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

Zhipeng Hu [a], Guangnan Ye [b], Guochen Jia [a], Xibin Chen [b], Qiong Hu [c], Kaihua Jiang [b], Yaowei Wang [a, d], Lei Qing [c], Yonghong Tian [a], Xihong Wu [b], Wen Gao [a]

[a] National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China

[b] Speech and Hearing Research Center, Peking University, Beijing 100871, China

[c] Key Lab of Intel. Inf. Proc., Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

[d] Department of Electronic Engineering, Beijing Institute of Technology, Beijing 100081, China

Abstract

In this paper, we describe our eSur system and experiments performed for the surveillance event detection task in TRECVID 2009. In the experiments, we addressed the detection problems of two categories of events: 1) single-actor events (e.g., PersonRuns and ElevatorNoEntry) that require only whole body modeling and no interaction with other persons, and 2) pair-activity events (e.g., PeopleMeet, PeopleSplitUp, Embrace) that need to explore the relationship between two active persons based on their motion information. Our contributions are three-folds. First, we designed effective strategies for background modeling, human detection and tracking. Second, we proposed an ensemble approach for both single-actor and pair-activity event analysis by fusing One-vs.-All SVM and rule-based classifier. Third, event merging and post-processing based on prior knowledge were applied to refine the system detection outputs, consequently reducing the false alarm in the event detection. We submitted three runs (i.e., p-eSur_1, p-eSur_2 and p-eSur_3), which were obtained by using different human detection and tracking modules. According to the TRECVid-ED formal evaluation, our prototype has yielded fairly promising results over TRECVid'09 dataset, with top Act.DCR of 1.023, 1.025, 1.02, and 0.334 for PeopleMeet, PeopleSplitUp, Embrace, and ElevatorNoEntry, respectively.

1. Introduction

Event detection in surveillance environments is a critical application in computer vision field. Although there are deep and extensive studies on this issue, the applicable system is still far away from our life due to the following challenges:(1) clutter scenes (2) illumination variations (3) partial and full occlusion between persons (4) different views of the same type of event (5) no clear event definition (6) no appropriate modeling method of event. (7) Real-time computation. In order to address part of the difficulties mentioned above, our team, PKU_IDM, participated in the retrospective events task of TRECVID 2009.

We chose five events, PersonRuns, ElevatorNoEntry, PeopleMeet, PeopleSplitUp and Embrace, in the ten retrospective events specified by the task. The five events are classified into two categories by whether a person has interactions with others. Thus the problem is formulated as detection of pair-activity events (e.g., PeopleMeet, PeopleSplitUp, Embrace), which needs to explore the relationship between two active persons based on their motion information, and detection of single-actor events (e.g., PersonRuns and ElevatorNoEntry) that requires only whole body modeling.

Following a typical machine learning framework, our system, called eSur, includes feature extraction, training and classification components. In [1], Ryan Rifkin claimed that one-vs.-all scheme would achieve better performance than one-for-all scheme. Thus, One-vs.-All SVM is selected to identify different events in our system. Also, for each camera, prior rules are fused with SVM classifiers to boost the classification results. Our contributions are three-folds.

First, we designed effective strategies for background modeling, human detection and tracking. For background modeling, we used Mixture of Gaussians (MoG) and Adaptive Block-wise PCA. The two background models can be used to distinguish different camera scenes and effectively accelerate the searching process while decreasing the false alarm in human detection. Within the extracted foreground region, we used the cascaded Histograms of Oriented Gradients (HoG) [2] for human body and head-shoulder detection. The online-boosting method is then used for tracking each detection part.

In addition, the tracking process is assisted by the detection result so that we can determine the initialization and termination of each tracklet, and dominant color information is used to solve the drifting problem.

Second, we proposed an ensemble approach for both single-actor and pair-activity event analysis by fusing One-vs.-All SVM and Rule-based classifier. For the pair-activity events, we first treated PeopleMeet, PeopleSplitUp and Embrace as the same event and employed One-vs.-All SVM to effectively separate them from the others. Then the three events are identified by object motion information and separated from each other by three One-vs.-All SVM classifiers. For single-actor events, both One-vs.-All SVM and Rule-based classifier are used in our experiments. In particular, elevator state detection and human disappearing modeling are used to effectively detect the ElevatorNoEntry event.

Third, event merging and post-processing based on prior knowledge are applied to refine the system detection results, consequently reducing the false alarm in the event detection effectively.

The remainder of this paper is organized as follows. In section 2, we present our system framework briefly. Background subtraction technology is described in section 3. In section 4, we describe our human detection and tracking approach. In section 5, we present our approach for detecting and identifying different events. Experimental results and analysis are given out in section 6. Finally, we conclude this paper in section 7.

