PKU-NEC @ TRECVid 2011 SED: Sequence-Based Event Detection in Surveillance Video*

Xiaoyu Fang [a], Hongming Zhang [b], Chi Su [a], Teng Xu [a], Feng Wang [b], Shaopeng Tang [b], Ziwei Xia [a], Peixi Peng [a], Guoyi Liu [b], Yaowei Wang [a], Wei Zeng [b], Yonghong Tian [a+]

[a] National Engineering Laboratory for Video Technology, School of EE & CS, Peking University

[b] NEC Laboratories, China

[+] Corresponding author: Phn: +86-10-62758116, E-mail: yhtian@pku.edu.cn

Abstract

In this paper, we describe our system for surveillance event detection task in TRECVid 2011. We focus on pair-wise events (e.g., PeopleMeet, PeopleSplitUp, Embrace) that need to explore the relationship between two active persons, and action-like events (e.g. ObjectPut and Pointing) that need to find the happenings of a person's action. Our team had participated in the TRECVid SED task in 2009 and 2010. This year the new improvements of our system are three-folds. First, we treat object detection and tracking as one problem, and integrate detection and tracking in one unified framework. That is mean "detection by tracking" and "tracking by detection". Also, we fuse multiple trackers to obtain a more accurate tracking result. Experimental results show that our system can achieve a much better precision and recall than our previous systems. Second, we propose sequence learning based method for pair-wise events detection. Visual features are extracted as a cubic feature representation and the discrimination is based on multiple relational and sequence kernels. Experimental results show that our system can detect more correct events with less false alarms. Third, a Markov-model based classifier is employed for action-like event detection. We define some states and learn the transition relation among these states to detect the event. Experimental results show our detectors are feasible and effective. Overall, we have submitted three versions of results, which are obtained by using different human detection, tracking and events detection modules. According to the results in the TRECVid SED formal evaluation, our experimental results are promising.

1. Introduction

This year we chose five events of two classes. One class is pair-wise events (e.g., PeopleMeet, PeopleSplitUp, Embrace) that need to explore the relationship between two active persons, the other is action-like events (e.g. ObjectPut and Pointing) that need to find the happening of a person's action. The diagram of our system is shown in Fig.1.

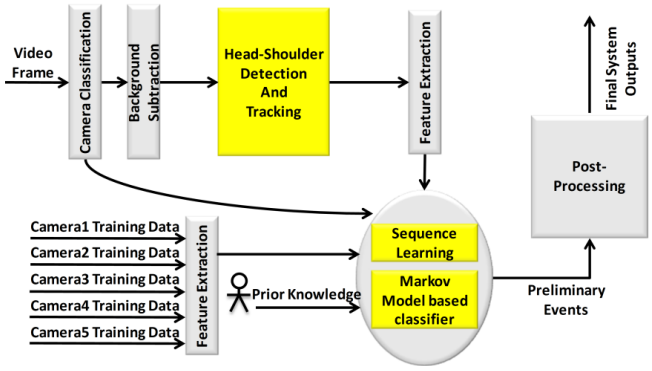

Fig.1 Diagram of our system

Three key improvements are made in the sytem than the 2010 and 2009 systems. First, we treat object detection and tracking as one problem, and integrate detection and tracking in one unified framework. That is mean "detection by tracking" and "tracking by detection". Also, we fuse multiple trackers to obtain a more ac-

* This work was cooperatively done by Peking University and NEC Laboratories, China. This work is partially supported by grants from the Chinese National Natural Science Foundation under contract No. 61035001, No. 60973055, No. 61072095 and No. 61003165, National Basic Research Program of China under contract No. 2009CB320906, and Fok Ying Dong Education Foundation under contract No. 122008. The authors would like to thank Mr. Yingkun Xu for MHT tracking method evaluation when he worked as intern in NEC Labs,China, and thank Dr. Guangyu Zhu for helpful discussion. They also would like to thank Mr. Huang Quan and Ms.Luo Yanlin from NEC Labs, China for large-scale video computing platform.

curate tracking result. Second, As the events videos are inherently sequential data, we propose sequence learning based method for pair-wise events detection. Visual features are extracted as a cubic feature representation. Instead of simply concatenating the features into a vector, we treat them as sequential data to exploit not only the discrete information from individual frames, but also the sequence and correlation information among frames. Therefore, a sequence discriminant learning method based on multiple relational and sequence kernels is employed in our system. Third, a Markov-model based classifier is employed for action-like event detection. We define some states and learn the transition relation among these states to detect the event. Experimental results show our sytem is feasible and effective. According to the results in the TRECVid SED formal evaluation, our experimental results are promising.

