Detecting Rare Actions and Events from Surveillance Big Data with Bag of Dynamic Trajectories

Yaowei Wang*, Yonghong Tian†, Limin Su‡, Xiaoyu Fang†, Ziwei Xia*, and Tiejun Huang†
* School of Information and Electronics, Beijing Institute of Technology, Beijing, China. 100081
Email: yaoweiwang@bit.edu.cn
† National Engineering Laboratory for Video Technology, School of EE & CS,
Peking University, Beijing, China. 100871
Email: yhtian@pku.edu.cn
‡ College of Information Technology, Beijing Union University, Beijing, China. 100101
Email: xxtlimin@buu.com.cn

***Abstract*—Surveillance video is increasingly becoming the "biggest big data". This presents an unprecedented challenge for analyzing and mining the meaningful information (e.g., rare actions or events) in such a huge amount of videos. Recent studies have shown that feature-trajectories-based methods are effective to encode motion information in video, consequently demonstrating superior performance in action and event detection. However, in existing methods, distance between two trajectories is often measured by linear models, which may be not robust enough when the lengths of trajectories are variable. Moreover, due to the rare distribution of target actions or events, the traditional classifier often tends to identify all samples as negative, consequently producing heavy performance bias. To address both two issues, this paper proposes a trajectory descriptor, BoDT (Bag of Dynamic Trajectories), and a multi-channel uneven SVM. By utilizing the DTW (dynamic time warping) algorithm to measure the similarity between two trajectories, BoDT is robust for variable-length trajectory representation. Meanwhie, as an extension of SVM with uneven margins, the proposed multi-channel uneven SVM can successfully identify rare events by adjusting a margin parameter to make the classification boundary properly moved away from the positive training examples. Extensive experiments on several benchmark datasets including KTH, YouTube, Olympic, MIT, QMUL and TRECVid demonstrate that our approach is feasible and effective.**

I. Introduction

Nowadays, surveillance cameras, especially high definition (HD) cameras, are widely deployed all over the world. No doubt this makes video quality better, while the amount of video data is explosively increasing at the same time. It is estimated that one single HD camera can approximately generate 0.7TB compressed video data per month. What an unimaginable huge amount of video data is generated by millions of HD cameras day and night! Therefore, surveillance video data is becoming the "biggest big data" [1], [2]. This presents an unprecedented challenge for analyzing, extracting and mining the meaningful information (e.g., rare actions or events) in such a huge amount of videos.

Normally, action and event detection is treated as a pattern recognition task. First, visual features are extracted from consecutive video frames; then, classifiers are utilized to determine whether the action or event happens. Obviously, motions are the most valuable visual clues for identifying an action or event. To effectively encode motion information in actions or events, feature-trajectories-based methods are proposed in recent studies [3]–[6] and have shown exciting performance in action and event detection. Basically, feature-trajectory-based methods track interesting points in video frames to obtain trajectories. Wang et al. [5] improved it remarkably by estimating dense trajectories with sampling dense points from video frames and tracking them based on displacement information from a dense optical flow field. After that, local features (such as histogram of gradients (HoG), histogram of optical flow (HoF) and motion boundary histogram (MBH)) are extracted from a 3D video volume along the trajectory, concatenated into a trajectory feature. Finally, the similarity between two trajectories is measured by linear method with $L2$ norm. However, we notice that lengths of feature trajectories are quite different (as shown in Fig. 1). Consequently, the commonly used linear method with $L2$ norm may

978-1-4799-8688-0/15 $31.00 © 2015 IEEE
DOI 10.1109/BigMM.2015.74

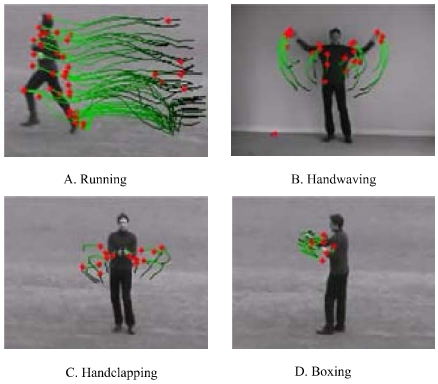


Fig. 1. Trajectories extracted from videos. Local descriptors are computed along trajectories, and then concatenated together as spatio-temporal features.