2. The eSur System Framework

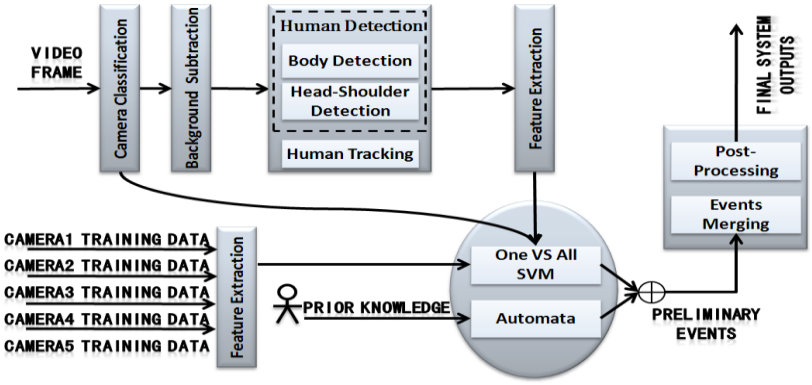


Fig.1 the diagram of our system, eSur

The diagram of our eSur (Event detection system on **SUR**veillance video) system is shown in Fig.1. Background model is generated for each camera and camera classification is performed based on these models. Background subtraction is employed to extract the foreground regions, thus reducing the searching space in human detection and tracking. In the human detection stage, human body detection results and head-shoulder detection results based on HoG are fused to obtain the final human detection results, followed by tracking using online-boosting and dominant color features. Meanwhile, human's motion information and three relation matrices are calculated from object tracking to interpret the object moving conditions. For each camera corpus, One-vs.-All SVM classifier is trained for each event to grasp the nature of the event and determine whether the event occurred in a given time span. An ensemble approach is also employed for single-actor event analysis by fusing One-vs.-All SVM and rule-based classifier. Finally, an event merging and post-processing procedure is applied to refine the system detection outputs.

As mentioned above, the motion characteristics of a human and his relationship with other persons can be represented by motion features and relation matrices. To identify an event, a window of 12 consecutive frames slides in the video. When the window is labeled TRUE by the classifier of some event, we then trace the corresponding person(s) backward or forward to find the event's start frame and end frame.

For the training corpus, all events are easily addressed by the ground truth data and we manually labeled the corresponding humans in each event with bounding boxes. We divided the training data to ten subsets, each for one event. As a result, performance of eSur is verified through a ten cross-validation.

3. Background Subtraction

In our framework, background subtraction is applied in two ways: (1) Reconstruct the stationary backgrounds for camera classification. (2) Extract foreground regions to accelerate the human detection process and decrease the detection false alarms effectively at the same time.

In [3],Stauffer and Grimson modeled the distribution of the pixel values by using Mixture of Gaussians. In [4], Elgammal applied non-parametric model using only recent observed pixel values in background subtraction. In [5], Kim et al quantized the background values into codebooks which represent a compressed form of background model. Background subtraction using motion information has also been explored in [6] and [7]. Eigenbackgrounds (also called PCA Background) [8] can also detect static objects. As most of the events in TRECVID corpus are somewhat related to static humans, we chose MoG and PCA with some simplification and improvement as our background subtraction technologies.

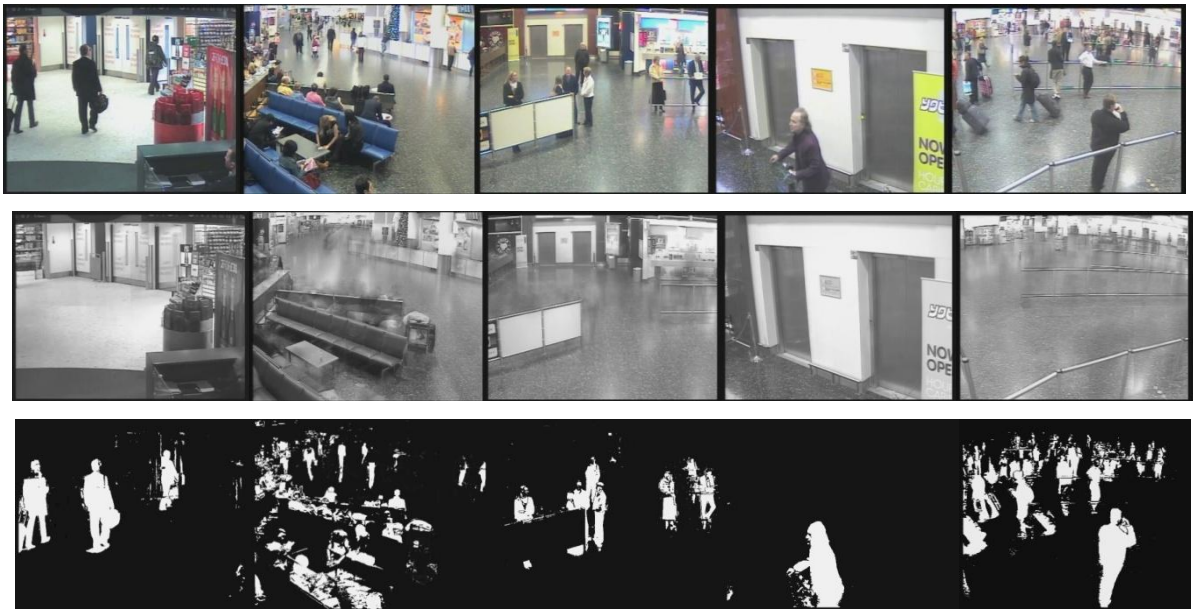
For MoG, we simplify the primary algorithm by eliminating the updating stage. 1000 frames in each camera are sampled and the foreground objects are manually labeled in each frame. Then EM algorithm is applied on the labeled data to construct a MoG background model of five Gaussian components.

