A SYSTEM BASED ON SEQUENCE LEARNING FOR EVENT DETECTION IN SURVEILLANCE VIDEO

Xiaoyu Fang[1], *Ziwei Xia*[2], *Chi Su*[1], *Teng Xu*[1], *Yonghong Tian*[1], *Yaowei Wang*[2], *Tiejun Huang*[1*]

[1]School of EE&CS, Peking University, Beijing, China
[2]Department of Electronic Engineering, Beijing Institute of Technology, Beijing, China

ABSTRACT

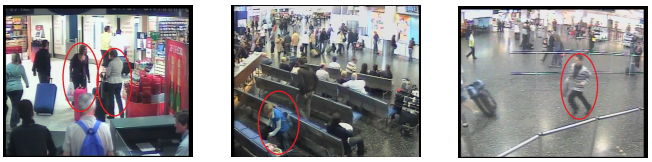
Event detection in crowded surveillance videos is a challenging yet important problem. In this paper, we present our eSur (Event detection system on SURveillance video) system, which is derived from TRECVid'12 surveillance tasks. Currently, eSur attempts to detect two categories of events: 1) pair-wise events (e.g., PeopleMeet, PeopleSplitUp and Embrace); 2) action-like events (e.g., ObjectPut, CellToEar, PersonRuns and Pointing). In eSur system, we first employ people detection and tracking algorithms to locate target persons in 3D space-time domain. Then the video sequences in which target persons occur are partitioned into several spatio-temporal cubes. Visual features (i.e. cubic feature and MoSIFT) are computed over these cubes. After that, a sequence learning method, (namely SVM with dynamic time alignment kernel), is employed to infer the existence of an event for the video sequence. According to the TRECVid SED formal evaluation, eSur has yielded fairly encouraging results on TRECVid'12 dataset.

Index Terms— Event detection, surveillance, sequence learning

1. INTRODUCTION

Nowadays, a lot of surveillance cameras are equipped for public security. Detecting interest events in real surveillance videos automatically could reduce the burden of human operators. The TRECVid surveillance event detection (SED) [1] evaluation campaign is initiated to evaluate systems that can detect instances of a variety of observable events in the airport surveillance domain. Unlike the well controlled laboratory environments (e.g. CAVIAR [2] and PETS [3]), scenes in TRECVid dataset are captured from the Gatwick airport of London, which are very complex with crowded conditions and frequent occlusions.

eSur is developed to meet the TRECVid SED requirements of automatically discovering predefined events, i.e., retrospective events task of TRECVid SED 2012. PeopleMeet, PeopleSplitUp and Embrace are pair-wise events and indicate interactions between different persons; ObjectPut, CellToEar, PersonRuns and Pointing are action-like events and indicate single person's movements. Some samples of video events defined in TRECVid SED task are shown in Fig.1. We apply different methods to each event category. For pair-wise events, a novel cubic feature is developed to describe the relationships between different persons. For action-like events, The MoSIFT [4] feature is employed to represent the person's motion and appearance. Meanwhile, we notice that events usually consist of temporal ordered movements. Therefore, we regard events as special sequential patterns occurring over a period of time. Motivated by the dynamic time warping method used for acoustical signal processing, we employ the SVM with dynamic time alignment kernel to infer the exitance of an event in a given video sequence.



(a) PeopleMeet (b) ObjectPut (c) PersonRuns

Fig. 1. Samples of events in TRECVid SED dataset.

The eSur system presented in this paper is an upgraded version of that introduced in [5]. The main differences with that system presented in [5] are summarized as follows. First, we target more events. ObjectPut, CellToEar, PersonRuns and Pointing could be detected using the upgraded eSur system now. Second, we employ the sequence learning method (SVM with dynamic time alignment kernel) to recognize sequential patterns in surveillance videos. Third, we use the multiple kernel learning (MKL) to combine different descriptors and improve the performance of pair-wise event detection. Experimental results have proved that the upgraded eSur system has significant improvements comparing with the former version presented in [5].

