TJU-TJUT@TRECVID 2015: Surveillance Event Detection

Yuting Su[1], Anan Liu[1*], Zan Gao[2], Weizhi Nie[1], Ning Xu[1], Fuwu Li[1]

[1] School of Electronic Information Engineering, Tianjin University, China
[2] School of Computer and Communication Engineering, Tianjin University of Technology, China
[*] Corresponding author: anan0422@gmail.com

Abstract

In this paper, we present an event detection system evaluated in 2015 TRECVID-SED. The system consists of two parts: automatic event detection (retrospective) and interactive event detection with human in the loop (interactive). For the retrospective part, an extended spatio-temporal features, MoSIFT, is extracted, and two types of temporal divisions (annotation partition and sliding window) are employed as the detection unit. BoW is used to encode low-level features as the representation of each video segment. In order to deal with the highly imbalanced nature of surveillance data, the system performs detections using the proposed Horizontal SVMs algorithm according to each specific event and decision-level post processing is used to combine multiple detection scores. For the interactive part, we designed and developed an interactive visual analytics system, which can enable effective rank detection results with score relations and utilize user feedbacks to improve surveillance event detection.

1 Introduction

Surveillance event detection (SED) addresses the need for automatic detection of events in large amounts of surveillance videos. It is a fundamental problem for a variety of highlevel applications of critical importance to public safety and security. Generally speaking, the task is to identify the temporal range of a specific event such as person running when it occurs in a video. TRECVID [1] provides the retrospective task and interactive task for Surveillance Event Detection to evaluate event detection in real-world surveillance settings. In TRECVID 2015 [2], SED provides a corpus of 144-hour videos under five camera views from the London Gatwick International Airport. This corpus is divided into development and evaluation subsets, in which $\sim$100 camera hours videos (10 days $\times$ 2 hours/day $\times$ 5 cameras) can be used as the training set with annotations of temporal extents and event labels, and $\sim$9 camera hours subset of the 2011 evaluation set for the testing material. Our system is evaluated on all the seven events, i.e., CellToEar, Embrace, ObjectPut, PeopleMeet, PeopleSplitUp, PersonRuns, and Pointing.

The rest of the paper is organized as follows. Section 2 introduces the overall system architecture. In Sections 3, we provide detailed procedures of retrospective event detection task, including

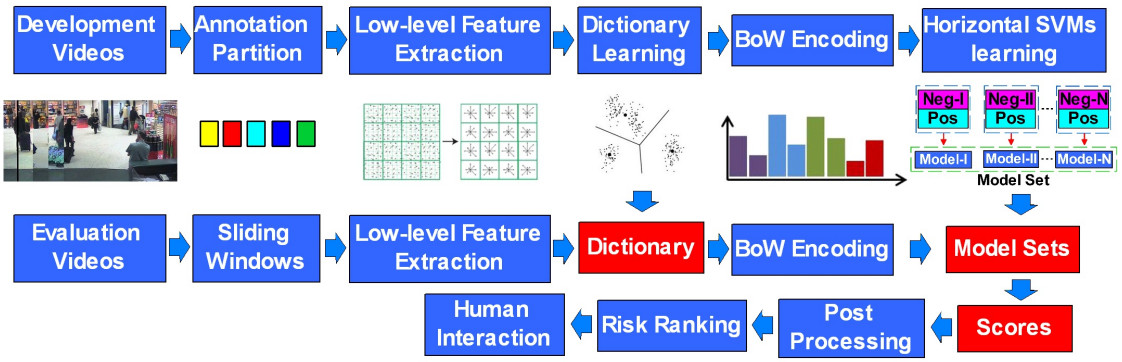


Figure 1: TJU-TJUT surveillance event detection system architecture.

subsequence generation, video representation, model learning and post processing. Especially, we proposed the Horizontal SVMs algorithm to overcome the high imbalance of data samples. Section 4 describes the interactive event detection task in which we design and develop an system focusing on the risk ranking to enable the feedback and improve surveillance event detection. Extensive experiment results and comparisons are reported in Section 5. Finally, we conclude in Section 6.