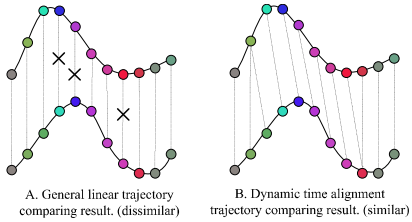


Fig. 2. The linear method compares points along trajectories one by one, as shown in (A). Two trajectories even have the similar patterns are considered to be dissimilar. The dynamic time alignment (DTW) method can align trajectories dynamically, and similar patterns can be recognized in a flexible way, see (B).

(a) Pointing (b) CellToEar

Fig. 3. Typical and rare event examples in the TRECVid dataset. The description of Pointing is that someone points, and CellToEar means that someone puts a cell phone to his/her head or ear. Rare event (red) usually co-occurs with numerous typical event (yellow).

be not robust enough to measure the similarity between two trajectories. It could be observed in Fig. 2(A) that even two sequences of exactly the same type of action cannot be matched if they are not aligned strictly.

Moreover, the rare distribution of target actions or events makes the traditional classifier failed to detect such rare actions or events. For example, in test set of TREVid ED [7] (as shown later in table I), the ratio between typical events and 'pointing' (Fig.3 (a)) is 306.58 $(122939 : 401)$, and the ratio between typical events and 'CellToEar' (Fig.3 (b)) reaches 1596.61 $(122939 : 77)$. When an action or event is rare, the classifier tends to bias toward the majority class (i.e., identify them all as typical). Over 99.9%, the classification accuracy is still very high, but it is not the result we want. To address this problem, a natural policy is data preprocessing [8], such as oversampling [9]–[11] and undersampling [12], to make the data distribution balancing. Another way is to map the original data to some other spaces where the distribution is not so unbalanced. Timothy et al. [13] proposed a weakly-supervised joint topic model (WS-JTM) to detect rare events and achieved state-of-the-art results. They introduced a multi-class topic model with partially shared latent structure and associated learning and inference algorithms. Actually, the original uneven-distributed data were mapped to a topic space, in which the rare event could be identified. However, due to the high model complexity, their method is difficult to scale to large datasets.

To address both two problems, this paper proposes BoDT (Bag of Dynamic Trajectories), a trajectory descriptor, to represent variable-length trajectories and a multi-channel uneven SVM to identify rare action or events. In our approach, trajectories are estimated by tracking densely sampled points in each frame [5] first. Then, different kinds of trajectory features are extracted at each point along a trajectory and the same kind of features are concatenated together. Subsequently, a set of trajectory models are learned as a codebook for each kind of trajectory features by an unsupervised DTW based k-medoids algorithm. As such, a compact representation of video, named BoDT (Bag of Dynamic Trajectories), can be constructed as a bag of feature representation of trajectories on the codebooks.

Such a BoDT descriptor can be used to identify actions or events. If the action or event is non-rare, a multi-channel SVM with χ^2 kernel can achieve good performance. But when the action or event is rare, a multi-channel uneven SVM, an extension of SVM with uneven margins [14], [15], is proposed. Note that the original SVM with uneven margins was proposed to deal with the unbalance data distribution in Chinese document categorization tasks. In this study, we extend it to multi-channels so as to detect rare video actions and events.

In order to evaluate the performance of our method, two kinds of tasks are performed in our experiments, one

for the non-rare action detection task while the other for the rare event detection task. Our objective is to evaluate the detection performance and generality of our method (including the BoDT descriptor and the MU-SVM method) in both non-rare and rare detection tasks. Extensive experiments are performed in six benchmark datasets, including KTH, YouTube, Olympic, MIT, Q-MUL and TRECVid. Experimental results demonstrate that our approach is feasible and effective, and outperforms several state-of-the-art methods.

The rest of the paper is organized as follows: Section 2 describes the proposed approach in detail. Experimental results are reported and analyzed in Section 3. Finally, conclusions are given in Section 4.

II. The Proposed Approach

The dense trajectories feature proposed in [5] has been proved effective in human action recognition on some benchmark datasets, i.e. KTH, YouTube, Hollywood2 and UCF sports. Therefore, our study is also to utilize the dense trajectories feature for action and event detection. First, we extract dense trajectories in 3D video volumes. Then, a set of trajectory models are learned by using dynamic time alignment (DTW) based k-medoids algorithm. Treating these trajectory models as codebooks, the BoDT descriptor can be constructed to represent video sequences. At last, a multi-channel SVM with χ^2 kernel is used for non-rare action detection and a multi-channel uneven SVM is used for rare event detection.