The remainder of this paper is organized as follows. In section 2, we describe our head-shoulder detection and tracking approach. In section 3, we present our approach for detecting different events in given surveillance video sequences. Experimental results and analysis are given out in section 4. Finally, we conclude this paper in section 5.

2. Detection and Tracking

2.1 Detection-by-Tracking and Tacking-by-Detection

Pedestrian Detection is an important step in this system. For there are many occlusions in the TRECVid corpus, we apply head-shoulder detection instead of human body detection. Many people in complex scenes will be occluded for a fairly long period. Thus, the human detection in individual frames and data-association of the detection results among several continuous frames are challenging and ambiguous. In [1] and [2], temporal coherency is involved to detection. In our system, we try to exploit temporal coherency by integrate detection and tracking in one unified framework. People-trajectories are extracted from a small number of consecutive frames and from those trajectories build models of the individual people.

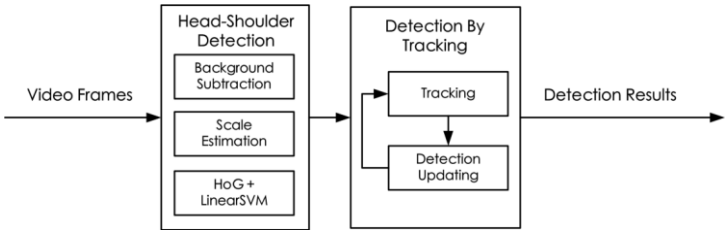

Fig.2 Framework of Detection-by-Tracking

Head-Shoulder Detection

In [3], Dalal and Triggs proved that Histograms of Oriented Gradients are powerful for pedestrian detection. In order to speed up, Zhu et al. [4] combined the cascaded rejection approach with HOG feature. They used AdaBoost to select the best features and constructed the rejection-based cascade.

In our system, we apply a simple and fast method to generate initial detection result. We use HOG feature to represent head-shoulder samples, and apply linear SVM classifier. 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 positions. Thus, the detection process is more efficient.

In practice, we labeled about 5000 head-shoulders as positive training samples, and collected hundreds of images without head-shoulders as the source to extract negative training samples.

Head-Shoulder Detection Update

The final probability of detection $\mathrm{p}(d_N)$ of current frame N will be predicted or updated with the following equation

$$\mathrm{p}(d_N) = w_1 C(d_N) + w_2 S_f(d_N, d_{N-1}) + w_3 S_l(d_N, d_{N-1}),$$

where w_1, w_2, and $w_3$are weights, d_N is the detection in frame N, $C(d_N)$ is confidence of d_N, $S_f(d_N, d_{N-1})$ is the appearance similarity (HOG) of d_N and d_{N-1}, and $S_l(d_N, d_{N-1})$ is the location and scale similarity of d_N and d_{N-1}. $S_l(d_N, d_{N-1})$ is defined by

$$S_l(d_N, d_{N-1}) = p_N\left(\frac{size_N - size_{N-1}}{size_N}\right) \times p_N(|d_N - d_{N-1}|),$$

where $size_N$ is the size of d_N, $p_N\left(\frac{size_N - size_{N-1}}{size_N}\right)$ is the scale similarity of d_N and d_{N-1}, and $p_N(|d_N - dN{-}1$ is the location distance of dN and $dN{-}1$.

We set diffrent weights for different scenes. Head-shoulder detection updating will terminate when the

tracking result change. Then, if the detection results have maximum $\mathrm{p}(d_N)$ and $\mathrm{p}(d_N) > Th$ (Th is the detection threshold) they are appended to the final detection results.

Particle Filter Tracking by Detection

In the TRECVid corpus, target appearance always changes significantly. This year we use a new framework for tracking process as described by Michael D. Breitenstein[5].

