# Interactive Event Detection in Crowd Scenes

Lei Qin
Key Lab of Intel. Inf. Proc., Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
lqin@jdl.ac.cn

Zhongwei Cheng
Graduate University of Chinese Academy of Sciences, Beijing, China
zwcheng@jdl.ac.cn

Qingming Huang
Graduate University of Chinese Academy of Sciences, Beijing, China
qmhuang@jdl.ac.cn

Junbiao Pang
Beijing Municipal Key Lab. of Multimedia and Intelligent Software Technology, Beijing Univ. of Tech., China
jbpang@jdl.ac.cn

## ABSTRACT

As an important aspect in video content analysis, event detection is still an open problem. In particular, the study on detecting interactive events in crowd scenes is still limited. In this paper, we investigate detecting interactive events between persons, e.g. PeopleMeet, PeopleSplitUp and Embrace in complex scenes using a sequence learning based approach. By sequence learning, the spatial-temporal context information is introduced in the learning stage. Experiments have been performed over TRECVid Event Detection 2010 dataset, which contains totally 144 hours surveillance video of London Gatwick airport. According to the TRECVid-ED 2010 formal evaluation, our approach obtains promising results, with the top performance (NDCR) for PeopleMeet and PeopleSplit-Up, and second-best performance (NDCR) for Embrace.

## Categories and Subject Descriptors

I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding – *Video analysis*

## General Terms

Algorithms, Design, Experimentation

## Keywords

Sequence learning, events detection, TRECVid.

## 1. INTRODUCTION

The analysis of events in videos is important for understanding the semantic content of videos. Therefore, many event based video analysis approaches have been reported in the literatures [1]. But the majority of previous studies focus on single person's action. Feature extraction is an important step for these methods. There are two types of widely-used low level features for action recognition. One is spatial-temporal interest points, such as 3D Harris detector[2][3] or separable linear filters[4]. Another is global shape feature, as Lin et al. [5]proposed a shape descriptor based on foreground segmentation, and Gorelick et al. [6]regarded human actions as three-dimensional shapes induced by the silhouettes in the space-time volume.

*ICIMCS '12,* September 09 - 11 2012, Wuhan, China
Copyright 2012 ACM 978-1-4503-1600-2/12/09…$15.00.

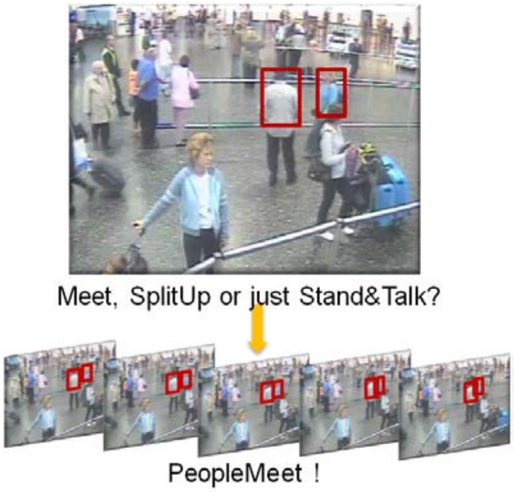

**Figure 1. It is easy to recognize interactive event in sequence level, while difficult in frame level. From the top image, we don't know whether the event is PeopleMeet, PeopleSplitUp or just Stand&Talk. However, from the bottom image sequence, we can judge the event is PeopleMeet.**

In this paper we focus on another event category, long-lasting interactive events in surveillance environment (e.g., PeopleMeet, PeopleSplitUp, Embrace), which needs to explore the relationship between active persons. In interactive event, the interactions between people are time variant holistic pattern. For the characteristic of comparative long duration and time variant, it is better to recognize the interactive events in sequence level than in frame level, as shown in Figure 1.

In typical surveillance video, due to the clustered background and the serious occlusion, the appearance features for classic action recognition are not reliable. Nevertheless, common interactive events, with overhead view and proper subject size, can be represented properly by the motion trajectories of people [7][8][9][10]. Ni et al. [9]proposed a promising approach based on trajectories. They applied filters on trajectories and took the responses as features for recognition. Wang et al. [7]extracted features such as position, velocity and motion direction from trajectories. However, these methods do not appropriately seize the sequential property of events. Therefore, in this paper, we introduce the sequence learning for event detection, using structural classifier to model the activity as sequence structure and exploring dynamics of the pattern within an event. The most similar work to ours is [8], but its target is activity retrieval. In addition to the need of query sequence, it is not easy to generalize [8] to model complex events in wild videos.