The remainder of this paper is organized as follows. In

*This work is partially supported by grants from the Chinese National Natural Science Foundation under contract No.61035001, No.61072095 and No.61171139, and National Basic Research Program of China under contract No.2009CB320906. Contact the author via yhtian@pku.edu.cn

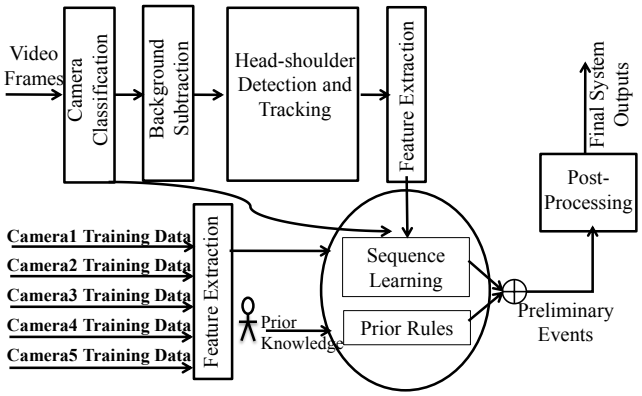


Fig. 2. Framework of eSur system.

Section 2, we present eSur system framework. Our event detection approach using sequence learning is described in Section 3. Experimental results are given in Section 4. Finally, we conclude this paper in Section 5.

2. ESUR SYSTEM FRAMEWORK

In our eSur system, we first perform a background subtraction algorithm [6] to get interest regions (foreground) in each frame. Then persons' positions and trajectories are detected over foreground regions using object detection and tracking algorithm [7, 8]. We integrate the object detection and tracking in one unified method (i.e., "detection by tracking" and "tracking by detection") [9] to give more accurate results. After that, the cubic feature and MoSIFT [4] are extracted to represent appearance, motion and trajectory information of video sequences. Finally, the SVM with dynamic time alignment kernel is used to recognize predefined events. Framework of eSur is illustrated in Fig.2.

Cubic feature is a novel feature developed to describe the relationships between object pairs (two persons). For a video subsequence in which an object pair coexists, we first partition it into k (variable) cubes, L (constant, e.g., 10) frames as one cube. Then statistical trajectory descriptor (i.e., mean distance one from another, mean relative speed magnitude, mean overlapped area of objects' regions) is extracted and spatio-temporal interest points [10] in target persons' regions are detected within each cube. After that we cluster these interest points and generate a histogram descriptor for each cube according to a visual vocabulary built off-line with training points. So, the object pair is represented with a sequence of k trajectory descriptors and a sequence of k BoW descriptors. Then the trajectory descriptor sequence and BoW descriptor sequence are fused into a cubic feature using MKL method [11]. The fused cubic feature is used for pair-wise event detection.

The MoSIFT [4] feature captures local appearance and motion information in videos by combining histogram of gradients and histogram of optical flow. We extract MoSIFT features within interest regions that detected and tracked by the detection and tracking algorithm. Also, we partition the video subsequence into k spatio-temporal cubes, and then perform the bag-of-feature method in each cube to form a sequence of BoW descriptors. The MoSIFT feature is used for action-like event detection.

We treat the event detection as a classification problem. And the one-vs.-all classifier is employed to classify event instances from the others. The SVM with dynamic time alignment kernel [12] is used to recognize sequential patterns in surveillance videos. Furthermore, some rules based on prior knowledge of the events are used to filter false detections. In the post-processing stage, event instances occurring in an overlapping time span are merged to one detection record, because only the locations of the events in time domain are checked in the TRECVid SED formal evaluation and repeated records increase false alarms.

3. EVENT DETECTION BASED ON SEQUENCE LEARNING

The pair-wise events (i.e., PeopleMeet, PeopleSplitUp and Embrace) involve the interaction of at least two persons. Therefore, the cubic feature is used to describe relationships between two persons coexisting in the same scene. However, the action-like events (i.e., ObjectPut, CellToEar, PersonRuns and Pointing) are just single person's special movements, so the MoSIFT [4] feature is used for detecting this kind of events. In the training and classification stage, the SVM with dynamic time alignment kernel is operated over both kinds of features to infer the happenings of two kinds of events respectively.

