Anomaly Detection via Local Coordinate Factorization and Spatio-temporal Pyramid*

Tan Xiao[1] [2], Chao Zhang[1] , Hongbin Zha[1], and Fangyun Wei[1]

[1] Key Laboratory of Machine Perception, Peking University, Beijing, P.R. China
[2] CRSC Communication & Information Corporation, Beijing, P.R. China
pkuxiaotan@pku.edu.cn, [chzhang,zha]@cis.pku.edu.cn, wei494300527@163.com

Abstract. Anomaly detection, which aims to discover anomalous events, defined as having a low likelihood of occurrence, from surveillance videos, has attracted increasing interest and is still a challenge in computer vision community. In this paper, we propose an efficient anomaly detection approach which can perform both real-time and multi-scale detection. Our approach can handle the change of background. Specifically, Local Coordinate Factorization is utilized to tell whether a spatio-temporal video volume (STV) belongs to an anomaly, which can effectively detect spatial, temporal and spatio-temporal anomalies. And we employ Spatio-temporal Pyramid (STP) to capture the spatial and temporal continuity of an anomalous event, enabling our approach to handle multi-scale and complicated events. We also propose an online method to update the local coordinates, which makes our approach self-adaptive to background change which typically occurs in real-world setting. We conduct extensive experiments on several publicly available datasets for anomaly detection, and the results show that our approach can outperform state-of-the-art approaches, which verifies the effectiveness of our approach.

1 Introduction

Recent years, surveillance system has been applied to almost everywhere in a city. However, current systems require human operators to watch a large number of screens[1] showing the content captured by differen cameras. One of the main tasks of human operators is to detect or discover suspicious and unusual individuals or events[2], or anomalies. However, with more cameras in city, more human efforts are required, and it's becoming more difficult for human operators and their performance may degrades significantly [3]. To address this problem, automatic anomaly detection approaches attract increasing interests in recent years. These techniques can automatically analyze video streams to warn, possibly in real-time, the human operators that an anomalous event is taking place.

In computer vision community, anomaly detection is defined as discovering events with low likelihood of occurrence. Recent works can be summarized

* This research was supported by National Key Basic Research Project of China (973 Program) 2011CB302400 and National Nature Science Foundation of China (NSFC Grant No.61071156 and 61131003).

into three categories based on how they construct their models: supervised [4–8], semi-supervised [9, 10] and unsupervised [11–17]. Actually, considering that anomalies are always rare and they can be quite different from each other with unpredictable variations, recent works [16, 17] concentrate more on unsupervised scenarios. Furthermore, since it's almost impossible to define all the anomalous events in advance, unsupervised approaches are more practical.

Several unsupervised approaches have been proposed. Trajectories based approaches [18–21] aim to track motions of objects and persons by their spatial location. But these methods only consider spatial deviations, thus abnormal appearance or motion of a target following a "normal" track is not detected. And they are difficult to cope with crowd scenes where precise segmentation of a target is nearly impossible. Optical flow has also been used to model typical motion patterns [11, 13, 12]. However, these methods perform unreliably in crowded scenes, as mentioned in [18]. Furthermore, two kinds of approaches above mainly focus on the motion of objects, i.e., they only considers anomalous motion while ignoring anomalous appearance. Instead, [16] and [17] propose to use densely sampled local spatio-temporal descriptor which represents both motion and appearance and possesses some degree of robustness to unimportant variations in data. A non-parametric statistic model is utilized in [16] to measure the degree of anomaly. And [17] proposes to organize spatio-temporal video volumes into large contextual graphs and decompose spatio-temporal contextual information into unique spatial and temporal contexts. Both methods achieve promising results for real-time anomaly detection on several publicly available datasets.

An effective real-time anomaly detection approach should have the following properties. 1) It should be unsupervised because it's almost impossible to define all anomalous events in advance and it's burdensome for human operators to do so. 2) It can detect both spatial and temporal anomalies. 3) It can detect multi-scale events. Actually, it's also hard to know in advance the range of an anomaly, e.g. how large the abnormal object is, how fast the abnormal object moves, or how long the abnormal event lasts. 4) It should be self-adaptive to scene change, both in appearance and motion, which has also emphasized in [16] and [17]. In fact, the appearance background is always changing in surveillance videos because of the lighting condition, weather, etc. 5) Of course, it should be able to effectively and efficiently detect the anomalies from surveillance videos.

In this paper, we propose a novel approach for anomaly detection in surveillance videos. Densely sampled spatio-temporal video volumes (STVs) with pixel-by-pixel analysis is utilized as the foundation of our approach. Specifically, each STV is represented by a local spatio-temporal descriptor which can capture both motion and appearance characteristic of STV. Motivated by the extensive study in employing STVs in the context of bag-of-video-words (BoVW), we propose to use Local Coordinate Factorization [22] to tell whether a STV belongs to an anomalous event. The local coordinates are updated continuously with coming surveillance videos, thus it requires no offline or supervised pre-training. This unsupervised method enables our approach to detect anomaly which hasn't been observed before. Furthermore, the updating procedure also ensures that our ap-

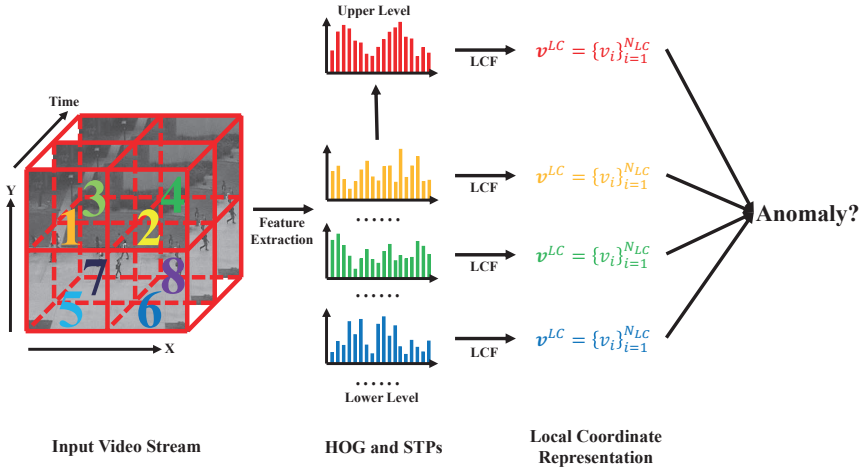

Fig. 1. Overview of our approach. This is an example of two-level spatio-temporal pyramid. The input is a video stream. Then a 3-D volume around a pixel is constructed represented by the outer red cube. Then it's segmented into 8 ($2 \times 2 \times 2$) smaller cubes denoted by different numbers in this figure. The smaller cubes form the lower but finer level of the pyramid. HOG features are extracted for each smaller cube. And the HOG features of upper level cube can be constructed efficiently from lower level cubes. Next, we apply Local Coordinate Factorization to each cube (both lower and upper level) to generate their local coordinate representation v. Finally, the anomaly judgement is given based on the combination of local coordinate representation of different levels.

