IBM-Northwestern@TRECVID 2014: Surveillance Event Detection

Yu Cheng[*] [†] Lisa Brown [*] Quanfu Fan[*] Jingjing Liu[*] Rogerio Feris[*]
Alok Choudhary[†] Sharath Pankanti[*]

1 Overview

We present a system for detecting events in surveillance videos and evaluate it in the SED task of TRECVID [1]. The system consists of two parts: automatic event detection (*retrospective*) and interactive event detection with human in the loop (*interactive*). The retrospective system jointly performs segmentation and classification of events in a video and applies the Sequence Memoizer [2] to capture long-range dependencies in the temporal context of a visual data sequence. For the interactive part, we designed and developed an interactive visual analytics system, which enables effective analysis of detection results and utilization of user feedback to improve surveillance event detection. In particular, we propose 1) an interactive approach to visualize data with temporal relations and 2) a novel risk ranking method to differentiate detection results and present more informative ones to the user for better interaction.

2 Automatic Event Detection

Given an input video $\mathbf{X}$, we first divide it into n temporal segments of a fixed length l_{seg}, i.e. $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$. We then compute the bag of words (BOW) feature for each segment upon motion SIFT key points. The segments are further clustered into k visual words using k-means, and each segment is assigned a visual word. Finally, the video is represented by a sequence of visual words $\mathbf{W} = \{w_1, w_2, \cdots, w_n\}$. In our experiments, l_{seg} was set invariantly to the total length of the video, and k usually ranges from 600 to 900 depending on the complexity of the data.

One immediate observation is that the same event tend to generate similar visual words. A visual word from one event may statistically interact with another one from a different event, even though the two words can be temporally distant. With the video representation described above, our goal is to partition $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ into m units and label each unit with an event of interest or a null event. Here a unit is a set of consecutive segments of $\mathbf{X}$. Let $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_m\}$ be such a partition where a unit $\mathbf{s}_i = \mathbf{X}_{t_i^1:t_i^2} = [\mathbf{x}_{t_i^1}, \ldots, \mathbf{x}_{t_i^2}]$ and t_i^1 and t_i^2 specify the start and end indices of the segments in $\mathbf{s}_i$. Also, let $\mathbf{Y} = \{y_1, y_2, \cdots, y_m\}$ where $y_i \in \mathbf{Y}$ is the event class label assigned to $\mathbf{s}_i$. To model temporal contexts in the data, we associate $\mathbf{S}$ with a visual sequence $\mathbf{Z} = \{z_1, z_2, \cdots, z_l\}$. The quality of the partition $\mathbf{S}$ with regard to event classification can then be evaluated by,

(2.1)

$$f(\mathbf{S}, \mathbf{Y}) = \sum_{i=1}^{m} \varphi(y_i|\mathbf{s_i}) + \mu \sum_{\substack{i=1 \\ 1 \le k \le i-1}}^{l} p(z_i|z_{i-k}, \cdots, z_{i-1})$$

where μ is a trade-off parameter learnt from data empirically. Note that $\mathbf{Z}$ can be of any visual data sequence created on top of $\mathbf{S}$. For example, a sequence of visual events or visual words.

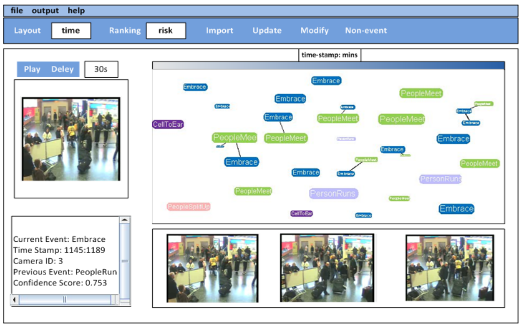
The first item $\varphi(y_i|\mathbf{s_i})$ in Eq. 2.1 measures the likelihood of the unit $\mathbf{s_i}$ being event y_i. We use the SVM classification score of $\mathbf{s_i}$ on event y_i for this item. The second item $p(z_i|z_{i-k}, \cdots, z_{i-1})$ is provided by our sequence model. To put it simple, it is the probability of predicting z_i as the next symbol after seeing the previous k symbols from z_{i-k} to z_{i-1}. The probability can be estimated by Sequence Memoizer [3]. The above objective function can be solved efficiently by dynamic programming.

