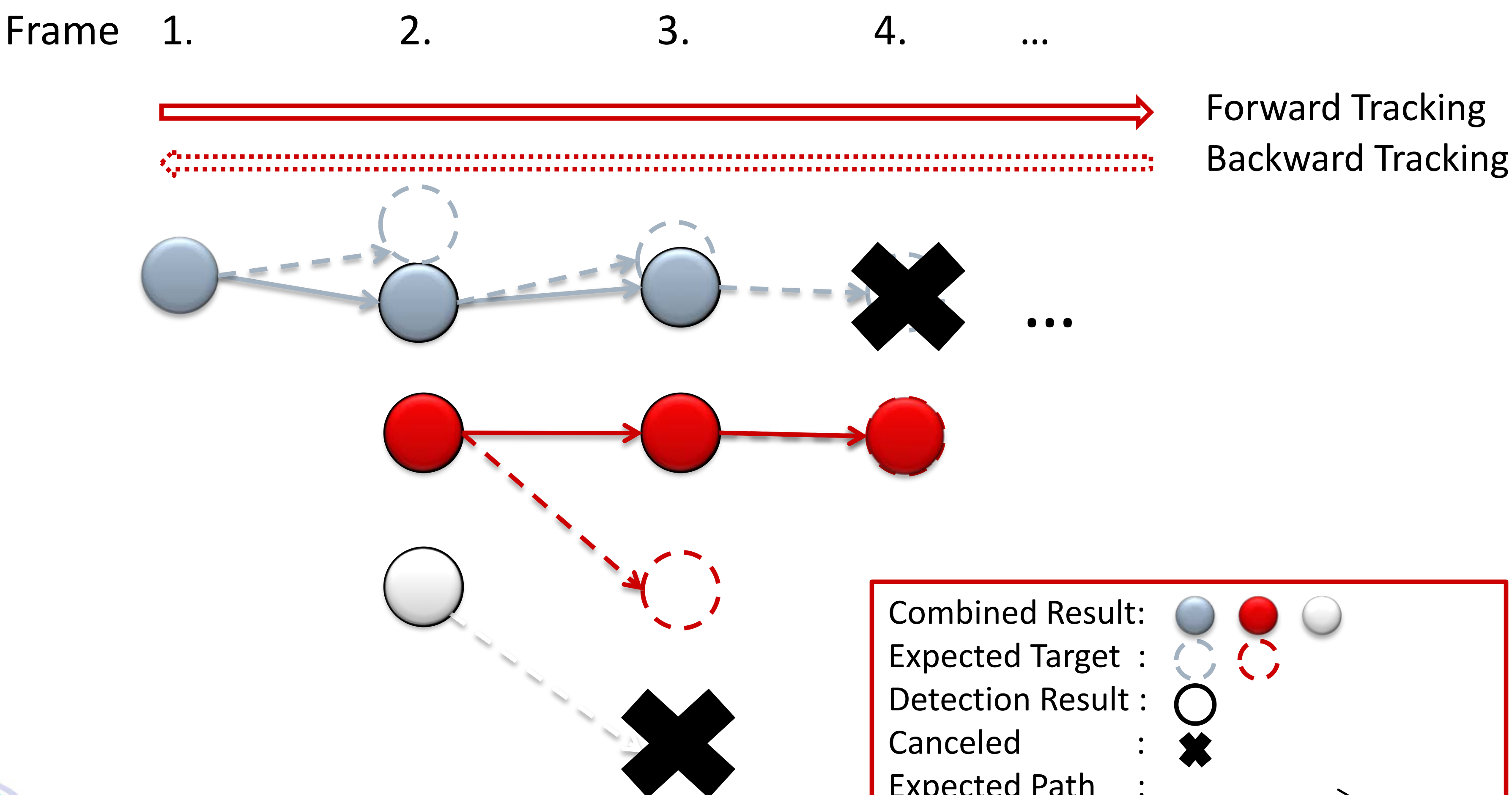




Detection Results

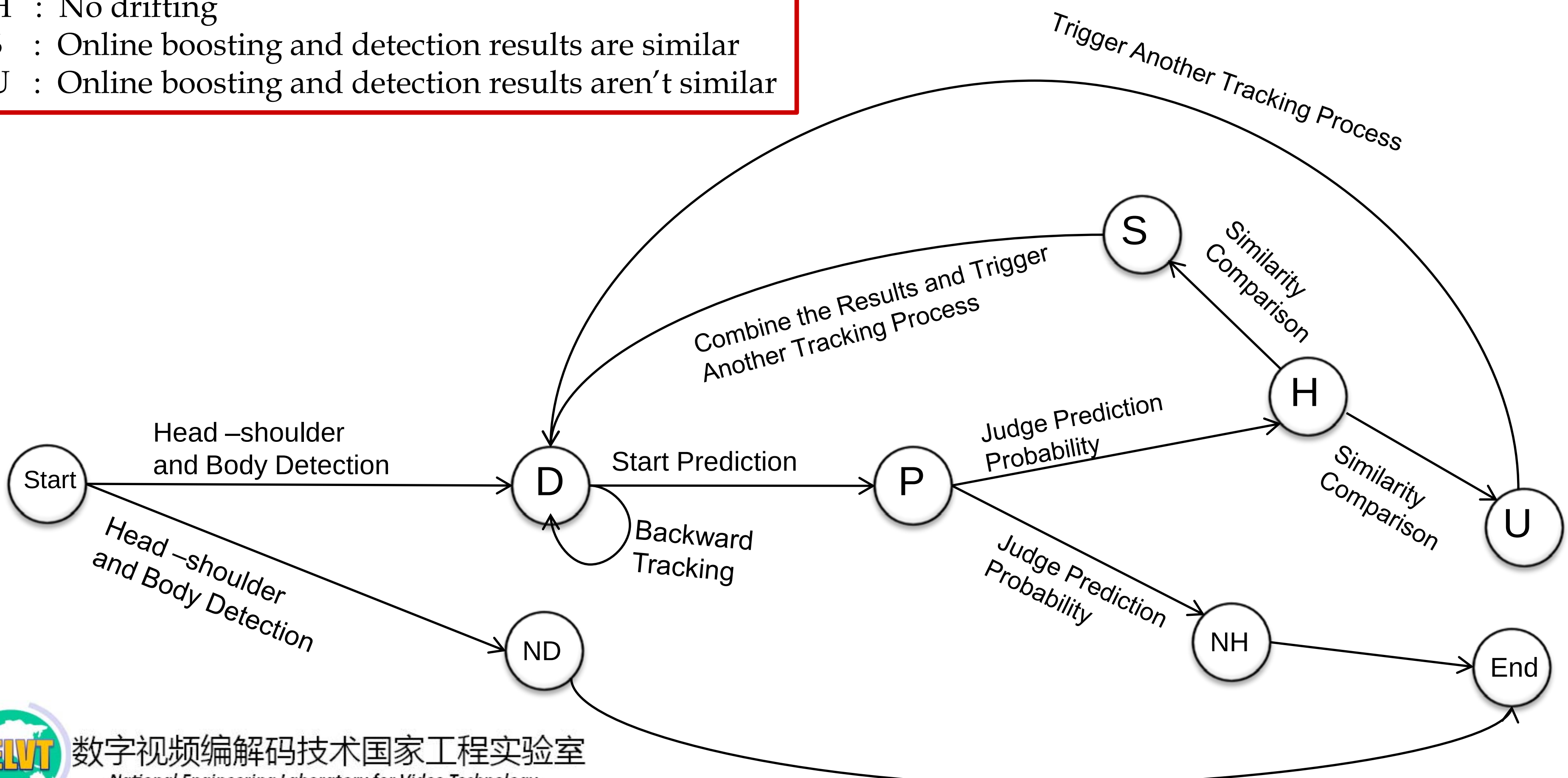


Tracking Process



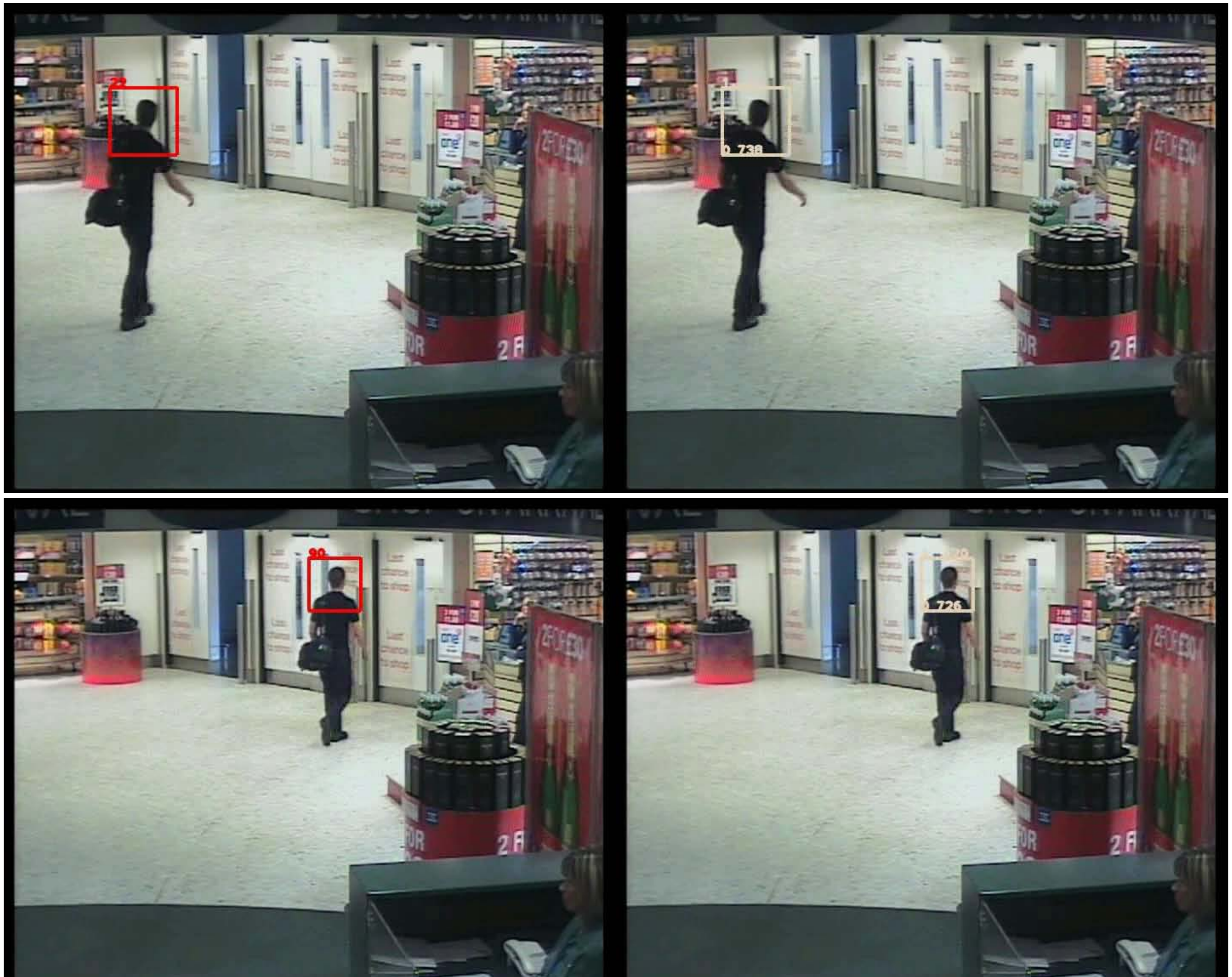
State Machine of Tracking

D : Detection existence
ND: No detection results
P : Online boosting prediction result
NH: Not human, drifting happens
H : No drifting
S : Online boosting and detection results are similar
U : Online boosting and detection results aren't similar



数字视频编解码技术国家工程实验室
National Engineering Laboratory for Video Technology

Detection and Tracking Results



Detection Results Tracking Results

Drift Reduction by Color Similarity

- Problem: Drifting
- Solution: Combine color similarity to refine tracking results

Tracking Result without Color Similarity Comparison

Tracking Result with Color Similarity Comparison

Our Solution (3):
Events Detection - Pair-activity