PCA is very different from MoG. Four hundred frames are uniformly sampled from each video and singular value deposition is computed to obtain eigenbackgrounds in the training stage. To get lower complexity and higher recall ratio, Adaptive Block-wise PCA is proposed. Each frame is segmented into several blocks and background subtraction is performed respectively for each block. First, backgrounds are reconstructed by projecting the block image onto each of the trained eigenbackgrounds. Then the best reconstructed background is selected for subtraction according to the minimized mean square error between the reconstructed background and the trained mean background as follows:

$$B = \arg\min_{B_i} \left\| B_i - \bar{I} \right\|^2 \tag{3.1}$$

$$B_i = \phi_i \phi_i^T I \quad 1 \le i \le M \tag{3.2}$$

where $\bar{I}$ is the trained mean background, ϕ_i is the i*th* eigenbackground, B_i is the i*th* reconstructed background and M is the number of eigenbackgrounds. The projection coefficients $\{\phi_i^T I\}_i$ can be used to identify different cameras in camera classification stage.



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

Fig.2 The background subtraction results using Adaptive Block-wise PCA for (a) Cam1, (b) Cam2, (c) Cam3, (d) Cam4 and (e) Cam5 respectively. From top to down, the figures show the source frame, reconstructed

background and the subtraction result

Experimental results demonstrate that the Adaptive Block-wise PCA is effective in our corpus. Fig.2 shows the background subtraction results using Adaptive Block-wise PCA for (a) Cam1, (b) Cam2, (c) Cam3, (d) Cam4 and (e) Cam5 respectively.

4. Detection and Tracking

Detection and Tracking module directly affects the recall and precision of the system. We used cascade of HoG framework for human detection and head-shoulder detection. The integration of the two detection results would solve the occlusion problem properly. For tracking, we first utilized the detection results for initializing the tracking trajectory. Online boosting method was then combined with the detection module to grow and terminate the trajectory. We used the method above frame by frame, called "Forward tracking". In order to make our system more robust, we also did it in a reverse direction for the video, called "Backward tracking". The whole framework is interpreted in Fig.3.

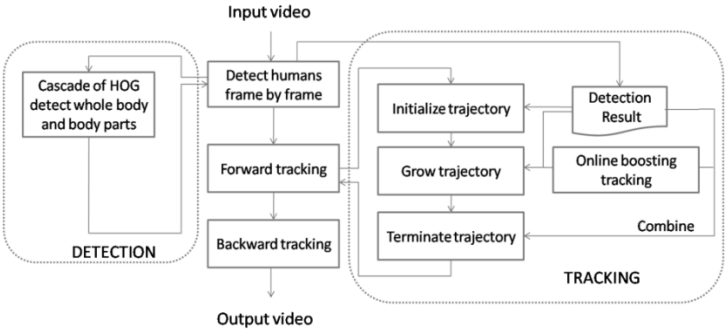


Fig.3 Detection and Tracking Diagram

4.1 Human body and head-shoulder detection

There are many occlusions in the TRECVID corpus, and sometimes the whole bodies of persons are not shown in the video scenes. For these situations, head-shoulder detection performs better. As a result, we integrate human detection and head-shoulder detection in our system.

Haar feature is widely used in face detection. Papageorgious and Poggio [9] first used Haar-based representation for human detection. Viola et al. [10] extended Haar-like feature for moving-human detection. After that, Dalal and Triggs [2] got a far better result with HOG feature. Zhu et al. [11] added a cascaded framework to improve HOG feature and it significantly increased the speed of the original method. Besides, Wu and Nevatia [12] also tried an edge like feature called Edgelet, and part-based approach is applied in their framework.