A. Dense Trajectories and Trajectory Features

For each frame of a video, a pyramid is constructed with different scales in each level and feature points are sampled on multiple scale levels. Then, each point $P_t = (x_t, y_t)$ at frame t is tracked to the frame $t+1$ by median filtering in a dense optical flow field $\omega = (u_t, v_t)$.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\overline{x}_t, \overline{y}_t)} \quad (1)$$

where M is the median filtering kernel, and $(\overline{x}_t, \overline{y}_t)$ is the rounded position of (x_t, y_t). So, a series of tracked points form a feature trajectory: $(P_t, P_{t+1}, P_{t+2}, ...)$. In our experiments, we directly use the toolbox of dense trajectories implemented by Wang et al. [5], which is available online [1].

Samilar with what Wang et al. have done, local appearance and motion patterns are encoded by four kinds of features, i.e. shape descriptor, histograms of oriented gradients (HOG) [16], histograms of optical flow (HOF) [17] and motion boundary histogram (MBH) [18] in our approach. We use the default parameters as Wang et al. used, except that we do not average the HOG, HOF and MBH features along each trajectory. That means features of each point will keep their original values in the trajectory feature.

[1] http://lear.inrialpes.fr/software

B. Similarity Estimation Between Trajectories

The DTW algorithm [19] is able to compare signal sequences or trajectories flexibly in applications of speaker identification and handwriting recognition. Inspired by this, we introduce the DTW distance to estimate the similarity between two trajectories.

Given two trajectory features, $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ and $Y = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m)$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ could be any kind of local descriptor (i.e. HoG, HoF and MBH) at point i, the DTW algorithm is to find the optimal warping path which minimizes the accumulated distance between X and Y. The DTW distance can be defined as follows:

$$DTW(X, Y) = \min_{\psi,\theta} \frac{1}{W_{\psi,\theta}} \sum_{i=1}^{N} w(i) d(\mathbf{x}_{\psi(i)}, \mathbf{y}_{\theta(i)}), \quad (2)$$

subject to

$$\psi(i) \leq \psi(i+1) \leq \psi(i) + 1$$
$$\theta(i) \leq \theta(i+1) \leq \theta(i) + 1$$
$$||\psi(i) - \theta(i)|| \leq Q$$

where Q is a locality constant, eithor ψ or θ stands for a warping path, N denotes 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 $d(\mathbf{x}_i, \mathbf{y}_j) = \sqrt{||\mathbf{x}_i - \mathbf{y}_j||^2}$ is the L2 distance between features. Obviously, trajectories are more similar when their DTW distance is lower.

C. Learning Feature Trajectory Models

Based on the DTW distance, k-medoids clustering algorithm is employed to learn the codebook (a set of trajectory models) for each kind of trajectory features. Given a trajectory set $S = \{s_1, s_2, ..., s_n\}$, where s_i is any of trajectory features, k trajectories are selected from S as models and they are considered to describe typical local motion patterns. Let $M = \{m_1, m_2, ..., m_k\}$ be the model set selected from S, and let $C = \{c_1, c_2, ..., c_k\}$ be the cluster set corresponding to M. Here M is generated as follows:

$$\arg\min_{M} \sum_{i=1}^{k} \sum_{s_j \in c_i} DTW(s_j, m_i), \quad (3)$$

where $c_i = \{s_p : DTW(s_p, m_i) \leq DTW(s_p, m_v), \forall 1 \leq v \leq k\}$, DTW is the DTW distance function defined in Eq.(2). This problem can be solved using the Partitioning Around Medoids (PAM) [20] method. In our implementation, the DTW distance between every pair of all trajectories is calculated first, and then used for finding new medoids at every iterative step [21]. This DTW-based k-medoids clustering algorithm is described in Algorithm 1.