proach can cope with the scene change both in appearance and motion. To detect multi-scale and complicated anomalies, we propose a Spatio-temporal Pyramid (STP), which is the temporal extension of spatial pyramid [23]. STP can describe videos by STV of different scales to detect multi-scale events. We also observe that an event is always associated with several STVs which are different in location or time or both. STP can be used to discover the relationship of different STVs associated to one event, which enables our approach to detect complicated events. Furthermore, upper level representation of STP can be easily constructed from lower level representation, which guarantees the efficiency.

The overview of our approach is summarized in Fig. 1. Given a video stream, initially it's densely sampled, i.e., sampled pixel by pixel, and a 3-D volume is constructed around a pixel. Then this volume is segmented without overlap into 8 smaller STVs. The large STV forms the upper and coarser level of STP, which can capture the overall information of an event. And the small STVs form the lower and finer level of STP, which can describe an event in detail. We can also segment any small STV into another smaller 8 STVs . But we find that a two-level STP is enough for anomaly detection. Then HOG features which can represent both motion and appearance are extracted for each STV. Interestingly, we find that the HOG of upper level STV can be easily constructed from lower level STVs, implying that the pyramid doesn't require too much extra computation. So it can be quite efficient which is essential for real-time

detection. Local Coordinate Factorization is then applied to the HOG features to obtain the local coordinate representation v for each STV. Furthermore, the local coordinates are updated automatically to adapt adapting themselves to scene changes. Finally, the anomaly judgement is given based on results of both levels.

The main contributions of this paper can be summarized as follows. 1) We propose a novel anomaly detection approach based on densely sampled STVs. Local Coordinate Factorization is applied to HOG features of STV to effectively judge whether this STV belongs to an anomalous event. 2) We propose to use Spatio-temporal Pyramid (STP) to capture the spatial and temporal continuity of an anomalous event. STP can also enable our approach to handle multi-scale and complicated events. 3) We propose an online method to continuously update the local coordinates so our method can adaptively learns the event patterns in the scene and thus can cope with scene changes. 4) We conduct extensive experiments on several public datasets to evaluate our approach for anomaly detection. The results show that our approach can significantly outperform several state-of-the-art approaches, which verifies the effectiveness of our approach.

2 Related Work

As mentioned above, trajectory analysis of objects are widely utilized in previous works. However, they require precise tracking methods [24, 25]. Unfortunately, tracking objects is time-consuming, especially in crowded scene where a lot of objects (or persons) are moving so that precise segmentation of targets which is the foundation of tracking is nearly impossible. Optical flow is also used in several works [11, 13, 12] but they also perform unreliably in crowded scenes [16].

Recent years, approaches not requiring object detection or tracking, focusing on local spatio-temporal features are proposed and have received increasing attention [26, 27]. These approaches describe the local characteristic at each pixel by low-level visual features such as color, texture and motion. Then a pixel-level background model and behavior template can be constructed [28–31]. Moreover, spatio-temporal video volumes in the context of bag-of-video-words are becoming popular [16, 12, 32, 33]. By ignoring the order of local features, probabilistic topic models like LDA [34] can be directly applied to analysis videos [35, 36]. But these methods often ignore the spatio-temporal relationship between STVs which is essential for scene understanding and event detection [37, 38]. Some works have made efforts to incorporate either spatial or temporal compositions of STVs into the probabilistic topic model. But they are highly time-consuming and computationally expensive, thus they can't be applied to online and real-time tasks [39]. Furthermore, some approaches [26, 29, 40, 41] propose to construct spatio-temporal behavior model and low-level local anomalous events can be detected by analyzing the spatio-temporal pattern of each pixel as a function of time. However, such as in [40], they independently process each pixel but ignore the relationships between pixel in space and time, thus leading to too local detection.

In [16], Bertini *et al.* propose a multi-scale and non-parametric approach to perform real-time anomaly detection and localization. To capture both appearance and motion of objects in the scene, dense local spatio-temporal features are extracted at each pixel. And they propose to use "overlapping" features to consider the relationship between pixels. Though they achieve promising results, their approach also face challenge of efficiency to achieve accurate multi-scale detection. In fact, our Spatio-temporal Pyramid is partially motivated by their "overlapping" features. But our STP can be constructed more efficiently and can achieve much better performance. And our STP can naturally cope with multi-scale detection while their approach actually treat different scales independently.

In addition, our Local Coordinate Factorization is similar to the Sparse Reconstruction method proposed in [15]. However, their reconstruction problem is formulated as an L_1-norm regularized least squares problem which can't be solved quite efficiently, thus their method can't be applied to real-time detection. But Local Coordinate Factorization can be solved by just few simple linear matrix operations which can be highly efficient for real-time detection.

