CCNY at TRECVID 2014: Surveillance Event Detection

Yang Xian‡, Xuejian Rong§, Xiaodong Yang§, and YingLi Tian‡§

‡Dept. Computer Science
The Graduate Center, CUNY

§Dept. Electrical Engineering
The City College, CUNY

INTRODUCTION

- We present a specific action detection system for CellToEar task and a generic event detection system for the rest events of Surveillance Event Detection (SED).
- Our generic system consists of four components: (1) low-level feature extraction, (2) video representation, (3) learning event model, and (4) post processing, as shown in Fig. 1.
- STIP-HOG/HOF, DT-Trajectory, and DT-MBH are used as the low-level features to represent human actions. The camera and event specific hot regions are employed to eliminate a large amount of irrelevant points from background.
- We employ Fisher Vector for further feature descriptor, which shares the benefits of both generative and discriminative models.

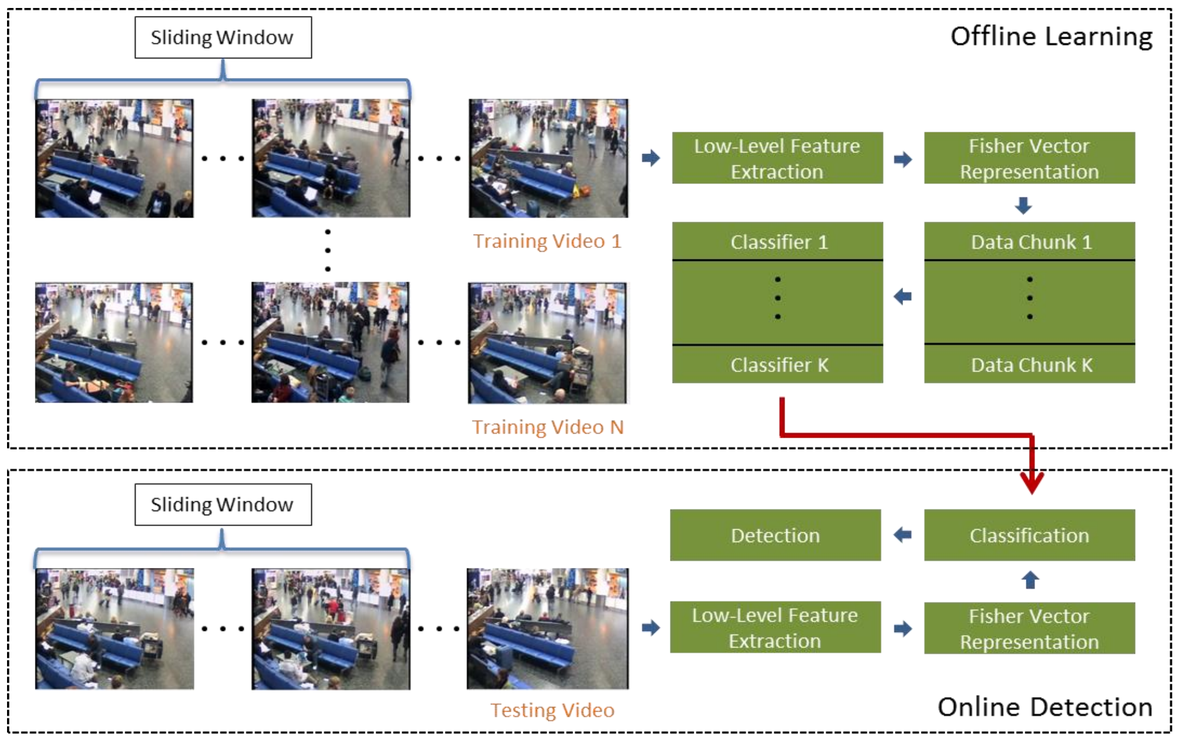
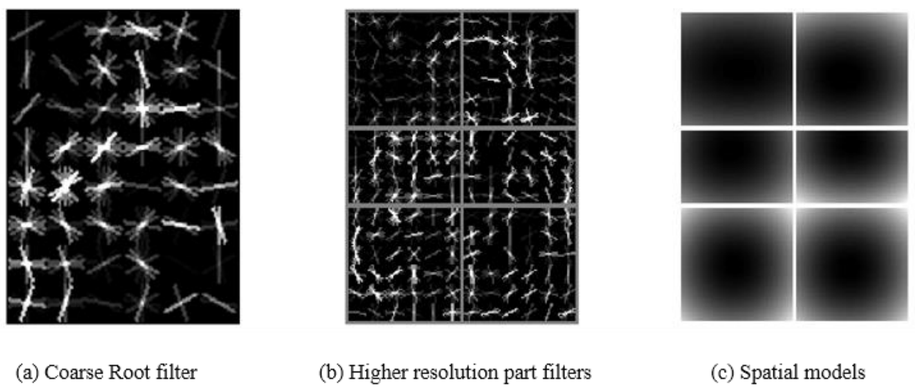


Figure 1: CCNY generic event detection system architecture.