Our tracking algorithm is based on estimating the distribution of each target state by a particle filter. We use a constant velocity motion model of each particle [6]. To compute the weight for a particle of the tracker, we estimates the likelihood of each particle. For this purpose, we combine information from different sources, the associated detection score, the preliminary detection results of the detection-by-tracking algorithm metioned in section 2.1, and the classifier outputs.

Considering most of the head-shoulder of pedestrians are small and blurred, we apply the online Multiple Instance Learning algorithm [6] instead of the Online Boosting algorithm in [5]. For each classifier, weak learners are selected using MIL Boost.

2.2 Head-Shoulder Detection Based on Gradient Tree Boosting and Tracking by MHT

We also propose another approach using Gradient Tree Boosting [7] to detect object with high accuracy and fast speed. The essential component of the proposed approach is a cascade Gradient Boosting Tree based object detector, which uses HoG features as object representation. In order to track multiple objects in Trecvid video, we adopt Multiple Hypothesis Tracking (MHT) Method.

Head-Shoulder Detection Based on Gradient Tree Boosting

Fig.3 shows the overall architecture of our object detection approach, which contains training stage and detection stage. The essential component of the proposed approach is a cascade Gradient Boosting Tree based object detector, which uses HoG (Histograms of Oriented Gradients) [4] features as object representation. During training stage, a lot of samples of object and negative images are used to select informative features and to train the object detector. The detection stage is the process to locate object instances in any given input image by using the object detector.

Gradient boosting method was invented by Jerome H. Friedman [8] in 1999 and can be used for classification problems by reducing them to regression with a suitable loss function. In our system, we use decision tree as base learner, and cascade gradient boosting as learning framework.

Multiple Hypothesis Tracking Method

MHT algorithm was invented by Reid [9] in the context of multi-target tracking, and was improved by Cox and Hingorani [10] by an efficient implementation. It uses statistical data association to deal with some tracking issues, such as track initiation, track termination, and track continuation. In our system, head-shoulder detection is incorporated with MHT tracking process to construct one integrated system. For any video, the track results are computed frame by frame. We tested the system on Trecvid dataset. Table 2 shows the evaluation results.

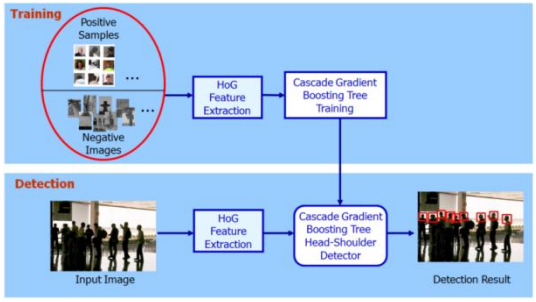

Fig.3 Object detection architecture based on Gradient Tree Boosting

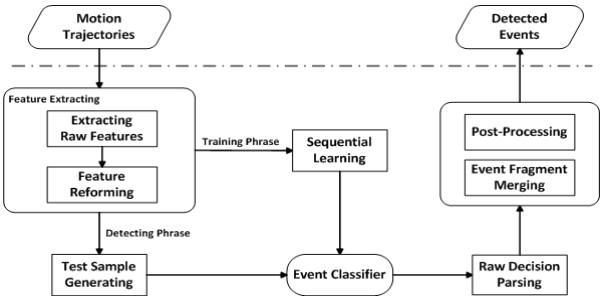

Fig.4 Flowchart of sequential learning based event detection

3. Event detection

3. 1 Pair-wise Event Detection

To detect the pair-wise events in this year's SED task, the interactive events, such as PeopleMeet, PeopleSplitUp, and Embrace, are considered as a time-variant holistic pattern, and spatio-temporal cubic feature and sequence discriminant learning method are introduced to serve the detection task.

The discriminative patterns for these three events in video sequences are inherently time sequential. How-

ever, most pervious activity recognition methods did not handle this properly with only modeling the patterns in single frames or simply concatenating them together. In our solution, the event is considered as a whole sequence and described by the spatio-temporal cubic feature. Specifically, we employ Support Vector Machine with dynamic time alignment kernel proposed in [11]. This method handles time series feature with varying length and the learning procedure is based on a maximum margin criterion. With the sequence discriminant learning method, the temporal correlations between different stages of the event are properly considered, and decisions based on integrated event sequences are reliable and semantically reasonable.