The remainder of the paper is organized as follows. In section 2, we present the proposed interactive event detection approach. In section 3, we provide the experimental results. Finally, we conclude this paper in section 4.

## 2. The approach

We propose an approach for the interactive events detection in this paper. The interactive events, such as PeopleMeet, PeopleSplitUp and Embrace, are considered as time-variant holistic patterns. Proper sequential model and structural classifier are introduced to serve the detection task.

Our approach consists of two stages: 1) feature extracting stage, 2) classification stage. In the feature extracting stage, features are extracted from a video fragment, which are based on the positions and motion trajectories of the objects. In the classification stage, the classifier is trained using the features and labels of video fragments. And the classifier outputs a label for every frame of a test video fragment. Then the output label sequence is transformed to a segment decision with voting. The flowchart of our approach is illustrated in Figure 2.

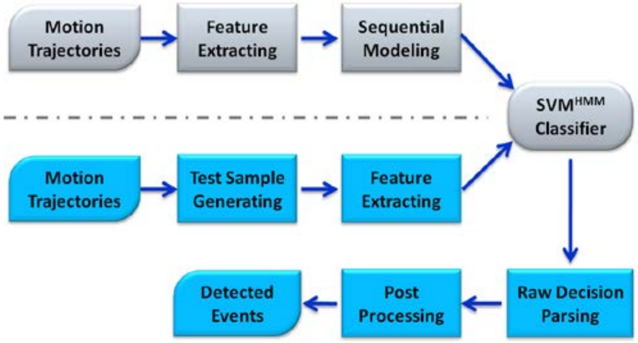

**Figure 2. The framework of our approach**

### 2.1 Feature extracting stage

The feature extracting stage consists of three elementary steps: pedestrian detection, tracking and visual features extraction. As there are many occlusions in the surveillance videos, such as TRECVid dataset[11], parts or even the whole body of the pedestrians are frequently invisible. For this reason, we apply head-shoulder detection instead of human body detection. We use HOG (Histogram of gradient) feature[12] to represent head-shoulder samples, and apply Multiple Pose Learning based RealBoost[13] to improve the performance of pedestrian detection. On the basis of detection results, an online boosting method[14] is employed to track the moving objects. Finally, features are extracted based on the motion trajectories generated by human detecting and tracking.

According to the locations of every person in frame *k-1* and frame *k*, we calculate the absolute velocity $V_{p_i}^k$ and the acceleration $A_{p_i}^k$ of person $p_i$, the distance and the angular separation of moving directions between each pair of people. The distance between two persons $p_i$ and $p_j$ in frame $k$ is measured by their Euclidean distance as (1), where *pos* is a person's centroid coordinate. The angular separation of two persons' moving directions is described in (2), where $\theta$ is the angle between a person's motion direction and horizontal axis.

$$Dist^k(p_i, p_j) = \left\| pos_{p_i}^k - pos_{p_j}^k \right\| \tag{1}$$

$$Ang^k(p_i, p_j) = \left\{ m \,\middle|\, \left| \theta_{p_i}^k - \theta_{p_j}^k \right| \in \left[ (m-1) * \frac{\pi}{4}, m * \frac{\pi}{4} \right), m \in Z^+ \right\} \tag{2}$$

These raw features describe the basic information of event. Then the features from the same video clips are transformed to structural sequence feature as follow.

$$SequenceFeatures = \{x_k \mid k = 1, \ldots, n\} \tag{3}$$

$$x_k = \{V_{p_i}^k, A_{p_i}^k, Dist^k(p_i, p_j), Ang^k(p_i, p_j)\} \tag{4}$$

The features are normalized separately before concatenation.

Furthermore, some statistics of raw features are also included into the reformed sequence features to explicitly employ the information of the temporal dependencies within a video fragment. The statistics features include the mean, variation and trend of distance between persons. The trend of distance between persons in a video fragment is defined as (5),

$$MD = \frac{1}{n-1} \sum_{k=1}^{n-1} \frac{1}{\overline{Dist}} (Dist^k - Dist^{k+1}) \tag{5}$$

where $n$ is the number of frames in the fragment, and $\overline{Dist}$ is the mean distance in the fragment.

