TRECVID 2008 Event Detection By MCG-ICT-CAS*

Junbo Guo, Anan Liu, Yan Song, Zhineng Chen, Lin Pang, Hongtao Xie, Leigang Zhang

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

{guojunbo, liuanan, songyan, chenzhineng, panglin, xiehongtao, zhangleigang}@ict.ac.cn

ABSTRACT

As for Event Detection in TRECVID 2008, we develop a surveillance system with two parts, the trajectory-based sub-system and the domain knowledge-based sub-system. The former focuses on the research of general methods for event discovery. Human detection and tracking are utilized to generate the trajectory and then novel three-level trajectory features are proposed to detect PersonRuns, PeopleMeet, PeopleSpiltUp, and Embrace. The latter focuses on the study of specified models to improve the results. Based on domain knowledge, three models are respectively constructed for PeopleMeet, Opposingflow, and ElevatorNoEntry. The results are separately shown in the submitted results, "MCG-ICT-CAS_2008_retroED_EVAL08_ENG_s-camera_p-baseline_1" and "MCG-ICT-CAS_2008_retroED_EVAL08_ENG_s-camera_p-Run2_1".

Keywords

Event detection, Surveillance, Trajectory, Domain Knowledge, Human Detection, Tracking.

1. Introduction

Event detection in video surveillance is very important for some public environments (e.g. communities, airport and shopping centers, etc.). However, large amount of video data in surveillance makes it an exhausting work for people to keep watching and finding abnormal events. Therefore, automatic event detection is urgently needed to make the objective, reliable and repeatable decision.

Event detection has been an active research field in recent years. There are mainly two kinds of methods. One is the fundamental method [1] consisting of human detection, tracking and behavior understanding. The current research on the three key problems [2-4] is usually separated and condition-constrained. Therefore, the algorithms are difficult to be implemented in the practical application ideally. The other constructs specific model for the event with spatio-temporal features and detects the event in the video volume [2]. Although machine learning methods [5-6] are widely used to improve the generalization, it is difficult to get a perfect model because of the diversity of patterns.

Since the surveillance video for this task is captured from airport. It is unconstrained and has the characteristics such as highly clutter, massive population flow, heavy occlusion and so on, we find that typical machine learning methods are unsuitable in this situation. As for this practical application, we develop a video surveillance system with two parts, the trajectory-based sub-system and the domain knowledge- based sub-system. The first one implements human detection and tracking to generate trajectory and three-level trajectory features are used to detect PersonRuns, PeopleMeet, PeopleSpiltUp and Embrace. The second one constructs specific models for PeopleMeet, Opposingflow, and ElevatorNoEntry depending on domain knowledge. Therefore, in our exploration for Event Detction in TRECVid 2008, we focus on both generality and specificity to develop a prototype system for video surveillance.

The remainder of the paper is organized as follows. We specifically present the trajectory-based sub-system and the domain knowledge- based sub-system in Section 2 and 3. The experimental results are presented in Section 4.

2. Trajectory-based Sub-system

In this section, we illustrate the trajectory-based sub-system for event detection in video surveillance in details.

2.1 Preprocessing

For each video, we only extracted I and P frame considering both the redundancy in temporal domain and low computational cost. The background subtraction algorithm and morphological operations followed by consist the preprocessing step, only keep those regions that contain more than 30 pixels.

2.2 Human-Detection

The cascade boosting object detection framework in [7] is used for human detection. Specifically, we independently

* This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), and National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), and National Nature Science Foundation of China (60873165, 60773056).

train two detectors, full-body and head-shoulder detectors using standard haar-like features. The detection result is made by joint decision of both two detectors. The training data is set as follows. For full-body detector, positive samples are public training data released by DCU, where 3749 people are labeled from 815 images. For head-shoulder detector, positive samples are 3000 frames manually labeled by our team, where 3140 head-shoulders, including frontal, rear and side views, are annotated. The negative samples for both two detectors are manually labeled by our team, which consists of 273 frames without human, collected from the training corpus.

2.3 Human-Tracking

We have tried several state-of-the-art tracking algorithms. Since occlusions happen frequently in limited camera scope, Particle filtering [8] achieves the best performance. Unfortunately, particle filtering is a time-consuming process, especially when the object tracked is large. It is difficult to complete the test on evaluation data within the limited time. Therefore, we adopt the data correlation method with the visual features of the center and color histogram of the detected bounding box.