3 Local Coordinate Factorization

3.1 Spatio-temporal features

Firstly, we need to describe a two-level STV (we can use any-level STV, but we find two level is enough) centered at pixel (x, y, t) by meaningful spatio-temporal features. Given a STV $v \in \mathbb{R}^{n_x \times n_y \times n_t}$ with the size $n_x \times n_y \times n_t$, where $n_x \times n_y$ is the size of spatial window and n_t is the depth of STV in time. In this paper, we find $10 \times 10 \times 10$ is a good choice. Then we calculate the histogram of the spatio-temporal gradient of the video in polar coordinates to describe the STV [16, 42, 17]. Denote the spatial gradients as $G_x(x, y, t)$, $G_y(x, y, t)$, and the temporal gradient as $G_t(x, y, t)$ respectively at pixel (x, y, t). To eliminate the effect of local texture and contrast, the spatial gradient is normalized as:

$$G_s(x, y, t) = \frac{\sqrt{G_x^2(x, y, t) + G_y^2(x, y, t)}}{\sum_{x', y', t' \in v} \sqrt{G_x^2(x', y', t') + G_y^2(x', y', t')} + \epsilon} \tag{1}$$

where $G_s(x, y, t)$ is the normalized spatial gradient and ϵ is a constant to avoid numeric instabilities. In this paper, we set $\epsilon = 0.01$. Then we can construct 3D normalized gradient represented in polar coordinates as below,

$$M_{3D}(x, y, t) = \sqrt{G_s^2(x, y, t) + G_t^2(x, y, t)} \tag{2}$$

$$\theta(x, y, t) = \tan^{-1}\left(\frac{G_y(x, y, t)}{G_x(x, y, t)}\right) \tag{3}$$

$$\phi(x, y, t) = \tan^{-1}\left(\frac{G_t(x, y, t)}{G_s(x, y, t)}\right) \tag{4}$$

where $M_{3D}(x, y, t)$ is the magnitude of 3D normalized gradient, and $\phi(x, y, t) \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\theta(x, y, t) \in [-\pi, \pi]$ are the orientations of the gradient respectively. Now for a given STV v, we can construct a histogram of oriented gradients (HOG) by quantizing each pixel in v into $n_\phi + n_\theta$ bins by their 3D normalized gradient. In this paper, we set $n_\phi = 8$ and $n_\theta = 16$. So the HOG features for STV v, denoted as h, has 24 dimension in this paper. From the feature extraction procedure, we can see that this HOG feature can capture the characteristics of both motion and appearance in the video so that we can detect both anomalous actions and objects. Furthermore, it's also robust to unimportant variations in the data such as texture and contrast. Though it's quite simple, it shows promising performance. Moreover, we need to mention that, as a histogram feature, each element in h is non-negative. This is an essential property for the Local Coordinate Factorization, which requires non-negative input.

We can also notice that it can be efficient to calculate this HOG feature. In fact, the gradient and 3D normalized gradient for all pixels can be computed in advance. And the histogram of pixel (x, y, t), denoted as $h(x, y, t)$ can also be precomputed by quantization. Then the histogram of a STV v around pixel (x, y, t) can be computed by simply sum up all the histogram in this STV as

$$h_v(x, y, t) = \sum_{(x',y',t')\in v} h(x', y', t') \tag{5}$$

And computing the HOG of a STV can use its neighbor's HOG which has been computed to save more computations. Denote the STV around (x, y, t) and $(x + 1, y, t)$ as v_1 and v_2 respectively. Then we have,

$$h_{v_2} = h_{v_1} - \sum_{(x',y',t')\in v_1\setminus v_2} h(x', y', t') + \sum_{(x',y',t')\in v_2\setminus v_1} h(x', y', t') \tag{6}$$

where "\" is the set minus operation. It's clear that $v_1 \setminus v_2$ is much smaller than v_1. Furthermore, the HOG feature of upper level STV can be computed by summing up the HOG features of lower level STVs in the same STP. Consider the outer red cube in Figure 1. Given the HOG of eight lower level STVs in this cube, the HOG of the upper level, i.e., the red cube, can be computed as follow,

$$h_{v_{\text{up}}} = \sum_{(x',y',t')\in v_{\text{up}}} h(x', y', t') = \sum_{i=1}^{8} \sum_{(x',y',t')\in v_i} h(x', y', t') = \sum_{i=1}^{8} h_{v_i} \tag{7}$$

The highly efficiency of spatio-temporal feature extraction, which is guaranteed by computation tricks above, is one of the requirements for real-time detection.

3.2 Local Coordinate Selection

When statio-temporal features are extracted for STV, we can perform Local Coordinate Factorization to tell whether this STV belongs to an anomalous event. But we need to construct local coordinates first. Actually, the local coordinates for our approach can be regarded as video words for BoVW, i.e., they

Algorithm 1 Local Coordinate Selection

Input: $\mathbf{H}$,$\lambda = 1$, $\mathbf{S}_0$ K, c
Output: $\mathbf{S}$;

Initialize $\mathbf{Z}_0 = \mathbf{S}_0$,$a_0 = 1$
for $k = 0, 1, 2, ..., K$ **do**
  $\mathbf{S}_{k+1} = \arg\min_{\mathbf{S}} : p_{\mathbf{Z}_k,L}(\mathbf{S}) = D_{\frac{\lambda}{L}}(\mathbf{Z}_k - \frac{1}{L}\nabla f(\mathbf{Z}_k))$
  while $f_0(\mathbf{S}_{k+1}) > p_{\mathbf{Z}_k,L}(\mathbf{S}_{k+1})$ **do**
    $L = L/c$
    $\mathbf{S}_{k+1} = \arg\min_{\mathbf{S}} : p_{\mathbf{Z}_k,L}(\mathbf{S}) = D_{\frac{\lambda}{L}}(\mathbf{Z}_k - \frac{1}{L}\nabla f(\mathbf{Z}_k))$
  end while
  $a_{k+1} = (1 + \sqrt{1 + 4a_k^2})/2$
  $\mathbf{Z}_{k+1} = (\frac{a_{k+1}+a_k-1}{a_{k+1}})\mathbf{S}_{k+1} - (\frac{a_k-1}{a_{k+1}})\mathbf{S}_k$
end for

are some points in the feature space. Analogous to video words, local coordinates can be generated by cluster the spatio-temporal features. The obtained cluster centroids are local coordinates. However, there are some parameters for clustering algorithm, for example, we need to specify k for kmeans clustering. And we find our approach is a little sensitive to this parameter.