Algorithm 1 Dynamic Time Alignment K-medoids Clustering Algorithm

Input: Cluster number k, trajectory set S
Output: Model set M

```
Randomly select k trajectories from S = {s_1, s_2, ..., s_n} as medoids, M = {m_1, m_2, ..., m_k};
while △M < threshold do
    for p = 1 to n do
        if DTW(s_p, m_i) ≤ DTW(s_p, m_v), ∀1 ≤ v ≤ k then
            set s_p ∈ c_i;
        end if
    end for
    for i = 1 to k do
        for s_j ∈ c_i do
            set m_i = s_j;
            cost = Σ_{s_p ∈ c_i} DTW(s_p, m_i);
        end for
        select m_i which minimizes cost;
    end for
end while
```

D. BoDT (Bag of Dynamic Trajectories)

Based on the learned codebook, a bag-of-feature representation, BoDT (Bag of Dynamic Trajectories), can be generated. Given a trajectory set $S_{vid} = \{s_1, s_2, ..., s_n\}$ extracted from video vid, its BoDT (histogram of trajectories) is defined as follows:

$$BoDT_{vid} =< h_1, h_2, ..., h_i, ..., h_k >, \quad h_i = \frac{\sum_{s_p \in S_{vid}} \mu(s_p, M, i)}{n}, \tag{4}$$

where μ is a function defined as:

$$\mu(s_p, M, i) = \begin{cases} 1 & s_p \in c_i \\ 0 & otherwise \end{cases}.$$

E. Multi-channel Uneven SVM

Generally speaking, the BoDT descriptor can be used in any action or event detection task on videos. Nevertheless, different classification methods should be adopted according to the different sample distributions.

For non-rare action and event detection, a non-linear SVM with a χ^2-kernel [22] can be employed. Like [23], different descriptors are combined in a multi-channel approach. The kernel function of multi-channel SVM is defined as

$$K(x_i, x_j) = exp\left(-\sum_c \frac{1}{A^c} D(x_i^c, x_j^c)\right), \tag{5}$$

where $D(x_i^c, x_j^c)$ is the χ^2 distance between x_i and x_j with respect to the c-th channel. A^c is the mean value of χ^2 distances between the training samples for the c-th channel [22]. The optimal problem of the multi-channel SVM is the same form as the standard SVM in Eq. (6) .

$$\min_{w,b,\xi} \frac{1}{2} w \circ w + C \sum_{i=1}^{l} \xi_i, \tag{6}$$

subject to

$$\begin{aligned} w \circ x^{(i)} + \xi_i + b &\geq 1, && \text{if} \quad y_i = +1; \\ w \circ x^{(i)} - \xi_i + b &\leq -1, && \text{if} \quad y_i = -1; \\ \xi_i &\geq 0, && \text{for} \quad i = 1, \ldots, m. \end{aligned}$$

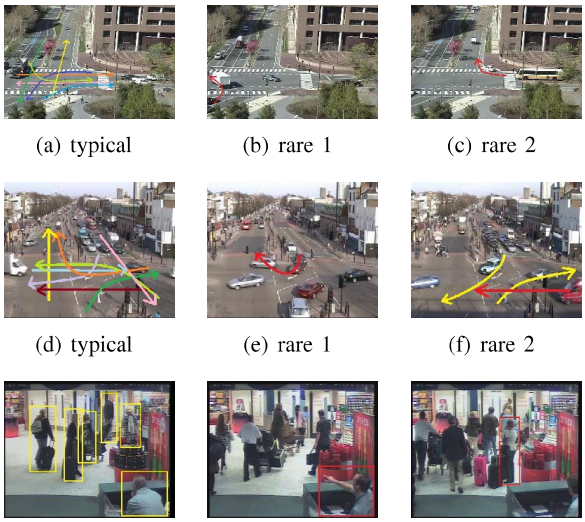
For rare event detection, we use the kernel function defined in Eq. (5), and introduce a margin parameter τ [14] to make the classification boundary properly moved away from the positive training examples. Thus, the optimize problem of the multi-channel uneven SVM (MU-SVM) is defined in Eq. (7), which is similar to Eq. (6) of the original SVM.

$$\min_{w,b,\xi} \frac{1}{2} w \circ w + C_\tau \sum_{i=1}^{l} \xi_i, \tag{7}$$

subject to

$$\begin{aligned} w \circ x^{(i)} + \xi_i + b &\geq 1, && \text{if} \quad y_i = +1; \\ w \circ x^{(i)} - \xi_i + b &\leq -\tau, && \text{if} \quad y_i = -1; \\ \xi_i &\geq 0, && \text{for} \quad i = 1, \ldots, m \end{aligned}$$

where $C_\tau = \frac{1+\tau}{2} C$, τ is the ratio of the negative margin to the positive margin of the classifier. For imbalanced classification tasks, set $0 < \tau < 1$ and the classification hyperplane will be close to the negative margin, thus improving the classification performance towards the minority (positive) samples. In [14], a theorem was