2.4 Event Detection

It is known that various features can be directly extracted from the trajectory. Then, we proposed a three-level trajectory features for event discovery. From bottom to top, they are individual feature, two-person feature, and crowd feature, as depicted in Fig. 1.

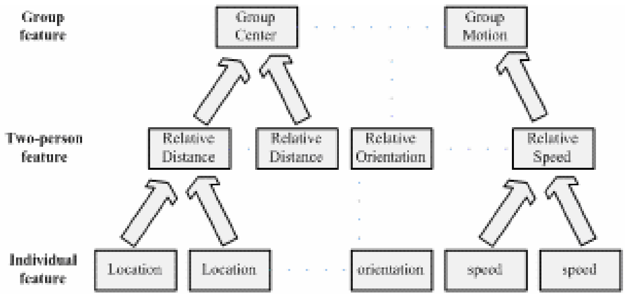

Fig. 1. The three-level hierarchical feature architecture

The individual features are speed and orientation directly extracted from one trajectory. The two-person features are relative speed, relative distance, relative orientation calculated between two persons in the same frame. Crowd features are combinations of two-person features, e.g. the center point of a group of people.

With these features and trajectory information, we set different rules to detect events, including PersonRuns, PeopleMeet, PeopleSpiltUp, and Embrace, as follows:

1) **PersonRuns**: Three types of speed, namely speed between people in two, three and four consecutive frames are extracted from each trajectory. We set three experienced thresholds respectively. Each speed exceeding its corresponding threshold is considered as an available speed. The number of available speed, as well as the corresponding trajectory points is recorded. The decision is made by jointly considering the trajectory length, location of trajectory points and percent of available speed.

2) **PeopleMeet**. PeopleMeet is detected using rules: a): Calculate relative distances between any two individuals in each frame. If there exist a relative distance smaller than a given threshold “D_t”, go to step b), otherwise, go to the next frame; b): If two persons appear in the following frames and satisfying rules: their distance decreases continuously, meanwhile, their relative orientation and speed are in reasonable range, which is represented as certain predefined constraints. We decide that the PeopleMeet occures. We go to step a) if relative distance of persons exceeds “D_t” in the above process. Since multi-person meet can be decomposed as some two-person meets, its start time and end time can be determined accordingly.

3) **PeopleSplitUp**. As PeopleSplitUp happens when one or more person separate from a group, our method consists of the following four steps: a): Detect the number of crowds in a frame, using a distance threshold “D_g”; b): Compute and update each crowd center in consecutive frames, recorde the number of frame that each person belongs to a specific crowd, which is called living-time; c): A person is decided to leave the corresponding crowd if the relative distance between him (her) and the crowd center is larger than “D_g”, and the living-time of the person is longer than a time threshold “T_g”; d): Track every person belonging to the seperated crowd and PeopleSplitUp event is detected only if there is at least one person coming off the frame.

4) **Embrace**. According to our observation, a large portion of Embrace events happen immediately after PeopleMeet. Therefore we use the trajectory location, relative distance and speed to detect Embrace as follows: Calculating relative distance and speed after PeopleMeet. Embrace is detected when relative distance and speed are respectively below given thresholds. Meanwhile, the trajectory locations of meet persons are nearly unchanged.

3 Domain Knowledge- based Sub-system

In the domain knowledge-based sub-system, we construct three specific methods for ElevatorNoEntry, OpposingFlow and PeopleMeet respectively.

1) **ElevatorNoEntry**. It is obvious that ElevatorNoEntry is related with both the state of the elevator door and the appearance of human. Therefore, we design one detector for the period of door open and close and another for human appearance. Because the elevator doors correspond to fixed regions in the frames and some specific regions change significantly during the period of door open and close, both periods can be detected with the changes of foreground in the door regions. Dynamic background construction and foreground segmentation [9] are adopted here to detect both periods. As for the period between door open and close, we implements human detection and tracking for door region

simultaneously. If the person exists during this period, the event of ElevatorNoEntry occurs.

2) **OpposingFlow**. We detect OpposingFlow as follows: a): Optical flow is calculated depending on Lucas-Kanade algorithm in [10] on a set of densely detected Harris corners, which is derived as low-level features with the post-processing of Guassian smoothing and de-noising; b): Orientation histogram of optical flow in the door region is calculated to represent the statistical feature of optical flow amplitude of corner points. If the value in the bin of reverse direction is over the pre-setted threshold, we mark this frame as a candidate frame; c): To avoid false detection, human detection is implemented in the candidate frame and human tracking is used foreward and backward. d): The candidate is decided to be positive only if the person in current region can be tracked back to last N frames and the trajectory spans over the inside and outside of the door.