2 System Overview

As demonstrated in Fig.1, the system consists of 5 major components: (1) low-level feature extraction, (2) subsequence (annotation partition and sliding window) representation, (3) event learning and prediction by Horizontal SVMs, (4) post processing to localize event temporal extents, and (5) human interaction.

In this system, Motion SIFT (MoSIFT) [3] is used as the low-level features to characterize human actions. Then, K-means and BoW encoding methods are used to aggregate the low-level features. With the video representations, the event model sets can be learned by the proposed Horizontal SVMs. Finally, post processing is performed over the positive classification results. Comparing to the typical search systems in previous TRECVID evaluations, a triage-based interaction style is utilized in our work. We design an interactive system based on risk ranking to effectively show the detection results to the end users for refinement.

3 Retrospective Event Detection Task

3.1 Subsequence Generation

We first temporally divide these continuous videos into short video segments. The development subset is divided relying on the event annotations. It generates quite imbalanced data as shown in Fig.2. Furthermore, we randomly select a large number of null event video segments as negative samples for the training process. Meanwhile, the sliding window scheme (25-frame window steps in every 20 frames) is applied on the evaluation subset and each continuous video is divided into small intervals with shot boundaries.

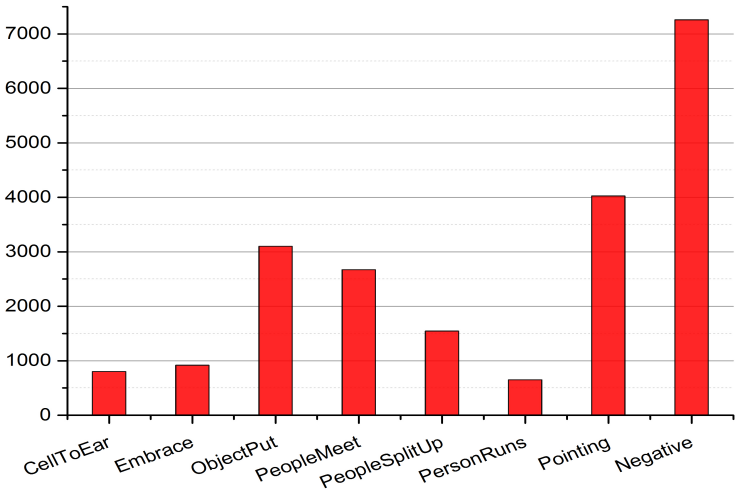


Figure 2: The number of video segments containing training events in the development subset.

3.2 Video representation

For motion features, we use MoSIFT which detects interest points and encodes not only their local appearance but also explicitly models local motion. This feature descriptor is based on the well-known SIFT descriptors to be robust to small deformations through grid aggregation. To limit the complexity, we set the step of temporal frames with 3.

K-means clustering algorithm is applied on these local descriptors. It is a method of vector quantization and aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. We use 1,000,000 randomly selected feature points from the training set to generate the feature pool and dictionary. The centers of the learned clusters is defined as the dictionary and we empirically set the number of dictionary size with 1000. This dictionary is further shared across the training and testing set of the TRECVID Challenge [2]. We employ Bag-of-Words (BoW) encoding to get the descriptors for each video segment which is mapped to a certain vocabulary-word through the clustering process and represented by the frequency histogram over the visual words.

3.3 Model Learning and Score Fusion

With the video representations, we can learn event models by non-linear SVMs solvers. However, the data is highly imbalanced because positive events are far less frequent than negative ones. Therefore, we propose a Horizontal SVMs algorithm to overcome this high imbalance. The model sets are learned according to each specific event.