Instead, we propose to construct local coordinates from data. Given a set of spatio-temporal features $\mathbf{H} = [h_1, ..., h_n] \in \mathbb{R}^{d\times n}$, where $d = 24$ is the dimension of feature, and n is the size of feature set. Actually we don't need a training set. This initial feature set $\mathbf{H}$ can be constructed by using the first one or two seconds of the video. Moreover, we can also randomly select some features to reduce n so that our selection algorithm is computationally feasible. In this paper, we tune n from $10,000$ to $20,000$ based on the resolution of videos. Then we need to select some features from $\mathbf{H}$ as the local coordinates. In our method, the number of local coordinates is determined automatically by the algorithm, which is more adaptive to the test data. Following the idea in [15], we'd like to select an optimal subset of $\mathbf{H}$ as local coordinate set, such that the rest of features can be well reconstructed from it. We can formulate this criterion as follows,

$$min_{\mathbf{S}} = \frac{1}{2}\|\mathbf{H} - \mathbf{HS}\|_F^2 + \lambda\|\mathbf{S}\|_{2,1} \tag{8}$$

where $\mathbf{S} \in \mathbb{R}^{n\times n}$ is the selection matrix, $\|\mathbf{S}\|_F = \sqrt{\sum_i \sum_j \mathbf{S}_{ij}^2}$ is the Frobenius norm of $\mathbf{S}$, $\|\mathbf{S}\|_{2,1} = \sum_{i=1}^n \|\mathbf{S}_{i.}\|_2$ is the $L_{2,1}$-norm, and λ is the model parameter and we set $\lambda = 1$ in this paper. Finally, by selecting features with $\|\mathbf{S}_{i.}\| > 0$, we can obtain the local coordinates. To solve this problem, we follow the method proposed in [43]. Consider an objective function $f_0(x) = f(x) + g(x)$ where $f(x)$ is convex and smooth and $g(x)$ is convex but non-smooth. The key step is to construct $p_{Z,L}(x) = f(Z) + \langle\nabla f(Z), x - Z\rangle + \frac{L}{2}\|x - Z\|_F^2 + g(Z)$ to approximate $f_0(x)$ at point Z. Obviously, we can define $f(\mathbf{S}) = \frac{1}{2}\|\mathbf{H} - \mathbf{HS}\|_F^2$ and $g(\mathbf{S}) = \|\mathbf{S}\|_{2,1}$. So we can construct $p_{\mathbf{Z},L}(\mathbf{S})$ as

$$p_{\mathbf{Z},L}(\mathbf{S}) = f(\mathbf{Z}) + \langle\nabla f(\mathbf{Z}), \mathbf{S} - \mathbf{Z}\rangle + \frac{L}{2}\|\mathbf{S} - \mathbf{Z}\|_F^2 + g(\mathbf{Z}) \tag{9}$$

And we can define another function $D_\tau(.) : \mathbf{M} \in \mathbb{R}^{n\times n} \mapsto \mathbf{N} \in \mathbb{R}^{n\times n}$

$$\mathbf{N}_{i.} = \begin{cases} 0, & \|\mathbf{M}_{i.}\| \leq \tau \\ (1 - \tau/\|\mathbf{M}_{i.}\|)\mathbf{M}_{i.}, & \text{otherwise} \end{cases} \tag{10}$$

Because of the limit of space, we can't show all the details to solve this problem. Instead we summarize the algorithm in Alg. 1.

3.3 Local Coordinate Factorization

As mentioned in [15] and [16], a normal STV may be close to a cluster while an anomalous STV may be an outlier. We also observe that a normal STV is always close to some local coordinates. Denote the local coordinates obtained above as $\mathbf{U} = [u_1, u_2, ..., u_m]$, where m is the number of local coordinates. The local coordinate representation $v \in \mathbb{R}^{m\times 1}$ of a STV represented by HOG feature $h \in \mathbb{R}^{d\times 1}$ can be calculated by minimizing the following objective function,

$$\mathcal{O}_h = \|h - \mathbf{U}v\|_F^2 + \mu \sum_{i=1}^{m} |v_i| \|u_i - h\|_F^2, \quad \text{s.t.} \quad v_i \geq 0, \forall i \tag{11}$$

where μ is a model parameter and we set $\mu = 10$ in this paper. The first term in Eq. (11) aims to reconstruct h by local coordinates. The second term requires that the local coordinates selected to reconstruct h should be close to h to preserve data locality, which is motivated by [44]. Furthermore, this term also leads to sparse v, i.e., h is reconstructed just by very few local coordinates. This is important because we have $m > d$. Without it, any h can be perfectly reconstructed because $\mathbf{U}$ is over-complete in feature space. Motivated by resent study in Non-negative Matrix Factorization [45–47], we constraint that v should be non-negative, thus h is reconstructed by addition but not substraction of $\mathbf{U}$, which will lead to better performance.

Actually, when reconstructed by just few local coordinates, i.e., v is quite sparse, the normal STV can be well reconstructed with less reconstruction error, while the reconstruction error for an anomalous STV will be quite large because it's always outliers so that it's far from all local coordinates. Moreover, the local coordinate representation for normal STV is also different from representation for anomalous STV. Generally, we observe that the length of v is close to 1 for most normal STV but far from 1 for anomalous STV. This is also reasonable because normal STV is always close to some local coordinates. To incorporate both observations above for anomaly detection, we compute the degree of anomaly as

$$d_{\mathrm{a}} = \|h - \mathbf{U}v\|_F^2 + \gamma|1 - \|v\|_F^2| \tag{12}$$

where γ is to balance the magnitude of both terms. Then d_{a} is utilized to determine whether a STV is anomalous based on its value compared to a threshold δ. STVs with d_{a} larger than δ are determined to be anomalous. δ is determined by the anomaly probability p_a depending on the user's need. A large p_a will lead to high true positive rate and high false positive rate while a small one will