In [2], Dalal and Triggs have proved that Histograms of Oriented Gradients is powerful enough to pedestrian detection. However, the method can only process 320×240 images with a low speed. In order to speed up their method, Zhu *et al.* combined the cascade of rejecter approach with the HOG feature. They used AdaBoost to select the best feature and constructed the rejecter-based cascade. This method significantly accelerated the speed of the system, and they have gotten a near real-time human detection result. In order to get a high performance system, we bring in Zhu's cascaded HOG framework and make a minor change. Compared with Zhu's system, we use different weak classifier for each layer, but not SVM as mentioned Zhu's paper.

For training the classifier, we labeled 5000 persons and head-shoulders in each camera. Human head-shoulder and whole body are detected separately. The two detection results for one person, head and the whole body, are combined by using Bayesian combination. Besides, with the coarse foreground regions extracted from background modeling module, we wipe out candidate regions that do not have enough foreground in them. Moreover, by using statistical data of each camera, we can simply estimate the possible size of person appeared in different position. By using this prior information, the

detection process is more efficient. Additionally, regions those have low possibility of events are pruned in searching process in order to accelerate the searching progress.

4.2 Tracking

Tracking problem is solved in several common ways. Appearance-based method such as Mean-Shift [13] uses color features. The motion condition of the object can be described through a Bayesian filtering function (e.g. Kalman filter and Particle filter). The second method is segmentation-based tracking. Graph Cut [14], as a method for segmentation, can be a way for solving the tracking problem. Furthermore, Helmut Grabner *et al*. [15] consider tracking as a classification problem. Online Boosting is such an example to extract the robust feature in the adaptive process. The fourth method is presented by Huang *et al*. [16]. In their framework, tracking is considered to make the association for each detection result.

In the TRECVID corpus, target object appearance always changes significantly. As a result, we use adaptive Online Boosting framework for tracking process. As it is described by Helmut Grabner, the detection result of each step is regarded as positive data, while the surrounding four blocks are regarded as negative data.

As it is described as Fig. 4, each tracklet is first triggered by the head-shoulder and body detection results. Throughout the Online Boosting method, prediction process is started and each detected object got an expected target in the next frame. Consecutively, each expected target is tested by detection module to confirm whether drifting is happening in tracking process. Based on the prediction probability, if the target object is drifted, termination module is triggered and the tracking process is ended. Else, similarity comparison is made between the new tracklet and other existent tracklets for further combination. Dominant color similarity is regarded as a method for this process. Moreover, Online Boosting can also provide us a similarity score. If any two tracklets are very similar, the two results are combined and continued to trigger another tracking process for the combined one. Otherwise, another tracking process is only triggered for the new tracklet. The whole process is doing continuously until the termination process is triggered. Finally, every trajectory is approached as long as possible based on tracklets, and the final result is refined. A backward tracking process is employed either. The process is the same as the forward tracking. By fusing the tracking result got from forward and backward, the result is further improved.

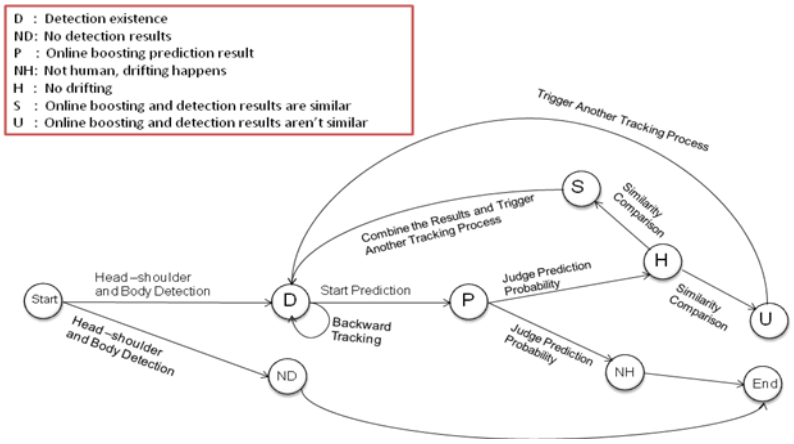


Fig. 4 State Machine of Tracking

5. Events detection

Automatic events detection is a highly active research area due both to the number of potential applications and its inherent complexity. Various methods such as HMM(Hidden Markov Model)[17], CFG(Context-free Grammar)[18], have been used to describe events. Luo[19] presented a dynamic Bayesian network to analysis and interpret human motion in sports video sequences, for the sake of simplicity and flexibility during training. The work similar to ours is that of [20], which used SVM to classify optical flow histogram and recognize actions. In our system, we extract object motion features from tracking results, then apply classifiers to detect specific events.