As shown in Fig.4, features are extracted based on the motion trajectories generated by human detecting and tracking module mentioned in previous sections. We first segment video sequences into several cubes, and then, according to the locations of every person in a frame, we calculate the mean absolute velocity, acceleration, distance between each pair of people and the angular separation of moving directions in each cube as the raw features. Then the extracted raw features from the same video clips (ground truth event samples for training and test samples for detecting) are transformed to structural sequence feature. Some statistics of raw features are also included into the reformed features to explicitly employ the information of the temporal dependencies over adjacent frames.

With the structural features, an appropriate implementation of SVM with dynamic time alignment kernel [8], is applied to train events classifiers and make decisions. As the raw decision is a sequence of binary decisions for each frame in a testing sample, we need to parse it into a single decision for the testing sample with the strategy like voting. As the detection task is actually transformed to a classification problem by using sliding window method to generate testing samples, the original results would be fragmental. So in the post-processing phrase, we merge the preliminary detections and introduce some prior knowledge based rules to filter out incredible detections. These rules are usually empirical restrictions such as a distance threshold between persons before “PeopleSplitUp” or after “PeopleMeet”.

3. 2 Action-like Event Detection

To detect “ObjectPut” and “Pointing”, a Markov-model based classifier is employed for action-like event detection. We first define some states and learn the transition relation among these states. Then a state transition model is constructed for each event. Base on the tracking results of objects, we use histogram of optical flow (HOF) for “ObjectPut” and MoSift for “Pointing” to represent their motions, which will cause transition of their states. Therefore, action-like events are recognized by classifying objects’ state transition process with their models.

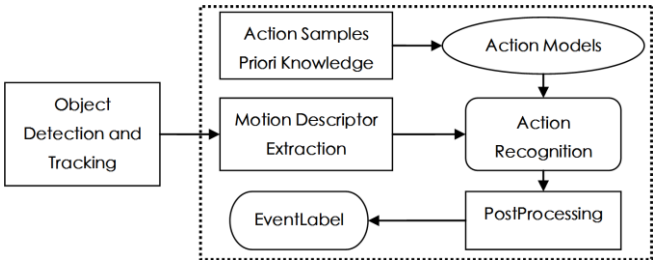

Fig.5 Action-like events detection

4. Experiment and results

Our team submitted three versions of results, which are obtained by using different human detection, tracking and events detection modules.

Table 1 Head-shoulder detection results of this year and last year

Camera1	Recall	Precision	F-score	Camera2	Recall	Precision	F-score
Last Year-SVM	0.511	0.832	0.6331	Last Year-SVM	0.373	0.615	0.4644
Last Year-MPL	0.539	0.796	0.6429	Last Year-MPL	0.560	0.773	0.6495
This Year-GTB	0.553	0.803	0.6550	This Year-GTB	0.356	0.727	0.4780
This Year-SVM	0.557	0.848	0.6724	This Year-SVM	0.372	0.785	0.5048
Camera3				**Camera5**			
Last Year-SVM	0.403	0.713	0.5149	Last Year-SVM	0.265	0.613	0.3700
Last Year-MPL	0.429	0.667	0.5222	Last Year-MPL	0.468	0.757	0.5783
This Year-GTB	0.294	0.801	0.4301	This Year-GTB	0.271	0.732	0.3755
This Year-SVM	0.423	0.756	0.5425	This Year-SVM	0.318	0.775	0.4510

Table 2 Tracking results of this year and last year

Camera1	MOTA	MOTP	Miss	FA	ID Switch
Last Year	0.321	0.591	0.510	0.134	0.035
This Year-MHT	0.368	0.571	0.486	0.134	0.012
This Year-PFT	0.364	0.567	0.472	0.154	0.010
Camera2					
Last Year	-0.135	0.599	0.791	0.317	0.027
This Year-MHT	0.151	0.601	0.680	0.160	0.009
This Year -PFT	0.213	0.607	0.644	0.132	0.011
Camera3					
Last Year	0.022	0.571	0.652	0.293	0.033
This Year-MHT	0.198	0.583	0.746	0.051	0.005
This Year-PFT	0.271	0.591	0.667	0.050	0.010
Camera5					
Last Year	-0.002	0.602	0.537	0.440	0.025
This Year-MHT	0.168	0.591	0.737	0.088	0.008
This Year-PFT	0.170	0.589	0.731	0.089	0.009