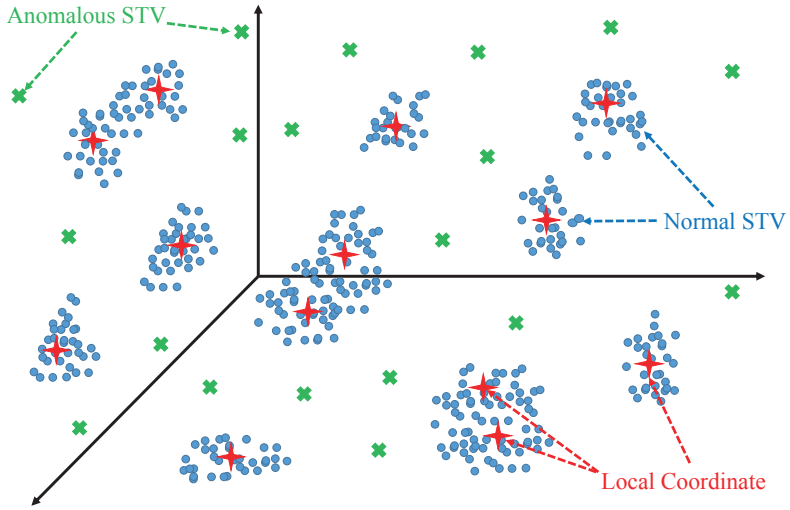

Fig. 2. Local Coordinate Factorization. There are a lot of local coordinates in feature space. Based on the construction method of local coordinates, normal STV can be well reconstructed by just few local coordinates, e.g., just one or two. However, there is large reconstruction error for anomalous STV. Furthermore, the local coordinate representations of normal STV and anomalous STV are also quite different.

lower both. Empirically, p_a can be set to 10^{-2},10^{-3},10^{-4}.... In this paper we set $p_a = 10^{-3}$. Then, as the local coordinate selection procedure, we can compute d_a for all STVs in the first one or two seconds in a test video, and set δ such that the ratio of STVs whose d_a are larger than δ is about p_a. So about p_a STVs will be treated as anomalies. We have a postprocessing step on the initial judgement to obtain better results, which will be introduced latter.

Now we will show how to solve Eq. (11) to obtain v for a STV represented by h. Following some simple algebraic steps, we can rewrite Eq. (11) as follows,

$$\mathcal{O}_h = \|h - \mathbf{U}v\|_F^2 + \mu \sum_{i=1}^{m} |v_i| \|u_i - h\|_F^2 = \|h - \mathbf{U}v\|_F^2 + \mu \|(h\mathbf{1}^T - \mathbf{U})\Lambda^{\frac{1}{2}}\|_F^2 \quad (13)$$

where $\Lambda = diag(v_1, ..., v_m) \in \mathbb{R}^{m \times m}$. Noticing that $\|A\|_F^2 = tr(AA^T)$, we have

$$\mathcal{O}_h = tr(hh^T + \mathbf{U}vv^T\mathbf{U}^T - 2hv^T\mathbf{U}^T + \mu(h\mathbf{1}^T\Lambda\mathbf{1}^Th^T - 2h\mathbf{1}^T\Lambda\mathbf{U}^T + \mathbf{U}\Lambda\mathbf{U}^T)) \quad (14)$$

Now let ϕ_i be the Lagrange multiplier for constraints $v_i \geq 0$, and define $\Phi = [\phi_i]$, then the Lagrange $\mathcal{L}$ is

$$\mathcal{L} = \mathcal{O}_h + tr(\Phi v^T) \quad (15)$$

Then we can calculate the partial derivative of $\mathcal{L}$ with respect to v as follows,

$$\frac{\partial \mathcal{L}}{\partial v} = 2\mathbf{U}^T\mathbf{U}v - 2\mathbf{U}^Th + \mu(\mathbf{C} - 2\mathbf{U}^Th + \mathbf{D}) + \Phi \quad (16)$$

where column vector $\mathbf{C} = [h^T h, ..., h^T h] \in \mathbb{R}^{m\times 1}$ and $\mathbf{D} = diag(\mathbf{U}^T\mathbf{U}) \in \mathbb{R}^{m\times 1}$. Using the KKT conditions $\phi_i v_i = 0$, we obtain the following equation,

$$2(\mathbf{U}^T\mathbf{U}v)_i v_i - 2(\mathbf{U}^T h)_i v_i + \mu(\mathbf{C} - 2\mathbf{U}^T h + \mathbf{D})_i v_i = 0 \tag{17}$$

Then the equation above lead to the following update rule for v

$$v_i \leftarrow v_i \frac{2(\mu+1)(\mathbf{U}^T h)_i}{(2\mathbf{U}^T\mathbf{U}v + \mu\mathbf{C} + \mu\mathbf{D})_i} \tag{18}$$

The update rules in Eq. (18) is guaranteed to converge and the final solution will be a local optimum, and similar proof of convergency based on [46] and [48] can be found in [47].

So to solve Eq. (11), we can randomly initialize v by some non-negative values and use Eq. (18) iteratively. Eq. (11) can converge in tens of iterations but we find that 5 to 10 iterations are enough to get satisfactory performance while guaranteing the efficiency, thus we set the maximum number of iterations to 5. Compared to [15] who proposes to solve a L_1-norm regularized least squares problem which can't be solved efficiently, our method just needs some simple matrix operations and few iterations which is quite efficient. Thus our approach can perform real-time detection while [15] can't.

3.4 Online Update

As discussed in [15–17], an anomaly detection system should be adaptive to the background change, both in appearance and motion. So we propose an update strategy to tune the local coordinates to capture the change in background. The basic idea is straightforward, i.e., the normal features can be well reconstructed by local coordinates, or the local coordinates should be as close as possible to normal features. Since the distribution of normal features is changing slightly all the time, we need to update the local coordinates simultaneously. Given a set of n normal features $\mathbf{H} = [h_1, ..., h_n] \in \mathbb{R}^{d\times n}$ and their local coordinate representations $\mathbf{V} = [v_1, ..., v_n] \in \mathbb{R}^{m\times n}$, the local coordinates $\mathbf{U}$ is updated by minimizing the following objective function,

$$\mathcal{O}_U = \|\mathbf{H} - \mathbf{U}\mathbf{V}\|_F^2 + \mu \sum_{i=1}^{n}\sum_{j=1}^{m} |v_{ji}| \|u_j - h_i\|_F^2 \quad \text{s.t.} \quad u_{ij} \geq 0, \forall i, j \tag{19}$$