CellToEar Specific System

▪ Part Models Training



(a) Coarse Root filter (b) Higher resolution part filters (c) Spatial models

Figure 2: Trained visual Deformable Part Models for CellToEar event.

- Models are trained for four scenes from different camera views, and the final training dataset contains ~15000 positive frames from *dev08* and ~6200 positive frames from *eval08*, with all bounding boxes manually labeled before.

▪Part Models based Detection

Figure 3: Initial detection bounding boxes including part models.

Generic System

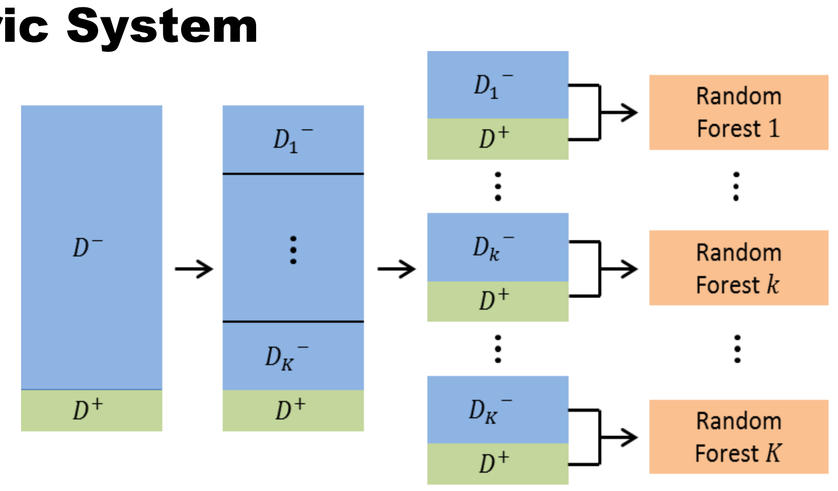


Figure 4: Illustration of data segmentation where within each data chunk a Random Forest is learned.

▪ Video Representation

- 60-frame sliding window which strides in every 15 frames. Highly imbalanced data in different events.
- Three low-level features are extracted from each sliding window, each generates a corresponding Fisher Vector.
- Each Fisher Vector is fed into a group of learned Random Forests, following classification and decision-level fusion as Fig. 5. shows.

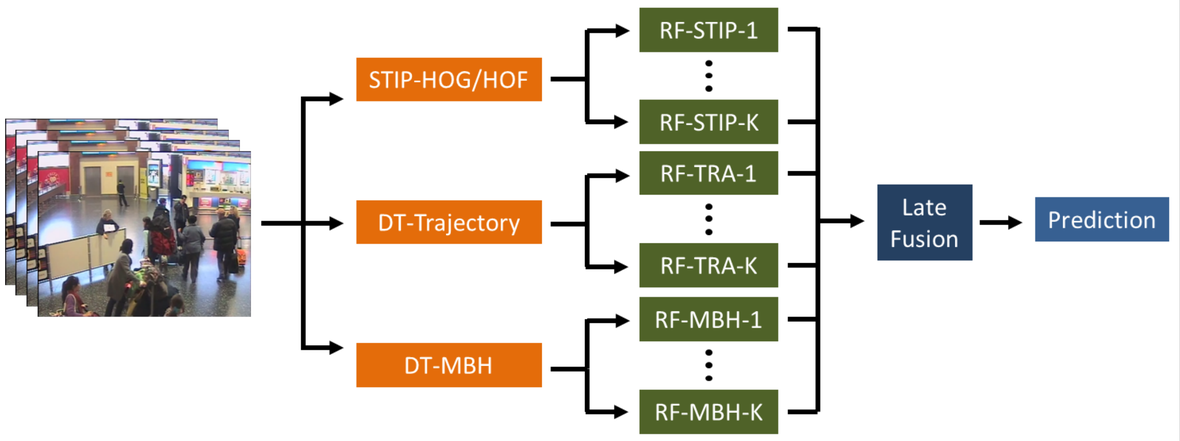


Figure 5: Illustration of late fusion in combining multiple low-level features.

▪ Decision-level Fusion

- The decision-level fusion combines outputs of multiple classifiers to make the final prediction.

▪ Post Processing

- Two positive predictions which have overlaps in their sliding windows can be merged together.

RESULTS

▪Comparisons between our system and the best systems in 2014 are listed in Table 2.

Event	Rank	ADCR of Other Best Systems	CCNY Primary Run				
			ADCR	MDCR	#CorDet	#FA	#Miss
CellToEar	3	0.9921	1.0257	1.0005	0	56	54
Embrace	4	0.8113	0.9611	0.9510	14	136	124
ObjectPut	3	0.9713	1.0177	1.0005	1	46	289
PeopleMeet	3	0.8587	0.9966	0.9901	11	86	245
PeopleSplitUp	2	0.8353	0.8698	0.8594	36	232	116
PersonRuns	1	0.8301	**0.8256**	0.8122	13	175	38
Pointing	4	0.9998	1.0547	1.0005	19	171	776

Table 1: Comparisons between CCNY and TRECVID SED best systems in 2014.