- Event Analysis using key frames
  - Key Frames: Frames characterize an event happening
  - “PeopleMeet” and “Embrace”
    - At the end of the event
  - “PeopleSplitUp”
    - At the beginning of the event

PeopleMeet

Embrace

PeopleSplitUp

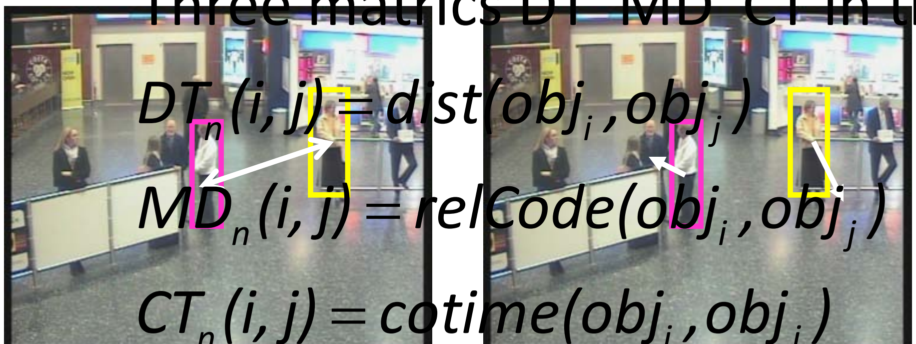
Events Detection - Pair-activity

- Relational Features

Distance

$$dist(obj_1,obj_2)=\left\|pos_{obj1}-pos_{obj2}\right\|$$

Three matrics DT, MD, CT in the *nth* frame:



$$DT_n(i,j)=dist(obj_i,obj_j)$$

$$MD_n(i,j)=relCode(obj_i,obj_j)$$

$$CT_n(i,j)=cotime(obj_i,obj_j)$$

Motion Direction Correlation

$$relCode(obj1,obj2)=\begin{cases}0 & if\ |\vartheta_1-\vartheta_2|<\alpha\\ 1 & if\ \pi-|\vartheta_1-\vartheta_2|<\alpha\\ 2 & otherwise\end{cases}$$

where ϑ_1 and ϑ_2 are the direction angles of the two persons

Feature vector between object$_i$ and object$_j$ in the *nth* frame :