Table 1 and 2 show the comparison detection and tracking results between the best outputs of our system this year and those of last year. It can be seen from the tables that detection result is improved greatly in recall with low or no decrease in the precision. Here we introduce Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) [8], metrics used in PETS 2009, to evaluate overall performance. These ID switches used in MOTA are calculated from the number of identity mismatches in a frame, from the mapped objects in its preceding frame. The MOTP is calculated from the spatiotemporal overlap between the ground truth tracks and the algorithm's output tracks. Conclusion can be drawn from table 2 that our performance is improved greatly.

According to the results in the TRECVid SED formal evaluation, our experimental results are promising this year, especially for the events PeopleMeet and Embrace. Table 3 shows the comparison results between the best outputs of our system this year and those of last year. It can be seen from the table that our eSur system is greatly improved by detecting more correct events. The number of correctly detected PeopleMeet and Embrace events is two times more than last year. Meanwhile, the false alarms do not rise too much and even dramatically decreased for PeopleMeet. Table 3 also shows results of ObjectPut and Pointing detection, which we participant and submit results for the first time this year. The correctly detected number of ObjectPut and Pointing is more than that of best results of last year, and DCR of our ObjectPut is even lower; and DCR of our Pointing is also comparable with the best of last year.

Table 3 Comparison results between the best outputs of eSur this year and last year

PeopleMeet	#Ref	#Sys	#CorDet	#FA	#Miss	Act.DCR
2010's eSur	449	156	12	144	437	1.02
2011's eSur	449	2382	24	108	425	0.9820
PeopleSplitUp						
2010's eSur	187	167	16	136	171	0.959
2011's eSur	187	2988	4	192	183	1.0416
Embrace						
2010's eSur	175	925	6	71	169	0.989
2011's eSur	175	5234	15	102	160	0.9477
ObjectPut						
2010's Best	621	8	1	7	620	1.001
2011's eSur	621	50	8	41	613	1.0006
Pointing						
2010's Best	1063	113	10	26	1053	0.999
2011's eSur	1063	2113	21	123	1042	1.0206

5. Conclusion

This year we improved our system significantly in head-shoulder detection and tracking where unified framework is employed and event detection where sequence discriminant learning method is used for pair-wise events detection and Markov-Model based classifier is used for the action-like event detection. The promising results of our system this year verify the effectiveness of these improvements. However, we believe there are still large improvement spaces for our system in exploring more effective and descriptive event models.

Reference

[1] Zhipeng Hu, Yaowei Wang, Yonghong Tian, Tiejun Huang, Selective Eigenbackgrounds Method for Background Subtraction in Crowed Scenes. ICIP 2010

[2]M. Andriluka, S. Roth, B. Schiele. People-tracking-by-detection and people-detection-by-tracking. Conference on Computer Vision and Pattern Recognition (CVPR), Page(s): 1–8, 2008.

[3] A. Garcia-Martin, A. Hauptmann, J.M. Martinez: People detection based on appearance and motion models. Advanced Video and Signal-Based Surveillance (AVSS), Page(s): 256 – 260, 2011

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

[5]Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, ShaiAvidan: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. CVPR (2) 2006: 1491-1498

[6] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, Luc Van Gool. Online Multi-Person Tracking-by-Detection from a Single, Uncalibrated Camera. PAMI, 2010.

[7]Yasemin Altun, Loannis Tsochantaridis and Thomas Hofmann. Hidden Markov Support Vector Machines. ICML, 2003.

[8] J. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Statist. 29(5), 2001, 1189-1232.

[9] D. Reid, An algorithm for tracking multiple targets, IEEE Transactions on Automatic Control, Volume: 24, Issue: 6, 843 – 854, 1979

[10] I.J. Cox, S.L. Hingorani, An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 18, Issue: 2, 138 – 150, 1996

[11] H. Shimodaira, et al, Dynamic Time-Alignment Kernel in Support Vector Machine, Proc. Advances in Neural Information Processing Systems, 14, vol.2, pp.921-928, 2001.