(a) typical (b) rare 1 (c) rare 2

(d) typical (e) rare 1 (f) rare 2

(g) typical (h) rare 1 (i) rare 2

Fig. 4. Typical and rare event examples in the (a)-(c) MIT, (d)-(f) QMUL and (g)-(i) TRECVid datasets.

proved to obtain the solution of the uneven SVM with any margin parameter $0 < \tau < 1$ from its corresponding solution of the standard SVM. Let (w_1^*, b_1^*, ξ_1^*) be the solution of Eq. (6). Then, the solution of Eq. (7), (w_2^*, b_2^*, ξ_2^*) could be obtained as follows [14]:

$$\begin{aligned} w_2^* &= \frac{1+\tau}{2} w_1^* \\ b_2^* &= \frac{1+\tau}{2} b_1^* + \frac{1-\tau}{2} \\ \xi_2^* &= \frac{1+\tau}{2} \xi_1^* \end{aligned} \tag{8}$$

III. Experimental Results

In this section, experimental results are reported and analyzed. Two kinds of tasks are performed in our experiments so as to evaluate the performance of our method, one for the non-rare action detection task while the other for the rare event detection task. Note that here our objective is to evaluate the detection performance and generality of our method (including the BoDT descriptor and the MU-SVM method) in both non-rare and rare detection tasks. It does not mean that our method cannot be used for the non-rare event detection or rare action detection task.

Non-rare Action Detection. For non-rare action detection, the multi-channel SVM with a χ^2-kernel with our BoDT descriptor is used. We evaluate our method on three datasets (i.e. KTH [24], YouTube [25] and Olympic sports [26]). The influence of different parameter settings in our method is also evaluated in the experiment. As what the related studies do, we report average accuracy over all classes of KTH and YouTube, and mean average precision (mAP) over all classes of Olympic sports.

TABLE I
Clip Number of Rare Events Detection Dataset

MIT			
Total	typical	rare 1	rare 2
Train	200	1	1
Test	280	16	18

QMUL			
Total	typical	rare 1	rare 2
Train	100	1	1
Test	198	10	4

TRECVid			
Total	typical	rare 1	rare 2
Train	2000	65	71
Test	122939	401	77

Rare Event Detection. For rare event detection, we evaluate the proposed Multi-channel Uneven SVM (MU-SVM) in three real-world surveillance datasets: MIT dataset (1.5 hours) [13], QMUL dataset (1 hour) [13], and TRECVid dataset (7 hours, randomly selected from the 100 hours TRECVid ED 09 training dataset) [7]. The rare events are as follows: left-turn (rare 1, Fig. 4(b)) and right-turn (rare 2, Fig. 4(c)) in the MIT dataset, U-turn (rare 1, Fig. 4(e)) and near-collision situation (rare 2, Fig. 4(f)) in the QMUL dataset, and Pointing (rare 1, Fig. 4(h)) and CellToEar (rare 2, Fig. 4(i)) in TRECVid dataset. The clip numbers of different events for training and testing are listed in Table I.

For the MIT and QMUL dataset, results are evaluated with classification confusion matrix (i.e., the mean along the diagonal of the normalized confusion matrix). But for the TRECVid dataset, we use Normalized Detection Cost Rate (NDCR) [7] to evaluate the algorithm performance. 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). The smaller the NDCR, the better the performance.

$$NDCR(S,E) = P_{Miss}(S,E) + Beta * R_{FA}(S,E) \tag{9}$$

where S is the evaluated system, E is the interest event and $Beta$ is composed of constant values that define the parameters of the surrogate application.

A. Parameter selection

The trajectory length L is an important factor which impacts the performance greatly. In [5], Wang et al. reports that an increase of L improves the performance up to a certain point (L=15 or 20), and then decreases slightly. We note that while L was set to 10, 15 or 20, dense trajectories give better results in Wang's work. Therefore, the three values are all kept to extract variable-length trajectories in our approach.