$$F_n(i,j)=\{<DT_f(i,j),MD_f(i,j),CT_f(i,j)>\}_{f=n+1-k,\dots,n}$$

Co-occurrence Span

person$_1$'s occurrence time span

person$_2$'s occurrence time span

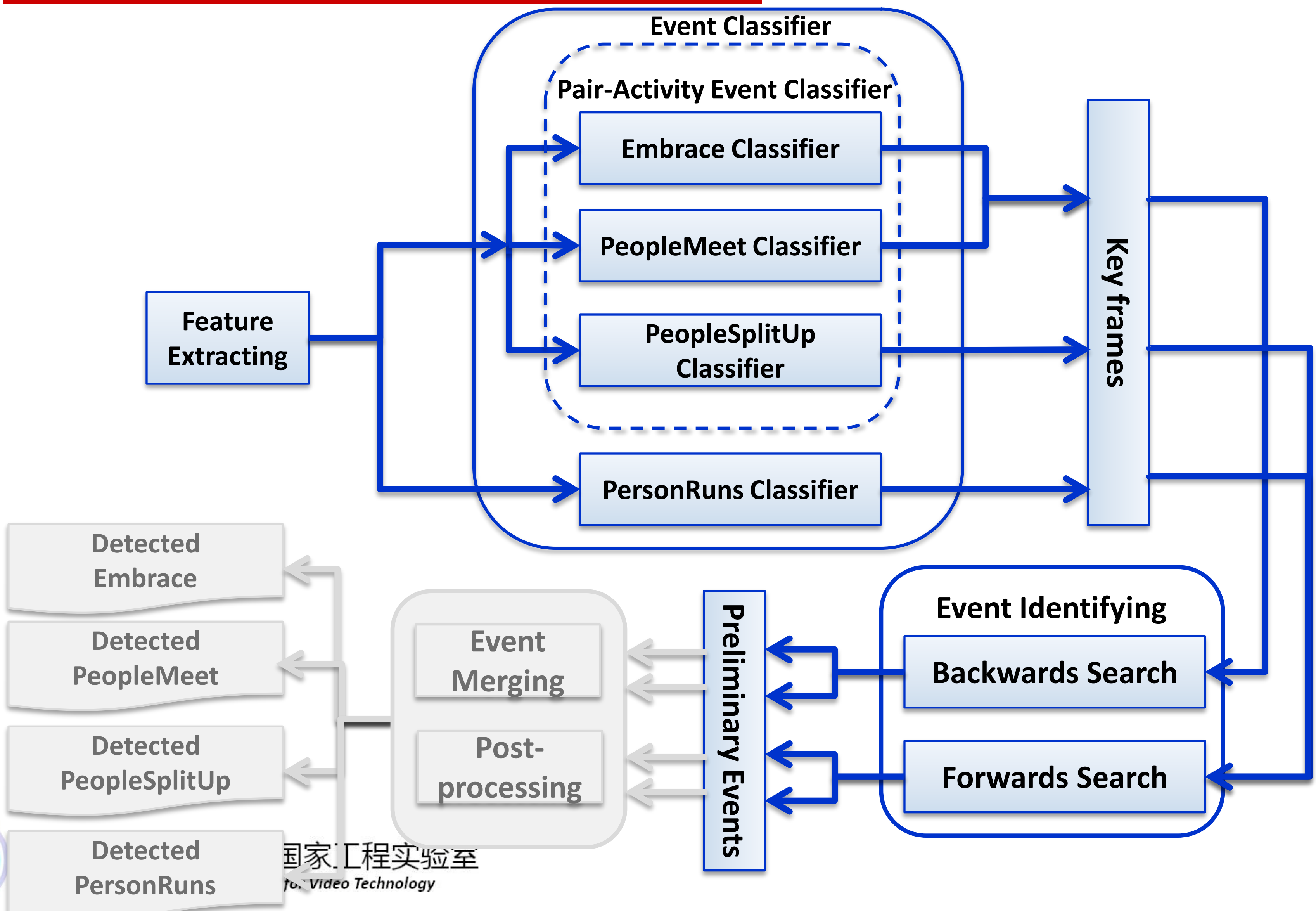
$$cotime(obj_1,obj_2)=$$
$$min(t_{end}(p1),t_{end}(p2))-max(t_{start}(p1),t_{start}(p2))$$

数字视频编解码技术国家工程实验室
National Engineering Laboratory for Video Technology

Events Detection - Single-Actor

- PersonRuns
  - Persons with higher velocity than others
  - Motion direction consistency
- ElevatorNoEntry
  - Elevator state detection
    - Keep close
    - Opening
    - Keep open
    - Closing
  - People state detection
    - PeopleExistence vs. No People
    - Enter, Leave and Waiting

Events Classification Framework



Classifiers Evaluation

- Single-level SVM classifier

Evaluated by NDCR

	Embrace	PeopleMeet	PeopleSplitUp	PersonRuns
MKL Classifier	1.088	1.324	1.117	1.068
Single-Level Classifier RBF Kernel	1.288	1.233	1.154	1.055

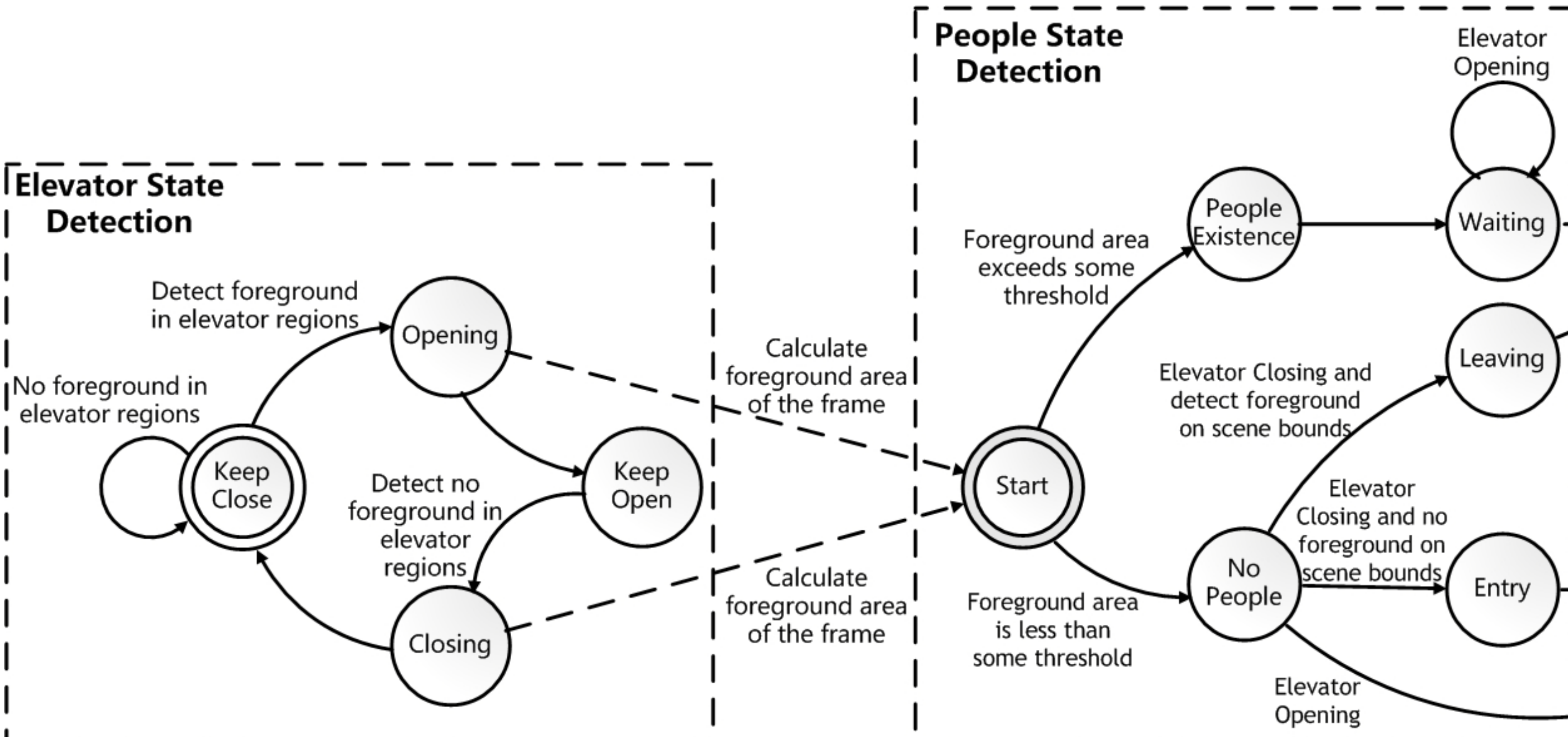
	Single-Level Classifier				Hierarchical	MKL
PersonRuns	0.813	0.708	0.728	0.815	0.700	0.590