### 2.2 Learning Method

The interactive events that we consider in this paper are time-variant and holistic. It is comprehensible that the discriminative patterns for these events in video sequences are inherently time sequential. By modeling the temporal sequential dependency within events, we expect to obtain more discriminative model. In our solution, the event is described by the stochastic sequential model and classified using structural support vector machines. Specifically, we employ the Markov Support Vector Machine proposed in[15][16], which is an extension of the SVM for sequential tagging. This model combines the advantage of SVM and HMM by discriminatively training models that are similar to hidden Markov models. In the following, we use SVM$^{hmm}$ to represent this model. SVM$^{hmm}$ handles dependencies between neighboring frames using Viterbi-like decoding and the learning procedure is based on a maximum margin criterion. With SVM$^{hmm}$, the temporal correlations between different stages of the event are properly considered, and decisions based on integrated event sequences are reliable and semantically reasonable.

We utilize a first-order transition in SVM$^{hmm}$. Given an input sequence feature vectors $F$ = ($x_1$ … $x_n$), SVM$^{hmm}$ predicts a sequence of labels $y$ = ($y_1$ … $y_n$) according to the following linear discriminant function[9]:

$$y = \arg\max_y \left\{ \sum_{i=1}^{n} \left[ \sum_{j=1}^{k} (x_i \bullet w_{y_{i-j},\ldots,y_i}) + \Phi_{trans}(y_{i-j},..,y_i) \bullet w_{trans} \right] \right\} \tag{6}$$

where $w_{yi}$ is an emission weight vector for the label $y_i$ and $w_{trans}$ is a transition weight vector for the transition between the label $y_{i-1}$ and $y_i$.

We train SVM$^{hmm}$ model from training samples for each event category and make sequence decisions for testing samples. As the raw decision is a sequence of label decisions for each frame in a testing sample, we need to parse it into a single category decision for the testing sample with some strategy, such as voting. There is an example shown in Figure 3.

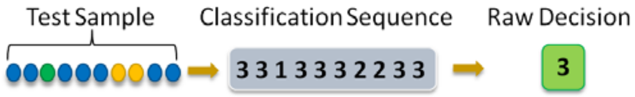

**Figure 3. Divide videos for detection into test samples using sliding fragment strategy. Sequential results are generated by SVM$^{hmm}$ classifiers, then transformed to raw decision with voting. Numbers stand for event class labels.**

As the detection task is actually transformed to a classification problem by using sliding fragment method to generate testing samples, the original results would be fragmental, and may contain false alarms. So in the post-processing phrase, we merge the preliminary detections and introduce some prior knowledge based rules to filter out improbable detections. A so-called “Events Merging” process deals with those fragmental events occurring in an overlapping time span. The other post-processing processes are applied to address several other forms of false alarms. For example, at the end of a PeopleMeet event, persons should not keep moving, and the distance between them should not be beyond a certain threshold.

# 3. Experiment results

To demonstrate the effectiveness of the proposed approach, we perform experiments on the TRECVid Event Detection dataset. The TRECVid-ED dataset is obtained from the Gatwick Airport which consists of 100 camera hours of video in the development set and 44 camera hours of video in the evaluation set. The three types of interactive events in the TRECVid dataset that we are interested in are: PeopleMeet, PeopleSplitUp, and Embrace. The description of these events can be found in[11]. There are about 190K frames in each video of the development set. The frame resolution is 720×576. NIST provides the preliminary annotations of the occurrences of events in the development set. We further label the precise locations of persons performing the actions with bounding boxes. Some samples of the 3 events are shown in Figure 4. We can see that there are large intra-class variations in each event.

With the detected event sequences, the TRECVid event detection task evaluates the performance by NDCR which is the normalized weighted linear combination of the missed detection probability and the false alarm rate. NDCR can be calculated by (7).

$$NDCR = \frac{N_{Miss}}{N_{Target}} + \frac{N_{FA}}{T_{Source}} \times \frac{Cost_{FA}}{Cost_{Miss} \times R_{Target}} \qquad (7)$$

where $N_{Miss}$ is the number of missd events, $N_{Target}$ is the system outputs number of detected events, $N_{FA}$ is the number of false alarms, $T_{Source}$ is the total duration of the evaluation video in hours, and $Cost_{FA}$, $Cost_{Miss}$, $R_{Target}$ are constants defined by[11], with the values of 1, 10 and 20, respectively. A smaller NDCR means better performance.