Suppose we have a training set $S = \{S^+, S^-\}$ for each event. The Horizontal algorithm averagely divides the negative set into a series of partitions S_i^- with the same size of $|S^+|$ and iteratively learns a group of binary SVMs classifiers M_i that favors to positive samples. These classifiers are combined as the event model $C = \{M_1, M_2, \ldots, M_{|C|}\}$. The outline of our proposed learning process is shown in Algorithm 1.

Algorithm 1 Learning event model by Horizontal SVMs

Input:

a training set $S := \{S^+, S^-\}$

Output:

the event model sets C

1: Initialization:
   $n = [|S^-|/|S^+|]$
   Divide the negative set into n partitions
   $C_0 := \{\}$,$w^+ := 1$,$w^- := 1$
2: **for** subset $i := 1$ to n **do**
3: M_i:=NON-LINEAR(S^+, S_i^-,w^+, w^-).
4: $C_i := C_{i-1} \cup M_i$
5: **end for**

We initialize both positive and negative weights as 1 and employ a non-linear support vector machine [4] with a χ^2-kernel [5]

$$K(H_p, H_q) = exp(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{pn} - h_{qn})^2}{h_{pn} + h_{qn}}) \quad (1)$$

where $H_p = \{h_{pn}\}$ and $H_q = \{h_{qn}\}$ are the frequency histograms of word occurrences and V is the vocabulary size. A is the mean value of distances between all training samples [6]. For multi-class classification, a simple, effective combination trains one-versus-rest classifiers (say, "one" positive, "rest" negative) for the special events of the SED.

Because of multiple classifiers operation used in our system, an video segment can obtain a set of scores by multiple classifiers for each special event. Therefore we execute score-level fusions after classification. The score-level fusion combines outputs of classifiers to make the final prediction. Popular score-level fusion methods include minimum, maximum, median, majority voting, weighted sum, and geometric mean [7]. In this system, the simple computation of mean is employed for score-level fusion.

3.4 Post processing

As introduced in Section 3.1, the sliding window scheme (25-frame window steps in every 20 frames) is employed to divide the evaluation subset into small intervals. However, the duration of observational events may range frome 25$\sim$175 frames. It is observed that most positive samples continuously last for a number of frames as temporal extents of most events cover several sliding windows. So neighboring positive predictions are grouped into a merged detection, which is assigned a higher confidence score than those isolated positive predictions. We fine-tuned the parameter of frame over the positive classification results to determine temporal localization of each event and further remove false alarms.

Because of the discriminative appearance for recognizing events in complex background, a fixed threshold is not able to verify all of the automatic detections generated by the system. We fine-tuned the threshold for each special event after the processes of score-fusion and frame-fusion. For

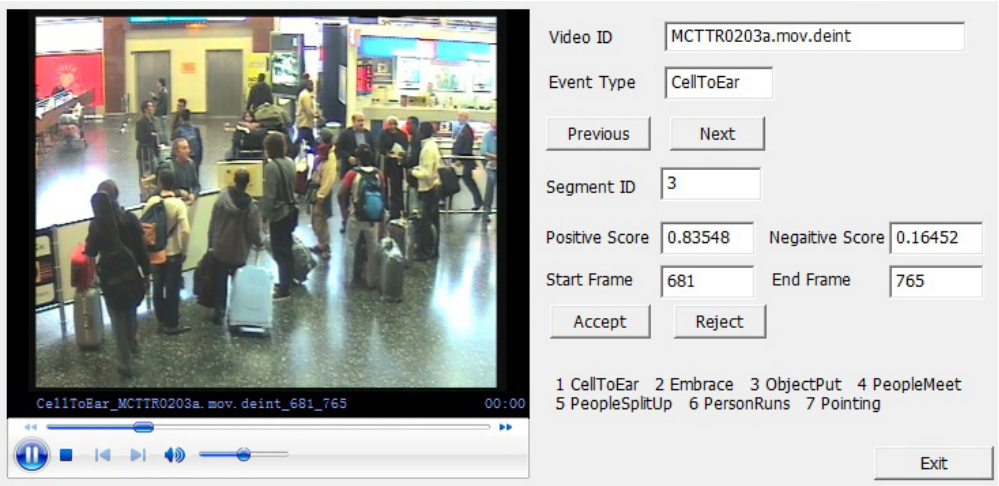


Figure 3: The interface of our interactive system for surveillance event detection.

individual video segment, we select the classes whose scores exceed the corresponding thresholds as the final detection events to further remove false alarms.