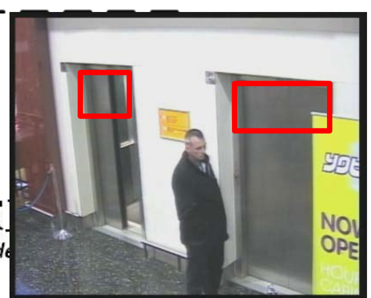
- Sample 10 hours data from TREVID-ED 2008 corpus
  - Multiple kernels (RBF, linear and poly kernels) are combined to enhance classifier performance
- Use detection and tracking results
- Sample 10 hours data from TREVID-ED 2008 corpus
- Single-level classifier is less time-consuming
- Manually label participating objects of each event

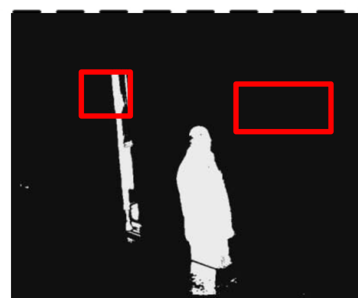
Classifiers	Single-level classifier	MKL classifier
1474071 vectors(for classification)	663.421s	15013.2s

State Machine of ElevatorNoEntry Detection

ElevatorNoEntry Event Detection







Results without Post-processing

- Data: 80 hours video from TRECVID-ED 2008

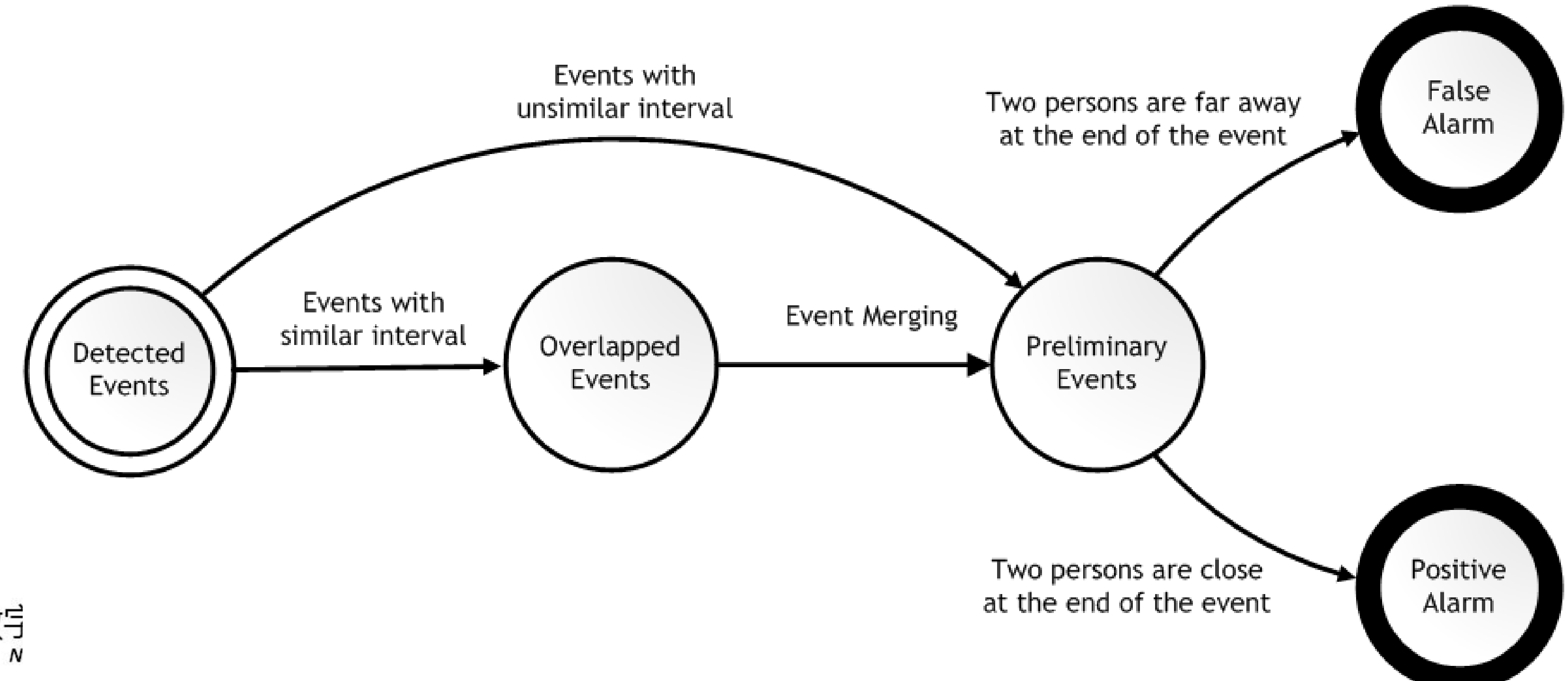
part of 2008 data	#Ref	#Sys	#Cor Det	#FA	#Miss	Act. DCR	Act. NDCR
PeopleMeet	796	1342	60	1282	736	433.949	1.245
People SplitUp	924	9505	176	9329	748	557.035	1.976
Embrace	279	1831	26	1805	253	422.891	1.208
PersonRuns	200	2731	18	2713	182	431.825	1.249

Too many false alarms!

Our Solution (4): Post-processing

- PeopleMeet and Embrace
  - Problem: False alarms shown as below
  - Solution:

    Final distance between the two persons must be less than some threshold.

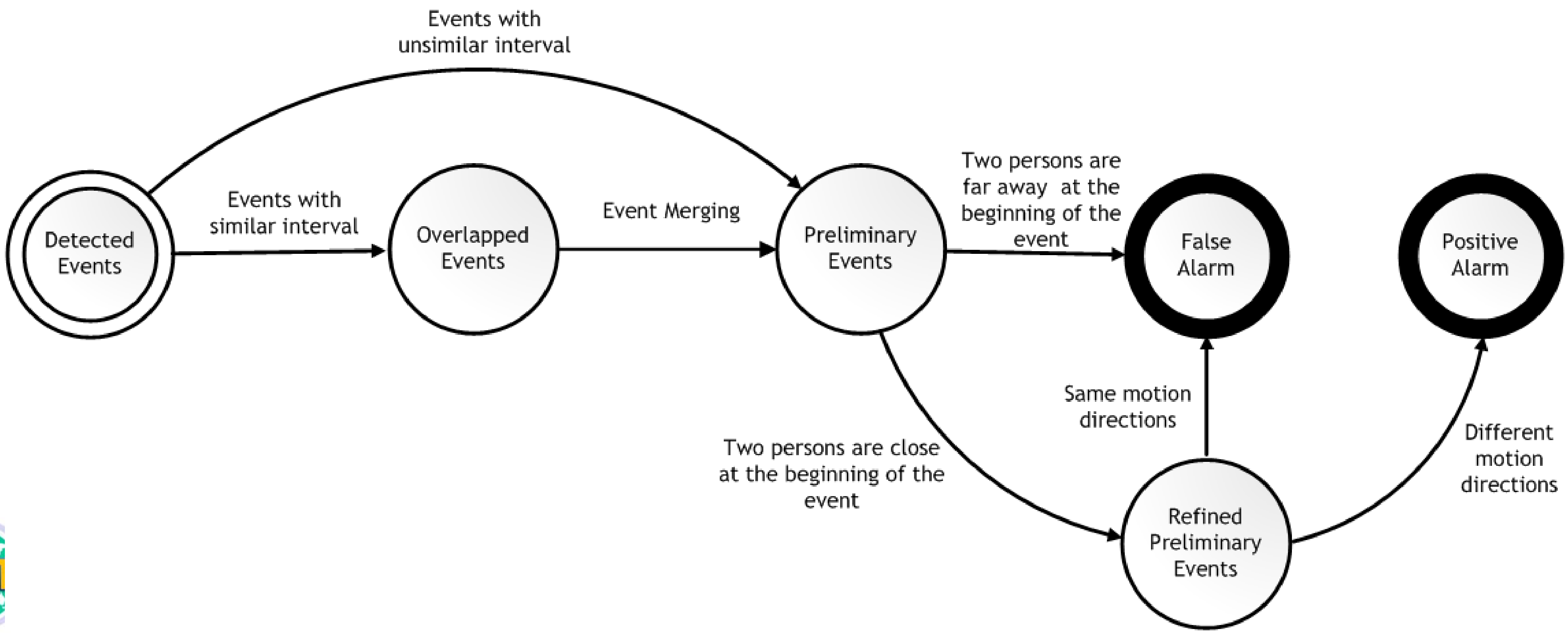


Post-processing

- PeopleSplitUp
  - Problem : False alarms shown below
  - Solution:

    (1) Original distance between the two persons must be less than some threshold

    (2) The two persons should not have the same motion direction



Results in TRECVID-ED 2009 (1)

EVENT : ElevatorNoEntry

Analysis Report	#Ref	#Sys	#CorDet	#FA	#Miss	Act. RFA	Act. PMiss	Act. DCR	Min RFA	Min PMiss	Min DCR
BUPT-MCPRL_6 / p-baseline_6	3	23	2	21	1	1.377	0.333	0.340	1.377	0.333	0.340
BUPT-PRIS_1 / p-baseline_1	3	4	1	1	2	0.066	0.667	0.667	0.066	0.667	0.667
CMU_3 / p-VCUBE_1	3	1041	3	1038	0	68.078	0.000	0.340	7.739	0.000	0.039
PKU-IDM_4 / p-eSur_1	**3**	**4**	**2**	**2**	**1**	**0.131**	**0.333**	**0.334**	**0.066**	**0.333**	**0.334**
PKU-IDM_4 / p-eSur_3	3	0	0	0	3	0.000	1.000	1.000	0.000	1.000	1.000
SJTU_3 / p-baseline_1	3	28	2	26	1	1.705	0.333	0.342	1.640	0.333	0.342
Toshiba_1 / p-cohog_1	**3**	**90**	**2**	**1**	**1**	**0.066**	**0.333**	**0.334**	**0.656**	**0.000**	**0.003**

Results in TRECVID-ED 2009 (2)

- **EVENT :** PeopleMeet

Analysis Report	#Ref	#Sys	#CorDet	#FA	#Miss	Act. RFA	Act. PMiss	Act. DCR	Min RFA	Min PMiss	Min DCR
CMU_3 / p-VCUBE_1	449	2130	58	2072	391	135.894	0.871	1.550	36.466	0.998	1.180
NHKSTRL_2 / p-NHK-SYS1_1	449	991	55	905	394	59.355	0.877	1.174	1.508	0.991	0.999
PKU-IDM_4 / p-eSur_1	**449**	**125**	**7**	**118**	**442**	**7.739**	**0.984**	**1.023**	**1.705**	**0.991**	**1.000**
PKU-IDM_4 / p-eSur_2	449	210	15	195	434	12.789	0.967	1.030	0.000	0.998	0.998
PKU-IDM_4 / p-eSur_3	449	210	15	195	434	12.789	0.967	1.030	0.000	0.998	0.998
SJTU_3 / p-baseline_1	449	19739	108	7706	341	505.404	0.759	3.287	1.443	0.996	1.003
TITGT_1 / c-EVAL_1	449	14884	362	14522	87	952.436	0.194	4.956	952.436	0.194	4.956
TITGT_1 / p-EVAL_1	449	14161	354	13807	95	905.542	0.212	4.739	905.542	0.212	4.739

Results in TRECVID-ED 2009 (3)

- **EVENT :** PeopleSplitUp

Analysis Report	#Ref	#Sys	#CorDet	#FA	#Miss	Act. RFA	Act. PMiss	Act. DCR	Min RFA	Min PMiss	Min DCR
CMU_3 / p-VCUBE_1	187	10184	28	10156	159	666.088	0.850	4.181	0.721	0.995	0.998
PKU-IDM_4 / p-eSur_1	**187**	**198**	**7**	**191**	**180**	**12.527**	**0.963**	**1.025**	**0.525**	**0.995**	**0.997**
PKU-IDM_4 / p-eSur_2	187	881	14	867	173	56.863	0.925	1.209	0.066	0.995	0.995
PKU-IDM_4 / p-eSur_3	187	881	14	867	173	56.863	0.925	1.209	0.066	0.995	0.995
SJTU_3 / p-baseline_1	187	22877	66	11690	121	766.697	0.647	4.481	1.705	0.984	0.993
TITGT_1 / c-EVAL_1	187	15007	186	14821	1	972.046	0.005	4.866	972.046	0.005	4.866
TITGT_1 / p-EVAL_1	187	14239	184	14055	3	921.807	0.016	4.625	921.807	0.016	4.625

Results in TRECVID-ED 2009 (4)

- **EVENT :** Embrace

Analysis Report	#Ref	#Sys	#CorDet	#FA	#Miss	Act. RFA	Act. PMiss	Act. DCR	Min RFA	Min PMiss	Min DCR
CMU_3 / p-VCUBE_1	175	20080	146	19934	29	1307.386	0.166	6.703	1.377	0.989	0.996
NEC-UIUC_2 / c-none_1	175	0	0	0	175	0.000	1.000	1.000	0.000	1.000	1.000
PKU-IDM_4 / p-eSur_1	**175**	**80**	**1**	**79**	**174**	**5.181**	**0.994**	**1.020**	**3.870**	**0.994**	**1.014**
PKU-IDM_4 / p-eSur_2	175	164	3	161	172	10.559	0.983	1.036	1.312	0.994	1.001
PKU-IDM_4 / p-eSur_3	175	164	3	161	172	10.559	0.983	1.036	1.312	0.994	1.001
SFU_1 / p-match_1	175	6712	28	650	147	42.631	0.840	1.053	1.968	0.989	0.998
SJTU_3 / p-baseline_1	175	14189	64	1919	111	125.859	0.634	1.264	0.328	0.994	0.996

Results in TRECVID-ED 2009 (5)

- **EVENT :** PersonRuns

Analysis Report	#Ref	#Sys	#CorDet	#FA	#Miss	Act. RFA	Act. PMiss	Act. DCR	Min RFA	Min PMiss	Min DCR
BUPT-MCPRL_6 / p-baseline_6	107	25275	78	25197	29	1652.563	0.271	8.534	1652.563	0.271	8.534
BUPT-PRIS_1 / p-baseline_1	107	39	2	14	105	0.918	0.981	0.986	0.656	0.981	0.985
CMU_3 / p-VCUBE_1	107	23721	87	23634	20	1550.053	0.187	7.937	2.427	0.991	1.003
NEC-UIUC_2 / c-none_1	107	0	0	0	107	0.000	1.000	1.000	0.000	1.000	1.000
NEC-UIUC_2 / p-UI_1	107	157	1	38	106	2.492	0.991	1.003	1.180	0.991	0.997
NHKSTRL_2 / p-NHK-SYS1_1	**107**	**468**	**15**	**339**	**92**	**22.234**	**0.860**	**0.971**	**21.053**	**0.860**	**0.965**
PKU-IDM_4 / p-eSur_2	**107**	**356**	**5**	**351**	**102**	**23.021**	**0.953**	**1.068**	**3.673**	**0.981**	**1.000**
SFU_1 / p-match_1	107	30948	22	3078	85	201.873	0.794	1.804	0.984	0.981	0.986
SJTU_3 / p-baseline_1	107	2217	19	1228	88	80.539	0.822	1.225	21.447	0.981	1.089
TITGT_1 / c-EVAL_1	107	11062	70	10992	37	720.918	0.346	3.950	720.918	0.346	3.950
TITGT_1 / p-EVAL_1	107	11019	70	10949	37	718.098	0.346	3.936	718.098	0.346	3.936
Toshiba_1 / p-cohog_1	107	8380	1	176	106	11.543	0.991	1.048	0.262	0.991	0.992
UAM_1 / p-baseline_1	107	0	0	0	107	0.000	1.000	1.000	0.000	1.000	1.000

Illustrative Results

PersonRuns

ElevatorNoEntry

All results are obtained from TRECVID-ED 2008(except ElevatorNoEntry) according to the ground truth.