3.1. Sequence Learning

Video events are inherently special sequential patterns. For example, the ObjectPut event indicates a person walks into the scene, stands still, bends down and puts something down. The whole process involves a sequence of movements: walking, standing, bending-down and putting-down. Each step of movement may last different lengths of time. The sequence learning method could align the sequential form features dynamically, and recognize sequential patterns regardless of each movement's time duration. Therefore, we first extract sequential form features by partitioning video sequence into spatio-temporal cubes and computing descriptors over these cubes. Then the SVM with dynamic time alignment kernel [12] is employed to detect both kinds of events.

Given two sequences of descriptors, $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$ and $V = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m)$, the dynamic time alignment kernel is able to find the optimal warping path that maximize the

accumulated similarity between them.

$$K_s(X,V) = \max_{\psi,\theta} \frac{1}{W_{\psi\theta}} \sum_{i=1}^{N} w(i) K(\mathbf{x}_{\psi(i)}, \mathbf{v}_{\theta(i)}) \quad (1)$$

subject to

$$1 \le \psi(i) \le \psi(i+1) \le |X|, \psi(i+1) - \psi(i) \le Q$$
$$1 \le \theta(i) \le \theta(i+1) \le |V|, \theta(i+1) - \theta(i) \le Q$$

where Q is a constant constraining the local continuity, ψ and θ stand for a warping path, N is the length of the warping path, $w(i)$ is a nonnegative weighting coefficient, $W_{\psi\theta} = \sum_i^N w(i)$ is a path normalizing factor, and $K(\mathbf{x}_{\psi(i)}, \mathbf{v}_{\theta(i)}) = exp(-\gamma||\mathbf{x}_{\psi(i)} - \mathbf{v}_{\theta(i)}||^2)$, that is Radial Basis Function (RBF) kernel. We set $w(i) = 1$, so $W_{\psi,\theta} = N$. Sequence learning method exploits not only the discrete information from individual descriptor, but also the sequence and correlation information among successive descriptors.

3.2. Pair-wise event detection

Statistical trajectory descriptor is developed to describe the relationships of two persons. Let $A_m = [a_1, ..., a_i, ..., a_T]$ and $B_n = [b_1, ..., b_i, ..., b_T]$ be motion trajectories of objects m and n, where a_i and b_i are tuples (x,y) of the object coordinates in 2D image plane at time i, and m, n are objects' identifiers. To represent the relationships of the objects in each cube, and remove the influences of occasional error caused by detection and tracking, statistical data is employed, such as mean distance one from another, mean relative speed magnitude, mean overlapped area of objects' regions. Meanwhile, the difference of these statistical data between current and next cubes is important as well. Therefore, trajectory descriptor of k^{th} cube is extracted as follows:

$$TD^k = \{c^k_{dis}, c^k_{sp}, c^k_{ov}, dc^k_{dis}, dc^k_{sp}, dc^k_{ov}\} \quad (2)$$

where c^k_{dis}, c^k_{sp} and c^k_{ov} is mean distance, mean relative speed magnitude and mean overlapped area within k^{th} cube respectively, and

$$\begin{cases} dc^k_{dis} = c^{k+1}_{dis} - c^k_{dis} \\ dc^k_{sp} = c^{k+1}_{sp} - c^k_{sp} \\ dc^k_{ov} = c^{k+1}_{ov} - c^k_{ov} \end{cases} \quad (3)$$

Cubic feature is used for detecting pair-wise events. The cubic feature includes a sequence of statistical trajectory descriptors and a sequence of BoW descriptors, expressed as $X^* = [X^{Tr}, X^B]$. Here, $X^{Tr} = (\mathbf{x}^{Tr}_1, \mathbf{x}^{Tr}_2, ..., \mathbf{x}^{Tr}_k)$ is a sequence of statistical trajectory descriptors extracted from spatio-temporal cubes, and $X^B = (\mathbf{x}^B_1, \mathbf{x}^B_2, ..., \mathbf{x}^B_k)$ is a sequence of BoW interest point [10] descriptors. We combine the two sequences together using MKL learning method. Assume that we have two cubic features $X^* = [X^{Tr}, X^B]$ and $V^* = [V^{Tr}, V^B]$. The sequence kernel with MKL feature fusion used for cubic feature is defined as follows.