## 3.1 Evaluation on TRECVid 2008 dataset

NIST provides the ground truth of TRECVid 2008 dataset. So we carry out some experiments on 10 hour video data of TRECVid 2008 dataset to test the validity of our approach. The method of Wang et al. gains the best results on interactive events detection task in TRECVid 2009 evaluation[7]. So we choose[7] as the state-of-the-art method, and compare the detection results of our method with those in[7]. The results are provided in Table 1 to Table 3. In the tables, #Ref stands for the number of ground truth events, #Sys is the number of system detected events, #CorDet is the number of correct detections, and #FA is the number of false alarms.

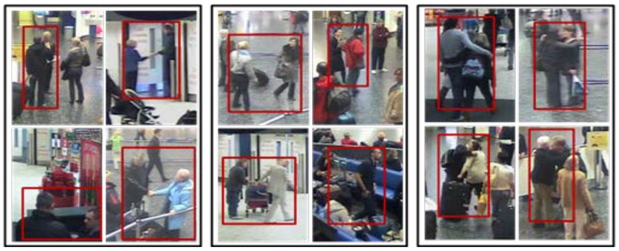
(a) Samples of PeopleMeet (b) Samples of PeopleSplitUp (c) Samples of Embrace

**Figure 4. Samples of the PeopleMeet, PeopleSplitUp, Embrace events in the TRECVid dataset**

**Table 1. The detection results of PeopleMeet event**

| Methods | #Ref | #Sys | #CorDet | #FA | NDCR |
|---|---|---|---|---|---|
| Our approach | 116 | 82 | 5 | 77 | 0.995 |
| [7] | 116 | 44 | 1 | 43 | 1.000 |

**Table 2. The detection results of PeopleSplitUp event**

| Methods | #Ref | #Sys | #CorDet | #FA | NDCR |
|---|---|---|---|---|---|
| Our approach | 298 | 54 | 7 | 47 | 1.000 |
| [7] | 298 | 29 | 2 | 27 | 1.007 |

**Table 3. The detection results of Embrace event**

| Methods | #Ref | #Sys | #CorDet | #FA | NDCR |
|---|---|---|---|---|---|
| Our approach | 152 | 81 | 7 | 74 | 0.991 |
| [7] | 152 | 21 | 0 | 21 | 1.011 |

From the experimental results, we can notice that #CorDets of both methods are relatively low. This is because that TRECVid dataset is too difficult. There are a large number of people, serious occlusion, large intra-class variations. However our approach obtains performance improvement on all the three events. Furthermore, our approach can obtain more #CorDet than that in [7]. These experimental results show that the spatial-temporal context information can help to improve the detection performance.

## 3.2 TRECVid-ED 2010 evaluation results

We took part in the TRECVid-ED 2010 with the proposed method. According to the TRECVid-ED formal evaluation, our approach achieves promising results[11], with the best NDCR of 1.02, 0.959 for PeopleMeet and PeopleSplitUp, and second best NDCR of 0.989 for Embrace as listed in Table 4 to Table 6. Systems with 0 correct detection are excluded. A smaller NDCR means better performance. The results also show that the sequence learning method can help to improve the interactive event detection performance.

**Table 4. The detection results of PeopleMeet event**

| | #Ref | #Sys | #CorDet | #FA | NDCR |
|---|---|---|---|---|---|
| **Our method** | **449** | **156** | **12** | **144** | **1.02** |
| CMU_2 / p-VCUBE_10 | 449 | 305 | 24 | 281 | 1.039 |
| CMU_2 / p-VCUBE_9 | 449 | 388 | 27 | 361 | 1.058 |

**Table 5. The detection results of PeopleSplitUp event**

| | #Ref | #Sys | #CorDet | #FA | NDCR |
|---|---|---|---|---|---|
| **Our method** | **187** | **167** | **16** | **136** | **0.959** |
| CMU_2 / p-VCUBE_8 | 187 | 281 | 17 | 264 | 0.996 |
| CMU_2 / p-VCUBE_9 | 187 | 42 | 3 | 39 | 0.997 |