Note that in model learning, the temporal sequence $\mathbf{Z}$ described above can be an event sequence from ground truth or a sequence of visual words obtained from unsupervised learning such as K-mean clustering on segments. More details about these types of modeling can be found in [2].

3 Interactive Event Detection

In this paper, we propose a novel interactive visual analytics system as shown in Figure 1, to improve the performance of SED with optimal feedback from human users. This platform is designed to provide effective interactions for the user and support interactive annotation (collecting training samples) and re-annotation. To make the most utilization of interaction available in a limited time, the system design was driven by considerations from two perspectives: 1) efficient visualization of intermediate detection results for

[*] IBM T.J. Watson Research Center, Yorktown Heights, NY 10582, USA
[†] EECS Department, Northwestern University, Evanston, IL 60208, USA



Current Event: Embrace
Time Stamp: 1145:1189
Camera ID: 3
Previous Event: PeopleRun
Confidence Score: 0.753

Figure 1: The interface of our interactive system for surveillance event detection.

user interaction; and 2) effective utilization of user feedback for performance boost. Our first major contribution is a novel visual design that presents the intermediate detections with temporality. We capture successive events from a sequence of detections and represent them as streaming belts, i.e. groups of events with temporal patterns. Compared to the linear presentation used by the currently existing systems, such a two-dimensional representation helps to facilitate data understanding and better portrays rich interactions that enables explorative analysis. Our second contribution is the development of a method based on risk ranking to present detection results effectively to the end user for analysis. Here the risk of a detection indicates the potential value or impact that the detection has on the system performance upon analysis by the user. We propose a novel way to measure the risk of a detection by combining several factors into an overall score, including the margin of top two candidate events for the detection, temporal relations between events and potential annotation costs.

Given a segment S_i, let the top two detections be the kth and k'th event with scores $\varphi^k(S_i)$ and $\varphi^{k'}(S_i)$, following the stream-based active learning [4], the formulation of the risk score for a single event can be expressed as:

(3.2)

$$R(S_i) = \frac{1 - (\varphi^k(S_i)p(k) - \varphi^{k'}(S_i)p(k'))}{||S_i||} \cdot \begin{cases} w_m \\ w_f \end{cases}$$

where w_m is the cost of a miss detection, w_f is the cost of a false alarm and $||S_i||$ is the length the segment S_i. $p(k), k \in 1, 2, ..., K$ is the occurrence prior of event k learnt from the ground truth. Similarly, we can develop the formulation for a pair of events, which is omitted here due to space limit.

4 Experimental Results

We applied our temporal modeling approach to the SED 2014 retrospective task. Table 1 shows our evaluation results provided by NIST (*Ours14_retro*), along with the best performance achieved by other participating teams (*Others14_best*). our approach outperforms others, in 6 out of 7 events, demonstrating the effectiveness of modeling temporal dynamics of events. Interested readers are referred to [2] for more comparisons of our approach with other event detection techniques on the SED development dataset.

Table 1: Retrospective and Interactive Evaluation Results

Method	*Others14_best*	*Ours14_retro*	*Ours14_inter*
CellToEar	1.0032	0.9835	**0.9749**
Embrace	0.7845	0.7456	**0.6662**
ObjectPut	1.0216	0.9914	**0.9900**
PeopleMeet	0.9125	0.8160	**0.7965**
PeopleSplitUp	0.8433	0.8278	**0.7869**
PeopleRuns	0.8339	0.8111	**0.8101**
pointing	1.0040	1.0050	**0.9788**

For the interactive task, following the NIST guidelines, we conducted a search for each event in 25 minutes using an interactive system we developed based on risk analysis. The results (*Ours14_inter*) are shown in Table 1. Clearly, an effective interactive process can significantly boost the retrospective results. More details about the design of our system can be found in [5].

References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot, "Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[2] Y. Cheng, Q. Fan, S. Pankanti, and A. N. Choudhary, "Temporal sequence modeling for video event detection," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 2014, pp. 2235–2242.

[3] F. Wood, C. Archambeau, J. Gasthaus, L. James, and Y. W. Teh, "A stochastic memoizer for sequence data," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1129–1136.

[4] Y. Cheng, Z. Chen, L. Liu, J. Wang, A. Agrawal, and A. Choudhary, "Feedback-driven multiclass active learning for data streams," in *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, 2013, pp. 1311–1320.

[5] *IEEE International Conference on Multimedia and Expo, ICME 2014, Chengdu, China, July 14-18, 2014*. IEEE Computer Society, 2014. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6882588