$$\begin{aligned} K^*_s(X^*, V^*) &= \beta^{Tr} K_s(X^{Tr}, V^{Tr}) + \beta^B K_s(X^B, V^B) \\ &= \beta^{Tr} \max_{\psi^{Tr},\theta^{Tr}} \frac{1}{N} \sum_{i=1}^{N} K(\mathbf{x}^{Tr}_{\psi^{Tr}(i)}, \mathbf{v}^{Tr}_{\theta^{Tr}(i)}) \\ &+ \beta^B \max_{\psi^B,\theta^B} \frac{1}{M} \sum_{i=1}^{M} K(\mathbf{x}^B_{\psi^B(i)}, \mathbf{v}^B_{\theta^B(i)}) \end{aligned} \quad (4)$$

where K_s is the Dynamic Time Alignment Kernel (DTAK) [12], K is RBF kernel, β^{Tr} and β^B are optimal combination parameters obtained in feature fusion step, (ψ^{Tr}, θ^{Tr}) and (ψ^B, θ^B) are warping paths of descriptor sequence pairs (X^{Tr}, V^{Tr}) and (X^B, V^B), N and M are lengths of the warping paths.

We first labeled a training set for each type of event, and partition the training samples into two folds. Then the MKL [13] learning method is applied to get the optimal combination parameters (β^{Tr}, β^B) over the two folds of samples. Using the optimal combination parameters, the classification hyperplane is decided and then used to classify unknown surveillance videos. Meanwhile, some rules based on prior knowledge of the pair-wise events are used to filter false detections. For PeopleMeet, Embrace and PeopleSplitUp, we define a contacting threshold Ct. Two persons having a meet (or) embrace, their mean distances from each other in spatio-temporal cubes must change from larger than Ct to smaller than Ct. For PeopleSplitUp, the rule is opposite, the mean distances change from smaller than Ct to larger than Ct. Ct is decided by the labeled training samples.

3.3. Action-like event detection

The MoSIFT feature is used for detecting single actor's action-like events. Also, we apply the SVM with dynamic time alignment kernel on MoSIFT features to find interest actions. To locate the precise range of time during which the events happened, we employ a slide window of W (e.g.,50) frames (in temporal domain) over an video clip bounded by the actor's region and existing time. Spatio-temporal cubes are acquired from videos sequence in the slide window. And then sequential features are constructed by computing a bag-of-features descriptor within each cube and concatenating these descriptors together.

We train a one-vs.-all classifier based on SVM with dynamic time alignment kernel for each type of action-like event. Also, some rules are used to improve the detection accuracy. For ObjectPut, a short period of downward optical flow should exist on the target person, which indicates the put-down action. As for CellToEar and Pointing, upward optical flow must be observed which indicates the raise-hand action. As for PersonRuns, running persons have a larger velocity

than others. We define a speed threshold St for PersonRuns, and prune those persons whose average speed is smaller than St from the positive set. St is obtained from the labeled training set using statistical method.

4. EXPERIMENTAL RESULTS

We use the Normalized Detection Cost Rate (NDCR) [1] as primary evaluation measure, which is the same with NIST formal evaluation. NDCR is a weighted linear combination of the system's Missed Detection Probability (P_{Miss}) and False Alarm Rate (R_{FA}) (measured per unit time).

$$NDCR(S, E) = P_{Miss}(S, E) + Beta * R_{FA}(S, E) \quad (5)$$

where S is the evaluated system, E is the interest event and $Beta$ is a constant value (default: 0.005). A **smaller** NDCR means **better** performance.

As listed in Table 1, our results of all events on TRECVid 2008 dataset outperform the reported best results of TRECVid 2008. Moreover, the upgraded eSur system obtains superior results on PeopleMeet, Embrace, PeopleSplitUp and PersonRuns over its former version [5]. The improvements are -0.345, -0.256, -1.041 and -0.279 respectively for four kinds of events. Comparing with the former version, the main updates of eSur system include: 1)sequential feature computation by partitioning video sequence into spatio-temporal cubes; 2)sequence learning using SVM with dynamic time alignment kernel. For event instances consist of temporal ordered motions, sequential feature is more informative by introducing temporal order information of movements. The dynamic time alignment kernel could align visual features according to similarity maximization principle (Eq.1). It is very effective in recognizing sequential patterns. Experimental results have proved that the sequence learning method significantly benefit the detection performance of eSur system.