An important factor of the BoDT descriptor is the constraint parameter Q of the DTW algorithm. Q adjusts the flexibility of trajectories alignment. Each point (or descriptor) $\mathbf{x_k}$ on trajectory X can be dynamically matched at any point (or descriptor) among $[k-Q, k+Q]$ along trajectory Y by meeting the requirement of distance-minimization. In this experiment, we empirically select the best value of Q in the three action detection datasets, which is then used in both action and event detection tasks. Comparison with different constraint parameter Q values is shown in Fig. 5. The performance climbs while Q increases up to a certain point ($Q = 5$ or 6). The best result is 97.6% on the KTH dataset when $Q = 5$. On the YouTube dataset, we get 86.8% at $Q = 5$. While $Q = 5, 6$, the maximum accuracy 77.7% is acquired on the Olympic sports dataset. Therefore, we use $Q = 5$ in the other experiments.

B. Experimental results on non-rare action detection

The proposed BoDT descriptor could robustly represent trajectories with variable lengths, which is important in action or event detection in video. In this experiment, we first compare the BoDT descriptor with the baseline algorithm, L2 k-means (dense trajectories by k-means with $L2$ norm distance), by using a multi-channel SVM with the χ^2-kernel [17]. The length of trajectories is set to 15. Then the BoDT descriptor is evaluated with variable-length trajectories. On all datasets, about 100K trajectories are randomly selected for training, and the number of clusters is fixed to 2,000 for both BoDT and baseline.

It can be seen from Table II that when using the same kind of trajectories with the fixed length of 15, BoDT outperforms L2 k-means for most kinds of features on the three datasets. Moreover, the performance of BoDT can be further improved when using variable-length trajectories (i.e., with the length of 10, 15 and 20). These results definitely confirm the advantage of the proposed BoDT descriptor.

Compared with the BoDT descriptor, the main disadvantage of the baseline algorithm, L2 k-means, is that its results are not robust when trajectories can not be aligned strictly. That is, if two trajectories are identical but one of them is shifted slightly along the time axis, then they may be judged as two different patterns by using the L2 distance. Instead, the BoDT descriptor can successfully overcome this limitation by making use of the DTW distance to evaluate the similarity between variable length trajectories. As a result, it can produce more robust trajectory representation for video sequences.

In Table III, we also compare BoDT with some state-of-the-art methods on the three datasets. Note that results on the KTH and YouTube datasets are reported in [5] using dense trajectories and L2 k-means method. We can see that, our BoDT shows the superior performances

TABLE II
COMPARISON OF L2 K-MEANS AND BODT FOR DIFFERENT FEATURES ON THE KTH, YOUTUBE AND OLYMPIC DATASETS. WE REPORT AVERAGE ACCURACY OVER ALL CLASSES FOR KTH AND YOUTUBE, AND MEAN AP OVER ALL CLASSES FOR OLYMPIC. (**SL** MEANS SOLO-LENGTH TRAJECTORIES OF 15 FRAMES, AND **ML** MEANS VARIABLE-LENGTH TRAJECTORIES OF 10, 15, AND 20.)

	KTH		
	SL+L2 k-means	SL+BoDT	ML+BoDT
Trajectory	89.4%	91.7%	**92.1%**
HOG	**85.6%**	83.8%	83.8%
HOF	93.5%	93.5%	**95.8%**
MBH	95.4%	96.3%	**96.8%**
Combined	95.8%	96.8%	**97.7%**
	YouTube		
	SL+L2 k-means	SL+BoDT	ML+BoDT
Trajectory	68.1%	70.2%	**72.1%**
HOG	75.4%	75.4%	**75.8%**
HOF	72.3%	72.6%	**74.3%**
MBH	83.9%	84.6%	**85.8%**
Combined	84.5%	85.6%	**86.8%**
	Olympic		
	SL+L2 k-means	SL+BoDT	ML+BoDT
Trajectory	61.8%	62.7%	**62.9%**
HOG	65.4%	65.7%	**66.2%**
HOF	58.3%	58.8%	**59.5%**
MBH	72.6%	72.4%	**73.1%**
Combined	74.4%	75.5%	**77.7%**

TABLE III
COMPARISON RESULTS WITH STATE-OF-THE-ART METHODS.