PeopleMeet

Embrace

PeopleSplitUp

[CLICK FOR PLAY]

Summary

- Our participation in TRECVID-ED 2009

- Events
  - PeopleMeet
  - PeopleSplitUp
  - Embrace
  - ElevatorNoEntry
  - PersonRun
  - CellToEar
  - ObjectPut
  - TakePicture
  - Pointing
  - OpposingFlow

- Submitted 5 event detection results
- 4 of them obtain significantly improvements over the best results of TRECVID-ED 2008

- Three-fold contributions:
  - Effective strategies for adaptive background modeling, human detection and tracking
  - An ensemble approach of one-vs.-all SVM and automata-based classifiers for both single-actor and pair-activity events
  - Post-processing to reduce the false alarm

Summary

- Future Work
  - Better human detection and tracking in crowd scenes
  - Better discriminative features such as temporally integrated spatial response (TISR) descriptor [Zhu, MM09]
  - More effective event classification models, such as MKL and sequence learning

G.Y. Zhu, M. Yang, K. Yu, W. Xu, Y.H. Gong, Detecting Video Events Based on Action Recognition in Complex Scenes Using Spatio-Temporal Descriptor. MM09

Thanks!

yhtian@pku.edu.cn

Our Team
(Main members
in the first row)

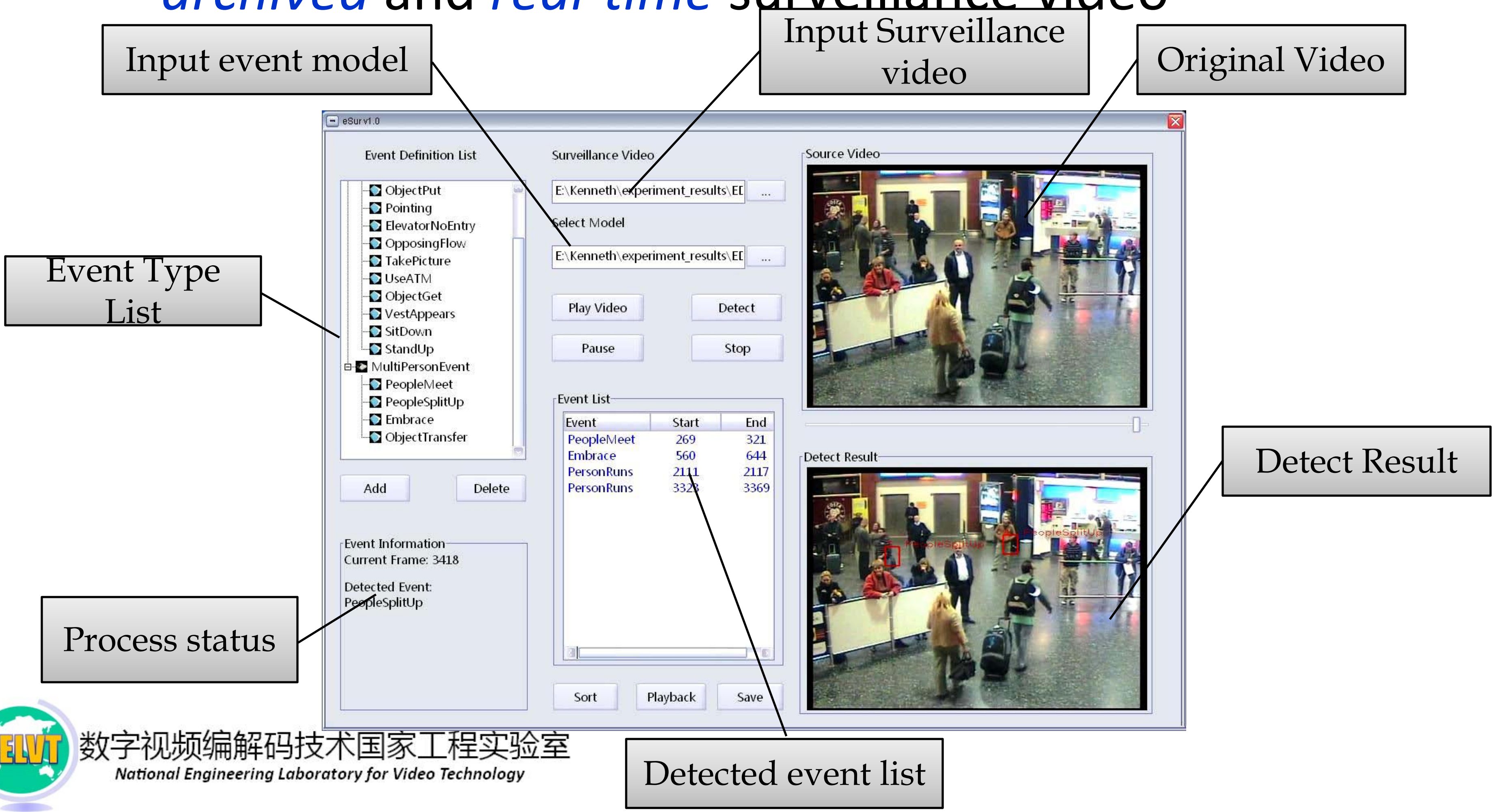
数字视频
National Eng

Our Solution:

The eSur System: An e-Sir for Event Surveillance

e-Sir: electronic policeman

- Towards an integrated system for analyzing *archived* and *real-time* surveillance video



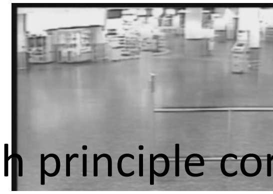
Basic Idea of Block-wise PCA

Select the best reconstructed background according to the MMSE to the mean background

$$B = \arg\min_{B_i} \left\| \bar{I} - B_i \right\|^2$$

Project the frame on each principle component to reconstruct background respectively

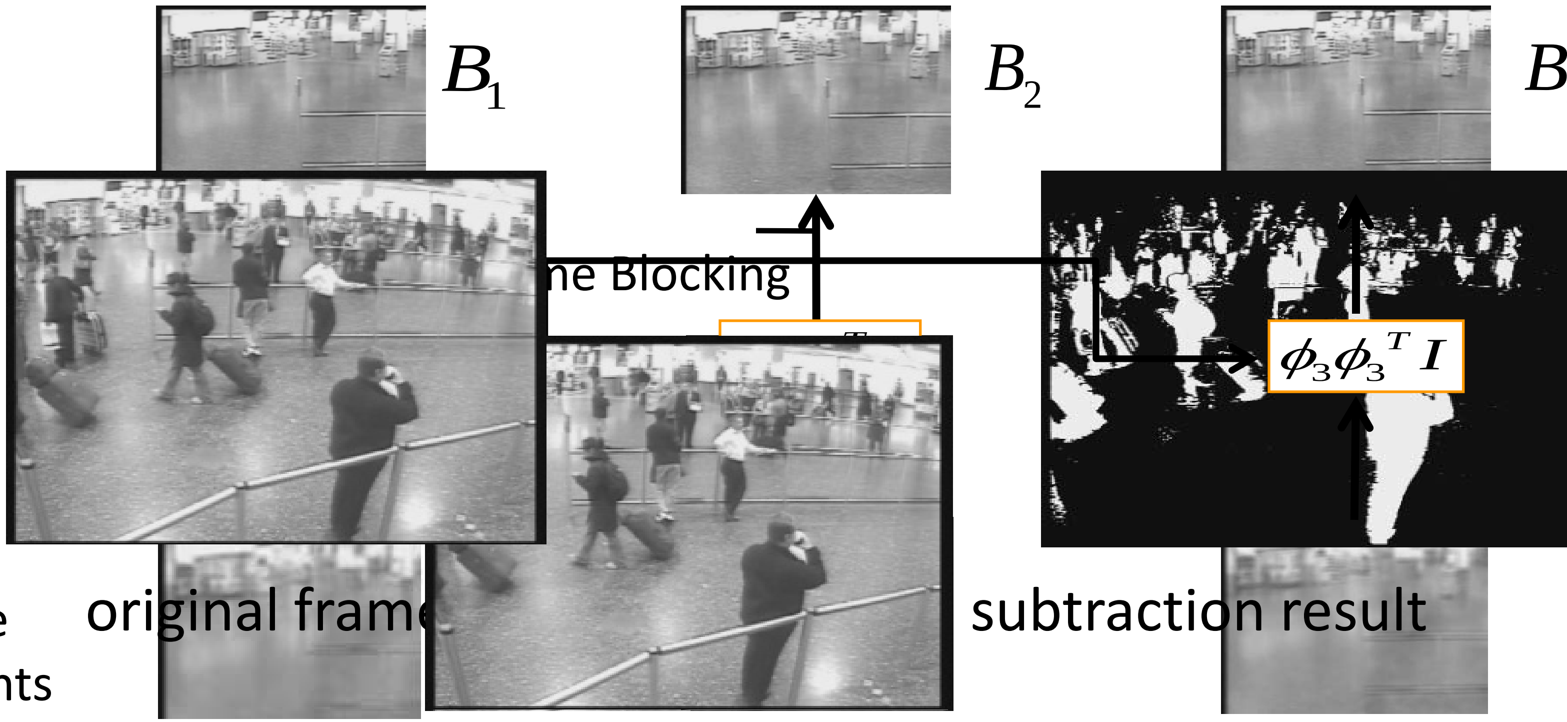
subtract with the best reconstructed background