Analogous to our strategy in Eq. (13)(14)(15), let $\Lambda_i = diag(v_{i1}, ..., v_{im})\mathbb{R}^{m\times m}$, ψ_{ij} be the Lagrange multiplier for constraints $u_{ij} \geq 0$, $\Psi = [\psi_{ij}]$, and the Lagrange $\mathcal{L}_U = \mathcal{O}_U + tr(\Psi\mathbf{U}^T)$. The partial derivatives of $\mathcal{L}_U$ to $\mathbf{U}$ is

$$\frac{\partial \mathcal{L}_U}{\partial \mathbf{U}} = 2\mathbf{U}\mathbf{V}\mathbf{V}^T - 2\mathbf{H}\mathbf{V}^T + \mu \sum_{i=1}^{n}(-2h_i\mathbf{1}^T\Lambda_i + 2\mathbf{U}\Lambda_i) + \Psi \tag{20}$$

and by using the KKT conditions $\psi_{ij}u_{ij} = 0$, we can get the update rule for $\mathbf{U}$ like Eq. (17)(18),

$$u_{ij} \leftarrow u_{ij} \frac{(\mathbf{H}\mathbf{V}^T + \mu\sum_{i=1}^{n} h_i\mathbf{1}^T\Lambda_i)_{ij}}{(\mathbf{U}\mathbf{V}\mathbf{V}^T + \mu\sum_{i=1}^{n}\mathbf{U}\Lambda_i)_{ij}} \tag{21}$$

Though the background may sharply change in long time, the change during a short time (e.g., two seconds) is always slight. So we just need to update $\mathbf{U}$ every two seconds. Further we can also do random sample among normal features to obtain a small n ($10,000$ to $20,000$ based on the resolution of videos) to guarantee real-time updating. Moreover, we can use the origin $\mathbf{U}$ as the input to Eq. (19) and execute Eq. (21) just once such that the updated $\mathbf{U}$ is different but close to origin one. As a result, the local coordinates can change all the time to adapt to the background change but the change in one updating step isn't too sharp. In addition, as our feature can capture both spatial and temporal properties of videos, the update strategy proposed here can be adaptive to background change in both appearance and motion.

4 Spatio-temporal Pyramid

In fact, there is noise in the video and our Local Coordinate Factorization method may sometimes give wrong judgement which may lower the true positive rate and lift the false positive rate. But we can observe that an anomalous event shows continuity in space and time, i.e., it's associated to a relatively large region and it lasts for a period of time. Thus considering the relationship of STVs in space and time can promote the detection performance. In this paper, we propose to use Spatio-temporal Pyramid as illustrated in Fig. 1. Specifically, we use two-level pyramid and we find that this setting can achieve satisfactory result.

As discussed in Section 3.1, the HOG feature of upper level STV can be constructed efficiently from lower level STVs. The upper level can capture the relationship of lower level STVs in space and time and global information of an event. And because STVs in different levels have different scales, the Spatio-temporal Pyramid can be utilized to detect multi-scale events. Given a STV in any scale (either upper level of lower level), we can tell whether it belongs to an anomalous event by Local Coordinate Factorization individually.

In our experiments, we find an interesting phenomenon. The judgement on upper level STV tends to have high precision but low recall, i.e., our approach can claim that a upper level STV is anomalous with high confidence but it may miss some anomalous STV. We think the reason is that the upper level STV can capture the global information of an event which fully considers the continuity of an event in space and time while it may ignore some important local details. On the contrary, the judgement on lower level STV tends to have high recall because it can capture the local details of anomalous event but low precision since it's too sensitive to local details and noise and ignores the relationship between STVs. So we propose to combine these two levels as Spatio-temporal Pyramid to consider both local details and global information as follows.

Firstly, when a upper STV is judged to be normal, it actually may be anomalous. So we should consider the results of 1) its six neighbors and 2) its lower STVs. In this paper, we consider an upper STV to be anomalous if 1) it's judged to be anomalous, 2) three or more of its neighbors are anomalous, and 3) five or more of its lower level STVs are anomalous. The first criterion is based on the

high-precision result for upper level STV. The second criterion is based on the continuity of events. The third criterion is based on a voting scheme, because it's reasonable to assume that though one lower STV may be influenced by noise or local details, it's difficult for most STVs to generate wrong judgement.

Secondly, as a lower level STV can't capture the continuity of events, so the judgement of a STV should be incorporated with its upper level STV and neighbors. So a lower level STV is considered to be anomalous if it's judged to be anomalous and 1) two of more of its neighbors are anomalous or 2) its upper level STV is anomalous.

Based on the Spatio-temporal Pyramid and criteria above, we take into consideration the continuity of events, the relationship between STVs in space and time, and the local details simultaneously which can promote the performance significantly. Furthermore, the Spatio-temporal Pyramid allow us to perform multi-scale detection.

5 Experiments

To validate the effectiveness of the proposed approach, we test it in the following two public datasets for anomaly detection: anomaly behavior detection dataset [49][3] and UCSD pedestrian dataset [14][4]. The evaluation and comparison of different approaches are presented in precision-recall, ROC curves and EER at both frame level and pixel level. As mentioned before, we use a two-level pyramid, and the the size of lower level STV is $10 \times 10 \times 10$. To extract HOG features, we set $n_\phi = 8$ and $n_\theta = 16$. We set $\lambda = 1$ for local coordinate selection in Eq. (8), $\mu = 10$ for Local Coordinate Factorization in Eq. (11) and online update in Eq. (19), and $p_a = 10^{-3}$. In fact, our method doesn't need training procedure. It just use the first two seconds of a test video to select initial local coordinates. Furthermore, we set that the local coordinates are updated every two seconds. We compare our approach to several state-of-the-art approaches for anomaly detection: Optical Flow [11], MDT [14], Sparse Reconstruction (Cong *et al.*) [15], spatio-temporal oriented energies [49], Dominant Behavior (Roshtkhari *et al.*) [17], Saligrama *et al.* [50], Reddy *et al.* [51] and Bertini *et al.* [16].