As mentioned in section 1, we address the problem in two categories, pair-activity events and single-actor events. To separate events from each other, we treat multiple events detection as a multi-class classification problem, so One-vs.-All scheme is employed. Three pair-activity events, PeopleMeet, PeopleSplitUp and Embrace, are identified by object motion

patterns and separated from others by three binary classifiers. For single-actor events, both One-vs.-All SVM and rule-based classifier are used. Events merging and post processing based on prior knowledge are applied to refine system outputs.

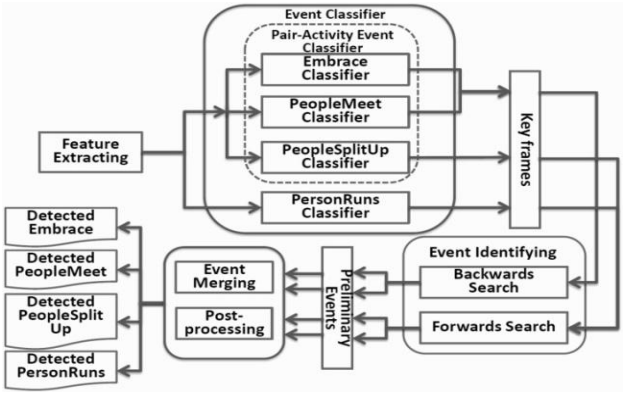


Fig.5 the diagram of event detection based on motion information

5.1 Different Classifier Strategies

Since we treat detection as a classification problem, there is a kind of needs to design and evaluate different classifier strategies. We tried SVM classifier, hierarchal classifiers and multi-kernel learning (MKL) classifier. Firstly, we chose SVM classifier proposal, which means training binary classifier for each event with others regarded as negative samples. Then we extended it to hierarchal classifier as follows: combine all required events as one event to distinguish from others; then choose similar events as one to further separate them; lastly recognize specific event respectively. We also used MKL method with rbf, linear and poly kernels to enhance classification performance. Comparison experiments were carried on TREVid-ED'08 data. In the end, SVM classifier was chosen for the sake of simplicity and efficiency.

5.2 Pair-activity events

Pair-activity events are those who involve two persons' interaction. In general, related works deal with objects through a long term video sequence, while we identify some key frames which characterize events. Recognizing such key frames could leads to valid event detection. In the "Embrace" and "PeopleMeet" cases, there is no significant beginning of events, but the pair-persons are remarkably close with each other at last. In other words, key frames are those ones at the end of the event. For the same reason, in "PeopleSplitUp" events, key frames are located at the beginning.

Our system picks distance, co-occurrence span and motion direction correlation of two persons as features, which are extracted in a sliding window with twelve consecutive frames. All these features are integrated into one feature vector with features depicted as follows:

$$dist(obj_1, obj_2) = \left\| pos_{obj1} - pos_{obj2} \right\| \tag{5.1}$$

$$relCode(obj_1, obj_2) = \begin{cases} 0 & if \ \ 0 < |\theta_1 - \theta_2| < \pi/6 \\ 1 & if \ \ 5\pi/6 < |\theta_1 - \theta_2| < \pi \\ 2 & \text{otherwise} \end{cases} \tag{5.2}$$

$$cotime(obj_1, obj_2) = min(t_{end}(obj_1), t_{end}(obj_2)) - max(t_{start}(obj_1), t_{start}(obj_2)) \tag{5.3}$$

All frames will be processed and output millions of feature vectors to pair-activity classifiers. In this way, all key frames will be extracted. As illustrated above, once we get such key frames of "Embrace" or "PeopleMeet", system will carry on backward search strategy to find response event's beginning. In contrast, a "PeopleSplitUp" prediction means a forward search will be implemented. After that, preliminary event interval is obtained. Consecutively, these are merged if they have similar time interval.

At last, some rules based on prior knowledge are introduced. The changing patterns of distance between correlated

persons are the key of pair-activity events. If the distance at the end of the event is greater than a threshold for “PeopleMeet” and ”Embrace”, or the distance at the beginning of the event is greater than a threshold for “PeopleSplitUp”, the detected event is identified as a false alarm. For PeopleSplitUp, we also assume that the motion directions of the two related persons should be different.

5.3 Single-actor events

For “PersonRuns” event, it is observed that a running person have a higher velocity than others, with motion direction consistency. Therefore, we integrate velocities and motion directions within a sliding window as the feature vector and use classifiers to identify “PersonRuns”. After that, objects labeled running are collected and stitched according to their identifications. Similarly, events that have overlap time intervals are merged to get preliminary results. Using statistical results, we can limit the object position and motion direction. For instance, people always run from left-bottom to top in the scene of camera1. Referring to this rule, we wiped out a lot of false alarms caused by tracking drifting.