4 Interactive Event Detection Task

4.1 Risk-ranking Strategy

Risk analysis is the key module to improve the performance of SED with optimal feedback from human users. We define the *risk score* for a detection as the potential value which differentiates detection results and present more informative ones to the user for better interaction. Popular risk ranking methods include confidence ranking, margin ranking and entropy ranking [8]. To make the most utilization of interaction available in a limited time (25 minutes), in this paper, we only compute confidence ranking based on the detection scores of events. Especially, we rank the video segments according to the detection scores from high to low and present ranking results effectively to the end user for analysis.

Meanwhile, the platform (detailed in Section 4.2) is designed to support a large number of interactive annotation, known as the collecting training samples. We extract the MoSIFT features and employ the BoW encoding on these augmentation samples. We perform the Horizontal SVMs for re-training event model sets and the updated detection scores are submitted as the results of interactive event detection task, which can be used as the measurement of ranking effectiveness.

4.2 Human Interaction

Fig.3 illustrates the user interface, which is simple yet effective in removing false events. The main goal here is to let the user focus on only one event at a time and make decision for the target event in a minimum amount of time. The event video segment is pre-generated and play back in the left portion at the original resolution. The user is able to pause/resume the timer if necessary. The *Previous* and *Next* buttons are mainly for browsing purposes. The video ID and event type are shown in the right portion. For the current event, its associated information is presented on the

right hand side, which includes the segment ID, machine generated score and segment boundary. The *Reject* and *Accept* buttons let the user to reject the events from the final list and accept it with its original boundary and score, respectively. When the user cannot make decision and do not want to spend more time on the current event, the user can directly push the *Next* button and split the single event without the consideration of the *Reject* and *Accept* buttons. It gives the user an option to revisit the difficult events later if time allows. All these action selections are arranged in a group, so the user can easily provide feedback for a sequence of events without much movement of the mouse.

5 Experimental Results

In TRECVID SED 2015, Evaluation Subset (EVAL15) and the new adding Group Dynamic Subset (SUB15) limited to Embrace, PeopleMeet and PeopleSplitUp events are provided as the evaluation sets.

We first compare our system to other best systems in TRECVID SED 2015 by the primary metric Actual Detection Cost Rate (ADCR) and the secondary metric Minimum Detection Cost Rate (MDCR) under the Retrospective Event Detection and Interactive Event Detection scenarios shown in Table 1,2,3,4. The rank column denotes our rankings among all participants in terms of ADCR. We achieve the best performance in five event detection tasks, i.e., CellToEar of the EVAL15 Interactive Evaluation, Embrace of the SUB15 Retrospective Evaluation and all events of the SUB15 Interactive Evaluation. The Detection Error Tradeoff (DET) curves of all events are shown in Fig.4. These curves represent event-averaged miss detection probabilities vs. false alarm rates through varying a detection threshold.

6 Conclusion

In this paper we have presented detailed implementation of our SED system participated in TRECVID 2015. Our system starts from extracting low-level features of MoSIFT from annotated event video segments. K-means and BoW are then employed to aggregate the low-level features. The proposed Horizontal SVMs algorithm is utilized to learn the detection models corresponding to the each specific event. The final scores of each testing segment are computed by fine-tuning the threshold after the score-level and frame-level fusions. We design and develop an interactive visual analytics system that focuses on confidence ranking to enable effective analysis of detection results and utilization of user feedback to improve surveillance event detection. In the primary run evaluations, our system ranks the top in 5 out of 20 event detection tasks and achieves top 50% performances in 6 event detection tasks.