Comparison results reported in TRECVid'12 SED formal evaluation are shown in Table 2. It is indicated that our results of PeopleMeet, ObjectPut, CellToEar and Pointing are encouraging, slightly better than some excellent teams (i.e., CMU-IBM and MediaCCN). However, for Embrace, PeopleSplitUp and PersonRuns, our results are inferior to the results reported by CMU-IBM. CMU-IBM team also uses the MoSIFT feature. However, they employ a bigram model of video code words based on tf-idf weight (term frequency-inverse document frequency) which is common in information retrieval and text classification. The bigram model obtains good results for detecting Emrace, PeopleSplitUp and PersonRuns. However, our sequence learning method still shows effectiveness comparing with the bigram model. Because, we employ MoSIFT feature only for detecting four kinds of events (i.e., ObjectPut, CellToEar, PersonRuns and Pointing). And using sequence learning, we get slightly better results on three of them (i.e., ObjectOut, CellToEar and Pointing).

Table 1. Comparison results with other methods on TRECVid 2008 data corpus using $NDCR$ measure.

Event	Best 2008	Wang [5]	Ours
PeopleMeet	1.337	1.245	**0.980**
Embrace	1.271	1.208	**0.952**
PeopleSplitUp	4.856	1.976	**0.935**
ObjectPut	1.004	N/A	**0.996**
CellToEar	0.999	N/A	**0.962**
PersonRuns	0.989	1.249	**0.970**
Pointing	1.080	N/A	**0.964**

Table 2. Comparison results with other methods in TRECVid 2012 SED tasks using $NDCR$ measure.

Event	CMU-IBM	MediaCCN	Ours
PeopleMeet	1.036	1.008	**0.980**
Embrace	**0.800**	0.955	0.951
PeopleSplitUp	**0.843**	0.984	0.978
ObjectPut	1.004	1.016	**0.998**
CellToEar	N/A	1.009	**1.004**
PersonRuns	**0.835**	0.970	0.975
Pointing	1.018	1.090	**0.994**

5. CONCLUSION

In this paper, we present a system based on sequence learning for event detection in surveillance videos. In our system, we extract sequential features by partitioning video sequences into spatio-temporal cubes and constructing visual descriptors within each cube. Furthermore, a sequence discriminant learning method (SVM with dynamic time alignment kernel) is used for detecting pair-wise and action-like events. Experimental results have shown that the presented eSur system is effective in detecting video events in real-world complex surveillance scenes. According to the TRECVid'12 formal evaluation, the presented eSur system obtains encouraging results comparing with several well-known methods in the literature.

6. REFERENCES

[1] NIST, "Trec video retrieval evaluation," `http://www-nlpir.nist.gov/projects/trecvid`.

[2] NIST, "Caviar: Comtext aware vision using image-based active recognition," `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`.

[3] Desurmont X., "Performance evaluation of frequent events detection systems," *IEEE international Work-*

shop Performance Evenuation of Tracking and Surveillance, 2006.

[4] M. Cheng and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," *CMU-CS-09-161*, 2009.

[5] Yaowei Wang, Yonghong Tian, Lingyu Duan, Zhipeng Hu, and Guochen Jia, "esur: a system for events detection in surveillance video," *ICIP*, pp. 2317–2320, September 2010.

[6] Z. Hu, Y. Wang, Y. Tian, and T. Huang, "Selective eigenbackgrounds method for background subtraction in crowed scenes," *ICIP*, 2010.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *CVPR*, 2005.

[8] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool, "Online multi-person tracking-by-detection from a single, uncalibrated camera," *PAMI*, 2010.

[9] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," *CVPR*, pp. 1–8, 2008.

[10] I. Laptev, "On space-time interest points," *IJCV*, 2005.

[11] F. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," *ICML*, 2004.

[12] Hiroshi Shimodaira, Ken ichi Nom, and Mitsuru Nakaiand Shigeki Sagayama, "Dynamic time-alignment kernel in support vector machine," *Neural Information Processing Systems*, pp. 921–928, 2001.

[13] Manik Varma and Bodla Rakesh Babu, "More generality in efficient multiple kernel learning," *ICML*, 2009.