For ElevatorNoEntry event, the first step is scene recognition. By observing the corpus, we can easily find that there is no elevator in scenes of camera one, two and five. Therefore, only if the scene is determined to come from camera 3 or 4, the following steps are executed.

The state transition diagram in Fig.6 presents our basic idea of ElvatorNoEntry Event Detection. Because elevators’ positions are absolutely fixed, we can know where the elevator is empirically. When the elevator is fully closed, the elevator regions are labeled as background after background subtraction while when the door is open, the elevator regions are detected as foreground. Thus, we can identify each elevator’s states (open or closed) by using Adaptive Block-Wise PCA. The time at which the elevator begins to open or is fully closed is detected and recorded.

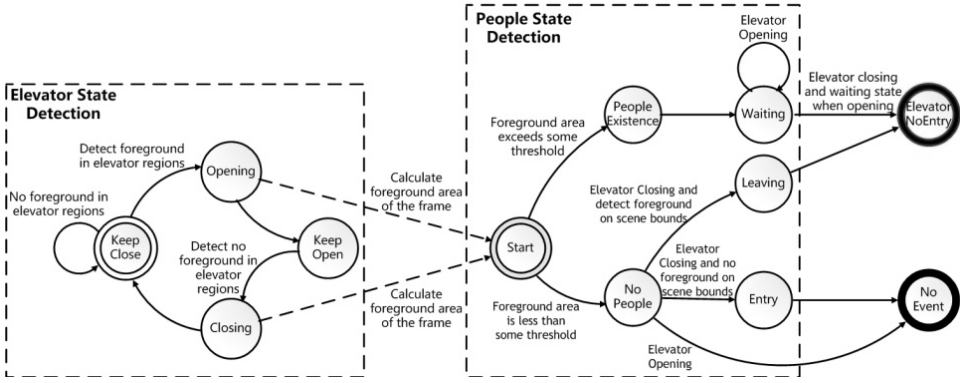


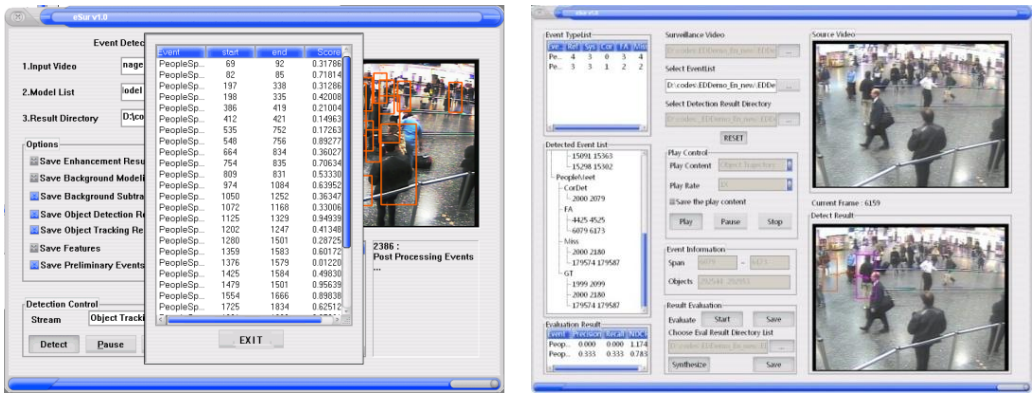
Fig.6 State transition diagram in ElevatorNoEntry detection

“ElevatorNoEntry” is defined as “elevator doors open with a person waiting in front of them, but the person does not get in before the doors close”. The foreground area is due to the number of persons in front of the elevators. Therefore, for a new frame, ElevatorNoEntry event can be detected according to the elevators’ states and human disappearing modeling. To be more exact, if we detect somebody is waiting in front of the elevators when the door opens and we still detect somebody waiting when the door is fully closed, an ElevatorNoEntry event is identified; otherwise, we think the person enters the elevator. However, there is the following exception: somebody is waiting when the door opens but he leaves the scene when the door closes. In this case, the human also disappears but the event will not be detected. To avoid this kind of miss, we will detect and record whether there is person passing the boundaries of the scene after the door opens and before it closes. If so, we believe the person leaves the scene and report an ElevatorNoEntry event. The criteria for determining whether there is person waiting in the scene or passing the scene boundaries is whether the foreground areas exceeds some thresholds.