Trained Mean Background

$$\bar{I} = \frac{1}{N}\sum_{i=1}^{N} I_{train}(i)$$

I



Trained Principle Components

aten

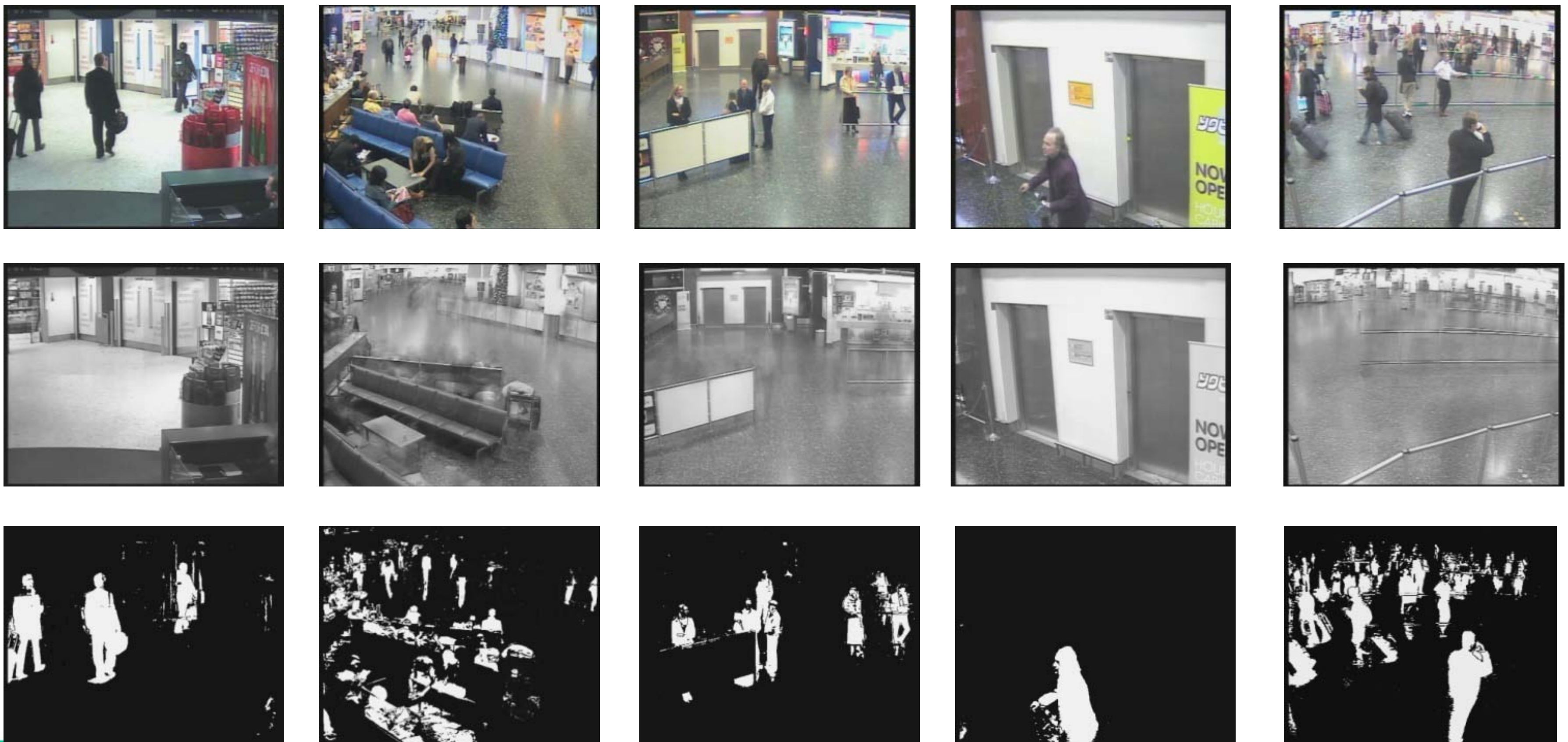
First
Principle Component
ϕ_1

Second
Principle Component
ϕ_2

Third
Principle Component
ϕ_3

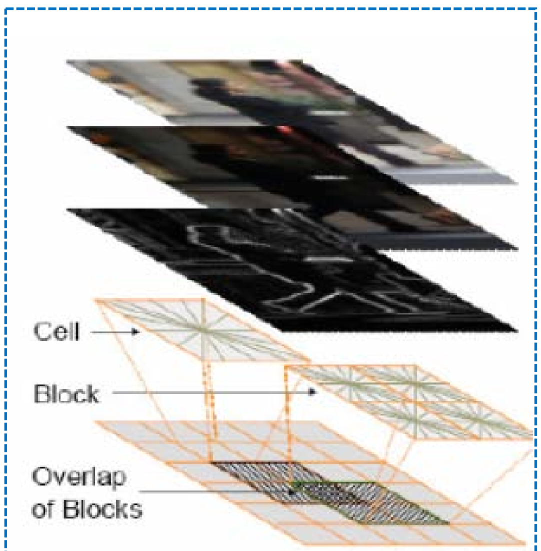
Results of Block-wise PCA

- The background subtraction results using block-wise PCA

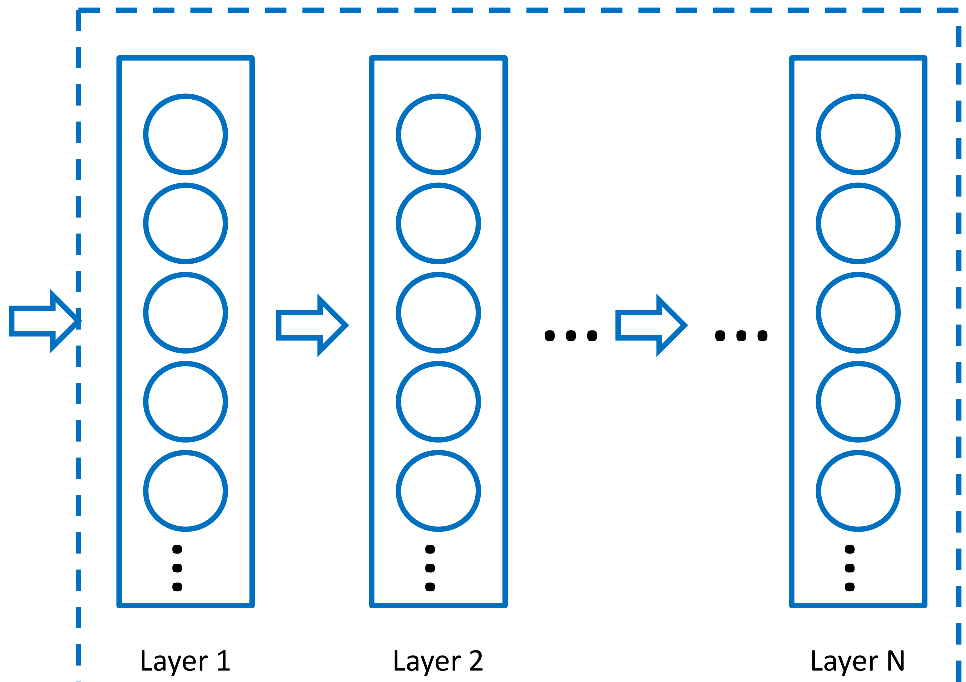


Detection

- HOG-based feature for both human and HS detection
  - AdaBoost for feature selection
- Cascaded structure
  - Different weak classifier for each layer for simplicity, but not SVM as in [Zhu, CVPR06]