3) **PeopleMeet**. Note that both the camera is fixed and the probability of PeopleMeet is varying with regions, we improve PeopleMeet detections by adding a post-process step to trajectory-based sub-system results that gives more weight to some regions containing more people activities when calculate the detectionscore.

Table 1 .Results of Baseline

Events	Ref	Act.P miss	Act.RFA	Recall (%)	Precision(%)	Act.D CR	Min.D CR
PersonRuns	314	0.9268	12.5043	7.32	3.474	0.9893	0.9724
PeopleMeet	1182	0.5000	239.5783	50.00	4.605	1.6979	1.0067
PeopleSpiltUp	671	0.5142	178.6417	48.58	3.4479	1.4074	0.9981
Embrace	401	0.8279	84.7907	17.21	1.567	1.2519	0.9993

Table 2.Results of Run2

Events	Ref	Act.P miss	Act.RFA	Recall (%)	Precision	Act.D CR	Min .D CR
ElevatorNoEntry	0	NA	0.1174	NA	0	NA	NA
OpposingFlow	12	0.4167	2.8962	58.33	4.516	0.4311	0.4307
PeopleMeet	1182	0.5964	180.8725	40.36	4.907	1.5008	1.0094

4 Experimental Results

The results of trajectory-based sub-system are considered as the baseline shown in Table 1 and domain knowledge-based sub-system are regarded as Run2 shown in Table 2.

From Table 1 we can see that recall is acceptable and precision is a little low in baseline. The reasons maybe lie in two aspects: 1) The surveillance video is in unconstrained condition and therefore the trajectory features can not perfectly represent the events; 2) The accumulation of errors in human detection, tracking and event detection can have great influence on the final decision. Besides, it is seen that the recalls of PeopleMeet and PeopleSpiltUp is higher than those of PersonRuns and Embrace. It is because the definitions of the former two are more clear and the rules are more robust.

From Table 2, we can see our specific model works well for OppositingFlow. Although the precision is low, the recall and DCR show that our method is effective. The ElevatorNoEntry result is difficult to analyze since there is no reference event annotation, however, our dryrun result as well as our test results on development corpus show that our model is effective. As for PeopleMeet, it is natural that the recall gets lower than that in trajectory-based sub-system, however, the precision only increased from 4.6% to 4.9%, which is lower than our expectation. The possible reason is that giving more weight to the regions containing more people activities also increase the probability of discarding the true detections in other places. To solution this problem, more complicated domain knowledge based rules is necessary.

5 REFERENCES

[1] I. Haritaoglu, D. Harwood, and L. S. Davis, "W: Real-time surveillance of people and their activities," IEEE Trans. Pattern Anal. Machine Intell., vol. 22, pp. 809–830, Aug. 2000.

[2] Actions as Space-Time Shapes. Blank M., Gorelick L., Shechtman E., Irani M., Basri, R. Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)Volume 2, 17-21 Oct. 2005 Page(s):1395 –1402.

[3] Human detection based on a probabilistic assembly of robust part detectors. K Mikolajczyk, C Schmid, A Zisserman. In Proc. of ECCV, volume 1, pages 69–82, 2004.

[4] Detection and Tracking of Humans by Probabilistic Body Part Assembly. A Micilotta, E Ong, R Bowden. British Machine Vision Conference, Oxford, UK, Sep 2005.

[5] A HMM based semantic analysis framework for sports game event detection. Gu Xu Yu-Fei Ma Hong-Jiang Zhang Shiqiang Yang Proceedings of the IEEE ICIP 2003 Volume: 1, On page(s): 25-28

[6] S Park, JK Aggarwal. A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Systems, vol.10, issue.2, papes: 164--179 2004.

[7] Voila P, Jones M. Rapid object detection using adaboosted cascade of simple features. In Proc of IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA, pages:511-518, 2001

[8] Arulampalam M S, Maskell S, Gordon N, Clapp T. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing, vol.50, issue.2, pages: 174-188, 2002.

[9] A.Monnet, A. Mittal, N. Paragios, Visvanathan R. Background modeling and subtraction of dynamic scenes. In the proceeding of ICCV, pages: 1305-1312 vol.2, 2003.

[10] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.