6. Experiment and results

The interface of our eSur system is shown in Fig.7. The left picture shows the interface of event detection. Throughout

the interface, system user can check the intermediate results during event detection, including source frame, enhanced frame, reconstructed background, subtracted foreground, detected humans, tracking result and the detected events (including start frame, end frame and decision score) as shown in Fig.7(a). We can also analyze our detection results according to the ground truths with the interface shown in Fig.7(b). First, the detected events are classified into three categories: Correct Detection, False Alarm and Miss Event. With the classification results, NDCR, the final evaluation criteria, can be calculated. Fig.7(b) gives out the analysis results in a short time span and the trajectories of the two persons related to a detected PeopleMeet event.



(a) Interface of Event Detection (b)Interface of Event Analysis

Fig.7 Interface of eSur system

Our team submitted three results, p-eSur_1, p-eSur_2 and p-eSur_3, which are obtained by using different human detection and tracking modules. In p-eSur_1, only head-shoulder detection is used. In p-eSur_2, background model is applied to accelerate the searching process and to refine detection results. In p-eSur_3, dominant color features are employed to improve tracking precision and to reduce drifting. Table 6.1 shows the formal evaluation results of TRECVID2009-ED announced by NIST. According to the results and the final evaluation criteria NDCR given in [21], our experimental results are promising. Among all submissions for the formal evaluation, we have four detection results (e.g., PeopleMeet, PeopleSplitUp, Embrace, and ElevatorNoEntry) with top NDCR.

Table 6.1 Formal Evaluation Results of TRECVID2009-ED

PeopleSplitUp

Participant	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
PKU-IDM_4 / p-eSur_1	187	198	7	191	180	1.025
PKU-IDM_4 / p-eSur_2	187	881	14	867	173	1.209
PKU-IDM_4 / p-eSur_3	187	881	14	867	173	1.209
CMU_3 / p-VCUBE_1	187	10184	28	10156	159	4.181
SJTU_3 / p-baseline_1	187	22877	66	11690	121	4.481
TITGT_1 / p-EVAL_1	187	14239	184	14055	3	4.625
TITGT_1 / c-EVAL_1	187	15007	186	14821	1	4.866

Embrace

Participant	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
NEC-UIUC_2 / c-none_1	175	0	0	0	175	1
PKU-IDM_4 / p-eSur_1	175	80	1	79	174	1.02
PKU-IDM_4 / p-eSur_2	175	164	3	161	172	1.036
PKU-IDM_4 / p-eSur_3	175	164	3	161	172	1.036
SFU_1 / p-match_1	175	6712	28	650	147	1.053
SJTU_3 / p-baseline_1	175	14189	64	1919	111	1.264
CMU_3 / p-VCUBE_1	175	20080	146	19934	29	6.703

PeopleMeet

Participant	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
PKU-IDM_4 / p-eSur_1	449	125	7	118	442	1.023
PKU-IDM_4 / p-eSur_2	449	210	15	195	434	1.03
PKU-IDM_4 / p-eSur_3	449	210	15	195	434	1.03
NHKSTRL_2 / p-NHK-SYS1_1	449	991	55	905	394	1.174
CMU_3 / p-VCUBE_1	449	2130	58	2072	391	1.55
SJTU_3 / p-baseline_1	449	19739	108	7706	341	3.287
TITGT_1 / p-EVAL_1	449	14161	354	13807	95	4.739
TITGT_1 / c-EVAL_1	449	14884	362	14522	87	4.956

ElevatorNoEntry

Participant	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
PKU-IDM_4 / p-eSur_1	3	4	2	2	1	0.334
Toshiba_1 / p-cohog_1	3	90	2	1	1	0.334
CMU_3 / p-VCUBE_1	3	1041	3	1038	0	0.34
BUPT-MCPRL_6 / p-baseline_6	3	23	2	21	1	0.34
SJTU_3 / p-baseline_1	3	28	2	26	1	0.342
BUPT-PRIS_1 / p-baseline_1	3	4	1	1	2	0.667
PKU-IDM_4 / p-eSur_3	3	0	0	0	3	1

PersonRuns

Participant	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
NHKSTRL_2 / p-NHK-SYS1_1	107	468	15	339	92	0.971
BUPT-PRIS_1 / p-baseline_1	107	39	2	14	105	0.986
NEC-UIUC_2 / c-none_1	107	0	0	0	107	1
UAM_1 / p-baseline_1	107	0	0	0	107	1
NEC-UIUC_2 / p-UI_1	107	157	1	38	106	1.003
Toshiba_1 / p-cohog_1	107	8380	1	176	106	1.048
PKU-IDM_4 / p-eSur_2	107	356	5	351	102	1.068

PersonRuns(continuous)