The first dataset is *Belleview*. It's a traffic scene where the lighting conditions changes during the day gradually. Cars running from top to bottom is normal event, while cars entering or exiting from the intersection from left or right and people in the lane is the anomalous event. The second is *Boat-river* dataset. The anomalous event is a boat that passing the scene. The third is *Train* dataset where anomalies are moving people. The results on three datasets above, including the anomalous regions detected by our approach (highlighted in red) and the precision-recall curves of different approaches, are shown in Fig. 3, Fig. 4 and Fig. 5 respectively. We can observe that our approach is superior to state-of-the-art methods, e.g., Zaharescu *et al.* and Roshtkhari *et al.*. Three main reasons are: 1) our approach can update model timely so that it's quite

[3] http://www.cse.yorku.ca/vision/research/
[4] http://www.svcl.ucsd.edu/projects/anomaly

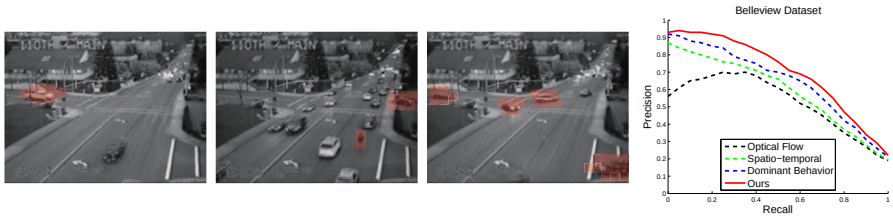

Fig. 3. Experiments on Belleview Dataset

Fig. 4. Experiments on Boat-Holborn Dataset

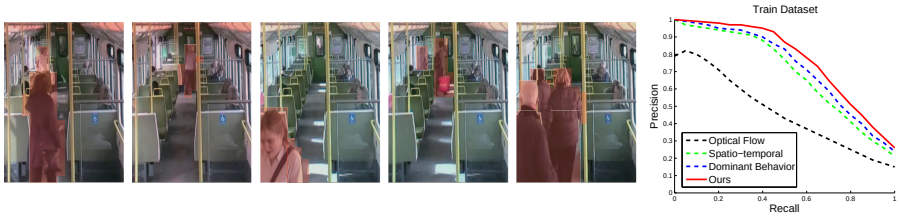

Fig. 5. Experiments on Train Dataset

robust to drastic background change, 2) our approach takes a full consideration of the relationship between neighbor STVs thus it's robust to local noise, and 3) our approach considers the continuity of anomalous event in space and time.

In the UCSD datasets (Ped1 and Ped2), the anomalies are non-pedestrian entities (e.g., cyclist, skaters, small carts) and pedestrians moving in anomalous motion. We follow the evaluation utilized in [14] and [16]. In the frame level, an anomalous frame is considered correctly detected if at least one pixel is detected as anomalous. In the pixel level, an anomalous frame is considered correctly detected only if at least 40% of the anomalous pixels are detected correctly. The anomalous regions detected and the ROC curves of other approaches for Ped1 and Ped2 datasets are shown in Fig. 6 and Fig. 7 respectively. And the Equal Error Rate (EER) for both frame level and pixel level detection of different approaches is shown in Table 1. The results show that our approach can outperform all other state-of-the-art approaches, both at frame level and pixel level. And we need to highlight that our approach is unsupervised which doesn't require any training data, and can perform real-time detection because it's quite efficient.

Table 1. Comparison of the proposed approach and the state-of-the-art for anomaly detection using Ped datasets. Approaches with * can perform real-time detection.

	Ped1		Ped2	
	EER (frame)	EER (pixel)	EER (frame)	EER (pixel)
Optical Flow* [11]	38%	76%	42%	80%
Saligrama *et al.* [50]	16%	-	19%	-
MDT [14]	25%	58%	25%	55%
Cong *et al.* [15]	19%	-	20%	-
Reddy *et al.** [51]	22.5%	32%	21%	31%
Bertini *et al.** [16]	31%	70%	30%	68%
Roshtkhari *et al.** [17]	15%	29%	17%	30%
Ours*	**12**%	**25**%	**13**%	**26**%

Fig. 6. Experiments on Ped1 Dataset

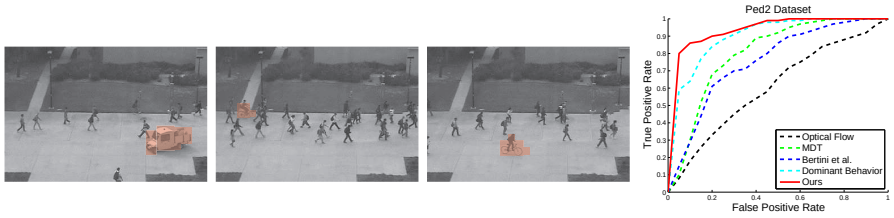

Fig. 7. Experiments on Ped2 Dataset

6 Conclusions

In this paper, we propose a novel approach to perform real-time and multi-scale anomaly detection. Specifically, we use spatio-temporal features to capture the characteristics of STV in both appearance and motion in order that our approach can detect spatial, temporal, and spatio-temporal anomalies. Then we utilize Local Coordinate Factorization to efficiently tell whether a SVT belongs to an anomaly. Then to consider the relationship between STVs, and the continuity of an event in space and time, we propose to use Spatio-temporal Pyramid, which can further support multi-scale detection. We also propose an efficient online method to update local coordinates such that our approach is self-adaptive to background change. Finally, we conduct extensive experiments on several public datasets for anomaly detection and compare our approach to state-of-the-art approaches. The results show that it achieve superior performance at both frame and pixel level and our approach can outperform state-of-the-art approaches.