KTH		YouTube		Olympic	
Laptev [17]	91.8%	Liu [25]	71.2%	Niebles [26]	72.1%
Kovashka [27]	94.5%	Ikizler [28]	75.2%	Zhou [29]	71.0%
Wang [5]	94.2%	Wang [5]	84.2%	Liu [30]	74.4%
Sadanand [31]	**98.2%**	-	-	Brendel [32]	77.3%
Our Method	97.6%	Our Method	**86.8%**	Our Method	**77.7%**

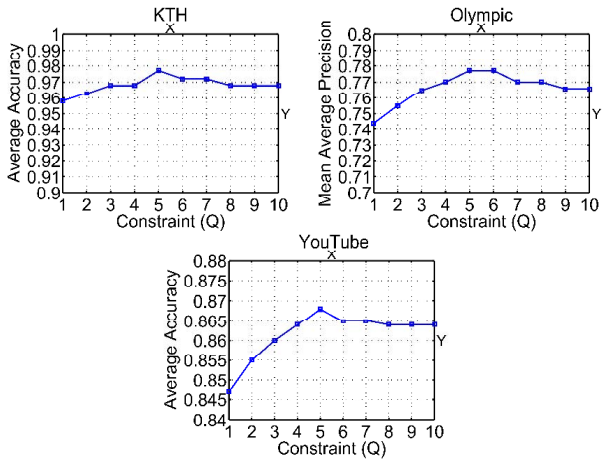


Fig. 5. Results for different constraint parameter values on the KTH, YouTube and Olympic datasets.

over these state-of-the-art methods on the YouTube and Olympic datasets, and only slightly performs worse that Sadanand et al. [31]. These results demonstrate that our BoDT descriptor is feasible and effective on the action detection task.

C. Experimental results on rare event detection

This experiment is to evaluate the performance of our method on the rare event detection task. In this experiment, our method is based on the *multi-channel uneven SVM* (MU-SVM) with the BoDT descriptor. The baseline system is the normal SVM with dense trajectories. A state-of-the-art method, WS-JTM [13], is also involved in this experiment.

The classification confusion matrices on the MIT and QMUL datasets are shown in Table IV and Table V respectively. It can be observed that MU-SVM shows better performance than other two approaches. Not surprisingly, the normal SVM has failed to identify any rare event, and identified almost every rare event as typical. It also should be noted that the performance of MU-SVM is lower than WS-JTM on the QMUL "rare 2" event (i.e., near-collision situation). The possible reason is that the QMUL "rare 2" event (as shown in Fig. 4(f)) typically denotes the situation that at least two trajectories, each representing the movement of a car, tend to collide with each other, while the proposed MU-SVM is only based the features from one trajectory while not specially designed to deal with the events that are related to the interaction between several trajectories.

Results on the TRECVid dataset are listed in Table VI and Table VII, where "Targ" means the number of

TABLE IV
CLASSIFICATION CONFUSION MATRICES AFTER ONE-SHOT LEARNING ON THE MIT DATASET. T: TYPICAL, R1: RARE 1, R2: RARE 2.

	t	r1	r2
t	.987	.002	.011
r1	.063	.937	.00
r2	.333	.00	.667

(a)MU-SVM

	t	r1	r2
t	.89	.07	.04
r1	.19	.81	.00
r2	.39	.00	.61

(b)WS-JTM

	t	r1	r2
t	1.0	.00	.00
r1	1.0	.00	.00
r2	1.0	.00	.00

(c)SVM

TABLE V
CLASSIFICATION CONFUSION MATRICES AFTER ONE-SHOT LEARNING ON THE QMUL DATASET. T: TYPICAL, R1: RARE 1, R2: RARE 2.

	t	r1	r2
t	.992	.003	.005
r1	.10	.90	.00
r2	.50	.00	.50

(a)MU-SVM

	t	r1	r2
t	.59	.26	.15
r1	.10	.90	.00
r2	.00	.33	.67

(b)WS-JTM

	t	r1	r2
t	1.0	.00	.00
r1	1.0	.00	.00
r2	1.0	.00	.00

(c)SVM

interest event instances, "Sys" is the number of system outputs, "CorDet" is the number of correct detected instances, "Miss" is the number of missed instances, and "NDCR" is the normalized detection cost rate. We can find that on the two rare events, the NDCR of MU-SVM is smaller than other two approaches. These results further validate the superior of the proposed MU-SVM and the BoDT descriptor.

IV. CONCLUSION

In this paper, we propose BoDT, a bag of feature trajectory descriptor to represent variable-length trajectories in video, and a multi-channel uneven SVM to detect rare action or event from surveillance video big data. Experimental results show that BoDT outperforms L2

TABLE VI
RESULTS OF "POINTING" EVENT DETECTION ON THE TRECVID DATASET.