7 AcknowLedgement

This work was supported in part by the National Natural Science Foundation of China(61472275, 61572356, 61170239, 61202168, 61572357), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200), the grant of Elite Scholar Program of Tianjin University (2014XRG-0046).

Event	Rank	ADCR of Other Best Systems	TJU-TJUT Research Primary Run				
			ADCR	MDCR	#CorDet	#FA	#Miss
CellToEar	5/5	1.0046	2.9094	1.0006	21	3999	33
Embrace	3/5	0.7909	0.961	0.8529	44	487	94
ObjectPut	4/6	1.012	2.0675	1.0006	31	2044	258
PeopleMeet	5/6	0.8939	1.1316	0.9978	20	365	236
PeopleSplitUp	4/6	0.8934	0.9892	0.9794	4	27	148
PersonRuns	5/6	0.5768	1.6269	1.0006	14	1578	36
Pointing	4/6	1.004	2.3752	1.0006	142	2704	652

Table 1: Comparison in EVAL15 Retrospective Evaluation.

Event	Rank	ADCR of Other Best Systems	TJU-TJUT Research Primary Run				
			ADCR	MDCR	#CorDet	#FA	#Miss
CellToEar	1/2	2.101	2.101	1.0006	6	2109	48
Embrace	2/4	0.854	0.9656	0.8539	44	495	94
ObjectPut	2/2	0.993	4.6685	1.0006	137	7208	152
PeopleMeet	3/3	0.9978	1.0523	1.0001	10	159	246
PeopleSplitUp	2/3	0.9164	0.9552	0.9477	9	25	143
PersonRuns	3/3	0.9411	1.563	0.9823	13	1432	37
Pointing	3/4	0.9939	1.1927	1.0006	35	412	759

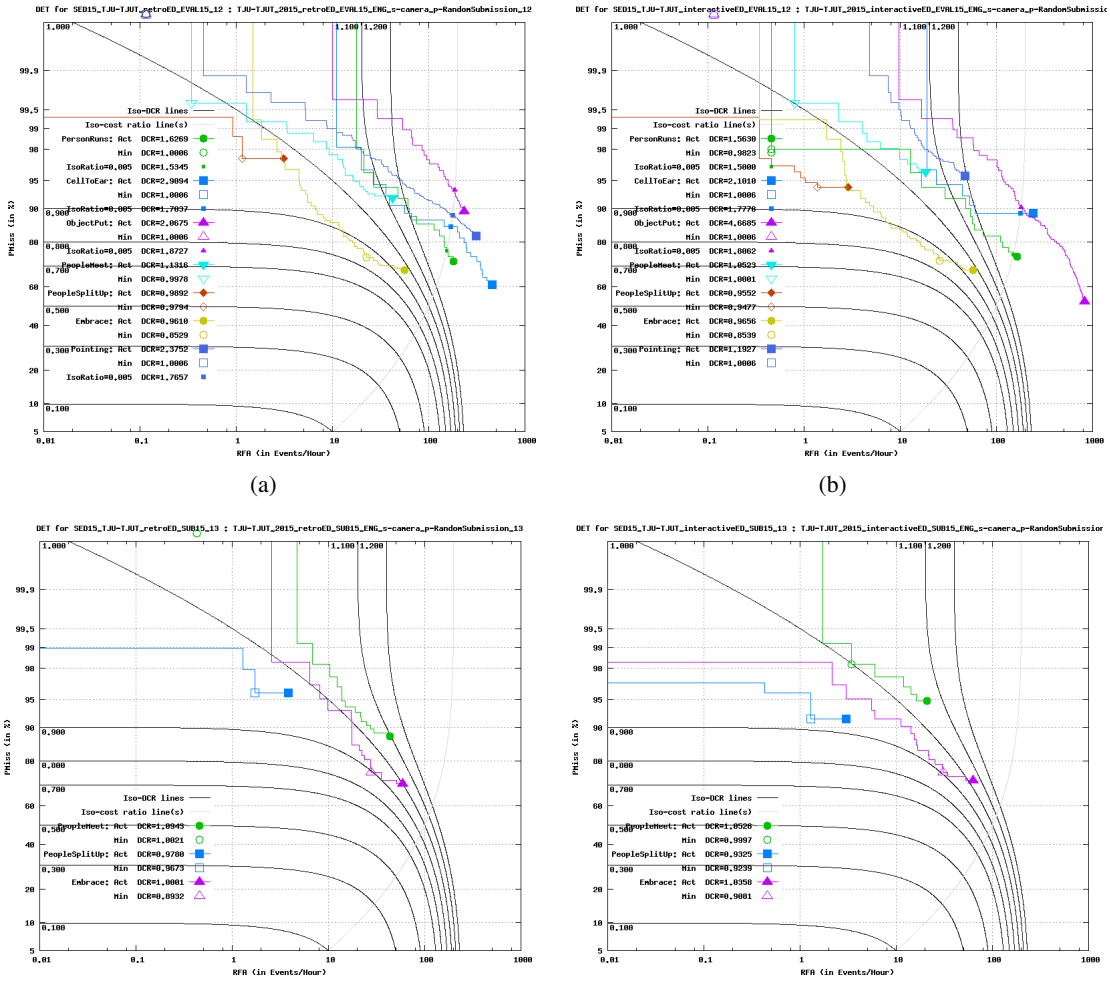
Table 2: Comparison in EVAL15 Interactive Evaluation.