HOG Feature
(Dalal,CVPR05)



Cascaded

Dalal. CVPR05. Histoarams of Oriented Gradients for Human Detection
Zhu, CVPR06 Fast Human Detection Using a Cascade of Histograms of Oriented Gradients

Final Tracking Result

数字视频编解码技术国家工程实验室
National Engineering Laboratory for Video Technology

[CLICK FOR PLAY]

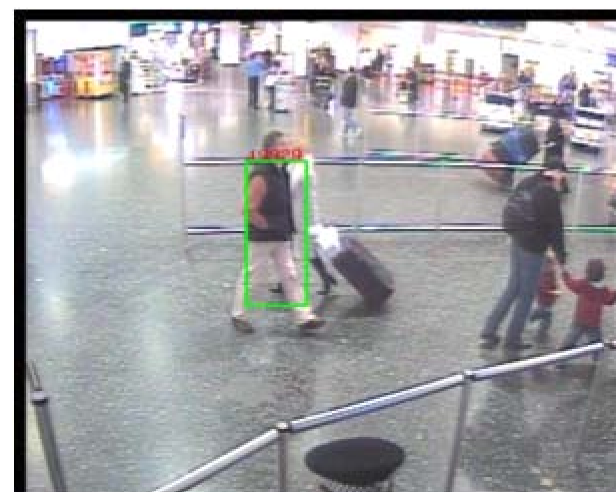
Illustrative Example – False Alarm

PersonRuns

ElevatorNoEntry

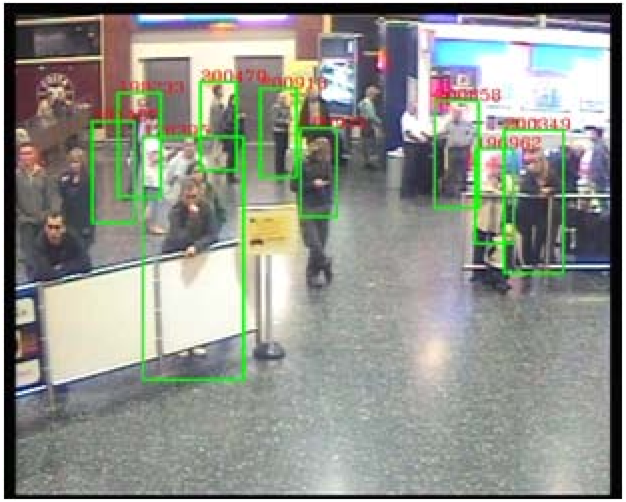
PeopleMeet

Embrace



PeopleSplitUp

Illustrative Example – Miss



PersonRuns

ElevatorNoEntry

PeopleMeet

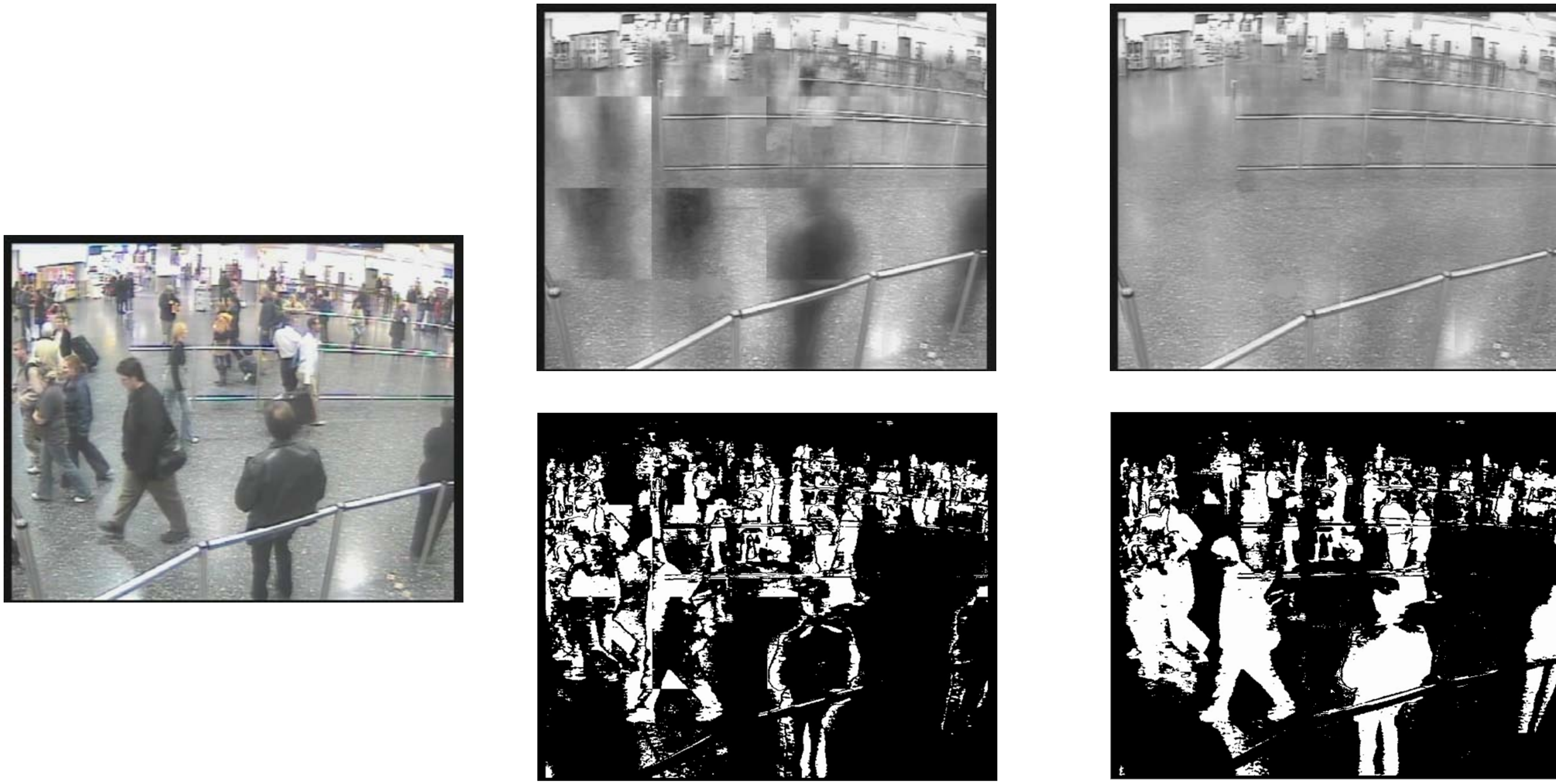
Embrace

PeopleSplitUp

National Engineering Laboratory for Video Technology

Comparative Results

- Block PCA vs. Block-wise PCA



original image　　Block PCA　　Block-wise PCA