Participant	#Ref	#Sys	#CorDet	#FA	#Miss	Act. DCR
SJTU_3 / p-baseline_1	107	2217	19	1228	88	1.225
SFU_1 / p-match_1	107	30948	22	3078	85	1.804
TITGT_1 / p-EVAL_1	107	11019	70	10949	37	3.936
TITGT_1 / c-EVAL_1	107	11062	70	10992	37	3.95
CMU_3 / p-VCUBE_1	107	23721	87	23634	20	7.937
BUPT-MCPRL_6 / p-baseline_6	107	25275	78	25197	29	8.534

However, there are some problems yet. Regarding the results, we have reduced the false alarms greatly by effective post-processing. Unfortunately, many correct detections are wiped off at the same time. In other words, the system recall is too low. Furthermore, we should make a good tradeoff between false alarms and system recall.

Another problem is system performance. To accomplish p-eSur_3, six computers that include four 8-core workstations, one 16-core server and one 4-core workstation had been running for more than a week. So, some optimization is necessary in the next year.

7. Conclusion

Event detection in surveillance video becomes a hot research topic of multimedia content analysis nowadays. And TRECVID ED task is a good benchmark evaluation for this topic. The difficulty of the task is due to the corpus, which includes many occlusions, un-uniform illuminations, appearance variations and scale variations. Another difficulty is that objects are too small to detect in some scenes (e.g. in camera 2 and 5).

Although the ACT.DCR of our results is promising, the low system precision indicates that the problem is very challenging and there are large improvement spaces for our system. Whatever, this year's experience shows our system framework is feasible to some extent. Clearly, a great amount of work needs to be done for reaching better results and performance.

References

[1]Ryan Rifkin, Aldebaro Klautau. In Defense of One-Vs-All Classification, Journal of Machine Learning Research 5 (2004) pp.101-141, January, 2004

[2]N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. Coference on Computer Vision and Pattern Recognition (CVPR), 2005.

[3]Stauffer C, Grimson W. E. L. Adaptive background mixture models for real-time tracking. in Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149). IEEE Comput. Soc. Part Vol. 2, 1999.
[4]Elgammal A, Harwood D, Davis L. Non-parametric model for background subtraction. ICCV Frame Rate Workshop, 1999, pp.246-252
[5]K. Kim, T. H. Chalidabhongse, D. Harwood and L. Davis, "Background Modeling and Subtraction by Codebook Construction" IEEE International Conference on Image Processing (ICIP) 2004
[6]D. Gutchess, M. Trajkovic, E. Cohen-Solal, D. Lyons and A. Jain, A background model initialization algorithm for video surveillance, International Conference on Computer Vision (2001), pp. 733–740.
[7]A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. Computer Vision and Pattern Recognition, 2:302–309, 2004.
[8]Oliver, N.M., Rosario, B., Pentland, A.P., A Bayesian computer vision system for modeling human interactions. IEEE Transactions PAMI, 2000, 22(8), pp.831- 843.
[9]Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian Detection Using Wavelet Templates. Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico (1997), pp.193–199
[10]P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. The 9th ICCV, Nice, France, volume 1, pages 734–741, 2003.
[11]Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, Shai Avidan: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. CVPR (2) 2006: 1491-1498
[12]Wu B., Nevatia R. Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors, IJCV(75), No. 2, November 2007, pp. 247-266.
[13]Cheng, Yizong. Mean Shift, Mode Seeking, and Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1995.
[14]Kolmogorov and Zabih. What Energy Functions can be Minimized via Graph Cuts? In: Proceedings of European Conference on Computer Vision. 2002.
[15]H. Grabner, T.T. Nguyen, B. Gruber, H. Bischof. On-line boosting-based car detection from arial images. ISPRS Journal of Photogrammetry & Remote Sencing, 2007,63(3),pp.382-396
[16]C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In ECCV'08, volume 2, pages 788–801
[17]Nguyen, N.T. Phung, D.Q. Venkatesh, S. Bui, H. Learning and detecting activities from movement trajectories using the Hierarchical Hidden Markov model. Computer Vision and Pattern Recognition.2005.
[18]RyooM S, Aggarwal J K. Recognition of composite human activities through context-free grammar based representation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2006: 1709-1718.
[19]Luo Ying, Wu Tzong-der, Hwang Jenq-neng. Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks. Computer Vision and Image Understanding, 2003, 92 (223):196-216.
[20]Zhu Guang-yu, Xu chang-sheng, Gao Wen. Action recognition in broadcast tennis Video using optical flow and support vector machine. In: Proceedings of European Conference on Computer Vision. 2006.
[21]National Institute of Standards and Technology, 2009 TRECVID Event Detection Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/trecvid/2009/doc/EventDet09-EvalPlan-v03.htm, 2009