Event	Rank	ADCR of Other Best Systems	TJU-TJUT Research Primary Run				
			ADCR	MDCR	#CorDet	#FA	#Miss
Embrace	1/4	1.0001	1.0001	0.8932	18	138	43
PeopleMeet	4/5	0.9417	1.0943	1.0021	14	101	101
PeopleSplitUp	3/5	0.9572	0.978	0.9673	4	9	93

Table 3: Comparison in SUB15 Retrospective Evaluation.

Event	Rank	ADCR of Other Best Systems	TJU-TJUT Research Primary Run				
			ADCR	MDCR	#CorDet	#FA	#Miss
Embrace	1/2	1.0358	1.0358	0.9081	17	147	44
PeopleMeet	1/1	1.0526	1.0526	0.9997	6	49	109
PeopleSplitUp	1/1	0.9325	0.9325	0.9239	8	7	89

Table 4: Comparison in SUB15 Interactive Evaluation.



(a) (b)

(c) (d)

Figure 4: The Detection Error Tradeoff (DET) curves of our system and each evaluation.

References

1. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and trecvid. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. New York, NY, USA: ACM Press, pp. 321–330. doi:http://doi.acm.org/10.1145/1178677.1178722.

2. Over P, Awad G, Michel M, Fiscus J, Sanders G, et al. (2015) Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2015. NIST, USA.

3. Ming-Yu Chen AH (2009) Mosift: Recognizing human actions in surveillance videos. CMU-CS-09-161 .

4. Chang C, Lin C (2011) LIBSVM: A library for support vector machines. ACM TIST 2: 27.

5. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. doi:10.1109/CVPR.2008.4587756. URL http://dx.doi.org/10.1109/CVPR.2008.4587756.

6. Zhang J, Marszalek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73: 213–238.

7. Atrey PK, Hossain MA, Saddik AE, Kankanhalli MS (2010) Multimodal fusion for multimedia analysis: a survey .

8. Cheng Y, Brown LMG, Fan Q, Feris R, Pankanti S, et al. (2014) Riskwheel: Interactive visual analytics for surveillance event detection. In: IEEE International Conference on Multimedia and Expo, ICME 2014, Chengdu, China, July 14-18, 2014. pp. 1–6. doi:10.1109/ICME.2014.6890286. URL http://dx.doi.org/10.1109/ICME.2014.6890286.