**Table 6. The detection results of Embrace event**

| | #Ref | #Sys | #CorDet | #FA | NDCR |
|---|---|---|---|---|---|
| IPG-BJTU_5 / p-SYS_1 | 175 | 64 | 9 | 55 | 0.967 |
| **Our method** | **175** | **925** | **6** | **71** | **0.989** |
| CMU_2 / p-VCUBE_9 | 175 | 551 | 26 | 525 | 1.024 |

## 4. Conclusions

Event detection in videos is an important clue for analyzing and understanding content of videos. In this paper, we present an effective approach to detect the interactive events in real-world surveillance videos. Sequence features are designed to capture the event patterns. And these sequence features contain both the basic information in frames and the statistics correlation within video fragments. Based on the proposed sequence features, the interactive event is detected by structural classifiers. The experimental results have validated that the sequence learning based approach can archive better performance for the interactive event detection. And our approach achieves promising results in TRECVid-ED 2010 evaluation.

The future work includes two directions. First, our approach is based on human detection and tracking. We will develop new methods to improve the accuracy of human detection and tracking. Second, more sophisticated sequence learning approach will be developed to learn the spatial-temporal context of events.

## 5. Acknowledgements

This work was supported in part by National Basic Research Program of China (973 Program):2009CB320906, in part by National Natural Science Foundation of China: 61025011, 61035001, 61133003 and 61003165, and in part by Beijing Natural Science Foundation: 4111003.

## 6. References

[1] Aggarwal, J. K. and Ryoo, M. S. 2011. Human Activity Analysis: A Review. *ACM Computing Surveys*. 43(April 2011).

[2] Laptev. On space-time interest points. 2005. *International Journal of Computer Vision*. 64(2):107–123.

[3] Hu, Q., Qin, L., Huang, Q., Jiang, S. and Tian. Q. 2010. Action Recognition Using Spatial-Temporal Context. In *Proceedings of the International Conference on Pattern Recognition*. 1521-1524.

[4] Dollar, P., Rabaud, V., Cottrell G. and Belongie, S. 2005. Behavior recognition via sparse spatio-temporal features. In VS-PETS. 65-72.

[5] Lin, Z., Jiang Z. and Davis, L. S. 2009. Recognizing Actions by Shape-Motion Prototype Trees. In *Proceedings of the International Conference on Computer Vision*. 444 – 451.

[6] Gorelick, L., Blank, M., Shechtman, E., Irani M. and Basri, R. 2007. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 29(12): 2247-2253.

[7] Wang, Y., Tian, Y., Duan, L., Hu Z. and Jia, G. 2010. ESUR: A system for events detection in surveillance video. In *Proceedings of the IEEE International Conference on Image Processing*. 2317 - 2320.

[8] Gaur, U., Song, B., Roy A. 2009. Query-based Retrieval of Complex Activities using "Strings of Motion-Words", In *Proceedings of the Workshop on Motion and Video Computing*. 1 – 8.

[9] Ni, B., Yan S. and Kassim, A. Recognizing human group activities with localized causalities. 2009. In *Proceedings of the Computer Vision and Pattern Recognition*. 1470- 1477.

[10] Cheng, Z., Qin, L., Huang, Q., Jiang, S. and Tian, Q. 2010. Group activity recognition by gaussian processes estimation. In *Proceedings of the International Conference on Pattern Recognition*. 3228-3231.

[11] National Institute of Standards and Technology (NIST), TRECVid 2010 Evaluation for Surveillance Event Detection, http://www.itl.nist.gov/iad/mig/tests/trecvid/2010

[12] Dalal, N. and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the Computer Vision and Pattern Recognition*. 886 – 893.

[13] Babenko, B., Dollar, P., Tu, Z. and Belongie, S. 2008. Simultaneous Learning and Alignment: Multi-Instance and Multi-Pose Learning. In *Proceedings of the European Conference on Computer Vision.*

[14] Grabner, H., Nguyen, T.T., Gruber, B. and Bischof, H. 2007. On-line boosting based car detection from aerial images. *Journal of Photogrammetry & Remote Sensing*. 63(3), 382-396.

[15] Altun, Y., Tsochantaridis, I. and Hofmann, T. 2003. Hidden Markov Support Vector Machines. In *Proceedings of the International Conference on Machine Learning.*

[16] Joachims, T. SVMHMM tool package. Software available at http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html