Pointing	Targ	Sys	CorDet	FA	Miss	NDCR
SVM	401	492	16	476	385	1.3001
WS-JTM	401	317	34	283	367	1.1174
MU-SVM	401	107	35	72	366	0.9641

TABLE VII
RESULTS OF "CELLTOEAR" EVENT DETECTION ON THE TRECVID.

CellToEar	Targ	Sys	CorDet	FA	Miss	NDCR
SVM	77	125	1	124	76	1.0756
WS-JTM	77	88	4	84	73	1.0081
MU-SVM	77	43	5	38	72	0.9622

k-means, especially when using variable-length trajectories. For the non-rare action detection task, our BoDT shows the superior performances over several state-of-the-art methods on the YouTube and Olympic datasets, and can achieve comparable performance on the KTH dataset; while for the rare event detection task, our MU-SVM, together with the BoDT descriptor, provides excellent detection performance on the MIT, QMUL and TRECVid datasets. These results demonstrate that our approach is feasible and effective.

ACKNOWLEDGMENT

This work is partially supported by grants from the National Basic Research Program of China under grant 2015CB351806, and the National Natural Science Foundation of China under contract No. 61390515and No. 61471042.

REFERENCES

[1] T. Huang, "Surveillance video: the biggest big data," *Computing Now*, vol. 7, no. 2, 2014.

[2] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the Future*, 2012.

[3] N. Moënne-Loccoz, E. Bruno, and S. Marchand-Maillet, "Local feature trajectories for efficient event-based indexing of video sequences," *Image and Video Retrieval*, pp. 82–91, 2006.

[4] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," *Proc. IEEE 12th Int'l Conf. Computer Vision*, pp. 104–111, 2009.

[5] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3169–3176, 2011.

[6] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," *Proc. IEEE Int'l Conf. Computer Vision Workshops*, pp. 514–521, 2009.

[7] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," *Proc. Int'l Workshop on Multimedia Information Retrieval*, pp. 321–330, 2006.

[8] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[9] D. Mease, A. J. Wyner, and A. Buja, "Boosted classification trees and class probability/quantile estimation," *J. Machine Learning Research*, vol. 8, pp. 409–439, 2007.

[10] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," *Proc. Advances in intelligent computing*, pp. 878–887, 2005.

[11] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," *Proc. IEEE World Cong. Computational Intelligence*, pp. 1322–1328, 2008.

[12] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.

[13] T. M. Hospedales, J. Li, S. Gong, and T. Xiang, "Identifying rare and subtle behaviors: A weakly supervised joint topic model," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2451–2464, 2011.

[14] Y. Li and J. Shawe-Taylor, "The svm with uneven margins and chinese document categorization," *Proc. Pacific Asia Conf. Language, Information and Computation*, pp. 216–227, 2003.

[15] W. Ni, J. Xu, Y. Huang, T. Liu, and J. Ge, "Acronym extraction using svm with uneven margins," *Proc. IEEE Symp. Web Society*, pp. 132–138, 2010.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, 2005.

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[18] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *Proc. European Conf. Computer Vision*, pp. 428–441, 2006.

[19] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." *Proc. AAAI Conf. Artificial Intelligence, Workshop on KDD*, vol. 1, pp. 359–370, 1994.

[20] S. Theodoridis and K. Koutroumbas, "Pattern recognition," 2006.

[21] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[22] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int'l J. Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[23] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," *Proc. European Conf. Computer Vision*, pp. 140–153, 2010.

[24] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," *Proc. Int'l Conf. Pattern Recognition*, vol. 3, pp. 32–36, 2004.

[25] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1996–2003, 2009.

[26] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," *Proc. European Conf. Computer Vision*, pp. 392–405, 2010.

[27] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2046–2053, 2010.

[28] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," *Proc. European Conf. Computer Vision*, pp. 494–507, 2010.

[29] Q. Zhou and G. Wang, "Atomic action features: a new feature for action recognition," *Proc. European Conf. Computer Vision*, pp. 291–300, 2012.

[30] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3337–3344, 2011.

[31] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1234–1241, 2012.

[32] W. Brendel and S. Todorovic, "Learning spatiotemporal graphs of human activities," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 778–785, 2011.