References

1. Troscianko, T., Holmes, A., Stillman, J., Mirmehdi, M., Wright, D., Wilson, A.: What happens next? the predictability of natural behaviour viewed through cctv cameras. In: Perception. (2004) 87–101
2. Keval, H., Sasse, M.: not the usual suspects: a study of factors reducing the effectiveness of cctv. In: Secur. J. (2010) 134–154
3. Haering, N., Venetianer, P., Lipton, A.: The evolution of video surveillance: an overview. In: Mach. Vis. Appl. (2008) 279–290
4. Brax, C., Niklasson, L., Smedberg, M.: Finding behavioural anomalies in public areas using video surveillance data. In: Proc. of 11th International Conference on Information Fusion. (2008)
5. Ivanov, I., Dufaux, F., Ha, T., Ebrahimi, T.: Towards generic detection of unusual events in video surveillance. In: Proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance. (2009)
6. Li, J., Gong, S., Xiang, T.: Discovering multi-camera behaviour correlations for on-the-fly global activity prediction and anomaly detection. In: Proc. of IEEE International Conference on Computer Vision Workshops. (2009)
7. Liu, C., Wang, G., Ning, W., Lin, X., Li, L., Liu, Z.: Anomaly detection in surveillance video using motion direction statistics. In: Proc. of IEEE International Conference on Image Processing. (2010)
8. Loy, C., Xiang, T., Gong, S.: Detecting and discriminating behavioural anomalies. In: Pattern Recogn. (2011)
9. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I.: Semi-supervised adapted hmms for unusual event detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2005)
10. R. Sillito, R.F.: Semi-supervised learning for anomalous trajectory detection. In: Proc. of British Machine Vision Conference. (2008)
11. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. In: IEEE Trans. Pattern Anal. Mach. Intell. (2008)
12. Kim, J., Grauman, K.: Observe locally, infer globally: a spacectime mrf for detecting abnormal activities with incremental updates. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2009)
13. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2009)
14. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2010)
15. Cong, Y., Yuan, J., , Liu, J.: Sparse reconstruction cost for abnormal event detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2011)
16. Bertini, M., Bimbo, A.D., , Seidenari, L.: Multi-scale and realtime non-parametric approach for anomaly detection and localization. In: CVIU. (2012)
17. Roshtkhari, M.J., Levine, M.D.: Online dominant and anomalous behavior detection in videos. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2012)
18. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2009)

19. Jiang, F., Wu, Y., Katsaggelos, A.: A dynamic hierarchical clustering method for trajectory-based unusual video event detection. In: IEEE Trans. Image Process. (TIP). (2009)
20. Khalid, S.: Activity classification and anomaly detection using m-medoids based modelling of motion patterns. In: Pattern Recogn. (2010)
21. Jiang, F., Yuan, J., Tsaftaris, S., Katsaggelos, A.: Activity classification and anomaly detection using m-medoids based modelling of motion patterns. In: CVIU. (2011)
22. Yu, K., Zhang, T., Gong, Y.: Nonlinear learning using local coordinate coding. In: Advances in Neural Information Processing Systems. (2009)
23. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2006)
24. Morris, B.T., Trivedi, M.M.: Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. In: IEEE Trans. Pattern Anal. Mach. Intell. (2011)
25. Ouivirach, K., Gharti, S., Dailey, M.N.: Incremental behavior modeling and suspicious activity detection. In: Pattern Recognition. (2013)
26. Benezeth, Y., Jodoin, P.M., Saligrama, V.: Abnormality detection using low-level co-occurring events. In: Pattern Recogn. Lett. (2011)
27. Hospedales, T., Gong, S., Xiang, T.: Video behaviour mining using a dynamic topic model. In: Int. J. Comput. Vision. (2012)
28. Benezeth, Y., Jodoin, P.M., Saligrama, V., , Rosenberger, C.: Abnormal events detection based on spatio-temporal co-occurences. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2009)
29. Ermis, E.B., Saligrama, V., Jodoin, P.M., , Konrad, J.: Motion segmentation and abnormal behavior detection via behavior clustering. In: Proc. of IEEE International Conference on Image Processing. (2008)
30. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Real-time foreground-background segmentation using codebook model. In: Real-Time Imaging. (2005)
31. Mittal, A., Monnet, A., , Paragios, N.: Scene modeling and change detection in dynamic scenes: A subspace approach. In: CVIU. (2009)
32. Boiman, O., Irani, M.: Detecting irregularities in images and in video. In: Int. J. Comput. Vision. (2007)
33. Zhu, X., Liu, Z.: Human behavior clustering for anomaly detection. In: Frontiers of Computer Science in China. (2011)
34. Blei, D., NG, A., Jordan, M.: Latent dirichlet allocation. In: Journal of Machine Learning Research. (2003)
35. Hospedales, T.M., Jian, L., Shaogang, G., Tao, X.: Identifying rare and subtle behaviors: A weakly supervised joint topic model. In: IEEE Trans. Pattern Anal. Mach. Intell. (2011)
36. J. Li, S.G., Xiang, T.: Learning behavioural context. In: Int. J. Comput. Vision. (2012)
37. Ricci, E., Zen, G., Sebe, N., , Messelodi, S.: A prototype learning framework using emd: Application to complex scenes analysis. In: IEEE Trans. Pattern Anal. Mach. Intell. (2012)
38. Roshtkhari, M.J., Levine, M.D.: A multi-scale hierarchical codebook method for human action recognition in videos using a single example. In: In Conf. Computer and Robot Vision. (2012)

39. Hospedales, T.M., Jian, L., Shaogang, G., Tao, X.: Identifying rare and subtle behaviors: A weakly supervised joint topic model. In: IEEE Trans. Pattern Anal. Mach. Intell. (2012)
40. P. Jodoin, J.K., Saligrama, V.: Modeling background activity for behavior subtraction. In: In Int. Conf. Distributed Smart Cameras. (2008)
41. Jodoin, P., Saligrama, V., , Konrad, J.: Behavior subtraction. In: IEEE Trans. Image Process. (TIP). (2012)
42. Scovanner, P., Ali, S., , Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: In International conference on Multimedia. (2007)
43. Nesterov, Y.: Gradient methods for minimizing composite objective function. In: CORE. (2007)
44. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2010)
45. Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. In: Nature. (1999)
46. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. (2001)
47. Cai, D., He, X., Han, J., , Huang, T.S.: Graph regularized nonnegative matrix factorization for data representation. In: IEEE Trans. Pattern Anal. Mach. Intell. (2011)
48. Lin, C.J.: On the convergence of multiplicative update algorithms for non-negative matrix factorization. In: IEEE Transactions on Neural Networks. (2007)
49. Zaharescu, A., Wildes, R.: Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In: ECCV. (2010)
50. Saligrama, V., Zhu, C.: Video anomaly detection based on local statistical aggregates. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (2012)
51. Reddy, V., Sanderson, C., Lovell, B.C.: Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2011)